Prédiction de situations anormales par apprentissage automatique pour la maintenance prédictive : approches en transport optimal pour la détection d'anomalies

Amina Alaoui Belghiti

université PARIS-SACLAY

Prédiction de situations anormales par apprentissage automatique pour la maintenance prédictive : approches en transport optimal pour la détection d'anomalies

Prediction of abnormal situations by machine learning in a predictive maintenance context : Optimal transport theory for anomaly detection

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Robotique
Graduate School Engineering and systems sciences. Référent : UVSQ

Thèse préparée dans l'unité de recherche **LISV(**Université Paris-Saclay, UVSQ), sous la direction de **Sylvain CHEVALLIER**, Maitre de conférences, HDR, le co- encadrement de **Eric Monacelli**, Professeur des universités

Thèse soutenue à Paris-Saclay, le 08 novembre 2021, par

Amina ALAOUI BELGHITI

Composition du Jury

Samia Bouchafa-Bruneau PU, Université Evry Val d'Essonne	Présidente
Gilles Gasso PU, INSA Rouen	Rapporteur & Examinateur
Guillaume Ginolhac PU, Université Savoie-Mont Blanc	Rapporteur & Examinateur
Eric Monacelli PU, Université Paris-Saclay	Examinateur
Julien Lefèvre MCF HDR, Université Aix Marseille	Examinateur
Sylvain Chevallier MCF HDR, Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT

NNT : 2021UPASG069

Je tiens à adresser mes sincères remerciements et ma profonde reconnaissance à mon directeur de thèse, M.Sylvain Chevallier, Maître de conférences HDR de l'université Paris Saclay pour m'avoir accueillie au sein de son équipe, pour le temps conséquent qu'il m'a accordé, pour ses qualités pédagogiques et scientifiques, sa franchise et sa sympathie. J'ai beaucoup appris à ses côtés et je lui adresse ma gratitude pour tout cela.

J'adresse de chaleureux remerciements à mon co-encadrant de thèse, M. Eric Monacelli, Professeur des universités de l'université Paris Saclay, pour son attention de tout instant sur mes travaux, pour ses conseils et son écoute. Son énergie et sa confiance ont été des éléments moteurs pour moi. J'ai pris un grand plaisir à travailler avec lui.

Je voudrais remercier les rapporteurs de cette thèse M. Gilles Gasso, Professeur des Universités de l'INSA Rouen, et M. Guillaume Ginolhac, Professeur des Universités de l'université Savoie-Mont Blanc, pour l'intérêt qu'ils ont porté à mon travail et leur lecture attentive.

J'associe à ces remerciements Mme. Samia Bouchafa-Bruneau, Professeure des universités de l'université Evry Val d'Essonne, et M. Julien Lefèvre, Maitre de conférences HDR de l'université Aix Marseille pour avoir accepté d'examiner mon travail.

Je tiens à remercier tous les membres de l'équipe Robotique interactive du laboratoire d'ingénierie des systèmes de Versailles, pour leur aide et leur bienveillance. Nous avons partagé de bons moments.

Je tiens à remercier également les équipes de la Business Unit Aerospace & Civil Systems de Hensoldt France, spécialement Frédéric Deville et Franck Nadalin pour leur soutien tout au long de la thèse.

Enfin, je remercie infiniment mes parents, mon mari, mon frère, mes soeurs et ma meilleure amie de m'avoir soutenue et encouragée au cours de ces années et sans lesquels je n'en serais pas là aujourd'hui.

Résumé

L'émergence de l'Industrie 4.0 et des systèmes intelligents entraîne une attention croissante pour les stratégies de maintenance prédictive qui peuvent réduire le coût et les temps d'arrêt et augmenter la disponibilité des équipements industriels. Dans cette thèse, nous présentons une vue d'ensemble des architectures de maintenance prédictive et nous nous intéressons à un pilier capital de ces architectures, la détection d'anomalies comme première étape de prise de décision dans une architecture de maintenance prédictive.

Nous apportons deux contributions à cette question de recherche. Une première méthode de classification semi-supervisée en transport optimal dans deux versions (paramétrique et non-paramétrique) pour la détection d'anomalies dans les séries temporelles. Les travaux expérimentaux de l'application de cette méthode sur des ensembles de données acoustiques synthétiques et réels prouvent la robustesse des métriques au sens transport optimal et démontrent en outre la supériorité des performances de la méthode par rapport aux algorithmes de l'état-de-l'art.

La deuxième contribution concerne une méthode non-supervisée de détection d'anomalies dans des données multidimensionnelles. Elle identifie les valeurs aberrantes locales dans un espace topologique non-euclidien en utilisant des métriques en transport optimal. Les résultats expérimentaux montrent l'efficacité de la méthode à remédier au problème de la malédiction de dimensionnalité et témoignent de la différence statistiquement significative de la méthode proposée par rapport aux méthodes de l'état de l'art évaluées.

Abstract

The emergence of Industry 4.0 and smart systems is leading to increasing attention to predictive maintenance strategies that can decrease the cost of downtime and increase the availability of industrial equipment. In this thesis, we present an overview of predictive maintenance architectures and we are interested in a capital pillar of these architectures, the anomaly detection as a first step of decision-making in a predictive maintenance architecture.

We provide two contributions to this research question. A first method of semi-supervised classification in optimal transport in two versions (parametric and non-parametric) for the detection of anomalies in time series. The experimental results of this method's application on synthetic and real acoustic data sets prove the robustness of the metrics derived from optimal transport and further demonstrate the superiority of the performance of the method over state-of-the-art algorithms.

The second contribution concerns an unsupervised anomaly detection method in multidimensional data. It identifies local outliers in a non-Euclidean topological space using optimal transport metrics. The experimental results revealed the effectiveness of the method in solving the dimensionality problem and testify to the statistically significant difference of the proposed method compared to the methods of the state of the art evaluated.

Table des matières

Table des figures

Liste des tableaux

Introduction

Contexte du travail

Aujourd'hui, la maintenance est un facteur stratégique pour garantir une productivité élevée des systèmes industriels, mais pour des raisons économiques les entreprises sont menées à réduire les dépenses de maintenance avec des conséquences critiques pour la fiabilité à long terme et développer des politiques de maintenance adaptées afin de garantir l'efficacité des usines de production, en termes de qualité et de disponibilité. Pour cette raison, le concept de maintenance lui-même a considérablement évolué au fil du temps, grâce à d'importants apports dans la recherche.

La maintenance prédictive est la dernière forme de maintenance conçue. Elle est adoptée dans de nombreux secteurs, en particulier dans ceux où la fiabilité est primordiale, tels que les centrales électriques, les services publics, les systèmes de transport, les systèmes de communication et les services d'urgence. Essentiellement, elle prévoit des défauts ou des défaillances dans un système qui se détériore afin d'optimiser les efforts de maintenance en évaluant principalement l'état du système au moyen de données historiques et de surveillance du système en question. Un programme de maintenance prédictive détecte principalement les premiers signes de panne, puis lance les procédures de maintenance au bon moment.

L'idée de base dans cette forme de maintenance remonte aux années 1940 lorsque les techniciens de maintenance expérimentés devaient utiliser leurs sens pour détecter un signe de problème dans les machines d'usine. Dans les stratégies actuelles, cette surveillance de l'état de la machine à toujours cours, le principal

changement concerne le remplacement des fonctions sensorielles humaines par l'utilisation de capteurs. Ces capteurs génèrent des données de mesure qui informent sur l'état du système afin de prédire les défaillances et anticiper et/ou optimiser les actions de maintenances à mener, d'où le nom de maintenance prédictive.

Une stratégie de maintenance prédictive se compose essentiellement de trois étapes : acquisition des données, traitement des données et prise de décision en matière de maintenance. Elle est le fruit des évolutions du 20e siècle liées au développement du secteur des télécommunications, et des systèmes de production informatisés, et l'ouverture sur les nouvelles avancées scientifiques du 21e siècle en traitement de l'information et exploitation des données.

Le sujet de maintenance prédictive s'inscrit dans l'axe de recherche de la société Hensoldt France en collaboration avec LISV afin de proposer une solution adaptée aux bancs de test qu'elle conçoit. Les développements liés à cette thèse seront intégrés à terme dans cette solution de maintenance prédictive.

Problématique

La prise de décision en matière de maintenance fait l'objet de plusieurs sujets de recherche aujourd'hui. Parmi ces sujets de recherche, la détection d'anomalie est une question scientifique ouverte qui préoccupe toujours les chercheurs en sciences des données.

La détection d'anomalies est un sujet autrefois étudié par la communauté statistique, dès le 19e siècle, qui est toujours d'actualité avec les différentes études menées aujourd'hui, et qui a même la vocation à s'ouvrir sur les méthodes d'intelligence artificielle et de traitement des données avancées. C'est un sujet qui fait référence au problème de la recherche, dans les données, de modèles non conformes au comportement attendu. Dans différents domaines d'application, ces modèles non conformes sont appelés aussi observations aberrantes, particularités, contaminants, etc. L'importance de la détection d'anomalies est due au fait que les anomalies dans les données se traduisent par des informations importantes, souvent critiques par

rapport au fonctionnement d'un système particulier. Les domaines d'application sont divers allant de la détection de fraude à la détection des dommages industriels, passant par la détection des anomalies médicales, la détection de la cyber-intrusion ou encore le traitement du texte et de l'image.

Pour donner une définition théorique, une anomalie est un modèle ou une observation qui n'est pas conforme au comportement normal attendu. Une approche simple de détection d'anomalie consiste donc, à définir une région représentant un comportement normal et à déclarer toute autre observation dans les données qui n'appartient pas à cette région normale comme une anomalie. Mais plusieurs facteurs rendent cette approche apparemment simple très difficile à mettre en œuvre. Dans de nombreux domaines, le comportement normal ne cesse d'évoluer et une représentation immédiate du comportement normal pourrait ne pas être suffisamment significative par la suite. Il est aussi très difficile de définir une frontière précise entre un comportement normal et un comportement anormal. Ainsi, des observations normales qui se trouvent près de la frontière peuvent être considérées comme anormales. De plus, selon leur nature, les données peuvent contenir du bruit qui tend à être similaire aux anomalies réelles et qui est donc difficile à distinguer et à éliminer. Considérant tous ces facteurs, les techniques de détection d'anomalies sont multiples et peuvent être divisées en trois grandes catégories : les méthodes statistiques, les techniques basées sur l'apprentissage automatique et les techniques basées sur l'exploration des données. L'usage d'une méthode ou d'une autre dépend du domaine d'application, du type des données et de la nature de la détection.

Dans le contexte industriel et dans le cadre d'une stratégie de maintenance prédictive des équipements, la détection d'anomalie constitue la première étape de la prise de décision en matière de maintenance. Il s'agit de détecter les anomalies de fonctionnement dans les données issues de la surveillance de l'état de l'équipement industriel et/ou dans des données historiques de fonctionnement qui sont souvent multidimensionnelles . Les données de surveillances peuvent être aussi différentes que le type de la machine à surveiller. Mais souvent, dans le cadre de l'industrie manufacturière, les usines se composent de grands parcs de machines

électro-mécaniques, et donc plusieurs sources d'informations à exploiter, telles que les vibrations, l'acoustique, la température ou encore la puissance utile. Ces informations sont enregistrées par des capteurs sur une période limitée ou en continu. Le flux des données issues des enregistrements conduit à des séquences de données appelées séries temporelles.

Principales contributions

La détection d'anomalie dans les series temporelles repose sur des méthodes qui modélisent explicitement le comportement normal de la série temporelle de sorte qu'un écart significatif par rapport à ce modèle soit considéré comme une valeur aberrante. La modélisation peut se faire à travers l'apprentissage automatique et la prédiction des données futures sur la base des données d'entrainement, ou un système reposant sur des règles qui encodent une expertise de connaissance sur le comportement des échantillons futurs de la serie temporelle. Compte tenu des échantillons prédits et des valeurs réelles observées, les anomalies sont détectées en s'appuyant sur la notion d'écart. La mesure la plus simple de l'écart est le calcul de l'erreur relative. Le seuillage de cette mesure permet distinguer les observations aberrantes et par conséquent, les valeurs de ces seuils déterminent la sensibilité de détection. Cependant, l'usage de l'erreur relative n'est pas toujours la façon optimale pour la détection d'anomalie dans une série temporelle donnée. Le choix de la métrique optimale dépend fortement de la nature de la série temporelle. Si la série est très régulière avec un modèle précis, l'usage de l'erreur est suffisant pour la détection d'anomalie, mais ce n'est pas le cas pour des séries plus complexes. Une catégorie de méthodes de détection d'anomalie plus efficace s'appuie sur la décomposition des séries temporelles. Dans la littérature, il s'agit de les décomposer en trois composantes : la tendance, la saisonnalité et le bruit. Cette décomposition peut être effectuée à la fois dans le domaine temporel via le lissage ou dans le domaine fréquentiel via la décomposition spectrale.

La détection d'anomalies dans les données multidimensionnelles devient un problème de recherche fondamentale qui a diverses applications dans le monde réel.

Cependant, de nombreuses techniques de détection d'anomalies existantes ne parviennent pas à conserver une précision suffisante. Ce phénomène, principalement dû à la difficulté de détecter des anomalies dans des espaces de grandes dimensions, est appelé la malédiction de la dimensionnalité, il affecte les techniques existantes en termes de performances et de précision.

Les travaux de ce manuscrit définissent deux nouvelles méthodes de détection d'anomalies. la première est une méthode de classification semi-supervisée pour la détection d'anomalies dans les séries temporelles qui repose sur la décomposition spectrale de la série temporelle et la quantification des variations au niveau des données dans le domaine fréquentiel grâce à des métriques en transport optimal. Le transport optimal qui est une théorie mathématique datant de la fin du 18e siècle qui a pris de l'ampleur récemment (2 médailles Fields dans les douze dernières années). C'est une théorie élégante qui renforce l'apprentissage automatique d'aujourd'hui et qui a fait ses preuves dans plusieurs sujets tels que la reconstruction de forme en infographie, le transfert de couleur en vision par ordinateur, l'apprentissage par transfert, le traitement d'images ou encore le démixage spectral pour les données musicales.

Afin de combler les lacunes des méthodes existantes liées à la malédiction de la dimensionnalité, ce manuscrit apporte une deuxième contribution. Une méthode non-supervisée de détection d'anomalies dans les données multidimensionnelles à travers l'identification des aberrations locales dans un espace topologique non-euclidien. Elle repose sur un concept de densité locale, où la localité est donnée par les voisins les plus proches, et dont les métriques en transport optimal sont utilisées pour estimer cette densité. En comparant la densité locale d'une instance de données à la densité locale de ses voisins, les régions de densité similaire, et les points qui ont une densité nettement plus basse que la densité de leurs voisins sont alors identifiés.

Plan de thèse

Afin de présenter les travaux, ce manuscrit de thèse est structuré en cinq chapitres. D'abord la présentation du contexte global du sujet de recherche et l'état de l'art associé. Ainsi, le chapitre 1 présente le concept de maintenance prédictive ainsi qu'une vue d'ensemble des architectures associées comme porte d'entrée des travaux. Le chapitre 2 présente les concepts fondamentaux pour la détection de situation anormale comme étape déterminante de l'efficacité opérationnelle d'une solution de maintenance prédictive. Ce chapitre se compose de deux volets complémentaires, le premier volet aborde l'état de l'art sur la détection d'anomalies et les mécanismes d'apprentissage automatique dédiés. Le second volet porte sur un outil puissant de la théorie d'optimisation, le transport optimal pour l'analyse et l'exploitation des distributions de données. Le chapitre 3 est le coeur du manuscrit, il est aussi le fil conducteur entre l'état de connaissances et les méthodes mises à contribution. Il aborde en détail les nouvelles méthodes proposées dans un cadre théorique. Le chapitre 4 définit la méthode de classification en transport optimal dans ses versions paramétrique et non-paramétrique pour la détection d'anomalies dans les séries temporelles. Le chapitre 5 présente la méthode non-supervisée de détection d'anomalie dans les données multidimentionnelles qui repose sur l'estimation optimale de la densité d'atteignabilité locale. Lors de ces deux chapitres de contribution, la mise en application de la méthode, les différentes expérimentations et les résultats d'évaluation des performances sont détaillés. Enfin, la conclusion générale fait le bilan des travaux et ouvre les perspectives de recherche.

Chapitre 1

Maintenance prédictive

Sommaire

La maintenance industrielle a connu de réels progrès au fil du temps. Elle était perçue uniquement comme une activité d'arrière-plan, dont l'utilité était considérée comme toute relative et à laquelle il ne fallait faire appel que lorsque la machine était tombée en panne. Son périmètre était très limité et axé sur l'électricité, la mécanique ou encore le graissage. Les notions de prévision ou de prévention n'avaient

pas encore fait leur apparition. Les stratégies alors employées s'appuyaient uniquement sur le dépannage et le déploiement de correctifs d'envergure. À ce stade de la maintenance industrielle, les exigences de production n'étaient pas aussi sévères qu'elles le sont aujourd'hui.

Après la seconde guerre mondiale, une nouvelle étape de l'histoire de la maintenance des équipements a commencé suite à la reconstruction de l'industrie. Un marché beaucoup plus compétitif s'est développé, obligeant les fabricants à augmenter leur production. Le surmenage des machines entraînait une augmentation des temps d'arrêt et une augmentation des coûts de réparation des machines. Cette augmentation de la production a requis de meilleures pratiques de maintenance, ce qui a conduit à une étape clé dans l'évolution des normes de l'industrie manufacturière : le développement de la maintenance préventive.

Avec l'évolution de l'industrie, les usines et les systèmes sont devenus encore plus complexes. Les exigences du marché concurrentiel et l'intolérance aux temps d'arrêt ont augmenté. Parallèlement à ces exigences, une nouvelle approche des processus de défaillances, des techniques de gestion améliorées et de nouvelles technologies ont permis une compréhension plus large du bon fonctionnement des machines et des composants.

Cette évolution des pratiques s'inscrit dans un mouvement plus large, que l'on peut rapprocher de l'évolution de concepts plus philosophiques. La question du rapport entre l'Humain et la machine est centrale depuis plusieurs siècles, comme en témoigne la théorie homme-machine de Descartes. L'assimilation du mécanisme cartésien et la formalisation de la pensée ont conduit à un rapprochement de la réflexion et de l'exécution, c'est la théorie du cerveau-machine Jeannerod, 1983. Après avoir délégué ses tâches manuelles à la machine, l'Humain cherche à déléguer certaines tâches de réflexions et il conçoit des machines plus complexes pour se faire.

C'est dans ce sens qu'une nouvelle forme de maintenance industrielle est en train d'émerger. Une maintenance intelligente reposant sur la surveillance de l'état des équipements, centrée sur la fiabilité et obtenue grâce à des techniques plus complexes. Ces techniques vont de l'utilisation de capteurs avancés à la prise en

compte semi-automatisé des réponses humaines. Le but de ces avancées est de prévenir les pannes et de mettre en oeuvre la maintenance uniquement en cas de présence d'une panne potentielle.

1.1 Vue d'ensemble des architectures de maintenance prédictive

La maintenance prédictive est la dernière forme de maintenance adoptée par de nombreux secteurs où la fiabilité est primordiale tels que les centrales électriques, les services du transport public, les systèmes de communication et les services d'urgence. L'idée originelle dans cette forme de maintenance vient des années 1940 lorsqu'un technicien expérimenté dans la maintenance avait utilisé tous ses sens pour détecter un signe de problème dans les machines. Aujourd'hui, ces informations peuvent être obtenues par la surveillance de l'état de la machine (Gorecky et al., 2014), en remplaçant les fonctions sensorielles humaines par l'utilisation de capteurs. Ces capteurs génèrent des données de mesures qui informent sur l'état du système afin de prédire les défaillances. Ils permettent d'anticiper et d'optimiser les actions de maintenance à mener, ce qui donne son nom à la maintenance prédictive. Les trois étapes d'une stratégie de maintenance prédictive sont : l'acquisition des données, le traitement des données et la prise de décision en matière de maintenance (Selcuk, 2017). Cette architecture peut être considérée comme une boucle perception-décision-action comme illustrée sur la Fig 1.1

Dans la partie perception, les données du capteur sont acquises via une unité d'acquisition puis traitées afin d'en extraire les informations utiles. Les données peuvent être aussi différentes que le type de l'équipement à maintenir. Dans le cas des équipements mécaniques ou électromécaniques, de nombreuses sources d'information sont à surveiller telles que les vibrations, les émissions acoustiques, la température et la puissance utile. Auxquelles s'ajoutent les données historiques des pannes et les données conceptuelles (Selcuk, 2017). L'ensemble de ces données permet de surveiller le comportement normal et d'aborder les questions d'analyse

de la connaissance et du raisonnement afin définir quelle approche d'analyse est utile pour quelle partie de l'aide à la décision.

Plusieurs méthodes d'analyse sont possibles dans la partie de décision, de la détection d'anomalies au pronostic, mais l'usage de la méthode dépend de la nature de la décision à prendre. Nous allons voir chacune des étapes et détailler les différentes approches.

Le diagnostic est la première étape de la partie décision. Il repose sur la détection d'anomalie comme pilier fondamental qui consiste à remonter les anomalies en comparant en continu le comportement courant de la machine avec son comportement de référence. Les résultats sont ensuite exploités pour déterminer les modes de défaillances possibles et identifier les défauts potentiels (Galar et al., 2015). Afin de résoudre les problématiques liées au diagnostic qui vont de la détection d'anomalies à la détermination des modes de défaillances, diverses méthodes sont à explorer. Elles incluent l'analyse des séries temporelles et la classification automatique.

Pour réaliser un pronostic ou une prédiction de défaut, trois méthodologies d'analyse sont possibles, une première axée sur les données, une deuxième axée sur les modèles et une troisième hybride (Bousdekis et al., 2019). Les méthodes statistiques reposent sur des modèles numériques tirés des données de capteurs. Les approches de type réseaux de neurones peuvent être utilisées pour créer un modèle utilisant sur les données historiques de la machine depuis le début de son fonctionnement jusqu'à sa défaillance et des approches stochastiques peuvent servir afin de gérer le niveau de prédiction.

La partie action inclut toutes les tâches à effectuer dans un système intelligent de gestion de maintenance, telles que l'optimisation des actions de maintenance et la mise à jour automatique et continu du calendrier de ces actions pour les communiquer au service de maintenance via une interface humain-machine.

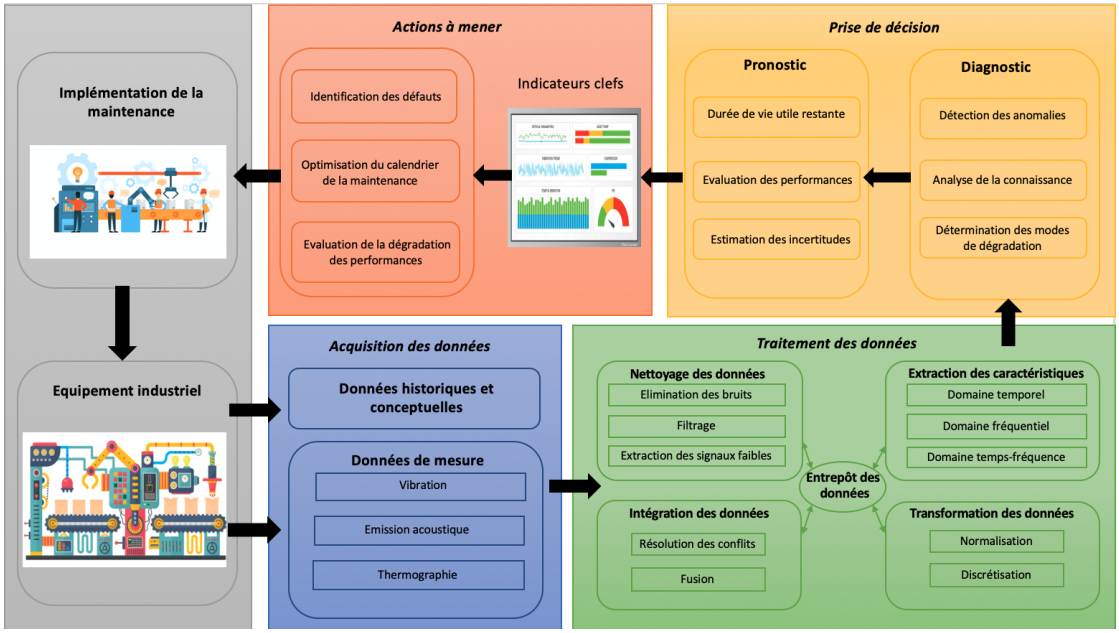


Figure 1.1: Maintenance prédictive : vue d'ensemble

1.2 Acquisition des données

Bien que les sens humains fournissent des informations précieuses liées à l'état du système qui doit être maintenu, la maintenance prédictive doit s'appuyer sur des capteurs manufacturés. La chaleur, le courant du moteur, la pression des fluides, les vibrations, le son et le débit sont de nombreux paramètres mesurables pour surveiller la santé d'un système. Dans le cadre typique d'une usine, une stratégie de maintenance prédictive complète doit inclure au moins la surveillance vibratoire, acoustique et de température (Kroll et al., 2014).

1.2.1 Surveillance vibratoire et acoustique

Surveillance vibratoire La plupart des usines étant constituées de systèmes électromécaniques, la surveillance des vibrations est un élément clé pour la maintenance prédictive. Elle est utilisée principalement pour les équipements rotatifs ou alternatifs, soit de façon continue soit sur des intervalles programmés. La nature des vibrations peut indiquer un déséquilibre ou un désalignement. Les niveaux

de vibrations concentrés sur des multiples de la vitesse de rotation peuvent indiquer des problèmes imminents au niveau des roulements. Cependant les vibrations ne sont pas toujours simples à interpréter, surtout pour les machines à courant alternatif. Elles peuvent être analysées avec des techniques d'apprentissage automatique (Henriquez et al., 2013). La surveillance vibratoire est particulièrement efficace pour détecter les problèmes liés au déséquilibre, à l'excentricité, aux mauvais alignements des accouplements et des roulements, aux fissures d'arbres ou encore à des engrenages et roulements usés ou endommagés, des courroies et des chaînes d'entraînement défectueuses ou mal réglées, etc. (Fedele, 2011).

Les instruments utilisés dans le cadre de la surveillance vibratoire comprennent les capteurs de déplacement, les capteurs de vitesse et les accéléromètres. Ces derniers sont le meilleur choix pour la plupart des équipements rotatifs industriels car ils sont simples, faciles à intégrer et très sensibles aux différentes vibrations (Dai et al., 2019). Les accéléromètres reposent généralement sur deux technologies : ils peuvent être piézoélectriques ou à système microélectromécanique (MEMS). Les capteurs de vibrations piézoélectriques, illustrés sur la Fig 1.2, utilisent l'effet de la contrainte mécanique causée par le mouvement de l'équipement pour détecter l'accélération et, par conséquent, les vibrations (Yaghootkar, Azimi et Bahreyni, 2017 ; Sinar et Knopf, 2020). Ils génèrent un flux d'impulsions et le signal peut être présenté comme une forme d'onde temporelle ou traité par FFT pour convertir les données en un spectre de fréquences. Il est judicieux de noter que ce genre de capteurs est intrinsèquement analogique, il nécessite donc une électronique de traitement supplémentaire pour numériser le signal comme représenté sur la Fig 1.3.

Les capteurs de vibrations utilisant les MEMS, visibles sur la Fig 1.4, sont fabriqués par lithographie, ce qui permet une intégration au niveau du capteur proprement dit avec l'électronique de support (Zhao et al., 2019 ; Zhang et al., 2019a). Ces avantages se traduisent par des systèmes de surveillance des vibrations compacts, robustes et économiques. Ces capteurs sont devenus de plus en plus populaires grâce aux progrès de l'analyse des données et du traitement du signal.

Figure 1.2: Capteur de vibration piézoélectrique

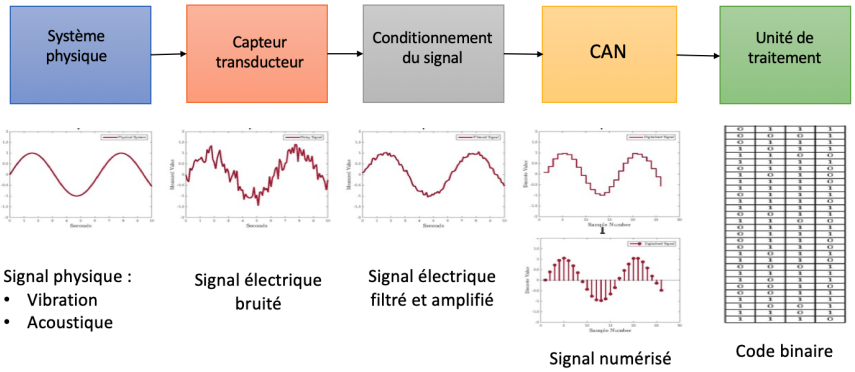


Figure 1.3: Schéma fonctionnel d'un système d'acquisition de données

Surveillance acoustique La plupart des machines créent des modèles sonores stables dans des conditions normales de fonctionnement. Si ces modèles sonores sont enregistrés comme références, les modifications des motifs de références permettent donc d'indiquer plusieurs types de détérioration des composants. Comme la surveillance vibratoire, le son est un sous-domaine de l'analyse du bruit. La différence entre les deux est la bande de fréquence à surveiller. Dans le cas de l'analyse des vibrations, la plage surveillée est comprise entre 1 Hz et 30 000 Hz ; les capteurs d'ultrasons visibles sur la Fig 1.5 mesurent les fréquences de bruit supérieures à 30 000 Hz. L'analyse des ultrasons permet de surveiller la santé du

Figure 1.4: Capteur de vibration MEMS

moteur dans des actifs complexes, en présence d'un bruit audible accru. Elle s'intéresse à des sons dans le spectre non audible, où l'amplitude des bruits est moindre. Les longueurs d'onde des signaux audibles dans les basses fréquences sont généralement comprises entre 1,7 cm et 17 m. Les longueurs d'onde des signaux dans les hautes fréquences vont d'environ 0,3 cm à 1,6 cm. Lorsque la fréquence de la longueur d'onde augmente, l'énergie augmente, ce qui rend les ultrasons plus directifs (Mehta, Werner et Mears, 2015). Ceci est extrêmement utile pour la localisation des défaillances au niveau des roulements ou pour la détection de fuites qui créent généralement un bruit à haute fréquence causé par l'expansion ou la compression de l'air, des gaz ou des liquides lorsqu'ils s'écoulent à travers l'orifice. Ils sont également très utiles pour la détection des défauts cachés dans les matériaux, en particulier dans les métaux (Moretti et al., 2020).

