Classification Automatique d'Images, Application à l'Imagerie du Poumon Profond

Chesner Desir

Laboratoire d'Informatique,
de Traitement de l'Information et des Systèmes
Université de Rouen
École Doctorale SPMII

THÈSE EN VUE DE L'OBTENTION DU DIPLÔME DE DOCTEUR DE L'UNIVERSITÉ DE ROUEN

Discipline : Informatique

présentée par

Chesner DESIR

Classification Automatique d'Images, Application à l'Imagerie du Poumon Profond

---

dirigée par Laurent HEUTTE

Soutenue le 10 juillet 2013 devant le jury composé de :

Mme Isabelle BLOCH	Télécom ParisTech	Présidente
M. Frédéric JURIE	Université de Caen	Rapporteur
M. Lionel PREVOST	Université Antilles-Guyane	Rapporteur
Mme Caroline PETITJEAN	Université de Rouen	Encadrante
M. Luc THIBERVILLE	Université de Rouen	Encadrant
M. Laurent HEUTTE	Université de Rouen	Directeur

Á Kaaj, ma famille.

“The world that we have made as a result of the level of thinking we have done thus far creates problems that we cannot solve at the same level as the level we created them at.”
Albert Einstein [a]

”I am enough of the artist to draw freely upon my imagination. Imagination is more important than knowledge. Knowledge is limited. Imagination encircle the world.”
Albert Einstein [b]

---

a. The Journal of Transpersonal Psychology Transpersonal Institute, 1969, 1-4, pp 124

b. The Saturday Evening Post, What Life Means to Einstein : An Interview by George Sylvester Viereck, 1929 October 26

Remerciements

Je remercie l'ADIR [1] et la Ligue contre le Cancer [2] pour leur soutien qui a permis la réalisation de ces travaux de thèse. Je remercie les membres du jury Mme Isabelle Bloch, M. Lionel Prevost, M. Frédéric Jurie pour avoir accepté d'évaluer mes travaux de thèse. Je remercie les rapporteurs MM. Lionel Prevost et Frédéric Jurie pour l'intérêt accordé à mes travaux, leurs points de vue et leurs remarques constructives dans les rapports de pré-soutenance.

Je tiens à remercier chaleureusement mon directeur de thèse, le Prof. Laurent Heutte qui a été à mon écoute, m'a accompagné, guidé durant toutes ces étapes, a su comprendre mes attentes, mon regard différent sur les études que nous avons menées. Merci pour toutes ces discussions constructives, rigoureuses, empreintes d'expériences et de savoir-faire sur des problématiques qui me tenaient à cœur et qui m'ont aidé à mieux comprendre et à avancer, merci de m'avoir fait confiance et permis d'explorer dans une grande autonomie des approches nouvelles. Je tiens à remercier chaleureusement également mon encadrante Mme Caroline PetitJean qui a toujours été présente, disponible pour répondre à mes questions, partager sa vision et ses expériences. Merci pour le temps précieux consacré à consolider mon esprit de synthèse, à souligner l'importance de la simplicité, du "straight-forward" [3], sans me ménager. Ce fut de longue haleine, sans baisser les bras mais avec du succès !
Merci pour tous ces échanges qui ont contribué à l'expérience que j'ai acquise durant toutes ces années de thèse. Merci à tous deux pour votre présence et votre chaleur humaine, notamment dans les moments personnels difficiles qui ont émaillé ces années de thèse où vous avez été à mon écoute et su vous rendre pleinement disponible.
Je remercie le Prof. Luc Thiberville pour son écoute, son regard pertinent et plein d'expériences sur la problématique médicale que j'avais à traiter, sa disponibilité pour répondre à mes questions, sa curiosité scientifique pour les méthodes que nous mettions en place. Merci pour toutes les démarches notamment administratives auprès de multiples organismes qui ont permis de mener à bien mes travaux de thèse. Je remercie Mathieu Salaün pour s'être rendu disponible au début de la thèse, durant ses propres travaux en interne avec le Prof. Luc Thiberville et m'avoir aidé dans l'obtention notamment de différentes données médicales nécessaires pour nos expérimentations.

Je remercie les membres du laboratoire pour leur accueil, leur chaleur humaine tout le long de ces années, ponctuées de "Alors comment ça avance ?" "On a connu ça aussi, ça va !", suivies de pauses café, de discussions amicales, de retours et partages d'expériences tant sur des sujets scientifiques, que des sujets philosophiques ou d'actualités. Merci pour les croissants de chaque anniversaire et tous les bons moments passés à la Kafèt' ! Merci Max (Maxime Bérar), un puits de savoir, de culture, musicien, danseur, collectionneur de Comics aussi :-), volontaire, toujours à l'écoute, merci pour les innombrables conseils, les partages d'expériences riches et les réflexions pleines de bon sens, critiques, constructives, toujours ouvert aux approches originales ou différentes ! Flo (Florian Yger), merci pour ton dynamisme, ton humour, ton "open-minded" (un terme de BCI-Spirit, nan ? ;-)), ta curiosité scientifique, ta pédagogie, ton écoute également, tes conseils et ton soutien sans faille dans les moments difficiles ; Kamel Ait-Mohand pour sa combativité, sa force de caractère, les expériences partagées, la rigueur de l'argumentation, son pragmatisme et son calme légendaire ! Tommy (Thomas Palfray), L'Ex (Alexandre Burnett) pour les discussions passionnantes, les résolutions de bugs, les conseils ; Vlad (Vladislavs Dovgalecs) avec qui on a refait le monde, souvent "from scratch", fait du "Think different", de l'"Extreme Learning" en acculant les algos, du "One more thing ..." et du Raspberry-3.141592653589793 ! Shishi (Simon Thomas), passionné de sport, "l'imbattable", avec qui j'ai partagé les derniers moments ...de rédaction, dans la bonne humeur ; Dada (David Hebert) mélomane, technophile, Android fanboy (à nous deux on a fait GS et 1 et 2 et 3 !) avec qui j'ai partagé passions pour ce système ouvert, problèmes de codes, idées innovantes et stress de rédaction ! Simon (Simon Bernard) qui a toujours montré de l'intérêt pour mon travail et avec qui les réunions ont été passionnantes et enrichissantes ! Merci Selma Belgacem pour sa force, son courage, son ouverture d'esprit, les discussions passionnantes souvent à l'encadrement même des portes de nos bureaux ! Merci Fabienne Bocquet pour sa bonne humeur, son dynamisme, son attention aux détails et ces innombrables démarches administratives assurant que tout se déroule bien du début jusqu'aux dernières semaines de ma thèse. Merci à Laurence Savouray pour sa joie, sa voix sans pareil et dont le couloir des doctorants se souvient encore, son humour, son accueil chaleureux (au passage, tu avais dit que j'allais oublier de t'envoyer une invitation pour la soutenance ;-)) ! Fabrice Hertel, passionné de technologies, volontaire toujours disponible et arrangeant pour les tracasseries matérielles ! Arnaud Citerin pour sa disponibilité, ses réponses rapides aux multiples problèmes de réseaux, d'identifiants, de mails, d'imprimantes et de config serveurs ! Bira (Ubiratan S. Freitas), pour d'innombrables discussions passionnantes sur l'innovation, la physique au quotidien, les nouvelles technologies, l'Open Source, la science-fiction, un Géo Trouve-tout, un puits de savoir ! Yousri Kessentini, pour son écoute, ses conseils, son attention ; quelle ne fut ma surprise de découvrir qu'on était ...voisins de palier ! Maryvonne Holzem pour ses combats, sa passion, sa culture, son ouverture d'esprit, son écoute, ses conseils.

Je remercie de nouveau tous les membres du labo pour leur bienveillance et leur accueil.

Je remercie mes amies et amis qui m'ont apporté leur soutien et leurs encouragements jusqu'aux derniers moments ! Kaaj, merci de m'avoir supporté durant toutes ces longues années et soutenu dans les moments difficiles.

1. Association d'assistance à Domicile aux Insuffisants Respiratoires : www.adir-assistance.com
2. www.ligue-cancer.net
3. C'est à présent un leitmotiv !

Résumé

Cette thèse porte sur la classification automatique d'images, appliquée aux images acquises par alvéoscopie, une nouvelle technique d'imagerie du poumon profond. L'objectif est la conception et le développement d'un système d'aide au diagnostic permettant d'aider le praticien à analyser ces images jamais vues auparavant. Nous avons élaboré, au travers de deux contributions, des méthodes performantes, génériques et robustes permettant de classer de façon satisfaisante les images de patients sains et pathologiques. Nous avons proposé un premier système complet de classification basé à la fois sur une caractérisation locale et riche du contenu des images, une approche de classification par méthodes d'ensemble d'arbres aléatoires et un mécanisme de pilotage du rejet de décision, fournissant à l'expert médical un moyen de renforcer la fiabilité du système. Face à la complexité des images alvéoscopiques et la difficulté de caractériser les cas pathologiques, contrairement aux cas sains, nous nous sommes orientés vers la classification one-class qui permet d'apprendre à partir des seules données des cas sains. Nous avons alors proposé une approche one-class tirant partie des mécanismes de combinaison et d'injection d'aléatoire des méthodes d'ensemble d'arbres de décision pour répondre aux difficultés rencontrées dans les approches standards, notamment la malédiction de la dimension. Les résultats obtenus montrent que notre méthode est performante, robuste à la dimension, compétitive et même meilleure comparée aux méthodes de l'état de l'art sur une grande variété de bases publiques. Elle s'est notamment avérée pertinente pour notre problématique médicale.

Mots-clefs : Alvéoscopie ; aide au diagnostic médical ; classification automatique ; extraction de caractéristiques ; méthodes d'ensemble ; arbre de décision ; injection d'aléatoire ; forêts aléatoires ; one-class ; out-of-class ; synthèse de données ; malédiction de la dimension

Abstract

This thesis deals with automated image classification, applied to images acquired with alveoscopy, a new imaging technique of the distal lung. The aim is to propose and develop a computer aided-diagnosis system, so as to help the clinician analyze these images never seen before. Our contributions lie in the development of effective, robust and generic methods to classify images of healthy and pathological patients. Our first classification system is based on a rich and local characterization of the images, an ensemble of random trees approach for classification and a rejection mechanism, providing the medical expert with tools to enhance the reliability of the system. Due to the complexity of alveoscopy images and to the lack of expertize on the pathological cases (unlike healthy cases), we adopt the one-class learning paradigm which allows to learn a classifier from healthy data only. We propose a one-class approach taking advantage of combining and randomization mechanisms of ensemble methods to respond to common issues such as the curse of dimensionality. Our method is shown to be effective, robust to the dimension, competitive and even better than state-of-the-art methods on various public datasets. It has proved to be particularly relevant to our medical problem.

Keywords : Alveoscopy ; computer aided-diagnosis ; automatic classification ; feature extraction ; ensemble methods ; decision tree ; randomization ; random forests ; one-class ; out-of-class ; data synthesis ; curse of dimensionality

Table des matières

Introduction générale

Le travail présenté dans cette thèse est le fruit de la collaboration entre le Service de Pneumologie du CHU de Rouen et l'équipe Quantif avec le Pr. Luc Thiberville (co-encadrant de la thèse) et l'équipe "Document et Apprentissage" du laboratoire LITIS EA4108 avec le Pr. Laurent Heutte (directeur de thèse) et Caroline Petitjean (co-encadrante de la thèse). Les travaux menés ont été soutenus par l'association médicale ADIR, la Ligue contre le Cancer et l'Université de Rouen. Le sujet de la thèse est la classification automatique d'images appliquée à l'imagerie du poumon profond. Il s'agit d'élaborer un système permettant de classer des images issues d'une nouvelle technique d'imagerie du poumon, aidant ainsi le praticien dans l'établissement de son diagnostic.

Les poumons constituent l'organe respiratoire principal et se composent de deux régions anatomiques et fonctionnelles : les voies respiratoires (la trachée, les bronches, les bronchioles) et la zone d'échange gazeux composée de sacs alvéolaires. Ces derniers ont pour fonction de permettre les échanges gazeux avec le sang. Si les voies de conduction aérienne sont bien connues, à l'inverse, le poumon profond, qui comprend le système alvéolaire, était jusqu'à peu inaccessible à l'analyse morphologique in-vivo. Le Pr. Luc Thiberville a développé en collaboration avec l'entreprise Maunakea Technologies une technique d'imagerie médicale non invasive basée sur l'application de la Microscopie Confocale Fibrée en Fluorescence (MCFF) pour l'exploration in-vivo des territoires alvéolaires pulmonaires, appelée alvéoscopie [4] [Thiberville et al., 2007a]. Le système d'acquisition, appelé *Cell-Vizio-LUNG*®, est basé sur le principe de la microscopie confocale, où l'objectif du microscope est remplacé par une mini-sonde flexible (de diamètre 1 mm) faite de milliers de microfibres optiques compactées les unes contre les autres (cf. Fig. 4) pouvant être introduites dans le canal fonctionnel d'un bronchoscope flexible. L'alvéoscopie permet d'obtenir une imagerie micro-structurale tridimensionnelle in vivo, en temps réel, des structures pulmonaires les plus profondes de l'arbre bronchique. Elle produit notamment des images de fluorophores endogènes avec une résolution latérale de 5μm et un champ de vision de 600 μm en diamètre, pour une résolution pixélique de 1 μm (cf Figure 1). Il est ainsi possible d'explorer la structure en élastine des poches alvéolaires et la structure interstitielle du poumon profond. Si le diagnostic des pathologies du poumon profond nécessitaient jusqu'à présent une intervention chirurgicale invasive via la biopsie, cette nouvelle technique de MCFF permettrait de dépister plus précocement et de manière non-invasive les pathologies respiratoires.

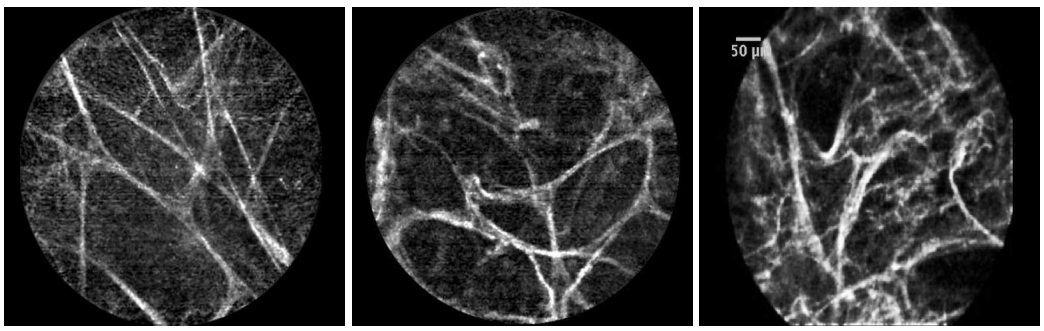
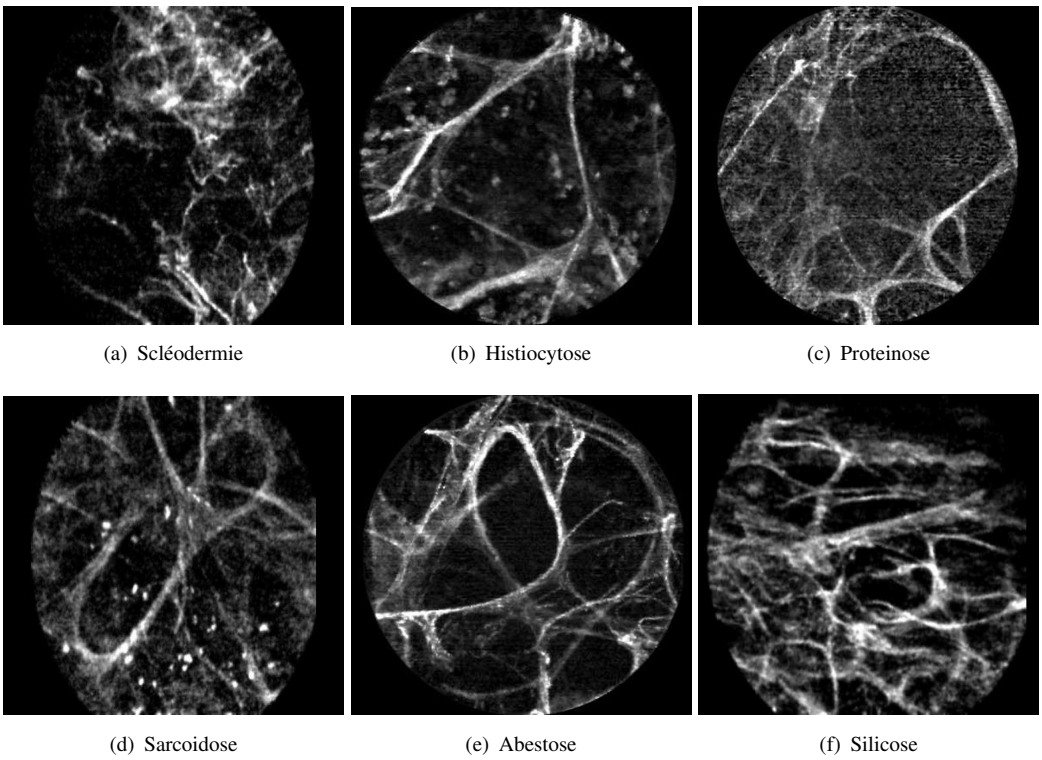


FIGURE 1 – Images alvéoscopiques d'un patient sain (à gauche) et de patients atteints de pathologies pulmonaires (à droite) ; le champ de vision est indiqué sur la dernière image avec l'échelle de 50 μm (en haut à gauche).

Sur les images alvéoscopiques, la structure des alvéoles apparaît sous forme d'un réseau linéique. Ce réseau peut être altéré par des pathologies pulmonaires comme la protéinose, la fibrose, la silicose ou la sclérodermie et apparaît sous une forme plus diffuse (cf. Figure 2). Notons que les images de patients fumeurs présentent des caractéristiques différentes de celles de patients non-fumeurs : les goudrons de tabac sont piégés dans les parois alvéolaires et les macrophages [5],

4. Référence du brevet entre Maunakea Technologies et l'Université de Rouen PCT/FR2007/001371
5. (Cellules ayant pour fonction de nettoyer les débris des tissus cellulaires.)

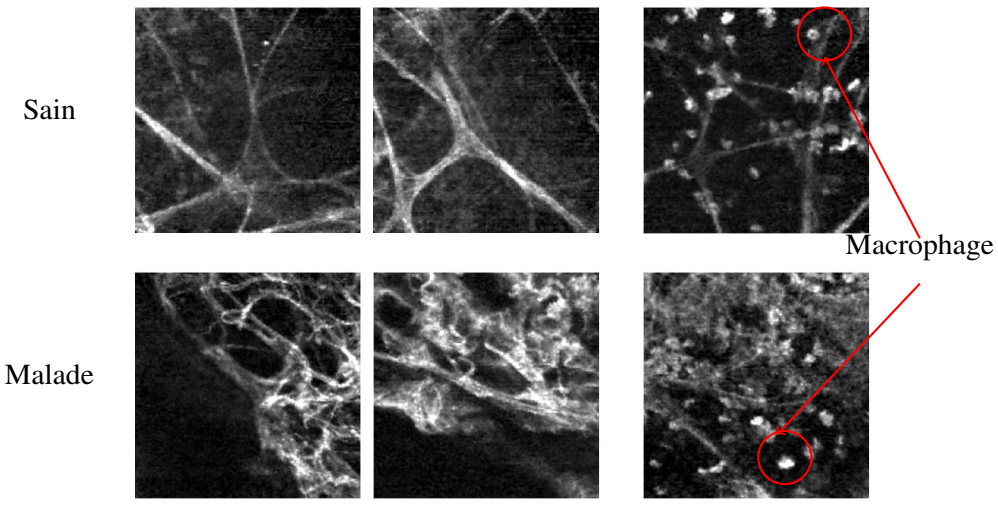
les rendant ainsi visibles. Les parois deviennent ainsi légèrement opaques et des tâches blanches (les macrophages) apparaissent dans l'image (cf. Figure 3). Devant la grande différence entre images de sujets fumeurs et non-fumeurs, les expériences ont été menées sur deux groupes séparés. Il est à noter que les images sont initialement circulaires (e.g. Figure 2) et que pour les besoins de l'étude nous avons extrait les fenêtres carrées inscrites (e.g. Figure 3) comme indiqué par les figures ci-après.



(a) Scléodermie (b) Histiocytose (c) Proteinose

(d) Sarcoidose (e) Abestose (f) Silicose

FIGURE 2 – Images alvéoscopiques associées à différentes pathologies pulmonaires



(a) Patient non-fumeur (b) Patient fumeur

FIGURE 3 – Images alvéoscopiques (MCFF) de patients non-fumeurs (a) et de patients fumeurs (b) sains et malades (images de la seconde ligne).

Les images dont nous disposons sont issues d'un essai clinique [6] en cours pour le diagnostic de sujets volontaires fumeurs et non fumeurs présentant des risques. Les images MCFF de patients atteints de différentes affections pulmonaires sont ainsi collectées pour notre analyse. Les premières images que nous avons eues à traiter ont été sélectionnées par deux experts médicaux et étiquetées "sain" si le patient n'a pas de signes pathologiques particuliers ou "pathologique" si (1) le patient présente une pathologie respiratoire diagnostiquée par ailleurs et (2) si les images ont été extraites à partir d'un segment du poumon présentant une anomalie au scanner thoracique.

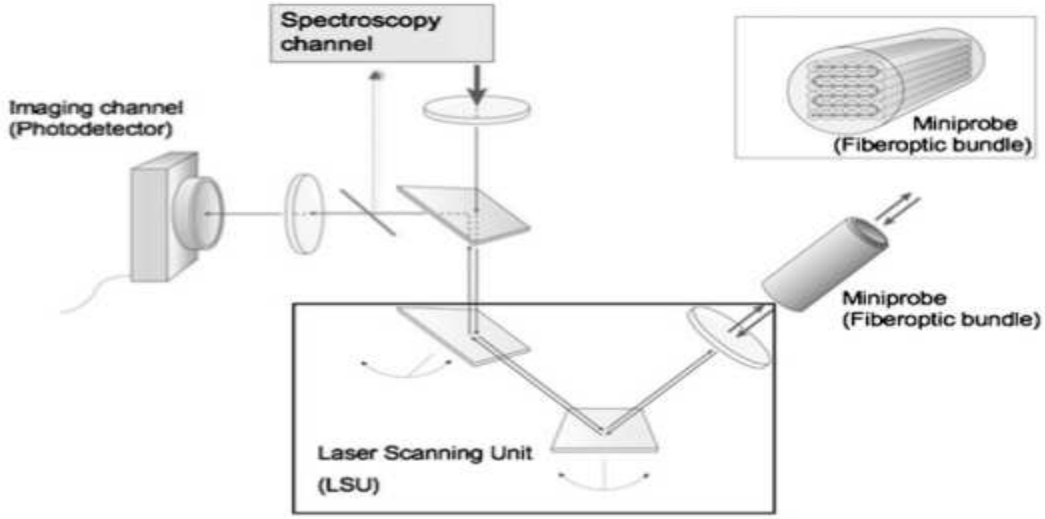


FIGURE 4 – Diagramme schématique du prototype F400/S (Mauna Kea Technologies) pour l'imagerie confocale fibrée [Thiberville et al., 2007b]

Ces images alvéoscopiques n'avaient jamais été visualisées auparavant et sont complexes à analyser et à interpréter par le praticien car il n'existe pas de description univoque ni de sémiologie claire pour ce type d'images. Ainsi il se révèle important de mettre en place une aide au diagnostic médical fournissant des outils quantitatifs pour évaluer l'état pathologique des images à traiter.

L'objectif principal de la thèse est ainsi de proposer un système d'aide au diagnostic médical dans le cadre de l'imagerie alvéoscopique. Plus précisément, il s'agit de mettre en place un système de classification automatique capable de différencier les images de patients pathologiques d'images de patients témoins considérés comme sains. Dans le cadre de ces travaux, nos contributions se situent à différents niveaux de la conception d'un système complet de classification.
Dans la première partie de la thèse, nous avons proposé un premier système pour la classification des images alvéoscopiques. Nous avons tout d'abord procédé à la recherche de descripteurs adaptés aux images alvéoscopiques. À la suite de l'évaluation de plusieurs jeux de caractéristiques composés de descripteurs de bas-niveau et de descripteurs de plus haut niveau à la fois dans une approche de description locale et globale de l'image, nous montrons que les descripteurs de plus haut niveau sont mieux adaptés à la description des images. En nous inspirant d'une approche originale d'extraction dense et locale de fenêtres de tailles aléatoires et à des positions aléatoires dans l'image proposée par Raphael Marée [7], nous avons proposé une description locale et riche, basée sur une approche multi-résolution de description de motifs texturés de l'image. Cette description dense de l'image est associée à une méthode d'ensemble d'arbres de décision de la famille des forêts aléatoires. Ce système générique s'est révélé performant sur les images alvéoscopiques.
Les images pathologiques demeurant difficiles à identifier en raison notamment de leur trop grande disparité dans notre espace de description, nous avons cherché une solution de minimisation du

6. Id :NCT00377338
7. Raphaël Marée, Classification automatique d'images par arbres de décision, PhD thesis from University of Liège - Electrical Engineering and Computer Science - February 2005

taux de non détection des cas pathologiques. Nous avons alors proposé un mécanisme permettant de quantifier le degré de confiance associée à une décision du système afin de piloter le rejet de décision. En effet, la décision de la forêt étant basée sur un consensus au sein des arbres la composant, les cas présentant un degré important d'ambiguïté obtiennent un degré de confiance faible qui peut être utilisé pour rejeter. Nous avons associé à ce mécanisme de pilotage du rejet une approche d'élagage des arbres avec une modification du mode de vote interne des arbres constituant la forêt. Nous avons montré que cette association permet de réduire davantage le taux de non détection. Cette contribution a notamment l'avantage de fournir à l'expert un moyen de renforcer le degré de fiabilité du système.

Dans la seconde partie de la thèse, nous nous sommes orientés vers l'approche one-class afin de répondre à une difficulté rencontrée par les experts médicaux : celle de définir l'état pathologique des images alvéoscopiques. Les images des cas sains étant les seules identifiées de manière certaine par le corps médical, l'approche one-class, qui permet d'apprendre avec les données d'une seule classe, a alors paru comme une solution naturelle à notre problème.
Les approches standards de la littérature utilisent naturellement des estimateurs de densité pour lesquels un seuil d'acceptation est défini manuellement ou par des procédés de validation mais ne traitent que très rarement de la problématique de la gestion des données outliers. D'ailleurs, une catégorisation largement adoptée des méthodes one-class existantes ne mentionne ni la gestion des données outliers, ni les apports possibles des méthodes d'ensembles. Cette catégorisation distingue ainsi les estimateurs de densité, les méthodes de reconstruction (un modèle de compression/restitution des données d'entrées) et les méthodes dites "frontière" cherchant à tracer une frontière automatiquement autour des données d'apprentissage. Cette catégorisation avait été proposée dans le cadre de la description de données ("data description") s'inscrivant davantage dans un cadre modélisant ou génératif et donc n'incluant pas systématiquement les données outliers (i.e. les données de l'autre classe).

Nous avons alors proposé une approche de catégorisation mettant en avant le positionnement des méthodes vis-à-vis de la gestion des données outliers. Les méthodes one-class de la littérature sont classées en fonction de la manière dont les données outliers sont prises en compte au sein de l'apprentissage de ces méthodes. Les approches discriminantes sont très peu mentionnées dans la littérature one-class en raison notamment des difficultés à générer les données de la classe non représentée. Ces méthodes ont en effet besoin de synthétiser les données outliers car elles tracent leur frontière de décision en s'appuyant sur les deux classes en présence. Ces difficultés sont liées particulièrement au problème de la malédiction de la dimension. Á l'exception notable de l'approche one-class SVM proposée par le Prof. Bernhard Scholkopf qui synthétise les données de la classe non représentée (un unique exemple de cette classe est placé à l'origine de l'espace de travail) pour l'apprentissage de la méthode, rares sont les publications traitant de la problématique de la synthèse des donnés outliers dans une approche de classification. De plus, les méthodes de la littérature font généralement l'impasse sur les mécanismes de combinaison à l'œuvre dans les méthodes d'ensemble largement adoptées dans le cas de la classification binaire standard et dont pourrait bénéficier la classification one-class.

Forts de ces constats, nous avons proposé un cadre de travail pour une approche one-class basée sur les méthodes d'ensembles. Nous avons élaboré, à partir de ce cadre, une approche one-class discriminante tirant partie des différents mécanismes de combinaison des méthodes d'ensembles et particulièrement des principes de randomisation à l'œuvre dans les ensembles d'arbres de décision de la famille des forêts aléatoires. La méthode est composée de quatres étapes essentielles : une étape d'extraction de connaissance à partir du set de données initiales disponibles ; une étape de randomisation issue des mécanismes de combinaison des méthodes d'ensembles ; une étape de génération des données outliers tenant compte de la connaissance extraite à la première étape et tirant partie des mécanismes de la seconde étape notamment en terme de réduction de la dimension et de sous-échantillonnage à la fois du set d'apprentissage et du set de caractéristiques disponibles ; une étape d'induction de chaque classifieur individuel utilisé de type arbre de décision. La phase

intégrant les mécanismes des méthodes d'ensemble a notamment permis d'apporter une solution aux problèmes liés à la génération de données en grande dimension dans les approches discriminantes. La phase d'extraction de connaissance a permis de tenir compte de la distribution des données disponibles afin de mieux estimer la distribution des données outliers. Il est à noter que notre approche est capable de faire de la détection de données différentes de celles qui auront été apprises et peut ainsi permettre, outre d'effectuer la tâche one-class, d'aider à caractériser la classe des données contre-exemples. Nous avons montré que cette approche était performante et générique, stable sur plusieurs problèmes de la littérature évalués et qu'elle se comportait de façon équivalente et même mieux que les méthodes de l'état de l'art sur certains problèmes. L'approche que nous avons proposée étant flexible, elle peut être améliorée et optimisée pour une problématique donnée. Ces différentes études ont ainsi conduit à la mise en place d'un système complet de classification basé sur des extracteurs génériques et une méthode d'ensemble également performante et générique pour l'aide au diagnostic dans le cadre des images alvéoscopiques.

Les travaux de thèse exposés présentent ces différentes contributions et sont structurés en 4 chapitres. Dans le premier chapitre nous proposons un état de l'art des systèmes de classification d'images en décrivant les différentes étapes constitutives de la chaine de classification et en présentant les contributions majeures de la littérature dans chacune des phases d'élaboration du système : 1) la phase de pré-traitement de l'image acquise permettant de nettoyer l'image avant son analyse ; 2) la phase d'extraction de caractéristiques discriminantes sur ces images dans laquelle nous proposons une catégorisation des méthodes de la littérature en fonction de l'échelle d'analyse et de la richesse de la description obtenue ; 3) la phase d'apprentissage pour élaborer la fonction de décision dans laquelle nous insistons notamment sur les méthodes d'ensembles en vogue en raison de leurs performances et de leur caractère générique ; puis en 4) la phase d'évaluation des performances de ce système. Cet état de l'art nous permet à la fin de ce chapitre de proposer un système de classification pour les images alvéoscopiques. Nous présentons plus en détail les images alvéoscopiques puis nous détaillons les méthodes retenues et les approches proposées pour chacune des phases mentionnées précédemment.

Dans le chapitre 2 nous mettons en œuvre le système de classification des images présenté dans le chapitre précédent. Nous montrons que le système proposé est performant comparé aux différentes approches de l'état de l'art et nous proposons une nouvelle contribution dans l'adaptation de ce système au mécanisme de pilotage de rejet permettant de renforcer la fiabilité de la décision du système. Cependant, face à la difficulté de caractériser les cas pathologiques au sein des images avéoscopiques contrairement aux cas des images de patients sains nous proposons une orientation one-class.

Dans le chapitre 3, nous présentons un état de l'art des méthodes one-class et nous proposons une nouvelle catégorisation de ces approches, plus à même d'étudier cet état de l'art du point de vue des données outliers. Cet état de l'art nous conduit à proposer un nouveau cadre de travail pour une approche one-class basée sur les ensembles de classifieurs et bénéficiant des mécanismes de combinaison et de randomisation de ces méthodes.

Dans le chapitre 4, nous mettons en œuvre une approche discriminante originale basée sur les méthodes d'ensemble d'arbres de décision de la famille des forêts aléatoires et s'inscrivant dans le cadre one-class que nous avons défini. Nous montrons dans ce chapitre que notre approche est performante, robuste et générique sur une grande variété de problèmes de la littérature, avec des performances proches ou souvent meilleures que celles obtenues par des classifieurs standards. Nous montrons particulièrement sur notre problématique médicale la pertinence de notre approche.

Nous concluons le manuscrit en présentant une synthèse des différents résultats obtenus et nous décrivons les perspectives à court-terme et à long terme pour les contributions issues de ces travaux pour la constitution d'un système de classification des images.

Chapitre 1

La classification d'images : un état de l'art

Sommaire

1.1 Introduction

La classification automatique d'images est une application de la reconnaissance de formes consistant à attribuer automatiquement une classe à une image à l'aide d'un système de classification. On retrouve ainsi la classification d'objets, de scènes, de textures, la reconnaissance de visages, d'empreintes digitales, de caractères parmi les applications courantes [Duda et al., 2000]. Il existe deux principaux types d'apprentissage dépendant des informations disponibles sur les données à classer : l'apprentissage supervisé et l'apprentissage non-supervisé. Dans l'approche supervisée, chaque image est associée à une étiquette qui décrit sa classe d'appartenance. Dans l'approche non-supervisée (ou clustering), les données disponibles ne possèdent pas d'étiquettes ; il appartient alors au système d'extraire une règle d'appartenance de chaque image à un groupe donné. Nous ne traiterons dans cet exposé que de l'approche supervisée et donc pour laquelle les images possèdent une étiquette.

Un système de classification automatique d'images est composé des étapes suivantes : l'étape de pré-traitement permettant de "nettoyer" les images ; la phase d'extraction de caractéristiques permettant de décrire l'information pertinente contenue dans l'image à l'aide d'opérateurs ou de descripteurs discriminants ; la phase d'apprentissage permettant de construire une frontière de décision pour identifier la classe d'une image présentée à l'entrée du système. Ces trois phases sont essentielles dans la construction du système de classification.

La description obtenue lors de la phase d'extraction de caractéristiques peut être locale si elle tient compte des informations locales comme des motifs dans l'image ou globale si elle tient compte de la totalité des informations présentes à l'échelle macroscopique [Mikolajczyk and Schmid, 2005, Schmid et al., 2000]. Les techniques d'extraction de caractéristiques peuvent également se scinder en caractéristiques bas-niveau utilisant par exemple l'information au niveau du pixel dans l'image et en caractéristiques de plus haut niveau avec des descripteurs utilisant notamment une représentation texturée de l'image [Due Trier et al., 1996].

La phase d'apprentissage consiste en la construction d'une règle de décision soit à partir d'un modèle, soit en élaborant une frontière dans l'espace de caractéristiques formé à l'étape précédente. Les méthodes standards de la littérature utilisent un seul classifieur pour la construction de la règle de décision. Ainsi, l'estimation de la fonction de décision se fait à partir d'une seule hypothèse. Cependant, des approches plus flexibles que la formulation d'une unique hypothèse considèrent plutôt une combinaison de classifieurs permettant d'améliorer les performances de ces mêmes classifieurs pris individuellement. Parmi ces approches de combinaison, on peut distinguer les méthodes d'ensembles combinant un même type de classifieurs construits en grand nombre. Particulièrement, les méthodes d'ensemble d'arbres de décision, regroupées dans le cadre plus général des forêts aléatoires [Breiman, 2001], ont suscité un intérêt croissant en raison des fondations théoriques solides dont elles bénéficient avec les travaux fondateurs de Breiman et les récents développements tant expérimentaux que théoriques présentés dans [Robnik-Sikonja, 2004, Geurts et al., 2006, Genuer et al., 2008, Biau et al., 2008, Biau, 2010].

Enfin, pour connaître les performances d'un système de classification d'images, il est nécessaire de définir une procédure d'évaluation consistant notamment dans le choix d'une mesure donnant les performances du système et des données d'évaluation, généralement constituées de bases de la littérature. La problématique de l'évaluation d'un système de classification est d'autant plus difficile qu'il n'existe pas de consensus sur la méthode, les mesures et les procédés d'évaluation.

Nous présentons dans les sections suivantes un état de l'art des différentes approches de la littérature pour chacune des étapes composant le système de classification d'images. Dans un premier temps nous parlons des méthodes d'extraction de caractéristiques. Dans un second temps nous abordons les méthodes d'apprentissage en présentant d'abord des méthodes à classifieur unique puis des méthodes de combinaison de classifieurs. Nous présentons ensuite les méthodes d'évaluation du système.
De cet état de l'art, nous proposons un système de classification pour les images alvéoscopiques,

basé sur une approche générique avec les méthodes d'ensembles d'arbres de décision. Ce système sera mis en œuvre au chapitre suivant.

1.2 Les motivations de la classification d'images

L'objectif de la classification d'images est d'élaborer un système capable d'affecter une classe automatiquement à une image. Ainsi, ce système permet d'effectuer une tâche d'expertise qui peut s'avérer coûteuse à acquérir pour un être humain en raison notamment de contraintes physiques comme la concentration, la fatigue ou le temps nécessité par un volume important de données images.

Les applications de la classification automatique d'images sont nombreuses et vont de l'analyse de documents à la médecine en passant par le domaine militaire [Duda et al., 2000]. Ainsi on retrouve des applications dans le domaine médical comme la reconnaissance de cellules [Keysers et al., 2001], de tumeurs dans les mammographies [Zane et al., 2002] ; dans l'agriculture comme la classification de pollen [Rodríguez-Damián et al., 2004], la reconnaissance du type de sol et des grains [Mäenpää et al., 2002, Mäenpää et al., 2003], la classification d'herbes [Ghazali et al., 2008] ; dans le domaine du document comme la reconnaissance d'écriture manuscrite pour les chèques, les codes postaux [LeCun et al., 1989], les cartes [Due Trier et al., 1996] ; dans le domaine urbain comme la reconnaissance de panneaux de signalisation [Lauziere et al., 2001], la reconnaissance de piétons [Suard, 2006, Suard et al., 2006, Suard et al., 2005, Hilario et al., 2005], la détection de véhicules [Negri et al., 2008], la reconnaissance de bâtiments [Shao et al., 2003] pour aider à la localisation ; dans le domaine de la biométrie comme la reconnaissance de visage [Viola and Jones, 2001, Viola and Jones, 2004], d'empreintes, d'iris [Rad et al., 2004, Guo and Jones, 2008]. Le point commun à toutes ces applications est qu'elles nécessitent la mise en place d'une chaîne de traitement à partir des images disponibles composée de plusieurs étapes afin de fournir en sortie une décision. Chaque étape de la mise en place d'un tel système de classification nécessite la recherche de méthodes appropriées pour une performance globale optimale ; à savoir la phase d'extraction de caractéristiques et la phase d'apprentissage. Typiquement, nous disposons de données images desquelles il nous faut extraire des informations pertinentes traduites sous formes de vecteurs numériques. Cette phase d'extraction nous permet de travailler dans un espace numérique. Il s'agit ensuite d'élaborer, dans la phase d'apprentissage, à partir de ces données initiales, une fonction de décision pour décider de l'appartenance d'une donnée nouvelle à l'une des classes en présence.

La phase d'extraction de caractéristiques peut être précédée d'une phase dite de pré-traitement. Cette phase a pour but de nettoyer l'image, c'est-à-dire d'isoler le contenu informatif ou d'intérêt dans l'image [Cocquerez and Philipp, 1995]. Cette opération permet ainsi d'occulter ou d'atténuer toute information susceptible de nuire à la description du contenu pertinent lors de la phase d'extraction de caractéristiques. On retrouve ainsi des techniques d'atténuation de bruits, de renforcement de contours, des techniques d'amélioration de l'image comme le réhaussement de contraste, la réduction de la dimension de l'image par la binarisation [Otsu, 1979], la réduction de l'image à ses primitives visuelles comme la squelettisation [Baja and Thiel, 1996, Blum, 1961, Lam et al., 1992] ou encore l'extraction de contours à l'aide de techniques de filtrage. Le lecteur peut se référer à un état de l'art des techniques de pré-traitement dans [Gonzalez and Woods, 2002, Cocquerez and Philipp, 1995].

1.3 Extraction de caractéristiques

La phase d'extraction de caractéristiques constitue généralement l'une des phases les plus importantes dans l'élaboration du système [Due Trier et al., 1996]. Il s'agit en effet de déterminer un espace numérique de description dans lequel les données images seront projetées et perme-

ttant une séparation optimale des classes en présence. Nous retrouvons des descripteurs de bas niveau s'intéressant à l'information contenue dans l'image au niveau du pixel et des descripteurs de plus haut niveau nécessitant une représentation intermédiaire de l'image plus adaptée [Srinivasan and Shobha, 2008]. Cette description peut être locale (e.g. description de motifs de textures) ou globale (e.g. histogramme des orientations de la distribution des gradients de toute l'image) selon la nature de l'information à prélever et se fait à l'aide d'opérateurs ou de descripteurs [Due Trier et al., 1996, Mikolajczyk and Schmid, 2005, Zhang et al., 2007].

Nous retrouvons deux types de taxonomies [Bay et al., 2006] : les méthodes classées selon le niveau d'information capturée (ou sémantique) et celles selon la résolution de capture (ou granularité). Dans le premier cas, l'information permettant de caractériser l'image peut être de bas niveau (intensité de pixel), d'un niveau intermédiaire (gradient) ou d'un haut niveau (description d'un pattern, d'un voisinage autour d'un point d'intérêt, d'une forme identifiée). Dans le second cas de type d'images à analyser, la granularité peut être adaptée pour chaque niveau d'information, i.e. une caractérisation globale peut être appliquée à un patch ou à une zone d'intérêt extraite de l'image. Nous présentons ci-après ces deux approches.

1.3.1 Extracteurs de bas-niveau vs extracteurs de plus haut niveau

Extracteurs de bas niveau

Les extracteurs bas niveau permettent de traduire l'information présente au niveau du pixel, sans tenir compte des formes ou des patterns présents dans l'image. Parmi les caractéristiques extraites, nous retrouvons l'intensité du pixel brut, l'histogramme des intensités de pixels, les statistiques sur cet histogramme (moyenne, entropie, variance, coefficient d'aplatissement, asymétrie), la densité de pixels. On rencontre également des extracteurs de niveau intermédiaire traduisant des informations comme des liaisons entre les pixels, des distances, leur localisation, le contraste dans l'image. C'est le cas des statistiques à partir des matrices de co-occurence [Haralick et al., 1973], de simples gradients de l'image ou des statistiques sur des sorties de filtres appliquées à l'image via une caractérisation spectrale.

Nous présentons dans cette section les statistiques d'histogrammes, de co-occurence, et la caractérisation spectrale, approches couramment utilisées dans la littérature.

Les statistiques d'histogramme

L'histogramme de l'image représente la distribution des intensités de pixels ; à chaque valeur d'intensité est associée le nombre de pixels ayant cette valeur dans l'image. Cette approche simple et générique est souvent utilisée pour l'analyse de texture [Cocquerez and Philipp, 1995]. Nous donnons ci-dessous quelques mesures statistiques couramment utilisées, g étant la valeur de l'intensité (e.g. $g \in [0;255]$), n le nombre de niveaux de gris, P_g étant la fréquence de l'intensité g dans une région R de l'image :

– la mesure de l'intensité moyenne :

$$\mu_R = \sum_{g=0}^{n-1} g \cdot P_g$$

– la variance mesurant le contraste moyen :

$$\sigma_R^2 = \sum_{g=0}^{n-1} (g - \mu_R)^2 \cdot P_g$$

– l'entropie mesurant le degré d'incertitude dans la distribution :

$$T_R = -\sum_{g=0}^{n-1} P_g \cdot log_2 P_g$$

– l'énergie :

$$E_R = -\sum_{g=0}^{n-1} P_g^2$$

– la dissymétrie :

$$D_R = \frac{1}{\sigma_R^3} \sum_{g=0}^{n-1} (g-\mu_R)^3 \cdot P_g$$

– l'aplatissement :

$$A_R = \frac{1}{\sigma_R^4} \sum_{g=0}^{n-1} (g-\mu_R)^4 \cdot P_g$$

Les statistiques des matrices de co-occurrence

Les matrices de co-occurrence [Haralick et al., 1973] sont un moyen de quantifier les relations spatiales entre les niveaux de gris des pixels de l'image en analysant la distribution jointe de paires de pixels. Cette approche est couramment utilisée pour la caractérisation des textures d'une image.

La matrice de co-occurence [1] $\mathbf{MC}_t$ d'une région $\mathbf{R}$ de l'image et associée à un vecteur de translation $\mathbf{t}$ contient les effectifs de paires de niveaux de gris entre un pixel et son translaté. Le vecteur de translation t est un paramètre de la matrice et est défini par une orientation et une longueur. Les éléments de cette matrice sont donnés par :

$$\mathbf{MC_t}(i,j) = card\{(\mathbf{s}, \mathbf{s}+\mathbf{t}) \in \mathbf{R} | I_\mathbf{s} = i, I_{\mathbf{s}+\mathbf{t}} = j\}$$

i,j étant les niveaux de gris de l'image, $I_\mathbf{s}$ désigne le niveau de gris du pixel $\mathbf{s}$ dans l'image, *card* étant la fonction cardinal retournant le nombre d'éléments de l'ensemble.
Afin d'extraire l'information contenue dans cette matrice, des indices statistiques sont ensuite définis. [Haralick et al., 1973] a initialement proposé 14 indices statistiques parmi lesquels ceux cités précédemment (adaptés à un histogramme bidimensionnel) et les indices suivants communément cités dans la littérature [Cocquerez and Philipp, 1995] :

– l'homogénéité, indice d'autant plus élevé que les zones sont homogènes ou périodiques :

$$\frac{1}{N_c^2} \sum_i \sum_j \mathbf{MC}_t(i,j)^2$$

N_c désignant le nombre de couples de la région de l'image analysée.

– le contraste

$$\frac{1}{N_c(L-1)^2} \sum_{k=0}^{L-1} k^2 \sum_{|j-i|=k} \mathbf{MC}_t(i,j)$$

L désignant le nombre de niveaux de gris dans l'image ; cet indice pondère le terme de la matrice par sa distance à la diagonale.

– l'entropie

$$1 - \frac{1}{N_c ln(N_c)} \sum_i \sum_j \mathbf{MC}_t(i,j) \cdot ln(\mathbf{MC}_t(i,j)) \cdot \mathbf{1}_{\mathbf{MC}_t(i,j)}$$

avec

$$\mathbf{1}_{\mathbf{MC}_t(i,j)} = \begin{cases} 1 \text{ si } \mathbf{MC}_t(i,j) \neq 0 \\ 0 \text{ sinon} \end{cases}$$

cet indice est faible si le même couple de pixels est redondant dans l'image ; ainsi il quantifie le niveau de désorganisation de la texture.

1. On retrouve le sigle GLCM pour "Gray-Level Cooccurrence Matrix"

– la corrélation

$$\frac{1}{N_c \sigma_x \sigma_y} |\sum_i \sum_j (i - m_x)(j - m_y) \mathbf{MC}_t(i,j)|$$

$m_x, m_y, \sigma_x, \sigma_y$ désignant respectivement la moyenne pondérée sur les lignes, les colonnes, la variance des lignes, des colonnes de $\mathbf{MC}_t$; nous en donnons les expressions ci-dessous :

$$\begin{cases} m_x = \frac{1}{N_c} \sum_i \sum_j i \cdot \mathbf{MC}_t(i,j) \\ m_y = \frac{1}{N_c} \sum_i \sum_j j \cdot \mathbf{MC}_t(i,j) \\ \sigma_x^2 = \frac{1}{N_c} \sum_i \sum_j (i - m_x)^2 \cdot \mathbf{MC}_t(i,j) \\ \sigma_y^2 = \frac{1}{N_c} \sum_i \sum_j (j - m_y)^2 \cdot \mathbf{MC}_t(i,j) \end{cases}$$

– l'homogénéité locale

$$-\frac{1}{N_c} \sum_i \sum_j \frac{1}{1 + (i-j)^2} \mathbf{MC}_t(i,j)$$

– la directivité détectant les pixels de même niveau de gris

$$\frac{1}{N_c} \sum_i \mathbf{MC}_t(i,i)$$

– l'uniformité

$$\frac{1}{N_c^2} \sum_i \mathbf{MC}_t(i,i)^2$$

Typiquement, les valeurs d'orientation du vecteur de translation sont prises dans $\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ et les longueurs sont généralement associées à une ou deux fois la longueur d'un pixel de l'image. Une approche multi-résolution est souvent utilisée, concaténant les différents espaces obtenus avec des vecteurs de translation différents. En effet, elle permet d'obtenir un set de descripteurs beaucoup plus large, à l'aide de différentes résolutions spatiales et angulaires permettant ainsi d'être invariant à la rotation et à la translation.

L'extracteur issu des caractéristiques spectrales

La caractérisation spectrale consiste à analyser l'image dans le domaine fréquentiel par exemple par la transformée de fourier ou la transformée en ondelettes. Dans l'approche par transformée de Fourier, l'approche de Fourier-Mellin est souvent citée en raison de ses propriétés d'invariance à la translation, à la rotation et au changement d'échelle [Reddy and Chatterji, 1996, Dahmen et al., 2000, Adam et al., 2001]. L'approche de Fourier-Mellin ayant des difficultés à décrire localement l'image, les transformées en ondelettes sont privilégiées.

Les ondelettes [Grossmann and Morlet, 1984, Cocquerez and Philipp, 1995] permettent une approche multi-résolution et ont fait l'objet de beaucoup de développements notamment dans l'analyse de textures [Iftene and Safia, 2004, Unser, 1995], l'imagerie médicale [Castellano et al., 2004], la reconnaissance faciale [Joutel et al., 2007, Negri et al., 2004, Zhang et al., 1998]. La transformée en ondelettes consiste à décomposer le signal d'entrée f(t) sur une série de fonctions $\Psi_{a,b}$ qui sont les translatées de paramètre b et dilatées de paramètre a d'une fonction mère Ψ, appelée ondelette, dont les coefficients $C_{a,b}$ de la série sont donnés par :

$$C_{a,b} = \int_{-\infty}^{\infty} f(t) \cdot \Psi_{a,b}(t) dt$$

de telle sorte que le signal original puisse être reconstruit comme suit :

$$\begin{cases} f(t) = \frac{1}{K_\Psi} \int_0^\infty \int_{-\infty}^\infty C_{a,b} \cdot \Psi_{a,b}(t) \frac{db \cdot da}{a^2} \\ K_\Psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{|\omega|} d\omega \end{cases}$$

avec $\Psi_{a,b}(t) = a^{\frac{1}{2}} \Psi(\frac{t-b}{a})$, a étant le facteur d'échelle (dilatation) et b la position du signal (translation).
Les ondelettes de Haar sont les plus utilisées dans la littérature en partie pour leur simplicité de mise en œuvre. Elles sont définies à partir de la fonction mère :

$$\Psi(t) = \begin{cases} 1 & \text{si } t \in [0;1/2[\\ -1 & \text{si } t \in [1/2;1[\\ 0 & \text{sinon.} \end{cases}$$

Ces ondelettes ont été popularisées par Schneiderman [Schneiderman and Kanade, 2005] pour la détection d'objets, pour la description de contours. Dans [Papageorgiou et al., 1998], les auteurs ont ainsi utilisé la décomposition en ondelettes de Haar pour la description de scènes en présence de piétons. L'espace des caractéristiques est obtenu à partir de la décomposition de l'image initiale en sous-bandes successives où l'image est d'abord filtrée horizontalement , sous-échantillonnée puis filtrée verticalement. Le descripteur est généralement constitué des coefficients d'ondelettes issus de chaque sous-bande. Mais des descripteurs standards appliqués aux images en sortie des filtres peuvent aussi servir à caractériser l'image originale.

Les filtres de Gabor [Gabor, 1946] sont également populaires dans la littérature pour notamment l'extraction de contours avec une orientation particulière [Tian et al., 2002, Zhang et al., 1998]. La fonction utilisée dans le filtre de Gabor est l'association d'une gaussienne et d'une sinusoïde donnée par :

$$G(x,y,\theta,\omega) = exp(-\frac{1}{2}(\frac{x_\theta^2}{\sigma_x^2} + \frac{y_\theta^2}{\sigma_y^2}) \cdot cos(2\pi \cdot \omega \cdot x_\theta)$$

où $x_\theta = x \cdot cos(\theta) + y \cdot sin(\theta)$ et $y_\theta = y \cdot cos(\theta) - x \cdot sin(\theta)$, θ étant l'orientation de la sinusoïde, ω sa fréquence, σ_x et σ_y étant les écarts-types de la gaussienne selon les deux axes du repère.
Ces filtres permettent d'isoler les contours d'image dont l'orientation est perpendiculaire à l'orientation de la sinusoïde. La détection des contours de l'image implique ainsi l'utilisation d'un ensemble de filtres de Gabor à différentes orientations appelé banc de filtres.
L'utilisation de la décomposition en ondelettes est cependant coûteuse et dans le cas des filtres de Gabor, un grand nombre de filtres est nécessaire pour décrire de faibles changements en fréquence et en orientation dans l'image [Mikolajczyk and Schmid, 2005] et des redondances dans la caractérisation des sorties du banc de filtres peuvent apparaitre [Unser, 1995].

Extracteurs de plus haut-niveau

Les méthodes d'extraction de plus haut niveau tiennent compte des formes et des structures dans l'image, des relations spatiales entre les pixels ou ces structures. Les propriétés les plus recherchées dans ces extracteurs sont, outre leur pouvoir descriptif, l'invariance et la robustesse à différentes transformations pouvant affecter l'image [Mikolajczyk and Schmid, 2004, Mikolajczyk and Schmid, 2005, Bay et al., 2006]. Ainsi, la description obtenue demeure relativement inchangée face à ces transformations pouvant plus ou moins affecter des contenus identiques ou similaires dans les images. On retrouve couramment les invariances au changement d'échelle, de perspective, aux transformations affines comme la translation, la rotation et la robustesse au changement de luminosité ou de contraste.
Les approches standards de la littérature d'extraction de caractéristiques dans l'image sont principalement à base de descripteurs de textures. On retrouve particulièrement les extracteurs issus du détecteur de Harris [Harris and Stephens, 1988, Schmid et al., 2000], des approches Local Binary Pattern (LBP) [Ojala et al., 1996, Ojala et al., 2000] et Scale Invariant Feature Transform (SIFT) [Lowe, 2004], présentées dans la littérature comme des approches génériques et parmi les plus

performantes [Mikolajczyk and Schmid, 2005, Bay et al., 2006].

L'extracteur issu du détecteur de Harris

Le détecteur de Harris (ou détecteur de Harris & Stephens) est une approche de description locale d'une région d'intérêt dans l'image en termes de coins et bords [Harris and Stephens, 1988]. Elle est basée sur l'analyse de gradients calculés localement ou dans une région d'intérêt (ROI) de l'image I, plus particulièrement sur la somme pondérée des différences quadratiques (Sum of Squared Differences) donnée par l'expression suivante :

$$SSD(x,y) = \sum_{(u,v)\in ROI} w(u,v)\cdot(I(u+x,v+y)-I(u,v))^2$$

où (x,y) représente un vecteur de translation, I étant la matrice d'image 2D (en niveaux de gris) et w une fenêtre de calculs contenant des coefficients de pondération (généralement un poids gaussien est choisi). Avec l'approximation de Taylor $I(u+x,v+y) \approx I(u,v)+I_x(u,v)x+I_y(u,v)y$, I_x et I_y désignant les dérivées partielles de I, on obtient, par substitution :

$$SSD(x,y) \approx \sum_{(u,v)\in ROI} w(u,v)\cdot(I_x(u,v)x+I_y(u,v)y)^2 = \begin{pmatrix} x & y \end{pmatrix} H \begin{pmatrix} x \\ y \end{pmatrix}$$

avec $H = \begin{pmatrix} < I_x^2 > & < I_x I_y > \\ < I_x I_y > & < I_y^2 > \end{pmatrix}$ désignant la matrice d'auto-corrélation de Harris, moyennant les dérivées de l'image, l'opérateur $< \cdot >$ désignant la somme pondérée sur $(u,v) \in ROI$.

L'analyse des valeurs propres de la matrice H renseigne sur la structure et la nature de la région analysée. Leurs coûts de calculs étant élevés, les auteurs ont proposé la mesure suivante :

$$r_H = det(H) - \kappa Trace(H)^2 = (\lambda_1\lambda_2) - \kappa(\lambda_1+\lambda_2)^2$$

H étant la matrice de Harris vue précédemment, λ_1 et λ_2 les valeurs propres de la matrice, κ une valeur empirique classiquement évaluée dans l'intervalle [0.04 ;0.15]. Il suffit ainsi, pour avoir la caractéristique de la région d'intérêt, de calculer le déterminant et la trace de la matrice H sans passer par le calcul des valeurs propres.
En effet, si on désigne par λ_1 et λ_2 les valeurs propres associées aux deux directions de l'espace propre de H, on a 3 situations :

- λ_1 et λ_2 ont des valeurs importantes : la région définit un coin et r_H a une valeur importante [2]
- l'une des deux valeurs seulement est importante : la région définit un bord et r_H est négatif [3]
- les deux valeurs propres sont proches de 0 : on a une zone homogène, sans caractérisation particulière et r_H est proche de 0 [4]

Cependant, l'approche n'est pas invariante à l'échelle et aux transformations affines [Schmid et al., 2000]

Scale Invariant Feature Transform (SIFT)

SIFT (Scale Invariant Feature Transform) proposé par [Lowe, 2004], est à la fois un détecteur et un descripteur de points d'intérêts dans l'image de la famille des histogrammes de gradients orientés. Les caractéristiques principales résident dans ses nombreuses invariances aux transformations affines, au changement d'échelles (ou résolution), de perspectives de faible distorsion et sa robustesse au bruit, à l'occlusion et à l'illumination dans l'image. SIFT est une approche générique et est considérée comme l'une des approches les plus performantes de la littérature [Mikolajczyk and Schmid, 2005, Ghazali et al., 2008, Negri et al., 2008].

2. $r_H \approx \lambda^2 - 4\cdot\kappa\lambda^2 = \lambda^2(1-4\kappa)$ avec $\lambda \sim \lambda_1 \sim \lambda_2 >> 1$.
3. $r_H \approx \lambda_1\cdot\lambda_2 - \kappa\lambda_1^2 = \lambda_1^2(\lambda_2/\lambda_1 - \kappa)$ avec par exemple $\lambda_1 >> 1$ et λ_2 proche de 0.
4. $r_H \approx \lambda^2(1-4\kappa)$ avec $\lambda \sim \lambda_1 \sim \lambda_2 << 1$

L'algorithme se compose d'abord d'une étape de détection des points d'intérêts puis d'une étape de description des points retenus. Dans l'étape de détection des points d'intérêts, SIFT identifie les extrema locaux dans l'espace échelle de l'image (constitué de différentes versions lissées de l'image) à partir de différences de gaussiennes. L'espace échelle L d'une image I est l'image lissée obtenue par la convolution entre une gaussienne et cette image et dont l'expression est :

$$L(x,y,\sigma) = G(x,y,\sigma)) * I(x,y)$$

où $G(x,y,\sigma) = \frac{1}{2\pi\cdot\sigma^2} \cdot e^{-(x^2+y^2)/(2\cdot\sigma^2)}$.

Les extrema sont ainsi détectés en construisant une pyramide gaussienne. En effet, plusieurs niveaux de filtrage gaussien sont appliqués à l'image, avec un paramètre d'échelle croissant choisi dans la suite $\{\sigma, k\cdot\sigma, k^2\cdot\sigma, ..., k^s\cdot\sigma\}$ appelée octave, s étant choisi de telle sorte que $k^s = 2$.
On effectue ensuite une différence entre les échelles adjacentes (ou "DoG" pour Difference of Gaussian) d'une octave, donnée par la formule suivante :

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,k\cdot\sigma) - G(x,y,\sigma)) * I(x,y) \\ &= L(x,y,k\cdot\sigma) - L(x,y,\sigma) \end{aligned}$$

Les maxima pour l'échelle courante sont ensuite obtenus par comparaison d'un point à ses 8 voisins à son échelle et les 18 plus proches voisins aux échelles adjacentes comme indiqué sur la Figure 1.1. Un premier set de points d'intérêts candidats est ainsi formé.
Le procédé précédent étant susceptible de générer un grand nombre de points, plusieurs critères ont été définis afin d'en réduire le nombre aux plus robustes d'entre eux. Tout d'abord, les points à faible contraste étant sensibles au bruit dans l'image, seuls les points avec un contraste suffisamment élevé seront conservés. Ainsi, un seuil sur le contraste a été défini empiriquement par l'auteur à 0.04. Ensuite, la stabilité du point d'intérêt candidat est évaluée. SIFT considère pour cela un offset $\mathbf{x}$ dans le voisinage du point candidat (x,y,σ) et évalue le DoG pour ce nouveau point. Un développement de Taylor à l'ordre 2 fournit l'approximation suivante (moyennant une translation, le point candidat (x,y,σ) est ramené à l'origine de l'espace échelle) :

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x}$$

avec ici $D = D(0)$. En dérivant l'expression précédente, on obtient la position de l'extremum de la fonction $D(\mathbf{x})$: $\hat{\mathbf{x}} = -(\frac{\partial^2 D}{\partial \mathbf{x}^2})^{-1}\cdot\frac{\partial D^T}{\partial \mathbf{x}}$.
Empiriquement, un seuil sur la longueur de l'offset est fixé à 0.5. Au delà de ce seuil, SIFT considère que le point d'intérêt évalué n'est pas stable. Un seuil est alors fixé sur la valeur du DoG au point offset lui-même. Une valeur empirique de 0.03 est proposée par l'auteur. Ainsi, ne sont retenus que les points candidats pour lesquels l'offset a une valeur de DoG inférieure à la valeur seuil. Le DoG étant très sensible aux éléments de contour de l'image, génère beaucoup de points candidats même dans le cas de faibles contours. SIFT utilise alors une approche similaire à celle utilisée dans l'algorithme de détection de bords de Harris-Stephens [Harris and Stephens, 1988], en évaluant la courbure principale des gradients au point considéré. Seuls les points ayant une forte réponse sur ces bords seront alors conservés.

Nous passons à présent à la phase de description du point d'intérêt retenu. A chaque point d'intérêt (x,y,σ) dans l'espace échelle $L(\cdot,\cdot,\sigma)$ de l'image est associé une longueur $m(x,y,\sigma)$ et une orientation $\theta(x,y,\sigma)$ du gradient principal $\nabla L(x,y,\sigma)$ obtenu par différences des pixels adjacents. Les valeurs de m et θ sont données par les expressions suivantes :

$$m(x,y) = ||\nabla L||_2 = \sqrt{(L(x+1,y,\sigma) - L(x-1,y,\sigma))^2 + (L(x,y+1,\sigma) - L(x,y-1,\sigma))^2}$$

$$\theta(x,y) = tan^{-1}(\frac{L(x,y+1,\sigma) - L(x,y-1,\sigma)}{L(x+1,y,\sigma) - L(x-1,y,\sigma)})$$

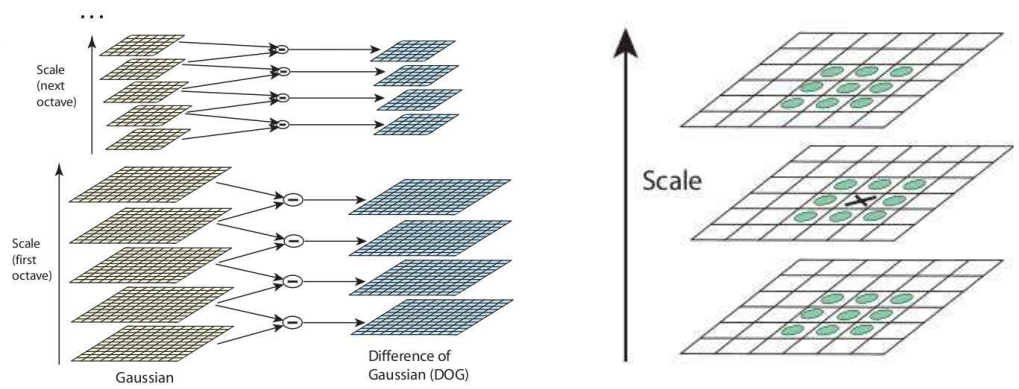


FIGURE 1.1 – Pyramide gaussienne et recherche d'extrema (source de l'image [Lowe, 2004]).

Attribuer une orientation principale au point d'intérêt permet de caractériser les orientation de son voisinage relativement à cette orientation et donc permet de rendre sa caractérisation invariante à la rotation.
Ensuite, un histogramme des orientations des gradients dans son voisinage est calculé. SIFT calcule le gradient de chaque pixel dans un voisinage du point d'intérêt de taille 16×16 empirique et pondéré par une gaussienne d'écart-type correspondant à l'échelle trouvée pour le point d'intérêt. Ce voisinage est ensuite divisé en 16 sous-régions 4×4 et pour chacune de ces 16 sous-régions, les champs de gradient sont synthétisés selon 8 orientations associées à 8 entrées de l'histogramme de gradients orientés. L'histogramme est composé des longueurs de chacun de ces gradients, pondérés par la norme du gradient principal. Les orientations calculées sont relatives à l'orientation principale du point d'intérêt. L'histogramme contient ainsi $4 \times 4 \times 8 = 128$ entrées. Il est ensuite normalisé afin de réduire les effets liés aux changements de luminosité dans l'image et donc de rendre le descripteur robuste à cette transformation. Une illustration proposée dans la Figure 1.2 montre les gradients utilisés pour le calcul de l'histogramme dans un voisinage réduit du point d'intérêt.
Plusieurs variantes de SIFT ont été proposées dans la littérature, améliorant notamment les coûts de calculs ou se restreignant à certaines invariances. On peut citer notamment les approches Dense-SIFT [Bosch et al., 2006] décrivant de façon exhaustive le contenu de l'image à partir d'une grille régulière (la phase de détection est ainsi évitée par rapport au SIFT standard) et Speeded Up Robust Feature (SURF) [Bay et al., 2006]. SURF a la particularité d'être moins coûteux que SIFT en utilisant notamment une méthode d'évaluation des gradients basée sur la technique de l'intégrale image [Viola and Jones, 2001] et les ondelettes de Haar.

Le Local Binary Pattern

Le Local Binary Pattern (LBP) ou "motif de texture binaire local" est un opérateur local d'analyse de texture introduit par Ojala et al. [Ojala et al., 1996, Ojala et al., 2000]. L'opérateur décrit la structure locale de la texture, par le biais de motifs circulaires (cf. Figure 1.5). Des contraintes, que nous détaillons ci-après, sont appliquées aux motifs afin de les rendre robustes et invariants. Ces motifs sont obtenus pour chaque pixel de l'image en seuillant chaque pixel du voisinage circulaire par rapport au pixel central. Ce voisinage est paramétré par sa distance R au pixel central et le nombre P de pixels présents (population du voisinage) comme on peut le voir sur la Figure 1.3. R est considéré comme un paramètre de résolution spatiale et P un paramètre de résolution angulaire. Il est donc possible de combiner plusieurs résolutions de l'opérateur en faisant varier à la fois la résolution spatiale et la résolution angulaire. Le LBP est capable ainsi de détecter les microstructures comme les coins, les bords comme présenté dans la Figure 1.4.
La distribution des motifs trouvés dans l'image est utilisé pour la description de l'image. Il a été montré que 90% des motifs extraits dans l'image sont constitués de ceux de la première ligne de la Figure 1.5 [Ojala et al., 2000, Mäenpää, 2003].

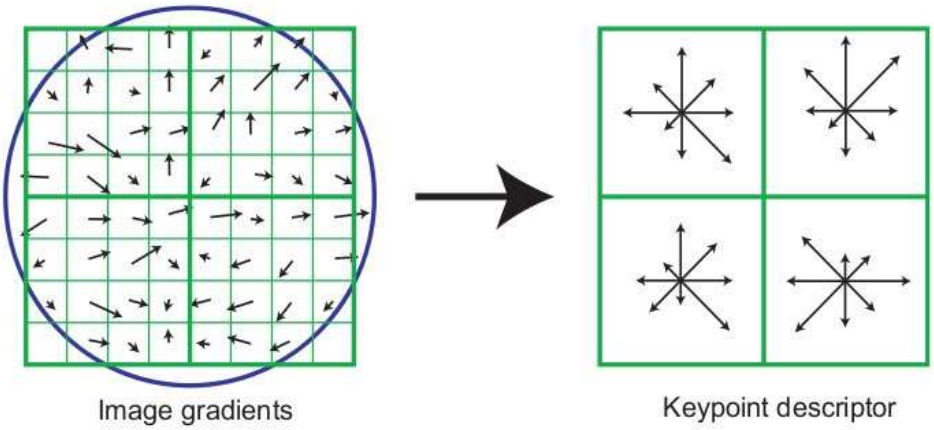


FIGURE 1.2 – Construction du descripteur SIFT dans le voisinage pondéré (avec un noyau gaussien) du point d'intérêt détecté ; dans cet exemple, un voisinage 8×8 est considéré plutôt que la valeur standard de 16×16 pour des raisons de clarté (source de l'image [Lowe, 2004]).

Pour le pixel c, l'opérateur a pour formulation :

$$LBP^{riu2}_{P,R}(c) = \begin{cases} \sum_k s(g_k - g_c) \text{ si } U(LBP^{ri}_{P,R}) \leq 2 \\ P+1 \text{ sinon} \end{cases}$$

avec :

$$\begin{cases} LBP^{ri}_{P,R} = min\ Cir\{LBP_{P,R}, i = 1..P\} \\ LBP_{P,R} = \sum_{k=0}^{P-1} 2^k \cdot s(g_k - g_c) \end{cases}$$

et

- la fonction s est la fonction qui vaut 0 si la différence $g_k - g_c$ est négative et 1 sinon
- g_k, g_c représentent respectivement l'intensité du k-ième élément du contour de coordonnées $(Rcos(k\frac{2\pi}{P}), -Rsin(k\frac{2\pi}{P}))$ et celle du pixel central
- la fonction Cir donne la valeur du LBP après avoir effectué une rotation du motif dans le sens horaire en passant d'un élément du contour à l'autre, l'opération étant répétée i fois ; on prend le minimum de cette valeur afin d'obtenir l'invariance par rotation (*ri* pour "rotation invariant")
- la fonction U (pour "Uniform") mesure le nombre de transitions d'une valeur binaire à une autre sur le contour (ne sont ainsi retenues que les transitions au plus égales à 2) :

$$U(LBP_{P,R}) = s(g_0 - g_{P-1}) + \sum_{k=1..P} s(g_k - g_{k-1})$$

- la valeur $P+1$ est attribuée à tous les autres motifs non représentées par les patterns $\{0...P-1\}$;

Le descripteur ainsi obtenu à partir de la distribution des motifs du LBP dans l'image est de dimension $P+2$.

