Classification partiellement supervisée par SVM : application à la détection d'événements en surveillance audio

Sébastien Lecomte

Thèse
de doctorat
de l'UTT

Sébastien LECOMTE

Classification partiellement supervisée par SVM. Application à la détection d'événements en surveillance audio

Spécialité :
Optimisation et Sûreté des Systèmes

2013TROY0031 **Année 2013**

utt
université de technologie
Troyes

THESE

pour l'obtention du grade de

DOCTEUR de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

Spécialité : OPTIMISATION ET SURETE DES SYSTEMES

présentée et soutenue par

Sébastien LECOMTE

le 9 décembre 2013

Classification partiellement supervisée par SVM. Application à la détection d'évènements en surveillance audio

JURY

M. C. JUTTEN	PROFESSEUR DES UNIVERSITES	Président
M. S. AMBELLOUIS	CHARGE DE RECHERCHE IFFSTAR	Examinateur
M. F. BIMBOT	DIRECTEUR DE RECHERCHE CNRS	Rapporteur
M. S. CANU	PROFESSEUR DES UNIVERSITES	Rapporteur
M. F. CAPMAN	DOCTEUR	Examinateur
M. R. LENGELLÉ	PROFESSEUR DES UNIVERSITES	Directeur de thèse
M. C. RICHARD	PROFESSEUR DES UNIVERSITES	Directeur de thèse

Mis en page avec la classe thloria.

Résumé

Cette thèse s'intéresse aux méthodes de classification Machines à Vecteurs de Support partiellement supervisées permettant la détection de nouveauté (SVM 1-classe). Celles-ci ont été étudiées dans le but de réaliser la détection d'événements audio anormaux pour la surveillance d'infrastructures publiques, en particulier dans les transports. Dans ce contexte, l'hypothèse « ambiance normale » est relativement bien connue (même si les signaux correspondants peuvent être très non stationnaires). En revanche, tout signal « anormal » doit pouvoir être détecté et, si possible, regroupé avec les signaux de même nature. Ainsi, un système de référence s'appuyant sur une modélisation unique de l'ambiance normale est présenté, puis nous proposons d'utiliser plusieurs SVM de type 1-classe mis en concurrence. La masse de données à traiter a impliqué l'étude de solveurs adaptés à ces problèmes. Les algorithmes devant fonctionner en temps réel, nous avons également investi le terrain de l'algorithmie pour proposer des solveurs capables de démarrer à chaud. Par l'étude de ces solveurs, nous proposons une formulation unifiée des problèmes à une et deux classes, avec et sans biais. Les approches proposées ont été validées sur un ensemble de signaux réels. Par ailleurs, un démonstrateur intégrant la détection d'événements anormaux pour la surveillance de station de métro en temps réel a également été présenté dans le cadre du projet Européen VANAHEIM.

Mots-clés: Traitement du signal - Analyse discriminante - Machines à vecteurs de support - Surveillance électronique

Abstract

This thesis addresses partially supervised Support Vector Machines for novelty detection (One-Class SVM). These have been studied to design abnormal audio events detection for supervision of public infrastructures, in particular public transportation environments. In this context, the null hypothesis ("normal" audio signals) is relatively well known (even though corresponding signals can be notably non stationary). Conversely, every "abnormal" signal should be detected and, if possible, clustered with similar signals. Thus, a reference system based on a single model of normal signals is presented, then we propose to use several concurrent One-Class SVM to cluster observed signals. Regarding the amount of data to process, special solvers have been studied. The proposed algorithms must be real time. This is the reason why we have also investigated algorithms with warm start capabilities. By the study of these algorithms, we have proposed a unified framework for One-Class and Binary SVM, with and without bias. The proposed approach has been validated on a database of real signals. The whole process applied to the monitoring of a subway station has been presented during the final review of the European Project VANAHEIM.

Keywords: Signal Processing - Discriminant Analysis - Support Vector Machines - Electronic Surveillance

Remerciements

Cette thèse n'aurait pas été ce qu'elle est sans le support, le soutien, l'amitié de nombreuses personnes. Qu'il ait s'agit de soigner la formule, d'orienter les recherches, de corriger l'orthographe, d'éveiller la curiosité ou simplement d'échanger, j'espère que chacun saura se retrouver dans ces quelques lignes.

Je remercie Christian JUTTEN, Professeur des Universités à l'Université Joseph Fourier de Grenoble, qui m'a fait l'honneur de présider mon jury de thèse. Je remercie Frédéric BIMBOT, Directeur de Recherche à l'IRISA Rennes, et Stéphane CANU, Professeur des Universités à l'INSA Rouen, pour avoir accepté de rapporter sur ce travail. Je remercie également chaleureusement Sébastien AMBELLOUIS d'avoir accepté d'examiner mes travaux.

J'adresse également mes sincères remerciements à Régis LENGELLE, Professeur des Universités à l'Université de Technologie de Troyes, et à Cédric RICHARD, Professeur des Universités à l'Université de Nice Sophia-Antipolis, pour m'avoir accompagné au cours de ces trois années de thèse. L'aventure a été belle et c'est sans nul doute grâce à vous, et vos remarques justes (même parfois moqueuses).

Merci mille fois également à François CAPMAN, Docteur chez Thales Communications & Security, sans qui tout cela n'aurait pas été possible. Merci à lui pour toutes ses idées qui auraient pu faire durer cette thèse dix ans de plus.

Je tiens également à remercier Bruno SOURDILLAT, responsable du Service Multi-Media Processing chez Thales Communications & Security, qui m'a accueilli au sein de son équipe depuis près de cinq ans maintenant.

Je salue et remercie tout particulièrement Bertrand RAVERA (Thales Communications & Security) et Paul HONEINE (Université de Technologie de Troyes) pour toutes les discussions que nous avons pu avoir, pour leurs idées, leur humeur, leur sympathie et leur amitié.

Je tiens ensuite à remercier tous mes compagnons de parcours à l'UTT : Maya, Zineb, Rémi, Cathel, et tous les autres. Un merci ému à ceux avec qui j'ai pu partager un bureau chez THALES (quand bureaux il y avait) : Benjamin, Cyril, Alexandre et Charles, et la moitié d'un Marc. Merci également à tous mes collègues pour leur sympathie : Rachid, Séverin, Yannick et Benoît. Je remercie également ceux que j'ai côtoyés avant qu'ils choisissent d'autres horizons : Didier, Marc et Erwann.

Merci également à tous ceux qui ont contribué à faire de ces trois années une aventure humaine. Je pense aux camarades Thales avec j'ai partagé des moments de modélisme ou d'œnologie, et qui ont contribué à développer mon envie de rester dans cette entreprise. Je pense aussi à toutes les personnes de l'administration de l'UTT qui me manqueront après ces dix années passées ensemble, je reviendrai !

Merci à ma famille, mes parents, mes amis, pour leurs encouragements et avoir supporté mes humeurs.

Merci à Vanessa pour qui il n'y a pas de mots assez justes pour qualifier tout ce qu'elle m'a apporté...

Avant-propos

Cette thèse a été réalisée avec le support de l'Agence Nationale pour la Recherche et la Technologie (ANRT) dans le cadre d'un contrat CIFRE (n°970/2009). Le service Multi-Media Processing (MMP) de Thales Communications & Security a accueilli le doctorant pour la réalisation des travaux. L'encadrement académique a été assuré par l'Institut Charles Delaunay de l'Université de Technologie de Troyes (UTT - ICD) au sein du Laboratoire de Modélisation et Sûreté des Systèmes.

Par ailleurs, les travaux ont bénéficié du soutien du projet collaboratif VANAHEIM (Video/Audio Networked surveillance system enhAncement through Human-cEntered adaptIve Monitoring, programme Européen FP7/2007-2013), qui s'intéresse aux composants innovants pour le contrôle autonome d'infrastructures complexes de surveillance audio/vidéo, telles que celles mises en place aujourd'hui dans les stations de métro.

Table des matières

Table des figures

Liste des tableaux

Introduction

Nos travaux s'intéressent à l'utilisation de la modalité audio pour la surveillance. Ce chapitre introductif débute donc par un rapide exposé sur les enjeux de la relation entre surveillance et modalité audio. Dans ce contexte, nous poursuivons par une section destinée à positionner les travaux réalisés par rapport à l'état de l'art et limiter le périmètre de nos recherches. Enfin, nous présentons la structure de ce manuscrit, indiquant en particulier les contributions principales de ces travaux.

1 Surveillance et modalité audio

1.1 Vidéo-surveillance : historique et enjeux

Dans les années 1970, le Royaume-Uni installe pour la première fois des caméras dans des lieux publics ou privés afin de visualiser à distance et d'enregistrer l'activité qui s'y déroule. L'objectif étant de lutter contre les attentats éventuels de l'IRA, les caméras stratégiquement placées permettent de surveiller et d'identifier des individus, d'enregistrer des preuves, et jouent également un rôle dissuasif. Encouragé par de bons résultats, le nombre de caméras augmente rapidement et la vidéo-surveillance se développe dans d'autres pays. Initialement utilisé pour la défense, l'outil évolue et les polices, les régies de transports et les municipalités développent leurs propres réseaux.

Depuis les années 90, des millions de caméras ont été déployés dans les villes, sur les routes et dans les transports. La vidéo surveillance s'est aussi développée dans les immeubles, les parkings, les magasins, et même dans la sphère privée. Cette explosion de caméras est en partie due à la chute du coût des réseaux et des capteurs. Cependant, ces systèmes de première génération [VV05] reposent sur des technologies analogiques, et le stockage et l'analyse des séquences vidéos est coûteux. Plus récemment l'arrivée de systèmes « tout numérique », dont le déploiement et la gestion à distance sont facilités, a encore accru le déploiement de la vidéo-surveillance.

A titre d'exemple, en 2007, les Etats-Unis comptaient plus de 30 millions de caméras de surveillance, générant près de 4 milliard de séquences vidéo par semaine [Vla08]. En France, le marché de la vidéo-surveillance est en croissance de près de 10% par an depuis 2003 (+600% pour le tout-numérique entre 2003 et 2007), et les transports occupent près 25% de parts de marché [MSI08]. On observe également une demande importante de la part des pays émergeants, en particulier les BRIC (Brésil, Russie, Inde, Chine), et les compétitions de dimension mondiale (Jeux Olympiques, Coupes du Monde, etc.) appuient en grande partie leur dispositif sécuritaire sur la surveillance vidéo. Enfin, le succès du projet récent *Ciudad Segura* à Mexico [Ciu12] témoigne de l'efficacité de tels systèmes.

Cependant, face à l'augmentation constante du nombre de caméras, un problème d'exploitation s'est posé. Noyes et Bransby [NB01] notent que l'attention d'un opérateur décroit avec l'augmentation du nombre d'écrans à contrôler ; les performances de détection passant pour un

opérateur de 85% pour 1 écran à 58% pour 6 écrans. De plus, la vigilance des opérateurs n'est pas constante. Les recherches montrent qu'une période de pleine attention ne dure jamais plus de 25 à 30 minutes [NB01]. Cela nécessite alors de mettre en place des relais entre opérateurs et de diversifier leurs tâches.

Ce constat justifie depuis les années 90 les efforts de recherche pour développer des systèmes capables d'automatiser une partie de la tâche de supervision. Les systèmes de surveillance de seconde génération [VV05] (ou tout numérique) sont ainsi caractérisés par une forte interaction avec les technologies de vision par ordinateur (*computer vision*). L'objectif de ces travaux n'est évidemment pas de substituer une machine au jugement d'un opérateur, mais de tout mettre en œuvre pour conserver un niveau de vigilance élevé. Par exemple, l'analyse automatique permet de limiter le nombre d'écrans en n'affichant que les scènes où se déroule une activité. Citons également la possibilité d'automatiser l'analyse des comportements afin d'attirer l'œil de l'opérateur sur des situations précises.

Ce développement s'est néanmoins accompagné d'un besoin constant de nouvelles méthodes afin de résoudre des problématiques liées à des changements d'angle de vue, de conditions d'éclairage ou de place d'objets, à l'apparition d'éléments temporaires dans les scènes, etc. Il existe aujourd'hui des systèmes de vidéo-surveillance avec analyse automatisée, mais coûteux, contraints et à l'efficacité limitée. Ainsi, ces systèmes sont souvent spécifiques à une tâche précise dans un environnement donné.

1.2 Modalité audio pour la surveillance

Les efforts de recherche actuels se concentrent sur l'intégration de capteurs complémentaires (caméras thermiques ou infra-rouges, détecteurs de mouvements, micros, capteurs sismiques, etc.) afin de développer une vigilance automatisée au sein des systèmes de surveillance. Ce sont les systèmes de surveillance de troisième génération [VV05]. La multiplication des capteurs permet de diversifier l'information disponible et de renforcer la qualité de la détection des situations anormales, y compris dans de mauvaises conditions d'observation. La modalité audio complète efficacement la vidéo car elle permet de capturer des événements qui se déroulent en dehors du champ des caméras, ou lorsque les conditions sont défavorables (luminosité trop faible par exemple). L'audio est également efficace pour décrire les activités humaines et les liens sociaux (discussion, cris, etc.).

De nombreux moyens ont été mis en œuvre pour automatiser l'analyse des signaux audio-numériques depuis leur apparition et l'essor qui s'en est suivi au cours des années 90. En effet, face à la quantité d'information croissante, la nécessité de disposer de traitements ne nécessitant pas d'intervention humaine s'est faite, comme pour la vidéo, de plus en plus présente. D'abord motivée par les besoins de recherche documentaire [CLZC03, CLH$^+$06] et d'analyse musicale [TC02, AP03], la caractérisation de signaux audio voit également son intérêt dans des applications d'identification, de surveillance et de sécurité.

Aujourd'hui, le domaine de l'analyse audio pour la surveillance est en pleine croissance. Les applications sont variées, comme en témoigne une littérature riche : reconnaissance d'événements donnés [CRE05, AMK06, VGT$^+$07], analyse des nuisances urbaines [Def05], télé-surveillance médicale [VIB$^+$04, ICV$^+$06], suivi de réunions [Tem07], surveillance des ascenseurs [KK11], reconnaissance d'animaux [GR10], suivi de trajectoires [ZDD01], reconnaissance des émotions pour les centres d'appel [CDR$^+$07], reconnaissance de locuteur [PB06], ou encore la surveillance des transports en commun [VBD$^+$06, RLA06, Yan09, PLB$^+$10].

De nombreux projets de recherche ont également été menés, visant, sur la base des signaux audio, à développer des algorithmes d'aide aux opérateurs pour identifier et sauvegarder l'in-

formation essentielle issue de capteurs de surveillance. Parmi les projets ayant trait au domaine des transports publics, qui nous intéresse plus particulièrement, nous pouvons citer PRISMATICA [PPaIsfSMbTA03], SAMSIT [VBD+06], SERKET [CE08], CARETAKER [CDR+06], SURTRAIN [Hee09] ou encore VANAHEIM (voir annexe B).

2 Positionnement des travaux

Nous l'avons vu, la détection et la description des activités dans un flux sonore est une fonctionnalité clé des systèmes de surveillance actuels et futurs. Dans ce contexte, on distingue deux familles de systèmes :

- les systèmes ayant vocation à décrire un environnement sonore, largement regroupés sous l'appellation *Auditory Scene Analysis* [Bre90], ont pour objectif de caractériser une zone surveillée (équipée de micros) à partir de l'observation d'un ensemble d'événements. La caractérisation dépend de l'application et des attentes opérationnelles : présence ou non d'activité en ville, reconnaissance d'un environnement à partir d'un enregistrement (urbain, forêt, plage, etc.), etc.
- les systèmes ayant vocation à identifier des événements sonores particuliers ont pour objectif, quel que soit l'environnement, de détecter la présence d'un son. Là encore, suivant les besoins opérationnels, la nature des événements détectés varie : tirs d'armes, défaillance d'un matériel, chutes, cris, etc.

De tels systèmes reposent sur deux tâches essentielles : la détection et la classification des événements sonores. La classification consiste à donner un sens à un événement qui a été préalablement isolé temporellement. La détection fait en revanche référence à l'identification et la localisation temporelle d'un événement dans un signal audio. Nous abordons principalement la thématique de la détection dans cette thèse.

2.1 Discussion sur les travaux antérieurs

La plupart des systèmes de détection d'événements sonores proposés dans la littérature sont supervisés. Ainsi, ils s'appuient sur un dictionnaire d'événements connus *a priori* et dont les modèles ont été préalablement appris. Cette approche présente notamment les contraintes suivantes :

- Il est nécessaire de spécifier au préalable l'ensemble des événements anormaux à détecter et de collecter une quantité suffisante de données représentatives de ces évènements.
- La correspondance entre les conditions d'enregistrement du dictionnaire et les conditions réelles influence en partie les performances du système, nécessitant de nouveaux signaux d'apprentissage à chaque déploiement.
- Les événements ne faisant pas partie du dictionnaire ne sont pas détectés.
- Certaines anormalités sont découvertes lorsqu'elles se produisent une fois et ne se reproduisent jamais ; elles ne peuvent donc pas être ajoutées au dictionnaire.
- Les approches supervisées souffrent de l'impossibilité de détecter des événements anormaux se produisant simultanément [HMEV13].

D'autre part, l'environnement (bruit ou ambiance) n'est généralement pas pris en compte ; ni sa nature, ni son évolution dans le temps. L'évaluation d'un système de détection repose ainsi souvent sur une modélisation de l'environnement par un bruit blanc à différents niveaux de rapport signal à bruit, non représentatif de la diversité des événements sonores dans des environnements réels. En effet, l'ambiance normale des environnements que nous considérons

(transports, milieux urbains, etc.) est un continuum non-stationnaire qui inclut des événements sonores considérés comme normaux (bruits ambiants récurrents).

Ces événements normaux évoluent également au cours du temps (heures pleines ou creuses, jour ou nuit, semaine ou weekend, etc.). Des approches récentes inspirées de la modalité vidéo [BGG09, KKC+09] permettent de spécifier les modèles d'événements en les entrainant pour un environnement donné [CLH+06]. Néanmoins, ces approches restent supervisées et la flexibilité de ces solutions est limitée par la connaissance nécessaire *a priori* de l'environnement et des événements.

Enfin, Valera et Velastin [VV05] considèrent que la nécessité d'un minimum d'information est un défi pour construire des systèmes de surveillance automatisés plus efficaces et plus intelligents. L'exploitation d'un dictionnaire prédéfini ou la conception de systèmes spécifiques à des environnements donnés n'est donc pas souhaitable. Nous considérons alors que les approches évoquées ci-dessus ne sont pas appropriées au contexte de la surveillance. Aussi, nous nous intéressons au développement de systèmes de détection non supervisés, dont l'entrainement ne repose sur aucune information préalable concernant les événements à détecter ou l'environnement d'exploitation.

2.2 Evénement sonore, normal ou anormal

On se place dans le contexte de la détection non supervisée d'événements sonores, normaux ou anormaux. En l'absence d'*a priori* sur les événements, nous définissons maintenant ce que nous allons chercher à détecter.

L'apprentissage consiste à réaliser un modèle de l'environnement sonore sous surveillance à partir de signaux enregistrés *in situ*, les signaux d'apprentissage. L'objectif est ensuite d'être capable de classer de nouveaux signaux dans l'une des catégories suivantes (des exemples de sons sont donnés pour le contexte de la surveillance d'un quai de station de métro) :

- « signal normal » : le signal correspond à l'ambiance sonore de l'environnement (par exemple : arrivées/départs de train, ventilations, discussions entre passagers, avertissements sonore de la fermeture des portes, annonces de service...),
- « signal anormal » : le signal correspond à un événement sonore non habituel pour l'ambiance normale (par exemple : coups de feu, bagarres, cris, vandalisme, bris de vitres, animaux, chahuts d'enfants...) ou à une ambiance anormale (mouvements de foule, pannes des ventilations...).

Nous faisons l'hypothèse qu'une grande quantité de signaux, représentative de l'environnement de déploiement, est disponible. On considère ensuite qu'un événement anormal est un événement qui se produit peu voire jamais. A l'inverse, nous considérons que les événements se produisant souvent, et donc largement représentés dans les signaux d'apprentissage, sont normaux. L'objectif de l'apprentissage est de déterminer un modèle correspondant à l'ensemble de ces signaux fréquents.

Ainsi, cette approche va considérer comme anormaux des événements normaux mais rares. Ces fausses alarmes, rares par définition, sont opérationnellement acceptables. Elles pourront éventuellement faire l'objet de modèles spécifiques lors d'un post-traitement des événements détectés destiné à les classifier. Elles seront alors détectées comme anormales, mais classées comme un événement ne nécessitant pas d'attention particulière de l'opérateur. On peut citer par exemple des opérations périodiques de maintenance nécessitant des machines particulières.

De façon opposée, cette approche considérera également comme normaux des événements anormaux mais fréquents. Là encore, une parade simple permettra d'éliminer ces ratés de détection en utilisant un système de détection supervisé en pré-traitement. L'événement anormal

étant par définition fréquent, il est possible d'obtenir une grande quantité de signaux d'apprentissage dans l'environnement et on s'appuie alors sur des méthodes de l'état de l'art. Néanmoins, notons qu'un événement anormal fréquent sera opérationnellement identifié par d'autres moyens que la surveillance audio, et le plus fréquemment traité pour ne plus se reproduire.

	Evénement normal	Evénement anormal
Evénement rare	Fausse alarme acceptable	Correctement classifié
Evénement fréquent	Correctement classifié	Ignoré, nécessite un détecteur spécifique

TABLE 1 – Classification des événements par le système visé

2.3 Objectifs

Les travaux présentés dans ce manuscrit visent l'étude d'un système non supervisé de détection d'événements anormaux. Celui-ci s'appuie sur la recherche, dans des signaux de test, des événements déviant du modèle d'ambiance « normale » appris par le système. L'apprentissage de ce modèle doit être indépendant de l'environnement et sans *a priori* sur les signaux anormaux, ne s'appuyant que sur des enregistrements *in situ*, et sans expertise préalable de ceux-ci.

En particulier, nous avons identifié cinq axes d'étude que nous explicitons dans les paragraphes suivants. Ceux-ci dénotent de la volonté de réaliser un système industrialisable à moyen terme [1]. Notons que l'ensemble des contraintes opérationnelles ne peuvent être satisfaites à l'issue des travaux dont la durée est limitée. Néanmoins, elles ont été prises en compte dans le développement du système proposé.

Le premier point concerne la possibilité pour Thales d'envisager un produit sur la base de ces travaux. Pour cela, l'apprentissage d'un modèle d'ambiance doit pouvoir être réalisé dans un temps raisonnable. De plus, le système doit avoir la capacité à suivre les évolutions de l'environnement (par exemple : surveillance d'une station métro aux heures creuses et aux heures de pointe). S'appuyant sur l'expertise du laboratoire, nous avons exploré des méthodes basées sur les machines à vecteurs de support (SVM). Ce choix a également été motivé par de récents travaux concernant d'une part des algorithmes rapides de résolution, et d'autre part des approches avec démarrage à chaud permettant de mettre à jour un modèle existant.

Ensuite, puisque le système de détection d'événements sonores anormaux ciblé est construit sans aucun *a priori* sur la nature des événements à détecter, nous nous intéressons également à la possibilité de détecter des événements de durées variées. Pour cela, l'analyse de signaux de test doit pouvoir bénéficier d'informations de segmentation automatique reflétant la structure temporelle du signal en événements successifs, de longueur variable.

Un autre besoin opérationnel pour l'industriel est la minimisation de l'expertise nécessaire lors de la mise en œuvre du système dans un nouvel environnement ; ceci afin de limiter les interventions auprès des clients et accroitre la capacité de ce dernier à redéployer le système (cas de surveillances itinérantes, camps ou chantiers). Pour cela, la possibilité d'automatiser la phase de configuration du système, lorsqu'il est placé dans son environnement d'utilisation, a motivé les choix techniques et l'orientation des travaux. Ceci concerne également la sélection des grandeurs représentant le signal audio, ou paramètres acoustiques.

1. Un démonstrateur des travaux a notamment pu être réalisé dans le cadre du projet VANAHEIM (voir B).

Le quatrième axe d'étude retenu est la gestion du compromis entre fausses alarmes et ratés de détection. En effet, l'équilibre entre ces deux erreurs est primordial pour que le système de détection soit opérationnellement accepté. Un taux de fausses alarmes trop élevé et le système sera ignoré par les opérateurs. Réciproquement, un taux de ratés de détection trop élevé et le système ne sera plus pertinent. Une attention particulière a notamment été portée aux méthodes d'évaluation afin de pouvoir quantifier ces erreurs.

Enfin, la normalité ou l'anormalité d'une situation peut ne pas résider uniquement dans la nature des événements sonores. En effet, un ensemble d'événements normaux se produisant de façon inhabituelle peut révéler un problème dans l'environnement sous surveillance. A titre d'exemple, toujours dans le cas d'une station de métro, il peut s'agir d'une rame qui reste trop longtemps à quai. Bien que ce type d'analyse ne soit pas traité au cours des travaux présentés, il s'agit d'une perspective en rupture avec les approches actuelles. Ainsi, nous avons étudié au travers de la modélisation non supervisée (*clustering*) des signaux d'apprentissage la possibilité d'identifier automatiquement les différentes classes d'événements normaux sous-jacentes.

3 Structure du document et résumé des contributions

La première partie de ce document expose quelques généralités concernant les domaines au confluent desquels s'inscrivent ces travaux. Ainsi, le chapitre premier donne un aperçu des méthodes couramment utilisées dans le contexte de l'analyse du signal audio. N'ayant pas d'*a priori* sur les signaux que nous traitons, nous veillons néanmoins à ce que ce chapitre reste le plus générique possible. Le chapitre suivant s'intéresse quant à lui à la problématique de l'apprentissage statistique. Cette discipline est vaste et nous en clarifions les éléments essentiels pour la suite de l'exposé.

La seconde partie du manuscrit présente notre modèle de détection. Le troisième chapitre introduit alors les méthodes SVM sur lesquelles nous nous appuyons : SVM 2-classes, avec ou sans biais et SVM 1-classe. Nous décrivons également dans ce chapitre une approche SVM 1-classe sans biais [LLR$^+$13]. Nous montrons que cette dernière conduit à un problème dual similaire au problème dual SVM 2-classe sans biais ; une écriture unifiée est formulée. Le chapitre quatre propose un algorithme générique de résolution pour l'ensemble des problèmes SVM introduits : 1 ou 2 classes, avec ou sans biais.

Nous consacrons la troisième partie du document à la présentation de nos résultats. Le cinquième chapitre expose notre protocole d'évaluation, spécifique au contexte audio-surveillance. Ce chapitre est l'occasion d'introduire un outil pour simuler des signaux audio destinés à l'évaluation d'un système de détection d'événements anormaux [Tha11, LLR$^+$11]. Le sixième chapitre est consacré à l'exposé des résultats pour différents détecteur. C'est pour nous l'occasion d'évoquer une procédure opérationnellement acceptable pour déterminer rapidement les paramètres du système. Le dernier chapitre de cette partie est consacré aux évaluations de l'approche sans biais proposée, d'abord sur des bases de données génériques puis sur des signaux audio.

La dernière partie du manuscrit est consacrée à une méthode de détection basée sur un regroupement (*clustering*) des observations par des modèles SVM 1-classe. Le chapitre huit expose l'approche qui a par ailleurs fait l'objet d'un dépôt de brevet [LCR$^+$12]. Puis, le chapitre neuf présente les résultats obtenus à l'aide d'une version simplifiée de la méthode. Ceux-ci permettent d'établir quelques perspectives parmi lesquelles la possibilité d'appliquer l'approche à une tâche d'audio-surveillance.

Nous concluons ce document par un bilan et l'exposé de perspectives pour des travaux futurs. Nous évoquons également les connaissances acquises concernant la problématique de

l'exploitation de la modalité audio pour la surveillance.

Cette thèse ayant également été l'occasion de collecter des signaux en environnement réel, une annexe est consacrée à l'étude d'une piste n'ayant pas donné satisfaction pour l'acquisition de signaux anormaux [Lec12]. Une seconde annexe est consacrée à la présentation du démonstrateur réalisé dans le cadre du projet VANAHEIM qui inclut le modèle de détection d'événements audio anormaux proposé dans la partie deux. L'annexe trois liste les événements anormaux considérés lors des évaluations. L'annexe quatre présente la méthode utilisée pour la segmentation des signaux audio. Enfin, la dernière annexe expose les résultats de notre approche comparée à une méthode de référence s'appuyant sur une modélisation GMM et développée par Thales [CR10, Tha11].

Première partie

Généralités

Chapitre 1

L'analyse numérique du signal sonore

Dans ce chapitre, nous rappelons dans un premier temps la manière dont le son est numérisé. Puis, nous nous intéressons à décrire la représentation du signal sonore dans un espace d'observation ou espace de paramètres acoustiques. Nous abordons ensuite le problème de la sélection automatique de ces paramètres. Enfin, nous décrivons et justifions l'approche que nous avons retenue au travers de cette thèse.

1.1 Du son au signal audio-numérique

1.1.1 Le monde des sons

Historiquement, l'intérêt pour les sons a été porté par la volonté de comprendre et maîtriser les phénomènes sonores dans le contexte de la musique. Les pythagoriciens se sont intéressés à la vibration des cordes puis Aristote, entre autres, au comportement ondulatoire des ondes. Très tôt donc les sons ont été associés à une onde, bien qu'il ait fallu attendre le XIIe siècle et les travaux de Mersenne et Galilée pour que l'acoustique devienne une véritable science. Physiquement, le son est alors décrit comme une onde mécanique de pression se propageant à travers un milieu.

Le Britannique Boyle, le Hollandais Huygens, les Suisses Euler et Bernoulli et le Turinois Lagrange ont contribué aux XIIe et XIIIe siècles à apporter les bases de l'étude des phénomènes sonores. Au XIXe, l'apparition de techniques expérimentales a permis à Regnault puis Helmotz d'améliorer la compréhension de l'acoustique, la perception des sons et l'analyse des sons complexes. Les théories les plus modernes reposent encore sur les résultats de cette époque.

En 1877, le futur prix Nobel lord Rayleigh a figé l'étendue des connaissances acquises jusque-là dans son ouvrage *Theory of Sound* [Ray77]. Rééditée à de nombreuses reprises, cette œuvre en deux volumes reste encore aujourd'hui une référence.

Dans son *Traité d'Acoulogie*[Chi04], Michel Chion, dresse un portrait original d'une discipline nouvelle au confluent de l'acoustique, la musique et la psychologie de l'écoute. A travers une revue de la philosophie, de la littérature, de la musique, du cinéma et des technologies, son essai montre que le monde des sons, souvent méconnu, reflète par bien des aspects le monde qui nous entoure. Au-delà de cette compréhension, la maîtrise technique permet également d'analyser l'environnement à partir des sons, naturels ou provoqués. Ainsi, l'étude des sons permet de traiter nombre de sujets, de la voix humaine au bang sonique en passant par la cartographie acoustique ou la thérapie ultra-sonore [PLS01].

L'analyse sonore est donc une discipline variée. Dans cette thèse, nous nous intéressons à l'analyse d'un environnement sonore afin de qualifier l'activité qui s'y déroule. Comme nous l'avons vu en introduction, nous nous limitons à qualifier comme normaux ou anormaux des

événements capturés en continu par un système. Nous laissons de côté les considérations acoustiques liées à la propagation des sons dans l'environnement. Enfin, nos scénarios excluent la surveillance d'environnements singuliers tels que le milieu marin, le contrôle non invasif ou encore l'écoute à champ ouvert sur de longues distances.

1.1.2 L'analyse du son dans le domaine numérique

Numérisation

L'onde acoustique est transformée en signal électrique par un système d'acquisition (microphone). Afin de permettre son analyse par un ordinateur, ce signal est numérisé via deux traitements distincts : l'échantillonnage et la quantification.

L'échantillonnage est une transformation du signal depuis un espace de temps continu vers un espace de temps discret. Les échantillons sont espacés d'une durée définie par la fréquence d'échantillonnage (par exemple pour les CD audio, celle-ci vaut 44100Hz, soit un échantillon toutes les $22,6\mu s$).

La quantification consiste à représenter la valeur de chaque échantillon dans l'espace des valeurs possibles numériquement. La précision, autrement dit la différence entre la valeur réelle et la valeur quantifiée, dépend des caractéristiques du système (nombre de bits utilisés, représentation en virgule flottante ou fixe, etc.).

Le signal numérique constitue la matière première de notre analyse. Par ailleurs, les signaux utilisés au cours de nos travaux sont échantillonnés sur 16 bits (quantification linéaire) à une fréquence de 16kHz.

Fenêtrage en trames

Afin d'analyser le signal audio numérisé, on définit des fenêtres temporelles de quelques dizaines à quelques centaines d'échantillons (soit de quelques milli-secondes à quelques secondes). Ces fenêtres d'analyse, communément appelées trames (*frames* en anglais), se recouvrent généralement les unes avec les autres et sont pondérées de façon à conserver l'amplitude d'origine aux points de recouvrement. On parle alors d'analyse fenêtrée avec recouvrement ou WOLA (*Windowed Over-Lapping Analysis*). La pondération est habituellement une fenêtre de Hanning ou de Hamming.

Notons que ce découpage peut être échantillonné, c'est à dire que l'on peut ne conserver qu'une partie des trames. La sélection des trames est réalisée soit par expertise humaine, soit par l'application d'algorithmes spécifiques.

Extraction des descripteurs acoustiques

Chaque trame est représentée par un ensemble de grandeurs physiques, calculées à partir des valeurs de ses échantillons. Ces grandeurs, appelées paramètres acoustiques ou descripteurs, rendent compte de l'information qu'elle contient. En particulier, on admet que le signal est stationnaire à l'horizon d'une trame ; le choix d'une fenêtre temporelle adaptée permettant généralement de vérifier cette hypothèse. Par extension, on nomme également trame un vecteur de descripteurs acoustiques, celui-ci constituant la représentation numérique élémentaire du signal sonore.

Le choix des descripteurs conditionne en grande partie les capacités d'un système. Ainsi, nous consacrons dans la suite une section à l'introduction de descripteurs acoustiques couramment utilisés dans la littérature. On peut également ne conserver qu'une partie des descripteurs extraits.

La sélection des paramètres peut, là encore, être réalisée soit par expertise humaine, soit par application d'algorithmes spécifiques. Nous consacrons également une section à ces algorithmes.

Espace d'observation

La décomposition en trames et l'extraction des paramètres permettent de construire une matrice. A chaque trame correspond une ligne de la matrice, tandis que les colonnes contiennent les valeurs prises pour chacun des descripteurs. Il s'agit de la représentation paramétrique du signal. Le contenu de cette matrice peut être représenté dans un espace multi-dimensionnel, appelé espace d'observation, dont chacune des dimensions correspond à un descripteur.

Chaque point ou observation dans l'espace d'observation correspond à une trame de signal. Il est alors possible de considérer des zones plus ou moins denses de l'espace acoustique. Des techniques d'apprentissage vont permettre de modéliser ces zones pour réaliser des tâches de détection ou classification des observations. Seule la représentation paramétrique est transmise aux algorithmes permettant de construire ces modèles.

Segmentation

La segmentation consiste à regrouper les trames par groupes cohérents de taille variable, appelés segments. Chaque segment peut être considéré comme l'ensemble des trames qui le compose (un ensemble de vecteurs de paramètres), ou bien comme le représentant de ces trames (un unique vecteur de paramètres, vecteur moyen ou médian) [2]. A l'issue de cette segmentation, on dispose d'une représentation structurée du signal. On peut alors distinguer des portions de signal contenant *a priori* des informations distinctes les unes des autres. Notons également que dans l'espace de représentation, où chaque trame est représentée par un point, les segments sont soit des points, soit des trajectoires.

La littérature fait état d'un nombre conséquent de méthodes permettant de réaliser une segmentation d'un flux audio. De façon générale, il s'agit d'évaluer un critère de similarité ou de cohérence, puis de regrouper les trames successives suivant ce critère. En particulier, nous pouvons citer l'utilisation de critères d'information, type BIC [CVR05], ou encore l'utilisation de méthodes d'analyse probabiliste [WTM+07, RCL11]. Nous avons cependant utilisé dans ces travaux une approche exploitant la distance euclidienne entre trames, inspirée de la segmentation du signal de parole [GZ88, HL96]. Cette méthode, brevetée par Thales [CR10], est décrite à l'annexe D.

1.2 Descripteurs acoustiques

Le choix de descripteurs acoustiques adaptés pour une tâche donnée permet d'extraire l'information discriminante, et facilite le travail de construction d'un algorithme de détection ou de classification. A l'inverse, un choix de descripteurs inadaptés entraîne un manque d'information qu'il est difficile de compenser lors de la construction d'un tel algorithme.

La littérature recèle de travaux mettant en avant les avantages d'ensembles de descripteurs pour une tâche donnée, voire proposant des descripteurs spécifiques. Nous dressons maintenant une liste des plus populaires pour les tâches de détection et classification de signaux audio. Cette liste, évidemment non exhaustive, reflète la richesse des informations qu'il est possible d'extraire du signal audio. La norme MPEG-7 [IEC, KMS05] inclut un certain nombre de ces descripteur.

2. En particulier, l'approche consistant à représenter le signal audio comme la distribution statistique à long terme des descripteurs locaux est appelée communément *Bag-of-Frames* (sac de trames).

Notons enfin que nous n'évoquons pas les descripteurs issus des techniques de codage, tels que les *Linear Predictive Coding Coefficients* [O'S98].

1.2.1 Domaine temporel

Descripteurs de forme du signal La forme d'onde permet d'obtenir une description compacte du signal audio en considérant son minimum et son maximum au sein de trames sans recouvrement. L'enveloppe temporelle [BL03] est un descripteur plus compact encore de la forme du signal. Celui-ci correspond à la valeur maximum de l'amplitude absolue d'une trame.

Puissance, énergie, intensité Comme souvent en traitement du signal, on s'intéresse dans le cadre de l'audio à l'énergie. Une mesure standard est la puissance sonore. Il s'agit de la puissance moyenne sur l'intervalle de la trame (en Watt ou décibels). L'énergie est simplement la somme des amplitudes présentes dans une trame [DZ07].

La racine des carrés moyens (*Root Mean Square*, RMS) de l'amplitude est parfois utilisée comme mesure de l'intensité sonore [WtBKW96, TC99, WLH00, WTM$^+$07], on parle alors de l'énergie du signal (ou puissance RMS). Dans la littérature, on retrouve également le terme d'énergie moyenne à court terme (*Short-Time Average Energy* [LSDM01]) [3]. L'énergie à court terme est adaptée pour la représentation des variations d'amplitude et permet de distinguer les moments de silence lorsque le rapport signal-bruit est élevé [ZK98].

Taux de passage à zéro Ce descripteur compte le nombre de passages par zéro du signal au sein d'une trame. Cette mesure donne une estimation grossière des propriétés spectrales du signal audio, et est corrélée au centroïde spectral [SS97]. Ce descripteur est efficace dans le domaine de la parole pour différencier la parole voisée de la parole non-voisée [WLH00] ou encore discriminer la musique de la parole [ZK98, KMS05]. Kedem [Ked86] a proposé une analyse complète des propriétés du ZCR et montre qu'il permet une mesure de « fréquence dominante » dans un signal. Ainsi, ce descripteur est corrélé au *pitch* [AP03] et est utile pour mesurer la « quantité de bruit » d'un signal [TEC01, Pee04].

Auto-Corrélation, périodicité L'auto-corrélation permet d'estimer la distribution fréquentielle du signal audio depuis le domaine temporel [Pee04, DZ07]. Cette analyse consiste à multiplier les échantillons d'une trame avec une copie d'eux-mêmes, après avoir appliqué un délai à cette copie. On recherche alors des pics parmi les premiers coefficients ainsi obtenus pour identifier les périodicités du signal. Chaque délai différent permet l'analyse d'une fréquence particulière.

1.2.2 Domaine fréquentiel

La représentation spectrale, ou spectre, d'une trame correspond à sa transformée de Fourier. On utilise cependant rarement le spectre comme un descripteur (vectoriel) en raison de sa grande dimension (généralement sur 512, 1024 ou 2048 coefficients, parfois plus). Celui-ci sert donc de base, comme la trame elle-même dans le domaine temporel, pour l'extraction des descripteurs fréquentiels.

3. Les mesures de volumes relèvent de l'acoustique, donc de la physique, et ne doivent pas être confondue avec les mesures psycho-acoustiques ou psychologiques d'intensité sonore perçue (appelées loudness, mesurées en phone ou sone)[FM33, WtBKW96, Pee04]. Voir 5.2.1

Energie par bancs de filtres L'énergie par bancs de filtres est une approche permettant de synthétiser l'information du spectre au travers d'un nombre restreint de coefficients. Afin de fournir une représentation compacte, on somme les coefficients du spectre par bandes de fréquence. Il est possible d'utiliser un spectre compressé (logarithme ou décibels). De plus, le choix des bancs de filtres est large : les bandes fréquentielles peuvent se superposer partiellement, être espacées linéairement ou selon une échelle perceptive (MEL, Bark, Warp). Enfin, les bandes de fréquence peuvent être fenêtrées de différentes manières : triangulaires, rectangulaire, *etc.*. La norme MPEG-7 décrit l'enveloppe du spectre audio (Audio Spectrum Envelope, ASE) comme le spectre de puissance logarithmique sommé au sein de bandes log-espacées suivant une base 2. La taille de ces bandes est fixée de 1/16 d'octave à 8 octaves.

MFCC Les coefficients cepstraux Mel-fréquence (*Mel-Frequency Cepstral Coefficients*, MFCC) sont une représentation compacte du spectre d'un signal (audio) qui tient compte de la perception non linéaire des fréquences par l'oreille humaine, à travers l'échelle Mel [BL03]. Le domaine cepstral est atteint par transformation de Fourier inverse du spectre log-compressé. Le premier coefficient concentre l'énergie de la trame, les coefficients suivants portent une information fréquentielle ; on ne conserve que les premiers échantillons. Ce descripteur est particulièrement utilisé dans le cadre de l'analyse de la parole.

Fréquence fondamentale, *pitch* La fréquence fondamentale d'un signal harmonique correspond à la fréquence dont les multiples entiers de son intégrale expliquent au mieux le contenu spectral du signal [Pee04]. La connaissance de cette fréquence est utile en analyse musicale pour déterminer les notes jouées, mais également en parole dans l'identification du locuteur ou la reconnaissance. Cette fréquence est estimée par l'analyse harmonique des pics spectraux ; pour chaque fréquence candidate, les premières harmoniques sont calculées, puis l'on vérifie si elles correspondent effectivement à des pics [DZ07].

Fréquence d'atténuation La fréquence d'atténuation spectrale (*Spectral Rolloff Frequency*) est définie comme la fréquence en dessous de laquelle sont concentrés p centiles de l'amplitude cumulée du spectre. Cette mesure est liée à la forme du spectre du signal. On trouve différentes valeurs de p dans la littérature, par exemple Tzanetakis et Cook [TC02] préconisent 85%, Li *et al* 92% [LSDM01], ou encore Wang *et al* comme Scheirer et Slaney proposent 95% [SS97, WLH00]. Cette mesure est utile pour la discrimination de parole voisée ou non-voisée [KMS05].

Centroïde, brillance Le centroïde spectral renvoie le centre de gravité, barycentre ou moyenne du spectre audio. Ce descripteur [SS97, LSDM01, TC99, AP03, Pee04] permet notamment de donner une mesure de la « brillance » du spectre : des valeurs élevées correspondent à des spectres plus brillants [DZ07, TC02]. En particulier, les sons impulsionnels, qui présentent de l'énergie en hautes fréquences, poussent le centroïde spectral vers de grandes valeurs [SS97]. Comme pour les bancs de filtre, on peut dériver de nombreux descripteurs de centroïde spectral, en utilisant des coefficients spectraux différents (amplitude, log-puissance, *etc.*) ou des échelles de fréquence différentes : logarithmique [IEC, KMS05], MEL, Bark [WTM+07] ou Warp.

Bande passante Aussi appelée étendue spectrale (*Spectrum Spread*), la bande passante instantanée correspond à la variance du spectre [WLH00], soit la différence moyenne entre le spectre et la brillance [DZ07]. Ce descripteur donne une indication de la concentration du spectre autour du centroïde spectral [WtBKW96, BL03]. La norme MPEG-7 [IEC, KMS05] décrit l'éten-

due spectrale (*Audio Spectrum Spread*, ASS) comme le moment centré d'ordre 2 du spectre log-fréquences.

Mesure d'harmonicité La platitude spectrale (*Spectral Flatness*) mesure la déviation du spectre en référence à un spectre plat au sein de bandes prédéfinies, autrement dit son niveau de similarité avec un bruit blanc. Ainsi, ce descripteur permet de distinguer les signaux bruités des signaux harmoniques. Il correspond au rapport des moyennes géométriques et arithmétiques du spectre d'énergie [Pee04]. Le facteur de crête spectrale (*Spectral Crest Factor*) constitue une autre mesure du caractère harmonique du spectre. Celui-ci est calculé comme le rapport de l'énergie spectrale maximum à la moyenne arithmétique du spectre d'énergie [Pee04].

1.2.3 Eléments d'ingénierie des descripteurs

Nous traitons maintenant de descripteurs nécessitant l'analyse de plusieurs trames. Il s'agit d'étudier à moyen ou long terme des descripteurs instantanés, tels que ceux précédemment décrits.

Statistiques Les descripteurs peuvent être enrichis d'informations complémentaires [BL03] : moyenne, variance, moyenne de la dérivée, variance de la dérivée, vitesse, accélération, *etc.*. Les moments centrés d'ordre 3 ou 4, appelés respectivement dérive (*skewness*) et kurtosis, sont également souvent utilisés. L'observation de cette évolution au cours du temps des descripteurs (quelques centaines de milli-secondes à plusieurs secondes ou minutes) est parfois décrite comme une analyse de la texture du signal audio [WLH00].

Taux de trames caractéristiques Un certain nombre de descripteurs définis précédemment donne une indication sur le caractère du signal : quantité d'énergie, silence, bruit, présence d'une fréquence fondamentale, etc. Un ensemble de descripteurs peut alors être construit par la mesure du rapport, sur un horizon donné, entre le nombre de trames présentant une caractéristique particulière et le nombre de trames totales : taux de faible énergie [LZJ02, BL03], taux de silence [LCTC05, BLLC06], taux de trames bruitées [LZJ02], ou encore taux de *pitch* [AEQG03, LCTC05].

Rapports d'énergie par bandes spectrales Le spectre peut également être normalisé par l'énergie totale de la trame [WLH00], ou fraction de cette énergie [WLH00, LSDM01]. En pratique, on peut considérer de nombreuses manières de découper le spectre en bandes et, selon l'application, ne conserver qu'une partie des coefficients. Cette approche regroupe un certain nombre de descripteurs comme, par exemple, le ratio parole à bruit qui s'intéresse à la fraction d'énergie de la bande $[300\text{Hz}; 3000\text{Hz}]$.

Flux spectral Ce descripteur, parfois appelé delta-spectre, mesure la différence spectrale entre deux trames successives [TC02], utile notamment en reconnaissance d'instruments monophoniques [AP03]. Il s'agit de la norme L_2 de la dérivation à l'ordre 1 d'une représentation du spectre (brute, compacte ou enveloppe). Tzanetakis et Cook [TC99] suggèrent d'utiliser les coefficients d'une transformée de Fourier à court-terme, normalisés en énergie. On constate notamment que pour la musique, les changement sont plus importants que pour la parole qui alterne des périodes relativement statiques (sons voisés) et des périodes de transition (sons non voisés) [SS97].

Centrage, réduction Enfin, les vecteurs peuvent être normalisés, c'est à dire centrés et réduits à l'aide d'une moyenne et d'une variance calculées sur un signal d'apprentissage, ou en ligne à l'horizon d'un nombre fini de trames, ou par filtrage. Cette opération permet généralement de comparer des signaux qui ne sont pas acquis dans les mêmes conditions d'enregistrement, ou dont les niveaux n'ont pas été normalisés.

1.3 Sélection de paramètres

Lorsque l'expertise ne permet pas d'identifier les descripteurs les plus pertinents, ou lorsque leur nombre est trop conséquent, on a recours à des méthodes de sélection de paramètres [LB97, DL97, GE03, LM08]. Ces approches doivent permettre de faciliter la construction des algorithmes de détection ou de classification [TK09]. En effet, limiter la dimensionnalité de l'espace d'observation permet de réduire le risque de sur-apprentissage [GE03] (voir 2.1.1), la complexité des modèles et le coût de calcul (temps d'apprentissage).

La réduction de dimensionnalité peut être séparée en deux familles d'algorithmes. D'une part, les algorithmes d'extraction ont pour objectif de capturer l'information de chaque descripteur au travers de leur combinaison. Géométriquement, cela revient à identifier un nombre réduit d'axes de projection dans l'espace d'observation. Ces approches souffrent néanmoins de la nécessité d'extraire l'ensemble des descripteurs pour opérer la réduction, et d'une perte d'interprétabilité des résultats lorsque des descripteurs hétérogènes sont combinés.

D'autre part, les algorithmes de sélection de paramètres, ou de descripteurs, ont pour objectif l'échantillonnage de l'ensemble des descripteurs initialement proposé. Cela revient à décomposer l'espace d'observation en sous-espaces distincts. Idéalement, on souhaite ainsi pouvoir isoler un sous-espace signal (les descripteurs d'intérêt), un sous-espace bruit (les descripteurs décorrélés de la tâche) et éventuellement un sous-espace de redondance s'il y a corrélation entre certains descripteurs. On s'intéresse à cette deuxième catégorie d'algorithmes.

1.3.1 Principe

La sélection de paramètres constitue un pré-traitement de la représentation. La procédure associée s'appuie sur deux étapes : la sélection de sous-ensembles et leur évaluation. La sélection s'appuie sur des heuristiques itératives, impliquant un état initial (ensemble vide, complet, pré-expertisé) et une stratégie de recherche. L'étape d'évaluation doit mettre en œuvre un critère corrélé à la pertinence d'un ou plusieurs descripteurs pour une tâche donnée.

L'alternance des étapes de sélection et d'évaluation permet de construire des sous-ensembles et de quantifier leur pertinence vis-à-vis de la structure des données ou de la tâche. Un critère d'arrêt doit permettre de mettre fin à la procédure. Il peut s'agir d'un seuil sur le critère d'évaluation, d'une limitation du nombre de sous-ensembles évalués, etc. Nous présentons dans les paragraphes suivants les stratégies de recherche, et les principaux paradigmes d'évaluation.

1.3.2 Stratégies

On distingue typiquement les stratégies de recherche optimales, assurant que le meilleur sous-ensemble sera sélectionné (pour un critère fixé), des stratégies sous-optimales. La figure 1.3.2 donne une taxonomie possible des stratégies que nous présentons.

Recherche exhaustive Cette stratégie évalue tous les sous-ensemble de descripteurs possibles. C'est l'assurance de trouver la meilleure combinaison de descripteurs, mais c'est également

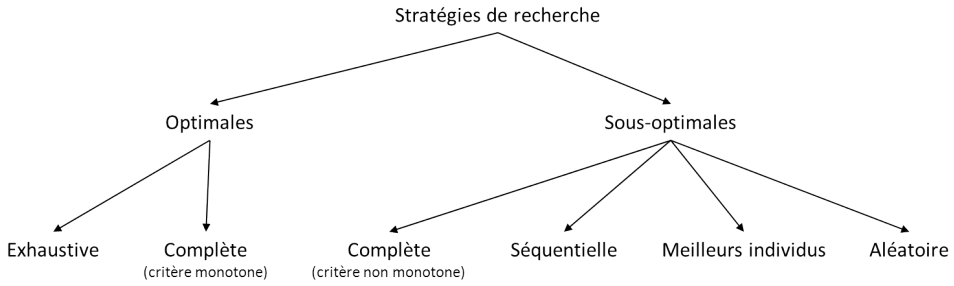


FIGURE 1.1 – Taxonomie des stratégies de sélection de paramètres.

coûteux en temps (par exemple, sélectionner 10 descripteurs parmi 50 représente 10 milliards de combinaisons). Cette stratégie permet néanmoins d'évaluer la qualité d'autres approches.

Recherche complète Cette stratégie repose sur une procédure de parcours en arbre des combinaisons de descripteurs qui peut éventuellement couvrir l'ensemble des possibilités. Néanmoins, suivant l'analyse du critère d'évaluation, certaines branches peuvent être abandonnées. Lorsque ce critère est monotone, cette stratégie est optimale. Dans le cas contraire, des heuristiques complémentaires autorisant à revenir à des branches abandonnées peuvent garantir de bons résultats [SP04, NC07].

Meilleurs individus Cette stratégie, la plus élémentaire, consiste à évaluer individuellement chacun des descripteurs pour ne conserver que les meilleurs. Si elle permet d'éliminer à moindre coût les descripteurs relevant du sous-espace bruit, elle n'identifie pas, en revanche, la redondance entre descripteurs [Web04].

Recherche séquentielle Il s'agit d'une stratégie itérative qui évalue des sous-ensembles par ajouts ou suppressions successifs de descripteurs. La procédure peut être « avant » (*Sequential Forward*, FS) ou « arrière » (*Sequential Backward*, BS) si la taille du sous-ensemble augmente ou diminue. Des approches généralisées complètent ou ôtent les descripteurs par groupes (*Generalized BS/FS*), ou combinent la recherche avant et arrière (*Floating BS/FS*). Là encore des heuristiques particulières conduisent à l'amélioration de la stratégie (*Improved Floating FS/BS*) [PFNJ94, KS00].

Recherche aléatoire Cette dernière stratégie consiste à tirer des combinaisons aléatoires de descripteurs. La densité de probabilité peut être fixée, ou itérativement contrôlée au fur et à mesure des évaluations. Les approches génétiques ou par essaims de particules relèvent de cette stratégie.

1.3.3 Paradigmes d'évaluation

Un critère peut être soit dépendant, soit indépendant, de la tâche à accomplir. Ces deux situations donnent naissance à deux paradigmes d'évaluation des sous-ensembles de descripteurs [DL97, KS00].

Filtre Cette approche évalue un sous-ensemble en tenant compte des qualités intrinsèques des données. Les critères indépendants les plus populaires sont :

- Les mesures de distance : séparabilité, divergence, discrimination. Différence entre les probabilités conditionnelles,
- Les mesures d'information : différence entre l'incertitude *a priori* et l'incertitude espérée *a posteriori*,
- Les mesures de dépendance : corrélation, similarité,
- Les mesures de consistance [4] : disposant de l'information sur les classes, on recherche le nombre minimum de paramètres qui séparent les classes avec autant de consistance que l'ensemble des paramètres d'origine (par exemple : critère de Fisher).

Wrapper Cette approche nécessite d'avoir spécifié la tâche à réaliser et l'algorithme utilisé. Le critère d'évaluation s'appuie sur les performances de ce dernier pour chaque sous-ensemble de paramètres. Les résultats sont meilleurs car ils dépendent de la tâche à accomplir.

Hybrides Cette approche exploite les avantages des deux précédents modèles.

1.4 Discussion

Comme cela a été mis en avant au cours de ce chapitre, le choix d'une représentation paramétrique particulière doit être guidée par 1- les caractéristiques des signaux et 2- une tâche particulière à réaliser. Ainsi, pour chaque besoin donné, une description « experte » du signal est proposée (MFCC pour l'analyse de la parole, descripteurs de timbre pour la reconnaissance des instruments de musique, *etc.*).

Dans le cas qui nous intéresse, nous ne souhaitons pas réaliser d'expertise sur les signaux. De plus, la tâche à résoudre n'est que partiellement définie puisque nous faisons l'hypothèse de ne disposer que des signaux de l'une des deux classes en jeu (hypothèse « normale »). Cette situation nous a d'abord conduits à extraire un très grand nombre de descripteurs acoustiques pour notre analyse ; filtrant ceux-ci en amont de l'algorithme de classification à l'aide d'approches de sélection automatique de paramètres. Cependant, ce degré de liberté dans le choix de la représentation paramétrique ajoute une complexité difficile à maîtriser ; en particulier dans le cadre de l'étude de l'apport des algorithmes de type SVM pour la détection.

Ainsi, nous avons fixé une représentation du signal pour l'ensemble des évaluations réalisées. Les descripteurs utilisés s'appuient sur les énergies en décibels à la sortie d'un banc de 32 filtres triangulaires. Ces filtres sont répartis sur l'ensemble de la bande passante (0-8kHz) suivant une échelle linéaire, avec recouvrement de 50% entre chaque. Par ailleurs, cette représentation permet d'identifier des caractéristiques audibles des zones de l'espace d'observation modélisées et facilite la compréhension du comportement des algorithmes de détection et de classification par l'analyse du spectrogramme des signaux traités [5].

4. Une définition *a contrario* : l'inconsistance est définie comme deux instances ayant les mêmes valeurs de paramètres et une étiquette de classe différente.

5. Cette même représentation a été utilisée dans les travaux de référence (projet CARETAKER [CDR+06, CR10]).

Chapitre 2

Apprentissage statistique

L'apprentissage statistique est une branche de l'apprentissage automatique (*Machine Learning*) dont l'objet est l'étude des méthodes permettant d'inférer (ou d'apprendre) une règle de décision à partir d'exemples numériques, ou encore d'améliorer la qualité de cette inférence. Nous formalisons dans ce chapitre les principes sur lesquels repose cette discipline. En particulier, nous allons tenter de répondre à la problématique suivante : *disposant d'un ensemble fini d'observations d'un ou plusieurs objets (des événements sonores par exemple) et de résultats associés (la nature de ces événements par exemple), comment construire efficacement une règle de décision capable de prédire le résultat associé à une nouvelle observation* ?