Une autre catégorie de capteurs pour la surveillance acoustique est le microphone à système microélectromécanique (MEMS), que l'on peut voir la Fig 1.6. Il contient un élément MEMS sur un PCB, généralement contenu dans un boîtier

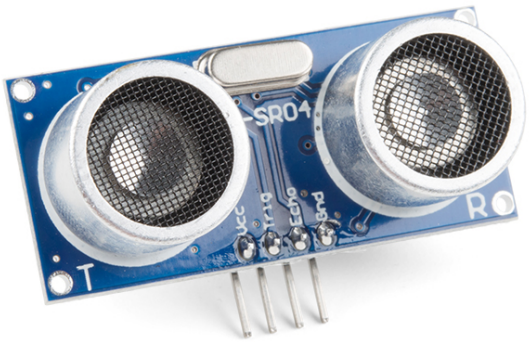


Figure 1.5: Capteur ultrason

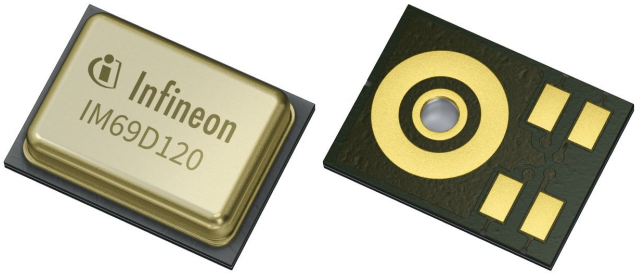


Figure 1.6: Microphone à système microélectromécanique (MEMS)

métallique avec un orifice inférieur ou supérieur pour permettre aux ondes de pression acoustique de se concentrer à l'intérieur (Murphy, 2020). Les microphones MEMS offrent un faible coût et une petite taille. Ce sont des moyens efficaces de détecter les défauts de la machine tels que l'état des roulements, l'engrènement des engrenages, le désalignement et le déséquilibre des pièces mécaniques. Les microphones MEMS sont un choix idéal pour les applications alimentées par batterie. Ils peuvent être situés à des distances importantes de la source de bruit et ne sont pas invasifs (Higashi et al., 2018).

1.2.2 Thermographie

La thermographie peut être utilisée pour surveiller l'état des machines, des structures et des systèmes d'usine en s'appuyant sur des instruments conçus pour surveiller l'émission d'énergie infrarouge (reliée principalement ici à la température de surface) afin de déterminer les conditions de fonctionnement (Huda et Taib, 2013a ; Bagavathiappan et al., 2013). En détectant les anomalies thermiques, c'est-à-dire les zones qui sont plus chaudes ou plus froides qu'elles ne devraient l'être, un diagnostic peut être fait pour localiser et définir une multitude de problèmes naissants au sein du système. Trois types d'instruments peuvent être utilisés dans le cadre de la surveillance thermographique : les thermomètres infrarouges, les scanners linéaires et les systèmes d'imagerie infrarouge.

Les thermomètres infrarouges sont conçus pour fournir la température réelle de la surface en un seul point, relativement petit, sur une machine. Dans le cadre la surveillance thermographique, le thermomètre infrarouge au point d'utilisation peut être utilisé en conjonction avec de nombreux instruments de vibration pour surveiller la température à des points critiques sur les machines, par exemple les températures des chapeaux de palier ou celles des enroulements du moteur (Sollai et al., 2016). Tandis que l'imagerie thermique obtenue avec des caméras infrarouges, telle que celle montrée sur la Fig 1.7, fournit un balayage plus large et offre un champ de vision plus grand. Elles permettent ainsi de scanner les émissions infrarouges des machines, des processus ou des équipements complets en très peu de temps.

Les caméras thermiques comprennent en général une lentille optique, un détecteur thermique et des circuits électroniques, éventuellement un refroidisseur pour le détecteur et un logiciel de traitement et de génération des images afin de les transmettre et les afficher (Minkina et Dudzik, 2009). Un schéma simple d'une caméra thermique infrarouge est représenté sur la Fig 1.8.

L'inclusion de la thermographie dans une stratégie de maintenance prédictive permet de surveiller l'efficacité thermique des processus critiques des systèmes. Ainsi, les techniques infrarouges peuvent être utilisées pour détecter les problèmes

Figure 1.7: Camera thermique

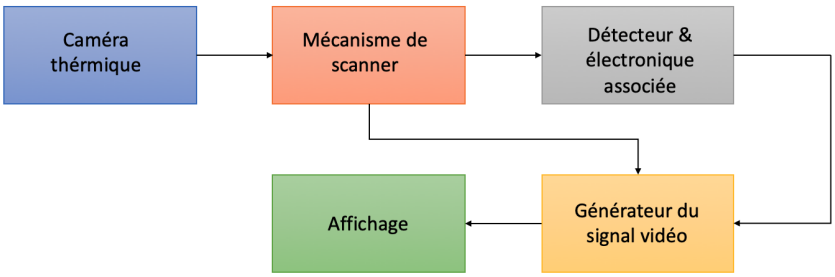


Figure 1.8: Schéma fonctionnel d'une caméra thermique

dans une variété de systèmes et d'équipements d'usine, y compris les appareillages électriques, les boîtes de vitesses, les sous-stations électriques, les transmissions, les panneaux de disjoncteurs, les moteurs et les roulements.

La pratique de la surveillance des données permet ainsi d'identifier les changements d'état de l'équipement et d'examiner ses parties internes en fonctionnement sans l'ouvrir physiquement. C'est une étape importante de la maintenance prédictive, dont une surveillance précise permet une détection précoce des anomalies et une identification correcte des types de défauts. Par conséquent, plus une stratégie de maintenance prédictive est précise et sensible, plus la décision de maintenance peut être précise, et plus il y a de la disponibilité pour planifier et effectuer la maintenance avant que la panne ne se produise.

1.3 Traitement des données

Après la collecte, la transmission et le stockage des données de mesure, cette partie de l'architecture permet de les transformer en caractéristiques et en indicateurs utiles pour la prise de décision.

La partie du traitement des données est divisée en deux étapes : un prétraitement et une analyse des données. L'étape de prétraitement supprime les données d'erreurs provenant de défauts de capteurs et l'étape d'analyse des données extrait des informations utiles à partir des signaux bruts, via un processus appelé extraction de caractéristiques. La sélection de la méthode à utiliser pour analyser les données dépend du type de données. Les données peuvent prendre la forme d'images pour la thermographie, des formes d'onde pour les vibrations ou pour les émissions acoustiques.

1.3.1 Traitement de l'imagerie thermique

Les images thermiques infrarouges sont en général bruitées et souffrent d'un rapport signal/bruit faible. Par conséquent, diverses techniques de traitement d'image sont utilisées pour améliorer les images thermiques acquises (Bagavathiappan et al., 2013).

Du point de vue de la surveillance de l'état d'un système, les principaux objectifs du traitement d'image sont la détection des points chauds et l'extraction des caractéristiques de défaut. À des fins d'amélioration de l'image, divers algorithmes d'opération ponctuelle tels que l'étirement du contraste ou l'égalisation d'histogramme peuvent être utilisés. L'objectif de ces algorithmes est d'étirer l'histogramme d'une image pour augmenter la plage dynamique de l'image et améliorer implicitement le contraste. L'utilisation des techniques avancées d'analyse du signal comme la reconstruction du signal thermographique (Balageas et al., 2015) et l'analyse en composantes principales (ACP), peut détecter les défauts de plus grande profondeur avec un contraste thermique très élevé. L'extraction des caractéristiques reposant sur l'analyse de texture dans les images thermiques s'est avéré

utile pour la classification des images et la détermination de la forme de l'objet (Pramanik, Bhattacharjee et Nasipuri, 2016 ; Huda et Taib, 2013b). Pour la détection des points chauds, une segmentation et un seuillage d'image sont généralement employés. Plusieurs algorithmes de segmentation et de seuillage peuvent être utilisés et le choix dépend de la nature de l'image et de l'objectif d'utilisation (Sharma, Mishra et Shrivastava, 2012). Le seuillage global, le seuillage flou, la segmentation morphologique, la segmentation à deux niveaux et la segmentation normalisée sont les algorithmes de segmentation et de seuillage d'image les plus utilisés (Mohd, Herman et Sharif, 2017).

1.3.2 Traitement des signaux vibratoires et acoustiques

L'analyse d'un signal de vibration ou d'émission acoustique implique la mesure d'onde telle que détectée à un endroit particulier. Le choix de l'emplacement du capteur doit être fait avec soin car cela affecte la quantité et la qualité des informations qui seront générées. Traditionnellement, les mesures de vibration sont effectuées en trois directions : deux directions radiales (X et Y) et la direction axiale. Ces mesures peuvent être effectuées à plusieurs endroits à l'aide de plusieurs accéléromètres ou à un seul endroit utilisant un accéléromètre triaxial. La mesure des émissions acoustiques nécessite une grande attention dans le choix de l'emplacement du capteur afin de réduire le nombre d'interférences possibles qui peuvent affecter les mesures du capteur. Les informations enregistrées par un capteur sont le résultat de multiples processus qui modifient le signal. Il est affecté par le milieu de propagation, le capteur et l'instrument électronique. Ainsi, le signal enregistré n'est pas le même que le signal source d'origine (Ono, 2018). Tous ces effets peuvent être représentés avec des fonctions de transfert individuelles. En identifiant les fonctions de transfert correspondant à tous les effets et en les supprimant de l'enregistrement signal, le signal source peut en principe être identifié (Kishawy et al., 2018). De même pour les vibrations, les signaux font face à un scénario similaire, bien que les effets dus au milieu de propagation soient moins significatifs que dans le cas des signaux d'émission acoustique. Il existe plusieurs méthodes qui peuvent être utilisées pour l'analyse des signaux d'ondes. Ces tech-

niques sont classées en fonction du domaine dans lequel l'analyse est effectuée. Dans l'analyse des signaux de vibration, les approches d'extraction de caractéristiques les plus simples reposent sur le calcul de la moyenne quadratique (RMS) et du facteur de crête. Le facteur de crête est le rapport de la valeur de crête à la valeur RMS de l'accélération. Les informations qui peuvent être obtenues sur d'éventuels défauts sont limitées, et la détection des défauts est difficile. Cependant, ces mesures peuvent être utilisées dans le cadre d'une analyse de tendance, dans laquelle des valeurs croissantes peuvent être considérées comme une indication de la détérioration de la machine (Castellani et al., 2017). Une autre approche de l'étude des signaux vibratoires dans le domaine temporel est liée à l'analyse des moments statistiques dont les plus courants sont :

— Premier moment : la valeur moyenne.
— Deuxième moment : la variance.
— Troisième moment : l'asymétrie.
— Quatrième moment : le Kurtosis.

Les moments impairs sont liés à la position de la valeur de crête de la fonction de densité de probabilité par rapport à la valeur moyenne et les moments pairs décrivent les caractéristiques de la répartition de la distribution (Georgiadis, Gong et Meier, 2018).

Les techniques d'analyse du domaine fréquentiel sont plus populaires pour l'analyse des vibrations et des émissions acoustiques car elles offrent la possibilité de capturer des événements périodiques produits par les défauts. L'utilisation des techniques du domaine fréquentiel est en outre facilitée par la transformée de Fourrier rapide (FFT), qui permet le calcul efficace du spectre de fréquences. Une technique simple d'analyse consiste à comparer directement le spectre du signal avec les fréquences caractéristiques du défaut. Une telle comparaison donne des indications qui peuvent être utilisées dans la localisation et l'identification des défauts. Cependant, cette comparaison directe ne donne de bons résultats que pour la détection de grands défauts (Malla et Panigrahi, 2019). Actuellement, la technique de résonance haute fréquence (HFRT), également connue sous le nom de technique d'enveloppe, est la méthode la plus utilisée dans l'analyse des vibrations

et des signaux d'émission acoustique. Dans cette méthode, la bande de fréquences dans laquelle se produisent les résonances est extraite par filtrage, redressement et démodulation des signaux d'onde. Une FFT est appliquée au signal résultant, et le signal de sortie est affiché dans le domaine fréquentiel. Cette technique est un moyen efficace pour identifier les défauts localisés mais l'est moins pour l'identification des défauts généralisés (Mishra et al., 2021 ; Segla, Wang et Wang, 2012).

La transformée en ondelettes est une technique dans laquelle le domaine temps-fréquence est considéré pour l'analyse du signal. Cette technique est utilisée à la fois pour l'analyse des vibrations et des signaux acoustiques. La motivation pour une approche temps-fréquence est l'hypothèse que les signaux présentent des non-stationnarités temporelles qui doivent être prises en compte pour assurer le bon diagnostic. D'autres techniques s'appuyant sur une représentation en domaine temps-fréquence, et qui sont utilisées pour l'analyse des signaux acoustique, comprennent l'aplatissement spectral et la cyclostationnarité (Vanhoy et Teku, 2017). Il a été démontré que l'aplatissement spectral donne un meilleur rapport signal sur bruit pour les roulements comparé à la décomposition en ondelettes. La cyclostationnarité a été prise en compte pour des capacités de détection améliorées par rapport à l'analyse de l'enveloppe spectrale reposant sur des signaux bruts ou des paramètres du domaine temporel (Camerini et al., 2019 ; Feng et al., 2019).

1.4 Prise de décision

L'étape décisionnelle fournit une analyse complète des résultats de la partie traitement des données afin de prendre des décisions liées aux actions de maintenance dans le but d'optimiser la durée de vie de l'équipement et d'éviter des pannes catastrophiques ou un arrêt brutal des équipements. Les techniques impliquées dans la phase de prise de décision sont divisées en deux grandes parties : le diagnostic et le pronostic.

1.4.1 Diagnostic

Le diagnostic est un travail complexe qui nécessite la corrélation de diverses méthodes, l'utilisation d'outils d'analyse approfondie et la modélisation du système, de son fonctionnement et de son comportement.

La première étape du diagnostic est la détection d'anomalies, une étape qui consiste à reconnaître la présence d'un état inhabituel dans le comportement du système, par rapport à certains modèles de comportements dits normaux qui peuvent être soit prédéfinis ou appris automatiquement. Cette étape aborde le problème de la découverte des modèles dans les données qui ne suivent pas le comportement prévisible. Ces modèles non conformes sont souvent dénotés comme étant des anomalies, aberrations, contaminants, discordants, exceptions, observations, valeurs aberrantes, particularités ou surprises selon le domaine d'application.

Si une instance de données est anormale dans un contexte défini (mais pas autrement), alors elle est étiquetée comme une anomalie contextuelle. La notion de contexte dépend de la structure de l'ensemble des données et doit être détaillée dans le cadre de la formulation du problème (Hayes et Capretz, 2015). Une instance de données est définie à l'aide des deux ensembles d'attributs (Marhas, Bhange et Ajankar, 2012) :

Attributs contextuels Ils sont utilisés pour définir le contexte. Par exemple, dans les données d'une série temporelle mesurant la température, le temps est un attribut contextuel qui définit l'emplacement d'une mesure sur la séquence entière.

Attributs comportementaux Ils décrivent les caractéristiques non contextuelles d'un fonctionnement. Dans le même exemple des données de température, le degré de température dans un contexte défini est un attribut comportemental.

Le comportement anormal est déterminé à l'aide des valeurs pour les attributs comportementaux dans un contexte spécifique. Ainsi, une instance de données peut être une anomalie contextuelle dans un contexte donné, mais une instance

de données identique pourrait être considérée comme régulière dans un contexte différent. Cette propriété est la clé de la classification des attributs contextuels et comportementaux pour une technique de détection d'anomalies. Un exemple explicatif est celui d'une série temporelle de température qui indique une température mensuelle dans une zone donnée. Une valeur spécifique peut être normale pendant l'hiver, mais la même valeur pendant l'été serait une anomalie.

La plupart des techniques de détection d'anomalies résolvent une formulation définie du problème. La formulation est induite par divers facteurs tels que la nature des données, la disponibilité des données étiquetées et le type d'anomalies à percevoir. Ces techniques font appel à des méthodes de diverses disciplines telles que l'analyse de données, la théorie de l'information, l'apprentissage automatique, les statistiques, la théorie spectrale, etc. (Patcha et Park, 2007).

La deuxième étape du diagnostic consiste à traiter la gravité des anomalies détectées et de déterminer la relation entre les causes et conséquences. Généralement, cette étape repose sur une base de connaissances. Cette base des connaissances peut être accumulée par simulation ou par des données historiques. Cependant, quelle que soit la manière dont les connaissances sont accumulées, la base de connaissances elle-même peut être incomplète et ne pas intégrer des causes et des conséquences imprévues (Pang et al., 2017).

Plusieurs méthodologies sont valables pour cette partie du diagnostic, du système expert à l'intelligence artificielle. On peut citer par exemple la régression du noyau auto-associatif, les réseaux de neurones auto-associatifs(Khaled, Hedi et Lotfi, 2011), les techniques d'estimation d'états multivariés auto-régressifs (Xing et Lv, 2019), des systèmes experts neuronaux, des systèmes neuro-flous (Viharos et Kis, 2015), ou encore des algorithmes génétiques (Cerrada et al., 2016). La plupart de ces méthodologies sont des outils de classification, ce qui signifie qu'elles sont incapables de faire face à des situations imprévues. Leur précision et leur robustesse diffèrent en fonction de leurs méthodes de prétraitement pour réduire les dimensions, transférer les domaines ou éliminer le bruit. Certaines d'entre elles se concentrent sur le mode de dégradation d'une composante spécifique, et d'autres essaient de trouver la composante entraînant une dégradation des performances

globales. En effet, pour que le module du diagnostic soit le plus efficace et le plus fiable possible, la base des connaissances doit être mise à jour périodiquement afin de couvrir les défauts imprévus. Un outil de clustering est donc nécessaire pour distinguer les nouveaux cas des cas déjà appris, plusieurs algorithmes, y compris le système expert pour déterminer les causes profondes, doivent être fournis en parallèle, sans oublier que la surveillance du système doit être effectuée à l'aide de capteurs avancés pour une meilleure précision.

1.4.2 Pronostic

Le module du pronostic a pour objectif d'analyser la durée de vie utile restante et de vérifier qu'elle est supérieure à une certaine limite de défaillance probabiliste. Tant que la limite de défaillance est raisonnablement convertie en une limite admissible en termes d'efficacité, les pronostics en ingénierie de fiabilité peuvent alors être appliqués (Jin et al., 2019). Ce module prévoit le modèle de perte incontrôlable pour les conditions d'exploitation futures. Il consiste à prédire la perte accumulée pendant un certain intervalle afin de comparer le rapport coût-bénéfice pour une gestion intelligente des actions de maintenances à mener. Trois catégories de méthodologies ont été suggérées en ingénierie de la fiabilité en fonction des informations disponibles pour la modélisation de la dégradation comme indiqué sur la Fig 1.9 : pronostic reposant sur les données de défaillance, pronostic reposant sur la fatigue et pronostic reposant sur l'effet. Parce que l'approche reposant sur les données de défaillance n'utilise que des données génériques issues de l'expérience industrielle, elle ne peut pas bénéficier de la surveillance de l'état (Halim et al., 2018). L'approche utilisant la fatigue fait un meilleur modèle de pronostic, mais elle estime toujours les états approximatifs de la perte incontrôlable en observant les conditions environnementales extérieures. De plus, il n'est pas facile de comprendre la relation entre de nombreux facteurs de fatigue et des pertes incontrôlables. En s'appuyant sur les effets, le pronostic caractérise la durée de vie d'une unité ou d'un système fonctionnant dans son état normal (Lei et al., 2018).

L'analyse de Weibull (Chanthery et Ribot, 2013), les modèles des risques proportionnels (Qiu, Gu et Chen, 2020), les modèles de physique des dé-

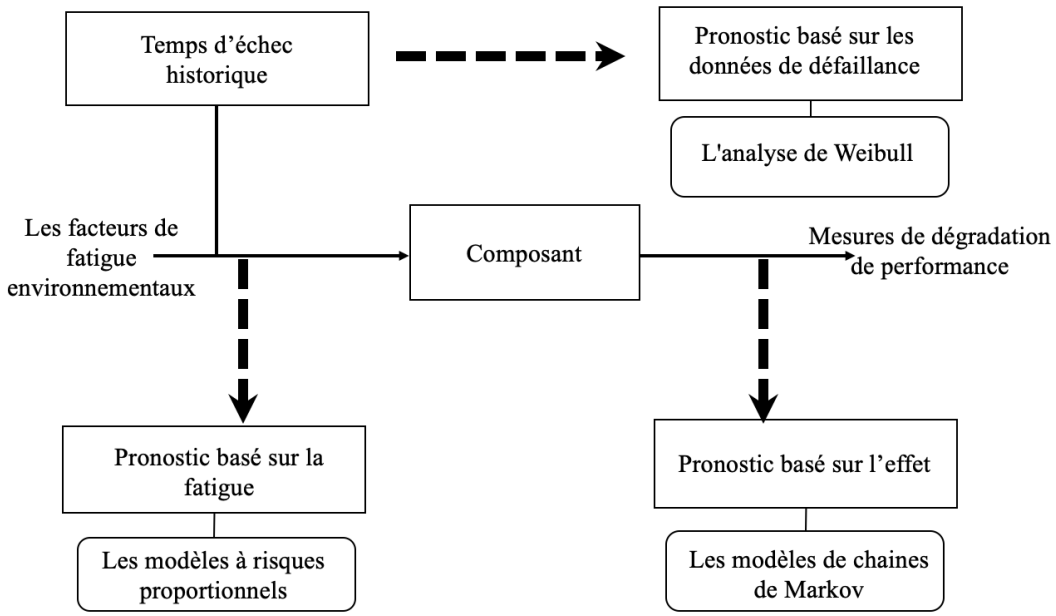


FIGURE 1.9: Méthodologies de pronostic

faillances (LU, LI et LIANG, 2018), l'analyse de régression (WANG et al., 2018), les modèles basés sur les chaînes de Markov (PENG et DONG, 2011 ; GALAGEDARAGE DON et KHAN, 2019), et les modèles de choc (ZHAO et al., 2017) ont tous été suggérés comme méthodologies de pronostic.

1.5 Conclusion

Au 21e siècle, la maintenance prédictive est impulsée par l'intégration des systèmes cyber-physiques avec l'évolution des infrastructures informatiques, des plateformes IoT, du cloud computing, de l'intelligence artificielle et l'analyse des données. Mais généralement, une approche générique de maintenance prédictive repose sur une bonne perception du système, une analyse des connaissances et un raisonnement en fonction du contexte.

La collecte des données significatives telles que les données de mesures (vibrations, thermographie, émissions acoustiques, etc.), les données historiques et les

données conceptuelles permettent de correctement surveiller le système. Pour les questions de la connaissance, la littérature fait la distinction entre deux méthodologies d'analyse, des approches reposant sur les données et d'autres s'appuyant sur les modèles. Les méthodes reposant sur les données sont considérées comme une boîte noire car elles reposent sur l'observation et l'analyse sans une connaissance préalable du système. Grâce à des techniques d'exploration de données, elles sont capables d'identifier la dérive du comportement normal de l'équipement, tandis que les approches reposant sur les modèles nécessitent une connaissance du système pour modéliser et représenter son fonctionnement normal. Quant au raisonnement, il consiste à déterminer la relation de correspondance entre les données et les méthodologies d'analyse.

Une vision globale sur la maintenance prédictive a été présentée dans ce chapitre dont l'architecture est composée de trois grands piliers : l'acquisition des données, le traitement des données et la prise de décision. Ce dernier pilier est un volet de recherche opérationnelle très vaste dont ce manuscrit fait partie, en abordant la détection d'anomalies comme étant une partie intégrante du processus de diagnostic dans le cadre de la prise de décision en matière de maintenance prédictive. De nouvelles contributions dans ce sens sont à suivre dans les chapitres suivants.

Enfin, l'intérêt d'adopter une stratégie de maintenance prédictive est d'éviter des pannes catastrophiques susceptibles de causer des dommages secondaires, des temps d'arrêt, des incidents de sécurité, une perte de production et des coûts élevés associés aux réparations. Mais elle n'est pas toujours évidente à mettre en œuvre car elle implique une planification du matériel, des logiciels, et de la formation des ressources humaines. En effet, la réussite de son implémentation repose sur plusieurs critères cruciaux. On peut citer entre autres : la détermination des composants vitaux de l'équipement, les paramètres indiquant leur détérioration, le choix de l'emplacement des capteurs et les seuils critiques pour chaque variable, la détermination de l'intervalle d'inspection ; soit en continu ou à des intervalles réguliers, qui peuvent être déterminés par avis d'expert ou par le fabricant, reposer sur des recommandations ou performances passées. Le choix d'algorithmes

ou méthodes d'analyse permet ainsi la prise des décisions optimales ainsi que la conception d'un système intelligent qui permet la gestion de cette stratégie de maintenance.

Chapitre 2

Concepts fondamentaux pour la prédiction de situation anormale

Sommaire

Ce chapitre présente une étude synthétique des concepts fondamentaux pour la détection de situations anormales, dans le cadre d'une procédure de diagnostic des équipements industriels. Dans un premier temps, ce chapitre donne un

aperçu global sur le principe de la détection d'anomalies par apprentissage automatique dans les données de perception d'un système donné. Les quatre aspects à prendre en compte lors de la conception ou l'utilisation d'une technique de détection d'anomalies sont détaillés, notamment la nature des données, les mécanismes d'apprentissage automatique, les méthodes d'analyse statistiques et probabilistes appropriées et enfin les techniques d'évaluation des performances et de la robustesse de la méthode utilisée.

Dans un second temps, ce chapitre vise à présenter la théorie du transport optimal comme outil d'analyse et d'exploitation de distributions des données afin de concevoir une méthode de détection d'anomalies précise, robuste et fiable. Il porte d'abord sur la présentation du problème du transport optimal, puis sa formulation pour les distributions des données discrètes, ensuite sa régularisation pour obtenir une solution approchée mais moins complexe à estimer.

2.1 État de l'art pour la détection d'anomalies

La détection d'anomalies est un sujet d'un grand intérêt pour des domaines d'applications divers. Elle vise à identifier à partir des données, les régions dont les comportements ou les modèles ne sont pas conformes aux valeurs attendues. Les comportements inattendus, qui sont significativement différents de ceux du reste des données fournies, sont communément appelés anomalies ou dénommés de différentes manières telles que données aberrantes, nouveautés, exceptions, attaques, erreurs ou violation de service, etc.

Le chapitre précédent a présenté la détection d'anomalies comme une étape fondamentale pour le diagnostic des équipements industriels dans le cadre d'une stratégie de maintenance prédictive. Ainsi, une anomalie dans le comportement du système pourrait signifier un défaut dans un ou plusieurs composants de l'équipement. Dans le domaine de la santé, une image IRM anormale pourrait indiquer la présence de tumeurs malignes (Alaverdyan, 2019 ; Govindaraj et al., 2020 ; Gonçalves et al., 2021). Dans le domaine des réseaux informatiques, un modèle

de trafic anormal pourrait signifier qu'un ordinateur piraté envoie des données sensibles vers une destination non autorisée (Perez et al., 2017 ; Almseidin et al., 2017 ; Vinayakumar et al., 2019), ou encore dans le domaine bancaire, des anomalies dans les données de transactions par cartes de crédit pourraient indiquer un vol de carte de crédits ou d'identité (Awoyemi, Adetunmbi et Oluwadare, 2017 ; Varmedja et al., 2019 ; Adewumi et Akinyelu, 2017).

En sciences des données, la détection d'anomalies se définit usuellement par l'action de discriminer dans un jeu de données caractérisant un système cible, des observations qui ne correspondent pas à la tendance globale représentée par la majorité des observations. Plusieurs facteurs tels que la nature des données d'entrée, la disponibilité ou l'indisponibilité des étiquettes ainsi que les contraintes et exigences induites par le domaine d'application rendent l'identification précise de cette tendance globale ou le problème de détection d'anomalies, dans sa forme la plus générale, difficile à résoudre (Kurniabudi et al., 2019). En effet, la plupart des techniques de détection d'anomalies existantes proposent une formulation précise du problème qui repose sur les facteurs cités ci-dessus en adoptant le cadre théorique de diverses disciplines telles que les statistiques, la théorie de l'information, l'apprentissage automatique et l'exploration de données.

La Fig 2.1 montre les composants clés associés à toute technique de détection d'anomalies. Chaque composant clé de la figure est détaillé dans cette section. D'abord, nous verrons la nature des données sur lesquelles agit un algorithme de détection d'anomalies, puis nous détaillerons les mécanismes pour la conception ou l'utilisation d'une méthode de détection en apprentissage automatique. En effet, l'apprentissage automatique est le domaine directement lié aux contributions principales de mon travail de thèse. Ensuite, ce chapitre présente les catégories de méthodes de détection d'anomalies du point de vue analyse et exploitation des données. Enfin, les outils et les métriques d'évaluations expérimentales des performances de ces méthodes sont présentés en fin du chapitre.

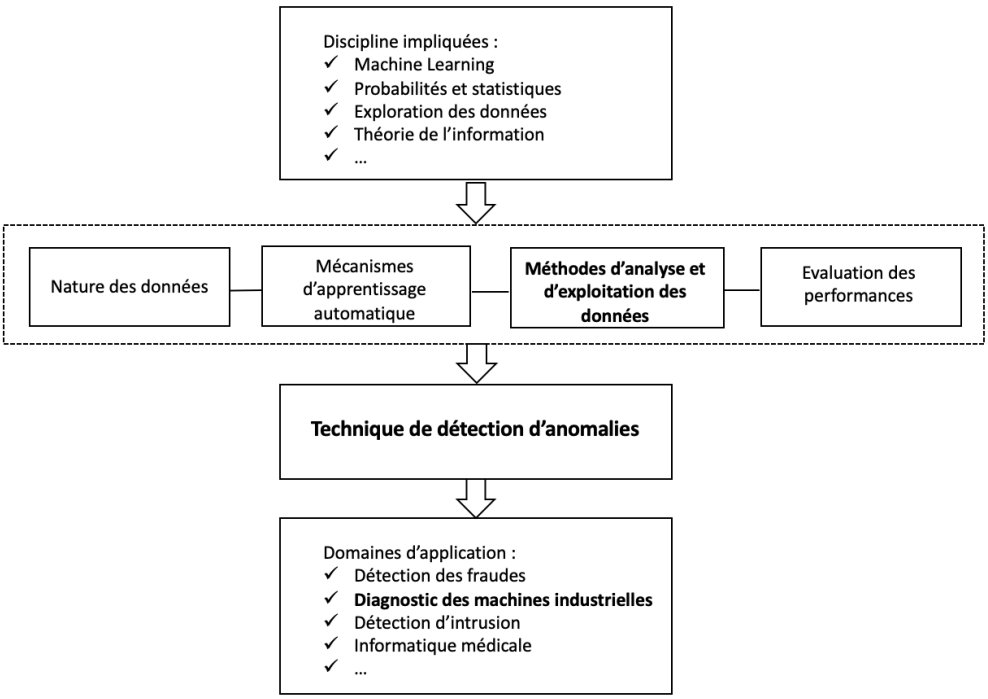


FIGURE 2.1: Cadre théorique de la détection d'anomalies

2.1.1 Formalismes des données

La nature des données d'entrée constitue un aspect majeur de toute technique de détection d'anomalies. L'entrée est généralement une collection d'instances de données, également appelées objets, observations, enregistrements, points, vecteurs, modèles, événements, cas, ou échantillons. Chaque instance de données peut être décrite à l'aide d'un ensemble d'attributs, que l'on appelle également variables, caractéristiques ou dimensions. Les attributs peuvent être de différents types tels que binaires, catégoriels ou continus. Chaque instance de données peut être constituée d'un seul attribut (univarié) ou de plusieurs attributs (multivarié). Dans le cas d'instances de données multivariées, tous les attributs peuvent être du même type ou peuvent être un mélange de différents types de données (FERNANDES et al., 2019).