Afin de rendre plus robuste la description, aux caractéristiques obtenues précédemment peut être ajoutée une caractérisation du contraste local associé au motif (à la résolution spatiale et angulaire considérée), en calculant la variance :

$$VAR_{P,R} = \frac{1}{P} \sum_{k=0}^{P-1} (g_k - \mu)^2$$

où μ est la moyenne des intensités des pixels du voisinage donnée par $\mu = \frac{1}{P} \sum_{k=0}^{P-1} g_k$.
L'utilisation complémentaire des opérateurs LBP et VAR fournit un descripteur de texture LBP/VAR robuste au contraste, à la rotation, aux changements monotones sur les niveaux de gris. L'opérateur

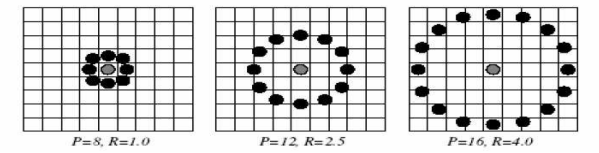


FIGURE 1.3 – Voisinage circulaire considéré pour le LBP (source de l'image [Mäenpää, 2003])

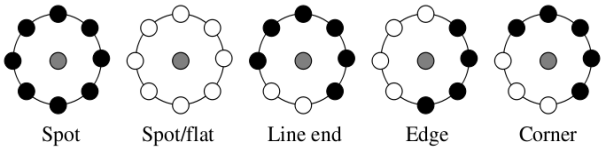


FIGURE 1.4 – Primitives visuelles détectées par l'opérateur LBP, allant du coin au bord, à la zone homogène ou au point (source de l'image [Mäenpää, 2003])

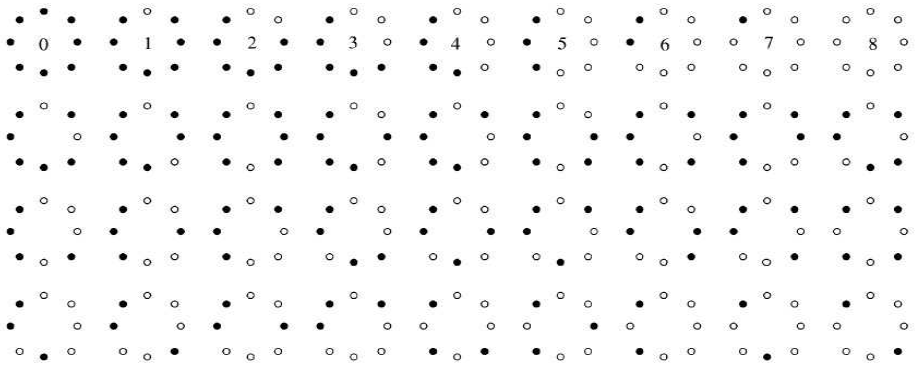


FIGURE 1.5 – Dans le cas du $LBP_{8,1}$, 36 configurations sont possibles à une rotation près. Seules celles ayant au plus 2 transitions -passage d'une valeur binaire à une autre sur le contour circulaire- sont finalement retenues pour la description, soit seulement les 9 configurations de la première ligne (source de l'image [Mäenpää, 2003])

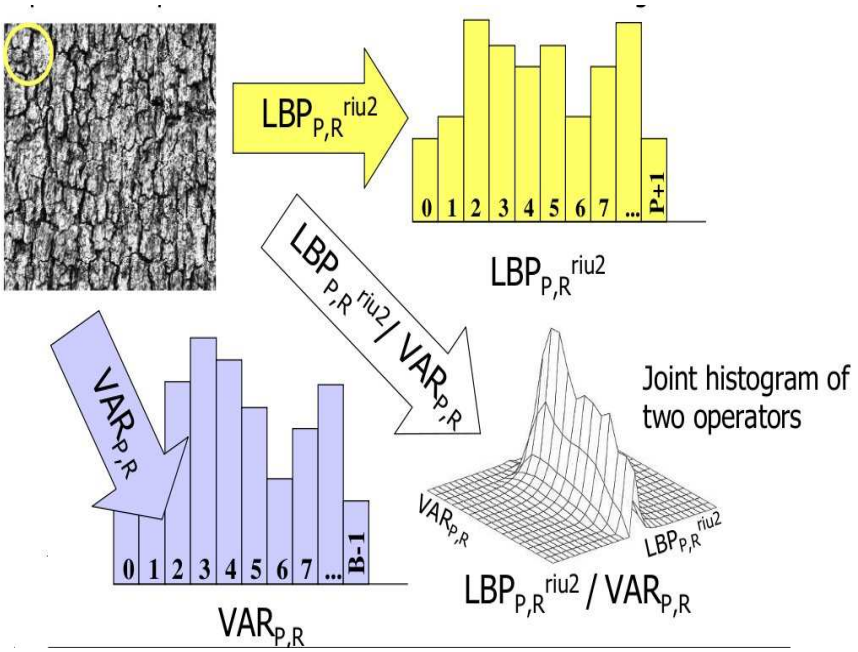


FIGURE 1.6 – Histogramme et quantification du LBP et de la variance locale pour un modèle de texture donné [Mäenpää, 2003]

LBP est quantifié avec la distribution des occurrences des différents motifs binaires ; l'opérateur *VAR* quant à lui possède des valeurs continues et doit être quantifié (cf. Figure 1.6).

L'opérateur LBP connaît de nombreux développements et extensions qui améliorent en général les coûts de calculs ou la richesse des motifs détectés. Ainsi, le Center-Symmetric-LBP (CS-LBP) [Heikkila et al., 2006] calcule des gradients sur le voisinage circulaire, non pas entre le pixel central et un pixel du voisinage mais entre deux pixels diamétralement opposés. L'approche est moins coûteuse que LBP et le descripteur obtenu en remplaçant le gradient dans le descripteur SIFT par l'opérateur CS-LBP permet d'obtenir des performances généralement meilleures que celles obtenues avec le descripteur par histogramme de motifs LBP mais en demeurant toutefois plus coûteux que ce dernier. D'autres approches consistent par exemple à extraire les motifs LBP à différentes échelles dans l'image (i.e. à différents degrés de lissage de l'image) ou à combiner différents descripteurs obtenus en faisant varier les paramètres de résolution spatiale, angulaire ou d'échelle [Maenpaa and Pietikainen, 2003].

1.3.2 Caractérisation locale vs. globale

On distingue pour l'extracteur de caractéristiques généralement trois niveaux de résolution de capture de l'information ou "granularité" présente dans l'image : l'échelle microscopique, l'échelle mésoscopique et l'échelle macroscopique. Au niveau de l'échelle microscopique, l'extracteur se trouve au plus près du pixel pour l'identification de micro-structures présentes dans l'image (e.g. l'intensité du pixel même ou pixel brut). Un tel descripteur est en général sensible au bruit présent dans l'image et une étape de pré-traitement peut s'avérer inévitable pour réduire les artefacts nuisant à la capture de l'information pertinente. Á l'échelle mésoscopique, l'extracteur, à l'exemple de LBP ou SIFT, capture des informations au niveau des motifs ou textons représentant une fraction non négligeable de l'image. Ces extracteurs en général identifient ou travaillent dans des régions d'intérêt de l'image (zones de forts contrastes notamment) ou tiennent compte du voisinage des point d'intérêts trouvés dans l'image. Au niveau macroscopique, on retrouve en général des approches dédiées au contenu de l'image permettant d'identifier les formes globales (e.g. formes géométriques) et donc fortement dépendantes du contenu de l'image. Ces extracteurs en général ne sont pas génériques.

Il est pertinent de définir l'échelle ou la taille des régions d'intérêt pour la capture de cette information. Plus la région d'intérêt est petite plus la caractérisation doit tenir compte des variations locales dans l'image. On peut remarquer cependant qu'il est possible de réduire la zone d'influence d'un extracteur macroscopique afin de le faire travailler à l'échelle mésoscopique par l'intermédiaire par exemple d'un découpage de l'image d'origine. Á l'inverse, on peut vouloir conserver la richesse des informations extraites aux échelles mésoscopiques et microscopiques en travaillant à l'échelle macroscopique. Il est possible d'obtenir un extracteur macroscopique en synthétisant les informations obtenues aux échelles microscopiques et mésoscopiques à l'aide par exemple d'un histogramme de la distribution des éléments individuels capturés (i.e. distribution des différents motifs identifiés, différentes orientations détectées). C'est le cas notamment de l'approche par vocabulaire visuel [Jurie and Triggs, 2005, Nowak and Jurie, 2007].

Une approche particulièrement originale pour la description à la fois locale et dense du contenu des images a été proposée par Marée et al. [Marée et al., 2003, Maree et al., 2005, Maree, 2005]. Il s'agit de décrire, à l'aide de la valeur d'intensité du pixel brut, un grand nombre de fenêtres extraites de l'image ayant des tailles aléatoires et prélevées à des positions aléatoires dans l'image. Ces fenêtres deviennent autant de représentants de l'image d'origine au sein d'une base artificiellement agrandie (composée uniquement de ces fenêtres). Cette technique permet l'extraction d'informations à la fois locales pour les fenêtres de petite taille et globales dans l'image pour celles de grande taille. Il est à noter qu'on retrouve une approche similaire à celle proposée par Marée et al. dans le cadre de la constitution d'un vocabulaire visuel à partir de sacs de mots visuels ("bag-of-features") pour l'identification d'objets dans l'image [Nowak et al., 2006, Moosmann et al., 2008]. Dans cette approche, les auteurs extraient des fenêtres de taille fixe dans l'image (ces fenêtres sont localisées soit de façon régulière sur une grille, soit aléatoire ou à l'aide au préalable de détecteurs de points d'intérêts) et les décrivent à l'aide d'un extracteur standard comme SIFT.

1.3.3 Conclusion

Nous avons vu dans cette section différentes techniques d'extraction de caractéristiques généralement proposées dans la littérature. Ces extracteurs sont souvent présentés dans le cadre d'une problématique donnée ou pour un type d'images spécifique. Cependant, la plupart d'entre eux sont des extracteurs de caractéristiques pour l'analyse de textures et peuvent ainsi se révéler génériques. Les extracteurs génériques sont intéressants car ils sont capables de décrire des images de types différents au moyen notamment d'invariances à certaines déformations courantes dans les images (étirement, rotation, retournement, changement d'échelle) et robustes aux occlusions ou aux transformations plus importantes grâce à une description invariante du voisinage de points d'intérêts.

Nous avons vu l'importance de choisir un extracteur en fonction du niveau d'information auquel on souhaite travailler (bas niveau si on souhaite travailler au niveau pixel ou plus haut niveau si l'information est contenue dans des régions ou structures ou micro-structures dans l'image) et l'échelle à laquelle se trouve cette information. Des combinaisons peuvent être construites afin d'obtenir une description plus riche de l'image en concaténant les descriptions obtenues à plusieurs niveaux et/ou plusieurs échelles de capture de l'information. Mais ce choix de combinaison est souvent effectué au détriment des coûts de calculs supplémentaires engendrés et produit des espaces de description de grandes dimensions souvent difficiles à gérer par les algorithmes de classification standard de la littérature.

Nous avons fait le choix dans cette étude, de travailler avec des extracteurs de haut niveau comme les deux approches couramment citées dans la littérature LBP et SIFT pour la description riche de structures et de micro-structures dans les images texturées que nous avons à traiter. Pour le choix de la résolution de capture de l'information, nous avons ainsi fait le choix de l'approche par extraction de fenêtres aléatoires car l'approche de Marée et al. permet une capture à la fois au niveau local et au niveau global de l'information dans l'image. Nous abordons dans la section

suivante les différentes méthodes d'apprentissage permettant d'identifier les différentes classes d'images en présence dans l'espace de description obtenu.

1.4 Méthodes de classification

Nous nous intéressons dans cette section aux méthodes d'apprentissage dans un espace de description tel que défini dans la section précédente. Le but de cette étape est de construire une fonction de décision à partir de données d'apprentissage projetées dans cet espace de description. On distingue principalement deux tendances dans les méthodes d'apprentissage : l'approche à classifieur unique et l'approche avec une combinaison de classifieurs agrégeant les réponses de plusieurs classifieurs (voir [Hastie et al., 2001] pour une référence exhaustive aux méthodes de classification).
La première approche consiste à élaborer une unique fonction de décision directement à partir des données d'apprentissage en faisant l'hypothèse d'une distribution particulière des classes (ces approches sont dites génératives) ou en traçant une frontière de décision séparant les classes en présence (ces approches sont dites discriminantes). L'obtention de la meilleure fonction de décision possible par cette approche peut être difficile pour un problème donné, voire impossible si on veut traiter plusieurs problèmes avec le même classifieur. C'est la raison pour laquelle des méthodes combinant les décisions de plusieurs classifieurs sont utilisées pour augmenter le nombre de solutions possibles.

Plus précisément, deux constats principaux ont conduit à l'exploration des méthodes de combinaison de classifieurs [Kuncheva, 2007] : le premier vient du fait que les méthodes standards avec un seul classifieur ont été beaucoup développées dans la littérature et que l'on tend vers des approches complexes n'apportant pas de véritables améliorations par rapport aux méthodes existantes [Ho, 2002] ; le second constat est un point de vue statistique faisant remarquer qu'il est difficile de trouver le classifieur optimal pour un problème donné en se basant notamment sur un unique apprentissage à partir des échantillons disponibles [Dietterich, 2000a] (typiquement il est difficile de trouver une méthode surclassant toutes les autres sur tous les problèmes [5]).
Pour ce dernier constat, l'espace des hypothèses avec un seul classifieur élaborant la fonction de décision optimale est plus restreint que celui des hypothèses pouvant être formées à partir de la combinaison d'hypothèses existantes (répondant ainsi au premier constat invitant à utiliser les méthodes existantes plutôt que d'en façonner de nouvelles, au détriment de la complexité). En outre, l'algorithme d'apprentissage du classifieur unique peut faire converger la fonction de décision vers une solution locale plutôt que de se rapprocher de la solution optimale [Dietterich, 2000a] (une approche par combinaison de classifieurs est plus à même d'éviter ces convergences locales).

Nous présentons dans un premier temps les approches courantes de la littérature dans le cadre de l'apprentissage d'un classifieur unique. Dans un second temps, nous présentons les approches par combinaison de classifieurs. Nous insistons particulièrement sur les méthodes d'ensembles qui sont une approche de combinaison n'utilisant que des classifieurs de même type, indépendants les uns des autres. Nous développons dans ce cadre les méthodes d'ensembles d'arbres de décision avec la famille des forêts aléatoires [Breiman, 2001] qui ont largement été adoptées en raison de leurs performances, leur généricité et leur simplicité d'induction et que nous allons utiliser aussi dans nos travaux.

1.4.1 Les classifieurs

Nous développons dans cette section les méthodes d'apprentissage standards de la littérature. Nous présentons l'algorithme classique des K-Plus Proches Voisins (KPPV), l'algorithme standard d'arbre de décision "Classification And Regression Trees" (CART) proposée par [Breiman et al., 1984], l'algorithme en vogue C4.5, une amélioration de CART proposée par [Quinlan, 1993, Wu et al., 2008],

5. Cette observation rejoint une des implications du théorème "no free lunch" [Wolpert and Macready, 1997] stipulant qu'on ne saurait trouver un classifieur unique apte à traiter tous les problèmes.

l'algorithme "Support Vector Machine" ou "Séparateur à Vaste Marge" (SVM) [Cortes and Vapnik, 1995, Vapnik, 1998], de la famille des méthodes dites "à noyaux" constituant aujourd'hui une approche des plus compétitives de l'état de l'art et enfin les réseaux de neurones [Bishop, 1995].

K-Plus Proches Voisins

L'algorithme des K-Plus Proches Voisins [Blum and Langley, 1997] (KPPV) est une approche basée sur l'estimation de distance. Il n'a pas besoin d'apprentissage car le modèle de classification est directement le set d'apprentissage (ce classifieur est ainsi dit à mémoire). Lors de la phase de décision, l'algorithme évalue la classe majoritaire des K individus du set d'apprentissage les plus proches de l'exemple à tester. K représente un paramètre important de l'approche car il conditionne la façon dont le classifier va s'adapter aux données en présence, la résolution du classifieur et sa généricité. Si la valeur de K est faible, on obtient un classifieur dit de bonne résolution (très proche des données, avec ainsi un biais d'autant plus faible que K est petit), il arrive à définir des fonctions complexes entre les classes mais est très sensible aux bruits sur les données (le classifieur généralise mal d'où une variance élevée) ; si en revanche K est élevé, on obtient un classifieur qui lisse la frontière de décision donc peu sensible au bruit, avec une variance faible mais un biais élevé. Il est démontré qu'une façon de réduire l'erreur d'un classifieur est de diminuer à la fois son biais et sa variance. Cependant, il est difficile de réduire l'un sans augmenter l'autre. Nous voyons donc apparaître un compromis biais-variance couramment évoqué dans la littérature [Hastie et al., 2001]. Plusieurs résultats théoriques ont été énoncés sur la stabilité et la bonne convergence asymptotique de la méthode sous des conditions générales, validant ainsi la solidité de la méthode [Cover and Hart, 1967, Stone, 1977, Devijver, 1977, Belaid and Belaid, 1992, Biau et al., 2009]. Malgré sa simplicité de mise en œuvre, les performances de l'algorithme K-PPV sont encore mal connues et les coûts sont importants en stockage et temps de classification. En effet, un parcours exhaustif de tous les exemples d'apprentissage est nécessaire lors de la phase de test du classifieur. Il est notamment sujet à la malédiction de la dimension [Bellman, 1961, Beyer et al., 1999], i.e. des difficultés à classer lorsque la dimension de l'espace de description devient trop importante par rapport au nombre de données disponibles. Plusieurs variantes et améliorations ont été proposées afin notamment de réduire l'influence de la malédiction de la dimension et améliorer ses performances. Ainsi nous retrouvons le Plus Proche Voisin Adaptatif Discriminant [Short and Fukunaga, 1981, Hastie et al., 2005] utilisant une ellipsoïde pour l'estimation du plus proche voisin, tenant compte ainsi de la répartition directionnelle des données d'apprentissage.

Arbre de décision

L'arbre de décision est une méthode de classification popularisée par Breiman et al. via la monographie [Breiman et al., 1984] dans laquelle l'algorithme "Classification And Regression Trees" (CART) a été introduit. L'algorithme est jusqu'à présent très populaire dans divers domaines avec des applications comme la détection de tumeurs musculaires [Decaestecker et al., 1998], la détection de tuberculose pulmonaire [Mello et al., 2006], la classification de l'émotion [Lee et al., 2009], la reconnaissance d'activités physiques réalisées par un individu [Bao and Intille, 2004] ou encore la classification de types d'occupation de sols dans le domaine de l'imagerie par satellite [Mishra et al., 2011]. L'approche a connu plusieurs améliorations tant au niveau de la complexité de l'algorithme que de sa généricité avec notamment les différentes contributions de [Quinlan, 1986, Quinlan, 1993, Quinlan, 1996b, Rakotomalala, 2005]. Il s'agit d'une méthode de classification qui décompose le domaine d'apprentissage en des sous-domaines les plus purs possibles, i.e. contenant une population la plus homogène possible en terme de classes, en utilisant une structure hiérarchique séquentielle partitionnant de façon récursive l'espace de caractéristiques [Murthy, 1996].

La méthode de construction de l'arbre (on parle alors d'induction de l'arbre) repose sur les étapes suivantes [Breiman et al., 1984] :

– le choix d'un attribut de partitionnement en chacun des nœuds de l'arbre

– le choix d'un point de coupure pour la variable sélectionnée précédemment
– le choix d'une mesure de score évaluant la qualité des coupures choisies afin de déterminer le meilleur partitionnement
– le choix d'un critère d'arrêt de partitionnement
– le choix d'une règle d'assignation d'une classe à un nœud terminal ou feuille.

Nous illustrons dans la Figure 1.7 et la Figure 1.8, un exemple de construction d'arbre de décision pour le problème classique de classification de variétés d'Iris proposé par [Fisher, 1936] et dont la base de données, dont nous représentons un échantillon dans le Tableau 1.1, est disponible dans le répertoire de l'UCI [Blake and Merz, 1998b]. L'arbre est induit sur 66% des données (100 exemples), les 33% restant (50 exemples) servant à évaluer ses performances. Il s'agit d'un problème à 4 dimensions (longueur et largeur des pétales et des sépales des iris), 3 classes (Iris setosa, versicolor et virginica) avec la classe setosa linéairement séparable des deux autres (voir Figure 1.8). Dans la Figure 1.8, nous voyons les exemples d'iris projetés dans l'espace des deux premiers attributs de partitionnement choisis pour l'induction de l'arbre.

TABLE 1.1 – Tableau des données de la base Iris formée des trois variétés I. setosa, I. versicolor et I. virginica [Fisher, 1936] ; l'espace de caractéristiques est de 4 dimensions dont les attributs sont la longueur de sépale (a_1), la largeur de sépale (a_2), la longueur de pétale (a_3) et la largeur de pétale (a_4).



	I. setosa				**I. versicolor**				**I. virginica**			
#	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
1	5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
2	4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4	4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
6	5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
7	4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
8	5	3.4	1.5	0.2	4.9	2.4	3.3	1	7.3	2.9	6.3	1.8
9	4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
10	4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
11	5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
12	4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
13	4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
14	4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
15	5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
...	...	...	...	...	...	...	...	...	...	...	...	...
45	5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
46	4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
47	5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
48	4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
49	5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
50	5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Beaucoup de travaux ont été menés sur l'optimisation de l'arbre de décision en travaillant à plusieurs niveaux [Breslow and Aha, 1997, Murthy, 1998, Brostaux, 2005] : la réduction de la complexité de l'arbre avec des algorithmes d'élagage ayant pour objectif de simplifier le clas-

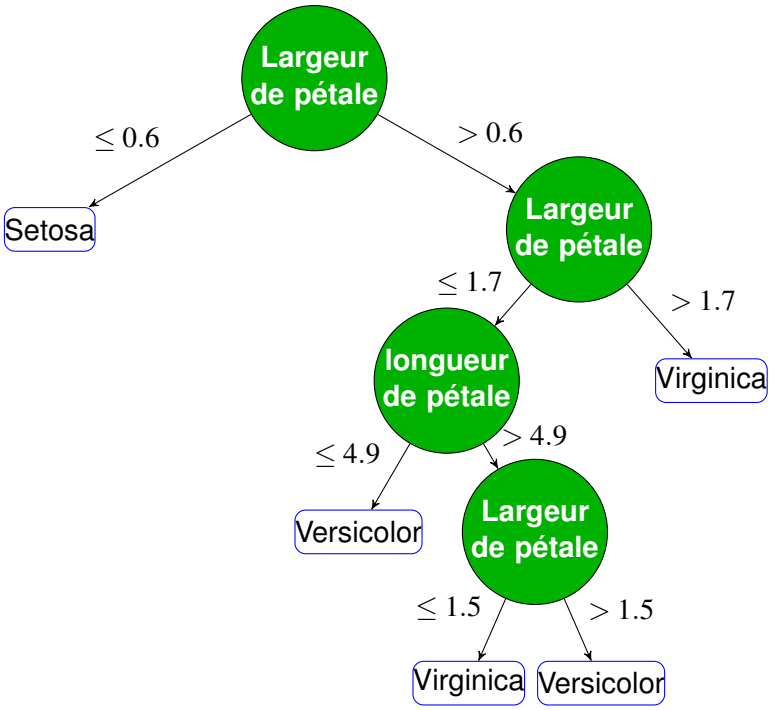


FIGURE 1.7 – Illustration de la construction d'un arbre de décision à partir des données de la base Iris [Fisher, 1936] ; on observe qu'avec cette règle de partitionnement, seules la longueur et la largeur des pétales suffisent à partitionner l'espace avec les trois variétés d'Iris.

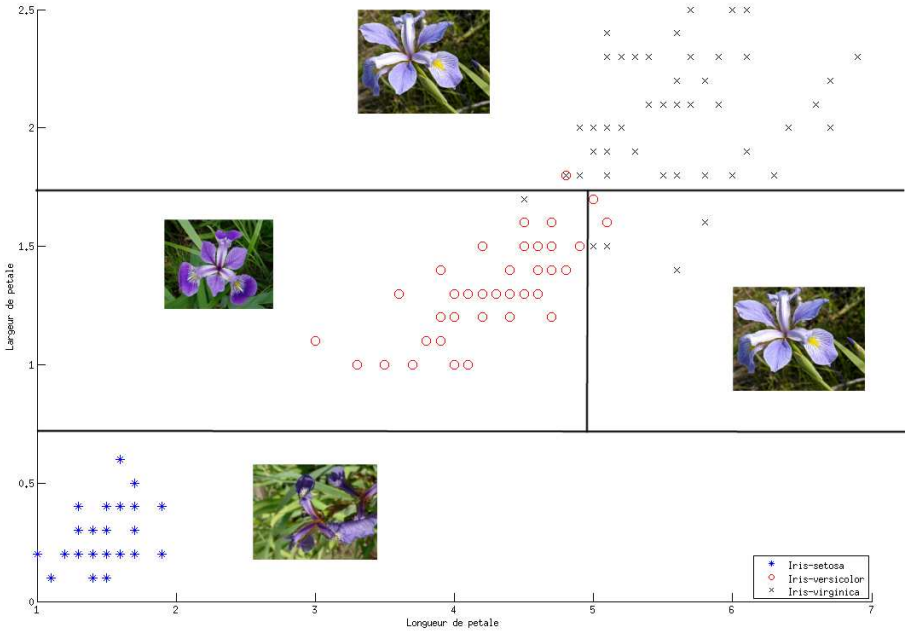


FIGURE 1.8 – Illustration du partitionnement de l'espace (formé par les deux attributs longueur de pétale -en abscisse et largeur de pétale -en ordonnée) effectué par l'arbre de la Figure 1.7.

sifieur afin d'augmenter sa capacité à généraliser ; la méthode de partitionnement des données tant au niveau du choix des attributs et des valeurs de coupure qu'au niveau du choix de la mesure de score [Quinlan, 1987, Mingers, 1989b, Mingers, 1989a, Lerman and Da Costa, 1996, Breslow and Aha, 1997, Murthy, 1998] ; la prise de décision en un nœud terminal [Brostaux, 2005].

Dans leur monographie, [Breiman et al., 1984] affirmaient déjà que les performances d'un arbre de décision reposaient principalement sur la détermination de sa profondeur. On reproche en effet aux arbres de produire des classifieurs trop proches des données (l'algorithme apprend parfaitement le set d'apprentissage). Les auteurs font ainsi remarquer que si la taille de l'arbre est trop grande, l'algorithme risque de faire du sur-apprentissage et donc aura de mauvaises performances sur des données nouvelles ; en revanche si l'arbre n'est pas assez profond, il ne sera pas assez discriminant. C'est la raison pour laquelle les auteurs préconisent l'élagage de l'arbre afin d'éviter que ce dernier ne se développe complètement et donc ne sur-apprenne. CART détermine ainsi une profondeur optimale par évaluation des performances de plusieurs "sous-arbres" obtenus par post-élagages à différents niveaux (i.e. l'arbre est complètement développé pour que ses feuilles ne comportent qu'une seule classe, puis est élagué récursivement). Les nœuds de l'arbre sont itérativement éliminés tant que les résultats obtenus ne sont pas significativement détériorés. En ce qui concerne la mesure de score, CART utilise l'indice de Gini évaluant la qualité de la réduction de l'impureté ou gain d'impureté au nœud courant. La mesure d'impureté au nœud courant est donnée par l'expression suivante :

$$I(noeud) = \sum_{j=1}^{c} \frac{n_{j.}}{n_{..}} \cdot (1 - \frac{n_{j.}}{n_{..}}) \tag{1.1}$$

où c est le nombre de classes en présence, $n_{j.}$ est l'effectif de la classe j au nœud courant, $n_{..}$ l'effectif total.
Le gain d'impureté est alors donné par la différence entre l'impureté $I(noeud)$ au nœud courant calculée précédemment et la somme de celle des nœuds fils engendrés, c'est à dire :

$$gain(A) = \Delta I = I(noeud) - \sum_{i=1}^{p} \frac{n_{.i}}{n_{..}} I(noeud_i) \tag{1.2}$$

où A est l'attribut de partitionnement utilisé, p le nombre de nœuds fils engendrés. L'affectation de la décision d'une classe pour une feuille se fait par la règle de la majorité. Cependant, de façon plus générale, une règle tenant compte du coût de mauvaise affectation est à privilégier [Rakotomalala, 2005].
Parmi les algorithmes d'arbres de décision aujourd'hui populaires, on retrouve, outre le CART, le C4.5 [6] qui est une amélioration de l'algorithme "Iterative Dichotomizer" (ID3) proposé par Quinlan [Quinlan, 1986] permettant de traiter davantage de problèmes (l'algorithme ne se limite plus aux seuls tests binaires) et apportant notamment une nouvelle mesure de score, le gain d'information, basée sur la quantité d'information initiale. Cette quantité d'information est évaluée par l'entropie de Shannon au nœud courant :

$$I(noeud) = \sum_{j=1..c} -P_j log_2 P_j \tag{1.3}$$

où c est le nombre de classes en présence, $P_j = \frac{n_{j.}}{n_{..}}$ est la fréquence observée de la classe j au nœud courant, n_{ji} l'effectif de la classe j au nœud fils i, $n_{j.}$ l'effectif de la classe j (somme des effectifs de la classe j de tous les nœuds fils), $n_{..}$ étant l'effectif total (somme des effectifs de toutes les classes.
Le gain d'information pour un attribut de partitionnement A est alors donné par la différence entre la quantité d'information initiale au nœud courant et la somme des quantités d'information des nœuds fils obtenus après un partitionnement donné :

$$gain(A) = \Delta I = I(noeud) - \sum_{i=1}^{p} \frac{n_{.i}}{n_{..}} I(noeud_i) \tag{1.4}$$

6. Le C5.0 améliore grandement les performances du C4.5 mais est utilisé dans un cadre commercial et les sources de cette version ne sont pas disponibles au grand public : "Data Mining Tools See5 and C5.0. RULEQUEST RESEARCH, St. Ives, NSW, Australia. http://www.rulequest.com/see5-info.html"

Ce gain a cependant tendance à privilégier les attributs ayant un grand nombre de valeurs possibles. Un terme de pondération $IV(A)$ est alors introduit apportée dans C4.5 par rapport à ID3 corrigeant ce biais et donnant ainsi le $gain_{ratio}$ dont l'expression est :

$$gain_{ratio}(A) = \frac{gain(A)}{IV(A)} \tag{1.5}$$

avec

$$IV(A) = \sum_{j=1}^{p} -\frac{n_{.i}}{n_{..}} log_2 \frac{n_{.i}}{n_{..}} \tag{1.6}$$

où p est le nombre de nœud fils engendrés lors du partitionnement, A étant l'attribut de partitionnement, $n_{.i}$ étant l'effectif du nœud fils i (en sommant les effectifs des classes de ce nœud). Une autre approche est la méthode "CHi-squared Automatic Interaction Detection" (CHAID) [Kass, 1980, Rakotomalala, 2005] permettant de traiter un grand volume de données. Elle propose notamment d'utiliser, pour évaluer la qualité de l'attribut de partitionnement, le χ^2 d'écart à l'indépendance, défini par :

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{p} \frac{(n_{ji} - \frac{n_{j.} \cdot n_{.i}}{n_{..}})}{\frac{n_{j.} \cdot n_{.i}}{n_{..}}} \tag{1.7}$$

où c désigne le nombre de classes, p le nombre de partitions, en reprenant les notations précédentes. Cependant, cette mesure, comme l'indice de Gini, avantage les attributs ayant de nombreuses modalités. C'est la raison pour laquelle la mesure t de Tschuprow est proposée en remplacement de la mesure originelle, défini par $t = \chi^2/(n\sqrt{(c-1)(p-1)})$.
On peut citer une dernière approche, l'arbre oblique, qui utilise une combinaison linéaire des attributs pour élaborer une droite de partitionnement plutôt qu'une coupure standard parallèle aux axes de l'espace [Murthy et al., 1994].

L'arbre de décision est très populaire du fait de sa rapidité, sa simplicité algorithmique et de l'interprétation transparente des règles de décision générées. De plus, le classifieur est à même de choisir les variables pertinentes pour la classification automatiquement à l'aide des sélections de variables faites lors de la construction de l'arbre [Hastie et al., 2001]. Cependant, c'est un classifieur instable : les règles de décision sont totalement modifiées lorsqu'un changement mineur intervient dans les données d'apprentissage. De plus, durant ces dernières années, les arbres de décision en tant que classifieur individuel n'ont pas connu d'avancées significatives en termes de performances par rapport aux algorithmes de référence CART, CHAID, C4.5.

Nous verrons dans la section suivante que les méthodes de combinaison de classifieurs offrent de nouvelles perspectives d'utilisation des arbres de décision et que notamment, l'inconvénient de l'instabilité du classifieur devient un avantage certain dans le cadre de ces méthodes en permettant particulièrement un gain de performances important.

SVM

Le "Support Vector Machine" ou "Séparateur à Vaste Marge" (SVM) [Cortes and Vapnik, 1995, Vapnik, 1998] est un classifieur discriminant paramétrique établissant un hyperplan séparateur de marge maximale entre les exemples représentants de chacune des classes des données d'apprentissage.
Sur la Figure 1.9, nous illustrons un problème de classification binaire en 2 dimensions. Il existe une infinité de droites séparatrices des deux classes en présence mais une seule maximise la distance aux exemples aux frontières des classes (i.e. maximise la distance z_2 à l'hyperplan) . Ces exemples frontières sont appelés vecteurs de support.

Le SVM est l'un des algorithmes les plus cités dans la littérature en raison de ses performances, de sa généricité et de ses fondations théoriques [Shawe-Taylor and Cristianini, 2004, Rakotomamonjy, 2003].

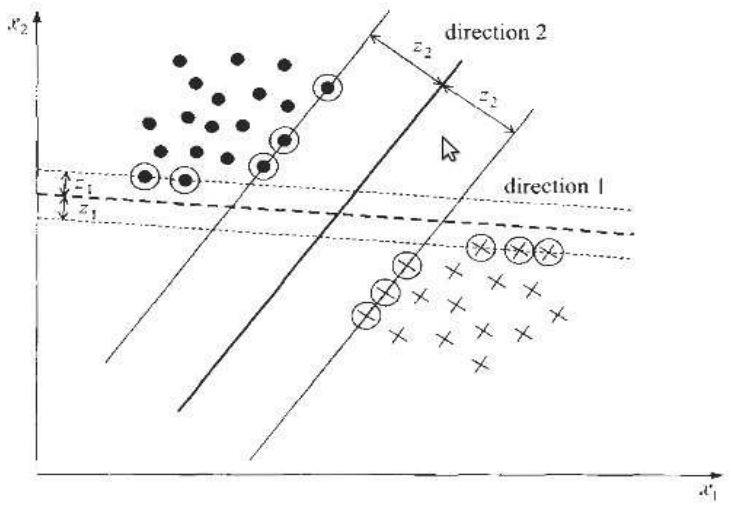


FIGURE 1.9 – Hyperplan séparateur de marge maximale ; deux jeux d'hyperplan, la direction 2 a la marge la plus forte ; les vecteurs de supports sont ceux situés sur les droites frontières.

L'hyperplan séparateur de paramètres $(\mathbf{w}, w_0)$ a pour équation :

$$H : h(x) = \mathbf{w}^t \cdot x + w_0$$

Il s'agit alors de maximiser la distance d'un point x à l'hyperplan H, donnée par :

$$dist(x,H) = \frac{|h(x)|}{||\mathbf{w}||}$$

où h est la fonction de projection sur H, $||\cdot||$ étant la norme euclidienne. Pour obtenir l'équation de l'hyperplan de marge maximale, on minimise alors la fonction objectif J :

$$J(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2$$

sous les contraintes :

$$u_i \cdot h(x_i) \geq 1, i = 1..n, u_i \in \{-1,1\} \text{ étant la classe des } x_i$$

Il s'agit alors d'un problème d'optimisation quadratique non linéaire avec une contrainte d'inégalité linéaire et dont la résolution fait appel aux multiplicateurs de Lagrange dans un espace équivalent dual :

$$\begin{cases} max_\lambda (\sum_{i=1}^n \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j u_i u_j x_i^t x_j) \\ \sum_{i=1}^n \lambda_i u_i = 0, \lambda \geq 0 \end{cases}$$

les λ_i étant les multiplicateurs de Lagrange. Il s'agit d'un problème de programmation quadratique convexe pour lequel des solutions algorithmiques sont disponibles [Canu et al., 2005, Chang and Lin, 2001, Platt, 1999]. On obtient alors :

$$\begin{cases} \mathbf{w} = \sum_{i=1}^n \lambda_i u_i x_i \\ w_0 = 1 - \mathbf{w}^t x_i, \, x_i \text{ étant n'importe quel vecteur de support, avec } h(x_i) = 1 \end{cases}$$

L'équation de l'hyperplan est alors :

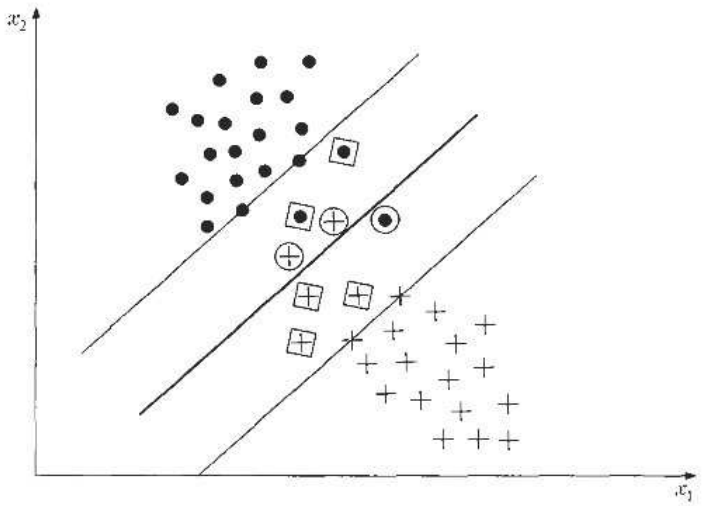


FIGURE 1.10 – La séparation linéaire n'est pas adaptée à ce problème ; les classes se chevauchent ; une représentation de ces données dans un espace adapté, à l'aide notamment de fonctions noyaux, est alors nécessaire pour obtenir une séparation linéaire

$$h(x) = \sum_{i=1}^{n} \lambda_i u_i x_i x + w_0$$

et la fonction de décision du SVM est $f(x) = signe(h(x))$. Il est à noter que les solutions trouvées ici contiennent des λ_i pouvant être nuls et ceux qui ne sont pas nuls sont ceux associés aux vecteurs de supports x_i. Ainsi, l'équation de l'hyperplan séparateur de marge maximale ne dépend que des vecteurs de support, soit généralement une faible portion du set d'apprentissage. Cette remarque est importante car elle montre que le SVM n'a pas besoin de l'intégralité des données d'apprentissage pour élaborer sa fonction de décision.
On a supposé dans nos calculs que les classes étaient linéairement séparables. En général ce n'est pas le cas comme nous le montre la Figure 1.10 où les données des deux classes ne sont pas linéairement séparables. Il est alors nécessaire de projeter ces données dans des espaces adaptés, de plus grande dimension et dans lesquels une séparation linéaire est possible.

On introduit pour cela des fonctions noyaux. Il s'agit de formes linéaires, plus complexes que le produit scalaire et utilisées en lieu et place du produit scalaire usuel que nous retrouvons dans l'équation de l'hyperplan qui devient alors :

$$h(x) = \sum_{i=1}^{n} \lambda_i u_i K(x_i, x) + w_0$$

Plusieurs noyaux sont utilisés dans la littérature. Le choix dépend essentiellement de l'application. On retrouve cependant couramment les noyaux linéaires, polynomiaux et "Radial Basis Function" (RBF) comme le noyau gaussien [Ben-Hur and Weston, 2010, Hastie et al., 2001]. L'utilisation de ces noyaux est illustrée dans la Figure 1.11.

Les réseaux de neurones

Les réseaux de neurones ont été proposés initialement comme modèle mathématique du fonctionnement connu des neurones physiologiques dès 1949. L'unité élémentaire du réseau est le neurone formel et dont le modèle le plus simple est le perceptron linéaire [Rosenblatt, 1957]. La sortie du réseau est évaluée par rapport à un seuil appliqué sur la somme pondérée des coefficients (potentiel post-synaptique) appliqués aux données en entrée :

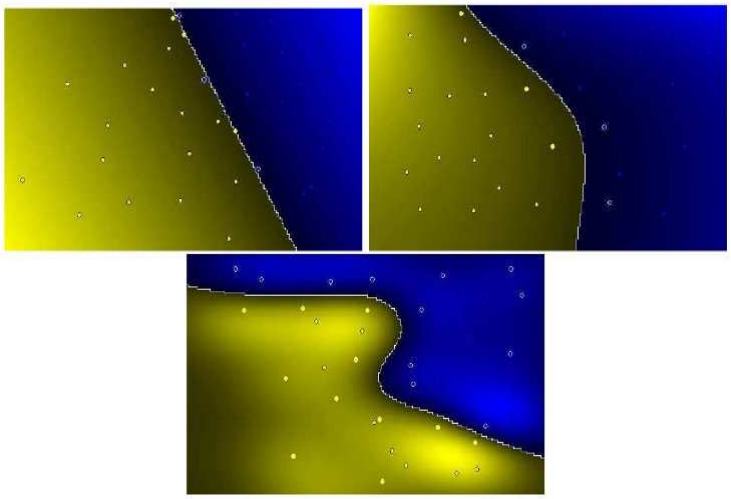


FIGURE 1.11 – Fonction de décision du SVM dans le cas de différentes fonctions noyaux : noyau linéaire, polynomial et RBF (source de l'image [Chatelain, 2006])

$$d_w(x) = \mathbf{w}^t x$$

$$sortie = f(d_w) = \begin{cases} +1 \text{ si } d_w > 0 \\ -1 \text{ sinon} \end{cases}$$

f est la fonction d'activation du neurone et peut être la fonction de Heaviside précédente, une sigmoïde ou une tangente hyperbolique. Il suffit que f soit une fonction à valeur réelle, non décroissante, bornée. On peut remarquer que l'équation :

$$d_w(x) = 0$$

définit un hyperplan de paramètre **w**, vecteur contenant les poids (les coefficients synaptiques) du réseau à déterminer durant l'apprentissage de ce dernier. Nous distinguons alors deux approches selon que les classes d'apprentissage sont linéairement séparables ou non.

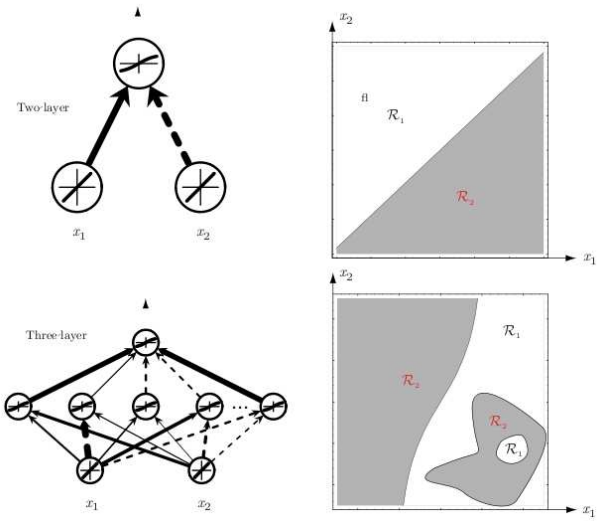


FIGURE 1.12 – Schéma d'un réseau de neurones à deux couches (haut) et à trois couches (bas) avec les surfaces de décision associées [Duda et al., 2000]

En général les classes ne sont pas linéairement séparables. On ajoute alors au réseau une ou plusieurs couches qui prennent elles-mêmes en entrée les sorties des couches précédentes (voir Figure 1.12). L'ajout de couches au réseau complexifie la frontière de décision lui permettant de traiter les cas non-linéaires. Des techniques d'apprentissage comme la rétro-propagation permettent de minimiser l'erreur quadratique empirique du réseau. L'algorithme de rétro-propagation propage dans le réseau, de la couche de sortie vers la couche d'entrée, l'erreur commise sur les exemples d'apprentissage et par le biais d'itérations successives permet d'optimiser le choix des pondérations. Le choix du nombre de couches, de nœuds demeure un problème actuel. En effet le dimensionnement du réseau est à figer avant la phase d'apprentissage des pondérations ; surtout, aucune règle ne permet de déterminer le nombre de neurones adéquat dans la couche cachée. L'heuristique communément utilisée est de considérer qu'un trop grand nombre de nœuds nuit aux performances en généralisation du modèle et trop peu de nœuds ne permet pas de tracer des frontières suffisamment flexibles pour apprendre les données du problème. Le contrôle du nombre d'itérations apparaît aussi comme essentiel dans la mesure où le réseau a tendance à sur-apprendre les exemples d'apprentissage [Bishop, 1995]. Il se pose également le problème des minima locaux vers lesquels peut converger la méthode de rétro-propagation.

Conclusion

L'approche standard de classification avec un classifieur unique est un domaine mâture où il est devenu difficile d'innover sans augmenter la complexité des algorithmes existants. Nous avons vu que l'arbre de décision est une méthode qui a suscité beaucoup d'intérêts tant au niveau théorique qu'expérimental. Beaucoup d'études ont été effectuées depuis l'introduction de l'approche CART. Ces récentes années ont vu l'émergence de méthodes nouvelles combinant des classifieurs existants. Particulièrement, les combinaisons d'arbres de décision sont devenues très populaires en raison des réponses favorables apportées aux inconvénients du classifieur individuel avec notamment un gain en performances et en stabilité.

Nous présentons dans la section suivante les approches de combinaison de classifieurs.

1.4.2 Méthodes d'ensemble de classifieurs

Dans la section précédente, nous avons discuté des approches de classification composées d'un classifieur unique. Des études menées essentiellement à la fin des années 1990 et plus abondamment par la suite ont montré que combiner plusieurs classifieurs pouvait apporter des améliorations en termes de flexibilité, de performances et de stabilité par rapport aux approches avec un classifieur unique [Dietterich and Bakiri, 1995, Dietterich, 1997, Dietterich, 2000a, Dietterich, 2002, Kuncheva, 2007, Banfield et al., 2007]. Plusieurs principes de combinaison ont alors été proposés agissant tant au niveau des données, des attributs, des étiquettes ou des classifieurs de la combinaison. On retrouve notamment les approches de combinaisons séquentielles et les approches de combinaisons parallèles.

Dans la combinaison séquentielle, la sortie de chaque classifieur individuel va influer sur l'induction du classifieur qui le suit dans la chaîne de traitement. L'ordre de placement de chacun des classifieurs est donc essentiel. La méthode de dopage ou "boosting" introduite par Freund et Schapire [Freund and Schapire, 1995, Freund and Schapire, 1996] met en œuvre l'énoncé indiquant qu'il est possible de convertir une règle de décision faible (e.g. un classifieur à peine plus performant que l'aléatoire) en une règle performante [Schapire, 1990]. Typiquement, les données d'un classifieur sont modifiées afin de mettre en avant les données mal classées par le classifieur qui le précède. Ainsi, le classifieur courant se spécialise sur les données difficiles en étant plus performant que ses prédécesseurs sur ce set. L'algorithme Adaboost est l'un des exemples les plus populaires de classifieur entrainé par dopage et a été utilisé notamment pour la détection des objets et des visages [Viola and Jones, 2001, Viola and Jones, 2004]. C'est une méthode largement utilisée pour le type d'architecture séquentielle.

Dans la combinaison parallèle, des classifieurs de même type sont construits en parallèle, indépendamment les uns des autres. Il est généralement sous-entendu un grand nombre de classifieurs ou "pool" de classifieurs. Dans ce cadre, il est montré que la diversité est un facteur important dans la construction de méthodes d'ensembles performantes [Dietterich, 2000b, Kuncheva, 2007, Brown, 2009, Brown and Kuncheva, 2010]. La création de cette diversité est notamment possible avec l'injection d'aléatoire tant dans le choix des données d'apprentissage que dans celui des attributs ou des paramètres d'induction des classifieurs. Des mécanismes de combinaison ont ainsi émergé travaillant à ces différents niveaux. On peut citer le Bagging [Breiman, 1996] injectant de l'aléatoire dans les données en sous-échantillonnant le set d'apprentissage ; le Random Subspace Method (RSM) [Ho, 1998] ou le Random Feature Selection (RFS) [Amit and Geman, 1997, Dietterich, 2000b, Breiman, 2001] travaillant au niveau des attributs en formant des sous-espaces de description complètement aléatoires. Dans le cas particulier de RSM ou RFS, la dimension des sous-espaces résultants peut être assez petite pour traiter favorablement les problèmes de grande dimension. Cependant, le gain de diversité peut être au détriment des performances des classifieurs, comme le fait remarquer [Brown and Kuncheva, 2010]. Parmi les méthodes d'ensemble, nous distinguons particulièrement la méthode d'ensemble d'arbres de décision, les forets aléatoires présentées par Breiman en 2001 [Breiman, 2001]. Dans cette famille de méthodes, différents mécanismes de randomisation interviennent comme le Bagging, le RSM ou le RFS. L'intérêt pour cette famille de méthodes a augmenté considérablement en raison des fondations théoriques solides dont elle a pu bénéficier avec les travaux fondateurs de [Dietterich, 1998, Ho, 1998, Dietterich, 2000b, Breiman, 2001] et ceux de [Robnik-Sikonja, 2004, Geurts et al., 2006, Genuer et al., 2008, Biau, 2010] tant au niveau expérimental que théorique.

Nous présentons ci-après ces mécanismes de randomisation et particulièrement la famille des forêts aléatoires.

1.4.2.1 Les mécanismes de randomisation

Le bagging

Le "bagging" a été introduit par Breiman en 1996 [Breiman, 1996] comme un algorithme réducteur de variance. Le Bagging signifiant "Bootstrap-AGGregatING" est un mécanisme de randomisation consistant à former différents jeux de données bootstrap et apprendre un pool de classifieurs individuels, chacun étant construit sur un jeu bootsrap distinct. Le prédicteur obtenu est alors une combinaison des prédicteurs du pool. Le jeu bootstrap est obtenu à partir d'un échantillonnage par tirage aléatoire avec remise du set initial des données d'apprentissage. Chaque set boostrap a alors le même effectif que le set initial. Ainsi, dans une combinaison parallèle, chaque classifieur individuel pourra bénéficier d'un jeu de données différent issu de la même distribution que le set initial. Statistiquement, lorsque le nombre n de sets boostrap est grand, on montre que $1-(1-\frac{1}{n})^n \sim 1-e^{-1}$, soit 62.3% de l'information initiale est communiquée aux classifieurs individuels. On retrouve l'utilisation du bagging particulièrement dans les forêts aléatoires [Breiman, 2001] que nous traitons dans la suite de l'exposé.L'approche par bagging a montré sa capacité à diminuer l'erreur en généralisation de classifieurs individuels de type arbre de décision, classifieur Bayes naïf, réseaux de neurones [Bauer and Kohavi, 1999, Breiman, 1996, Dietterich, 2000a, Drucker, 1997, Quinlan, 1996a, Schapire et al., 1998]. Des améliorations systématiques sont constatées notamment lorsque le classifieur individuel est instable [Breiman, 1996, Poggio et al., 2004, Elisseeff et al., 2006, Bousquet and Elisseeff, 2002]. Un classifieur est dit instable lorsqu'une faible perturbation du set d'apprentissage entraine des modifications importantes dans la structure du classifieur. Il est notamment montré que le bagging permet de rendre stable un classifieur initialement instable [Grandvalet, 2006, Elisseeff et al., 2006].

Le Random Subspace Method

Le Random Subspaces Method (RSM) est introduit par Ho et al. [Ho, 1998] comme une méthode d'ensemble. Elle consiste à projeter les données d'apprentissage dans des sous-espaces de même dimension m, choisis complètement aléatoirement et à construire des classifieurs individuels sur chacun de ces sous-espaces. L'auteur a alors présenté le classifieur "Random Trees" basé sur cette approche où le classifieur individuel de l'ensemble est un arbre de décision. La dimension des sous-espaces créés est un paramètre de la méthode et est évalué empiriquement à $m \approx \frac{d}{2}$ où d est la dimension de l'espace d'origine.

L'un des apports majeurs du RSM est la réduction drastique de la dimensionalité d'un problème donné, tout en demeurant très simple d'utilisation (e.g. pas de techniques de sélection de variables). Il permet alors de traiter les problématiques en grande dimension et permet de réduire la complexité des classifieurs. L'auteur précise d'ailleurs que la méthode est particulièrement recommandée dans le cas de grands volumes de données, de grands espaces de caractéristiques et dans le cas d'attributs redondants.

Le Random Feature Selection

Le Random Feature Selection (RFS) [Amit and Geman, 1997, Dietterich, 2000b, Breiman, 2001] a été introduit dans le cadre de l'induction d'arbres de décision et utilisé pour former des ensembles d'arbres aléatoires. Elle consiste à sélectionner aléatoirement un sous-ensemble de l'espace de caractéristiques initial en chaque nœud de l'arbre lors de son induction. Tout comme le RSM vu précédemment, le RFS permet de réduire drastiquement la dimensionalité d'un problème donné. La dimension des sous-espaces crées est un paramètre de la méthode. De nombreuses études ont été menées pour déterminer la valeur optimale de ce paramètre. Plusieurs valeurs empiriques ont alors été proposées sans qu'il ne se dégage de valeur universelle.

Nous parlons plus spécifiquement dans la section suivante des algorithmes basés sur des ensembles d'arbres de décision.

1.4.2.2 Les forêts aléatoires

Les forêts aléatoires constituent une famille de méthodes d'ensembles de classifieurs à base d'arbres de décision. Le terme et le formalisme des forêts aléatoires ont été proposés par Breiman [Breiman, 2001]. Ces méthodes utilisent les mécanismes de randomisation afin de produire un set de classifieurs individuels les plus diversifiés possibles. Le bagging [Breiman, 1996] ou le Random Subspace Method [Ho, 1998] peuvent être cités comme des principes forts de randomisation sur lesquels se base l'induction des forêts aléatoires. Les forêts aléatoires sont un principe général de combinaison de L classifieurs à base d'arbres $h(x,\Theta_i), i = 1,...L$ où Θ_i est une famille de vecteurs aléatoires indépendants et identiquement distribués et $\mathbf{x}$ la donnée d'entrée. Ainsi chaque classifieur individuel est induit à partir d'un vecteur aléatoire de paramètres.

Une des propriétés importantes des forêts est leur convergence avec un nombre suffisant d'arbres ; elles évitent donc le sur-apprentissage. De plus, elles sont capables de traiter naturellement les problèmes en grande dimension en se focalisant notamment sur les variables importantes du problème [Genuer et al., 2008, Biau, 2010].

Les forêts ont suscité beaucoup de travaux tant au niveau théorique qu'expérimental [Biau, 2010, Geurts et al., 2006, Genuer et al., 2008, Robnik-Sikonja, 2004, Bernard et al., 2008, Breiman, 2001] afin de comprendre notamment l'origine de leurs performances et contrôler leur comportement. Dans la plupart de ces travaux, les forêts apparaissent comme l'une des méthodes les plus efficaces pour une grande majorité de problèmes et notamment compétitive avec les algorithmes Adaboost et SVM [Robnik-Sikonja, 2004, Cutler and Zhao, 2001, Rodriguez et al., 2006]. Les différentes approches issues de la famille des forêts aléatoires tirent partie de la flexibilité de l'algorithme au niveau du choix effectué tant au niveau de l'échantillonnage des données que de celui des caractéristiques.

Nous développons ci-après les algorithmes phares de la famille des forêts aléatoires, à savoir les forest-Random Input (forest-RI) et les Extremly Randomized Trees (extra-trees).

Forests-Random Inputs

L'un des premiers algorithmes proposé par Breiman est Forests-RI (Random Forests with Random Inputs [7]) [Breiman, 2001] où deux mécanismes de randomisation sont utilisés : le bagging pour manipuler les données du set d'apprentissage et le RFS pour sous-échantillonner en chaque nœud des arbres les caractéristiques disponibles. Le bagging a la particularité de fournir un jeu de validation, différent du jeu d'apprentissage bootstrap appelé "out-of-bag", constitué des données non sélectionnées lors du tirage aléatoire avec remise. Ce set permet notamment de mesurer les performances du système durant la phase d'apprentissage.

En chaque nœud des arbres de la forêt, la variable de partitionnement est choisie aléatoirement parmi K_{RFS} attributs tirés aléatoirement puis un point de coupure optimal sur cette variable est déterminé après un parcours exhaustif des coupures possibles. La valeur de K_{RFS} contrôle le degré d'aléatoire utilisé lors du partitionnement. En effet, de faibles valeurs de K_{RFS} vont avoir pour effet de rendre la randomisation plus forte en rendant ainsi la structure de l'arbre moins dépendante des valeurs des classes du set d'apprentissage.

En revanche, en prenant des valeurs de K_{RFS} plus grandes, l'étape d'optimisation de score va orienter le partitionnement de l'arbre vers une meilleure séparation des données des classes en chaque nœud. L'algorithme a ainsi la possibilité, à cette étape, de filtrer les variables non pertinentes pour la discrimination des classes. La valeur de K_{RFS} communément adoptée dans la littérature est la valeur empirique $K_{RFS} = \sqrt{M}$, où M est la dimension de l'espace de caractéristiques. Des études plus approfondies ont été menées sur le choix de K_{RFS} mais elles concluent toutes qu'il n'y a pas de valeur universelle [Geurts et al., 2006, Bernard et al., 2008, Bernard et al., 2009].

Extremely Randomized Trees

Les Extremely Randomized Trees (Extra-Trees) sont des ensembles d'arbres aléatoires de la famille des forêts aléatoires, introduite par [Geurts et al., 2006], dans lesquelles les principes de randomisation sont renforcés. Contrairement à Forest-RI utilisant le bagging pour alimenter ses arbres, une forêt d'extra-trees utilise tout le set d'apprentissage. Les auteurs nt montré que cette approche permet de réduire le biais de l'ensemble. Les arbres extra-tree sont construits sur le même set d'apprentissage en utilisant des mécanismes fortement randomisés. En effet, en chaque nœud des arbres de la forêt, une variable de partitionnement est choisie aléatoirement et un point de coupure est défini sur cette variable de façon aléatoire également. La procédure est répétée K fois et le critère de partitionnement choisi est alors associé au point de coupure maximisant un score qui est une variante normalisée du gain d'information de Shannon [Wehenkel, 1998]. L'exemple de $K = 1$ montre que l'étape de maximisation de score n'intervient plus dans le choix du partitionnement qui est complètement guidé par l'aléatoire. L'arbre formé est d'ailleurs appelé par les auteurs "Totally Randomized Tree".
Concernant le compromis biais-variance, les extra-trees ont la propriété de réduire la variance plus fortement que la méthode Forest-RI du fait de la randomisation plus forte injectée lors de l'induction. L'algorithme diminue également le biais en se servant de la totalité du set d'apprentissage au lieu d'un échantillon bootstrap. Comme il n'y a pas de recherche exhaustive du meilleur point de coupure, les extra-trees ont un coût calculatoire moindre que les Forest-RI. Les expériences menées par les auteurs ont de plus montré que les extra-trees se comportaient de façon favorable par rapport aux algorithmes de la littérature [Marée et al., 2005].

De ces méthodes de l'état de l'art, nous constatons que les extra-trees constituent un compromis intéressant pour l'élaboration de notre système de classification. En effet, cette méthode de forêts aléatoires obtient les performances de l'état de l'art, se comparant notamment à l'algorithme standard Forest-RI, en étant générique et tout en ayant un coût de calcul faible par rapport aux

7. "Forêts aléatoires à variables d'entrées aléatoires"

autres méthodes notamment de la famille des forêts.

1.4.3 Conclusion

Nous avons introduit dans cette section les méthodes de combinaison de classifieurs, leurs avantages pour pousser plus loin les limites des performances des systèmes individuels comme les arbres de décision. Nous avons présenté notamment les méthodes d'ensembles d'arbres de décision avec les forêts aléatoires qui suscitent beaucoup d'intérêts tant au niveau des performances, des propriétés statistiques intéressantes et une implémentation parallélisable pour la diminution drastique des coûts de calculs. L'un des avantages de cette famille de méthodes est d'avoir des fondations statistiques solides avec notamment des résultats de convergence favorables. Elles sont pour cela considérées comme étant l'une des approches les plus performantes, génériques et robustes à ce jour.

1.5 Notre problématique : la classification des images alvéoscopiques

Notre problématique est la classification des images alvéoscopiques. Notre objectif est ainsi la mise en place d'un système de classification automatique pour ces images. Nous voyons après l'étude de l'état de l'art précédent que plusieurs approches se dégagent pour la constitution d'un tel système.
Nous avons notamment retenu la pertinence d'une approche générique, tant au niveau de la description des images qu'au niveau de la méthode d'apprentissage de la fonction de décision permettant d'identifier les cas pathologies et les cas sains. Ces choix sont notamment motivés par l'absence de définition et de sémiologie claire pour les cas pathologiques associés à ces images. En effet, la vérité terrain sur laquelle nous pouvons nous appuyer est l'état effectivement pathologique du patient et non les structures présentes dans l'image qui peuvent à de nombreuses reprises se rapprocher visuellement des structures de la classe de patients sains. Ces choix génériques sont également intéressants pour le praticien du fait qu'il est possible d'en déduire a posteriori les caractéristiques pertinentes ayant conduit à la résolution du problème.

A partir de cet état de l'art, nous proposons un premier système de classification des images alvéoscopiques. Notre système de classification est composé des deux étapes essentielles. Tout d'abord, nous avons l'étape d'extraction de caractéristiques. Nous avons vu que plusieurs approches étaient proposées dans la littérature : des approches bas niveau, de plus haut niveau, des approches globales ou locales. Nous avons fait le choix de l'approche d'extraction de fenêtres aléatoires de Marée et al. [Maree et al., 2005] en raison de sa capacité à travailler tant au niveau local qu'au niveau global dans l'image. En effet, les images alvéoscopiques présentent à la fois des micro-structures et des structures plus grandes. Le contenu informatif se trouve donc à différentes résolutions dans l'image. Ces images étant texturées, nous avons fait le choix de décrire les fenêtres extraites avec des extracteurs de haut niveau (plutôt que la valeur de l'intensité du pixel brut initialement proposée par l'auteur), parmi lesquels SIFT et LBP présentés dans ce chapitre comme étant performants et robustes tout en étant génériques. La résolution spatiale ici est importante car dans les images alvéoscopiques, des structures semblables à différentes résolutions sont distinctes. Une approche multi-résolution de ces différents extracteurs est envisagée afin de tenir compte des différents niveaux de détails présents dans les images.

En ce qui concerne la classification, nous avons montré que les méthodes d'ensembles s'imposent comme de véritables alternatives aux approches standards. Particulièrement, les méthodes d'ensemble d'arbres de décision sont largement adoptées en raison notamment de leur bonne gestion des problèmes en grande dimension, leur performance, leur moindre coût calculatoire et leur capacité à traiter efficacement des problèmes divers. De plus, ces méthodes, et typiquement la famille des forêts aléatoires, bénéficient de solides fondations théoriques statistiques et expérimentales avec notamment des résultats de convergence statistique [Breiman, 2001, Biau et al., 2008, Biau, 2010]. Ainsi, il est intéressant d'analyser les différents principes de randomisation possibles au travers de

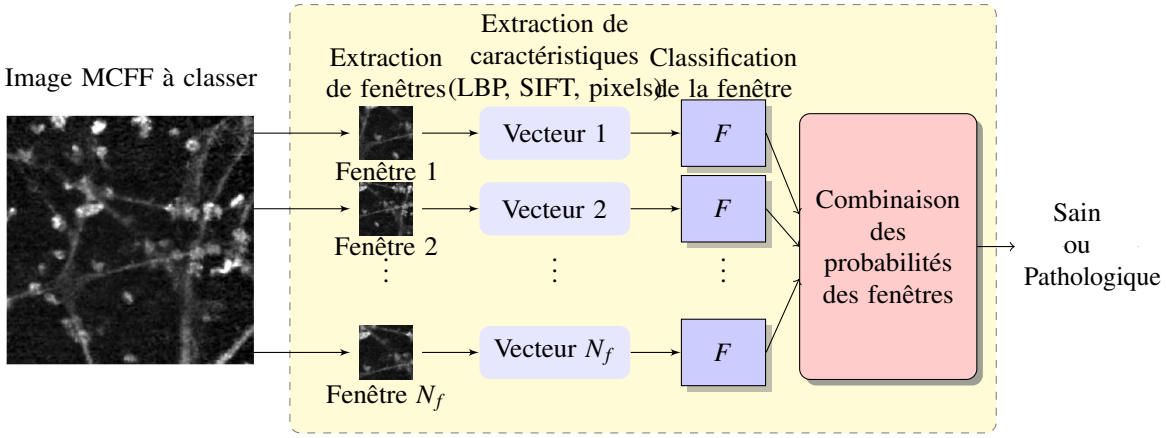


FIGURE 1.13 – Utilisation des sous-fenêtres dans la classification par extra-trees ; F désigne une forêt de L arbres de type extra-trees.

ces méthodes, permettant alors de contrôler notamment la flexibilité du modèle.

Nous proposons alors un système de classification tirant partie de ces différentes approches, tant au niveau de la phase d'extraction de caractéristiques localement dans l'image qu'au niveau de la phase d'apprentissage. Nous décrivons ce système dans la Figure 1.13.

1.6 Conclusion

Dans ce chapitre, nous avons abordé un état de l'art sur les systèmes de classification automatique d'images. Nous avons présenté les différentes méthodes couramment utilisées dans la littérature pour chacune des étapes composant le système de classification, à savoir notamment l'étape d'extraction de caractéristiques et l'étape d'apprentissage permettant d'établir la fonction de décision du classifieur. Nous avons noté la tendance grandissante de se tourner vers les méthodes de combinaison de classifieurs afin d'envisager des approches différentes des approches standards. Nous avons notamment insisté sur les récentes approches des méthodes d'ensembles d'arbres de décision avec la famille des forêts aléatoires qui possèdent une fondation solide tant au niveau théorique qu'au niveau expérimental. Ces approches sont en effet considérées comme l'une des plus performantes, génériques et robustes de la littérature.

Cet état de l'art nous a permis d'avoir un panorama des méthodes utilisées dans la littérature. En nous appuyant sur l'état de l'art, nous avons proposé en fin de chapitre, un système de classification pour les images alvéoscopiques, suivant des contraintes de généricité liées à l'application. Nous nous proposons, dans le chapitre suivant, de mettre en œuvre cette approche et de l'évaluer par rapport aux autres approches possibles de l'état de l'art. Nous détaillons les expérimentations menées dans le chapitre suivant.

Chapitre 2

Un système de classification des images alvéoscopiques

Sommaire

2.1 Introduction

2.1.1 Description du système

Notre système de classification est composé de quatre étapes : une étape d'extraction d'informations locales à l'aide de fenêtres aléatoires extraites dans les images ; une étape de caractérisation riche de ces fenêtres au moyen d'opérateurs de texture ; une étape d'apprentissage avec la méthode de forêts aléatoires "extra-trees" ; une étape de décision dans laquelle une image est classée selon la classe majoritaire attribuée aux fenêtres extraites de cette image.

Nous observons dans les images alvéoscopiques à la fois des micro-structures et des structures plus importantes formées par les fibres d'élastine. Ainsi, dans l'étape d'extraction d'informations dans l'image, afin de caractériser à l'information contenue dans l'image à différentes résolutions spatiales, nous adoptons l'approche originale proposée par Marée [Maree et al., 2005] consistant à extraire de façon aléatoire des fenêtres dans les images. Ces fenêtres ont des tailles aléatoires et sont extraites à des positions aléatoires dans l'image. L'extraction aléatoire de fenêtres dans les images par rapport à une extraction sur une grille régulière avec des fenêtres de taille fixe a pour avantage d'être indépendante de l'image à caractériser et particulièrement simple à mettre en oeuvre, i.e. sans a priori ni paramétrisation. Nous voyons bien par exemple que le choix d'une taille fixe de fenêtres à extraire nécessiterait une étude préalable afin de déterminer une taille optimale pour ce problème en particulier.
Dans l'étape de caractérisation de ces fenêtres extraites, nous adoptons une approche multi-résolution du LBP afin d'identifier le contenu dans l'image à différents niveaux de résolutions à la fois spatiales et angulaires. Chaque pixel de l'image est alors caractérisé par plusieurs motifs. Cette caractérisation est ainsi à la fois riche et générique. Pour l'étape de classification, nous tirons partie de l'approche de forêt aléatoire "extra-trees" [Geurts et al., 2006] en raison de ses performances générales, de sa généricité et de ses coûts de calculs moindres par rapport aux méthodes de l'état de l'art et même d'autres méthodes de forêts comme la méthode standard forest-RI.

Nous illustrons sur les schémas détaillés en Figure 2.1 et en Figure 2.2 les différentes étapes du système complet que nous proposons. La Figure 2.1 illustre l'apprentissage : l'étape d'extraction de fenêtres aléatoires, puis l'étape de caractérisation et enfin l'étape d'apprentissage de la forêt. La Figure 2.2, présentée dans le Chapitre 1, illustre l'étape de décision du système.

2.1.2 La base d'images alvéoscopiques

La base d'images alvéoscopiques fournie par les pneumologues du CHU de Rouen est composée de 226 images en niveaux de gris, 133 issues de patients non-fumeurs et 93 de patients fumeurs (cf. tableau 2.1). Parmi les images de patients non-fumeurs, 102 sont pathologiques et 31 sont saines. Pour les patients fumeurs, 33 des images sont pathologiques et 60 sont saines. Les images alvéoscopiques sont carrées, de côté variant entre 244 et 428 pixels (moyenne 262 pixels).

TABLE 2.1 – Effectifs des images MCFF pour les deux groupes fumeur et non-fumeur

Groupes / Classes	Non-fumeur	Fumeur
Sain	31	60
Pathologique	102	33
Total	133	93

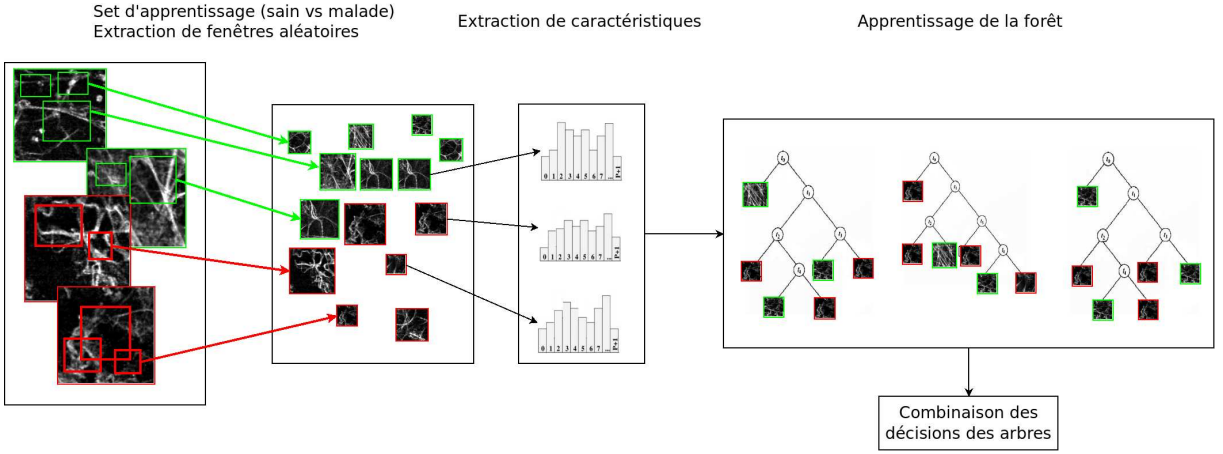


FIGURE 2.1 – Illustration de la phase d'apprentissage du système de classification d'images proposé : 1) extraction de fenêtres aléatoires dans les images MCFF, 2) description de ces fenêtres à l'aide de LBP, SIFT, co-occurrence, 3) apprentissage d'un ensemble d'arbre de décision à partir du set formé des fenêtres extraites. La Figure 2.2 illustre la phase de décision du classifieur.