Dans un premier temps, nous présentons une description informelle et intuitive. Puis, nous définissons les notions de prédicteur et de risque. Enfin, nous évoquons le principe de minimisation du risque structurel et la régularisation, deux concepts permettant d'assurer et de quantifier les capacités de généralisation d'un prédicteur.

2.1 Principe de l'apprentissage par l'exemple

2.1.1 Modélisation et généralisation

On considère des observations issues de réalisations successives d'une expérimentation, ou de la capture répétée de l'état d'un système. Celles-ci peuvent être associées à une étiquette qui exprime la sortie attendue de la règle de décision que l'on souhaite inférer [6]. L'ensemble des observations, et leurs étiquettes associées, est appelé ensemble d'apprentissage. Enfin, on appelle classe un objet ou un ensemble d'objets qui correspondent à l'ensemble des réalisations possibles dont l'étiquette attendue est identique.

L'apprentissage par les machines a pour objectif d'inférer des règles de décision à partir d'un ensemble d'apprentissage, sous l'hypothèse que les observations de cet ensemble sont issues d'une même loi de probabilité, inconnue. En pratique, cela revient à sélectionner, au sein d'une famille de règles de décision, la règle qui optimise un critère choisi. Il s'agit, par exemple, de déterminer les paramètres d'une fonction d'un type spécifié *a priori* (linéaire, quadratique, *etc.*) qui minimisent l'erreur commise sur l'ensemble d'apprentissage. Les algorithmes d'apprentissage recherchent, au sein d'une famille, la règle de décision optimale en s'appuyant sur l'évaluation de risques et d'erreurs. Ces grandeurs sont présentées à la section 2.2.

Notons tout de suite que plus il y a d'exemples disponibles, plus l'apprentissage peut être de bonne qualité. A l'inverse, si le nombre d'observations est faible alors il y a un risque important

6. L'association des étiquettes aux observations résulte, en général, d'un processus d'expertise.

que les objets ayant été observés ne soient pas représentatifs de la classe, ou des classes, et que le modèle inféré ne puisse pas s'appliquer pertinemment à de nouvelles observations. Une attention particulière doit ainsi être portée à l'étude du pouvoir de généralisation d'une méthode d'apprentissage. Ceci se traduit formellement par la notion de risque abordée à la section 2.3.

Un autre enjeu de l'apprentissage statistique, pour les méthodes que nous étudions, est lié à la complexité de la famille de règles de décision au sein de laquelle nous évoluons. Une famille trop complexe est en mesure de capturer toute l'information disponible dans l'ensemble d'apprentissage, y compris les erreurs d'observation (bruit). La règle est alors inefficace sur de nouvelles observations ; on parle de sur-apprentissage. A l'inverse, une famille trop simple ne permet pas d'assimiler l'information présente dans l'ensemble d'apprentissage ; on parle alors de sous-apprentissage. La théorie de Vapnik-Chervonenkis apporte un cadre formel permettant de quantifier l'efficacité d'une famille de règles de décision. Celle-ci est également présentée à la section 2.3.

2.1.2 Espace d'observation

Les observations sont décrites par un ensemble de mesures associées à des caractéristiques des objets observés. Il s'agit par exemple des descripteurs acoustiques pour un signal audio. Il est alors possible de représenter les observations dans un espace d'observation, noté $\mathcal{X}$, dont les dimensions correspondent à ces caractéristiques. Généralement, les descripteurs sont à valeurs réelles et $\mathcal{X} \subset \mathbb{R}^D$.

Le choix de l'espace d'observation $\mathcal{X}$ nécessite de l'attention car les performances des règles de décision obtenues en dépendent en partie. Il reflète l'intelligence, ou expertise, que l'on peut mettre pour assimiler l'information contenue dans le signal, s'assurant du caractère discriminant de celle-ci. La sélection de paramètres, évoquée au paragraphe précédent, est un outil fondamental pour élaborer l'espace d'observation.

2.1.3 Espace de décision

Les résultats sont définis dans un espace de décision noté $\mathcal{Y}$. Selon la nature de $\mathcal{Y}$ (mono ou multi-varié, fini ou infini), nous distinguons différentes tâches de l'apprentissage statistique (voir la figure 2.1.3). Dans ces travaux, nous nous intéressons à la classification, cas où $\mathcal{Y}$ est un espace de variables catégorielles.

Lorsqu'il n'y a que deux classes, par exemple $\mathcal{Y} = \{-1, +1\}$ ou $\mathcal{Y} = \{0, 1\}$, on parle de problème 2-classes. Dans ce cas, l'espace d'observation est partagé en deux sous-espaces complémentaires réalisant une partition de $\mathcal{X}$.

Lorsque l'on ne s'intéresse qu'à une seule et unique classe, il s'agit d'un problème 1-classe, parfois appelé estimation de densité ou détection de nouveauté. Dans ce cas, l'objectif est de définir l'enveloppe des observations de la classe à modéliser. On parle d'estimation de densité [SPST$^+$01] car les lignes de niveaux définies par le classificateur peuvent être liées à la densité des observations modélisées. Le terme détection de nouveauté (ou d'anormalité) provient du fait que toute observation étant classée en dehors du volume modélisé doit correspondre à un objet nouveau (inconnu ou anormal).

Dans le cas des classificateurs multi-classes ($Q > 2$), il est possible de traiter le problème globalement, ou en utilisant des méthodes 2-classes accompagnées de règles de décision locales (un contre un, un contre tous, *etc.*). Il est possible également d'utiliser des approches 1-classe en construisant un classificateur pour chaque classe à modéliser [TL11]. Cette dernière approche est celle utilisée au cours des travaux présentés. Elle présente l'avantage de ne pas réaliser de

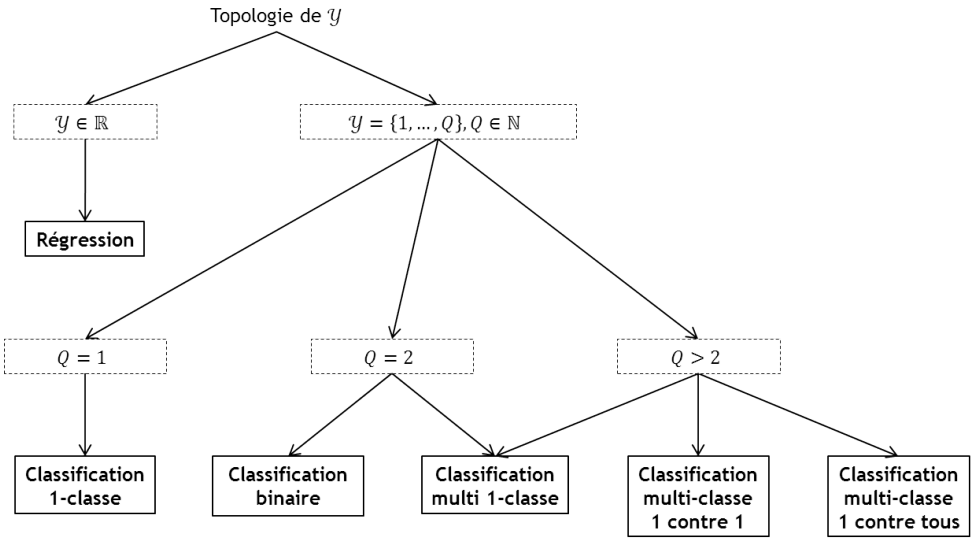


FIGURE 2.1 – Problèmes de l'apprentissage statistique en fonction de la topologie de l'espace de décision $\mathcal{Y}$

partition de l'espace d'observation et ainsi de conserver la capacité à détecter des objets inconnus ou anormaux.

On note $y_i \in \mathcal{Y}$ le résultat associé à la i-ème observation. On définit alors différents paradigmes d'apprentissage suivant l'ensemble des y_i disponibles dans l'ensemble d'apprentissage. Si tous les y_i sont observés, alors l'apprentissage est dit supervisé. Si aucun des y_i n'est connu, alors l'apprentissage est dit non supervisé. Enfin, si seule une partie des observations y_i est disponible, alors l'apprentissage est dit semi-supervisé.

2.2 Problème d'estimation : prédicteur et risque

Soit l'observation d'un ensemble $\mathcal{S}$ de N exemples, ou couples de variables aléatoires, $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N$ indépendantes et identiquement distribuées (iid) suivant une même loi $\mathbb{P}$. On appelle prédicteur [7] une fonction f permettant de déterminer le résultat y à partir d'une réalisation observée $\mathbf{x}$. L'idée de l'apprentissage statistique est de déterminer ce prédicteur à partir de l'ensemble d'observations $\mathcal{S}$, appelé ensemble d'apprentissage, sans nécessairement connaître $\mathbb{P}$. Ainsi, $f : (\mathcal{X} \times \mathcal{Y})^N \times \mathcal{X} \mapsto \mathcal{Y}$ exprime la dépendance fonctionnelle qui existe entre les observations $\mathbf{x}$ et y, et déterminée à partir de l'ensemble $\mathcal{S}$.

2.2.1 Risque fonctionnel

L'un des défis de l'apprentissage réside dans le pouvoir de généralisation de f, c'est-à-dire se capacité à correctement prédire le résultat y_i associé à une observation $\mathbf{x}_i$. La qualité d'un prédicteur se mesure par le résultat d'une fonction $R_c(f)$, appelée erreur de généralisation ou

7. Prédicteur est un terme générique en apprentissage statistique. Lorsque la tâche à résoudre est un problème de classification, on appelle le prédicteur un classificateur ; dans le cas de la régression, on parle de régresseur.

risque fonctionnel :

$$R(f) = \mathbb{E}\left[c\left(f, \mathbf{x}, y\right)\right] = \int_{\mathcal{X}\times\mathcal{Y}} c\left(f, \mathbf{x}, y\right) \mathbb{P}\left(\mathbf{x}, y\right) d\mathbf{x}dy \tag{2.1}$$

où $\mathbb{E}[X]$ est l'espérance mathématique d'une variable aléatoire X, et $c\left(f, \mathbf{x}, y\right)$ définit une fonction perte (ou fonction de contraste/coût), qui représente l'erreur entre le résultat prédit $f(\mathbf{x})$ et le résultat attendu y. Plus cette quantité est faible, plus le prédicteur est performant. On appelle prédicteur de Bayes f_{Bayes} le meilleur prédicteur au sens de la fonction perte :

$$f^*_{c,Bayes} = \inf_{f\in\mathcal{Y}} R_c(f) \tag{2.2}$$

Dans la suite, on suppose que la fonction coût est fixée et l'indice c est omis pour des raisons de compacité des notations. On remarque que si f est aléatoire (ce qui est le cas car f dépend de l'ensemble S, un tirage aléatoire d'observations), alors $c\left(f, \mathbf{x}, y\right)$ est aussi une variable aléatoire. De plus les observations de cette variable sont également indépendantes et identiquement distribuées.

2.2.2 Risque empirique : un estimateur consistant du risque fonctionnel

En pratique, seules les données d'apprentissage étant disponibles, on cherche à minimiser l'erreur de généralisation, conditionnellement à $\mathcal{S}$:

$$R(f, \mathcal{S}) = \mathbb{E}_{(\mathbf{x},y)}\left[c\left(f, \mathbf{x}, y\right)|S\right] = \int_{\mathcal{X}\times\mathcal{Y}} c\left(f, \mathbf{x}, y\right) d\mathbb{P}\left(\mathbf{x}, y|S\right) \tag{2.3}$$

Cependant, la distribution $\mathbb{P}\left(\mathbf{x}, y\right)$ étant inconnue, il est impossible de minimiser cette erreur. On utilise l'approximation stochastique suivante, appelée erreur empirique, ou risque empirique :

$$R_{emp}(f, \mathcal{S}) = \frac{1}{N}\sum_{i=1}^{N} c\left(f, \mathbf{x}_i, y_i\right) \tag{2.4}$$

Définition 2.1 (Estimateur consistant) *L'estimateur $\hat{\theta}_N$ d'une variable aléatoire est dit faiblement consistant s'il converge en probabilité vers θ lorsque N tend vers l'infini :*

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{\theta}_N\right] = \theta$$

Ce même estimateur est dit fortement consistant s'il converge presque sûrement vers θ :

$$\lim_{N\to\infty} \hat{\theta}_N \overset{p.s.}{=} \theta$$

Théorème 2.1 (Loi forte des grands nombres) *Si θ_n, $n > 0$ est une suite de variables aléatoires indépendantes et identiquement distribuées, on a équivalence entre :*

- *$\mathbb{E}\left(|\theta_1|\right) < +\infty$*
- *la suite $\frac{\theta_1+\cdots+\theta_n}{n}$ converge presque sûrement.*

De plus, si l'une de ces deux conditions équivalentes est remplie, alors la suite $\frac{\theta_1+\cdots+\theta_n}{n}$ converge presque sûrement vers la constante $\mathbb{E}\left(\theta_1\right)$.

Le risque empirique est donc un estimateur consistant (fortement) de l'erreur de généralisation :

$$\lim_{N\to\infty} R_{emp}\left(f, \mathcal{S}\right) \overset{p.s.}{=} R\left(f\right) \tag{2.5}$$

2.2.3 Erreurs de modélisation, d'estimation et d'approximation

La recherche du prédicteur minimisant le risque empirique pour un ensemble $\mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^N$ donné est un problème toutefois mal posé. Il existe en effet une infinité de solutions $f \in \mathcal{Y}$ garantissant un apprentissage exact. De plus, il n'y a aucune garantie sur le fait que ces solutions soient performantes sur de nouvelles données, pourtant issues de la même distribution $\mathbb{P}$. Afin de trouver une solution qui « généralise », l'espace de recherche est restreint à un espace de fonctions régulières dans $\mathcal{Y}$, noté $\mathcal{F}$. Dans cet espace restreint, on note f^* la fonction minimisant le risque empirique :

$$f^* = \arg\min_{f \in \mathcal{F}} R_{emp}(f, \mathcal{S}) \tag{2.6}$$

De plus, rien ne garantit que le prédicteur optimal f_{Bayes} soit inclus dans l'espace restreint $\mathcal{F}$. Il existe donc un prédicteur optimal dans $\mathcal{F}$ non nécessairement optimal au sens de Bayes que l'on note $f^*_{\mathcal{F}}$:

$$f^*_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} R(f) \tag{2.7}$$

Seule la fonction f^* peut être déterminée au cours de l'apprentissage, néanmoins il est intéressant de qualifier cette fonction en la comparant aux fonctions f^*_{Bayes} et $f^*_{\mathcal{F}}$. On définit l'erreur d'estimation, également appelée variance inductive, comme la différence entre risque réel et risque empirique dans $\mathcal{F}$:

$$E_{estim} = R(f^*) - R(f^*_{\mathcal{F}}) \tag{2.8}$$

Cette erreur dépend du processus d'apprentissage, de l'ensemble $\mathcal{S}$ et du choix de l'espace fonctionnel $\mathcal{F}$ et il s'agit d'une quantité aléatoire (dépendance à S, un ensemble de variables aléatoires). Idéalement, l'erreur d'estimation tend vers 0 lorsque N augmente. A l'inverse, plus $\mathcal{F}$ est un ensemble riche de fonctions (moins l'espace est contraint), plus on risque d'augmenter cette erreur. D'autre part, on définit l'erreur d'approximation, également appelé biais inductif, comme la différence entre risque réel dans $\mathcal{F}$ et dans $\mathcal{Y}$:

$$E_{approx} = R(f^*_{\mathcal{F}}) - R\left(f^*_{Bayes}\right) \tag{2.9}$$

Cette erreur est à relier à la pertinence du choix de $\mathcal{F}$ dont elle dépend. Plus l'ensemble est riche, plus cette erreur tend vers 0, autrement dit moins le prédicteur $f^*_{\mathcal{F}}$ est contraint plus il tend vers le prédicteur de Bayes. On définit enfin l'excès de risque ou l'erreur de modélisation comme l'erreur commise en choisissant le prédicteur f^* (que l'on peut calculer) plutôt que le prédicteur f_{Bayes} :

$$E_{model} = E_{estim} + E_{approx} = R(f^*) - R\left(f^*_{Bayes}\right) \tag{2.10}$$

Notons également que ces erreurs sont toutes, par définition, positives ou nulles.

La minimisation du risque de modélisation ressemble au compromis biais-variance (inductifs). En effet, les erreurs d'estimation et d'approximation évoluent de façon opposée lorsque $\mathcal{F}$ croît ou décroît. Il convient de s'assurer d'un choix correct de l'ensemble $\mathcal{F}$. En pratique, on peut seulement déterminer f^* et son risque empirique $R_{emp}(f^*)$. Bien que la convergence de $R_{emp}(f)$ vers $R(f)$ soit assurée, seule une convergence uniforme peut garantir la convergence de leurs minimums respectifs. Ce point est traité par la théorie de Vapnik et Chervonenkis [VC71] qui fait l'objet du paragraphe suivant. Celle-ci permet d'une part de qualifier une famille de fonctions $\mathcal{F}$ pour l'apprentissage, et d'autre part, de borner la différence entre risque de généralisation et risque empirique.

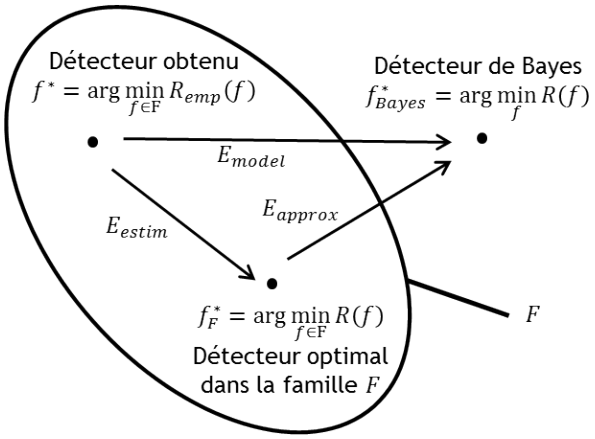


FIGURE 2.2 – Erreurs de modélisation, d'estimation et d'approximation

2.3 Minimisation du risque et généralisation

La capacité d'une famille de prédicteurs à correctement capturer l'information d'un ensemble d'apprentissage est liée à sa complexité. En effet, plus une famille est complexe, plus elle offre de possibilités pour capturer les spécificités d'un ensemble d'observations. Dans un premier temps, nous montrons comment la théorie de Vapnik et Chervonenkis concernant la Minimisation du Risque Stucturel (MRS) apporte un cadre formel à ce constat [Vap95]. Nous évoquons ensuite le problème de la régularisation, utile pour rendre bien posé un problème *a priori* mal-posé, tel que celui de la minimisation du risque structurel.

2.3.1 Minimisation du risque structurel

Définition 2.2 (Dimension de Vapnik-Chervonenkis) *On dit qu'un espace de fonctions $\mathcal{F}$ pulvérise un ensemble de données $\mathcal{S} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1 \dots N\}$ si quelles que soient les valeurs prises par les étiquettes y_i, il existe toujours un prédicteur $f \in \mathcal{F}$ tel que f ne commet aucune erreur sur $\mathcal{S}$. On définit la dimension de Vapnik-Chervonenkis, notée VC-dim, pour un espace $\mathcal{F}$ comme le cardinal N du plus grand ensemble de points $\mathcal{S}$ qui peut être pulvérisé par $\mathcal{F}$. Soit :*

$$VC\text{-}dim = \max \{k | \operatorname{card}(\mathcal{S}) = k \text{ et } f \in \mathcal{F} \text{ pulvérise } \mathcal{S}\} \tag{2.11}$$

La VC-dim offre donc une mesure quantifiable de la complexité d'un ensemble de fonctions $\mathcal{F}$. La figure 2.3 donne deux exemples illustrant ce principe. Il est important de noter qu'un mauvais choix dans la structure des prédicteurs, autrement dit un choix de $\mathcal{F}$ inadapté, peut conduire aux situations de sur-apprentissage et de sous-apprentissage que nous avons déjà évoquées.

Ce phénomène est particulièrement vrai lorsque la famille de prédicteurs est complexe, c'est-à-dire lorsque la VC-dim de $\mathcal{F}$ est élevée. Le prédicteur a alors une trop grande capacité à assimiler les informations de l'ensemble d'apprentissage. Le sur-apprentissage est caractérisé par le fait qu'un prédicteur modélise très bien l'ensemble d'apprentissage mais présente une faible capacité de généralisation. En classification, le recouvrement entre classes est connu pour être une source de sur-apprentissage [TK09]. En d'autres termes, un prédicteur étudié pour présenter une erreur trop faible sur l'ensemble d'apprentissage présentera un taux d'erreur important sur de nouvelles données.

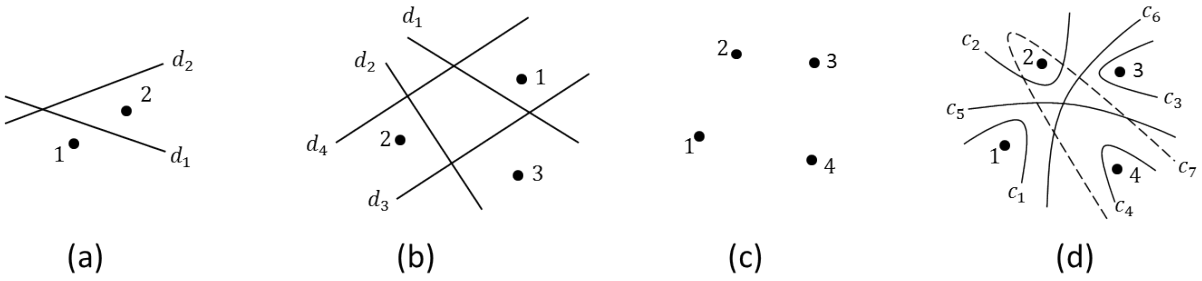


FIGURE 2.3 – La dimension de Vapnik-Chervonenkis, pour une famille de fonctions et un espace fixé, indique la capacité à pulvériser (réaliser toutes les dichotomies possibles) un ensemble de $VC-dim$ observations. Dans le plan, les droites pulvérisent 3 points au maximum ($VC-dim = 3$ pour les fonctions affines dans $\mathbb{R}^2$).

A l'inverse, un ensemble $\mathcal{F}$ de fonctions trop simples ne permettra pas d'assimiler l'information utile pour réaliser la tâche de classification désirée. En pratique, un bon compromis est donc à trouver.

A partir de la VC-dim, Vapnik et Chervonenkis ont montré que, avec une probabilité $1-\varepsilon$:

$$R(f) < R_{emp}(f) + \sqrt{\frac{h\ln\left(\frac{2N}{h}+1\right) - \ln\frac{\varepsilon}{4}}{N}} \tag{2.12}$$

où h est la VC-dim de la famille $\mathcal{F}$ tel que $f \in \mathcal{F}$. Le principe de minimisation du risque structurel (MRS) suggère donc de minimiser par rapport à h le risque garanti défini par la borne supérieure $R_{emp}(f) + \Phi(N, h, \varepsilon)$, plutôt que le risque empirique. La figure 2.4 illustre ce principe. On constate alors qu'une condition nécessaire et suffisante pour garantir la convergence du minimum du risque empirique vers le minimum du risque réel, lorsque $N \to \infty$, est que la VC-dim h soit finie. [Gei12] fournit une introduction complète à l'apprentissage statistique où le lecteur intéressé peut trouver les démonstrations associées.

Définition 2.3 (Malédiction de la dimensionnalité) *La Malédiction de la dimensionnalité ou fléau de la dimension (*Curse of dimensionality*) désigne l'augmentation explosive du volume de données nécessaire pour occuper un espace avec une densité constante lorsque des dimensions supplémentaires sont ajoutées. [Bel61]*

D'après (2.12), le risque empirique devient exploitable lorsque les bases de données utilisées pour l'apprentissage sont grandes (N grand). Or, il faut également prendre en compte la dimensionnalité de ces bases (nombre de descripteurs utilisés), car plus celle-ci est importante, plus le nombre d'observations nécessaires à une bonne estimation est important. Ainsi, le risque empirique peut devenir un estimateur fortement biaisé de l'erreur de généralisation [EP03] et même un prédicteur à la structure très simple (un classificateur linéaire par exemple) peut se trouver dans ce cas en situation de sur-apprentissage.

2.3.2 Estimateurs de l'erreur de généralisation

Si la minimisation du risque empirique, conjointement à la restriction de $\mathcal{F}$ suivant le principe de MRS, permet d'estimer un prédicteur satisfaisant, le risque empirique ne permet pas, en

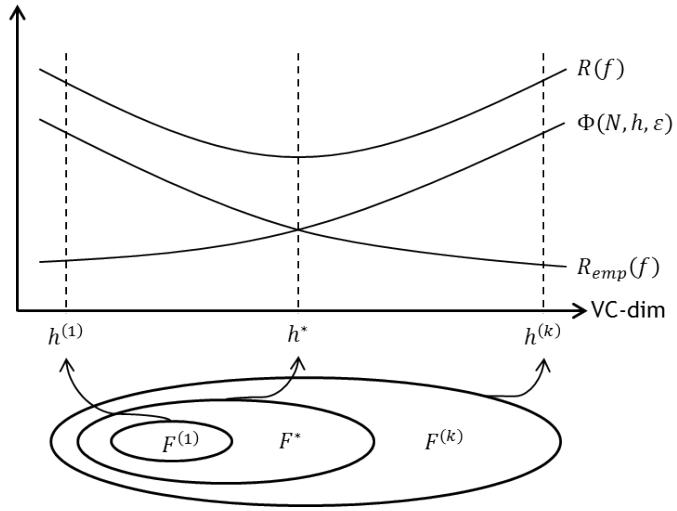


FIGURE 2.4 – Minimisation du risque structurel

général, d'estimer l'erreur de généralisation. Nous reprenons dans cette section les estimateurs d'erreur les plus connus dans la littérature tels que Tohmé les a présentés [Toh09].

1. *Jackknife* : [Que49] a introduit cette méthode pour estimer le biais d'une statistique ou d'un estimateur donné. Cette méthode consiste à supprimer une observation de l'ensemble des données et à calculer la valeur de l'estimateur sur l'ensemble des données restantes. Pour un ensemble de N observations, à chaque itération, le *Jackknife* utilise $N-1$ observations et calcule la statistique qui nous intéresse. La moyenne de cette statistique est comparée à celle calculée à partir de l'ensemble total des observations dans le but d'estimer le biais de cette dernière.
2. *Bootstrap* : Cette méthode, introduite par [Efr79], est basée sur ce qu'on appelle « ré-échantillonnage ». Soit un échantillon de N observations, le principe de la méthode du *Bootstrap* est de prélever une série d'échantillons aléatoires avec remise de N observations dans l'échantillon initial. A chaque tirage, on estime les statistiques de l'échantillon initial et on calcule l'erreur. D'autres formes plus sophistiquées de *Bootstrap* [ET94] ont été développées plus récemment pour estimer non seulement l'erreur de généralisation mais aussi l'intervalle de confiance de la sortie.
3. *Cross-validation* : Pour la *n-fold cross-validation* (validation croisée), on divise les données en n sous-ensembles. A chaque itération, on retire un de ces sous-ensembles que l'on utilise pour l'estimation [DK82]. Cependant, quand ce sous-ensemble est de taille 1, *n-fold cross-validation* sera nommé *Leave-One-Out* puisque nous utilisons l'intégralité de l'ensemble d'apprentissage sauf une observation.
4. *Leave-One-Out* : L'estimateur d'erreur *Leave-One-Out* (LOO) apparaît dans plusieurs articles [Lac67, LB69]. Il est défini par :

$$R_{LOO}(f, \mathcal{S}) = \frac{1}{N} \sum_{i=1}^{N} c(f^i, \mathbf{x}_i, y_i) \tag{2.13}$$

où f^i est la fonction déterminée à partir de $\mathcal{S}^i$, qui est l'ensemble $\mathcal{S}$ duquel est retiré l'individu i : $\mathcal{S}^i = \mathcal{S} - \{(x_i, y_i)\}$. L'estimateur LOO est un bon estimateur de l'erreur de

généralisation et est connu pour être « presque » non biaisé. [LB69] montre cette propriété à travers le résultat suivant :

$$\mathbb{E}_S\left[R_{LOO}\left(f\right)\right] = \mathbb{E}_{\mathcal{S}^i}\left[R\left(f\right)\right] \tag{2.14}$$

Autrement dit, l'erreur LOO est informative sur l'erreur de généralisation obtenue en utilisant le même algorithme d'apprentissage sur $N-1$ points.

2.3.3 Régularisation

Nous avons vu au cours de ce chapitre que l'apprentissage d'un prédicteur passe par la minimisation d'une quantité, appelée risque empirique, mesurable pour un ensemble de données d'apprentissage $\mathcal{S}$ de taille N. Ce problème étant mal posé du fait de la non unicité de la solution, l'espace de recherche du prédicteur a été limité, selon le principe de minimisation du risque structurel, à un ensemble fonctionnel $\mathcal{F} \subset \mathcal{H}$ dont la $VC-dim$ est finie. Cependant, nous n'avons jusqu'à présent pas défini de moyen permettant, en pratique, de limiter cet espace. On appelle régularisation une telle tâche, consistant à privilégier une solution dotée de caractéristiques particulières dans un problème d'optimisation. On va alors utiliser la régularisation pour contraindre la complexité de $\mathcal{F}$, et éviter les situations de sur-apprentissage.

La régularisation, un outil mathématique élégant permettant de rendre un problème bien posé, est étudiée depuis les années 1960. On présente maintenant trois méthodes de régularisation adaptées à la minimisation du risque empirique [Phi62, Iva76, Tik63]. Soit $\mathcal{S}$ un ensemble d'apprentissage de taille N et $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ un espace de fonctions $\mathcal{F} \subset \mathcal{H}\mathcal{X} \mapsto \mathbb{R}$, auquel appartient le prédicteur recherché f^* :

Régularisation de Phillips La régularisation prend la forme d'un terme, appelée fonctionnelle de régularisation, qu'il s'agit de minimiser sous la contrainte que le risque empirique reste inférieur à un seuil donné.

$$\min_{f\in\mathcal{F}} \lambda\Omega(f) \text{ sous contrainte } R_{emp}(f,\mathcal{S}) \leq \tau_1 \tag{2.15}$$

Régularisation de Ivanov A l'inverse, il s'agit de minimiser le risque empirique, avec une contrainte sur la fonctionnelle de régularisation.

$$\min_{f\in\mathcal{F}} R_{emp}(f,\mathcal{S}) \text{ sous contrainte } \lambda\Omega(f) \leq \tau_2 \tag{2.16}$$

Régularisation de Tikhonov Il s'agit d'une régularisation sans contrainte.

$$\min_{f\in\mathcal{F}} \lambda\Omega(f) + R_{emp}(f,\mathcal{S}) \tag{2.17}$$

Les seuils τ_1 et τ_2 sont des réels fixés, et $\lambda \in \mathbb{R}$ contrôle le compromis entre risque empirique et régularité de la solution. Ce dernier est communément appelé paramètre de régularisation. Dans le contexte qui nous intéresse, la fonctionnelle de régularisation est, par restriction, la norme quadratique dans l'espace de $\mathcal{H}$: $\Omega(f) = \|f\|^2_{\mathcal{H}}$. De plus, on montre que ces trois formes de régularisation sont équivalentes [Muk04]. Aussi, on ne s'intéressera qu'à la régularisation de Thikhonov dont la solution (alors unique) s'exprime comme :

$$f^* = \arg\min_{f\in\mathcal{H}} \left[\lambda\|f\|^2_{\mathcal{H}} + \frac{1}{N}\sum_{i=1}^{N} c\left(f, \mathbf{x}_i, y_i\right)\right] \tag{2.18}$$

Par ailleurs, la théorie de la régularisation s'est généralisée depuis qu'elle a été introduite dans le contexte de l'apprentissage statistique [PG90, Vap95]. Ainsi, soit $g : \mathbb{R} \mapsto \mathbb{R}^+$ une application monotone croissante, alors le problème (2.18) se généralise :

$$f^* = \arg\min_{f \in \mathcal{H}} \left[\lambda g \left(\|f\|_{\mathcal{H}}^2 \right) + \frac{1}{N} \sum_{i=1}^{N} c\left(f, \mathbf{x}_i, y_i\right) \right] \tag{2.19}$$

Trouver une telle solution est alors réalisable à l'aide de la technique des multiplicateurs de Lagrange [TA77] qui sera exposée en détail au chapitre suivant.

2.4 Synthèse

L'apprentissage statistique consiste en l'estimation d'une fonction permettant de relier des observations à des décisions. Cette fonction est choisie au sein d'une famille de fonctions fixée *a priori*. Les paramètres inconnus, permettant de sélectionner une fonction au sein de cette famille, sont solution d'un problème d'optimisation, étant donné un ensemble d'apprentissage. Si on considère que la quantité d'exemples est suffisamment grande, la minimisation du risque empirique est alors le problème privilégié pour déterminer ces paramètres.

Le choix d'un coût approprié dans l'expression du risque empirique permet de déterminer la tâche à réaliser. En particulier, nous exploitons cette approche pour la classification 1 ou 2 classes. Les fonctions coût utilisées sont introduites par la suite.

Nous avons également proposé plusieurs mesures du risque de généralisation qui permettent d'établir ou de mesurer la capacité d'un prédicteur à se généraliser à de nouvelles données. Celles-ci sont utiles pour la détermination des paramètres non obtenus par la résolution du problème d'optimisation, la largeur du noyau Gaussien par exemple.

Deuxième partie

Modèle et algorithme SVM 1-classe pour la détection

Chapitre 3

Machines à vecteurs de support

Nous introduisons dans ce chapitre les machines à vecteurs de support (SVM). En particulier, nous décrivons cette approche pour servir les besoins de la classification 1-classe (problème 1-classe) et de la classification 2-classes (problème discriminant). Nous inscrivons notre présentation dans le contexte général des espaces de Hilbert à noyau reproduisant dont les propriétés permettent aux SVM de s'adapter à une grande variété de non-linéarités. Nous décrivons dans ce chapitre la méthode des multiplicateurs de Lagrange qui est privilégiée pour la résolution des problèmes SVM du fait de la nature du problème dual obtenu.

A travers ce chapitre, nous présentons enfin une approche originale des problèmes SVM 1-classe et 2-classes. Après avoir introduit dans un premier temps les approches de l'état de l'art, avec biais, nous présentons une approche statistique des SVM, s'affranchissant du biais. Nous montrons ensuite que les problèmes associés, 1-classe ou 2-classes, avec ou sans biais, peuvent être mis sous la forme d'un problème dual unifié.

3.1 Méthodes à noyaux

Les SVM sont des méthodes de classification qui ont été proposées dans le cas de la classification linéaire. Elles ont été étendues au cas non linéaire par la projection des données dans un espace transformé. L'astuce du noyau permet, grâce à l'utilisation d'un espace image particulier, de ne pas exprimer directement la projection, mais une fonction des données d'origine, appelée noyau.

Dans cette section, nous donnons les notions utiles pour l'introduction des machines à vecteurs de support. Nous rappelons dans un premier temps la définition des espaces de Hilbert à noyau reproduisant. Nous introduisons ensuite le théorème du représentant qui garantit l'existence d'une solution au problème SVM non linéaire exposé par la suite.

3.1.1 Exemple introductif

Soient deux classes de données réparties suivant deux cercles concentriques distincts tels que sur la figure 3.1. La représentation des données dans le plan ne permet pas de définir un séparateur linéaire entre les deux classes. Cependant, il est possible de transformer les données

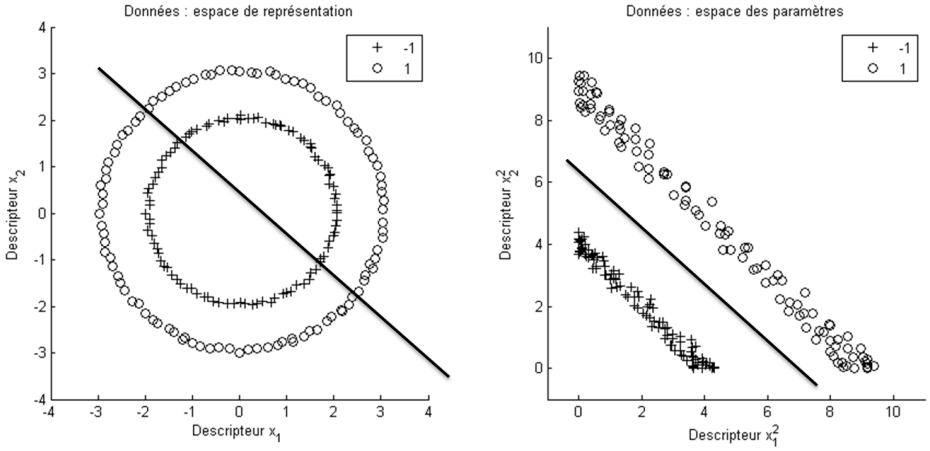


FIGURE 3.1 – Images des données dans un espace de paramètres : deux classes de données non linéairement séparables en considérant les descripteurs x_1 et x_2 (à gauche) le deviennent en considérant les monômes à l'ordre 2 (à droite, les monômes x_1^2 et x_2^2)

par des monômes à l'ordre 2 des dimensions originales. Une telle application ϕ s'exprime [8] :

$$\begin{array}{rccc} \phi : & \mathbb{R}^2 & \mapsto & \mathbb{R}^4 \\ & \mathbf{x}_i = (x_{i,1}, x_{i,2}) & \rightarrow & \left(x_{i,1}^2, x_{i,2}^2, x_{i,1}x_{i,2}, x_{i,2}x_{i,1}\right) \end{array} \tag{3.1}$$

et il est alors possible dans ce nouvel espace de séparer linéairement les données. En effet, dans le plan constitué par les nouveaux descripteurs x_1^2 et x_2^2, les données sont linéairement séparables, comme l'illustre la figure 3.1.

Intéressons-nous maintenant aux produits scalaires entre deux éléments, à la fois dans l'espace d'observation et dans l'espace des paramètres. On remarque que, pour deux éléments quelconques i et j :

$$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = x_{i,1}^2 x_{j,1}^2 + x_{i,2}^2 x_{j,2}^2 + 2x_{i,1}x_{i,2}x_{j,1}x_{j,2} = (x_{i,1}x_{j,1} + x_{i,2}x_{j,2})^2 = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 \tag{3.2}$$

Il est alors possible, pour la transformation ϕ proposée, de déterminer le produit scalaire de deux éléments image sans nécessairement devoir calculer explicitement $\phi(\mathbf{x}_i)$ et $\phi(\mathbf{x}_j)$, mais directement à partir du seul produit scalaire des éléments originaux. Soit κ la fonction réalisant cette opération :

$$\begin{array}{rccc} \kappa : & \mathbb{R}^2 & \mapsto & \mathbb{R} \\ & (\mathbf{x}_i, \mathbf{x}_j) & \rightarrow & \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 \end{array} \tag{3.3}$$

La fonction κ réalise bien un produit scalaire dans l'espace image. On note que la fonction κ résume à elle seule la structure de cet espace. Il s'agit d'un raccourci calculatoire intéressant lorsque la dimensionnalité de cet espace est grande, voire indispensable lorsque celle-ci est infinie. La théorie des noyaux reproduisants permet d'apporter un cadre formel à ce résultat et de l'étendre à un grand nombre de transformations.

8. On note $x_{i,d}$ la valeur prise par le d-ième descripteur d'une observation i.

3.1.2 Transformée vers un espace de dimensionnalité élevée

L'idée générale pour traiter un problème non linéaire est de projeter les données de l'espace d'observation $\mathcal{X}$ de dimension D vers un nouvel espace $\mathcal{H}$ de dimension $K > D$, appelé espace des paramètres (*feature space* selon la terminologie anglo-saxonne) :

$$\mathcal{X} \ni \mathbf{x} \mapsto \phi(\mathbf{x}) \in \mathcal{H} \tag{3.4}$$

On espère alors que les données image sont linéairement séparables. Il nous appartient de sélectionner la fonction ϕ appropriée. Le théorème de Cover apporte une justification théorique à cette approche dans le cadre de la classification. Il stipule qu'un problème de classification projeté non linéairement dans un espace de grande dimensionnalité est vraisemblablement mieux séparable que dans un espace de faible dimensionnalité, pourvu que cet espace ne soit pas densément peuplé [Cov65].

Théorème 3.1 (Théorème de Cover) *Soient N points aléatoirement distribués dans un espace de dimensionnalité D et une fonction de projection de ces points vers un espace de dimensionnalité $K > D$. Alors la probabilité de localiser ces points, dans l'espace image, au sein de groupes linéairement séparables tend vers* 1 *quand K tend vers l'infini.*

Bien que cette approche semble intéressante, nous sommes limités dans la pratique par l'expression de cette application ϕ. Le calcul peut même être impossible lorsque la dimensionnalité K de l'espace image tend vers l'infini. Comme on l'a vu précédemment, il est néanmoins possible de réaliser des opérations dans $\mathcal{H}$ sans expliciter ϕ dès lors que les calculs s'expriment sous forme de produits scalaires. Cette astuce est notablement exploitable lorsque $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant (dénoté par l'abréviation RKHS, *Reproducing Kernel Hilbert Spaces* selon la terminologie anglo-saxonne).

3.1.3 Espaces de Hilbert à noyau reproduisant

Noyau et théorème de Mercer

Historiquement [9], dans le cadre de la théorie des équations intégrales, Mercer a introduit la notion de noyau défini positif $\kappa(\cdot,\cdot)$ défini sur $\mathcal{X} \times \mathcal{X}$, lequel satisfait le théorème de Mercer [Mer09] (voir ci-dessous). Dans la continuité de ces travaux, Moore a montré qu'il existe une classe de fonctions $\mathcal{H}$ dotée du produit scalaire $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ pour laquelle un tel noyau κ possède la propriété dite reproduisante [Moo16] :

$$f(\mathbf{x}) = \langle f, \kappa(\cdot, \mathbf{x})\rangle_{\mathcal{H}} \tag{3.5}$$

pour toute fonction $f \in \mathcal{H}$ et observation $\mathbf{x} \in \mathcal{X}$. Suivant cette approche, l'espace $\mathcal{H}$ est introduit a posteriori dans le cadre de l'étude des noyaux définis positifs. A l'inverse, Zaremba a introduit les RKHS en s'intéressant aux classes de fonctions $\mathcal{H}$, alors que le noyau n'était qu'un outil pour l'étude [Zar07]. Bergman et Aronszajn ont également suivi cette seconde approche [Ber22, Aro50], s'intéressant à l'aspect reproduisant des noyaux et non à l'aspect défini positif.

Les propriétés des RKHS et des noyaux ont ensuite été étudiées conjointement au cours de la seconde moitié du XX^{e} siècle, l'article de Aronszajn [Aro50] étant le point de départ de

9. Nous introduisons les RKHS dans le contexte de la classification. Bien que la suite de l'exposé reste dans ce cadre, il est intéressant de préciser que les méthodes à noyaux, reposant sur cet outil mathématique, sont également utilisées à d'autres fins comme l'analyse en composantes principales [SSM97], l'analyse discriminante de Fisher [MRW^{+}99], la régression [Nad64, Wat64], le traitement des images et bien d'autres encore [Par62, Yao67].

ces travaux. Ceux-ci reposent sur l'analyse fonctionnelle dans les espaces de Hilbert et sur le théorème de Mercer pour formaliser le lien entre la fonction noyau κ et la fonction de projection ϕ. En particulier, il a été montré l'équivalence entre espace de noyaux définis positifs et espace de Hilbert à noyau reproduisant.

Définition 3.1 (Noyau défini positif) *On appelle noyau défini positif une fonction symétrique $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ telle que :*

$$\int_{\mathcal{X}\times\mathcal{X}} g(\mathbf{x}_i)\,\kappa(\mathbf{x}_i,\mathbf{x}_j)\,g(\mathbf{x}_j)\,d\mathbf{x}_i d\mathbf{x}_j \geq 0 \tag{3.6}$$

pour tout $g \in \mathcal{L}^2_{\mathcal{X}}$, ou alternativement, dans le cas où $\mathcal{X}$ est un espace échantillonné, telle que :

$$\sum_{i,j=1}^{N} c_i c_j \kappa(\mathbf{x}_i,\mathbf{x}_j) \geq 0 \qquad \forall \mathbf{x}_1,\ldots,\mathbf{x}_N \in \mathcal{X} \qquad \forall c_1,\ldots,c_N \in \mathbb{R} \tag{3.7}$$

On note également que pour un ensemble de N points, il est possible de construire une matrice de l'ensemble des éléments $\kappa(\mathbf{x}_i,\mathbf{x}_j)$. On appelle cette matrice la matrice de Gram, notée $\mathbf{K}$, où $\forall i = 1,\ldots,N$ et $\forall j = 1,\ldots,N$, $\mathbf{K}_{i,j} = \kappa(\mathbf{x}_i,\mathbf{x}_j)$. Pour un noyau défini positif, la matrice de Gram est également définie positive [10].

Théorème 3.2 (Théorème de Mercer) *Soit $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ un noyau défini positif continu. Soit $T_\kappa : \mathcal{L}^2_{\mathcal{X}} \mapsto \mathcal{L}^2_{\mathcal{X}}$ l'opérateur intégral défini par :*

$$[T_\kappa(f)](\mathbf{x}) = \int_{\mathcal{X}} \kappa(t,\mathbf{x})\,f(t)dt \tag{3.8}$$

Alors :

- *les valeurs propres λ_k de T_κ sont des réels positifs,*
- *il existe une base de fonctions propres orthogonales ψ_k de $\mathcal{L}^2_{\mathcal{X}}$,*
- *les fonctions propres correspondant à des valeurs propres strictement positives sont continues sur $\mathcal{X}$ et*

$$\int_{\mathcal{X}} \kappa(t,\mathbf{x})\,\psi_k(t)dt = \lambda_k \psi_k(\mathbf{x}) \tag{3.9}$$

- *κ admet la décomposition suivante :*

$$\kappa(\mathbf{x}_i,\mathbf{x}_j) = \sum_{k=1}^{\infty} \lambda_k\,\psi_k(\mathbf{x}_i)\psi_k(\mathbf{x}_j) \tag{3.10}$$

On montre alors qu'un noyau κ satisfaisant le théorème de Mercer peut servir de produit scalaire dans un espace de paramètres $\mathcal{H}$. Il suffit d'écrire :

$$\phi(\mathbf{x}) = \begin{pmatrix} \sqrt{\lambda_1}\psi_1(\mathbf{x}) \\ \sqrt{\lambda_2}\psi_2(\mathbf{x}) \\ \cdots \end{pmatrix} \tag{3.11}$$

10. En analyse matricielle, on dira qu'elle est semi-définie positive. Dans le contexte des méthodes à noyau, on conserve la terminologie de Mercer, soit défini positif.

Espace de Hilbert à noyau reproduisant

L'espace $\mathcal{H}$ induit par la projection (3.11) est, sous certaines conditions sur le noyau κ, un espace de Hilbert à noyaux reproduisant.

Définition 3.2 (Espace de Hilbert) *Un espace de Hilbert, noté $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$, est un espace vectoriel normé $\mathcal{H}$ sur un sous-corps de $K \in \mathbb{C}$ (typiquement, $K = \mathbb{C}$ ou $K = \mathbb{R}$), complet pour la distance issue de sa norme, et dont cette norme $\|\cdot\|$ découle d'un produit scalaire $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ par la formule $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_{\mathcal{H}}}$.*

Définition 3.3 (Espace de Hilbert à noyau reproduisant) *Soit $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ un espace de Hilbert de fonctions de $\mathcal{X}$ dans $\mathbb{R}$. La fonction $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}$ est le noyau reproduisant de $\mathcal{H}$, sous réserve que celui-ci en admette un, si et seulement si :*
- *la fonction $\kappa(\cdot, \mathbf{x})$ appartient à $\mathcal{H}$, quel que soit $\mathbf{x} \in \mathcal{X}$,*
- *on a $f(\mathbf{x}) = \langle f, \kappa(\cdot, \mathbf{x})\rangle_{\mathcal{H}}$ pour tout $\mathbf{x} \in \mathcal{X}$ et $f \in \mathcal{H}$ (propriété reproduisante).*

On dit alors que $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant (RKHS).

On peut alors exprimer formellement ϕ :

$$\begin{array}{rcl} \phi : \mathcal{X} & \mapsto & \mathcal{H} \\ \mathbf{x} & \to & \kappa(\cdot, \mathbf{x}) \end{array} \tag{3.12}$$

et dans ces conditions, on vérifie alors la propriété fondamentale des méthodes à noyaux :

$$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle_{\mathcal{H}} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \tag{3.13}$$

Le théorème de Mercer assure donc l'existence d'une fonction ϕ dans un espace de Hilbert à noyau reproduisant. De plus, il apparait que le noyau $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ associé à cet espace $\mathcal{H}$ est en réalité le produit scalaire entre les images $\phi(\mathbf{x}_i)$ et $\phi(\mathbf{x}_j)$. Ainsi, les traitements réalisés dans l'espace image $\mathcal{H}$, mis sous forme de produits scalaires, peuvent être effectués sans le calcul des images des éléments. En particulier, il est possible de travailler dans $\mathcal{H}$ car sa structure est accessible à travers sa norme, directement liée au produit scalaire qui, lui, est calculable grâce à la fonction noyau. En se basant sur le théorème de Cover, cet espace $\mathcal{H}$ de grande dimension est plus favorable pour effectuer les calculs. Cette astuce est appelée astuce du noyau ou, suivant la terminologie anglo-saxonne, *kernel trick*.

Théorème du représentant

Nous avons décrit une procédure permettant d'exprimer un problème d'estimation de prédicteur de manière formelle. Cependant, il ne nous est pour le moment pas possible de calculer directement la solution d'un problème tel que (2.19) dans le cadre des méthodes à noyaux. Afin de rendre opérationnelle cette approche, il est nécessaire de s'appuyer sur le Théorème du Représentant 3.3 issu des travaux de Kimeldorf et Wahba dans le domaine de la théorie de l'approximation [KW71, Wah90].

Théorème 3.3 (Théorème du représentant) *Soient :*
- *S un ensemble d'observations $\{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \quad i = 1, \ldots, N\}$,*
- *$\mathcal{L}$ une fonction perte arbitraire,*
- *g une fonction monotone croissante sur $\mathbb{R}^+$,*
- *$\mathcal{H}$ un RKHS induit par le noyau κ défini positif sur $\mathcal{X}$.*

Alors toute fonction $f^* \in \mathcal{H}$ minimisant la fonctionnelle de coût

$$\lambda g\left(\|f\|_{\mathcal{H}}^2\right) + \mathcal{L}\left(\left(f\left(\mathbf{x}_1\right), y_1\right), \ldots, \left(f\left(\mathbf{x}_N\right), y_N\right)\right) \tag{3.14}$$

peut s'écrire sous la forme

$$f^*\left(\cdot\right) = \sum_{i=1}^{N} \beta_i \kappa\left(\mathbf{x}_i, \cdot\right) \tag{3.15}$$

Le problème (2.19) est un cas particulier de la procédure (3.14) ci-dessus. L'optimisation (minimisation) conduisant à l'estimation d'un prédicteur se ramène donc à la résolution d'un problème à N variables : la détermination des coefficients $\beta_1, \ldots, \beta_N$.

Des exemples de fonction perte comme mentionnés dans le théorème 3.3 sont :

L'erreur quadratique moyenne (utilisée pour la régression) :

$$\mathcal{L}\left(\left(f\left(\mathbf{x}_1\right), y_1\right), \ldots, \left(f\left(\mathbf{x}_N\right), y_N\right)\right) = \frac{1}{N} \sum_{i=1}^{N} \left(f(\mathbf{x}_i) - y_n\right)^2 \tag{3.16}$$

L'erreur totale en norme l_1 (utilisée pour la régression) :

$$\mathcal{L}\left(\left(f\left(\mathbf{x}_1\right), y_1\right), \ldots, \left(f\left(\mathbf{x}_N\right), y_N\right)\right) = \sum_{i=1}^{N} \left|f(\mathbf{x}_i) - y_n\right| \tag{3.17}$$

La perte charnière (utilisée pour la classification 2-classes) :

$$\mathcal{L}\left(\left(f\left(\mathbf{x}_1\right), y_1\right), \ldots, \left(f\left(\mathbf{x}_N\right), y_N\right)\right) = \sum_{i=1}^{N} \max(0, 1 - y_n f(\mathbf{x}_i)) \tag{3.18}$$

Résumé

A travers la description des RKHS, deux caractéristiques fondamentales qui justifient l'utilisation de cet outil ont été mises en exergue. Soit la transformation $\mathcal{X} \ni \mathbf{x} \mapsto \phi(\mathbf{x}) \in \mathcal{H}$ où $\mathcal{H}$ est un RKHS, alors les propriétés suivantes sont vraies :

Astuce du noyau (*kernel trick*) Soient $\mathbf{x}_1$ et $\mathbf{x}_2$ des observations dans l'espace $\mathcal{X}$. Comme $\mathcal{H}$ est un espace de Hilbert, il existe un produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ associé à cet espace. Dans le cadre des RKHS, ce produit scalaire est unique et défini par une fonction noyau $\kappa(\cdot, \cdot)$ (*kernel function*) :

$$\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}} = \kappa(\mathbf{x}_1, \mathbf{x}_2) \tag{3.19}$$

Cette propriété permet alors d'utiliser dans un espace image, sans exprimer explicitement l'application, un algorithme mis sous forme de produit scalaire.

Théorème du représentant (*representer theorem*) Soit une fonction coût de la forme :

$$J(f) := \lambda g(\|f\|_{\mathcal{H}}^2) + \frac{1}{N} \sum_{i=1}^{N} c(f, \mathbf{x}_i, y_i) \tag{3.20}$$

Si l'on contraint la solution f dans un RKHS $\mathcal{H}$ muni d'une fonction noyau $\kappa(\cdot, \cdot)$, alors chaque minimiseur de (3.20) peut se mettre sous la forme :

$$f^*(\cdot) = \sum_{i=1}^{N} \beta_i \kappa(\cdot, \mathbf{x}_i) \tag{3.21}$$

La théorie garantit donc une solution au problème de recherche d'un prédicteur de coût minimal.

Notre intérêt doit alors se porter sur le choix d'une fonction perte $\mathcal{L}$ (associant une fonction coût c) et d'une transformation ϕ adaptées. Le choix de la première est déterminé en partie par la tâche à réaliser (régression ou classification). Un bon choix de la fonction perte limite également les risques de sur-apprentissage. Le choix de la transformation est intrinsèquement lié au RKHS retenu, implicitement déterminée par le choix du noyau κ.

3.1.4 Noyaux courants

Nous proposons ici une revue non exhaustive de quelques noyaux couramment utilisés dans le cadre de l'apprentissage statistique.

Noyau linéaire

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{3.22}$$

Il s'agit du produit scalaire usuel, ce noyau permet l'utilisation traditionnelle d'algorithmes, sans astuce du noyau.

Noyau polynomial homogène

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{1}{d}\mathbf{x}_i^T \mathbf{x}_j\right)^p \tag{3.23}$$

Ce noyau permet de transformer des algorithmes linéaires en algorithmes polynomiaux (voir exemple introductif). Tous les monômes d'ordre p sont inclus à la transformation ϕ.

Noyau polynomial inhomogène

$$\left(\frac{c}{d}\mathbf{x}_i^T \mathbf{x}_j + 1\right)^p \tag{3.24}$$

Dans ce cas, la constante additive a pour effet l'ajout de l'ensemble des monômes d'ordre inférieur ou égal à p dans la transformation ϕ.

Noyau Gaussien RBF

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{d\sigma^2}\right) \tag{3.25}$$

On appelle fonction à base radiale, notée RBF (du terme anglo-saxon *Radial Basis Function*), une fonction qui ne dépend que de la distance entre ses arguments. Le noyau Gaussien RBF applique alors une échelle gaussienne sur la distance entre les exemples. L'espace transformé est dans ce cas de dimension infinie. On note en outre que pour le noyau Gaussien RBF, les exemples sont placés sur une sphère de rayon unité dans l'espace des paramètres : $\|\phi(\mathbf{x})\|^2 = \kappa(\mathbf{x}, \mathbf{x}) = 1$, $\forall \mathbf{x}$. Ce noyau est dit de norme unité.

Ces noyaux sont repris de [SBV95] où la dimensionnalité d de l'espace d'observation est compensée par un facteur de normalisation.

3.1.5 Ingénierie des noyaux

Il existe une grande quantité d'autres noyaux dans la littérature [STC04]. De plus il est possible de construire de nouveaux noyaux à partir de noyaux élémentaires.

Propriété 3.1 *Soient κ_1 et κ_2 deux noyaux sur $\mathcal{X} \times \mathcal{X}$, $a \in \mathbb{R}^+$, une application $f : \mathcal{X} \mapsto \mathbb{R}$, une transformation $\phi : \mathcal{X} \mapsto \mathcal{H}$, κ_3 un noyau sur $\mathcal{H} \times \mathcal{H}$, et M une matrice carrée symétrique semi-définie positive. Alors les fonctions suivantes sont des noyaux sur $\mathcal{X} \times \mathcal{X}$:*

1. $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa_1(\mathbf{x}_i, \mathbf{x}_j) + \kappa_2(\mathbf{x}_i, \mathbf{x}_j)$
2. $\kappa(\mathbf{x}_i, \mathbf{x}_j) = a\kappa_1(\mathbf{x}_i, \mathbf{x}_j), a > 0$
3. $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa_1(\mathbf{x}_i, \mathbf{x}_j)\kappa_2(\mathbf{x}_i, \mathbf{x}_j)$
4. $\kappa(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i)f(\mathbf{x}_j)$
5. $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa_3\left(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\right)$
6. $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i M \mathbf{x}_j$

$\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$

Des noyaux peuvent également être construits à partir des observations. [STC04] en donne les clés que nous ne détaillons pas ici. [SS02] constitue également une excellente référence pour approfondir la théorie des noyaux appliquée à l'apprentissage statistique. Pothin [Pot07] évoque également l'interprétation géométrique des différentes règles de construction d'un noyau.