La représentation des instances de données détermine l'applicabilité des techniques de détection d'anomalies (DJENOURI et al., 2019). Elles sont représentées

sous trois formes : représentations métriques, représentations évolutives et représentations multistructurées. Les représentations métriques sont la forme la plus courante de représentation des données, où chaque objet d'un ensemble de données possède un certain ensemble d'attributs qui permet de fonctionner avec des notions de distance et de proximité. Les représentations évolutives sont des objets bien étudiés : les séquences discrètes, les séries temporelles et les flux de données multidimensionnels. La troisième forme est la représentation multistructurée des données, sous cette forme de représentation les instances de données peuvent être non structurées, semi-structurées ou structurées. Les données graphiques et les données textuelles sont parmi les représentations les plus courantes de ce formulaire de données. La tâche la plus courante avec ce type de données est d'extraire des attributs qui permettent d'utiliser des métriques adaptées (Stupnikov et Kalinichenko, 2018).

Représentations métriques

Les représentations métriques étudient les objets dans un espace choisi. Les méthodes de détection d'anomalies considérant cette forme de données reposent le plus souvent sur des métriques spécifiques telles que la distance entre les objets, la corrélation entre eux et leur distribution dans l'espace. Il s'agit de déterminer les points irréguliers en fonction de ces métriques. Cette forme de représentations est largement utilisée, principalement en raison du fait que presque toutes les entités peuvent être représentées comme un objet structuré, un ensemble d'attributs, et donc comme un point dans un espace particulier (Pang et al., 2018).

Méthodes reposant sur la distance : L'ensemble de base des méthodes qui utilisent la notion de distance comprend les méthodes de clustering, les k plus proches voisins et leurs dérivées. Nous aborderons en détail ces méthodes dans la suite du manuscrit.

Méthodes de corrélations : L'idée des méthodes agissant sur cette forme de données repose sur le concept de corrélation entre les attributs de données (Lever, Krzywinski et Altman, 2017). Cette situation est courante pour les données

réelles car différents attributs peuvent être générés par les mêmes processus. Ainsi, cet effet permet d'utiliser des modèles linéaires et des méthodes reposant sur ces modèles comme la régression linéaire.

Méthodes probabilistes : Pour les méthodes s'appuyant sur des données distribuées de manière probabiliste, l'approche principale consiste à supposer que les données satisfont une loi de distribution. Ainsi, les objets anormaux peuvent être définis comme des objets ne satisfaisant pas à cette distribution (Zhang et al., 2019b). Un exemple classique de ces méthodes est l'EM (Expectation- Maximisation).

Méthodes catégorielles : Les variables catégorielles représentent des types de données qui peuvent être divisés en groupes. Différentes approches probabilistes peuvent être utilisées pour le traitement des données catégorielles. Certaines méthodes utilisant des distances peuvent être partiellement modifiées pour pouvoir s'appliquer sur ce type de données. Cependant, pour les méthodes de détection d'anomalies, une approche appropriée agissant sur des données catégorielles consiste à traduire les attributs catégoriels en attributs continus (Taha et Hadi, 2019).

Représentations évolutives

Les données en évolution se distinguent par leur représentation temporelle. La caractéristique temporelle peut être discrète ou continue, de sorte que les données peuvent être présentées en séquences ou en séries temporelles.

Séquences discrètes : Les séquences discrètes ou symboliques sont des ensembles ordonnés d'événements (Das, Matthews et Lawrence, 2011). Souvent, on s'intéresse à détecter des anomalies dans des séquences discrètes pour trouver d'éventuelles intrusions, fraudes, failles, ou contaminations. De nombreux problèmes nécessitent la détection d'anomalies dans des séquences discrètes, comme les séquences biologiques telles que les séquences d'ADN (Kukita et al., 2013), les séquences de protéines (Nasr et al., 2021), les données des capteurs des systèmes opérationnels (Kyritsis et al., 2017). Les séquences sont collectées pendant

le fonctionnement du système grâce à de multiples capteurs discrets. Il est par exemple possible d'enregistrer les séquences de clics de navigation à partir de sites internet. Les symboles dans de telles séquences correspondent à des liens cliqués ou à des catégories auxquelles les liens appartiennent. Il existe plusieurs façons de déterminer une valeur aberrante dans les données présentées sous forme de séquence discrète. Dans l'étude Chandola, Banerjee et Kumar (2010), les méthodes sont divisées en trois groupes : méthodes reposant sur des séquences, reposant sur des sous-séquences contiguës et reposant sur des motifs. Le premier groupe comprend les techniques utilisant des fonctions noyaux, les techniques de fenêtrage, les techniques markoviennes. Les méthodes de sous-séquence contiguë incluent les techniques d'annotation des fenêtres et les techniques utilisant la segmentation. Les méthodes reposant sur des motifs regroupent les techniques de correspondance de sous-chaîne, de correspondance de sous-séquence et de correspondance de permutation.

Séries temporelles : Une série temporelle est le signal produit par un système en fonctionnement. Elle peut être utilisée pour représenter des données de capteurs, des formes d'objets ou des trajectoires d'objets en mouvement, elle peut en outre être appliquée pour identifier des objets et des phénomènes dans des applications telles que la vidéosurveillance, la reconnaissance de signaux, la surveillance des systèmes industriels, etc. Si les données dépendent fortement du temps, alors la prédiction des données à venir et l'analyse des tendances actuelles s'imposent. La façon la plus courante de déterminer une valeur aberrante ou une nouveauté est un changement de tendance abrupte (Wang et Wang, 2019). Les algorithmes de détection d'anomalies courants apprennent des modèles de normalité en ajustant des modèles à des instances d'apprentissage considérées comme normales. Les instances précédemment inédites sont testées en mesurant leur score d'anomalie par rapport aux modèles appris, et un seuil est utilisé pour déterminer les nouveautés. L'objectif des algorithmes de détection d'anomalies est de trouver des limites de décision efficaces des normalités qui donnent la meilleure précision pour les nouveautés prédites. La détection d'anomalies dans les séries temporelles sera abordée en détail dans le chapitre suivant.

Données multi-structurées

Parfois, les données sont présentées sous une forme plus complexe qu'un tableau numérique "attribut/valeur". Dans ce cas, il est important de comprendre ce qu'est une anomalie en utilisant une méthode d'analyse appropriée. Les formes de représentation les plus courantes dans cette catégorie de données sont les données textuelles et les graphes.

Graphes : De nombreux types de données contiennent des relations temporelles ou spatiales entre des éléments qui seraient mieux représentés sous forme de graphes. Par exemple, en utilisant un graphe représentant des transactions par carte de crédit, il est possible de créer des relations entre les transactions se produisant à moins d'un kilomètre ou à une seconde les unes des autres. Ces types de relations s'avèrent utiles pour certaines applications et seraient difficiles à représenter sous une autre forme (Zhong et al., 2019). On peut distinguer deux groupes dans ces méthodes, selon qu'elles s'appliquent sur des graphes statiques ou dynamiques. La détection d'anomalies dans les graphes statiques consiste à repérer les entités de réseau anormales (par exemple, les noeuds, les arêtes, les sous-graphes) compte tenu de la structure entière du graphe. La détection d'événements dans les graphes dynamiques est fonction du type d'événement à détecter : les méthodes peuvent utiliser des caractéristiques (par exemple, des nœuds, des arêtes, des poids d'arête), s'appuyer sur une décomposition, utiliser du clustering ou utiliser des fenêtres d'analyse. Les méthodes de détection pour les deux groupes sont alors constituées des définitions générales des anomalies et de leurs techniques de détection proposées pour les données métriques.

Données textuelles : Plusieurs tâches majeures peuvent être distinguées pour la détection d'anomalies dans l'analyse de textes : la recherche des textes anormaux (Mahapatra, Srivastava et Srivastava, 2012) comme la détection des spams et la recherche des textes non standards, ou bien la détection des nouveautés. La question principale pour ce formalisme de données est de représenter les textes en données métriques pour que les méthodes définies précédemment soient utilisables.

2.1.2 Mécanismes de l'apprentissage automatique

Modélisation d'un problème d'apprentissage automatique

Un algorithme d'apprentissage automatique est un algorithme capable d'apprendre à partir des données. Goodfellow et al. (2016) propose la définition suivante : *"On dit qu'un programme informatique apprend de l'expérience E en ce qui concerne une classe de tâches T et une mesure de performance P, si ses performances aux tâches de T, telles que mesurées par P, s'améliorent avec l'expérience E."* Ainsi on peut dire qu'un programme apprend de ses expériences afin de maximiser la probabilité d'obtenir le résultat attendu. Les algorithmes d'apprentissage automatique peuvent être globalement découpés en familles supervisées, semi-supervisées ou non-supervisées en fonction du type d'expérience qui est mis à disposition pendant le processus d'apprentissage.

Les algorithmes d'apprentissage supervisé font l'expérience d'observer un ensemble de données contenant des caractéristiques sous forme de vecteurs $\mathbf{x}$ associés à des valeurs $\mathbf{y}$ sous l'appellation d'étiquettes ou de cibles, afin d'apprendre de prédire $\mathbf{y}$ en fonction de $\mathbf{x}$, souvent en estimant $p(\mathbf{x}|\mathbf{y})$. Les algorithmes d'apprentissage semi-supervisé supposent que les données d'apprentissage ont des instances étiquetées uniquement pour la classe normale. Les algorithmes d'apprentissage non supervisé disposent d'un ensemble de données contenant uniquement les caractéristiques $\mathbf{x}$, puis apprennent la distribution $\mathbb{D}(\mathbf{x})$ ou certaines propriétés utiles de cette distribution. Le terme apprentissage supervisé provient de la vision de la cible $\mathbf{y}$ fournie par un instructeur qui montre au système d'apprentissage automatique ce qu'il faut faire. Dans l'apprentissage non supervisé, il n'y a pas d'instructeur, et l'algorithme utilise des hypothèses sur la distribution des données pour faire des prédictions.

Comme indiqué par Goodfellow et al. (2016), l'apprentissage non supervisé et l'apprentissage supervisé ne sont pas des termes formellement définis. De nombreuses approches d'apprentissage automatique peuvent être utilisées pour effectuer les deux tâches. C'est le cas pour les approches probabilistes utilisant la formule des probabilités composées pour un vecteur $\mathbf{x} \in \mathbf{R}^n$. La distribution jointe

peut être décomposée en :

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1}) \ . \tag{2.1}$$

Cette décomposition permet de transformer un problème non supervisé de la modélisation $p(\mathbf{x})$ en n problèmes d'apprentissage supervisé qui peuvent être résolu séparément. Par ailleurs, le problème d'apprentissage supervisé $p(\mathbf{x}|\mathbf{y})$ peut être résolu en utilisant des approches d'apprentissage non supervisé pour apprendre la distribution conjointe $p(\mathbf{x}, \mathbf{y})$ et en déduisant :

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{\sum_{y'}(\mathbf{x}, y')} \tag{2.2}$$

De nombreux types de tâches peuvent être résolus avec l'apprentissage automatique dont le résultat en sortie du modèle d'apprentissage dépend du problème posé. Ce peut être une régression, une classification, une transcription, une traduction, une détection, etc.

Afin de mieux caractériser le concept de modélisation d'un problème d'apprentissage automatique, l'exemple le plus simple est la régression linéaire. Le même exemple servira à présenter différents mécanismes d'apprentissage automatique dans cette partie. Le but d'un modèle de régression est de construire un système qui peut prendre un vecteur $\mathbf{x} \in \mathbf{R}^n$ en entrée et prédire la valeur d'un scalaire $\hat{y} \in \mathbf{R}$ en sortie. Dans le cas de la régression linéaire, la sortie est une fonction linéaire de l'entrée. Soit la valeur prédite par le modèle d'apprentissage dont les poids sont les paramètres de celui-ci. La sortie du modèle est définie par :

$$\hat{y} = w^T * \mathbf{x} \ , \tag{2.3}$$

où w est le vecteur des paramètres.

Après avoir abordé l'expérience E (les différents types de données) et la tache T (les approches supervisées ou non-supervisées), il reste à définir la performance P. Pour un modèle donné, l'apprentissage vise à déterminer les valeurs pour toutes

les pondérations et le biais à partir d'exemples étiquetés pour optimiser un critère choisi. Dans l'apprentissage supervisé, un algorithme crée un modèle en examinant de nombreux exemples, puis en tentant de trouver un modèle qui minimise une fonction de coût. Ce processus est appelé minimisation du risque empirique. La fonction de coût évalue la qualité de la prévision du modèle pour un exemple donné. Si la prédiction du modèle est parfaite, le coût est nul. Sinon, le coût est supérieur à zéro. Le but de l'entraînement d'un modèle est de trouver un ensemble de pondérations et de biais pour lesquels la perte est faible en moyenne sur tous les exemples.

Une façon de mesurer la performance P est la perte quadratique, autrement appelée erreur quadratique moyenne (MSE), qui s'exprime comme :

$$MSE = \frac{1}{N} \sum_{(\mathbf{x},y)\in S} (y - \hat{y})^2 \ , \tag{2.4}$$

avec S est un ensemble de données contenant de nombreux exemples étiquetés, qui sont des paires $(\mathbf{x}, y)$ et N le nombre d'exemples dans S. Bien que l'erreur MSE soit couramment utilisée en apprentissage automatique, ce n'est ni la seule fonction possible, ni la meilleure en toutes les circonstances.

Capacité, sous-apprentissage et sur-apprentissage

Un autre mécanisme important de l'apprentissage automatique est la capacité d'un modèle d'apprentissage à construire des connaissances pertinentes sur la base d'un ensemble initial d'exemples d'entraînement.

Conceptuellement, le *sous-apprentissage* est associé à l'incapacité d'un algorithme d'apprentissage automatique à déduire des connaissances valides à partir des données d'entraînement initiales. Contrairement à cela, le *sur-apprentissage* est associé à des modèles qui créent des hypothèses bien trop spécialisées ou abstraites pour être valides en pratiques (Aalst et al., 2010).

Un scénario d'apprentissage automatique typique repose sur un ensemble de données initial utilisé pour entraîner et tester les performances d'un algorithme,

souvent réparti en 80% pour l'entraînement du modèle et 20% pour le test. Pendant la phase d'entraînement, le modèle produit un certain écart par rapport aux données d'entraînement, appelé souvent erreur d'entraînement. De même, l'écart produit pendant la phase de test est appelé erreur de test. De ce point de vue, les performances d'un modèle d'apprentissage automatique peuvent être jugées sur sa capacité à accomplir deux tâches simultanément :

— réduire l'erreur d'entraînement,

— réduire l'écart entre les erreurs d'entraînement et de test.

Fondamentalement, un sous-apprentissage se produit lorsque le modèle échoue à la première tâche et n'est pas capable d'obtenir une erreur suffisamment faible à partir de l'ensemble d'apprentissage. Le sur-apprentissage se produit alors lorsqu'un modèle échoue à la deuxième règle et que l'écart entre les erreurs de test et d'entraînement est trop important (Koehrsen, 2018). Ceci est illustré dans la Fig 2.2.

Dans cet exemple, les fonctions non linéaires sont approximées par des caractéristiques polynomiales de différents degrés dans le cadre d'une regression linéaire. La figure montre une fonction sinusoïdale à approximer. Ainsi, les échantillons de la fonction réelle et les approximations des différents modèles sont affichés. Le premier cas (à gauche) illustre un sous-apprentissage, le modèle ici ne peut décrire que des droites, il n'est pas assez complexe car c'est un polynôme du premier degré. Donc, cette fonction n'est pas suffisante pour modéliser les échantillons d'apprentissage. Dans le cas du milieu, un polynôme de degré 4 se rapproche presque parfaitement de la vraie fonction, ce bon degré de complexité du modèle permet un apprentissage correct. Cependant, pour des degrés plus élevés comme sur le dernier cas de la figure (à droite), le modèle explique extrêmement bien la plupart des points appartenant aux données d'entraînement, c'est-à-dire qu'il a appris le bruit des données d'entrainement, il s'agit donc d'un sur-apprentissage.

Optimisation des hyper-paramètres

Un hyper-paramètre est une configuration externe au modèle dont la valeur ne peut être estimée à partir des données. Il est souvent utilisé dans les proces-

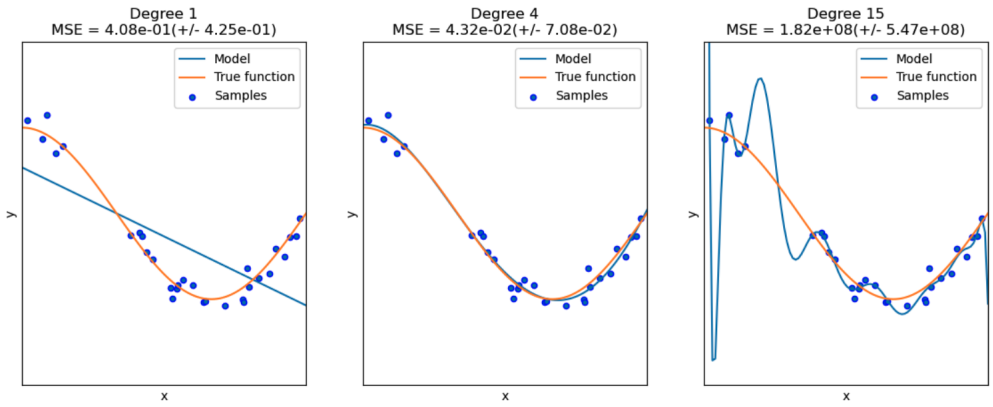


FIGURE 2.2: Illustration du sous-apprentissage et sur-apprentissage dans le cas d'une regression linéaire - Image tirée de la documentation de scikit-learn

sus pour aider à estimer les paramètres du modèle d'apprentissage, son optimisation consiste donc, à contrôler le processus d'apprentissage c'est-à-dire, trouver la configuration qui offre les meilleures performances mesurées sur un ensemble de validation (GOODFELLOW et al., 2016).

Il est courant de séparer un jeu de données en trois ensembles : un ensemble d'entraînement pour entraîner le modèle, un ensemble de test pour l'évaluer et un ensemble de validation pour fixer les hyper-paramètres. Cette approche est tout à fait adéquate avec ce qui a été abordé précédemment sur la modélisation d'un problème en apprentissage automatique. Un hyper-paramètre ne peut pas être choisi sur un ensemble d'entraînement puisqu'il va minimiser l'erreur sur cet ensemble de données, conduisant ainsi à un sur-apprentissage. Il ne peut pas non plus être choisi sur un ensemble de test car ce dernier n'est utilisé que pour estimer l'erreur que fait un modèle sur des données qu'il n'a jamais vues. Cependant, les jeux de données sont parfois trop petits pour pouvoir être séparés en trois ensembles. Un ensemble d'entraînement trop petit mène à un sous-apprentissage. Un ensemble de test trop petit ne permet pas de généraliser l'erreur de test et un ensemble de validation trop petit ne permet pas de s'assurer que les hyper-paramètres choisis sont optimaux. Pour remédier à ce problème, une méthode courante est d'avoir recours à la validation croisée (ARLOT, CELISSE et al., 2010 ;

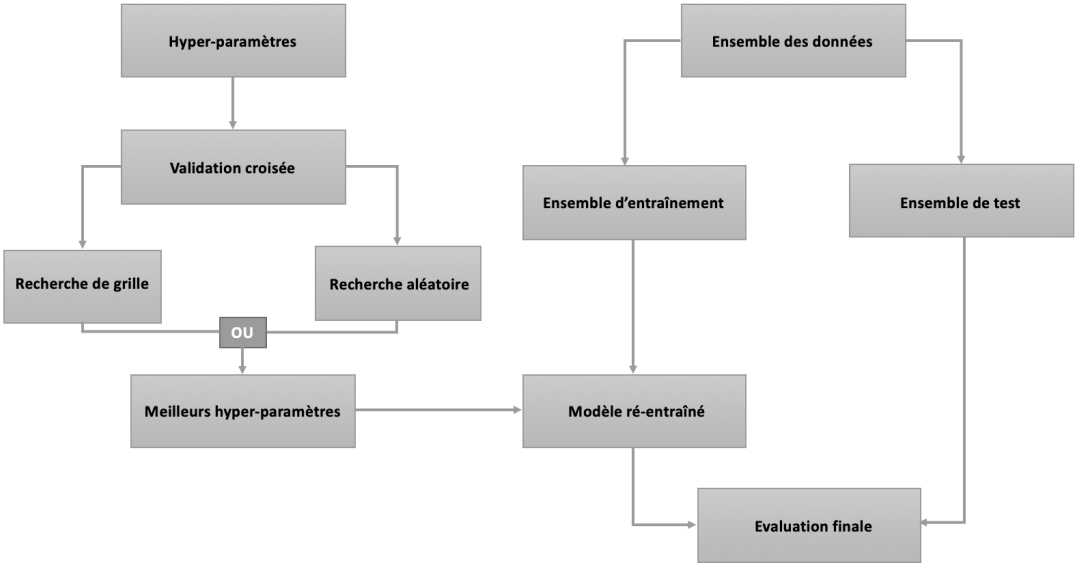


FIGURE 2.3: Répartition d'un ensemble des données avec une validation croisée pour une optimisation des hyper-paramètres - Adaptée de la documentation Scikit-learn

KALE, KUMAR et VASSILVITSKII, 2011 ; BERGMEIR et BENÍTEZ, 2012), qui est illustrée sur la Fig 2.3.

Dans ce processus, les données sont divisées en k sous-ensembles, de sorte qu'à chaque fois, l'un des k sous-ensembles est utilisé comme ensemble de test/ensemble de validation et les autres $k-1$ sous-ensembles sont réunis pour former un ensemble d'entraînement tel qu'illustré sur la Fig 2.4 . L'estimation de l'erreur est moyennée sur tous les k essais pour obtenir l'efficacité totale du modèle. Chaque point de données se trouve dans un ensemble de validation exactement une seule fois et se trouve dans un ensemble d'apprentissage $k-1$ fois. L'échange des données entre ensembles d'apprentissage et de test améliore l'efficacité de cette méthode.

Pour revenir aux questions d'optimisation des hyper-paramètres, les modèles peuvent avoir de nombreux hyper-paramètres à fixer, l'idée donc est de trouver la combinaison la plus optimale pour une meilleure performance du modèle d'apprentissage. Deux stratégies sont possibles : la recherche exhaustive et la recherche aléatoire.

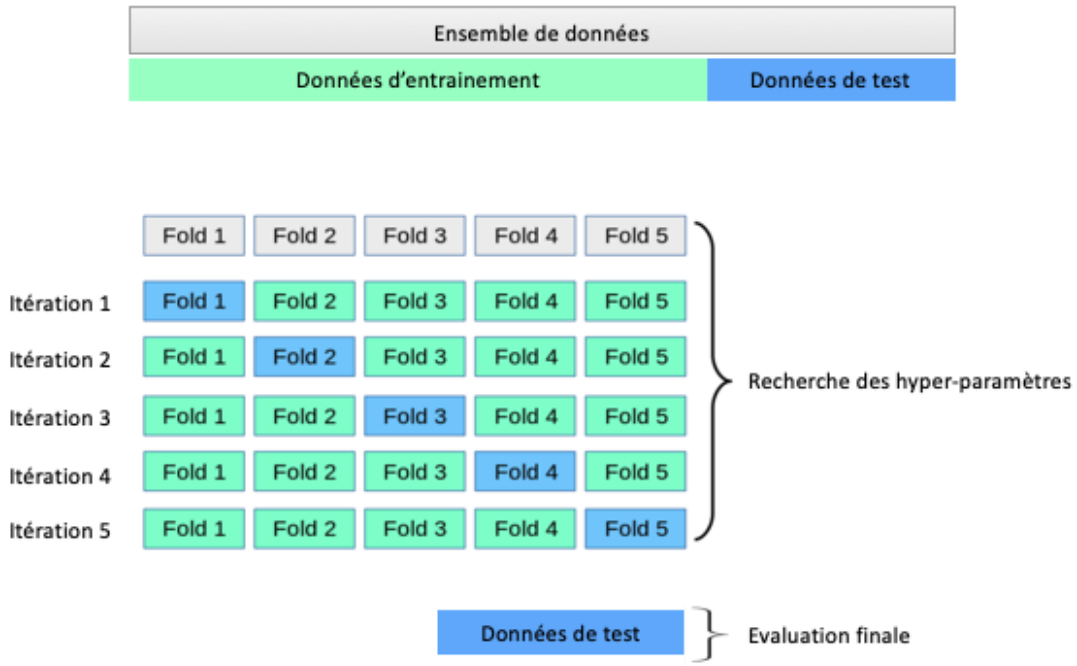


FIGURE 2.4: Approche de la validation croisée - Adaptée de la documentation Scikit-learn

Recherche exhaustive d'hyper-paramètres : C'est une approche de sélection d'hyper-paramètres qui construit et évalue méthodiquement un modèle pour chaque combinaison de paramètres spécifiés. Elle génère de manière combinatoire des candidats à partir d'un ensemble de valeurs de paramètres, ensuite toutes les combinaisons possibles de valeurs de paramètres sont évaluées et la meilleure combinaison est retenue (KRSTAJIC et al., 2014).

Recherche aléatoire d'hyper-paramètres : L'idée de recherche aléatoire d'hyper-paramètres a été proposée par BERGSTRA et BENGIO (2012). Elle diffère d'une recherche par grille dans le sens où une recherche aléatoire sur les paramètres repose sur la définition d'une distribution statistique pour chaque hyper-paramètre à partir de laquelle les valeurs peuvent être échantillonnées aléatoirement. Cela présente deux avantages principaux par rapport à une recherche de grille (ROBERTS et al., 2017) :

— Le coût de calcul peut être choisi indépendamment du nombre de para-

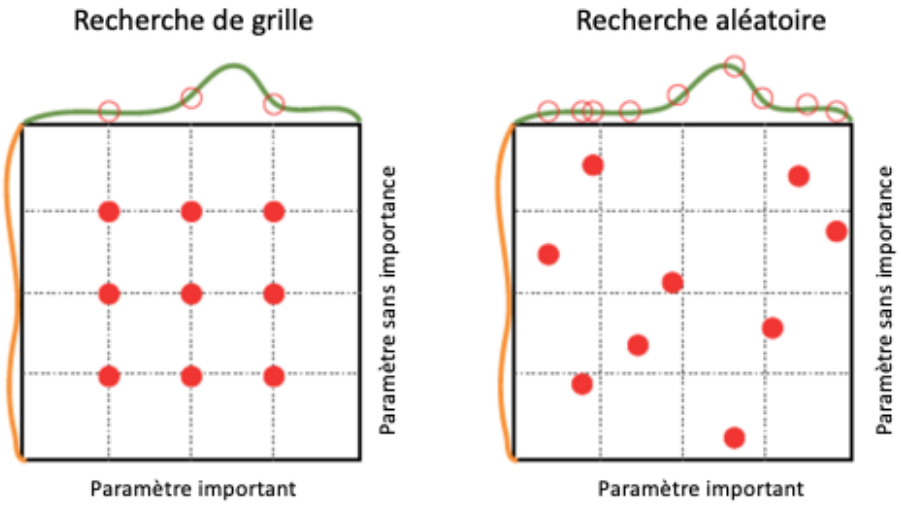


Figure 2.5: Techniques de recherche d'hyper-paramètres

mètres et de valeurs possibles.

— L'ajout de paramètres qui n'influencent pas les performances de l'algorithme ne diminue pas l'efficacité.

La Fig 2.5 montre la recherche de l'hyper-paramètre qui a le plus d'influence sur l'optimisation du score d'un modèle d'apprentissage dans un espace d'hyper-paramètres. Les distributions affichées sur chaque axe représentent le score du modèle. Dans chaque cas, neuf modèles différents sont évalués. La stratégie de recherche de grille passe un temps redondant à explorer un même paramètre qui n'a pas d'importance, car chaque hyper-paramètre est isolé pour trouver la meilleure valeur possible tout en maintenant constants les autres hyper-paramètres. Pour les cas où l'hyper-paramètre étudié a peu d'effet sur le score du modèle résultant, cela entraîne une perte d'efficacité. Tandis que la recherche aléatoire a un pouvoir exploratoire bien amélioré et peut se concentrer sur la recherche de la valeur optimale pour l'hyper-paramètre critique.

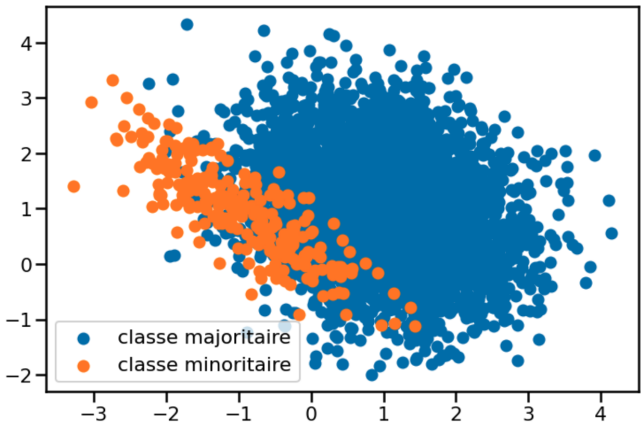


Figure 2.6: Exemple de classes déséquilibrées

Apprentissage sur des classes non-équilibrées

Les classes déséquilibrées sont un problème courant voir inévitable en apprentissage automatique. Le déséquilibre se produit lorsqu'une ou plusieurs classes ont des proportions très faibles dans les données d'apprentissage par rapport aux autres classes. Autrement dit, un ensemble de données est déséquilibré si les catégories de classification ne sont pas également représentées. La classe avec des exemples abondants est appelée classe majeure ou majoritaire, tandis que la classe avec peu d'exemples est appelée classe mineure ou minoritaire. Cela rend le modèle d'apprentissage automatique biaisé en faveur de la classe majoritaire. Ce type de situation provoque une mauvaise classification des classes minoritaires (Krawczyk, 2016). Ce problème peut entraver le bon fonctionnement des algorithmes d'apprentissage. Un exemple de classes déséquilibrées est montré dans la Fig 2.6.