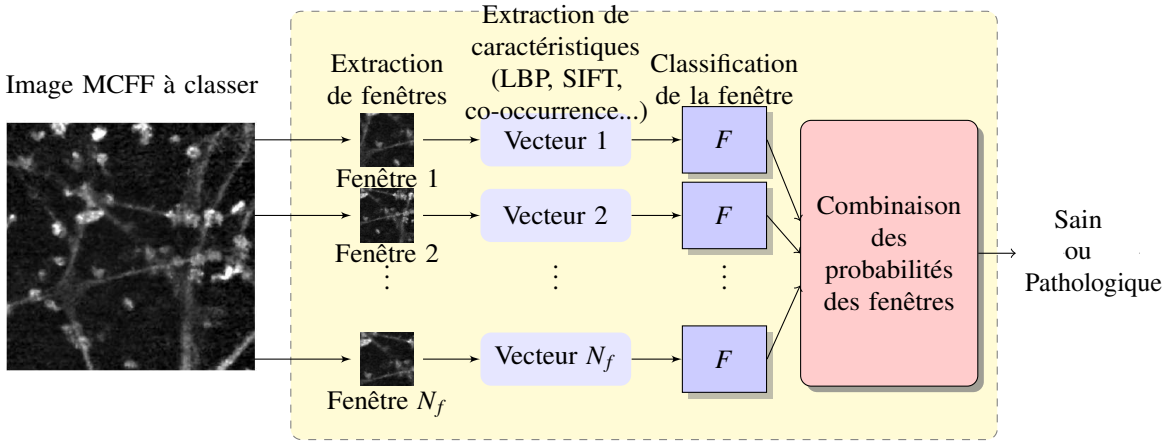


FIGURE 2.2 – Illustration de la phase de décision du système de classification d'images proposé. F désigne une forêt de L arbres de type extra-trees issue de l'apprentissage de la Figure 2.1.

2.1.3 Plan des expérimentations

Nos expérimentations se déroulent en trois points :

- Nous comparons dans un premier temps les différentes approches de description afin de déterminer la plus adaptée à nos images. Sont évalués : le LBP et ses variantes, à savoir l'opérateur Center-Symmetric LBP (CS-LBP) et l'opérateur de variance associé au LBP, les détecteur et descripteur SIFT et les statistiques de co-occurrence. Afin d'étudier le pouvoir discriminant des extracteurs utilisés dans notre comparatif, il est classique d'utiliser le classifieur standard 1-Plus Proche Voisin dans une procédure d'évaluation en leave-one-out [Hastie et al., 2001].
- Dans la seconde partie, nous présentons l'évaluation de la méthode d'apprentissage associée à l'extracteur choisi précédemment. Nous comparons les résultats obtenus avec des classifieurs à l'état de l'art (SVM et forest-RI).
- Enfin, nous proposons un mécanisme de pilotage du rejet avec les extra-trees permettant notamment de réduire la non détection. Nous proposons dans un second temps un mécanisme d'élagage des arbres de la forêt permettant d'améliorer notre approche de gestion du rejet en modifiant le mode de vote des arbres individuels et celui de la forêt, montrant par là même, de manière plus pertinente, l'influence du mode de vote sur les performances du classifieur.

2.2 Évaluation des différents descripteurs

Dans cette section, l'extracteur issu de l'opérateur *LBP* selon plusieurs résolutions et plusieurs configurations est comparé à d'autres descripteurs de texture : la variance *VAR* sur les motifs du LBP, la combinaison LBP/VAR, le "Center-Symetric" LBP ($CS-LBP$), SIFT, Dense-SIFT, les statistiques sur les matrices de co-occurrence. Nous comparons également l'approche par LBP avec une approche de description dite "Ad-Hoc", basée sur une inspection visuelle des images alvéoscopiques issue des travaux préliminaires dans l'équipe [Petitjean et al., 2009].

2.2.1 Les descripteurs évalués

Descripteurs LBP

Plusieurs résolutions de l'opérateur LBP sont considérées afin de capturer les structures de différentes tailles présentes dans les images alvéoscopiques. Nous étudions pour cela les résolutions individuelles $r_1 = \{(8,1)\}$, $r_2 = \{(16,2)\}$ et $r_3 = \{(24,3)\}$ mais aussi une approche multi-résolution combinant les résolutions précédentes, i.e. $r_{1,2} = \{(8,1),(16,2)\}$, $r_{2,3} = \{(16,2),(24,3)\}$, $r_{1,3} = \{(8,1),(24,3)\}$ et celle de 3 niveaux avec $r_{1,2,3} = \{(8,1),(16,2),(24,3)\}$. Par souci de simplicité, nous désignerons par LBP à la fois l'opérateur et le descripteur de motifs associé. Les dimensions des différents descripteurs sont indiquées dans le détail dans le tableau 2.2.

Descripteurs dérivés du LBP

Il existe plusieurs variantes autour de l'opérateur LBP comme *VAR* et le "Center-Symetric LBP" ($CS-LBP$), tous deux basés sur la description du même voisinage circulaire du LBP. Les auteurs de l'approche LBP [Ojala et al., 2002] décrivent le descripteur *VAR* comme complémentaire au LBP car il fournit une information de contraste que ne traduit pas le LBP seul. Nous étudions également l'apport de la variance *VAR* au LBP standard à l'aide du descripteur LBP/VAR réunissant ces deux approches de description. L'opérateur de variance étant continu, les valeurs de variance sont discrétisées en 15 intervalles contigus. Nous n'avons pas cherché à optimiser le nombre d'intervalles en lui attribuant une valeur de l'ordre de grandeur de la dimension du vecteur de caractéristiques issues de l'opérateur $LBP_{P,R}$. Nous noterons $VAR_{P,R}$ et $CS-LBP_{P,R}$ les descripteur VAR et CS-LBP pour le paramètre (P,R). Pour l'opérateur $CS-LBP_{P,R}$, nous ne considérons pas la résolution r_3

en raison de la dimension importante de l'espace de caractéristiques obtenu. Nous indiquons les dimensions des différents descripteurs dans le tableau 2.2.

Descripteur "Ad-Hoc"

Ce jeu de caractéristiques, dit "Ad-hoc" [Petitjean et al., 2009], est basé sur une analyse visuelle des images alvéoscopiques. Ce jeu de caractéristiques a été développé au sein de l'équipe dans le cadre de travaux préliminaires sur ces images.
Ce set est composé de cinq types de caractéristiques comprenant : 1) 5 caractéristiques issues des statistiques d'histogramme (la moyenne, la variance, la valeur du coefficient d'asymétrie, celle du coefficient d'aplatissement et celle de l'entropie) ; 2) la valeur de la densité de pixels blancs sur une image binarisée par la méthode adaptative d'Otsu [Otsu, 1979] ; 3) la somme des gradients dans l'image ; 4) le nombre de points de jonction entre les segments du squelette de l'image obtenu par la squelettisation [Baja and Thiel, 1996] et 5) 140 valeurs de statistiques extraites à partir des matrices de co-occurrence pour différentes configuration. Les caractéristiques 1-2-3 sont utilisées pour la caractérisation du contraste dans l'image et les caractéristiques 4 et 5 sont utilisées pour décrire les structures observées.

Statistiques sur les matrices de co-occurrence

Chaque matrice de co-occurrence est construite à partir de deux paramètres qui sont le nombre de niveaux de gris considéré et le vecteur de translation (caractérisé par une orientation et une longueur) pour le calcul des paires de pixels appariés, traduisant également la résolution spatiale et angulaire du descripteur. Le nombre de niveaux de gris est fixé à 8 comme valeur standard et nous considérons les 10 translations définies par les vecteurs suivants : $t_1 = (0,1)$, $t_2 = (-1,1)$, $t_3 = (-1,0)$, $t_4 = (-1,-1)$, $t_5 = (0,2)$, $t_6 = (-1,2)$, $t_7 = (-1,-2)$, $t_8 = (-2,1)$, $t_9 = (-2,0)$, $t_{10} = (-2,-1)$.

Pour chacune de ces 10 configurations, 14 valeurs de statistiques classiques sont calculées [Haralick et al., 1973] : l'énergie f_1, le contraste f_2, la corrélation f_3, la variance f_4, l'homogénéité f_5 [Pratt, 1991], l'entropie f_6, la dissimilarité f_7, la somme de moyennes f_8, la somme de l'entropie f_9, la somme de la variance f_{10}, la différence d'entropie f_{11}, la différence de variance f_{12}, deux statistiques d'information de corrélation f_{13} et f_{14}.
Le descripteur final est obtenu par la concaténation de ces différentes statistiques pour les 10 configurations adoptées de la matrice de cooccurence. Le vecteur de description obtenu est de dimension 140.

Descripteurs SIFT et Dense-SIFT

Nous avons utilisé le détecteur et le descripteur SIFT [Lowe, 2004] par le biais de la toolbox "VLFeat" [Vedaldi and Fulkerson, 2008] avec les valeurs empiriques standards proposés dans [Lowe, 2004] : la grille de calcul du descripteur d'un point d'intérêt est 16×16 ; l'histogramme des orientations est calculé à partir de 8 orientations de 45 degrés ; le seuil sur le contraste du point d'intérêt est 0.04. Avec ces paramètres standards, le vecteur décrivant l'image est de dimension 128. Le descripteur SIFT fournit un histogramme pour chaque point d'intérêt trouvé dans l'image. Afin de synthétiser la description obtenue, nous avons choisi de faire la moyenne de ces histogrammes, moyenne des différents descripteurs de chaque point d'intérêt détecté dans l'image. Cette approche de somme des histogrammes SIFT a notamment été utilisée dans [Ghazali et al., 2008] pour l'identification d'herbes. Nous avons représenté sur la figure 2.3 les points d'intérêt déterminés par le descripteur SIFT sur quelques exemples d'images alvéoscopiques. Nous voyons sur cette image que le descripteur identifie comme points robustes essentiellement les bords de structures d'élastine qui sont saillantes dans l'image.
Le descripteur Dense-SIFT reprend la paramétrisation du descripteur SIFT à l'exception de la phase

de recherche de points d'intérêts qui est simplement absente. En effet, le descripteur est calculé pour chaque pixel de l'image. Le descripteur a ainsi la même dimension 128.

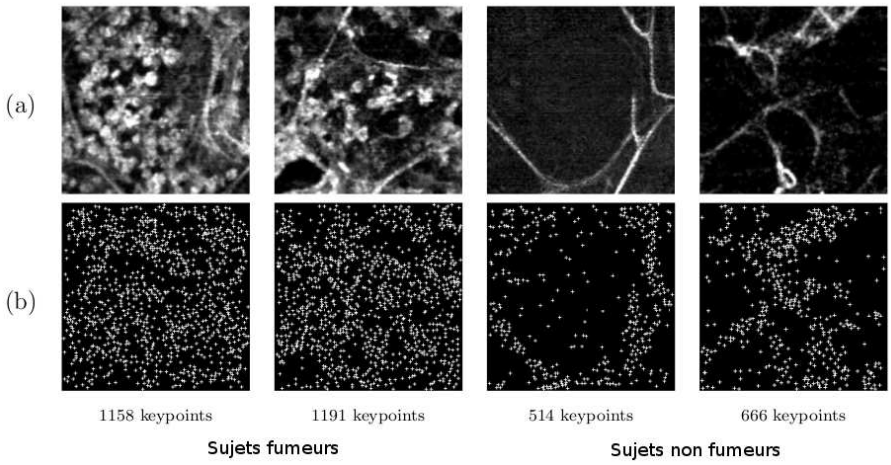


FIGURE 2.3 – (a) Images alvéoscopiques (MCFF) avec (b) les points d'intérêt trouvés par SIFT

TABLE 2.2 – Dimension des vecteurs de caractéristiques associées à chaque extracteur : $LBP_{P,R}$, $VAR_{P,R}$, $LBP/VAR_{P,R}$, $CS-LBP_{P,R}$, SIFT, Dense-SIFT et le descripteur des matrices de co-occurrence

Résolution (P,R)	$LBP_{P,R}$	$VAR_{P,R}$	$LBP/VAR_{P,R}$	$CS-LBP_{P,R}$
8,1	10	15	25	16
16,2	18	15	33	256
24,3	26	15	41	-
8,1 + 16,2	28	30	58	272
8,1 + 24,3	36	30	66	-
16,2 + 24,3	44	30	74	-
8,1 + 16,2 + 24,3	54	45	99	-

(a)

	SIFT	Dense-SIFT	Cooccurrence
Dimension	128	128	140

(b)

2.2.2 Protocole expérimental

Nous utilisons le classifieur standard 1ppv pour évaluer le pouvoir discriminant des différents descripteurs. La procédure de classification utilisée est le leave-one-out dans laquelle, tour à tour, un unique exemple de la base d'images est utilisé pour former la base de test et les exemples restant constituent la base d'apprentissage. Nous utilisons comme mesure de performances le taux de reconnaissance global (noté "accr" [1]), le taux de reconnaissance sur les images de patients sains (taux de vrais négatifs ou spécificité noté "tnr" [2]) et celui sur les images de patients malades (taux de vrais positifs ou sensibilité noté "tpr" [3]). Si on désigne par TP le nombre de cas malades identifiés correctement par le 1PPV, TN pour les cas pathologiques, FP le nombre de cas sains

1. "accr" est utilisé pour le terme anglais "accuracy rate".
2. "tnr" est utilisé pour "true negative rate".
3. "tpr" est utilisé pour "true positive rate".

identifiés comme pathologiques, FN le nombre de cas pathologiques identifiés comme sains, alors les expressions de accr, tnr et tpr sont données par les expressions ci-après :

$$\begin{cases} accr = \frac{TP+TN}{TP+TN+FN+FP} \\ tpr = \frac{TP}{TP+FN} \\ tnr = \frac{TN}{TN+FP} \end{cases} \tag{2.1}$$

2.2.3 Résultats et analyse

TABLE 2.3 – Performances pour la base non-fumeur des extracteurs Ad-hoc, statistiques de co-occurrence, SIFT et Dense-SIFT en termes de taux de reconnaissance globale (accr), spécificité (tpr) et sensibilité (tnr) ; les meilleurs résultats obtenus sont mis en gras.

Descripteur	accr	tpr	tnr
Ad-hoc	0.6165	0.7549	0.1612
Cooccurrence	0.7969	0.8921	0.4838
SIFT	**0.8345**	**0.9215**	**0.5483**
Dense-SIFT	0.7368	0.8333	0.4193

TABLE 2.4 – Performances pour la base non-fumeur des extracteurs $LBP_{P,R}$, $VAR_{P,R}$, $LBP/VAR_{P,R}$ et $CS-LBP_{P,R}$

Résolution (P,R)	Mesure	$LBP_{P,R}$	$VAR_{P,R}$	$LBP/VAR_{P,R}$	$CS-LBP_{P,R}$
8,1	accr	**0.9323**	0.6917	0.9172	0.7744
	tpr	0.9607	0.8333	0.9411	0.8137
	tnr	0.8387	0.2258	0.8387	0.6451
16,2	accr	**0.9323**	0.6691	0.8947	0.8345
	tpr	0.9509	0.8039	0.9313	0.9019
	tnr	0.8709	0.2258	0.7741	0.6129
24,3	accr	0.8947	0.7368	**0.9172**	-
	tpr	0.9215	0.8627	0.9411	
	tnr	0.8064	0.3225	0.8387	
8,1 + 16,2	accr	**0.9248**	0.6541	0.879	0.8345
	tpr	0.9509	0.7941	0.9215	0.8921
	tnr	0.8387	0.1935	0.7419	0.6451
8,1 + 24,3	accr	**0.9398**	0.6992	0.9172	-
	tpr	0.9607	0.8235	0.9607	
	tnr	0.8709	0.2903	0.7741	
16,2 + 24,3	accr	0.9097	0.6992	**0.9248**	-
	tpr	0.9411	0.8235	0.9509	
	tnr	0.8064	0.2903	0.8387	
8,1 + 16,2 + 24,3	accr	**0.9398**	0.6917	0.9323	-
	tpr	0.9607	0.8235	0.9509	
	tnr	0.8709	0.2580	0.8709	

Base non-fumeurs Les résultats obtenus pour cette base sont présentés dans les tableaux 2.3 et 2.4. On observe que les meilleurs résultats sont obtenus avec l'extracteur $LBP_{P,R}$ pour différentes résolutions (toutes les résolutions sauf $r_3 = \{(24,3)\}$ et $r_{2,3} = \{(16,2),(24,3)\}$). Le bon comportement de l'opérateur LBP est illustré dans la figure 2.4, qui donne pour chaque pixel son association avec un pattern de bord ou un pattern de zone homogène. Sur cette figure, on voit

TABLE 2.5 – Performances pour la base fumeur des extracteurs Ad-hoc, statistiques de co-occurrence, SIFT et Dense-SIFT en termes de taux de reconnaissance globale (accr), spécificité (tpr) et sensibilité (tnr)

Descripteur	accr	tpr	tnr
Ad-hoc	0.6989	0.5757	0.7666
Cooccurrence	**0.8817**	**0.8787**	**0.8833**
SIFT	0.7849	0.7272	0.8166
Dense-SIFT	0.6666	0.5151	0.75

TABLE 2.6 – Performances pour la base fumeur des extracteurs $LBP_{P,R}$, $VAR_{P,R}$, $LBP/VAR_{P,R}$ et $CS-LBP_{P,R}$

Résolution (P,R)	Mesure	$LBP_{P,R}$	$VAR_{P,R}$	$LBP/VAR_{P,R}$	$CS-LBP_{P,R}$
8,1	accr	0.9462	0.6129	**0.9569**	0.6881
	tpr	0.9697	0.5151	0.9697	0.5454
	tnr	0.9333	0.6666	0.9500	0.7666
16,2	accr	0.9354	0.6559	**0.9569**	0.8602
	tpr	0.9393	0.5151	0.9697	0.7878
	tnr	0.9333	0.7333	0.9500	0.9
24,3	accr	0.9354	0.6021	**0.9569**	-
	tpr	0.9090	0.3939	0.9697	
	tnr	0.95	0.7166	0.9500	
8,1 + 16,2	accr	0.9354	0.6881	**0.9569**	0.8387
	tpr	0.9393	0.6060	0.9697	0.7878
	tnr	0.9333	0.7333	0.9500	0.8666
8,1 + 24,3	accr	0.9354	0.7311	**0.9569**	-
	tpr	0.9393	0.6363	0.9697	
	tnr	0.9333	0.7833	0.9500	
16,2 + 24,3	accr	0.9462	0.6451	**0.9569**	-
	tpr	0.9090	0.3636	0.9697	
	tnr	0.9666	0.8	0.9500	
8,1 + 16,2 + 24,3	accr	0.9354	0.6989	**0.9569**	-
	tpr	0.9393	0.5151	0.9697	
	tnr	0.9333	0.8	0.9500	

bien que les motifs de bords coïncident avec les structures d'élastine dans l'image, et que l'image de patient sain présente des caractéristiques différentes de celle du patient pathologique.

L'extracteur issu de l'opérateur de variance $VAR_{P,R}$ obtient de très mauvais résultats pour l'identification des cas pathologiques comparativement aux autres extracteurs. On remarque alors l'apport de l'opérateur de variance en terme de performance dans la distribution jointe pour l'extracteur $LBP/VAR_{P,R}$ dans le cas notamment de la résolution $r_3 = \{(24,3)\}$ tant pour les images de patients sains que pour les images pathologiques. Les performances des extracteurs $LBP_{P,R}$ et $LBP/VAR_{P,R}$ sont généralement proches avec une légère avance pour LBP. Cependant $LBP/VAR_{P,R}$ est plus coûteux que $LBP_{P,R}$ en raison de l'ajout de l'opérateur de variance. Donc, à performances proches, nous privilégions l'extracteur LBP, moins gourmand. Ensuite, on constate qu'à résolution équivalente, l'extracteur $LBP_{P,R}$ est plus performant que $CS-LBP_{P,R}$. Le descripteur SIFT quant à lui, moins performant que $LBP_{P,R}$ en terme de taux de reconnaissance globale, se place en tête des autres descripteurs pour la base non-fumeur, dont le descripteur de co-occurrence.

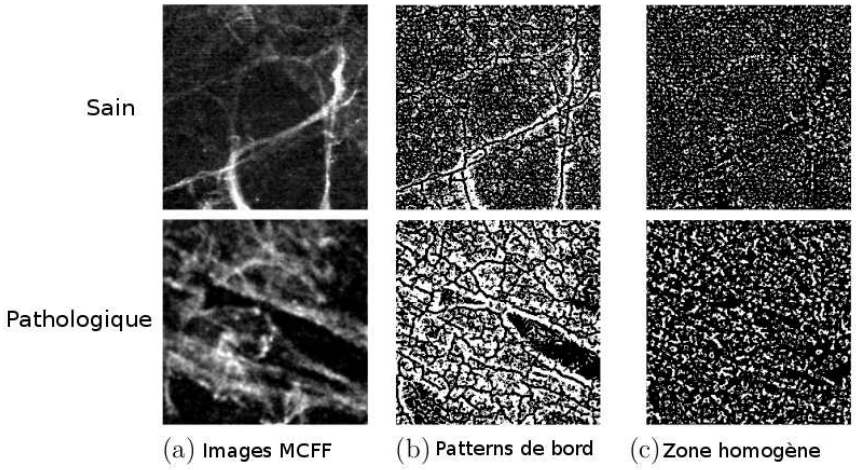


FIGURE 2.4 – (a) Images alvéoscopiques avec (b) les patterns de bord et (c) les patterns de zones homogènes de l'opérateur LBP $r_{1,2,3} = \{(8,1),(16,2),(24,3)\}$

Base fumeurs Les résultats obtenus pour cette base sont présentés dans les tableaux 2.5 et 2.6. On observe les bonnes performances de l'opérateur $LBP/VAR_{P,R}$ pour toutes les résolutions. Les résultats identiques constatés dans le tableau 2.6 sont dus au choix d'un unique voisin en phase de test pour le classifieur 1-PPV. Nous avons en effet évalué cet extracteur avec un nombre différent de voisins et nous avons constaté une tendance à la baisse des performances lorsque le nombre de voisins augmente, quelle que soit la résolution considérée. Ces résulats complémentaires figurent à l'Annexe 4.5 pour les deux groupes, fumeur et non-fumeur. Les performances de l'extracteur $LBP_{P,R}$ sont proches de celles de $LBP/VAR_{P,R}$, montrant de nouveau l'apport de l'opérateur de variance. En revanche, pour la base fumeur, c'est le descripteur de co-occurrence, tout en étant moins performant que *LBP* et LBP/VAR qui se comporte mieux que les autres approches, suivi par SIFT. Cependant, l'approche par co-occurrence est plus coûteuse que SIFT lorsque l'on tient compte des 14 statistiques, les implémentations courantes ne reprenant que les 5 premières.

Nous constatons notamment que les résultats obtenus avec *LBP* sont nettement meilleurs que ceux obtenus avec le descripteur Ad-hoc. Cette baisse de performances par rapport aux précédents résultats obtenus avec le descripteur Ad-hoc peut être notamment due à l'agrandissement de la base de données d'images alvéoscopiques utilisée initialement au sein de l'équipe (passant d'un effectif de 33 à 133 pour la base non-fumeur, de 27 à 93 pour la base fumeur) soulignant un problème de généricité du descripteur constitué. Le *LBP* quant à lui montre l'apport en termes à la fois de performances et de généricité de la description par la distribution de primitives visuelles simples ou motifs binaires dans l'image, en étant discriminant à la fois sur les images de patients fumeurs et non-fumeurs.

Conclusion : choix de l'opérateur LBP En considérant les résultats obtenus sur les deux bases alvéoscopiques fumeur et non-fumeur, nous choisissons pour la suite l'extracteur LBP combinant les deux résolutions $r_{1,2} = \{(8,1),(16,2)\}$. Afin de souligner ce caractère discriminant, des matrices montrant les distances euclidiennes inter-classes et intra-classes dans l'espace LBP sont représentées à la figure 2.5, pour les deux groupes fumeur et non-fumeur. Les zones claires sont associées aux plus fortes valeurs de distance. Ces zones indiquent des images éloignées dans l'espace des caractéristiques et donc susceptible d'appartenir à des classes différentes.
On observe que les régions sombres sont généralement associées aux individus de la même classe (blocs diagonaux) et que les régions plus claires sont associées aux données inter-classes. Ceci indique bien que l'espace de caractéristiques choisi a un bon pouvoir discriminant. On constate que la région intra-classe du groupe non-fumeur est beaucoup plus sombre que dans le cas fumeur,

indiquant ainsi que les images de la même classe sont proches l'une de l'autre. On observe cependant que la région interclasse est cependant moins claire chez le non-fumeur que chez le fumeur, indiquant une plus grande confusion possible dans l'identification des cas non-fumeurs. Nous constatons en effet que les performances sont généralement meilleures dans le cas fumeur que dans le cas non-fumeur. La présence de macrophages et la visibilité des parois alvéolaires jour peut-être un rôle dans la discrimination d'images saines et pathologiques dans ce cas.

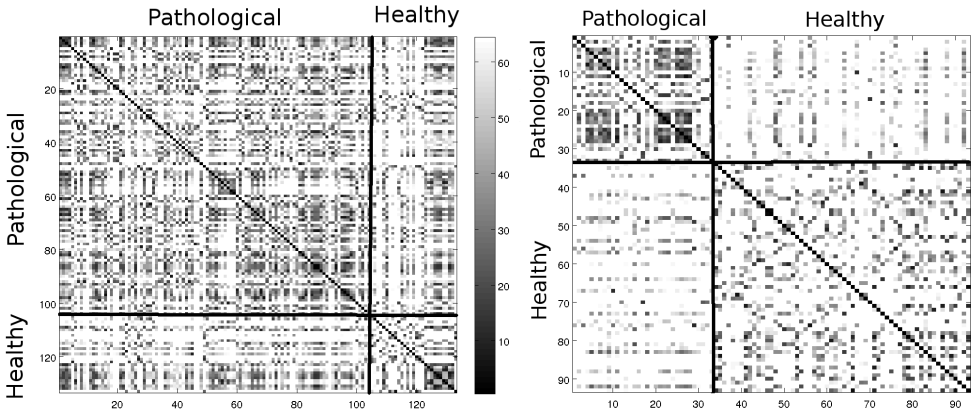


FIGURE 2.5 – Matrices des distances entre les individus des deux classes sain et pathologique pour le groupe non-fumeur (à gauche, 133 images) et fumeur (à droite, 93 images) dans l'espace LBP. Les régions claires sont associées à des paires d'images éloignées dans l'espace des caractéristiques ; les régions sombres indiquent des paires d'individus susceptibles d'appartenir à la même classe

2.3 Évaluation du système de classification

Nous étudions dans cette section la méthode de forêt extra-trees présentée comme performante et générique et la comparons aux approches de classification standards de la littérature, à savoir les forest-RI et le SVM. La comparaison des classifieurs est faite sur deux caractérisations possibles de l'image : une caractérisation locale par l'extraction de fenêtres aléatoires dans l'image (décrites par LBP ou le pixel brut comme dans la description originelle) vs une caractérisation globale.

2.3.1 Protocole expérimental

Caractérisation des images

Dans la première partie de cette expérimentation, les images sont caractérisées de façon globale, comme précédemment, avec le descripteur $LBP_{(8,1)/(16,2)}$ de dimension 28, noté par simplification *LBP*. Dans la seconde partie, les images sont caractérisées de façon locale : des fenêtres sont extraites aléatoirement dans les images ; ces fenêtres, constituant autant de représentants de la même classe que l'image dont elles ont été extraites, sont alors décrites par LBP. Nous avons opté pour le choix par défaut de 80000 fenêtres pour toute la base d'apprentissage, équiréparties entre les classes (i.e. 40000 fenêtres par classes) et 100 fenêtres par image de test.

Paramétrisation des classifieurs

SVM

Pour le SVM, nous avons utilisé l'implémentation de la Kernel Methods Matlab Toolbox [Canu et al., 2005]. Nous avons fait le choix standard d'un noyau polynomial, dont le degré 3 a été

choisi empiriquement. Le paramètre de coût du SVM est fixé à $C = 10^5$ comme valeur par défaut de la toolbox.

forest-RI, extra-trees

Nous rappelons que la différence fondamentale entre les extra-trees et les forest-RI est que les extra-trees injectent beaucoup plus d'aléatoire dans la construction de la règle de partitionnement en ne faisant pas de parcours exhaustif du point de coupure pour un attribut donné et utilisent tout le set d'apprentissage au lieu d'un sous-échantillon bootstrap du set initial dans le cas de forest-RI.

Les paramètres intervenant dans la construction des extra-trees et des forest-RI sont le nombre L d'arbres de la forêt, le nombre d'attributs K_{RFS} sélectionnés aléatoirement en chaque nœud lors de l'induction d'un arbre et le critère d'arrêt de partitionnement d'un nœud lors de l'induction d'arbre qui est ici le nombre n_{min} de données minimales en présence. Nous avons opté pour des valeurs standards de la plupart de ces paramètres [Geurts et al., 2006] :

- le nombre d'arbres de la forêt est $L = 30$; ce choix est empirique car nous avons observé une convergence des extra-trees dès la valeur $L = 15$;
- $K_{RFS} = \sqrt{M}$ où M est la dimension de l'espace de caractéristiques sélectionné ;
- $n_{min} = 2$ (dans ce cas chaque arbre est complètement développé).

Protocole d'évaluation

Le protocole repose sur une validation croisée 10-fold stratifiée présentée comme un standard notamment en présence de bases de données de faibles effectifs [Hastie et al., 2001]. Avec cette procédure de validation croisée, 10 partitions sont constituées, les classes étant réparties de façon égale dans chacune des partitions. Tour à tour, une partition est utilisée pour former la base de test et les 9 partitions restantes sont utilisées pour constituer la base d'apprentissage. Ainsi, 90% des données de chaque classe est utilisée pour former la base d'apprentissage et les 10% restants servant à tester le classifieur.
La mesure des performances est donnée comme dans l'expérimentation précédente en termes de taux de reconnaissance global (accr), taux de reconnaissance sur les cas sains (tnr) et les cas pathologiques (tpr). Il est à noter que dans le cas de l'étude de l'approche par extraction de fenêtres aléatoires, 100 représentants de l'image de test sont évalués ; le classifieur attribue alors la classe majoritaire à l'image testée comme suggéré par Marée.

2.3.2 Résultats et analyse

2.3.2.1 Approche par caractérisation globale

Nous présentons dans le tableau 2.7 les résultats obtenus avec les extra-trees (ET) que nous comparons avec ceux du SVM. Nous constatons que les deux approches ont des taux de reconnaissance globale très proches sur les deux bases, voire même identiques pour la base de données fumeur. Sur cette dernière, nous pouvons observer des différences importantes en analysant les taux de reconnaissance sur les cas sains et les cas pathologiques. En effet, 2 cas sains ont été identifiés comme pathologique pour le SVM contre 4 cas pour les extra-trees ; inversement, le SVM présente 3 cas de non-détection de la pathologie (des cas pathologiques identifiés comme sains) contre 1 seul cas pour les extra-trees. Le faible effectif de la base d'apprentissage fait aboutir au même résultat de reconnaissance globale pour les deux classifieurs.

La non-détection des cas pathologiques est très préjudiciable pour notre application ; nous tenons donc compte de cette observation pour la suite de notre étude. Nous observons également une explosion de l'écart-type pour la détection des cas pathologiques pour la base fumeur dans le cas du SVM, les extra-trees se révélant plus stable pour cette base. Tout en ayant des performances assez proches, les extra-trees demeurent beaucoup moins coûteux que le SVM. En effet, si on note

N la taille de l'échantillon d'apprentissage, les complexités de ces deux algorithmes sont de l'ordre de :

– $L \cdot N \cdot log_2 N$ pour la forêt, où L est le nombre d'arbres,

– $N^3 + (M+1) \cdot N^2$ pour le SVM, où M est la dimension de l'espace de caractéristiques [4]

Ainsi, pour notre paramétrisation, il existe un facteur 100 entre les complexités des deux classifieurs. Cette différence est notamment accentuée lors de l'étude de l'approche de caractérisation locale que nous présentons ci-après où le nombre d'éléments en apprentissage est beaucoup plus important.

TABLE 2.7 – Taux de reconnaissance globale des extra-trees (ET) et du SVM pour les bases fumeur et non-fumeur avec le descripteur LBP dans le cadre d'une caractérisation globale de l'image.

		Non-fumeur	Fumeur
ET	accr	$91.53 \pm 08.46\%$	$94.44 \pm 07.85\%$
	tpr	92.0 ± 10.32	96.66 ± 10.54
	tnr	$90.0 \pm 16.0O$	93.33 ± 08.60
SVM	accr	$91.82 \pm 08.24\%$	$94.44 \pm 07.85\%$
	tpr	91.33 ± 08.34	90.0 ± 22.49
	tnr	93.33 ± 14.05	96.6 ± 07.02

2.3.2.2 Approche par caractérisation de fenêtres dans l'image

Dans cette partie, un grand nombre de fenêtres sont extraites de façon totalement aléatoire dans la base d'images afin d'évaluer l'apport d'une approche locale de la caractérisation du contenu de l'image.

Comparaison des extra-trees avec le SVM

Avec l'implémentation du SVM de la Toolbox utilisée, il ne nous a pas été possible de générer les 80000 fenêtres en apprentissage que nous avons fixé par défaut dans la paramétrisation. Afin de maintenir notre comparaison avec le SVM, nous avons abaissé ce nombre à 10000. Les résultats obtenus pour cette valeur sont affichés dans le tableau 2.8. Nous pouvons constater en premier lieu que les performances sont généralement améliorées pour les deux classifieurs, notamment dans le cas de la base des patients fumeurs.

Nous voyons ainsi l'apport de l'approche d'extraction locale d'informations dans l'image. Les extra-trees se révèlent de façon générale meilleurs que le SVM, avec notamment de meilleurs performances sur la base non-fumeur. Nous avons tracé les courbes ROC des deux approches SVM et extra-trees pour les deux bases fumeur et non fumeur dans la figure 2.6. Les deux courbes ne se croisent pas et les extra-trees ont une aire sous la courbe plus importante que SVM. Cela confirme les bonnes performances de l'approche que nous proposons par extra-trees par rapport au SVM.
Outre les bonnes performances générales, les extra-trees sont beaucoup moins coûteux à entrainer que le SVM, avec cette fois, pour $N = 10000$, un facteur 10^5 d'écart entre les complexités des deux méthodes. Le SVM s'avère alors prohibitif pour une plus grande valeur de N.
Nous poursuivons le reste de l'étude avec les extra-trees et avec N =80000 comme fixé dans le protocole.

Influence de l'aléatoire dans la taille des fenêtres extraites

Nous avons comparé ici deux approches d'extraction des fenêtres pour la classification par extra-trees. Tout en maintenant une extraction des fenêtres à des positions aléatoires, l'utilisation d'une taille de fenêtre aléatoirement choisie ou fixée à l'avance est étudiée. L'idée est de montrer qu'avec une taille aléatoire de fenêtres, le balayage ainsi effectué au niveau à la fois local (fenêtres

4. [Bottou and Lin, 2007] indique par ailleurs que la complexité générale des solveurs SVM croît avec N^2 si le paramètre de coût est faible et avec N^3 si ce dernier a une valeur importante.

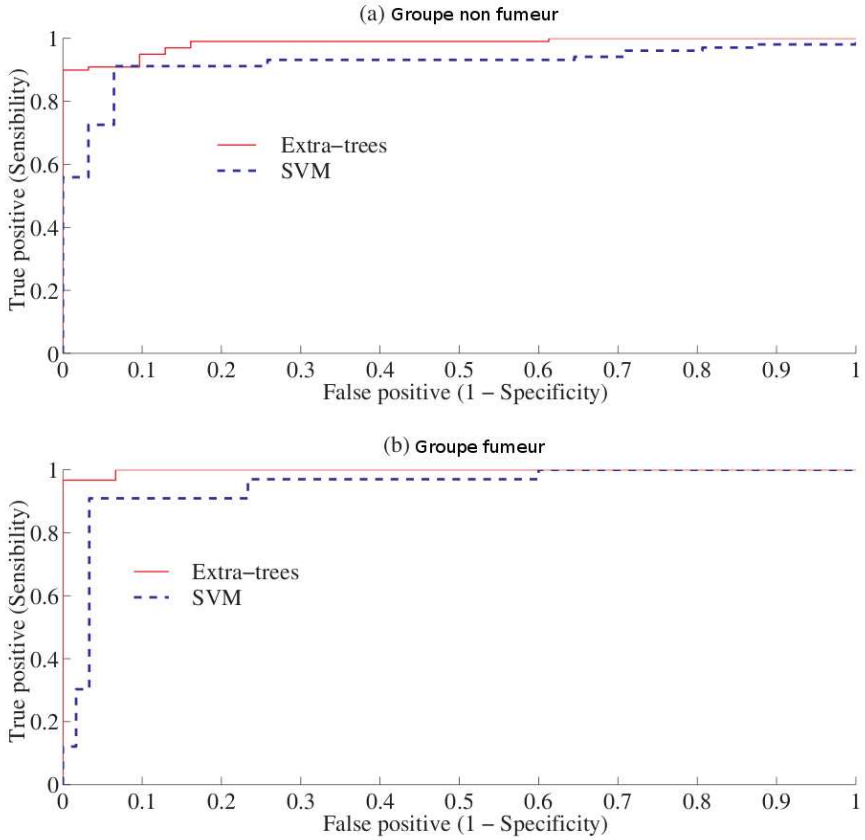


FIGURE 2.6 – Courbes ROC pour les extra-trees et le SVM sur les deux bases (a) non fumeur et (b) fumeur

TABLE 2.8 – Taux de reconnaissance globale des extra-trees (ET) et du SVM pour les bases fumeur et non-fumeur avec une caractérisation locale par extraction de fenêtres (nombre de fenêtres abaissé à 10000 pour des raisons de coûts de calculs trop importants pour le SVM de notre implémentation)

		Non-fumeur	Fumeur
ET	accr	$93.84 \pm 07.06\%$	$97.77 \pm 04.68\%$
	tpr	92.25 ± 08.72	97.5 ± 07.90
	tnr	100.0 ± 0.0	98.57 ± 04.51
SVM	accr	$90.76 \pm 08.73\%$	$98.88 \pm 03.51\%$
	tpr	90.34 ± 09.80	100.0 ± 0.0
	tnr	94.66 ± 11.67	97.5 ± 07.90

de petites tailles) et global (fenêtres de grandes tailles) est suffisamment important pour capturer le contenu informatif des structures et micro-structures de l'image. Les résultats de cette étude sont présentées dans la figure 2.7.

Nous observons que l'approche par extraction de fenêtres de tailles aléatoires permet effectivement d'atteindre des performances comparables à l'approche avec la taille optimale de fenêtres dans l'image. En outre, le choix d'une taille fixe de fenêtres à extraire se fait en amont de l'apprentissage et la valeur retenue dépend très fortement de la nature du problème à traiter. Sélectionner l'approche aléatoire permet de ne pas dépendre de ce paramètre tout en ayant une description à différentes résolutions du contenu de l'image.

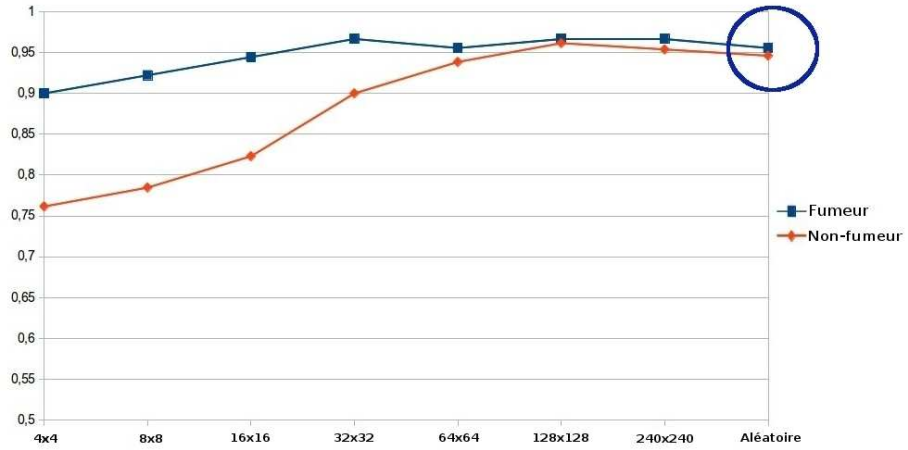


FIGURE 2.7 – Performances des extra-trees en fonction de tailles fixes de fenêtres de 4x4 jusqu'à la taille maximale pour les images que nous avons à traiter, soit 240x240. Le dernier point cerclé en bleu indique une taille de fenêtre aléatoire (indiqué par le cercle en dernière position).

Apport d'une description riche des fenêtres extraites

La caractérisation des fenêtres par pixels bruts initialement proposée par Marée est comparée à la caractérisation LBP dans le tableau 2.9 (résultats obtenus avec les extra-trees pour $N = 80000$). La caractérisation par pixels bruts a des performances inférieures comparée au LBP. On observe ainsi pour la base fumeur une différence de 23% entre la caractérisation par pixels bruts et celle par LBP. Cela peut s'expliquer par le fait que la description par LBP est beaucoup plus riche en terme de capture d'informations dans l'image que la description par pixels bruts et que le contenu informatif des images semble résider dans des structures et des micro-structures que le LBP est capable de capturer. Nous montrons des résultats équivalents à l'Annexe 4.5, confortant cette approche sur la base d'images médicales publiques Hela-Cells [Murphy et al., 2000]. Ainsi, la description plus riche des fenêtres extraites dans les images alvéoscopiques est fortement favorable dans notre étude.

TABLE 2.9 – Taux de reconnaissance globale des extra-trees pour les images alvéoscopiques de patients fumeurs et non-fumeurs pour la caractérisation par l'intensité du pixel brut initialement proposée par Marée et pour la caractérisation par LBP proposée.

	Non-Fumeur	Fumeur
Pixels bruts [Maree et al., 2005]	$80.0 \pm 9.2\%$	$73.3 \pm 13.3\%$
$LBP_{(8,1)(16,2)}$	$93.5 \pm 7.4\%$	$96.8 \pm 5.4\%$

Comparaison entre extra-trees et forest-RI

Les résultats des extra-trees sont comparés à ceux obtenus par forest-RI dans le tableau 2.10. Pour ces deux méthodes de forêts aléatoires, le temps moyen de calcul par arbre est également indiqué. Nous observons que les deux méthodes présentent des résultats très proches, avec un avantage en moyenne pour les extra-trees. La différence essentielle réside dans les coûts de calculs : les extra-trees sont deux fois plus rapides que les forest-RI.

Analyse de la matrice de confusion des extra-trees

En observant plus en détail la matrice de confusion pour les extra-trees sur les deux bases d'images (cf. tableau 2.11), nous constatons que les extra-trees obtiennent de très bonnes performances sur les images des cas sains (96.7% pour le cas non-fumeur et 96.1% pour le cas fumeur). La non-détection du cas pathologique est cependant plus importante chez le non-fumeur que chez le

TABLE 2.10 – Taux de reconnaissance et temps de calcul par arbre des deux méthodes de forêts aléatoires sur les deux bases d'images alvéoscopiques fumeur et non-fumeur pour la caractérisation par $LBP_{(8,1)(16,2)}$.

	Non-Fumeur	Temps	Fumeur	Temps
forest-RI	$92.3 \pm 4.8\%$	48.6 s	$95.5 \pm 7.3\%$	46.4 s
ET	$93.5 \pm 7.4\%$	23.4 s	96.8 ± 5.4 %	24.7 s

fumeur et s'élève à 9.6% (correspondant dans notre expérimentation à 9 cas pathologiques). Nous présentons dans la section suivante des résultats complémentaires permettant d'étudier plus en détail la non-détection et nous proposons des approches permettant de diminuer sa valeur.

TABLE 2.11 – Matrice de confusion des cas pathologiques (P) et sains (S) avec les extra-trees

		Non-fumeur		Fumeur	
		Sortie classifieur		Sortie classifieur	
		P	S	P	S
Vérité terrain	P	90.4%	9.6%	97.5%	2.5%
	S	3.3%	96.7%	3.9%	96.1%

Conclusion

Dans cette section, nous avons comparé l'approche par méthode d'ensemble d'arbres de décision extra-trees aux approches standards SVM et forest-RI. Nous avons également étudié l'apport en termes de performances d'une description locale riche des images par la méthode d'extraction de fenêtres et l'influence de l'aléatoire dans cette dernière approche. Nous avons montré, à l'issue de ces expérimentations d'une part que les extra-trees constituaient un bon choix de classifieur en raison de leurs performances plus importantes et leurs coûts beaucoup plus faibles comparées au SVM et en raison essentiellement du coût plus faible comparé aux forest-RI ; d'autre part, nous avons montré qu'une caractérisation plus riche des fenêtres extraites des images alvéoscopiques permettait d'obtenir de meilleures performances comparées à une approche de description bas-niveau.

Cependant, nous avons constaté que le taux de non détection demeurait élevé et nous cherchons à le minimiser dans la section suivante.

2.4 Réduction de la non-détection

Dans le domaine médical, la non détection des cas pathologiques a un coût important en raison des conséquences néfastes pour un patient malade qui serait pris en charge tardivement. Il est alors nécessaire d'avoir une procédure de décision en deux étapes : une étape d'évaluation de la confiance associée à cette décision et une étape de gestion du rejet de la décision dans le cas d'un trop faible niveau de confiance. Cette seconde étape consistera par exemple à analyser plus en détail par un opérateur expert ou avec un classifieur plus spécifique les exemples rejetés.

Nous proposons un mécanisme de pilotage du rejet en modifiant la fonction de décision des extra-trees. Cette modification consiste essentiellement à obtenir un taux de confiance dans l'attribution d'une classe à une image de test à partir des votes des arbres. Un exemple difficile à classer aura ainsi un taux de confiance faible en l'absence d'un consensus clair au sein des arbres. Nous avons cherché également à approfondir et améliorer ce mécanisme en modifiant le mode de vote des arbres.

2.4.1 Mécanisme de rejet avec les extra-trees

Nous développons et évaluons dans cette partie un mécanisme de rejet basé sur le consensus entre les arbres au sein des extra-trees pour l'attribution d'une classe à une image. Nous souhaitons

montrer l'efficacité de cette approche pour réduire la non détection en proposant un mécanisme de seuillage basé sur le niveau de confiance qui pourra être choisi par le praticien afin de fiabiliser la réponse du système de classification.

Dans notre application médicale, nous avons vu que chaque fenêtre de test reçoit un vote lui attribuant sa classe d'appartenance. Puis l'ensemble des votes reçus pour toutes les fenêtres d'une image de test sont rassemblés pour attribuer la classe majoritaire à l'image de test. Formellement, si ω_S désigne la classe saine, I l'image de test, x_i la i-ème fenêtre extraite de I, l_i^S le nombre d'arbres de la forêt ayant voté pour la classe saine ω_S et L le nombre total d'arbres de la forêt, nous obtenons la probabilité conditionnelle :

$$p(\omega_S|x_i) = \frac{l_i^S}{L}$$

On en déduit alors la probabilité de la classe ω_S conditionnellement à l'image d'origine I :

$$p(\omega_S|I) = \frac{1}{N_f} \cdot \sum_i \frac{l_i^S}{L}$$

où N_f désigne le nombre de fenêtres extraites de l'image de test. La valeur de probabilité $p(\omega_S|I)$ représente une mesure de confiance pour l'attribution de la classe ω_S à l'image I. Une valeur proche de 0.5 signifie qu'il n'existe pas de consensus réel au sein de la forêt et donc que l'exemple à classer est particulièrement difficile. Ainsi un seuillage adapté sur la valeur de probabilité $p(\omega_S|I)$ permettra d'écarter les images de test qui obtiendraient de faibles valeurs de probabilité.

Protocole expérimental

Ce protocole se base sur celui de la section précédente. Nous modifions le jeu de test standard (que nous appelons ci-après "images standards" par soucis de simplification) en considérant les fenêtres extraites comme autant d'images indépendantes à tester. Ainsi, ce second jeu est virtuellement plus important que le jeu de départ car il contient à présent autant d'images que de fenêtres ayant été extraites. Nous appelons ce second jeu simplement "fenêtres". Nous rappelons que dans le protocole précédent, nous avons extrait 100 fenêtres de chaque image de test ; à l'issue du vote des arbres, la classe majoritaire parmi celles attribuées à ces fenêtres est choisie comme classe de l'image testée. Dans ce protocole, notre "jeu de fenêtres" est composé donc de 100 images nouvelles à classer, indépendamment les unes des autres. Ce second jeu a été choisi pour disposer de davantage d'images dans la phase de test afin d'étudier l'impact du mécanisme de rejet que nous mettons en place.

Nous précisons dans le tableau 2.12 les valeurs des différents paramètres de la méthode et les effectifs en apprentissage et en test des deux jeux considérés.

TABLE 2.12 – Effectifs des bases d'apprentissage et test pour les jeux "images standards" et "fenêtres" et rappel des paramètres des extra-trees.

Nombre de fenêtres en apprentissage	$N_{app} = 80000$
Nombre de fenêtres en apprentissage par classe	40000
Nombre d'images en test (jeu "images standards"')	$N_{test} = 130$ (fumeur) $N_{test} = 90$ (non-fumeur)
Nombre d'images en test (jeu "fenêtres")	$N_{test} = 13000$ (fumeur) $N_{test} = 9000$ (non-fumeur)
Nombre d'attributs pour $LBP_{(8,1)/(16,2)}$	$M = 28$
Nombre d'attributs aléatoires par arbre	$K_{RFS} = \sqrt{M}$
Nombre d'arbres par forêt	$L = 30$

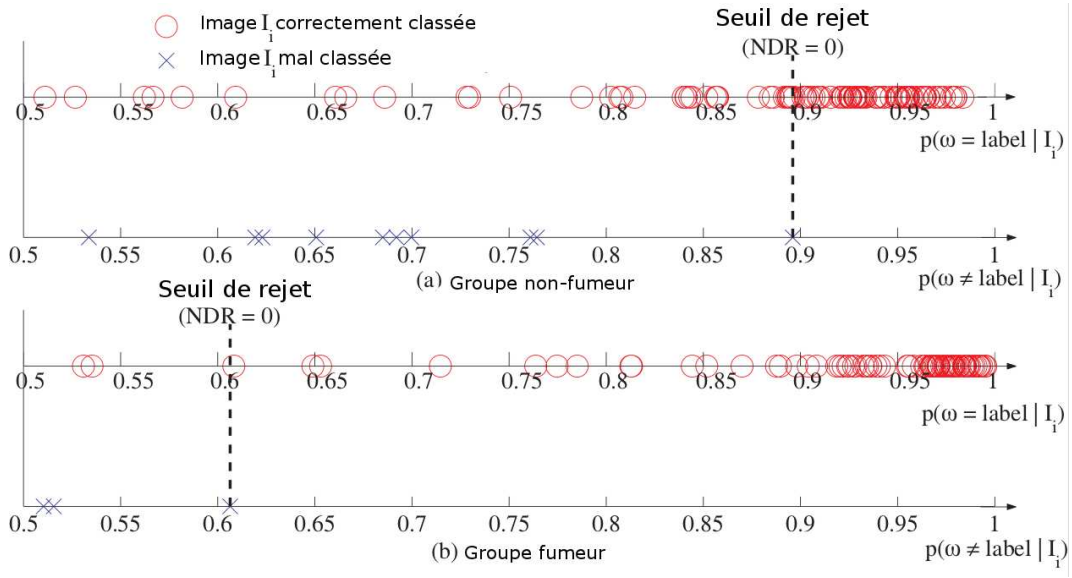


FIGURE 2.8 – Probabilités conditionnelles associées aux images de test ; la première ligne (a) correspond au cas non-fumeur et la seconde (b) au cas fumeur ; NDR ou "non-detection rate" correspond à la valeur de seuil pour laquelle le taux de non-détection est égale à zéro.

Résultats et analyses

Nous présentons dans la figure 2.8 les différentes valeurs associées aux probabilités conditionnelles $p(\omega_S|I)$ pour les deux groupes fumeur et non-fumeur dans le cas du set de test "images standards". Ces valeurs varient de 0.51 à 0.98 pour le groupe non-fumeur et de 0.53 à 0.99 pour le groupe fumeur. Pour les deux classes sain et pathologique, nous constatons que les valeurs de probabilité associées aux images mal classées varient de 0.53 à 0.89 pour le groupe non-fumeur et de 0.52 à 0.61 pour le groupe fumeur. Ainsi, l'application d'un seuil de rejet adapté (0.89 pour le cas non-fumeur et 0.61 pour le cas fumeur) permet au système de ne commettre aucune erreur de classification. En particulier, ce seuil permet d'obtenir un taux de non-détection nul (indiqué par une ligne pointillée dans la figure 2.8. Notons cependant qu'il faut rejeter 36.1% d'images correctement classées (soit 48/133) pour la classe non-fumeur (première ligne) et 5.4% (soit 5 images sur 93) pour la classe fumeur.

Dans le tableau 2.13 sont présentées les matrices de confusion pour les jeux "images standards" (a) et "fenêtres" (b). Comme attendu, les résultats sont légèrements inférieurs pour les fenêtres.

TABLE 2.13 – Matrices de confusion des cas pathologiques (P) et sains (S) pour les jeux "images standards" (a) et "fenêtres" (b). Pour le jeu "fenêtres", nous rappelons que les effectifs sont 100 fois plus importants et qu'une fenêtre extraite est considérée comme une image de test complètement indépendante.

		Non-fumeur		Fumeur	
		Sortie classifieur		Sortie classifieur	
		P	S	P	S
Vérité terrain	P	90.4%	9.6%	97.5%	2.5%
	S	3.3%	96.7%	3.9%	96.1%

(a) Jeu d'images standards

		Non-fumeur		Fumeur	
		Sortie classifieur		Sortie classifieur	
		P	S	P	S
Vérité terrain	P	86.78	13.22%	92.2%	7.78%
	S	7.06%	92.94%	6.38%	93.6%

(b) Jeu de fenêtres

Les variations des taux de non-détection et de fausse alarme en fonction du taux de rejet sont données, le cas du jeu “fenêtres‘, dans les figures 2.9 et 2.10 pour les groupes non-fumeur et fumeur, respectivement. Dans le cas fumeur, on peut observer que pour 10% des fenêtres rejetées (soit 900 fenêtres sur les 9000), le taux d’erreur (que ce soit la non-détection ou la fausse alarme) est presque divisé par deux. Dans le cas non-fumeur, diviser l’erreur par 2 nécessite de rejeter 20% de fenêtres (soit 2600 fenêtres sur les 13000). On observe également que dans les deux cas, le taux d’erreur ne peut être annulé complètement car nous ne pouvons pas rejeter au-delà de 30% des données. En effet, cela s’explique par le fait qu’au delà de 30% de rejet, toutes les données ont un taux de confiance supérieur à 99%, notamment des données mal classées. Ainsi, dans le groupe fumeur, 121 fenêtres sur les 1299 qui ont été mal classées ont une probabilité égale à 1 indiquant que ces fenêtres ne peuvent être rejetées.

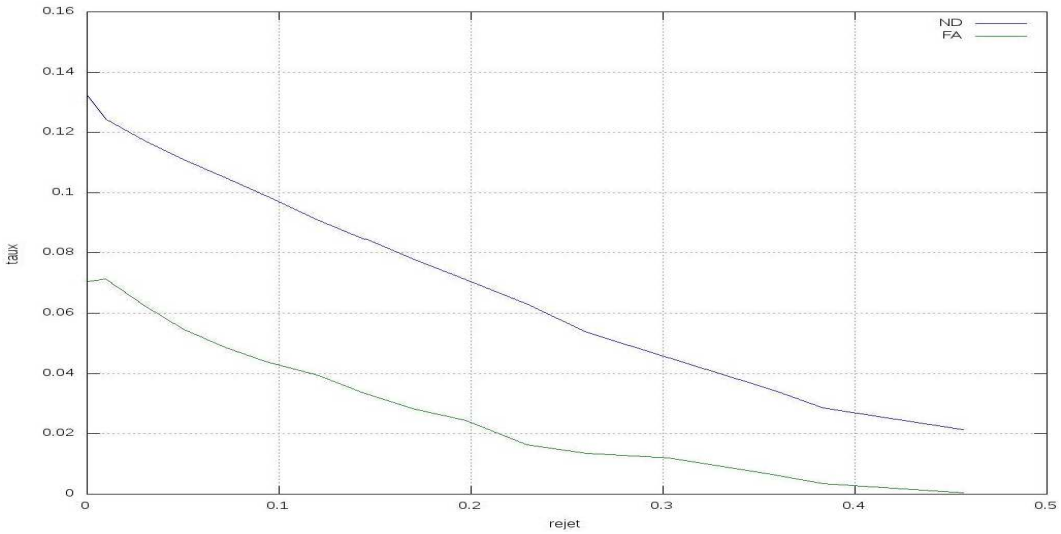


FIGURE 2.9 – Taux d’erreur de non-détection (ND) et fausse alarme (FA) pour le groupe non-fumeur

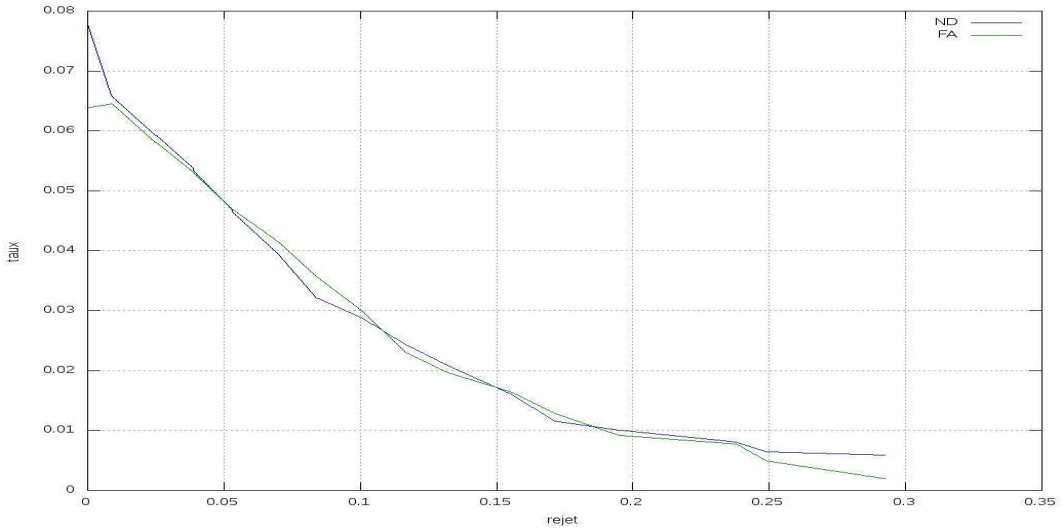


FIGURE 2.10 – Taux d’erreur de non-détection (ND) et fausse alarme (FA) pour le groupe fumeur

Conclusion

Nous constatons que la forêt classique, avec le mode de vote binaire, ne peut tenir compte de l’homogénéité de la distribution des populations au sein des feuilles des arbres. Nous proposons

ainsi de modifier le mode de vote interne des arbres en adoptant au lieu d'un vote binaire classique (0 ou 1), un vote basé sur la fréquence de chacune des populations en présence. Pour rendre cette approche significative, il est impératif de disposer d'un échantillon suffisant de données dans les feuilles des arbres. C'est la raison pour laquelle nous proposons, dans la section suivante, d'étudier les effets de l'élagage des arbres sur le pilotage du rejet. Nous avons étudié les premiers résultats obtenus avec la mise en place de notre mécanisme de rejet basé sur le taux de confiance du classifieur extra-trees associé à une donnée de test. Nous avons utilisé le jeu de fenêtres plutôt que le jeu d'images standards afin de mieux apprécier l'impact du mécanisme de rejet sur les résultats de bonne classification.
Nous étudions donc dans la section suivante, le mode de vote interne des arbres, i.e. la prise en compte de la population à chaque nœud terminal de l'arbre lors de son induction. La taille maximale de la population en chaque nœud terminal [5] peut être déterminé par le critère lié au paramètre n_{min}, le partitionnement du nœud courant ne s'effectuant que si la taille de la population en ce nœud est supérieure à n_{min}.

2.4.2 Élagage des extra-trees et mécanisme de vote des arbres

Nous avons vu dans les expérimentations précédentes que la considération d'une mesure de confiance associée à la décision de la forêt a un impact favorable sur la qualité de la décision. Nous avons ainsi introduit la notion de rejet permettant d'écarter des exemples difficiles à classer par la forêt. Nous cherchons maintenant à exploiter tout le mécanisme d'induction de la forêt afin de combiner et ainsi fusionner les taux de confiance associés à chacune des décisions de l'arbre, dès la phase d'apprentissage.
Nous savons que chaque feuille de l'arbre de décision est associée à une décision binaire indiquant la classe majoritaire présente au sein de la feuille. Il est tout à fait possible par conséquent, de façon similaire à ce qui a été fait précédemment de considérer une décision qui ne serait plus forcément binaire, via un vote majoritaire sur la population présente au sein du nœud terminal, mais une décision fréquentielle. Nous étudions deux mécanismes de vote : le mode de vote binaire classique de l'arbre et un mode de vote fréquentiel. Le vote binaire est l'approche standard de décision où on attribue la classe majoritaire à chaque nœud terminal de l'arbre. Dans le mode de vote fréquentiel proposé, chaque nœud terminal de l'arbre se voit attribuer un couple réel correspondant à la fréquence de chacune des classes en présence. Cela permet donc de tenir compte de l'entropie de la population présente dans chacun des nœuds terminaux de l'arbre, laquelle est susceptible d'influer sur le degré de confiance pour la classe attribuée lors du vote final de la forêt.

Le vote binaire classique correspond à une décision que l'on peut qualifier de "certaine" égale à 0 ou à 1 et un vote fréquentiel prenant ses valeurs dans l'intervalle réel [0;1]. Nous conservons l'expression de la probabilité conditionnelle $p(\omega_S|x_i) = \frac{l_i^S}{L}$, excepté que l_i^S n'est plus le nombre d'arbres de la forêt ayant voté pour la classe ω_S pour l'exemple x_i comme dans le cas précédent dit "binaire", mais la somme des fréquences associées à la classe ω_S par chacun des arbres. Nous avons ainsi $l_i^S = \sum_{k=1}^{L} f_{ik}^S$, où f_{ik}^S est la fréquence de la population associée à la classe ω_S de la feuille de l'arbre k dans laquelle est tombée l'exemple x_i. Lorsque l'arbre est complètement développé, comme c'est le cas par défaut pour la méthodes des forêts, les feuilles de l'arbre sont presque toutes pures ou bien ont une mesure de confiance élevée, i.e. f_{ik} proche de 1 pour la classe majoritaire. Ainsi, dans le cas des arbres non élagués (i.e. complètement développés), on ne s'attendra pas à une différence notable entre le vote binaire et le vote fréquentiel dans les résultats. C'est la raison pour laquelle nous introduisons l'élagage de l'arbre. En effet, l'élagage des arbres permet d'augmenter la population au sein de chaque nœud terminal offrant par conséquent un échantillon plus large pour le calcul des fréquences. Nous étudierons l'impact des deux modes de vote sur les performances du classifieur en fonction du degré d'élagage que nous appliquerons en faisant varier le paramètre n_{min}, le nombre de données minimal pour la coupure en un nœud.
Il est à noter que Breiman et al. [Breiman et al., 1984] préconisent l'élagage de l'arbre dans le

5. Á l'exception d'un nœud pur dans lequel toutes les données sont de la même classe

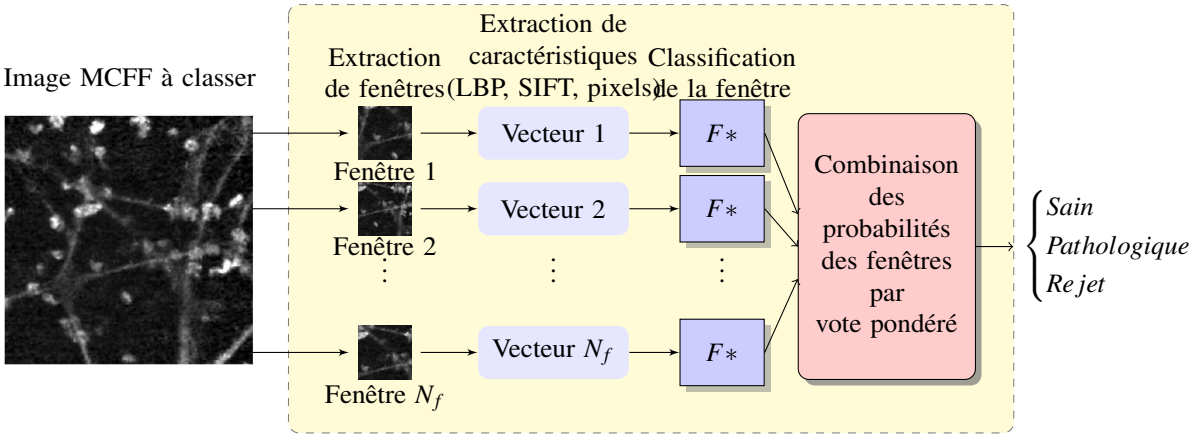


FIGURE 2.11 – Système de classification des images alvéoscopiques avec mécanisme de pilotage du rejet ; $F*$ désigne la forêt des extra-trees élagués ; la règle de décision de la forêt est basée sur le choix d'un seuil a priori : au dessus de ce seuil, la classe sain ou malade est attribuée à l'image, autrement elle est rejetée et peut être traitée par une approche plus spécifique.

cadre d'un unique classifieur car un arbre trop développé aura tendance à sur-apprendre le set d'apprentissage et par conséquent aura tendance à augmenter son taux d'erreur en généralisation. L'amélioration des performances de la forêt par l'introduction de l'élagage est de plus rendue difficile en raison du fait que dans l'absolu, la forêt ne sur-apprend pas [Breiman, 2001, Biau, 2010]. Il est à noter qu'une étude similaire de l'influence de l'élagage sur les performances de la forêt a été menée par Geurts et al. [Geurts et al., 2006] et avait conclu que le choix de l'élagage (avec notamment le paramètre n_{min}) peut être orienté en fonction du degré de données bruitées présentes dans la base d'apprentissage. En effet, sur des données extrêmement bruitées, les extra-trees peuvent sur-apprendre. Ainsi, plus les données semblent bruitées, plus grande semble être la valeur optimale n_{min}. Enfin, le choix de l'élagage ou non des arbres ne changent pas les propriétés de convergence statistique des forêts.

L'approche que nous proposons répond à deux objectifs : fiabiliser une mesure de confiance associée à un mécanisme de rejet et diminuer le taux d'erreur en généralisation de la forêt. Nous illustrons les modifications proposées sur la Figure 2.11, avec l'élagage des arbres, la combinaison des probabilités des fenêtres par vote pondéré (pondération en raison du vote fréquentiel) et la règle de décision impliquant le mécanisme de rejet.

Pour des raisons purement techniques, nous considérons dans les expériences suivantes de nouveau le jeu "fenêtres" afin de disposer de davantage de données pour établir notre analyse.

Protocole expérimental

Nous travaillons dans cette expérience avec les données du protocole précédent d'évaluation des extra-trees (cf. tableau 2.12). Nous considérons l'extraction locale et dense des fenêtres sur les images alvéoscopiques et leur description par l'extracteur issu du LBP. Les arbres sont post-élagués, c'est à dire qu'ils sont développés complètement puis de bas en haut, nous transformons en feuille les nœuds parents. L'avantage de cette approche par rapport à une approche plus en amont de pré-élagage (i.e. de haut en bas en empêchant l'arbre de se développer complètement) est le fait de pouvoir tenir compte, à chaque niveau d'élagage de tout le contenu informatif que l'arbre a pu capturer. En effet, dans le cas d'une approche de haut en bas comme le pré-élagage, un nœud peut être pauvre en terme de qualité du critère de partitionnement trouvé alors que le fils peut obtenir un critère de plus grande qualité (score plus important). Nous considérons comme mesure d'élagage le nombre de données minimales présentes à un nœud pour autoriser le partitionnement. Ce nombre de données minimales sera contrôlé par le paramètre n_{min} qui vaut 2 à l'état d'un arbre complètement

développé. Nous faisons varier ce paramètre dans l'intervalle $[2;1520]$ afin de contrôler le degré d'élagage des arbres de la forêt.
Afin de mieux percevoir les effets de l'élagage sur le pilotage du rejet, nous utilisons le jeu "fenêtres" défini précédemment. Ainsi, au lieu que les fenêtres d'une image dans le set de test contribuent à classer cette image, nous considérons les fenêtres extraites comme étant totalement indépendantes et donc comme autant d'images constituant le set de test. Nous perdons cependant la robustesse de la décision pour l'image de test d'origine mais notre objectif est de montrer la pertinence du mécanisme de pilotage du rejet en disposant d'un grand nombre d'échantillons en test.

Résultats et analyses

Les taux de non-détection et de fausse alarme en fonction du taux de rejet sont donnés dans les Figures 2.12 et 2.13 (fumeurs) et 2.12 et 2.15 (non-fumeurs), respectivement, pour les 2 types de vote, binaire (gauche) et fréquentiel (droite), pour différentes valeurs de n_{min}. Dans le cas fumeur, l'analyse des Figures 2.12 et 2.13 indiquent que ce rejet plus important permet de diminuer le taux d'erreur dès 4% de données rejetées, notamment l'erreur due de la non-détection. Dans ce dernier cas, le taux d'erreur peut même être réduit à zéro dès 20% de données rejetées, ce qui n'est pas possible dans le cas du mode de vote binaire. Nous voyons également, dans le cas du mode de vote fréquentiel, qu'un faible élagage est suffisant pour réduire l'erreur globale. En effet, dès $n_{min} = 80$ et jusqu'à $n_{min} = 320$ les plus bas taux d'erreur sont obtenus (les courbes intermédiaires confirment ce résultat ; elles ne sont pas affichées par soucis de lisibilité). Cependant, un fort taux d'élagage (e.g. $n_{min} = 1520$) ne permet pas d'augmenter les performances dans le cas fumeur et est même quelque peu défavorable dans le cas non-fumeur. Nous constatons que seul le mode de vote fréquentiel permet de réaliser des gains de performances avec l'élagage. Une analyse similaire peut être faite pour le cas non-fumeur. En effet, en analysant les Figures 2.14 pour la non-détection et 2.15 pour la fausse alarme, nous voyons que l'approche par mode de vote fréquentiel permet de nouveau de réduire davantage l'erreur de classification.
Nous avons ainsi montré, dans le cadre de notre application, pour les deux bases fumeur et non-fumeur, que l'introduction du vote fréquentiel combiné avec de l'élagage permet de réduire le taux d'erreur global en procédant par rejet des exemples ayant un niveau de confiance trop faible. Notre méthode permet même l'annulation de cette erreur dans le cas du groupe fumeur, ce qui n'est possible ni avec le mode binaire ni avec l'approche classique dans laquelle l'arbre est complètement développé.