Noyau normalisé

On note enfin qu'un noyau quelconque peut, à l'instar du noyau Gaussien RBF, être transformé en noyau de norme unité. On parle alors de noyau normalisé. Il suffit pour cela de poser :

$$\phi_{norm}(\mathbf{x}) \leftarrow \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|} \tag{3.26}$$

En pratique, cela revient à utiliser le noyau κ_{norm} suivant :

$$\kappa_{norm}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right)}{\sqrt{\kappa\left(\mathbf{x}_i, \mathbf{x}_i\right)\kappa\left(\mathbf{x}_j, \mathbf{x}_j\right)}} \tag{3.27}$$

Noyau centré

Soulignons enfin qu'une opération de centrage des observations dans l'espace de représentation peut être effectuée. Il s'agit, comme le suggère Pothin [Pot07] de transformer les $\phi(\mathbf{x})$ en $\overline{\phi}(\mathbf{x})$ par translation d'un vecteur obtenu comme combinaison linéaire des observations de l'ensemble d'apprentissage. Formellement :

$$\overline{\phi}(\mathbf{x}) \triangleq \phi(\mathbf{x}) - \sum_{i=1}^{N} \gamma_i \phi(\mathbf{x}_i) \tag{3.28}$$

où $\gamma_i \in \mathbb{R}$. Pothin montre que le produit scalaire issu de cette transformation est lié à un noyau $\overline{\kappa}$ défini positif. Dans le cas particulier $\gamma_i = \frac{1}{N}$, la méthode est identique au centrage des données proposé par Cristianini et Shawe-Taylor [CST00, STC04]. Pour un problème 2-classes, choisir

$$\gamma_i = \begin{cases} \frac{1}{2N^+} & \text{si } y_i = +1 \\ \frac{1}{2N^-} & \text{si } y_i = -1 \end{cases} \tag{3.29}$$

où N^+ (respectivement N^-) correspond au nombre d'observations étiquetées $y_i = +1$ (respectivement $y_i = -1$), permet de positionner l'origine des données au centre d'inertie des deux classes. Enfin, il apparait que la matrice de Gram $\overline{\mathbf{K}}$ associée au noyau centré $\overline{\kappa}$ s'exprime :

$$\overline{\mathbf{K}} = \mathbf{K} - \Gamma^T\mathbf{K} - \mathbf{K}\Gamma + \Gamma^T\mathbf{K}\Gamma \tag{3.30}$$

où $\Gamma = [\boldsymbol{\gamma}, \ldots, \boldsymbol{\gamma}] \in \mathbb{R}^{N\times N}$ avec $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^T$.

3.2 Optimisation par la méthode des multiplicateurs de Lagrange

Nous présentons maintenant quelques éléments de théorie de l'optimisation [11] permettant de résoudre les problèmes SVM comme ceux définis ci-après. En particulier, nous allons suivre pour ce rappel le même chemin que celui emprunté par Cristianini et Shawe-Taylor au chapitre 5 de [CST00].

On appelle problème primal un problème qui s'exprime sous la forme générale suivante comprenant une fonction objectif, des contraintes d'égalité et d'inégalité :

$$\begin{array}{rll} \min_{\mathbf{w}} & f(\mathbf{w}), & \mathbf{w} \in \Omega \\ \text{sous les contraintes} & g_i(\mathbf{w}) \leq 0, & i = 1, \ldots, k \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m \end{array} \tag{3.31}$$

où f, g_i et h_i sont des fonctions définies sur le domaine $\Omega \subseteq \mathbb{R}^d$. Ici, $\mathbf{w}$ est un vecteur de variables, dites primales. Lorsque la fonction objectif est quadratique et les contraintes linéaires, ce problème est appelé programme quadratique. De plus, si Ω, la fonction objectif et l'ensemble des contraintes sont convexes, le problème est convexe [12] et sa solution est unique.

La méthode, initialement développée par Lagrange en 1797, est une généralisation du théorème de Fermat en présence de contraintes d'égalité. Elle s'appuie sur une fonctionnelle, appelée Lagrangien et notée L, qui inclut les informations de la fonction objectif et des contraintes :

$$L(\mathbf{w}, \boldsymbol{\beta}) = f(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w}) \tag{3.32}$$

On appelle les coefficients β_i multiplicateurs de Lagrange ou encore variables duales. Le problème exprimé en fonction de ces variables introduites a posteriori est appelé problème dual.

Théorème 3.4 (Théorème de Fermat (1629)) *Une condition nécessaire pour que $\mathbf{w}^* \in \Omega$ soit un minimum de $f(\mathbf{w})$ où $f \in \mathcal{C}^1_\Omega$, est $\frac{\partial f(\mathbf{w}^*)}{\partial \mathbf{w}} = \mathbf{0}$. Cette condition est suffisante lorsque f est convexe.*

Théorème 3.5 (Théorème de Lagrange (1797)) *Des conditions nécessaires pour que $\mathbf{w}^* \in \Omega$ soit un minimum de $f(\mathbf{w})$ sous les contraintes $h_i(\mathbf{w}) = 0$, $i = 1, \ldots, m$, où $f \in \mathcal{C}^1_\Omega$, sont qu'il existe un vecteur β^* tel que*

$$\frac{\partial L(\mathbf{w}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{w}} = \mathbf{0}$$

$$\frac{\partial L(\mathbf{w}^*, \beta_i)}{\partial \beta_i} = 0$$

Ces conditions sont suffisantes lorsque $L(\mathbf{w}, \boldsymbol{\beta})$ est convexe en $\mathbf{w}$.

11. « La théorie de l'optimisation est la branche des mathématiques qui s'intéresse à la caractérisation des solutions de problèmes donnés, et au développement d'algorithme pour les trouver »[CST00]

12. Rappel sur la notion de convexité :

Fonction convexe Une application $f(\mathbf{w})$ dans $\mathbb{R}$ est dite convexe pour $\mathbf{w} \in \mathbb{R}^d$ si, $\forall \mathbf{w}, \mathbf{u} \in \mathbb{R}^d$, et pour tout $\theta \in (0, 1)$, on a $f(\theta\mathbf{w} + (1-\theta)\mathbf{u}) \leq \theta f(\mathbf{w}) + (1-\theta) f(\mathbf{u})$.

Ensemble convexe Un ensemble $\Omega \subseteq \mathbb{R}^d$ est dit convexe si, $\forall \mathbf{w}, \mathbf{u} \in \Omega$, et pour tout $\theta \in (0, 1)$, le point $(\theta\mathbf{w} + (1-\theta)\mathbf{u}) \in \Omega$.

En 1951, Kuhn et Tucker étendent ce résultat au cas le plus général incluant également des contraintes d'inégalité [KT51]. Ainsi, le Lagrangien généralisé s'exprime :

$$L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{w}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w}) \tag{3.33}$$

et nous pouvons alors définir le dual du problème (3.31) :

$$\begin{array}{rl} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} & \inf_{\mathbf{w}\in\Omega} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{sous les contraintes} & \boldsymbol{\alpha} \geq 0 \end{array} \tag{3.34}$$

On appelle écart de dualité (*duality gap*) la différence entre les valeurs des fonctionnelles primale et duale au point d'optimalité $\mathbf{w}^*$.

Dans le cas d'un problème d'optimisation convexe, cet écart est nul lorsque les conditions de Karush-Kuhn-Tucker sont réunies. Par extension, l'écart de dualité désigne également la différence entre deux solutions primales et duales non optimales ; cette seconde définition, comme nous le verrons plus loin, permet de définir un critère d'arrêt pour un algorithme d'optimisation.

Théorème 3.6 (Théorème de Kuhn-Tucker (1951)) *Soit le problème d'optimisation primal* (3.31) *où*

- $\Omega \in \mathbb{R}^d$ *est un domaine convexe,*
- $f \in \mathcal{C}^1_\Omega$ *est une fonction convexe,*
- *les fonctions* g_i *et* h_i *sont affines* $\forall i$.

Alors les conditions nécessaires et suffisantes pour qu'un point $\mathbf{w}^* \in \Omega$ *soit optimal sont l'existence de coefficients* $\boldsymbol{\alpha}^*$ *et* $\boldsymbol{\beta}^*$ *tels que*

$$\begin{array}{rl} \frac{\partial L(\mathbf{w}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)}{\partial \mathbf{w}} = \mathbf{0}, & \\ \frac{\partial L(\mathbf{w}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)}{\partial \beta_i} = 0, & \\ \alpha_i^* g_i(\mathbf{w}^*) = 0, & i = 1, \ldots, k, \\ g_i(\mathbf{w}^*) \leq 0, & i = 1, \ldots, k, \\ \alpha_i^* \geq 0, & i = 1, \ldots, k, \end{array}$$

La littérature fait souvent référence à ces conditions par la terminologie de conditions de Karush-Kuhn-Tucker (KKT). En particulier, la troisième condition est plus connue sous le nom de condition de complémentarité de Karush-Kuhn-Tucker. Elle implique que pour un élément i donné, soit la contrainte est active ($g_i(\mathbf{w}^*) = 0$) et alors le multiplicateur correspondant vaut $\alpha_i^* \geq 0$, soit la contrainte est inactive ($g_i(\mathbf{w}^*) < 0$) et alors le multiplicateur correspondant vaut $\alpha_i^* = 0$.

3.3 SVM à marge maximale

Nous allons maintenant définir trois problèmes SVM [13] : linéaire à marge dure (problème (3.38)), linéaire à marge souple (problème (3.40)) et enfin la forme « canonique » non linéaire à marge souple (problème (3.42)). Notons que la notion de souplesse de la marge traduit la possibilité que des données de l'ensemble d'apprentissage $S = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N\}$, où $\mathcal{Y} = \{-1, +1\}$ soient situées au-delà de cette marge, dont nous donnons la définition par la suite.

13. *Support Vector Machines* traduit Machines à vecteur de support ou Séparateurs à Vaste Marge.

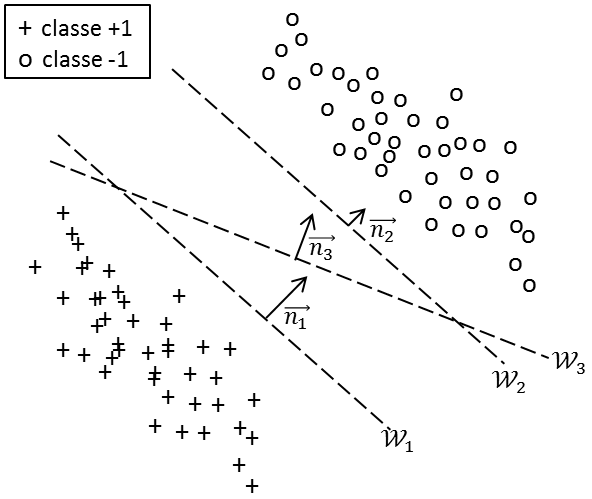


FIGURE 3.2 – Les hyperplans $\mathcal{W}_1$, $\mathcal{W}_2$ et $\mathcal{W}_3$ permettent tous les trois de séparer sans erreur les données des deux classes. Les vecteurs normaux n_1 et n_2 sont également colinéaires.

3.3.1 Cas des SVM linéaires à marge dure

On suppose que l'ensemble S des observations est linéairement séparable. Il existe alors dans $\mathcal{X}$ au moins un hyperplan [14] $\mathcal{W}$ séparant sans erreur les données des classes $+1$ et -1. L'existence d'un tel hyperplan est certaine, par définition de S. En revanche, l'unicité n'est pas garantie (voir figure 3.2). Le principe de marge maximale introduit en 1963 par Vapnik et Lerner [VL63] pose un problème d'optimisation contraint afin de déterminer les paramètres d'un hyperplan séparateur.

L'hyperplan $\mathcal{W}$ est défini par l'équation $\langle \mathbf{w}, \mathbf{x} \rangle_{\mathcal{X}} + b = 0$ où $\mathbf{w}$ est la normale à cet hyperplan et $\frac{b}{\|w\|}$ la distance de $\mathcal{W}$ à l'origine de l'espace $\mathcal{X}$ (b est appelé biais). La fonction de décision, ou prédicteur, prend alors la forme de la fonction $\text{sign}(f(\mathbf{x}))$ où f est l'application suivante :

$$\begin{array}{rccl} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \to & \langle \mathbf{w}, \mathbf{x} \rangle_{\mathcal{X}} + b \end{array} \tag{3.35}$$

On définit la distance d_i entre une observation $(\mathbf{x}_i, y_i)$ et l'hyperplan $\mathcal{W}$ et on appelle marge la valeur M telle que $d_i \geq M$, $\forall i$. On a alors le lien suivant :

$$d_i = \frac{y_i f(\mathbf{x}_i)}{\|\mathbf{w}\|} \geq M \qquad \forall i \tag{3.36}$$

Afin de définir un séparateur de manière unique, on impose :

$$M\|\mathbf{w}\| = 1 \tag{3.37}$$

Le choix de l'hyperplan est alors lié uniquement à la maximisation de la marge $\frac{1}{\|\mathbf{w}\|}$, principe fondamental des SVM. En pratique, il s'agit alors de minimiser $\|\mathbf{w}\|$ et le problème primal

14. Le terme hyperplan est une généralisation en dimension d quelconque de la notion de plan dans l'espace ou encore de droite dans le plan. Autrement dit, cela correspond à une frontière réalisant une partition complète et exclusive de cet (hyper-)espace de dimension d.

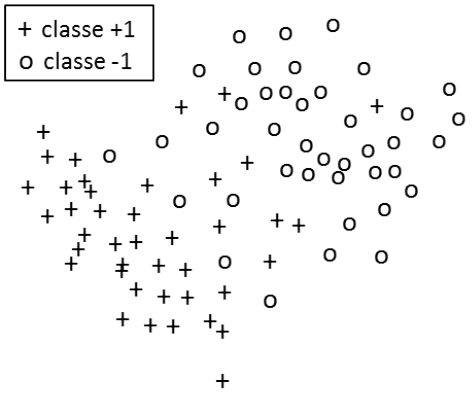


FIGURE 3.3 – Exemple de données non linéairement séparables

de recherche d'un hyperplan séparateur à marge dure, noté HM-SVM (*Hard-Margin* SVM), s'exprime :

$$\begin{array}{rl} \min_{\mathbf{w}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{sous les contraintes} & y_i f(\mathbf{x}_i) \geq 1, \quad \forall i = 1, \ldots, N \end{array} \tag{3.38}$$

3.3.2 Cas des SVM linéaires à marge souple

On suppose maintenant que les données de S ne sont pas linéairement séparables (voir figure 3.3). En conséquence, la contrainte du problème (3.38) ne peut pas être vérifiée et il n'existe alors aucune solution au problème. Pour pallier cette limite, les SVM à marge souple assouplissent cette contrainte par l'introduction de variables de relâchement ξ_i [Smi68, BM92] :

$$y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, N \tag{3.39}$$

Cependant, il existe maintenant pour chaque observation $(\mathbf{x}_i, y_i)$ une valeur de $\xi_i \geq 0$ permettant de vérifier cette nouvelle contrainte, quel que soit $\mathcal{W}$. Le problème primal SM-SVM (*Soft-Margin* SVM), relaxé [15] par les variables ξ_i, devient :

$$\begin{array}{rl} \min_{\mathbf{w}} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{sous les contraintes} & y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, N \\ & \xi_i \geq 0 \end{array} \tag{3.40}$$

Le paramètre C gère le compromis entre maximisation de la marge et minimisation des erreurs. On désigne généralement ce compromis par « souplesse » de la marge. Plus C est élevé, plus la pénalisation est forte pour une donnée mal classée. Lorsque C est grand, on cherche à minimiser le terme d'erreur (quelle que soit la marge) tandis que quand C est petit, on cherche à maximiser la marge (quelles que soient les erreurs).

3.3.3 Extension aux cas des SVM non linéaires

Bien que le théorème de Mercer fut énoncé en 1909 [Mer09], ce n'est qu'en 1964 que Aizerman *et al*[ABR64] utilisèrent une interprétation géométrique des noyaux comme produits scalaires

15. Les variables de relâchement (de la contrainte) sont à ce titre parfois appelées variables de pénalisation (de la fonctionnelle).

d'un espace de paramètres pour l'apprentissage statistique. Il faudra encore attendre plusieurs décennies avant que ces notions soient appliquées aux séparateurs à vaste marge [BGV92, CV95].

Le problème des SVM non linéaires à marge maximale s'exprime en remplaçant le produit scalaire entre éléments de l'espace d'observation $\mathcal{X}$ par le produit scalaire entre les images de ces éléments dans l'espace transformé $\mathcal{H}$, autrement dit par le noyau κ. L'hyperplan séparateur (linéaire) est alors défini dans $\mathcal{H}$ et conduit à une surface de décision non linéaire dans $\mathcal{X}$. La fonction de décision devient l'application suivante :

$$\begin{array}{rccl} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \to & \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \end{array} \tag{3.41}$$

Ici, à la différence du cas linéaire, la normale $\mathbf{w}$ est un élément de l'espace de Hilbert $\mathcal{H}$. Le problème primal SVM non linéaire, noté C-SVM, s'exprime alors :

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 + C\sum_{i=1}^{N}\xi_i \tag{3.42a}$$

$$\text{sous les contraintes} \quad \begin{array}{l} y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, N \\ \xi_i \geq 0 \end{array} \tag{3.42b}$$

3.3.4 Problème dual C-SVM

Nous allons maintenant appliquer au problème SVM (3.42) la méthode des multiplicateurs de Lagrange présentée à la section 3.2. Dans le cas des SVM, on traite un problème d'optimisation quadratique convexe dont seules les contraintes d'inégalité s'appliquent. En pratique, on va exploiter les conditions d'optimalité données par le théorème de Kuhn-Tucker 3.6 pour exprimer le problème C-SVM dual en fonction des seuls multiplicateurs de Lagrange.

Le problème considéré fait apparaître deux types de contraintes d'inégalité, d'où l'ajout d'un second vecteur de multiplicateurs de Lagrange noté $\boldsymbol{\eta}$. De plus, nous avons trois variables primales : $\mathbf{w}$, b et $\boldsymbol{\xi}$. Le Lagrangien correspondant s'exprime alors :

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i\left(\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle_{\mathcal{H}} + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{N}\eta_i\xi_i \tag{3.43}$$

La première condition de KKT indique que les équations suivantes doivent être satisfaites au point d'optimalité :

$$\frac{\partial L\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N}\alpha_i y_i \phi(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \phi(\mathbf{x}) \tag{3.44}$$

$$\frac{\partial L\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right)}{\partial b} = -\sum_{i=1}^{N}\alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{3.45}$$

$$\frac{\partial L\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right)}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \quad \Rightarrow \quad \eta_i = C - \alpha_i \tag{3.46}$$

On reformule le problème C-SVM dual en insérant (3.44) dans (3.43) :

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\right\rangle_{\mathcal{H}} \\ & -\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\right\rangle_{\mathcal{H}} \\ & -b\sum_{i=1}^{N}\alpha_i y_i + \sum_{i=1}^{N}\alpha_i + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\eta_i\xi_i \end{aligned} \tag{3.47}$$

qui devient à l'aide des relations (3.45) et (3.46), et de l'astuce du noyau (3.13) :

$$L(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) + \sum_{i=1}^{N}\alpha_i \tag{3.48}$$

D'après les contraintes du problème C-SVM dual (3.34), les inégalités $\alpha_i \geq 0$ et $\eta_i \geq 0$ doivent être vérifiées, $\forall i$. En considérant la relation (3.46), il apparaît que ces contraintes peuvent se résumer par l'expression $0 \leq \alpha_i \leq C$. La formulation duale du problème C-SVM (3.42) s'exprime alors :

$$\max_{\boldsymbol{\alpha}} \; L(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{3.49a}$$

$$\text{sous les contraintes} \quad 0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, N \tag{3.49b}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

La seconde contrainte est directement issue de la condition (3.45). Pour des raisons de compacité, on peut également mettre ce problème C-SVM sous forme matricielle :

$$\max_{\boldsymbol{\alpha}} \; L(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \mathbf{1}_N^T\boldsymbol{\alpha} \tag{3.50a}$$

$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N C \tag{3.50b}$$

$$\boldsymbol{\alpha}^T\mathbf{y} = 0$$

avec $\mathbf{H}$ l'opposée de la matrice de Gram *polarisée*[16] :

$$\mathbf{H}_{i,j} = -y_i y_j \mathbf{K}_{i,j} = -y_i y_j \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{3.51}$$

Nous n'allons pas aborder d'algorithme de résolution du problème dual maintenant. Néanmoins, nous pouvons noter que disposant d'une solution $\boldsymbol{\alpha}^*$, il est possible, à partir de la relation (3.44), d'exprimer la fonction de décision (3.41) de la manière suivante :

$$\begin{array}{rccc} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \to & \displaystyle\sum_{i=1}^{N}\alpha_i^* y_i \kappa\left(\mathbf{x}_i, \mathbf{x}\right) + b^* \end{array} \tag{3.52}$$

16. Le terme « polarisé » fait référence au fait que la valeur de chaque élément de la matrice tient compte de l'étiquette (ou signe) $+1$ ou -1 associée aux observations en question. Le choix de cette terminologie est à rapprocher des travaux de Baram [Bar05] sur la polarisation des noyaux.

Intéressons-nous un instant à l'activation des contraintes $y_i f(\mathbf{x}_i) \geq 1 - \xi_i$ et $\xi_i \geq 0$ du problème C-SVM (3.42). L'exploitation des conditions KKT permet de dresser le tableau 3.1 duquel on peut extraire une méthode de calcul de b. Ceci nous permet de définir les vecteurs de support comme étant les observations pour lesquelles $\alpha_i > 0$. La fonction de décision repose alors sur ces seules observations (voir figure 3.4). Soit MSV (*Margin Support Vectors*, vecteurs de support sur la marge) l'ensemble de points $\{k, \alpha_k \in]0, C[\}$, alors le biais s'exprime :

$$b^* = y_k - \sum_{i=1}^{N} \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}_k) \quad \forall k \in MSV \tag{3.53}$$

Valeur α_i	Valeur η_i	Contrainte $y_i f(\mathbf{x}_i) \geq 1 - \xi_i$	Contrainte $\xi_i \geq 0$	Observation
$\in]0, C[$	$\in]0, C[$	active	active	Le point $\mathbf{x}_i$ est situé sur la marge. On appelle ce point un vecteur de support sur la marge (MSV).
C	0	active	inactive	Le point $\mathbf{x}_i$ est situé du mauvais côté de la marge. On appelle ce point un vecteur de support au-delà de la marge ou borné (BSV).
0	C	inactive	active	Le point $\mathbf{x}_i$ est situé du bon côté de la marge. Ce point n'est pas vecteur de support (NSV).

TABLE 3.1 – Définition des vecteurs de support en lien avec les contraintes actives ou non du problème SVM dual.

3.4 SVM à volume minimal, ou 1-classe

Schölkopf *et al* ont proposé en 2001 une extension naturelle des SVM permettant d'estimer le support de la distribution d'une loi de probabilité [SPST+01]. Pour cela, ils proposent un algorithme permettant d'estimer une fonction f dont la valeur est positive dans une région de l'espace d'observation qui capture la plupart des données observées, et négative ailleurs. Plus le volume de cette région est petit, tout en garantissant une probabilité qu'une observation dans ce volume appartienne à la classe à apprendre, meilleure est la description ; on appelle cette approche, par opposition aux SVM à marge maximale, les SVM à volume minimal.

Dans le cadre de la classification, il s'agit d'une approche 1-classe [MH96]. Les approches 1-classe ont de nombreuses applications comme la détection de points aberrants (appelés *outliers*), la détection d'anomalies ou encore la détection de nouveauté.

Outre les travaux de Schölkopf [SPST+01], Tax *et al* [TD99, Tax01, TD04] ont également proposé une méthode de modélisation 1-classe basée sur l'approche SVM appelée *Support Vector Data Description* (SVDD). Les approches de Schölkopf et Tax sont équivalentes sous certaines conditions. Nous présentons rapidement la SVDD dans cette section. Notons que notre intérêt

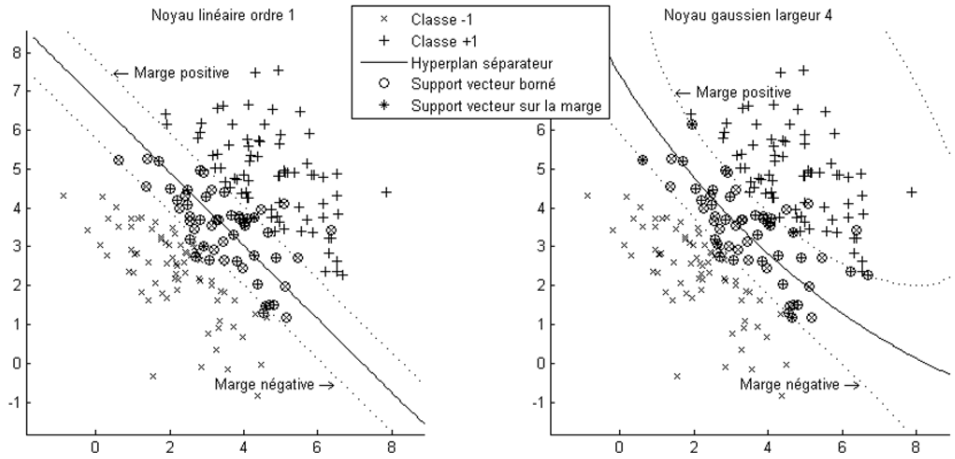


FIGURE 3.4 – Exemple de mise en œuvre de l'algorithme SVM sur des données de test non linéairement séparables avec mise en avant des points vecteurs de support sur ou hors marge. A gauche, utilisation d'un SVM linéaire. A droite, utilisation d'un SVM non linéaire avec noyau Gaussien.

pour ces méthodes 1-classe s'appuie sur les bons résultats obtenus en discrimination ou détection d'anomalie dans des contextes proches de l'application visée : séparation parole/musique [SAH07], la détection de phonèmes [GBT09] ou encore la détection d'événements sonores anormaux [RDR+07].

3.4.1 Principe et formulation du problème SVM 1-classe

La modélisation SVM 1-classe (ou *One-Class SVM*, d'où l'abréviation OC-SVM) proposée par Schölkopf *et al* cherche à estimer la distribution d'une classe de données. Soit un ensemble d'apprentissage $S = \{\mathbf{x}_i, i = 1, \ldots, N\} \in \mathcal{X}$ de N vecteurs d'observation de la classe dont on veut estimer la distribution. Soit Γ la région de volume minimal qui englobe une fraction $(1-\nu)$ des données de cet ensemble. On souhaite alors estimer la fonction $f : \mathcal{X} \mapsto \mathbb{R}$ telle que :

$$\begin{cases} f(\mathbf{x}) & \geq & 0 \text{ si } \mathbf{x} \in \Gamma \\ f(\mathbf{x}) & < & 0 \text{ sinon} \end{cases} \tag{3.54}$$

où f est la fonction de décision suivante :

$$\begin{array}{rccl} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \rightarrow & \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}} - b \end{array} \tag{3.55}$$

La stratégie de l'algorithme consiste à injecter les observations dans un espace $\mathcal{H}$ (RKHS) puis à les isoler par un hyperplan dont la distance à l'origine $\frac{b}{\|\mathbf{w}\|}$ est maximale (voir figure 3.5). Sachant que le problème SVM minimise déjà $\|\mathbf{w}\|$, maximiser cette distance est équivalent à

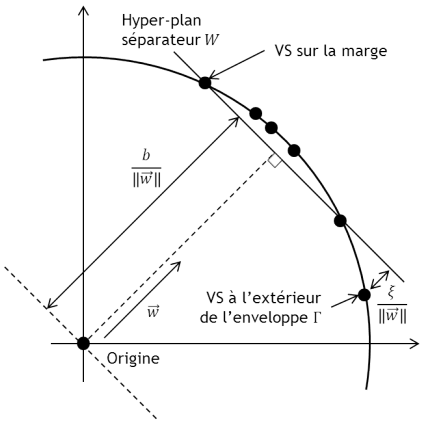


FIGURE 3.5 – Principe des SVM 1-classe : l'hyperplan séparateur optimal est celui qui sépare les données projetées dans l'espace des paramètres de l'origine de cet espace avec un biais maximale.

minimiser $-b$. Le problème SVM 1-classe au sens de Schölkopf (OC-SVM) s'exprime alors :

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 - b + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i \\ \text{sous les contraintes} \quad & f(\mathbf{x}_i) + \xi_i \geq 0, \quad \forall i = 1,\ldots,N \\ & \xi_i \geq 0 \end{aligned} \tag{3.56}$$

On note que le problème (3.56) est similaire au problème (3.42) dans lequel C est remplacé par $\frac{1}{\nu N}$. Ainsi, ce terme dépend, dans le problème de Schölkopf, explicitement de la taille de l'ensemble d'apprentissage et d'un terme $\nu \in [0,1]$ dont le rôle sera interprété par la suite. Enfin, la fonctionnelle de décision issue du problème (3.56) est de la forme (3.52) et le biais b peut être déterminé *a posteriori*.

En suivant le raisonnement décrit précédemment (section 3.3.4), le problème dual OC-SVM s'exprime :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\kappa(\mathbf{x}_i,\mathbf{x}_j) \tag{3.57a}$$

$$\text{sous les contraintes} \quad 0 \leq \alpha_i \leq \frac{1}{\nu N}, \qquad \forall i = 1,\ldots,N \tag{3.57b}$$

$$\sum_{i=1}^{N}\alpha_i = 1$$

ou encore sous forme matricielle :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} \tag{3.58a}$$

$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N\frac{1}{\nu N} \tag{3.58b}$$

$$\boldsymbol{\alpha}^T\mathbf{1}_N = 1$$

avec $\mathbf{H} = -\mathbf{K}$.

Intéressons-nous maintenant au paramètre ν. Lorsque la contrainte $f(\mathbf{x}_i) + \xi_i \geq 0$ est active, alors $\xi_i > 0$ est inactive et $\alpha_i = \frac{1}{\nu N}$. La contrainte $\sum_{i=1}^{N} \alpha_i = 1$ impose de plus qu'il ne peut y avoir que νN points au maximum vérifiant cet état des contraintes, et ν représente donc bien le nombre maximum de points autorisés à être en dehors du volume Γ. A ce titre, les SVM 1-classe, notés dans la suite OC-SVM (*One-Class*), sont parfois référencés dans la littérature comme ν-SVM.

Nous pouvons là encore noter que, disposant d'une solution $\boldsymbol{\alpha}^*$, il est possible d'exprimer la fonction de décision (3.55) de la manière suivante :

$$\begin{array}{rccc} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \rightarrow & \displaystyle\sum_{i=1}^{N} \alpha_i^* \kappa\left(\mathbf{x}_i, \mathbf{x}\right) - b \end{array} \tag{3.59}$$

Support Vector Data Description (minimum enclosing ball)

Alors que Schölkopf *et al* s'intéressent à l'estimation de la distribution d'une classe de données, Tax et Duin introduisent la méthode *Support Vector Data Description* [TD99, Tax01, TD04]. Celle-ci s'appuie sur des boules permettant de décrire les données dans l'espace des paramètres. L'idée est, là encore, de disposer le plus de données de la classe à modéliser dans une boule de plus petit volume possible. Le problème se pose alors sous la forme suivante :

$$\min_{R,c,\boldsymbol{\xi}} R^2 + C \sum_{i=1}^{N} \xi_i$$

$$\text{sous les contraintes } \|\phi(\mathbf{x}_i) - c\|^2 \leq R^2 + \xi_i, \quad \forall i = 1, \ldots, N$$
$$\xi_i \geq 0$$

où R et c sont respectivement le rayon et le centre de la boule recherchée. La solution est alors de la forme :

$$c = \sum_{i=1}^{N} \beta_i \phi(\mathbf{x}_i) \tag{3.60}$$

et R^2 la distance (au carré) entre c et n'importe quel vecteur $\phi(\mathbf{x}_i)$ situé sur la frontière de la boule (calculé *a posteriori*). Il apparait alors une équivalence entre SVDD et OC-SVM dans le cas où $\kappa(\mathbf{x}, \mathbf{x})$ est constant. Schölkopf a montré que c'est le cas pour l'ensemble des noyaux qui dépendent seulement de $\mathbf{x}_i - \mathbf{x}_j$, tel que le noyau Gaussien RBF [SPST$^+$01]. On montre aisément également qu'un noyau quelconque peut satisfaire cette condition si ce dernier est normalisé (voir l'expression (3.27)).

La figure 3.6 illustre cette équivalence. Cette dernière justifie que l'on ne traite dans la suite du document que de la méthode OC-SVM.

3.4.2 Extension au cas SVM 1-classe avec contraintes binaires

Tohmé a proposé une extension du problème SVM 1-classe prenant en compte les contraintes du problème SVM de classification binaire, noté OC2-SVM [TL11]. Dans cette optique, les observations de la classe d'intérêt sont étiquetées $+1$ tandis que les autres observations sont étiquetées -1. Le bénéfice attendu est de trouver pour une classe donnée le classificateur 1-classe qui isole au mieux les observations de la classe d'intérêt tout en rejetant les données

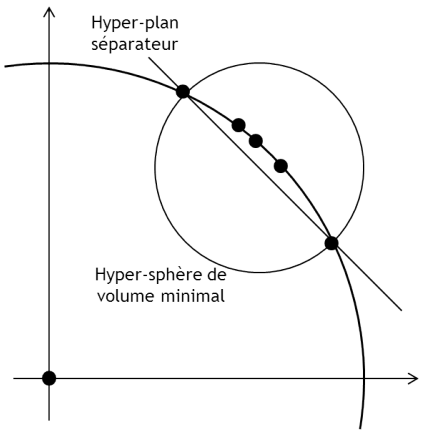


FIGURE 3.6 – Equivalence entre OC-SVM et SVDD. Le sous-espace délimité est identique du point de vue de l'hyperplan séparateur de marge maximale et de la boule de volume minimal.

aberrantes et les observations des autres classes. Le problème primal OC2-SVM correspondant est :

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 - b + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i \tag{3.61a}$$

$$\text{sous les contraintes } f(\mathbf{x}_i)y_i + \xi_i \geq 0, \quad \forall i = 1,\ldots,N \tag{3.61b}$$

$$\xi_i \geq 0$$

Il est intéressant de noter ici que ce problème OC2-SVM est en fait une généralisation du problème SVM 1-classe tel qu'il avait été proposé par Schölkopf *et al*[SPST+01]. Il suffit en effet de ne considérer que des données de la classe d'intérêt, étiquetées +1, et donc de poser $y_i = 1 \forall i$, ou encore $\mathbf{y} = \mathbf{1}_N$, pour retrouver le problème initial. Ceci signifie qu'un algorithme de résolution pour le problème SVM 1-classe avec contraintes binaires peut s'appliquer également au problème SVM 1-classe sans contraintes binaires. Ce constat est le point de départ pour la recherche d'un problème SVM unifié tel qu'il est décrit par la suite.

Le problème dual OC2-SVM s'exprime :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \kappa(\mathbf{x}_i,\mathbf{x}_j)$$

$$\text{sous les contraintes} \quad 0 \leq \alpha_i \leq \frac{1}{\nu N}, \qquad \forall i = 1,\ldots,N$$

$$\sum_{i=1}^{N}\alpha_i y_i = 1$$

ou encore sous forme matricielle :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} \tag{3.62a}$$

$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N\frac{1}{\nu N} \tag{3.62b}$$

$$\boldsymbol{\alpha}^T\mathbf{y} = 1$$

avec **H** l'opposée de la matrice de Gram polarisée telle que définie à l'équation (3.51). Là encore, la fonction de décision (3.55) introduite *a priori* peut s'exprimer en fonction des variables duales :

$$\begin{array}{rccc} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \rightarrow & \sum_{i=1}^{N} \alpha_i^* y_i \kappa\left(\mathbf{x}_i, \mathbf{x}\right) - b \end{array} \tag{3.63}$$

Enfin, on note qu'il est possible de déterminer b à partir de l'ensemble des points situés sur la marge. Soit MSV cet ensemble de points : $MSV = \{k, \alpha_k \in]0, C[\}$. Alors le biais s'exprime :

$$b = \sum_{i=1}^{N} \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_k) \quad \forall k \in MSV \tag{3.64}$$

3.5 Problèmes SVM sans biais

Les approches SVM 2-classes et 1-classe introduites jusqu'à présent s'appuient sur une définition *a priori* de la fonction de décision. En particulier, l'expression de celle-ci est motivée par des considérations géométriques puisqu'elle s'appuie sur la normale $\mathbf{w}$ et la distance à l'origine ou biais b, d'un hyperplan séparateur défini par $\{\langle \mathbf{w}, \phi(\cdot)\rangle_{\mathcal{H}} + b = 0\}$. Le cadre formel assure ensuite l'unicité de la solution recherchée. On dit de cette approche qu'elle est « avec biais ».

Il existe néanmoins une approche différente des méthodes SVM. Celle-ci s'appuie sur les éléments d'apprentissage statistique présentés au chapitre 2. Dans ce contexte, la fonction de décision recherchée est la solution d'un problème de minimisation d'un risque empirique et de régularisation, posé *a priori*. L'expression de la fonction de décision est alors donnée *a posteriori* par le théorème du représentant 3.3, et ne fait pas apparaître de terme de biais. On dit de cette approche qu'elle est « sans biais ».

3.5.1 Biais et SVM

Dans les problèmes SVM 2-classes et 1-classe, le terme de biais issu des fonctions de décision (3.41) et (3.55) a pour effet, à travers la condition de KKT (3.45), d'ajouter une contrainte d'égalité aux problèmes duaux 2-classes (3.49) et 1-classe (3.57). Cette contrainte rend l'optimisation par décomposition des problèmes cités délicate. En effet, la réalisation d'une étape d'optimisation ne peut se faire qu'en modifiant au minimum deux coefficients α_i (multiplicateurs de Lagrange) simultanément afin de conserver l'égalité.

S'affranchir du terme de biais dans les problèmes SVM a été largement exploré dans la littérature récente [KVH03, GAAVO08, DGI11, SHS11]. Bien qu'il soit admis que cela conduise à des solutions avec plus de vecteurs de support [HK04], cela permet d'exploiter des techniques de résolution par décomposition tels que l'ISDA [HK04], ou d'améliorer le fonctionnement du populaire SMO [Pla98] en optimisant simultanément un nombre variable (1, 2 ou plus) de paramètres [SHS11].

Une autre approche ayant pour objectif de s'affranchir de la contrainte d'égalité dans le problème dual a également été proposée, sans supprimer le terme de biais du problème primal. Celle-ci consiste à ajouter un terme $\frac{b^2}{2}$ dans la fonctionnelle de coût du problème primal [FCC98, MM99]. Cette pénalisation du biais permet d'obtenir également une expression analytique du terme b. Cette approche a été utilisée avec succès pour proposer une approche d'optimisation séquentielle, l'algorithme SMGO [TL09]. Utilisant la même astuce, le logiciel BSVM [HL02] se positionne comme une alternative au populaire SVM$^{\text{light}}$ [Joa98].

Il apparait à travers cette littérature que le terme de biais, et la contrainte sous-jacente d'égalité, est limitant pour la résolution d'un problème SVM. Steinwart a remis en cause cette

approche [Ste03] puis proposé une méthode SVM sans biais adaptée au cas 2-classes [SHS11]. Nous inspirant de cette approche, nous montrons maintenant que les problèmes SVM 2-classes et 1-classe peuvent s'exprimer sous la forme d'un problème sans biais.

3.5.2 Approche sans biais du problème SVM 2-classes

On généralise le problème (2.19) de minimisation du risque empirique pour la classification pondérée :

$$f^* = \arg\min_{f\in\mathcal{H}} \lambda g\left(\|f\|_{\mathcal{H}}^2\right) + \frac{1}{N}\sum_{i=1}^{N} \omega_i c\left(f, \mathbf{x}_i, y_i\right) \tag{3.65}$$

avec $\mathcal{H}$ un RKHS, où les ω_i sont des poids associés aux observations $(\mathbf{x}_i, y_i)$. Typiquement, dans le cadre de la classification, on pose :

$$\omega_i = \begin{cases} \omega_{pos} \text{ si } y_i = +1 \\ \omega_{neg} \text{ si } y_i = -1 \end{cases} \tag{3.66}$$

avec $\omega_{pos} > 0$ et $\omega_{neg} > 0$. Afin de retrouver une formulation proche des problèmes SVM déjà exposés, on reformule (3.65) :

$$f^* = \arg\min_{f\in\mathcal{H}} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \sum_{i=1}^{N} C_i c\left(f, \mathbf{x}_i, y_i\right) \tag{3.67}$$

où les $C_i = \frac{\omega_i}{2\lambda N}$ contrôlent le poids associé aux données, et le compromis entre maximisation de la marge et minimisation des erreurs.

D'après le théorème du représentant 3.3, une solution $f^* \in \mathcal{H}$ au problème (3.67) s'exprime sous la forme :

$$f^*(\cdot) = \sum_{i=1}^{N} \beta_i \kappa\left(\mathbf{x}_i, \cdot\right) \tag{3.68}$$

Cette solution, sans biais, est équivalente aux fonctions de décision (3.55) et (3.41) lorsque $b = 0$.

L'utilisation d'une fonction perte adaptée va permettre d'immerger le concept SVM dans l'approche sans biais de l'apprentissage.

La fonction perte du problème SVM 2-classes (3.42) est implicitement exprimée au travers des contraintes (3.43). Sa forme explicite est la suivante :

$$c(f, \mathbf{x}_i, y_i) = \xi_i = \max(0, 1 - y_i f(\mathbf{x}_i)) \tag{3.69}$$

Cette fonction, appelée perte charnière (*hinge loss* selon la terminologie anglo-saxonne), est représentée sur la figure 3.7.

On substitue la perte charnière (3.69) dans le problème (3.67) au travers des contraintes d'inégalité sur les variables de relâchement ξ_i. On peut alors formuler le problème primal sans biais, noté WO-SVM (*Without Offset SVM*) en référence à l'article de Steinwart [SHS11] :

$$\min_{f\in\mathcal{H}} \quad \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \sum_{i=1}^{N} C_i \xi_i \tag{3.70a}$$

$$\text{sous les contraintes} \quad \xi_i \geq 1 - y_i f(\mathbf{x}_i), \quad \forall i = 1, \ldots, N \tag{3.70b}$$
$$\xi_i \geq 0$$

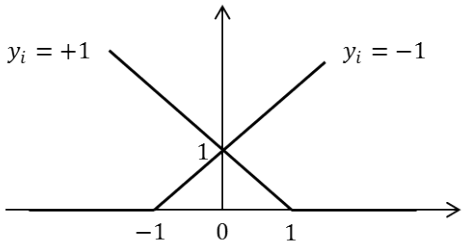


FIGURE 3.7 – Représentation de la perte charnière pour SVM discriminants $c(f, \mathbf{x}_i, y_i) = \max(0, 1 - y_i f(\mathbf{x}_i))$

Le Lagrangien correspondant s'exprime alors :

$$L(f, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \sum_{i=1}^{N} C_i \xi_i - \sum_{i=1}^{N} \alpha_i (\xi_i - 1 + y_i f(\mathbf{x}_i)) - \sum_{i=1}^{N} \eta_i \xi_i \tag{3.71}$$

où $\boldsymbol{\alpha}$ et $\boldsymbol{\eta}$ sont les vecteurs de variables duales.

Afin d'appliquer le théorème de Kuhn-Tucker (théorème 3.6), la notion de dérivabilité d'une fonctionnelle $f \in \mathcal{H}$ doit être précisée. On se contente [17] de s'appuyer sur la propriété reproduisante (3.5) des RKHS $f(\mathbf{x}) = \langle f, \kappa(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$, qui conduit au sens de Fréchet à :

$$\nabla_f f(\mathbf{x}) = \kappa(\cdot, \mathbf{x}) \tag{3.72}$$

Dès lors, il est possible de dériver une fonctionnelle dépendant de $f(\mathbf{x})$. L'analyse du Lagrangien (3.71) permet d'obtenir les conditions d'optimalité de KKT suivantes :

$$\nabla_f L\,(f, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = f - \sum_{i=1}^{N} \alpha_i y_i \kappa(\cdot, \mathbf{x}) = 0 \qquad \Rightarrow \qquad f = \sum_{i=1}^{N} \alpha_i y_i \kappa(\cdot, \mathbf{x}) \tag{3.73}$$

$$\nabla_{\boldsymbol{\xi}} L\,(f, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = C - \alpha_i - \eta_i = 0 \qquad \Rightarrow \qquad \eta_i = C - \alpha_i \tag{3.74}$$

qui nous permettent d'exprimer le Lagrangien en fonction des seules variables duales :

$$L(\boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \kappa\,(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N} \alpha_i$$

La formulation duale du problème (3.70) s'exprime alors :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \kappa\,(\mathbf{x}_i, \mathbf{x}_j) \tag{3.75a}$$

$$\text{sous les contraintes} \quad 0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, N \tag{3.75b}$$

On peut également mettre ce problème sous forme matricielle :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} + \mathbf{1}_N^T \boldsymbol{\alpha} \tag{3.76a}$$

$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N C \tag{3.76b}$$

17. Nous invitons le lecteur intéressé à se référer à la littérature concernant la dérivée de Fréchet pour plus de détails ou l'ouvrage de Debnath et Mikusinski [DM99].

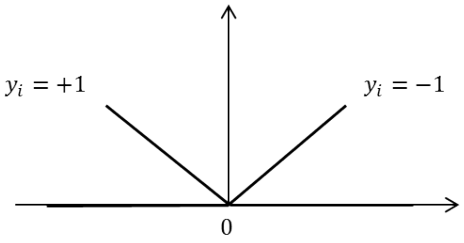


FIGURE 3.8 – Représentation de la perte charnière pour SVM 1-classe $c(f, \mathbf{x}_i, y_i) = \max(0, -y_i f(\mathbf{x}_i))$

avec $\mathbf{H}$ l'opposée de la matrice de Gram polarisée :

$$\mathbf{H}_{i,j} = -y_i y_j \mathbf{K}_{i,j} = -y_i y_j \kappa(\mathbf{x}_i, \mathbf{x}_j)$$

On constate qu'à l'exception de la contrainte d'égalité, ce problème est identique au problème dual SVM 2-classes avec biais (3.50).

Enfin, notons que les équations (3.72) et (3.68) nous permettent d'exprimer *a posteriori* la fonction de décision, en fonction d'une solution $\boldsymbol{\alpha}^*$:

$$\begin{array}{rccc} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \to & \displaystyle\sum_{i=1}^{N} \alpha_i^* y_i \kappa(\mathbf{x}, \mathbf{x}_i) \end{array} \tag{3.77}$$

Ce résultat est conforme au résultat attendu (3.68) et on identifie $\beta_i = \alpha_i y_i$, $\forall i$.

3.5.3 Application au cas 1-classe de l'approche sans biais

De la même manière que précédemment, il est possible d'exprimer la fonction perte associée au problème SVM 1-classe (3.61) à partir des contraintes suivantes :

$$\begin{cases} y_i f(\mathbf{x}_i) + \xi_i \geq 0 \\ \xi_i \geq 0 \end{cases}$$

La forme explicite de la fonction perte est alors la suivante :

$$c(f, \mathbf{x}_i, y_i) = \xi_i = \max(0, -y_i f(\mathbf{x}_i)) \tag{3.78}$$

Il s'agit là encore d'une fonction perte charnière. Celle-ci est représentée sur la figure 3.8.

Malheureusement, il n'est pas possible d'exploiter directement cette fonction de perte dans le cadre SVM 1-classe comme nous l'avons fait pour le problème SVM 2-classes. En effet, rappelons que l'objectif est ici de minimiser un volume, ce qui se traduit dans la fonctionnelle d'optimisation (3.61) par la maximisation du terme de biais b. Or le problème d'apprentissage (3.65) n'inclut pas ce dernier. Nous proposons alors de réécrire la fonction de perte SVM 1-classe de la manière suivante :

$$c(f, \mathbf{x}_i, y_i) = \max(0, -y_i f(\mathbf{x}_i)) = \max(0, y_i - y_i(f(\mathbf{x}_i) + 1)) \tag{3.79}$$

Cette astuce introduit une fonction charnière, fonction de $f(\mathbf{x}_i) + 1$, permettant alors de maximiser une marge au sens de Vapnik entre l'origine de $\mathcal{H}$ et les données de la classe à modéliser. Cette fonction charnière est représentée sur la figure 3.9.

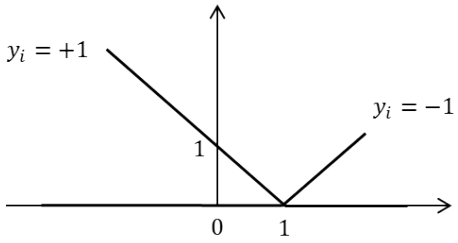


FIGURE 3.9 – Représentation de la perte charnière pour SVM 1-classe avec approche marge maximale $c(f, \mathbf{x}_i, y_i) = \max(0, y_i - y_i(f(\mathbf{x}_i) + 1))$

Maximiser cette marge est strictement équivalent à la maximisation du biais. Ainsi, l'échelle étant implicitement fixée telle que la distance de la marge à l'hyperplan valle 1, il en résulte immédiatement que la distance de l'hyperplan à l'origine sera 1. Il n'y a donc plus de biais à calculer.

Soit $\tilde{f}(\cdot)$ la fonction $f(\cdot) + 1$. On substitue alors la perte charnière (3.79) dans le problème (3.67) à travers les contraintes d'inégalité sur les variables de relâchement ξ_i. Le problème OC2-SVM sans biais, noté WOOC2-SVM (*Without Offset OC2-SVM*), s'exprime alors :

$$\min_{f \in \mathcal{H}} \quad \frac{1}{2}\|\tilde{f}\|_{\mathcal{H}}^2 + \sum_{i=1}^{N} C_i \xi_i \tag{3.80a}$$

$$\text{sous les contraintes} \quad \xi_i \geq y_i - y_i \tilde{f}(\mathbf{x}_i), \quad \forall i = 1, \ldots, N \tag{3.80b}$$

$$\xi_i \geq 0$$

Le Lagrangien s'exprime alors :

$$L(f, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{1}{2}\|\tilde{f}\|_{\mathcal{H}}^2 + \sum_{i=1}^{N} C_i \xi_i - \sum_{i=1}^{N} \alpha_i (\xi_i - y_i + y_i \tilde{f}(\mathbf{x}_i)) - \sum_{i=1}^{N} \eta_i \xi_i \tag{3.81}$$

où $\boldsymbol{\alpha}$ et $\boldsymbol{\eta}$ sont les vecteurs de variables duales. On obtient ensuite les conditions d'optimalité suivantes :

$$\begin{aligned} \nabla_f L\left(f, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right) = \tilde{f} - \sum_{i=1}^{N} \alpha_i y_i \kappa(\cdot, \mathbf{x}) = 0 \quad &\Rightarrow \quad \tilde{f} = \sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i) \\ \nabla_{\boldsymbol{\xi}} L\left(f, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right) = C - \alpha_i - \eta_i = 0 \quad &\Rightarrow \quad \eta_i = C - \alpha_i \end{aligned}$$

qui nous permettent d'exprimer le Lagrangien en fonction des seules variables duales :

$$L(\boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) + \sum_{i=1}^{N} \alpha_i y_i$$

La formulation duale du problème WOOC2-SVM (3.80) s'exprime alors :

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i y_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \tag{3.82a}$$

$$\text{sous les contraintes} \quad 0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, N \tag{3.82b}$$

On peut également mettre ce problème sous forme matricielle :

$$\max_{\boldsymbol{\alpha}} \; L(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{H}\boldsymbol{\alpha} + \mathbf{y}^T\boldsymbol{\alpha} \tag{3.83a}$$

$$\text{sous les contraintes} \;\; \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N C \tag{3.83b}$$

avec $\mathbf{H}$ l'opposée de la matrice de Gram polarisée (3.51).

Rappelons enfin que l'astuce introduite permet de déterminer $\tilde{f} = f(\cdot)+1$. Soit $\boldsymbol{\alpha}^*$ la solution obtenue, la fonction de décision à appliquer dans le cas 1-classe est alors $f(\cdot) = \tilde{f} - 1$:

$$\begin{array}{rccl} f: & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \to & \displaystyle\sum_{i=1}^{N} \alpha_i^* y_i \kappa(\mathbf{x}, \mathbf{x}_i) - 1 \end{array} \tag{3.84}$$

Chapitre 4

Résolution d'un problème SVM unifié

Au cours du chapitre précédent, nous avons introduit 5 problèmes SVM différents : C-SVM, OC-SVM, OC2-SVM, WO-SVM et WOOC2-SVM. Nous montrons dans ce chapitre qu'il existe une forme unifiée à l'ensemble des problèmes duaux associés.

Une fois le problème SVM unifié présenté, nous développons un algorithme de résolution. Cet algorithme unique permet de traiter facilement l'ensemble des problèmes étudiés. Nous montrons que cet algorithme permet de réaliser un démarrage à chaud. Cette propriété apporte la possibilité de passer d'un problème à un autre (changement de paradigme ou encore d'ensemble d'apprentissage) sans nécessairement relancer un apprentissage depuis une condition initiale standard.

4.1 Proposition d'un problème SVM unifié

4.1.1 Avant-propos

Les problèmes SVM que nous avons traités au chapitre précédent sont par essence assez différents. Néanmoins, ils appartiennent tous à une même famille de problèmes : l'optimisation quadratique (*quadratic programming*). D'autre part nous avons, pour l'ensemble de ces problèmes, exploité la même technique des multiplicateurs de Lagrange afin de les exprimer sous une forme dite duale. Finalement, les formulations de ces problèmes duaux ne sont pas très différentes.

Cette ressemblance entre problèmes duaux a déjà été mise en évidence dans la littérature. Cependant, il existe une grande quantité de problèmes dérivés de l'approche SVM et la proposition d'un problème généralisé ou unifié ne sous-entend pas qu'il s'agit d'un problème universel [GAVC05, TMK12]. Il s'agit généralement de trouver une forme canonique à un sous-ensemble de problèmes afin de les comparer ou de les faire cohabiter. Notons que Kivinen *et al* ont exploré une formulation unifiée des problèmes SVM 2-classes, 1-classe et régression au travers de l'approche statistique et en se basant notamment sur l'étude de la fonction perte [KSW01, KSW10]. C'est ce dernier point qui a motivé les travaux que nous présentons dans ce chapitre.

4.1.2 Problème SVM unifié

Le tableau 4.1 rappelle les formulations duales des problèmes d'optimisation présentés au cours du chapitre précédent. Ces problèmes consistent tous en la maximisation d'une fonction-

nelle de coût sous contraintes d'inégalité, et pour certains sous contrainte d'égalité également. Les similarités ne s'arrêtent pas là et nous montrons qu'il est possible de réunir ces cinq problèmes sous une forme unique.

Le premier constat réalisé est la stricte équivalence entre les contraintes d'inégalité de l'ensemble des problèmes considérés. Dans les cas OC-SVM et OC2-SVM, il suffit de poser $C = \frac{1}{\nu N}$ pour s'en convaincre. Le second constat immédiat est la stricte équivalence entre les fonctionnelles de coût des formes duales des problèmes C-SVM et WO-SVM.

On s'intéresse maintenant à l'équivalence entre les problèmes OC-SVM et OC2-SVM. Dans le cas OC-SVM, seuls les éléments de la classe à modéliser sont disponibles. Autrement dit, l'étiquette associée à chacune des observations est identique. Trivialement, il suffit de poser $y_i = 1$, $\forall i$ dans le problème OC2-SVM pour retrouver l'ensemble des éléments du problème OC-SVM. Le problème OC-SVM n'a donc plus lieu d'être étudié sous cette forme.

L'ensemble de ces problèmes SVM recherche un extremum de la fonctionnelle de coût. L'ajout d'une constante à l'une de ces fonctionnelles ne modifie en rien la solution du problème d'optimisation. Ainsi, la fonctionnelle du problème OC2-SVM peut se mettre sous la forme $\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + 1$.

Problème	Fonctionnelle de coût	Contraintes	Fonction de décision
C-SVM [CV95]	$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \mathbf{1}_N^T\boldsymbol{\alpha}$	$\mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N C$ $\boldsymbol{\alpha}^T\mathbf{y} = 0$	$\sum_{i=1}^N \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}) + b$ $b = y_k - \sum_{i=1}^N \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_k), \forall k \in MSV$
OC-SVM [SPST+01]	$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha}$	$\mathbf{0}_N \leq \boldsymbol{\alpha} \leq \frac{\mathbf{1}_N}{\nu N}$ $\boldsymbol{\alpha}^T\mathbf{1}_N = 1$	$\sum_{i=1}^N \alpha_i^* \kappa(\mathbf{x}_i, \mathbf{x}) - b$ $b = \sum_{i=1}^N \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}_k) \quad \forall k \in MSV$
OC2-SVM [TL11]	$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha}$	$\mathbf{0}_N \leq \boldsymbol{\alpha} \leq \frac{\mathbf{1}_N}{\nu N}$ $\boldsymbol{\alpha}^T\mathbf{y} = 1$	$\sum_{i=1}^N \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}) - b$ $b = \sum_{i=1}^N \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_k) \quad \forall k \in MSV$
WO-SVM [SHS11]	$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \mathbf{1}_N^T\boldsymbol{\alpha}$	$\mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{C}$	$\sum_{i=1}^N \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x})$
WOOC2-SVM	$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \mathbf{y}^T\boldsymbol{\alpha}$	$\mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{C}$	$\sum_{i=1}^N \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}) - 1$

TABLE 4.1 – Synthèse des problèmes SVM : C-SVM, OC-SVM, OC2-SVM, WO-SVM et WOOC2-SVM.