De nombreux problèmes de classification peuvent avoir un important déséquilibre dans la distribution des classes. C'est particulièrement le cas dans les tâches de détection d'anomalies. Bien que cette description puisse sembler impliquer que les déséquilibres existent uniquement pour les problèmes de classification binaires, il faut noter qu'il existe des données multiclasses dans lesquelles des déséquilibres existent (Sun, Kamel et Wang, 2006 ; Zhou et Liu, 2010 ; Chen, Lu et Kwok, 2006 ; Abe, Zadrozny et Langford, 2004). La suite de la section se concentre plutôt sur le problème d'apprentissage déséquilibré à deux classes.

La littérature propose plusieurs directives pour travailler avec des ensembles de données déséquilibrés ; le choix d'une bonne métrique de performance et les méthodes de rééchantillonnage sont les plus courantes (Hernandez, Carrasco-Ochoa et Martínez-Trinidad, 2013 ; Bunkhumpornpat, Sinapiromsaran et Lursinsap, 2012 ; Lemaître, Nogueira et Aridas, 2017). Nous abordons dans cette partie les méthodes de rééchantillonage alors que les métriques de performance viendront dans la sous-section 2.1.4.

Pour une présentation claire, nous considérons un ensemble de données d'apprentissage S avec N exemples. Nous définissons $S = \{(\mathbf{x}_i, y_i)\}_{i=1,\dots,N}$, avec $\mathbf{x}_i \in X$ est une instance dans l'espace de caractéristiques à n dimensions $X = \{f_1, f_2, \dots, f_n\}$, et $y_i \in Y = \{1, \dots, C\}$ est la classe d'étiquette associée à l'instance $\mathbf{x}_i$. $C = 2$ dans le cas de l'apprentissage à deux classes. De plus, nous définissons les sous-ensembles $S_{min} \subset S$ et $S_{maj} \subset S$ avec S_{min} le sous-ensemble de la classe minoritaire et S_{maj} le sous-ensemble de la classe majoritaire dans S. Enfin, tous les ensembles générés à partir des procédures d'échantillonnage sur S sont étiquetés E, avec des sous-ensembles disjoints E_{min} et E_{maj} représentant respectivement les échantillons minoritaires et majoritaires de E.

Généralement, l'utilisation de méthodes d'échantillonnage dans des applications d'apprentissage consiste à modifier un ensemble de données déséquilibré par certains mécanismes afin d'assurer une répartition équilibrée. Un large choix de méthodes de rééchantillonnage est fourni par la littérature, chacune offrant ses propres avantages et inconvénients. Ici, nous allons nous concentrer sur quelques-unes des méthodes les plus populaires et qui sont utilisées pour les données de l'industrie.

Sur-échantillonnage aléatoire : Le mécanisme de sur-échantillonnage aléatoire vise à ajouter un ensemble E échantillonné à partir de la classe minoritaire. Pour un ensemble d'exemples minoritaires choisis aléatoirement dans S_{min}, une augmentation de l'ensemble original S est faite en dupliquant les exemples sélectionnés et en les ajoutant à ce dernier. C'est ainsi que le nombre total d'exemples dans S_{min} est augmenté de $|E|$ et l'équilibre de distribution des classes dans S est ajusté en conséquence. Ceci fournit un mécanisme pour faire

varier le degré d'équilibre de la distribution des classes à n'importe quel niveau souhaité (Junsomboon et Phienthrakul, 2017).

Sous-échantillonnage alétaoire : Contrairement au sur-échantillonnage, le sous-échantillonnage aléatoire supprime les données de l'ensemble de données d'origine. En particulier, il permet de sélectionner aléatoirement un ensemble d'exemples de la classe majoritaire S_{maj} et retirer ces échantillons de S de sorte que $|E| = |S_{maj}| + |S_{min}| - |E|$. Par conséquent, le sous-échantillonnage fournit une autre méthode simple pour ajuster l'équilibre de l'ensemble de données d'origine S.

Échantillonnage synthétique avec génération de données : La technique de sur-échantillonnage des minorités synthétiques (SMOTE) est une méthode puissante qui a connu un grand succès dans diverses applications (He et Garcia, 2009). L'algorithme SMOTE crée des données artificielles en s'appuyant sur les similitudes d'espace de caractéristiques entre les exemples minoritaires existants. Spécifiquement, pour le sous-ensemble $S_{min} \in S$, les K voisins les plus proches pour chaque exemple $\mathbf{X}_i \in S_{min}$ sont considérés pour un entier spécifié K ; les K plus proches voisins sont définis comme les K éléments de S_{min} dont la distance euclidienne entre k et x_i présente la plus petite amplitude pour les n dimensions de l'espace des caractéristiques X. Afin de créer un échantillon synthétique, il faut sélectionner aléatoirement l'un des K plus proches voisins, puis multiplier la différence de vecteur des caractéristiques correspondante par un nombre aléatoire compris entre [0,1] et ajouter ce vecteur à x_i. L'instance synthétique résultante selon l'équation 2.5 est un point le long du segment de droite joignant x_i et le K plus proche voisin x_i sélectionné aléatoirement.

$$x_{nv} = \mathbf{x}_i + (\hat{\mathbf{x}}_i - \mathbf{x}_i)\delta \ , \tag{2.5}$$

avec $\hat{\mathbf{x}}_i$ est l'un des K plus proches voisins pour $\mathbf{x}_i$, $\hat{\mathbf{x}}_i \in S_{min}$, et $\delta \in [0, 1]$ est un nombre aléatoire.

Échantillonnage reposant sur les clusters : Les algorithmes d'échan-

tillonnage reposant sur les clusters sont particulièrement intéressants de par leur flexibilité, qui n'est pas égalée dans les algorithmes d'échantillonnage simples et synthétiques. Ces algorithmes peuvent être adaptés pour cibler des problèmes très spécifiques. Parmi ces algorithmes, l'algorithme de sur-échantillonnage basé sur les clusters (CBO) en utilisant la technique de clustering K-means. Cette procédure prend un ensemble aléatoire de K exemples de chaque cluster (pour les deux classes) et calcule le vecteur caractéristique moyen de ces exemples, qui est désigné comme le centre du cluster. Ensuite, les exemples d'apprentissage restants sont présentés un par un et pour chaque exemple, le vecteur de distance euclidienne entre celui-ci et chaque centre du cluster est calculé. Chaque exemple d'apprentissage est ensuite affecté au cluster qui présente la plus petite magnitude de vecteur de distance. Enfin, toutes les moyennes de cluster sont mises à jour et le processus est répété jusqu'à ce que tous les exemples soient épuisés. Le détail de cet algorithme est fourni par Jo et Japkowicz (2004).

2.1.3 Vue d'ensemble méthodes de détection

La Fig 2.7 présente une représentation des principales approches de détection d'anomalies à savoir, les techniques statistiques, la classification, les algorithmes de plus proches voisins, le clustering, la théorie spectrale et la théorie de l'information. Dans cette section nous nous intéressons aux quatre premières catégories de méthodes pour la suite du manuscrit.

Techniques statistiques

Les techniques statistiques utilisent un modèle statistique pour classer les instances de données. Le modèle est construit pour refléter la distribution des données d'apprentissage et les nouvelles d'instances sont classées en fonction de leur adéquation avec le modèle. Une instance de données générée à partir du même processus stochastique que les données d'apprentissage s'adaptera bien au modèle, tandis que les instances de données générées à partir d'un processus différent ne s'adapteront pas au modèle et seront considérées comme des anomalies. Il existe

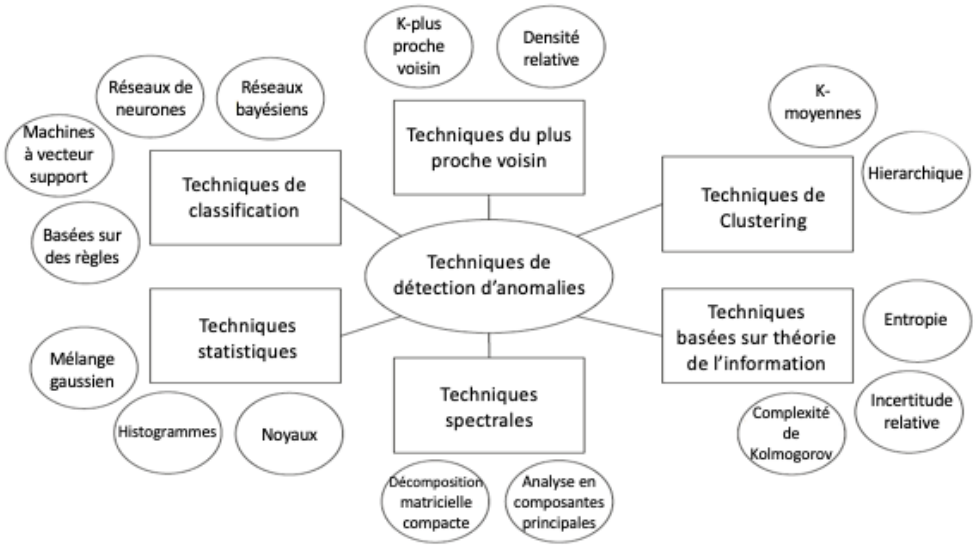


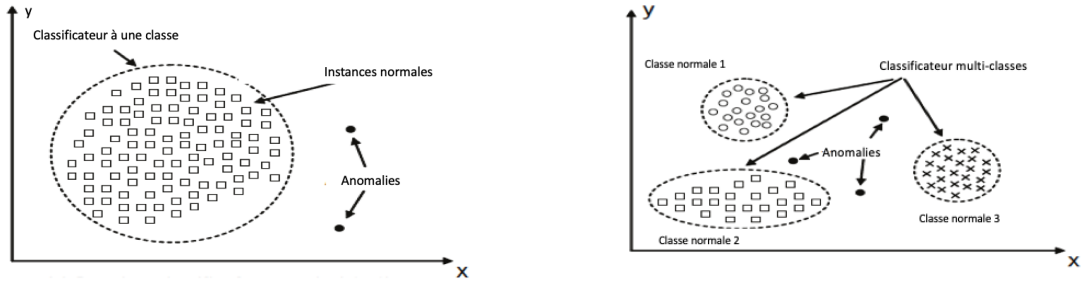
Figure 2.7: Vue d'ensemble des techniques de détection d'anomalies

deux groupes de techniques statistiques pour ce contexte (Chandola, Banerjee et Kumar, 2010) : paramétriques et non-paramétriques. Lors de l'utilisation d'une technique paramétrique, la distribution sous-jacente des données doit être connue ou supposée connue. Si l'hypothèse concernant la distribution est incorrecte, les performances de la technique diminueront. Il existe un certain nombre de distributions courantes – Gaussienne, Poisson, binomiale, etc. – qui peuvent être appliquées à de nombreux types d'ensembles de données. Des exemples de techniques paramétriques incluent les modèles de régression, les modèles utilisant des fonctions noyaux et les techniques reposant sur un mélange de distributions paramétriques. Si la distribution des données est inconnue, des techniques non paramétriques peuvent être utilisées. Une technique non paramétrique ne suppose pas de modèle a priori. Au lieu de cela, le modèle est construit à partir des instances de données dans l'ensemble de données d'apprentissage. Les histogrammes sont un exemple de techniques non paramétriques, où la fréquence des instances de données dans chaque groupe est utilisée pour classer les nouvelles instances de données.

Techniques de classification

La classification est la tâche d'assigner un objet à une des catégories déjà prédéfinies. Cela se fait généralement en apprenant une fonction de classification qui associe chaque ensemble d'attributs à une étiquette de classe prédéfinie. Des techniques de classification peuvent être utilisées de deux manières : pour construire des modèles descriptifs et pour construire des modèles de prédictifs. Un modèle descriptif peut être utilisé pour expliquer les différences entre les instances de données appartenant à différentes classes, tandis que les modèles prédictifs peuvent attribuer une étiquette de classe à une instance de données en fonction des attributs. Dans les techniques de détection d'anomalies reposant sur la classification, des modèles prédictifs sont construits à partir de données d'apprentissage étiquetées. Ces modèles sont ensuite utilisés pour classer les nouvelles instances de données comme appartenant à la classe normale ou à la classe d'anomalies. La littérature fournit des techniques de détection d'anomalies reposant sur la classification qui peuvent être considérées comme des classificateurs à une ou plusieurs classes. La Fig 2.8 montre des exemples des deux types. La principale différence est que pour les classificateurs à une classe, toutes les instances de données d'apprentissage sont supposées être étiquetées normales, tandis que pour les classificateurs à plusieurs classes, les instances de données dans les données d'apprentissage peuvent avoir un certain nombre d'étiquettes normales différentes. Une autre technique souvent utilisée pour la classification multi-classes est celle des réseaux bayésiens (Chaitra et Kumar, 2018). Lorsqu'il s'agit de données univariées et conditionnellement indépendantes, le classifieur naïf de Bayes peut être utilisé (Bielza et Larranaga, 2014). Si les attributs dans les ensembles de données sont conditionnellement dépendants, des réseaux bayésiens plus complexes peuvent être utilisés (Wang et al., 2019 ; Krishnakumar et Abdou, 2020). Les machines à vecteur support (SVM) est une autre technique qui peut être utilisée comme classificateur à une classe (One-class SVM). Un one-class SVM peut apprendre une région qui encapsule toutes les instances de données normales dans l'ensemble de données d'apprentissage. Dans l'étape de test, les nouvelles instances de données sont classées comme normales si elles tombent dans la région apprise, sinon elles sont classées comme

anormales. Une autre catégorie de techniques de classification sont les techniques à base de règles, elles sont couramment utilisées pour de nombreuses applications et ont été appliquées à la fois dans des environnements à une classe et à plusieurs classes. Comme pour la plupart des techniques décrites précédemment, les techniques à base de règles fonctionnent en deux étapes. Dans la première étape, un algorithme d'apprentissage de règles tel que les arbres de décisions (KOTSIANTIS, 2013) est utilisé pour générer un certain nombre de règles décrivant l'ensemble de données d'apprentissage. À l'étape suivante, les nouvelles instances de données sont évaluées par chaque règle pour trouver la règle qui capture le mieux l'instance de données. Chaque règle a une valeur de support et de confiance correspondante.



(a) Classification à une classe (b) Classification multi-classes

FIGURE 2.8: Deux approches de classification pour la détection d'anomalies

Techniques utilisant les plus proches voisins

Les techniques utilisant les plus proches voisins reposent sur l'hypothèse que les instances de données normales se produisent dans des zones denses, tandis que les instances de données anormales se produisent dans les zones de faible densité loin des autres instances de données. Un concept central dans les techniques utilisant le plus proche voisin est la mesure de distance ou de similarité. La mesure de distance décrit à quel point deux objets sont proches l'un de l'autre. Plus les objets sont proches, plus ils se ressemblent. Il est donc très important de choisir une mesure de distance appropriée au problème et à la nature de l'ensemble de données. Pour les ensembles de données avec des attributs continus, il est courant d'utiliser des

mesures de distance telles que la distance euclidienne, la distance de Mahalanobis, etc. Il existe deux classes principales de techniques de détection d'anomalies utilisant les plus proches voisins. La première classe utilise la distance entre une instance de données et son k^{e} voisin le plus proche comme un degré d'anomalie. La deuxième classe utilise la densité relative de chaque instance de données pour calculer le degré d'anomalie. La densité relative est calculée en mesurant la distance entre une instance de données et son k^{e} plus proche voisin. Par exemple, Local outlier factor(LOF) utilise la densité des instances de données dans un voisinage. Un avantage important des techniques de détection d'anomalies utilisant les plus proches voisins est qu'elles ne nécessitent pas d'instances de données étiquetées. Elles peuvent être utilisées pour une détection d'anomalies non-supervisée.

Techniques de clustering

Le but de l'analyse de cluster est de trouver des groupes d'instances de données qui sont étroitement liées. Les techniques de détection d'anomalies utilisant le clustering peuvent être classées en trois groupes. Dans le premier groupe, l'hypothèse est que les instances de données normales appartiennent à un cluster, tandis que les instances de données anormales finissent en dehors de tous les clusters. Des techniques de clustering telles que DBSCAN (Birant et Kut, 2007), ROCK (Guha, Rastogi et Shim, 2000) ont déjà été utilisées pour ce type de détection d'anomalies. Le deuxième groupe de techniques repose sur l'hypothèse que les instances de données normales finissent près du centroïde du cluster le plus proche, tandis que les instances de données anormales finissent loin du centroïde du cluster le plus proche. Pour ce groupe de techniques, la détection d'anomalies est effectuée en deux étapes. Tout d'abord, les instances de données de l'ensemble de données d'apprentissage sont regroupées. Ensuite, la distance entre chaque instance de données, dans l'ensemble de données de test, et le cluster le plus proche est calculée et utilisée comme score d'anomalie. Les techniques de clustering populaires utilisant cette hypothèse sont les Self-Organizing Maps (SOM), le clustering K-means et les méthodes de types Expectation-Maximization (EM). Dans le troisième groupe de techniques, l'hypothèse est que les instances de données normales sont regroupées

en clusters larges et denses, tandis que les instances de données anormales sont soit regroupées en petits clusters soit de réparties façon éparse. He, Xu et Deng (2003) ont proposé une technique appelée FindCBLOF qui satisfait l'hypothèse ci-dessus. La méthode attribue un score d'anomalie appelé Cluster-Based Local Outlier Factor (CBLOF) à chaque instance de données. Le score combine à la fois la distance entre l'instance de données et son cluster le plus proche et la taille du cluster le plus proche.

Les techniques de détection d'anomalies reposant sur le clustering partagent des propriétés avec les techniques utilisant les plus proches voisins : les deux s'appuient sur une mesure de distance et leurs performances dépendent de la manière dont la mesure de distance correspond aux propriétés de l'ensemble de données. Selon Chandola, Banerjee et Kumar (2010), une différence importante entre les deux approches est que les techniques reposant sur le clustering utilisent le cluster le plus proche comme référence, tandis que les techniques utilisant les voisins les plus proches utilisent le voisinage local comme référence lors de l'évaluation de nouvelles instances de données. Les techniques reposant sur le clustering peuvent être utilisées pour la détection d'anomalies de façon non-supervisée, mais elles peuvent être coûteuses en temps de calcul.

2.1.4 Évaluation des performances sur des classes déséquilibrées

L'évaluation de la performance des algorithmes de détection d'anomalies est compliquée. D'une part, la vérité terrain des anomalies n'est pas évidente à obtenir car les anomalies réelles sont rares par nature. D'autre part, les algorithmes de détection d'anomalies produisent souvent un score anormal pour chaque observation. Les observations avec des scores d'anomalies relativement importants sont considérées comme des anomalies si elles sont supérieures à un seuil donné. Il est relativement difficile de définir à l'avance un seuil approprié pour chaque application. Si le seuil est trop grand, les vraies anomalies sont manquées ; sinon, certaines observations qui ne sont pas de véritables anomalies sont prises à tort comme des anomalies potentielles.

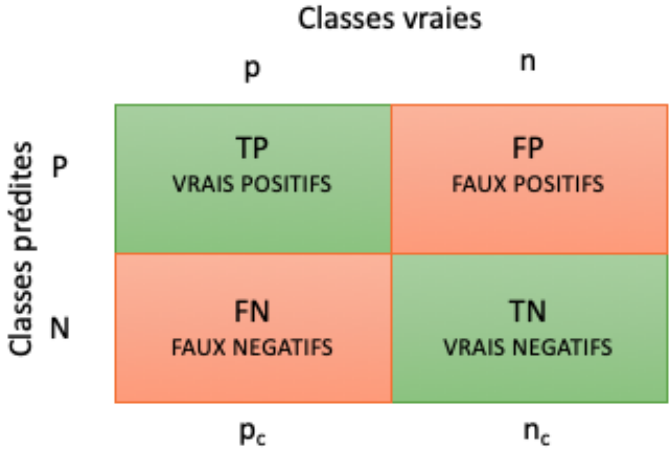


FIGURE 2.9: Matrice de confusion

Métriques d'évaluation adaptées

Les métriques les plus utilisées sont l'*Accuracy*[1] et l'*ErrorRate*. Considérons un problème de classification à deux classes avec $\{p, n\}$ les vraies étiquettes de classe positive et négative et $\{P, N\}$ les étiquettes de classes positive et négative prédites. Ensuite, une représentation des performances de classification peut être formulée par une matrice de confusion comme illustrée sur la Fig 2.9, qui introduit les étiquettes $\{TP, TN, FP, FN\}$. Nous utilisons la classe minoritaire comme classe négative et la classe majoritaire comme classe positive. Ainsi, les mesures *Accuracy* et *ErrorRate* sont définies comme suit :

$$Accuracy = \frac{TP + TN}{p_c + n_c}, \qquad ErrorRate = 1 - Accuracy \tag{2.6}$$

Ces métriques fournissent un moyen simple de décrire les performances d'un classificateur sur un ensemble de données. Cependant, ils peuvent être trompeurs dans certaines situations et sont très sensibles aux changements de données. Dans la situation la plus simple, si un ensemble de données comprend 5% d'exemples de

1. Le terme *Accuracy* est utilisé dans ce manuscrit pour éviter la confusion en français avec le terme anglophone *Precision*

classe minoritaire et 95% d'exemples majoritaires, une approche naïve consistant à classer chaque exemple comme un exemple de classe majoritaire fournirait une précision de 95%. Prise à sa valeur nominale, une précision de 95% sur la totalité de l'ensemble de données semble excellente cependant, de la même manière, cette description ne reflète pas le fait que 0% des exemples minoritaires sont identifiés. C'est-à-dire que la métrique de précision dans ce cas ne fournit pas d'informations adéquates sur la qualité d'un classificateur par rapport au type de classification requis. La littérature indique de nombreux travaux représentatifs sur l'inefficacité de la métrique *Accuracy* dans le scénario d'apprentissage déséquilibré.

Le problème fondamental peut être expliqué en évaluant la matrice de confusion de la Fig 2.9 : la colonne de gauche représente les instances positives de l'ensemble des données et la colonne de droite représente les instances négatives. Par conséquent, la proportion des deux colonnes est représentative de la distribution de classe de l'ensemble des données et de toute métrique qui utilise les valeurs des deux colonnes seront intrinsèquement sensible aux déséquilibres. L'équation 2.6 montre que la métrique *Accuracy* utilise les informations des deux colonnes ; par conséquent, à mesure que la distribution des classes varie, les mesures de la performance changeront même si la performance fondamentale sous-jacente du classificateur ne change pas. Comme nous pouvons l'imaginer, cela peut être très problématique lors de la comparaison des performances de différents algorithmes d'apprentissage sur différents ensembles des données en raison de l'incohérence de la représentation des performances. En d'autres termes, en présence de données déséquilibrées, il devient difficile de faire une analyse relative lorsque les métriques d'évaluation sont sensibles aux distributions de données. Pour remédier à cette problématique, d'autres mesures d'évaluation sont fréquemment adoptées afin de fournir des évaluations complètes des problèmes d'apprentissage déséquilibrés, à savoir les mesures de $Precision$, $Recall$, $F-score$ et $G-mean$.

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

$$F - score = \frac{(1 + \beta)^2 \times Precision \times Recall}{\beta^2 \times Recall + Precision} \ , \tag{2.9}$$

où β est un facteur réel positif choisi de telle sorte que le *Recall* soit considéré β fois plus important que la *Precision*.

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{2.10}$$

La *Precision* indiquée dans l'Eq. (2.7) est une mesure d'exactitude, c'est-à-dire combien d'exemples étiquetés comme positifs sont réellement étiquetés correctement, tandis que le *Recall* de l'Eq. (2.8) est une mesure d'exhaustivité, c'est-à-dire combien d'exemples de la classe positive ont été étiquetés correctement. Ces deux mesures, tout comme l'*Accuracy* et l'*ErrorRate*, partagent une relation inverse entre elles. Cependant, contrairement à l'*Accuracy* et l'*ErrorRate*, la *Precision* et le *Recall* ne sont pas tous deux sensibles aux changements dans la distribution des données. La littérature révèle que la *Precision* est sensible aux distributions des données, alors que le *Recall* ne l'est pas. En revanche, le fait que le *Recall* ne dépende pas de la distribution des données est presque superflu car une conclusion reposant uniquement sur le *Recall* est ambiguë, car le *Recall* ne donne aucune idée du nombre d'exemples étiquetés à tort comme positifs. De même, la *Precision* ne permet pas de mesurer combien d'exemples positifs sont mal étiquetés. Lorsqu'ils sont utilisés ensemble, la *Precision* et le *Recall* peuvent évaluer efficacement les performances de classification dans des scénarios d'apprentissage déséquilibrés. Une métrique plus précise est le $F - score$ détaillé dans l'Eq. (2.9), qui combine la *Precision* et le *Recall* en tant que mesure d'efficacité de la classification en termes de rapport de l'importance pondérée sur le *Recall* ou la *Precision* tel que déterminé par le coefficient choisi par l'utilisateur. Le $F - score$ fournit plus d'informations sur la fonctionnalité d'un classificateur que la métrique *Precision*, tout en restant sensible aux distributions des données. Une autre métrique est la

$G - mean$ 2.10, elle évalue le degré de biais inductif en termes de ratio de précision positive et de précision négative. Bien que $F - score$ et $G - mean$ soient de grandes améliorations par rapport à la $Precision$, elles sont toujours inefficaces pour répondre à des questions plus génériques sur l'évaluation de la classification.

La courbe ROC

La courbe ROC, pour *Receiver Operating Characteristics*, est une mesure de performance pour les problèmes de classification à divers réglages de seuil. C'est une courbe de probabilité qui utilise la proportion de deux métriques d'évaluation regroupées sur une seule colonne, à savoir le taux des vrais positifs TPR et le taux des faux positifs FPR, qui sont définis comme suit :

$$TPR/Recall/Sensitivity = \frac{TP}{p_c} = \frac{TP}{TP + FN} \ , \tag{2.11}$$

$$Specificity = \frac{TN}{n_c} = \frac{TN}{TN + FP} \ , \tag{2.12}$$

$$FPR = 1 - Specificity = \frac{FP}{TN + FP} \ . \tag{2.13}$$

La mesure AUC (*Area Under Curve*) représente le degré de séparabilité. Elle indique dans quelle mesure le modèle est capable de distinguer les classes. Plus l'AUC est élevée, meilleur est le modèle pour prédire correctement les étiquettes des différentes instances de données. Un exemple de la courbe ROC et la mesure AUC est montré sur la Fig 2.10.

La sensibilité et la spécificité dénommées respectivement $Sensitivity$ et $Specificity$ sont inversement proportionnelles l'une à l'autre. Ainsi, lorsque la sensibilité augmente, la spécificité diminue et vice-versa. Comme la courbe ROC est formée en traçant le TPR en fonction du FPR, lorsque nous diminuons le seuil, nous obtenons plus de valeurs positives, ce qui augmente la sensibilité et diminue la spécificité. De même, lorsque nous augmentons le seuil, nous obtenons plus de

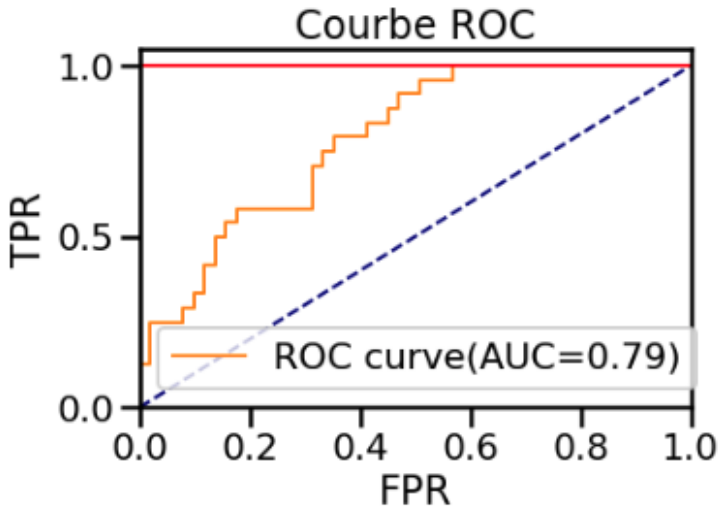


Figure 2.10: Exemple de courbe ROC : la classification parfaite est en rouge et le niveau de chance en bleu

valeurs négatives, nous obtenons donc une spécificité plus élevée et une sensibilité plus faible. Ainsi, lorsque nous augmentons le *TPR*, le *FPR* augmente également.

La courbe ROC est utile car elle fournit une représentation visuelle des compromis relatifs entre les avantages (reflétés par les vrais positifs) et les coûts (reflétés par les faux positifs) de la classification en ce qui concerne les distributions des données. Un excellent modèle a une AUC proche de 1, ce qui signifie qu'il a une bonne mesure de séparabilité. Un modèle qui a une AUC proche de 0 signifie qu'il prédit systématiquement la mauvaise classe, mais de façon extrêmement fiable. Et lorsque l'AUC est de 0,5, cela signifie que le modèle n'a aucune capacité de séparation des classes et qu'il est donc incapable de faire des prédictions fiables. La mesure AUC possède la propriété d'invariante d'échelle, c'est-à-dire qu'elle mesure la qualité du classement des prédictions plutôt que leurs valeurs absolues. Elle est aussi indépendante des seuils de classification. Elle mesure la qualité des précisions du modèle quel que soit le seuil de classification sélectionné. En revanche, ces avantages comportent des limites qui peuvent réduire la pertinence de l'AUC dans certains cas d'utilisation, notamment en ce qui concerne l'indépendance vis-à-vis des seuils de classification. Ce n'est en effet pas toujours souhaitable dans le cas de

disparités importantes de coût entre les faux négatifs et les faux positifs. Il peut être essentiel de minimiser l'un des types d'erreurs de classification.

La courbe Precision-Recall (PR)

Bien que les courbes ROC fournissent des méthodes puissantes pour visualiser l'évaluation des performances, elles ont également leurs limites. Dans le cas d'ensemble de données très asymétriques, il est observé que la courbe ROC peut fournir une évaluation trop optimiste des performances d'un algorithme. Dans de telles situations, les courbes PR peuvent fournir une représentation plus informative de l'évaluation des performances (Davis et Goadrich, 2006). Étant donné une matrice de confusion comme dans la Fig 2.9 et la définition de la *Precision* de l'Eq (2.7) et du *Recall* de l'Eq (2.8), la courbe PR est définie en traçant le taux de *Precision* sur le taux du *Recall*, comme le montre la Fig 2.11. Les courbes PR présentent une forte correspondance avec les courbes ROC : une courbe domine dans l'espace ROC si et seulement si elle domine dans l'espace PR (Davis et Goadrich, 2006). Cependant, un algorithme qui optimise l'AUC dans l'espace ROC n'est pas garanti pour optimiser l'AUC dans l'espace PR (Davis et Goadrich, 2006). De plus, alors que l'objectif des courbes ROC est d'être dans le coin supérieur gauche de l'espace ROC, une courbe PR dominante réside dans le coin supérieur droit de l'espace PR. L'espace PR caractérise également des courbes analogues à l'enveloppe convexe dans l'espace ROC (Davis et Goadrich, 2006). Par conséquent, l'espace PR a tous les avantages de l'espace ROC, ce qui en fait une technique d'évaluation efficace.