La complexité de l'arbre de décision binaire correspond à sa profondeur, i.e. au nombre de feuilles construites. Nous avons tracé le nombre de nœuds total moyen construits par arbre lors de l'induction des extra-trees à chaque niveau d'élagage sur la figure 2.16 à la fois pour le cas fumeur et le cas non fumeur, dans le cas du mode de vote fréquentiel. Nous constatons une chute exponentielle du nombre de nœuds pour les deux groupes, passant de 12000 pour l'arbre complètement développé à 1000 nœuds pour un faible élagage (3% de l'élagage total) dans le cas fumeur. Cette chute va de 20000 à 2000 nœuds pour le cas non-fumeur. Des résultat similaires sont obtenus pour le mode de vote binaire. Ainsi, avec un faible degré d'élagage, la complexité des extra-trees est drastiquement réduite.

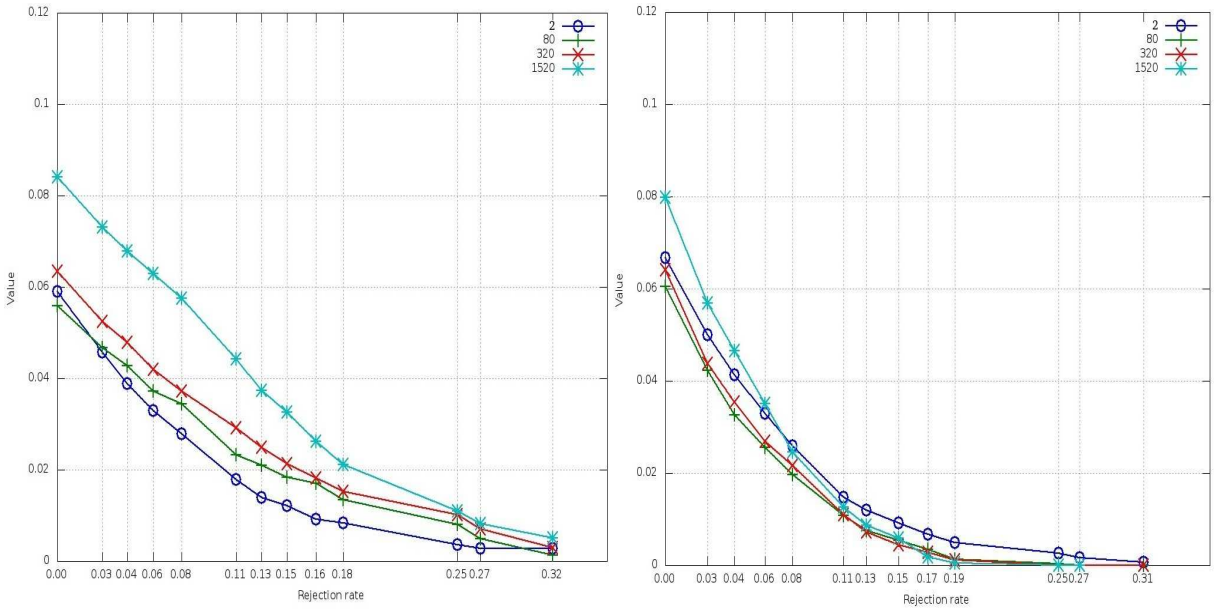


FIGURE 2.12 – Taux d'erreur de non-détection (faux négatifs) pour le mode de vote binaire (à gauche) et le mode fréquentiel (à droite) pour le groupe fumeur, en fonction du taux de données rejetées, pour différentes valeurs de n_{min}.

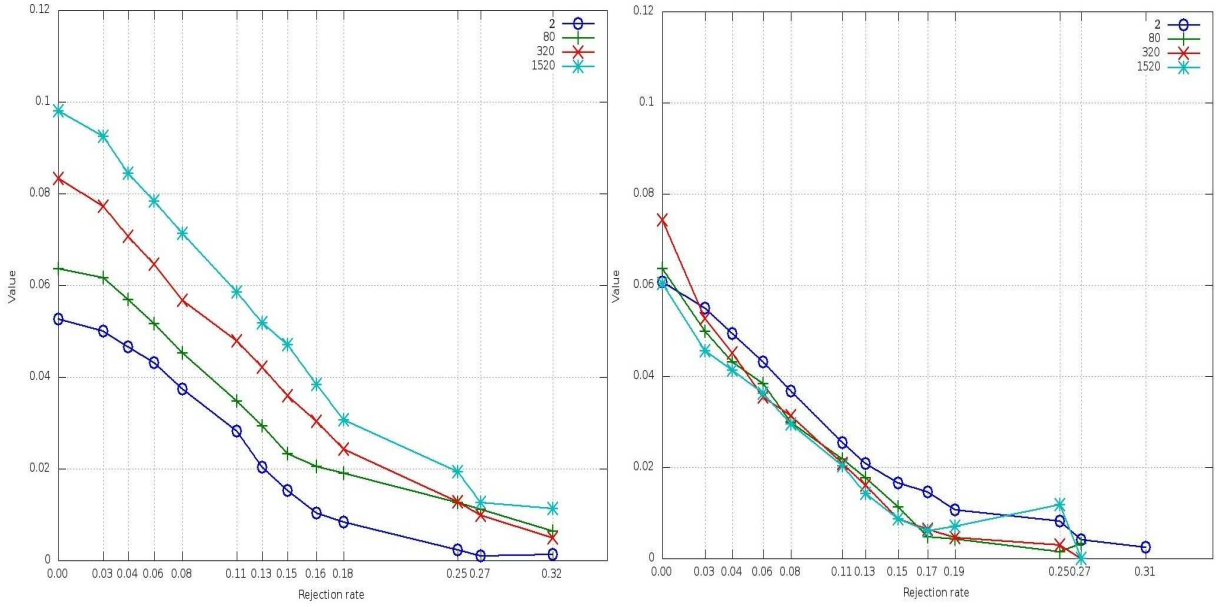


FIGURE 2.13 – Taux d'erreur de fausse alarme (faux positifs) pour le mode de vote binaire (à gauche) et le mode fréquentiel (à droite) pour le groupe fumeur, en fonction du taux de données rejetées.

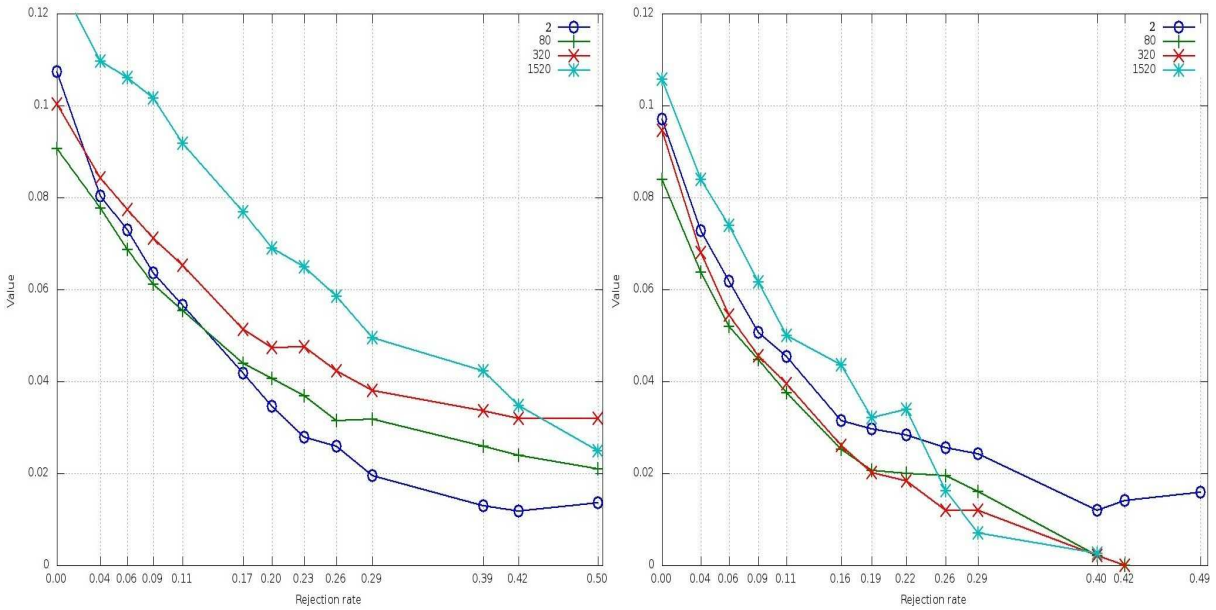


FIGURE 2.14 – Taux d'erreur de non-détection (faux négatifs) pour le mode de vote binaire (à gauche) et le mode fréquentiel (à droite) pour le groupe non-fumeur, en fonction du taux de données rejetées, pour différentes valeurs de n_{min}.

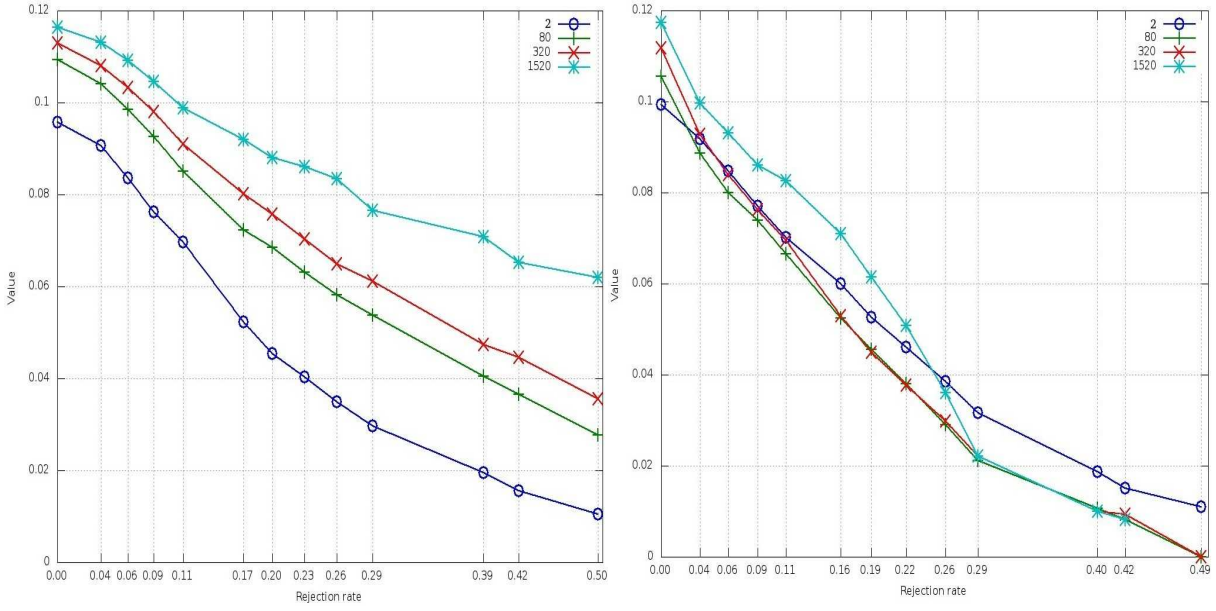


FIGURE 2.15 – Taux d'erreur de fausse alarme (faux positifs) pour le mode de vote binaire (à gauche) et le mode fréquentiel (à droite) pour le groupe non-fumeur, pour différentes valeurs de n_{min}.

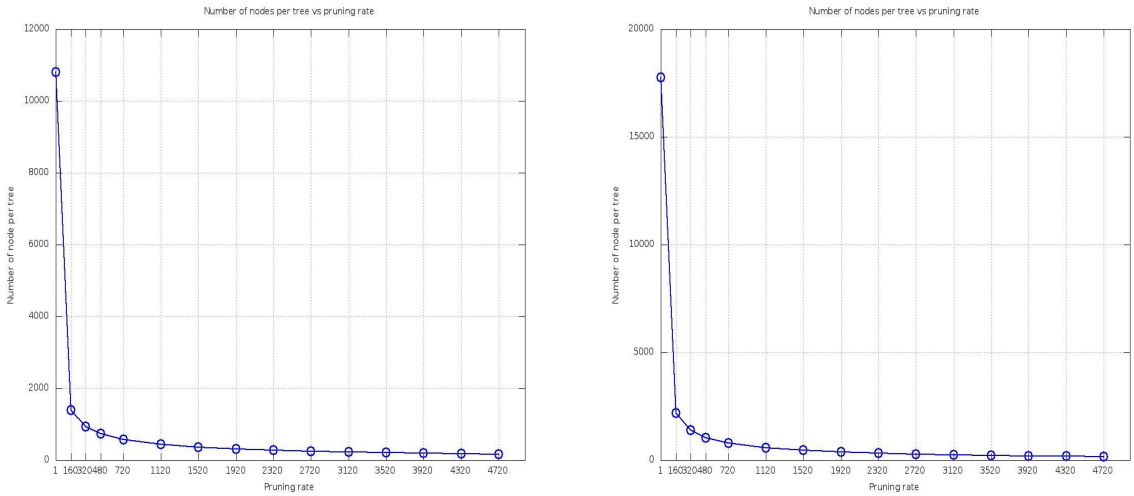


FIGURE 2.16 – Nombre de nœuds moyens par arbre extra-tree en fonction du niveau d'élagage pour le groupe fumeur (gauche) et le groupe non fumeur (droite) pour le mode de vote séquentiel.

2.5 Conclusion

Dans ce chapitre, nous avons proposé et mis en œuvre un système de classification d'images pour traiter notre problématique de classification des images alvéoscopiques. Ce système s'est révélé performant et robuste. Dans la phase d'extraction de caractéristiques, nous nous sommes inspirés d'une approche d'extraction locale de fenêtres dans l'image pour proposer une description plus riche des images alvéoscopiques à l'aide de l'extracteur issu du LBP. Puis dans la phase d'apprentissage, nous avons proposé, en nous basant sur l'approche de forêts aléatoires extra-trees, un mécanisme de pilotage du rejet permettant de quantifier le degré de confiance associée à la décision du classifieur. Enfin, nous avons proposé une modification du mode interne de vote des arbres de la forêt d'extra-trees en y associant un mécanisme d'élagage et permettant d'améliorer le pilotage du rejet. Ces mécanismes mis en place ont permis de répondre favorablement au problème de non détection élevée que nous avions rencontré dans la première phase d'évaluation du système de classification.

Dans la suite, nous envisageons une approche plus adaptée à la détection des cas pathologiques. Nous ne disposons pas à l'heure actuelle de sémiologie univoque déterminée par les pneumologues permettant d'associer à une image donnée une pathologie pulmonaire. De plus, nous avons constaté des difficultés d'étiquetage des images de patients malades pour des raisons médicales car des territoires sains peuvent être traversés lors de l'exploration de territoires pathologiques chez un patient malade. Nous avons alors conclu avec les experts médicaux que seules les images de patients sains peuvent être étiquetées comme étant saines avec certitude et servir de référence pour élaborer un modèle de décision.
En partant de cette analyse, nous envisageons une approche one-class, dans laquelle le classifieur n'est en présence que d'une seule classe qui va servir de référence en phase d'apprentissage (dans notre cas cette classe de référence n'est composée que des images des cas sains). La méthode doit alors distinguer cette classe des autres classes dites "outlier" qui lui seraient présentées lors de la phase de décision. Nous présentons dans le Chapitre 3 un état de l'art des approches one-class nous permettant de sélectionner, parmi les approches existantes, les méthodes ou les approches intéressantes pour l'élaboration du système de classification one-class que nous proposons dans le Chapitre 4.

Chapitre 3

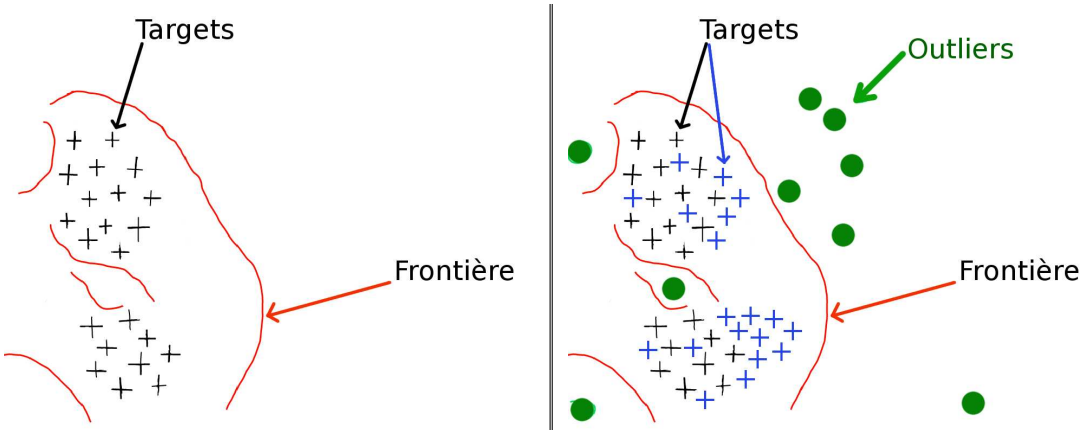
L'approche one-class

Sommaire

3.1 Introduction

Les expérimentations menées dans la section précédente ont montré que la classe des données pathologiques demeurait encore difficile à identifier. La solution de rejet ne répondant que partiellement à ce problème, nous proposons dans ce chapitre de considérer notre problématique sous l'angle one-class. Dans une tâche de classification à deux classes, le classifieur élabore sa frontière de décision en s'appuyant sur la distribution des classes en présence. Dans la classification one-class, le classifieur ne dispose que d'une seule classe correctement définie dite "target", l'autre classe, dite "outlier" étant absente. Il s'agit alors d'élaborer une frontière de décision en ne partant que de cette seule classe d'apprentissage.
La problématique one-class est fréquente dans la littérature notamment dans les domaines où une classe de données à identifier n'est pas représentée car impossible à obtenir pour des raisons de coûts ou pour des raisons pratiques. Par exemple l'obtention de données de pannes d'une machine industrielle ne peut se faire qu'en détériorant de toutes les façons possibles cette machine. Le but de l'approche one-class est donc de pallier à l'absence de ces données outliers en tirant partie uniquement des données targets disponibles.

Nous illustrons sur la figure 3.1 la problématique one-class. Dans cette figure, nous représentons la classe target constituée d'objets "+". L'objectif est de pouvoir placer une frontière de décision (trait en rouge sur la figure) dans l'espace de description en l'absence de données outliers ("ronds" en vert). Nous montrons particulièrement sur cette figure la difficulté de cette tâche en indiquant un outlier situé en plein milieu du domaine des targets, absent lors de la phase de construction de cette frontière.



(a) Données d'apprentissage + frontière (b) Données de test

FIGURE 3.1 – Illustration de la problématique one-class dans laquelle seules les instances targets (symbole "+") sont disponibles lors de la phase d'apprentissage ; le classifieur doit apprendre la frontière de décision (trait continu en rouge) à partir de ces seules données disponibles (figure a), i.e en l'absence des données outliers (figure b) ; en phase de test (figure b), les instances des deux classes target ("+" bleu) et outlier ("rond" vert) sont disponibles.

Différentes catégorisations des approches one-class existent dans la littérature [Chandola et al., 2009a, Chandola et al., 2009b, Tax and Duin, 2004]. Nous retrouvons communément la catégorisation suivante en trois groupes [Tax and Duin, 2004, Hodge and Austin, 2004] : les méthodes par estimation de densité qui établissent un modèle de densité de la classe des données targets ; les méthodes par reconstruction qui reproduisent en sortie du classifieur la donnée d'entrée et les méthodes frontières cherchant à circonscrire les données targets dans l'espace de description.
Dans ce chapitre nous avons choisi une catégorisation différente, se focalisant principalement sur notre problématique de gestion des données outliers. Ainsi, nous présenterons les méthodes de la littérature en fonction de la manière dont elles tiennent compte de la classe des outliers lors de la phase d'apprentissage. Nous mettrons alors en lumière la difficulté d'utilisation des approches

discriminantes pour traiter le problème one-class, puisque ces approches nécessitent la synthèse des données outliers durant la phase d'apprentissage. Nous observons que les méthodes de combinaison et particulièrement les méthodes d'ensembles, n'ont été que très rarement mentionnées dans la littérature pour traiter la problématique one-class. La plupart des méthodes proposées sont des combinaisons d'approches déjà one-class n'exploitant pas les mécanismes de combinaison et les divers principes de randomisation que proposent notamment les méthodes d'ensembles comme le Bagging [Breiman, 1996] ou le Random Subspace Method [Ho, 1998] pour élaborer une méthode d'ensemble dédiée au one-class. Nous traitons des méthodes existantes et discutons des principes de combinaison dont pourraient bénéficier notamment les approches discriminantes au sein d'une combinaison de classifieurs. Ainsi, à l'issue de cet état de l'art, nous proposons un cadre de travail dédié au one-class permettant de tirer partie des différents mécanismes de combinaison des méthodes d'ensembles et particulièrement des principes de randomisation des méthodes de forêts aléatoires afin d'élaborer une nouvelle approche one-class que nous mettons en œuvre dans le chapitre suivant.

Le chapitre est structuré comme suit : nous proposons dans la première section une nouvelle approche de catégorisation des méthodes one-class de l'état de l'art en classant les méthodes selon le positionnement adopté vis-à-vis des outliers. Cette catégorisation différente a pour objectif de mettre en avant la manière dont les approches de l'état de l'art traitent la problématique des données outliers, absentes lors de la phase de construction de la méthode de classification. Nous abordons notamment l'utilisation très peu faite des méthodes de combinaison de classifieurs pour traiter le problème one-class. Enfin, dans la dernière section, en nous basant sur l'étude de l'état de l'art, nous proposons un système de classification one-class utilisant les méthodes performantes et génériques d'ensembles d'arbres de décision et tirant pleinement partie des mécanismes d'ensembles que sont notamment la randomisation ou l'injection d'aléatoire et la combinaison des décisions individuelles des classifieurs.

3.2 Catégorisation des méthodes one-class

L'approche one-class est souvent mentionnée dans des tâches impliquant des données déséquilibrées, dans le domaine médical notamment où les données de patients malades sont difficiles et coûteuses à échantillonner et dans l'industrie lourde (moteurs d'avions, ponts, machines industrielles).

Un nombre important d'études et de revues ont été publiées dans la littérature durant les 10 dernières années au sujet de l'approche one-class [Khan and Madden, 2010, Mazhelis, 2006]. Ces études concernent principalement la détection d'outlier [Chandola et al., 2009b, Hodge and Austin, 2004], la détection d'anomalies [Chandola et al., 2009a], la détection de nouveauté dans [Marsland, 2003] et [Markou and Singh, 2003].

Le terme one-class semble avoir été mentionné pour la première fois dans Moya et al. [Moya et al., 1993] et repris dans diverses revues générales de la problématique one-class [Moya and Hush, 1996, Markou and Singh, 2003]. On retrouve cependant plusieurs synonymes se différenciant principalement au niveau de l'application ou du domaine traité [Tax et al., 1999, Sillito and Fisher, 2007, Hempstalk and Frank, 2008]. On retrouve ainsi les termes "détection de nouveauté" ("novelty detection") [Japkowicz, 1999, Markou and Singh, 2003, Tarassenko et al., 1995], "détection d'intrusion" dans les réseaux informatiques [Fan et al., 2004], "détection de défauts" dans les machines industrielles [Taylor and Addison, 2000], "détection d'anomalies" dans le domaine médical [Tarassenko et al., 1995], "détection d'outliers" dans l'industrie [Bishop, 1994, Taylor and Addison, 2000, King et al., 2002, Carpenter et al., 1997] et le terme "data description" [Tax and Duin, 1999] ; on retrouve le terme d'origine "one-class" dans la reconnaissance de visages [Mamloukf et al., 2003] et la détection d'expressions faciales d'émotions spontanées [Zeng et al., 2006], l'identification de cibles radars [Moya et al., 1993], dans la reconnaissance de l'écriture manuscrite dans les codes postaux (US) [Scholkopf et al., 2001].

Outre les applications citées ci-dessus, nous pouvons ajouter les applications pour lesquelles l'approche one-class apparaît incontournable [Banhalmi et al., 2009] : la détection de plagiat ("au-

thorship verification“) [Koppel and Schler, 2004], la reconnaissance d'utilisateur par analyse du rythme de frappe au clavier (”typist recognition“) [Hempstalk et al., 2008], la reconnaissance du locuteur [Brew et al., 2007], la surveillance de l'intégrité d'un bâtiment, d'une structure ou du bon état de fonctionnement d'une machine industrielle [Toivola et al., 2010, Tarassenko et al., 2009, Taylor and Addison, 2000, King et al., 2002, Carpenter et al., 1997, Bishop, 1994]. Le point commun à toutes ces applications est la mise en œuvre d'un système de reconnaissance capable d'identifier une unique classe d'intérêt ou de détecter des exemples d'une classe mal définie ou difficile à échantillonner.

Une des propriétés principalement recherchées pour un algorithme d'apprentissage est sa capacité à généraliser. C'est-à-dire sa capacité à traiter correctement des données nouvelles. Moya et al. [Moya et al., 1993] considèrent que le problème one-class ne peut se traiter qu'en vertu de la résolution de trois types de généralisation. Les auteurs ont donné ces considérations dans le cadre des réseaux de neurones mais elles peuvent s'étendre à d'autres types d'approches de classification. Les auteurs considèrent i) la généralisation intra-classe où les éléments appartenant à une même classe sont bien identifiés, ii) la généralisation inter-classe permettant de bien distinguer la frontière entre deux classes distinctes et enfin iii) la généralisation dite ”out-of class” (littéralement en dehors des classes) où les éléments n'appartenant à aucune des classes apprises ou disponibles sont bien identifiés. Cette distinction est importante notamment pour la généralisation “out-of-class“ qui se distingue de la généralisation inter-classe dans la mesure où la classe des outliers est à considérer comme un tout et non pas comme une instance particulière du concept à apprendre [1].

Dans les différentes catégorisations des approches one-class [Chandola et al., 2009a, Chandola et al., 2009b, Tax and Duin, 2004], nous retrouvons communément la catégorisation suivante en trois groupes [Tax and Duin, 2004, Hodge and Austin, 2004] : les méthodes par estimation de densité ; les méthodes par reconstruction et les méthodes frontières. Les méthodes par estimation de densité établissent un modèle de la classe target et déterminent un seuil adapté pour construire la frontière de décision. Les méthodes par reconstruction sont essentiellement des méthodes de reproduction des données targets à la sortie du classifieur. Un seuil est alors fixé sur l'erreur dans la reproduction des données d'entrée pour déterminer si la donnée présentée en test provient bien de la classe target. Les méthodes frontières regroupent principalement des méthodes discriminantes construisant directement la frontière de décision à partir des données d'apprentissage. Une approche discriminante est cependant plus difficile à mettre en œuvre dans le cadre one-class car deux classes sont requises pour construire cette frontière de décision. Comme les données de la seconde classe (ou “out-of-class“) ne sont pas disponibles, il est donc nécessaire pour ce type d'approche de les synthétiser. Nous distinguons deux façons de synthétiser ces données outliers : la génération physique selon une distribution donnée ou leur simulation.

Il n'existe pas à notre connaissance d'étude sur la catégorisation des méthodes one-class discriminantes insistant sur ces différents modes de synthèse des données outliers. C'est la raison pour laquelle nous proposons une telle catégorisation, classant les méthodes one-class selon la manière dont les outliers sont pris en compte durant la phase d'apprentissage. Ainsi, nous répartissons les approches one-class selon trois critères : (i) les approches nécessitant la présence physique de données outliers dans le set d'apprentissage et donc leur génération ; (ii) les approches ne nécessitant pas la génération des données outliers mais seulement l'hypothèse de leur présence dans l'espace de caractéristiques selon une distribution à définir (cela revient à la simulation de données outliers) ; (iii) les méthodes ne faisant aucune hypothèse sur les données outliers.

Comme nous le verrons par la suite, les approches génératives par estimation de densité [Duda et al., 2000] comme l'estimateur gaussien, celui de Parzen ou la mixture de gaussiennes, les approches par reconstruction et certaines approches de type “data description“ comme l'algorithme SVDD de Tax et al. [Tax and Duin, 1999, Tax and Duin, 2004] basé sur le SVM se placent dans la catégorie (iii) car le seuil sur la probabilité ou la distance au modèle ou l'erreur en reconstruction peut être choisi empiriquement sans nécessiter l'hypothèse de données outliers ; une approche

1. La définition ici de l'outlier repose sur la description de ce qu'il n'est pas, i.e. ce n'est pas un élément target

discriminante standard ayant besoin des deux classes pour tracer la frontière de décision sera classée généralement dans la catégorie (i) ; une approche faisant l'hypothèse d'une distribution spécifique des outliers sans les générer, uniforme dans le cas de Liu et al. [Liu et al., 2000], réduite à un point dans le cas du one-class-svm [Scholkopf et al., 2001], sera classée dans la catégorie (ii).

Nous détaillons ci-après les méthodes associées aux catégories que nous avons proposées.

3.2.1 Méthodes sans outliers

Parmi les méthodes ne nécessitant pas la génération d'outliers, nous retrouvons naturellement les approches génératives et notamment les estimateurs de densité. En effet, ces approches construisent un modèle de la classe target et définissent un seuil permettant de faire du rejet de distance au modèle. Ce seuil est en général défini par des mécanismes de validation croisée comme le leave-one-out ou par une valeur déduite du rejet d'un faible pourcentage de données pourtant targets (légitimes) mais considérées comme outliers pour les besoins de la méthode. Ces méthodes, immédiatement mises en œuvre pour le one-class, présentent cependant des difficultés d'adaptation aux propriétés des bases de données disponibles. En effet, ces méthodes ne sont pas efficaces en grande dimension en raison du nombre important de données requises pour obtenir une estimation fiable de la distribution des données en présence. Nous discutons des méthodes comme le SVDD proposé par [Tax and Duin, 2004], des méthodes adaptées au one-class basées sur l'évaluation d'une distance comme le plus proche voisin et des approches dites par reconstruction dans lesquelles le classifieur reproduit à sa sortie la donnée d'entrée, un seuil sur l'erreur de reproduction permettant d'identifier les données outliers.

3.2.1.1 Approches par estimation de densité

Les approches génératives construisent un modèle qui se veut le plus proche possible de la distribution des données d'apprentissage. C'est donc tout naturellement que ces méthodes sont proposées pour traiter l'approche one-class. Parmi les approches génératives, les méthodes par estimation de densité sont les plus citées. Cependant, elles ne doivent pas se présenter comme des solutions systématiques pour des tâches et des domaines aussi divers et variés que ceux que nous avons cités plus haut dans l'exposé. Notamment, les méthodes par estimation de densité s'adaptent difficilement à la grande dimension [Markou and Singh, 2003, Nairac et al., 1997, Tax and Duin, 1999].

Ce problème est notamment connu sous le nom de malédiction de la dimension [Bellman, 1961]. Typiquement, le nombre d'observations doit être considérablement plus grand que la dimension de l'espace de description des données, sous peine de ne pouvoir fiabiliser le modèle construit à partir des données disponibles. Ce problème liée à la dimension est encore plus délicat lorsque l'ont tient compte du phénomène appelé "phénomène d'espace vide" ou "empty space phenomenon" [Verleysen et al., 2003, Donoho, 2000]. Il s'agit de la création de zones vides (sparses) dans l'espace qui est la conséquence de l'agrandissement du volume du domaine avec la dimension et non suivie de l'augmentation en terme exponentiel du nombre d'exemples observés dans ce même domaine. Une approche travaillant sur la totalité de ce type d'espaces doit donc pouvoir s'adapter à ce phénomène. Particulièrement, les approches semi-paramétriques ou non-paramétriques sont coûteuses par exemple pour l'estimation de coefficients dans le cas des mixtures de modèles, pour la sélection de la largeur de bande dans le cas des fenêtres de Parzen. Ainsi, ces approches nécessitent un grand nombre de données afin d'obtenir une bonne estimation de la distribution des données targets [Duda and Hart, 1973]. Dans le cas des approches paramétriques, il est nécessaire d'opérer un choix de modèle. En ce sens, il s'agit de méthodes d'estimation de paramètres d'une fonction de densité connue. Ce modèle choisi peut ne pas être adapté à la distribution (inconnue) des données du problème et entraîner ainsi une mauvaise estimation de la distribution des données.

Outre le choix d'un modèle adapté, il est nécessaire d'établir un seuil sur la probabilité calculée à l'aide de ce modèle afin de classer toute donnée nouvelle (cf. figure 3.2). Ce seuil, qui peut être fixé globalement ou bien calculé localement pour tenir compte des informations locales présentes

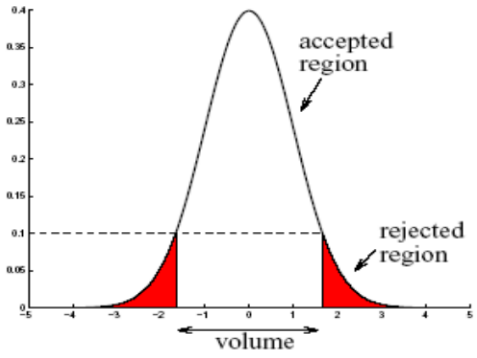


FIGURE 3.2 – Approche par estimation de densité de la classe cible et seuil de rejet des outliers ; on obtient ainsi une région cible et une région de rejet.

dans les données [Bishop, 1994, Tax and Duin, 1998, Tarassenko et al., 2009], est soit à définir par l'utilisateur, soit estimé à partir des données d'apprentissage. Il est à noter que la valeur de ce seuil est déterminante pour les performances du classifieur car il détermine la valeur à partir de laquelle une donnée nouvelle ne vérifie pas le modèle établi. Cependant, ce choix de seuil est plus difficile à faire dans le cas one-class que dans le cas de classification binaire standard en raison de l'absence des données de la seconde classe. Plusieurs approches ont été proposées dans la littérature pour fixer la valeur de ce seuil, basées pour la plupart sur l'hypothèse qu'une fraction des données targets n'est pas représentative des données de la classe target, cette fraction est alors considérée comme des outliers (e.g. les données de plus faibles probabilités) [Tax and Duin, 1999, Tax and Duin, 2002]. Dans ce dernier cas, la fraction de données targets rejetées constitue une estimation de l'erreur qui sera commise lors de la phase d'identification des données targets. L'un des avantages de cette approche est de pouvoir déterminer la valeur de ce seuil durant la phase d'apprentissage de la méthode et donc cette valeur est adaptée à la configuration actuelle des données en présence. Ce seuil est bien entendu également ajustable en phase de prédiction.

Nous présentons ci-après les approches par estimation de densité suivantes : l'estimateur gaussien, la mixture de gaussiennes et les fenêtres de Parzen.

L'estimateur gaussien

Dans le modèle gaussien, il s'agit de modéliser les données cibles selon une distribution normale en dimension d en estimant les deux paramètres que sont la matrice de variance-covariance Σ et le vecteur moyen μ [Parra et al., 1996]. L'expression de la fonction de densité (ou noyau gaussien) associée à cette distribution est donnée par :

$$\tilde{p}_G(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\}$$

Le nombre de paramètres (ou degrés de liberté) de la méthode vaut $N_{ddl} = d + d(d+1)/2$ obtenu en additionnant le nombre de valeurs à estimer pour le vecteur moyen (sa dimension d) et pour la matrice de variance-covariance qui est symétrique (d(d+1)/2). Le modèle gaussien suppose cependant que les données targets ne sont réparties dans l'espace que selon un unique groupe compact, i.e. une distribution unimodale. Les données réelles ne vérifiant généralement pas cette hypothèse, un modèle multi-modal, plus flexible, est envisagé avec la mixture de gaussiennes.

La mixture de Gaussiennes

L'approche par mixture de gaussiennes (ou "MoG" pour "Mixture of Gaussians") permet de rendre plus flexible le modèle utilisé pour estimer la distribution des données targets [Duda and Hart, 1973]. La méthode utilise la combinaison linéaire de plusieurs noyaux gaussiens afin de modéliser ces

données targets. Les paramètres sont le nombre K de noyaux avec les coefficients de pondération associés ainsi que les paramètres internes standards de chacun des noyaux, vecteurs moyens et matrices de variance-covariance comme précédemment. Ces paramètres sont estimés à partir des données d'apprentissage. Le nombre de noyaux est fixé en général par des procédés de validation croisée. Les paramètres internes de chacune des gaussiennes peuvent être estimés par la maximisation de la log-vraisemblance avec des techniques courantes comme l'algorithme Espérance-Maximisation (EM) [Dempster et al., 1977, Roberts and Tarassenko, 1994].

La mixture de K gaussiennes est donnée par la combinaison linéaire des densités gaussiennes ci-dessous :

$$\tilde{p}_{MoG_K}(\mathbf{x}) = \frac{1}{K} \sum_{i=1..K} \alpha_i \tilde{p}_{G_i}(\mathbf{x}; \mu_i, \Sigma_i)$$

Les coefficients α_i du mélange sont des paramètres aussi à estimer. Le nombre total de paramètres à estimer vaut alors vaut $N_{ddl} = (d + d(d+1)/2 + 1) \cdot K$. On opère généralement une simplification du modèle afin de réduire le nombre de paramètres à estimer en supposant par exemple que la matrice de variance-covariance est diagonale (on passe de d(d+1)/2 valeurs à estimer à seulement d valeurs et $N_{ddl} = (d + d + 1) \cdot K$), ou encore diagonale avec une même variance pour toutes les dimensions de l'espace. Dans ce dernier cas, il n'y a plus qu'une seule valeur de paramètre à estimer pour la variance covariance et on a, avec $\Sigma_k = \sigma_k \dot{I}_d$, I_d étant la matrice identité : $N_{ddl} = (d + 1 + 1) \cdot K$.
L'approche par mixture de gaussiennes offre l'avantage de fournir un modèle mieux adapté à la distribution des données targets plutôt que l'approche avec un unique noyau de l'estimateur gaussien. Cependant, en dimension élevée, les mixtures se révèlent très coûteuses dans la mesure où beaucoup de données sont nécessaires pour estimer les différents paramètres des noyaux du modèle [Markou and Singh, 2003]. Particulièrement, en présence d'un nombre insuffisant de données, l'approche par fenêtres de Parzen est privilégiée.

Les fenêtres de Parzen

L'estimateur de Parzen est une approche non-paramétrique souvent citée comme référence dans les approches one-class [Tarassenko et al., 2009, Cohen et al., 2008, Tarassenko et al., 1995, Bishop, 1994]. Il s'agit d'approximer la distribution réelle des données par une combinaison linéaire de densités construites autour de chaque exemple en apprentissage [Parzen, 1962]. Un noyau gaussien centré de covariance simplifiée ($\Sigma_k = h\dot{I}$) est souvent utilisé comme densité de base du mélange. L'algorithme calcule un même noyau pour chaque donnée d'apprentissage et en fait la somme. Ainsi, la densité obtenue est proche des données d'apprentissage mais la technique s'avère très coûteuse car elle nécessite le parcours de tous les exemples d'apprentissage pour chaque évaluation de la densité. Un noyau gaussien est utilisé par défaut et la largeur de bande est souvent estimée par une procédure en leave-one-out [Duin, 1976, Kraaijveld and Duin, 1991]. Les fenêtres de Parzen sont à conseiller au détriment des approches par mixture de gaussiennes lorsque le nombre de données disponibles est faible ; mais a contrario, l'estimateur de Parzen est coûteux et est moins favorable si un grand nombre de données est présent.

Le paramètre h contrôlant la largeur d'un noyau donné (ou fenêtre) est le seul paramètre à estimer de ce modèle simplifié. L'expression de la densité résultante, en reprenant les notations précédentes, est donnée par :

$$\tilde{p}_{Parzen_N}(\mathbf{x}) = \frac{1}{N} \sum_{i=1..N} \tilde{p}_{G_i}(\mathbf{x} - \mathbf{x_i}; 0, h \cdot I)$$

où x_i est un des exemples disponibles dans la base d'apprentissage et I la matrice identité.

De façon plus générale, on définit dans un premier temps une fonction noyau ϕ, vérifiant les conditions de densité :

- $\phi(u) \geq 0, \forall u \in \Re^d$
- $\int_{\Re^d} \phi(u) du = 1$

Il est courant de prendre le noyau suivant qui vaut 1 dans l'hypercube unité centré à l'origine et 0 ailleurs :

$$\phi(u) = \begin{cases} 1 & \text{si } ||u|| \leq 1/2 \\ 0 & \text{sinon} \end{cases}$$

Puis, on définit un hypercube de coté h, centré en la donnée x_i, qui a pour équation $\phi(\frac{x-x_i}{h}) = 1$, pour volume $V = h^d$ et un nombre d'observations k_N à l'intérieur de cet hypercube ($k_N = \sum_{j=1}^{N} \phi(\frac{x_j - x_i}{h_N})$), on a :

$$\tilde{p}_{Parzen_N}(\mathbf{x}) = \frac{k_N}{N \cdot V} = \frac{1}{N \cdot h^d} \cdot \sum_{i=1}^{N} \phi(\frac{x - x_i}{h_N})$$

La méthode est très sensible à la représentativité des données et les coûts de calculs sont prohibitifs : le calcul de la valeur de probabilité d'un nouvel exemple nécessite le parcours de tous les exemples d'apprentissage. Cette approche est utilisée par exemple dans [Cohen et al., 2008] pour traiter le problème de la surveillance des infections nosocomiales dans un cadre one-class. Les auteurs ont utilisé le noyau gaussien classique et ont évalué leur système avec plusieurs valeurs du seuil de rejet des éléments d'apprentissage pour la fonction de densité résultante. Ce paramètre de seuil est difficile à évaluer comme nous l'avons expliqué précédemment et a priori hautement dépendant du problème à traiter. Les auteurs ont cependant constaté qu'un seuil de rejet a priori de 50% donnait les meilleurs résultats pour leur problème à traiter.

3.2.1.2 Approches par estimation de distance

Le plus proche voisin a également été proposé pour traiter le problème one-class en considérant une estimation de distance [Markou and Singh, 2003, Tax and Duin, 2004]. L'approche proposée est une version adaptée de l'algorithme du plus proche voisin classique [Duda et al., 2000] ne tenant compte que des exemples targets disponibles. Pour cela, les auteurs définissent deux distances : la distance d'un nouvel exemple $\mathbf{x}$ à son plus proche voisin dans la base d'apprentissage notée $NN(\mathbf{x})$ et la distance de ce plus proche voisin à son plus proche voisin notée donc $NN(NN(\mathbf{x}))$. On peut définir alors la mesure :

$$\rho(\mathbf{x}) = \frac{||\mathbf{x} - NN(\mathbf{x})||}{||NN(\mathbf{x}) - NN(NN(\mathbf{x}))||}$$

où $||.||$ est une norme utilisée pour le calcul de la distance. La distance euclidienne est couramment utilisée. Si la première distance est plus grande que la seconde, on a $\rho > 1$, cela signifie que le nouvel exemple n'est pas aussi proche de son plus proche voisin que ne l'est ce dernier des autres exemples en apprentissage. Le nouvel exemple peut alors être considéré comme outlier.

Plusieurs inconvénients cependant affectent cette approche. L'inconvénient majeur, quoique non spécifique au one-class, réside dans la nécessité de parcourir tous les exemples d'apprentissage pour déterminer la classe d'un nouvel exemple. Ensuite, sa difficulté à gérer la grande dimension est bien connue avec notamment la perte du caractère significatif des valeurs de distance en grande dimension lorsque la distance au plus proche voisin n'est pas significativement différente de la distance au plus lointain voisin si la métrique n'est pas adaptée comme précisé dans [Hinneburg et al., 2000, Donoho, 2000, Beyer et al., 1999].
En effet, l'approche par estimation de distance est confrontée au phénomène plus général appelé "concentration de la mesure" [Verleysen et al., 2003, Donoho, 2000, Milman, 1998] dans laquelle la variance des mesures (distance, norme) sur un échantillon demeure fixe alors que leur moyenne augmente naturellement avec la dimension. Ainsi, la norme de tous les vecteurs de l'espace semble constante (concentrée autour de la moyenne de la distribution des normes) et c'est le cas aussi particulièrement de la norme euclidienne de la différence de deux vecteurs (e.g. la distance entre un exemple et son plus proche voisin et celle entre le même exemple et son plus lointain voisin). Enfin, spécifiquement dans cette version modifiée, un seuil sur la mesure ρ, plus adapté à la configuration

des données disponibles est à définir [Tax and Duin, 2004, Tax and Duin, 2000]. Il est à noter que pour rendre la méthode plus robuste notamment aux bruits dans la base de données, davantage de voisins peuvent être utilisés. Dans ce cas, on peut considérer la valeur moyenne des distances pour induire la mesure ρ ou encore, dans le cas du kNN, considérer seulement le k-ième voisin.

Nous pouvons également citer parmi les méthodes par estimation de distance les méthodes k-means [MacQueen et al., 1967] et k-centres qui sont initialement des approches de clustering mais qui peuvent être aisément adaptées au one-class en étiquetant par exemple les données présentes comme targets. Dans ces deux approches, les données targets sont décrites par k partitions ou clusters. L'objectif est d'optimiser la position des centres de ces k clusters.
Pour k-means, il s'agit de minimiser la moyenne des distances à ces centres. Pour k-centres, il s'agit de minimiser la distance maximale au centre des clusters (cf. [Bishop, 1995] pour des algorithmes d'optimisation pour ces méthodes). Dans les deux cas, un nouvel exemple est étiqueté outlier si la distance au plus proche cluster est plus grande qu'un certain seuil fixé à l'avance, autrement il est étiqueté target. La formulation de la fonction de décision pour les deux méthodes est identique et on a :

$$f(\mathbf{x}) = \begin{cases} target \text{ si } \min_{i=1..k} (\mathbf{x} - \mathbf{c}_i)^2 \leqslant \theta \\ outlier \text{ sinon} \end{cases}$$

où les $\mathbf{c}_i$ sont les k vecteurs moyens des clusters, θ étant la valeur du paramètre de seuil.
Les difficultés principales de ces deux approches résident dans le choix d'un seuil adapté et la gestion des problèmes liés à la dimension notamment pour les phénomènes évoqués précédemment.

3.2.1.3 Approches par reconstruction

L'objectif des approches par reconstruction est d'encoder ou "compresser" les données afin de les reproduire le plus fidèlement possible en phase de test. Ainsi, le classifieur est optimisé pour reproduire le plus fidèlement possible les données d'apprentissage fournies en entrée du système. Lors de la phase de test, les exemples ayant une erreur de reconstruction élevée pourront être classés comme outlier. La principale difficulté de ces approches est de définir la procédure d'optimisation pour l'encodage ou la compression des données et le choix d'un seuil approprié sur l'erreur de reconstruction.
Les réseaux de neurones sont les plus cités dans les approches par reconstruction. L'approche est basée sur l'erreur de reconstruction du réseau sur les données targets pour construire le modèle de la classe target [Nairac et al., 1999, Bishop, 1994, Hodge and Austin, 2004]. Moya et al. [Moya and Hush, 1996] proposent un réseau de neurones à plusieurs fonctions objectives (dites fonctions d'activation) formant une surface de décision ellipsoïdale, compacte et fermée, modélisant la taille, la forme et l'orientation des données targets. [Desforges et al., 1988, Bishop, 1994] proposent un perceptron multi-couches en tant qu'auto-encodeur combiné avec un estimateur de Parzen de la distribution des targets. Dans cette approche, les données ayant une faible valeur de probabilité sont vraisemblablement des exemples outliers ou des exemples d'une nature nouvelle, différente de celle des données d'apprentissage. Ces exemples sont ainsi filtrés et le réseau se chargera alors des exemples plus complexes. Le système est ainsi plus robuste [Bishop, 1994]. Le réseau diabolo est également un exemple de classifieur auto-encodeur [Hinton, 1989] étudié par Japkowicz et al. [Japkowicz et al., 1995, Japkowicz, 1999]. La particularité de ce réseau est de posséder une couche cachée composée d'un très faible nombre de nœuds, créant ainsi un goulot d'étranglement (ou zone de compression avec perte), encodant ainsi les données par projection dans un sous-ensemble plus petit entraînant nécessairement une perte informative. Cette approche permet de rendre plus robuste le classifieur, le rendant ainsi moins sensible au sur-apprentissage. Les données d'entrée ayant une projection faible sur cette couche cachée auront ainsi vraisemblablement une erreur de reconstruction élevée.

3.2.1.4 Le Support Vector Data Description (SVDD)

Dans une approche de type description des données (”data description“), Tax et Duin [Tax and Duin, 2004, Tax and Duin, 1999] proposent le Support Vector Data Description (SVDD). Il s'agit d'englober les clusters targets au sein d'une boule de centre a et de rayon R rendue flexible par l'utilisation de fonctions noyaux, de façon similaire au SVM standard [Cortes and Vapnik, 1995, Vapnik, 1998].

Il s'agit alors de résoudre le problème d'optimisation suivant :

$$\min_{R\in\Re,\xi\in\Re^l} R^2 + C\cdot\sum_i \xi_i$$

sous la contrainte :

$$||\phi(\mathbf{x_i}) - a||^2 \leq R^2 + \xi_i,\ \xi_i \geq 0,\ i = 1..l$$

les ξ_i étant des variables de relaxation, C un paramètre de coût représentant un compromis entre le volume de la boule et l'erreur commise en acceptant des données outliers dans la description des données targets. La résolution de ce problème d'optimisation donne la fonction de décision suivante :

$$f(\mathbf{x}) = signe(R^2 - \sum_{i,j}\alpha_i\cdot\alpha_j\cdot k(\mathbf{x_i},\mathbf{x_j}) + 2\sum_i \alpha_i\cdot k(\mathbf{x_i},\mathbf{x}) - k(\mathbf{x},\mathbf{x}))$$

les α_i étant les coefficients issus du Lagrangien [Bishop, 1995] et associés aux vecteurs de support de la distribution des données, $k(\mathbf{x_i},\mathbf{x_j})$ étant une fonction noyau, généralement gaussien.

L'approche SVDD présente la flexibilité d'utiliser des fonctions noyaux exactement comme dans le cas du SVM standard. Mais comme le one-class-SVM [Scholkopf et al., 2001] que nous verrons dans les approches discriminantes en simulation des données outliers, la méthode est sensible à la dimension [Evangelista et al., 2006].

3.2.2 Méthodes générant des outliers

Dans cette partie nous parlons des approches nécessitant la synthèse des données outliers absentes lors de la phase d'apprentissage. C'est le cas généralement des méthodes discriminantes nécessitant la présence de données de toutes les classes en présence afin de tracer la frontière de décision entre ces classes. Nous présentons les difficultés de synthétiser la classe des données outliers en l'absence d'a priori sur leur distribution. En effet, il est généralement considéré dans la littérature que les données outliers sont uniformément réparties dans l'espace de description. Cette hypothèse forte pose alors le problème de la mise en pratique d'une méthode efficace de génération de ces données en raison notamment de la malédiction de la dimension. Nous analysons d'autres hypothèses de distribution des données outliers permettant de résoudre ce problème avec notamment l'approche originale proposée par [Hempstalk et al., 2008] proposant de calquer la distribution des outliers sur celle des données targets.

Ces approches consistent en la génération des données outliers muni d'hypothèses souvent fortes concernant la distribution de ces données, leur quantité, leur localisation par rapport aux données targets [Hempstalk et al., 2008, Fan et al., 2004, Banhalmi, 2008]. La génération des outliers est une approche intuitive pour pallier le manque de données outlier. En effet, elle permet de créer, artificiellement, le jeu de données absent et ainsi transformer l'approche one-class en une approche à deux classes pouvant être traitée par tout classifieur binaire. Les problèmes majeurs soulevés par la génération de données outliers réside dans le choix de la distribution et la gestion de la dimension. Le choix notamment d'une distribution uniforme des outliers se confronte au problème de la malédiction de la dimension [Hinneburg and Keim, 1999, Bellman, 1961]. En effet, le nombre de données nécessaires est exponentiel en la dimension si on veut couvrir l'espace ou le domaine de description, i.e. un nombre de données suffisamment important pour avoir une densité suffisante des données outliers. Ce nombre grimpe facilement à plusieurs milliards dès que la dimension dépasse 10. Il devient alors nécessaire de réduire la dimension du problème ou bien adopter des hypothèses

plus adaptées pour pallier ces effets inhérents aux problèmes de grande dimension.
Le domaine de définition a également son importance. Si on prend l'exemple d'une hyperbox, une grande partie des données générées uniformément et indépendamment les unes des autres vont se retrouver avec une forte probabilité dans les coins de cette hyperbox plutôt que d'être équi-répartis dans tout le volume. Dans le cas d'une hypersphère, les données outliers générées sont beaucoup plus compactes [Tax and Duin, 2002, Luban and Staunton, 1988].

L'approche qu'on retrouve alors le plus souvent dans la littérature est une approche basée sur la distribution originale des targets. Ces approches sont dites "distribution-based". On effectue une transformation sur les données targets afin de générer les données outliers. Par exemple les outliers peuvent être générés en modifiant une unique variable des données targets, laissant les autres variables inchangées [Fan et al., 2004, Wang et al., 2009]. Ces méthodes ont l'avantage de générer des outliers en grand nombre, proches des données et suffisamment diversifiés. D'autres auteurs préconisent l'identification de régions sparses faiblement peuplées en données targets pour y générer, seulement à ces endroits, beaucoup plus de données outliers [Fan et al., 2004]. Ces mêmes régions sparses peuvent constituer des patterns caractérisant une présence possible des données outliers.
Dans [Hempstalk et al., 2008], les auteurs utilisent une approche originale combinant à la fois une génération de données outliers et une estimation de probabilité de classes calculée avec un arbre de décision standard conduisant à une estimation de la densité des données targets. Les auteurs utilisent la règle de Bayes pour montrer que la densité des données outliers et celle des données targets sont simplement équivalentes à un facteur près. Ce facteur peut être obtenu avec un algorithme discriminant standard capable d'estimer la probabilité des classes (comme l'arbre de décision). Dans leur implémentation, les auteurs ont ainsi choisi une distribution normale, dont les paramètres ont été estimés sur les données targets, pour caractériser les données outliers et ont ensuite utilisé les sorties d'un arbre de décision pour établir les probabilités des classes (afin de constituer le facteur mentionné plus haut). L'expression de la probabilité des données targets est alors donnée par :

$$P(X|T) = \frac{(1-P(T)) \cdot P(T|X)}{P(T) \cdot (1-P(T|X))} \cdot P(X|A)$$

où T est la classe target, A la classe outlier, X une donnée d'entrée, $P(T|X)$ la probabilité estimée par le biais d'un classifieur que la donnée X soit de classe target.

3.2.3 Méthodes simulant des outliers

Dans cette section, nous retrouvons des approches qui contournent les problèmes mentionnés précédemment concernant la mise en pratique d'approches de génération des outliers, notamment ceux liés à la gestion de la dimension. En effet, dans ces méthodes simulant des outliers, l'absence des données outliers est compensée par la simulation de leurs distributions dans l'espace de description. Ces méthodes, par exemple, font l'hypothèse d'une distribution uniforme des données outliers selon une grille régulière de l'espace de description et tiennent compte de cette localisation régulière des données outliers dans l'élaboration de la fonction de décision. L'apprentissage est alors possible sans la nécessité de stocker physiquement les données simulées. C'est le cas notamment de l'approche proposée par [Scholkopf et al., 2001] basée sur le SVM et dans laquelle la classe des outliers est représentée par un unique point placé à l'origine de l'espace de description. C'est également le cas de l'approche de [Liu et al., 2000] dans laquelle les données outliers nécessaires à l'établissement du critère de partitionnement choisi au sein d'un arbre de décision sont supposées uniformément réparties sur une grille de l'espace de description. En effet, le calcul de ce critère ne nécessite pas la présence en mémoire des données de chacune des classes mais seulement de leur fréquence respective au sein du nœud considéré. Nous décrivons dans un premier temps l'approche originale proposée par [Aggarwal and Yu, 2001] permettant de caractériser, à l'aide d'une mesure de sparsité, les régions susceptibles de contenir des données outliers.

3.2.3.1 Mesure de sparsité

Une hypothèse communément admise est de considérer que si une région de l'espace de description est vide en données targets, alors cette région est susceptible de contenir des données de l'out-of-class [Aggarwal and Yu, 2001, Bishop, 1994]. Dans [Aggarwal and Yu, 2001], les auteurs proposent une mesure de sparsité [2] capable d'évaluer le degré de présence des données targets dans un sous-espace et un sous-domaine sélectionné dans cet espace. Ainsi, les faibles valeurs de cette mesure traduisent un sous-espace faiblement peuplé en données targets et donc favorable à la présence de données outliers. La principale difficulté de cette méthode consiste à trouver ces sous-espaces sparses tout en évitant la recherche exhaustive dans un grand nombre de combinaisons possibles (i.e ; la catastrophe combinatoire). Les auteurs proposent alors un algorithme évolutionniste capable de trouver rapidement des combinaisons de caractéristiques conduisant à des sous-espaces sparses. Les auteurs n'ont pas appliqué d'algorithmes de classification en particulier, leur approche restant qualitative dans le choix de ces sous-espaces en ayant toutefois obtenu un partitionnement de l'espace dans lequel certaines régions peuvent être étiquetées "outlier". Ainsi, de nouvelles données ayant des projections non négligeables dans les sous-espaces sélectionnés pourront être évaluées comme atypiques ou anormales.
Les auteurs estiment le nombre de targets présents dans un sous-espace de type hypercube en dimension k en partant de l'hypothèse d'une distribution uniforme de ces données targets dans l'espace d'origine. Tout d'abord ils définissent un pas de discrétisation régulier pour chacune des dimensions. Φ intervalles de valeurs d'attributs sont ainsi définis pour chacune des dimensions du problème, chaque intervalle contenant ainsi une fraction $f = \frac{1}{\Phi}$ des données. En formant un hypercube de dimension k à partir de k intervalles pris aléatoirement dans l'espace d'origine, la fraction de données targets attendues dans cet espace est de f^k. La présence ou l'absence d'une donnée target suit une loi de Bernouilli de paramètre f^k et si on suppose l'échantillon target i.i.d., le nombre de données targets attendues dans ce k-hypercube est de $N \cdot f^k$, avec une variance de $\sqrt{N \cdot f^k \cdot (1 - f^k}$. Le nombre de données targets peut alors être approximé par une loi normale. Les auteurs définissent alors la mesure de sparsité suivante :

$$S(D) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}} \tag{3.1}$$

où D est l'hypercube de dimension k, N le nombre initial de données targets dans l'espace d'origine, f la fraction des données présentes dans un des intervalles de discrétisation pour chacune des dimensions. Ainsi les valeurs de sparsité négatives sont indicatrices de régions sparses du domaine initial. La forme de la mesure de sparsité, si on suppose une distribution normale du nombre réel d'outliers dans l'hypercube, permet d'évaluer statistiquement le caractère sparse du sous-espace étudié, avec la tabulation de la loi normale, i.e. la probabilité que le nombre de données réellement présentes dans l'hypercube soit significativement différente de la valeur attendue. Le critère i.i.d. n'est cependant généralement pas respecté, mais cette mesure de sparsité fournit néanmoins une approche statistique du caractère sparse des données targets. L'une des restrictions de la méthode est l'"hypothèse de la distribution uniforme des données targets. Cette approche peut être vue comme une méthode d'extraction d'informations a priori (sur la localisation des données outliers) à partir de la base d'apprentissage et servir pour l'étape de construction d'un classifieur.

3.2.3.2 Le CLustering Tree

Liu et al. 2000 [Liu et al., 2000] introduisent l'algorithme discriminant CLTree (pour CLustering based decision Trees) au sein duquel ils proposent de simuler (sans génération) les données outliers, selon une distribution uniforme, en chacun des nœuds d'un arbre de décision (appelé cluster tree), permettant ainsi, dynamiquement, de conduire l'induction de ce dernier. L'approche proposée

2. La sparsité est le degré de vide d'un espace donné ; on pourra aussi parler de la densité d'une population donnée dans une région de cet espace (ainsi un espace est "sparse" s'il est très peu dense)

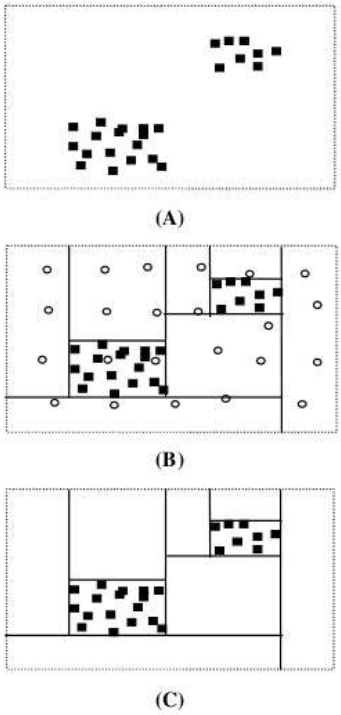


FIGURE 3.3 – Dans l'approche du Clustering Tree [Liu et al., 2000], les outliers sont générés en chacun des nœuds de l'arbre ; le partitionnement récursif peut alors se dérouler de façon classique par recherche du meilleur point de coupure ; (A) population initiale de données targets, (B) génération des données outliers et partitionnement classique en présence de données de cette seconde classe et (C) identification des régions targets et outliers.

est traitée initialement sous l'angle du clustering : le groupe de données des clusters initiaux est étiqueté target et les données simulées autour de ces targets peuvent alors être considérées comme outliers (cf. figure 3.3). L'objectif de cette approche est d'identifier à l'aide du partitionnement naturel de l'arbre de décision et sans a priori sur les données targets, les régions de l'espace identifiées comme contenant des clusters target et les régions vides contenant potentiellement des donnés outliers. Ces outliers sont simulés plutôt que générés car le critère de partitionnement utilisé au sein d'un nœud ne nécessite pas la présence physique de ces outliers. Cette approche de simulation possède toutefois l'inconvénient de continuellement remplir l'espace autour des données targets ; il devient alors nécessaire de déterminer un critère d'arrêt. L'auteur effectue pour cela un pré-élagage de l'arbre en spécifiant un nombre minimal de données target présent en chaque nœud pour continuer le partitionnement (ce nombre est ici spécifié en termes de pourcentage du set initial de données).

3.2.3.3 Le One-class SVM (OCSVM)

Le one-class SVM (OCSVM) ou $\upsilon - SVM$ proposé par Scholkopf et al. [Scholkopf et al., 2001], tout comme le SVDD (Support Vector Data Description) proposé par Tax et al. [Tax and Duin, 2004] et présenté précédemment, sont des exemples d'algorithmes s'inspirant de classifieurs multi-classes standards, particulièrement le SVM ici, pour s'adapter à la tâche one-class. OCSVM est souvent mentionné comme un classifieur de référence dans la classification one-class comme alternative aux approches par estimation de densité comme le Parzen ou la mixture de gaussiennes. Il dérive, tout comme le SVDD, de l'algorithme discriminant standard SVM [Vapnik, 1998] et a donc la propriété de pouvoir apprendre en l'absence d'une seconde classe. Son principe de base consiste à séparer à l'aide d'un hyperplan avec une marge maximale les données targets du point origine, unique représentant des données de l'out-of-class. Cette approche est intéressante et originale car

elle permet de simuler un approche de classification binaire standard et donc de résoudre la tâche de classification one-class à l'aide d'une formulation identique à celle du SVM standard. On montre par ailleurs que la fonction de décision de OCSVM a la même formulation que celle du SVDD pour le noyau gaussien et des données normalisées.

Les auteurs proposent le OCSVM comme une extension naturelle des SVM aux données non étiquetées. Comme pour le SVM, il est possible d'utiliser l'astuce du noyau pour projeter les données targets dans un espace adapté à l'aide d'une fonction adaptée ϕ. La fonction de décision devra renvoyer "+1" dans les régions contenant les targets et "-1" partout ailleurs. La séparation des données de l'origine passe par la résolution du problème de programmation quadratique suivant :

$$\min_{\omega \in F, \xi \in \Re^d, \rho \in \Re} \frac{1}{2} \cdot ||\omega||^2 + \frac{1}{\upsilon d} \cdot \sum_i \xi_i - \rho$$

sous la contrainte :

$$\omega \cdot \phi(\mathbf{x_i}) \geq \rho - \xi_i, \, \xi_i \geq 0, \, i = 1..l$$

où d est la dimension de l'espace, ρ un paramètre d'offset de l'hyperplan, w le vecteur de poids de l'hyperplan de telle sorte que $\rho/||\omega||$ représente la distance entre l'hyperplan et le point origine ; $\upsilon \in [0:1]$ est un paramètre contrôlant l'erreur du modèle, les ξ_i étant des variables de relaxation. Les auteurs montrent que le paramètre υ représente une borne supérieure sur la fraction de données outliers présentes dans le set d'apprentissage et une borne inférieure sur le nombre de vecteurs de support. La fonction de décision f s'écrit :

$$f(\mathbf{x}) = signe(\sum_i \omega \cdot \phi(\mathbf{x}) - \rho)$$

positive sur les données targets et négatives sur les données outliers. En utilisant les multiplicateurs α_i et γ_i, on introduit le Lagrangien :

$$L(\omega, \xi, \rho, \alpha, \gamma) = \frac{1}{2} \cdot ||\omega||^2 + \frac{1}{\upsilon l} \cdot \sum_i \xi_i - \rho) - \sum_i \alpha_i \cdot (\omega \cdot \phi(\mathbf{x_i}) - \rho + \xi_i - \sum_i \gamma_i \cdot \xi_i$$

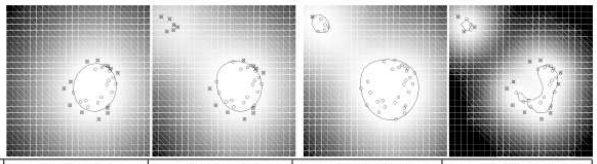
dont la résolution fournit, comme pour le SVM l'expression :

$$f(\mathbf{x}) = signe(\sum_i \alpha_i \cdot \mathbf{k}(\mathbf{x_i}, \mathbf{x}) - \rho)$$

où $\mathbf{k}(\mathbf{x}, \mathbf{x_i}) = (\phi(x)^T \cdot \phi(x_i))$ est le noyau du modèle, s'apparentant à un produit scalaire. Ce noyau est généralement choisi de type RBF ("Radial basis Function" pour fonction à base radiale) comme le noyau gaussien donné par $\mathbf{k}(\mathbf{x}, \mathbf{y}) = exp(-\gamma \cdot ||\mathbf{x} - \mathbf{y}||^2)$, où γ est un paramètre lié à la largeur de bande du noyau.
Un exemple de frontière de décision sur deux problèmes artificiels est donné sur la Figure 3.4. Cette fonction est construite avec un noyau gaussien d'écart-type c. Nous pouvons observer l'influence du paramètre υ : une faible valeur de υ permet de tenir compte du cluster des points targets situés en haut à gauche des configurations 2 et 3 ; en réduisant le paramètre de la gaussienne, l'algorithme peut être contraint de tenir compte des points outliers situés également en haut à gauche des configurations 2, 3 et 4.

3.2.4 Méthodes d'ensembles

Les méthodes d'ensembles ont été très peu mentionnées dans la littérature pour traiter l'approche one-class [Khan and Madden, 2010, Tax and Duin, 2001, Shieh and Kamm, 2009, Evangelista et al., 2006] malgré les avancées tant théoriques qu'expérimentales dont elles ont bénéficié depuis la fin des années 1990 [Dietterich, 2000a] et leur adoption rapide dans la classification standard mentionnée dans le premier chapitre.



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|w\|$	0.84	0.70	0.62	0.48

FIGURE 3.4 – Exemples de fonctions de décision pour le one-class SVM avec un noyau gaussien de largeur c sur deux problèmes artificiels différents en 2D (figure 1 et figures suivantes) obtenues en faisant varier c et υ ; dans la seconde ligne du tableau sont indiqués la fraction de vecteurs de support (SV), et celle des données outliers (OL) ; cette expérimentation rend compte de l'influence du paramètre υ sur le nombre de vecteurs de support et le nombre de données considérées comme outliers [Scholkopf et al., 2001].

Face aux difficultés des approches one-class standards génératives et discriminantes où notamment la malédiction de la dimension ou les phénomènes d'espaces vides et de concentration de la mesure doivent être pris en compte [Verleysen et al., 2003, Donoho, 2000], il est surprenant de voir la faible adoption des méthodes d'ensemble pour le traitement de l'approche one-class. En effet, les ensembles offrent davantage de flexibilité, par rapport aux classifieurs standards, pour apprendre des frontières de décision complexes au sein des données, en raison de l'agrandissement de l'espace de solutions rendu possible par la combinaison de plusieurs solutions existantes. De plus, il a été démontré que les méthodes d'ensembles sont généralement plus performantes que les approches standards. Dans [Khan and Madden, 2010], les auteurs observent ainsi que l'exploitation des principes de combinaison de classifieurs comme le bagging et le boosting pourraient apporter un réel avantage aux méthodes one-class existantes.

Un exemple de combinaison de classifieurs one-class est donné dans Tax et Duin [Tax and Duin, 2001]. Les auteurs proposent plusieurs approches pour combiner des classifieurs one-class de natures différentes et munis de sets d'apprentissage distincts et des espaces de description distincts. Les classifieurs de la combinaison proposée sont des approches par estimation de densité (estimateur gaussien, fenêtre de Parzen et mixture de gaussienne), des approches basées sur des distances (K-means, K-centres [Ypma and Duin, 1998]), une approche par reconstruction (réseau de neurones auto-encodeur [Japkowicz, 1999]) et une approche frontière (SVDD [Tax and Duin, 1999]). La combinaison de Parzen identiques dans des espaces distincts permet d'obtenir les meilleurs résultats. Les auteurs font alors l'hypothèse que les bonnes performances obtenues en utilisant différents espaces de description pour chacun des classifieurs viendraient d'une relative indépendance des espaces de description. Nous avons vu en effet au chapitre 1 que la diversité au sein du pool de classifieurs de l'ensemble était favorable pour l'obtention de bonnes performances.
Les auteurs avaient précédemment proposé une méthode d'ensemble basée sur le degré de désaccord existant au sein des classifieurs de la combinaison [Tax and Duin, 1998]. Lorsque les classifieurs ne sont pas d'accord sur la classe d'un exemple en test, ce dernier est vraisemblablement un outlier. [Shieh and Kamm, 2009] utilisent une combinaison de classifieurs OCSVM en se servant du bagging [Breiman, 1996] comme mécanisme de randomisation de l'ensemble. Ainsi, chaque classifieur OCSVM apprend individuellement sur un jeu bootstrap des données targets disponibles. Les auteurs ont montré l'apport non-négligeable de l'approche par méthode de combinaison, notamment en terme de stabilité comparé à l'approche par OCSVM seule. En effet, OCSVM construit une surface de décision qui est sensible à la configuration des données et qui provoque une sur-estimation de la frontière des données targets [Hoffmann, 2007, Tax and Juszczak, 2002]. On observe en effet que pour des valeurs de υ faible, OCSVM est instable et le gain apporté par l'approche par méthode d'ensemble est significatif pour différents jeux de données réelles, montrant ainsi une des propriétés de ces méthodes vis-à-vis de l'amélioration des performances des classifieurs instables. Toutefois, l'approche proposée par les auteurs se révèle très coûteuse et avec une paramétrisation standard

(typiquement $\upsilon = 0.1$) la différence entre l'approche par OCSVM seule et l'ensemble de OCSVM n'excède pas 3%, ayant même une baisse de performance pour une des bases testées.