Compte-tenu de la contrainte d'égalité qui s'applique au problème, cette dernière est donc équivalente à la fonctionnelle de coût du problème WOOC2-SVM.

Soit $\mathbb{1}_{OC}$ une variable qui vaut 1 lorsque le problème traité est un problème 1-classe et 0 dans le cas 2-classes. On pose également $\boldsymbol{\delta} = \mathbb{1}_{OC}\mathbf{y} + (1 - \mathbb{1}_{OC})\mathbf{1}_N$; ainsi $\delta_i = y_i$ dans le cas 1-classe et $\delta_i = 1$ dans le cas 2-classes. Le problème SVM unifié, noté UNI-SVM, s'exprime alors :

$$\max_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\delta}^T\boldsymbol{\alpha} \tag{4.1a}$$

$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{C} \qquad \text{dans tous les cas} \tag{4.1b}$$

$$\text{et} \quad \boldsymbol{\alpha}^T\mathbf{y} = \mathbb{1}_{OC} \qquad \text{si SVM avec biais} \tag{4.1c}$$

Enfin, les fonctions de décision peuvent se mettre sous une forme unique :

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}) + b \tag{4.2}$$

avec :

$$b = \begin{cases} y_k - b_0 & \text{dans le cas C-SVM} \\ -b_0 & \text{dans les cas OC-SVM et OC2-SVM} \\ 0 & \text{dans le cas WO-SVM} \\ -1 & \text{dans le cas WOOC2-SVM} \end{cases} \tag{4.3}$$

où $b_0 = \sum_{i=1}^{N} \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_k), \forall k \in MSV$.

4.2 Algorithme de résolution pour les SVM

Lors de la résolution du problème (4.1), nous sommes confrontés à la manipulation et au stockage de la matrice $\mathbf{H}$ de dimension $N \times N$, où N est le nombre d'observations de l'ensemble d'apprentissage. Afin de limiter l'occupation de l'espace mémoire, on s'intéresse aux méthodes par décomposition, reposant sur l'optimisation successive d'un nombre réduit de multiplicateurs de Lagrange (coefficients α_i). L'idée est de décomposer le vecteur $\boldsymbol{\alpha}$ en deux sous-ensembles : un jeu de variables libres et un jeu de variables fixes. On résout ensuite successivement des sous problèmes liés à des ensembles de variables différents jusqu'à ce que l'optimum global soit atteint.

L'un des solveurs les plus populaires est le SMO [Pla98]. Celui-ci consiste à optimiser deux variables duales à chaque itération afin de pouvoir satisfaire la contrainte d'égalité du problème avec biais. [TL11] a proposé un solveur adapté au problème SVM 1-classe avec biais et contraintes binaires, appelé *Fast*-OC_2. Enfin, [TL09] et [SHS11] ont proposé deux solveurs permettant la résolution de problème SVM sans biais, très semblables dans leur formulation, que nous nommons respectivement SMGO et SMO-*like*. Nous proposons dans ce qui suit un algorithme basé sur l'optimisation séquentielle d'un nombre restreint de coefficients (variables duales) de la solution, qui de même que le problème SVM unifié, est une synthèse des approches *Fast*-OC_2, SMGO et SMO-*like*.

4.2.1 Principe de l'algorithme

L'algorithme que nous décrivons est, à l'image de ceux évoqués précédemment, un algorithme séquentiel. L'objectif est de réitérer certaines étapes permettant de faire évoluer une

solution courante $\boldsymbol{\alpha}$ vers la solution optimale $\boldsymbol{\alpha}^*$. Ainsi, à chaque itération de l'algorithme, nous considérons que la solution est mise à jour suivant l'expression suivante :

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \lambda \mathbf{u} \tag{4.4}$$

avec $\lambda > 0$ un coefficient à déterminer, appelé pas, et $\mathbf{u}$ une direction à choisir pour l'optimisation. De plus, on appelle ensemble de travail (*working-set* selon la terminologie anglo-saxonne) l'ensemble des éléments non-nuls de $\mathbf{u}$.

Le problème (4.1) consiste en la recherche d'un maximum global d'une fonctionnelle de coût quadratique. Dans ce contexte, le gradient est un outil essentiel : sa direction correspond à la pente maximale. Son étude nous permet de déterminer la direction $\mathbf{u}$ à suivre pour mettre à jour la solution. Pour le problème à traiter, le gradient s'exprime :

$$\mathbf{g} = \mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\delta} \tag{4.5}$$

S'agissant d'un algorithme séquentiel, les différentes grandeurs doivent être mises à jour à chaque itération. Ainsi, l'incrément de la fonction coût vaut :

$$\Delta L = L(\boldsymbol{\alpha} + \lambda \mathbf{u}) - L(\boldsymbol{\alpha}) = \frac{1}{2}\lambda^2 \mathbf{u}^T \mathbf{H} \mathbf{u} + \lambda \mathbf{u}^T \mathbf{g} \tag{4.6}$$

et l'incrément de son gradient vaut :

$$\Delta \mathbf{g} = \mathbf{g}(\boldsymbol{\alpha} + \lambda \mathbf{u}) - \mathbf{g}(\boldsymbol{\alpha}) = \lambda \mathbf{H} \mathbf{u} \tag{4.7}$$

On note également qu'il est possible de déduire de l'équation (4.6) une expression analytique du pas optimal λ^*, pour lequel ΔL est maximum. En effet, d'après le théorème 3.4, $\Delta L(\lambda)$ est maximal lorsque sa dérivée s'annule :

$$\frac{\partial \Delta L}{\partial \lambda}(\lambda^*) = \lambda^* \mathbf{u}^T \mathbf{H} \mathbf{u} + \mathbf{u}^T \mathbf{g} = 0 \quad \Rightarrow \quad \lambda^* = -\frac{\mathbf{u}^T \mathbf{g}}{\mathbf{u}^T \mathbf{H} \mathbf{u}} \tag{4.8}$$

L'algorithme de résolution proposé consiste à substituer (4.8) dans (4.4), puis de mettre à jour le gradient (4.5) à l'aide de l'équation (4.7) afin de calculer un nouveau pas. Ainsi, la complexité de calcul pour une itération est directement liée au calcul de la matrice $\mathbf{H}$. Une manière efficace de contrôler cette complexité est alors d'imposer que le vecteur direction $\mathbf{u}$ soit parcimonieux ou *sparse*[18].

4.2.2 Ensemble de travail et direction d'optimisation

Les expressions (4.7) et (4.8) peuvent donc être calculées en ne disposant que des seules colonnes de la matrice $\mathbf{H}$ correspondant aux éléments de l'ensemble de travail. Ce dernier doit être choisi de manière optimale à chaque itération afin d'apporter un gain important à la fonction coût en ne modifiant qu'un nombre restreint de coefficients α_i de la solution. Le choix de l'ensemble de travail détermine alors la direction $\mathbf{u}$ à suivre pour optimiser la solution $\boldsymbol{\alpha}$. En particulier, on s'intéresse à l'ensemble des éléments pouvant évoluer dans la direction du gradient :

$$S = \{k : g_k > 0, \alpha_k < C\} \cup \{k : g_k < 0, \alpha_k > 0\} \tag{4.9}$$

Les conditions $\alpha_k < C$ et $\alpha_k > 0$ correspondent à la recherche de directions réalisables, c'est-à-dire pour lesquelles on ne risque pas de violer les contraintes d'inégalité après modification des coefficients α_k, si petite puisse être cette modification.

18. On dit qu'une matrice ou un vecteur est parcimonieux lorsque seuls quelques-uns de ses éléments sont non-nuls, la terminologie anglo-saxonne *sparse* étant largement répandue, y compris dans les textes francophones, nous utiliserons ce terme par la suite.

Cas sans biais

On choisit au sein de l'ensemble S les q coefficients pour lesquels le gradient (en valeur absolue) est maximal. La valeur q, qui détermine la taille de l'ensemble de travail, est fixée *a priori*. L'ensemble de travail S_{ws} est alors défini comme :

$$S_{ws} = \{ \text{ indices des } q \text{ plus grands éléments } |g_k|, k \in S\} \tag{4.10}$$

et on détermine les coefficients u_k de la direction $\mathbf{u}$ tels que :

$$u_k = \begin{cases} g_k \text{ si } k \in S_{ws} \\ 0 \text{ sinon} \end{cases} \tag{4.11}$$

Cas avec biais

Dans le cas avec biais, la contrainte d'égalité (4.1c) doit être vérifiée avant et après la mise à jour de la solution (4.4). Cette approche nous permet de déduire la relation suivante :

$$\begin{aligned} \boldsymbol{\alpha}^T\mathbf{y} &= (\boldsymbol{\alpha} + \lambda\mathbf{u})^T\mathbf{y} \\ \Rightarrow \lambda\mathbf{u}^T\mathbf{y} &= 0 \\ \Rightarrow \quad \mathbf{u}^T\mathbf{y} &= 0 \end{aligned} \tag{4.12}$$

Dans l'algorithme présenté, nous proposons, dans ce cas, de ne modifier que deux coefficients de la solution. Soit i et j ces coefficients, on simplifie également la direction d'optimisation en fixant $|u_i| = |u_j| = 1$.

Le choix du coefficient i est basé sur l'approche sans biais, avec $q = 1$. Ainsi :

$$i = \arg\max_k |g_k|, k \in S \tag{4.13}$$

et on fixe $u_i = \text{sign}(g_i)$.

On recherche ensuite un élément j, s'il existe, parmi ceux maximisant l'accroissement de la fonction objectif et on fixe $u_j = \text{sign}(g_j)$. La contrainte (4.12) impose également que j doit être choisi tel que : $u_i y_i + u_j y_j = \text{sign}(g_i)y_i + \text{sign}(g_j)y_j = 0$. La recherche de l'indice j s'exprime alors :

$$j = \arg\max_k |g_k|, k \in S \cap \{l, \text{sign}(g_i)y_i = -\text{sign}(g_l)y_l\} \tag{4.14}$$

Si aucun élément n'est trouvé par l'équation (4.14), alors j doit être choisi parmi l'ensemble des éléments dégradant la fonction objectif. On s'intéresse alors à l'ensemble des éléments pouvant évoluer dans la direction opposée au gradient :

$$S_{opp} = \{k : g_k < 0, \alpha_k < C\} \cup \{k : g_k > 0, \alpha_k > 0\} \tag{4.15}$$

Dans ce cas, j doit être l'élément qui minimise la dégradation de la fonction objectif. On fixe $u_j = -\text{sign}(g_j)$, alors la contrainte (4.12) impose que $\text{sign}(g_i)y_i - \text{sign}(g_j)y_j = 0$ et la recherche de l'indice j s'exprime :

$$j = \arg\min_k |g_k|, k \in S_{opp} \cap \{l, \text{sign}(g_i)y_i = \text{sign}(g_l)y_l\} \tag{4.16}$$

On pose enfin :

$$u_k = \begin{cases} \text{sign}(g_i)y_i y_k & \text{si } k = i \\ -\text{sign}(g_i)y_i y_k & \text{si } k = j \\ 0 & \text{sinon} \end{cases} \tag{4.17}$$

4.2.3 Respect des contraintes d'inégalité

Les contraintes d'inégalité, qui apparaissent lors de la résolution d'un problème SVM avec ou sans biais, influent également sur le pas λ^* optimal (4.8). En effet, il s'agit qu'à l'issue de la mise à jour (4.4) de la solution, la nouvelle solution respecte toujours les contraintes d'inégalité (4.1b) :

$$\mathbf{0}_N \leq \boldsymbol{\alpha} + \lambda \mathbf{u} \leq \mathbf{C} \tag{4.18}$$

Le pas optimal est déterminé à l'équation (4.8). Cependant, celui-ci doit être seuillé s'il conduit l'un des coefficients de la solution à sortir du domaine acceptable $[\mathbf{0}_N, \mathbf{C}]$. Les bornes inférieures λ_{inf} et supérieures λ_{sup} sur λ sont déterminées à partir des inégalités suivantes :

$$0 \leq \alpha_k + \lambda u_k \leq C_i$$

$$\Rightarrow \begin{cases} -\frac{\alpha_k}{u_k} \leq \lambda \leq \frac{C_k - \alpha_k}{u_k} & \text{si } u_k > 0 \\ \frac{C_k - \alpha_k}{u_k} \leq \lambda \leq -\frac{\alpha_k}{u_k} & \text{si } u_k < 0 \\ \lambda \text{ quelconque} & \text{si } u_k = 0 \end{cases}$$

Ces bornes s'expriment alors :

$$\lambda_{inf} = \max \left(\max_{k:u_k>0} \left(-\frac{\alpha_k}{u_k} \right), \max_{k:u_k<0} \left(\frac{C_k - \alpha_k}{u_k} \right) \right) \tag{4.19}$$

$$\lambda_{sup} = \min \left(\min_{k:u_k>0} \left(\frac{C_k - \alpha_k}{u_k} \right), \min_{k:u_k<0} \left(-\frac{\alpha_k}{u_k} \right) \right) \tag{4.20}$$

4.2.4 Critère d'arrêt

L'algorithme proposé étant un algorithme itératif, un critère d'arrêt doit être fixé.

Critère d'arrêt basé sur les conditions KKT

On propose dans un premier temps de s'intéresser aux conditions KKT (voir théorème 3.6), nécessaires et suffisantes pour l'atteinte de l'optimal. Dans notre cas, on sait (tableau 3.1) que lorsque $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, les propriétés suivantes sont vraies :

$$\text{Si} \quad 0 < \alpha_i^* < C \quad \text{alors} \quad y_i f(\mathbf{x}_i) = 1 - \mathbb{1}_{OC} \tag{4.21}$$

$$\text{Si} \quad \alpha_i^* = 0 \quad \text{alors} \quad y_i f(\mathbf{x}_i) < 1 - \mathbb{1}_{OC} \tag{4.22}$$

$$\text{Si} \quad \alpha_i^* = C \quad \text{alors} \quad y_i f(\mathbf{x}_i) > 1 - \mathbb{1}_{OC} \tag{4.23}$$

Il est alors possible de construire un test d'optimalité sur ces propriétés : si l'une d'entre elles n'est pas vérifiée, alors l'optimum n'est pas atteint. En pratique, compte-tenu de la précision finie des calculateurs numériques, on introduit deux seuils sur les tests : ε_α et ε_{KKT}.

Ainsi, le premier critère d'arrêt que nous proposons, basé sur les conditions KKT prend la forme suivante :

$$\begin{array}{lllll} \text{Si } \exists i : & (\varepsilon_\alpha < \alpha_i < C - \varepsilon_\alpha) & ET & (|y_i f(\mathbf{x}_i) - 1 + \mathbb{1}_{OC}| > \varepsilon_{KKT}) \\ \quad \text{OU} & (\alpha_i \leq \varepsilon_\alpha) & ET & (y_i f(\mathbf{x}_i) - 1 + \mathbb{1}_{OC} < \varepsilon_{KKT}) \\ \quad \text{OU} & (\alpha_i \geq C - \varepsilon_\alpha) & ET & (y_i f(\mathbf{x}_i) - 1 + \mathbb{1}_{OC} > \varepsilon_{KKT}) \\ \text{alors } \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* & & & \\ \text{finSi} & & & \end{array}$$

L'inconvénient de l'utilisation d'un tel critère d'arrêt dans le cadre des SVM avec biais est que l'évaluation de $f(\mathbf{x})$ nécessite le biais. Ainsi, ce dernier devrait être calculé à chaque itération. De plus, l'évaluation de $f(\mathbf{x})$ peut s'avérer coûteuse dans le cas d'un nombre important de vecteurs de support.

Critère basé sur le Lemme de Keerthi

Keerthi *et al* [KSBM01] ont montré qu'il est possible d'obtenir, dans le cas 2-classes, un critère d'arrêt rendant inutile la connaissance du biais b. Nous présentons ci-dessous le Lemme de Keerthi. Nous suivons l'auteur dans son raisonnement, et nous faisons le lien entre les conditions de KKT et le gradient de la fonction objectif. Ce lien permet de montrer qu'il est possible d'évaluer ce critère sans calcul supplémentaire, bénéficiant avantageusement des grandeurs évaluées pour le choix de l'ensemble de travail et pour la mise à jour de la solution.

Lemme 4.1 (Lemme de Keerthi) *Soient b_{low} et b_{up} deux seuils pour caractériser l'optimum du problème C-SVM tels que :*

$$b_{low} = \max\{F_i : i \in I_0 \cup I_3 \cup I_4\} \tag{4.24}$$
$$b_{up} = \min\{F_i : i \in I_0 \cup I_1 \cup I_2\} \tag{4.25}$$

où $F_i = \sum_j \alpha_j y_j \kappa(x_i, x_j) - y_i$, et

$$\begin{aligned} I_0 &= \{i : 0 < a_i < C\} \\ I_1 &= \{i : y_i = 1, a_i = 0\} \\ I_2 &= \{i : y_i = -1, a_i = C\} \\ I_3 &= \{i : y_i = 1, a_i = C\} \\ I_4 &= \{i : y_i = -1, a_i = 0\} \end{aligned}$$

Alors une condition nécessaire et suffisante d'optimalité est :

$$b_{low} \leq b_{up} \tag{4.26}$$

L'apprentissage par décomposition peut alors s'effectuer tant que cette condition n'est pas vérifiée. En pratique, de manière à prendre en compte les erreurs numériques, l'inégalité précédente est remplacée par :

$$b_{low} \leq b_{up} + 2\varepsilon_{KKT}$$

où ε_{KKT} est le paramètre de tolérance sur le respect des conditions KKT. De ces éléments, nous pouvons en déduire qu'il y a violation de l'optimalité tant qu'il existe un couple (i, j) tel que :

$$F_i > F_j + 2\varepsilon_{KKT} \quad (i \in I_0 \cup I_3 \cup I_4, j \in I_0 \cup I_1 \cup I_2) \tag{4.27}$$

L'algorithme prend alors fin lorsqu'il n'existe plus aucun couple définissant une telle violation.

Tohmé [Toh09] a noté que $F_i = -y_i g_i$, et en déduit que :

$$g_i - g_j = b_{low} - b_{up}$$

où

$$g_i = \max_k g_k : k \in I_a$$

$$g_j = \min_l g_l : l \in I_b$$

et

$$I_a = i : y_i = 1, 0 \leq \alpha_i < C \cup y_i = -1, 0 < \alpha_i \leq C$$

$$I_b = i : y_i = 1, 0 < \alpha_i \leq C \cup y_i = -1, 0 \leq \alpha_i < C$$

Ce résultat permet de définir le critère d'arrêt suivant :

$$g_i - g_j \leq 2\varepsilon_{KKT}$$

pour les algorithmes Fast-SVC [Toh09] et Fast-OC2 [TL11]. De plus, on montre que les éléments i et j correspondent, dans le cas de la résolution de problèmes avec biais, aux éléments i et j de l'ensemble de travail (équations (4.13), (4.14) et (4.16)). L'évaluation du critère d'arrêt ne nécessite donc aucun calcul complémentaire.

Dans le cas des problèmes sans biais, et en particulier pour le SMGO, Tohmé montre un lien entre le critère proposé par Keerthi, la direction choisie $\mathbf{u}$ et la taille de l'ensemble de travail. Le critère d'arrêt s'exprime alors :

$$\frac{\mathbf{u}^T \mathbf{g}}{\operatorname{card} S_{ws}} \leq \varepsilon_{KKT} \tag{4.28}$$

Enfin, remarquons également que dans le cas avec biais, Keerthi a noté que la valeur vers laquelle b_{low} et b_{up} convergent pour vérifier l'optimalité est strictement équivalente au biais b [KSBM01]. Dans le cas du critère d'arrêt proposé, une estimation correcte du biais (dans la tolérance fixée par ε_{KKT}) sera la moyenne entre b_{low} et b_{up}. En se basant sur les résultats de Tohmé [Toh09], le biais b peut alors être estimé sans calcul supplémentaire lorsque les conditions d'optimalité sont réunies :

$$b = \frac{y_j g_j + y_i g_i}{2} \tag{4.29}$$

où i et j sont les indices des éléments de l'ensemble de travail (équations (4.13), (4.14) et (4.16)) au moment où les conditions d'optimalité (approximative) sont vérifiées.

4.2.5 Initialisation et algorithme

L'algorithme doit enfin être initialisé avec une solution réalisable. Dans le cas sans biais, $\boldsymbol{\alpha} = \mathbf{0}_N$ est une solution possible, de même que dans le cas de la classification 2-classes avec biais. L'initialisation du gradient est alors immédiate compte-tenu de l'équation (4.5). En revanche, dans le cas d'un problème 1-classe avec biais (OC-SVM ou OC2-SVM), l'initialisation doit se faire en tenant compte des contraintes d'égalité et d'inégalité. L'idée est alors d'initialiser νN coefficients α_i à la valeur $\frac{1}{\nu N}$ (νN supposé entier). Le gradient initial est alors calculé en évaluant (4.5). L'ensemble de ces éléments nous permet de décrire les étapes de l'algorithme et de présenter le pseudo-code 4.2.5 correspondant.

L'algorithme proposé permet avantageusement de résoudre le problème (4.1) soit l'ensemble des problèmes C-SVM, OC-SVM, OC2-SVM, WO-SVM et WOOC2-SVM. On remarque que l'algorithme est strictement identique pour l'ensemble des problèmes, seuls l'initialisation et le calcul éventuel du biais changent. Le choix de la direction d'optimisation est également différent suivant les cas avec ou sans biais, mais néanmoins basé sur la même approche d'analyse du gradient.

Entrées: $\boldsymbol{\alpha}$
1: Initialiser $\mathbf{g}$ à l'aide de l'équation (4.5)
2: **while** Critère d'arrêt $<$ Seuil **do**
3: Choisir l'ensemble de travail S_{ws}
4: Déterminer la direction $\mathbf{u}$ à l'aide des équations (4.11) ou (4.17)
5: Calculer le pas λ^* à l'aide de l'équation (4.8)
6: Seuiller le pas λ^* à l'aide des bornes (4.19) et (4.20)
7: Mettre à jour la solution $\boldsymbol{\alpha}$ à l'aide de la formule (4.4)
8: Mettre à jour le gradient $\mathbf{g}$ à l'aide de la formule (4.7)
9: Mettre à jour le coût L à l'aide de la formule (4.6)
10: Calculer le critère d'arrêt (4.28)
11: **end while**
12: Calculer le biais (4.29) si nécessaire
Sorties: $\boldsymbol{\alpha}$

4.3 Démarrage à chaud

L'algorithme proposé dispose de capacités de démarrage à chaud. En effet, disposant d'une solution réalisable $\boldsymbol{\alpha}$, l'utilisateur peut bénéficier du travail déjà accompli sans nécessairement reprendre l'optimisation depuis le départ.

Rappelons qu'un avantage majeur de l'approche sans biais réside dans l'absence de contrainte d'égalité. Ainsi, l'équilibre imposé par les conditions KKT dans le cas avec biais n'est plus à maintenir. Ce dernier induit en effet une complexité importante lorsqu'il s'agit de mettre à jour une solution existante. On parle alors de changements adiabatiques. Ce type d'approche a été popularisé suite aux travaux de Cauwenberghs et Poggio concernant l'apprentissage incrémental et décrémental des SVM [CP00]. Ces travaux ont été appliqués notamment dans le cadre de la segmentation de signal audio [GD03] ou la détection d'événements anormaux [DDGD05]. Bien que Laskov *et al* aient proposé une implémentation rapide, stable et robuste afin de mettre en œuvre ce type de mise à jour [LGKM06], il s'avère que le maintien de la contrainte d'égalité est une tâche difficile, particulièrement lorsque l'ensemble d'apprentissage est grand.

Inspirés par l'absence de biais inhérente aux approches statistiques des SVM, Kivinen *et al* [KSW01, KSW10], d'une part, et Steinwart *et al* [SHS11] d'autre part, ont proposé différentes stratégies de mise à jour d'une solution. La solution proposée par les premiers, s'immergeant dans le contexte de l'apprentissage en ligne, ne se prête pas à la mise à jour des paramètres d'apprentissage. La solution proposée par les seconds partage l'esprit de l'approche présentée ici.

4.3.1 Applications

Dans un premier temps, remarquons qu'il est possible d'améliorer une solution existante. Disposant d'une solution pour un seuil fixé sur le critère d'arrêt, il est possible de réduire ce seuil et de reprendre le travail d'optimisation depuis la solution connue. En effet, on montre aisément que la solution obtenue pour un seuil élevé d'un critère d'arrêt est une solution intermédiaire à celle obtenue pour un seuil plus faible du même critère d'arrêt. D'un point de vue pratique cela peut permettre de prototyper un système pour la recherche de l'hyper-paramètres optimal C (ou ν dans le cas 1-classe), puis lorsque ce dernier est identifié de bénéficier du résultat déjà obtenu pour faire converger l'algorithme vers une solution performante. Une procédure de mise à jour de la solution est néanmoins nécessaire, et celle-ci est décrite au paragraphe suivant.

On remarque également que les solutions de l'approche SVM avec biais sont des solutions possibles aux problèmes SVM sans biais, il est alors envisageable a posteriori de relâcher la contrainte d'égalité du problème (4.1). Pour autant, on constate en pratique que cette approche nécessite souvent un nombre d'itérations plus important que de reprendre l'apprentissage depuis une solution initiale standard.

Enfin, et il s'agit sûrement de l'une de ses propriétés les plus importantes, le démarrage à chaud va permettre de refléter dans une solution existante la modification de l'ensemble d'apprentissage sans nécessairement réaliser un nouvel apprentissage. Ceci va, en particulier, permettre de répondre d'une part à des besoins d'apprentissage en ligne (ajout de nouvelles données, oubli d'anciennes données), et d'autre part à des problématiques d'apprentissage non supervisé (en permettant la modification des étiquettes attribuées a priori aux données). Nous allons également montrer que le démarrage à chaud de l'algorithme permet aussi un gain de temps lorsque l'hyper-paramètre $\mathbf{C}$ est déterminé à partir d'une grille de recherche (lors d'une stratégie par validation croisée par exemple).

La mise en œuvre des deux premières applications proposées (amélioration d'une solution et bascule d'une approche avec biais vers une approche sans biais) est évidente. La modification de l'ensemble d'apprentissage nécessite par contre une mise à jour de la solution précédemment obtenue avant de la soumettre à nouveau au processus d'entrainement. Nous allons maintenant détailler les procédures à réaliser avant d'effectuer un démarrage à chaud.

4.3.2 Procédures de mise à jour

Comme cela a été souligné à la section précédente, seuls les vecteurs $\mathbf{g}$ et $\boldsymbol{\alpha}$ doivent être connus pour initialiser l'algorithme proposé. Par ailleurs, ces deux vecteurs sont connus à l'issue du processus d'optimisation et contiennent l'ensemble de l'information nécessaire à la représentation du SVM souhaité, y compris dans le cas avec biais compte-tenu du résultat (4.29). Nous allons maintenant décrire les procédures de mise à jour de ces vecteurs lorsque l'ensemble d'apprentissage est modifié ou encore lorsque les paramètres du SVM sont modifiés.

On rappelle le lien entre le gradient $\mathbf{g}$ et le vecteur solution $\boldsymbol{\alpha}$:

$$\mathbf{g} = \mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\delta}$$

On distingue dans un premier deux type de mises à jour différentes du vecteur $\boldsymbol{\alpha}$: une mise à jour additive et une mise à l'échelle . On montre alors aisément que les procédures correspondantes de mise à jour du vecteur solution $\boldsymbol{\alpha}$ et du vecteur gradient $\mathbf{g}$ suivent les procédures suivantes :

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha} \tag{4.30}$$

$$\mathbf{g} \leftarrow \mathbf{g} + \mathbf{H}\Delta\boldsymbol{\alpha} \tag{4.31}$$

pour la mise à jour additive et :

$$\boldsymbol{\alpha} \leftarrow \theta\boldsymbol{\alpha} \tag{4.32}$$

$$\mathbf{g} \leftarrow \theta\mathbf{g} + (1-\theta)\boldsymbol{\delta} \tag{4.33}$$

pour la mise à l'échelle de la solution.

Soit maintenant la décomposition du vecteur $\boldsymbol{\alpha}$ en deux sous-vecteurs $\boldsymbol{\alpha}_1$ et $\boldsymbol{\alpha}_2$, de même que la décomposition de la matrice $\mathbf{H}$ en trois sous-matrices $\mathbf{H}_{1,1}$, $\mathbf{H}_{1,2}$ et $\mathbf{H}_{2,2}$. On peut alors

décomposer le gradient $\mathbf{g}$ suivant :

$$\mathbf{g} = \begin{bmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} \\ \mathbf{H}_{1,2}^T & \mathbf{H}_{2,2} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{H}_{1,2}\boldsymbol{\alpha}_2 \\ \mathbf{H}_{1,2}^T\boldsymbol{\alpha}_1 \end{pmatrix} \tag{4.34}$$

où $\mathbf{g}_1 = \mathbf{H}_{1,1}\boldsymbol{\alpha}_1 + \boldsymbol{\delta}_1$ et $\mathbf{g}_2 = \mathbf{H}_{2,2}\boldsymbol{\alpha}_2 + \boldsymbol{\delta}_2$. Nous pouvons alors extraire de cette décomposition les procédures de mise à jour des vecteurs $\boldsymbol{\alpha}$ et $\mathbf{g}$ lorsque des éléments sont ajoutés à l'ensemble d'apprentissage :

$$\boldsymbol{\alpha} \leftarrow \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}_{new} \end{pmatrix} \tag{4.35}$$

$$\mathbf{g} \leftarrow \begin{pmatrix} \mathbf{g} \\ \mathbf{g}_{new} \end{pmatrix} + \begin{pmatrix} \mathbf{H}_{new}\boldsymbol{\alpha}_{new} \\ \mathbf{H}_{new}^T\boldsymbol{\alpha} \end{pmatrix} \tag{4.36}$$

ou ôtés de l'ensemble d'apprentissage :

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}_{\backslash del} \tag{4.37}$$

$$\mathbf{g} \leftarrow \mathbf{g}_{\backslash del} - \mathbf{H}_{del}\alpha_{del} \tag{4.38}$$

avec :

- S l'ensemble des éléments existants ou conservés, *del* l'ensemble des éléments ôtés et *new* l'ensemble des nouveaux éléments,
- $\boldsymbol{\alpha}_{new}$ l'ensemble des coefficients α pour les éléments ajoutés,
- $\mathbf{H}_{new}$ la matrice de Gram polarisée partielle composée des coefficients $\{h_{i,j}, i \in S, j \in new\}$,
- $\mathbf{g}_{new}$ est le gradient pour les éléments de *new*,
- $\boldsymbol{\alpha}_{\backslash del}$ le vecteur $\boldsymbol{\alpha}$ privé des éléments de *del*,
- $\mathbf{g}_{\backslash del}$ le vecteur $\mathbf{g}$ privé des éléments de *del*,
- $\mathbf{H}_{del}$ la matrice de Gram polarisée partielle composée des coefficients $\{h_{i,j}, i \in S, j \in del\}$,
- $\boldsymbol{\alpha}_{del}$ l'ensemble des coefficients α de l'ensemble *del*.

4.3.3 Stratégies de mise à jour d'une solution

On rappelle la contrainte d'inégalité du problème (4.1) :

$$\mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{C}$$

Disposant des procédures de mise à jour ci-dessus, nous proposons les stratégies suivantes lorsque l'hyper-paramètre du problème SVM est modifié entre de deux processus d'optimisation :

Stratégie S0 : ne rien faire $\mathbf{C} \leftarrow \mathbf{C}_{new}$ avec $\mathbf{C}_{new} > \mathbf{C}$, on ne modifie ni $\boldsymbol{\alpha}$, ni $\mathbf{g}$.

Stratégie S1 : expansion des supports vecteurs bornés $\mathbf{C} \leftarrow \mathbf{C}_{new}$ avec $\mathbf{C}_{new} > \mathbf{C}$, on modifie l'ensemble des supports vecteurs $\{\alpha_k, \alpha_k = C\}$. On utilise alors la procédure de mise à jour additive avec :

$$\Delta\boldsymbol{\alpha}_k = \begin{cases} \mathbf{C}_{new} - \mathbf{C} & \text{si } \alpha_k = \mathbf{C} \\ 0 & \text{sinon} \end{cases}$$

Cette stratégie met en œuvre l'idée qu'en augmentant C, une proportion importante des observations qui sont vecteurs de support va le rester.

Stratégie S2 : *clipping*[19] des supports vecteurs saturés $\mathbf{C} \leftarrow \mathbf{C}_{new}$ avec $\mathbf{C}_{new} < \mathbf{C}$, on modifie l'ensemble des supports vecteurs $\{\alpha_k, \boldsymbol{\alpha}_k > C_{new}\}$. On utilise alors la procédure de mise à jour additive avec :

$$\Delta\boldsymbol{\alpha}_k = \begin{cases} \mathbf{C}_{new} - \alpha_k & \text{si } \alpha_k > \mathbf{C}_{new} \\ 0 & \text{sinon} \end{cases}$$

Stratégie S3 : mise à l'échelle $\mathbf{C} \leftarrow \tau\mathbf{C}$, on utilise alors la procédure de mise à jour par mise à l'échelle avec $\theta = \tau$.

On s'intéresse maintenant au cas où l'ensemble d'apprentissage S est modifié. Dans ce cas, on peut soit ajouter de nouvelles données, soit ôter des données existantes. Dans ce dernier cas, il n'y a pas de stratégie particulière à adopter, la procédure de suppression de données est appliquée telle que définie au paragraphe précédent. Lorsque des données sont ajoutées à l'ensemble d'apprentissage, les coefficients multiplicateurs α_i correspondant à ces données doivent être initialisés. On propose alors les stratégies suivantes :

Stratégie A0 : les nouvelles données sont initialisées avec des valeurs nulles $\boldsymbol{\alpha}_{new} = 0$, dans ce cas la procédure d'ajout de données est utilisée. On remarque en particulier que le gradient n'évolue pas pour les données déjà présentes dans l'ensemble d'apprentissage tandis que le gradient associé aux nouvelles données vaut $\mathbf{g}_{new} + \mathbf{H}_{new}^T\boldsymbol{\alpha}$.

Stratégie A1 : initialisation avec des valeurs saturées $\boldsymbol{\alpha}_{new} = C$, dans ce cas la procédure d'ajout de données est utilisée telle que définie ci-dessus.

Stratégie A2 : initialisation avec des valeurs saturées pour les données mal classées $\boldsymbol{\alpha}_k = C$ si $k \in new$ et $y_k f(\mathbf{x}_k) < 0$, 0 sinon. Dans ce cas la procédure d'ajout de données est utilisée.

On considère maintenant le cas où certaines étiquettes associées à des données existantes sont modifiées, c'est-à-dire $y_k \leftarrow -y_k$. Dans ce cas la procédure à appliquer est une combinaison de la suppression et de l'ajout de données. Néanmoins le calcul de certains éléments n'est pas nécessaire. Soit *del* l'ensemble des données dont les étiquettes sont modifiées, alors $\mathbf{H}_{new} = -\mathbf{H}_{del}$. La détermination des valeurs des coefficients multiplicateurs pour les données modifiées peut s'effectuer suivant les stratégies A0, A1 ou A2. Ainsi la procédure sera la suivante :

$$\mathbf{y} \leftarrow \begin{pmatrix} \mathbf{y}_{\setminus del} \\ -\mathbf{y}_{del} \end{pmatrix} \tag{4.39}$$

$$\boldsymbol{\alpha} \leftarrow \begin{pmatrix} \boldsymbol{\alpha}_{\setminus del} \\ \boldsymbol{\alpha}_{new} \end{pmatrix} \tag{4.40}$$

$$\mathbf{g} \leftarrow \begin{pmatrix} \mathbf{g}_{\setminus del} \\ \mathbf{g}_{new} \end{pmatrix} - \begin{pmatrix} \mathbf{H}_{del}(\boldsymbol{\alpha}_{del} + \boldsymbol{\alpha}_{new}) \\ \mathbf{H}_{del}^T\boldsymbol{\alpha}_{\setminus del} \end{pmatrix} \tag{4.41}$$

On remarque également que dans le cas des SVM 1-classe, le paramètre C est fixé à $\frac{1}{\nu N}$. L'augmentation ou la réduction de la taille de l'ensemble d'apprentissage implique donc nécessairement une modification de cet hyper-paramètre. La stratégie retenue pour la modification de C (S0, S1, S2, ou S3) sera systématiquement appliquée avant d'effectuer la procédure d'ajout ou de suppression de données. Dans le cas 1-classe avec contrainte binaire, l'hyper-paramètre est modifié uniquement lorsque le nombre de données associées à la classe +1 est modifié.

Enfin, il est important de noter que compte-tenu de l'expression sous-forme matricielle de l'ensemble des procédures exposées il est possible d'ajouter, supprimer ou modifier une ou plusieurs données à la fois.

Troisième partie

Application à la surveillance basée sur la modalité audio

Chapitre 5

Protocole d'évaluation

Dans ce chapitre nous décrivons les outils utilisés pour évaluer et présenter les performances d'un système de détection d'événements anormaux dans le contexte de la surveillance audio. Il n'existe pas de bases de données publiques pour l'évaluation de tels systèmes, d'une part à cause de la confidentialité des données collectées, et d'autre part parce que les événements anormaux sont par définition trop rares pour être correctement représentés. En l'absence de base commune, les outils de mesure de performances ne sont pas non plus identifiés dans ce cadre. Ainsi, nous décrivons le protocole d'évaluation utilisé chez Thales, celui-ci ayant déjà permis d'évaluer d'autres systèmes développés en interne [CAR08, CR10].

Nous décrivons dans un premier temps l'outil proposé pour générer des séquences de signaux incluant des événements anormaux à des fins d'évaluation du système de surveillance. Cette thèse a contribué à l'amélioration de l'outil par l'analyse de l'influence des différentes pondérations fréquentielles dans la mesure du rapport signal à bruit des événements insérés. Nous présentons ensuite quelques critères d'évaluation, en particulier les courbes DET, adaptées au contexte opérationnel de notre application. Enfin, nous proposons une approche nouvelle de la fonction de décision SVM permettant de construire des familles de fonctions de décision.

5.1 Rappel de la tâche

L'approche proposée dans cette thèse consiste en l'apprentissage d'un modèle d'ambiance normale (hypothèse H_0) en se basant sur une grande quantité de signaux enregistrés *in situ* et sans expertise. Le postulat initial sur lequel repose cette approche considère que seule une quantité très faible de ces signaux ne relève en réalité pas de H_0. On s'appuie alors sur une modélisation SVM 1-classe, bénéficiant du paramètre de contrôle ν pour spécifier une fraction des données relevant de l'hypothèse complémentaire H_1. Ainsi, l'approche est dite « partiellement supervisée » car cette fraction de données à rejeter (les *outliers*) ne sont pas explicitement définis mais découverts au cours de l'apprentissage.

Le problème de modélisation de l'ambiance normale est donc traité par une optimisation SVM sur un ensemble d'observations issues d'enregistrements. L'évaluation consiste ensuite à mesurer la capacité du modèle à correctement identifier des événements anormaux quelconques insérés dans un signal d'ambiance. Ce dernier est enregistré également *in situ* dans les mêmes conditions que les signaux d'apprentissage. Il n'est donc pas nécessaire, dans cette approche, de disposer d'une grande quantité de signaux anormaux puisque aucun modèle de ceux-ci n'a à être construit. En revanche, nous utilisons des événements de différentes nature (impulsionnels ou stationnaires, harmoniques ou stochastiques, etc.) afin de pouvoir garantir les performances

indépendamment des événements sonores rencontrés une fois le système déployé.

5.2 Génération de séquences audio anormales

Dans le contexte d'environnements sous surveillance audio tels que les gares de transports publics, les centres événementiels ou encore les zones urbaines, l'ambiance sonore est un signal complexe, composé de centaines d'événements qui peuvent être considérés comme normaux : discussions, klaxons, arrivées/départs de trains, etc. Il se peut également qu'une structure temporelle soit naturellement présente : passages réguliers des trains, heures pleines/creuses, jours de la semaine, etc. Bien qu'il ne soit pas possible de synthétiser ces types d'environnements, il demeure possible de les enregistrer. Cependant, il sera difficile de disposer de réalisations d'événements anormaux. En effet, ces événements étant rares par nature, une trop grande quantité de signaux devrait être enregistrée pour réaliser une base d'apprentissage ; sans compter le temps d'analyse et d'étiquetage de ces dits signaux, long et coûteux.

Dans cette section, nous décrivons un cadre d'évaluation dont l'objectif est d'inclure des événements anormaux au sein d'ambiances sonores enregistrées, de manière la plus réaliste possible. Afin de qualifier la robustesse et le pouvoir de généralisation du système de surveillance à évaluer, nous proposons également de contrôler le niveau sonore des événements anormaux en termes de rapport signal à bruit (RSB). Cependant, là où des approches classiques mesurent ce ratio uniformément sur l'ensemble du spectre audio, nous proposons d'améliorer cette mesure par l'utilisation d'un RSB pondéré, adapté aux signaux étudiés. Notre approche présente les avantages suivants : contrôle précis du RSB et de la position des événements, génération rapide de bases de données réalistes (incluant des événements anormaux en nombre suffisant), et finalement, ne nécessitant pas d'intervenir dans un environnement réel une fois l'ambiance enregistrée.

D'abord, nous donnons quelques clés ayant rapport à la mesure du niveau de bruit. Ensuite, considérant des signaux audio issus d'environnements sous surveillance, nous discutons de résultats empiriques afin de déterminer une pondération fréquentielle adaptée pour la mesure de RSB. Enfin, s'appuyant sur ce contrôle adapté des niveaux de mélange ambiance/événement anormal, nous présentons l'outil mis en œuvre pour générer des bases de données de signaux pour l'évaluation de systèmes de surveillance. Nous évoquons également d'autres approches étudiées au cours des travaux qui, n'ayant pas donné satisfaction, sont décrites à l'annexe A.

5.2.1 La mesure du niveau sonore

Nous ciblons l'étude d'un système capable d'automatiser une tâche réalisable par l'Homme : la surveillance audio. L'objectif est de comprendre les phénomènes qui rendent l'oreille humaine performante pour la tâche à accomplir. En particulier, l'oreille et le cerveau réalisent un certain nombre de traitements afin de clarifier le signal acoustique. Ces traitements permettent en particulier de focaliser l'attention d'un auditeur sur certains sons, améliorant sa capacité à qualifier comme normal ou anormal une situation.

Nous nous intéressons dans un premier temps à la perception de ce que nous appelons le niveau d'intensité d'un son. Nous montrons en particulier qu'une échelle linéaire de mesure de l'amplitude de l'onde acoustique n'est pas la plus adaptée pour comparer le niveau sonore de différents événements. Ensuite, nous focalisons brièvement notre attention sur la différence de perception suivant le contenu fréquentiel d'un son. Ceci nous amène à adopter une approche singulière pour notre analyse du signal audio, bénéficiant des avantages reconnus de la perception humaine.

De la mesure physique...

Le son est une onde mécanique de pression. Dans l'air, cette onde fait varier la pression de proche en proche jusqu'à stimuler l'oreille d'un auditeur et provoquer la sensation d'audition. Les acousticiens et physiciens mesurent alors l'intensité sonore (SPL, pour *Sound Pressure Level*) à partir de la variation moyenne quadratique de la pression acoustique (en $\frac{N}{m^2}$ ou Pa).

Le décibel (dB), une unité de référence pour exprimer le rapport entre deux puissances sous une forme logarithmique, est utilisé pour refléter le caractère logarithmique de la sensation auditive [TS99]. La mesure en dB de l'intensité sonore s'effectue relativement au seuil d'audition à $1kHz$, soit $20 \times 10^{-6} Pa$ [20] :

$$SPL(dB) = 10 \log_{10} \frac{P_{\text{mesurée}}}{20 \times 10^{-6}} \tag{5.1}$$

Cependant, cette mesure, notée $dB(SPL)$, est une information acoustique (liée à une grandeur physique) qui ne reflète pas suffisamment la manière dont sont perçus les sons par l'oreille humaine.

... à la mesure pondérée

Adapter la mesure du niveau sonore à une tache donnée est un problème déjà largement abordé dans la littérature. Les premiers travaux menés pour mettre au point une métrique adaptée à la perception humaine datent des années 1920. Cependant, les approches proposées furent d'abandonnées à cause de données contradictoires, jusqu'aux travaux de Fletcher et Munson [FM33]. Ceux-ci introduisent alors des courbes d'isosonie permettant de relier l'intensité sonore (en dB SPL) et le niveau de sonorité. Ce dernier terme décrit l'intensité d'une sensation auditive et est exprimé en phones [21]. En 1956, Robinson et Dadson [RD56] révisent ces courbes, puis en 2003 celles-ci sont normalisées par l'*International Organization for Standardization* (ISO226 :2003) [Int03b].

D'autre part, la BBC a commencé en 1968 [Bri68] des études concernant la mesure du niveau de bruit dans les équipements électroniques de radio-diffusion afin d'en améliorer la qualité sonore. Une des principales motivations de ces travaux était l'inappropriation des pondérations existantes aux bruits de nature aléatoire. Ces recherches ont conduit en 1986 à la recommandation R468-4 du CCIR, aujourd'hui référencée ITU-R468 [Int90]. La pondération issue de cette recommandation est notamment utilisée dans les populaires standards Dolby A et B.

Voici une description qualitative des pondérations issues des courbes d'isosonie d'une part, et de la recommandation ITU d'autre part (le tracé de ces courbes est présenté à la figure 5.1) :

- Type-A : originellement utilisée pour simuler la réponse de l'oreille humaine à 40 phones. Elle est particulièrement recommandée pour la mesure subjective des bruits ambiants (norme IEC-61672 :2003 [Int03a]).
- Type-C : atténuation des fréquences en dehors de la bande 200-1250 Hz (norme IEC-61672 :2003 [Int03a]).
- ITU-R468 : accentuation des fréquences dans la bande 1-12,5kHz (jusqu'à +12,2dB à 6,3kHz) et atténuation progressive des fréquences en dehors de cette bande (norme ITU-R468 [Int90]). Originellement conçue pour mesurer l'impact subjectif des bruits aléatoires large bande (dans les appareils de diffusion professionnels).

20. Plutôt que la pression, la puissance (W) ou l'intensité ($\frac{W}{m^2}$) peuvent être utilisées. La référence pour la mesure en dB est alors $1 \times 10^{-12} W$ et $1 \times 10^{-12} W.m^{-2}$. [TS99]

21. A titre indicatif, 1 phones égale 1dB SPL à 1kHz.

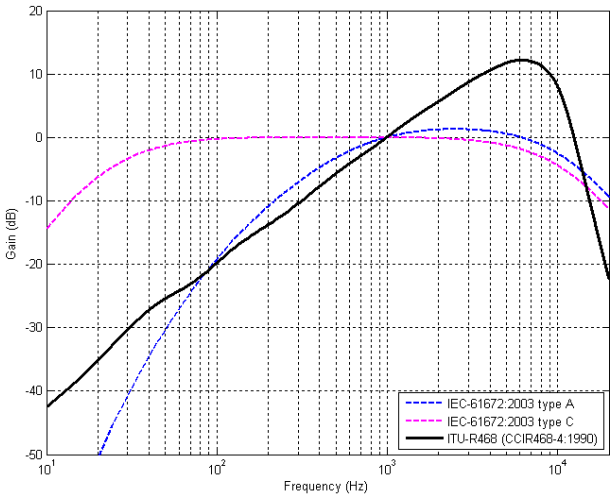


FIGURE 5.1 – Représentation des courbes de pondération fréquentielle type-A, type-C et ITU-R468.

– Type-Z : absence d'atténuation fréquentielle (zéro).

Les pondérations A et C (ainsi que les désormais obsolètes B et D abandonnées lors du passage de la norme IEC-60651 à IEC-61672 :2003) sont nommées arbitrairement utilisant les premières lettres de l'alphabet. La dénomination type-Z a récemment été normée (bien qu'il ne s'agisse pas réellement d'une pondération). Notons enfin que les pondérations type-A et type-C sont quasi-systématiquement disponibles sur les sonomètres.

Mesure relative de l'intensité sonore

La mesure absolue d'un niveau sonore nécessite une référence. Il n'est pas toujours possible de disposer d'une telle référence, particulièrement dans le cas de signaux enregistrés. D'autre part, les signaux d'intérêt sont souvent immergés dans un environnement sonore ou bruit ambiant, et il est alors plus significatif de mesurer l'intensité relative à ce bruit ambiant.

Une mesure permettant d'évaluer le niveau sonore relatif d'un son (événement sonore) dans un environnement sous surveillance est le rapport signal à bruit (RSB). Dans notre contexte, l'événement est le signal et l'environnement ou ambiance est le bruit. Le RSB mesure la log-différence ou rapport de l'intensité moyenne de ces deux signaux. Cette différence peut se calculer si l'on dispose de l'événement et de l'ambiance séparés, ou s'estimer si l'on ne dispose que du mélange des deux.

Le RSB permet de donner une indication sur la difficulté à détecter l'événement anormal. Plus le RSB est important, plus l'événement est fort relativement à l'environnement et plus la détection devra être facile ; inversement, plus le RSB est faible, plus l'environnement masquera l'événement, le rendant alors plus délicat à détecter. Cette mesure est donc particulièrement utile pour construire des signaux d'évaluation.

Il est admis à travers la littérature que le RSB est une mesure adaptée à l'évaluation de systèmes de détection d'événements audio. En particulier, les chercheurs concentrent leurs efforts au développement de solutions performantes dans le cas de RSB faibles, voire négatifs. L'évaluation de ces performances est généralement mise en œuvre en immergeant l'événement

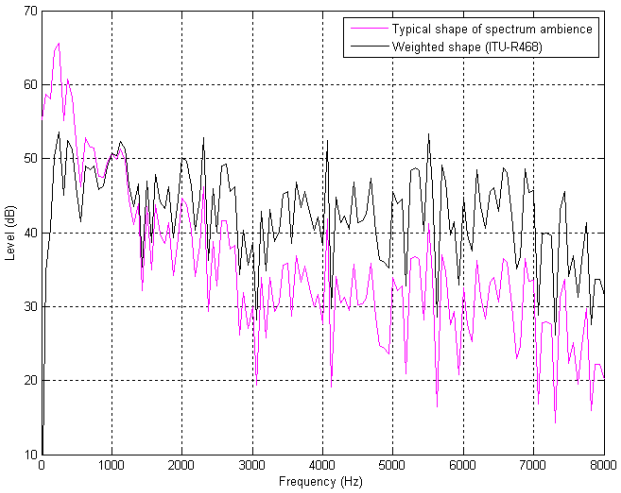


FIGURE 5.2 – Spectre fréquentiel typique d'un signal d'ambiance, avec (noir) et sans (rouge) pondération ITU-R468

à détecter, le signal, dans un bruit de niveau de plus en plus élevé. Le gain étant contrôlé, les performances peuvent alors s'exprimer en fonction du RSB.

5.2.2 Mesure pondérée du RSB dans le contexte audio-surveillance

Il n'est pas courant dans la littérature d'utiliser, dans ce cadre, une mesure de RSB pondérée fréquentiellement. Cependant cette approche permet de tenir compte de la spécificité des signaux étudiés, en ne mesurant le RSB que sur une partie utile du spectre, celle où signal et bruit ont même support. De plus, la littérature utilise souvent comme bruit un bruit blanc [22] qui n'est pas représentatif des bruits attendus dans des conditions réelles.

Dans des signaux d'ambiance réels, on observe qu'une part importante de l'énergie du signal se situe dans les basses fréquences (voir figure 5.2) tandis que l'énergie des signaux anormaux se répartit sur l'ensemble du spectre, voire sur les hautes fréquences. Sans pondération, ce phénomène conduit à une évaluation biaisée du RSB lorsque l'on souhaite qualifier un événement par rapport à une ambiance. En effet, le calcul du rapport s'effectue à partir d'énergies issues de segments disjoints du spectre. Afin de minimiser cet effet, nous proposons d'utiliser l'une des pondérations précédemment introduites. Celle-ci aura alors pour effet de renforcer la partie supposée utile du signal.

Exploitant l'outil décrit à la section suivante, nous avons mélangé 96 événements anormaux aux 18 ambiances de la base Caretaker (voir annexe E). Chaque événement a été inséré à 50 positions différentes dans chacune des ambiances. Ceci a permis de mélanger ambiances et événements dans des conditions variées : pendant les arrivées/départs de trains, avec/sans la présence de voyageurs, etc. Nous avons ensuite calculé le RSB moyen mesuré à l'aide des pondérations présentées précédemment, pour l'ensemble des 86400 réalisations d'événements anormaux ainsi observées. Ces résultats empiriques (mais jugés représentatifs compte-tenu du nombre de réali-

22. On appelle bruit blanc un son aléatoire dont l'énergie spectrale est la même pour toutes les fréquences. Ce son s'apparente à un souffle.

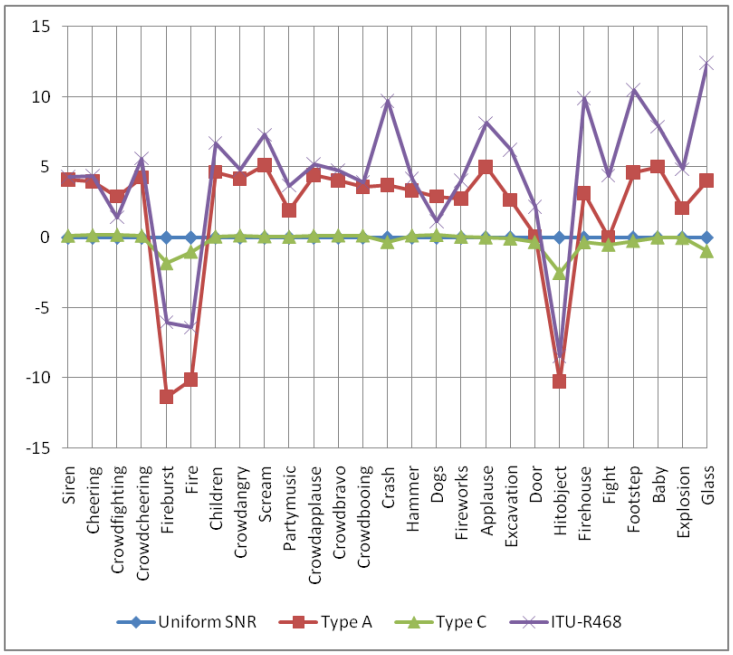


FIGURE 5.3 – Variation du Rapport Signal-à-Bruit (événement à ambiance) suivant la pondération utilisée et pour différents types d'événements

sations) sont présentés au tableau 5.1. Ils montrent que l'utilisation de la pondération ITU-R468 donne des résultats de RSB plus élevés. Cette surestimation de $4,3$dB (en moyenne) constitue un point de vue intéressant pour qualifier un système de surveillance audio. En effet, ces résultats signifient que cette mesure capture plus de différence entre événement et ambiance, autrement dit, elle cible particulièrement bien la partie utile du signal.

Pondération	type-Z	type-A	type-C	ITU-R468
RSB	$10,77$dB	$12,61$dB	$10,5$dB	$15,09$dB

TABLE 5.1 – Mesure de RSB moyen en utilisant différentes pondérations

Nous avons également analysé ces résultats par types d'événements. Dans ce cas, ces derniers ont été mélangés avec un RSB global cible de 0dB type-Z, puis le RSB local a été mesuré à l'aide des différentes pondérations. La figure 5.3 présente les résultats obtenus. Ces résultats vont dans le sens de nos premières constatations, favorisant là encore l'utilisation de la norme ITU-R468 pour les évaluations. Les événements *fire* et *fireburst* (enregistrements de bruits de feux) sont particulièrement riches en basses fréquences, tandis que les événements *hitobjects* sont riches en hautes fréquences ; ainsi la partie utile du spectre pour une mesure correcte de RSB ne se situe plus dans la bande rehaussée par les pondérations pré-sélectionnées. Cependant, compte-tenu du résultat global et des résultats obtenus pour l'ensemble des autres événements anormaux, la pondération ITU-R468 est utilisée pour l'ensemble des résultats présentés.

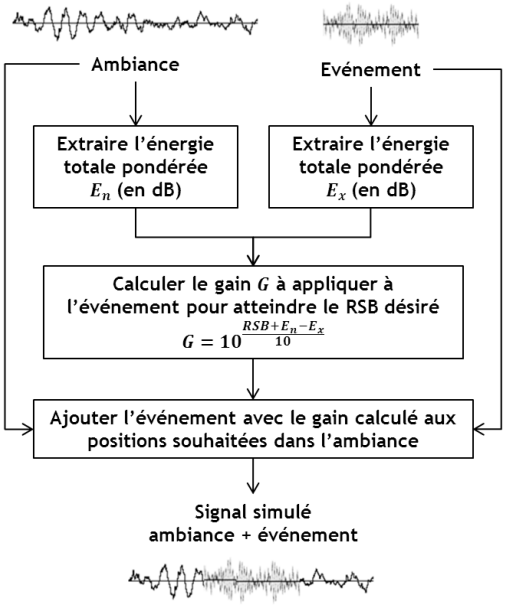


FIGURE 5.4 – Schéma fonctionnel de l'outil de simulation de signaux pour l'évaluation de système de surveillance audio

5.2.3 Simulation de signaux pour l'évaluation de système audio-surveillance

Nous présentons maintenant l'outil réalisé afin de générer des signaux dans le but d'évaluer et qualifier les performances d'un système de surveillance basé sur la modalité audio. En effet, comme aucune base de données spécifique n'existe pour notre contexte (surveillance des stations de métro), nous avons développé un outil permettant de simuler les signaux appropriés.