En général, la courbe PR peut fournir des représentations plus informatives sur l'évaluation des performances pour des données déséquilibrées. Un simple exemple est celui d'une distribution où les exemples négatifs dépassent de manière significative le nombre d'exemples positifs, c'est-à-dire $n_c > p_c$. Dans ce cas, si la performance d'un classificateur a un grand changement dans le nombre de faux positifs, cela ne changera pas de manière significative le *FPR* puisque le dénominateur n_c est très grand. Par conséquent, la courbe ROC ne parviendra pas à capturer

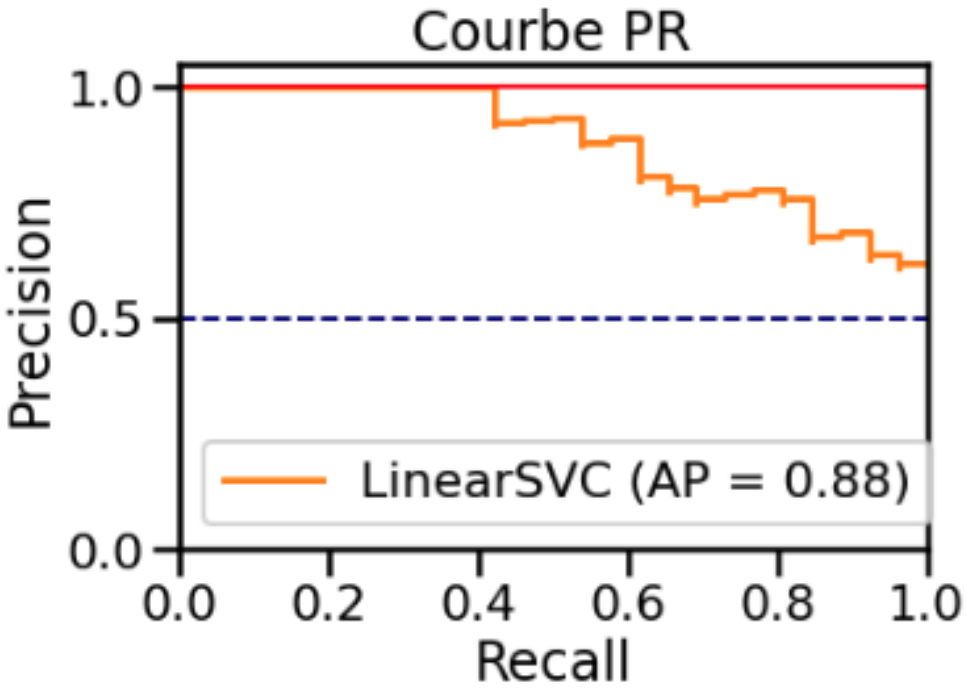


Figure 2.11: Exemple de courbe PR : classification parfaite en rouge, le niveau de la chance est représenté en bleu

ce phénomène. D'une autre part, la métrique *Precision* considère le rapport de TP par rapport à $TP + FP$, elle peut capturer correctement les performances du classificateur lorsque le nombre de faux positifs change radicalement.

2.2 Théorie du transport optimal pour l'apprentissage automatique

La théorie du Transport Optimal se situe à l'intersection de divers domaines, notamment de la théorie des probabilités, de la géométrie et de l'optimisation. Il a connu une progression soutenue à partir de la formulation du problème pour la première fois par Monge en 1781. Le Transport Optimal a bénéficié récemment d'une attention croissante au sein de la communauté de l'apprentissage automatique en raison sa formulation élégante dans divers contextes et surtout pour sa capacité à

aborder des scénarios d'apprentissage difficiles comme la réduction de dimensionnalité et les problèmes de prédiction structurée qui impliquent des histogrammes, des estimations de densités paramétriques ou encore des modèles génératifs dans des problèmes de grande dimension.

2.2.1 Problème du transport Monge–Kantorovich

La théorie du Transport Optimal (OT) a été introduite par le géomètre français Gaspard Monge. Dans son article, *Mémoire sur la théorie des déblais et des remblais* (Monge, 1781), Monge a posé la question : "Comment puis-je déplacer un tas de terre (une ressource naturelle) vers un emplacement cible avec le moins d'effort ou de coût ?" L'idée était de trouver le meilleur moyen d'optimiser ce coût en évitant de parcourir toutes les permutations possibles du fournisseur par rapport au destinataire et à choisir celui qui avait le coût le plus bas. L'une des percées majeures à la suite des travaux de Monge a été réalisée par le mathématicien Leonid Vitaliyevich Kantorovich, le fondateur de la programmation linéaire. Ses recherches sur l'allocation optimale des ressources, qui lui ont valu son prix Nobel d'économie, l'ont amené à étudier le couplage optimal et son problème dual. Il a ainsi revisité certaines parties du problème du transport optimal en un problème de programmation linéaire. Les travaux de Kantorovich ont conduit à renommer le couplage optimal entre deux mesures de probabilité en problème de Monge-Kantorovich.

Problème de Monge

Nous pouvons reformuler le questionnement de Monge ainsi : "Chercher un plan de transport m qui déplacera la masse entre la distribution de masse source μ_s et la distribution cible μ_t, à condition que ce plan de transport soit optimal par rapport à une fonction de coût C donnée qui donne l'effort nécessaire pour déplacer une unité de masse entre deux positions dans l'espace."

Le problème peut alors être exprimé formellement comme suit :
Considérant deux mesures de probabilité μ_s et μ_t et une fonction de coût c :

$\Omega \times \Omega \rightarrow [0, +\infty]$, où $c(x, y)$ mesure le coût de transport d'une unité de masse $x \in \Omega_s$ à $y \in \Omega_t$, la formulation de Monge du transport optimal cherche à trouver un plan de transport m tel que :

$$\inf \int_{\Omega_s} (x, m(x)) \mu_s \, dx \;\; | \, m\#\mu_s = \mu_t \;, \tag{2.14}$$

où $\#$ est l'opérateur qui déplace la masse d'une distribution donnée en utilisant un plan de transport m tel que pour tout sous-ensemble de Borel mesurable $A \in \Omega_t$, la masse est préservée par le plan de transport :

$$\mu_t(A) = \mu_s(m^{-1}(A)) = m\#\mu_s(A) \;. \tag{2.15}$$

Monge a initialement envisagé le problème avec le coût $C(x, y) = |x - y|$. Ce problème est significativement plus difficile qu'avec le coût $C(x, y) = \|x - y\|^2$. En fait, la première preuve correcte avec le coût $C(x, y) = |x - y|$ ne remonte qu'à 1999 (Evans et Gangbo, 1999) et nécessite des hypothèses plus solides que la formule du coût quadratique. En général, le problème de Monge est difficile en raison de la non-linéarité de la contrainte de l'Eq. (2.15), voir (Santambrogio, 2010) pour plus de détail.

Formulation de Kantorovich

La formulation du problème de transport optimal de Monge a été améliorée par Kantorovich en adoptant l'idée de trouver une distribution conjointe entre la source et la cible permettant de définir la manière dont la masse est allouée au lieu de rechercher un plan de transport comme illustré sur la Fig 2.12. Ceci en considérant une mesure de probabilité sur $\Omega_s \times \Omega_t$ qui permet d'atteindre :

$$\inf \int_{\Omega_s \times \Omega_t} C(x, y) \, d\gamma(x, y) \;\; | \, \gamma \in \Gamma(\mu_s, \mu_t) \;, \tag{2.16}$$

où $\Gamma(\mu_s, \mu_t)$ est appelé espace des plans de transport et dont le détail est donné ci-après. Il désigne l'ensemble de toutes les mesures de probabilité sur $\Omega_s \times \Omega_t$

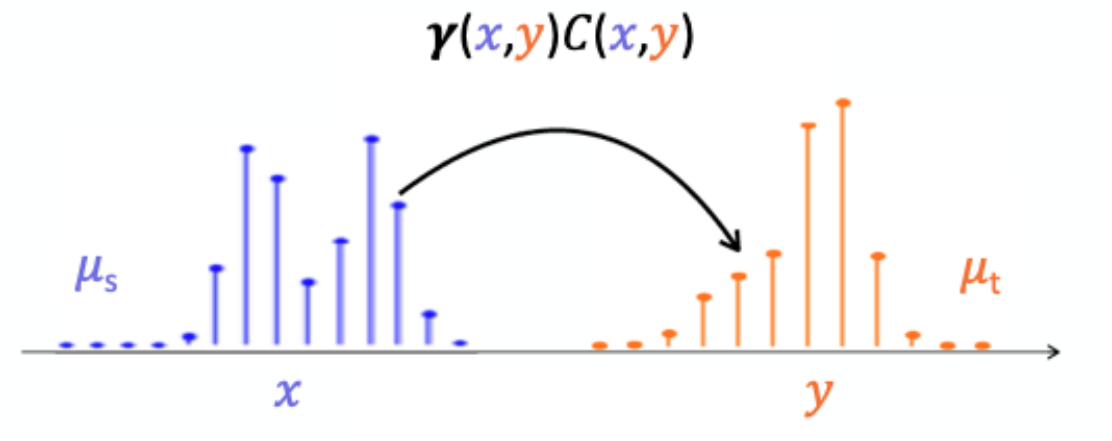


Figure 2.12: Problème du transport optimal

avec des marginales μ_s sur Ω_s et μ_t sur Ω_t permettant de minimiser le coût du transport :

$$\Gamma(\mu_s, \mu_t) = \{\gamma \in P(\Omega_s, \Omega_t) : \int \gamma(x, y)\, dy \ = \mu_s(x), \int \gamma(x, y)\, dx \ = \mu_t(y)\} \ . \tag{2.17}$$

Ce problème d'optimisation est un programme linéaire. Il est convexe et admet une solution unique lorsque la fonction de coût C est semi-continue inférieurement et $\Gamma(\mu_s, \mu_t)$ est une collection compacte de mesures.

Kantorovich a montré que la minimisation du problème fonctionnel linéaire admet un problème dual, cette dualité est donnée par le théorème de Rockefeller-Fenchel :

$$\sup \int \rho(x)\, d\mu_s(x) \ + \int \psi(y)\, d\mu_t(y) \quad | \quad \rho(x) + \psi(y) \leq C(x, y) \ , \tag{2.18}$$

où les deux fonctions scalaires $\rho : \Omega_s \to \mathbb{R}$ et $\psi : \Omega_t \to \mathbb{R}$ sont les variables duales du problème d'optimisation. Une preuve de ce théorème, ainsi qu'une discussion plus approfondie peut être trouvée dans (Villani, 2003).

Distance de Wasserstein

Considérant l'espace des mesures de probabilité $\Omega_s \subset \mathbb{R}^d$ avec un p^{e} moment comme suit :

$$P_p(\Omega_s) := \mu_s \in P(\Omega_s) : \int_{\Omega_s} |x|^p \, d\mu_s(x) \; < +\infty \tag{2.19}$$

La distance de Wasserstein est la p^{e} racine du minimum du problème de transport optimal de Kantorovich pour la fonction de coût $c(x, y) = \|x - y\|^p$.

$$d_{w^p}(\mu_s, \mu_t) = \min_{\gamma \in \Gamma(\mu_s, \mu_t)} \Big(\int_{\Omega_s \times \Omega_t} \|x - y\|^p \, d\gamma(x, y) \Big)^{\frac{1}{p}} \tag{2.20}$$

Cette distance est connue aussi sous le nom de *Earth Mover's Distance* lorsque $p = 1$.

Par exemple, il est possible de définir ainsi la distance de Wasserstein entre deux distributions normales. Considérons $\mu_s = \mathcal{N}(m_s, C_s)$ et $\mu_t = \mathcal{N}(m_t, C_t)$ deux mesures gaussiennes (de distributions normales) sur $\mathbb{R}^n$, avec les valeurs attendues respectives m_s et $m_t \in \mathbb{R}^n$ et C_s et $C_t \in \mathbb{R}^{n \times n}$ des matrices de covariance semi-définies positives symétriques.
La distance 2-Wasserstein entre μ_s et μ_t est :

$$d^2_{w^2}(\mu_s, \mu_t) = \|m_s - m_t\|_2^2 + \operatorname{tr}\left(C_s + C_t - 2\sqrt{C_s^{\frac{1}{2}} C_t C_s^{\frac{1}{2}}} \right) \tag{2.21}$$

Barycentre de Wasserstein

Le barycentre est un outil central en géométrie affine qui permet de caractériser et d'étudier les sous-espace affines, les applications affines et la convexité. Dans (Agueh et Carlier, 2011), les auteurs ont présenté son analogue dans l'espace de Wasserstein, prouvant son existence, son unicité et fournissant ses caractéristiques.

C'est une interpolation entre les distributions $\{\mu_i\}_i$ minimisant la distance de

Wasserstein et qui est définie par :

$$\overline{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \lambda_i d_{w^p}^p(\mu_i, \mu) \ , \tag{2.22}$$

où les λ_i sont connues sous le nom des coordonnées barycentriques avec $\lambda_i > 0$ et $\sum_{i=1}^{n} \lambda_i = 1$.

2.2.2 De la probabilité à la géométrie discrète

Nous nous intéressons dans cette partie au problème du transport optimal pour les distributions discrètes car c'est la situation la plus courante en apprentissage automatique.

Problème primal

Considérons les distributions discrètes suivantes :

$$\mu = \Sigma_{i=1}^{n} a_i \delta_{x_i}, \quad \mu_s = \Sigma_{i=1}^{n_s} a_i \delta_{x_i^s}, \quad \mu_t = \Sigma_{j=1}^{n_t} b_j \delta_{x_j^t} \ , \tag{2.23}$$

où $x_i^s, x_j^t \in \Omega^2$, $a \in \Sigma_{n_s}$, $b \in \Sigma_{n_t}$ et $\Sigma_n = \{(a_i)_i \geq 0, \ \ \Sigma_{i=1}^{n} a_i = 1\}$, et enfin les échantillons x_i peuvent être stockés dans des matrices $X = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$, de même pour X_s et X_t. Cette formulation peut représenter à la fois la formulation lagrangienne où le support x_i et les poids a_i sont libres (espace quotient : Ω_n, Σ_n) et la formulation eulérienne où le support x_i est fixe et l'information est encodée dans a (espace quotient : Σ_n) (Flamary, 2019).

Le problème du transport optimal entre les distributions μ_s et μ_t repose alors sur deux éléments : la matrice C des distances par paires entre les éléments de X_s et X_t avec $C_{i,j} = c(x_i^s, x_j^t)$, qui agit comme un paramètre de coût et le plan de transport $U(a, b)$, qui agit comme un ensemble réalisable et qui est défini comme l'ensemble de $n_s \times n_t$ matrices non négatives telles que leurs marginales de ligne et de colonne sont respectivement égales à a et b. En écrivant respectivement $\mathbf{1}_{n_s}$

et $\mathbf{1}_{n_t}$ pour les vecteurs des uns à n_s-dimensions, et n_t-dimensions,

$$U(a,b) = \left\{ P \in \mathbb{R}_+^{n_s \times n_t} \mid P\mathbf{1}_{n_t} = a \ et \ P^T\mathbf{1}_{n_s} = b \right\} \tag{2.24}$$

et le problème du transport optimal s'écrit :

$$L_C(a,b) = \underset{P \in U(a,b)}{\arg\min} \langle P, C \rangle = \sum_{i,j} C_{i,j} P_{i,j} \ , \tag{2.25}$$

avec $\langle \cdot, \cdot \rangle$ le produit scalaire euclidien.

Problème dual

Le problème primal de Kantorovich est un problème de minimisation convexe contraint, et en tant que tel, il peut être naturellement associé à un problème dit dual, qui est un problème de maximisation concave contraint. Il admet donc la forme duale :

$$L_c(a,b) = \max_{(f,g) \in \mathbb{R}^{ns} \times \mathbb{R}^{nt}} \langle f, a \rangle + \langle g, b \rangle \qquad \text{tel que : } f_i + g_j \leq C_{i,j} \quad \forall i,j \tag{2.26}$$

Ces variables duales f et g sont appelées potentiels de Kantorovich. La figure 2.13 illustre les solutions primale et duale résultant du même problème de transport. À gauche le problème de transport optimal entre deux mesures discrètes α et β, représentées respectivement par des points bleus et des carrés rouges. La surface de ces marqueurs est proportionnelle au poids à chaque emplacement. Cette figure affiche également le transport optimal P^* en utilisant un coût euclidien quadratique. Les potentiels de Kantorovich f^* et g^* correspondants sont affichés sur la partie droite. Puisqu'il y a un f_i pour chaque point de α et un g_i pour chaque point de β, la couleur de chaque point représente la valeur obtenue en utilisant la carte des couleurs à droite. Le coût du transport optimal est égal à la somme des carrés des longueurs de tous les arcs de gauche pondérés par leur épaisseur ou, alternativement, en utilisant la formulation duale, à la somme des valeurs (codées avec des couleurs) multipliée par l'aire de chaque marqueur sur la tracé à droite. Pour

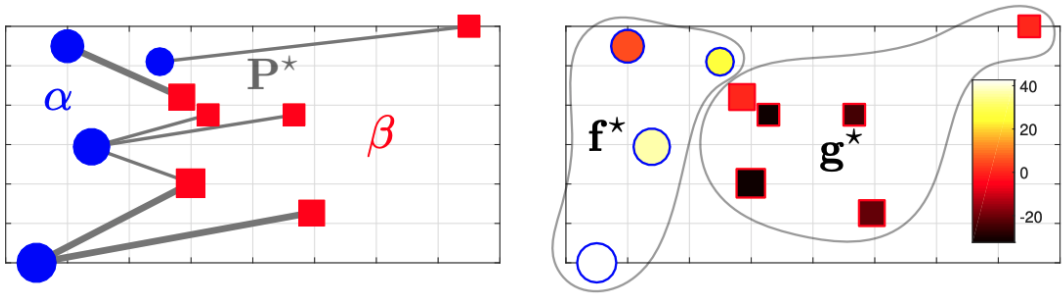


Figure 2.13: Solutions primale et duale d'un problème de transport optimal - image tirée de (Peyré, Cuturi et al., 2019)

des questions de coût computationnel, on résout le problème dual pour obtenir la solution au problème initial.

Distance de Wasserstein

Une caractéristique importante du transport optimal est qu'il définit une distance entre les histogrammes et les mesures de probabilité dès que la matrice de coût satisfait certaines propriétés appropriées ; La distance de Wasserstein prouve que le transport optimal fournit une distance valide entre les histogrammes.

Nous supposons que $n = n_s = n_t$ et que pour $p \geq 1$, $C = D^p = D^p_{i,j} \in \mathbb{R}^{n \times n}$ est une matrice de distance. La distance de p-Wasserstein sur Σ_n s'écrit :

$$d_{w_p}(a, b) = L_{D^p}(a, b)^{\frac{1}{p}} \tag{2.27}$$

2.2.3 Régularisation entropique du transport optimal

Une avancée majeure dans le transport optimal a connu le jour grâce à Marco Cuturi (Cuturi, 2013) qui a permis son utilisation dans de nombreuses applications d'apprentissage automatique, en ajoutant une pénalité de régularisation entropique au problème d'origine, réduisant ainsi sa complexité.

Régularisation entropique

La régularisation entropique est une technique qui permet un compromis explicite entre la précision et l'efficacité de calcul. Cette approche s'est révélée particulièrement prometteuse dans le régime où une estimation approximative du transport est suffisante.

Considérons le terme d'entropie discrète d'une matrice de couplage, qui est défini comme :

$$H(P) = -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1). \tag{2.28}$$

Cette définition fait deux hypothèses qui sont nécessaires pour travailler avec l'entropie. La première c'est qu'une mesure de probabilité admet une distribution et qu'elle est non nulle partout, et la deuxième c'est que $H(p)$ est une fonction concave qui mesure approximativement le flou d'une distribution. Une entropie faible indique qu'une distribution a un pic marqué sur quelques points, tandis qu'une entropie élevée indique qu'elle est plus uniformément distribuée dans l'espace. L'idée d'utiliser l'entropie $-H(P)$ pour régulariser le problème du transport optimal est pour obtenir une solution approchant celle problème de transport initial.

$$L_C^\epsilon(a,b) = \underset{P \in U(a,b)}{\arg\min} \langle P, C \rangle - \epsilon H(P) \ , \tag{2.29}$$

avec ϵ le paramètre de régularisation. Cette régularisation rend le problème fortement convexe et moins sensible aux petits changements dans la distribution. Cela rendra le problème d'optimisation moins compliqué à résoudre. Cette formulation est équivalente à la divergence de Kullback-Leibler KL, une mesure de type distance (mais asymétrique) de la similarité entre P et le noyau K défini par $(K_\epsilon)_{i,j} = e^{\frac{-C_{i,j}}{\epsilon}}$. La définition de K_ϵ est singulière lorsque $\epsilon = 0$, indiquant que la connexion à KL n'est possible que dans le régime $\epsilon > 0$.

Sinkhorn-Knopp et projections de Bregman

La matrice K ne satisfait pas vraiment aux contraintes du problème de transport régularisé de l'Eq. (2.29). En considérant KL grossièrement comme une mesure de distance, l'idée consiste à trouver la projection la plus proche (par rapport à KL) de K sur l'ensemble des P satisfaisant les contraintes $\sum_j P_{i,j} = x_i^s$ et $\sum_i P_{i,j} x_j^t$.

Si nous reprenons le problème (2.29), nous aurons :

$$\begin{aligned}
\underset{P\in U(a,b)}{\arg\min} \langle P, C\rangle - \epsilon H(P) &= \sum_{i,j} P_{i,j} C_{i,j} - \epsilon \sum_{i,j} P_{i,j} \log(P_{i,j}) \\
&= \epsilon \sum_{i,j} P_{i,j} \Big(\frac{C_{i,j}}{\epsilon} - \log P_{i,j} \Big) \\
&= \epsilon \sum_{i,j} P_{i,j} \log(\frac{e^{\frac{-C_{i,j}}{\epsilon}}}{P_{i,j}}) \\
&= \epsilon KL(P|K_\epsilon) \ .
\end{aligned} \tag{2.30}$$

Le problème du transport peut alors être reformulé en :

$$\min_{P\in U(a,b)} KL(P|K_\epsilon) \ , \tag{2.31}$$

où $KL(P|K_\epsilon)$ est la divergence de Kullback-Leibler entre les matrices :

$$\begin{aligned}
P = \mathrm{Diag}(\mathbf{u}) K_\epsilon \mathrm{Diag}(\mathbf{v}) \\
\text{et} \\
K = e^{\frac{-C}{\epsilon}} \ .
\end{aligned} \tag{2.32}$$

L'algorithme de Sinkhorn, qui s'appuie sur la projection de Bregman, peut alors être appliqué en projetant alternativement la matrice P au sens KL sur les marginales gauche et droite. Il s'agit ainsi de mettre à jour itérativement $\mathbf{u}$ et $\mathbf{v}$ jusqu'à convergence.

L'algorithme de Sinkhorn se distingue en optimisation numérique par ses avantages. Au-delà de sa facilité d'implémentation, cet algorithme est construit à partir d'opérations d'algèbre linéaire simples (multiplications matrice-vecteur et arithmétique par élément) qui se parallélisent bien et peuvent être effectuées extrêmement rapidement sur du matériel de traitement moderne CPU ou GPU. Une version moderne de Sinkhorn montre comment réduire encore plus les calculs tout en préservant une bonne vitesse de convergence.

Calcul rapide du transport optimal régularisé

La régularisation du problème du transport optimal avec une pénalité entropique permet de mieux aborder les applications à l'intersection de la théorie du transport optimal et de l'apprentissage automatique, car elle garantit des accélérations efficaces dans le calcul. Les travaux de Cuturi (2013) ont permis une plus large adoption par la mise à disposition d'un calcul rapide du transport optimal régularisé. Il a approché la distance de Wasserstein dans le cas du transport optimal régularisé par l'algorithme Sinkhorn-Knopp.

La distance de Sinkhorn est donnée par :

$$d_c^{\epsilon}(a,b) = \langle P^{\epsilon}, C\rangle \qquad \text{avec } P^{\epsilon} = \arg\min \langle P, C\rangle - \frac{1}{\epsilon} H(p) \ , \tag{2.33}$$

d_c^{ϵ} est appelée aussi la divergence dual-Sinkhorn et peut être calculée à un coût bien plus faible que le problème du transport optimal classique pour des valeurs de ϵ raisonnables. L'algorithme 1 montre le calcul de $d_c^{\epsilon}(a,b)$ en utilisant l'itération à

point fixe de Sinkhorn-Knopp.

Algorithm 1: Algorithme Sinkhorn-Knopp

Inputs : C,ϵ,a,b
$I = (a > 0); a = a(I); C = C(I, :); K = \exp(-C/\epsilon)$;
$x = \text{ones}(\text{length}(a), size(b, 2))/\text{length}(a)$;
while x *changes* **do**
    $x = \text{diag}(1./a) * K * (b. * (1./(K^T * (1./x))))$;
end
$u = 1./x$;
$v = b. * (1./(K^T * u))$;
$d_c^\epsilon = \text{sum}(u. * ((K. * C) * v))$;

2.2.4 Apprentissage automatique avec du transport optimal

Récemment, le transport optimal a connu un essor dans les applications d'apprentissage automatique. Cela a commencé par le traitement d'images en utilisant des histogrammes des couleurs d'images et la distance de Wasserstein pour calculer la similitude entre les images. Ensuite, il a été utilisé pour la reconnaissance de forme (Gangbo et McCann, 2000 ; Ahmad, 2004). Par exemple, Haker et al. ont introduit une méthode de calcul des cartes d'alignement et de déformation élastique basée sur la théorie de Monge-Kantorovich du transport optimal (Haker et Tannenbaum, 2003 ; Haker et al., 2004). En raison du rôle important de la factorisation matricielle dans l'apprentissage automatique, il a été rapidement appliqué comme composante de divergence de la factorisation matricielle non négative (NMF) (Sandler et Lindenbaum, 2011), basée sur la *Earth Mover's Distance (EMD)* introduite par Rubner et al dans le papier *A Metric for Distributions with Applications to Image Databases* (Rubner, Tomasi et Guibas, 1998). En 2014, Solomon et al., ont examiné les applications du transport optimal dans l'apprentissage semi-supervisé dans leur article *Wasserstein Propagation for Semi-Supervised Learning* (Solomon et al., 2014). Une autre étude récente a été publiée sur l'utilisation de la métrique de Wasserstein en inférence variationnelle, qui est au coeur

de l'apprentissage automatique (Ambrogioni et al., 2018). En outre, le domaine du transport optimal en apprentissage automatique est maintenant plus actif que jamais, les chercheurs étendant les théories qui fonctionnent pour les problèmes en apprentissage de faibles dimensions aux problèmes de grandes dimensions, soulevant plusieurs questions théoriques et algorithmiques complexes, grâce aux derniers progrès dans la théorie du transport optimal menant à des méthodes d'approximations des problèmes de Kantorovich.

Apprentissage supervisé et semi-supervisé avec du transport optimal

Classification utilisant la distance de Sinkhorn La distance de Wasserstein peut être utilisée comme terme d'ajustement des données pour l'apprentissage d'un prédicteur probabiliste puisque sa sortie est un histogramme. L'utilisation du transport optimal est particulièrement intéressante dans le cas de l'apprentissage par transfert et l'adaptation de domaines. Dans (Chobola, Vašata et Kordík, 2021), les auteurs ont montré comment transformer des vecteurs de caractéristiques en de meilleures distributions de type gaussiennes. En appliquant un algorithme de transport optimal itératif pour estimer les centres de classe de manière empirique, leur méthode a gagné une amélioration significative par rapport aux autres méthodes de classification basées sur les mélanges de gaussiennes, les plus proches voisins et d'autres classificateurs.

Propagation de Wasserstein pour l'apprentissage semi-supervisé Solomon et al. (2014) ont proposé une approche efficace et théoriquement solide pour l'apprentissage semi-supervisé reposant sur des graphes des distributions de probabilités. La stratégie repose sur l'utilisation de la distance 2-Wasserstein entre les distributions pour construire un terme de régularisation mesurant le coût de l'affectation d'une distribution de probabilité à chaque noeud du graphe $G = (V, E)$. L'affectation finale est produite en optimisant l'énergie de Dirichlet tout en ajus-

tant les prédictions de l'histogramme aux noeuds étiquetés :

$$\xi_D[\rho] = \sum_{(v,w)\in E} d_{w_2^2}(\rho_v, \rho_w) \ , \tag{2.34}$$

avec $d_{w_2^2}$ est la distance 2-Wasserstein entre les distributions de probabilités ρ_v et ρ_w.

Apprentissage non supervisé avec du transport optimal

Co-clustering avec du transport optimal régularisé Dans (Laclau et al., 2017), les auteurs ont présenté une approche d'apprentissage non-supervisé de co-clustering utilisant le transport optimal régularisé par l'entropie. La méthode consiste à trouver un couplage probabiliste des mesures empiriques définies à partir des instances de données simulées suivant le processus génératif des modèles de blocs latents gaussiens et de leurs caractéristiques. Selon les auteurs, cette procédure peut être considérée comme un problème d'inférence variationnelle et la distribution inférée peut être utilisée pour obtenir les hyperplans de séparation. En plus de la précision de l'algorithme résultant, il est rapide et capable de détecter automatiquement le nombre de co-clusters. Une version plus étendue de l'algorithme qui utilise la distance de transport optimale définie sur des matrices de similarité associées aux mesures empiriques des lignes et des colonnes a été présentée dans cette étude.

Transport spectral optimal Proposé par Flamary et al. (2016), ce nouveau modèle pour la transcription musicale spectrale repose sur l'utilisation d'un dictionnaire surcomplet. La méthode consiste à utiliser la distance de Wasserstein pour le démixage linéaire du spectre audio. Comme la conception de la matrice de coût est primordiale pour le calcul de la distance de Wasserstein, leur méthode repose sur une nouvelle forme de matrice des coûts de transport qui prend en compte la structure harmonique inhérente aux signaux musicaux. La matrice de coût de transport proposée permet d'utiliser un dictionnaire simpliste composé de

vecteurs de Dirac placés aux fréquences fondamentales cibles, éliminant ainsi le problème du choix d'un dictionnaire significatif. Les résultats expérimentaux ont montré la robustesse et la précision de cette l'approche et la structure particulière du dictionnaire proposé a permis un algorithme simple qui est bien plus rapide que les approches de type factorisation matricielle non négative.