Nous observons que les approches one-class existantes d'ensembles de classifieurs utilisent des combinaisons de méthodes déjà one-class sans exploiter pleinement les mécanismes proposés par la théorie des ensembles de classifieurs dans le but d'engendrer de meilleures approches one-class. Les principes de randomisation comme le bagging, le Random Subspace Method (RSM) [Ho, 1998] ou le Random Feature Selection (RFS) [Breiman, 2001, Dietterich, 2000b] peuvent être utilisés pour résoudre des problèmes particuliers du one-class comme la gestion de la grande dimension. Ainsi, la difficulté de générer convenablement des données outliers peut être drastiquement réduite en considérant à la fois le bagging et le RSM pour injecter de la diversité dans le mécanisme de génération de ces outliers et réduire la dimension de l'espace de génération. Deux difficultés sont donc attaquées : l'apport de diversité dans le pool de classifieurs et la gestion de la grande dimension. Ainsi l'exploitation de ces mécanismes de randomisation offre de nouvelles perspectives pour traiter le problème one-class avec des approches nouvelles et prometteuses.

3.2.5 Conclusion

Dans cette section, nous avons présenté les algorithmes one-class de la littérature selon le positionnement adopté dans leur mise en œuvre pour le traitement des données de l'out-of-class. Nous avons vu la difficulté pour les approches génératives de déterminer un seuil adapté à l'identification des deux classes targets et outliers et pour traiter les problèmes de grande dimension en raison des coûts importants en grande dimension liés notamment à l'estimation des paramètres ou au parcours exhaustif des données disponibles. En outre, dans la quasi totalité des problèmes réels à traiter, les données requises sont en nombre insuffisant pour l'obtention d'une estimation fiable de la distribution des classes en présence et ce, même pour des dimensions aussi petites que 15. D'autres approches one-class comme celles basées sur l'estimation de distance se heurtent également aux problèmes de la dimension en raison de phénomènes connus sous le nom de "phénomène d'espace vide" créant des espaces de plus en plus sparses difficiles à traiter par des méthodes travaillant dans la totalité de l'espace de description ou encore de "concentration de la mesure" où par exemple le plus proche voisin n'est pas significativement différent du voisin le plus lointain avec la dimension. Nous avons vu que les approches discriminantes sont confrontées quant à elles à la nécessité de synthétiser les données de l'out-of-class pour établir la frontière de décision. Deux approches ont été proposées : la génération physique de ces données outliers ou leur simulation. Dans les deux cas, plusieurs problèmes apparaissent quant à la définition de la nature et des propriétés de ces données à synthétiser. L'un des problèmes majeurs rencontrés est de nouveau la gestion de la dimension et la représentativité des données synthétisées vis-à-vis de l'out-of-class. Les mécanismes et principes utilisés dans les méthodes d'ensembles n'ont été que très peu mentionnés dans la littérature one-class. Leur exploitation, pourtant conseillée dans la littérature, offre par leurs bonnes propriétés et leurs performances reconnues dans le cas de la classification multi-classe standard, des perspectives intéressantes pour une mise en œuvre dans l'approche one-class.
Nous proposons alors d'aborder la classification one-class sous l'angle des méthodes d'ensemble en nous intéressant particulièrement à la famille de méthodes d'ensemble d'arbres de décision, les forêts aléatoires (RF). Ces méthodes bénéficient en effet pleinement des mécanismes évoqués précédemment et sont bien connues pour être parmi les plus performantes de la littérature comme nous l'avons mentionné dans les chapitres 1 et 2.

3.3 Une approche par Forêts Aléatoires pour la classification one-class

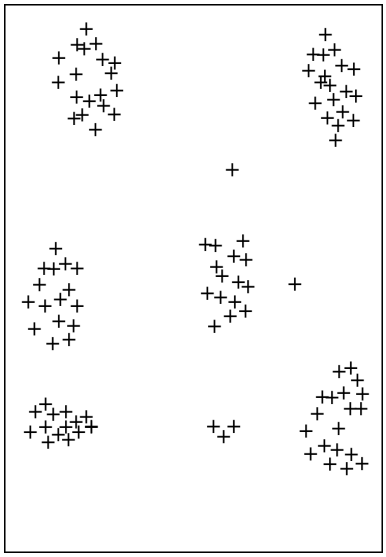
Parmi les méthodes d'ensembles, on distingue les forêts aléatoires [Breiman, 2001] qui sont une famille d'ensembles d'arbres de décision, considérées comme l'une des meilleures approches

proposées dans le cadre de la classification multi-classe standard comme nous l'avons décrit dans les deux premiers chapitres de la thèse. Ces méthodes se sont révélées performantes et robustes, capables particulièrement de gérer les problèmes de grande dimension.
Les forêts aléatoires sont une approche discriminante pour laquelle les données outliers doivent être synthétisées. Les trois mécanismes de randomisation bagging, RSM et RFS peuvent être utilisés conjointement dans les forêts pour rendre le procédé de génération des outliers plus efficace, en permettant notamment la gestion de données de grande dimension. En effet, le bagging, en sous-échantillonnant le set des données d'apprentissage, permet de réduire le nombre d'outliers à générer lorsque la distribution de ces outliers dépend du nombre d'éléments en apprentissage. Puis le RSM ou le RFS, en sous-échantillonnant le set des caractéristiques des données d'apprentissage permet la réduction drastique de la dimension (e.g. constante ou d/2 ou $\sqrt{d}$ où d est la dimension) et donc la forte diminution des coûts de génération liés à la grande dimension. De plus, les forêts sont suffisamment flexibles en permettant de générer les outliers à différents moments de leur induction : dans le set boostrap, dans le sous-espace RSM ou en chaque nœud de l'arbre (dans le sous-espace RFS déterminé en chaque nœud).
Outre l'utilisation des mécanismes de randomisation des forêts aléatoires, nous tenons compte du fait que les données outliers ne doivent pas systématiquement suivre une distribution uniforme (choix "naturel"). En effet, nous avons mentionné à la section 3.2.3 l'existence d'approches de génération d'outliers basées sur l'identification de régions ou sous-espaces sparses de l'espace de description et donc susceptibles d'héberger des outliers. Avec de telles approches, complémentaires aux mécanismes de randomisation, il est possible de réduire davantage le nombre d'outliers à générer dans les sous-espaces RSM ou RFS. Nous illustrons sur la figure 3.5 une configuration en 2 dimensions d'un problème one-class dans laquelle les outliers sont générés dans les zones sparses de l'espace de description. Initialement, le set est composé d'objets de la classe target ("+"). Les zones sparses sont identifiées au moyen des distributions des données targets (régions vides en données targets), puis les outliers y sont générés. Nous montrons ensuite un partitionnement effectué par un arbre de décision sur ces mêmes données. Nous voyons qu'avec un arbre de décision, il n'est pas nécessaire de remplir tout l'espace de description en données outliers car des instances outliers aux frontières des données targets sont suffisantes.

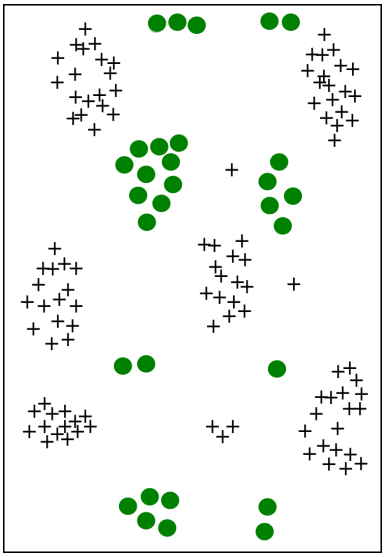
En nous basant sur les remarques et la discussion précédentes, nous proposons une approche one-class discriminante par méthodes d'ensemble d'arbres de décision, appelée one-class random forest (OCRF), tirant partie des mécanismes de randomisation des méthodes d'ensemble pour le traitement de la classification one-class. L'arbre étant par essence discriminant, les données outliers doivent être synthétisées. Nous avons décidé de synthétiser ces données outliers en les générant au sein du set d'apprentissage de chacun des arbres afin de disposer de données physiquement présentes dans le set d'apprentissage lors de l'induction de chaque arbre de la forêt.

La génération d'outliers pose des problèmes majeurs en grande dimension en raison principalement de la malédiction de la dimension. Les mécanismes de randomisation comme le RSM et le RFS constituent des réponses favorables à ces difficultés car ces méthodes permettent de travailler dans des sous-espaces plus petits que l'espace initial. Notamment, le RSM permet d'obtenir un espace plus petit pour générer les données outliers. Le RFS est un mécanisme largement utilisé au sein des arbres des forêts aléatoires permettant notamment de créer de la diversité favorable aux forêts et que nous avons choisi de conserver. Le bagging est un autre mécanisme d'ensemble important car il permet de conserver la distribution initiale des données targets tout en créant également de la diversité. Nous faisons le choix de combiner les deux mécanismes bagging et RSM pour la génération des données outliers afin de bénéficier à la fois des propriétés du bagging et du RSM dans le cadre des forêts aléatoires et la possibilité d'attaquer avec le RSM les difficultés de génération des outliers en grande dimension.

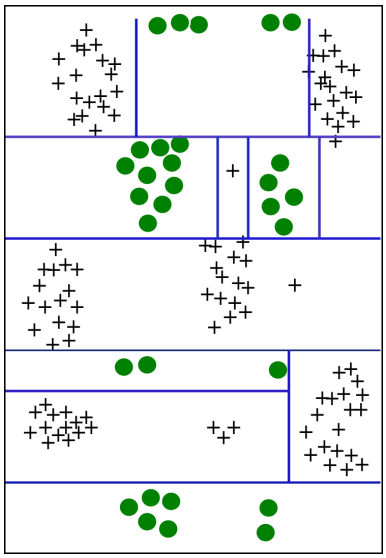
Il se pose naturellement la question de la distribution des outliers. Plusieurs approches sont possibles : une génération uniforme ; une génération selon un modèle gaussien ; une génération distribution-based, i.e. basée sur la distribution des targets. Nous avons fait le choix de la dernière



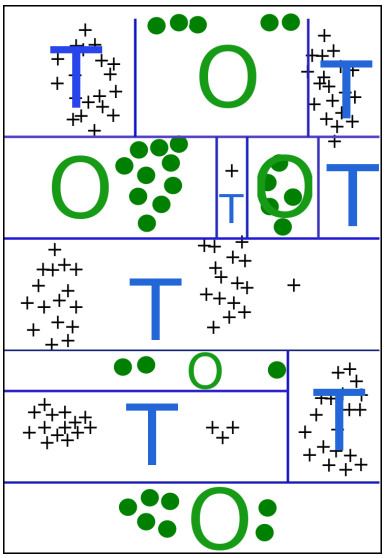
(a) Données Targets



(b) Données Targets + Outliers générés



(c) Frontière de décision d'un arbre



(d) Étiquettes des feuilles de l'arbre

FIGURE 3.5 – Illustration de la problématique de la génération des outliers dans des zones sparses de l'espace des targets. Les symboles "+" (noir) représentent des instances targets (a). La distribution en (a) des données targets permet d'identifier des régions sparses pour y générer des outliers (figure b) ; les objets "ronds" (vert) représentent des instances de l'out-of-class générées alors dans ces zones sparses ; l'arbre de décision est classiquement induit à l'aide de ces nouvelles données (c) ; les étiquettes des feuilles de l'arbre sont indiquées pour chacune des partitions : "*T*" pour une feuille target et "*O*" pour une feuille étiquetée outlier.

approche afin de tenir compte au mieux de la spécificité des données en présence, ce qui permet aussi de rendre notre approche flexible. Ceci nécessite cependant de caractériser les données en présence et donc d'extraire des informations a priori du set d'apprentissage.

Notre approche est ainsi basée sur les trois principes de randomisation bagging, RSM et RFS et comprend les étapes essentielles suivantes :

(1) Extraction d'informations a priori du set d'apprentissage afin de guider le processus de génération au point suivant (2) ; ces informations concernent notamment la localisation possible des outliers sous formes de contraintes imposées à la distribution finale des outliers ;
(2) Génération des outliers lors des mécanismes de randomisation bagging et RSM suivant les contraintes imposées en (1) ;
(3) Induction classique de chaque arbre individuel sur le set bootstrap + RSM généré en (2) ;
(4) Règle de combinaison classique par vote à la majorité ;

Nous illustrons ces différentes étapes à travers le schéma de la figure 3.6. Á partir de ce cadre de travail, nous proposons dans le chapitre suivant notre approche OCRF, tirant partie des mécanismes de combinaison et des principes de randomisation évoqués dans cette section.

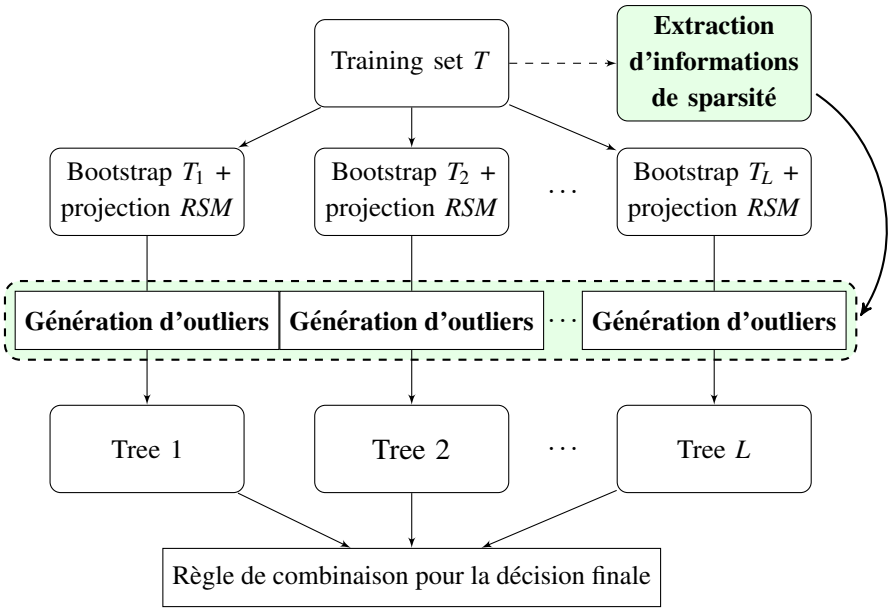


FIGURE 3.6 – Schéma des différentes étapes de l'approche one-class que nous proposons basé sur les forêts aléatoires.

3.4 Conclusion

Dans ce chapitre, nous avons étudié les différentes approches one-class couramment mentionnées dans la littérature. Nous avons proposé une nouvelle catégorisation des méthodes one-class mettant en avant la façon dont ces méthodes prennent compte des données out-of-class.Nous avons vu les difficultés auxquelles sont confrontées notamment les approches par estimation de densité ou de distance et les approches discriminantes avec particulièrement la malédiction de la dimension et leurs conséquences comme le phénomène d'espace vide et la concentration de mesure. Face à ces difficultés, nous avons observé que les méthodes d'ensembles, et particulièrement les forêts aléatoires, rarement mentionnées dans la littérature one-class, possèdent des mécanismes permettant de répondre favorablement à ces problèmes posés.

Les forêts aléatoires sont des approches discriminantes qui par conséquent nécessitent la synthèse des données de l'out-of-class. Les méthodes de synthèse de ces données sont actuellement difficiles à mettre en oeuvre en raison principalement de la grande dimension. Les mécanismes de

randomisation et de combinaison mis en oeuvre dans les forêts aléatoires permettent d'apporter des solutions aux problèmes rencontrés par les approches one-class standards. Nous proposons une approche one-class par méthodes d'ensemble reprenant ces mécanismes et détaillons dans le chapitre suivant la mise en oeuvre de cette nouvelle approche.

Chapitre 4

Les forêts aléatoires one-class

Sommaire

4.1 Introduction

Nous avons présenté dans le chapitre précédent un panorama des approches one-class couramment citées dans la littérature. L'approche one-class apparaît non seulement comme une solution aux problèmes de données déséquilibrées, aux problèmes où une classe critique est mal définie mais aussi comme une alternative aux approches multi-classe standards. Nous avons étudié les différentes heuristiques proposées pour catégoriser les méthodes one-class et avons évalué les difficultés qui sont associées à chacune des méthodes citées. Nous avons insisté sur le fait que les difficultés des méthodes génératives résidaient notamment dans la nécessité du choix d'un modèle a priori de la distribution des targets qui restait à valider, la difficulté de l'adaptation au facteur d'échelle liée à la malédiction de la dimension en raison d'une part des coûts calculatoires souvent importants (liés à l'estimation des paramètres du modèle au moyen d'algorithmes souvent coûteux) mais aussi du manque de données targets pour avoir une estimation fiable de la distribution des targets. Nous avons également détaillé ce qui pouvait être un frein à l'adoption des approches discriminantes pour traiter la problématique one-class en présentant les difficultés liées à la synthèse des données outliers nécessaires pour ce type d'approches. Ainsi, résoudre ce problème permettrait de bénéficier d'une catégorie de méthodes beaucoup plus large avec des perspectives d'amélioration des approches existantes.
Dans ce chapitre, nous avons choisi de nous orienter vers les méthodes d'ensemble d'arbres de décision et d'exploiter au mieux les principes de combinaison et de randomisation proposés pour élaborer un système de classification pour les images alvéoscopiques. Ainsi nous proposons la méthode des forêts aléatoires one-class ou one-class random forests (OCRF) basée sur les ensembles d'arbres de décision. Plus précisément, la méthode OCRF tire partie des principes de randomisation proposés dans l'algorithme Forest-RI [Breiman, 2001], à savoir le Bagging [Breiman, 1996], le Random Feature Selection [Breiman, 2001] et auquel nous ajoutons le Random Subspace Method [Ho, 1998]. Nous expliquons dans ce chapitre notre démarche pour l'élaboration des OCRF en présentant les principes mis en œuvre.

Les différents principes de randomisation dont nous nous servons agissent à la fois au niveau du set d'apprentissage avec le bagging en sous-échantillonnant les données et le Random Subspace Method en projetant ces mêmes données dans un sous-espace de l'espace de caractéristiques. Les raisons pour lesquelles nous avons choisi ces mécanismes sont essentiellement liées à la problématique de la malédiction de la dimension. En effet, au lieu de considérer l'espace complet de caractéristiques, les données outliers sont générées dans le sous-espace RSM tout en nous basant sur la distribution initiale des données targets. L'étude de cette distribution nous permet alors d'identifier les régions de l'espace faiblement peuplées en données targets. Deux nouveaux paramètres permettent alors de contrôler la génération des données outliers et influent donc sur le comportement de notre méthode : le premier contrôle le nombre de données générées et le second le domaine de génération.
Nous étudions ensuite ces paramètres et analysons leurs impacts sur la performance des OCRF. Pour cela, nous conduisons, dans un premier temps, des expérimentations sur des bases de données artificielles. Les bases artificielles permettent en effet d'analyser en finesse l'influence d'une configuration donnée des bases d'apprentissage en modifiant par exemple un des paramètres de la distribution véritable fixée des données en présence. Ensuite nous analysons les performances des OCRF sur des bases réelles publiques couramment utilisées dans la littérature afin de valider le caractère générique des OCRF et leur robustesse avec un paramétrage par défaut déterminé à l'aide de l'étude précédente. En raison du déséquilibre naturel des bases one-class couramment utilisées dans la littérature, nous présentons nos résultats à l'aide notamment du coefficient de corrélation de Matthew [Matthews, 1975]. Il s'agit d'une mesure de corrélation se basant sur la matrice de confusion, très peu mentionnée dans la littérature et que nous montrons pourtant plus à même de rendre compte des performances des approches one-class évaluées que par exemple le taux de reconnaissance global [Baldi et al., 2000]. Puis, nous comparons statistiquement les résultats des OCRF avec ceux obtenus sur les mêmes problèmes avec des approches one-class de la littérature,

à savoir notamment les approches par estimation de densité et une approche de type SVM. Nous appliquons dans un dernier temps notre approche OCRF avec le paramétrage par défaut aux images alvéoscopiques.

Dans la première section nous présentons les OCRF et décrivons chacune des étapes constitutives de la méthode en insistant particulièrement sur la génération des données outliers. Dans la seconde section nous analysons le comportement de notre méthode vis-à-vis de paramètres déterminants identifiés, en détaillant les expérimentations sur plusieurs bases synthétiques. Dans la troisième section nous présentons les résultats obtenus sur les bases réelles et dans la dernière section nous proposons une discussion sur la méthode one-class proposée et plus généralement l'approche par méthodes d'ensemble pour traiter le problème one-class.

4.2 Les forêts aléatoires one-class (OCRF)

Nous avons vu dans le premier chapitre la famille des forêts aléatoires (RF) [Breiman, 2001], ensemble de méthodes d'ensemble d'arbres de décision. Nous avons également vu que les RF sont particulièrement compétitives avec le SVM, l'une des approches les plus couramment citées [Robnik-Sikonja, 2004] et ont des performances comparables voire meilleures que celles de l'algorithme Adaboost utilisant le principe du boosting [Kuncheva and Rodríguez, 2007, Breiman, 2001, Breiman, 1998, Freund and Schapire, 1996, Cutler and Zhao, 2001, Rodriguez et al., 2006]. Les OCRF bénéficient ainsi des bonnes propriétés statistiques des forêts aléatoires tout en apportant une solution générique pour le traitement de la problématique one-class à l'aide d'une approche discriminante. Nous décrivons et détaillons ci-après les différents mécanismes intervenant dans la construction des OCRF et analysons les performances de la méthode avec la mise en œuvre de ces mécanismes.

4.2.1 Principes des OCRF

Les mécanismes mis en oeuvre dans la construction des OCRF se situent à différents niveaux. S'agissant d'une approche discriminante, les OCRF requièrent les données des deux classes à discriminer. Ainsi, les données de l'out-of-class doivent être synthétisées afin de construire le classifieur. Nous proposons une approche de génération des données de l'out-of-class en nous servant d'une part de la distribution des données targets disponibles (la distribution des outliers est alors dite "distribution-based") et d'autre part des mécanismes de combinaison et de randomisation des méthodes d'ensemble d'arbres.
Tout d'abord nous considérons un mécanisme d'extraction de connaissance avec lequel nous obtenons de l'information du set d'apprentissage en évaluant la distribution des données target disponibles. Ces informations sont utilisées dans la phase de génération des données outliers afin de mieux contraindre la distribution des outliers. Ensuite nous appliquons les mécanismes de randomisation et de combinaison bagging et Random Subspace Method (RSM) pour générer de façon efficace les données de l'out-of-class. Puis, chaque arbre de la forêt est induit sur les sets alors augmentés (contenant à la fois les targets et les outliers) ; nous bénéficions à nouveau de la randomisation injectée à l'aide de la sélection d'attributs en chacun des nœuds de l'arbre (Random Feature Selection).

Nous reprenons sur la Figure 4.1 le schéma de la fin du chapitre précédent pour illustrer notre propos. Nous développons ci-après les différents mécanismes mis en œuvre.

4.2.2 Mécanismes d'extraction de connaissances et synthèse des données de l'out-of-class

Nous avons ainsi fait le choix d'une méthode d'extraction de l'information a priori à partir du set de données disponible en apprentissage, le choix d'un modèle de distribution des données générées de l'out-of-class avec comme paramètres immédiats la quantité de données à générer et le

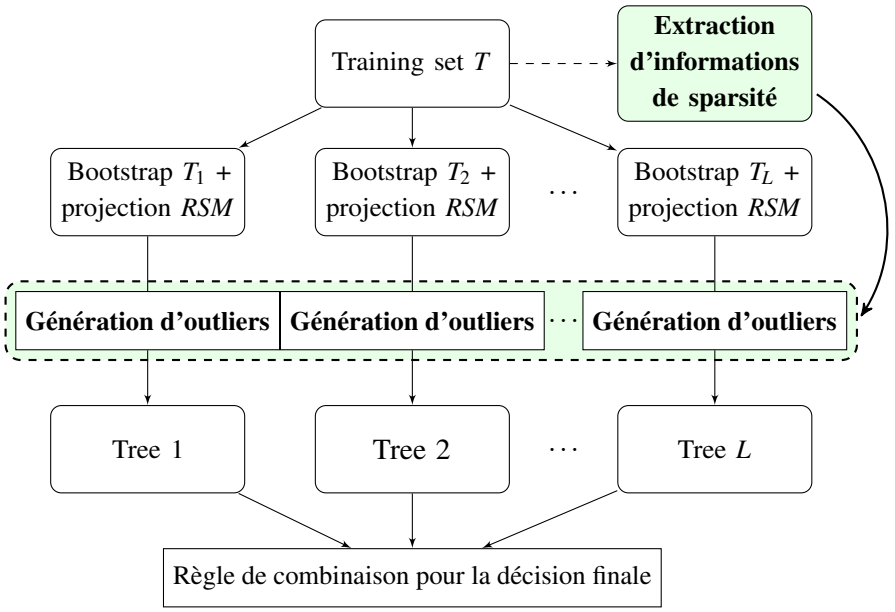


FIGURE 4.1 – Schéma des différentes étapes de l'approche one-class que nous proposons, basée sur les forêts aléatoires.

domaine de génération.

Extraction de connaissances

Dans la première phase de notre approche, nous extrayons de l'information a priori du set d'apprentissage initial des targets. Nous calculons pour cela l'histogramme de la distribution des targets pour chacune des dimensions. Nous faisons l'hypothèse que les données de l'out-of-class se trouvent majoritairement dans les régions faiblement peuplées en données targets (régions sparses). Ainsi, l'information a priori extraite à partir de l'histogramme des données targets permet d'identifier les régions sparses de l'espace susceptibles donc de contenir des données de l'out-of-class. Si H_{target} désigne l'histogramme normalisé en une dimension des données targets sur un attribut donné et si on définit $H_{outlier} = 1 - H_{target}$ celui des outliers (complémentaire de celui des targets), on voit que les maxima de cet histogramme correspondent aux minima des effectifs targets, i.e. une zone sparse, faiblement peuplée en données targets. On est donc en mesure, au moyen de cette information a priori, d'identifier les régions propices de l'espace pour y localiser des données outliers.

Afin donc de contraindre la génération des données de l'out-of-class dans ces régions sparses, nous calculons pour chaque attribut, comme suggéré précédemment, et en amont de l'apprentissage, l'histogramme de la distribution des données outliers comme complémentaire de celui des données targets calculé.

Injection d'aléatoire

Après l'étape d'extraction d'informations a priori, nous sous-échantillonnons le set d'apprentissage tant au niveau des données avec le bagging que des attributs avec le Random Subspace Method (RSM). Cette étape permet, à chaque itération, à la fois de diversifier le set d'apprentissage (il s'agit en effet d'un set boostrap projeté dans un sous-espace de l'espace d'origine), de réduire le nombre d'outliers à considérer en fonction du nombre de targets présents et de drastiquement diminuer les coûts de calculs inhérents à la génération en grande dimension. Dans le cadre des méthodes d'ensemble, cette étape est doublement intéressante à la fois en termes de gains possibles de perfor-

mances (nous savons en effet que la diversité est favorable pour l'obtention de bonnes performances [Breiman, 2001, Kuncheva, 2007, Brown and Kuncheva, 2010]) et en termes de coûts.

Á l'étape suivante, nous générons des données outliers dans ce set bootstrap projeté dans le sous-espace RSM afin de pouvoir induire l'arbre de décision.

Synthèse des données de l'out-of-class

Plusieurs questions sont posées à l'étape de génération des outliers : quand génère t-on les données, selon quelle distribution, avec quelles valeurs de paramètres pour la distribution choisie, dans quel domaine (ou sous-espace délimité), en quelle quantité ? Nous soulignons les points caractéristiques de notre approche en répondant à ces questions.

Á quel moment se fait la génération des données outliers ? La génération des données outliers peut intervenir à plusieurs moments de la construction de l'ensemble d'arbres : (i) en amont de l'apprentissage de l'ensemble, (ii) dans le set d'apprentissage de chaque arbre de décision ou (iii) en chaque nœud des arbres.
La solution (i) peut s'appliquer à n'importe quel classifieur binaire. Cependant, cette approche de génération ne bénéficie pas des mécanismes d'ensemble de la forêt mentionnées précédemment et se retrouve confrontée naturellement aux difficultés des approches de la littérature. La solution (iii) bénéficie pleinement des mécanismes d'ensemble ; cependant, un très grand nombre d'outliers est généré avec cette approche et la complexité de l'arbre est augmentée. De plus un critère d'arrêt est à définir sous peine de continuellement remplir l'espace autour des données targets. L'approche se retrouve alors confrontée aux même difficultés que l'approche Clustering Tree décrite dans le chapitre précédent [Liu et al., 2000].
Nous avons fait le choix de la génération des données outliers selon l'approche (ii), i.e. après l'injection d'aléatoire avec le bagging et le RSM afin de bénéficier des mécanismes de randomisation pour l'introduction de la diversité dans l'induction des arbres comme rappelé précédemment et afin de palier les difficultés inhérentes aux grandes dimensions.

Quelle est la distribution des outliers ? Dans la littérature, la distribution uniforme des outliers est couramment mentionnée comme distribution par défaut des outliers. Cette approche est cependant difficilement réalisable dans la pratique. En considérant la première étape d'extraction d'informations du set d'apprentissage décrite plus haut, nous proposons une approche "distribution-based", i.e. basée sur la distribution des données targets présentes plutôt qu'une approche uniforme. Notre approche consiste alors à générer les outliers suivant la distribution complémentaire des targets $H_{outlier} = 1 - H_{target}$. Ainsi les données outliers sont générées dans un voisinage des targets et non dans tout l'espace de description. Pour des raisons techniques, nous considérons dans une première étude des histogrammes complémentaires en une seule dimension. Une étude en plus grande dimension est envisagée comme une perspective de ces travaux.

La procédure que nous utilisons pour l'obtention de l'histogramme des outliers s'inspire de la théorie mathématique des jeux avec la roue de la fortune biaisée.
Il s'agit de générer des données uniformément sur la surface d'un disque dont les portions sont biaisées. En effet, les tailles des portions du disque sont proportionnelles aux valeurs de l'histogramme. On retrouve la procédure de roue de la fortune biaisée notamment dans les algorithmes génétiques où la sélection d'un individu est proportionnellement conditionnée à sa capacité d'adaptation. L'approche proposée nous permet donc de biaiser la génération uniforme en modulant la probabilité de générer une donnée dans un intervalle donné. Ainsi, il s'agit pour nous de transformer une distribution uniforme en une distribution obéissant à des contraintes de localisation. Avec cette procédure, nous générons davantage de données outliers dans les zones sparses de l'espace (i.e. les zones associées aux fortes valeurs de l'histogramme des outliers). Nous illustrons ce principe de génération des outliers sur la figure 4.2.

L'algorithme 1 résume les différentes étapes de réalisation de l'approche one-class OCRF.

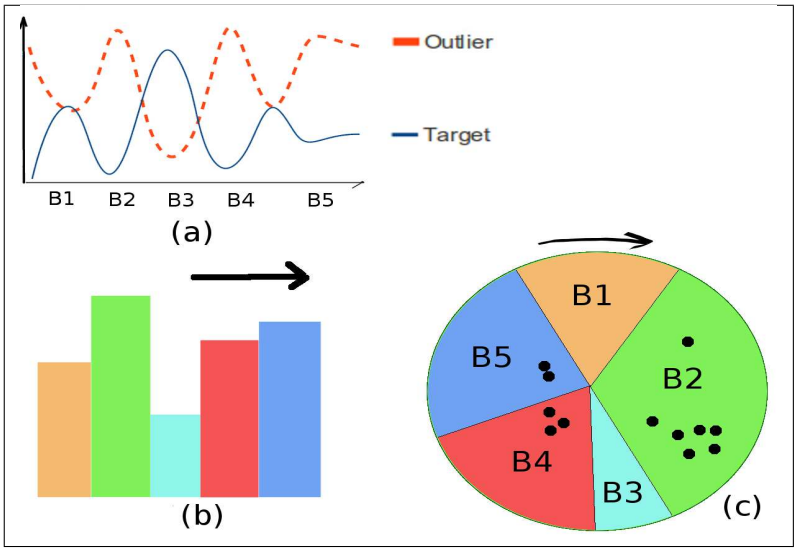


FIGURE 4.2 – Approche de génération des outliers dans les zones sparses de l'espace de description avec la procédure de roue de la fortune biaisée ; la distribution des outliers est déduite de celle des targets en identifiant les zones de faible population (a) ; un histogramme est formé à partir de la distribution des outliers (b) ; la génération se fait alors selon le principe de la roue (c), des outliers sont générés en plus grand nombre dans les portions plus grandes de la roue, associées aux zones de faible population en données targets.

Algorithm 1 Apprentissage des OCRF

Require : Une base d'apprentissage T, le nombre d'outliers à générer $N_{outlier}$, le domaine de définition des outliers $\Omega_{outlier}$, le nombre d'arbres de la forêt L, la dimension des sous-espaces RSM K_{RSM}

Ensure : Un classifieur de type forêt aléatoire

(A) ***Extraction des informations a priori***
Calculer H_{target}, histogramme normalisé des données targets pour chaque attribut
Calculer $H_{outlier}$, histogramme de la distribution des outliers, tel que $H_{outlier}$ soit le complément de H_{target}, i.e. $H_{outlier} = 1 - H_{target}$

(B) ***Génération des outliers et induction de la forêt***
for $l = 1$ to L **do**
    (i) Sélectionner un échantillon bootstrap T_l à partir du set d'apprentissage
    (ii) Projeter ce set bootstrap dans un sous-espace de dimension K_{RSM} aléatoirement choisi par la méthode random subspace ;
    (iv) Générer $n_{outlier}$ données outliers selon la distribution complémentaire définie par l'histogramme complémentaire $H_{outliers}$ dans le domaine $\Omega_{outlier}$, de telle façon que la probabilité de générer un outlier dans un intervalle donné soit proportionnelle à la valeur associée aux bins de l'histogramme $H_{outlier}$ associés à cet intervalle.
    (v) Induire un arbre de décision classique sur ce set augmenté composé des éléments targets et des outliers générés.
end for
return Un modèle de forêt aléatoire one-class

4.2.3 Discussion sur la paramétrisation de la méthode

La procédure de génération des données outliers fait intervenir deux paramètres : le premier contrôle le nombre de données à générer et le second permet d'étendre le domaine de génération. En ce qui concerne le domaine de génération, nous faisons l'hypothèse que le domaine des données de l'out-of-class englobe celui des données targets avec par exemple un hypercube ou une boule englobante. Nous discutons de ces deux paramètres ci-après.

Pour étudier l'influence du nombre d'outliers dans la génération, nous définissons le facteur β contrôlant le nombre $N_{outlier}$ de données outliers par rapport au nombre N_{target} de données targets initialement présents, avec $N_{outlier} = \beta \cdot N_{target}$. Concernant le paramètre lié au domaine, nous définissons un facteur α permettant de contrôler la taille du domaine $\Omega_{outlier}$ des outliers (côté de l'hypercube ou diamètre de la boule englobante), proportionnelle à Ω_{target}, domaine des targets, avec $\Omega_{outlier} = \alpha \cdot \Omega_{target}$.
Nous voyons que si α est choisi trop large, le volume du domaine des données sera grand et les outliers seront sparses dans l'espace de description et essentiellement localisés entre le domaine englobant des données targets et celui des outliers, i.e. les outliers seront faiblement présents entre les clusters du domaine des targets. Intuitivement, plus le domaine de génération est élargi plus la valeur du paramètre β doit être augmentée afin de maintenir un échantillon d'outliers suffisamment représentatif de la distribution. Si α a une valeur trop faible, les données outliers peuvent recouvrir tout ou partie des données targets, le voisinage du domaine des targets demeurant non représenté. Une estimation de la distribution des outliers devient alors difficile pour le classifieur en raison de ce recouvrement.

De même, si β a une valeur trop faible, trop peu de données outliers seront générées pour permettre l'identification par le classifieur des zones sparses, la distribution des outliers ne pourra être estimée de façon fiable. Au contraire, si β est trop important, le risque immédiat est d'obtenir un recouvrement trop important des données target, engendrant une trop forte confusion entre les deux classes.
Nous voyons ainsi se dégager un compromis entre les valeurs de α et β. Nous faisons l'hypothèse que les valeurs de α et β peuvent être dépendantes du problème et de la base de données à analyser, dans une mesure que nous étudions dans les expérimentations menées dans ce chapitre. Les OCRF comportent également les paramètres classiques liés par construction aux forêts aléatoires comme la dimension du sous-espace RSM (K_{RSM}), le nombre d'attributs sélectionnés aléatoirement en chaque nœud de l'arbre lors de son induction (K_{RFS}), le nombre minimal de données requises pour le partitionnement d'un nœud (n_{min}) et le nombre d'arbres composant la forêt (L). Nous utiliserons, sauf mention contraire, les valeurs standards de la littérature [Breiman, 2001, Geurts et al., 2006, Bernard et al., 2009] en ce qui concerne le nombre d'arbres de la forêt, K_{RFS} et n_{min} que nous préciserons ci-après.

Nous menons dans les sections suivantes une étude approfondie des OCRF afin d'analyser le comportement de la méthode à la fois sur des données synthétiques, des bases réelles publiques et les bases d'images alvéoscopiques de notre problématique médicale.

4.3 Étude des paramètres des OCRF

Nous avons évoqué dans la section précédente la présence de deux paramètres importants pour les OCRF contrôlant la génération des données outliers. Nous étudions dans cette section l'influence de ces paramètres. Étant donné que les OCRF se basent sur les forêts aléatoires, naturellement les paramètres des forêts sont également à définir comme évoqué précédemment. L'objectif de cette étude est d'analyser le comportement de la méthode en fonction des valeurs choisies pour ces paramètres.
Pour cela, nous choisissons de travailler dans un premier temps avec des bases de données artificielles. Nous disposons ainsi de plusieurs jeux de données dont nous connaissons les propriétés, les distributions véritables et dont nous pouvons mesurer plus finement l'impact sur les OCRF. Nous

avons généré cinq problèmes avec 5 distributions distinctes : une distribution gaussienne, trois clusters gaussiens, une distribution en forme de banane pour étudier l'influence de la convexité sur le tracé de la frontière de décision du classifieur, une distribution en forme de beignets pour observer l'influence de l'imbrication d'une classe dans une autre sur les performances de la méthode et enfin une distribution elliptique pour étudier l'adaptation à l'échelle de la méthode (cf Figure 4.3).

Nous étudions dans un premier temps le paramètre α, puis dans un second temps le paramètre β. Nous montrons que les valeurs optimales de ces deux paramètres sont effectivement dépendantes des problèmes analysés mais que pour une grande partie de ces problèmes, des valeurs de compromis peuvent être trouvées. Avec ces paramètres de compromis, nous validons dans un premier temps l'approche de génération par roue de la fortune biaisée choisie en comparant les résultats obtenus sur ces bases avec une distribution uniforme. Nous montrons ainsi que la localisation des outliers dans les zones sparses permet de mieux rendre compte de la distribution des outliers et permet donc d'améliorer la fiabilité de l'estimation de cette distribution.
Puis nous comparons la performance des OCRF à celles de plusieurs méthodes one-class de l'état de l'art comme l'estimateur gaussien, la mixture de gaussiennes, l'estimateur de Parzen et le SVDD. Ces algorithmes standards sont implémentés dans "Pattern Recognition/Data Description Toolbox" (PRTools/DDTools) [Tax, 2005, Duin, 2000].

Nous détaillons ci-après le plan des expérimentations menées et analysons les résultats obtenus.

4.3.1 Étude du paramètre α : contrôle de l'extension du domaine de génération

Le paramètre α contrôle l'extension du domaine de génération des outliers par rapport à celui des targets. Nous faisons varier le paramètre α, les autres paramètres étant fixés. Nous définissons ci-après le protocole expérimental utilisé.

4.3.1.1 Protocole expérimental

Nous présentons dans le détail ci-dessous les bases analysées, leur paramétrisation et la procédure d'évaluation et les mesures utilisées pour le calcul des performances.

Les bases de données artificielles

Nous considérons les 5 distributions suivantes : une distribution gaussienne (Gauss), trois clusters gaussiens (Three_gauss), une distribution en forme de banane (Banana), une distribution en forme de beignets (Donut) et une distribution elliptique (Ellipse). Pour ces 5 distributions, nous définissons la dimension de l'espace de génération, la taille des échantillons d'apprentissage et de test, les paramètres internes à chacune des distributions. Les outliers de test sont générés par défaut selon une distribution uniforme dans un hypercube englobant celui des données targets, avec $\Omega_{outlier-test} = \gamma \cdot \Omega_{target}$, avec $\gamma = 1.5$ comme suggéré dans la toolbox [Tax, 2005, Duin, 2000]. Pour chacune de ces bases, nous faisons varier la dimension avec les valeurs $m \in [2;160]$. $N_{target-app} =$ 2000 données targets sont générées dans la phase d'apprentissage. Le set de test est composé de $N_{outlier-test} = N_{target-test} = 10000$ données targets et outliers.

Une illustration de ces différentes distributions est présentée dans la Figure 4.3. Nous montrons également sur cette figure la distribution des outliers générés en utilisant l'approche par roue de la fortune biaisée.

Distribution gaussienne (Gauss) Chaque exemple de la distribution gaussienne est généré selon une loi normale centrée réduite, i.e. centré à l'origine de l'espace de description et de matrice de variance-covariance l'identité.

Distribution des trois clusters gaussiens (three_gauss) Pour les trois clusters gaussiens, trois distributions gaussiennes sont générées comme précédemment, avec des centres séparés d'une distance au moins égale à $10 \cdot \sqrt{m}$, où m est la dimension de l'espace de description.

Distribution elliptique Les données de la distribution elliptique sont générées à partir d'une distribution uniforme dans la boule unité comme expliqué dans [Tax and Duin, 2002, Duin, 2000]. De cette distribution uniforme, on applique un facteur d'échelle égale à 2 à un quelconque des attributs de l'espace afin de créer une élongation des valeurs dans la direction de cet attribut.

Distribution en forme de beignet La distribution en forme de beignets (appelée "Donut") est obtenue en translatant les points d'une distribution normale centrée réduite. Chaque point $\mathbf{x}$ de la distribution en forme de beignet est obtenu en translatant d'un vecteur t un point x_0 de la distribution gaussienne, la longueur du vecteur de translation dépendant inversement de la norme de $\mathbf{x}_0$:

$$\mathbf{x} = \mathbf{x_0} \cdot (1 + \frac{1}{||\mathbf{x_0}||})$$

où $||\cdot||$ est la norme euclidienne.

Distribution en forme de banane La construction de cette distribution se fait à partir d'un schéma 2D auquel on superpose une distribution normale centrée réduite. Une distribution en dimension m paire est obtenue en concaténant ces distributions 2D intermédiaires [Duin, 2000, Tax, 2005].

Paramètres des OCRF

Pour la synthèse des données outliers, nous générons la distribution complémentaire des targets selon la roue de la fortune biaisée et nous faisons le choix immédiat comme domaine de génération d'un hypercube englobant celui des données targets. Nous faisons varier le paramètre α dans l'intervalle [0.5 ;3], avec $\beta = 1$ fixé pour toute cette expérience. Cette valeur indique simplement que nous générons autant de données outliers que de données targets présentes initialement dans le set d'apprentissage et donc $N_{outlier} = N_{target}$. Le nombre de bins utilisés pour le calcul des histogrammes des données targets et outliers est fixé à 50.

Les paramètres standards des OCRF sont les suivants :

- le nombre d'arbres de la forêt est $L = 200$; ce nombre est considéré comme suffisant pour un grand nombre de problèmes [Geurts et al., 2006]
- la valeur du paramètre du RFS est $k_{RFS} = \sqrt{M}$
- celle du RSM est $k_{RSM} = 10$ ou $k_{RSM} = M$ si $M < 10$; ce nombre a été fixé empiriquement, comparativement à la valeur $M/2$ standard dans la littérature [Ho, 1998]

Procédure d'évaluation et mesure des performances

En étant dans une configuration idéale avec un grand nombre de données, nous constituons un seul jeu d'apprentissage et de test pour chacune des bases étudiées. Les performances des OCRF sont indiquées en termes de taux de reconnaissance globale ("accr"), taux de reconnaissance sur les données targets ("T") et celui sur les données outliers ("O").

4.3.1.2 Résultats et analyse

Les résultats de cette expérience sont présentés sur les graphiques de la Figure 4.4 pour les 5 bases étudiées en fonction de α.

On observe, dans la plupart des courbes, trois phases : une première phase de faibles performances pour de faibles valeurs de α puis une seconde phase avec une montée brutale suivie d'une dernière phase où les performances se stabilisent (accompagnées d'une très légère baisse). En outre, plus la dimension est grande, meilleures sont les performances de la méthode. Cette augmentation des performances avec la dimension peut être due au fait que le problème devient de plus en plus "facile" pour les OCRF car les partitions créées de l'espace deviennent de plus en plus volumineuses et que les outliers de tests, générés uniformément, n'introduisent pas suffisamment de confusion au sein

(a) Données targets et outliers (b) Histogramme 1D des targets (c) Données outliers générées

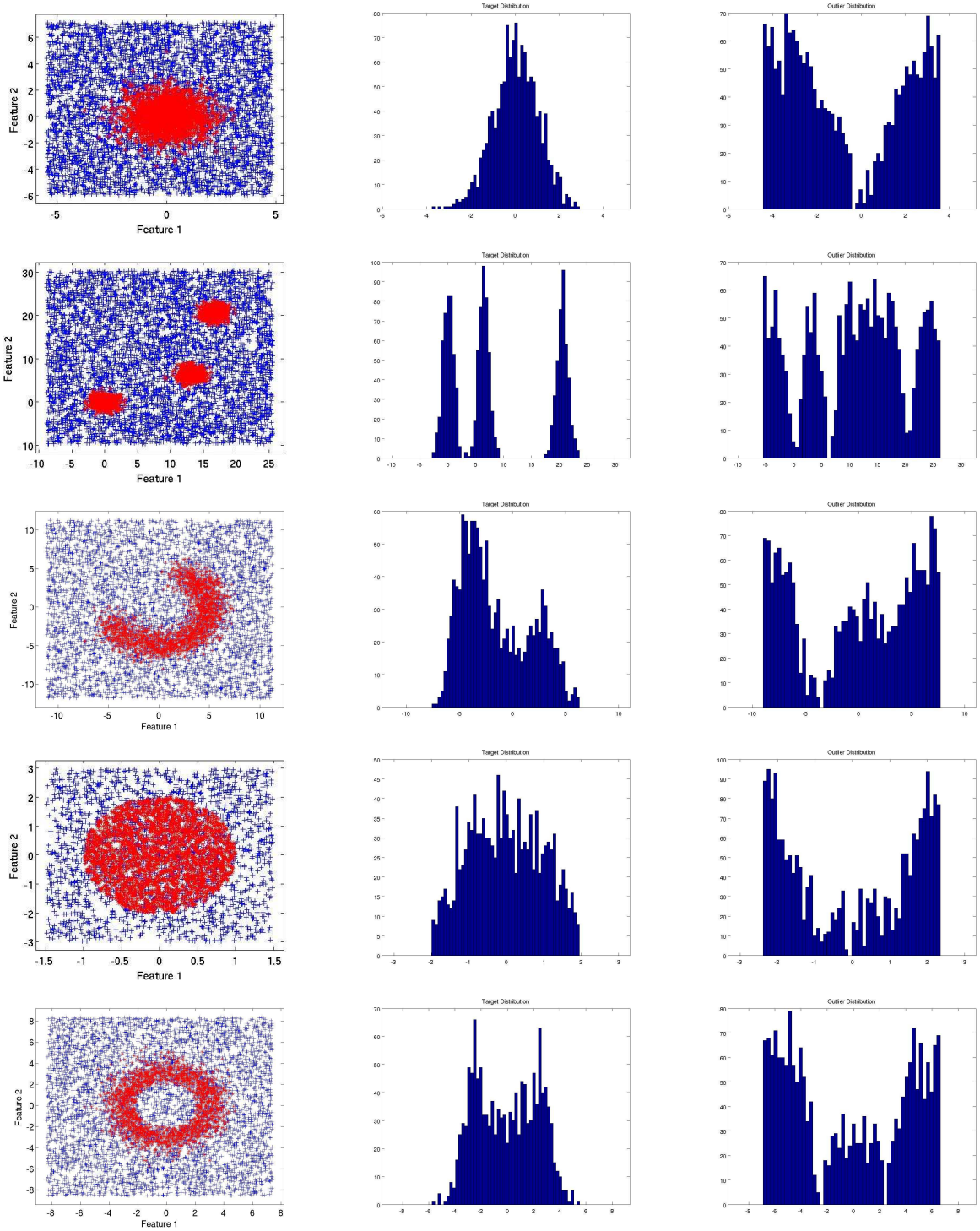


FIGURE 4.3 – Illustration 2D des distributions artificielles étudiées en (a), l'histogramme 1D des données targets en (b) et celui de la distribution des données outliers générées par les OCRF par la roue de la fortune biaisée en (c) ; la première ligne correspond à la distribution gaussienne ; la seconde aux trois clusters gaussiens, puis à la distribution en forme de banane, en forme d'ellipse et enfin en forme de beignet à la dernière ligne ; concernant la distribution elliptique, on observera que l'élongation est sur l'axe des ordonnées avec un facteur 2.

de ces partitions.
Dans la première phase, nous avons de petites valeurs de α, typiquement $\alpha \in [0.5;0.75]$, avec un taux de reconnaissance globale autour de 50%. Cette faible performance peut être due au fait que pour des petites valeurs, les données outliers générées recouvrent les données targets et entrainent ainsi une confusion élevée. Dans la dernière phase, nous avons de grandes valeurs de α, typiquement $\alpha > 1.5$. La légère décroissance observée peut s'expliquer par le volume trop important du domaine englobant dans lequel les outliers sont générés. Le nombre d'outliers générés n'est pas suffisant pour maintenir l'estimation rencontrée aux maxima des courbes rencontrés dans la seconde phase. Dans cette seconde phase, typiquement $\alpha \in [0.75;1.5]$, on observe, pour toutes les distributions, un gain de 40 à 50%, avec un maxima très proche de la valeur $\alpha = 1$. Il se dégage ainsi un bon compromis pour le choix de α dans la partie droite de cet intervalle avec $\alpha \in]1;1.5]$.

Nous tenons compte de cette plage de compromis pour la suite de nos expériences.

4.3.2 Étude du paramètre β : contrôle du nombre d'outliers à générer

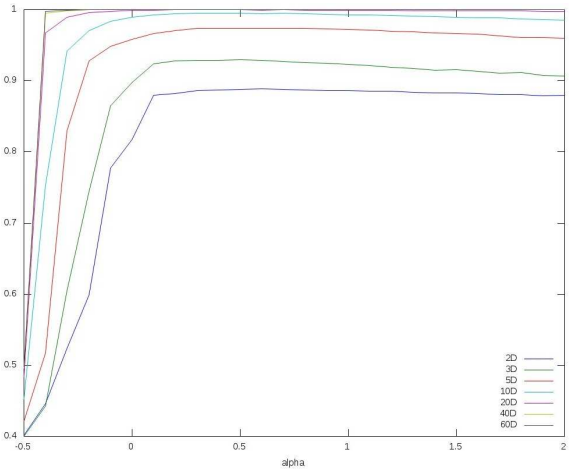
De même que dans l'étude précédente, nous évaluons dans cette expérimentation le paramètre β afin d'identifier l'existence d'une valeur optimale et si possible générique pour les bases évaluées. Nous reprenons exactement le même protocole que le précédent évaluant l'influence de α, à l'exception que dans cette expérience, on choisit $\alpha = 1.2$ et on fait varier β dans l'intervalle $]0;10]$. Nous présentons les résultats obtenus sur les graphiques de la Figure 4.5. Nous observons sur ces figures des courbes en cloche indiquant la présence de maxima pour tous les problèmes considérés. Nous voyons également l'existence d'un compromis dans le voisinage des maxima en prenant β dans l'intervalle $[1;10]$. Le début et la fin de la cloche sont caractérisés par une baisse importante des performances pour la plupart des problèmes comme Banana et Gauss. Ceci peut s'expliquer par deux effets extrêmes. D'une part, pour de faibles valeurs de β, le nombre d'outliers n'est pas suffisant pour une bonne estimation de la distribution des outliers. D'autre part, pour de grandes valeurs de β, trop de données outliers sont générées, engendrant un recouvrement des données targets. On remarque également que les conclusions présentées ici sont en écho de celles formulées pour le paramètre α ; ceci tend à montrer l'interdépendance de ces deux paramètres. L'existence de valeurs de compromis pour ces deux paramètres permet de fixer des valeurs standards qui rendent les OCRF génériques.

Une analyse plus fine du comportement à la fois des targets et des outliers est présentée sur la Figure 4.6 pour $m = 3$. On observe un comportement inverse entre les targets et les outliers. Lorsque $\beta << 1$, on constate que les OCRF reconnaissent bien les targets au détriment des outliers. Lorsque $\beta >> 1$, le contraire se produit. Cette observation confirme bien l'analyse précédente.

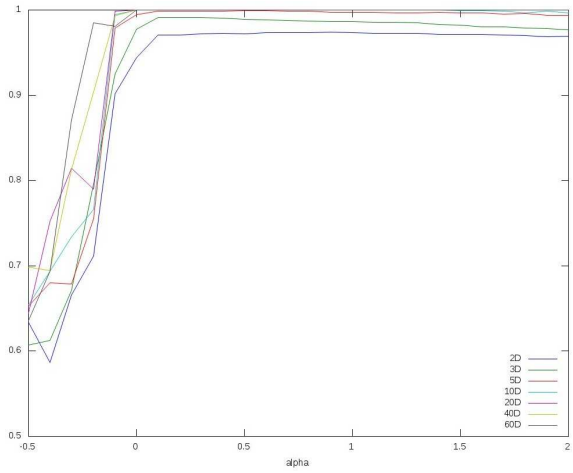
Nous illustrons sur les graphiques de la Figure 4.7 un exemple de partitionnement d'un arbre des OCRF pour le problème Banana en 2D pour différentes valeurs de β. Sur cette figure, nous voyons qu'avec une faible valeur de β, le partitionnement est grossier par manque de données (a) puis s'affine en décrivant mieux les données targets en (b) et (c). Nous remarquons la capacité de l'arbre à traiter les espaces ouverts comme outliers même en présence de peu d'outliers comme la zone définie par $\{(x,y), x < 0, y > 0\}$.

4.3.3 Validation de la distribution par roue de la fortune biaisée vs distribution uniforme

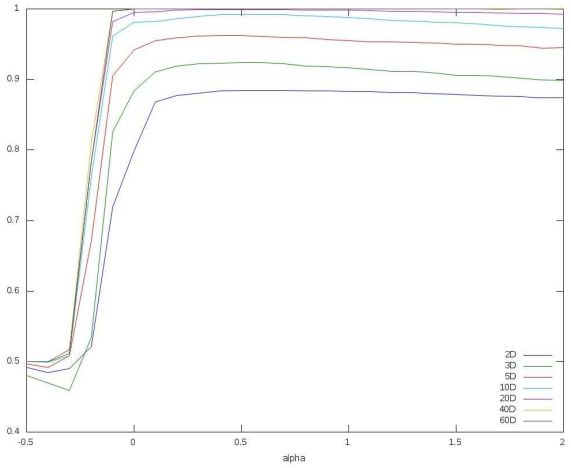
Nous voulons valider dans cette expérience le bon choix d'une distribution tenant compte des informations a priori sur la localisation des outliers par rapport à une distribution uniforme. Nous reprenons exactement le protocole décrit à la section 4.3.1, à l'exception des paramètres α et β pour lesquels nous pouvons désormais fixer les premières valeurs. Nous choisissons $\alpha = 1.2$ et $\beta = 1$.
Pour différencier les OCRF avec notre approche de génération par roue de la fortune biaisée et l'approche avec une distribution uniforme des outliers, nous appelons cette dernière OCRF-U, U étant pour "Uniforme". Les résultats obtenus sont présentés dans le Tableau 4.3.3. Nous voyons



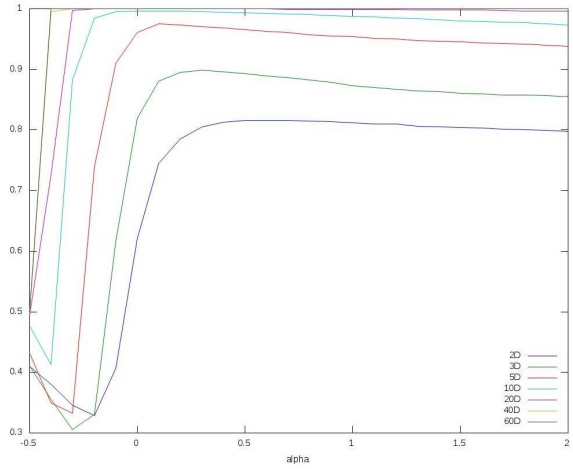
(a) Distribution gaussienne



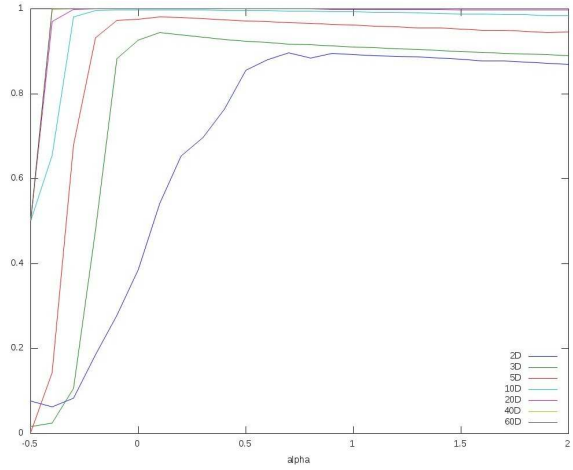
(b) Distribution de 3 clusters gaussiens



(a) Distribution en forme de banane

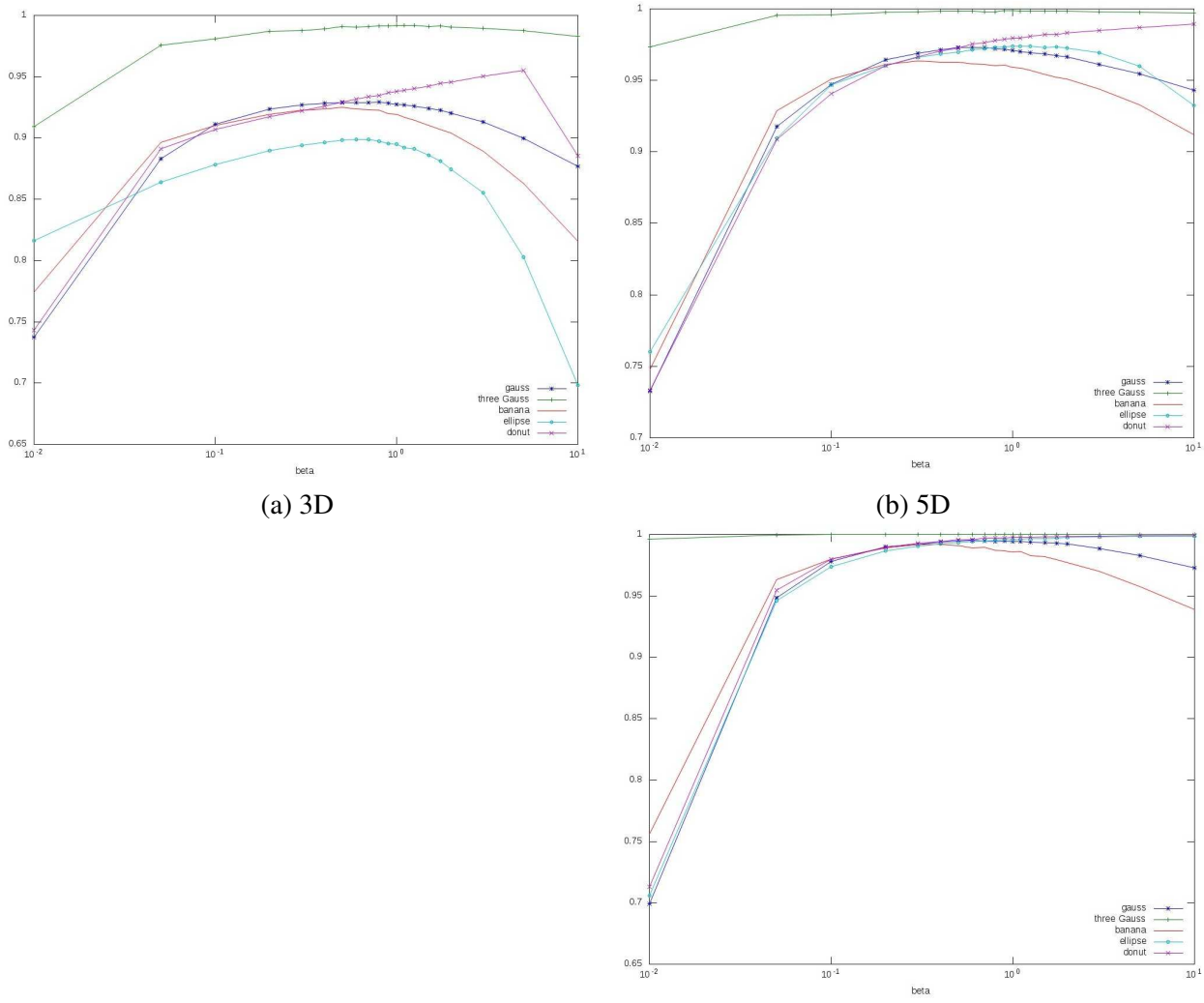


(b) Distribution elliptique



(b) Distribution en forme de beignet

FIGURE 4.4 – Performance globale des OCRF en fonction de α contrôlant le diamètre du domaine de génération des outliers ($\Omega_{outlier} = \alpha\Omega_{target}$; Ω_{target} est actuellement l'hypercube englobant des données targets. Ici en abscisse est représentée la valeur a telle que $a = \alpha - 1$

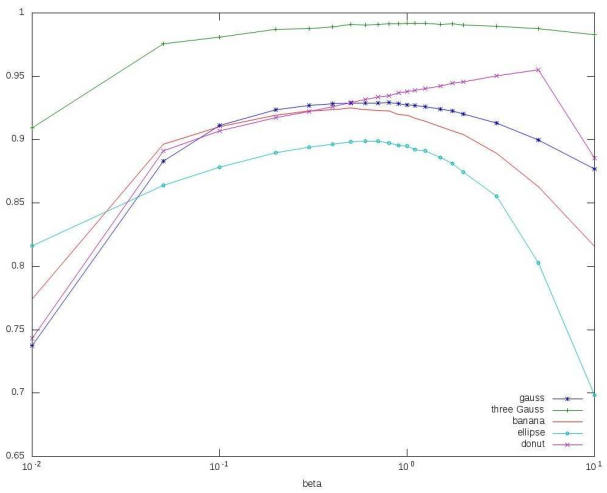


(a) 3D

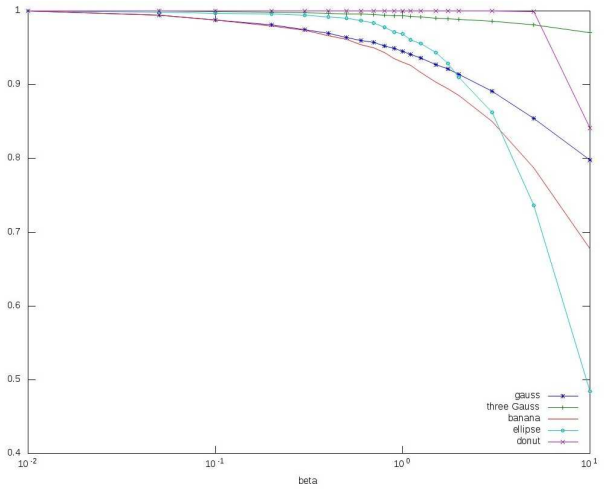
(b) 5D

(c) 10D

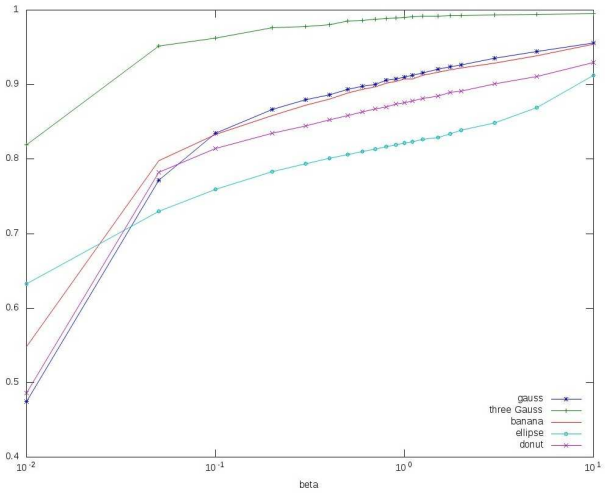
FIGURE 4.5 – Performance des OCRF sur les différentes bases artificielles en fonction du nombre d'outliers générés contrôlé par β, pour plusieurs dimensions ; sur chaque graphique sont affichés les courbes de performances associées à chacune des bases ; chaque graphique correspond à une dimension évaluée : 3D en (a), 5D en (b) et 10D en (c).



(a) Taux de reconnaissance global



(b) Taux de reconnaissance sur les targets



(c) Taux de reconnaissance sur les outliers

FIGURE 4.6 – Taux de reconnaissance des OCRF sur les données targets et outliers pour la dimension m=3, pour les 5 bases artificielles ; le graphique en (a) reprend le taux de reconnaissance globale ; en (b) le taux de reconnaissance sur les targets et en (c) celui sur les outliers.

que les OCRF sont plus performantes que les OCRF-U sur toutes les bases. Nous observons que les OCRF arrivent particulièrement à mieux identifier les targets que les OCRF-U avec par exemple dans le cas de la base Ellipse 5D un gain de près de 10%.

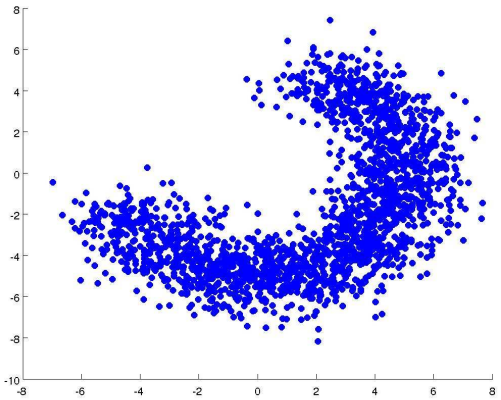
Ces résultats valident bien notre approche de génération des outliers par localisation des zones sparses dans l'espace de description en étant plus à même de représenter les données outliers à la frontière des données targets sans créer un recouvrement comme dans le cas de la génération uniforme.

4.3.4 Comparaison OCRF vs classifieurs one-class standards

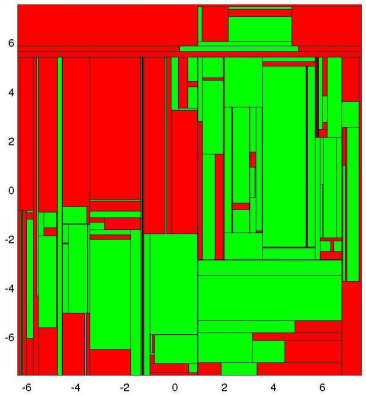
Dans cette expérience, nous comparons les performances des OCRF avec celles de classifieurs standards de la littérature. Nous reprenons le protocole de la section 4.3.1 en fixant les valeurs de $\alpha = 1.2$ et $\beta = 1$ et en ajoutant la paramétrisation des classifieurs de notre comparatif.

Protocole expérimental

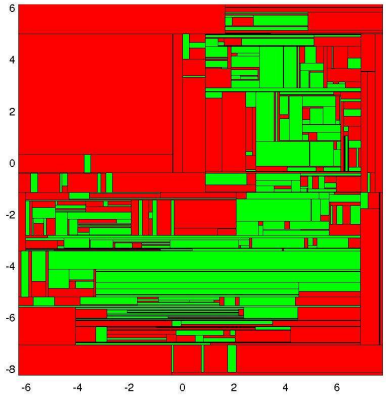
Les classifieurs utilisés dans ce comparatif sont l'estimateur gaussien (Gauss), la mixture de gaussiennes (MoG), l'estimateur de Parzen qui sont implémentés dans la toolbox PRTools



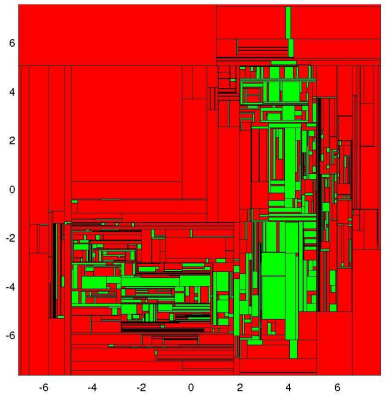
(a) Distribution des targets



(b) $\beta = 0.1$



(c) $\beta = 1$



(d) $\beta = 10$

FIGURE 4.7 – Illustration du partitionnement effectué par un arbre des OCRF selon plusieurs valeurs de β pour la base de données Banana en 2D.

TABLE 4.1 – Performances des OCRF comparées à celles des OCRF-U utilisant une distribution uniforme des outliers ; les meilleurs résultats en terme de taux de reconnaissance globale sont en gras.