L'outil développé est présenté à la figure 5.4. L'idée principale est de mélanger des ambiances enregistrées par un système de surveillance et une sélection d'événements sonores anormaux. Afin de qualifier les performances du système de surveillance, l'outil permet un contrôle du volume des événements sonores au sein de l'ambiance en spécifiant un RSB (avec ou sans pondération). Précisons également que le RSB est calculé à partir de l'énergie totale du fichier d'ambiance. Ainsi quel que soit la position de l'événement anormal dans cette ambiance, le gain appliqué est le même. Ceci implique des variations locales du RSB mais augmente le réalisme des évaluations : par exemple, l'intensité sonore d'un coup de feu ne sera pas plus ou moins forte si le niveau sonore de l'environnement sous surveillance est élevé ou non.

Nous nous sommes intéressés à différentes approches permettant de produire des événements anormaux dans une ambiance réelle. Le tableau 5.2 résume nos principales conclusions. Les approches « acter et enregistrer » et « diffuser et enregistrer » constituent une possibilité intéressante pour construire une base d'événements anormaux réalistes, assurant le respect de l'impact de l'environnement acoustique sur les événements anormaux. Bien que nous ayons travaillé dans ce sens, la qualité des signaux recueillis et le temps nécessaire à leur post-traitement ne nous a pas permis de les exploiter. Ajoutons également que les approches « acter et enregistrer » et « diffuser et enregistrer » ne peuvent pas être réalisées pendant les heures d'ouverture des stations. L'environnement sonore étant différent la nuit, cette limitation ternit l'avantage espéré par l'utilisation de ces deux méthodes. Ainsi, nous avons mis en œuvre un outil « enregistrer

et mélanger » pour générer nos signaux.

Méthode	Description	Avantages	Inconvénients
Acter et enregistrer	Enregistrer depuis le système de surveillance des situations anormales actées dans l'environnement réel.	Les signaux les plus réalistes.	Coût important, certains événements irréalisables (explosions), contrôle du RSB impossible.
Diffuser et enregistrer	Diffuser dans l'environnement à l'aide d'un système de restitution des événements préalablement enregistrés.	L'acoustique de l'environnement est respectée pour les événements anormaux. Une grande variété d'événements anormaux peut être utilisée.	Déplacement sur place nécessaire, contrôle du RSB limité. Voir annexe A.
Enregistrer et mélanger	Ajouter des événements anormaux préalablement enregistrés à des ambiances également enregistrées.	Coût très faible, contrôle précis du RSB et de la position (temporelle) des événements et rapidité de réalisation d'une base de données incluant de nombreux événements et ambiances.	Les propriétés acoustiques de l'environnement ne sont pas préservées pour les événements anormaux.

TABLE 5.2 – Méthodes pour simuler des signaux de surveillance avec événements anormaux : description, avantages et inconvénients

5.3 Critères d'évaluation

5.3.1 Généralités

L'évaluation d'un système de classification ou de détection s'appuie sur quatre grandeurs élémentaires. Par convention, les observations sont dites positives lorsqu'elles appartiennent à la classe pour laquelle les performances sont évaluées et négatives dans le cas contraire.

Les deux premières grandeurs comptent le nombre d'observations positives, respectivement négatives, qui sont classées positives, respectivement négatives ; il s'agit des vrais positifs, notés TP (*true positives*), et des vrais négatifs, notés TN. Les deux autres grandeurs comptent le nombre d'observations mal détectées ou classifiées. Le nombre d'observations négatives mal classées, les faux positifs notés FP, constituent les erreurs de type I. Enfin, le nombre d'observations positives mal classées, les faux négatifs notés FN, constituent les erreurs de type II [23].

On s'appuie en général également sur le nombre total d'observations positives ou négatives (notés P et N) et le nombre total d'observations pour lesquelles le système prend une décision

23. On distingue les erreurs de type I de celles de type II car les premières ont des conséquences généralement plus importantes ; par exemple, décider qu'un médicament est efficace alors qu'il n'est est rien.

	Observation positive	**Observation négative**	Total
Décision positive	Vrai positif TP	Faux positif FP (I)	P'
Décision négative	Faux négatif FN	Vrai négatif TN (II)	N'
Total	P	N	

TABLE 5.3 – Grandeurs élémentaires pour l'évaluation des performances des systèmes de détection et classification

positive ou négative (notées P' et N'). Ces grandeurs peuvent déjà être déduites des 4 premières. Le tableau 5.3 synthétise l'ensemble de ces grandeurs.

Dans le cas de la classification, les erreurs peuvent être ventilées par classes dans une matrice, appelée matrice de confusion. Les faux positifs et faux négatifs sont alors répartis suivant les colonnes et les lignes de cette matrice et on retrouve l'ensemble des vrais positifs, pour chaque classe, le long de la diagonale.

Une grande partie des mesures qualifiant des résultats de détection ou de classification s'appuient sur ces valeurs ou sont dérivées de celles-ci [Faw04]. Nous allons maintenant présenter les mesures que nous avons utilisées dans le cadre des résultats présentés par la suite.

5.3.2 Probabilité de bonne classification

La probabilité de bonne classification mesure la proportion des observations testées qui ont été correctement classées. Cette mesure simple permet d'évaluer très rapidement les performances d'un détecteur sur une base de test équilibrée. On l'utilise en particulier lors des étapes de validation croisée afin d'estimer les paramètres d'un modèle.

On détermine la probabilité de bonne classification par l'expression suivante :

$$Pbc = \frac{TP + TN}{P + N} \tag{5.2}$$

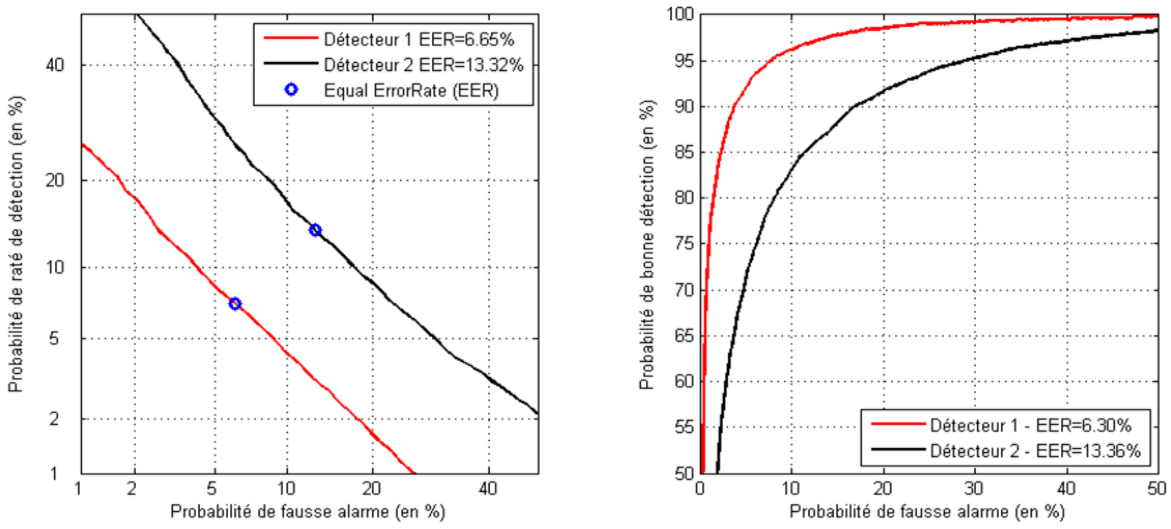
Il ne faut pas confondre la probabilité de bonne classification avec la probabilité de bonne détection. La première considère l'ensemble des observations quelle que soit la classe à laquelle elles appartiennent. La seconde ne s'intéresse qu'aux observations d'une classe donnée, et donc à la détection correcte de cette classe (TP/P).

5.3.3 Courbes DET (*Detection Error Trade-off*)

Afin d'évaluer les performances complètes d'un système de détection d'événements anormaux, nous utilisons les courbes DET (*Detection Error Tradeoff*) [MDK+97]. Ces courbes sont notamment couramment utilisées dans la communauté de la reconnaissance du locuteur.

Les courbes DET confrontent les deux types d'erreurs d'un détecteur. Elles illustrent ainsi le compromis entre probabilité de faux positifs (fausse-alarme) :

$$Pfa = \frac{FP}{N} \tag{5.3}$$



(a) (b)

FIGURE 5.5 – Comparaison de deux classificateurs à l'aide (a) de courbes DET et (b) de courbes ROC. La représentation linéaire des courbes DET (dues à la transformation des axes) permet de mettre naturellement en exergue la zone d'information du graphique.

et probabilité de faux négatifs (ratés de détection) :

$$Pmd = \frac{FN}{P} \tag{5.4}$$

Les axes non linéaires utilisés pour leur tracé suivent la fonction quantile (fonction de répartition inverse) d'une distribution normale. Les courbes sont généralement obtenues en prenant différentes configurations d'un même algorithme ou en faisant varier le seuil de détection. Elles permettent ainsi sur une même figure de représenter différents détecteurs et leurs performances respectives, ou encore les performances d'un même détecteur sur des bases d'évaluation différentes (événements anormaux à différents SNR par exemple).

Ces courbes constituent une alternative aux populaires courbes ROC (*Receiver Operating Characteristics*). Dans le cas des courbes DET, la région d'intérêt occupe néanmoins une place plus importante du graphique comme l'illustre la figure 5.5. Les résultats peuvent ainsi être comparés avec plus de facilité [Faw04].

Ajoutons que la lecture des courbes DET afin de choisir un point de fonctionnement du détecteur est immédiate. En effet, les deux axes représentent des erreurs concédées. Enfin, nous représentons également sur certaines courbes le point de taux d'erreurs égales (EER, *Equal Error Rate*), un point de fonctionnement singulier du détecteur appris.

5.4 Familles de fonctions de décision

Nous l'avons évoqué en introduction de ce document, le contrôle du compromis entre fausses-alarmes et ratés de détection est un aspect opérationnel important pour un système de surveillance. Nous présentons maintenant une approche permettant à partir d'une fonction de décision, issue de la résolution d'un problème SVM 1-classe, d'engendrer une famille de fonctions de

décision. Cette approche va permettre de modifier le point de fonctionnement du système sans nécessairement réaliser un nouvel apprentissage.

Dans le contexte des SVM 1-classe, le paramètre ν est lié à la fraction des données d'apprentissage supposées normales mais néanmoins rejetées, donc à la fausse-alarme. Cependant, cette approche est limitée car la modification de ce paramètre oblige à réaliser un nouvel apprentissage du modèle SVM, souvent long.

Soit C_0 la classe correspondante aux données de l'ensemble d'apprentissage et Γ le volume estimé autour de ces données. La formulation SVM 1-classe considère l'unique hypothèse H_0 : « l'observation appartient à la classe C_0 ». A cette hypothèse est associée la fonction de décision suivante : si $f(\mathbf{x}) \geq 0$ alors $\mathbf{x}$ vérifie H_0 (décision D_0). Nous proposons de définir une hypothèse H_1 « l'observation n'appartient pas à C_0 » et d'introduire le seuil $\lambda \in \mathbb{R}$ afin de construire une famille de règles de décision :

$$\begin{cases} \text{si } f(\mathbf{x}) \geq \lambda, \text{ alors } \mathbf{x} \in C_0 \ (D_0) \\ \text{si } f(\mathbf{x}) < \lambda, \text{ alors } \mathbf{x} \notin C_0 \ (D_1) \end{cases} \tag{5.5}$$

Nous pouvons alors définir les probabilités de non-détection et de fausse-alarme, respectivement $P(D_0|H_1)$ et $P(D_1|H_0)$, qui déterminent le point de fonctionnement de notre détecteur.

Dans cette formulation, le seuil λ, qui paramètre une translation de l'hyperplan W dans l'espace de représentation $\mathcal{H}$ (espace des *features*), permet de contrôler le compromis entre fausse-alarme et non-détection sans avoir à réaliser un nouvel apprentissage. La frontière de l'enveloppe Γ qui en résulte dans l'espace d'observation $\mathcal{X}$ est une ligne de contour de la fonction de décision $f(\mathbf{x})$. On ajuste λ expérimentalement en fonction des besoins opérationnels.

Le choix de ν est un problème délicat car, pour des valeurs faibles, Γ est estimé dans des régions où la densité de probabilité des données d'apprentissage est très faible (variance d'estimation élevée). A l'inverse, pour des valeurs de ν élevées, un biais important dans l'estimation de Γ pourrait conduire à une représentation sous-optimale de la distribution des données normales. Dans notre formulation le paramètre ν est indépendant des besoins opérationnels et est uniquement conditionné par le signal d'apprentissage : c'est une estimation de la fraction de données qui doivent être exclues du domaine Γ.

Les résultats présentés au chapitre 6 participent à valider opérationnellement cette formulation.

Chapitre 6

Détection par SVM 1-classe avec biais

Dans ce chapitre, nous présentons les performances de l'approche SVM 1-classe dans le cas applicatif qui nous intéresse, la surveillance de zone par modalité audio. En particulier, nous présentons dans un premier temps les données puis le protocole expérimental nous ayant conduit à fixer le paramètre de noyau σ (noyau Gaussien) et le paramètre ν. Nous nous intéressons ensuite à la validation sur nos données de l'approche par famille de fonctions de décision introduite au chapitre 5. Enfin, nous étudions l'influence sur les performances en détection de la prise en compte d'une information temporelle. Ces résultats sont par ailleurs exprimés pour différents niveaux de rapport signal à bruit (RSB).

6.1 Protocole

6.1.1 Données

Les résultats présentés dans ce chapitre exploitent les signaux audio enregistrés à la station *XVIII diciembre* du métro de Turin, dans le cadre du projet VANAHEIM. Ces signaux sont numérisés à une fréquence d'échantillonnage de 16kHz, et linéairement quantifiés sur 16 bits. Deux flux sont utilisés : le premier correspond aux enregistrements provenant d'un micro situé au-dessus de l'un des quais (*mic07*), le second correspond aux enregistrements mélangés de deux micros situés autour des tourniquets permettant l'accès à la station (*mic12*). La durée totale des signaux utilisés est de 1 heure pour l'apprentissage et 30 minutes pour l'évaluation. Ces créneaux se suivent, il s'agit d'enregistrements en semaine, dans l'après-midi. Un ensemble de 118 événements sonores extraits d'une base de données commerciale [Sou12] a été utilisé pour générer des événements anormaux (voir annexe C). Les signaux d'évaluation ont été construits en exploitant l'approche décrite dans le chapitre 5. En particulier, nous avons créé 150 réalisations de chaque événement, aux RSB de 0, 5, 10, 15 et 20 dB. La mesure de RSB a notamment été pondérée suivant la norme ITU-R468 (cf. chapitre 5).

Le signal audio est analysé à l'aide de descripteurs acoustiques spectraux. Il s'agit des énergies en sortie d'un banc de 16 filtres triangulaires avec recouvrement de 50% répartis suivant une échelle linéaire sur l'ensemble de la bande passante du signal (0-8kHz). Nous avons choisi d'utiliser cette représentation pour les raisons suivantes :

- une représentation spectrale permet d'interpréter plus facilement les résultats,
- ce choix de descripteur ne reflète aucun *a priori* concernant les signaux,

– l'objectif premier étant d'évaluer l'algorithme de détection, on se réserve la possibilité d'adapter les descripteurs aux conditions grâce aux méthodes décrites au chapitre 2,

De l'ensemble de ces choix, il résulte alors une base de données comprenant :

– 360160 vecteurs dans l'ensemble d'apprentissage, soit 1 heure de signal,
– 180080 vecteurs dans l'ensemble de test correspondant à une ambiance normale, soit 30 minutes de signal,
– 1973121 vecteurs dans l'ensemble de test correspondant à un événement anormal, soit $5,5$ heures de signal.

La base de données a également été normalisée suivant le vecteur moyen et la variance calculés sur l'ensemble d'apprentissage. On dispose alors initialement de quatre conditions d'évaluation correspondantes à :

– 2 zones : quai (*mic07*) et tourniquets (*mic12*),
– 2 représentations : données brutes et données normalisées.

6.1.2 Normalisation et largeur de noyau, choix des paramètres

Rappelons que l'une des contraintes opérationnelles identifiées en introduction est l'automatisation du choix des paramètres du système de surveillance. Dans ce contexte, nous avons étudié deux méthodes afin de déterminer la largeur du noyau Gaussien. Nous présentons dans les paragraphes qui suivent ces deux approches, avant de conclure sur le choix d'une méthode pour la suite des évaluations.

La première approche s'appuie sur une stratégie de validation croisée. Nous avons estimé la probabilité de fausse alarme du système avec ν fixé et σ variant de 2^1 à 2^{13}. La validation croisée s'opère sur 10 sous-ensembles de l'ensemble d'apprentissage. Les résultats, pour deux valeurs $\nu = .1$ et $\nu = .01$ et pour chaque condition d'évaluation, sont présentés aux figures 6.1 et 6.2.

La seconde approche s'appuie sur la suggestion de Li *et al* [LTKZ09]. Leur méthode consiste à rechercher une valeur de σ au sein de l'ensemble $\left\{0,25\sqrt{\gamma}; 0,5\sqrt{\gamma}; \sqrt{\gamma}; 2\sqrt{\gamma}; 4\sqrt{\gamma}\right\}$ où γ est la distance moyenne entre chaque paire d'observations. L'avantage de cette approche est de pouvoir estimer l'ordre de grandeur de σ de manière indépendante de tout apprentissage. Les valeurs déterminées par cette approche sont :

– $\sqrt{\gamma} = 24,31$ pour les signaux *mic07*,
– $\sqrt{\gamma} = 21,25$ pour les signaux *mic12*.
– $\sqrt{\gamma} = 5,68$ pour les signaux *mic07* normalisés.
– $\sqrt{\gamma} = 5,60$ pour les signaux *mic12* normalisés.

Nous comparons maintenant les résultats obtenus pour différentes configurations d'apprentissage. En particulier, nous nous intéressons à l'apport de la normalisation des données et au choix définitif du paramètre de noyau σ parmi $\sigma = \sqrt{\gamma}$ et $\sigma = 4\sqrt{\gamma}$. L'algorithme utilisé est FastOC2 ; il résout le problème SVM 1-classe avec biais.

La figure 6.3 détaille les performances des différents détecteurs pour les signaux *mic07*. L'étude de ces résultats montre que la normalisation des données, le choix d'une valeur σ faible et d'une valeur de ν élevée apportent des gains sensibles ; le gain observé entre les deux extrema est de près de 40%.

La figure 6.4 détaille les performances des différents détecteurs pour les signaux *mic12*. Bien que l'amélioration paraisse moins nette (gain de 20% au mieux), ces résultats confirment notre précédente conclusion concernant l'apport de la normalisation et le choix d'une valeur de ν élevé. Notons cependant que le choix d'une valeur de σ faible apporte un gain faible dans le cas d'un ν élevé et dégrade très légèrement les performances dans le cas d'un ν faible. Ce comportement, ainsi que les performances globalement meilleures par rapport aux signaux *mic07*, s'explique

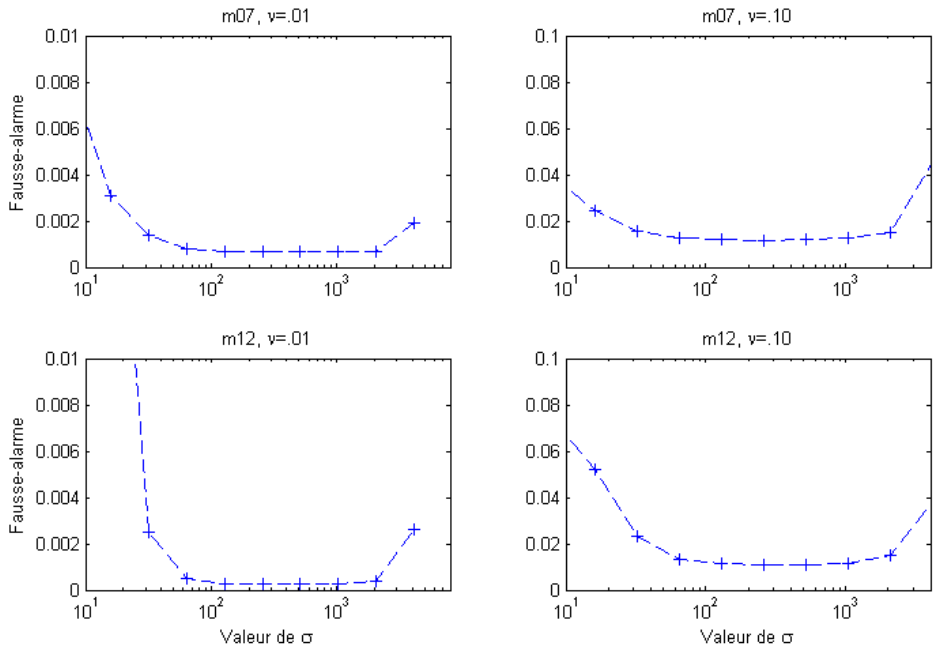


FIGURE 6.1 – Recherche du paramètre de noyau Gaussien par validation croisée pour les données non normalisées

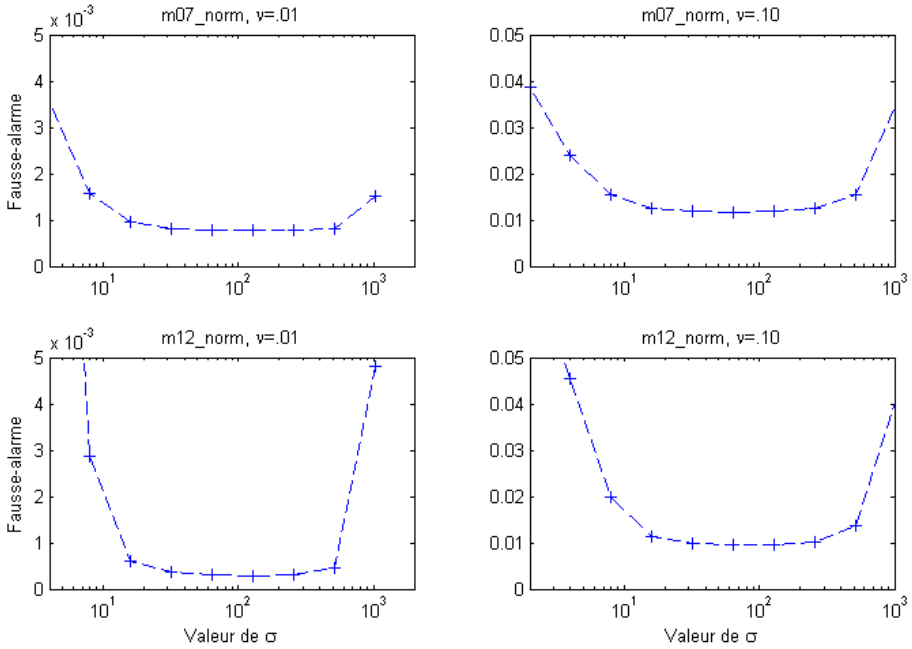


FIGURE 6.2 – Recherche du paramètre de noyau Gaussien par validation croisée pour les données normalisées.

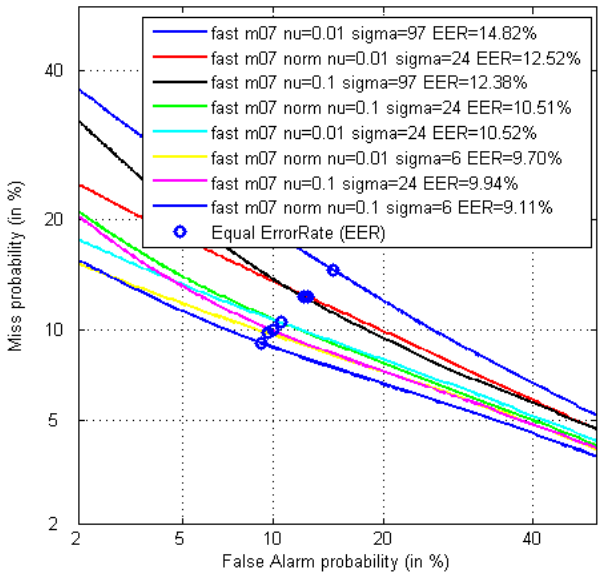


FIGURE 6.3 – Courbes DET pour différents détecteurs appliqués aux signaux *mic07*.

par une faible variabilité dans les signaux *m12* et un environnement de nature différente (moins riche en basses fréquences).

Cette expérimentation nous conduit à certaines conclusions. Tout d'abord, le fait de normaliser les données apporte un gain significatif aux performances en détection. On explique notamment ce comportement par une meilleure adaptation du noyau aux données lorsque la dynamique de celles-ci est homogène. Ensuite, les résultats obtenus pour des environnements différents nous permettent d'affirmer que les performances d'un système de détection non supervisé ne peuvent être exprimées indépendamment de l'environnement dans lequel il est évalué. Enfin, si le choix du paramètre de noyau σ reste délicat, l'approche proposée satisfait un besoin d'automatisation, tout en garantissant des performances acceptables. A ce titre, la valeur $\sigma = \sqrt{\gamma}$ est retenue pour l'ensemble des expérimentations à suivre.

Les résultats que nous avons présentés permettent également de conclure sur l'amélioration des performances qu'implique le choix d'une valeur de ν élevée. Cependant, comme le rapporte le tableau 6.1, cette amélioration se traduit par une augmentation importante du temps d'apprentissage. Nous allons étudier plus particulièrement l'influence de ce paramètre dans l'expérimentation suivante.

6.2 Familles de fonctions de décision

Nous présentons maintenant les résultats liés aux expérimentations menées pour valider l'approche présentée à la section 5.4 de ce manuscrit. Pour rappel, celle-ci consiste à s'appuyer sur la fonction de décision SVM pour construire une famille de fonctions de décision. Cette famille est parcourue en faisant varier le seuil de décision, noté λ. L'avantage majeur de cette approche réside dans la possibilité de modifier le point de fonctionnement (rapport entre fausse-alarme et ratés de détection) du détecteur sans réaliser de nouvel apprentissage. Le risque en

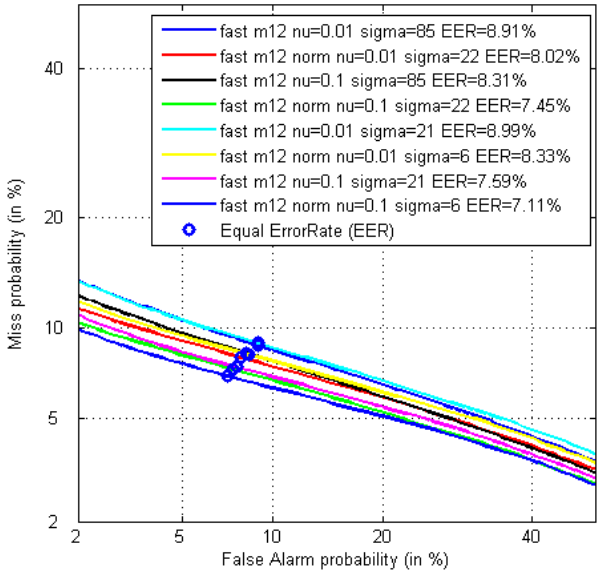


FIGURE 6.4 – Courbes DET pour différents détecteurs appliqués aux signaux *mic12*.

revanche est que les propriétés de la fonction de décision SVM ne soient plus garanties dès que ce seuil est modifié. Nous allons montrer dans les paragraphes et illustrations qui suivent que dans le cas de nos signaux ce risque est négligeable d'une part, et que l'approche améliore les résultats attendus d'autre part.

La construction des expérimentations est la même que précédemment. Les descripteurs acoustiques du signal sont des énergies en sortie d'un banc de 16 filtres linéairement répartis sur la bande passante des signaux (0-8kHz) ; ceux-ci sont par ailleurs centrés-réduits à l'aide de la moyenne et de la variance des signaux d'apprentissage. Nous étudions là encore les deux sources de signaux nommées *m07* et *m12*. Enfin, conformément à nos précédentes conclusions, le paramètre de largeur de noyau est fixé à $\sigma = 5,68$ pour *m07* et $\sigma = 5,60$ pour *m12*.

Nous allons étudier les points de fonctionnement obtenus par la fonction de décision SVM ($\lambda = 0$) pour différentes valeurs du paramètre ν. Pour rappel, ce paramètre gère le compromis entre maximisation de la marge et minimisation des erreurs ; il correspond à la borne supérieure de la probabilité de fausses alarmes sur les données d'apprentissage. Ces points de fonctionnement sont comparés aux familles de fonctions de décision obtenues pour chacun des modèles (variations de λ). On rappelle que modifier ν nécessite de réaliser un nouvel apprentissage ; ce n'est pas le cas de λ.

Les figures 6.5(a) et 6.5(b) présentent respectivement les résultats de l'expérimentation sur les signaux *m07* et *m12*. Dans les deux cas, le choix d'une valeur de ν faible, impliquant une variance d'estimation élevée, conduit à une famille de fonctions de décision moins performante (courbes bleues). Par ailleurs, dans cette situation, des régions de l'espace d'observation dans lesquelles l'ensemble d'apprentissage présente des observations en faible densité sont incluses dans le volume estimé ; le risque encouru est alors un sur-apprentissage et ce phénomène se traduit par une violation de la borne supérieure sur la probabilité de fausse alarme. A l'inverse, le choix d'une valeur de ν élevée conduit à de meilleures performances. Notons cependant que

ν	Descripteurs	Source *m07*		Source *m12*	
		σ	t (s)	σ	t (s)
$0,01$	Bruts	97	$771,3$	85	$777,2$
		24	$712,0$	21	$656,3$
	Normalisés	24	$774,8$	22	$788,2$
		6	$711,1$	6	$683,1$
$0,1$	Bruts	97	$8291,4$	85	$7465,1$
		24	$7440,9$	21	$6263,4$
	Normalisés	24	$7764,0$	22	$6609,8$
		6	$7554,5$	6	$6543,6$

TABLE 6.1 – Temps d'apprentissage (en secondes) des modèles correspondant à différentes conditions.

dans le cas d'une valeur très élevée de ν, la structure des données peut être mal capturée et la famille de fonctions de décision apporte des performances inégales ; c'est le cas pour les signaux *m07*.

Ainsi, bien que l'approche proposée apporte une souplesse de fonctionnement, elle n'éclipse pas pour autant le choix délicat d'une valeur de ν. Celui-ci doit par ailleurs se faire en gardant à l'esprit que plus cette valeur est élevée, plus le temps d'apprentissage est conséquent. A ce titre, la figure 6.6 rapporte les temps d'apprentissage des différents modèles utilisés dans cette expérimentation.

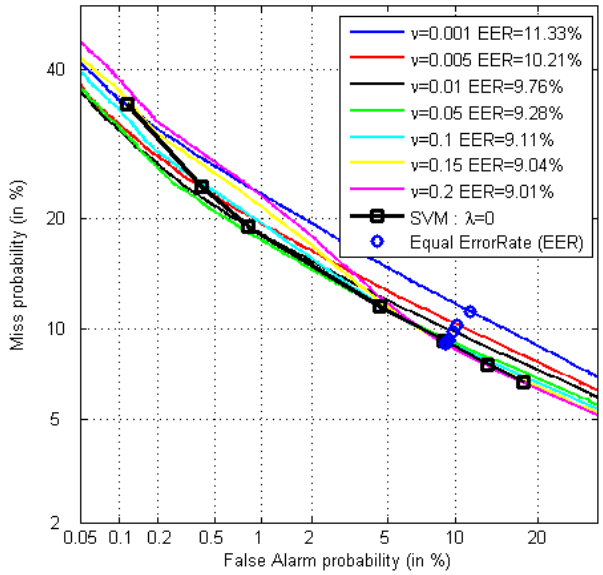
Les résultats présentés dans ces paragraphes confirment que l'approche proposée se prête aux signaux audio et particulièrement aux signaux de surveillance tels que ceux qui nous préoccupent. De plus, il s'avère que pour un taux de fausse-alarme fixé, notre approche permet d'obtenir de meilleurs résultats que par la modification de la valeur de ν. Enfin, ces travaux nous permettent pour la suite de fixer notre paramètre $\nu = 0,05$. En pratique, sur de nouveaux signaux, on pourra envisager de tester différentes valeurs afin d'assurer un fonctionnement optimal. En particulier, pour un système où l'on concède plus de fausses alarmes afin de réduire les ratés de détection, un choix de ν élevé peut s'avérer intéressant sous réserve d'un temps d'apprentissage raisonnable.

6.3 Performances du système proposé

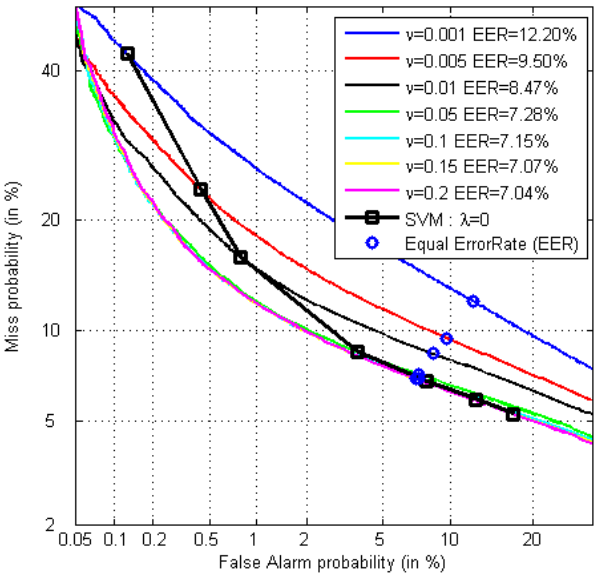
Nous présentons dans les paragraphes suivants les performances du système de détection d'événements anormaux réalisé et dont les paramètres ont été déterminés suivant le protocole précédemment décrit. Dans un premier temps, nous étudions les résultats en fonction du rapport signal à bruit des événements anormaux. Dans un second temps, nous illustrons l'apport de la prise en compte d'informations temporelles pour intégrer les résultats de décision.

6.3.1 Influence du SNR

Le système de détection d'événements anormaux proposé a été évalué sur l'ensemble des signaux d'évaluation générés (RSB de 0 à 20dB ITU-R468). Les modèles SVM 1-classe sont appris avec une valeur $\nu = 0,05$ et des paramètres de largeur de noyau $\sigma = 5,68$ (signaux *m07*) et $\sigma = 5,60$ (signaux *m12*). La figure 6.7 présente les résultats obtenus pour les signaux *m07* (quai de métro) tandis que la figure 6.8 présente ceux obtenus pour les signaux *m12* (tourniquets).



(a)



(b)

FIGURE 6.5 – Familles de fonctions de décision obtenues pour différentes valeurs de ν ((a) *m07* et (b) *m12*). Chaque famille est représentée par une courbe, ensemble des points de fonctionnement possibles en faisant varier λ après un apprentissage (ν fixé). Sur chaque courbe, un symbole « □ » localise le point de fonctionnement correspondant à la fonction de décision SVM 1-classe ($\lambda = 0$).

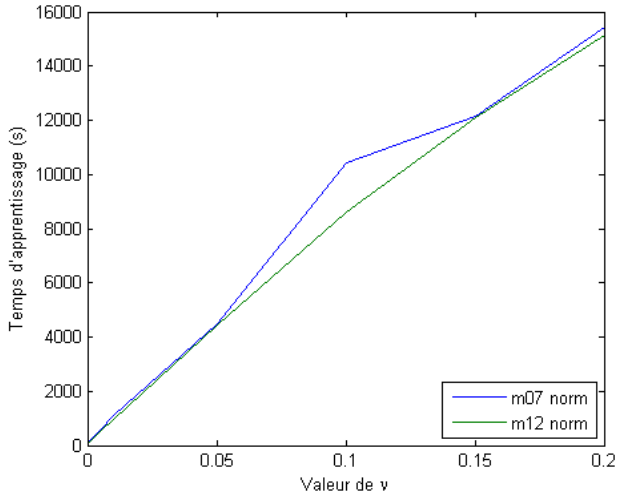


FIGURE 6.6 – Temps d'apprentissage des modèles SVM 1-classe avec l'algorithme FastOC2 sur les signaux *m07* et *m12*.

Notons que ces résultats sont obtenus sur la base de signaux audio collectés dans le cadre du projet VANAHEIM. A titre indicatif, l'annexe E présente des résultats obtenus au début de la thèse sur une autre base de données, issue du projet CARETAKER [CAR08]. Ces résultats préliminaires avaient été comparés à l'approche *Gaussian Mixture Models* développée chez Thales [CR10] et faisant référence au début des travaux.

6.3.2 Prise en compte d'informations temporelles

Nous montrons maintenant qu'il est possible d'améliorer les scores de détection en intégrant temporellement la statistique de décision. Comme les événements que nous souhaitons détecter sont de taille variable, nous utilisons des informations de segmentation du signal audio afin de procéder à l'intégration. Enfin, dans les résultats qui suivent, un événement anormal est considéré comme étant correctement détecté si le score d'au moins l'un des segments analysés recouvrant totalement ou partiellement l'événement franchit le seuil de détection λ.

L'approche choisie a été proposée dans [CR10] et est présentée à l'annexe D. Les paramètres acoustiques utilisés résultent d'une analyse spectrale non pondérée : il s'agit des énergies en sortie d'un banc de 32 filtres triangulaires avec recouvrement de 50% répartis suivant une échelle linéaire sur l'ensemble de la bande passante du signal (0-8kHz). Enfin, le score d'un segment correspond à la moyenne des scores des trames qui constituent ce segment.

Afin de simplifier la lecture des courbes nous ne présentons dans cette section que les cas des SNR les plus défavorables, soit 0dB, 5dB et 10dB (ITU-R468). La figure 6.9 compare les résultats obtenus avec ou sans segmentation pour les signaux *m07* tandis que la figure 6.10 compare ces mêmes résultats pour les signaux *m12*.

6.4 Performances du système proposé

Au travers de ce chapitre, nous avons mis en évidence les bénéfices d'un ensemble d'éléments associés à l'approche que nous proposons.

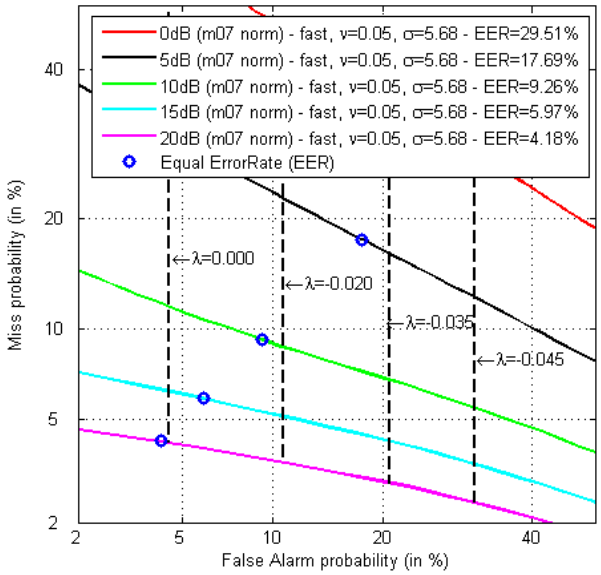


FIGURE 6.7 – Performances du détecteur pour différents SNR sur les signaux *m07*, les points de fonctionnement correspondant à des valeurs de λ fixées sont représentés par des segments pointillés noirs.

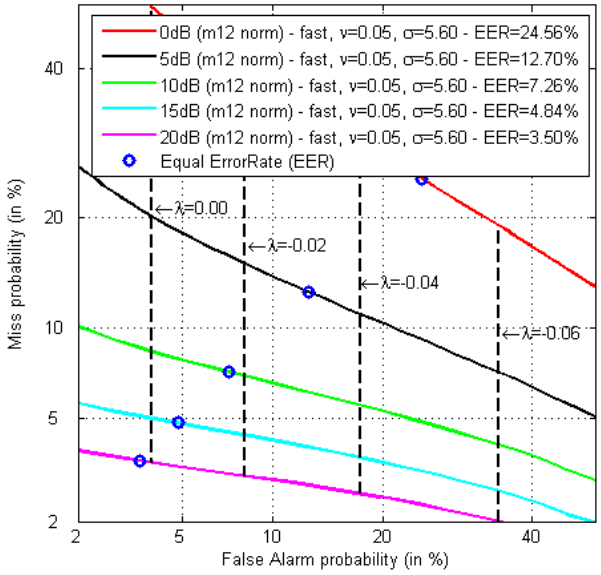


FIGURE 6.8 – Performances du détecteur pour différents SNR sur les signaux *m12*, les points de fonctionnement correspondant à des valeurs de λ fixées sont représentés par des segments pointillés noirs.

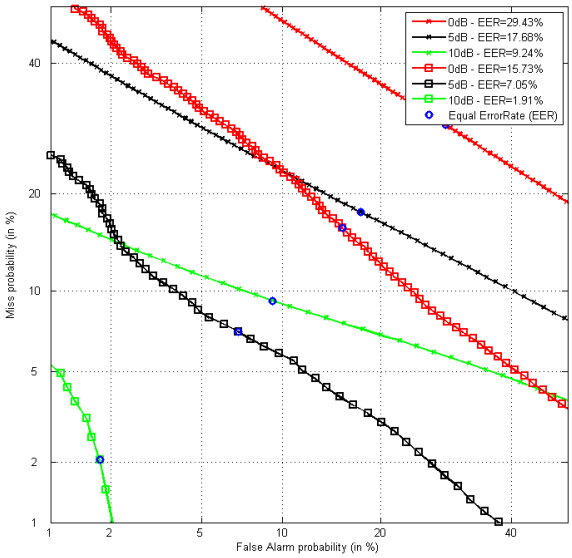


FIGURE 6.9 – Apports de l'intégration des scores par segments sur les performances en détection pour les signaux *m07*.

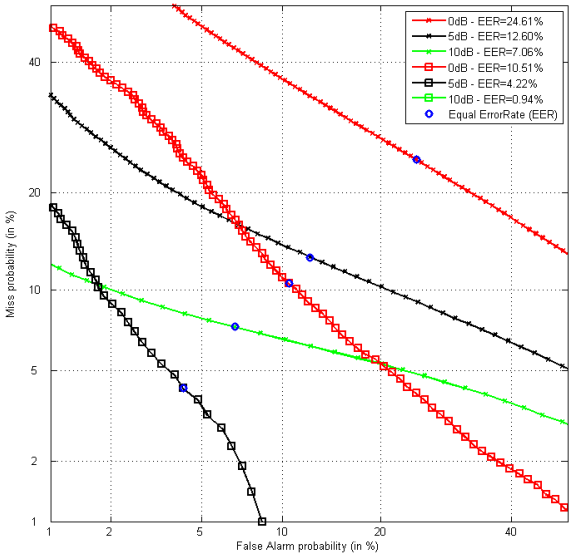


FIGURE 6.10 – Apports de l'intégration des scores par segments sur les performances en détection pour les signaux *m12*.

Dans un premier temps, nous nous sommes intéressés aux hyper-paramètres du problème SVM. Nous avons montré que la normalisation des données, associée à un calcul déterministe du paramètre σ de largeur du noyau gaussien, apporte des résultats satisfaisants. En particulier, l'expression analytique utilisée pour calculer ce paramètre conduit à un choix proche de celui que nous aurions fait lors d'une recherche par grille, sur les signaux étudiés. L'analyse de ces résultats montre également que le choix d'une valeur de ν (le second hyper-paramètre) élevée permet d'obtenir de meilleures performances au détriment de l'augmentation du temps d'apprentissage. Ce compromis plus finement analysé dans le contexte de l'algorithme sans biais au paragraphe suivant.

Dans un second temps, nous avons validé pour les données sur lesquelles nous travaillons l'intérêt de construire des familles de fonction de décision, exploitant l'approche présentée au chapitre précédent. Les résultats montrent notamment que indépendamment de l'apprentissage réalisé et du niveau de rapport signal à bruit des événements anormaux considérés, le choix du paramètre λ permet de spécifier une fausse-alarme attendue du système. Ainsi, disposant uniquement de signaux d'ambiance différents de ceux utilisés pour l'apprentissage, notre approche permet de contrôler le point de fonctionnement du système. La probabilité de ratés de détection alors associée à ce point de fonctionnement dépend quant à elle du rapport signal à bruit des événements anormaux en présence.

Enfin, le signal audio, et particulièrement une ambiance telle que celles considérées dans notre étude est un continuum temporel. Les performances extrêmement bonnes obtenues après intégration de la statistique de décision sur des informations temporelles issue d'une segmentation du signal permettent de valider d'une part cette hypothèse, et d'autre part la méthode de segmentation en ligne choisie.

A l'issue de ce chapitre, nous pouvons considérer disposer d'une solution adaptée au problème de détection d'événements anormaux. Celle-ci consiste à exploiter une modélisation SVM 1-classe des signaux d'apprentissage pour laquelle nous avons proposé une manière rapide de spécifier les hyper-paramètres. Le contrôle du point de fonctionnement du détecteur réalisé peut également être réalisé après apprentissage par l'ajustement d'un paramètre dont le choix est lié à la fausse-alarme attendue. Enfin, s'appuyant sur les caractéristiques du signal audio, nous avons montré qu'il était possible d'exploiter une information temporelle afin de lisser la statistique de décision et obtenir d'excellentes performances. Le principal inconvénient de cette approche réside dans le temps d'apprentissage conséquent du modèle.

Chapitre 7

Evaluation des algorithmes SVM 1-classe avec biais et sans biais

Dans ce chapitre, nous présentons les résultats comparatifs des algorithmes avec biais et sans biais. Dans un premier temps, nous rapportons les expérimentations conduites sur des bases de données simples. Celles-ci permettent de valider l'approche et qualifier les gains de l'algorithme smgo. Dans un second temps, nous reprenons les expérimentations du chapitre 6. En particulier, nous présentons les résultats obtenus en utilisant l'algorithme sans biais en lieu et place de l'algorithme avec biais utilisé dans les chapitres précédents. Nous étudierons les performances, la vitesse de convergence et le nombre de vecteurs de support résultant des apprentissages.

7.1 Résultats préliminaires

Dans cette section, nous décrivons des évaluations comparatives des algorithmes sans biais et avec biais (identifiés respectivement smgo et fast) . Pour chaque expérience, nous comparons les résultats des deux algorithmes, avec et sans les contraintes du problème 2-classes (identifiés respectivement OC2 et OC).

Pour l'ensemble des expériences, nous choisissons un noyau Gaussien RBF dont le paramètre σ est estimé suivant [LTKZ09] ($\sigma = \sqrt{\gamma}$) et un pré-traitement est appliqué pour centrer-réduire les données.

7.1.1 Données de synthèse

Les données sont issues de deux distributions normales de variance 1 ; de moyenne $[0; 0]$ pour la classe $+1$ (données normales) et de moyenne $[3; 3]$ pour la classe -1 (données anormales). Les résultats sont moyennés sur 100 réalisations de chaque expérience.

Dans un premier temps, on s'intéresse à la vitesse de convergence lorsque la taille de l'ensemble d'apprentissage augmente. On génère 100 à 10000 observations pour la classe $+1$ et 100 pour la classe -1. Le paramètre ν est fixé à 10^{-2}. Dans un second temps, on s'intéresse à la vitesse de convergence lorsque le paramètre ν diminue. Cette étude relève d'un besoin opérationnel car ν est à relier au taux de fausses alarmes cible du détecteur d'anormalités [SPST$^+$01]. Dans cette seconde expérience, le nombre d'observations de la classe $+1$ est fixé à 2000 échantillons et le paramètre ν varie de 2.10^{-1} à 2.10^{-3}. La figure 7.1 présente les résultats obtenus.

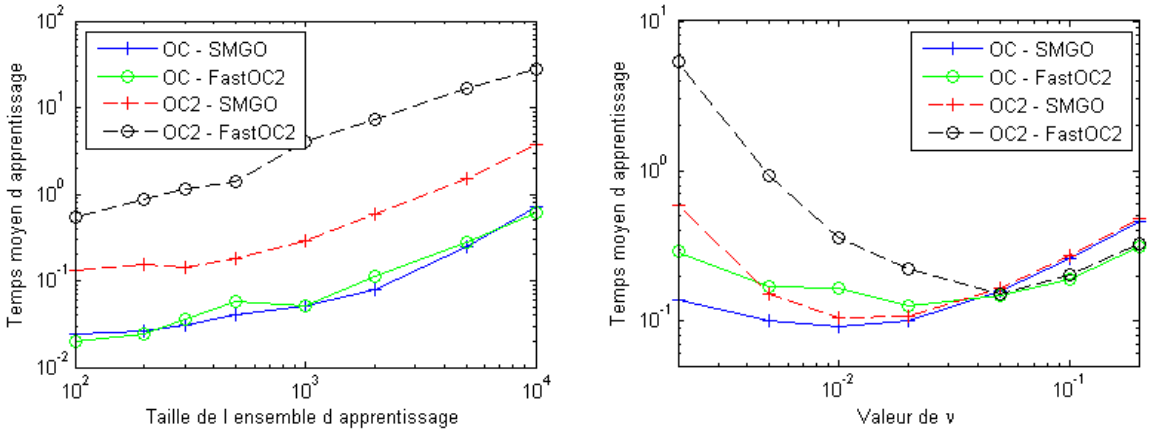


FIGURE 7.1 – Evolution du temps moyen de convergence en fonction de la taille de l'ensemble d'apprentissage (à gauche) et en fonction de la valeur de ν (à droite). Les échelles sont logarithmiques afin d'en clarifier la lecture.

7.1.2 Données réelles

On utilise les bases de données Cancer, Crab, Glass, Iris et Wine du répertoire UCI [24]. La stratégie multi-classe appliquée est celle présentée dans [TL11] et les performances sont évaluées suivant une procédure *leave-one-out*. Enfin, ν est fixé à 10^{-3}.

Les taux de bonne classification sont présentés au tableau 7.1. Le tableau 7.2 rapporte lui les temps moyens de convergence et le nombre moyen de vecteurs de support. Ces résultats sont donnés pour l'ensemble des classes.

7.1.3 Commentaires

Les résultats montrent que l'algorithme sans biais (smgo) est plus rapide que l'algorithme avec biais (fast). Le gain est particulièrement marqué pour des ensembles d'apprentissage de grande taille ou pour des faibles valeurs de ν. Cette conclusion s'applique indépendamment d'une approche avec ou sans les contraintes du problème 2-classes. L'analyse des résultats sur les données réelles montre que les approches avec ou sans biais ont des performances comparables, sans concéder d'augmentation du nombre de vecteurs de support.

7.2 Evaluations sur des données de surveillance audio

Nous proposons maintenant de comparer les performances des deux algorithmes avec et sans biais dans le contexte de la surveillance audio. Pour cela, nous reprenons les données (après normalisation) présentées au chapitre 6 : *m07* ($\sigma = 5.68$) et *m12* ($\sigma = 5.60$). Dans un premier temps, nous reprenons les évaluations de la section précédente, à savoir comparer les temps d'apprentissage et le nombre de vecteurs de support. Dans un second temps, on étudie les performances obtenues à l'aide de l'algorithme sans biais pour différentes valeurs de ν, comparativement aux résultats du chapitre précédent, avec biais.

24. http ://archive.ics.uci.edu/ml/datasets.html.

TABLE 7.1 – Probabilités de bonne classification (en %)

	OC		OC2	
Données	fast	smgo	fast	smgo
Cancer	96,42	**96,85**	95,85	95,85
Crab	85,50	85,00	**96,50**	94,00
Glass	78,04	48,60	**94,86**	**94,86**
Iris	89,33	88,00	94,67	**95,33**
Wine	92,70	89,33	**97,19**	96,63

TABLE 7.2 – Temps de convergence moyen (en ms) et nombre moyen de vecteurs de support (entre parenthèses)

	OC		OC2	
Données	fast	smgo	fast	smgo
Cancer	0,7 (27)	0,7 (28)	1,8 (99)	1,3 (98)
Crab	2,6 (14)	2,5 (26)	8,9 (30)	3,3 (34)
Glass	2,4 (18)	2,4 (18)	2,5 (43)	2,5 (44)
Iris	6,7 (8)	6,7 (10)	11,3 (25)	8,9 (24)
Wine	5,6 (23)	5,6 (25)	5,6 (32)	5,6 (34)

Les figures 7.2 et 7.3 rapportent les temps d'apprentissages et le nombre de vecteurs de support obtenus pour les deux algorithmes lorsque le paramètre ν varie. Sur la première figure, il apparaît que le temps d'apprentissage est sensiblement équivalent pour les deux algorithmes. On note également que, comme dans le cas des données de synthèse étudiées précédemment, l'algorithme sans biais est légèrement moins performant pour des valeurs de ν élevées. En revanche, l'étude de la seconde figure montre que l'algorithme sans biais nécessite après convergence environ deux fois plus de vecteurs de supports, à valeur de ν identique. Dans ces conditions, nous étudions les performances des différents modèles. En particulier, se pose la question de savoir si les performances sont corrélées au nombre de vecteurs de support ou à la valeur de ν. En effet, dans le cas de l'algorithme sans biais, ν ne constitue plus une borne supérieure pour la fausse-alarme.

La lecture de la figure 7.3 montre qu'un nombre équivalent de vecteurs de support est obtenu pour d'une part l'algorithme avec biais avec $\nu = 0,05$ et d'autre part l'algorithme sans biais avec $\nu = 0,02$. Les figures 7.4 et 7.5 comparent les performances des détecteurs suivants, respectivement sur les signaux *m07* et *m12* :

- approche avec biais (fast) pour $\nu = 0,05$: détecteur de référence,
- approche sans biais (smgo) pour $\nu = 0,05$: détecteur à ν équivalent,
- approche sans biais (smgo) pour $\nu = 0,02$: détecteur à nombre de vecteurs de support équivalents.

Dans tous les cas, les résultats sont présentés après la prise en compte d'informations temporelles pour l'intégration des scores (voir l'approche au paragraphe 6.3.2).

Les performances des détecteurs évalués montrent que les résultats de l'algorithme sans biais pour $\nu = 0,02$ sont équivalents à ceux obtenus par l'algorithme avec biais pour $\nu = 0,05$. En conclusion, l'approche proposée permet, à nombre de vecteurs de support égal et performances équivalentes, de réduire le temps d'apprentissage.

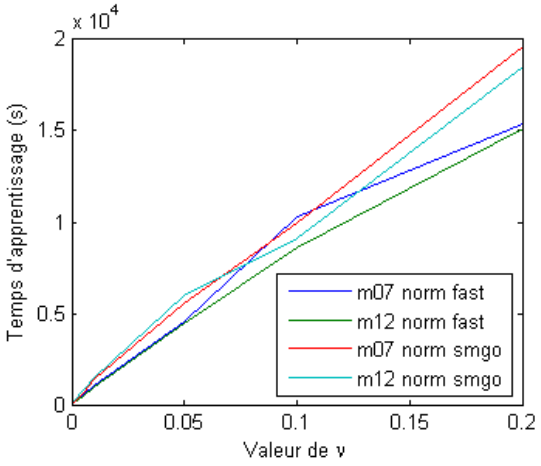


FIGURE 7.2 – Comparaison des temps d'apprentissage entre les algorithmes avec biais (fast) et sans biais (smgo) pour les signaux *m07* et *m12*.

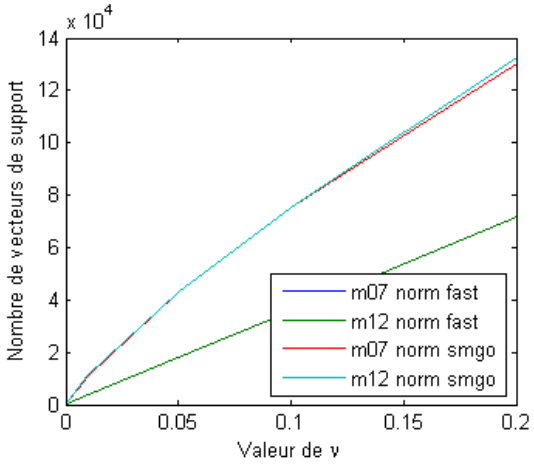


FIGURE 7.3 – Comparaison du nombre de vecteurs de support entre les solutions des algorithmes avec biais (fast) et sans biais (smgo) pour les signaux *m07* et *m12*.

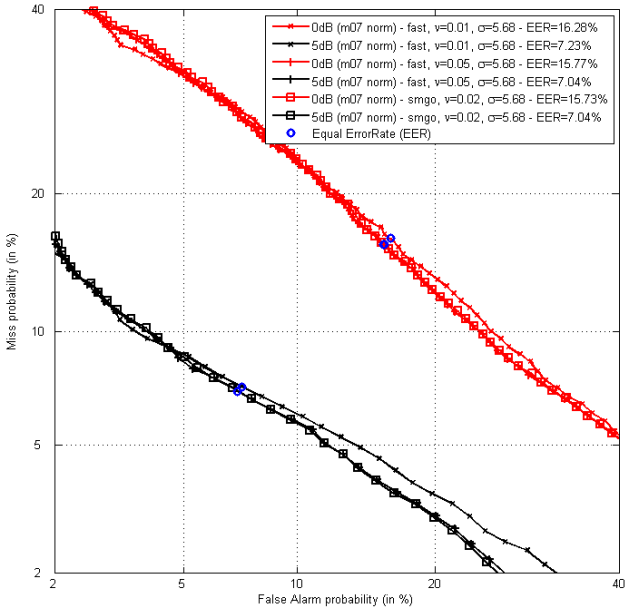


FIGURE 7.4 – Performances comparatives de détecteurs issus d'une approche avec ou sans biais pour les signaux *m07*.