2.3 Conclusion

Tout au long de ce chapitre, nous avons tenté de fournir une vue d'ensemble sur les éléments importants pour la prédiction de situation anormale dans les équipements industriels. La première partie du chapitre a abordé l'état de l'art sur la compréhension théorique du problème de détection d'anomalies par apprentissage automatique au regard des différents aspects qui y sont liés. Une discussion a été menée sur ce qu'est une anomalie et l'identification de ses manifestations les plus courantes en fonction de leur nature, regroupées en anomalies contextuelles et anomalies comportementales. Cette première partie a relevé aussi l'importance d'identifier la nature des données à exploiter car leur représentation peut prendre la forme de représentations métriques, évolutives ou multistructurées. Cette représentation détermine l'applicabilité des techniques de détection d'anomalies, que ce soit pour utiliser ces techniques ou en concevoir de nouvelles. Il est judicieux de prendre en compte quelques mécanismes importants en apprentissage automatique, que nous avons vu dans la suite du chapitre.

En fonction de l'étiquetage des données, un problème d'apprentissage peut être modélisé comme supervisé, semi-supervisé ou non-supervisé. Ce modèle d'apprentissage doit avoir la capacité à construire des connaissances pertinentes sur la base d'un ensemble initial d'exemples d'entraînement en évitant un sur-apprentissage ou un sous-apprentissage. La plupart des algorithmes d'apprentissage automatique ont des hyper-paramètres qui permettent de contrôler le processus d'apprentissage et de trouver la configuration qui offre les meilleures performances sur un ensemble de validation. Nous avons montré la façon de définir la sélection de modèles en utilisant la validation croisée. Nous avons souligné également le problème des classes

déséquilibrées qui est très présent dans les situations de détection d'anomalies. De plus, nous avons présenté quatre catégories de méthodes de détection d'anomalies, chacune d'entre elles a ses avantages et ses inconvénients en fonction du contexte d'application et des mécanismes d'apprentissage automatique cités ci-dessus. Enfin, nous avons vu une synthèse sur les métriques d'évaluation des performances des algorithmes de détection d'anomalies, en particulier dans le cas des classes non-équilibrées.

Dans la deuxième partie du chapitre, nous avons montré que le transport optimal bénéficie d'une attention croissante au sein de la communauté de l'apprentissage automatique en raison de son applicabilité dans plusieurs domaines. Bien que le transport optimal soit de plus en plus accepté dans l'apprentissage automatique, il est profondément enraciné dans les mathématiques et la théorie de l'optimisation. Nous avons donc extrait les sujets les plus importants afin permettre une compréhension de haut niveau du problème du transport et de mettre l'accent uniquement sur les aspects reliés à l'apprentissage automatique. Ces aspects incluent le problème de Monge, la formulation duale de Kantorovich, les distances de Wasserstein et la divergence de Kullback-Leibler. Bien que les applications du transport optimal couvrent un large éventail de domaines, elles sont limitées par leur calculabilité. C'est dans ce sens que nous avons exploré la notion de la régularisation entropique et comment elle a mené à la conception d'un algorithme permettant le calcul rapide des distances de Wasserstein qui a rendu à son tour les problèmes de transport optimal utilisables en pratique. Enfin, nous avons complété le chapitre par quelques applications du transport optimal en apprentissage automatique pour des modèles supervisés, semi-supervisés et non-supervisés.

Chapitre 3

Présentation des contributions pour la détection d'anomalies

Sommaire

Ce chapitre propose des éléments de réponses à la question scientifique de la détection d'anomalies dans le cadre des processus de diagnostic des machines industrielles. Il présente les méthodes de détection d'anomalies mises à contribution dans ce manuscrit. La première section aborde le contexte applicatif des méthodes

existantes, notamment la surveillance de l'état de santé des machines industrielles à travers des systèmes de capteurs qui délivrent des signaux sous forme de séries temporelles. Une liste non exhaustive des travaux de recherche menés sur cette question est proposée dans le but de positionner les contributions de ce manuscrit par rapport aux méthodes existantes agissant dans un contexte applicatif similaire. Ainsi, les algorithmes de détection d'anomalies les plus répandus dans la littérature sont rappelés dans cette section en vue d'une étude comparative ultérieure entre ces derniers et les contributions décrites dans ce manuscrit.

La deuxième section présente la première contribution du manuscrit ; une nouvelle méthode de classification à une classe pour la détection d'anomalies dans les séries temporelles qui s'appuie sur les avancées scientifiques de la théorie du transport optimal pour le calcul de distance entre spectres de fréquences. Elle montre d'abord comment le transport optimal permet d'étudier la similarité entre les séries temporelles puis présente les algorithmes qui implémentent les méthodes paramétrique et non-paramétrique que je propose.

La troisième section présente la deuxième contribution du manuscrit. Il s'agit d'une méthode de détection d'anomalies de la catégorie des plus proches voisins en se s'appuyant sur métriques de distances en transport optimal. Cette section porte d'abord sur l'impact de la fonction de distance dans les approches du plus proche voisin. Elle présente le nouvel algorithme de détection d'anomalies, appelé LOFO, qui est inspiré de l'algorithme LOF mais qui calcule la densité d'atteignabilité locale par des métriques en transport optimal. Il est appliqué à plusieurs types de données et évalué expérimentalement.

3.1 Détection d'anomalies pour les séries temporelles

Une série temporelle est une série de points de données répertoriés dans l'ordre temporel. C'est une collection d'observations obtenues par des mesures répétées dans le temps. Nous nous intéressons dans cette section à cette forme de données. Nous étudions le problème de la détection d'anomalies dans une série temporelle donnée, par rapport à un ensemble de référence constitué de séries temporelles

normales. Cette formulation du problème est hautement applicable dans les domaines du diagnostic des machines industrielles, de la détection de fraude par carte de crédit, de la détection de conditions anormales dans les données ECG, de la détection d'anomalies de forme, de la détection de courbes de lumière aberrantes dans les données astronomiques, etc.

3.1.1 Diagnostic des défauts dans les machines industrielles

La surveillance de l'état de santé des machines industrielles dépend fortement des systèmes de capteurs utilisés. Ils doivent être capables de fonctionner dans diverses conditions environnementales et être capables de transmettre des signaux vers une unité de traitement et d'analyse, tout en agissant de manière neutre sur les performances globales du fonctionnement de la machine. Les signaux issus des systèmes de capteurs sont collectés sous forme de séries temporelles. Dans l'étude de Abdul-Aziz et al. (2012), les auteurs traitent le diagnostic d'un disque de rotor semblable à un moteur à turbine à l'aide de techniques combinées de détection d'anomalies expérimentales et de celles s'appuyant sur les données de séries temporelles issues de la surveillance de l'état de santé de l'engin. Leur approche consistait à effectuer des tests de rotation sur des rotors de type turbo-machine à différents niveaux de charge en rotation avec et sans une encoche induite artificiellement. Les tracés de Bode des données de ces tests ont fourni des indications sur les différences induites par la faille. En parallèle une évaluation des données de surveillance a été effectuée à l'aide de trois algorithmes de détection d'anomalies : ORCA qui est un algorithme de détection des valeurs aberrantes basée sur la distance euclidienne du plus proche voisin (Bay et Schwabacher, 2003), OC-SVM (présenté dans le chapitre précédent) et IMS qui est une méthode de modélisation reposant sur les clusters (Iverson, 2004).

Dans une autre étude, Purarjomandlangrudi, Ghapanchi et Esmalifalak (2014) appliquent une méthode de détection d'anomalies pour diagnostiquer les défauts précoces des roulements d'éoliennes. Cette méthode utilise des techniques de classification pour distinguer les anomalies des données

normales sur la base de deux caractéristiques, l'aplatissement et le score de non-gaussianité (NGS). Jin et al. (2016) ont développé une méthode de détection d'anomalie de roulement et de pronostic de défaut. La méthode détecte les anomalies des roulements et prédit ensuite leur durée de vie utile restante. Cette méthode repose sur un modèle autorégressif pour filtrer les signaux non liés aux défauts, et un indice de santé est développé pour indiquer les conditions de santé des roulements à partir d'une décomposition en ondelettes. Les anomalies des roulements sont détectées en choisissant un seuil approprié à l'aide du calcul des distances de Mahalanobis en considérant les corrélations entre les caractéristiques par rapport aux caractéristiques normales. Une transformation de Box-Cox permet de convertir les distances de Mahalanobis en variables distribuées normales, de sorte que les propriétés de la distribution normale peuvent être utilisées pour déterminer les plages de distances de Mahalanobis correspondant à différentes conditions de santé. Le même algorithme a été déjà utilisé par les mêmes auteurs sur des données expérimentales d'un ventilateur de refroidissement et d'un moteur à induction (Jin et Chow, 2013).

Pittino et al. fournissent une étude comparative des méthodes de détection automatique des anomalies dans le fonctionnement d'un roulement rotatif dans une machine commerciale de fabrication de semi-conducteurs entre les méthodes statistiques telles que les cartes de contrôle et les méthodes de classification (Pittino et al., 2020). Les différentes méthodes offrent des performances très similaires en utilisant des données très différentes, certaines méthodes étant semi-supervisées et d'autres supervisées. Les méthodes se sont avérées capables de fournir un moyen flexible et précoce de détecter les anomalies, avec une flexibilité et une robustesse remarquables.

Tian, Azarian et Pecht (2014) ont présenté une méthode de détection d'anomalies qui repose les plus proches voisins avec des cartes auto-organisatrices (SOM). Ils appliquent leur méthode aux données issues de la surveillance de l'état de santé des systèmes mécaniques et électroniques. Les meilleures unités correspondantes (BMU) du SOM entraînées avec les données d'entraînement normales sont extraites en tant que références. Pour chacune des observations de données de

test, la distance euclidienne aux voisins les plus proches dans les BMU est calculée comme la valeur de son indicateur de santé.

Dans une autre étude, afin de résoudre les problèmes de la recherche globale coûteuse et inefficace des voisins les plus proches, Lu et al. (2021) ont proposé une nouvelle méthode de diagnostic des défauts des machines tournantes qui effectue une extraction de caractéristiques via des méthodes non supervisées. La détection d'anomalies se fait via les voisins les plus proches déterminés automatiquement. Selon les auteurs, la particularité de leur méthode est sa capacité à effectuer une recherche plus efficace à l'aide des vecteurs de corrélation obtenus pour tester la reconstruction d'échantillons.

Cette liste n'est pas exhaustive, mais permet de fournir un aperçu objectif de l'orientation de la littérature sur la détection d'anomalies dans les séries temporelles dans un contexte de surveillance de l'état de santé des machines industrielles.

3.1.2 Positionnement des méthodes de détection existantes

Dans le chapitre précédent, nous avons regroupé les méthodes de détection d'anomalies en plusieurs catégories telles qu'illustrées sur la Fig 2.8. Nous nous sommes concentrés sur les quatre catégories de méthodes les plus répandues pour résoudre le problème d'identification des défauts de fonctionnement dans un contexte industriel, notamment les méthodes de classification, les méthodes statistiques et les méthodes reposant sur les plus proches voisins. Cette section met l'accent sur des algorithmes spécifiques de ces catégories de méthodes. Il s'agit d'algorithmes d'apprentissage semi-supervisé ou non supervisé (en fonction de la disponibilité des étiquettes) qui tentent de modéliser des échantillons normaux afin de classer les échantillons de l'ensemble de test comme normaux ou anormaux. Ces algorithmes sont souvent employés pour une classification à une classe, c'est-à-dire une classification binaire avec une distribution de classe fortement asymétrique. Ces algorithmes se sont avérés efficaces pour les ensembles de données déséquilibrés où il n'y a pas ou très peu d'échantillons de la classe minoritaire, ou pour les ensembles de données où il n'y a pas de structure cohérente pour séparer les

classes qui pourraient être apprises par un algorithme supervisé.

Machines à vecteurs de support

L'algorithme de la machine à vecteurs de support a pour objectif de trouver un hyperplan dans un espace à N dimensions qui classe distinctement les points de données. Pour séparer les deux classes de points de données, il existe de nombreux hyperplans possibles qui pourraient être choisis, mais l'objectif est de trouver un plan qui maximise un critère de marge, c'est-à-dire qui maximise la distance maximale entre la marge et les points de données des deux classes. Ce processus d'optimisation de la distance de la marge fournit un certain renforcement afin que les futurs points de données puissent être classés avec plus de confiance. Les hyperplans sont des limites de décision qui aident à classer les points de données. Les points de données tombant de chaque côté de l'hyperplan peuvent être attribués à différentes classes. La dimension de l'hyperplan dépend du nombre de caractéristiques ou d'attributs. Pour construire un hyperplan à l'aide de ce critère de marge, il est nécessaire d'avoir des vecteurs de support, c'est-à-dire les points de données les plus proches de l'hyperplan et qui influencent sa position et son orientation. En effet, l'utilisation de ces vecteurs de support maximise la marge du classificateur et leur suppression modifie la position de l'hyperplan. La Fig 3.1 montre un exemple illustratif de SVM.

Dans un classificateur SVM, il est facile de déterminer un hyperplan linéaire entre deux classes, si les points de données sont linéairement séparables ce qui n'est pas souvent le cas en pratique. Dans une telle situation, les SVM utilisent une reformulation en utilisant une fonction noyau pour transformer l'espace d'entrée en un espace de dimension supérieure, comme indiqué sur la Fig 3.2. Cette figure montre une transformation non linéaire de l'espace d'entrée 2D en espace 3D, appelé espace de redescription, pour rechercher l'hyperplan. L'utilisation d'une fonction noyau (ou *kernel trick*) permet de chercher une séparation linéaire en projetant le problème dans un espace de grande dimension.

Quatre noyaux sont populaires dans le contexte des SVM : le noyau linéaire,

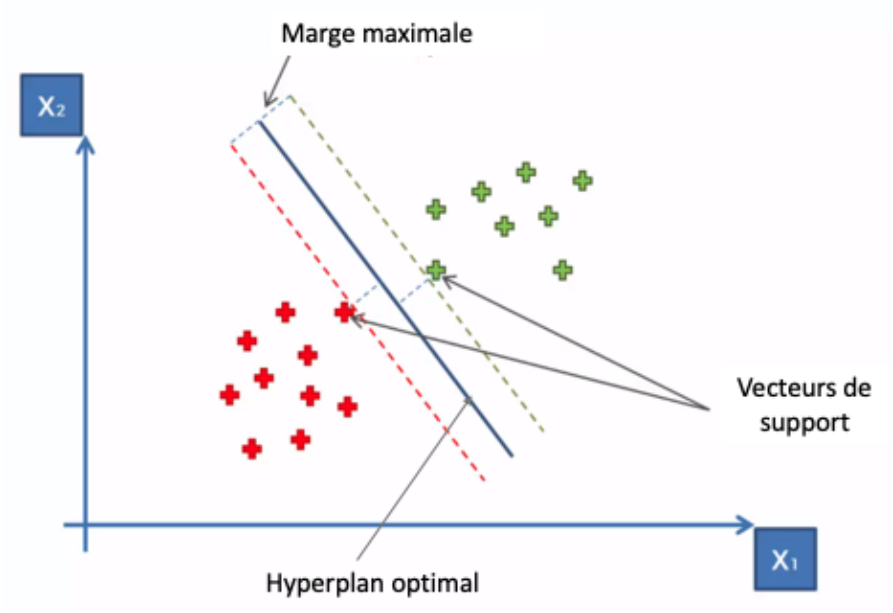


Figure 3.1: illustration d'un classificateur SVM

le noyau polynomial, le noyau radial (RBF) appelé aussi noyau gaussien et noyau sigmoïde. Le noyau linéaire est utilisé lorsque les données sont séparables linéairement. C'est l'un des noyaux les plus couramment utilisés. Le noyau polynomial représente la similarité des échantillons d'entraînement dans un espace de redescription. Le noyau polynomial examine non seulement les caractéristiques données des échantillons d'entrée pour déterminer leur similitude, mais également les combinaisons des échantillons d'entrée. Le noyau radial est un noyau à usage général. Il est fréquemment utilisé lorsqu'il y a peu de connaissance préalable des données. Enfin le noyau sigmoïde qui a son origine dans les réseaux de neurones. Nous pouvons l'utiliser comme une approximation de réseaux de neurones.

Le *One-Class SVM* est une variation de l'algorithme SVM. La principale différence entre les deux est que la SVM est entraînée de façon supervisée, alors que pour la One-Class SVM, il est possible de faire l'entraînement de façon semi-supervisé. Un hyper-paramètre permet de contrôler la sensibilité de l'algorithme et doit être réglé sur le rapport approximatif de valeurs anormales dans les données.

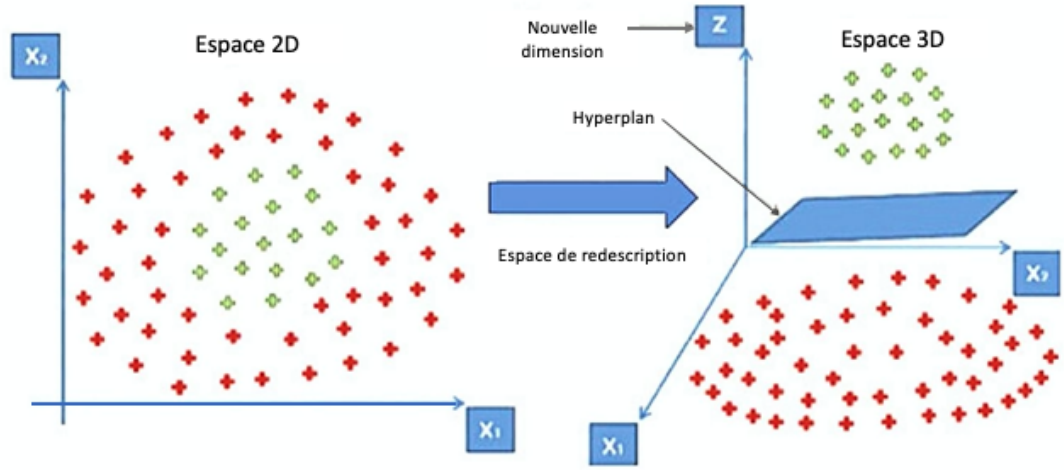


FIGURE 3.2: Recherche d'hyperplan dans un espace de redescription

Méthode de l'enveloppe elliptique

La méthode du Déterminant de Covariance Minimum (ou *MCD*) est un estimateur très robuste de la localisation et de la dispersion de données multivariées qui s'appuie sur un algorithme d'estimation très efficace. Étant donné que l'estimation de la matrice de covariance est la pierre angulaire de nombreuses méthodes statistiques multivariées, la méthode du MCD est particulièrement utile dans le cadre des données bruitées. C'est un outil performant pour la détection d'anomalies en raison de sa résistance aux observations aberrantes. À partir de la méthode du MCD, il est possible d'estimer une enveloppe elliptique en faisant des hypothèses fiables sur la distribution des données. Bien qu'il ait déjà été introduit en 1984, son utilisation principale n'a commencé que depuis la construction de l'algorithme de calcul rapide proposé par ROUSSEEUW et DRIESSEN (1999). La méthode du MCD a également été utilisée pour développer de nombreuses techniques multivariées robustes, parmi lesquelles l'analyse en composantes principales robuste, l'analyse factorielle et la régression multiple. L'approche consiste à définir une hypersphère (ellipsoïde) qui couvre les données normales, et les données qui tombent en dehors

de cette forme sont considérées comme anomalies. Statistiquement, une distribution multivariée est dite elliptiquement symétrique et unimodale s'il existe une fonction réelle g strictement décroissante telle que la densité s'écrit sous la forme :

$$f(x) = \frac{1}{\sqrt{|x|}} g(d^2(x, \overline{x}, \Sigma) \ , \tag{3.1}$$

où l'ellipse de tolérance classique est définie comme l'ensemble des points x de p-dimensions dont la distance de Mahalanobis MD est donnée par :

$$MD(x) = d(x, \overline{x}, \Sigma) = \sqrt{(x - \overline{x})^T \Sigma^{-1} (x - \overline{x})} \ , \tag{3.2}$$

avec la moyenne $\overline{x}$ et la matrice de covariance Σ. Alors que l'ellipse de tolérance robuste repose sur les distances robustes telles que :

$$RD(x) = d(x, \hat{\mu}_{MCD}, \hat{\Sigma}_{MCD}) \ , \tag{3.3}$$

avec l'estimation de l'emplacement $\hat{\mu}_{MCD}$ et l'estimation de la covariance $\hat{\Sigma}_{MCD}$. Le détail est fourni par Hubert, Debruyne et Rousseeuw (2018). La Fig 3.3 montre la différence entre une ellipse de tolérance classique en rouge et l'ellipse robuste en bleu qui est beaucoup plus petite et ne contient que les points de données réguliers dans un contexte de détection d'anomalies.

Forêt d'isolement

La technique de forêt d'isolement ou *Isolation Forest* a été introduite par Liu, Ting et Zhou (2008). L'idée principale de cette technique et qui lui donne sa particularité par rapport aux autres techniques de détection d'anomalies est qu'elle identifie explicitement les anomalies au lieu de profiler les points de données normaux. La forêt d'isolement, comme toute méthode d'ensemble d'arbres, est construite à partir d'arbres de décisions mais avec une notion d'isolement qui repose sur la séparation d'un point de données du reste des points en mesurant la susceptibilité des points individuels à être isolés. Les anomalies sont les points qui ont la susceptibilité la plus élevée. Pour réaliser cet isolement, les partitions

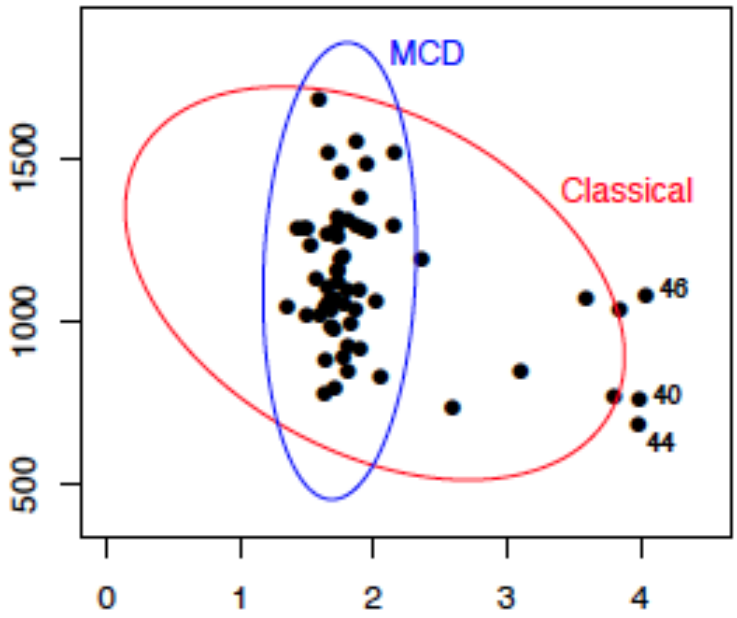


Figure 3.3: Ellipse de tolérance classique et robuste - Image tirée de (Hubert, Debruyne et Rousseeuw, 2018)

sont créées dans des structures arborescentes de manière récursive en sélectionnant d'abord au hasard un attribut, puis en sélectionnant une valeur de division aléatoire entre la valeur minimale et maximale de l'attribut sélectionné. Le nombre de partitions nécessaires pour isoler un point est équivalent à la traversée de la longueur du chemin du noeud racine à un noeud de terminaison. Cette longueur de chemin, moyennée sur une forêt d'arbres aléatoires, est une mesure de la normalité et est utilisée comme fonction de décision (Liu, Ting et Zhou, 2012). En effet, le partitionnement aléatoire produit des chemins sensiblement plus courts pour les anomalies. Ainsi, lorsqu'une forêt d'arbres aléatoires produit des trajets plus courts pour des points particuliers, il est très probable qu'il s'agisse d'anomalies. La Fig 3.4 montre un exemple de détection d'anomalies avec l'algorithme *IsolationForest* de la bibliothèque scikit-learn sur des données synthétiques.

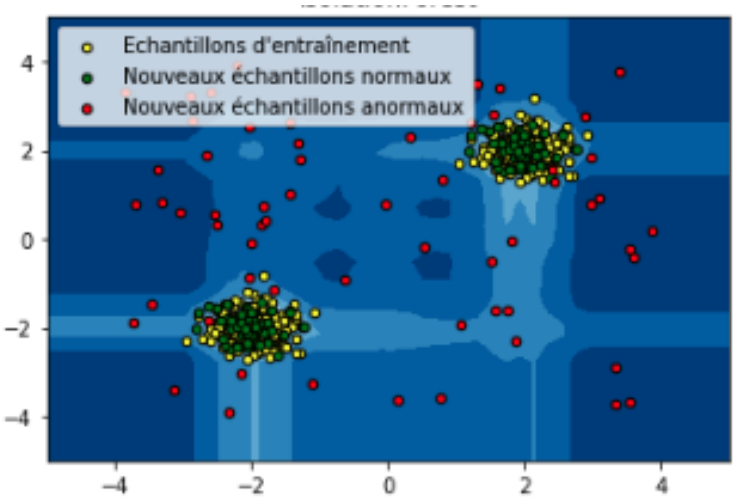


Figure 3.4: Détection d'anomalies avec l'algorithme IsolationForest

Facteur de valeurs localement aberrantes

Le facteur de valeurs localement aberrantes ou *local outlier factor* (LOF) est une technique de détection d'anomalies reposant sur les méthodes de plus proches voisins. C'est un algorithme qui produit un score d'anomalie afin de représenter les valeurs aberrantes dans l'ensemble de données. Pour ce faire, il mesure l'écart de densité local d'un point de données par rapport aux points de données à proximité. La densité locale est déterminée en estimant les distances entre les points de données qui sont voisins (k-plus proches voisins). Ainsi, pour chaque point de données, une densité locale peut être calculée. En les comparant, il est possible de distinguer entre les points de données qui ont des densités similaires et ceux qui ont une densité inférieure à leur ses voisins. En effet, les points de données avec les densités les plus faibles sont considérés comme des anomalies (Breunig et al., 2000). L'algorithme commence par calculer les k-distances, c'est-à-dire les distances calculées pour chaque point afin de déterminer ses k-plus proches voisins. En fonction de ces distances, le k^{e} point le plus proche est dit le k^{e} voisin le plus proche du point. La Fig 3.5 montre comment sont représentées les k-distances dans le cluster du point A.

Cette distance est utilisée pour calculer la distance d'atteignabilité qui est

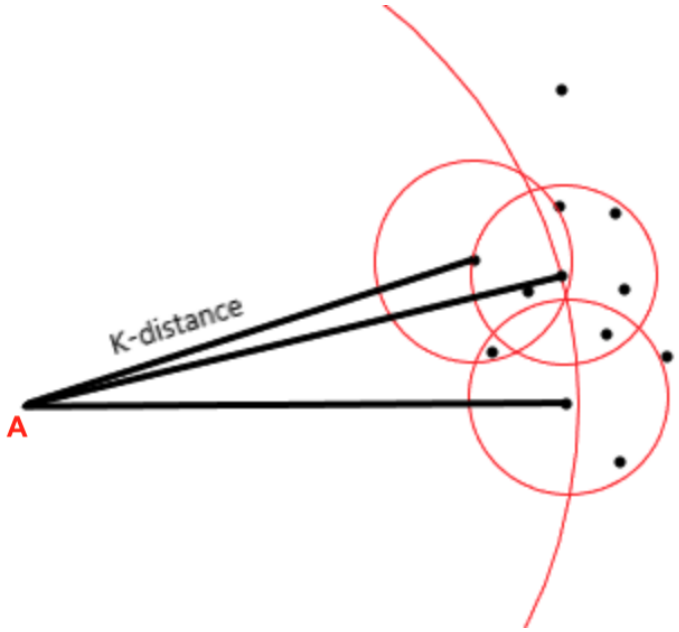


FIGURE 3.5: Illustration des k-distances pour le LOF

définie comme le maximum de la distance entre deux points et la k-distance de ce point. Elle s'écrit :

$$\text{reach-distance}_k(A, B) = \max(k\text{-distance}(B), d(A, B)) \ . \tag{3.4}$$

La Fig 3.6 montre la distance d'atteignabilité. Ainsi, les points A et B ont la même distance d'atteignabilité, alors que le point D n'est pas un k plus proche voisin pour $k = 3$.

Les distances d'atteignabilité de tous les k plus proches voisins d'un point sont calculées pour déterminer la densité d'atteignabilité locale (lrd) de ce point qui est une mesure de la densité des k points les plus proches autour d'un point. Elle est calculée en prenant l'inverse de la somme de toutes les distances d'atteignabilité de tous les k plus proches voisins. Par conséquent, plus les points sont proches, plus la distance est petite, et plus la densité est grande. La densité d'atteignabilité

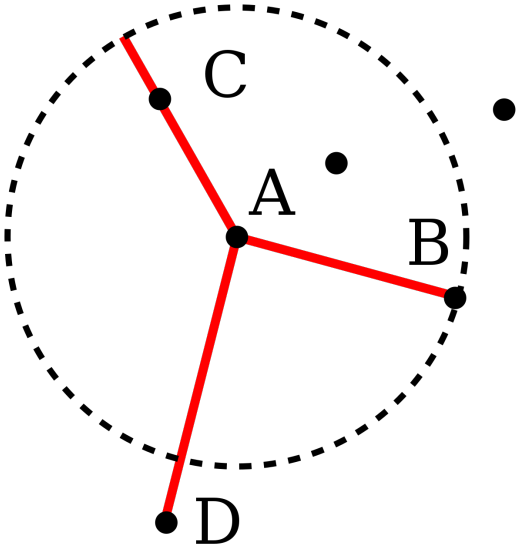


Figure 3.6: Illustration de la distance d'atteignabilité pour le LOF

locale s'écrit :

$$\mathrm{lrd}_k(A) = \frac{1}{\frac{1}{|N_k(A)|}\sum_{B \in N_k(A)} \text{reach-distance}_k(A, B)} \,. \tag{3.5}$$

Le calcul du facteur de valeur localement aberrante se fait en prenant le rapport de la moyenne des distances d'atteignabilité du nombre k des voisins d'un point sur la distance d'atteignabilité de ce point. Il s'écrit :

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{\mathrm{lrd}_k(B)}{\mathrm{lrd}_k(A)}}{|N_k(A)|} = \frac{1}{|N_k(A)|\,\mathrm{lrd}_k(A)} \sum_{B \in N_k(A)} \mathrm{lrd}_k(B) \,. \tag{3.6}$$

Si les densités des voisins et du point sont presque égales, les points sont supposés être similaires. Alors que si la densité des voisins est inférieure à la densité du point, le point est considéré comme un *inlier*, c'est-à-dire un point à l'intérieur du cluster. Si la densité des voisins est supérieure à la densité du point, il est considéré comme un *outlier*, c'est-à-dire une anomalie. La Fig 3.7 montre un exemple d'application de l'algorithme LOF possible avec la classe *LocalOutlierFactor* de la bibliothèque Scikit-learn sur un ensemble de données synthétiques.