Base	OCRF-U	OCRF
gauss_2D	85.60|79.76|91.44	**88.52**|87.96|89.08
gauss_5D	90.26|81.22|99.29	**94.41**|90.40|98.41
gauss_10D	95.85|91.72|99.98	**98.94**|98.05|99.84
gauss_50D	99.97|99.94|100.00	**100.00**|100.00|100.00
3_gauss_2D	62.81|26.45|99.16	**63.93**|28.87|98.98
3_gauss_5D	64.36|28.71|100.00	**65.57**|31.13|100.00
3_gauss_10D	65.24|30.48|100.00	**66.35**|32.70|100.00
3_gauss_50D	66.66|33.32|100.00	**66.67**|33.33|100.00
ellipse_2D	67.26|52.84|81.68	**73.86**|72.74|74.98
ellipse_5D	91.63|86.34|96.91	**95.91**|95.93|95.89
ellipse_10D	98.31|96.75|99.88	**99.74**|99.99|99.49
ellipse_50D	100.00|100.00|100.00	100.00|100.00|100.00
banana_2D	86.04|84.04|88.04	**88.12**|90.04|86.20
banana_5D	90.23|82.01|98.44	**92.59**|87.45|97.72
banana_10D	94.56|89.22|99.90	**97.84**|95.98|99.70
banana_50D	99.90|99.81|100.00	**100.00**|100.00|100.00
donut_2D	58.24|30.70|85.78	**60.11**|37.34|82.88
donut_5D	96.15|93.82|98.47	**96.94**|96.06|97.81
donut_10D	98.01|96.05|99.97	**99.50**|99.16|99.84
donut_50D	100.00|99.99|100.00	100.00|100.00|100.00

[Duin, 2000, Tax, 2005] et le classifieur one-class SVM (OCSVM) qui est disponible dans la toolbox LibSVM [Chang and Lin, 2001].

Les différentes méthodes de la PRTools définissent une mesure de densité ou de distance permettant d'identifier les données outliers à partir d'un seuil sur ces mesures. Pour déterminer ce seuil, un paramètre *fracrej* est introduit représentant la fraction de données targets en apprentissage qui seront considérées comme des données outliers. Plus précisément, les différentes valeurs de sortie des classifieurs pour les données targets en apprentissage sont calculées et triées ; le seuil choisi correspond à la valeur pour laquelle la fraction *fracrej* de données targets auront été rejetées. Cette fraction doit être fournie par l'utilisateur. Dans toutes ces méthodes, le paramètre *fracrej* est fixé par défaut dans la toolbox à 0.05.

L'estimateur gaussien L'estimateur gaussien est calculé à partir de la distance de Mahalanobis avec une matrice de variance-covariance Σ et une espérance μ estimée sur les données d'apprentissage. La fonction de densité fournit alors, pour tout vecteur d'entrée $\mathbf{x}$:

$$f(\mathbf{x}) = (\mathbf{x}-\mu)^T \cdot \Sigma^{-1} \cdot (\mathbf{x}-\mu)$$

La Mixture de Gaussiennes (MoG) La mixture de gaussiennes est une combinaison linéaire de noyaux gaussiens. Par défaut $N_{mog} = 5$ noyaux sont considérés et les coefficients de la combinaison et les paramètres des noyaux sont estimés par l'algorithme Espérance-Maximisation (EM) [Dempster et al., 1977, Bishop, 1995, Hastie et al., 2001]. La fonction de densité est donnée par :

$$p_{mog}(\mathbf{x}) = \frac{1}{N_{mog}} \cdot \sum_{j=1..N_{mog}} \alpha_j \cdot g(\mathbf{x}; \mu_j, \Sigma_j)$$

où g est une densité gaussienne, en dimension m :

$$g(\mathbf{x}; \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{m/2} \cdot |\Sigma_j|^{1/2}} exp\{-\frac{1}{2}(\mathbf{x}-\mu_j)^T \cdot \Sigma_j^{-1} \cdot (\mathbf{x}-\mu_j)\}$$

L'estimateur de Parzen L'estimateur de Parzen est une combinaison linéaire de n noyaux, où n est le nombre d'éléments targets en apprentissage. Le noyau utilisé par défaut est un noyau gaussien et on a :

$$p_{parzen}(\mathbf{x}) = \frac{1}{nh^m} \cdot \sum_{j=1..n} g(\mathbf{x}; \mathbf{x}_j, h \cdot I)$$

où I est la matrice identité, m la dimension de l'espace, g la gaussienne définie précédemment et h le paramètre de la largeur de la fenêtre estimé par maximisation de la vraisemblance avec une procédure en leave-one-out [Duin, 1976, Kraaijveld and Duin, 1991].

OCSVM Nous rappelons que OCSVM trace un hyperplan séparant les données targets d'un côté et l'unique représentant de l'out-of-class placé à l'origine de l'espace (cf. chapitre 3, section 2 et [Scholkopf et al., 2001]). Ce séparateur linéaire est rendu plus flexible par l'utilisation de noyaux, à l'instar du SVM. Nous reprenons la paramétrisation par défaut de la LibSVM, à l'exception du paramètre υ dont la valeur est une estimation de la borne inférieure de l'erreur sur les données targets. Ce paramètre est fixé à $\upsilon = 0.5$ dans la LibSVM par défaut alors qu'une valeur couramment utilisée dans la littérature est $\upsilon = 0.1$ que nous utiliserons. Le noyau utilisé par défaut est gaussien de paramètre de variance $s = \frac{1}{m}$, m étant la dimension de l'espace. Le paramètre de coût est quant à lui fixé à $C = 1$.

Nous analysons ci-après les résultats obtenus pour ces différents classifieurs.

Résultats et analyse

Nous présentons les résulats obtenus par les OCRF sur les bases artificielles et comparons avec ceux des classifieurs de l'état de l'art dans le Tableau 4.2. Nous constatons que les OCRF sont en général plus performantes que les méthodes de la PRTools, avec une forte progression lorsque la dimension augmente.
Nous observons que la mixture de gaussiennes s'adapte favorablement à la dimension et que généralement les classifieurs de la PRTools reconnaissent mieux les outliers que les targets, à l'exception des bases en petite dimension, typiquement 2 et 3. Nous voyons clairement l'effet de la dimension sur la plupart des classifieurs dont les performances pour la plupart convergent vers 50%. C'est le cas de Parzen pour toutes les bases, le cas de MoG pour toutes les bases à l'exception de Donut et le cas d'OCSVM pour la base Banana. Dans tous ces cas, à l'exception de la base Three_gauss pour MoG, le classifieur ne prédit que la classe outlier. On remarquera aussi l'absence de valeurs pour les deux bases Ellipse en 140 et 160D pour lesquelles la procédure d'optimisation de MoG n'a pas abouti. En revanche, on observe que OCRF et Gauss sont les approches les plus performantes sur ces bases. Gauss cependant n'arrive pas à traiter la base Ellipse pour laquelle, avec la dimension, il reconnait de moins en moins bien les données targets. Cela peut être du au fait que Gauss ne parvient pas à modéliser les données situées dans la direction de l'élongation de l'ellipse et les considère ainsi comme autant de données outliers du modèle, ce phénomène étant amplifié avec la dimension.

Nous voyons que les OCRF parviennent aisément à résoudre ces différentes tâches en identifiant correctement les données des deux classes, en étant même meilleurs avec la dimension. Ceci montre bien la capacité des OCRF à traiter les problèmes en dimension élevée là où les approches standards discriminantes et génératives que nous avons évaluées ont des difficultés. On ajoutera également la généricité de l'approche, capable de s'adapter à ces différentes problématiques.

4.3.5 Conclusion sur l'étude des paramètres des OCRF

Les expérimentations sur les bases artificielles nous ont permis d'analyser le comportement des OCRF en situation idéale. Nous avons pu étudier l'impact de la dimension, du domaine de génération des outliers, de leur distribution, leur nombre. Nous avons ainsi obtenu les résultats suivants :

- α est le paramètre contrôlant le degré d'extension du domaine de génération des outliers par rapport au domaine des targets : il se dégage un compromis dans la partie droite de l'intervalle]1 ;1.5] correspondant presque à une zone de plateau avec une très légère baisse des performances ; on a constaté qu'une valeur proche de 1 pouvait être instable avec un gradient gauche très grand par rapport au gradient droit ; une valeur trop grande de α fait chuter les performances au delà d'un certain seuil ; ce seuil semble cependant dépendre de la quantité d'outliers générés.
- β est le paramètre contrôlant le nombre de données outliers générées : la valeur optimale de β dépend fortement de la nature des données ; on observe en effet des maxima distincts pour différents problèmes. Toutefois, une valeur dans la partie droite de l'intervalle [1 ;10] se dégage également comme un bon compromis ;
- l'approche de génération des ouliers en extrayant des informations de localisation des zones sparses dans l'espace de description apporte un gain non négligeable aux performances des OCRF et rend la génération plus efficace qu'avec une approche uniforme, même en bénéficiant des mêmes mécanismes de randomisation des méthodes d'ensemble ;
- les OCRF se comportent favorablement en grande dimension sur des données artificielles générées selon plusieurs distributions ayant des propriétés géométriques et statistiques différentes ; elles se révèlent ainsi génériques et performantes pour ces bases avec un paramétrage par défaut.

Nous abordons dans la section suivante les expérimentations menées sur les données réelles en partant des conclusions énoncées dans cette section concernant la paramétrisation des OCRF afin d'évaluer et de consolider cette approche.

TABLE 4.2 – Performances des OCRF comparées à celles des classifieurs one-class sur les bases artificielles et mesurées en taux de reconnaissance globale (Accr), taux de reconnaissance sur les targets ("*T*") et les outliers ("*O*") ; les meilleures performances sont indiquées en gras.

Base	OCRF	OCSVM	Gauss	Parzen	MoG
	Accr|T|O	Accr|T|O	Accr|T|O	Accr|T|O	Accr|T|O
gauss_2D	87.72|91.18|84.26	**94.98**|89.96|100.0	88.51|94.81|82.21	88.33|93.88|82.78	88.47|94.61|82.33
gauss_3D	92.47|94.26|90.67	**95.06**|90.12|100.0	92.67|94.76|90.57	92.40|93.24|91.56	92.59|94.45|90.73
gauss_5D	**97.24**|97.83|96.65	94.96|89.92|100.0	97.04|95.96|98.11	94.07|89.28|98.85	96.72|95.19|98.24
gauss_10D	**99.38**|99.48|99.28	94.43|88.88|99.98	97.31|94.64|99.98	65.98|31.95|100.0	96.30|92.62|99.98
gauss_20D	**99.98**|99.99|99.97	93.51|87.01|100.0	96.63|93.26|100.0	50.00|00.00|100.0	92.96|85.92|100.0
gauss_40D	**100.0**|100.0|100.0	93.92|87.84|100.0	96.42|92.83|100.0	50.00|00.00|100.0	90.13|80.26|100.0
gauss_60D	**100.0**|100.0|100.0	94.46|88.92|100.0	95.78|91.55|100.0	50.00|00.00|100.0	89.81|79.62|100.0
gauss_80D	**100.0**|100.0|100.0	93.76|87.52|100.0	95.21|90.41|100.0	50.00|00.00|100.0	77.62|55.24|100.0
gauss_100D	**100.0**|100.0|100.0	93.21|86.42|100.0	93.29|86.58|100.0	50.00|00.00|100.0	72.39|44.78|100.0
gauss_120D	**100.0**|100.0|100.0	93.43|86.86|100.0	92.42|84.83|100.0	50.00|00.00|100.0	84.17|68.34|100.0
gauss_140D	**100.0**|100.0|100.0	93.79|87.57|100.0	90.82|81.64|100.0	50.00|00.00|100.0	76.60|53.20|100.0
three_gauss_2D	**98.14**|98.83|97.45	94.45|88.91|100.0	92.26|95.74|88.78	96.52|94.70|98.35	96.76|95.19|98.32
three_gauss_3D	**99.24**|99.15|99.32	95.09|90.19|100.0	94.60|95.16|94.05	95.91|92.01|99.82	97.41|95.07|99.74
three_gauss_5D	**99.89**|99.87|99.91	94.04|88.08|100.0	97.26|94.56|99.97	91.36|82.72|100.0	97.63|95.25|100.0
three_gauss_10D	**100.0**|100.0|100.0	92.29|84.59|100.0	97.41|94.83|100.0	58.68|17.35|100.0	96.96|93.92|100.0
three_gauss_20D	**100.0**|100.0|100.0	92.64|85.29|100.0	97.50|95.00|100.0	50.00|00.00|100.0	97.52|95.03|100.0
three_gauss_40D	**100.0**|100.0|100.0	91.33|82.67|100.0	97.30|94.61|100.0	50.00|00.00|100.0	97.28|94.57|100.0
three_gauss_60D	**100.0**|100.0|100.0	91.29|82.58|100.0	97.10|94.21|100.0	50.00|00.00|100.0	97.40|94.80|100.0
three_gauss_80D	**100.0**|100.0|100.0	91.31|82.63|100.0	97.76|95.52|100.0	50.00|00.00|100.0	97.59|95.17|100.0
three_gauss_100D	**100.0**|100.0|100.0	89.47|78.95|100.0	96.96|93.92|100.0	50.00|00.00|100.0	50.00|100.0|00.00
three_gauss_120D	**100.0**|100.0|100.0	89.98|79.96|100.0	97.08|94.17|100.0	50.00|00.00|100.0	50.00|100.0|00.00
three_gauss_140D	**100.0**|100.0|100.0	89.76|79.52|100.0	97.55|95.09|100.0	50.00|00.00|100.0	50.00|100.0|00.00
banana_2D	86.83|90.23|83.42	**92.34**|84.68|100.0	81.51|95.67|67.35	88.05|94.66|81.44	88.11|95.40|80.81
banana_3D	90.73|92.09|89.37	**92.20**|84.39|100.0	88.02|95.04|80.99	90.97|91.22|90.72	91.35|93.27|89.43
banana_5D	**95.70**|95.45|95.95	86.37|72.73|100.0	93.69|95.13|92.25	85.02|71.00|99.04	95.45|94.34|96.56
banana_10D	**98.32**|97.34|99.30	54.39|8.78|100.0	97.01|94.69|99.33	50.02|0.03|100.0	96.81|93.94|99.68
banana_20D	**99.93**|99.88|99.98	50.00|0.00|100.0	97.20|94.39|100.0	50.00|00.00|100.0	95.24|90.47|100.0
banana_40D	**100.0**|100.0|100.0	50.00|0.00|100.0	96.27|92.54|100.0	50.00|00.00|100.0	87.67|75.33|100.0
banana_60D	**100.0**|100.0|100.0	50.00|0.00|100.0	95.26|90.51|100.0	50.00|00.00|100.0	79.55|59.10|100.0
banana_80D	**100.0**|100.0|100.0	50.00|0.00|100.0	94.29|88.58|100.0	50.00|00.00|100.0	72.15|44.29|100.0
banana_100D	**100.0**|100.0|100.0	50.00|0.00|100.0	91.52|83.04|100.0	50.00|00.00|100.0	63.85|27.69|100.0
banana_120D	**100.0**|100.0|100.0	50.00|0.00|100.0	89.94|79.88|100.0	50.00|00.00|100.0	60.92|21.84|100.0
banana_140D	**100.0**|100.0|100.0	50.00|0.00|100.0	88.26|76.51|100.0	50.00|00.00|100.0	50.88|01.75|100.0
ellipse_2D	78.48|88.74|68.21	**95.36**|90.71|100.0	80.67|94.96|66.38	77.84|86.46|69.21	80.68|95.36|65.99
ellipse_3D	89.67|96.32|83.01	**95.35**|90.70|100.0	89.45|94.84|84.05	85.56|85.71|85.41	88.66|93.53|83.78
ellipse_5D	**97.11**|99.35|94.86	94.29|88.57|100.0	95.49|93.83|97.14	78.69|59.16|98.21	94.82|92.73|96.91
ellipse_10D	**99.56**|100.0|99.11	94.45|88.90|100.0	96.94|93.88|99.99	50.00|00.00|100.0	94.37|88.76|99.98
ellipse_20D	**99.98**|100.0|99.96	95.03|90.06|100.0	95.76|91.52|100.0	50.00|00.00|100.0	89.50|79.00|100.0
ellipse_40D	**100.0**|100.0|100.0	94.87|89.74|100.0	92.75|85.49|100.0	50.00|00.00|100.0	72.82|45.63|100.0
ellipse_60D	**100.0**|100.0|100.0	94.94|89.87|100.0	88.90|77.80|100.0	50.00|00.00|100.0	57.87|15.73|100.0
ellipse_80D	**100.0**|100.0|100.0	95.36|90.72|100.0	83.67|67.33|100.0	50.00|00.00|100.0	52.94|05.88|100.0
ellipse_100D	**100.0**|100.0|100.0	94.56|89.12|100.0	75.62|51.23|100.0	50.00|00.00|100.0	50.44|00.87|100.0
ellipse_120D	**100.0**|100.0|100.0	94.89|89.78|100.0	70.53|41.06|100.0	50.00|00.00|100.0	50.09|00.17|100.0
ellipse_140D	**100.0**|100.0|100.0	94.61|89.22|100.0	63.60|27.19|100.0	50.00|00.00|100.0	-|-|-
donut_2D	72.38|64.08|80.68	**96.90**|93.80|100.0	87.44|100.0|74.87	89.88|100.0|79.75	88.85|100.0|77.70
donut_3D	93.82|100.0|87.63	**98.19**|96.38|100.0	94.39|100.0|88.77	93.16|95.94|90.38	94.43|100.0|88.86
donut_5D	97.65|100.0|95.30	87.75|75.50|100.0	**98.65**|100.0|97.30	88.06|77.78|98.33	98.64|100.0|97.27
donut_10D	99.68|100.0|99.35	**99.99**|100.0|99.98	99.97|100.0|99.94	66.68|33.36|100.0	99.97|100.0|99.94
donut_20D	99.99|100.0|99.98	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0
donut_40D	100.0|100.0|100.0	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0
donut_60D	100.0|100.0|100.0	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0
donut_80D	100.0|100.0|100.0	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0
donut_100D	100.0|100.0|100.0	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0
donut_120D	100.0|100.0|100.0	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0
donut_140D	100.0|100.0|100.0	100.0|100.0|100.0	100.0|100.0|100.0	50.00|00.00|100.0	100.0|100.0|100.0

4.4 Évaluation des OCRF sur des bases réelles

Nous avons observé dans la section précédente le comportement des OCRF sur des problèmes synthétiques pour différentes valeurs des paramètres contrôlant la méthode. Nous avons vu que pour une majorité des problèmes une plage de valeurs apparaissait comme un compromis pour obtenir des performances souvent optimales. Nous évaluons dans cette section les performances de la méthode sur des bases réelles de la littérature, issues du répertoire de l'UCI. Nous terminons par l'application de la méthode sur notre base d'images alvéoscopiques. Nous comparons à nouveau les résultats obtenus avec ceux des méthodes one-class standards de la littérature, à savoir les classifieurs précédents : le one-class-SVM, les estimateurs de densité gaussienne, les fenêtres de Parzen et la Mixture de Gaussiennes.
Comme dans les expérimentations précédentes, nous utilisons les valeurs par défaut des paramètres fournis par les deux toolboxes PRTools et LibSVM, sauf mention contraire. Nous insistons sur le fait que nous n'avons pas cherché à optimiser ces paramètres pour chacune des méthodes et chacune des bases testées, l'objectif étant d'évaluer notre méthode, d'étudier son comportement sur plusieurs bases de données de la littérature et d'analyser à travers une étude comparative, le comportement des méthodes de l'état de l'art dans une configuration identique (à paramètres standards fixes pour toutes les expérimentations).

4.4.1 Expérimentations sur des bases publiques

4.4.1.1 Protocole expérimental

Les bases utilisées dans cette expérimentation sont des bases publiques du répertoire de l'UCI [Blake and Merz, 1998a] et les bases des images alvéoscopiques. Nous présentons d'abord la configuration des bases de données utilisées, puis les paramètres des méthodes évaluées et ensuite le procédé d'évaluation ainsi que la mesure de performances adoptée.

4.4.1.1.1 Bases de données

Il est difficile de collecter des bases purement one-class. Les échantillons outliers peuvent être rares ou durs à échantillonner ou pas suffisamment représentatifs de l'"out-of-class". Il n'existe pas à notre connaissance de répertoire publique permettant d'évaluer les approches one-class. Lorsque le problème à traiter est initialement multi-classe, les auteurs découpent généralement la base de données pour former autant de jeux de données binaires qu'il n'a de classes. Une des classes est choisie en tant que classe target et les autres données sont regroupées pour former la classe outlier [Hempstalk and Frank, 2008, Hempstalk et al., 2008, Tax and Duin, 2002, Tax, 2001, Scholkopf et al., 2001]. D'autres auteurs font exactement le contraire [Tian and Gu, 2010, Oliveira et al., 2008, Ratle et al., 2007, Spinosa and Carvalho, 2005, Cao et al., 2003, Tax et al., 1999]. Notre problématique médicale consiste à identifier des images de patients pathologiques en ayant une connaissance certaine pour une seule classe, la classe des images de patients sains. Pour nous mettre dans les mêmes conditions, nous allons utiliser le premier découpage. Nous allons donc utiliser le premier découpage consistant à prendre une classe en tant que target et les autres regroupés en outliers, seules instances disponibles de l'"out-of-class".
Pour valider notre approche, nous utilisons les bases de l'UCI (répertoire de bases de données pour l'apprentissage automatique de l'Université de Carolina Irvine) [Blake and Merz, 1998b]. Nous avons traité les 14 bases listées dans le Tableau 4.3. Ces bases sont couramment mentionnées dans la littérature et leur utilisation pourra permettre ainsi d'établir, autant que possible, des comparaisons avec des résultats déjà mentionnés dans l'état de l'art.

Les bases de données sont associées à des applications diverses parmi lesquelles on retrouve des applications militaires comme "sonar", l'analyse de scènes de crime avec "glass", des applications météorologiques avec "ionosphere", botaniques avec "iris" et "musk", médicales avec "breast-cancer", "diabetes" et des bases de données liées à la reconnaissance de chiffres manuscrits comme "Optical digits" (optdigits), "Pen based" (pendigits), "multiple features dataset" (MFeat)

[Tax and Duin, 2001]. La base "Mfeat" est composée de chiffres manuscrits scannés et est en fait une série de 5 bases de données obtenues pour 5 espaces de description différents (les coefficients de fourier ; les coefficients de la transformation de karhunen-loeve ; les caractéristiques de transformations morphologiques ; la valeur d'intensité des pixels ; les moments de Zernike ; les facteurs de corrélation). Les datasets "mfeat-fourier" et "mfeat-pixel" n'ont pas pu être évalués dans nos expérimentations en raison de calculs n'ayant pas abouti pour MoG. La base "Mfeat" a été notamment utilisée dans [Tax and Duin, 2001] pour étudier différentes approches de combinaison de classifieurs one-class. La base de données "Glass" possède 7 classes décrivant les différents types de verres souvent retrouvés sur des scènes de crimes. Le type 4 est complètement absent de la base d'origine et la classe 6 possède seulement 9 éléments. Pour ces raisons, nos expérimentations n'évaluent pas la classe 4 et la classe 6 n'est pas utilisée comme target mais ajoutée systématiquement à la classe des outliers. Le dataset "Breast Cancer" contient 16 valeurs manquantes qui ont été remplacées arbitrairement par la valeur 0. Nous n'avons pas connaissance d'une étude précédente exhaustive de type one-class sur chacune des classes de ces bases.

Nos expérimentations couvrent des données de dimensions et de natures diverses. Ainsi, la dimension des données est dans l'intervalle [4 ;216], le nombre de classes d'origine dans [2 :10] et le nombre de données disponibles dans l'intervalle [150 :11000]. Comme chacune des classes d'origine sert tour à tour de classe target dans un jeu de données indépendant, nous avons réalisé nos tests sur 78 bases (cf. Tableau 4.3). Nous n'avons effectué aucun pré-traitement sur les données, i.e. pas de normalisation, pas de réduction ou de transformation de nature à mieux représenter les données comme une analyse en composantes principales par exemple.

TABLE 4.3 – Statistiques sur les bases de données utilisées lors de nos expérimentations ; ces données sont issues de l'UCI [Blake and Merz, 1998b]

	Nombre		
Dataset	caractéristiques	classes	instances totales
Sonar	60	2	208
Glass	9	5	214
Ionosphere	34	2	351
OptDigits	64	10	5620
Iris	4	3	150
Musk	166	2	6598
Breast Cancer W.	9	2	699
PenDigits	16	10	10994
Diabetes	8	2	768
Mfeat-factors	216	10	2000
Mfeat-karhunen	64	10	2000
Mfeat-zernike	47	10	2000
Mfeat-morphological	6	10	2000
Datasets "pseudo one-class"		**78**	

Méthodes d'évaluation

Il n'existe pas à notre connaissance d'approches particulières ou communément admises pour évaluer les méthodes one-class. Une approche a été présentée dans [Hempstalk and Frank, 2008] mettant en oeuvre l'utilisation de procédures 10-folds pour des bases multi-classes transformées en bases pseudo one-class. Nous optons pour une procédure 10-folds similaire et pour chaque base de données, une classe est choisie comme target et les données des autres classes sont regroupées dans la classe outlier. Le découpage en folds se fait de la même manière que dans le cas multi-classe : les données des deux classes target et outlier sont partitionnées en 10 sous-ensembles ou folds ; 9 folds serviront à l'apprentissage (les données outliers sont alors exclues de l'apprentissage)

et le dernier fold servira de base de test (contenant à la fois les données targets et outliers). La procédure 10-folds est stratifiée (i.e. les données target et outlier sont dans les mêmes proportions) puis avons répété l'échantillonnage 5 fois (5x10-folds stratifiés) et avons finalement moyenné les résultats obtenus sur ces 50 jeux de données. La procédure 10-folds est souvent utilisée comme un estimateur de la moyenne de l'erreur en généralisation et est un bon compromis dans le cas de bases de données de faibles effectifs [Hastie et al., 2005, Kohavi, 1996]. Les performances des classifieurs sont moyennées sur les 50 jeux de données mais la variance est calculée pour chacune des 5 procédures 10-folds (au lieu de la variance des performances sur les 50 jeux). Avec les caractéristiques mentionnées ci-dessus, nous avons calculé 3900 forêts dans cette étude (une forêt par classe de données et par fold).

Beaucoup de travaux ont été publiés dans la littérature avec différentes méthodes pour traiter les bases que nous nous proposons d'étudier. Il est cependant difficile de trouver un standard en termes méthodologiques pour comparer différents classifieurs sur ces bases. Il est difficile en outre de choisir une mesure de performance adaptée à un problème donné [Baldi et al., 2000]. Beaucoup de mesures ont été proposées comme le taux de performance globale, la sensibilité, la spécificité, la précision, le rappel, les courbes ROC, l'aire sous la courbe ROC (AUC) [Bradley, 1997, Hand, 2009], l'AUC pondéré ou "Weighted-AUC" [Hempstalk and Frank, 2008]. De plus plusieurs critiques ont été émises concernant l'utilisation de l'AUC pour la comparaison de performances des classifieurs. En effet, certaines études n'indiquent pas les courbes ROC associées à ces valeurs d'AUC, permettant notamment de choisir un point de fonctionnement donné des méthodes évaluées. Ainsi, si les courbes ROC des méthodes évaluées se croisent, les AUC peuvent ne pas indiquer clairement si une méthode est préférable à une autre [Lobo et al., 2008, Hand, 2009]. Nous utiliserons dans nos évaluations le coefficient de corrélation de Matthews [Matthews, 1975]. Ce coefficient se base sur les effectifs de la matrice de confusion pour évaluer les performances de la méthode. Elle est indiquée dans le cas de données déséquilibrées et est un bon compromis d'utilisation dans le cas général [Baldi et al., 2000, Yousef et al., 2010]. Cette mesure est souvent mentionnée dans les publications biomédicales dans lesquelles on retrouve des problèmes de données déséquilibrées ; elle est aussi associée aux mesures de précision et sensibilité afin d'identifier la classe impactant les performances.
Le coefficient de Matthew (MCC) est donné par la formule :

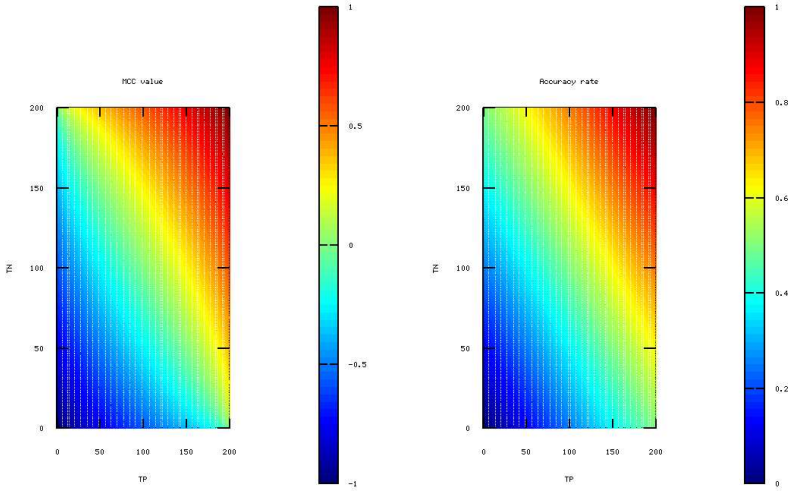
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

TP (true positive) représente l'effectif de données outliers bien identifiées par le classifieur, TN (true negative) celui des données targets correctement identifiées, FN (false negative ou non-détection) l'effectif des données outliers prises pour des données targets et FP (false positive ou fausse alarme) le nombre de données targets identifiées comme des données outliers. MCC mesure le degré de corrélation entre les classes observées et les sorties du classifieur. Ses valeurs sont comprises entre +1 et -1, +1 étant le cas d'une prédiction parfaite, -1 celui d'une inversion des classes par le classifieur, 0 représentant une classification aléatoire ou bien le classifieur ne prédit qu'une seule classe. Cette mesure s'adapte bien à l'estimation non optimale des performances des méthodes sur des données déséquilibrées [Baldi et al., 2000]. MCC est plus à même de tenir compte du déséquilibre des classes en raison de la pondération effectuée tendant à contrebalancer ce déséquilibre comme nous le voyons sur la Figure 4.8. Sur cette figure, nous illustrons les différences entre les mesures MCC et le taux global de reconnaissance. En cas de déséquilibre des classes, nous voyons que la valeur renvoyée par MCC est plus stable que celle du taux global.

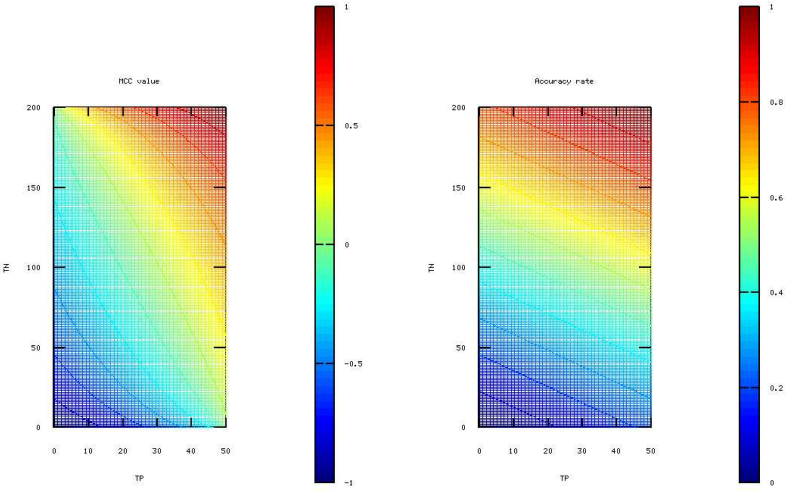
4.4.1.1.2 Comparaison statistique de plusieurs classifieurs

Il n'est pas simple d'évaluer et de comparer différents classifieurs sur de multiples bases de données [Zheng, 1993, Duin, 1996, Hand, 2006, Jamain and Hand, 2008, Demsar, 2006]. Pour la comparaison de classifieurs, plusieurs techniques ont été proposées. Des techniques basées

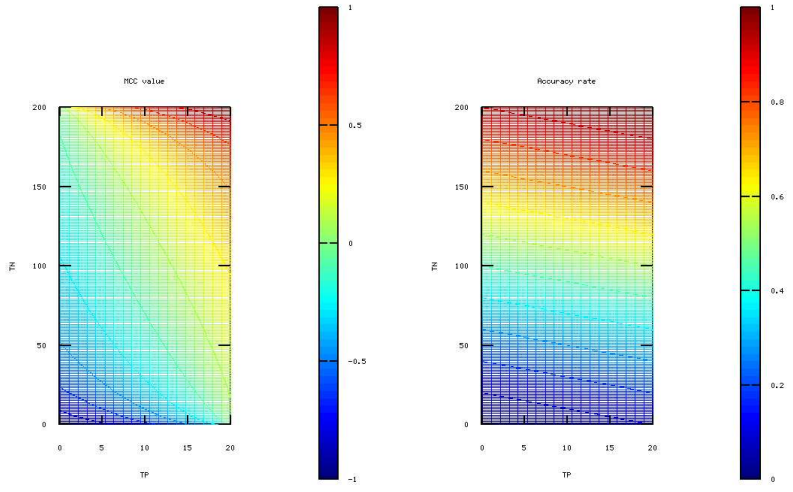
FIGURE 4.8 – Mesures de performances avec le coefficient de corrélation de Matthews (à gauche) vs le taux global de reconnaissance (à droite) pour trois exemples de configuration de bases de données plus ou moins déséquilibrées ; la barre d'échelle à droite des graphiques indique la valeur croissante de la mesure ; on observe une différence croissante entre les deux approches d'autant plus forte que les classes sont déséquilibrées ; en (a) les deux mesures sont à même de fournir des résultats équivalents ; avec le déséquilibre croissant, on voit que le taux global a tendance à favoriser la classe la plus représentée (glissement vers le rouge de la partie haute à gauche des graphiques) alors que c'est plutôt stable pour MCC (la partie rouge demeure dans le coin supérieur droit).



(a) Cas de 200 données targets et 200 données outliers



(b) Cas de 50 données targets et 200 données outliers



(c) Cas de 20 données targets et 200 données outliers

sur le rang moyen ont été proposées [Brazdil and Soares, 2000]. [Dietterich, 1998] propose des tests statistiques pour comparer deux classifieurs : le t-test avec de la validation croisée par re-échantillonnage, le t-test avec du 10-folds, le t-test corrigé tenant compte du recouvrement des données entre les différents jeux de tests générés [Nadeau and Bengio, 2003], le t-test avec 5x2CV, le test de McNemar. Il est plus difficile de définir les tests appropriés dans le cadre de la comparaison de plus de deux classifieurs [Dietterich, 1998, Menke and Martinez, 2004] sur de multiples bases de données. [Demsar, 2006] suggère deux approches pour répondre à ce problème : le test paramétrique ANOVA [Fisher, 1959] et une approche non paramétrique avec le test statistique de Friedman [Friedman, 1937, Friedman, 1940] associé au test post-hoc de Nemenyi [Nemenyi, 1963]. ANOVA est un test paramétrique pour évaluer la différence entre les performances de différents classifieurs. Mais ce test utilise des hypothèses fortes sur les données à analyser (les résultats de performance) qui peuvent ne pas être vérifiées dans le cas courant des bases de données réelles : les données doivent suivre une distribution normale ; les variables aléatoires issues des performances pour un classifieur sur les différentes bases de données doivent avoir la même variance. Ces restrictions nous ont conduit à choisir l'approche non-paramétrique basée sur le test de Friedman. Le test de Friedman est un test non-paramétrique utilisant les rangs de chaque classifieur pour chaque base de données. L'hypothèse nulle stipule que les classifieurs comparés ne sont pas significativement différents. Pour k classifieurs et N datasets, on définit R_j le rang moyen du classifieur j pour toutes les bases de données ; le test de Friedman est alors donné par :

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right)$$

avec k-1 degrés de liberté. Une version améliorée de ce test, moins conservative, a été proposée par Iman et Davenport [Iman and Davenport, 1979] :

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2}$$

qui suit la distribution de Fisher avec (k-1) et (k-1)(N-1) degrés de liberté. Si l'hypothèse nulle est rejetée, un test post-hoc est alors mené comme celui de Nemeyi [Nemenyi, 1963] lorsque les classifieurs sont comparés les uns aux autres ou le test de Bonferroni Dunn [Dunn, 1961] lorsqu'un classifieur est comparé aux autres. Nous avons dans cette étude utilisé le test de Nemeyi donné par la Différence Critique (Critical Difference) :

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

avec q_α valeur critique tabulée.

4.4.1.1.3 Les classifieurs

Nous reprenons les classifieurs suivants pour la comparaison : le one-class SVM [Scholkopf et al., 2001] (OCSVM) adapté de la toolbox LibSVM [Chang and Lin, 2001] ; 3 estimateurs de densité vus dans l'expérimentation sur les bases artificielles (estimateur gaussien Gauss, fenêtres de Parzen, mixtures de gaussiennes MoG) implémentés à partir de la librairie Pattern Recognition Toolbox (PRTools) [Duin, 2000, Tax, 2005]. OCSVM est utilisé avec les paramètres communément fixés dans la toolbox LibSVM (i.e. le coefficient de coût $C = 1$, un noyau RBF, $\gamma = \frac{1}{m}$ pour la largeur du noyau, m étant la dimension de l'espace de description...), à l'exception du coefficient υ. Il est démontré dans [Scholkopf et al., 2001] que ce coefficient représente une borne supérieure pour la fraction de données targets susceptibles d'être rejetées par le modèle et aussi une borne inférieure pour la fraction de vecteurs de support du SVM. Cette valeur est fixée par défaut à 0.5 dans la toolbox LibSVM alors qu'une valeur de 0.1 est davantage utilisée dans la littérature. OCSVM est particulièrement connu pour être instable vis-à-vis du choix du noyau (gaussien par défaut) et notamment pour le choix du paramètre υ [Hempstalk et al., 2008, Manevitz and Yousef, 2002, Tax and Duin, 2002,

Scholkopf et al., 2001]. Nous avons donc choisi d'utiliser la valeur couramment mentionnée dans la littérature, i.e. $\upsilon = 0.1$. Pour les estimateurs de densité, nous avons utilisé les paramètres définis par défaut dans la toolbox PRTools (seuils sur la probabilité, largeur de bande des noyaux, matrice de covariance, nombre de clusters gaussiens). La PRTools définit notamment le paramètre $fracrej = 0.05$, correspondant à la fraction de données targets considérées comme des données outliers dans la phase d'apprentissage des méthodes. Ce paramètre permet notamment de définir le seuil sur la probabilité en sortie de l'estimateur. Comme lors de l'expérimentation sur les bases artificielles, nous choisissons de ne pas faire varier ce paramètre.

Nous n'avons pas cherché à optimiser ces paramètres présentés comme standards dans la littérature sur les bases que nous avons testées. Nous insistons sur le fait que les résultats auraient sans doute pu être meilleurs avec des paramètres optimisés pour ces bases par des procédures de type validation croisée mais notre objectif ici est d'observer le comportement des méthodes en nous plaçant dans un cadre suffisamment générique pour traiter un grand nombre de problèmes différents. Nous avons appliqué la même restriction pour notre approche en fixant les mêmes valeurs de paramètres pour toutes les bases de données testées. Ces valeurs de paramètres sont essentiellement celles déclarées dans les précédentes expérimentations que nous reprenons ci-après :

- le nombre d'arbres de la forêt vaut $L = 200$;
- les arbres sont complètement développés avec $n_{min} = 2$;
- la dimension des sous-espaces RSM vaut $K_{RSM} = 10$ ou $K_{RSM} = m$ si $m < 10$, m étant la dimension de l'espace ;
- le nombre d'attributs sélectionnés aléatoirement en chaque nœud de l'arbre est $K_{RFS} = \sqrt{m}$;
- le facteur d'extension du domaine hypercube des targets pour former le domaine de génération des outliers vaut $\alpha = 1.2$ reprenant une valeur de compromis trouvée à la section 4.3.1 ; on a $\Omega_{outlier} = \alpha \cdot \Omega_{target}$;
- le facteur contrôlant le nombre d'outliers en fonction du nombre de targets disponibles vaut $\beta = 10$ en tenant compte de la suggestion de se rapprocher de la partie droite de l'intervalle de compromis [1 ;10] déterminé à la section 4.3.2 ; on a $N_{outlier-app} = \beta \cdot N_{target-app}$;

4.4.1.2 Résultats et analyse

Afin de montrer l'intérêt de la mesure MCC pour l'analyse des performances, nous présentons des cas d'études dans le Tableau 4.4. Dans ce tableau, nous détaillons les résultats de trois classifieurs one-class (OCRF, OCSVM et Gauss) obtenus sur deux bases de données (*opdigits* et *mfeat-factors*) avec les mesures MCC, taux global de reconnaissance (Accr), taux de reconnaissance des données targets (T) et des données outliers ("O"). Nous rappelons que le coefficient MCC permet de tenir compte du déséquilibre des bases de données one-class : plus la valeur du coefficient est proche de 1, meilleur est le classifieur ; une valeur nulle indique généralement que le classifieur ne prédit qu'une seule des deux classes en présence, la seconde n'étant pas identifiée correctement ; plus la valeur est proche de -1, moins bonnes sont les performances du classifieur. Les résultats montrent que MCC est plus à même de fournir une mesure fiable des performances et du comportement du classifieur que la mesure du taux de reconnaissance global.

Nous retrouvons dans le Tableau 4.4 trois cas illustrant l'observation précédente :

(i) MCC et Accr ont des valeurs élevées lorsque le classifieur obtient de bonnes performances sur les deux classes target et outlier (e.g. OCRF sur la base *optdigits_0*)

(ii) MCC et Accr ont tous deux de faibles valeurs lorsque le classifieur obtient de mauvais résultats sur les deux classes à la fois (e.g. OCRF sur *optdigits_1*)

(iii) MCC a une valeur nulle ou proche de 0 alors que Accr a une valeur élevée (e.g. OCSVM sur toutes les bases présentées dans ce tableau) ; dans ce dernier cas, avec la valeur élevée de Accr, on pourrait croire que le classifieur a de bonnes performances globales alors que ce n'est pas le cas.

Donc nous constatons que le taux de reconnaissance global ne permet pas de rendre compte des performances d'un classifieur lorsque les données targets ou outliers n'ont pas été correctement

identifiées par le classifieur alors que MCC permet de mieux rendre compte de ce comportement en ayant une valeur proche de 0. Ainsi, la suite des résultats de ces expérimentations ne sera présentée qu'en terme de MCC.

L'ensemble des résulats obtenus sur toutes les bases et tous les classifieurs one-class est présenté dans le Tableau 4.6 avec la valeur moyenne du MCC. Nou résumons ce tableau avec le Tableau 4.5. On voit ainsi que les OCRF sont performantes en général avec de fortes valeurs de MCC (typiquement $MCC > 0.5$) pour 45 bases de données sur les 78 évaluées. De plus, la méthode est la meilleure (rang = 1) sur 23 problèmes. Ces premiers résultats montrent le caractère générique des OCRF, i.e. leur capacité à traiter raisonnablement bien des problèmes divers. Comparés aux OCRF, MoG n'arrive en tête que dans 9 cas, OCSVM dans 7 cas et Parzen dans 4 cas. Gauss est la meilleure des méthodes évaluées en arrivant en tête 35 fois sur 78. Ainsi, les OCRF et Gauss se montrent plus performants que les autres méthodes évaluées. Ce résultat sera confirmé par l'étude statistique que nous menons ci-après.
Nous remarquons que les OCRF n'obtiennent à aucun moment de valeurs négatives ou nulles, contrairement à Gauss et les autres méthodes : Gauss a 2 valeurs négatives sur les bases "Ionosphere bad" et "diabete negative", MoG 1 valeur négative et 12 valeurs nulles, Parzen a 2 valeurs négatives et 37 valeurs nulles et OCSVM a 1 valeur négative pour 32 valeurs nulles. Nous rappelons que les valeurs MCC négatives indiquent que le classifieur a inversé les classes véritables et ainsi n'a pas du tout pu reconnaitre les données. Les valeurs faibles correspondent en général aux performances très faibles sur les deux classes et les valeurs nulles sont obtenues lorsque le classifieur prédit les classes d'appartenance de façon complètement aléatoire ou ne prédit qu'une seule des deux classes. Les OCRF montrent ainsi leur robustesse sur des problématiques variées par rapport aux autres méthodes évaluées.
On observe par exemple dans le Tableau 4.6 que OCSVM et Parzen prédisent toujours la classe outlier dans le cas des bases pendigits, optdigits et mfeat-zernike et que Parzen et MoG prédisent toujours la classe target pour la base mfeat-factor. Ainsi, en comparant les méthodes one-class sur leur capacité à bien traiter les deux classes target et outlier, on voit que les deux méthodes les mieux placées sont OCRF et Gauss. Nous voyons ainsi que les OCRF ont un comportement plus stable que les approches de l'état de l'art, performantes (avec les meilleurs résultats dans 23 cas), générique et compétitive par rapport à Gauss qui s'est révélé plus performant en général. Nous confirmons ces résultats statistiquement dans la section suivante.

TABLE 4.4 – Études de cas pour OCRF, OCSVM et Gauss pour les deux mesures MCC et taux de reconnaissance globale sur les bases (a) *optdigit* (*opt_N*) et (b) *mfeat-factors* (*fact_N*) ; les meilleurs résultats sont en gras ; T désigne le taux de reconnaissance sur les données targets et O celui sur les données outliers. On observe en général une meilleure synthèse des performances des classifieurs avec MCC plutôt qu'avec le taux global de reconnaissance.

		OCRF	OCSVM	Gauss
opt_0	MCC	0,776	0,165	**0,954**
	Acc	0,94	0,90	0,99
	T	0,99	0,04	0,92
	O	0,94	1,00	1,00
opt_1	MCC	0,147	0,054	**0,937**
	Acc	0,26	0,90	0,99
	T	1,00	0,01	0,90
	O	0,18	1,00	1,00
opt_2	MCC	0,143	0,000	**0,953**
	Acc	0,26	0,90	0,99
	T	1,00	0,00	0,92
	O	0,18	1,00	1,00
opt_3	MCC	0,121	0,000	**0,914**
	Acc	0,22	0,90	0,98
	T	1,00	0,00	0,92
	O	0,13	1,00	0,99
opt_4	MCC	0,077	0,000	**0,905**
	Acc	0,16	0,90	0,98
	T	1,00	0,00	0,92
	O	0,06	1,00	0,99
opt_5	MCC	0,041	0,000	**0,954**
	Acc	0,12	0,90	0,99
	T	1,00	0,00	0,92
	O	0,02	1,00	1,00
opt_6	MCC	0,410	0,026	**0,956**
	Acc	0,70	0,90	0,99
	T	1,00	0,00	0,92
	O	0,67	1,00	1,00
opt_7	MCC	0,264	0,000	**0,933**
	Acc	0,48	0,90	0,99
	T	1,00	0,00	0,91
	O	0,42	1,00	1,00
opt_8	MCC	0,043	0,000	**0,719**
	Acc	0,12	0,90	0,94
	T	1,00	0,00	0,91
	O	0,02	1,00	0,94
opt_9	MCC	0,077	0,000	**0,860**
	Acc	0,15	0,90	0,97
	T	1,00	0,00	0,90
	O	0,06	1,00	0,98

(a)

		OCRF	OCSVM	Gauss
fact_0	MCC	**0,844**	0,000	0,737
	Acc	0,97	0,90	0,96
	T	0,86	0,00	0,58
	O	0,98	1,00	1,00
fact_1	MCC	**0,873**	0,000	0,712
	Acc	0,98	0,90	0,95
	T	0,79	0,00	0,54
	O	1,00	1,00	1,00
fact_2	MCC	**0,879**	0,000	0,740
	Acc	0,98	0,90	0,96
	T	0,85	0,00	0,58
	O	0,99	1,00	1,00
fact_3	MCC	**0,887**	0,017	0,695
	Acc	0,98	0,90	0,95
	T	0,87	0,00	0,51
	O	0,99	1,00	1,00
fact_4	MCC	**0,884**	0,000	0,743
	Acc	0,98	0,90	0,96
	T	0,82	0,00	0,58
	O	1,00	1,00	1,00
fact_5	MCC	**0,843**	0,013	0,738
	Acc	0,97	0,90	0,96
	T	0,82	0,00	0,58
	O	0,99	1,00	1,00
fact_6	MCC	**0,910**	0,068	0,770
	Acc	0,98	0,90	0,96
	T	0,85	0,02	0,62
	O	1,00	1,00	1,00
fact_7	MCC	**0,879**	0,017	0,841
	Acc	0,98	0,90	0,97
	T	0,83	0,00	0,73
	O	1,00	1,00	1,00
fact_8	MCC	0,613	0,000	**0,647**
	Acc	0,91	0,90	0,94
	T	0,83	0,00	0,45
	O	0,91	1,00	1,00
fact_9	MCC	**0,866**	0,026	0,751
	Acc	0,98	0,90	0,96
	T	0,85	0,01	0,60
	O	0,99	1,00	1,00

(b)

TABLE 4.5 – Synthèse des résultats du tableau 4.6 ; le rang 1 correspond à la meilleure place en terme de performance ; pour le nombre de valeurs de MCC plus grandes que 0.5, la valeur seuil 0.5 est un exemple indiquant des performances généralement élevées (cf. figure 4.8)

	OCRF	OCSVM	Gauss	Parzen	MoG
#MCC négatifs	0	1	2	2	1
#MCC nuls	0	32	0	37	12
$\#MCC > 0.5$	45	12	60	8	39
# rang=1	23	7	35	4	9

TABLE 4.6 – Résultats des OCRF en terme de du coefficient MCC comparé à tous les classifieurs sur l'ensemble des bases de données ; le taux de performance global est précisé entre parenthèses et les meilleures valeurs MCC sont indiquées en gras. *bwc est la base breast-cancer du Wisconsin

Dataset	OCRF	OCSVM	Gauss	Parzen	Mog
iris_versicolour	0,579 (81,5)	0,897 (95,3)	**0,903** (95,6)	0,685 (85,6)	0,607 (82,9)
iris_virginica	0,614 (82,7)	**0,900** (95,5)	0,813 (90,9)	0,716 (87,3)	0,604 (82,5)
iris_setosa	0,722 (87,1)	0,903 (95,6)	**0,921** (96,4)	0,799 (90,9)	0,643 (83,3)
*bwc_benign	**0,919** (96,2)	0,848 (92,1)	0,902 (95,3)	0,709 (83,2)	0,867 (93,3)
*bwc_malignant	**0,629** (81,3)	0,208 (68,2)	0,179 (46,3)	0,273 (69,1)	0,084 (49,6)
ionosphere_good	0,683 (83,3)	**0,785** (89,5)	0,781 (89,3)	0,180 (40,8)	0,584 (75,4)
ionosphere_bad	**0,169** (56,7)	-0,348 (28,2)	-0,410 (26,0)	0,106 (64,7)	-0,346 (33,2)
musk_0	0,071 (84,6)	0,103 (21,0)	**0,264** (83,6)	0,180 (30,6)	0 (84,6)
musk_1	0,306 (49,3)	0,049 (84,7)	**0,818** (95,1)	0,495 (88,9)	0 (15,4)
sonar_mines	0,048 (53,3)	**0,882** (93,6)	0,342 (65,9)	0 (46,2)	0,222 (47,8)
sonar_rocks	0,179 (59,0)	**0,889** (94,0)	0,120 (56,3)	0 (53,8)	0,274 (56,1)
diabetes_positive	0,139 (46,4)	0 (65,2)	0,147 (35,2)	0,188 (55,3)	0,219 (39,2)
diabetes_negative	0,241 (68,7)	0 (34,8)	-0,046 (66,5)	0,064 (53,9)	0,020 (68,3)
pendigits_0	**0,976** (99,6)	0 (89,6)	0,970 (99,4)	0,100 (89,7)	0,961 (99,3)
pendigits_1	0,585 (85,8)	0 (89,6)	0,652 (90,0)	0,212 (90,1)	**0,835** (96,6)
pendigits_2	0,835 (96,3)	0 (89,6)	**0,957** (99,2)	0 (89,6)	0,956 (99,2)
pendigits_3	0,918 (98,5)	0 (90,4)	**0,969** (99,5)	0,092 (90,4)	0,949 (99,1)
pendigits_4	0,961 (99,3)	0 (89,6)	**0,969** (99,4)	0 (89,6)	0,953 (99,1)
pendigits_5	0,756 (94,1)	0 (90,4)	0,880 (97,8)	0,092 (90,4)	**0,942** (99,0)
pendigits_6	**0,985** (99,7)	0 (90,4)	0,970 (99,5)	0 (90,4)	0,954 (99,2)
pendigits_7	0,887 (97,6)	0 (89,6)	0,887 (97,7)	0 (89,6)	**0,937** (98,8)
pendigits_8	0,634 (89,3)	0 (90,4)	0,716 (93,2)	0 (90,4)	**0,951** (99,2)
pendigits_9	0,577 (85,9)	0 (90,4)	0,577 (86,9)	0,093 (90,4)	**0,936** (98,9)
optdigits_0	0,776 (94,2)	0,165 (90,5)	**0,954** (99,2)	0 (90,1)	0,745 (95,9)
optdigits_1	0,147 (26,2)	0,054 (89,9)	**0,937** (98,9)	0 (89,8)	0,803 (96,7)
optdigits_2	0,143 (25,8)	0 (90,1)	**0,953** (99,2)	0 (90,1)	0,755 (96,0)
optdigits_3	0,121 (21,7)	0 (89,8)	**0,914** (98,4)	0 (89,8)	0,727 (95,5)
optdigits_4	0,077 (15,6)	0 (89,9)	**0,905** (98,3)	0 (89,9)	0,766 (96,1)
optdigits_5	0,041 (11,5)	0 (90,1)	**0,954** (99,2)	0 (90,1)	0,738 (95,8)
optdigits_6	0,410 (70,3)	0,026 (90,1)	**0,956** (99,2)	0 (90,1)	0,778 (96,3)
optdigits_7	0,264 (48,2)	0 (89,9)	**0,933** (98,8)	0 (89,9)	0,777 (96,3)
optdigits_8	0,043 (11,7)	0 (90,1)	**0,719** (93,6)	0 (90,1)	0,696 (95,2)
optdigits_9	0,077 (15,2)	0 (90,0)	**0,860** (97,4)	0 (90,0)	0,739 (95,7)
...suite des résultats sur la page suivante					

Dataset	OCRF	OCSVM	Gauss	Parzen	Mog
mfeat_factors_0	**0,844** (97,2)	0 (90,0)	0,737 (95,8)	0 (10,0)	0 (10,0)
mfeat_factors_1	**0,873** (97,8)	0 (90,0)	0,712 (95,4)	0 (10,0)	0 (10,0)
mfeat_factors_2	**0,879** (97,9)	0 (90,0)	0,740 (95,8)	0 (10,0)	0 (10,0)
mfeat_factors_3	**0,887** (98,0)	0,017 (90,0)	0,695 (95,1)	0 (10,0)	0 (10,0)
mfeat_factors_4	**0,884** (98,0)	0 (90,0)	0,743 (95,8)	0 (10,0)	0 (10,0)
mfeat_factors_5	**0,843** (97,3)	0,013 (90,0)	0,738 (95,8)	0 (10,0)	0 (10,0)
mfeat_factors_6	**0,910** (98,5)	0,068 (90,2)	0,770 (96,2)	0 (10,0)	0 (10,0)
mfeat_factors_7	**0,879** (97,9)	0,017 (90,0)	0,841 (97,3)	0 (10,0)	0 (10,0)
mfeat_factors_8	**0,613** (90,6)	0 (90,0)	0,647 (94,5)	0 (10,0)	0 (10,0)
mfeat_factors_9	**0,866** (97,6)	0,026 (90,1)	0,751 (96,0)	0 (10,0)	0 (10,0)
mfeat_karhunen_0	**0,807** (96,4)	0,363 (91,6)	0,784 (96,5)	0 (90,0)	0,302 (90,1)
mfeat_karhunen_1	0,750 (95,5)	0,248 (90,9)	**0,765** (96,1)	0 (90,0)	0,247 (90,3)
mfeat_karhunen_2	0,755 (95,5)	0,222 (90,7)	**0,776** (96,3)	0 (90,0)	0,213 (90,1)
mfeat_karhunen_3	0,703 (93,8)	0,239 (90,8)	**0,759** (96,0)	0,213 (90,1)	0,213 (90,1)
mfeat_karhunen_4	**0,813** (96,5)	0,192 (90,6)	0,794 (96,6)	0 (90,0)	0,257 (90,1)
mfeat_karhunen_5	0,622 (91,6)	0,167 (90,5)	**0,730** (95,7)	0,213 (90,1)	0,229 (90,2)
mfeat_karhunen_6	0,684 (93,8)	0,255 (90,9)	**0,790** (96,5)	0,224 (90,2)	0,232 (90,2)
mfeat_karhunen_7	**0,864** (97,7)	0,532 (93,2)	0,849 (97,5)	0,213 (90,1)	0,257 (90,4)
mfeat_karhunen_8	0,407 (78,5)	0,030 (90,1)	**0,713** (95,4)	0 (90,0)	0,213 (90,0)
mfeat_karhunen_9	0,752 (95,4)	0,315 (91,2)	**0,770** (96,2)	0,213 (90,1)	0,220 (90,1)
mfeat_zernike_0	0,697 (93,5)	0 (90,0)	**0,944** (99,0)	0 (90,0)	0,637 (94,4)
mfeat_zernike_1	0,663 (92,7)	0 (90,0)	**0,908** (98,4)	0 (90,0)	0,686 (95,0)
mfeat_zernike_2	0,679 (93,5)	0 (90,0)	**0,903** (98,3)	0 (90,0)	0,512 (93,0)
mfeat_zernike_3	0,365 (77,3)	0,017 (90,0)	**0,674** (91,8)	0,213 (90,1)	0,617 (94,2)
mfeat_zernike_4	0,461 (84,2)	0 (90,0)	**0,908** (98,3)	0 (90,0)	0,653 (94,6)
mfeat_zernike_5	0,322 (72,4)	0,013 (90,0)	**0,721** (94,1)	0,213 (90,1)	0,535 (93,2)
mfeat_zernike_6	0,413 (79,7)	0,068 (90,2)	**0,551** (86,3)	-0,036 (86,4)	0,321 (87,4)
mfeat_zernike_7	0,796 (96,5)	0,013 (90,0)	**0,925** (98,7)	0,213 (90,1)	0,647 (94,6)
mfeat_zernike_8	0,548 (87,3)	0 (90,0)	**0,908** (98,4)	0 (90,0)	0,598 (93,9)
mfeat_zernike_9	0,455 (83,5)	0,026 (90,1)	**0,578** (87,5)	-0,045 (86,6)	0,337 (87,6)
mfeat_morph_0	0,698 (91,6)	0,136 (90,4)	0,682 (91,6)	**0,765** (94,5)	0,764 (94,5)
mfeat_morph_1	0,304 (56,5)	0 (90,0)	0,345 (65,2)	0,375 (82,2)	**0,395** (71,3)
mfeat_morph_2	0,291 (54,0)	0 (90,0)	0,400 (71,9)	**0,457** (81,2)	0,407 (72,9)
mfeat_morph_3	**0,335** (63,5)	0,030 (90,1)	0,326 (63,0)	0,298 (71,5)	0,328 (63,2)
mfeat_morph_4	0,294 (56,8)	0 (90,0)	0,432 (75,4)	**0,443** (87,2)	0,430 (75,3)
mfeat_morph_5	0,378 (67,4)	0,013 (90,0)	**0,468** (78,6)	0,388 (86,5)	**0,468** (78,7)
mfeat_morph_6	**0,637** (88,7)	0,057 (90,1)	0,397 (71,7)	0,398 (75,6)	0,416 (74,0)
mfeat_morph_7	0,398 (70,0)	0,026 (90,1)	0,524 (82,8)	0,505 (88,0)	**0,540** (84,0)
mfeat_morph_8	**0,943** (98,9)	0,013 (90,0)	0,682 (91,6)	0,666 (91,7)	0,645 (89,9)
mfeat_morph_9	**0,456** (76,7)	0,013 (90,0)	0,389 (70,9)	0,395 (74,1)	0,398 (71,9)
glass_1	0,403 (66,2)	**0,896** (95,4)	0,465 (67,0)	0,484 (77,7)	0,509 (77,8)
glass_2	0,229 (56,5)	**0,880** (94,4)	0,212 (49,7)	0,322 (64,8)	0,365 (65,5)
glass_3	0,064 (69,0)	**0,908** (98,6)	0,179 (73,7)	0,145 (92,1)	0,091 (92,6)
glass_5	0,498 (90,9)	0,465 (96,6)	**0,964** (96,1)	0,307 (94,1)	0,823 (94,4)
glass_7	0,813 (95,0)	0,703 (95,4)	0,308 (67,2)	**0,877** (96,4)	0,749 (93,1)

4.4.1.3 Étude de rangs et comparaison statistique des classifieurs

Dans cette section, nous étudions statistiquement les performances des 5 classifieurs de notre comparatif afin de voir leur positionnement l'un par rapport à l'autre à l'aide du test statistique de

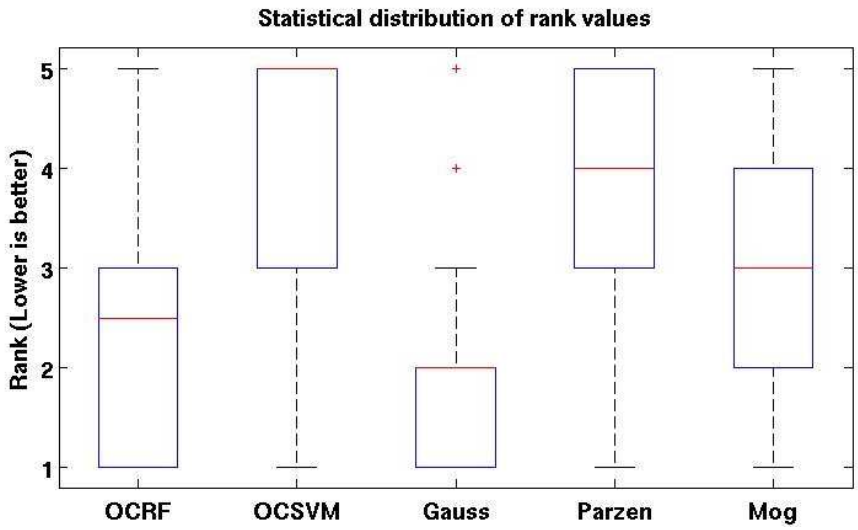


FIGURE 4.9 – Statistiques sur la valeur des rangs pour les 78 bases de données avec la distribution des valeurs selon une boîte à moustache ; la ligne rouge indique la valeur médiane qui dans les cas de Gauss et OCSVM est confondue avec le 3ème quartile soit 75% des valeurs de rangs.

Friedman présenté à la section 4.4.1.1.2. Ce test est basé sur la mesure du rang de chaque classifieur sur chacune des bases évaluées. Les valeurs de rang vont de 1 à 5, le nombre de classifieurs de notre comparaison. Le rang 1 est associé au classifieur ayant les meilleures performances (meilleures valeurs de MCC) tandis que la valeur 5 est associée au classifieur ayant la plus faible valeur de MCC pour une base donnée. Le rang moyen de chaque classifieur et la variance associée sur l'ensemble des 78 bases sont présentés dans le Tableau 4.7. Nous observons que la meilleure méthode vis-à-vis de l'analyse du rang est Gauss avec la valeur de rang moyen la plus basse (R=1.90) suivi par les OCRF (R=2.43) ; MoG est positionné après les OCRF avec une valeur de rang moyen de 2.79 ; OCSVM et Parzen sont alors les méthodes les moins performantes vis-à-vis du rang moyen avec des valeurs au-dessus de 3.

Dans la Figure 4.9, nous pouvons remarquer que 75% des rangs des OCRF est en dessous de 3, i.e. les OCRF sont dans le "Top 3" des meilleures méthodes pour 75% des 78 bases de données (i.e. 52 bases sur 78). Nous pouvons observer également que le rang de Gauss ne dépasse pas 3, tout en étant premier ou second sur 75% des bases évaluées confirmant ainsi ses bonnes performances générales alors que les méthodes OCSVM et Parzen sont parmi les moins performantes pour 75% des bases avec un rang entre 3 et 5. La figure 4.10 permet de visualiser où se situent les meilleures méthodes en fonction du nombre de bases évaluées. Ainsi, nous voyons que les OCRF font partie du "Top 1" et donc des meilleures méthodes pour 30% des bases de données, Gauss étant meilleur pour 50% des bases.

TABLE 4.7 – Valeur moyenne et variance des rangs pour tous les classifieurs sur les 78 bases de données.

	OCRF	OCSVM	Gauss	Parzen	Mog
Averaged rank	2.43 ± 1.16	4.04 ± 1.28	1.90 ± 1.15	3.83 ± 1.12	2.79 ± 1.08

Nous allons maintenant appliquer le test de Friedman afin d'affiner ce positionnement. Nous avons évalué N=78 bases avec k=5 classifieurs. La valeur de la statistique de Friedman indique la différence entre le rang moyen du classifieur actuel et le rang moyen $R = \frac{1+2+3+4+5}{5} = 3$. Ainsi, nous avons :

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right) = 104.48$$

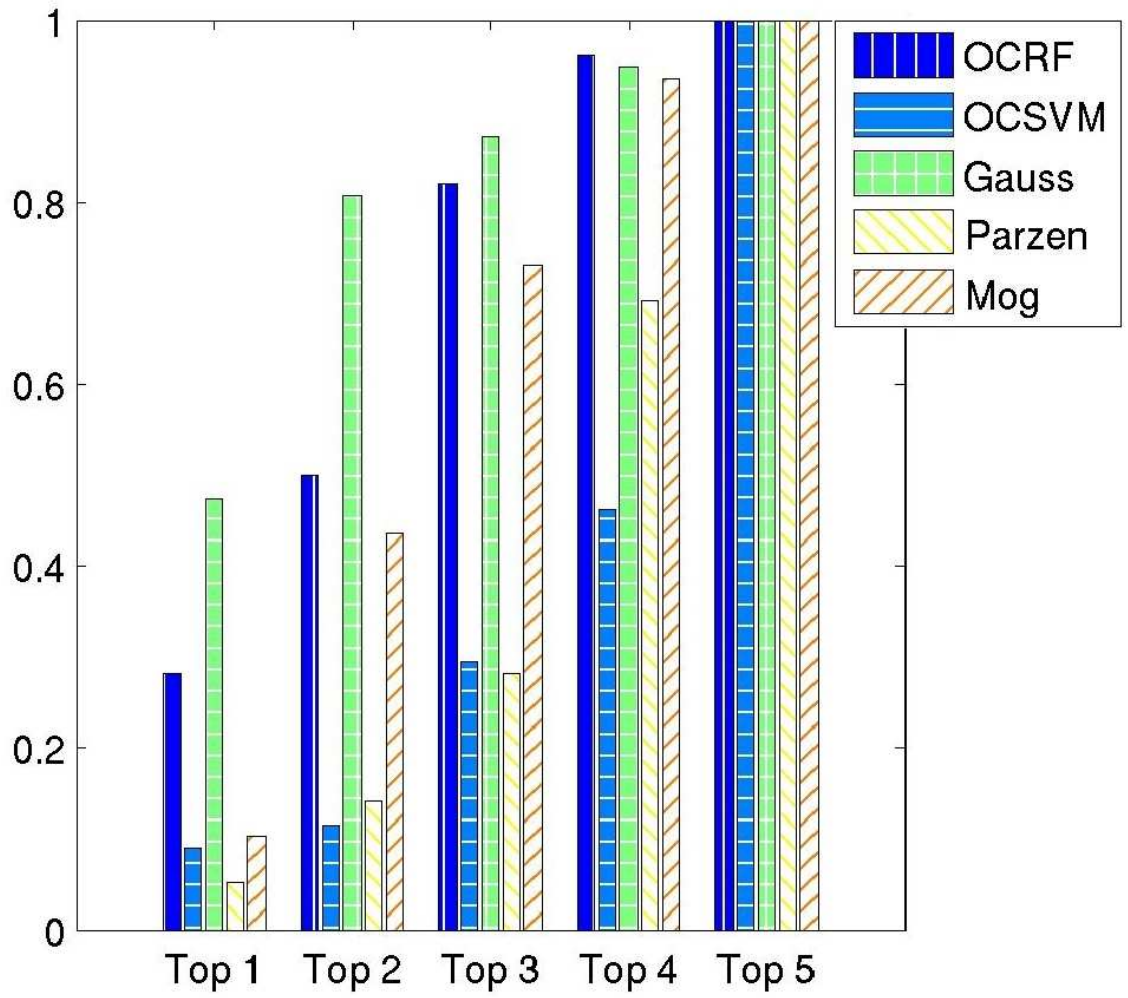


FIGURE 4.10 – Statistiques sur la valeur des rangs pour les 78 datasets : fraction de bases de données associée aux rangs cumulés pour chaque classifieur

puis en calculant la statistique de Iman et Davenport associée :

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} = 38.768$$

A partir des valeurs tabulées de la distribution de Fisher avec $k-1=4$ et $(k-1)(N-1)=308$ degrés de liberté et avec un risque de $\alpha_F = 0.05$, nous trouvons $F(4,308) \approx 2.37 < 38.768$. Ainsi, l'hypothèse nulle est rejetée, les rangs évalués sont significativement différents. Nous utilisons ensuite le test post-hoc de Nemenyi qui permet de détailler cette différence entre les classifieurs en fournissant la différence critique :

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} = 0.690$$

avec un risque de 5%, $q_\alpha = 2.728$. Nous pouvons ainsi conclure que les OCRF sont significativement plus performantes que l'estimateur de Parzen ($R_{Parzen} - R_{OCRF} = 3.83 - 2.43 > CD$) et OCSVM ($R_{OCSVM} - R_{OCRF} = 4.04 - 2.43 > CD$) pour le paramétrage standard choisi pour ces méthodes. On ne peut rien dire en revanche en ce qui concerne OCRF *vs* Gauss, ni OCRF *vs* MoG car la différence de rang est inférieure à la valeur critique CD du test de Nemenyi. Davantage de données de rangs sont alors nécessaires pour pouvoir comparer ces classifieurs. En effectuant les mêmes calculs, on trouve que l'estimateur gaussien est plus performant que les 3 classifieurs MoG, Parzen et OCSVM ; MoG est significativement plus performant que les deux classifieurs OCSVM et Parzen. On ne peut cependant rien conclure entre les classifieurs Parzen et OCSVM. Nous synthétisons dans le Tableau 4.8 les résultats que nous venons de commenter. On peut donc en conclure que dans le cadre de notre expérimentation, le classifieur Gauss s'est révélé le plus performant, suivi des OCRF puis MoG, Parzen et OCSVM.

TABLE 4.8 – Comparaison statistique des performances des méthodes. Le symbole '+' indique que la méthode dans la ligne correspondante est statistiquement plus performante que la méthode dans la colonne correspondante. Un symbole '-' indique au contraire que la méthode est moins performante significativement que la méthode en colonne. Le symbole '0' indique quant à lui que rien ne peut être conclu entre les deux méthodes en ligne et colonne.