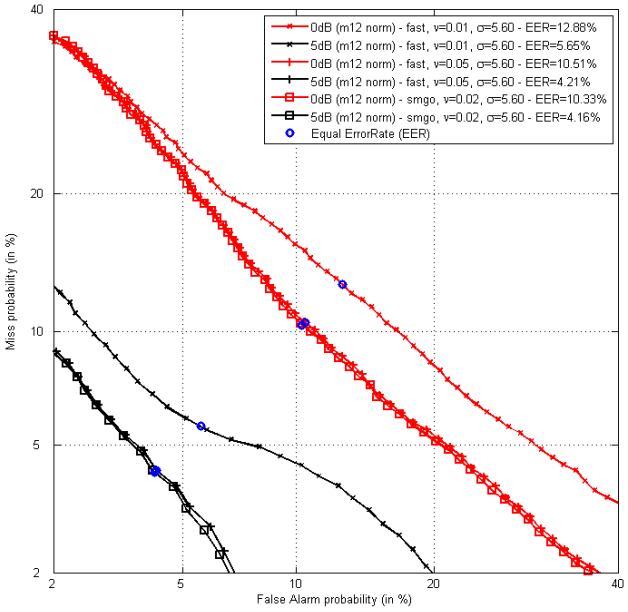


FIGURE 7.5 – Performances comparatives de détecteurs issus d'une approche avec ou sans biais pour les signaux *m12*.

7.3 Performances de l'approche sans biais

L'objet de ce chapitre était l'évaluation des gains en termes de performance (rapidité de convergence en particulier) grâce à l'approche sans biais.

Dans un premier temps, nous avons étudié, sur des données de synthèse (une classe à modéliser et une classe de rejet), la vitesse de convergence de l'algorithme en fonction de la taille de l'ensemble d'apprentissage et du choix de la valeur de l'hyper-paramètre ν. Ces premiers résultats nous ont conduits à penser que l'approche sans biais présente un intérêt non négligeable pour améliorer la vitesse de convergence d'une modélisation SVM 1-classe. Afin de s'assurer que les performances obtenues par cette approche sont comparables à celles obtenues par l'approche avec biais, nous avons étudié les performances de notre algorithme sur des données réelles simples. Là encore les temps d'apprentissage sont améliorés dans le cas de l'approche avec biais, et ce avec des performances de bonne classification équivalentes.

Dans un second temps, fort de ces premières conclusions, nous avons mené une évaluation des approches avec et sans biais sur un problème de détection d'événements anormaux. Les résultats obtenus témoignent de l'intérêt de l'approche sans biais dans le cadre des signaux audio étudiés. Cette conclusion, empirique, est valable pour les enregistrements sur lesquels nous avons travaillé. Il est important de noter qu'en l'absence d'une étude théorique de la convergence de l'algorithme sans biais proposé, une validation empirique sera nécessaire si d'autres signaux devaient être traités.

Quatrième partie

Clustering de type SVM 1-classe appliqué à la détection

Chapitre 8

Classification non supervisée par multiples SVM 1-classe

Les modélisations SVM 1-classe décrites aux chapitres précédents conduisent à l'apprentissage d'un unique volume minimal englobant les observations d'une classe à modéliser. Lorsque l'ensemble des données d'apprentissage est considéré comme appartenant à l'hypothèse H_0 « ambiance normale », une tâche de détection d'anormalité peut être conduite ; il s'agit alors de tester si les observations à évaluer se trouvent à l'intérieur ou à l'extérieur du volume appris. Dans le cas où la classe considérée occupe des régions disjointes de l'espace de représentation cette approche peut conduire à inclure dans l'enveloppe estimée des régions où la densité des observations est faible, voire nulle. Cette situation se présente notamment lorsque la classe considérée est en réalité une réunion de sous-classes qu'il n'est pas toujours possible d'observer indépendamment.

Nous présentons dans ce chapitre une approche de type *clustering* (classification non supervisée) par modélisation 1-classe. L'idée générale est, en s'appuyant sur les SVM 1-classe, d'apprendre un ensemble de volumes minimaux englobant les observations de la classe à modéliser. Chaque volume ainsi appris décrira localement la classe considérée, ou sous-classe. La réunion de ces volumes décrira quant à elle la classe H_0 dans son ensemble. La fonction de décision ne repose alors plus sur un unique modèle SVM 1-classe mais sur un ensemble de tels modèles.

Dans un premier temps, nous exposons le principe des méthodes de l'état de l'art permettant de réaliser une classification non supervisée à l'aide de SVM. S'appuyant sur ses éléments différenciants, nous introduisons ensuite notre approche multi-SVM 1-classe. Enfin, nous discutons des avantages attendus grâce à l'utilisation de cette approche et énonçons des perspectives de travail.

8.1 Clustering et SVM

Une part importante de l'effort de recherche est consacrée aux problèmes de classification non supervisée, appelée *clustering* [JD88, Jai10]. On rappelle la taxonomie classique des problèmes de classification :

- apprentissage supervisé : étant donné un ensemble de données d'apprentissage, chacune associée à une étiquette, on cherche à donner une définition des zones de l'espace (*clusters*) associées à ces étiquettes afin de pouvoir étiqueter de nouvelles observations.
- apprentissage semi-supervisé : dans ce cas, seule une partie des étiquettes est disponible. L'objectif est d'une part de modéliser les *clusters* étant donnés les étiquettes disponibles,

mais également d'estimer les étiquettes des autres observations afin de définir plus précisément les modèles.
- apprentissage contraint (avec contraintes par paires, *pairwise*, par exemple) : les informations sont des liens entre une partie observations : *must-link* (les observations appartiennent au même cluster) ou *cannot-link* (les observations appartiennent à des *clusters* différents). Le problème d'optimisation doit alors estimer les étiquettes en plus de la description des *clusters*, tout en s'appuyant sur les contraintes.
- apprentissage non supervisé : dans cette situation, seules les observations sont disponibles. Il s'agit alors conjointement d'optimiser la modélisation des *clusters* et l'attribution des étiquettes.

Dans la littérature, les deux premières approches sont usuellement référencées comme *classification* et les deux dernières comme *clustering*.

Intéressons-nous maintenant à la qualification des problèmes sous-jacents. Le premier problème, la modélisation des *clusters* s'exprime souvent comme un problème quadratique. C'est le cas notamment lorsque ce problème est traité à l'aide d'une approche SVM. En revanche, l'optimisation des étiquettes est un problème combinatoire non convexe difficile.

Pour résoudre le problème d'optimisation conjointe de la description des *clusters* et de l'attribution des étiquettes, deux familles d'algorithmes ont émergé, chacune basée sur l'un des points clé des SVM. La première famille, Support Vector Clustering (SVC) [BHHSV01], s'appuie sur les vecteurs de support pour déterminer les *clusters*, la seconde famille, Maximum Margin Clustering (MMC) [XNLS04], s'intéresse à optimiser le problème de marge maximum en intégrant l'optimisation des labels.

8.1.1 Méthodes SVC

Introduites par Ben-Hur *et al* [BHHSV01], les méthodes SVC séparent l'optimisation en deux problèmes distincts :
- La description des *clusters*. Ce problème est toujours résolu par une modélisation SVM 1-classe des observations [SPST$^+$01] ou son équivalent *Support Vector Data/Domain Description* [TD99, TD04]. Les lignes de contours obtenues à l'issue de l'une de ces modélisations servent à initialiser les *clusters*, les vecteurs de support sur la marge permettant d'identifier les frontières. Les contrôles du nombre de *clusters* et de leur qualité sont réalisés à l'aide des paramètres de la modélisation : rejet (fraction d'*outliers*) et noyau (largeur dans le cas gaussien).
- L'attribution des étiquettes (*cluster labeling*). Pour une modélisation donnée, les observations sont analysées, généralement par des méthodes graphiques, pour attribuer les étiquettes. Parmi ces méthodes on trouve la *Graphical Connected-component Method* (GCM) [BHHSV01] et la *Modified* GCM [NS05] qui procèdent toutes deux en s'assurant qu'aucun élément du chemin entre deux points d'un même *cluster* ne sorte de ce *cluster* ; dans le cas contraire, les points sont attribués à des *clusters* différents. Ces approches nécessitent d'échantillonner le chemin entre chaque paire de points de l'ensemble d'apprentissage. [LL05] propose de regrouper les observations autour de points d'équilibres avant d'appliquer la GCM, réduisant le coût calculatoire. Enfin, [Lee06] propose la méthode du *Cone Cluster Labeling* (CCL) (utilisation d'hyper-cônes dans l'espace image pour regrouper les observations entre vecteurs de support) et [YECC02] une approche basée sur les graphes de proximité.

Ces approches sont naturellement multi-classes car elles s'appuient sur les lignes de contour de modélisation 1-classe et non sur une partition binaire de l'espace.

8.1.2 Méthodes MMC

Initialement introduite en 2004 dans [XNLS04], l'approche *Maximum Margin Clustering* propose d'étendre directement la théorie SVM à un usage non supervisé. [XNLS04] avance en effet qu'il manque aux méthodes de *clustering* existantes, même les plus récentes (SVC, spectral clustering [Lux07], *etc.*), une connexion avec les autres types de problèmes d'apprentissage (semi-supervisé, supervisé, contraint). Pour retrouver ce lien, [XNLS04] introduit un cadre général d'optimisation simultanée de l'hyperplan discriminant (problème quadratique convexe classique) et des étiquettes (problème combinatoire non convexe, difficile) et montre que le problème global peut être reformulé comme un *Convex Integer Program*, puis moyennant la relaxation de la contrainte entière, comme un *Semi-Definite Program* pour lequel il existe des méthodes de résolution.

S'appuyant sur la formulation SVM classique, cette approche est naturellement 2-classes. Bien que [XS05] étende le problème 2-classes au cas multi-classes en s'inspirant des travaux de [CS01], la formulation MMC exclut le biais (donc les hyperplans séparateurs passent par l'origine, ce qui nécessite un pré-traitement coûteux des données) et le SDP reste un outil lourd et coûteux en temps ($O(n^2)$, les données traitées ne contiennent que quelques centaines d'observations au mieux). En 2007, [VJ07] généralise les MMC et réduit la complexité ($O(n)$, prise en compte du biais et optimisation simultanée du noyau). Dans cette nouvelle formulation, [VJ07] montre une équivalence avec le *spectral clustering* par *normalized cut*.

Zhao propose d'utiliser une approche *Cutting Plan Algorithm* pour résoudre le problème MMC (cas 2-classes [ZWZ08b, WZZ10], multi-classes [ZWZ08c, WZZ10] ou encore semi-supervisé [ZWZ08a]). Dans ce cas le SDP est remplacé par une *Constrained Concave-Convex Procedure* (CCCP), plus efficace (convergence garantie et complexité réduite). [HWYH08] étend l'approche *pairwise constraints* au cas multi-classes. [WWL09] propose également d'inclure un a priori géométrique (*data manifolds*) à cette méthode pour lisser la solution.

Zhang [ZTK07, ZTK09] propose une résolution par optimisation alternée basée sur une résolution itérative par SVM, *Least-Square* SVM ou plus efficacement SVR (cependant sans garantie sur la convergence). [GPK09] étudie une approche évolutionnaire grâce à une astuce calculatoire permettant d'évaluer la qualité d'une solution intermédiaire (évolution des labels) ; il obtient les résultats les plus performants de l'état de l'art (taux d'erreur le plus faible). [LTKZ09] propose une relaxation du problème MMC différente, plus précise que celle de [XS05] (pas de minima locaux, résolution par *Cutting Plane Algorithm*, lien avec *Multiple Kernel Learning*). Le cas multi-classes est toujours une extension du cas 2-classes basée sur la méthode de [CS01]. Plus récemment, [ZC11] a proposé une approche par apprentissage contraint en alternant descente de gradient et projection, il prétend également améliorer la fonction coût pour les liens *cannot-link* par rapport à [HWYH08].

8.2 Modélisation concurrentielle par modèles SVM 1-classe

L'approche SVC implique l'apprentissage d'un unique modèle SVM 1-classe. Ainsi, la description des *clusters* ne peut pas être localement adaptée (choix des hyper-paramètres) en fonction des données. Si l'approche MMC permet cet ajustement, elle réalise néanmoins un apprentissage conjoint des étiquettes et des modèles. L'optimisation simultanée d'un problème quadratique et d'un problème combinatoire est une tâche complexe. Les algorithmes utilisés sont donc particulièrement lents. S'inspirant d'approches par mixture d'experts, nous proposons alors une approche mettant en compétition plusieurs modèles SVM 1-classe afin de modéliser un l'ensemble d'apprentissage. L'idée principale est de considérer la classe à apprendre comme une réunion de

« sous-classes », chacune pouvant être caractérisée par un SVM 1-classe (local) différent.

8.2.1 Principe

L'optimisation des modèles consiste en la résolution d'un ensemble de problèmes (quadratiques) SVM 1-classe, un pour chaque sous-classe. Cette étape peut être réalisée indépendamment pour chacune des sous-classes ; on résout alors un problème SVM 1-classe pour chaque sous-classe en réduisant l'ensemble d'apprentissage aux seules données appartenant à cette sous-classe. Alternativement, les modèles peuvent être entraînés suivant une stratégie 1-contre-tous, telle que Tohmé l'a proposé dans le cadre de la classification en s'appuyant sur le problème OC2-SVM (3.62) [TL11]. Cette stratégie, que nous proposons d'étendre également au problème WOOC2-SVM (3.83), permet de minimiser le recouvrement entre les sous-classes.

L'optimisation des étiquettes est un problème combinatoire. Au cours de ce processus, il est également possible de rejeter des données, supposées aberrantes ; celle-ci ne se voient alors attribuer aucune sous-classe. Lorsque les modèles SVM 1-classe ne sont pas disponibles, cette étape est réalisée à l'aide d'un algorithme de *clustering* du type k-moyennes. Dans le cas contraire, on utilise les fonctions de décision locales, associées aux modèles SVM 1-classe.

Ces deux optimisations peuvent être répétées, réalisant alors un processus d'optimisation itératif par directions alternées. Ce processus prend fin à l'atteinte d'un critère d'arrêt préalablement fixé : nombre limite d'itérations, critère de qualité des modèles, critère d'information sur les sous-classes, etc.

Enfin, on construit une règle de décision fusionnant les résultats des fonctions de décision locales. Celle-ci permet de détecter si une nouvelle donnée est incluse dans au moins l'un des volumes minimaux déterminés - et on prend la décision qu'il s'agit d'une observation normale - ou si elle est en dehors de ces enveloppes - il s'agit alors d'une observation anormale.

Dans les paragraphes suivants, nous allons apporter un cadre formel à ces quatre étapes : apprentissage des modèles, *clustering* des observations, critère d'arrêt et fusion de décision.

8.2.2 Modélisation par multiples SVM 1-classe

Cas sans contraintes binaires

Soit Q le nombre de sous-classes à l'aide desquelles nous souhaitons modéliser un ensemble d'apprentissage S de taille N. Soit $\mathbf{l} \in \{1, \ldots, Q\}^N$ un vecteur d'étiquettes l_i attribuant une classe à chacune des observations. L'approche que nous proposons consiste alors à résoudre Q problèmes OC-SVM ou WOOC-SVM. Chacun de ces problèmes est associé à :

- un ensemble d'apprentissage $S_q = \{\mathbf{x}_i \in S, l_i = q\}$ de taille N_q,
- une matrice de Gram $\mathbf{H}_q$ composée des éléments $-\kappa(\mathbf{x}_i, \mathbf{x}_j), (\mathbf{x}_i, \mathbf{x}_j) \in (S_q)^2$, et de taille $N_q \times N_q$,
- un vecteur de multiplicateurs de Lagrange α_q, de taille N_q,
- dans le cas de problèmes OC-SVM, un biais b_q.

pour $q = 1, \ldots, Q$.

La résolution de ces Q problèmes conduit à l'obtention de Q vecteurs solutions $\boldsymbol{\alpha}_q$, soit aux Q fonctions de décision définies par :

$$\begin{array}{rccl} f_q : & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \rightarrow & \displaystyle\sum_{i=1}^{N_q} \alpha^*_{q,i} \kappa\left(\mathbf{x}_i, \mathbf{x}\right) + b_q \quad \mathbf{x}_i \in S_q \end{array} \tag{8.1}$$

où α_q^* est la solution du problème q et b_q le biais déterminé selon le problème tel que défini au chapitre 4.

Cas avec contraintes binaires

Il s'agit ici de la même approche que précédemment, en considérant cette fois la résolution de problèmes OC2-SVM ou WOOC2-SVM. Dans le premier cas, cela revient à exploiter la modélisation proposée par Tohmé [TL11]. Chacun de ces problèmes est associé à :

- l'ensemble d'apprentissage S de taille N,
- un vecteur de décisions $\mathbf{y}_q$ tel que :

$$y_{q,i} = \begin{cases} +1 & \text{si } l_i = q \\ -1 & \text{sinon} \end{cases} \qquad i = 1, \ldots, N \tag{8.2}$$

- une matrice de Gram polarisée $\mathbf{H}_q$, composée des éléments $-y_{q,i}y_{q,j}\kappa(\mathbf{x}_i, \mathbf{x}_j), (\mathbf{x}_i, \mathbf{x}_j) \in (S_q)^2$, et de taille $N \times N$,
- un vecteur de multiplicateurs de Lagrange $\boldsymbol{\alpha}_q$ de taille N,
- dans le cas de problèmes OC2-SVM, un biais b_q.

pour $q = 1, \ldots, Q$.

Ainsi, lors de l'apprentissage d'une sous-classe, on recherche un volume minimal autour des données associées à cette sous-classe (décision $+1$), tout en rejetant les observations associées aux autres sous-classes (décision -1). Ceci permet d'optimiser localement la représentation en minimisant le recouvrement entre sous-classes.

Là encore, la résolution de ces Q problèmes conduit à l'obtention de Q vecteurs solutions $\boldsymbol{\alpha}_q$, soit aux Q fonctions de décision définies par :

$$\begin{array}{rccl} f_q : & \mathcal{X} & \mapsto & \mathbb{R} \\ & \mathbf{x} & \rightarrow & \displaystyle\sum_{i=1}^{N} \alpha_{q,i}^* \kappa\left(\mathbf{x}_i, \mathbf{x}\right) + b_q \quad \mathbf{x}_i \in S \end{array} \tag{8.3}$$

où α_q^* est la solution au problème q et b_q le biais déterminé selon le problème tel que défini au chapitre 4.

8.2.3 Attribution des étiquettes

Initialisation

Le regroupement initial par sous-classes des observations d'apprentissage s'effectue en l'absence de modèles SVM 1-classe. Il est donc nécessaire de faire appel à un algorithme tiers pour initialiser notre méthode. Nous présentons ici l'approche k-moyennes décrite dans [CR10]. Celle-ci est privilégiée pour ses résultats opérationnels satisfaisants sur l'application audio qui nous intéresse (résultats Thales).

L'élément différenciant de cette approche réside dans le positionnement initial des centroïdes, déterministe et uniformément distribué dans l'espace des observations. Les centroïdes initiaux sont sélectionnés parmi les observations de l'ensemble d'apprentissage de la manière suivante :

1. Calcul du vecteur moyen à partir de l'ensemble des données d'apprentissage,
2. Sélection de l'observation la plus éloignée du vecteur moyen comme premier centroïde ($Q = 1$),

3. Itérations permettant d'incrémenter le nombre de classes ($Q = Q + 1$) : on recherche le point qui maximise la distance cumulée aux centroïdes identifiés à l'itération précédente,
4. Critère d'arrêt : soit lorsque le nombre prédéfini de classes Q est atteint, soit lorsque la distance entre le point trouvé et les centroïdes identifiés à l'itération précédente est inférieure à un seuil.

Un algorithme estimation-maximisation [DLR77] est ensuite utilisé pour déterminer les centroïdes optimaux. Chaque centroïde, ainsi que les observations qui lui sont associées, se voit attribué une étiquette (sous-classe) différente.

Classifier les données d'apprentissage à partir des modèles obtenus

Lorsqu'une modélisation a été réalisée, nous disposons des Q fonctions de décision f_q, $q = 1, \ldots, Q$. Nous utilisons alors la règle de décision suivante pour déterminer le vecteur $\mathbf{l}$:

$$l_i = \arg\max_q f_q(\mathbf{x}_i) \tag{8.4}$$

Cette règle simple permet d'attribuer une étiquette à chacune des N observations de S, y compris à celles qui ne sont incluses dans aucune des Q enveloppes apprises.

Cas avec rejet

Il est également possible d'exclure certaines observations en les associant à une classe 0 par la règle suivante :

$$l_i := 0 \text{ si } f_q(\mathbf{x}_i) < 0, \forall q = 1, \ldots, Q \tag{8.5}$$

Dans ce cas, les données aberrantes (*outliers*) seront exclues des observations à modéliser. Dans le cas sans contraintes binaires, elles seront simplement ignorées ; dans le cas avec contraintes binaires, elle seront rejetées par chacun des modèles réalisés, la décision leur étant associée vaut -1 pour chaque problème.

Le bénéfice attendu est la réduction du nombre de vecteurs de support, ces données aberrantes étant au-delà de la marge et associées à des multiplicateurs de Lagrange saturés. La valeur du paramètre ν peut alors être ajustée en fonction du nombre d'observations rejetées suivant : $\nu = \nu_0 \frac{N - n_r}{N}$ où n_r est le nombre d'observations rejetées et ν_0 la valeur initiale. Cet ajustement permet de conserver constant le nombre maximum de vecteurs de supports au-delà de la marge, fixé par définition à $\nu_0 N$.

Notons que dans la pratique, on utilisera la règle suivante :

$$l_i := 0 \text{ si } f_q(\mathbf{x}_i) < \varepsilon, \forall q = 1, \ldots, Q \tag{8.6}$$

avec $\varepsilon < 0$ un seuil fixé. Cette règle limite le rejet d'observations proches de la marge, bien que en dehors de l'enveloppe apprise. En particulier, lorsque ν est fixé à une valeur élevée, ce seuil limite le rejet d'un trop grand nombre d'observations. Notons enfin que, disposant des scores obtenus par l'ensemble des observations de l'ensemble d'apprentissage, ce seuil peut être déterminé automatiquement pour rejeter une fraction prédéterminée des observations au-delà de la marge.

8.2.4 Procédure itérative et critère d'arrêt

Les étapes d'apprentissage des modèles SVM 1-classe et d'optimisation des étiquettes sont successivement répétées. On parle de processus d'optimisation par méthode des « directions alternées » car deux optimisations successives sont réalisées suivant d'une part les vecteurs solution α_q, et d'autre part le vecteur d'étiquettes $\mathbf{l}$. Ainsi, notre approche contourne la difficulté d'une optimisation conjointe, un problème difficile du fait que ces deux optimisations ne soient pas du même type (problème quadratique pour les uns, combinatoire pour l'autre).

Le processus d'optimisation prend fin lorsqu'un critère d'arrêt prédéfini est rencontré. Différents critères d'arrêt peuvent être mis en œuvre, par exemple :

- un nombre maximum d'itérations est atteint,
- la fraction des données qui changent d'étiquette est inférieure à un seuil prédéfini,
- un critère d'information sur les données et la modélisation de chaque groupe atteint un seuil prédéfini (exemple : critère BIC, critère de Fisher, etc.).

Procédure 1 Clustering par multiples SVM 1-classe

Entrées: $\mathcal{X} = \{\mathbf{x}_i : i = 1, \ldots, l\}$, un ensemble de données non étiquetées, et Q, le nombre de classes attendues.

Sorties: $\mathbf{l} \in \{1, \ldots, Q\}^N$ un vecteur d'étiquettes et $\{(\boldsymbol{\alpha}_q, b_q), q = 1, \ldots, Q\}$ un ensemble de modèles SVM 1-classe.

1: Initialiser le vecteur $\mathbf{l}$ à partir d'une méthode tierce.
2: *Description :* résoudre les Q problèmes SVM 1-classe, étant donné le vecteur $\mathbf{l}$.
3: *Attribution des étiquettes :* mettre à jour le vecteur $\mathbf{l}$ à partir des modèles obtenus.
4: Si aucun critère d'arrêt n'est atteint, aller à l'étape 2. Sinon, terminer.

8.2.5 Fonction de décision pour la détection

A l'issue de la procédure itérative, nous disposons d'un ensemble de Q fonctions de décision f_q. Afin de pouvoir détecter si une nouvelle observation est issue de la classe (ensemble de sous-classes) normale ou non, une règle de décision doit être construite afin de fusionner les scores obtenus pour cette observation à travers ces Q fonctions.

La règle de fusion la plus simple consiste à considérer une nouvelle observation comme normale si elle se situe à l'intérieur d'au moins une des enveloppes définies par ces fonctions de décision. Cette règle s'exprime :

$$y = \begin{cases} +1 & \text{si } \max_q f_q(\mathbf{x}) > 0 \\ -1 & \text{sinon} \end{cases} \tag{8.7}$$

En pratique, cette règle peut utiliser la condition $\max_q f_q(\mathbf{x}) > \lambda$ où $\lambda \in \mathbb{R}$ est un seuil à déterminer suivant les besoins opérationnels (voir 6.2).

8.3 Discussion

Dans les paragraphes qui suivent, nous évoquons quelques perspectives de mise en œuvre d'un détecteur s'appuyant sur une modélisation par multiples SVM 1-classe telle que définie précédemment. Ces variantes ont notamment été décrites dans le brevet proposé au cours des travaux de thèse concernant cette approche [LCR+12]. Nous discutons également du problème de convergence de cette modélisation qui reste ouvert.

8.3.1 Mise à jour des modèles et apprentissage en ligne

Par la nature de la formulation du problème SVM 1-classe présentée précédemment, l'approche proposée ne nécessite pas de réaliser un apprentissage complet de chaque SVM 1-classe à chaque itération du processus de modélisation concurrentielle (apprentissage). En effet, il est possible de mettre à jour une solution obtenue afin de tenir compte de la modification de la classe associée à une (ou plusieurs) donnée(s) de l'ensemble d'apprentissage (modification du vecteur d'étiquettes $\mathbf{y}$). De la même manière, il peut également être envisagé l'ajout et/ou la suppression de données dans l'ensemble d'apprentissage. Ceci permet de prendre en compte de nouvelles observations pour raffiner un modèle préalablement appris par exemple. En effet, en exploitant les capacités de mise à jour de solution de l'algorithme de résolution SVM 1-classe proposé, on minimise le coût calculatoire

Notons que les observations peuvent être traitées en ligne, c'est-à-dire au fil de l'eau, ou bien hors ligne, c'est-à-dire en une seule fois. Généralement, l'ensemble d'apprentissage est traité pour tout ou partie hors ligne tandis que les nouvelles observations font l'objet d'un traitement en ligne, l'objet du système étant de détecter une éventuelle anormalité avec un minimum de délai. Lorsque peu de données sont modifiées, ajoutées ou supprimées, la mise à jour peut être suffisamment rapide pour envisager un apprentissage en ligne. Le modèle créé lors de l'étape d'apprentissage est alors mis à jour en continu à partir de nouvelles observations, une fois celles-ci classées comme normales ou anormales. Des heuristiques peuvent être utilisées afin de déterminer quelles données doivent être intégrées à ou ôtées de l'ensemble d'apprentissage afin de limiter son occupation mémoire.

8.3.2 Evolution du nombre de classes et structuration des signaux

Dans la description de l'approche proposée, le nombre de classes de la modélisation est déterminé *a priori*. Il s'agit par exemple du paramètre k si l'initialisation se fait à l'aide d'un algorithme k-moyennes. Néanmoins il est possible de faire évoluer cette valeur au cours de l'étape d'apprentissage ou au cours d'une mise à jour des modèles telle que décrite au paragraphe ci-dessus.

Pour cela, on envisage la définition d'heuristiques permettant de fusionner des classes sur des critères de dispersion inter et intra classes des données. De même, on peut scinder en deux une classe. On agit ainsi au cours de l'étape de modification des étiquettes. Enfin, il est possible de créer une classe pour un groupe de données rejetées, scinder une classe dont le nombre d'observations est trop grand ou de supprimer une classe dont le nombre d'observations est trop peu significatif (alors ces données peuvent être associées à des groupes existants ou à une classe anormale).

On rappelle également que dans l'application de détection :

- les données de l'ensemble d'apprentissage appartiennent toutes à la classe normale et l'objectif est de détecter les anormalités,
- les sous-classes de normalité sont déterminées de façon non supervisée.

Par conséquent, une donnée (ou plusieurs données) peut (ou peuvent) changer de classe sans conséquence et le nombre de classes n'est pas contraint. On affine ainsi la description. D'une part, les données correctement modélisées continuent d'appartenir à des classes de la normalité alors que les données aberrantes sont identifiées et isolées. D'autre part les modèles sont appris sur des données de moins en moins polluées (les données aberrantes étant rejetés) et les modèles sont de plus en plus précis.

Notons que cette approche peut permettre de faire émerger une structure sous-jacente à

l'ensemble d'apprentissage. En effet, une classe peut contenir des événements « identiques » (par exemple le claquement d'une porte métallique), mais aussi des regroupements d'événements (les claquements de portes en général) ou encore une quantité illimité d'événements (l'ensemble des claquements). Dans ce dernier cas, notre approche est particulièrement intéressante. Lorsque nous sommes confrontés à de vastes classes d'événement hétérogènes (l'ensemble des événements normaux sur un quai de métro par exemple), il est souvent utile d'utiliser plus d'un modèle.

8.3.3 La convergence, un problème ouvert

Il est enfin important de noter que la convergence de l'approche proposée n'a pas été étudiée. Il s'agit d'un défi à relever, tant du point de vue de l'étude de la modélisation SVM 1-classe que du point de vue de la construction des heuristiques visant à établir le vecteur d'étiquettes $\mathbf{l}$.

Chapitre 9

Application de la méthode pour la détection

Dans ce chapitre, on s'intéresse à deux types de résultats de la modélisation par multiples SVM 1-classe : la découverte des classes par le modèle et les performances en détection comparées à une approche s'appuyant sur un unique SVM 1-classe.

Un premier problème basé sur la reconnaissance de chiffres manuscrits est étudié. Ce choix est motivé par un dimensionnement des évaluations adapté au temps dont nous disposons et par la possibilité d'analyser le lien entre les classes sous-jacentes (les chiffres) et les classes découvertes.

Dans un second temps, un échantillon des données utilisées dans les précédents chapitres pour la détection d'événements sonores anormaux est confronté à cette modélisation. L'échantillonnage de la base de données a été réalisé afin de réduire la quantité d'observations et mener les expériences dans le temps imparti pour la thèse.

9.1 Résultats sur une base de données structurée

9.1.1 Mise en œuvre expérimentale

Données et problème de détection

Nous utilisons dans cette section la base de données *ZIP Code* [HTF09] [25]. Il s'agit d'un ensemble de chiffres manuscrits, scannés par le service postal américain [26]. Les images originales ont été normalisées en taille (16×16) et en orientation, et les valeurs des pixels correspondent à 256 niveaux de gris. La base de données est étiquetée par chiffre (0-9) et décomposée en deux ensembles, 7291 échantillons d'entrainement et 2007 de test. L'ensemble de test est identifié comme « notoirement » difficile par ses auteurs. La figure 9.1 donne quelque exemples d'observations issus de cette base.

Initialement, cette base de données a été utilisée pour l'évaluation d'algorithmes de classification. Dans notre cas, nous souhaitons exploiter celle-ci afin de construire un problème de détection. L'intérêt de travailler avec cette base est la structure sous-jacente par classes de chiffres distincts afin de qualifier le comportement de l'algorithme de *clustering* et étudier la découverte automatique de ces classes. Ainsi, nous construisons un problème de détection de nouveauté (ou d'*outliers*) en réalisant un apprentissage après avoir exclu l'une des classes, le chiffre 0. Lors du test, les observations de cette classe « exclue » devront donc être considérées comme nouvelles,

25. Disponible en ligne sur http ://www-stat.stanford.edu/ tibs/ElemStatLearn/
26. A ce titre *ZIP Code* est parfois nommé *USPS* dans la littérature, il s'agit de la même base de données

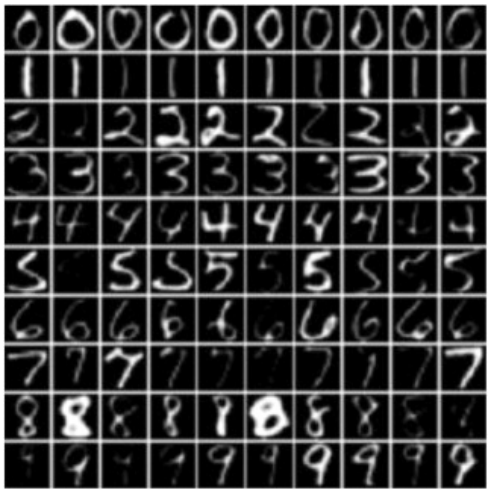


FIGURE 9.1 – Exemple de chiffres manuscrits issus de la base *ZIP Code*.

ou anormales, tandis que les observations des classes 1-9 devront être reconnues par le modèle appris. Précisons enfin que pour limiter la complexité de calcul les données sont projetées dans un espace de dimension 5 par analyse en composantes principales.

Construction du détecteur

Nous proposons d'évaluer un algorithme construit tel que présenté dans le chapitre précédent : un ensemble de modèles SVM 1-classe modélisent des *clusters* déterminés de façon non supervisée. Cependant, cette approche faisant appel à nombre relativement conséquent de paramètres, nous avons contraint son fonctionnement. En particulier, la convergence de l'algorithme n'ayant pas été étudiée, nous ne mettons pas en œuvre la procédure itérative ; ainsi, le détecteur est construit à partir de modèles SVM 1-classe appris depuis l'initialisation des étiquettes ; ladite initialisation étant réalisée à l'aide d'un algorithme k-moyennes. On étudie en partie l'influence du paramètre de noyau σ, mais ni celle de la normalisation des données, ni celle du choix du paramètre ν dont la valeur sera fixée *a priori* et identique pour tous les SVM 1-classe. Enfin, nous comparons les résultats utilisant les problèmes avec biais (*fast*) ou sans biais (*smgo*), avec ou sans l'utilisation des contraintes du problème 2-classes (respectivement OC2 et OC).

L'algorithme k-moyennes nécessite par ailleurs une initialisation des centroïdes. Celle-ci conditionne généralement la qualité des centroïdes obtenus après convergence. Nous étudions trois stratégies d'initialisation différentes. La première consiste en un tirage aléatoire de k observations au sein de l'ensemble d'apprentissage. Par cette technique, nous pourrons par la suite situer la qualité d'autres méthodes d'initialisation relativement à la distribution des performances obtenues pour un nombre important d'initialisations aléatoires. La seconde technique, déterministe, consiste à choisir comme dictionnaire initial l'ensemble des k observations les plus proches de la moyenne calculée sur l'ensemble d'apprentissage. Enfin, nous considérons une initialisation incrémentale qui sélectionne itérativement les points les plus distants du dictionnaire. Cette approche est également déterministe et il s'agit alors de sélectionner les points qui maximisent leur distance minimale aux centroïdes, le premier centroïde étant le point le plus éloigné de la moyenne. Les figures suivantes font référence à la seconde approche par « voisinage moyenne », et à la troisième par « MaxMin dictionnaire ».

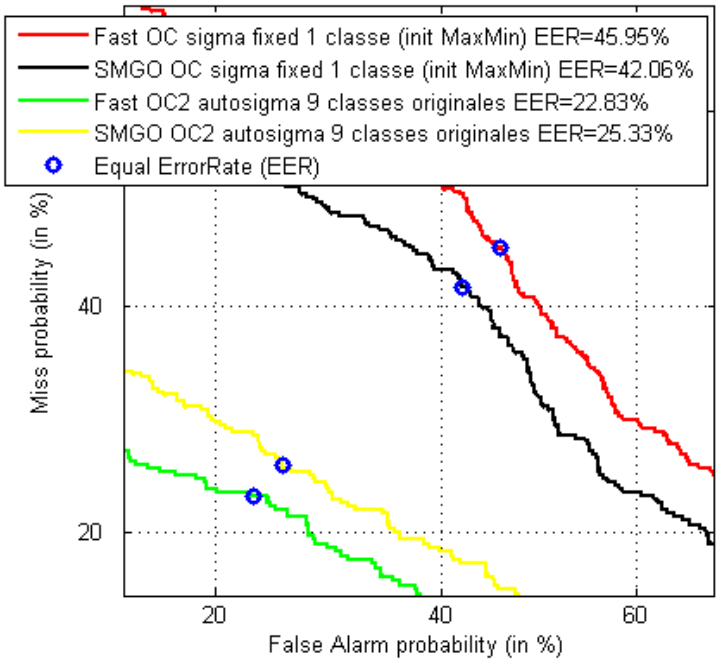


FIGURE 9.2 – Détecteurs de référence pour l'évaluation de l'approche *clustering*.

9.1.2 Résultats

Choix des paramètres et résultats de référence

L'approche proposée s'appuyant sur un ensemble de modèles SVM 1-classe, il convient de choisir les deux paramètres ν et σ ; le dernier étant la largeur du noyau Gaussien RBF retenu pour nos expériences. S'appuyant sur les résultats présentés au précédent chapitre 6 (voir section 6.1.2), nous fixons arbitrairement les valeurs des paramètres $\nu = 0,05$ et $\sigma = \sqrt{\gamma}$. Notons que γ dépend des données d'apprentissage de la classe positive (la classe normale contient les chiffres 1 à 9 dans notre cas). Il est alors possible dans notre modèle de réévaluer le paramètre de noyau localement pour chaque SVM 1-classe. On fait référence à cette approche par l'appellation « auto-sigma » dans la suite. Dans le cas contraire, la même valeur de σ est conservée pour l'ensemble des modèles ; il s'agit de celle estimée pour l'ensemble d'apprentissage complet.

Compte-tenu de ces éléments, nous présentons maintenant les résultats de références. Ceux-ci sont issus d'une part, d'une modélisation 1-classe à l'aide d'un unique SVM, à l'image des résultats du chapitre précédent, et d'autre part, d'une modélisation « experte » s'appuyant sur les étiquettes connues des données d'apprentissage pour apprendre 9 SVM 1-classe. Cette dernière approche, supervisée, constitue un objectif en termes de performances. La figure 9.2 illustre, au travers d'un réseau de courbes DET, les performances des meilleurs détecteurs de référence obtenus pour chaque algorithme (avec ou sans biais). [27]

27. Notons que dans le cas 1-classe, ni la prise en compte des contraintes binaires, ni l'adaptation automatique de σ ne change le détecteur puisqu'il n'y a qu'un seul SVM modélisant l'ensemble des données d'apprentissage.

Résultats avec la méthode par classification non supervisée

Chaque expérience a été réalisée 100 fois avec autant d'initialisations différentes. La première initialisation considère l'approche « MaxMin dictionnaire », la seconde l'approche « voisinage moyenne », et les autres initialisations sont des tirages aléatoires. Sur chacune des figures présentées ci-après, on trace les taux d'erreurs aux points de fonctionnement EER des détecteurs obtenus avec les deux premières initialisations. On trace également les taux d'erreur EER minimum, maximum et moyens pour l'ensemble des initialisations. Ces courbes sont fonctions du nombre de classes, variant de 1 à 25. Notons que les résultats des initialisations aléatoires permettent simplement de positionner les détecteurs issus des initialisations déterministes.

Nous étudions les performances obtenues pour différentes conditions d'entrainement. Ces dernières sont rapportées au tableau 9.1, ainsi que le numéro de figure correspondant. Nous évaluons ainsi l'apport de la sélection automatique du paramètre σ, de la prise en compte des contraintes du problème 2-classes ou encore du choix de l'algorithme avec biais (fast) ou sans biais (smgo). Le tableau 9.1 rapporte par ailleurs le temps total nécessaire à la réalisation de chacune de ces expériences, soit les 100 entrainements et évaluations pour 1 à 25 classes.

Biais	Contraintes du problème 2-classes	Choix du paramètre σ	Temps (s) de réalisation	Figure
avec	Non	Fixe	2185	9.3
avec	Non	Auto	2847	9.4
avec	Oui	Fixe	34889	9.5
avec	Oui	Auto	43281	9.6
sans	Non	Fixe	1718	9.7
sans	Non	Auto	2363	9.8
sans	Oui	Fixe	18122	9.9
sans	Oui	Auto	20749	9.10

TABLE 9.1 – Conditions d'entrainement des différents détecteurs construits par clustering.

Discussion

L'approche proposée donne des résultats plus performants que par l'utilisation d'un seul SVM 1-classe. Les gains sont également plus marqués lorsque l'on procède à l'adaptation locale du paramètre σ, ou lorsque l'on prend en compte les contraintes du problème 2-classes. Ce constat confirme l'intérêt de construire un détecteur sur la base de modèles localement optimisés dans l'espace d'observation.

D'autre part, ces expériences nous permettent de justifier le choix des deux initialisations déterministes proposées. Lorsque le nombre de classes est faible (particulièrement jusqu'à 6 *clusters*), l'approche au voisinage de la moyenne permet de construire des détecteurs aux performances très proches des meilleures performances obtenues. En revanche, lorsque le nombre de classes augmente, l'approche « MaxMin » permet d'obtenir de meilleures performances que la première approche, sans toutefois être aussi proche des meilleures performances obtenues. Nous expliquons ce comportement par le fait qu'augmenter le nombre de classes nécessite d'identifier des *clusters* isolés et/ou de faible densité dans l'espace d'observation, ce que l'approche « voisinage moyenne » ne permet pas tandis que l'autre approche identifie ces zones en priorité.

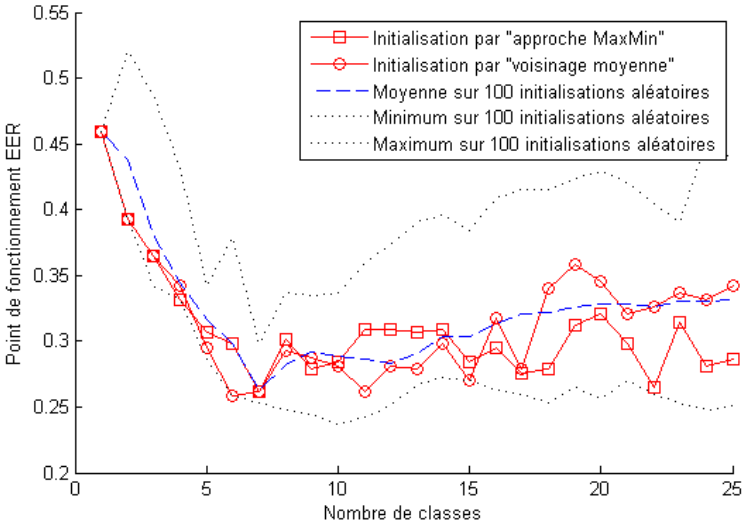


FIGURE 9.3 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme avec biais sans les contraintes du problème 2-classes et choix fixe du paramètre σ.

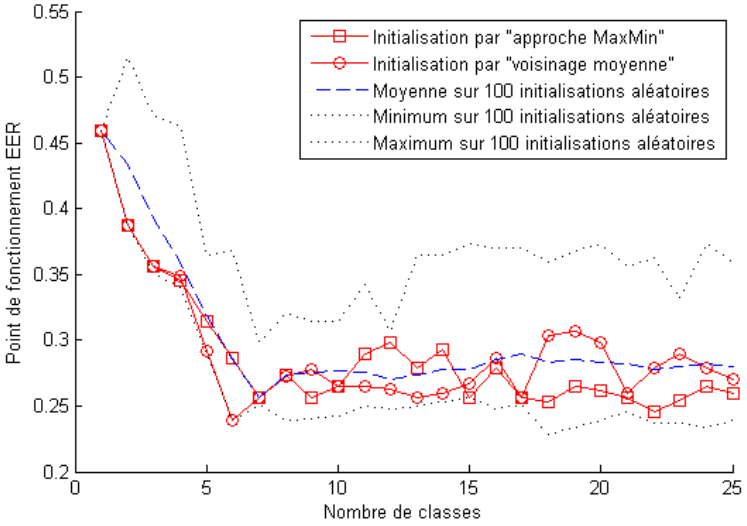


FIGURE 9.4 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme avec biais sans les contraintes du problème 2-classes et choix automatique du paramètre σ.

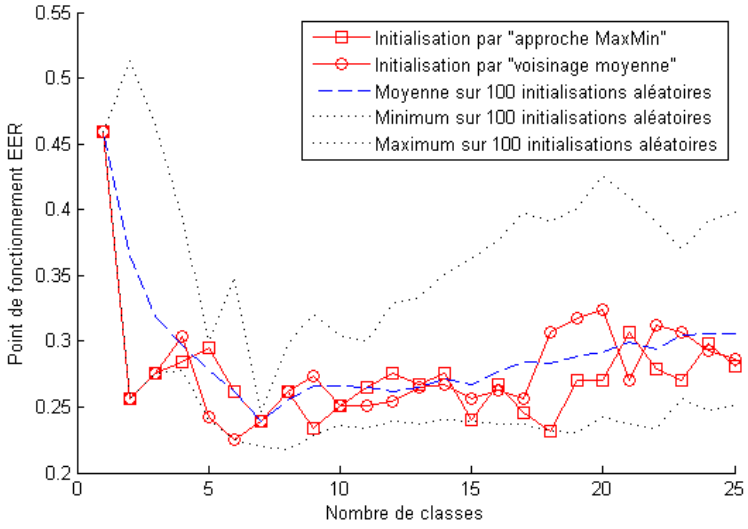


FIGURE 9.5 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme avec biais avec les contraintes du problème 2-classes et choix fixe du paramètre σ.

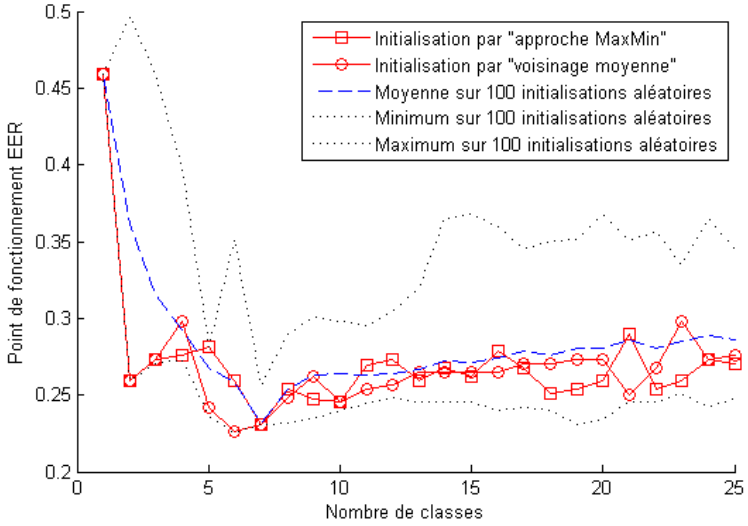


FIGURE 9.6 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme avec biais avec les contraintes du problème 2-classes et choix automatique du paramètre σ.

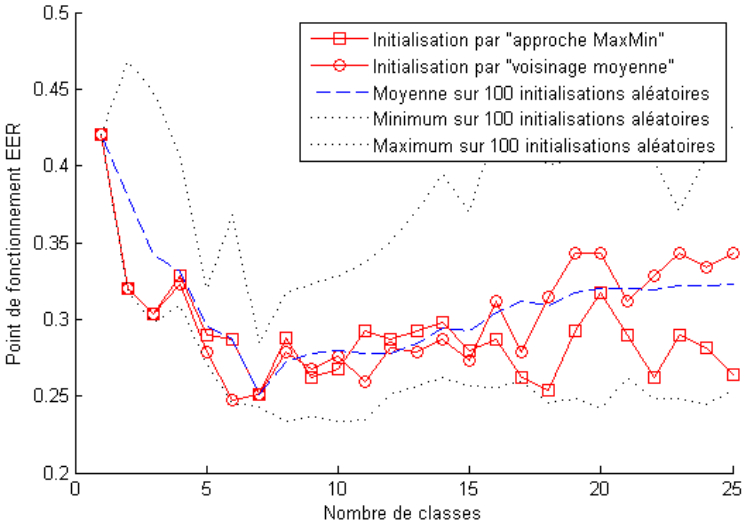


FIGURE 9.7 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme sans biais sans les contraintes du problème 2-classes et choix fixe du paramètre σ.

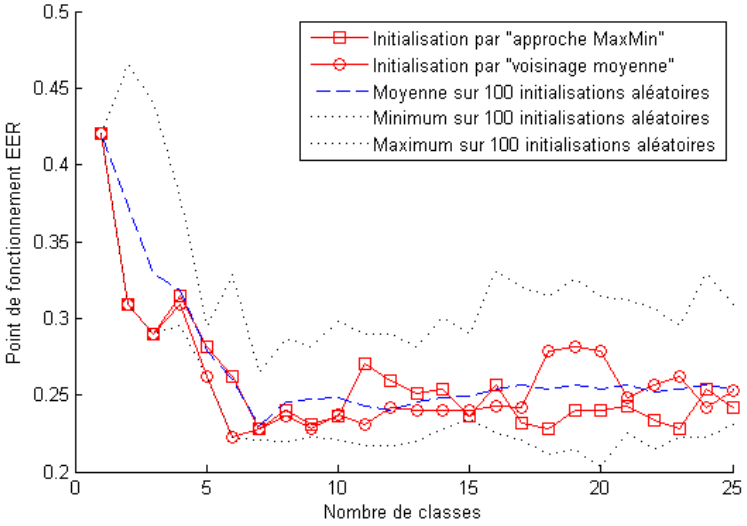


FIGURE 9.8 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme sans biais sans les contraintes du problème 2-classes et choix automatique du paramètre σ.

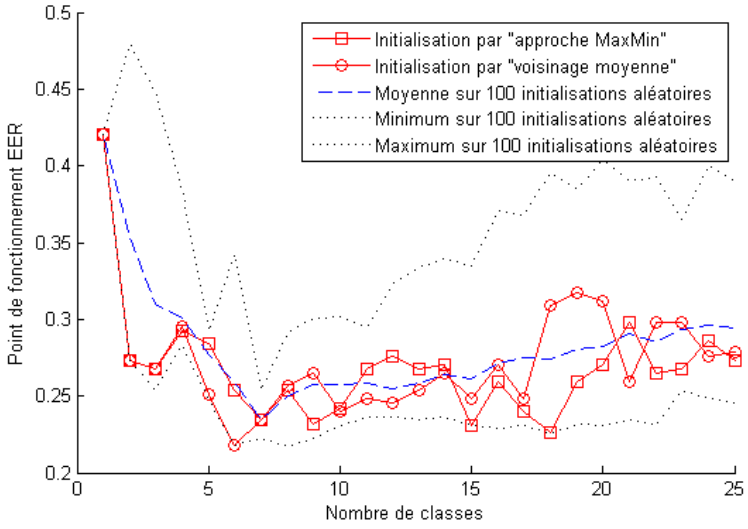


FIGURE 9.9 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme sans biais avec les contraintes du problème 2-classes et choix fixe du paramètre σ.

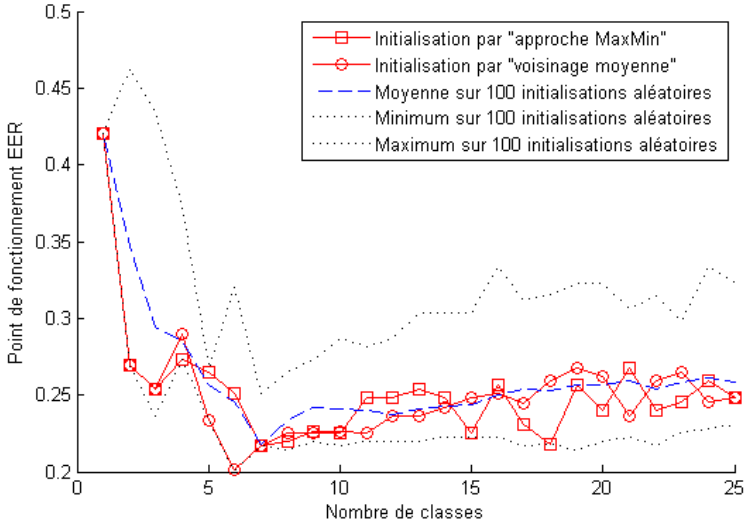


FIGURE 9.10 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, algorithme sans biais avec les contraintes du problème 2-classes et choix automatique du paramètre σ.

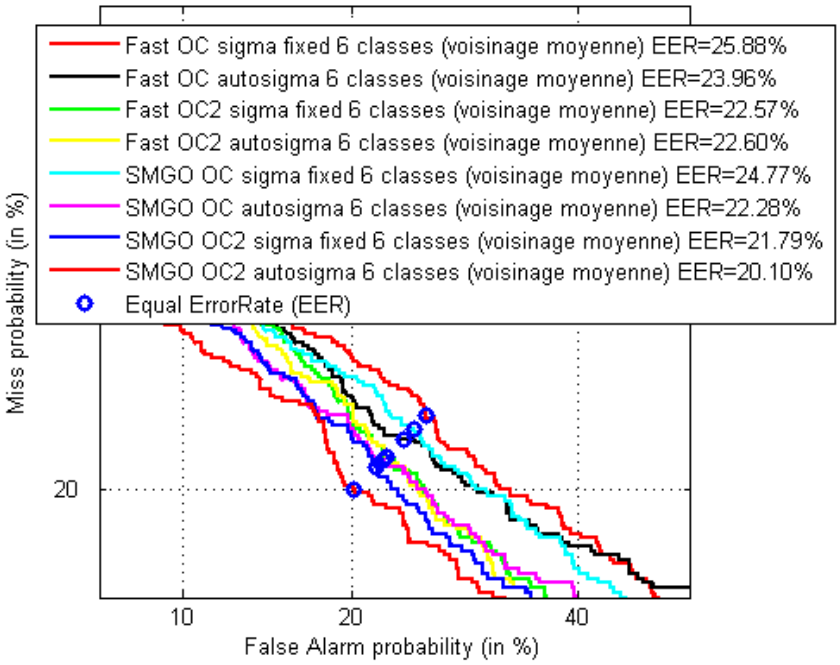


FIGURE 9.11 – Meilleurs détecteurs obtenus à l'aide de l'approche *clustering*.

Notons également que l'approche proposée met en évidence que le nombre de classes optimal ne correspond pas à l'expertise initiale (un *cluster* pour chaque chiffre). La méthode permet notamment de découvrir une répartition des données différente de celle-ci amenant à de meilleures performances. La figure 9.11 présente le réseau de courbes DET correspondant aux meilleurs détecteurs obtenus pour chacune des conditions d'apprentissage (choix d'algorithme, contraintes du problème 2-classes ou non, avec ou sans mise à jour locale de σ). Ces détecteurs correspondent tous à une configuration à 6 classes. Ils présentent également des performances nettement supérieures à celles d'une approche à l'aide d'un seul SVM 1-classe, et 5 d'entre eux sont meilleurs que ceux obtenus par approche supervisée.

Enfin, les performances obtenues à l'aide des méthodes d'initialisation exposées sont satisfaisantes. Celles-ci peuvent dès lors être utilisées pour initialiser l'algorithme proposé lorsque la quantité de données ne permet pas de réaliser plusieurs tirages aléatoires. En revanche, exploitant ces résultats, il reste possible d'étudier de nouvelles méthodes, notamment s'appuyant sur des critères d'information, afin de perfectionner l'initialisation et de s'approcher d'une méthode déterministe permettant d'atteindre le meilleur dictionnaire initial.

Cette étude nous a permis d'apprécier les bénéfices de la spécialisation des SVM constituant notre modèle. L'adaptation systématique du paramètre de noyau σ aux données permet d'obtenir de meilleures performances que via l'utilisation d'un paramètre global. Au-delà d'une meilleure description globale des observations, qui conduit à des détecteurs plus performants, l'approche proposée réalise un partitionnement non supervisé des données qui ne correspond pas à l'expertise attendue bien qu'elle soit plus performante que cette dernière. Ces éléments valident l'intérêt de l'approche *clustering* de type SVM 1-classe que nous pourrons expérimenter sur des données audio pour lesquelles aucune expertise n'a été réalisée.

9.2 Résultats sur des données audio

Cette section est l'occasion de présenter des résultats préliminaires de l'approche classification non supervisée par multiples SVM 1-classe appliquée à des données audio. Il s'agit de d'illustrer le potentiel de l'approche proposée pour traiter le problème de détection d'événements sonores anormaux abordé dans cette thèse.

9.2.1 Mise en œuvre expérimentale et résultats

Le protocole expérimental suivi s'appuie sur les résultats précédents pour la construction du détecteur et sur les expérimentations présentées aux chapitres précédents pour la constitution de la base de données. Cependant, la quantité de données comme les degrés de liberté du détecteur ont été limités afin d'obtenir ces résultats dans le temps imparti par la thèse.