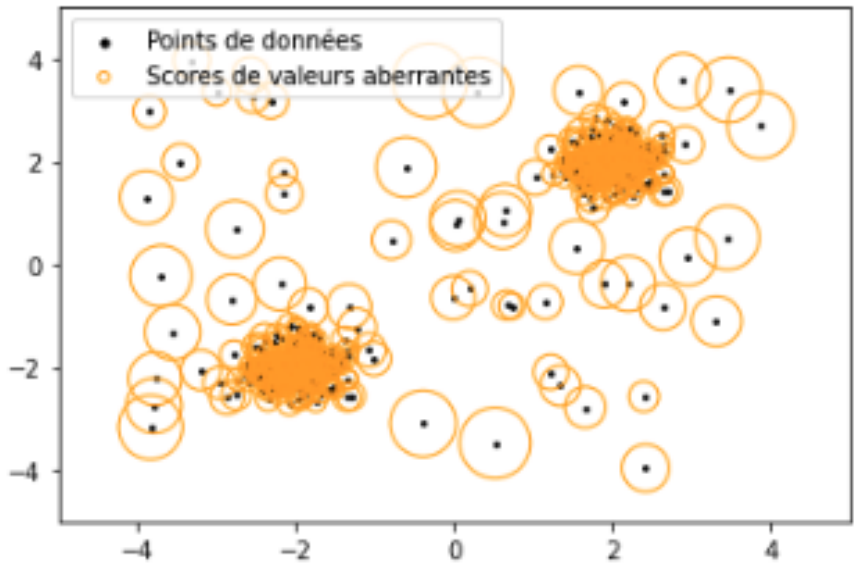


Figure 3.7: Scores d'anomalies avec l'algorithme LOF

3.2 Contribution à la détection d'anomalies dans les séries temporelles

La plupart des approches de détection d'anomalies dans les séries temporelles nécessitent une base d'apprentissage ne contenant que des points de données normaux. Leur objectif commun est de détecter si une série temporelle de test est anormale par rapport à cette base d'apprentissage. Ce processus de détection s'inscrit dans le cadre de la détection semi-supervisée et repose principalement sur la modélisation de la normalité à partir de la base d'apprentissage. Cela permet une analyse de similarité et une comparaison des séries temporelles en mesurant les distances des séries de test par rapport aux modèles appris, et un seuil est utilisé pour déterminer les anomalies.

L'objectif des algorithmes est de trouver des frontières de décision efficaces des observations normales qui donnent le meilleur rappel et la meilleure précision pour les anomalies prédites. Dans cette partie, nous présentons une nouvelle méthode de détection d'anomalies semi-supervisée dans les séries temporelles qui repose sur des métriques de transport optimal.

3.2.1 Transport optimal pour l'exploration des séries temporelles

L'étude de la similarité est abordée dans la littérature liée à la théorie du transport optimal pour l'analyse des séries temporelles. L'étude de Zhang, Tang et Corpetti (2020) propose un nouveau cadre de mesure de la similarité de séries temporelles appelé *Time Adaptive Optimal Transport* qui repose sur une mesure de distance en transport optimal pour les histogrammes et les distributions de probabilité. Il hérite de plusieurs propriétés puissantes du transport optimal pour résoudre les problèmes des méthodes d'analyse de similarité classiques reposant sur la déformation temporelle dynamique ou *Dynamic Time Warping* (DTW).

Une autre étude concernant la décomposition des flux de données des séries temporelles en segments disjoints, incluant le regroupement de segments non adjacents similaires et la détection du point de changement, a été proposée par Cheng et al. (2020). Les auteurs apportent un nouvel algorithme qui repose sur la caractérisation théorique d'une distance de Wasserstein "lissée" pour aborder le problème du test non-paramétrique de deux variables aléatoires (Ramdas, Trillos et Cuturi, 2017). Ce test utilise le transport optimal pour rejeter l'hypothèse nulle, c'est-à-dire décider si deux fonctions de densité de probabilité empiriques proviennent de la même distribution. Cheng et al. (2020) utilisent le résultat de ce test comme base pour un test de fenêtre glissante afin d'identifier les points de changement dans une série temporelle scalaire.

On trouve des applications du transport optimal pour trois problèmes de fusion des données de télédétection dans l'étude de Courty et al. (2016) : l'adaptation de domaine, le moyennage de séries temporelles et la détection de changement dans les données LIDAR. Une autre étude de comparaison des distributions dans les séries temporelles a été menée par Marti et al. (2016) dans le cadre d'une méthodologie de regroupement des séries temporelles multivariées. Une méthode de clustering est ainsi proposée, qui s'appuie sur des copules de distributions codant la structure de dépendance entre plusieurs variables aléatoires. La prise en compte des informations de dépendance est assurée par le calcul de la distance de Wasserstein entre les couples de distributions.

Cependant, la plupart des méthodes impliquant le transport optimal pour l'analyse des séries temporelles, en particulier l'étude de similarité et la comparaison des distributions de probabilités des séries temporelles considèrent les séquences de données dans le domaine temporel.

Généralement pour analyser statistiquement une série temporelle, les éléments de la séquence des données sont considérés comme un ensemble de variables aléatoires. Il faut donc supposer que la structure du processus stochastique qui génère les observations est essentiellement invariante dans le temps. Les hypothèses conventionnelles se résument à la condition de stationnarité. Dans sa forme forte, cette condition exige que deux segments de longueur égale qui sont extraits de la série temporelle doivent avoir des densités de probabilité identiques. La condition de faible stationnarité exige seulement que les éléments de la série temporelle aient une valeur attendue finie commune et que l'autocovariance de deux éléments ne dépende que de leur séparation temporelle.

Cette hypothèse de stationnarité pousse à examiner les séries temporelles dans une autre base de représentation, celle du domaine spectral où les données sont représentées en termes de répartition de leur énergie sur les fréquences. Ainsi, le contenu spectral d'un signal est donné par sa densité spectrale de puissance (PSD). La comparaison des séries temporelles à travers leurs PSD est de plus en plus courante principalement en raison des avancées récentes sur les possibilités d'estimation spectrale robustes au bruit, aux valeurs manquantes et aux artefacts d'acquisition généraux (Choudhuri, Ghosal et Roy, 2004 ; Wang, Khardon et Protopapas, 2012 ; Tobar, 2018 ; Turner et Sahani, 2014).

Cependant, les distances sur lesquelles repose l'étude de comparaison entre les fonctions ne sont d'aucune utilité lorsque les supports spectraux des fonctions diffèrent. Par conséquent, le développement de métriques solides pour comparer les PSD reste essentiel. Prenons l'exemple de la distance euclidienne ou même la divergence de Kullback-Leibler (KL). Lorsqu'elles sont utilisées pour la comparaison des distributions de densités de puissance, elles n'ont un sens que si ces distributions partagent l'ensemble des composants fréquentiels qui sont présents dans la série temporelle. Autrement dit, elles sont des divergences temporelles, elles comparent

les valeurs point par point de deux distributions pour une bande de fréquence spécifique.

L'objectif est de percevoir la façon dont la densité spectrale est distribuée sur les différentes fréquences, d'où la nécessité d'une distance fréquentielle. Notre approche de détection d'anomalies repose alors sur la quantification des variations locales de décalage des séries temporelles quelque soient leurs supports spectraux à l'aide d'une distance fréquentielle s'appuyant sur la théorie du transport optimal.

3.2.2 Méthodes de classification utilisant le transport optimal

Dans le but de concevoir une approche robuste de détection d'anomalies dans les séries temporelles, nous nous sommes intéressés à une distance spectrale qui soit sensible aux variations de puissance des séries temporelles. Cette approche repose sur la théorie du transport optimal et en particulier la distance de Wasserstein donnée par le coût optimal du transport entre deux distributions de probabilité. Même si la distance de Wasserstein a déjà fait ses preuves dans des applications en apprentissage automatique grâce à la régularisation entropique, qui permet une estimation rapide, c'est la première contribution dans ce domaine pour la détection d'anomalies dans les séries temporelles.

Nous présentons dans cette partie deux algorithmes de classification à une classe des séries temporelles, le premier est paramétrique et le deuxième est non paramétrique. Dans l'approche proposée, nous considérons les distributions discrètes suivantes :

$$\mu_s = \Sigma_{i=1}^{n_s} a_i \delta_{x_i^s}, \quad \mu_t = \Sigma_{j=1}^{n_t} b_j \delta_{x_j^t} \ , \tag{3.7}$$

où $x_i^s, x_j^t \in \Omega^2$, $a \in \Sigma_{n_s}$ $b \in \Sigma_{n_t}$, et $\Sigma_n = \{(a_i)_i \geq 0, \ \Sigma_{i=1}^n a_i = 1\}$, et les échantillons x_i et x_j peuvent être stockés dans des matrices X_s et X_t telles que $X = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$.

Nous considérons donc le problème du transport entre les distributions μ_s et μ_t avec la matrice de coût C et le plan de transport $U(a, b)$. Le plan de transport est défini dans l'Eq (2.24) et la distance de Wasserstein est la solution de l'Eq (2.25).

Comme nous l'avons vu dans le chapitre précédent, il est intéressant d'approcher la solution de ce problème en utilisant un critère de régularisation entropique, comme définie dans l'Eq. (2.28). La distance de Wasserstein incluant cette régularisation est définie dans l'Eq. (2.29).

L'unique solution de ce problème est de la forme $P_{i,j} = u_i K_{i,j} v_j$ avec $u_i, v_j \in \mathbb{R}_+^n$. L'algorithme Sinkhorn pour le calcul rapide de la distance de Wasserstein est résolu en appliquant itérativement les fonctions de mise à jour suivantes pour l'itération $l+1$:

$$u^{(l+1)} = \frac{a}{Kv^{(l)}} \text{ et } v^{(l+1)} = \frac{b}{K^T u^{(l)}}, \tag{3.8}$$

Le détail du calcul de la distance Sinkhorn est fourni dans l'algorithme 1, cette distance nous permet d'étudier la similarité des séries temporelles.

L'approche de détection d'anomalies proposée consiste à modéliser le comportement normal d'une série temporelle issue d'un signal capteur à partir d'un ensemble de données d'apprentissage et identifier les anomalies à partir des séries temporelles de test. Le signal anormal peut être défini par la distance entre sa

propre représentation et la représentation du modèle de référence.

Algorithm 2: Méthode paramétrique pour la détection d'anomalies

Data: set of reference signals X, signal to evaluate $\hat{X}$

Result: binary classification, 1 if normal signal, -1 if abnormal

$F(\bar{X}) \leftarrow \frac{1}{k} \sum_k F(X_k)$

for $i \leftarrow 1$ **to** k **do**

  $d_i \leftarrow d_C^\epsilon(F(\bar{X}), F(X_i))$

end

Set threshold ϑ from LogNormal fit on $\{d_i\}_{i=1...k}$

$\hat{d} \leftarrow d_C^\epsilon(F(\bar{X}), F(\hat{X}))$

if $\hat{d} > \vartheta$ **then**

  return -1

end

else

  return 1

end

Méthode paramétrique

Considérons un ensemble k de signaux initiaux $X = \{X_i\}_{i=1,...,k}$, avec $X \in \mathbb{R}^t$ et le signal à évaluer $\hat{X}$. Les signaux sont considérés dans le domaine fréquentiel, en estimant la densité spectrale de puissance. Pour chaque signal, la densité spectrale de puissance est calculée avec l'estimateur de Welch $F(\cdot)$, c'est-à-dire que les signaux sont divisés en plusieurs segments partiellement superposés, une fonction de fenêtrage est appliquée (ici Hamming) avant de calculer leur transformée de Fourier, et de moyenner les résultats pour chaque segment. Les signaux résultants sont dénotés $F(X) \in \mathbb{R}_+^n$.

Les signaux initiaux sont ensuite moyennés, pour obtenir un barycentre servant de référence $F(\overline{X}) = \frac{1}{k} \sum_k (X_k)$. Les distances entre les PSD individuels $F(X_k)$ et le barycentre de référence $F(\overline{X})$ sont calculées avec la distance Sinkhorn $d_C^\epsilon(F(\overline{X}), F(X_k))$, en utilisant une fonction de coût de Chebyshev qui est équi-

valente à la limite de la métrique L_p lorsque $p \to \infty$. Il en est de même pour les signaux de test $d_C^\epsilon(F(\overline{X}), F(\hat{X})$. Une distribution suivant une loi log-normale est alors obtenue à partir de l'histogramme des distances. Cette distribution nous permet de fixer un seuil de distance à partir duquel un classificateur est conçu afin de prédire le signal testé comme étant une anomalie ou non. Les différentes étapes de la méthode sont présentées dans l'algorithme 2.

Méthode non-paramétrique

Une version plus robuste de l'algorithme paramétrique est proposée ci-après. L'hypothèse selon laquelle la distribution des distances entre le barycentre et les signaux d'apprentissage suit une distribution log-normale peut être violée. En cas de distribution différente telle que binomiale ou autre, une décision reposant sur une mauvaise hypothèse peut induire une décision erronée. Un autre problème se pose lorsque l'anomalie à détecter est restreinte à une bande passante spécifique. Dans ce cas, l'anomalie peut ne pas être détectée car la variation induite est diluée dans tout le spectre de fréquences.

Pour atténuer ces problèmes, nous nous sommes appuyés sur une approche non paramétrique estimée sur une décomposition de signal en banc de filtres, c'est-à-dire un ensemble de filtres passe-bandes qui décomposent le signal en plusieurs composants, chacun portant une seule sous-bande de fréquence. La PSD du signal est analysée indépendamment pour f bandes de fréquences différentes $B = b_1, \ldots, b_f$, cela permet de détecter des variations anormales se produisant dans des bandes passantes étroites. Les bornes inférieure et supérieure de la distribution des distances pour chaque bande de fréquences b sont estimées comme le premier et le dernier centile, $p^b_{0.01} = d_C^\epsilon(F(\bar{X}^b), F(X^b_{0.01}))$ et $p^b_{0.99} = d_C^\epsilon(F(\bar{X}^b), F(X^b_{0.99}))$.

Le score d'anomalie pour chaque bande de fréquence b est calculé comme :

$$A^b_{\text{lower}} = \frac{d_C^\epsilon(F(\bar{X}^b), F(\tilde{X}^b))}{p_{0.01}}, A^b_{\text{upper}} = \frac{d_C^\epsilon(F(\bar{X}^b), F(\tilde{X}^b))}{p_{0.99}} \ . \tag{3.9}$$

Une valeur supérieure à 1 indique un échantillon anormal pour la bande consi-

dérée. La fonction de décision $g(\tilde{X})$ repose sur une combinaison du score pour toute la bande de fréquence f :

$$g(\tilde{X}) = \begin{cases} -1 \text{ si } \frac{1}{f}\sum_i A^{b_i}_{\text{lower}} > 1 \text{ ou si } \frac{1}{f}\sum_i A^{b_i}_{\text{upper}} > 1 \\ 1 \text{ par ailleurs} \end{cases} \tag{3.10}$$

Les différentes étapes de la méthode non paramétrique sont détaillées dans l'algorithme 3.

Algorithm 3: Méthode non paramétrique

Data: set of reference signals X, signal to evaluate $\hat{X}$

Result: binary classification, 1 if normal signal, -1 if abnormal

for $j \leftarrow 1$ **to** f **do**

  $F(\bar{\mathbf{X}}^{b_j}) \leftarrow \frac{1}{k}\sum_k F(X^{b_j}_k)$

  Get $p^{b_j}_{0.01}$ and $p^{b_j}_{0.99}$

  Compute A^b_{lower} and A^b_{upper} from Eq. (3.9)

end

return $g(\tilde{X})$, as in Eq. (3.10)

3.3 Contributions aux méthodes des plus proches voisins utilisant le transport optimal

Les techniques de détection d'anomalies des plus proches voisins supposent que les instances de données normales se produisent dans des zones de voisinage denses tandis que les anomalies se produisent loin de leurs voisins les plus proches. Elles nécessitent une mesure de distance ou de similarité définie entre deux instances de données. La distance ou la similitude peuvent être calculées de différentes manières. La distance d'une instance de données à son voisin le plus proche où la densité relative de chaque instance de données est définie comme le score d'anomalie dans les techniques de détection d'anomalies reposant sur le plus proche voisin.

Il existe plusieurs variantes de cette technique de base ; Elles ont été abordées par la communauté scientifique de trois manières différentes. La première manière

consiste à modifier la définition du score d'anomalie, la deuxième manière consiste à sélectionner différentes mesures de distance ou de densité pour les différents types de données, et la troisième manière consiste à réduire la complexité du calcul. Dans ce manuscrit, nous nous intéressons à la deuxième manière, c'est-à-dire les techniques des k plus proches voisins pour la détection d'anomalies en utilisant les mesures de distance. En effet, nous discuterons de l'impact du choix de la fonction de distance sur les performances des approches kNN et nous mettrons à contribution une nouvelle méthode de détection d'anomalies inspirée de la méthode LOF mais utilisant une métrique de distance robuste.

3.3.1 Effet de la fonction de distance sur les approches k-NN

L'approche des k-plus proches voisins (kNN) est l'une des méthodes les plus courantes pour les applications de détection d'anomalies, ses performances dépendent principalement de la mesure de similitude entre les données d'apprentissage et les données de test à travers le calcul de distance entre ces dernières.

Ceci soulève une question majeure sur les fonctions de distance à adopter dans une approche kNN parmi un grand nombre de métriques de distances et de similarité disponibles. La littérature témoigne de quelques revues abordant cette question en évaluant les performances des approches kNN en fonction de différentes métriques de distance sur des données diversifiées. L'étude de Abu Alfeilat et al. (2019) a évalué les performances du kNN en fonction de huit familles de distances, notamment les distances L_p de Minkowski, les mesures L_1 qui induisent un critère de parcimonie, les mesures de distance de produit interne, les mesures de distance de la corde au carré (SCD), les mesures de distance euclidienne L_2, les mesures de distance d'entropie de Shannon incluant la divergence de Kullback-Leibler, et bien d'autres qui sont considérées comme des métriques respectant des propriétés qui incluent la non-négativité, l'identité des indiscernables, la symétrie et l'inégalité triangulaire. Il est à noter que la divergence de Kullback-Leibler n'est pas une métrique, car elle n'est pas symétrique et elle ne satisfait pas la propriété d'inégalité triangulaire, elle est donc appelée quasi-distance.

D'autres études ont été menées mais dans des cadres applicatifs spécifiques, comme (Hu et al., 2016), où les auteurs ont analysé l'effet des mesures de distance sur un kNN pour les ensembles de données du domaine médical. Leurs expériences étaient basées sur trois types différents d'ensembles de données médicales contenant des types de données catégoriques, numériques et mixtes, choisis dans le référentiel d'apprentissage automatique (UCI), et quatre mesures de distance, notamment la distance euclidienne, la Cosine, la χ^2 et les distances de Minkowski. Ils ont divisé chaque ensemble de données en 90% des données d'entraînement et 10% en tant qu'ensemble de test, avec des valeurs k allant de 1 à 15. Les résultats expérimentaux ont montré que la fonction de distance du χ^2 était le meilleur choix pour les trois différents types d'ensembles de données. Cependant, les métriques de distance Cosine, euclidienne et Minkowski ont donné des résultats avec la plus faible précision sur le type mixte d'ensembles de données.

Dans leur étude, Mulak et Talhar (2015) ont évalué les performances du classificateur kNN à l'aide de mesures de distances Chebychev, euclidienne et Manhattan sur un ensemble de données numériques qui contient 41 caractéristiques et 2 classes. L'ensemble de données a été normalisé avant de mener l'expérience. Pour évaluer les performances du kNN, des mesures de précision, de sensibilité et de spécificité ont été calculées pour chaque distance. Les résultats rapportés indiquent que l'utilisation de la distance Manhattan surpasse les autres distances testées. Cette liste des travaux n'est pas exhaustive mais elle permet de donner un aperçu sur l'importance du choix de la mesure de distance dans une approche kNN.

Cependant, très peu d'études abordent la question du bruit dans les données et son impact sur les performances du kNN. Ce type de résultats peut permettre l'utilisation des métriques de distance les moins affectées par le bruit. Ce questionnement nous a mené à pousser la réflexion sur le travail de (Chazal, Cohen-Steiner et Mérigot, 2011) qui présente l'inférence géométrique pour les mesures de probabilité. L'objectif principal est de définir des fonctions de distance des points à des nuages de points qui sont robustes aux valeurs aberrantes, afin que les résultats montrent une bonne stabilité et puissent être utilisés pour faire de l'inférence topologique. Le coeur de leur contribution consiste à remplacer la

fonction de distance habituelle par une autre notion de fonction de distance robuste à l'ajout d'un certain nombre de valeurs aberrantes ou du bruit en général. Pour ce, ils ont changé l'interprétation des nuages de points. Au lieu de les considérer seulement comme des objets purement géométriques, ils les considèrent aussi porteurs d'une notion de masse. Formellement, ils remplacent les sous-ensembles compacts de $\mathbb{R}^d$ par des mesures de probabilité finies dans un espace métrique et calculent la distance entre deux mesures de probabilité par la distance de Wasserstein. Ceci leur permet de quantifier le coût minimal de transport d'une mesure à l'autre. Les résultats expérimentaux ont montré que le changement de la distance euclidienne par la distance de Wasserstein a conservé toutes les propriétés requises pour étendre les résultats d'inférence au cas où les données peuvent être corrompues par des valeurs aberrantes ou du bruit.

Ce constat nous a encouragé à adopter le même principe pour les approches kNN afin d'étendre leur efficacité et leur robustesse à des données bruitées. Nous proposons donc, d'utiliser la distance de Wasserstein régularisée par l'entropie grâce à l'algorithme Sinkhorn pour mesurer la distance entre les données d'apprentissage et les données de test dans une approche k-plus proches voisins. Nous appliquons cette approche dans le cadre de l'algorithme LOF où la distance de Wasserstein nous servira pour estimer la densité d'atteignabilité locale donnée par les k-plus proches voisins. Le détail de la méthode est présenté dans la partie suivante.

3.3.2 LOFO - Amélioration de la méthode LOF

L'algorithme LOF consiste à attribuer un score d'anomalie appelé LOF à chaque point de données, ces scores sont ensuite comparés entre eux afin de trouver les valeurs aberrantes. Plus la valeur du LOF d'un point de données est élevée, plus il y a de chances qu'il s'agisse d'une valeur aberrante. Pour calculer les valeurs LOF, plusieurs paramètres importants sont utilisés :

— k-distance,

— k-plus proches voisins,

— distance d'atteignabilité,
— densité d'atteignabilité locale,

Le détail de ces paramètres est donné dans la section 3.1.2. Plusieurs variantes de l'algorithme LOF ont été proposées depuis son apparition. Un facteur aberrant basé sur la distance locale (LDOF) utilisant la distance relative d'un objet à ses voisins est proposé pour la détection des valeurs aberrantes dans des ensembles de données dispersés (Zhang, Hutter et Jin, 2009), et un score de valeur aberrante INFluenced (INFLO) est proposé en tenant compte à la fois des voisins et de l'inverse des voisins d'un objet lors de l'estimation de sa distribution de densité relative (Jin et al., 2006). Considérant les modèles sous-jacents des données, Tang et al. (2001) ont proposé un schéma de facteur aberrant basé sur la connectivité (COF). Plusieurs autres méthodes basées sur la densité ont été proposées sur la base de l'estimation de la densité par noyau (Schubert, Zimek et Kriegel, 2014 ; Gao et al., 2011 ; Latecki, Lazarevic et Pokrajac, 2007), comme le facteur de densité locale (LDF) et l'intégrale de la corrélation des valeurs aberrantes locales (LOCI).

Dans ce manuscrit, nous proposons une méthode de détection des valeurs aberrantes inspirée de l'algorithme LOF mais qui repose sur l'estimation de la densité du noyau local et la théorie du transport optimal pour le calcul de la distance d'atteignabilité afin d'assurer une détection des valeurs aberrantes robuste quelque soit la nature des données à exploiter.

Estimation de la densité du noyau local

Nous utilisons une méthode de Parzen pour évaluer la densité à l'emplacement d'un objet en fonction de l'ensemble de données. Étant donné un ensemble d'objets $X = \{X_1, X_2, \ldots, X_n\}$, où $X_i \in \mathbb{R}^d$ pour $i = 1, 2, \ldots, n$, la méthode de Parzen estime la densité comme suit :

$$P(X) = \frac{1}{n}\Sigma_{i=1}^{n}\frac{1}{h^d}K(\frac{X - X_i}{h}) \ , \tag{3.11}$$

avec $K(\frac{X-X_i}{h})$ est la fonction noyau définie avec la largeur du noyau h, qui satisfait les conditions suivantes :

$$\int K(u)du = 1, \int uK(u)du = 0 \text{ et } \int u^2K(u)du > 0 \tag{3.12}$$

Une fonction à noyau gaussien multivarié couramment utilisée est donnée par :

$$K\Big(\frac{X-X_i}{h}\Big)_{Gaussien} = \frac{1}{(2\pi)^{d/2}} \exp\big(-\frac{|X-X_i|^2}{2h^2}\big) , \tag{3.13}$$

avec $|X-X_i|^2$ la distance euclidienne entre X et X_i. L'estimation de la distribution dans l'équation 3.11 offre de nombreuses propriétés intéressantes, telles que sa propriété non paramétrique, sa continuité et sa différentiabilité. C'est aussi un estimateur asymptotique sans biais de la densité.

Pour estimer la densité à l'emplacement d'un objet X_i, nous considérons ses k-plus proches voisins comme des noyaux au lieu d'utiliser tous les objets de l'ensemble de données. Ceci est pour deux raisons : premièrement, l'estimation de la densité utilisant l'ensemble de données complet peut perdre la différence locale de densité et ne pas détecter les valeurs aberrantes locales, deuxièmement, la détection des valeurs aberrantes calcule le score pour chaque objet, et l'utilisation de l'ensemble de données complet entraînerait un coût de calcul élevé.

Pour mieux estimer la distribution de densité au voisinage d'un objet X_i, nous utilisons les k-plus proches voisins à travers les données de la $kdistance(X_i)$ à partir desquelles la fonction de densité est estimée à l'aide d'un noyau gaussien.

Calcul de la densité d'atteignabilité locale utilisant le transport optimal

La densité d'atteignabilité locale d'un point de données X_i est l'inverse de la distance d'atteignabilité moyenne de X_i à partir de son voisinage. Fondamentalement, elle mesure la proximité des points de données de voisinage de X_i de celui-ci. Plus la densité est faible, plus X_i est éloigné de ses voisins. Dans l'algorithme LOF classique, la distance d'atteignabilité d'un point de données X_j à partir de X_i est

le maximum de la $kdistance(X_j)$ et la distance réelle entre X_i et X_j.
Dans notre approche, nous définissons la distance d'atteignabilité autrement. En effet, cette distance est calculée par la métrique de Wasserstein régularisée par l'entropie entre la densité locale de la distribution des données de la $kdistance(X_i)$ de l'objet X_i et la densité moyenne au sens du transport optimal des objets X_j du voisinage noté $N_k(X_i)$.

Formellement, cette distance se définit comme :

$$rd^w(X_i) = d_C^\epsilon(P_{kdistance}(X_i), \bar{P}_{kdistance}(X_j)) \ , \tag{3.14}$$

où $X_j \in N_k(X_i)$, et $\bar{P}_{kdistance}(X_j)$ est le barycentre de Wasserstein régularisé par l'entropie des densités des distributions des k-distances du voisinage $N_k(X_i)$.

À partir de l'Eq (3.14), la densité d'atteignabilité locale peut être définie comme :

$$lrd^w(X_i) = \frac{1}{\Sigma_{X_j \in N_k(X_i)} \left\{ \frac{rd^w(X_i)}{|N_k(X_i)|} \right\}} \ . \tag{3.15}$$

Par conséquent, le facteur de valeur aberrante pour chaque objet X_i s'écrit :

$$LOFO(X_i) = \frac{\Sigma_{X_j \in N(X_i)} lrd^w(X_j)}{|N(X_i)|} \times \frac{1}{lrd^w(X_i)} \tag{3.16}$$

La formule du facteur de valeur aberrante montre la relation entre la densité d'atteignabilité locale $lrd^w(X_i)$ notée A et la densité d'atteignabilité locale moyenne des points de données au voisinage de X_i $\frac{\Sigma_{X_j \in N(X_i)} lrd^w(X_j)}{|N(X_i)|}$ notée B.

— $LOFO(X_i)$ est grand si A est petit et B est grand (*outlier*).
— $LOFO(X_i)$ est petit si A est grand et B est petit (*inlier*).

La classification se fait alors grâce aux scores $LOFO(X_i)$:

— $LOFO(X_i) < 1$ signifie une densité plus élevée que celle des voisins (*inlier*).
— $LOFO(X_i) > 1$ signifie densité inférieure à celle des voisins (valeur aberrante).

Les différentes étapes de la méthode sont présentées dans l'algorithme 4.