	Gauss	OCRF	MoG	Parzen	OCSVM
Gauss		0	+	+	+
OCRF	0		0	+	+
MoG	-	0		+	+
Parzen	-	-	-		0
OCSVM	-	-	-	0	

Afin de mieux comprendre les raisons des bonnes performances de l'estimateur gaussien, nous avons vérifié l'hypothèse de multi-normalité des bases testées à l'aide du test standard de Mardia d'aplatissement (kurtosis) et d'asymétrie multi-variés [Mardia, 1974]. Ce test est considéré comme l'une des meilleures approches pour évaluer à quel point des données multi-variées dévient de l'hypothèse de multi-normalité [Von Eye and Bogat, 2004]. Les résultats se sont cependant révélés négatifs. Ces résultats montrent que l'hypothèse de multi-normalité est rejetée pour 58 bases et acceptée pour seulement 3 bases. Pour 17 bases, les calculs du test n'ont pas pu aboutir en raison de matrices de variance-covariance singulières. Ainsi les bons résultats de Gauss ne peuvent être expliqués par la normalité des données d'apprentissage. Des travaux supplémentaires sont donc nécessaires pour mieux comprendre le bon fonctionnement de l'estimateur sur ces configurations. Le détail du test et des résultats mentionnés figure à l'Annexe 4.5.

Nous avons voulu pousser plus loin l'analyse du comportement des OCRF en fonction de la dimension de l'espace de description et comparer leurs performances aux mêmes approches de

l'état de l'art. Cette étude fait l'objet de la section suivante.

4.4.1.4 Étude de la robustesse des OCRF par rapport à la dimension

L'un des avantages des OCRF est leur caractère générique et leur capacité à traiter des problèmes de grande dimension. Elles héritent naturellement des mécanismes des forêts aléatoires qui ont la propriété de sélectionner en interne les variables importantes du problème [Breiman, 2001, Biau, 2010].

Nous évaluons donc la robustesse des OCRF par rapport à la dimension du problème. Nous avons montré lors des expérimentations sur les bases artificielles le bon comportement des OCRF. Cependant, nous avons montré les difficultés à générer des problèmes en grande dimension : nous avons vu qu'au delà de 50 dimensions, les problèmes générés à l'aide de distributions standards devenaient trop "faciles" pour les OCRF. Ceci peut s'expliquer par le phénomène d'espace vide et la concentration de la mesure en grande dimension [Verleysen et al., 2003, Donoho, 2000]. Ces phénomènes impactent la géométrie des distributions en les rendant possiblement singulières générant soit des problèmes faciles à traiter lorsque les distributions des deux classes sont séparées, soit des problèmes très difficiles lorsqu'il n'est plus possible de distinguer les deux distributions. C'est le cas par exemple de deux distributions gaussiennes de mêmes moyennes et de variances distinctes. Le problème est difficile en petite dimension car les deux distributions se recouvrent. En revanche, en grande dimension, l'intérieur de ces distributions se vide et les données se concentrent dans un mince anneau circulaire dont le rayon dépend de la dimension. Ce phénomène de concentration peut s'expliquer par l'analyse de la distribution des distances euclidiennes des données normales. Ces distances suivent en effet une loi du χ^2 dont la moyenne est la dimension de l'espace. La grande majorité des problèmes réels en très grande dimension concernent des problématiques de sélection de variables comme par exemple l'analyse de gène ou la catégorisation de texte où l'hypothèse principale est que le nombre de variables informatives (dimension intrinsèque) est très petit par rapport à la dimension initiale de l'espace de description [Jimenez and Landgrebe, 1998, Guyon et al., 2004, Tax and Duin, 2005]. Il est ainsi difficile de trouver des problèmes réels en grande dimension où toutes les variables sont informatives.

Nous avons donc considéré la base MFeat pour laquelle plusieurs jeux de caractéristiques ont été extraits formant plusieurs espaces distincts (MFeat-factors, MFeat-kahrunen, MFeat-Zernike, MFeat-Morphological) mais construits à partir des mêmes images. Nous choisissons alors de concaténer ces espaces afin d'obtenir un espace de dimension 333, raisonnablement plus grand que nous appelons MFeat-FKZM reprenant les initiales des caractéristiques concaténées. Nous avons suivi le même protocole que précédemment avec l'ajout d'une procédure de sélection aléatoire d'un sous-espace de dimension croissante afin d'étudier l'influence de la dimension sur les performances des OCRF. Nous avons donc 10 problèmes one-class (une classe est considérée target et les autres forment les outliers) et pour chacun de ces problèmes, nous faisons évoluer la dimension de l'espace de description $m \in \{2, 3, 5, 10, 15, 20, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 333\}$. Pour chaque dimension m choisie, nous effectuons 10 tirages afin de calculer une valeur moyenne et une variance des performances en terme de MCC.

Les résultats obtenus pour chacune des 10 bases et pour chaque classifieur sont résumés sur la Figure 4.12 avec des courbes de valeurs moyennes de MCC en fonction de la dimension. Sur ces figures, chaque classifieur est représenté par une courbe pour chacune des 10 bases MFeat-FKZM formées. Nous détaillons les résultats pour la base du chiffre 1 appelée ici "Digit 1" dans le Tableau 4.9 en termes de valeur moyenne et variance du MCC. Pour des raisons de clarté, seule la base Digit 1 est présentée dans le détail avec la Figure 4.11 associée. Comme nous pouvons le voir sur les graphiques de la Figure 4.12, des résultats similaires ont été obtenus. Les tableaux des 9 autres bases Digits figurent à l'Annexe 4.5.

Nous observons sur la Figure 4.11 et la Figure 4.12 un comportement stable des OCRF avec la dimension, contrairement aux 4 autres méthodes qui ne parviennent pas à maintenir leurs performances obtenues pour des dimensions plus petites. On peut noter que ces résultats de stabilité

pour les OCRF sont renforcés par de faibles valeurs d'écart-type pour toutes les dimensions au delà de $m = 50$, i.e. plus petite que 5%, contrairement par exemple à Gauss où pour les mêmes dimensions l'écart-type peut avoisiner 14%. Il est important de noter, comme lors des expérimentations précédentes que certains classifieurs comme OCSVM ont des valeurs MCC nulles en prédisant uniquement une seule des deux classes. Ainsi, même si Gauss est quelquefois meilleur que OCRF sur certains problèmes, nous voyons avec ces expériences supplémentaires que le choix de ce classifieur peut être au prix d'une plus grande instabilité pour des dimensions plus grandes.

Ces expériences supplémentaires donnent de meilleures indications sur le comportement des OCRF en grande dimension. Alors que les méthodes de l'état de l'art affichent une certaine instabilité lorsque la dimension augmente, nous montrons ainsi que les OCRF demeurent stables et performantes.

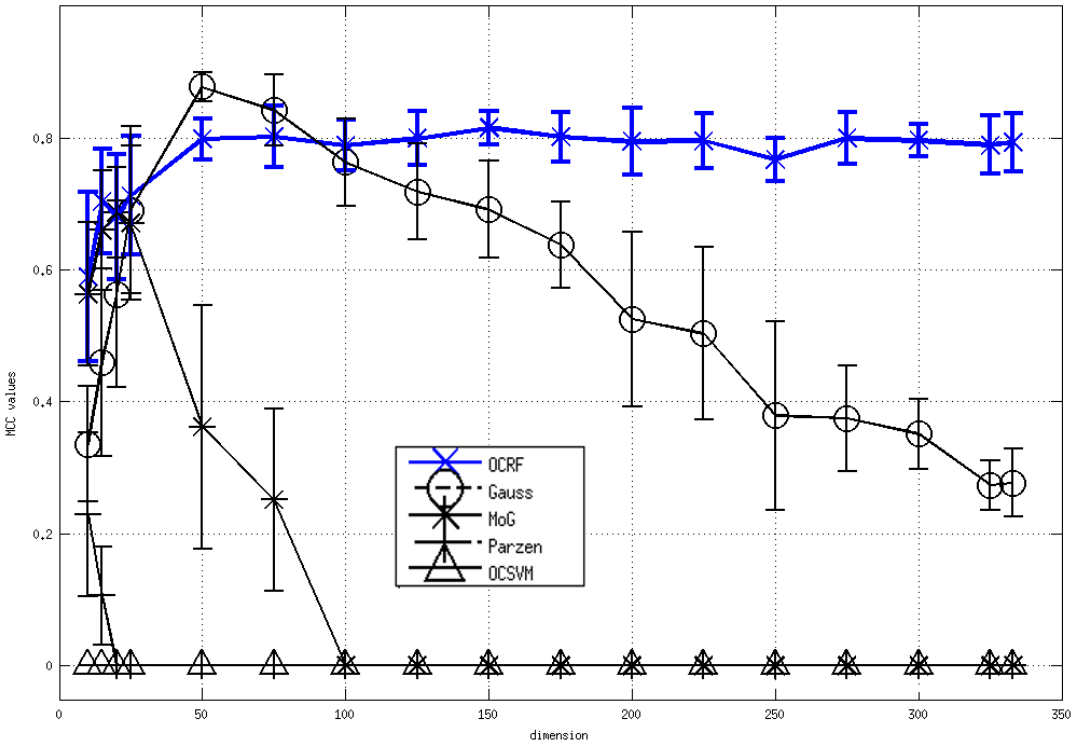
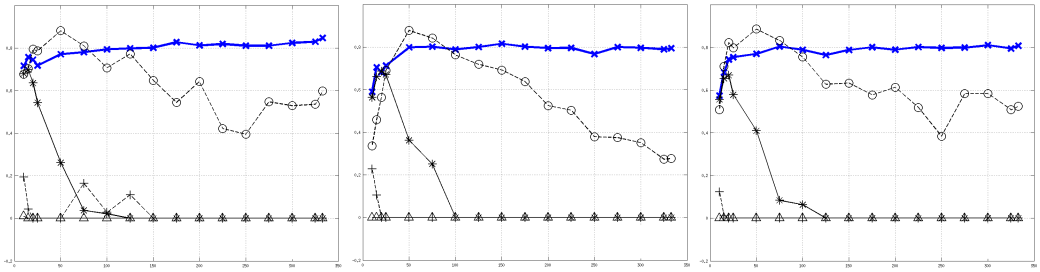


FIGURE 4.11 – Valeurs MCC en fonction de la dimension pour les classifieurs OCRF, OCSVM, Gauss, Parzen et MoG, pour la base *Digit 1*.

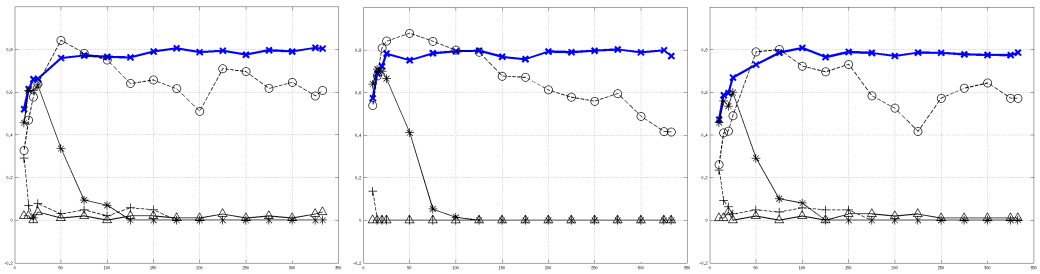
Conclusions sur l'étude des bases publiques

Nous avons montré que les OCRF obtenaient des performances stables pour une grande majorité de bases publiques et se positionnaient parmi les meilleurs classifieurs sur les 78 bases évaluées, en étant compétitives avec l'estimateur gaussien et souvent meilleures. Avec des paramètres désormais standards, nous avons montré que les OCRF sont performantes, génériques et particulièrement robustes à la dimension par rapport aux autres approches évaluées. Nous avons proposé une large évaluation des bases couramment citées dans la littérature en les abordant sous l'angle one-class et avons montré l'apport significatif des OCRF pour le traitement de ces problématiques variées. Plus généralement, nous avons montré que le cadre one-class que nous avons proposé, basé sur les méthodes d'ensemble et mis en œuvre avec les OCRF, tirant partie des mécanismes de combinaison et de randomisation de ces méthodes d'ensemble apporte une réponse favorable à la problématique one-class.

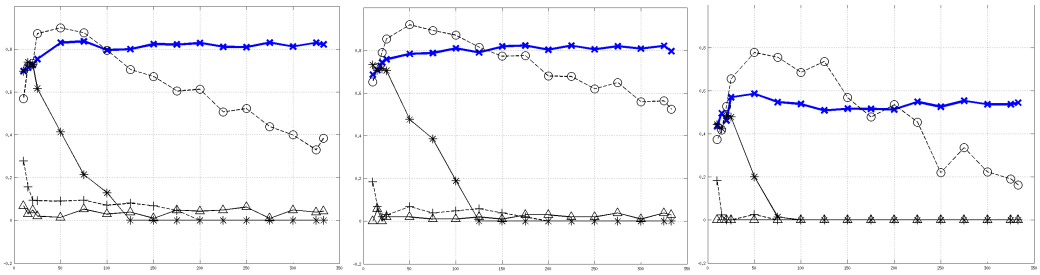
Nous proposons dans l'étude suivante d'appliquer les OCRF sur les bases alvéoscopiques afin d'analyser le comportement de l'approche one-class sur la problématique médicale que nous avons



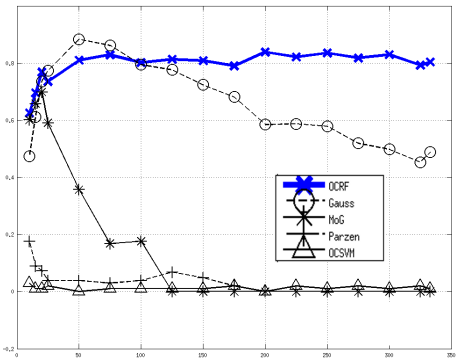
(a) Digit 0 (b) Digit 1 (c) Digit 2



(d) Digit 3 (e) Digit 4 (f) Digit 5



(g) Digit 6 (h) Digit 7 (i) Digit 8



(j) Digit 9

FIGURE 4.12 – Valeurs MCC par rapport à la dimension pour les classifieurs OCRF (”croix”), OCSVM (”triangle”), Gauss (”cercle”), Parzen (”plus”) et MoG (“étoile“) sur la base MFeat-FKZM pour tous les digits.

TABLE 4.9 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 1

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.15 (± 0.09)	**0.53** (± 0.16)	0.10 (± 0.06)	0.13 (± 0.05)	0.12 (± 0.05)
3	0.20 (± 0.13)	**0.39** (± 0.31)	0.11 (± 0.09)	0.17 (± 0.12)	0.15 (± 0.10)
5	**0.39** (± 0.09)	0.00 (± 0.00)	0.19 (± 0.07)	0.30 (± 0.10)	0.26 (± 0.06)
10	**0.59** (± 0.13)	0.00 (± 0.00)	0.34 (± 0.09)	0.23 (± 0.12)	0.56 (± 0.11)
15	**0.70** (± 0.08)	0.00 (± 0.00)	0.46 (± 0.14)	0.11 (± 0.07)	0.66 (± 0.09)
20	0.68 (± 0.09)	0.00 (± 0.00)	0.56 (± 0.14)	0.00 (± 0.00)	**0.69** (± 0.07)
25	**0.71** (± 0.09)	0.00 (± 0.00)	0.69 (± 0.13)	0.00 (± 0.00)	0.67 (± 0.12)
50	0.80 (± 0.03)	0.00 (± 0.00)	**0.88** (± 0.02)	0.00 (± 0.00)	0.36 (± 0.18)
75	0.80 (± 0.05)	0.00 (± 0.00)	**0.84** (± 0.05)	0.00 (± 0.00)	0.25 (± 0.14)
100	**0.79** (± 0.04)	0.00 (± 0.00)	0.76 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
125	**0.80** (± 0.04)	0.00 (± 0.00)	0.72 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
150	**0.82** (± 0.03)	0.00 (± 0.00)	0.69 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
175	**0.80** (± 0.04)	0.00 (± 0.00)	0.64 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.79** (± 0.05)	0.00 (± 0.00)	0.53 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.80** (± 0.04)	0.00 (± 0.00)	0.50 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.77** (± 0.03)	0.00 (± 0.00)	0.38 (± 0.14)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.80** (± 0.04)	0.00 (± 0.00)	0.37 (± 0.08)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.80** (± 0.02)	0.00 (± 0.00)	0.35 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.79** (± 0.04)	0.00 (± 0.00)	0.27 (± 0.04)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.79** (± 0.04)	0.00 (± 0.00)	0.28 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)

traitée dans les deux premiers chapitres.

4.4.2 Expérimentations sur les images alvéoscopiques

Dans cette section, nous expérimentons les OCRF sur les bases MCFF. Nous détaillons ci-dessous le protocole pour les deux bases fumeur et non fumeur et présentons séparément les différents résultats obtenus.

Protocole expérimental

Bases de données Nous étudions les bases d'images MCFF (fumeur F et non-fumeur NF) décrites par le jeu de caractéristiques double résolution $LBP_{(8,1)/(16,2)}$, formant un espace de description de 28 dimensions, sélectionné à l'issue du second chapitre section 2.2. Nous avons choisi de travailler directement sur les images plutôt que sur des fenêtres extraites afin de valider l'approche des OCRF sans effectuer de procédure supplémentaire pour l'évaluation (e.g. le vote à la majorité sur les fenêtres extraites d'une image en phase de test). Nous rappelons dans le Tableau 4.10 les effectifs des deux bases alvéoscopiques.

TABLE 4.10 – Effectifs des images FCFM pour les deux groupes fumeur et non-fumeur

Groupes / Classes	Non-fumeur	Fumeur
Sain	31	60
Pathologique	102	33
Total	133	93

Paramétrisation et méthodes d'évaluation Nous reprenons le paramétrage des OCRF de l'expérimentation précédente sur les bases publiques. La procédure d'évaluation est ici modifiée en raison du faible effectif des données MCFF. L'évaluation sera faite selon une procédure de

re-échantillonnage (“resampling“) avec un découpage 50/50 (50% des données serviront pour la base d'apprentissage en excluant les données de la classe des patients malades et les 50% restants serviront pour la phase de test incluant les deux classes). La procédure est répétée 100 fois. Les résultats des performances seront alors ensuite moyennés sur les 100 jeux.

Résultats et analyse

Les résultats des expérimentations sur les bases MCFF avec les OCRF et les classifieurs one-class standards sont présentés dans le Tableau 4.11. Nous pouvons remarquer les bonnes performances des OCRF dans la reconnaissance des outliers et une bonne performance générale sur la base non-fumeur. Nous voyons que les OCRF obtiennent les meilleures performances sur la base non-fumeur (MCC=0.71) lorsque les autres méthodes échouent complètement dans l'identification d'une des classes, à l'exception de OCSVM qui identifie une partie des targets (27%). Cependant, la méthode OCRF éprouve des difficultés à correctement identifier la classe target dans le cas fumeur (48%) alors que Gauss obtient les meilleures performances avec 90% de taux de reconnaissance sur cette classe ; on observe par ailleurs que la classe target est la plus difficile à classer pour tous les classifieurs, à l'exception de Gauss. Afin de mieux comprendre ces faibles résultats pour la base fumeur, des résultats supplémentaires ont été obtenus en générant moins de données outliers, i.e. avec $\beta = 1$, le nombre de données générées en apprentissage est $N_{outlier/app} = 60$ au lieu de 600. Les résultats obtenus pour la base fumeur sont présentés dans le Tableau 4.12. Nous avons alors constaté que les résultats pour les données targets pour la base fumeur ont été améliorés, au détriment d'une faible baisse de performances pour la reconnaissance des données outliers. En effet, MCC passe de 0.42 à 0.56, Accr de 63% à 80% et le taux de reconnaissance des données targets passe de 48% à 86%, avec une baisse de 8% sur le taux de reconnaissance des données outliers. Ainsi, il semble que lorsque $\beta = 10$, les données outliers générées recouvrent de façon trop importante les données targets. Nous concluons que le nombre de données outliers à générer, quoique la valeur obtenue avec le coefficient $\beta = 10$ soit un bon compromis en général, peut être ajusté pour mieux s'adapter aux particularités de certaines bases à faibles effectifs notamment.

Les résultats de cette expérience montrent les bonnes performances générales de notre approche sur les images alvéoscopiques, notamment pour l'identification des cas pathologiques. Ces résultats sont perfectibles avec une configuration plus fine des paramètres α et β et font l'objet des études en cours et des perspectives de ces travaux de thèse.

TABLE 4.11 – Résultats des OCRF sur les bases d'images alvéoscopiques et comparaison avec les classifieurs de la littérature ; les performances sont indiquées en termes de MCC, taux de reconnaissance globale (Accr), taux sur les targets (T) et taux sur les outliers (O) ; les meilleurs résultats sont mis en gras.

		OCRF	OCSVM	Gauss	Parzen	MoG
Alveo_S	MCC	0.42±0.03	0.50±0.01	**0.76±0.00**	0.31±0.00	0.36±0.04
	Accr	0.63±0.02	0.67±0.01	0.89±0.00	0.50±0.00	0.54±0.04
	T	0.48	0.49	0.90	0.23	0.30
	O	0.93	1.00	0.87	1.00	1.00
Alveo_NS	MCC	**0.71±0.03**	0.51±0.07	0.00±0.00	0.00±0.00	0.00±0.00
	Accr	0.88±0.02	0.82±0.03	0.76±0.00	0.76±0.00	0.76±0.00
	T	0.81	0.27	0.00	0.00	0.00
	O	0.91	1.00	1.00	1.00	1.00

4.4.3 Discussion

Nous observons que les OCRF sont compétitives par rapport aux méthodes de l'état de l'art voire même meilleures sur certains problèmes. Les OCRF sont donc une approche générique et

TABLE 4.12 – Comparaison des résultats obtenus par les OCRF sur la base d'images alvéoscopiques fumeur pour les deux valeurs du paramètre β ; les performances sont indiquées en termes de MCC, taux de reconnaissance globale (Accr), taux sur les targets (T) et taux sur les outliers (O).

		$\beta = 1$	$\beta = 10$ (standard)
Alveo_S	MCC	0.56±0.09	0.42±0.03
	Accr	0.80±0.04	0.63±0.02
	T	0.86	0.48
	O	0.85	0.93

robuste pour la plupart des problèmes évalués. Ainsi, nous avons montré que l'utilisation des mécanismes de combinaison et de randomisation des méthodes d'ensemble et particulièrement des forêts aléatoires permettaient de répondre favorablement aux problèmes posés aux approches one-class de l'état de l'art et particulièrement aux approches discriminantes, pour la synthèse des données outliers notamment en grande dimension, à la fois en terme de quantité et de distribution. Concernant la paramétrisation des OCRF, nous observons que les résultats ont été obtenus avec des valeurs prises par défaut et fixes pour toutes les bases étudiées. Cependant, nous avons également constaté les contraintes structurelles liées aux valeurs des paramètres α et β. Typiquement, lorsque α est trop petit, le modèle des OCRF ne généralise pas ; lorsque α est grand, les OCRF peuvent généraliser à condition de générer un nombre suffisant d'outliers. Il est ainsi possible d'obtenir des résultats meilleurs avec une paramétrisation adaptée à l'application. Le nombre de données outliers à générer semble dépendre des propriétés statistiques de la base étudiée indiquant une amélioration possible soit dans le mode de génération des outliers, soit dans le critère de partitionnement en chaque nœud de l'arbre avec par exemple un partitionnement plus fin dans les régions sparses de l'espace.

4.5 Conclusion et perspectives

Nous avons proposé dans ce chapitre une approche one-class discriminante, générique et robuste capable de traiter un grand nombre de problèmes dans la littérature. Les résultats obtenus sont pour la plupart équivalents ou meilleurs que ceux obtenus par des classifieurs couramment cités. Nous avons pu répondre favorablement aux difficultés auxquelles sont confrontées les approches discriminantes en raison notamment de la malédiction de la dimension rendant la synthèse pourtant nécessaire des outliers difficile voire preque impossible dans la pratique. L'approche proposée utilise des arbres de décision au sein de la méthode d'ensemble mais d'autres approches discriminantes peuvent être envisagées comme des réseaux de neurones, des SVM.

Les OCRF permettent la génération de données outliers dans les zones sparses de l'espace de description en utilisant une distribution des outliers basée sur celles des données targets en présence. De plus, en utilisant les principes de combinaison et de randomisation des forêts, les OCRF bénéficient des avantages des méthodes de cette famille, à savoir notamment les bonnes performances, la bonne gestion de la dimension, le fort pouvoir généralisant. Afin de valider l'efficacité des OCRF, nous avons évalué leurs performances à la fois sur des bases synthétiques et sur des bases réelles publiques communément utilisées dans la littérature. Des comparaisons statistiques avec des classifieurs de la littérature ont montré que les OCRF se comportaient favorablement et souvent obtenaient de meilleurs résultats que ces derniers. Le résultat sans doute le plus remarquable est la stabilité de notre méthode. En effet, au travers de différentes applications, différents jeux de données et de caractéristiques, les OCRF offrent des performances favorables.

Plusieurs perspectives se dessinent pour la continuité des travaux sur le one-class. Grâce à la flexibilité du cadre one-class par méthodes d'ensemble proposé et en raison de la diversité des approches par méthodes d'ensemble, plusieurs améliorations peuvent être proposées pour chacune des étapes :

– une meilleure gestion de la construction de la distribution des outliers comme distribution

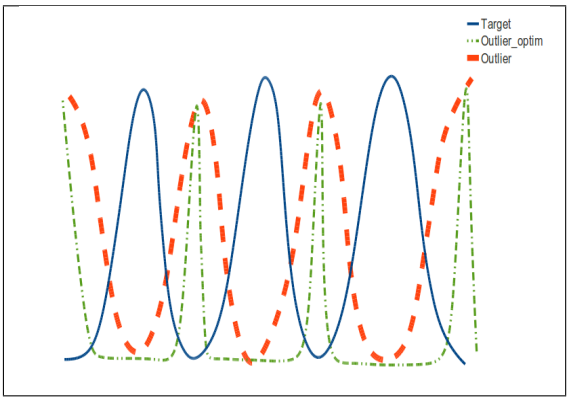


FIGURE 4.13 – Approche de génération des outliers à la frontière des données targets en évitant le recouvrement ; la distribution des targets est représentée en bleu (trait plein), la distribution actuelle des outliers est représentée en rouge (trait pointillé) ; celle proposée vérifiant l'équation 4.1 est représentée en vert (trait fin en pointillé) ;

complémentaire des targets (toujours "distribution-based") peut être envisagée en extrayant davantage d'informations du set d'apprentissage comme par exemple la localisation en amont de l'apprentissage des sous-espaces pertinents pour le RSM à l'instar de ce qui est fait dans [Aggarwal and Yu, 2001] à l'aide d'un algorithme de recherche évolutionniste et d'une mesure de sparsité ;

- une analyse des effets de la dimension sur les performances est à approfondir car la grande dimension engendre notamment le phénomène d'espace vide réduisant par exemple le volume de la boule unité à 0 ou encore agrandissant le volume des coins de l'hypercube ;
- une meilleure gestion des données outliers générées est possible avec l'introduction de termes de pénalisation pour la génération de données outliers au sein des données targets ; l'histogramme complémentaire 1D actuel $H_{outlier} = 1 - H_{target}$ pourrait être transformé en :

$$H_{outlier} = max(0, 1 - \gamma \cdot H_{target}) \tag{4.1}$$

$\gamma > 0$ permettant de contrôler le degré de remplissage des zones où des données targets sont présentes (une illustration est présentée sur la figure 4.13) ; pour cette dernière approche avec un γ très élevé, on peut générer de façon déterministe, i.e. sans aléatoire, exactement un outlier dans la zone sparse dans le cas d'utilisation d'arbres de décision.

- on peut éviter la génération physique des données outliers avec une variante des OCRF sans génération de données outliers, simulant leur présence en chacun des nœuds de l'arbre ; cette approche estime en chaque nœud de l'arbre la distribution des données targets et les points de coupure sont alors choisis dans les régions de faibles densités, ces régions étant échantillonnées aléatoirement ; de l'information a priori prélevée en amont de l'apprentissage peut être injectée au moment de la construction du critère de partitionnement ; la difficulté principale est de définir un critère d'arrêt basé actuellement sur le nombre d'échantillons testés et la valeur de densité retournée ; nous décrivons plus en détail cette approche à l'Annexe 4.5.
- il est possible d'utiliser le principe général des ensembles one-class avec d'autres classifieurs différents de l'arbre de décision comme un réseau de neurones MLP ou un SVM

Ces perspectives sont à plus ou moins long termes et montrent les possibilités nouvelles offertes par l'approche proposée pour traiter des problématiques one-class mais également des problématiques multi-classes et du clustering.

Conclusion générale

Nous avons présenté dans cet exposé les travaux de thèse réalisés dans le cadre de l'aide au diagnostic médical des images issues de l'alvéoscopie. La problématique considérée concerne la classification de ces images alvéoscopiques en identifiant les images issues de patients sains et celles issues de patients atteints d'une pathologie pulmonaire. Elles sont complexes à analyser et il n'existe pas à ce jour de description univoque de ces images. De plus, elles sont très hétérogènes et pour une pathologie donnée, plusieurs caractéristiques visuelles distinctes peuvent être observées. Nous avons proposé, au travers de deux contributions, des méthodes performantes d'aide au diagnostic fournissant au praticien des outils quantitatifs pour l'évaluation de l'état pathologique des images analysées.

Notre première contribution concerne la mise en place d'un système complet de classification des images alvéoscopiques permettant de classer de façon satisfaisante, au vu des experts médicaux, les images de patients sains et pathologiques. Ces images étant difficiles à analyser, nous avons avant tout cherché à mettre en place une approche générique afin de ne pas être dépendant du contenu de l'image. Ce système est basé sur une extraction locale et une caractérisation riche des informations des images et une approche de classification par méthodes d'ensembles d'arbres et un mécanisme de pilotage du rejet permettant de renforcer la fiabilité du système.

Nous avons montré que l'approche de caractérisation locale par LBP est la plus performante des approches évaluées. Elle est notamment plus discriminante que les autres approches d'extraction de textures comme SIFT ou les statistiques des matrices de cooccurrence, en tenant compte des structures et microstructures présentes dans l'image à différents niveaux de résolution spatiale.

Pour l'étape de classification, nous avons adopté une approche performante et générique de la littérature avec les méthodes d'ensembles d'arbres de décision Extremely Randomized Trees (extra-trees), une méthode de forêts aléatoires qui s'est révélée meilleure que le SVM dans nos expérimentations et aussi performante que l'approche standard forest-RI avec des coûts de calculs divisés par deux. Les images issues de patients pathologiques sont cependant difficiles à classer et le système obtient un taux de non-détection des cas pathologiques relativement élevé par rapport à celui des images issues de patients sains. Les modifications que nous avons alors opérées dans l'algorithme des forêts extra-trees, tant en phase d'apprentissage qu'en phase de décision, ont permis de renforcer la fiabilité de la décision du classifieur en mettant en place un mécanisme de pilotage du rejet des images de faible niveau de confiance. Ce mécanisme inclut une approche d'élagage des arbres de la forêt, la modification à la fois du mode de vote interne des arbres que nous avons appelé "vote fréquentiel" permettant de tenir compte des populations présentes en chaque nœud de l'arbre et du mode de vote de la forêt tenant compte de la contribution de chacun des arbres. Ces modifications permettent ainsi de mieux évaluer le degré de consensus entre les arbres de la forêt pour l'attribution d'une classe.

Nous avons ensuite proposé une approche one-class, plus adaptée à notre problématique et que nous avons appelée one-class random-forests (OCRF). Cette seconde contribution nous a permis de répondre favorablement au problème de définition de la classe des données pathologiques en ne considérant pour l'apprentissage de la méthode que les données de la classe de patients sains pour lesquelles nous avons une connaissance certaine. L'approche OCRF est une approche one-class discriminante basée sur les forêts aléatoires et un mécanisme efficace de génération des outliers. Les méthodes proposées dans la littérature sont principalement confrontées au problème des données en grande dimension et aux phénomènes plus complexes comme le phénomène de l'espace vide ou de la concentration de la mesure affectant notamment les méthodes par estimation de densité ou de distance.

En grande dimension, en raison particulièrement de la malédiction de la dimension, le manque de données rend difficile une estimation fiable des paramètres de ces méthodes ou de la distribution des classes en présence et dans le cas particulier des approches discriminantes, rend difficile la synthèse des données outliers. L'approche OCRF est originale car elle répond favorablement à ces problèmes posés en tirant non seulement partie des mécanismes des méthodes d'ensembles mais également, avec les forêts aléatoires en particulier, elle travaille dans des sous-espaces pertinents de l'espace original.

Nous avons proposé un algorithme efficace de génération des outliers tirant partie des mécanismes de randomisation des méthodes d'ensembles, à savoir le bagging et le random subspace method (RSM). Il a fallu pour cela répondre à plusieurs questions concernant la synthèse des données outliers : où générer les outliers, à quel moment de l'induction du système les générer, avec quelle distribution, en quelle quantité. Le nombre de données outliers à générer a été fixé comme paramètre et nous avons observé qu'une valeur proche du nombre de données targets présents pouvait être définie comme standard. Tenant compte du problème de la malédiction de la dimension, nous avons tiré parti à la fois du bagging pour le re-échantillonnage des données d'apprentissage et du RSM pour la projection de ces mêmes données dans des sous-espaces de dimensions beaucoup plus petites que celle de l'espace d'origine et choisies aléatoirement. Nous avons ainsi généré utilement les outliers dans les sets boostraps projetés dans les sous-espaces RSM, juste avant l'induction de chacun des arbres de la forêt.

Nous avons proposé une approche d'évaluation exhaustive afin de valider notre approche, en utilisant des bases pseudo one-class obtenues à partir de bases de l'UCI (bases utilisées pour la classification standard) en prenant tour à tour chacune des classes en présence comme étant la classe target, les autres classes étant regroupées dans celle des outliers. Les OCRF se sont révélées génériques, performantes, compétitives par rapport aux méthodes de l'état de l'art et même meilleures sur un grand nombre de bases de données publiques, avec une paramétrisation encore non optimale. La méthode est notamment robuste à la dimension. L'application de cette approche aux images alvéoscopiques a montré qu'en l'absence des images issues de patients pathologiques, en ne se basant donc que sur les images de patients sains, les OCRF sont capables de bien identifier les images pathologiques, en étant compétitive avec notre premier système notamment dans le cas des images de patients non-fumeurs. Cette nouvelle approche ouvre ainsi la possibilité d'appréhender la forte diversité constatée au sein des images issues de pathologies identifiées mais également celles issues de nouvelles configurations de pathologies nouvelles ou existantes permettant ainsi de s'adapter à de nouvelles données.

La méthode proposée est nouvelle et n'est pas parfaite dans le sens où elle est capable de traiter un grand nombre de problèmes mais pour certains problèmes des baisses de performances sont observées par rapport aux approches de l'état de l'art. Ces baisses de performances, dont les origines ne sont pas encore déterminées avec certitude pour le moment, peuvent être en partie liées au mode de génération des outliers et la paramétrisation non optimale de notre approche. La paramétrisation par défaut que nous avons choisie nous a permis d'avoir un premier aperçu des performances de notre système. Pour certains problèmes, cette paramétrisation standard ne semble pas adaptée, notamment en ce qui concerne le nombre d'outliers à générer dont la valeur optimale serait moins grande que celle que nous avons proposée. De même, la valeur par défaut choisie permettant de contrôler l'extension du domaine de génération des outliers pourrait être réajustée pour certains problèmes.

La méthode actuelle de génération des outliers induit naturellement un léger recouvrement des données targets, aux frontières de ces dernières. L'impact et le degré de ce recouvrement n'a pas encore été mesuré. De plus, les problèmes artificiels standards que nous avons générés ne présentaient pas pour la plupart de difficultés pour les OCRF, notamment en grande dimension. D'autres distributions sont en cours d'investigation afin d'analyser davantage de propriétés géométriques et spatiales. Ainsi, des travaux restent encore à approfondir afin de mieux appréhender notre approche.

Á la suite de ces travaux, plusieurs perspectives ont été considérées, certaines d'entre elles étant déjà mises en œuvre. Un des problèmes sur lesquels nous travaillons est de savoir en quoi certaines bases de données sont difficiles pour les OCRF et plus particulièrement pourquoi certaines bases ont des valeurs optimales de paramètres différentes des valeurs standards que nous avons utilisées. Nous avons vu que le mode de génération des outliers entraînait un léger recouvrement des données targets par les outliers pouvant impacter négativement la classification des données targets en particulier. La dégradation des performances sur les données targets que nous avons observée en fonction du nombre de données outliers générées semble appuyer cette hypothèse. Une solution

envisagée est de pénaliser fortement la génération des outliers dans les zones même faiblement peuplées en données targets afin de ne considérer que les zones essentiellement vides en données targets. De plus, certaines zones vides ne sont pas atteignables par la méthode actuelle parce que nous générons les outliers en nous basant sur les projections 1D de la distribution des targets en amont de l'apprentissage. Il est alors possible de considérer un espace de projection plus important afin d'augmenter les chances de l'algorithme de découvrir ces espaces vides. En circonscrivant ces zones vides, nous envisageons également d'y générer beaucoup moins de données outliers tout en maintenant les performances de la méthode. En effet, les outliers supplémentaires dans cette même zone tombent principalement dans la même feuille de l'arbre induit et ne sont donc pas indispensables. Toute la difficulté de cette approche est de découvrir ces zones vides en les circonscrivant et y générer un nombre nécessaire d'outliers pas trop loin des frontières.

Nous avons opté pour les méthodes d'ensembles en raison des mécanismes internes favorables au traitement de la problématique one-class notamment dans le cas du discriminant. Nous avons choisi les arbres de décision comme classifieurs individuels mais d'autres choix sont possibles avec par exemple le SVM ou le MLP ou encore des méthodes génératives comme les estimateurs de densité en conservant ces mêmes mécanismes qui ont rendu possible l'application des méthodes d'ensembles au one-class. Les choix que nous avons alors effectués pour la génération des outliers ne sont pas uniques et nous pouvons envisager des approches dérivées des OCRF modifiant par exemple le mécanisme de génération des outliers, leur distribution, le moment où ces derniers sont générés ou en renforçant l'aléatoire injecté dans les mécanismes de randomisation à l'exemple des Extra-trees.

Avec cette seconde contribution, il est possible d'envisager le traitement de problématiques proches du one-class comme la classification de données déséquilibrées avec l'incorporation ou non des données disponibles de la seconde classe ou le clustering où les clusters présents sont naturellement découverts lors du partitionnement automatique de l'espace de description. Particulièrement, en étant une approche one-class capable de travailler en grande dimension et en tirant partie des propriétés générales des forêts aléatoires permettant de mesurer l'importance des variables, il est possible avec les OCRF de faire de la sélection de variables pour les problématiques one-class mais aussi celles du clustering et des données déséquilibrées en grande dimension.

En ce qui concerne les images alvéoscopiques, nous avons notamment travaillé avec les experts de l'équipe de pneumologie du CHU de Rouen sur une dizaine d'items qui sont des caractéristiques cliniques visuelles extraites manuellement d'images de patients pathologiques. Ces caractéristiques, une fois traduites numériquement et extraites de façon automatisée, pourront être ajoutées à la caractérisation générique initiale que nous avons proposée afin de renforcer la description fournie aux OCRF pour ce type d'images en particulier et tenter ainsi d'améliorer les performances de la méthode pour cette problématique. Nous disposons également d'une base de données vidéos d'alvéoscopie nous permettant notamment d'étendre considérablement la quantité d'informations disponibles pour une étude plus poussée des systèmes que nous avons mis en place et les nouvelles approches que nous envisageons, en incorporant par exemple de l'information temporelle dans la classification one-class.

Publications de l’auteur

(1) C. Desir, S. Bernard, C. Petitjean, and L. Heutte, Oneclass Random Forest, *Pattern Recognition*, volume 46, issue 12, year 2013, pp. 3490 - 3506

(2) C. Desir, C. Petitjean, L. Heutte, M. Salaun, and L. Thiberville. Classification of endomicroscopic images of the lung based on random subwindows and extra-trees. *Biomedical Engineering, IEEE Transactions on*, 59(9) :2677–2683, 2012.

(3) C. Desir, C. Petitjean, L. Heutte, M. Salaun, and L. Thiberville. An svm distal lung image classification using texture descriptors. *Computerized Medical Imaging and Graphics*, Volume 36, pp. 264-270, 2012.

(4) C. Desir, S. Bernard, C. Petitjean, and L. Heutte. A new random forest method for one-class classification. *IAPR International Workshop on Statistical Techniques in Pattern Recognition*, SPR 2012, Hiroshima, Japan, LNCS 7626, pp 282–290, 2012.

(5) C. Desir, S. Bernard, C. Petitjean, and L. Heutte. A random forest based approach for one-class classification in medical imaging. 3rd *MICCAI International Workshop on Machine Learning in Medical Imaging*, MLMI 2012 Nice, France, LNCS 57588, pp 250-257, 2012.

(6) D. Hebert, C. Desir, C. Petitjean, L. Heutte, and L. Thiberville. Detection of pathological condition in distal lung images. *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1603–1606. IEEE, 2012.

(7) C. Desir, C. Petitjean, L. Heutte, and L. Thiberville. Using a priori knowledge to classify in vivo images of the lung. *In International Conference on Intelligent Computing*, LNAI 6216,Changsha, China, pages p. 207–212, Aug. 2010.

(8) A. Saint-Requier, B. Lelandais, C. Petitjean, C. Desir, and L. Heutte. Characterization of endomicroscopic images of the distal lung for computer-aided diagnosis. *In International Conference on Intelligent Computing*, Ulsan, South Korea, volume 5754 of Lecture Notes in Computer Science, pages 994-1003. Springer, 2009.

(9) C. Petitjean, C. Desir, M. Salaun, L. Thiberville, and L. Heutte. Automatic classification of in vivo distal lung images for computer-aided diagnosis. *In Proceedings of Medical Image Understanding and Analysis (MIUA)*, pp 99-103, Kingston, UK, 2009.

(10) B. Lelandais, C. Desir, C. Petitjean, L. Heutte, and L. Thiberville. Classification d'images alveoscopiques par extra-trees. *Reconnaissance des Formes et Intelligence Artificielle*, 2010.

(11) C. Petitjean, C. Desir, L. Thiberville, G. Lettier, L. Heutte, Manual vs. automatic classification of endomicroscopic images of the distal lung, *Proceedings 25th International Congress and Exhibition on Computer Assisted Radiology and Surgery (CARS'2011)*, Berlin, Germany, 2011

Annexes

Annexe au chapitre 2

Classification par KPPV

Nous présentons dans les Tableaux A.13 et A.14 des résultats complémentaires obtenus avec le classifieur KPPV, $K \in [1:10]$, pour les deux bases fumeur et non-fumeur avec la procédure en leave-one-out. On constate que les performances des différents extracteurs considérés sont relativement stables avec K le nombre de voisins considérés.

TABLE A.13 – Résultats du KPPV sur la base d'images alvéoscopiques fumeur; $r_1/r_2 = (8,1)/(16,2)$ désigne la double résolution de l'opérateur LBP, variance ou CS-LBP

K	Cooccurrence	LBP_{r_1/r_2}	$LBP/VAR_{r_1/r_2}$	VAR_{r_1/r_2}	$CS-LBP_{r_1/r_2}$	Ad-Hoc	SIFT
1	0.8817	0.9355	**0.9570**	0.6882	0.8602	0.6989	0.7850
2	0.8387	**0.9462**	0.9355	0.7419	0.8280	0.6774	0.7742
3	0.8817	0.9462	**0.9677**	0.7419	0.8602	0.6882	0.7850
4	0.8495	0.9140	**0.9462**	0.7097	0.8602	0.7097	0.7850
5	0.8710	0.9355	**0.9570**	0.6882	0.8710	0.6129	0.7634
6	0.8387	**0.9570**	0.9355	0.6882	0.8925	0.6882	0.7527
7	0.8602	0.8925	**0.9140**	0.6882	0.8602	0.6022	0.7742
8	0.8495	0.8817	**0.9140**	0.6667	0.8602	0.6774	0.7312
9	0.8817	0.9140	**0.9247**	0.6774	0.8817	0.6237	0.7312
10	0.8710	0.9247	0.9247	0.6882	0.8817	0.7312	0.7419

TABLE A.14 – Résultats du KPPV sur la base alvéoscopique non-fumeur; $r_1/r_2 = (8,1)/(16,2)$ désigne la double résolution de l'opérateur LBP, variance ou CS-LBP

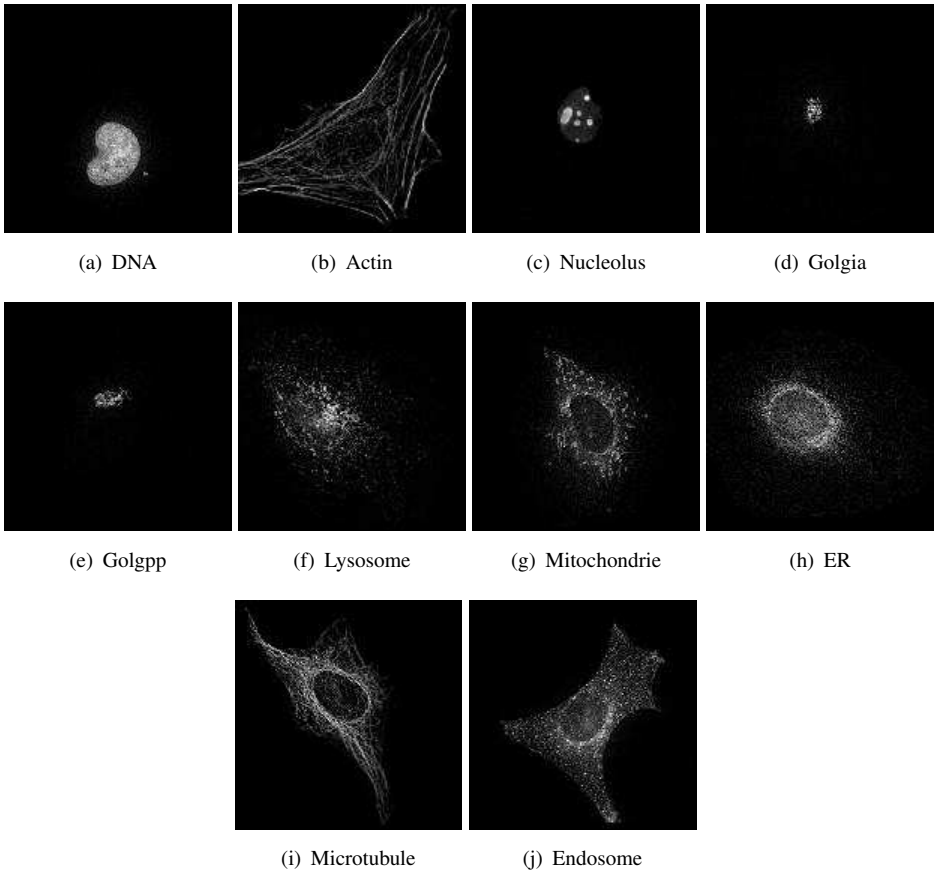
K	Cooccurrence	LBP_{r_1/r_2}	$LBP/VAR_{r_1/r_2}$	VAR_{r_1/r_2}	$CS-LBP_{r_1/r_2}$	Ad-Hoc	SIFT
1	0.7970	**0.9248**	0.8797	0.6541	0.8346	0.6165	0.8346
2	0.7970	**0.9248**	0.9098	0.7820	0.8271	0.7143	0.8571
3	0.8120	0.9023	**0.9173**	0.7368	0.8421	0.6767	0.8722
4	0.7820	**0.9098**	0.8872	0.7594	0.8346	0.7143	0.8647
5	0.7970	**0.9098**	0.8872	0.7368	0.8797	0.6466	0.8797
6	0.8045	**0.9023**	0.8797	0.7293	0.8647	0.7218	0.8571
7	0.7895	**0.9248**	0.8571	0.7368	0.8872	0.6692	0.8571
8	0.7820	**0.9248**	0.8797	0.7444	0.8797	0.7068	0.8421
9	0.7895	**0.9173**	0.8947	0.7444	0.8722	0.6541	0.8722
10	0.7594	**0.9098**	0.8722	0.7594	0.8196	0.7293	0.8346

Évaluation sur la base Hela Cells

Notre approche de description riche des fenêtres combinée avec la méthode de classification par extra-trees a été évaluée sur la base Hela Cells [Murphy et al., 2000] afin de la positionner par rapport aux méthodes de l'état de l'art sur un problème publique connu pour sa difficulté.

La base compte 862 images en niveau de gris et de taille 512x382 réparties en 10 classes équilibrées, les effectifs variant entre 73 et 98 images. Les classes sont le noyau cellulaire et 9 protéines figurant dans le cytoplasme de la cellule : les filaments d'actine, l'endosome, ER, Golgi gia, Golgi gpp, le lysosome, les microtubules, la mitochondrie et le nucléole. Les protéines ont un rôle essentiel dans la vie des cellules et leur absence peut être dûe à une altération du génome. Elles sont actuellement identifiées par leur séquence d'acides aminés et l'intégration d'un mécanisme de classification automatique de ces différentes protéines est d'un grand intérêt pour le domaine médical dans le cadre de l'aide au diagnostic et du dépistage. Les images de ces différentes classes

sont présentées dans la figure A.14. Notons que pour certaines classes, la structure n'est pas sans rappeler celle des images de la base alvéoscopique (notamment les classes Actin, Mitochondrie et Microtubule).



(a) DNA (b) Actin (c) Nucleolus (d) Golgia

(e) Golgpp (f) Lysosome (g) Mitochondrie (h) ER

(i) Microtubule (j) Endosome

FIGURE A.14 – Les 10 classes d'images de la base Hela cells

Le protocole utilisé est identique à celui de la Section 2.3.2.2 avec une caractérisation riche des fenêtres extraites. Cependant, la base de données étant plus importante, nous avons opté pour l'extraction d'un plus grand nombre de fenêtres avec $N_{app} = 120000$, 12000 fenêtres étant extraites sur chacune des 10 classes de la base, dans une approche identique à [Maree, 2005].

Nous avons comparé les résultats obtenus par notre approche avec ceux des classifieurs de la littérature. Ces résultats sont résumés dans le Tableau A.15. Dans un premier temps, nous observons qu'un taux de 83% a été obtenu sur cette base en classification manuelle, i.e. par l'œil humain, montrant ainsi la difficulté importante de cette tâche de classification. Ces résultats montrent que l'approche "LBP+ET" que nous proposons, tout en demeurant générique, se positionne favorablement pour le traitement de ce problème avec un taux de 90.9%. Ainsi, ce taux de reconnaissance est meilleur que celui obtenu via la caractérisation par pixels bruts et celui obtenu par la classification manuelle. La meilleure méthode présentée dans le tableau obtient cependant 95.3% de taux de reconnaissance, la meilleure performance connue à ce jour sur cette base [Chebira et al., 2007], utilisant une approche adaptative multi-résolution à l'aide de filtres de Haar.

TABLE A.15 – Comparaisons des performances de différents systèmes de classification pour la base Hela Cells.

SIFT + ET (*notre approche*)	$75.4\% \pm 4.7\%$
Classification manuelle [Murphy et al., 2003]	83.0%
Pixels bruts + ET [Marée et al., 2007]	$83.3\% \pm 2.7\%$
$LBP_{(8,1)(16,2)}$ + ET (*notre approche*)	$90.9\% \pm 1.7\%$
Méthode dédiée [Chebira et al., 2007]	95.3%

Annexe au chapitre 4

Une approche dérivée des OCRF

Nous étudions actuellement une approche dérivée des OCRF simulant la présence des données outliers dans le set d'apprentissage sans les générer et que nous appelons $OCRF_{DENSE}$. Dans cette approche, la règle de partitionnement est établie lors de l'évaluation d'une densité de probabilité estimée sur les données targets en chacun des nœuds de l'arbre. Elle ne formule pas d'hypothèses particulières sur la distribution des outliers hormis l'idée que ces derniers sont situés dans les zones faiblement peuplées en données targets.

Étant une approche de simulation, elle nécessite la modification du critère de partitionnement classique utilisé lors de l'induction de l'arbre. Il s'agit d'estimer en chaque nœud de l'arbre la distribution des données targets. Pour ce faire, nous avons utilisé l'estimateur standard non-paramétrique des fenêtres de Parzen, qui en un point de coupure x_{cut} sur un attribut de partitionnement donné fournit la valeur de probabilité :

$$P_j(x_{cut}) = \frac{1}{n \cdot h} \cdot \sum_{i=1..n} g(\frac{x_{cut} - x_i}{h})$$

avec g un noyau gaussien centré en les x_i, x_i étant les n données d'apprentissage présentes au nœud courant et h la largeur de la fenêtre de Parzen. On évalue un nombre limité k_{parzen} de points de coupures choisis aléatoirement pour une variable de partitionnement choisie également aléatoirement parmi un set de K_{RFS} variables. Si aucun des points de coupure n'a une valeur de probabilité inférieure à un seuil θ donné, alors le nœud courant est une feuille qui sera étiquetée target. Aucune zone sparse n'a donc pu être identifiée avec l'estimateur. Si au contraire toutes les valeurs sont en dessous du seuil, alors il s'agit d'une zone sparse et le nœud courant est transformé en une feuille outlier. Si des valeurs de coupure sont trouvées de part et d'autre du seuil, la plus faible valeur est alors choisie comme valeur de coupure pour le nœud courant. De l'information a priori obtenue en amont de l'apprentissage sur la distribution des outliers peut être injectée lors du choix du point de coupure comme dans le cas des OCRF classiques (e.g. localisation des zones sparses de l'espace de description à l'aide de l'histogramme de la distribution complémentaire des données targets). L'algorithme de $OCRF_{DENSE}$ est présenté dans le Tableau 2.

Nous avons fait face à plusieurs difficultés lors de l'élaboration de cette approche. La première réside dans l'estimation de la largeur h de la fenêtre de Parzen. Plusieurs valeurs empiriques sont proposées dans la littérature, généralement de la forme $h \approx a \cdot \sigma \cdot n^{-b}$, avec $b = 1/(m+4)$ un coefficient dépendant de la dimension m de l'espace d'analyse, σ la variance estimée sur l'échantillon et a étant un coefficient proche de 1 [Silverman Bernard, 1986, Scott and Sain, 2005]. Des approches de paramétrisation par validation croisée ont également été proposées mais ne sont pas pertinentes pour notre problématique où un modèle de la distribution des targets est construit en chacun des nœuds de l'arbre [1]. Des approches adaptatives ont également été proposées, notamment en présence d'un faible nombre de données, dans lesquelles la largeur de la fenêtre

1. Une approche par évaluation de la qualité du partitionnement obtenu au nœud courant est cependant possible dans une idée similaire à la validation croisée

dépend du nombre de voisins en chacun des points de la distribution [Scott and Sain, 2005]. La seconde difficulté réside dans le choix du seuil θ permettant de décider de la classe d'une zone échantillonnée. Les résultats obtenus à l'aide de notre approche sont très dépendants de ces deux hyperparamètres et une étude approfondie est nécessaire pour mieux comprendre son influence.

L'estimation de la distribution des données targets nécessite un nombre suffisant de données. Or, l'arbre étant complètement développé, plus on avance dans le partitionnement de l'espace, plus le nombre de données targets se réduit. La fiabilité de l'estimation devient ainsi plus faible. Une solution que nous proposons est d'élaguer l'arbre de décision, afin d'assurer la présence d'un nombre minimal de données targets.

Nous avons choisi l'estimateur non-paramétrique des fenêtres de Parzen mais d'autres choix sont possibles comme l'estimateur gaussien ou la mixture de gaussiennes ou un simple histogramme. Les premières expérimentations que nous avons menées ont montré que l'approche était faisable. Cependant, les résultats sont très instables en raison d'une paramétrisation qui s'avère difficile. Une étude plus approfondie des valeurs des hyperparamètres est en cours pour mieux appréhender leur influence sur les performances. Néanmoins, cette approche ouvre de nouvelles perspectives aux forêts one-class.

Algorithm 2 Apprentissage des $OCRF_{DENSE}$: une approche simulant les données outliers par estimation de la distribution des données targets en chacun des nœuds des arbres de la forêt

Require : base d'apprentissage de données target T , nombre d'arbres de la forêt L, la dimension des sous-espaces RSM K_{RSM}, le nombre de valeurs de coupure à évaluer K_{parzen} (nombre de tirages en chaque noeud), seuil sur la probabilité de présence d'outliers θ, largeur des fenêtres de Parzen h.

Ensure : le classifieur de type forêt aléatoire

for $l = 1$ to L **do**
  (i) Sélectionner un échantillon bootstrap T_l du set d'apprentissage T ;
  (ii) Projeter cet set bootstrap dans un sous-espace aléatoirement choisi par la méthode random subspace et de dimension K_{RSM} puis commencer l'induction de l'arbre sur ce dernier set ;
  En chacun des nœuds de l'arbre, faire (A) et (B) :
  (A) ***Critère de partitionnement***
  Tirer aléatoirement K_{parzen} couples (a_j, x_{cut_i}) parmi tous les couples caractéristique/point de coupure possibles ;
  S'il existe au moins un couple vérifiant $P(x_{cut_i}) \geq \theta$ alors sélectionner un couple $(a_j, x_{cut}*)$ vérifiant $P(x_{cut}*) < \theta$ pour partitionner le nœud courant, avec :

$$P(x) = \frac{1}{n \cdot h} \cdot \sum_{i=1..n} g(\frac{x - x_i}{h}),$$ n le nombre de données targets au nœud courant, g le noyau gaussien.

  Si les conditions précédentes ne peuvent être vérifiées, alors le nœud courant est transformé en une feuille selon le critère d'arrêt qui suit.
  (B) ***Critère d'arrêt***
  Si pour tout i des K_{parzen} tirages $P(x_{cut_i}) < \theta$, transformer le nœud courant en feuille étiquetée "outlier".
  Si pour tout i des K_{parzen} tirages $P(x_{cut_i}) \geq \theta$, transformer le nœud courant en feuille étiquetée "target".
end for
return OCRF

Test de multinormalité de Mardia

Le test de multi-normalité de Mardia [Mardia, 1974, Von Eye and Bogat, 2004] utilise deux moments statistiques multivariés, le coefficient d'aplatissement (ou kurtosis, moment d'ordre 4) et le coefficient d'asymétrie (moment d'ordre 3) afin de tester indépendamment si les valeurs obtenues vérifient l'hypothèse de multi-normalité. Les données évaluées ont des valeurs compatibles avec une distribution multi-normale seulement si l'hypothèse de multi-normalité n'est pas rejetée pour les deux tests.

L'expression du coefficient d'asymétrie pour un échantillon de taille N dans un espace de dimension d est donné par :

$$\gamma_{1,d} = \frac{1}{N^2} \sum_{i \leq N, j \leq N} m_{ij}^3 \tag{A.2}$$

et le coefficient d'aplatissement est donné par :

$$\gamma_{2,d} = \frac{1}{N} \sum_{i \leq N} m_{ii}^2 \tag{A.3}$$

avec $m_{i,j} = (x_i - \bar{x})^T S^{-1} (x_j - \bar{x})$, x_i étant une donnée de l'échantillon, $\bar{x}$ la valeur moyenne et S la matrice de variance-covariance de l'échantillon.

Sous l'hypothèse de multi-normalité, la statistique $N \cdot \gamma_{1,d}/6$ suit asymptotiquement la distribution du χ^2 avec $d(d+1)(d+2)/6$ degrés de liberté. Si l'estimation obtenue à l'aide de l'équation A.2 dévie significativement de la valeur de référence tabulée, alors on peut conclure que les données ne proviennent vraisemblablement pas d'une distribution multi-normale. De même, la statistique $\gamma_{2,d}$ suit une distribution normale de moyenne $d(d+2)$ et de variance $8d(d+2)/N$. La valeur estimée pour la statistique centrée réduite associée à $\gamma_{2,d}$ peut alors être comparée aux valeurs tabulées de la distribution normale.

Nous présentons dans le Tableau A.16 les résultats détaillés obtenus pour les bases de l'UCI dans le cadre one-class. Nous rappelons également dans ce tableau les résultats obtenus par notre approche OCRF et l'estimateur gaussien. L'hypothèse de multi-normalité est acceptée (A) si les valeurs absolues des statistiques pour le test du coefficient d'asymétrie (Ms) et pour celui du coefficient d'aplatissement (Mk) sont plus petites que leurs valeurs critiques (CVs et CVk respectivement). Autrement l'hypothèse est rejetée (R).

Les valeurs présentées ont été calculées en utilisant l'implémentation publique de A. Trujillo-Ortiz et R. Hernandez-Walls [2].

Dataset	Gauss	OCRF	M	Hs	Hk	Ms	CVs	Mk	CVk
iris_versicolour	0.903	0.579	A	A	A	23.70	31.41	-1.03	1.64
iris_virginica	0.813	0.614	A	A	A	24.73	31.41	-0.34	1.64
iris_setosa	0.921	0.722	A	A	A	24.22	31.41	0.81	1.64
bw_benign	0.902	0.919	R	R	R	17792	195.97	262.98	1.64
bw_malignant	0.179	0.629	R	R	A	379.05	195.97	0.19	1.64
ionosphere_good	0.781	0.683	-	-	-	-	7337.70	-	1.64
ionosphere_bad	-0.410	0.169	-	-	-	-	7337.70	-	1.64
musk_0	0.264	0.071	R	R	R	7427800	778270	759.52	1.64
musk_1	0.818	0.306	R	R	R	2510300	778270	400.30	1.64
sonar_mines	0.342	0.048	R	R	R	46219	38274	11.72	1.64
sonar_rocks	0.120	0.179	R	R	R	42422	38274	4.29	1.64
diabetes_positive	0.147	0.139	R	R	R	834.50	146.57	16.28	1.64
diabetes_negative	-0.046	0.241	R	R	R	2036.70	146.57	35.90	1.64
pendigits_0	0.970	0.976	R	R	R	35883	883.57	276.09	1.64
pendigits_1	0.652	0.585	R	R	R	25181	883.57	145.08	1.64
pendigits_2	0.957	0.835	R	R	R	29982	883.57	181.53	1.64

La suite des résultats à la page suivante.

2. Trujillo-Ortiz, A. and R. Hernandez-Walls. (2003). Mskekur : Mardia's multivariate skewness and kurtosis coefficients and its hypotheses testing. A MATLAB file. URL : http ://www.mathworks.com/matlabcentral/fileexchange/3519

Dataset	Gauss	OCRF	M	Hs	Hk	Ms	CVs	Mk	CVk
pendigits_3	0.969	0.918	R	R	R	93759	883.57	501.62	1.64
pendigits_4	0.969	0.961	-	-	-	-	883.57	-	1.64
pendigits_5	0.880	0.756	R	R	R	9777	883.57	47.19	1.64
pendigits_6	0.970	0.985	R	R	R	79284	883.57	423.84	1.64
pendigits_7	0.887	0.887	R	R	R	16781	883.57	72.32	1.64
pendigits_8	0.716	0.634	R	R	R	15720	883.57	71.33	1.64
pendigits_9	0.577	0.577	R	R	R	22974	883.57	117.02	1.64
optdigits_0	0.954	0.776	-	-	-	-	46259	-	1.64
optdigits_1	0.937	0.147	-	-	-	-	46259	-	1.64
optdigits_2	0.953	0.143	-	-	-	-	46259	-	1.64
optdigits_3	0.914	0.121	-	-	-	-	46259	-	1.64
optdigits_4	0.905	0.077	-	-	-	-	46259	-	1.64
optdigits_5	0.954	0.041	-	-	-	-	46259	-	1.64
optdigits_6	0.956	0.410	-	-	-	-	46259	-	1.64
optdigits_7	0.933	0.264	-	-	-	-	46259	-	1.64
optdigits_8	0.719	0.043	-	-	-	-	46259	-	1.64
optdigits_9	0.860	0.077	-	-	-	-	46259	-	1.64
mfeat_factors_0	0.737	0.844	R	A	R	1182300	1706100	-197.80	1.64
mfeat_factors_1	0.712	0.873	R	A	R	1294000	1706100	-181.54	1.64
mfeat_factors_2	0.740	0.879	R	A	R	1307400	1706100	-174.76	1.64
mfeat_factors_3	0.695	0.887	R	A	R	1274800	1706100	-190.40	1.64
mfeat_factors_4	0.743	0.884	R	A	R	1292000	1706100	-182.07	1.64
mfeat_factors_5	0.738	0.843	R	A	R	1403200	1706100	-122.57	1.64
mfeat_factors_6	0.770	0.910	R	A	R	1261500	1706100	-203.41	1.64
mfeat_factors_7	0.841	0.879	R	A	R	1322600	1706100	-172.74	1.64
mfeat_factors_8	0.647	0.613	R	A	R	1313900	1706100	-172.14	1.64
mfeat_factors_9	0.751	0.866	R	A	R	1259600	1706100	-199.69	1.64
mfeat_karhunen_0	0.784	0.807	R	R	R	60984	46259	21.57	1.64
mfeat_karhunen_1	0.765	0.750	R	R	R	62413	46259	23.83	1.64
mfeat_karhunen_2	0.776	0.755	R	R	R	59144	46259	17.76	1.64
mfeat_karhunen_3	0.759	0.703	R	R	R	63683	46259	23.30	1.64
mfeat_karhunen_4	0.794	0.813	R	R	R	62594	46259	23.28	1.64
mfeat_karhunen_5	0.730	0.622	R	R	R	56194	46259	12.32	1.64
mfeat_karhunen_6	0.790	0.684	R	R	R	63072	46259	24.79	1.64
mfeat_karhunen_7	0.849	0.864	R	R	R	70505	46259	38.11	1.64
mfeat_karhunen_8	0.713	0.407	R	R	R	54754	46259	9.40	1.64
mfeat_karhunen_9	0.770	0.752	R	R	R	59354	46259	18.11	1.64
mfeat_zernike_0	0.944	0.697	R	R	R	45559	18741	65.53	1.64
mfeat_zernike_1	0.908	0.663	R	R	R	52521	18741	81.09	1.64
mfeat_zernike_2	0.903	0.679	R	R	R	31692	18741	32.08	1.64
mfeat_zernike_3	0.674	0.365	R	R	R	39827	18741	54.49	1.64
mfeat_zernike_4	0.908	0.461	R	R	R	42016	18741	59.67	1.64
mfeat_zernike_5	0.721	0.322	R	R	R	33583	18741	36.38	1.64
mfeat_zernike_6	0.551	0.413	R	R	R	42811	18741	65.64	1.64
mfeat_zernike_7	0.925	0.796	R	R	R	34880	18741	42.32	1.64
mfeat_zernike_8	0.908	0.548	R	R	R	38553	18741	52.73	1.64
mfeat_zernike_9	0.578	0.455	R	R	R	45975	18741	71.96	1.64
mfeat_morph_0	0.682	0.698	-	-	-	-	74.47	-	1.64
mfeat_morph_1	0.345	0.304	R	R	R	10122	74.47	218.95	1.64
mfeat_morph_2	0.400	0.291	R	R	R	10030	74.47	201.90	1.64
mfeat_morph_3	0.326	0.335	-	-	-	-	74.47	-	1.64
mfeat_morph_4	0.432	0.294	R	R	R	7960.40	74.47	168.16	1.64
mfeat_morph_5	0.468	0.378	-	-	-	-	74.47	-	1.64
mfeat_morph_6	0.397	0.637	R	R	R	12871.00	74.47	300.58	1.64
mfeat_morph_7	0.524	0.398	-	-	-	-	74.47	-	1.64
mfeat_morph_8	0.682	0.943	R	R	R	4434.20	74.47	90.45	1.64
mfeat_morph_9	0.389	0.456	R	R	R	10504	74.47	248.09	1.64

La suite des résultats à la page suivante.

Dataset	Gauss	OCRF	M	Hs	Hk	Ms	CVs	Mk	CVk
glass_1	0.465	0.403	R	R	R	1005.50	195.97	18.55	1.64
glass_2	0.212	0.229	R	R	R	1397.60	195.97	28.80	1.64
glass_3	0.179	0.064	R	A	R	143.56	195.97	-2.33	1.64
glass_5	0.964	0.498	R	A	R	109.14	195.97	-3.25	1.64
glass_7	0.308	0.813	R	R	R	314.42	195.97	2.74	1.64

TABLE A.16 – Détails des résultats du test statistique de Mardia pour le coefficient d'asymétrie et le coefficient d'aplatissement multivariés [Von Eye and Bogat, 2004, Mardia, 1974] sur les bases de l'UCI [Blake and Merz, 1998a]. Les valeurs du coefficient de corrélation de Matthews (MCC) [Matthews, 1975] pour les classifieurs OCRF et Gauss sont présentés dans les 2 premières colonnes ; M représente le résultat du test de Mardia, Hs celui du test pour le coefficient d'asymétrie, Hk pour celui du coefficient d'aplatissement, Ms la valeur du coefficient d'asymétrie, Mk celle du coefficient d'aplatissement, CVs la valeur critique du coefficient d'asymétrie et CVk celle du coefficient d'aplatissement ; A indique que l'hypothèse de multi-normalité a été acceptée et R qu'elle a été rejetée. Les valeurs indiquées par le trait d'union "-" n'ont pu être obtenues en raison de singularités présentes dans les matrices de variance-covariance empêchant le calcul des valeurs des statistiques utilisées pour les deux tests.

Résultats des OCRF sur MFeat-FKZM

Nous présentons dans cette section les résultats détaillés de l'étude de la robustesse des OCRF par rapport à la dimension sur la base MFeat-FKZM (cf. Section 4.4.1.4). Nous rappelons que cette base a été obtenue à partir de la concaténation de différents espaces de description déjà mentionnés pour cette base : MFeat-Fourier, MFeat-Karhunen, MFeat-Zernike, MFeat-Morphological. Nous comparons dans les différents tableaux ci-après (Tableau A.17 à A.26) les performances des OCRF à celles des autres approches OCSVM, Gauss, Parzen et MoG pour les différentes classes MFeat (Digit 0, 1, 2...,9). Á travers ces différents tableaux, nous constatons que les OCRF se comportent favorablement par rapport à la dimension comparé aux autres méthodes. Ces résultats confirment la robustesse à la dimension de l'approche oneclass que nous proposons.