Concernant la base de données, les observations ont été décimées d'un facteur 50, aussi bien pour l'apprentissage que pour le test. Par ailleurs, les centroïdes permettant d'initialiser le détecteur construit sont initialisés à l'aide de l'approche « voisinage moyenne » proposée à la section précédente. Le paramètre du noyau gaussien retenu est ici encore $\sigma = \sqrt{\gamma}$ (cf. 6.1.2). Deux valeurs $\nu = 0,02$ et $\nu = 0,05$ sont également choisies comme hyper-paramètre SVM. Enfin, nous présentons des résultats avec (OC2) et sans (OC) utilisation des contraintes du problème 2-classes ; et seul l'algorithme sans biais est utilisé.

Les figures 9.12 à 9.15 présentent les performances EER obtenues pour ces différentes expérimentations en fonction du nombre de sous-classes considérées, pour les signaux *m07* et *m12*.

Ces résultats montrent l'intérêt de l'algorithme avec utilisation des contraintes du problème 2-classes. En effet, sans ces contraintes, les performances sont moins intéressantes que celles obtenues en utilisant qu'un seul SVM 1-classe. Ces résultats indiquent une tendance en s'inté-

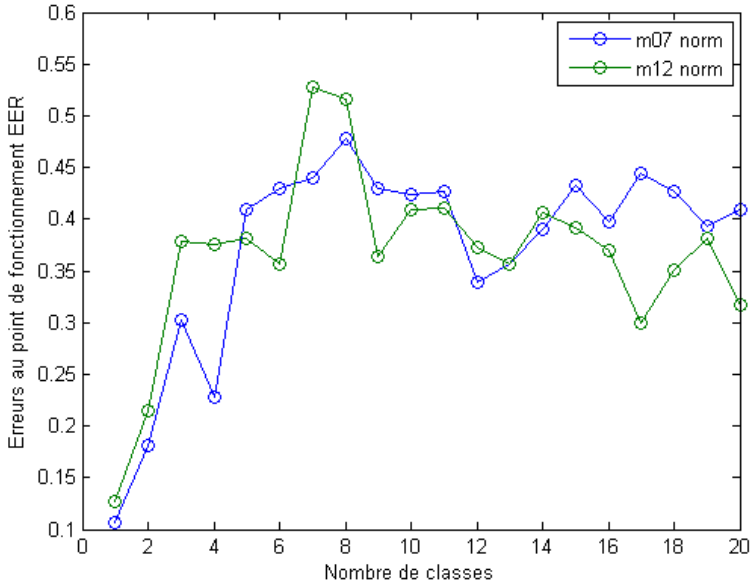


FIGURE 9.12 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, sans les contraintes du problème 2-classes et $\nu = 0,02$.

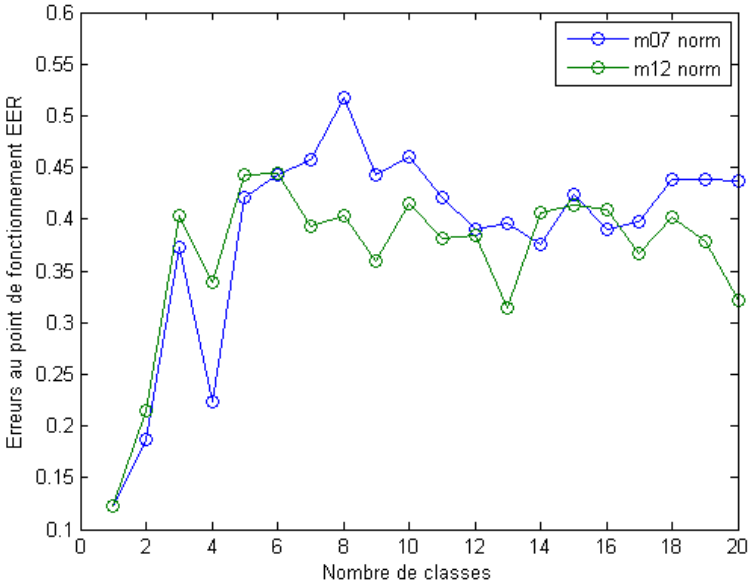


FIGURE 9.13 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, sans les contraintes du problème 2-classes et $\nu = 0,05$.

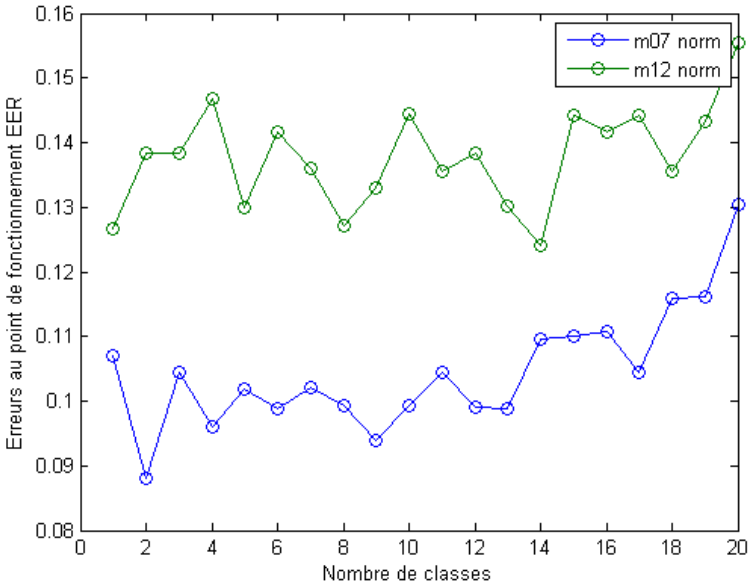


FIGURE 9.14 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, avec les contraintes du problème 2-classes et $\nu = 0,02$.

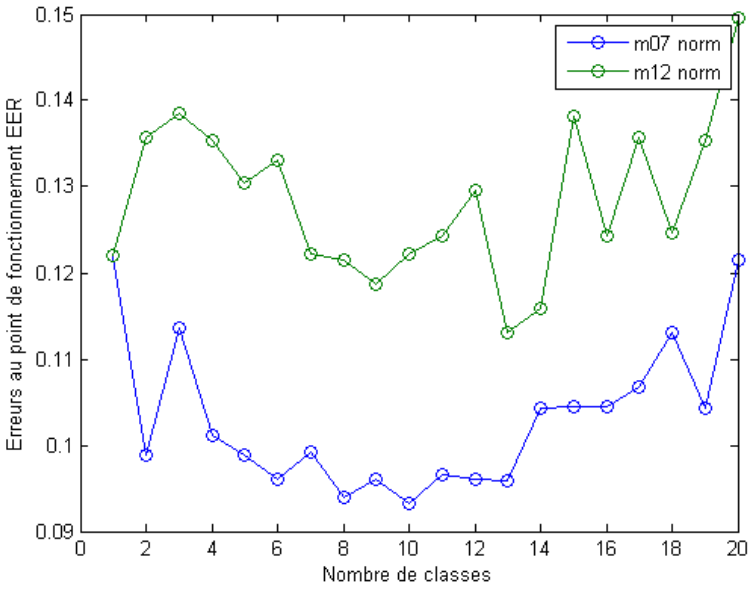


FIGURE 9.15 – Taux d'erreurs au point de fonctionnement EER des détecteurs proposés en fonction du nombre de classes, avec les contraintes du problème 2-classes et $\nu = 0,05$.

ressant au point de fonctionnement EER, mais ne permettent pas de juger globalement le gain pour l'ensemble des points de fonctionnement des détecteurs.

Ainsi, nous nous intéressons maintenant aux courbes DET associées aux détecteurs obtenus. En particulier, nous considérons ceux pour lesquels les performances semblent les plus avantageuses : avec contraintes du problème 2-classes, et dans le cas où $\nu = 0,05$. Ces courbes sont représentées sur les figures 9.16 pour les signaux *m07* et 9.17 pour les signaux *m12*.

9.2.2 Discussion

Malgré l'apparente complexité des résultats présentés, il est aisé de remarquer que la hiérarchie des détecteurs change en fonction du point de fonctionnement considéré. Ainsi, un détecteur optimal pour une probabilité de fausse alarme donnée n'est pas nécessairement optimal pour une autre valeur de probabilité de fausse-alarme. Ce constat impose donc la plus grande prudence dans l'analyse des courbes précédemment présentées qui ne considèrent qu'un unique point de fonctionnement à l'EER.

Cependant, indépendamment de la remarque précédente, il apparait qu'un ensemble de détecteurs exploitant plusieurs SVM 1-classes sont globalement meilleurs que celui obtenu par l'utilisation d'un seul SVM 1-classe ; leurs courbes respectives se situant sous celle établie pour ce dernier. Ce constat est particulièrement visible sur la figure **??**. De plus, ces résultats permettent d'envisager de considérer plusieurs détecteurs parmi lesquels choisir en fonction de la plage de fonctionnement souhaitée. Notamment, sur la figure **??**, si le détecteur obtenu pour 13 classes est acceptable pour une probabilité de fausse-alarme tolérée au-delà de $7,5\%$, d'autres détecteurs sont meilleurs sous ce seuil.

La figure 9.18 représente quelques courbes DET correspondantes aux détecteurs obtenus dans les conditions de la figure 9.14 ($\nu = 0,02$) pour les signaux *m07*. Celle-ci illustre dans

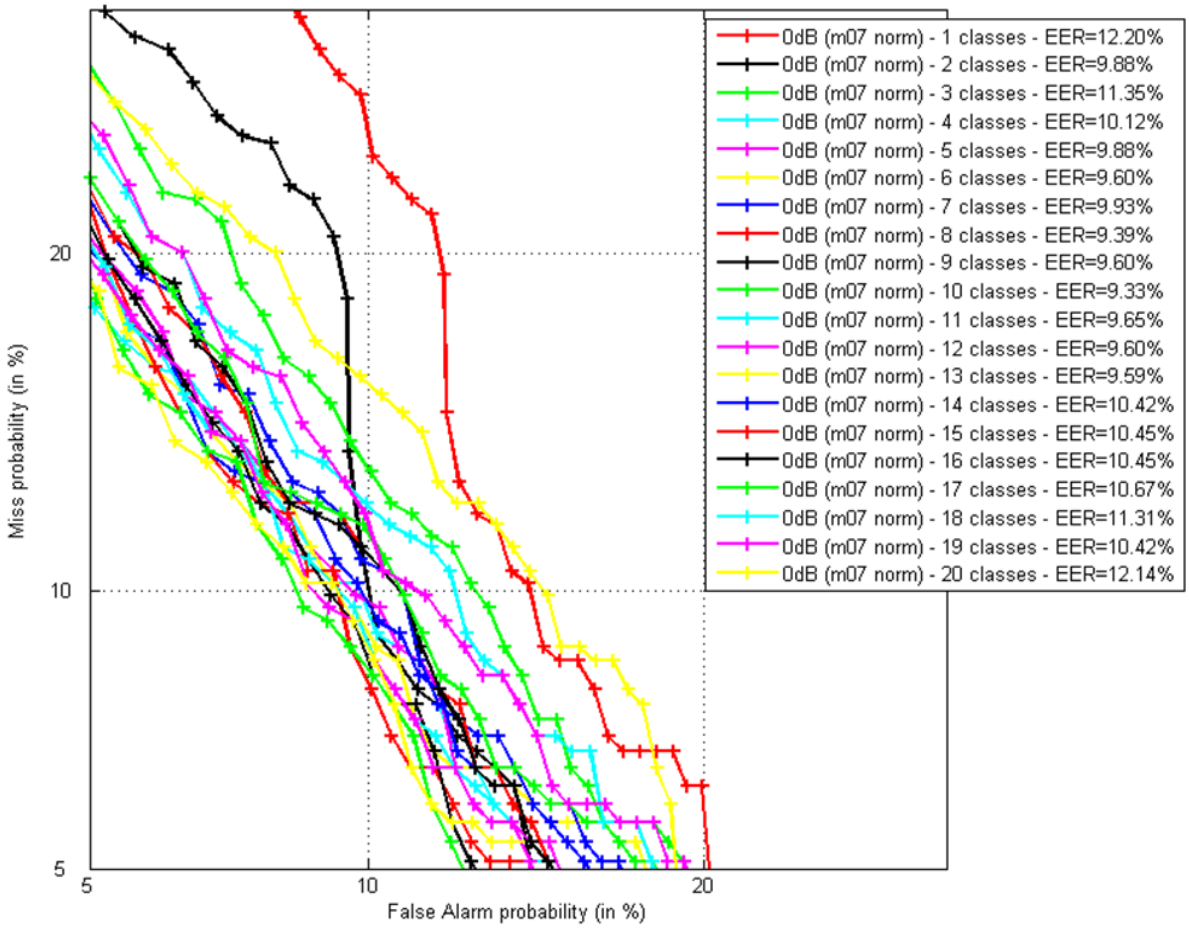


FIGURE 9.16 – Réseau de courbes DET des détecteurs proposés en fonction du nombre de classes pour les signaux *m07*, avec les contraintes du problème 2-classes et $\nu = 0,05$.

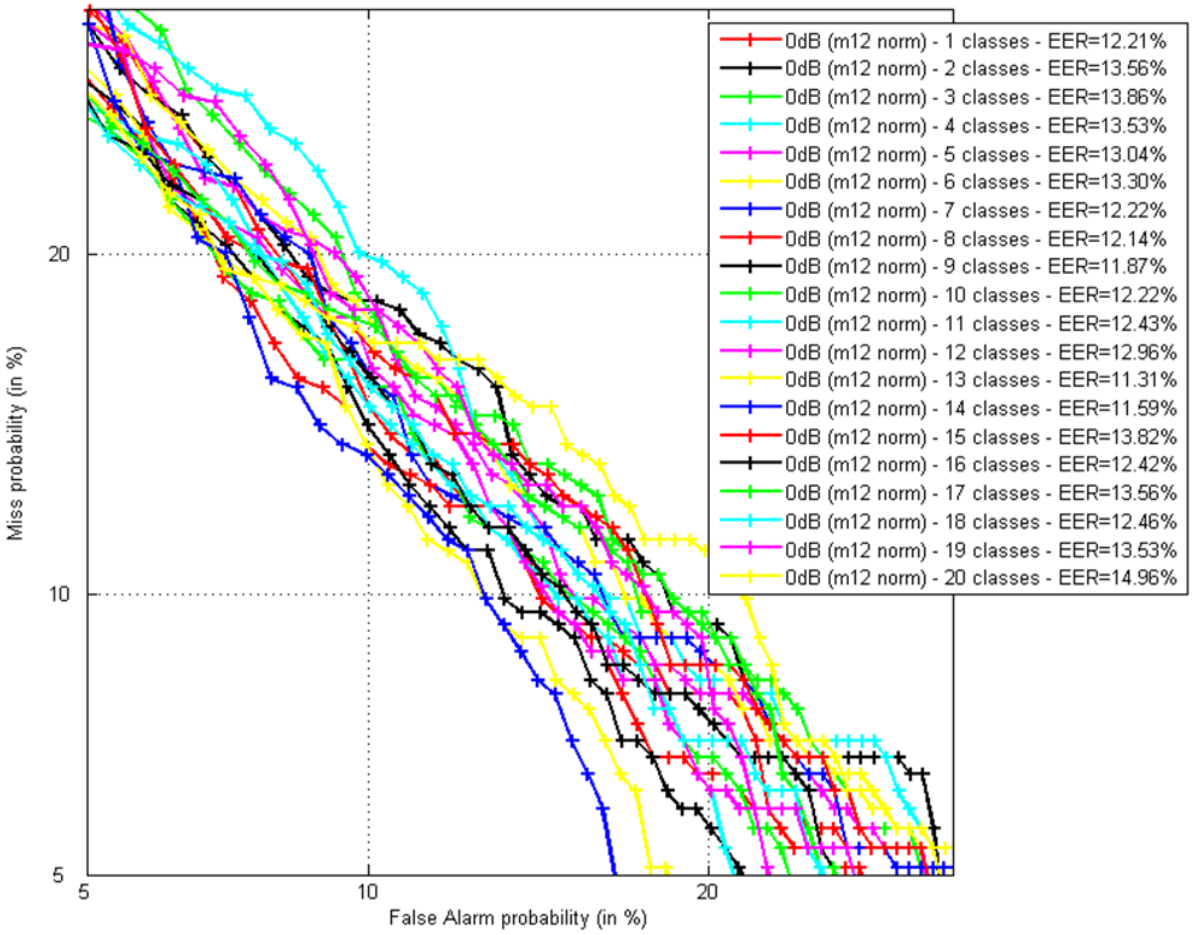


FIGURE 9.17 – Réseau de courbes DET des détecteurs proposés en fonction du nombre de classes pour les signaux *m12*, avec les contraintes du problème 2-classes et $\nu = 0,05$.

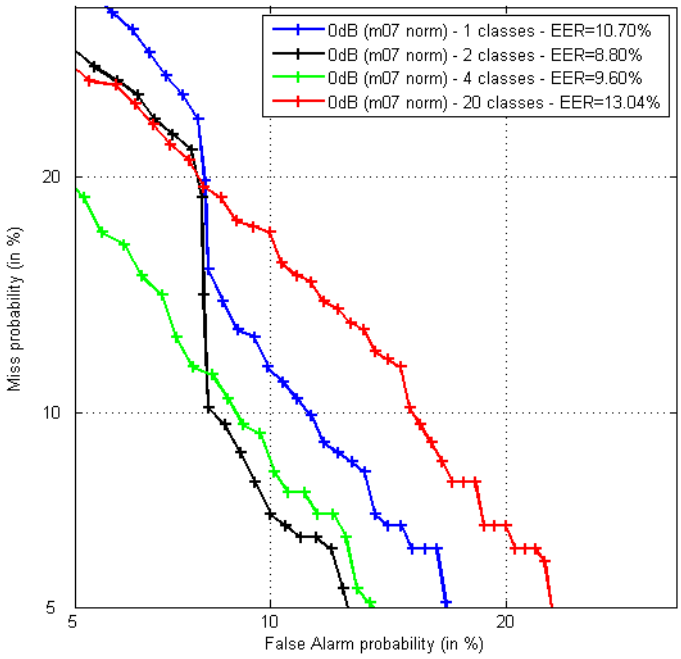


FIGURE 9.18 – Extrait du réseau de courbes DET des détecteurs proposés en fonction du nombre de classes pour les signaux *m07*, avec les contraintes du problème 2-classes et $\nu = 0,02$.

le détail les deux remarques précédentes. Le détecteur obtenu pour 2 classes (courbe noire), bien qu'il présente l'EER optimal, n'est pas le plus performant pour l'ensemble des plages de fonctionnement. En revanche, le détecteur obtenu à l'aide de 4 classes (courbe verte) constitue un compromis intéressant, bien plus performant que le détecteur original n'utilisant qu'un unique SVM 1-classe (courbe bleue).

9.3 Conclusion

Les résultats expérimentaux de la première partie de ce chapitre supportent les travaux présentés dans cette thèse et démontrent l'intérêt de l'approche proposée. En effet, la meilleure configuration obtenue met en œvre une approche de type *clustering* exploitant un ensemble de modèles 1-classe sans biais et avec contraintes du problème 2-classes, dont le paramètre de noyau est individuellement adapté.

Forts de ces résultats encourageants, nous avons utilisé cette approche des signaux de surveillance audio. Faute de temps, ces évaluations n'ont pu être conduites complètement au cours de la thèse et ces travaux se poursuivent. Néanmoins, les résultats préliminaires présentés dans la seconde partie du chapitre tendent à conforter l'intérêt de l'approche pour traiter le problème de détection d'événements sonores anormaux. Le fait d'avoir considérablement décimé la base d'évaluation nous impose toutefois d'être prudent dans cette interprétation qui devra être confortée par de plus larges expérimentations.

Conclusion

Ce dernier chapitre est l'occasion de dresser un bilan des travaux réalisés dans le cadre de cette thèse. Dans un premier temps, nous rappelons les motivations de cette étude ; celles-ci nous ayant amenés à étudier différents aspects du problème posé. Dans un second temps, nous listons les pistes sur lesquelles nous avons travaillé, les choix qui ont été faits et les perspectives liées à la poursuite des recherches dans ces différentes voies. Enfin, nous proposons de nouveaux axes de recherche auxquels des travaux futurs pourront être consacrés.

1 Rappel des motivations

La garantie de la sécurité des biens et des personnes est une problématique qui suscite un intérêt croissant dans le monde contemporain. Les technologies liées au marché de la surveillance, particulièrement vidéo, sont devenues incontournables dans tout environnement urbain moderne. Face à l'augmentation des zones sous surveillance, à la densification des caméras, et à une réduction constante des effectifs sur zone, la tâche des télé-opérateurs a considérablement évolué. Afin de les assister, l'automatisation des systèmes de surveillance est en plein essor. Celle-ci passe notamment par le déploiement de modalités nouvelles pour la surveillance, assistant efficacement la prise de décision, parmi lesquelles la modalité audio.

La communauté scientifique s'intéresse à l'utilisation de l'audio pour la surveillance de sites depuis le début des années 2000. Les algorithmes d'analyse proposés ont d'abord suivi un paradigme d'apprentissage totalement supervisé. Les événements à détecter sont alors complètement déterminés *a priori* et l'entrainement des classificateurs est réalisé une base de données de ces événements. Dans le contexte de la surveillance d'infrastructures de transport, il s'agit typiquement de cris, de coups de feu, ou encore d'actes de vandalisme (bris de glace, bombes de peinture aérosols, etc).

Les premières approches proposées ne prennent pas en compte l'environnement sonore, ou ambiance, de la zone sous surveillance. Cependant, la nature de ces environnements, complexe et parfois non stationnaire, augmente considérablement la difficulté de la tâche d'analyse. Aussi, des algorithmes intégrant une modélisation de l'ambiance ont été plus récemment mis au point. La tâche à réaliser reste en revanche la détection d'événements sonores pré-définis, dont un dictionnaire de modèles aura été préalablement appris.

En 2009, Capman et Ravera [CR10] ont proposé de ne s'intéresser qu'à la modélisation de l'ambiance. Ils montrent que leur approche, par modèles de mélange de gaussiennes (GMM), permet de détecter des événements anormaux sans nécessairement spécifier la nature de ceux-ci. Cette approche est partiellement supervisée, considérant uniquement des signaux de l'ambiance normale pour l'apprentissage et sans le besoin de construire un dictionnaire d'événements anormaux. La thèse s'est inscrite dans la continuité de ces travaux. En particulier, nous nous sommes intéressés à l'utilisation des algorithmes SVM 1-classe non supervisés pour adresser le problème

de la modélisation d'une ambiance sonore normale.

2 Réalisations et perspectives

2.1 Choix de la représentation

Nous avons illustré la richesse des espaces de représentation possibles du signal audio. Néanmoins, nous avons adopté, pour l'ensemble des travaux, un ensemble de descripteurs fréquentiels (énergies en sortie d'un banc de filtres) fixé *a priori*, et indépendant des signaux analysés. L'interprétabilité est une des raisons essentielles de ce choix. En effet, ces descripteurs constituent une version compacte du spectre. Ainsi, il a été possible d'analyser les résultats en se référant à une représentation temps-fréquence (spectrogramme) des signaux audio. Dans un second temps, figer la représentation a limité les degrés de liberté de notre étude. Comme nous avons orienté nos recherches sur la mise en œuvre de méthodes SVM adaptées au problème, le choix d'un meilleur espace de représentation n'a pas constitué un axe de recherche majeur. Enfin, parce que cette représentation a été celle utilisée au cours du projet CARETAKER qui a constitué la référence pour nos travaux, il a été naturel de s'appuyer sur celle-ci afin de comparer les approches.

Néanmoins, la modification de l'espace de représentation constitue une perspective d'amélioration des systèmes proposés. D'une part, un choix adapté de cet espace permet d'accroitre les performances globales ; en témoigne l'ensemble des représentations spécifiques étudiées dans la littérature. D'autre part, des travaux connexes, consistant à classifier des signaux détectés, ont montré tout le bénéfice que pouvait apporter cette liberté d'espace de représentation lorsque le problème est complètement posé (pour des signaux anormaux pré-définis par exemple) [Lec09]. Enfin, dans le contexte de l'approche proposée au chapitre 8, le choix d'un espace de représentation propre à chacun des SVM 1-classe pourrait affiner encore le modèle de l'ambiance normale. Cela permettrai d'améliorer la description globale, et, dans le même temps, d'analyser finement les signaux concernés par cette description locale.

2.2 Détection à l'aide de SVM 1-classe

S'intéressant aux méthodes par machines à vecteurs de support (SVM) pour la modélisation d'événements anormaux, nous avons proposé un premier modèle de détection s'appuyant sur les SVM 1-classe. L'idée est de tester l'hypothèse H_0 « ambiance normale » de laquelle relèvent les signaux d'apprentissage. Les besoins opérationnels de gestion du compromis entre les erreurs commises, fausse-alarme et non détection, nous ont amenés à proposer une méthode de construction de familles de fonctions de décision sur la base d'un modèle SVM entraîné. Cette approche, validée sur des données réelles, permet d'adapter les performances d'un détecteur sans réaliser de nouvel apprentissage. De surcroît, les résultats montrent qu'il est possible, au travers de cette construction, d'obtenir des détecteurs plus performants que ne l'aurait permis une approche SVM 1-classe standard. Nous avons également démontré la pertinence de l'intégration temporelle de la statistique de décision, sur la base d'une information de segmentation automatiquement extraite du signal audio.

L'ensemble de ces travaux a donné lieu à plusieurs publications dans des conférences nationales et internationales, et contributions dans le cadre du projet VANAHEIM. Les performances obtenues ont confirmé l'intérêt pour l'approche SVM, en améliorant les résultats d'un précédent projet, établis à l'aide d'une approche GMM. Néanmoins, pour aller plus loin dans cette approche, nous proposons trois pistes d'étude.

Dans un premier temps, l'approche SVM souffre d'une statistique de décision difficile à interpréter opérationnellement. En effet, le score obtenu n'a pas de sens probabiliste ; l'adjonction d'une méthode telle que proposée par Platt [Pla99] est alors envisageable.

Ensuite, l'approche s'appuie sur l'hypothèse que les signaux d'apprentissage relèvent de l'hypothèse H_0. En réalité, ces signaux n'étant pas expertisés, il est fort probable que, lorsque leur quantité augmente, ceux-ci présentent une part non négligeable d'événements devant être considérés comme anormaux. Si l'approche SVM 1-classe rejette naturellement une fraction d'*outliers*, rien ne garantit que ces derniers correspondent bien aux données anormales. La robustesse des modèles ainsi entraînés face à la présence d'événements anormaux dans les signaux d'apprentissage devrait donc être étudiée également.

Enfin, nous avons montré le gain apporté par l'exploitation d'une information temporelle pour compresser la statistique de décision. L'approche consiste en effet à considérer une unique valeur pour un ensemble de trames constituant un segment ; typiquement la statistique moyenne. De plus, l'étape d'apprentissage ne bénéficie pas de cette information. Une piste possible, outre la confrontation de différentes méthodes de segmentation automatique des signaux, est de considérer une information non compressée. Ainsi, chaque segment serait représenté par l'ensemble des observations qui le composent, devenant alors une trajectoire dans l'espace d'observation ; l'entraînement et le test des modèles SVM pouvant être réalisés sur ces trajectoires.

2.3 Travaux algorithmiques

La quantité de signaux disponibles pour l'apprentissage, le besoin d'algorithmes rapides pour un déploiement aisé et la possibilité de faire évoluer les modèles au cours du temps (apprentissage en ligne ou avec démarrage à chaud) sont autant d'éléments nous ayant amenés à investiguer le terrain de l'algorithmie.

Nous avons d'abord proposé un problème SVM 1-classe sans biais. Ce travail a été motivé par de récents travaux concernant les approches sans biais dans le cadre des SVM 2-classes d'une part, et l'approche SVM 1-classe avec contraintes du problème 2-classes d'autre part. Nous avons ensuite montré qu'il existe un problème SVM unifié, couvrant les approches avec ou sans biais, 2-classes ou 1-classe, et dans ce dernier cas, avec ou sans les contraintes du problème 2-classes. Partant de cette forme unifiée, nous avons ensuite démontré que l'algorithme SMGO peut traiter l'ensemble des problèmes SVM évoqués. Les expériences conduites sur des données de synthèse et des données réelles ont démontré les performances de cet algorithme face à des approches de l'état de l'art. Ces travaux ont également donné lieu à une publication dans une conférence nationale.

En terme de perspectives, le principal apport de ces travaux réside dans la possibilité, à l'avenir, de faire bénéficier l'ensemble des problèmes SVM des améliorations algorithmiques du solveur rapide SMGO. Par ailleurs, le SVM 1-classe proposé ne présente plus la ν-propriété du problème 1-classe introduit par Schölkopf. Cette propriété relie le paramètre ν à la fois à la borne supérieure du nombre de vecteurs de support au-delà de la marge, et, à la borne inférieure du nombre total de vecteurs de support. Réinjecter cette propriété au problème proposé nous semble constituer un défi intéressant. Enfin, si l'algorithme proposé se prête - il s'agissait d'une contrainte initiale - au démarrage à chaud, nous n'avons pas étudié, au cours de la thèse, l'apprentissage en ligne d'une ambiance normale. Construire un système capable d'évoluer au cours du temps reste donc un défi à relever.

2.4 Une approche de type *clustering*

A la recherche d'une méthode afin de décrire le plus précisément les signaux d'apprentissage, nous avons étudié une approche de type *clustering*. Celle-ci, s'appuyant là encore sur les méthodes SVM, a pour objectif de décrire des données d'apprentissage à l'aide d'un ensemble de modèles localement optimisés, à l'instar d'une modélisation de type GMM. Ainsi, nous avons proposé une approche permettant de construire un ensemble de modèles SVM 1-classe concurrents. Celle-ci a été l'objet d'un brevet au cours de la thèse.

Les résultats présentés sur un problème de détection simple, spécifiquement construit, permettent d'attester de l'intérêt de cette approche face à une modélisation à l'aide d'un unique SVM 1-classe. De surcroît, ces résultats ont conduit à de meilleures performances que celles obtenues au travers de l'utilisation d'un algorithme supervisé. Cette approche nécessite néanmoins d'y consacrer encore des efforts afin de la rendre pleinement opérationnelle dans le cadre applicatif qui nous intéresse. D'une part, les degrés de liberté de celle-ci sont nombreux et leurs effets doivent être quantifiés. D'autre part, l'approche présentant une composante itérative (optimisation par directions alternées), la convergence de la procédure doit être étudiée.

2.5 Aspects opérationnels

Au travers de cette thèse, nous avons également contribué à la définition d'un protocole d'évaluation des performances pour un système de détection dans le cadre de la surveillance audio. Cette étude a permis de préciser certaines contraintes liées à l'acquisition des signaux d'apprentissage, ou d'éventuels événements anormaux dans le but de construire une base d'évaluation. En particulier, nous avons constaté la nécessité, y compris dans le cas d'approches partiellement supervisées, de parfaitement maitriser l'environnement et le système d'acquisition des signaux audio.

Toutefois, cette thèse a permis de démontrer en situation opérationnelle un système de surveillance s'appuyant sur la modalité audio. En effet, les résultats ont été pour partie intégrés dans un démonstrateur temps-réel. Ces travaux, réalisés dans le cadre du projet VANAHEIM, ont notamment fait l'objet d'une présentation en environnement réel (station de métro *Bibliothèque François Mitterrand*, RATP). La problématique de l'industrialisation de la solution proposée au cours de cette thèse CIFRE a donc été abordée. Il reste cependant une partie des travaux, la modélisation de type *clustering* par exemple, dont l'intérêt devra être démontré dans des conditions réelles.

3 Travaux futurs

S'appuyant sur le concept de détection d'événements anormaux, nous avons abouti à une solution performante, opérationnellement efficace dans le contexte de l'audio-surveillance. Naturellement, un tel système devrait à terme être capable d'identifier la nature des événements anormaux détectés. Il s'agit de classification, dont le but est d'attribuer une étiquette (ou classe) à un événement isolé temporellement. Ainsi, un système à deux étages - détection et classification - pourra être construit en utilisant les algorithmes proposés dans cette thèse. La réalisation d'un tel système permettra également de comparer l'approche proposée à des approches standard, complètement supervisées.

Le démonstrateur proposé utilise la modalité audio de deux façons différentes. D'un côté, il indique, pour chaque micro dans la zone sous surveillance, un niveau d'anormalité constaté au travers d'une jauge. D'un autre côté, couplé à un algorithme de sélection de flux, il sélectionne

automatiquement le micro (ou la caméra lui étant associée) le plus pertinent pour un opérateur. Cependant, nous pensons que la mise en place d'une analyse multi-modale, où l'information issue du signal audio est fusionnée avec celle issue d'autres capteurs (vidéo au premier chef), sera une des clés du déploiement de tels systèmes. Enfin, alors qu'un opérateur seul peut visualiser plusieurs écrans de contrôle, il n'est pas envisageable d'écouter plusieurs flux audio. Ainsi la problématique de la représentation de l'information audio dans le cadre de la surveillance reste pleinement à traiter.

Nous avons mis en exergue également la nature fortement non stationnaire des signaux de surveillance audio. Néanmoins, des cycles apparaissent au cours du temps et à différentes échelles (arrivées/départs de trains, heures pleines/creuses, jours ouvrés/weekends, etc). Si les algorithmes mis en œuvre au cours de la thèse permettent le suivi de l'ambiance au cours du temps (apprentissage en ligne), nous pensons que des modèles indépendants les uns des autres doivent pouvoir être activés en fonction de ces cycles ; simplifiant l'effort de mise à jour, et dans le même temps, par l'analyse de ces modèles, retournant une information sur la nature des différents cycles.

Enfin, si par les besoins du partenaire industriel nous avons limité l'application de nos travaux au contexte de la surveillance audio, les modèles proposés sont applicables à l'ensemble des problèmes de type « détection de nouveauté » ou « *clustering* ». Ainsi, nous espérons pouvoir évaluer ces approches dans d'autres contextes parmi lesquels, sans s'y limiter, la détection d'intrusion dans les réseaux, le contrôle non destructif, ou plus généralement la détection de rupture dans des séries temporelles.

Cinquième partie

Annexes

Annexe A

Enregistrements *in situ*

Dans le cadre du projet VANAHEIM, des acquisitions spécifiques de signaux anormaux dans un environnement de station de métro ont été organisées. Cette annexe rapporte le travail réalisé dans l'optique de constituer une base de données la plus réaliste possible à des fins d'évaluation. Dans les deux premières sections, nous décrivons les motivations qui nous ont conduits à effectuer ces travaux ainsi que le protocole expérimental mis au point. La troisième section rapporte des travaux d'analyse des signaux enregistrés, qui se sont révélés très bruités. Deux approches visant à exploiter ces signaux ont été explorées : la réduction de bruit et l'extraction des caractéristiques de l'environnement. La seconde n'ayant pas relevé des travaux de cette thèse n'est pas présentée.

A.1 Motivations

En juillet 2011, il a été décidé de réaliser des acquisitions spécifiques de signaux audio dans la station "XVIII diciembre" du métro de Turin. Cette action a notamment été motivée par :

- le besoin de réaliser une inspection des installations, les signaux acquis jusque-là étant d'une qualité inexploitable,
- l'enregistrement d'événements anormaux (station fermée) afin d'être utilisés par l'outil de simulation (voir chapitre 5 ; ceci devra permettre de synthétiser des signaux d'évaluation les plus réalistes possibles, à la fois pour des travaux de détection et des travaux de classification,
- mesurer des propriétés acoustiques de l'environnement de la station (réverbération, niveaux sonores, etc.) pour une meilleure compréhension des résultats obtenus sur des signaux enregistrés.

Ces travaux ont été réalisés par Thales avec le support de GTT (régie des transports de Turin). La constitution d'une base de données de signaux étant l'un des axes proposés pour cette thèse, j'ai donc été associé à la préparation et à la réalisation de la mission qui s'est déroulée du 26 au 28 juillet 2011.

A.2 Protocole

A.2.1 Principe

L'idée de base est de jouer des événements anormaux à un niveau sonore cohérent au sein de la station de métro. Ainsi, les signaux audio sont impactés par les effets acoustiques de l'environnement (atténuations fréquentielles, réverbération, etc.). L'ensemble de l'activité acoustique est

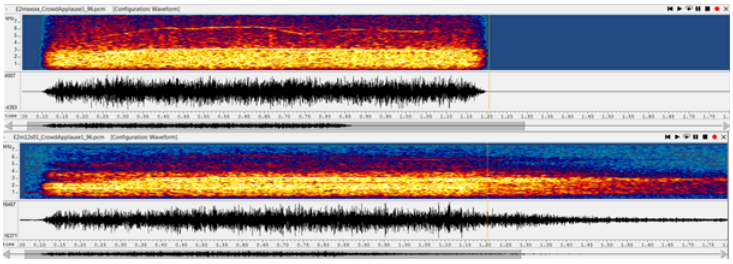


FIGURE A.1 – Spectrogramme et forme d'onde d'un événement sonore anormal (foule en liesse) joué (en haut) et enregistré (en bas) dans la station. On note immédiatement un filtrage fréquentiel en hautes fréquences ainsi qu'une réverbération marquée.

capturé par le système d'acquisition audio et les événements spécifiques peuvent être retrouvés hors ligne au sein des enregistrements. Enfin, ces enregistrements d'événements anormaux sont utilisés dans l'outil de simulation développé afin de générer des bases de données plus réalistes.

Les événements anormaux sont joués depuis différentes positions car, pour chaque couple source/microphone, des fonctions de transfert acoustique différentes s'appliquent. Les événements sont concaténés à un niveau sonore approprié au sein d'une séquence audio unique. Ceci permet une maîtrise parfaite de l'intervalle entre deux événements et accélère les traitements d'extraction depuis les enregistrements.

A.2.2 Séquences d'événements anormaux

Le choix des événements anormaux doit être représentatif d'une grande variabilité spectrale dans les signaux anormaux à détecter. Comme les événements ne seront pas utilisés pour de la classification, le réalisme de la présence d'un événement spécifique dans une station de métro n'est pas une prérogative. Cette variabilité sera utile pour identifier les caractéristiques des signaux qui pourraient être particulièrement difficiles à détecter. Cependant, certaines catégories particulières comme des coups de feu, des complaintes, des aboiements ou des cris pourront être utilisés pour achever des résultats préliminaires en classification ou pour illustrer nos travaux. Ainsi, la séquence audio utilisée est constituée de 77 événements de différentes catégories : enfants, foule, vandalisme, cris, travaux, coups de feu, explosions, claquements de portes, chutes, pas, pleurs, chiens, bagarres, musique, sirènes ou encore téléphones.

Cette sélection est représentative de caractéristiques variées : signaux hautes fréquences, pleine bande, basse fréquence, stationnaires ou impulsifs. Nous avons également adapté les niveaux sonores des événements ; la figure A.2 illustre les niveaux sonores moyens pour différents sons. En particulier, la séquence audio inclut 3 réalisations de chaque événement : au niveau sonore souhaité, à +6dB et à −6dB. En complément, une sinusoïde pure à 1kHz à un niveau de référence de 94dB et un signal de parole à un niveau de 67, 7dB ont été inclus pour référence à la séquence.

A.2.3 Protocole expérimental

Le système de surveillance audio installé dans la station de métro "XVIII diciembre" de Turin utilise 7 microphones parmi lesquels deux sont physiquement couplés : 3 micros à l'entrée de la station autour des tourniquets et 2 micros au-dessus de chacun des deux quais. Deux (micros 04 et 06) ne sont pas utilisés pour des raisons techniques.

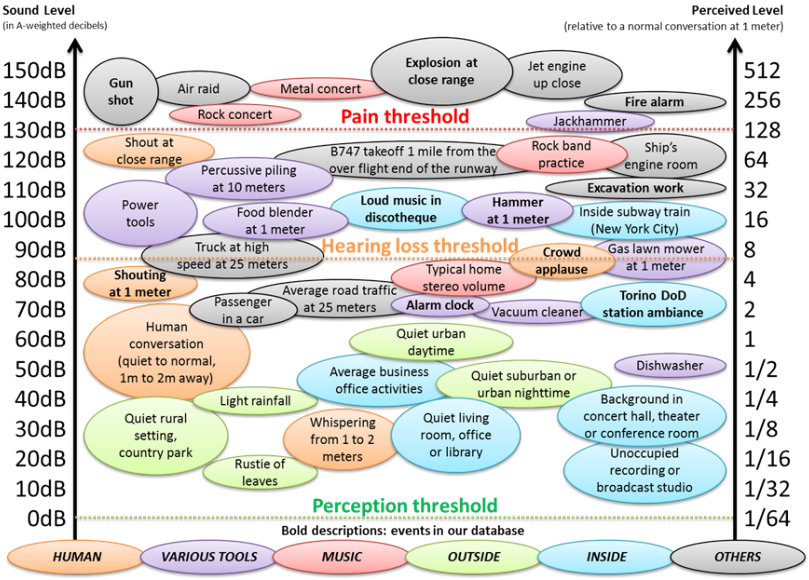


FIGURE A.2 – niveaux sonores moyens pour des événements sonores variés.

Grâce à un système mobile, la séquence d'événements anormaux a été jouée à différents endroits de la station, sous les microphones 01, 02, 03, 05 et 07. Le niveau sonore a été réglé afin de refléter les niveaux attendus. Afin de ne capturer que les événements anormaux, les acquisitions ont été réalisées de nuit alors que la station était close.

A.3 Analyse des signaux corrompus par le bruit

Dans cette section, nous décrivons l'analyse des signaux enregistrés au cours de la session à Turin dont le protocole a été décrit ci-dessus. Toutes les analyses ont été réalisées hors ligne. Nous avons particulièrement porté notre intérêt sur la corruption par le bruit des signaux car celui-ci ne nous permettait pas d'utiliser les signaux comme nous le souhaitions.

A.3.1 Identification du bruit

Nous avons constaté sur l'ensemble des signaux enregistrés la présence d'un fort bruit de fond. Deux causes ont été identifiées :

- la numérisation des signaux par le système est perturbée par l'environnement électromagnétique de la station (parasite des courants forts sur les alimentations en courant faible des boitiers de traitement audio),
- le bruit ambiant de la station, en particulier des ventilations.

Les figures A.3 et A.4 illustrent ces altérations sur une portion de signal d'environ 5 secondes.

Ces figures permettent d'identifier les sources de bruit :

- une structure harmonique correspondant au bruit de numérisation : multiples de 50Hz particulièrement marqués en dessous de 600Hz et au-delà de 4200Hz,
- les pics fréquentiels correspondants au système de ventilation à 930Hz, 1120Hz et 5000Hz.

Lors de l'analyse des signaux enregistrés, nous avons mesuré des RSB variant de 5 à 40 dB (ITU-R468) avec un SNR médian de 25dB.

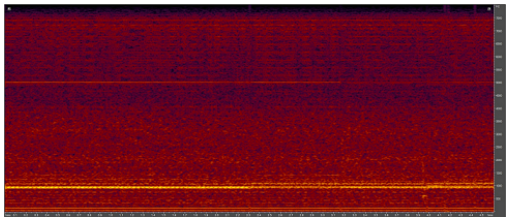


FIGURE A.3 – Spectrogramme d'une séquence de bruit.

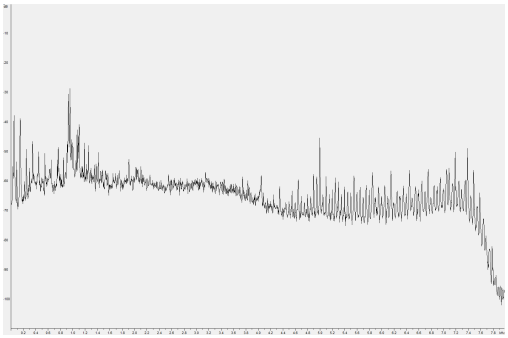


FIGURE A.4 – Spectre long terme d'une séquence de bruit.

A.3.2 Conséquences du bruit

Notre principal objectif est l'enregistrement d'événements anormaux *in situ* pour améliorer la qualité de notre outil de simulation de signaux. Cet outil permet notamment de contrôler le rapport signal à bruit (RSB), qui dans notre cas est un rapport événement à ambiance. On distingue maintenant trois signaux : l'ambiance (A), l'événement anormal (E) et le bruit de fond identifié (B) qui corrompent E. On note alors les RSB correspondants : REA, REB et RBA.

La station agissant comme un filtre, l'ajout d'événements issus de la base de données fait apparaître des fréquences naturellement atténuées dans la station de métro, rendant alors plus facile leur détection. A l'inverse, si nous incluons les événements enregistrés dans la station il y aura une bonne correspondance. De plus, pour l'inclusion de signaux de type coups de feu ou cris, nous produisons des signaux d'un grand réalisme.

Conséquence de la corruption des signaux par le bruit décrit ci-dessus, nous n'avons pas pu utiliser ceux-ci directement dans notre outil de simulation. En effet, le REB des enregistrements était si faible que lors de l'inclusion de ceux-ci le système de détection détectait également le bruit de fond.

Nous avons utilisé notre outil de simulation, ciblant un REA de 10dB pour les événements enregistrés. On constate que pour 12% des événements anormaux ainsi inclus, le RBA observé est supérieur à 0dB ; pour un REA cible de 25dB, ce score monte à 50%. Or, comme l'illustrent les figures A.5 et A.6, le système de détection détecte de manière performante l'inclusion du bruit de fond à 0dB. Ainsi, le système sera vraisemblablement plus à même de détecter le bruit de fond que l'événement anormal inclus, faussant considérablement notre évaluation. La figure A.7 montre la fraction d'événements dont le RBA est supérieur à une valeur donnée pour un REA fixé.

A.3.3 Utilisation des signaux enregistrés

En définitive, les signaux enregistrés ne pouvaient pas être utilisés directement dans l'outil de simulation. Nous avons donc exploré deux approches afin d'en tirer néanmoins profit :

- Débruitage des signaux pour une utilisation suivant le protocole initial : il s'agit d'estimer et de supprimer les composants du bruit de fond sans altérer l'événement enregistré, ni l'information supplémentaire gagnée par cette approche (réverbération, trajectoires acoustiques multiples, filtrage du signal, etc.),
- Extraire les caractéristiques de l'environnement depuis les signaux enregistrés : cette approche consiste à estimer à l'aveugle la réponse impulsionnelle de l'environnement, per-

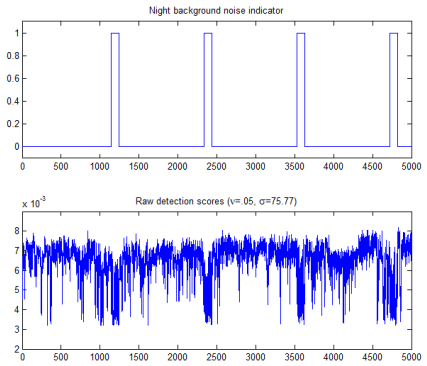


FIGURE A.5 – Résultat de détection du bruit de fond inclus à un RBA de 0dB. Les SNR locaux mesurés sont −3.66dB, −3.09dB, −0.9dB et −1.6dB.

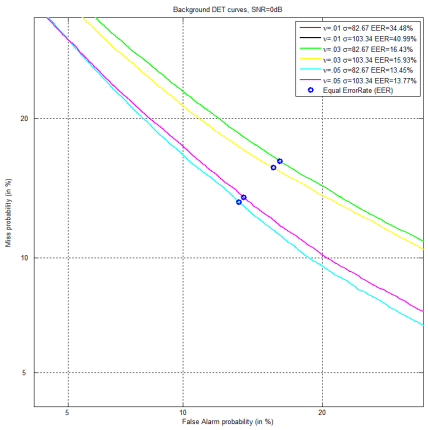


FIGURE A.6 – Détection d'un bruit de fond, courbes DET pour différentes configurations du système.

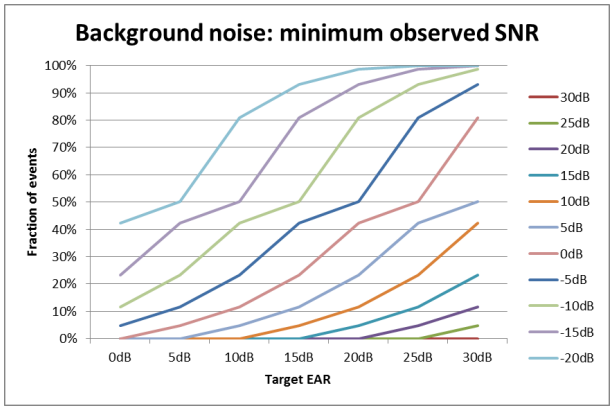


FIGURE A.7 – Fraction des événements dont le RBA est supérieur à une valeur donnée (de −20dB à +30dB) à différents niveaux de REA cibles (0dB à 30dB).

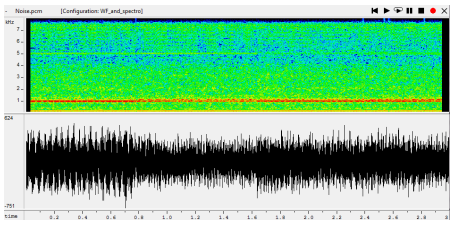


FIGURE A.8 – Séquence de bruit pour l'estimation des filtres et profil de débruitage.

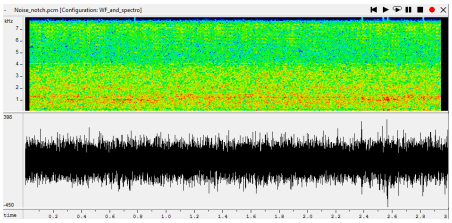


FIGURE A.9 – Séquence de bruit sur laquelle a été appliquée une combinaison de filtres à encoche.

mettant ainsi de simuler des signaux enregistrés depuis des signaux bruts.
Comme évoqué plus tôt, nous ne présentons que les travaux concernant la première approche.

A.4 Réduction de bruit

Dans cette section, nous décrivons les travaux de réduction du bruit de fond des signaux d'événements anormaux enregistrés *in situ* à Turin. Notre objectif est de s'assurer que le bruit de fond de ces signaux n'augmente pas artificiellement les scores de détection. A l'inverse, nous devons également garder à l'esprit que si les signaux sont trop modifiés par la réduction de bruit, ils sont dénaturés et nous perdons en réalisme. En particulier, nous avons exploré trois approches pour débruiter les signaux :

- Générer des filtres à encoche appropriés pour réduire des fréquences spécifiques,
- Soustraction du bruit de fond après estimation statistique d'un profil fréquentiel,
- Une combinaison de ces deux premières approches.

Dans cette section, nous présentons et illustrons les résultats obtenus pour chacune des approches. La figure A.8 montrent la séquence de bruit utilisée pour notre analyse (estimation de profil ou de filtres). Il s'agit d'une concaténation de différents instants capturés au sein d'une séquence audio de 13 minutes.

A.4.1 Approche par filtres à encoche

Les filtres à encoche sont paramétrés par une fréquence centrale, une bande passante et un gain. La combinaison de 20 à 30 filtres, construits à partir d'une analyse des pics spectraux sur le spectre long-terme du signal de bruit, est nécessaire pour obtenir une réduction significative du bruit. La figure A.9 montre le résultat de ce filtrage sur la séquence de bruit. La figure A.10 présente un enregistrement d'événement anormal, et la figure A.11 le résultat du filtrage sur cet événement.

Cette approche conduit à une réduction moyenne du rapport événement à bruit (REB) de $6,25$dB. Cependant, elle ne traite que des fréquences ciblées et ne traite pas le souffle du système de ventilation.

A.4.2 Approche par estimation d'un profil de bruit

Cette approche modélise statistiquement le bruit de la séquence de référence. Celle-ci est plus adaptée au traitement de bruits aléatoires tel que le souffle du système de ventilation. Cependant, elle est également plus complexe à paramétrer. Nous avons utilisé pour cette étude préliminaire l'outil de réduction de bruit fourni pas le logiciel Adobe Audition.

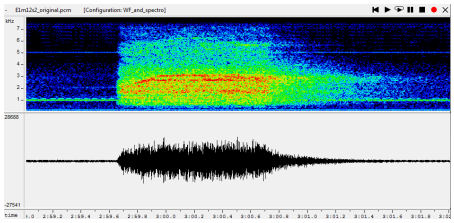


FIGURE A.10 – Evénement anormal enregistré.

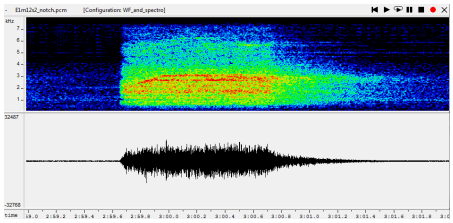


FIGURE A.11 – Evénement anormal sur lequel a été appliquée une combinaison de filtres à encoche.

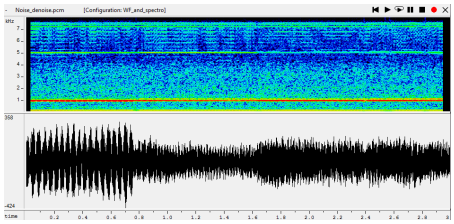


FIGURE A.12 – Séquence de bruit de laquelle a été soustrait le profil de bruit estimé.

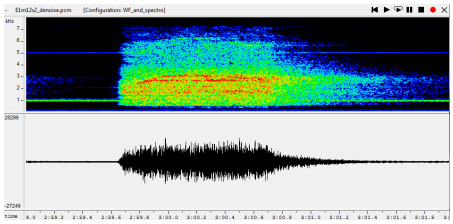


FIGURE A.13 – Evénement anormal duquel a été soustrait le profil de bruit estimé.

Les figures A.12 et A.13 montrent les résultats de cette approche sur les signaux de référence : bruit et événement. Nous avons, durant nos recherches, utilisé différents paramétrages, ces résultats sont les meilleurs obtenus. Bien que cette approche soit particulièrement performante pour atténuer le bruit de ventilation, il apparait que les fréquences dont le niveau sonore est trop élevé ne sont pas traitées.

A.4.3 Combinaison des approches

Nous avons ensuite exploré une approche combinant les deux premières. Notre motivation est de bénéficier des avantages de chacune de ces méthodes. Pour cela, nous appliquons dans un premier temps l'approche par estimation d'un profil de bruit. Le signal ainsi obtenu permet de définir avec précision les fréquences à supprimer par le banc de filtres à encoche. La figure A.14 montre les résultats de cette approche combinée sur le signal de bruit de référence, et la figure A.15 sur l'événement enregistré.

La combinaison de ces approches apporte une amélioration significative. En particulier, nous avons estimé un gain de $18,88$dB (ITU-R468) du rapport événement à bruit (REB). La figure A.16 présente les résultats actualisés de la figure A.7, après application du débruitage.

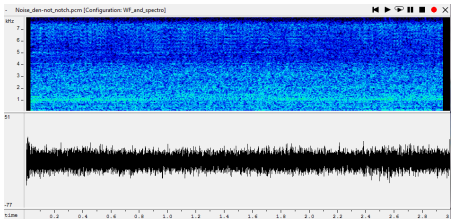


FIGURE A.14 – Séquence de bruit débruitée par la combinaison des approches.

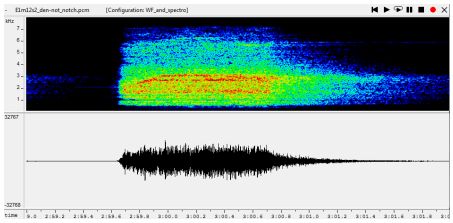


FIGURE A.15 – Evénement anormal débruité par la combinaison des approches.

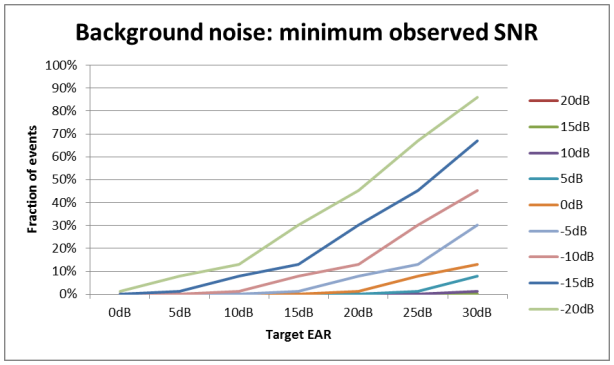


FIGURE A.16 – Fraction des événements débruités dont le RBA est supérieur à une valeur donnée (de −20dB à +30dB) à différents niveaux de REA cibles (0dB à 30dB).

A.5 Perspectives

La mission réalisée à Turin ainsi que les travaux d'analyse qui ont suivi ont permis d'améliorer notre connaissance d'un exemple d'environnement réel dans lequel déployer un système de surveillance basé sur la modalité audio. Un protocole expérimental a été établi afin de collecter des signaux spécifiques dans le but de simuler des signaux d'évaluation plus réalistes. Les conditions réelles n'ayant pas permis d'exploiter ces signaux comme prévu initialement, ceux-ci étant trop fortement bruités. Nous avons donc étudié différentes méthodes de débruitage qui ont permis d'améliorer considérablement la qualité de ces signaux. Ceux-ci pourront donc être exploités à l'avenir.

L'objectif principal étant la réduction de la différence entre les conditions réelles et les signaux d'apprentissage, d'autres approches peuvent également être étudiées. Nous avons évoqué l'estimation des caractéristiques acoustiques de l'environnement à partir de signaux enregistrés. Il est également possible d'exploiter d'autres méthodes provenant du domaine de la reconnaissance de parole ; en effet, la grande variété de canaux de transmission et de capteurs a confronté cette communauté de chercheurs à la problématique de l'adaptation des signaux d'entrainement.

Annexe B

Démonstrateur VANAHEIM

Dans le cadre du projet VANAHEIM, un démonstrateur implémentant les travaux réalisés dans cette thèse a été réalisé. Celui-ci a été présenté à Turin lors de l'EXPO-Ferroviera en mars 2012 et a été présenté à Paris en septembre 2013 au cours d'une démonstration en conditions réelles pour la clôture du projet.

B.1 Présentation du système d'intégration VAIF-AVAS

Le système d'intégration d'outils analytiques vidéo VAIF (*Video Analytics Integration Framework*) est développé par les équipes du domaine Sécurité et Transport (Thales Italia S.P.A.). L'API (*Application Programming Interface*) du système d'analyse audio/vidéo AVAS (*Audio/Video Analysis System*) s'intègre au sein du VAIF. Cette API a pour objectif de simplifier les développements de modules analytiques pour la vidéo et l'audio, mettant à disposition un ensemble de fonctions de base pour l'acquisition et le décodage des signaux, la gestion des modules et la communication entre modules ou vers l'extérieur d'AVAS.

Le VAIF dans son ensemble inclut les éléments de contrôle et d'interface AVAS (gestionnaire de service, machine à état, bus de messages, interfaces de commande, d'événements ou de méta-données), les modules analytiques (audio, vidéo, méta-données), un système d'enregistrement, un mur d'image et une interface utilisateur. Ce système peut fonctionner de manière distribuée sur plusieurs machines partageant un même gestionnaire (partage des ressources).

Le mur d'image (*video-wall*) est une application logicielle qui permet de visualiser les flux vidéos en direct ou enregistrés, écouter les flux audio, montrer des images fixes et afficher des informations en surimpression (typiquement issues des modules d'analyse). Il est composé d'un ensemble d'écrans virtuels indépendants et configurables. Une interface homme-machine simple, sous forme d'application web, est fournie pour les besoins des tests. Elle permet de contrôler l'essentiel des éléments du VAIF, de visualiser son état, les événements, les alarmes et de gérer les modules analytiques.

La couche d'intégration AVAS comprend des composants logiciels bas-niveau pour le développement des modules analytiques audio et vidéo. Ces composants sont mis à disposition sous forme de librairies. Ils abstraient les interfaces bas-niveau vers les ressources externes (vidéo, audio, méta-données), rendent opérables les modules depuis la couche de contrôle et fournissent des fonctions de base, notamment pour la mise en œuvre de traces. AVAS supporte entre autre de manière transparente l'acquisition et le décodage de flux vidéo MPEG-4 sur RTP, MPEG-4/H.264 sur RTSP, fichiers AVI,... et de flux audio AAC sur RTSP. Enfin des interfaces sont mises à disposition pour créer des événements, des méta-données (entiers, flottants, chaines de

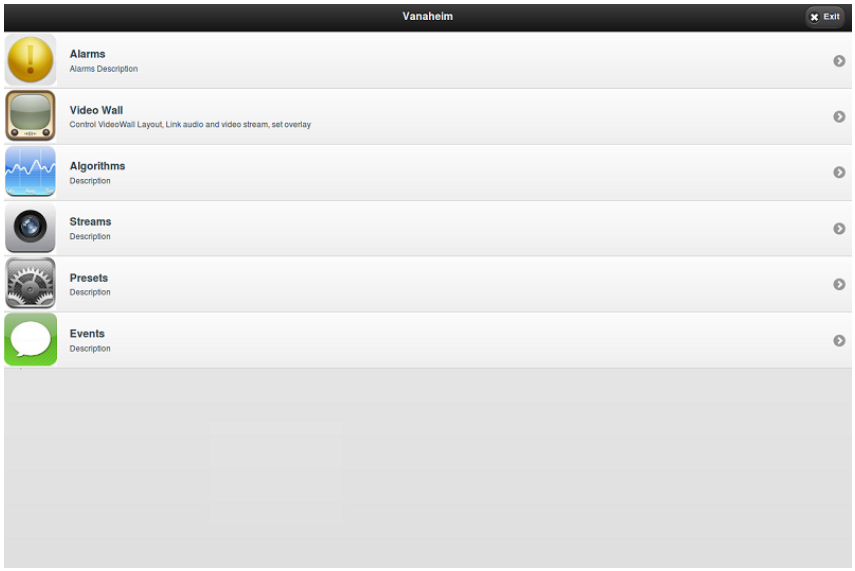


FIGURE B.1 – Interface utilisateur VAIF-AVAS

caractères, listes, etc.) ou des objets à afficher (cercles, lignes, polygones, texte, etc.). Enfin, des interfaces externes sont également mises à disposition pour contrôler le VAIF ou AVAS depuis des programmes tiers (intégration système).

B.2 Lien avec le projet VANAHEIM

VANAHEIM (*Video/Audio Networked surveillance system enhAncement through Human-cEntered adaptIve Monitoring*) est un projet financé par la Communauté Européenne [28] qui vise l'étude et l'intégration d'outils d'analyse audio et vidéo innovants sur des plates-formes de vidéo-surveillance typiquement utilisées dans les environnements urbains (stations de métro, centres commerciaux, etc). L'objectif du projet est d'étudier des composants pour la supervision autonome de ces infrastructures complexes. En particulier, il rassemble huit partenaires (instituts de recherche, industriels, opérateurs publics) de six pays différents et aux compétences complémentaires :

- vision par ordinateur et analyse audio,
- conception de systèmes de surveillance,
- opérateurs de transports publics,
- comportement humain (ethnologistes).

Thales Italia S.P.A. a été impliqué dans le projet VANAHEIM au titre d'intégrateur. Ainsi l'ensemble des algorithmes d'analyse automatique ont pu être portés dans AVAS pour être testés dans des conditions réelles. En particulier, les capacités de sélection automatique de caméra basé sur l'analyse des signaux de surveillance vidéo comme audio ont pu être démontrées au travers d'un démonstrateur. Celui-ci a été mis en place sur deux sites tests : station *XVIII diciembre* à Turin (GTT) et station *Bibliothèque François Mitterrand* à Paris (RATP).

28. VANAHEIM est un projet collaboratif financé par le septième programme cadre de la Communauté Européenne FP7/2007-2013 - Challenge 2- Cognitive Systems, Interaction, Robotics - under grant agreement n° 248907.

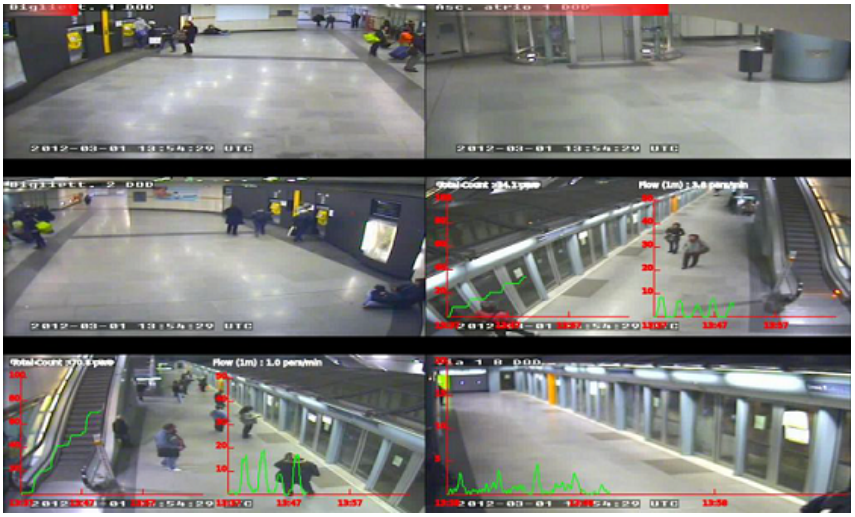


FIGURE B.2 – Interface utilisateur VAIF-AVAS

B.3 Implication de Thales Communications & Security

Thales Communications & Security a eu la charge, entre autre, au sein du projet VANAHEIM de réaliser les recherches en lien avec l'analyse audio. Trois modules ont ainsi été réalisés et intégrés au démonstrateur.

B.3.1 Détection d'événements anormaux

Le module de détection d'événements anormaux a pour objectif la détection de signaux audio inhabituels qui doivent dès lors être considérés comme anormaux. Ce module utilise des modèles entraînés de façon non supervisée sur des signaux enregistrés. L'algorithme d'apprentissage utilisé est une machine à vecteurs de support (SVM) 1-classe développé spécifiquement, bénéficiant d'une complexité réduite et de capacité de mise à jour en ligne. Le module extraits en ligne des descripteurs acoustiques depuis le flux audio en direct, par trames de quelques dizaines de milli-secondes. Dans le même temps, le module réalise une segmentation automatique du flux. Les résultats de détection d'anormalité, calculés pour chaque trame étant donné un modèle, sont ensuite intégrés par segments. Le résultat ainsi obtenu peut être affiché sous forme de jauge en surimpression d'une vidéo ou transmis sous forme de méta-données à un autre module tel que le module de sélection de flux.

Ce module exploite des modèles appris à l'aide de l'algorithme SVM 1-classe avec biais proposé. Les performances obtenues en conditions réelles et lors des démonstrations *live* du projet ont apporté satisfaction. Fort de ces résultats, nous avons pu d'une part étudier en détail le comportement de détection sur des situations réelles, et d'autre part bénéficier d'une vitrine pour l'approche défendue dans cette thèse. Cette dernière, nous l'espérons, permettra d'envisager un possible déploiement de l'analyse audio dans le cadre de systèmes de surveillance fournis par Thales.

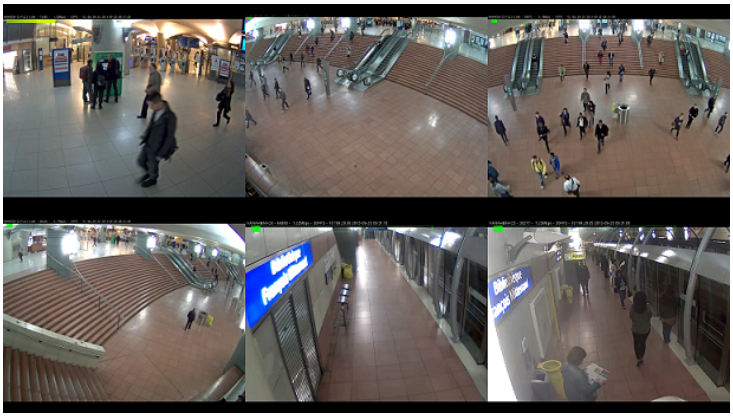


FIGURE B.3 – Détection d'énévements anormaux au sein du VAIF AVAS. Les écrans affichent l'image des caméras les plus proches pour chacun des 6 micros en place sur le site test « Bibliothèque François Mitterrand » (RATP, Paris). Le score d'anormalité est représenté par une jauge en haut de chaque écran. Une jeune femme court avec des talons à droite de la première scène, faisant réagir le détecteur du micro correspondant.

B.3.2 Sélection de flux

Le module de sélection de flux réalisé par TCS a pour objet d'identifier un flux vidéo ou audio dans une zone sous surveillance en s'appuyant sur les niveaux d'anormalité. Le module précédemment décrit peut fournir cette information. Cependant, comme les modèles de chaque capteur sont indépendants, la dynamique de la décision et la variance d'estimation du niveau d'anormalité sont décorrélées. Le module réalisé projette alors les scores sur une distribution statistique pré-selectionnée qui peut être soit normale, soit empiriquement extraite à partir de l'un des capteurs en compétition. L'apprentissage de cette projection et de sa distribution sont réalisées hors ligne sur des signaux enregistrés. En fonctionnement, le module retourne le flux correspondant au capteur dont le score projeté présente le niveau le plus élevé. Cette sortie peut se traduire par l'affichage du nom du flux sélectionné, ou la transmission d'une méta-donnée.

B.3.3 Audio situational awareness

Le module d'attention (localisation d'activité) exploite une analyse statistique du signal audio. Une caractérisation court-terme du signal est comparée à une caractérisation long-terme. Un niveau de « surprise audio » est obtenu en mesurant la distance statistique entre les deux distributions extraites. Les études ont montré que cette mesure est corrélée aux changements d'activité dans l'environnement surveillé. Par le choix d'une méthode d'intégration temporelle appropriée, le module renvoie alors une mesure de l'activité autour de chaque capteur audio. Le module retourne finalement une méta-donnée contenant l'état d'activité de chaque microphone, laquelle est ensuite interprétée par un module dédié à l'affichage du niveau d'activité autour des différents capteurs (vidéo ou audio) dans la station.

Annexe C

Liste des événements anormaux utilisés

Les signaux d'évaluation générés dans le cadre de cette thèse exploitent pour la partie test des événements anormaux. Ceci permet de qualifier les performances des systèmes proposés. En particulier, l'outil de construction de ces signaux contenant des événements anormaux est décrit au chapitre 5. Nous dressons dans cette annexe la liste des événements utilisés. Il est à noter que nous avons extrait la seconde la plus énergétique de l'événement pour nos tests. Enfin, ces événements sont tous extraits d'une base de données commerciale destinée aux effets spéciaux, *The Series 6000 "The General" Sound Effect Library*, éditée par *Sound Ideas* [Sou12].

Le choix de ces événements a été motivé par la volonté de couvrir une grande variété de conditions différentes : événements impulsionnels ou stationnaires, richesse fréquentielle dans différentes bandes spectrales, etc. Le réalisme de la présence de ces événements dans une station de métro importe peu, l'ensemble du système de détection étant non supervisé. Néanmoins cette variété nous a permis d'identifier et caractériser les événements les plus difficiles à détecter par notre système afin de l'adapter (choix des paramètres de modèles et des descripteurs acoustiques). Enfin rappelons que la notion d'anormalité est très relative. Il s'agit d'événements pouvant être distingués de l'ambiance, par exemple la présence d'enfants. Cette information bien qu'elle ne reflète pas de risque direct est importante pour un opérateur. Notons enfin que certains événements sont extraits d'ambiances vues comme anormales dans le cas de notre étude (mouvements de foule, tremblement de terre, etc.).

Les événements, répartis suivant 23 catégories, sont :

- Applaudissements : 7 événements
- Chutes : 3 événements
- Acclamations : 10 événements
- Enfants : 4 événements
- Mouvements de foule : 9 événements
- Aboiements : 4 événements
- Claquements de portes : 5 événements
- Tremblement de terre : 6 événements
- Travaux : 4 événements
- Explosions : 4 événements
- Incendies : 10 événements
- Feux d'artifices : 3 événements
- Pas : 3 événements

- Bris de verre : 5 événements
- Arme à feu : 1 événement
- Marteaux : 8 événements
- Pleurs : 4 événements
- Musique de fête : 4 événements
- Cris : 4 événements
- Sirènes : 3 événements
- Téléphones : 4 événements
- Bagarres : 4 événements
- Craquements de bois : 9 événements

Annexe D

Segmentation en ligne

Une des difficultés évidentes lors du traitement en ligne de signaux audio (pour la détection ou la classification) réside dans la diversité des propriétés des sons étudiés, particulièrement dans le cas de la surveillance. Dans ce contexte, un prétraitement de segmentation du signal permet d'identifier des portions cohérentes du signal. Nous présentons dans cette annexe la méthode proposée par Thales [CR10], telle qu'elle a été présentée dans [Tha11]. Cette approche a été étudiée pour n'exploiter aucune information forte *a priori* sur le signal audio ; cette méthode dépend uniquement des paramètres audio utilisés comme entrée.

A l'instar de [CVR05], la plupart des algorithmes de segmentation audio proposés dans la littérature s'appuient sur des critères d'information tels que BIC (*Bayesian Information Criterion*). Malheureusement, il est difficile de mettre en œuvre un module de segmentation doté de bonne capacité de généralisation lorsque des types d'événements variés sont considérés. Ainsi, la solution proposée est basée sur une approche de segmentation multi-niveau par mesure de similarité entre trames telle que suggérée dans le domaine de la reconnaissance de parole [GZ88, HL96]. L'objectif est de structurer le signal audio en segments successifs présentant des segments spectralement similaires, d'après une analyse trame par trame.

La détermination des segments acoustiques est réalisée en ligne en utilisant un *buffer* de vecteurs (paramètres acoustiques). Lorsque le signal analysé est de la parole, ce *buffer* est au moins de la taille de la plus longue unité de parole (phonème) attendue, soit 200 à 300 millisecondes. Dans le contexte de l'audio pour la surveillance, aucune hypothèse n'existe concernant la durée des segments attendus, bien que ceux-ci puissent être significativement plus long qu'un phonème. Il est alors souhaitable d'accroitre la taille du *buffer* pour ces signaux. Néanmoins, le système devant rester réactif, il faut veiller à ne pas retarder la décision. Ainsi, quelques secondes est une taille réaliste pour le *buffer*. Notons que le dernier segment de chaque *buffer* sera considéré comme premier segment du *buffer* suivant (voir figure D.1).

En pratique, des descripteurs spectraux sont extraits pour chaque trame de signal (énergies en sortie d'un banc de filtres). Lorsque le *buffer* défini est complet, le processus de segmentation est réalisé par un algorithme de construction de dendrogramme via un regroupement des trames dit *bottom-up*. Au cours de cette phase, chaque trame est considérée comme un segments, puis les segments sont regroupés par paires suivant un critère de similarité pré-défini. Le plus couramment ce critère est la distance euclidienne entre vecteurs de descripteurs représentatifs de chaque segment (moyenne des trames composant le segment). Une segmentation optimale doit ensuite être extraite du dendrogramme ainsi construit. Si dans le cas de la parole [GZ88], les résultats de transcription déterminent le choix optimal, une approche particulière doit être utilisée pour les signaux audio qui nous intéressent. Le choix de la segmentation optimale s'opère alors par

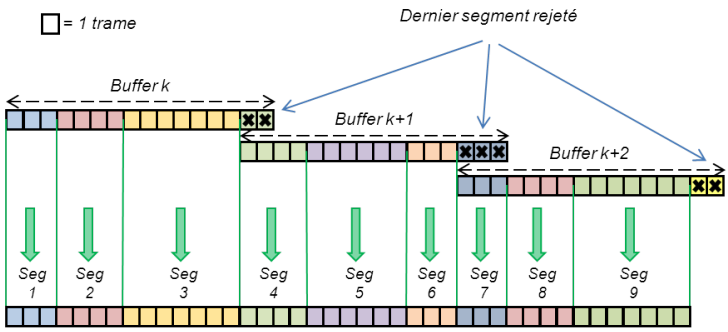


FIGURE D.1 – Principe de *buffer* pour la segmentation en ligne.

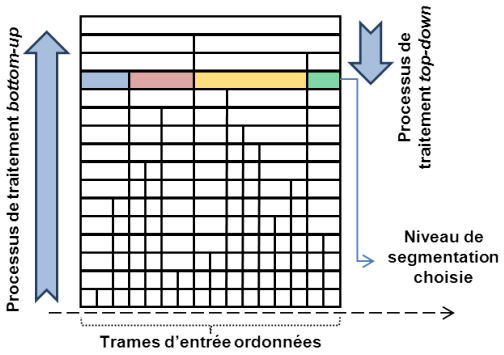


FIGURE D.2 – Segmentation par regroupement hiérarchique (dendrogramme) des trames.

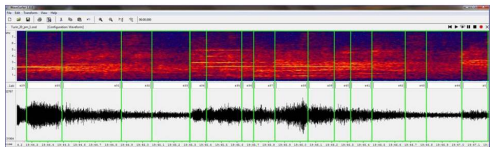


FIGURE D.3 – Exemple de résultat de segmentation.

l'analyse du coefficient de corrélation intra-segment (entre les deux segments qui composent le segment de niveau supérieur). Nous suggérons alors de choisir la segmentation pour laquelle la corrélation intra-segment excède un seuil pré-défini pour tous les segments considérés. Ce principe est illustré par la figure D.2.

La figure D.3 donne un exemple de segmentation obtenu lors de l'analyse d'un signal audio capturé dans la station de métro *XVIII diciembre* à Turin.

Annexe E

Résultats sur la base données CARETAKER

Le projet CARETAKER (*Content Analysis and REtrieval Technologies to Apply Knowledge Extraction to massive Recording*, projet Européen FP6, 2006-2008)a permis de mettre en évidence la possibilité d'exploiter des méthodes non supervisées pour l'analyse d'environnements dans les contextes de la surveillance. Il s'agissait en effet de s'intéresser aux anormalités pouvant être identifiées par l'apprentissage sur des signaux en l'absence de signaux anormaux. En particulier, des modèles par mélanges de gaussiennes (GMM) ont été utilisés pour modéliser l'ambiance normale et démontrer la possibilité de détecter des événements sonores anormaux. Cette approche avait notamment fait l'objet d'un brevet déposé par THALES [CR10].

Au cours de ce projet, une base de données de signaux audio acquis dans une station de métro a été constituée à des fins d'évaluation. Celle-ci constituait la seule base sur laquelle travailler au début des travaux de thèse. Ainsi, elle a servi à l'évaluation des modèles SVM 1-classe initialement proposés. De plus, l'utilisation d'une base commune a permis de comparer les approches GMM et SVM. Nous montrons dans cette annexe que cette seconde approche est bien plus performante. Ainsi, pour une même base de signaux, la figure E.1 illustre les résultats obtenus par l'approche GMM (projet Caretaker) et la figure E.2 illustre ceux obtenus par l'approche SVM 1-classe (travaux de cette thèse). Notons que le protocole expérimental est le même que celui présenté au chapitre 5.

Les gains en termes de performance de détection vont jusqu'à 7% bruts et sont de l'ordre de 40 à 50% relatifs. Le tableau E.1 rapporte les gains observés aux points de fonctionnement EER des détecteurs pour des conditions d'événements anormaux de RSB similaire.

RSB	EER GMM	EER SVM	Gain relatif
10dB	16,59 %	9,69 %	41 %
15dB	10,25 %	5,22 %	49 %
20dB	5,96 %	3,22 %	46 %
25dB	3,37 %	2,09 %	38 %

TABLE E.1 – Gains en termes de performances EER de l'approche SVM par rapport à l'approche GMM à différents niveaux de RSB

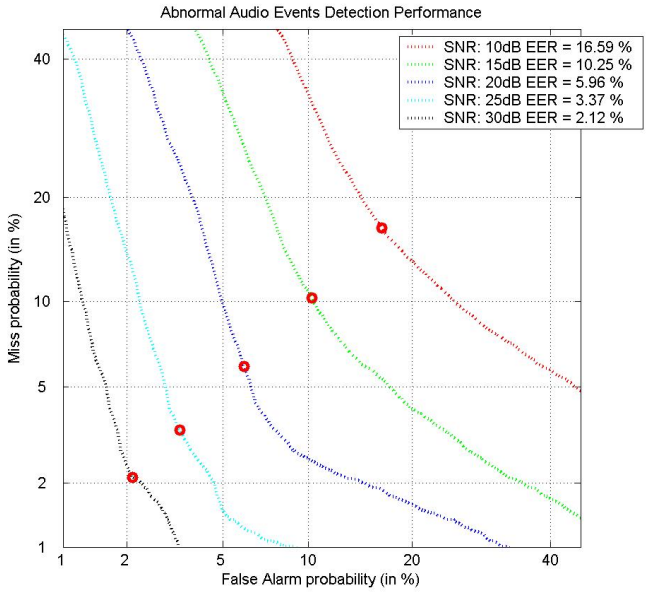


FIGURE E.1 – Performances des détecteurs obtenus sur la base de données CARETAKER par une approche GMM pour des événements anormaux à différents niveaux de RSB.

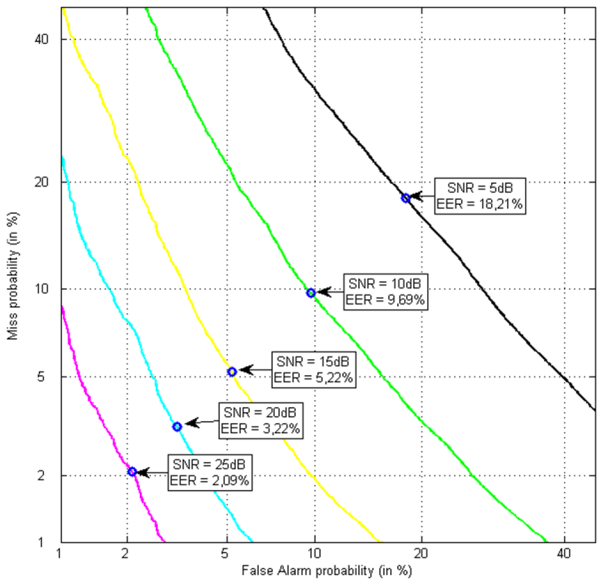


FIGURE E.2 – Performances des détecteurs obtenus sur la base de données CARETAKER par une approche SVM 1-classe pour des événements anormaux à différents niveaux de RSB.

Bibliographie

[ABR64] Mark A. Aizerman, Emmanuil M. Braverman et Lev I. Rozonoer : Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[AEQG03] Ahmad R. Abu-El-Quran et Rafik A. Goubran : Pitch-based feature extraction for audio classification. *In Proceedings of Haptic IEEE Internatioal Workshop on Audio and Visual Environments and Their Applications (HAVE)*, 2003.

[AMK06] Pradeep Atrey, Namunu Maddage et Mohan Kankanhalli : Audio based event detection for multimodal surveillance. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5:V, 2006.

[AP03] Jean-Julien Aucouturier et François Pachet : Representing musical genre : a state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[Aro50] Nachman Aronszajn : Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[Bar05] Yoram Baram : Learning by kernel polarization. *Neural Computation*, 17(6): 1264–1275, 2005.

[Bel61] Richard Bellman : *Adaptive Control Processes : a Guided Tour*. Princeton University Press, New Jersey, 1961.

[Ber22] Stefan Bergman : Über die entwicklung der harmonischen funktionen der ebene und des raumes nach orthogonalfunktionen. *Matematische Annalen*, 86:238–271, 1922.

[BGG09] Michael Breitenstein, Helmut Grabner et Luc Van Gool : Hunting nessie - real-time abnormality detection from webcams. *In International Conference on Computer Vision*, 2009.

[BGV92] Bernhard Boser, Isabelle Guyon et Vladimir Vapnik : A training algorithm for optimal margin classifiers. *In COLT'92 : Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[BHHSV01] Asa Ben-Hur, David Horn, Hava T. Siegelmann et Vladimir Vapnik : Support vector clustering. *Journal of machine learning research*, 2:125–137, 2001.

[BL03] Juan Rosé Burred et Alexander Lerch : A hierarchical approach to automatic musical genre classification. *In Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03)*, 2003.

[BLLC06] Liang Bai, Song-Yang Lao, Hu-Xiong Liao et Jian-Yun Chen : Audio classification and segementation for sports video structure extraction using support vector machine. *In Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, 2006.

[BM92] Kristin Bennett et Olvi Mangasarian : Robust linear programming discrimination of two linearly inseparable set. *Optimization Methods and Software*, 1:23–34, 1992.

[Bre90] Al Bregman : *Auditory Scene Analysis : The Perceptual Organization of Sound*. Bradford Books, MIT Press, 1990.

[Bri68] British Broadcasting Corporation : *The Assessment of Noise in Audio Frequency-Circuits - EL17*. Engineering Division Research Report, 1968.

[CAR08] CARETAKER Project : Content Analysis REtrieval Technologies to Apply Knowledge Extraction to massive Recording : *FP6 IST 4-027231*. 2006-2008.

[CDR+06] Cyril Carincotte, X. Desurmont, Bertrand Ravera, François Brémond, J. Orwell, S.A. Velastin, Jean-Marc Odobez, B. Corbucci, J. Palo et J. Cernocky : Toward generic intelligent knowledge extraction from video and audio : The eu-funded caretaker project. *The Institution of Engineering and Technology Conference on CRIME AND SECURITY, Imaging for Crime Detection and Prevention (ICDP)*, pages 470–475, 2006.

[CDR+07] Chloé Clavel, Laurence Devillers, Gaël Richard, Iona Vasilescu et Thibault Ehrette : Detection and analysis of abnormal situations through fear-type acoustic manifestations. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[CE08] Chloé Clavel et Thibault Ehrette : *Risk Analysis VI*, chapitre Fear-type emotion recognition and abnormal events detection for an audio-based surveillance system, pages 471–479. WIT Transactions on Information and Communication Technologies, 2008.

[Chi04] Michel Chion : *Le son : traité d'acoulogie*. Armand Colin, 2004.

[Ciu12] Ciudad segura. En ligne, 2012.

[CLH+06] Rui Cai, Lie Lu, Alan Hanjalic, Hong-Jjiang Zhang et Lian-Hong Cai : A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1026–1039, 2006.

[CLZC03] Rui Cai, Lie Lu, Hong-Jiang Zhang et Liun-Hong Cai : Highlight sound effects detection in audio stream. *In Proceedings of International Conference on Multimedia and Expo*, 2003.

[Cov65] Thomas M. Cover : Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.

[CP00] Gert Cauwenberghs et Tomaso Poggio : Incremental and decremental support vector machine learning. *Neural Information Processing Systems*, 2000.

[CR10] François Capman et Bertrand Ravera : Système et méthode pour détecter des événements audio anormaux, 2010.

[CRE05] Chloé Clavel, Gaël Richard et Thibault Ehrette : Events detection for an audio-based surveillance system. *In IEEE International Conference on Multimedia and Expo*, 2005.

[CS01] K. Crammer et Y. Singer : On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Res.*, 2:265–292, 2001.

[CST00] Nello Cristianini et John Shawe-Taylor : *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[CV95] Corinna Cortes et Vladimir Vapnik : Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[CVR05] Mauro Cettolo, Michele Vescovi et Romeo Rizzi : Evaluation of bic-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2):147–170, 2005.

[DDGD05] Manuel Davy, Frédéric Desobry, Arthur Gretton et Christian Doncarli : An online support vector machine for abnormal events detection. *Journal on Signal Processing*, pages 2009–2025, 2005.

[Def05] Boris Defréville : *Caractérsisation de la qualité sonore de l'environnement urbain : une approche physique et perceptive basée sur l'identification des sources sonores*. Thèse de doctorat, Université de Cergy-Pontoise, France, 2005.

[DGI11] Ürün Dogan, Tobias Glasmachers et Christian Igel : Fast training of muli-class support vector machines. Rapport technique, Faculty of Science, University of Copenhagen, 2011.

[DK82] Pierre Devyver et Josef Kittler : *Pattern Recognition : a Statistical Approach*. 1982.

[DL97] M Dash et H Liu : Feature selection for classification. *Intelligent Data Analysis - An International Journal*, 1(3):131–156, 1997.

[DLR77] Arthur Dempster, Nan Laird et Donald Rubin : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[DM99] Lokenath Debnath et Piotr Mikusinski : *Introduction to Hilbert spaces with Applications*. Academic, 2nd édition, 1999.

[DZ07] Qian Ding et Nian Zhang : Classification of recorded musical instruments sounds based on neural networks. *IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP)*, 2007.

[Efr79] Bradley Efron : Bootstrap methods : another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[EP03] André Elisseeff et Massimiliano Pontil : Leave-one-out error and stability of learning algorithms with applications. *Advances in Learning Theory : Methods, Models and Applications*, NATO Science Series III : Computer and Systems Sciences, 190, 2003.

[ET94] Bradley Efron et Robert Tibshirani : *An Introduction to Bootstrap*. Chapman & Hall, 1994.

[Faw04] Tom Fawcett : Roc graphs : Notes and practical considerations for researchers. *ReCALL*, 31:1–38, 2004.

[FCC98] Thilo-Thomas Friess, Nello Cristianini et Colin Campbell : The kernel-adatron algorithm a fast and simple learning procedure for support vector machines. *Proceedings of the Fifteenth International Conference of Machine Learning Research (ICML)*, pages 188–196, 1998.

[FM33] Harvey Fletcher et Wilden A. Munson : Loudness, its definition, measurement and calculation. *Journal of Acoustical Society of America*, 5:82–108, 1933.

[GAAVO08] Luis Gonzalez-Abril, Cecilio Angulo, Francisco Velasco et Juan Antonio Ortega : A note on the bias in svms for multi-classification. *IEEE Transactions on Neural Networks*, 19(4):723–725, 2008.

[GAVC05] Luis Gonzalez, Cecilio Angulo, Francisco Velasco et Andreu Catala : Unified dual for bi-class svm approaches. *Pattern Recognition*, 38(10):1772–1774, 2005.

[GBT09] Gabor Gosztolya, Andras Banhalmi et Laszlo Toth : Using one-class classification techniques in the anti-phoneme problem. *Iberian Conference on Pattern Recognition and Image Analysis*, pages 433–440, 2009.

[GD03] Arthur Gretton et Frédéric Desobry : On-line one-class support vector machines. an application to signal segmentation. *International Conference on Audio, Speech and Signal Processing, IEEE ICASSP*, 2003.

[GE03] Isabelle Guyon et André Elisseeff : An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[Gei12] Matthieu Geist : Précis introductif à l'apprentissage statistique. Rapport technique, Supélec - groupe de recherche IMS, 2011-2012.

[GPK09] Fabian Gieseke, Tapio Pahikkala et Oliver Kramer : Fast evolutionary maximum margin clustering. *In Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 361–368, New York, NY, USA, 2009. ACM.

[GR10] S. Gunasekaran et K. Revathy : Content-based classification and retrieval of wild animal sounds using feature selection algorithm. *Second International Conference on Machine Learning and Computing*, pages 272–275, 2010.

[GZ88] James Glass et Victor Zue : Multi-level acoustic segmentation of continuous speech. *In Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, volume 1, pages 429–432, 1988.

[Hee09] Frédéric Heer : Surveillance des transports par l'analyse de l'image et du son. *In Séminaire de valorisation GO 2 - Gouvernance et transports en commun*, 2009.

[HK04] Te Ming Huang et Vojislav Kecman : Bias term b in svms again. *In Proceedings of European Symposium on Artificial Neural Networks (ESANN)*, pages 441–448, 2004.

[HL96] Jean-Luc Husson et Yves Laprie : A new search algorithm in segmentation lattices of speech signals. *In Proceedings of the 4th International Conf. on Spoken Language Processing*. IEEE Computer Society, 1996.

[HL02] Chih-Wei Hsu et Chih-Jen Lin : A simple decomposition method for support vector machines. *Journal of Machine Learning*, 46(1-3):291–314, 2002.

[HMEV13] Toni Heittola, Annamaria Mesaros, Antti Eronen et Tuomas Virtanen : Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 1, 2013.

[HTF09] Trevor Hastie, Robert Tobshirani et Jerome Friedman : *The Elements of Statistical Learning*. Springer-Verlag, 2nde édition édition, 2009.

[HWYH08] Yang Hu, Jingdong Wang, Nenghai Yu et Xian-Sheng Hua : Maximum margin clustering with pairwise constraints. *In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 253–262, Washington, DC, USA, 2008. IEEE Computer Society.

[ICV+06] Dan Istrate, Eric Castelli, Michel Vacher, Laurent Besacier et Jean-François Serignat : Information extraction from sound for medical telemonitoring. *Information Technology in Biomedicine, IEEE Transactions on*, 10(2):264–274, 2006.

[IEC] Information technology – multimedia content description interface – part 4 : Audio.

[Int90] Internation Telecommunication Union : *Recommandation 468-4 : Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting*. Recommandation, Broadcasting Service, 1990.

[Int03a] International Electrotechnical Commission : *IEC-61672-2 Sound Level Meters - Part 2 : Pattern Evaluation Tests*. 2003.

[Int03b] International Organization for Standardization : *ISO-226 Acoustics - Normal Equal-Loudness-Level Contours*. ISO Standards, 2003.

[Iva76] Viktor Vladimirovich Ivanov : *The theory of approximate methods and their application to numerical solution of singular integral equations*. Noordhoff International, 1976.

[Jai10] Anil K. Jain : Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666, 2010.

[JD88] Anil K. Jain et Richard C. Dubes : *Algorithms for Clustering Data*. Prentice Hall, 1988.

[Joa98] Thorsten Joachims : *Advances in Kernel methods - Support Vector learning*, chapitre Making large scale SVM learning practical, pages 41–56. MIT Press, 1998.

[Ked86] Benjamin Kedem : Spectral analysis and discrimination using zero-crossings. *In Proceedings of the IEEE*, volume 74, 1986.

[KK11] Kwangyoun Kim et Hanseok Ko : Hierarchical approach for abnormal acoustic event classification in an elevator. *In 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2011.

[KKC+09] Jogile Kuklyte, Philip Kelly, Ciaran O Conaire, Noel E. O'Connor et Li-Qun Xu : Anti-social behavior detection in audio-visual surveillance systems. *In Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis*, 2009.

[KMS05] Hyoung-Gook KIM, Nicolas MOREAU et Thomas SIKORA : *MPEG-7 Audio and Beyond : Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

[KS00] Mineichi KUDO et Jack SKLANSKY : Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41, 2000.

[KSBM01] Sathiya KEERTHI, Shirish SHEVADE, Chiranjib BHATTACHARYYA et K.R. Krishna MURTHY : Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13:637–649, 2001.

[KSW01] Jyrki KIVINEN, Alexander SMOLA et Robert WILLIAMSON : Online learning with kernels. *Neural Information Processing Systems*, 2001.

[KSW10] Jyrki KIVINEN, Alexander SMOLA et Robert WILLIAMSON : Learning online with kernels. *IEEE Transactions on Signal Processing*, 100(10), 2010.

[KT51] Harold KUHN et Albert TUCKER : Nonlinear programming. *In* Univ. of CALIF. PRESS, éditeur : *Second Berkeley Symposium on Mathematics Statistics and Proba.*, pages 481–492, 1951.

[KVH03] Vojislav KECMAN, Michael VOGT et Te Ming HUANG : On the equality of kernel adatron and sequential minimal optimization in classification and regression tasks and alike algorithms for kernel machines. *In Proc. of ESANN 2003, 11 th European Symposium on Artificial Neural Networks*, 2003.

[KW71] George KIMELDORF et Grace WAHBA : Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[Lac67] Peter A. LACHENBRUCH : An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 23(4):639–645, 1967.

[LB69] Aleksandr LUNTZ et Viktor BRAILOVSKY : On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3, 1969.

[LB97] Avrim L. LANGLEY et Pat BLUM : Selection of relevant features and examples in machine learning. *Artificial Intelligence Journal, Special Issue on Relevance.*, 97:245–271, 1997.

[LCR+12] Sebastien LECOMTE, François CAPMAN, Bertrand RAVERA, Régis LENGELLÉ et Cédric RICHARD : Procédé et système pour détecter des événements sonores dans un environnement donné, 2012.

[LCTC05] Chien-Chang LIN, Shi-Huang CHEN, Trieu-Kien TRUONG et Yukon CHANG : Audio classification and categorization based on wavelets and support vector machine. *IEEE Transactions on Speech and Audio Processing*, 13(5):644–651, 2005.

[Lec09] Sébastien LECOMTE : Etude et mise en œuvre de méthodes de sélection de paramètres pour la classification d'événements sonores. Mémoire de D.E.A., Université de Technologie de Troyes, 2009.

[Lec12] Sébastien LECOMTE : *Perceptive audio features extraction and multimodal activity modelling (v2)*, chapitre Specific Audio Acquisitions. VANAHEIM - Video/Audio Networked surveillance system enhAncement through HumancEnteredadaptIve Monitoring, 2012.

[Lee06] Sei-Hyung Lee : Cone cluster labeling for support vector clustering. *In In Proceedings of 6th SIAM Conference on Data Mining*, pages 484–488, 2006.

[LGKM06] Pavel Laskov, Christian Gehl, Stefan Krüger et Klaus-Robert Müller : Incremental support vector learning : Analysis, implementation and applications. *Journal of Machine Learning Research*, 7:1909–1936, 2006.

[LL05] Jaewook Lee et Daewon Lee : An improved cluster labeling method for support vector clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:461–464, 2005.

[LLR+11] Sébastien Lecomte, Régis Lengellé, Cédric Richard, François Capman et Bertrand Ravera : Abnormal events detection using unsupervised one-class svm - application to audio surveillance and evaluation. *8th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 124–129, 2011.

[LLR+13] Sébastien Lecomte, Régis Lengellé, Cédric Richard, François Capman et Bertrand Ravera : One-class svm sans biais. *Colloque GRETSI*, 2013.

[LM08] Huan Liu et Hiroshi Motoda : *Computational Methods of Feature Selection - Data Mining and Knowledge Discovery Series.* Taylor & Francis Group, LLC, 2008.

[LSDM01] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova et Tom McGee : Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[LTKZ09] Yu-Feng Li, Ivor W. Tsang, James T. Kwok et Zhi-Hua Zhou : Tighter and convex maximum margin clustering. *In 12th Int. Conf. on Aratificial Intelligence and STATisticS*, 2009.

[Lux07] Ulrike Von Luxburg : A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[LZJ02] Lie Lu, Hong-Jiang Zhang et Hao Jiang : Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7), 2002.

[MDK+97] A G Martin, G Doddington, T Kamm, M Ordowski et M Przybicki : The det curve in assessment of detection task performance. *EUROSPEECH 1997*, pages 1895–1898, 1997.

[Mer09] James Mercer : Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, 209:415–446, 1909.

[MH96] Mary Moya et Don Hush : Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[MM99] Olvi Mangasarian et David Musicant : Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10(5): 1032–1037, 1999.

[Moo16] Eliakim H. Moore : On properly positive hermitianmatrices. *Bulletin of American Mathematical Society*, 23:59, 1916.

[MRW+99] Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf et Klaus-Robert Müller : Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing*, pages 41–48, 1999.

[MSI08] MSI ÉTUDES : Le marché de la vidéosurveillance en france. en ligne, Avril 2008.

[Muk04] Sayan MUKHERJEE : Statistical learning : Algorithms and theory. Course notes for STA270, 2004.

[Nad64] E. A. NADARAYA : On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

[NB01] Jan NOYES et Matthew BRANSKY : *People in control : Human factors in control room design*. Institution of Engineering and Technology, 2001.

[NC07] Songyot NAKARIYAKUL et David P. CASASENT : Adaptive branch and bound algorithm for selecting optimal features. *Pattern Recognition Letters*, 28: 1415–1427, 2007.

[NS05] J. Saketha NATH et Shirish Krishnaj SHEVADE : An efficient clustering scheme using support vector methods. *In IICAI*, pages 3436–3448, 2005.

[O'S98] Douglas O'SHAUGHNESSY : Linear predictive coding. *IEEE Potentials*, 7(1): 29–32, 1998.

[Par62] Emanuel PARZEN : Extraction and detection problems and reproducing kernel hilbert spaces. *Journal of Society for Industrial and Applied Mathematics*, 1:35–62, 1962.

[PB06] Alexandre PRETI et Jean-François BONASTRE : Unsupervised adaptation for speaker verification. *Interspeech*, 2006.

[Pee04] Geoffroy PEETERS : A large set of audio features for sound description (similarity and classification) in the cuidado project. Rapport technique, IRCAM, 2004.

[PFNJ94] P. PUDIL, F.J. FERRI, J. NOVOVICOVA et Kittler J. : Floating search methods for feature selection with nonmonotonic criterion functions. *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 2:279–283, 1994.

[PG90] Tomaso POGGIO et Frederico GIROSI : Networks for approxomation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[Phi62] David PHILLIPS : A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Association for Computing Machinery*, 9(1):84–97, 1962.

[Pla98] John PLATT : *Advances in Kernel Methods - Support Vector Learning*, chapitre Sequential Minimal Optimization : a Fast Algorithm for Training Support Vector Machines. MIT Press, 1998.

[Pla99] John PLATT : *Advances in Large Margin Classifiers*, chapitre Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press, 1999.

[PLB+10] Quoc-Cuong PHAM, Agnès LAPEYRONNIE, Christelle BAUDRY, Laurent LUCAT, Patrick SAYD, Sébastien AMBELLOUIS, David SODOYER, Amaury FLANCQUART, Alain-Claude BARCELO, Frédéric HEER, Fabrice GANANSIA et Vincent DELCOURT : Audio-video surveillance system for public transportation. *Image Processing Theory, Tools and Applications*, pages 47–53, 2010.

[PLS01] Le monde des sons - dossier hors-série pour la science, Juil./Oct 2001.

[Pot07] Jean-Baptiste Pothin : *Décision par méthodes à noyaux en traitement du signal - Techniques de sélection et d'élaboration de noyaux adaptés*. Thèse de doctorat, Université de Technologie de Troyes, 2007.

[PPaIsfSMbTA03] Institutional PRISMATICA PRo-active Integrated systems for Security Management by Technological et Communication Assistance : D13 - key findings and results. Rapport technique, EC project under the Transport R&D programme of the Fifth Framework Programme, 2003.

[Que49] Maurice Quenouille : Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B (Methodological)*, 11(1):68–84, 1949.

[Ray77] John William Strutt Rayleigh : *Theory of Sound*. 1877.

[RCL11] Bertrand Ravera, François Capman et Sébastien Lecomte : Système et procédé non supervisé d'analyse et de structuration thématique multi-résolution de flux audio, 2011.

[RD56] D. W. Robinson et R. S. Dadson : A re-determination of the equal-loudness relations for pure tones. *Journal of Applied Physics*, 7(5):166–181, 1956.

[RDR+07] Asma Rabaoui, Manuel Davy, Stéphane Rossignol, Zied Lachiri et Noureddine Ellouze : Improved one-class svm classifier for sounds classification. *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 117 –122, 2007.

[RLA06] Jean-Luc Rouas, Jérôme Louradour et Sébastien Ambellouis : Audio events detection in public transport vehicle. *IEEE Intelligent Transportation Systems Conference*, pages 17–20, 2006.

[SAH07] Seyed Omed Sadjadi, Seyed Mohammad Ahadi et Oldooz Hazrati : Unsupervised speech music classification using one-class support vector machines. *International Conference on Information, Communications & Signal Processing*, pages 1–5, 2007.

[SBV95] Bernhard Schölkopf, Chris Burges et Vladimir Vapnik : Extracting support data for a given task. *In* AAAI Press, éditeur : *First International Conference on Knowledge Discovery & Data mining*, pages 252–257, 1995.

[SHS11] Ingo Steinwart, Don Hush et Clint Scovel : Training svms without offset. *Journal of Machine Learning Research*, 12:141–202, 2011.

[Smi68] Fred Smith : Pattern classifier design by linear programming. *IEEE Transactions on Computers*, C-17(4):367–372, 1968.

[Sou12] Sound Ideas : The series 6000 "the general" sound effect library. en ligne, http ://www.sound-ideas.com/, 2012.

[SP04] Petr Somol et Pavel Pudil : Fast branch & bound algorithms for optimal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):900–912, 2004.

[SPST+01] Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola et Robert Williamson : Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13:1443–1471, 2001.

[SS97] Eric Scheirer et Malcolm Slaney : Construction and evaluation of a robust multi-feature speech/music discriminator. *In IEEE int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, 1997.

[SS02] Bernhard Schölkopf et Alexander Smola : *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*. Massachusetts Institute of Technology, 2002.

[SSM97] Bernhard Schölkopf, Alexander Smola et Klaus-Robert Müller : Kernel principal component analysis. *Artificial Neural Networks*, 1327:583–588, 1997.

[STC04] John Shawe-Taylor et Nello Cristianini : *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[Ste03] Ingo Steinwart : Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

[TA77] Andreï Tikhonov et Vasiliy Arsenin : *Solutions of Ill Posed Problems*. John wiley, 1977.

[Tax01] David Tax : *One-class classification*. Thèse de doctorat, Technische Universiteit Delft, 2001.

[TC99] Georges Tzanetakis et Perry Cook : Multi-feature audio segmentation for browsing and annotation. *In Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

[TC02] Georges Tzanetakis et Perry Cook : Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[TD99] David Tax et Robert Duin : Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199, 1999.

[TD04] David Tax et Robert Duin : Support vector data description. *Machine Learning*, 54:45–66, 2004.

[TEC01] Georges Tzanetakis, Georg Essl et Perry Cook : Automatic musical genre classification of audio signals. *In Proc. of ISMIR'2001*, 2001.

[Tem07] Andriy Temko : *Acoustic Event Detection and Classification*. Thèse de doctorat, Universitat Politècnica de Catalunya, 2007.

[Tha11] Thales Communications & Security : First report on audio features extraction and multimodal activity modelling (v1). Rapport technique, VANAHEIM - Video/Audio Networked surveillance system enhAncement through Human-cEnteredadaptIve Monitoring, 2011.

[Tik63] Andreï Tikhonov : Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady*, 4:1035–1038, 1963.

[TK09] Sergios Theodoridis et Konstantinos Koutroumbas : *Pattern Recognition*. Academic Press, fourth edition édition, 2009.

[TL09] Mireille Tohmé et Régis Lengellé : Sequential maximum gradient optimization for support vector detection. *In 17th European Signal Processing Conference (EUSIPCO)*, 2009.

[TL11] Mireille Tohmé et Régis Lengellé : Maximum margin one class support vector machines for multiclass problems. *Pattern Recognition Letters*, 32: 1652–1658, 2011.

[TMK12] Akiko Takeda, Hiroyuki Mitsugi et Takafumi Kanmori : A unified robust classification model. *In 29th Intl. Conf. on Machine Learning*, 2012.

[Toh09] Mireille Tohmé : *Tests multivariés de comparaison de groupes basés Reconnaissance des Formes : Application à la pharmaco-EEG*. Thèse de doctorat, Université de Technologie de Troyes, 2009.

[TS99] Michael Talbot-Smith : *Audio Engineer's Reference Book*. Focal Press, 2nd édition, 1999.

[Vap95] Vladimir Vapnik : *The nature of statistical learning theory*. Springer-Verlag, second edition édition, 1995.

[VBD+06] Van-Thinh Vu, François Brémond, Gabriele Davini, Monique Thonnat, Quoc-Cuong Pham, Nicolas Allezard, Patrick Sayd, Jean-Luc Rouas, Sébastien Ambellouis et Amaury Flancquart : Audio-video event recognition system for public transport security. *In IEEE International conference on Crime Detection and Prevention (ICDP'2006)*, 2006.

[VC71] Vladimir Vapnik et Alexey Chervonenkis : On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[VGT+07] Giuseppe Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci et A. Sarti : Scream and gunshot detection and localization for audio-surveillance systems. *In Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26. IEEE Computer Society, 2007.

[VIB+04] Michel Vacher, Dan Istrate, Laurent Besacier, Jean-François Serignat et Eric Castelli : Sound detection and classification for medical telesurvey. *IASTED Biomedical Conference*, pages 395–399, 2004.

[VJ07] Hamed Valizadegan et Rong Jin : Generalized maximum margin clustering and unsupervised kernel learning. *In* B. Schölkopf, J. Platt et T. Hoffman, éditeurs : *Advances in Neural Information Processing Systems 19*, pages 1417–1424. MIT Press, Cambridge, MA, 2007.

[VL63] Vladimir Vapnik et A. Lerner : Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.

[Vla08] J. Vlahos : Welcome to the planopticon. *Popular Mechanics*, 1(1):64–69, 2008.

[VV05] Maria Valera et Sergio Velastin : Intelligent distributed surveillance systems : a review. *Vision, Image and Signal Processing, IEE Proceedings -*, 152(2):192–204, 2005.

[Wah90] Grace Wahba : *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[Wat64] Geoffrey S. Watson : Smooth regression analy. *The Indian Journal of Statistics*, 26(4):359–372, 1964.

[Web04] Andrew Webb : *Statistical Pattern Recognition*. Wiley, 2004.

[WLH00] Yao Wang, Zhu Liu et Jin-Cheng Huang : Multimedia content analysis - using both audio and visual clues. *IEEE Signal Processing Magazine*, pages 12–36, 2000.

[WtBKW96] Erling Wold, thom Blum, Douglas Keislar et James Wheaton : Content-based classification, search, and rretrieval of audio. *Journal of IEEE Multi-Media*, 3(3):27–36, 1996.

[WTM+07] Gordon Wichern, Harvey Thornburg, Brandon Mechtley, Alex Fink, Kai Tu et Andreas Spanias : Robust multi-feature segmentation and indexing for natural sound environments. *IEEE Workshop on Content-Based and Multimedia Indexing*, pages 69–76, 2007.

[WWL09] Fei Wang, Xin Wang et Tao Li : Maximum margin clustering on data manifolds. *In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 1028–1033, Washington, DC, USA, 2009. IEEE Computer Society.

[WZZ10] Fei Wang, Bin Zhao et Changshui Zhang : Linear time maximum margin clustering. *Neural Networks, IEEE Transactions on*, 21(2):319–332, 2010.

[XNLS04] Linli Xu, James Neufeld, Bryce Larson et Dale Schuurmans : Maximum margin clustering. *In Advances in Neural Information Processing Systems 17*, pages 1537–1544. MIT Press, 2004.

[XS05] Linli Xu et Dale Schuurmans : Unsupervised and semi-supervised multi-class support vector machines. *In AAAI-05, The Twentieth National Conference on Artificial Intelligence*, pages 904–910, 2005.

[Yan09] Zhenke Yang : *Multi-Modal Aggression Detection in Trains*. Thèse de doctorat, Technische Universiteit Delft, 2009.

[Yao67] K. Yao : Applications of reproducing kernel hilbert spaces - band limited signal models. *Journal of Information and Control*, 11(4):429–444, 1967.

[YECC02] Jianhua Yang, Vladimir Estivill-Castro et Stephan K. Chalup : Support vector clustering through proximity graph modelling. *In Proceedings, 9th International Conference on Neural Information Processing (ICONIP02)*, pages 898–903, 2002.

[Zar07] Stanislaw Zaremba : L'equation biharmonique et une classe remarquable de fonctions fondamentales harmoniques. *Bulletin International de l'Académie des Sciences de Cracovie*, pages 147–196, 1907.

[ZC11] H. Zeng et Y. Cheung : Semi-supervised maximum margin clustering with pairwise constraints. *Knowledge and Data Engineering, IEEE Transactions on*, 2011.

[ZDD01] Dmitry Zotkin, Ramani Duraiswami et Larry Davis : Multimodal 3-d tracking and event detection via the particle filter. *In Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[ZK98] Tong Zhang et C.-C. Jay Kuo : Hierarchical system for content-based audio classification and retrieval. *In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98)*, 1998.

[ZTK07] Kai Zhang, Ivor W. Tsang et James T. Kwok : Maximum margin clustering made practical. *In Proceedings of the 24th international conference on*

Machine learning, ICML '07, pages 1119–1126, New York, NY, USA, 2007. ACM.

[ZTK09] Kai Zhang, Ivor W. Tsang et James T. Kwok : Maximum margin clustering made practical. *Trans. Neur. Netw.*, 20:583–596, April 2009.

[ZWZ08a] Bin Zhao, Fei Wang et Changshui Zhang : Cuts3vm : a fast semi-supervised svm algorithm. *In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 830–838, New York, NY, USA, 2008. ACM.

[ZWZ08b] Bin Zhao, Fei Wang et Changshui Zhang : Efficient maximum margin clustering via cutting plane algorithm. *In SIAM International Conference on Data Mining*, pages 751–762, 2008.

[ZWZ08c] Bin Zhao, Fei Wang et Changshui Zhang : Efficient multiclass maximum margin clustering. *In Proc. of the 25th Int. Conf. on Machine Learning*, 2008.

Sébastien LECOMTE

Doctorat : Optimisation et Sûreté des Systèmes

Année 2013

Classification partiellement supervisée par SVM. Application à la détection d'événements en surveillance audio

Cette thèse s'intéresse aux méthodes de classification par Machines à Vecteurs de Support (SVM) partiellement supervisées permettant la détection de nouveauté (One-Class SVM). Celles-ci ont été étudiées dans le but de réaliser la détection d'événements audio anormaux pour la surveillance d'infrastructures publiques, en particulier dans les transports. Dans ce contexte, l'hypothèse « ambiance normale » est relativement bien connue (même si les signaux correspondants peuvent être très non stationnaires). En revanche, tout signal « anormal » doit pouvoir être détecté et, si possible, regroupé avec les signaux de même nature. Ainsi, un système de référence s'appuyant sur une modélisation unique de l'ambiance normale est présenté, puis nous proposons d'utiliser plusieurs SVM de type One Class mis en concurrence. La masse de données à traiter a impliqué l'étude de solveurs adaptés à ces problèmes. Les algorithmes devant fonctionner en temps réel, nous avons également investi le terrain de l'algorithmie pour proposer des solveurs capables de démarrer à chaud. Par l'étude de ces solveurs, nous proposons une formulation unifiée des problèmes à une et deux classes, avec et sans biais. Les approches proposées ont été validées sur un ensemble de signaux réels. Par ailleurs, un démonstrateur intégrant la détection d'événements anormaux pour la surveillance de station de métro en temps réel a également été présenté dans le cadre du projet Européen VANAHEIM.

Mots clés : traitement du signal - machines à vecteurs de support - analyse discriminante – surveillance électronique.

Partially Supervised Classification Based on SVM. Application to Audio Events Detection for Surveillance

This thesis addresses partially supervised Support Vector Machines for novelty detection (One-Class SVM). These have been studied to design abnormal audio events detection for supervision of public infrastructures, in particular public transportation systems. In this context, the null hypothesis ("normal" audio signals) is relatively well known (even though corresponding signals can be notably non stationary). Conversely, every "abnormal" signal should be detected and, if possible, clustered with similar signals. Thus, a reference system based on a single model of normal signals is presented, then we propose to use several concurrent One-Class SVM to cluster new data. Regarding the amount of data to process, special solvers have been studied. The proposed algorithms must be real time. This is the reason why we have also investigated algorithms with warm start capabilities. By the study of these algorithms, we have proposed a unified framework for One Class and Binary SVMs, with and without bias. The proposed approach has been validated on a database of real signals. The whole process applied to the monitoring of a subway station has been presented during the final review of the European Project VANAHEIM.

Keywords: signal processing - support vector machines - discriminant analysis – electronic surveillance.

Thèse réalisée en partenariat entre :

THALES

Ecole Doctorale "Sciences et Technologies"