TABLE A.17 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 0

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.24 (± 0.20)	**0.71** (± 0.18)	0.19 (± 0.11)	0.20 (± 0.09)	0.19 (± 0.11)
3	0.36 (± 0.21)	**0.49** (± 0.31)	0.32 (± 0.15)	0.34 (± 0.17)	0.35 (± 0.15)
5	**0.50** (± 0.15)	0.11 (± 0.18)	0.41 (± 0.13)	0.42 (± 0.11)	0.46 (± 0.14)
10	**0.72** (± 0.10)	0.01 (± 0.03)	0.68 (± 0.09)	0.19 (± 0.15)	0.68 (± 0.06)
15	**0.76** (± 0.08)	0.00 (± 0.00)	0.70 (± 0.10)	0.04 (± 0.07)	0.70 (± 0.08)
20	**0.74** (± 0.08)	0.00 (± 0.00)	0.79 (± 0.08)	0.00 (± 0.00)	0.64 (± 0.08)
25	**0.72** (± 0.09)	0.00 (± 0.00)	0.79 (± 0.08)	0.00 (± 0.00)	0.54 (± 0.12)
50	**0.77** (± 0.07)	0.00 (± 0.00)	0.88 (± 0.03)	0.00 (± 0.00)	0.26 (± 0.18)
75	**0.78** (± 0.06)	0.00 (± 0.00)	0.81 (± 0.07)	0.16 (± 0.35)	0.04 (± 0.09)
100	**0.79** (± 0.05)	0.00 (± 0.00)	0.71 (± 0.09)	0.03 (± 0.09)	0.02 (± 0.07)
125	**0.80** (± 0.05)	0.00 (± 0.00)	0.77 (± 0.13)	0.11 (± 0.23)	0.00 (± 0.00)
150	**0.80** (± 0.02)	0.00 (± 0.00)	0.65 (± 0.23)	0.00 (± 0.00)	0.00 (± 0.00)
175	**0.83** (± 0.03)	0.00 (± 0.00)	0.54 (± 0.27)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.81** (± 0.03)	0.00 (± 0.00)	0.64 (± 0.24)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.82** (± 0.04)	0.00 (± 0.00)	0.42 (± 0.40)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.81** (± 0.03)	0.00 (± 0.00)	0.39 (± 0.33)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.81** (± 0.04)	0.00 (± 0.00)	0.55 (± 0.30)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.82** (± 0.04)	0.00 (± 0.00)	0.53 (± 0.29)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.83** (± 0.03)	0.00 (± 0.00)	0.54 (± 0.23)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.85** (± 0.02)	0.00 (± 0.00)	0.60 (± 0.23)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.18 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 1

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.15 (± 0.09)	**0.53** (± 0.16)	0.10 (± 0.06)	0.13 (± 0.05)	0.12 (± 0.05)
3	0.20 (± 0.13)	**0.39** (± 0.31)	0.11 (± 0.09)	0.17 (± 0.12)	0.15 (± 0.10)
5	**0.39** (± 0.09)	0.00 (± 0.00)	0.19 (± 0.07)	0.30 (± 0.10)	0.26 (± 0.06)
10	**0.59** (± 0.13)	0.00 (± 0.00)	0.34 (± 0.09)	0.23 (± 0.12)	0.56 (± 0.11)
15	**0.70** (± 0.08)	0.00 (± 0.00)	0.46 (± 0.14)	0.11 (± 0.07)	0.66 (± 0.09)
20	0.68 (± 0.09)	0.00 (± 0.00)	0.56 (± 0.14)	0.00 (± 0.00)	**0.69** (± 0.07)
25	**0.71** (± 0.09)	0.00 (± 0.00)	0.69 (± 0.13)	0.00 (± 0.00)	0.67 (± 0.12)
50	0.80 (± 0.03)	0.00 (± 0.00)	**0.88** (± 0.02)	0.00 (± 0.00)	0.36 (± 0.18)
75	0.80 (± 0.05)	0.00 (± 0.00)	**0.84** (± 0.05)	0.00 (± 0.00)	0.25 (± 0.14)
100	**0.79** (± 0.04)	0.00 (± 0.00)	0.76 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
125	**0.80** (± 0.04)	0.00 (± 0.00)	0.72 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
150	**0.82** (± 0.03)	0.00 (± 0.00)	0.69 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
175	**0.80** (± 0.04)	0.00 (± 0.00)	0.64 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.79** (± 0.05)	0.00 (± 0.00)	0.53 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.80** (± 0.04)	0.00 (± 0.00)	0.50 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.77** (± 0.03)	0.00 (± 0.00)	0.38 (± 0.14)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.80** (± 0.04)	0.00 (± 0.00)	0.37 (± 0.08)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.80** (± 0.02)	0.00 (± 0.00)	0.35 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.79** (± 0.04)	0.00 (± 0.00)	0.27 (± 0.04)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.79** (± 0.04)	0.00 (± 0.00)	0.28 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.19 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 2

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.11 (± 0.07)	**0.67** (± 0.19)	0.15 (± 0.06)	0.13 (± 0.04)	0.14 (± 0.06)
3	**0.26** (± 0.16)	**0.26** (± 0.30)	0.22 (± 0.10)	0.23 (± 0.15)	0.24 (± 0.13)
5	**0.44** (± 0.09)	0.12 (± 0.19)	0.36 (± 0.11)	0.36 (± 0.08)	0.41 (± 0.11)
10	**0.57** (± 0.10)	0.00 (± 0.00)	0.51 (± 0.10)	0.12 (± 0.08)	0.56 (± 0.11)
15	0.68 (± 0.05)	0.00 (± 0.00)	**0.71** (± 0.16)	0.01 (± 0.03)	0.65 (± 0.05)
20	0.74 (± 0.08)	0.00 (± 0.00)	**0.82** (± 0.08)	0.00 (± 0.00)	0.67 (± 0.08)
25	0.75 (± 0.05)	0.00 (± 0.00)	**0.80** (± 0.16)	0.00 (± 0.00)	0.58 (± 0.11)
50	0.77 (± 0.05)	0.00 (± 0.00)	**0.89** (± 0.05)	0.00 (± 0.00)	0.41 (± 0.22)
75	0.80 (± 0.04)	0.00 (± 0.00)	**0.83** (± 0.09)	0.00 (± 0.00)	0.08 (± 0.12)
100	**0.79** (± 0.04)	0.00 (± 0.00)	0.75 (± 0.11)	0.00 (± 0.00)	0.06 (± 0.17)
125	**0.76** (± 0.05)	0.00 (± 0.00)	0.63 (± 0.14)	0.00 (± 0.00)	0.00 (± 0.00)
150	**0.79** (± 0.02)	0.00 (± 0.00)	0.63 (± 0.20)	0.00 (± 0.00)	0.00 (± 0.00)
175	**0.80** (± 0.03)	0.00 (± 0.00)	0.58 (± 0.21)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.79** (± 0.06)	0.00 (± 0.00)	0.61 (± 0.28)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.80** (± 0.04)	0.00 (± 0.00)	0.52 (± 0.31)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.80** (± 0.06)	0.00 (± 0.00)	0.38 (± 0.31)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.80** (± 0.03)	0.00 (± 0.00)	0.58 (± 0.18)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.81** (± 0.02)	0.00 (± 0.00)	0.58 (± 0.06)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.79** (± 0.05)	0.00 (± 0.00)	0.51 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.81** (± 0.03)	0.00 (± 0.00)	0.52 (± 0.06)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.20 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 3

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.14 (± 0.09)	**0.67** (± 0.19)	0.15 (± 0.05)	0.15 (± 0.05)	0.15 (± 0.04)
3	0.22 (± 0.10)	**0.40** (± 0.32)	0.16 (± 0.09)	0.18 (± 0.09)	0.19 (± 0.09)
5	**0.32** (± 0.11)	0.07 (± 0.15)	0.22 (± 0.07)	0.25 (± 0.09)	0.27 (± 0.08)
10	**0.52** (± 0.15)	0.02 (± 0.04)	0.33 (± 0.14)	0.29 (± 0.10)	0.46 (± 0.17)
15	0.61 (± 0.07)	0.02 (± 0.04)	0.47 (± 0.11)	0.07 (± 0.06)	0.61 (± 0.05)
20	**0.66** (± 0.11)	0.00 (± 0.00)	0.58 (± 0.18)	0.01 (± 0.03)	0.61 (± 0.07)
25	**0.66** (± 0.10)	0.04 (± 0.05)	0.64 (± 0.18)	0.08 (± 0.04)	0.63 (± 0.12)
50	0.76 (± 0.04)	0.01 (± 0.03)	**0.84** (± 0.06)	0.03 (± 0.05)	0.34 (± 0.15)
75	0.77 (± 0.06)	0.02 (± 0.04)	**0.78** (± 0.04)	0.05 (± 0.05)	0.09 (± 0.09)
100	**0.76** (± 0.03)	0.00 (± 0.00)	0.75 (± 0.10)	0.02 (± 0.04)	0.07 (± 0.22)
125	**0.76** (± 0.04)	0.02 (± 0.04)	0.64 (± 0.14)	0.06 (± 0.05)	0.00 (± 0.00)
150	**0.79** (± 0.04)	0.02 (± 0.04)	0.66 (± 0.19)	0.05 (± 0.05)	0.00 (± 0.00)
175	**0.80** (± 0.03)	0.01 (± 0.03)	0.62 (± 0.20)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.79** (± 0.03)	0.01 (± 0.03)	0.51 (± 0.22)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.79** (± 0.04)	0.03 (± 0.05)	0.71 (± 0.18)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.77** (± 0.03)	0.01 (± 0.03)	0.70 (± 0.06)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.80** (± 0.03)	0.02 (± 0.04)	0.62 (± 0.19)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.79** (± 0.04)	0.01 (± 0.03)	0.65 (± 0.04)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.81** (± 0.03)	0.03 (± 0.05)	0.58 (± 0.16)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.80** (± 0.03)	0.04 (± 0.05)	0.61 (± 0.04)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.21 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 4

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.17 (± 0.10)	**0.72** (± 0.11)	0.22 (± 0.19)	0.24 (± 0.19)	0.23 (± 0.19)
3	0.25 (± 0.18)	**0.42** (± 0.34)	0.17 (± 0.11)	0.23 (± 0.12)	0.22 (± 0.13)
5	**0.49** (± 0.17)	0.03 (± 0.11)	0.33 (± 0.15)	0.36 (± 0.17)	0.41 (± 0.19)
10	0.57 (± 0.08)	0.00 (± 0.00)	0.54 (± 0.13)	0.14 (± 0.05)	**0.64** (± 0.13)
15	0.69 (± 0.11)	0.00 (± 0.00)	0.68 (± 0.11)	0.00 (± 0.00)	**0.71** (± 0.10)
20	0.72 (± 0.07)	0.00 (± 0.00)	**0.81** (± 0.07)	0.00 (± 0.00)	0.70 (± 0.08)
25	0.78 (± 0.06)	0.00 (± 0.00)	**0.84** (± 0.06)	0.00 (± 0.00)	0.67 (± 0.11)
50	0.75 (± 0.03)	0.00 (± 0.00)	**0.88** (± 0.04)	0.00 (± 0.00)	0.41 (± 0.17)
75	0.78 (± 0.05)	0.00 (± 0.00)	**0.84** (± 0.08)	0.00 (± 0.00)	0.05 (± 0.07)
100	0.79 (± 0.05)	0.00 (± 0.00)	**0.80** (± 0.09)	0.00 (± 0.00)	0.01 (± 0.04)
125	**0.80** (± 0.05)	0.00 (± 0.00)	0.79 (± 0.10)	0.00 (± 0.00)	0.00 (± 0.00)
150	**0.77** (± 0.05)	0.00 (± 0.00)	0.68 (± 0.12)	0.00 (± 0.00)	0.00 (± 0.00)
175	**0.76** (± 0.05)	0.00 (± 0.00)	0.67 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.79** (± 0.05)	0.00 (± 0.00)	0.61 (± 0.20)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.79** (± 0.04)	0.00 (± 0.00)	0.58 (± 0.21)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.80** (± 0.05)	0.00 (± 0.00)	0.56 (± 0.21)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.80** (± 0.04)	0.00 (± 0.00)	0.60 (± 0.06)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.79** (± 0.04)	0.00 (± 0.00)	0.49 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.80** (± 0.03)	0.00 (± 0.00)	0.42 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.77** (± 0.03)	0.00 (± 0.00)	0.42 (± 0.06)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.22 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 5

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.13 (± 0.10)	**0.67** (± 0.17)	0.12 (± 0.07)	0.14 (± 0.08)	0.14 (± 0.07)
3	0.21 (± 0.06)	**0.38** (± 0.31)	0.15 (± 0.05)	0.18 (± 0.04)	0.18 (± 0.05)
5	**0.25** (± 0.07)	0.08 (± 0.18)	0.17 (± 0.04)	0.17 (± 0.05)	0.21 (± 0.05)
10	**0.47** (± 0.13)	0.01 (± 0.03)	0.26 (± 0.05)	0.23 (± 0.07)	0.46 (± 0.10)
15	**0.59** (± 0.11)	0.01 (± 0.03)	0.41 (± 0.11)	0.09 (± 0.08)	0.56 (± 0.11)
20	**0.60** (± 0.12)	0.04 (± 0.05)	0.42 (± 0.15)	0.06 (± 0.06)	0.53 (± 0.07)
25	**0.67** (± 0.09)	0.00 (± 0.00)	0.49 (± 0.08)	0.03 (± 0.05)	0.60 (± 0.09)
50	0.73 (± 0.03)	0.02 (± 0.04)	**0.79** (± 0.10)	0.05 (± 0.05)	0.29 (± 0.12)
75	0.79 (± 0.03)	0.00 (± 0.00)	**0.80** (± 0.07)	0.04 (± 0.05)	0.10 (± 0.09)
100	**0.81** (± 0.03)	0.02 (± 0.04)	0.72 (± 0.07)	0.06 (± 0.05)	0.08 (± 0.14)
125	**0.76** (± 0.04)	0.00 (± 0.00)	0.70 (± 0.13)	0.05 (± 0.05)	0.00 (± 0.00)
150	**0.79** (± 0.04)	0.03 (± 0.05)	0.73 (± 0.20)	0.05 (± 0.05)	0.00 (± 0.00)
175	**0.78** (± 0.02)	0.03 (± 0.05)	0.58 (± 0.26)	0.00 (± 0.00)	0.00 (± 0.00)
200	**0.77** (± 0.04)	0.02 (± 0.04)	0.53 (± 0.30)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.79** (± 0.05)	0.03 (± 0.05)	0.42 (± 0.27)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.79** (± 0.03)	0.01 (± 0.03)	0.57 (± 0.29)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.78** (± 0.04)	0.01 (± 0.03)	0.62 (± 0.22)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.78** (± 0.02)	0.01 (± 0.03)	0.64 (± 0.04)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.77** (± 0.04)	0.01 (± 0.03)	0.57 (± 0.08)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.79** (± 0.04)	0.01 (± 0.03)	0.57 (± 0.08)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.23 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 6

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.19 (± 0.19)	**0.66** (± 0.22)	0.16 (± 0.11)	0.19 (± 0.13)	0.17 (± 0.12)
3	0.25 (± 0.13)	**0.49** (± 0.33)	0.25 (± 0.09)	0.27 (± 0.12)	0.27 (± 0.10)
5	**0.55** (± 0.21)	0.13 (± 0.14)	0.36 (± 0.08)	0.42 (± 0.18)	0.42 (± 0.11)
10	0.70 (± 0.09)	0.07 (± 0.07)	0.57 (± 0.12)	0.28 (± 0.12)	0.70 (± 0.11)
15	0.71 (± 0.08)	0.03 (± 0.05)	0.72 (± 0.09)	0.16 (± 0.23)	**0.74** (± 0.06)
20	0.73 (± 0.09)	0.05 (± 0.05)	0.73 (± 0.12)	0.09 (± 0.05)	0.73 (± 0.07)
25	0.75 (± 0.09)	0.02 (± 0.04)	**0.87** (± 0.05)	0.09 (± 0.06)	0.62 (± 0.11)
50	0.83 (± 0.05)	0.01 (± 0.04)	**0.90** (± 0.04)	0.09 (± 0.05)	0.41 (± 0.18)
75	0.84 (± 0.04)	0.05 (± 0.06)	**0.88** (± 0.06)	0.09 (± 0.05)	0.21 (± 0.14)
100	**0.80** (± 0.03)	0.03 (± 0.05)	0.79 (± 0.07)	0.07 (± 0.06)	0.13 (± 0.24)
125	**0.80** (± 0.03)	0.04 (± 0.05)	0.71 (± 0.07)	0.08 (± 0.06)	0.00 (± 0.00)
150	**0.82** (± 0.04)	0.01 (± 0.03)	0.67 (± 0.09)	0.07 (± 0.05)	0.00 (± 0.00)
175	**0.82** (± 0.04)	0.05 (± 0.05)	0.60 (± 0.17)	0.05 (± 0.06)	0.00 (± 0.00)
200	**0.83** (± 0.03)	0.04 (± 0.06)	0.61 (± 0.17)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.81** (± 0.04)	0.05 (± 0.05)	0.51 (± 0.16)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.81** (± 0.05)	0.06 (± 0.06)	0.52 (± 0.15)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.83** (± 0.04)	0.01 (± 0.03)	0.44 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.81** (± 0.04)	0.05 (± 0.05)	0.40 (± 0.12)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.83** (± 0.04)	0.04 (± 0.05)	0.33 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.82** (± 0.04)	0.04 (± 0.06)	0.38 (± 0.07)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.24 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 7

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.25 (± 0.12)	**0.63** (± 0.23)	0.29 (± 0.18)	0.30 (± 0.15)	0.29 (± 0.17)
3	0.32 (± 0.19)	**0.52** (± 0.33)	0.19 (± 0.11)	0.25 (± 0.15)	0.22 (± 0.13)
5	**0.50** (± 0.18)	0.04 (± 0.08)	0.40 (± 0.21)	0.43 (± 0.16)	0.46 (± 0.18)
10	0.69 (± 0.10)	0.00 (± 0.00)	0.65 (± 0.10)	0.18 (± 0.04)	**0.73** (± 0.07)
15	0.71 (± 0.08)	0.06 (± 0.05)	**0.72** (± 0.12)	0.07 (± 0.05)	0.71 (± 0.06)
20	0.74 (± 0.03)	0.00 (± 0.00)	**0.79** (± 0.05)	0.02 (± 0.04)	0.71 (± 0.09)
25	0.76 (± 0.04)	0.02 (± 0.04)	**0.85** (± 0.06)	0.03 (± 0.05)	0.70 (± 0.09)
50	0.78 (± 0.03)	0.02 (± 0.04)	**0.92** (± 0.03)	0.07 (± 0.05)	0.48 (± 0.19)
75	0.79 (± 0.04)	0.01 (± 0.03)	**0.89** (± 0.04)	0.04 (± 0.05)	0.39 (± 0.29)
100	0.81 (± 0.05)	0.01 (± 0.03)	**0.87** (± 0.07)	0.05 (± 0.05)	0.19 (± 0.33)
125	0.79 (± 0.03)	0.02 (± 0.04)	**0.81** (± 0.05)	0.06 (± 0.05)	0.00 (± 0.00)
150	**0.82** (± 0.04)	0.01 (± 0.03)	0.77 (± 0.07)	0.04 (± 0.05)	0.00 (± 0.00)
175	**0.82** (± 0.04)	0.03 (± 0.05)	0.78 (± 0.06)	0.02 (± 0.04)	0.00 (± 0.00)
200	**0.80** (± 0.04)	0.03 (± 0.05)	0.68 (± 0.11)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.82** (± 0.05)	0.02 (± 0.04)	0.68 (± 0.12)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.80** (± 0.03)	0.02 (± 0.04)	0.62 (± 0.12)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.82** (± 0.03)	0.04 (± 0.05)	0.65 (± 0.10)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.81** (± 0.02)	0.01 (± 0.03)	0.56 (± 0.15)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.82** (± 0.03)	0.04 (± 0.05)	0.56 (± 0.11)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.80** (± 0.03)	0.03 (± 0.05)	0.52 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.25 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 8

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.13 (± 0.07)	**0.64** (± 0.17)	0.18 (± 0.05)	0.20 (± 0.04)	0.19 (± 0.04)
3	0.28 (± 0.10)	**0.35** (± 0.30)	0.21 (± 0.07)	0.27 (± 0.06)	0.26 (± 0.07)
5	**0.39** (± 0.21)	0.10 (± 0.19)	0.25 (± 0.08)	0.27 (± 0.06)	0.30 (± 0.08)
10	0.44 (± 0.10)	0.00 (± 0.00)	0.37 (± 0.09)	0.18 (± 0.09)	0.44 (± 0.09)
15	**0.50** (± 0.16)	0.00 (± 0.00)	0.42 (± 0.08)	0.02 (± 0.05)	0.42 (± 0.08)
20	0.46 (± 0.08)	0.00 (± 0.00)	**0.53** (± 0.06)	0.00 (± 0.00)	0.48 (± 0.09)
25	0.57 (± 0.17)	0.00 (± 0.00)	**0.66** (± 0.11)	0.00 (± 0.00)	0.48 (± 0.09)
50	0.59 (± 0.10)	0.00 (± 0.00)	**0.78** (± 0.10)	0.03 (± 0.08)	0.20 (± 0.13)
75	0.55 (± 0.09)	0.00 (± 0.00)	**0.76** (± 0.03)	0.00 (± 0.00)	0.01 (± 0.04)
100	0.54 (± 0.07)	0.00 (± 0.00)	**0.69** (± 0.10)	0.00 (± 0.00)	0.00 (± 0.00)
125	0.51 (± 0.06)	0.00 (± 0.00)	**0.74** (± 0.14)	0.00 (± 0.00)	0.00 (± 0.00)
150	0.52 (± 0.05)	0.00 (± 0.00)	**0.57** (± 0.18)	0.00 (± 0.00)	0.00 (± 0.00)
175	**0.52** (± 0.07)	0.00 (± 0.00)	0.48 (± 0.28)	0.00 (± 0.00)	0.00 (± 0.00)
200	0.51 (± 0.05)	0.00 (± 0.00)	**0.54** (± 0.24)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.55** (± 0.06)	0.00 (± 0.00)	0.45 (± 0.21)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.53** (± 0.05)	0.00 (± 0.00)	0.22 (± 0.25)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.55** (± 0.06)	0.00 (± 0.00)	0.34 (± 0.16)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.54** (± 0.04)	0.00 (± 0.00)	0.22 (± 0.15)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.54** (± 0.03)	0.00 (± 0.00)	0.19 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.54** (± 0.06)	0.00 (± 0.00)	0.16 (± 0.13)	0.00 (± 0.00)	0.00 (± 0.00)

TABLE A.26 – Performances des OCRF comparées à celles des classifieurs OCSVM, Gauss, Parzen et MoG classifiers sur la base MFeat-FKZM pour Digit 9

m	OCRF	OCSVM	Gauss	Parzen	MoG
2	0.16 (± 0.08)	**0.59** (± 0.20)	0.18 (± 0.08)	0.20 (± 0.09)	0.18 (± 0.07)
3	0.28 (± 0.13)	**0.36** (± 0.30)	0.23 (± 0.09)	0.25 (± 0.10)	0.27 (± 0.11)
5	**0.40** (± 0.10)	0.09 (± 0.18)	0.26 (± 0.10)	0.29 (± 0.14)	0.32 (± 0.12)
10	**0.62** (± 0.10)	0.03 (± 0.05)	0.47 (± 0.11)	0.18 (± 0.09)	0.60 (± 0.10)
15	**0.70** (± 0.11)	0.01 (± 0.03)	0.61 (± 0.12)	0.09 (± 0.13)	0.66 (± 0.09)
20	**0.77** (± 0.08)	0.01 (± 0.03)	0.74 (± 0.11)	0.07 (± 0.15)	0.70 (± 0.10)
25	0.74 (± 0.06)	0.02 (± 0.04)	**0.77** (± 0.08)	0.04 (± 0.05)	0.59 (± 0.08)
50	0.81 (± 0.04)	0.00 (± 0.00)	**0.88** (± 0.03)	0.04 (± 0.05)	0.36 (± 0.23)
75	0.83 (± 0.06)	0.01 (± 0.03)	**0.86** (± 0.03)	0.03 (± 0.05)	0.17 (± 0.21)
100	**0.80** (± 0.03)	0.01 (± 0.03)	0.79 (± 0.07)	0.04 (± 0.05)	0.18 (± 0.32)
125	**0.81** (± 0.04)	0.01 (± 0.03)	0.78 (± 0.11)	0.07 (± 0.05)	0.00 (± 0.00)
150	**0.81** (± 0.04)	0.01 (± 0.03)	0.72 (± 0.16)	0.05 (± 0.05)	0.00 (± 0.00)
175	**0.79** (± 0.05)	0.02 (± 0.04)	0.68 (± 0.19)	0.02 (± 0.04)	0.00 (± 0.00)
200	**0.84** (± 0.02)	0.00 (± 0.00)	0.58 (± 0.20)	0.00 (± 0.00)	0.00 (± 0.00)
225	**0.82** (± 0.07)	0.02 (± 0.04)	0.59 (± 0.24)	0.00 (± 0.00)	0.00 (± 0.00)
250	**0.84** (± 0.05)	0.01 (± 0.03)	0.58 (± 0.21)	0.00 (± 0.00)	0.00 (± 0.00)
275	**0.82** (± 0.04)	0.02 (± 0.04)	0.52 (± 0.19)	0.00 (± 0.00)	0.00 (± 0.00)
300	**0.83** (± 0.06)	0.01 (± 0.03)	0.50 (± 0.21)	0.00 (± 0.00)	0.00 (± 0.00)
325	**0.79** (± 0.03)	0.02 (± 0.04)	0.45 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)
333	**0.80** (± 0.04)	0.01 (± 0.03)	0.49 (± 0.05)	0.00 (± 0.00)	0.00 (± 0.00)

Bibliographie

[Adam et al., 2001] Adam, S., Ogier, J., Cariou, C., Mullot, R., Gardes, J., and Lecourtier, Y. (2001). Utilisation de la transformée de fourier-mellin pour la reconnaissance de formes multi-orientées et multi-échelles : application à l'analyse automatique de documents techniques. *Traitement du signal*, 18(1) :17. (cf. 20.)

[Aggarwal and Yu, 2001] Aggarwal, C. and Yu, P. (2001). Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, page 46. ACM. (cf. 81, 82, and 129.)

[Amit and Geman, 1997] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7) :1545–1588. (cf. 39 and 40.)

[Baja and Thiel, 1996] Baja, G. S. D. and Thiel, E. (1996). Skeletonization algorithm running on path-based distance maps. *Image and Vision Computing*, 14 :47–57. (cf. 17 and 49.)

[Baldi et al., 2000] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics*, 16(5) :412–424. (cf. 92 and 112.)

[Banfield et al., 2007] Banfield, R., Hall, L., Bowyer, K., and Kegelmeyer, W. (2007). A comparison of decision tree ensemble creation techniques. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1) :173–180. (cf. 38.)

[Banhalmi, 2008] Banhalmi, A. (2008). One-class classification methods via automatic counter-example generation. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, pages 58–63. Citeseer. (cf. 80.)

[Banhalmi et al., 2009] Banhalmi, A., Busa-Fekete, R., and Kegl, B. (2009). A one-class classification approach for protein sequences and structures. *Bioinformatics Research and Applications*, pages 310–322. (cf. 73.)

[Bao and Intille, 2004] Bao, L. and Intille, S. (2004). Activity recognition from user-annotated acceleration data. *Pervasive Computing*, pages 1–17. (cf. 30.)

[Bauer and Kohavi, 1999] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms : Bagging, boosting, and variants. *Machine learning*, 36(1) :105–139. (cf. 39.)

[Bay et al., 2006] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf : Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417. (cf. 18, 21, 22, and 24.)

[Belaid and Belaid, 1992] Belaid, A. and Belaid, Y. (1992). *Reconnaissance des formes : méthodes et applications*. Collection : Informatique intelligence artificielle. (cf. 30.)

[Bellman, 1961] Bellman, R. E. (1961). Adaptive control processes : a guided tour. Technical report, Princeton University Press. (cf. 30, 75, and 80.)

[Ben-Hur and Weston, 2010] Ben-Hur, A. and Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology*, 609 :223–239. (cf. 36.)

[Bernard et al., 2008] Bernard, S., Heutte, L., and Adam, S. (2008). Forest-rk : A new random forest induction method. *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pages 430–437. (cf. 40 and 41.)

[Bernard et al., 2009] Bernard, S., Heutte, L., and Adam, S. (2009). Influence of hyperparameters on random forest accuracy. *Multiple Classifier Systems*, pages 171–180. (cf. 41 and 97.)

[Beyer et al., 1999] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful ? *Database Theory-ICDT 99*, pages 217–235. (cf. 30 and 78.)

[Biau, 2010] Biau, G. (2010). Analysis of a random forests model. *Arxiv preprint arXiv :1005.0208*. (cf. 16, 39, 40, 42, 64, and 123.)

[Biau et al., 2009] Biau, G., Cérou, F., Guyader, A., et al. (2009). Sur la vitesse de convergence de l'estimateur du plus proche voisin baggé. (cf. 30.)

[Biau et al., 2008] Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9 :2015–2033. (cf. 16 and 42.)

[Bishop, 1994] Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4) :217–222. (cf. 73, 74, 76, 77, 79, and 82.)

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for pattern recognition*. Clarendon Press Oxford. (cf. 30, 38, 79, 80, and 106.)

[Blake and Merz, 1998a] Blake and Merz, C. J. (1998a). UCI repository of machine learning databases. (cf. 110 and 144.)

[Blake and Merz, 1998b] Blake, C. and Merz, C. (1998b). Uci repository of machine learning databases [http ://www. ics. uci. edu/˜ mlearn/mlrepository. html]. irvine, ca : University of california. *Department of Information and Computer Science*, 55. (cf. 31, 110, and 111.)

[Blum and Langley, 1997] Blum, A. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *In Artificial Intelligence*. (cf. 30.)

[Blum, 1961] Blum, H. (1961). An associative machine for dealing with the visual field and some of its bilogical implications. *Biological Prototypes and synthetic systems. 2nd Annual Bionics Symposium, Cornell Univ., E. E. Bernard and M. R, Kare eds., Plenum Press, New-York*, pages 1 :244–260. (cf. 17.)

[Bosch et al., 2006] Bosch, A., Zisserman, A., and Munoz, X. (2006). Scene classification via plsa. *Computer Vision–ECCV 2006*, pages 517–530. (cf. 24.)

[Bottou and Lin, 2007] Bottou, L. and Lin, C.-J. (2007). Support vector machine solvers. *Large scale kernel machines*, pages 301–320. (cf. 56.)

[Bousquet and Elisseeff, 2002] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2 :499–526. (cf. 39.)

[Bradley, 1997] Bradley, A. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7) :1145–1159. (cf. 112.)

[Brazdil and Soares, 2000] Brazdil, P. and Soares, C. (2000). A comparison of ranking methods for classification algorithm selection. *Machine Learning : ECML 2000*, pages 63–75. (cf. 114.)

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26(2) :123–140. (cf. 39, 40, 73, 85, and 92.)

[Breiman, 1998] Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, Vol. 26(3) :801–849. (cf. 93.)

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, Vol. 45 (1) :5–32. (cf. 16, 29, 39, 40, 41, 42, 64, 86, 92, 93, 95, 97, and 123.)

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. *Wadsworth and Brooks, Monterey, CA*. (cf. 29, 30, 33, and 63.)

[Breslow and Aha, 1997] Breslow, L. A. and Aha, D. W. (1997). Simplifying decision trees : A survey. *The Knowledge Engineering Review*, 12(01) :1–40. (cf. 31 and 32.)

[Brew et al., 2007] Brew, A., Grimaldi, M., and Cunningham, P. (2007). An evaluation of one-class classification techniques for speaker verification. *Artificial Intelligence Review*, 27(4) :295–307. (cf. 74.)

[Brostaux, 2005] Brostaux, Y. (2005). Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'interaction. (cf. 31 and 32.)

[Brown, 2009] Brown, G. (2009). Ensemble learning. (cf. 39.)

[Brown and Kuncheva, 2010] Brown, G. and Kuncheva, L. (2010). Good and bad diversity in majority vote ensembles. *Multiple Classifier Systems*, pages 124–133. (cf. 39 and 95.)

[Canu et al., 2005] Canu, S., Grandvalet, Y., Guigue, V., and Rakotomamonjy, A. (2005). SVM and kernel methods matlab toolbox. *Perception Systemes et Information, INSA de Rouen, France*. (cf. 35 and 54.)

[Cao et al., 2003] Cao, L., Lee, H., and Chong, W. (2003). Modified support vector novelty detector using training data with outliers. *Pattern Recognition Letters*, 24(14) :2479–2487. (cf. 110.)

[Carpenter et al., 1997] Carpenter, G., Rubin, M., and Streilein, W. (1997). ARTMAP-FD : familiarity discrimination applied to radar target recognition. In *Proc. International Conference on Neural Networks*, volume 3, pages 1459–1464. (cf. 73 and 74.)

[Castellano et al., 2004] Castellano, G., Bonilha, L., Li, L., and Cendes, F. (2004). Texture analysis of medical images. *Clinical radiology*, 59(12) :1061–1069. (cf. 20.)

[Chandola et al., 2009a] Chandola, V., Banerjee, A., and Kumar, V. (2009a). Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, 41(3) :1–58. (cf. 72, 73, and 74.)

[Chandola et al., 2009b] Chandola, V., Banerjee, A., and Kumar, V. (2009b). Outlier detection : A survey. *ACM Computing Surveys*, pages 1–72. (cf. 72, 73, and 74.)

[Chang and Lin, 2001] Chang, C. and Lin, C. (2001). Libsvm : a library for support vector machines. (cf. 35, 106, and 114.)

[Chatelain, 2006] Chatelain, C. (2006). *Extraction de sequences numeriques dans des documents manuscrits quelconques*. PhD thesis, Universite de Rouen. (cf. 37.)

[Chebira et al., 2007] Chebira, A., Barbotin, Y., Jackson, C., Merryman, T., Srinivasa, G., Murphy, R., and Kovačević, J. (2007). A multiresolution approach to automated classification of protein subcellular location images. *BMC bioinformatics*, 8(1) :210. (cf. 139 and 140.)

[Cocquerez and Philipp, 1995] Cocquerez, J.-P. and Philipp, S. (1995). *Analyse d'images :filtrage et segmentation*. Masson :enseignement de la physique. (cf. 17, 18, 19, and 20.)

[Cohen et al., 2008] Cohen, G., Sax, H., Geissbuhler, A., et al. (2008). Novelty detection using one-class parzen density estimator. an application to surveillance of nosocomial infections. volume 136, page 21. IOS Press ; 1999. (cf. 77 and 78.)

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3) :273–297. (cf. 30, 34, and 80.)

[Cover and Hart, 1967] Cover, T. and Hart, P. E. (1967). Nearest neighbour pattern classification. *IEE Transactions Information Theory*, 13 :21–27. (cf. 30.)

[Cutler and Zhao, 2001] Cutler, A. and Zhao, G. (2001). PERT-perfect random tree ensembles. *Computing Science and Statistics*, 33 :490–497. (cf. 40 and 93.)

[Dahmen et al., 2000] Dahmen, J., Hektor, J., Perrey, R., and Ney, H. (2000). Automatic classification of red blood cells using gaussian mixture densities. volume 2000, pages 331–335. (cf. 20.)

[Decaestecker et al., 1998] Decaestecker, C., Remmelink, M., Salmon, I., Camby, I., Goldschmidt, D., Petein, M., Van Ham, P., Pasteels, J.-L., and Kiss, R. (1998). Methodological aspects of using decision trees to characterise leiomyomatous tumors. *Cytometry*, 24(1) :83–92. (cf. 30.)

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statisfical Society*, Vol. 39(1) :pp. 1–38. (cf. 77 and 106.)

[Demsar, 2006] Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7 :1–30. (cf. 112 and 114.)

[Desforges et al., 1988] Desforges, M., Jacob, P., and Cooper, J. (1988). Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C : Journal of Mechanical Engineering Science*, 212(8) :687–703. (cf. 79.)

[Devijver, 1977] Devijver, P. (1977). *Reconnaissance des formes par la méthode des plus proches voisins*. PhD thesis, Université Pierre et Marie-Curie. (cf. 30.)

[Dietterich, 1997] Dietterich, T. (1997). Machine learning research : four current directions. *AI MAG*, 18(4) :97–136. (cf. 38.)

[Dietterich, 2000a] Dietterich, T. (2000a). Ensemble methods in machine learning. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, volume 1857 of *LNCS*, pages 1–15. Springer. (cf. 29, 38, 39, and 84.)

[Dietterich, 2000b] Dietterich, T. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees : Bagging, boosting, and randomization. *Machine learning*, 40(2) :139–157. (cf. 39, 40, and 86.)

[Dietterich and Bakiri, 1995] Dietterich, T. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2 :263–286. (cf. 38.)

[Dietterich, 1998] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7) :1895–1924. (cf. 39 and 114.)

[Dietterich, 2002] Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, pages 405–408. (cf. 38.)

[Donoho, 2000] Donoho, D. (2000). Aide-memoire. high-dimensional data analysis : The curses and blessings of dimensionality. *American Math. Society Lecture-Math Challenges of the 21st Century*. (cf. 75, 78, 85, and 123.)

[Drucker, 1997] Drucker, H. (1997). Improving regressors using boosting techniques. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 107–115. Citeseer. (cf. 39.)

[Duda and Hart, 1973] Duda, R. and Hart, P. (1973). *Pattern classification and scene analysis*. John Wiley and sons. (cf. 75 and 76.)

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*. Wiley-interscience. (cf. 16, 17, 37, 74, and 78.)

[Due Trier et al., 1996] Due Trier, O., Jain, A., and Taxt, T. (1996). Feature extraction methods for character recognition-a survey. *Pattern recognition*, 29(4) :641–662. (cf. 16, 17, and 18.)

[Duin, 1976] Duin, R. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *Computers, IEEE Transactions on*, 100(11) :1175–1179. (cf. 77 and 107.)

[Duin, 1996] Duin, R. (1996). A note on comparing classifiers. *Pattern Recognition Letters*, 17(5) :529–536. (cf. 112.)

[Duin, 2000] Duin, R. (2000). PRTools version 3.0 : A matlab toolbox for pattern recognition. In *Proc. of SPIE*. Citeseer. (cf. 98, 99, 106, and 114.)

[Dunn, 1961] Dunn, O. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, pages 52–64. (cf. 114.)

[Elisseeff et al., 2006] Elisseeff, A., Evgeniou, T., and Pontil, M. (2006). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1) :55. (cf. 39.)

[Evangelista et al., 2006] Evangelista, P., Embrechts, M., and Szymanski, B. (2006). Taming the curse of dimensionality in kernels and novelty detection. *Applied Soft Computing Technologies : The Challenge of Complexity*, pages 425–438. (cf. 80 and 84.)

[Fan et al., 2004] Fan, W., Miller, M., Stolfo, S., Lee, W., and Chan, P. (2004). Using artificial anomalies to detect unknown and known network intrusions. *Knowledge and Information Systems*, 6(5) :507–527. (cf. 73, 80, and 81.)

[Fisher, 1936] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2) :179–188. (cf. 31 and 32.)

[Fisher, 1959] Fisher, S. (1959). *Statistical methods and scientific inference*, volume 1959. Oliver and Boyd. (cf. 114.)

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. *In International Conference on Machine Learning*, pages 148–156. (cf. 38 and 93.)

[Freund and Schapire, 1995] Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *In European Conference on Computational Learning Theory*, pages pages 23–37. (cf. 38.)

[Friedman, 1937] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200) :675–701. (cf. 114.)

[Friedman, 1940] Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1) :86–92. (cf. 114.)

[Gabor, 1946] Gabor, D. (1946). Theory of communication. part 1 : The analysis of information. *Electrical Engineers-Part III : Radio and Communication Engineering, Journal of the Institution of*, 93(26) :429–441. (cf. 21.)

[Genuer et al., 2008] Genuer, R., Poggi, J.-M., and Tuleau, C. (2008). Random Forests : some methodological insights. (RR-6729). (cf. 16, 39, and 40.)

[Geurts et al., 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1) :3–42. (cf. 16, 39, 40, 41, 46, 55, 64, 97, and 99.)

[Ghazali et al., 2008] Ghazali, K. H., Mustafa, M. M., Hussain, A., and Razali, S. (2008). Scale invariant feature transform technique for weed classification in oil palm plantation. *Journal of Applied Sciences*, 8 :1179–1187. (cf. 17, 22, and 49.)

[Gonzalez and Woods, 2002] Gonzalez, R. C. and Woods, R. E. (2002). *Digital image processing*. Prentice-Hall, 2. ed. edition. (cf. 17.)

[Grandvalet, 2006] Grandvalet, Y. (2006). Stability of bagged decision trees. In *Proceedings of the XLIII Scientific Meeting of the Italian Statistical Society*, pages 221–230. (cf. 39.)

[Grossmann and Morlet, 1984] Grossmann, A. and Morlet, J. (1984). Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15(4) :723–736. (cf. 20.)

[Guo and Jones, 2008] Guo, G. and Jones, M. (2008). Iris Extraction Based on Intensity Gradient and Texture Difference. In *IEEE Workshop on Applications of Computer Vision, 2008. WACV 2008*, pages 1–6. (cf. 17.)

[Guyon et al., 2004] Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2004). Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems*, 17 :545–552. (cf. 123.)

[Hand, 2006] Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1) :1–14. (cf. 112.)

[Hand, 2009] Hand, D. (2009). Measuring classifier performance : a coherent alternative to the area under the roc curve. *Machine learning*, 77(1) :103–123. (cf. 112.)

[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6) :610–621. (cf. 18, 19, and 49.)

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK. (cf. 21, 22, and 23.)

[Hastie et al., 2005] Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning : data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2) :83–85. (cf. 30 and 112.)

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer Verlag. (cf. 29, 30, 34, 36, 48, 55, and 106.)

[Heikkila et al., 2006] Heikkila, M., Pietikainen, M., and Schmid, C. (2006). Description of interest regions with center-symmetric local binary patterns. *LNCS*, 4338 :58. (cf. 27.)

[Hempstalk and Frank, 2008] Hempstalk, K. and Frank, E. (2008). Discriminating against new classes : One-class versus multi-class classification. *AI 2008 : Advances in Artificial Intelligence*, pages 325–336. (cf. 73, 110, 111, and 112.)

[Hempstalk et al., 2008] Hempstalk, K., Frank, E., and Witten, I. (2008). One-class classification by combining density and class probability estimation. *Machine Learning and Knowledge Discovery in Databases*, pages 505–519. (cf. 74, 80, 81, 110, and 115.)

[Hilario et al., 2005] Hilario, C., Collado, J. M., Armingol, J. M., and de la Escalera, A. (2005). Pedestrian detection for intelligent vehicles based on active contour models and stereo vision. In Moreno-Díaz, R., Pichler, F., and Quesada-Arencibia, A., editors, *EUROCAST*, volume 3643 of *Lecture Notes in Computer Science*, pages 537–542. Springer. (cf. 17.)

[Hinneburg and Keim, 1999] Hinneburg, A. and Keim, D. (1999). Optimal grid-clustering : Towards breaking the curse of dimensionality in high-dimensional clustering. In *Proceedings of the 25th International Conference on Very Large Data Bases*, pages 506–517. Citeseer. (cf. 80.)

[Hinneburg et al., 2000] Hinneburg, E., Aggarwal, C., Keim, D., and Hinneburg, A. (2000). What is the nearest neighbor in high dimensional spaces ? (cf. 78.)

[Hinton, 1989] Hinton, G. (1989). Connectionist learning procedures. *Artificial intelligence*, 40(1-3) :185–234. (cf. 79.)

[Ho, 1998] Ho, T. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8) :832–844. (cf. 39, 40, 73, 86, 92, and 99.)

[Ho, 2002] Ho, T. (2002). Multiple classifier combination : Lessons and next steps. *Series in Machine Perception and Artificial Intelligence*, 47 :171–198. (cf. 29.)

[Hodge and Austin, 2004] Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2) :85–126. (cf. 72, 73, 74, and 79.)

[Hoffmann, 2007] Hoffmann, H. (2007). Kernel pca for novelty detection. *Pattern Recognition*, 40(3) :863–874. (cf. 85.)

[Iftene and Safia, 2004] Iftene, T. and Safia, A. (2004). Comparaison entre la matrice de cooccurrence et la transformation en ondelettes pour la classification texturale des images hrv (xs) de spot. *Télédétection*, 4(1) :39–49. (cf. 20.)

[Iman and Davenport, 1979] Iman, R. and Davenport, J. (1979). Approximations of the critical region of the friedman statistic. Technical report, Sandia Labs., Albuquerque, NM (USA) ; Texas Tech Univ., Lubbock (USA). (cf. 114.)

[Jamain and Hand, 2008] Jamain, A. and Hand, D. J. (2008). Mining supervised classification performance studies : A meta-analytic investigation. *J. Classif.*, 25 :87–112. (cf. 112.)

[Japkowicz, 1999] Japkowicz, N. (1999). *Concept-learning in the absence of counter-examples : an autoassociation-based approach to classification*. PhD thesis, Citeseer. (cf. 73, 79, and 85.)

[Japkowicz et al., 1995] Japkowicz, N., Myers, C., and Gluck, M. (1995). A novelty detection approach to classification. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 518–523. (cf. 79.)

[Jimenez and Landgrebe, 1998] Jimenez, L. O. and Landgrebe, D. A. (1998). Supervised classification in high-dimensional space : Geometrical, statistical, and asymptotical properties of multivariate data. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, 28(1) :39–54. (cf. 123.)

[Joutel et al., 2007] Joutel, G., Eglin, V., Bres, S., and Emptoz, H. (2007). Extraction de caractéristiques dans les images par transformée multi-échelle. *GRETSI, Groupe dÉtudes du Traitement du Signal et des Images*. (cf. 20.)

[Jurie and Triggs, 2005] Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. *In IEEE International Conference on Computer Vision*, volume 1 :pages 604–610. (cf. 28.)

[Kass, 1980] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, pages 119–127. (cf. 34.)

[Keysers et al., 2001] Keysers, D., Dahmen, J., and Ney, H. (2001). Invariant classification of red blood cells : A comparison of different approaches. In *Bildverarbeitung für die Medizin*, pages 367–371. (cf. 17.)

[Khan and Madden, 2010] Khan, S. and Madden, M. (2010). A survey of recent trends in one class classification. *Artificial Intelligence and Cognitive Science*, pages 188–197. (cf. 73, 84, and 85.)

[King et al., 2002] King, S., King, D., Astley, K., Tarassenko, L., Hayton, P., and Utete, S. (2002). The use of novelty detection techniques for monitoring high-integrity plant. In *Proceedings of the 2002 International Conference on Control Applications, 2002*, volume 1. (cf. 73 and 74.)

[Kohavi, 1996] Kohavi, R. (1996). Wrappers for performance enhancement and oblivious decision graphs. (cf. 112.)

[Koppel and Schler, 2004] Koppel, M. and Schler, J. (2004). Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM. (cf. 74.)

[Kraaijveld and Duin, 1991] Kraaijveld, M. and Duin, R. (1991). A criterion for the smoothing parameter for parzen-estimators of probability density functions. Technical report, Delft University of Technology. (cf. 77 and 107.)

[Kuncheva, 2007] Kuncheva, L. (2007). Combining pattern classifiers : Methods and algorithms (kuncheva, li ; 2004)[book review]. *Neural Networks, IEEE Transactions on*, 18(3) :964–964. (cf. 29, 38, 39, and 95.)

[Kuncheva and Rodríguez, 2007] Kuncheva, L. and Rodríguez, J. (2007). An experimental study on rotation forest ensembles. *Multiple Classifier Systems*, pages 459–468. (cf. 93.)

[Lam et al., 1992] Lam, L., Lee, S.-W., and Suen, C. Y. (1992). Thinning methodologies, a comprehensive survey. *Pattern Analysis and Machine Intelligence*, page 14(9). (cf. 17.)

[Lauziere et al., 2001] Lauziere, Y., Gingras, D., and Ferrie, F. (2001). A model-based road sign identification system. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1. IEEE Computer Society ; 1999. (cf. 17.)

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4) :541–551. (cf. 17.)

[Lee et al., 2009] Lee, C.-C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2009). Emotion recognition using a hierarchical binary decision tree approach. In *Proc. Interspeech*, pages 320–323. (cf. 30.)

[Lerman and Da Costa, 1996] Lerman, I.-C. and Da Costa, J. F. P. (1996). Coefficients d'association et variables à très grand nombre de catégories dans les arbres de décision ; application à l'identification de la structure secondaire d'une protéine. *Publication interne- IRISA*, (RR-2803). (cf. 32.)

[Liu et al., 2000] Liu, B., Xia, Y., and Yu, P. (2000). Clustering through decision tree construction. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 20–29. ACM New York, NY, USA. (cf. 75, 81, 82, 83, and 95.)

[Lobo et al., 2008] Lobo, J., Jiménez-Valverde, A., and Real, R. (2008). Auc : a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2) :145–151. (cf. 112.)

[Lowe, 2004] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110. (cf. 21, 22, 24, 25, and 49.)

[Luban and Staunton, 1988] Luban, M. and Staunton, L. P. (1988). An efficient method for generating a uniform distribution of points within a hyperspace. *Computers in Physics*, 2(6) :55–60. (cf. 81.)

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA. (cf. 79.)

[Mäenpää, 2003] Mäenpää, T. (2003). *The local binary pattern approach to texture analysis : extensions and applications*. Acta Universitatis Ouluensis : Technica. University of Oulu. (cf. 24, 26, and 27.)

[Maenpaa and Pietikainen, 2003] Maenpaa, T. and Pietikainen, M. (2003). Multi-scale binary patterns for texture analysis. *Proc. 13th Scandinavian Conference on Image Analysis*, pages 885–892. (cf. 27.)

[Mäenpää et al., 2002] Mäenpää, T., Pietikäinen, M., and Viertola, J. (2002). Separating color and pattern information for color texture discrimination. In *ICPR (1)*, pages 668–671. (cf. 17.)

[Mäenpää et al., 2003] Mäenpää, T., Viertola, J., and Pietikäinen, M. (2003). Optimising colour and texture features for real-time visual inspection. *Pattern Anal. Appl.*, 6(3) :169–175. (cf. 17.)

[Mamloukf et al., 2003] Mamloukf, A., Kimf, J., Barthf, E., Brauckmann, M., and Martinetzf, T. (2003). One-class classification with subgaussians. *Pattern recognition : Magdeburg, Germany, September 10-12, 2003*, page 346. (cf. 73.)

[Manevitz and Yousef, 2002] Manevitz, L. M. and Yousef, M. (2002). One-class svms for document classification. *J. Mach. Learn. Res.*, 2 :139–154. (cf. 115.)

[Mardia, 1974] Mardia, K. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya : The Indian Journal of Statistics, Series B*, pages 115–128. (cf. 122, 142, and 144.)

[Maree, 2005] Maree, R. (2005). *Classification automatique d'images par arbres de décision*. Electrical engineering and computer science, University of Liège. (cf. 28 and 139.)

[Marée et al., 2005] Marée, R., Geurts, P., Piater, J., and Wehenkel, L. (2005). Biomedical image classification with random subwindows and decision trees. In *Proc. of ICCV Workshop on Computer Vision for Biomedical Image Applications*, volume 3765 of *LNCS*, pages 220–229. (cf. 41.)

[Maree et al., 2005] Maree, R., Geurts, P., Piater, J., and Wehenkel, L. (2005). Random subwindows for robust image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 1. (cf. 28, 42, 46, and 58.)

[Marée et al., 2007] Marée, R., Geurts, P., and Wehenkel, L. (2007). Random subwindows and extremely randomized trees for image classification in cell biology. *BMC Cell Biology*, 8(Suppl-1) :S–2. (cf. 140.)

[Markou and Singh, 2003] Markou, M. and Singh, S. (2003). Novelty detection : a review–part 1 : statistical approaches. *Signal Processing*, 83(12) :2481–2497. (cf. 73, 75, 77, and 78.)

[Marsland, 2003] Marsland, S. (2003). Novelty detection in learning systems. *Neural computing surveys*, 3 :157–195. (cf. 73.)

[Marée et al., 2003] Marée, R., Geurts, P., Visimberga, G., Piater, J., and Wehenkel, L. (2003). An empirical comparison of machine learning algorithms for generic image classification. (cf. 28.)

[Matthews, 1975] Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2) :442–451. (cf. 92, 112, and 144.)

[Mazhelis, 2006] Mazhelis, O. (2006). One-class classifiers : A review and analysis of suitability in the context of mobile-masquerader detection. *South African Computer Journal (SACJ), ARIMA & SACJ Joint Special Issue on Advances in End-User Data-Mining Techniques*, 36 :29–48. (cf. 73.)

[Mello et al., 2006] Mello, F. C., Bastos, L. G., Soares, S. L., Rezende, V. M., Conde, M. B., Chaisson, R. E., Kritski, A. L., Ruffino-Netto, A., and Werneck, G. L. (2006). Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression : a cross-sectional study. *BMC Public Health*, 6(1) :43. (cf. 30.)

[Menke and Martinez, 2004] Menke, J. and Martinez, T. (2004). Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pages 1331–1335. IEEE. (cf. 114.)

[Mikolajczyk and Schmid, 2004] Mikolajczyk, K. and Schmid, C. (2004). Scale and affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1) :63–86. (cf. 21.)

[Mikolajczyk and Schmid, 2005] Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10) :1615–1630. (cf. 16, 18, 21, and 22.)

[Milman, 1998] Milman, V. (1998). Surprising geometric phenomena in high-dimensional convexity theory. *Progress in Mathematics - Boston*, 169 :73–91. (cf. 78.)

[Mingers, 1989a] Mingers, J. (1989a). An empirical comparison of pruning methods for decision-tree induction. *Machine Learning*, pages 4 :227–243. (cf. 32.)

[Mingers, 1989b] Mingers, J. (1989b). An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, pages 3 :319–342. (cf. 32.)

[Mishra et al., 2011] Mishra, P., Singh, D., and Yamaguchi, Y. (2011). Land cover classification of palsar images by knowledge based decision tree classifier and supervised classifiers based on sar observables. *Progress In Electromagnetics Research B*, 30 :47–70. (cf. 30.)

[Moosmann et al., 2008] Moosmann, F., Nowak, E., and Jurie, F. (2008). Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9) :1632–1646. (cf. 28.)

[Moya and Hush, 1996] Moya, M. and Hush, D. (1996). Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3) :463–474. (cf. 73 and 79.)

[Moya et al., 1993] Moya, M., Koch, M., and Hostetler, L. (1993). One-class classifier networks for target recognition applications. *NASA STI/Recon Technical Report N*, 93 :24043. (cf. 73 and 74.)

[Murphy et al., 2003] Murphy, R., Velliste, M., and Porreca, G. (2003). Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *The Journal of VLSI Signal Processing*, 35(3) :311–321. (cf. 140.)

[Murphy et al., 2000] Murphy, R. F., Boland, M. V., and Velliste, M. (2000). Towards a systematics for protein subcellular location : Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. In Bourne, P. E., Gribskov, M., Altman, R. B., Jensen, N., Hope, D. A., Lengauer, T., Mitchell, J. C., Scheeff, E. D., Smith, C., Strande, S., and Weissig, H., editors, *ISMB*, pages 251–259. AAAI. (cf. 58 and 138.)

[Murthy, 1996] Murthy, K. V. S. (1996). *On growing better decision trees from data*. PhD thesis, The John Hopkins University, Baltimore, Maryland. AAI9617584. (cf. 30.)

[Murthy, 1998] Murthy, S. K. (1998). Automatic construction of decision trees from data : A multi-disciplinary survey. *Data mining and knowledge discovery*, 2(4) :345–389. (cf. 31 and 32.)

[Murthy et al., 1994] Murthy, S. K., Kasif, S., and Salzberg, S. (1994). A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2 :1–32. (cf. 34.)

[Nadeau and Bengio, 2003] Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3) :239–281. (cf. 114.)

[Nairac et al., 1997] Nairac, A., Corbett-Clark, T., Ripley, R., Townsend, N., and Tarassenko, L. (1997). Choosing an appropriate model for novelty detection. In *Artificial Neural Networks, Fifth International Conference on (Conf. Publ. No. 440)*, pages 117–122. IET. (cf. 75.)

[Nairac et al., 1999] Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., and Tarassenko, L. (1999). A system for the analysis of jet engine vibration data. *Integrated Computer-Aided Engineering*, 6(1) :53–66. (cf. 79.)

[Negri et al., 2008] Negri, P., Clady, X., Hanif, S., and Prevost, L. (2008). A cascade of boosted generative and discriminative classifiers for vehicle detection. *EURASIP Journal on Advances in Signal Processing*, 8(2). (cf. 17 and 22.)

[Negri et al., 2004] Negri, P., Clady, X., Milgram, M., and LISIF-PARC, U. (2004). Perception visuelle du geste de préhension : application à la robotique manipulatrice. *18ème Journée des JJCR, Douai, France*. (cf. 20.)

[Nemenyi, 1963] Nemenyi, P. (1963). *Distribution-free multiple comparisons*. PhD thesis, Princeton University. (cf. 114.)

[Nowak and Jurie, 2007] Nowak, E. and Jurie, F. (2007). Learning visual similarity measures for comparing never seen objects. In *CVPR*. IEEE Computer Society. (cf. 28.)

[Nowak et al., 2006] Nowak, E., Jurie, F., and Triggs, B. (2006). Sampling strategies for bag-of-features image classification. pages 490–503. (cf. 28.)

[Ojala et al., 1996] Ojala, T., Pietikainen, M., and Harwood, F. (1996). A comparative study of texture measures with classification based on feature distribution. *Pattern Recognition*, 29(1) :55–59. (cf. 21 and 24.)

[Ojala et al., 2000] Ojala, T., Pietikäinen, M., and Mäenpää, T. (2000). Gray scale and rotation invariant texture classification with local binary patterns. *Computer Vision-ECCV 2000*, pages 404–420. (cf. 21 and 24.)

[Ojala et al., 2002] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7) :971–987. (cf. 48.)

[Oliveira et al., 2008] Oliveira, A., Costa, F., and Clovis Filho, O. (2008). Novelty detection with constructive probabilistic neural networks. *Neurocomputing*, 71(4-6) :1046–1053. (cf. 110.)

[Otsu, 1979] Otsu, N. (1979). A threshold selection method from gray-level histogram. *IEEE Trans. on Systems, Man and Cybernetics*, 9 :62-66. (cf. 17 and 49.)

[Papageorgiou et al., 1998] Papageorgiou, C., Oren, M., and Poggio, T. (1998). A general framework for object detection. *In Proceedings of ICCV*. (cf. 21.)

[Parra et al., 1996] Parra, L., Deco, G., and Miesbach, S. (1996). Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Comput.*, 8(2) :260–269. (cf. 76.)

[Parzen, 1962] Parzen, E. (1962). On the estimation of a probability density function and mode. *Ann. Math. Stat.*, 33 :pp. 1065–1076. (cf. 77.)

[Petitjean et al., 2009] Petitjean, C., Benoist, J., Thiberville, L., Salaün, M., and Heutte, L. (2009). Classification of In-vivo Endomicroscopic Images of the Alveolar Respiratory System. In *IAPR Conference on Machine Vision Applications (MVA)*, pages 471–474, Yokohama, Japon. (cf. 48 and 49.)

[Platt, 1999] Platt, J. (1999). Fast training of support vector machines using sequential minimal imization. advances in kernel methods - support vector learning. *MIT Press*, pages 185–208. (cf. 35.)

[Poggio et al., 2004] Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, 428(6981) :419–422. (cf. 39.)

[Pratt, 1991] Pratt, W. (1991). *Digital Image Processing, 2nd Edition*. John Wiley & Sons. (cf. 49.)

[Quinlan, 1986] Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1) :81–106. (cf. 30 and 33.)

[Quinlan, 1987] Quinlan, J. (1987). Simplifying decision trees. *International Journal of Machine Studies*, 27 :221–234. (cf. 32.)

[Quinlan, 1993] Quinlan, J. (1993). *C4. 5 : programs for machine learning*. Morgan Kaufmann. (cf. 29 and 30.)

[Quinlan, 1996a] Quinlan, J. (1996a). Bagging, boosting, and C4. 5. In *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*. (cf. 39.)

[Quinlan, 1996b] Quinlan, J. (1996b). Improved use of continuous attributes in C4. 5. *Arxiv preprint cs.AI/9603103*. (cf. 30.)

[Rad et al., 2004] Rad, A., Safabakhsh, R., Qaragozlou, N., and Zaheri, M. (2004). Fast iris and pupil localization and eyhelid removal using gradient vector pairs and certainty factors. In *The Irish Machine Vision and Image Processing Conf*, pages 82–91. (cf. 17.)

[Rakotomalala, 2005] Rakotomalala, R. (2005). Arbres de décision. *Revue Modulad*, 33 :163–187. (cf. 30, 33, and 34.)

[Rakotomamonjy, 2003] Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Reseach*, pages 3 :1357–1370. (cf. 34.)

[Ratle et al., 2007] Ratle, F., Kanevski, M., Terrettaz-Zufferey, A., Esseiva, P., and Ribaux, O. (2007). A comparison of one-class classifiers for novelty detection in forensic case data. *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, pages 67–76. (cf. 110.)

[Reddy and Chatterji, 1996] Reddy, B. S. and Chatterji, B. (1996). An fft-based technique for translation, rotation, and scale-invariant image registration. *Image Processing, IEEE Transactions on*, 5(8) :1266–1271. (cf. 20.)

[Roberts and Tarassenko, 1994] Roberts, S. and Tarassenko, L. (1994). A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6(2) :270–284. (cf. 77.)

[Robnik-Sikonja, 2004] Robnik-Sikonja, M. (2004). Improving random forests. *Machine Learning : ECML 2004*, pages 359–370. (cf. 16, 39, 40, and 93.)

[Rodriguez et al., 2006] Rodriguez, J., Kuncheva, L., and Alonso, C. (2006). Rotation forest : A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10) :1619–1630. (cf. 40 and 93.)

[Rodríguez-Damián et al., 2004] Rodríguez-Damián, M., Cernadas, E., Formella, A., and de Sá-Otero, P. (2004). Pollen classification using brightness-based and shape-based descriptors. In *ICPR (2)*, pages 212–215. (cf. 17.)

[Rosenblatt, 1957] Rosenblatt, F. (1957). *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory. (cf. 36.)

[Schapire et al., 1998] Schapire, R., Freund, Y., Bartlett, P., and Lee, W. (1998). Boosting the margin : A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5) :1651–1686. (cf. 39.)

[Schapire, 1990] Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2) :197–227. (cf. 38.)

[Schmid et al., 2000] Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of computer vision*, 37(2) :151–172. (cf. 16, 21, and 22.)

[Schneiderman and Kanade, 2005] Schneiderman, H. and Kanade, T. (2005). A statistical method for 3d object detection applied to faces and cars. *In IEEE International Conference on Computer Vision and Pattern Recognition*, pages 746–751. (cf. 21.)

[Scholkopf et al., 2001] Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7) :1443–1471. (cf. 73, 75, 80, 81, 83, 85, 107, 110, 114, and 115.)

[Scott and Sain, 2005] Scott, D. W. and Sain, S. R. (2005). Multi-dimensional density estimation. *Handbook of Statistics*, 24 :229–261. (cf. 140 and 141.)

[Shao et al., 2003] Shao, H., Svoboda, T., and Gool, L. V. (2003). ZuBuD — Zürich buildings database for image based recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology. (cf. 17.)

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge Univ Pr. (cf. 34.)

[Shieh and Kamm, 2009] Shieh, A. and Kamm, D. (2009). Ensembles of one class support vector machines. *Multiple Classifier Systems*, pages 181–190. (cf. 84 and 85.)

[Short and Fukunaga, 1981] Short, R. D. and Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory*, 27(5) :622–626. (cf. 30.)

[Sillito and Fisher, 2007] Sillito, R. and Fisher, R. (2007). Incremental one-class learning with bounded computational complexity. *Artificial Neural Networks–ICANN 2007*, pages 58–67. (cf. 73.)

[Silverman Bernard, 1986] Silverman Bernard, W. (1986). Density Estimation for Statistics and Data Analysis. (cf. 140.)

[Spinosa and Carvalho, 2005] Spinosa, E. and Carvalho, A. (2005). Support vector machines for novel class detection in bioinformatics. *Genet Mol Res*, 4(3) :608–15. (cf. 110.)

[Srinivasan and Shobha, 2008] Srinivasan, G. and Shobha, G. (2008). Statistical texture analysis. In *Proceedings of world academy of science, engineering and technology*, volume 36, pages 1264–1269. (cf. 18.)

[Stone, 1977] Stone, C. (1977). Consistent nonparametric regression. *The annals of statistics*, 5(4) :595–620. (cf. 30.)

[Suard, 2006] Suard, F. (2006). *Methodes a noyaux pour la detection de pietons*. PhD thesis, LITIS - INSA de Rouen. (cf. 17.)

[Suard et al., 2005] Suard, F., Guigue, V., Rakotomamonjy, A., and Benshrair, A. (2005). Pedestrian detection using stereo-vision and graph kernels. In *IEEE Intelligent Vehicles Symposium, 2005. Proceedings*, pages 267–272. (cf. 17.)

[Suard et al., 2006] Suard, F., Rakotomamonjy, A., Bensrhair, A., and Broggi, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients. pages 206–212. (cf. 17.)

[Tarassenko et al., 2009] Tarassenko, L., Clifton, D., Bannister, P., King, S., and King, D. (2009). Novelty detection. *Encyclopedia of Structural Health Monitoring*. (cf. 74, 76, and 77.)

[Tarassenko et al., 1995] Tarassenko, L., Hayton, P., Cerneaz, N., and Brady, M. (1995). Novelty detection for the identification of masses in mammograms. In *Fourth International Conference on Artificial Neural Networks*, pages 442–447. (cf. 73 and 77.)

[Tax, 2001] Tax, D. (2001). *One-class Classification, concept-learning in the absence of counter-examples*. PhD thesis, Delft University of Technology, ACSI Dissertation. (cf. 110.)

[Tax, 2005] Tax, D. (2005). Ddtools, the data description toolbox for matlab. *Delft University of Technology ed*. (cf. 98, 99, 106, and 114.)

[Tax and Duin, 1998] Tax, D. and Duin, R. (1998). Outlier detection using classifier instability. *Lecture Notes in Computer Science*, pages 593–601. (cf. 76 and 85.)

[Tax and Duin, 1999] Tax, D. and Duin, R. (1999). Support vector domain description. *Pattern Recognition Letters*, 20(11-13) :1191–1199. (cf. 73, 74, 75, 76, 80, and 85.)

[Tax and Duin, 2001] Tax, D. and Duin, R. (2001). Combining one-class classifiers. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, pages 299–308. Springer. (cf. 84, 85, and 111.)

[Tax and Duin, 2002] Tax, D. and Duin, R. (2002). Uniform object generation for optimizing one-class classifiers. *The Journal of Machine Learning Research*, 2 :155–173. (cf. 76, 81, 99, 110, and 115.)

[Tax and Duin, 2004] Tax, D. and Duin, R. (2004). Support vector data description. *Machine learning*, 54(1) :45–66. (cf. 72, 74, 75, 78, 79, 80, and 83.)

[Tax and Duin, 2005] Tax, D. and Duin, R. (2005). Characterizing one-class datasets. In *Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 21–26. (cf. 123.)

[Tax and Juszczak, 2002] Tax, D. and Juszczak, P. (2002). Kernel whitening for one-class classification. *Pattern Recognition with Support Vector Machines*, pages 855–873. (cf. 85.)

[Tax et al., 1999] Tax, D., Ypma, A., and Duin, R. (1999). Support vector data description applied to machine vibration analysis. In *Proc. 5th Annual Conference of the Advanced School for Computing and Imaging (Heijen, NL*. Citeseer. (cf. 73 and 110.)

[Tax and Duin, 2000] Tax, D. M. and Duin, R. P. (2000). Data description in subspaces. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 672–675. IEEE. (cf. 79.)

[Taylor and Addison, 2000] Taylor, O. and Addison, D. (2000). Novelty detection using neural network technology. In *COMADEM 2000 : 13 th International Congress on Condition Monitoring and Diagnostic Engineering Management*, pages 731–743. (cf. 73 and 74.)

[Thiberville et al., 2007a] Thiberville, L., G.Bourg-Heckly, M. Salaün, S. D., and Moreno-Swirc, S. (2007a). Human in-vivo confocal microscopic imaging of the distal bronchioles and alveoli. *Chest Journal*, 132(4) :426. (cf. 10.)

[Thiberville et al., 2007b] Thiberville, L., Moreno-Swirc, S., Vercauteren, T., Peltier, E., Cave, C., and Bourg-Heckly, G. (2007b). In vivo imaging of the bronchial wall microstructure using fibered confocal fluorescence microscopy. *American Journal of Respiratory and Critical Care Medicine*, 175 :pp. 178–187. (cf. 12.)

[Tian and Gu, 2010] Tian, J. and Gu, H. (2010). Anomaly detection combining one-class svms and particle swarm optimization algorithms. *Nonlinear Dynamics*, pages 1–8. (cf. 110.)

[Tian et al., 2002] Tian, Y., Kanade, T., and Cohn, J. (2002). Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 229–234. IEEE. (cf. 21.)

[Toivola et al., 2010] Toivola, J., Prada, M., and Hollmén, J. (2010). Novelty detection in projected spaces for structural health monitoring. *Advances in Intelligent Data Analysis IX*, pages 208–219. (cf. 74.)

[Unser, 1995] Unser, M. (1995). Texture classification and segmentation using wavelet frames. *Image Processing, IEEE Transactions on*, 4(11) :1549–1560. (cf. 20 and 21.)

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley Interscience. (cf. 30, 34, 80, and 83.)

[Vedaldi and Fulkerson, 2008] Vedaldi, A. and Fulkerson, B. (2008). VLFeat : An open and portable library of computer vision algorithms. (cf. 49.)

[Verleysen et al., 2003] Verleysen, M. et al. (2003). Learning high-dimensional data. *Nato Science Series III, Computer And Systems Sciences*, 186 :141–162. (cf. 75, 78, 85, and 123.)

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In IEEE Conference on Computer Vision and Pattern Recognition*, 1 :511–518. (cf. 17, 24, and 38.)

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*. (cf. 17 and 38.)

[Von Eye and Bogat, 2004] Von Eye, A. and Bogat, G. (2004). Testing the assumption of multivariate normality. *Psychology Science*, 46 :243–258. (cf. 122, 142, and 144.)

[Wang et al., 2009] Wang, C.-K., Ting, Y., Liu, Y.-H., and Hariyanto, G. (2009). A novel approach to generate artificial outliers for support vector data description. In *Industrial Electronics, 2009. ISIE 2009. IEEE International Symposium on*, pages 2202–2207. IEEE. (cf. 81.)

[Wehenkel, 1998] Wehenkel, L. (1998). *Automatic learning techniques in power systems*. Kluwer Academic Publishers. (cf. 41.)

[Wolpert and Macready, 1997] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1) :67–82. (cf. 29.)

[Wu et al., 2008] Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1) :1–37. (cf. 29.)

[Yousef et al., 2010] Yousef, M., Najami, N., and Khalifa, W. (2010). A comparison study between one-class and two-class machine learning for microrna target detection. *Journal of Biomedical Science and Engineering*. (cf. 112.)

[Ypma and Duin, 1998] Ypma, A. and Duin, R. (1998). Support objects for domain approximation. In *ICANN*, volume 98, pages 719–724. Citeseer. (cf. 85.)

[Zane et al., 2002] Zane, O. R., Antonie, M.-L., and Coman, A. (2002). Mammography classification by an association rule-based classifier. In Simoff, S. J., Djeraba, C., and Zane, O. R., editors, *MDM/KDD*, pages 62–69. University of Alberta. (cf. 17.)

[Zeng et al., 2006] Zeng, Z., Fu, Y., Roisman, G., Wen, Z., Hu, Y., and Huang, T. (2006). Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5) :1–8. (cf. 73.)

[Zhang et al., 2007] Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories : A comprehensive study. *International Journal of Computer Vision*, 73(2) :213–238. (cf. 18.)

[Zhang et al., 1998] Zhang, Z., Lyons, M., Schuster, M., and Akamatsu, S. (1998). Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multilayer perceptron. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 454–459. IEEE. (cf. 20 and 21.)

[Zheng, 1993] Zheng, Z. (1993). A benchmark for classifier learning. In *Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence*, pages 281–286. Citeseer. (cf. 112.)