Algorithm 4: Algorithme LOFO : Local Outlier Factor with Optimal transport

Data: Data X, number of kNN $n\_neighbors$

Result: binary classification, 1 if inlier, -1 if outlier

for *each object* $X_i \in X$ **do**
  for $k \leftarrow 1$ **to** $n\_neighbors$ **do**
    Get $k - distance(X_i)$
  end
  Estimate KDE of $kdistance(X_i) : P_{kdistance}(X_i)$
  Calculate the entropic regularized Wasserstein barycenter of neighborhood KDEs : $\{\bar{P}_{kdistance}(X_j), X_j \in N_k(X_i)\}$
  Calculate reachability distance from eq. (3.14) : $rd^w(X_i)$
  Calculate Local reachability density from eq. (3.15) : $lrd^w(X_i)$
  Calculate LOFO score from eq. (3.16) : $LOFO(X_i)$
end
for *each object* $X_i \in X$ **do**
  if $LOFO(X_i) > 1$ **then**
    return -1
  end
  else
    return 1
  end
end

3.4 Conclusion

Le comportement normal des machines et des structures est décrit par des données qui suivent des modèles temporels réguliers. Inversement, un comportement anormal perturbe la régularité et provoque des déviations du modèle temporel régulier. La détection d'anomalies est l'ensemble des processus et méthodes mis

en place pour reconnaître automatiquement les modèles anormaux. La détection d'anomalies concerne un grand nombre d'applications dans le cadre de la surveillance de l'état de santé des machines et avec l'explosion de l'utilisation des capteurs, elle est devenue d'une importance capitale. Par exemple, elle est utilisée pour détecter les dommages structurels ou pour détecter les surcharges ou les souscharges. Les données sont collectées sous forme de séquences ou de séries temporelles. Par exemple, des séquences d'accélérations, de déplacements, de déformations, etc. Elles sont ensuite enregistrées pendant le fonctionnement. Un défaut entraîne des lectures anormales dans les séquences collectées à partir d'un ou plusieurs capteurs. Dans ce chapitre nous avons présenté plusieurs travaux dédiés à la détection d'anomalies dans le cadre du diagnostic des machines industrielles, particulièrement ceux qui reposent sur l'analyse des séries temporelles. Nous avons rappelé également les algorithmes de détection d'anomalies les plus répandus dans les applications de surveillance et de diagnostic des machines industrielles à savoir les machines à vecteurs de support, la covariance robuste, la forêt d'isolement et le facteur de valeurs localement aberrantes qui appartiennent à de grandes catégories de méthodes détaillées dans le deuxième chapitre, notamment les méthodes de classification, les méthodes statistiques et les méthodes des plus proches voisins. Ces algorithmes peuvent être entraînés de manière semi-supervisée en modélisant le comportement normal reposant sur la disponibilité des étiquettes positives ou de manière non supervisée où il y a une absence complète d'étiquettes. La détection d'anomalies dans les séries temporelles est souvent une classification binaire avec une distribution de classe fortement asymétrique d'où le problème d'apprentissage à partir de données déséquilibrées. Ce problème pourrait être résolu par certains mécanismes présentés dans le deuxième chapitre qui consistent à modifier un ensemble de données déséquilibrée en une répartition équilibrée.

Concernant les approches de détection d'anomalies semi-supervisées dans les séries temporelles, il s'agit le plus souvent de modéliser la normalité à partir de la base d'apprentissage (étiquettes positives) pour une analyse de similarité et une comparaison des séries temporelles en mesurant les distances des séries de test par rapport aux modèles appris, et un seuil est utilisé pour déterminer les anoma-

lies. Sur la base de ce principe, nous avons présenté dans ce chapitre une nouvelle méthode de détection d'anomalies semi-supervisée qui repose sur des métriques en transport optimal en vue de l'étude de similarité des séries temporelles. Notre approche se distingue par rapport aux méthodes existantes par l'utilisation d'une distance spectrale horizontale significative et sensible aux variations de puissance des séries temporelles. Deux algorithmes ont surgi de cette approche, un paramétrique et un non-paramétrique. Les développements initiaux ont mené à version paramétrique de la méthode de classification. En effet, l'étude de comparaison entre les séries temporelles d'apprentissage et celles de test s'est faite à travers le calcul de la distance de Wasserstein entre leurs densités spectrales de puissance. Cette première version de la méthode a pu se distinguer par rapport aux méthodes existantes, chose que nous démontrons dans le chapitre suivant. Cependant cette version paramétrique présente des limites, notamment lorsque l'anomalie à détecter est restreinte à une bande passante spécifique, c'est-à-dire la variation induite par l'anomalie ne peut pas être détectée car elle est diluée dans tout le spectre de fréquences. Pour remédier à ce problème, nous avons réfléchi à une version non paramétrique de la méthode en s'appuyant sur des statistiques non paramétriques calculées sur une décomposition de signal en banque de filtres, et la densité spectrale de puissance de la série temporelle est analysée indépendamment pour chaque bande de fréquence. C'est donc une version plus robuste de la méthode qui a pu devancer l'algorithme paramétrique et les algorithmes de l'état-de-l'art.

Les dernières avancées scientifiques sur la théorie du transport optimal et son implication dans le domaine de l'apprentissage automatique nous a poussé à explorer d'autres méthodes de détection d'anomalies et d'apporter des améliorations à l'existant. Dans cette perspective, nous nous sommes intéressé aux méthodes de détection d'anomalies non supervisées qui reposent sur les techniques du plus proche voisin. Les performances de ces techniques dépendent principalement de la métrique de distance utilisée pour la mesure de similarité entre les instances de données. Ce en supposant que les instances de données normales se produisent dans des zones de voisinage denses tandis que les anomalies se produisent loin de leurs voisins les plus proches. Nous avons mis donc à contribution une nouvelle

méthode facteur de valeur aberrante locale en transport optimal (LOFO) inspirée de l'algorithme facteur à valeur aberrante locale (LOF) en se référent à la distance et le barycentre de Wasserstein pour calculer la distance d'atteignabilité entre les densités de probabilités des distributions des k-distances entre les instances de données et leurs voisinages. Contrairement à la première contribution de ce manuscrit qui est réservée aux séries temporelles, LOFO est conçu pour agir plusieurs types de données. Le cinquième chapitre présentera les expérimentations et les résultats de l'algorithme LOFO sur des ensembles de données multidimensionnelles.

Chapitre 4

Méthodes de classification semi-supervisées en transport optimal pour la détection d'anomalies

Sommaire

Ce chapitre se concentre sur le premier volet des contributions du manuscrit, les méthodes de classification semi-supervisées pour la détection d'anomalies dans les séries temporelles. Nous rappelons dans la première section les mo-

dèles théoriques des méthodes de détection paramétrique OT et non-paramétrique *multiband-OT* déjà abordées en détail dans le chapitre 3. Ensuite, nous présentons dans la deuxième section l'analyse expérimentale de l'application de ces méthodes sur des données acoustiques. Nous décrivons d'abord les ensembles de données acoustiques synthétiques et réels, puis nous évaluons les méthodes proposées sur ces ensembles de données par des courbes ROC et nous les comparons aux algorithmes de l'état de l'art, SVM à une classe, facteur de valeur aberrante et forêt d'isolement par des métriques singulières.

4.1 Aperçu des méthodes de détection

L'approche de détection d'anomalies proposée est conçue pour modéliser les comportements normaux et identifier les comportements anormaux. Nous nous concentrons dans ce chapitre sur les séries temporelles des signaux acoustiques et/ou vibratoires, car il est facile d'intégrer de tels capteurs dans un système de surveillance non invasif. Néanmoins, l'approche en question pourrait être adaptée à divers problèmes, car toute distribution d'échantillons pourrait servir de caractéristique d'entrée. Un signal anormal peut être défini par la distance entre sa propre représentation et la représentation d'un autre signal défini comme référence. En effet, les signaux bruyants affichent la distance la plus élevée par rapport à la représentation de référence.

Considérons un ensemble k de signaux initiaux $X = \{X_i\}_{i=1,\dots,k}$, avec $X \in \mathbb{R}^t$ et le signal à évaluer $\tilde{X} = X + \eta N$ où N est la composante anormale et η est le niveau de cette composante. L'ensemble d'entraînement ne contient que des échantillons étiquetés positifs, tel est le cas des données de séries temporelles issues de la surveillance de l'état de santé des machines industrielles. Compte tenu du contexte applicatif des méthodes de détection d'anomalies, il est à noter que les algorithmes d'apprentissage s'appliquent sur des données déséquilibrées. Ce problème est abordé en détail dans la section 2.1.2. Dans cette section, nous rappelons brièvement les méthodes de classification mises à contribution à travers ce manuscrit et qui sont présentées en détail dans la section 3.2.2.

4.1.1 Méthode de classification paramétrique : OT

Le diagramme de la Fig 4.1 synthétise les différentes étapes de l'algorithme de classification paramétrique en transport optimal. Pour une détection semi-supervisée, les signaux d'entraînement sont considérés comme référence normale de fonctionnement. La densité spectrale de puissance est estimée pour chaque signal d'entrée par l'estimateur de Welch $F(\cdot)$ avec un fenêtrage de Hamming. Ces densités spectrales de puissance sont ensuite moyennées pour obtenir un barycentre de référence. Les distances Sinkhorn $d_C^\epsilon(F(\overline{X}), F(X_k))$ entre les PSD individuels $F(X_k)$ et le barycentre de référence $F(\overline{X})$ sont calculées en se référant à l'équation(**??**). La même procédure est appliquée sur les signaux de test $d_C^\epsilon(F(\overline{X}), F(\tilde{X}_k)$. Une distribution suivant une loi log-normale est ensuite obtenue de l'histogramme des distances. À partir de cette distribution, un seuil de distance a été déduit pour concevoir un classificateur capable de prédire le signal testé comme étant strictement inférieur à la normale ou anormal.

4.1.2 Méthode de classification non-paramétrique : multiband-OT

Dans la version non-paramétrique de la méthode de classification en transport optimal représentée par le diagramme de la Fig 4.2, des statistiques non paramétriques sont calculées sur une décomposition du signal fréquentiel en banc de filtres, La densité spectrale de puissance du signal est analysée indépendamment pour f bandes de fréquences différentes $B = b_1, \ldots, b_f$ d'où l'appellation de l'algorithme *multiband-OT*, cela permet de détecter des variations anormales se produisant dans des bandes passantes étroites. Les bornes inférieure et supérieure de la distribution des distances de Sinkhorn pour chaque bande de fréquences b sont estimées comme le premier et le dernier centile, $p^b_{0.01}$ et $p^b_{0.99}$. Les signaux de test subissent aussi le même traitement afin de définir les scores d'anomalies A^b_{lower} et A^b_{upper} pour chaque bande de fréquence b suivant l'équation (3.9). La fonction de décision $g(\tilde{X})$ permet de classifier les signaux de test comme normaux ou contenant des anomalies selon l'équation (3.10).

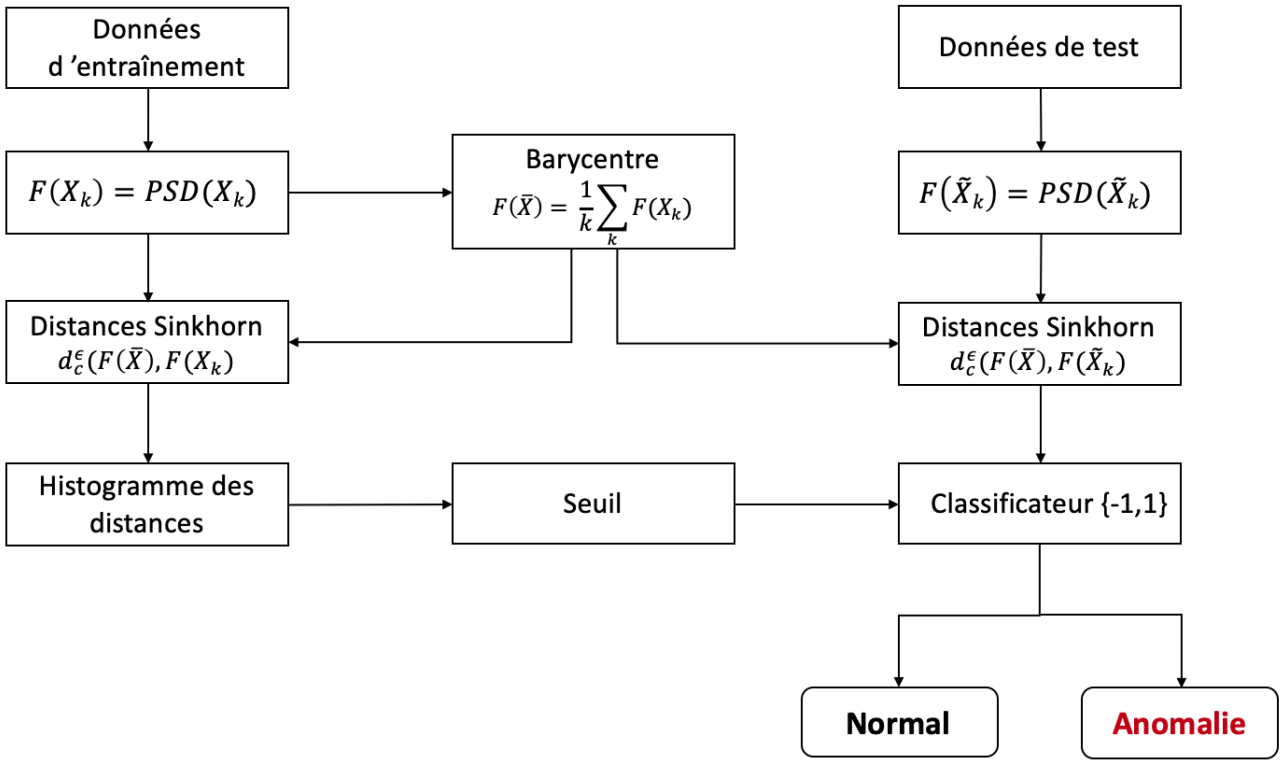


Figure 4.1: Diagramme représentatif de l'algorithme de classification paramétrique : OT

4.2 Analyse expérimentale et résultats

Dans cette partie, nous présentons l'analyse expérimentale et les résultats correspondants des méthodes de classification en transport optimal mises à contribution.

4.2.1 Description des ensembles de données

Afin d'assurer la reproductibilité de notre étude, et pour des raisons de confidentialité des données industrielles du projet dans lequel s'inscrit cette thèse, nous avons décidé d'appliquer notre approche sur des données publiques. Un premier lot de données est choisi pour démontrer la robustesse de la méthode. Il offre la

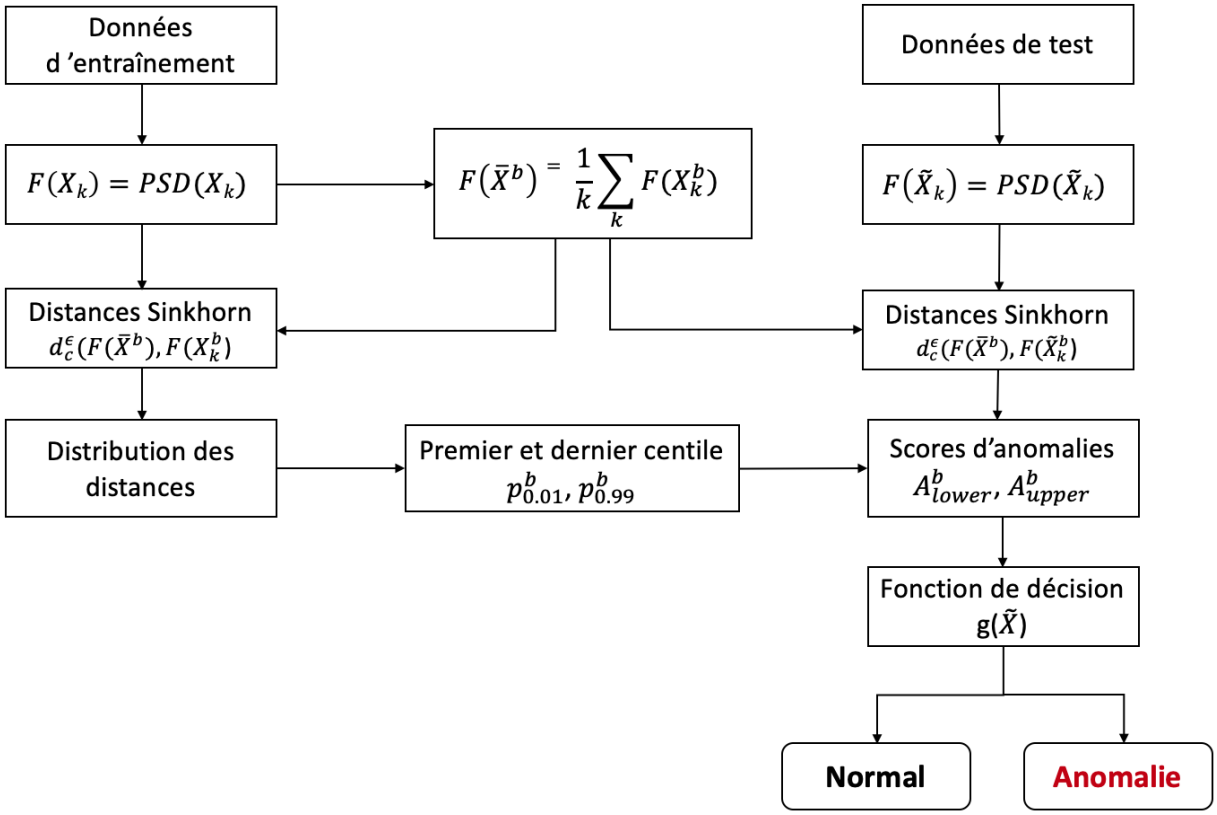


Figure 4.2: Diagramme représentatif de l'algorithme de classification non-paramétrique : multiband-OT

possibilité de vérifier que la distance en transport optimale a une réelle influence positive. Nous avons sélectionné un enregistrement sonore d'une journée de travail dans un espace ouvert de travail comme données de référence et différents niveaux de bruit rose sont mélangés pour simuler des données bruitées de manière contrôlée.

Pour tester davantage notre approche sur des données réalistes, nous avons également évalué la méthode sur un ensemble de données similaire aux données industrielles privées. Nous avons sélectionné un enregistrement du son d'une machine industrielle tournante pour le signal acoustique de référence, qui est proche de celui rencontré en situation industrielle. Nous avons également sélectionné des enregistrements de certains sons d'événements anormaux pour représenter les ano-

malies qui sont également similaires à celles se produisant sur le banc d'essai de simulation des pannes mécaniques que nous avons utilisé.

4.2.2 Déploiement des méthodes de détection

Dans un premier temps, la méthode de classification paramétrique OT est évaluée sur un enregistrement monorail codé à 44100 Hz de sons d'un espace ouvert de travail. L'enregistrement acoustique est d'une durée de 15 minutes. La référence est calculée sur les sept premières minutes, l'audio restant est soit laissé tel quel, soit corrompu avec du bruit rose. Deux types différents de bruit anormal sont considérés : un bruit à large bande est ajouté au signal ou un bruit de crête pointu. Cette première expérience a pour objectif de tester la robustesse de la méthode et de comparer les métriques en transport optimal aux métriques euclidiennes.

Afin d'évaluer les méthodes de classification paramétrique OT et non-paramétrique *multiband-*OT et de les comparer aux algorithmes de l'état de l'art, un signal de référence issu des émissions acoustiques d'un banc d'essai de simulation des pannes mécaniques est choisi. Le signal est enregistré en mono, à 44100 Hz pendant 15 min. Deux types qualitativement différents de pièces mécaniques défectueuses sont considérés : Le son d'un léger sifflement aigu (dataset 1) et un son cyclique grave (dataset 2), semblable à un roulement à billes défectueux. La Fig 4.3 montre les signaux considérés dans le domaine fréquentiel. Chaque ligne est une mesure dans le domaine fréquentiel, estimée avec la méthode de Welch.

Une validation croisée répétée k-fold est utilisée pour séparer l'ensemble de données d'entraînement (500 échantillons) et les données de test (500 échantillons). Les modèles de détection d'anomalies sont calibrés sur des données d'apprentissage pour calculer le signal de référence $F(\overline{X})$. L'ensemble de données comprend 3 niveaux de sons anormaux à détecter (bruit), le comportement normal de la machine étant mélangé à un bruit mécanique défectueux. Plus le niveau de bruit est élevé, plus il est facile à détecter.

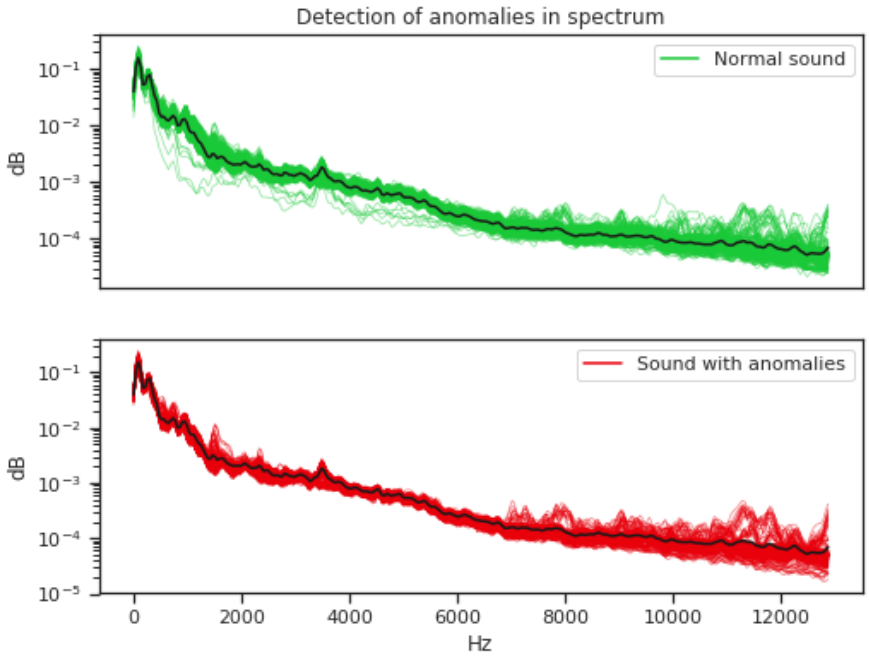


Figure 4.3: Exemples de sons normaux (en haut) et anormaux (en bas) extraits du premier ensemble de données

4.2.3 Évaluation des performances

Évaluation de la robustesse de la méthode paramétrique sur un ensemble de données synthétiques

L'algorithme 2 est évalué avec différents niveaux de bruit pour calculer une estimation AUC-ROC. Comme le montre la Fig 4.4, lorsque la puissance du signal de bruit est forte (SNR 3,65 dB), l'algorithme détecte facilement l'anomalie dans le signal. Cependant, la détection est plus difficile lorsque le pic de bruit est plus faible, avec un SNR de 0,55, l'aire sous la courbe (AUC) est de 0,62 ce qui est au-dessus du niveau aléatoire (0,5) mais d'une précision limitée. La deuxième expérience repose sur le même type de procédure, mais avec un bruit à large bande. La Fig 4.5 montre qu'à SNR équivalent, l'algorithme proposé est capable de détecter un événement anormal plus facilement que dans le cas du bruit de crête. Ceci se reflète sur les valeurs AUC qui sont de 0,72 pour un faible SNR de 0,55 dB, de 0,79 pour un SNR de 1,71 dB et de 0,88 pour 3,65 dB. Cette expérience sur un

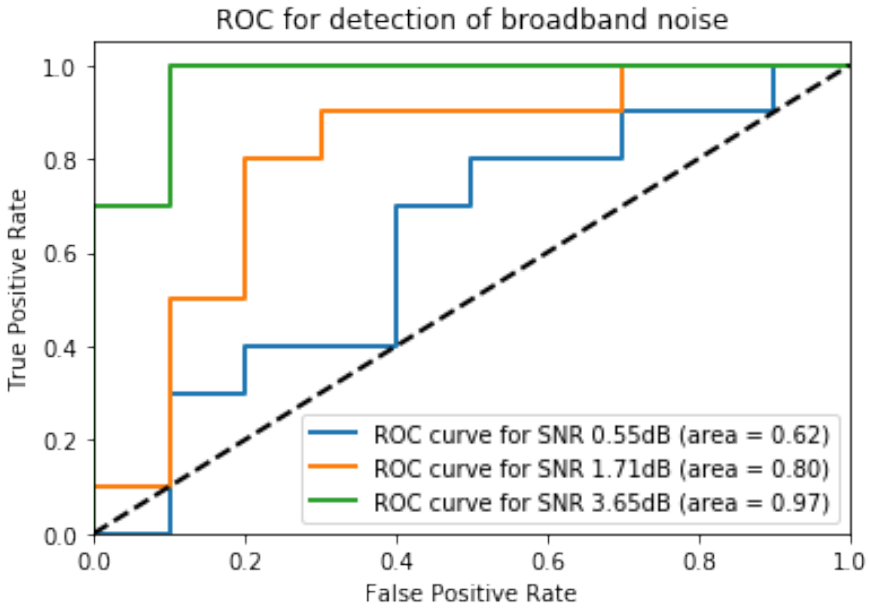


Figure 4.4: Estimation ROC pour la détection d'un bruit de crête

jeu de données synthétique démontre la faisabilité de la détection d'anomalies avec des métriques en transport optimal. Ces résultats montrent qu'il est possible de détecter un bruit lorsque l'algorithme est calibré avec un son réel. L'algorithme OT montre une sensibilité limitée aux anomalies concentrées sur un pic de fréquence étroit, chose qui a été améliorée en introduisant l'algorithme *multiband-*OT, mais fonctionne bien avec un changement de large bande, même si les modifications sont subtiles.

Évaluation de l'algorithme de classiification OT sur des ensembles de données réels

Afin de démontrer l'impact des métriques de Wasserstein sur la méthode de détection, nous avons comparé l'algorithme à une base euclidienne. Dans cette base, le même algorithme OT a été utilisé mais en remplaçant les distances de Sinkhorn par des distances euclidiennes. Par la même occasion, ces deux versions de l'algorithme (base Sinkhorn et base euclidienne) ont été comparées à une méthode de classification de pointe pour la détection d'anomalies semi-supervisée

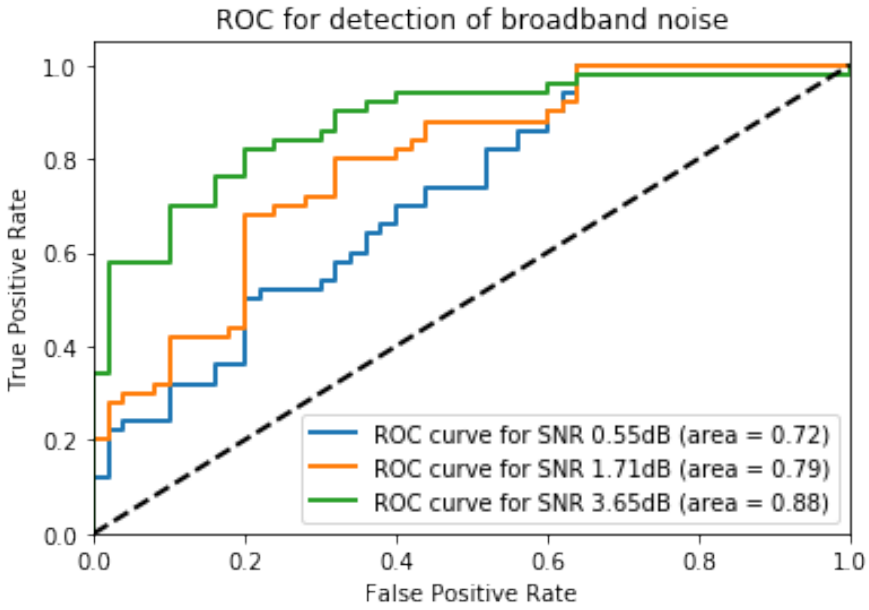


FIGURE 4.5: Estimation ROC pour la détection du bruit à large bande

qui est la *OC-SVM*. Les ensembles de données sont séparés en entraînement (500 échantillons) et en données de test (500 échantillons) à l'aide d'une division k-fold répétée, les données d'entraînement sont utilisées pour entraîner la *OC-SVM* et calculer le signal de référence $F(\overline{X})$. Les algorithmes (*OC-SVM*, *OT* et Euclidien) sont ensuite évalués sur les données de test : la moitié des données de test sont mélangées avec le son des mécanismes défectueux des ensembles de données 1 et 2.

Les résultats sont présentés sur la Fig 4.6, pour l'ensemble de données 1 avec des anomalies de sifflement léger et aigu. Comme les résultats sont qualitativement similaires pour l'ensemble de données 2, les résultats pour les ensembles de données 1 et 2 sont résumés dans le tableau 5.1. La partie gauche de la Fig 4.6 montre la précision pour 3 niveaux de bruit différents. Le SVM à une classe *OC-SVM* obtient de bons résultats mais manque plusieurs anomalies abaissant ainsi son score autour de 70-75%. La méthode euclidienne montre des performances inférieures à celles de la *OC-SVM* pour un faible SNR, mais surpasse la *OC-SVM* pour un SNR élevé. L'algorithme *OT* donne les résultats les plus élevés, avec une précision de

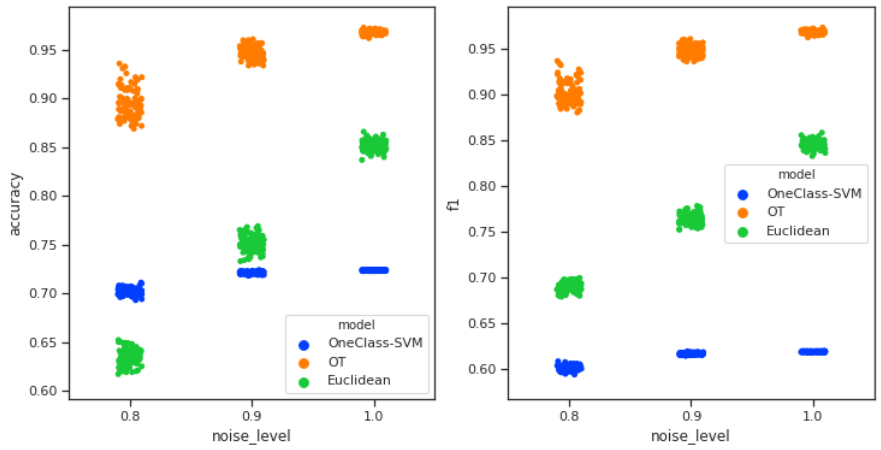


Figure 4.6: Estimation des métriques Accuracy et F1 sur le premier ensemble de données pour différents niveaux de bruit. *OC-SVM*, *OT* et base euclidienne sont évalués sur cet ensemble de données

90 à 98%. Nous avons également évalué le score F1 pour prendre en compte la précision et le rappel de l'anomalie. Ces scores sont affichés sur la partie droite où l'algorithme *OT* surpasse toutes les méthodes.

		Accuracy			F1		
	Niveau de bruit	OT	OC-SVM	Euclidien	OT	OCSVM	Euclidien
Dataset 1	0,8	**0,89**	0,70	0,63	**0,90**	0,60	0,69
	0,9	**0,95**	0,72	0,75	**0,95**	0,62	0,76
	1,0	**0,97**	0,72	0,85	**0,97**	0,62	0,84
Dataset 2	1,0	**0,54**	0,5	0,5	**0,90**	0,60	0,62
	1,5	**0,64**	0,5	0,5	**0,94**	0,61	0,62
	2,0	**0,84**	0,5	0,5	**0,97**	0,62	0,62

Table 4.1: Les métriques Accuracy et F1 pour deux ensembles de données d'enregistrement acoustique, corrompus par des sons mécaniques défectueux.

Amélioration de la méthode de détection avec l'algorithme *multiband-OT* et comparaison avec les méthodes de l'état de l'art

La version non-paramétrique de la méthode de classification a fait ses preuves. Ceci en la comparant à la version paramétrique OT montrant ainsi l'amélioration des performances de détection lorsque les scores d'anomalies sont calculés indépendement pour plusieurs bandes de fréquences étroites du PSD du signal et aux algorithmes de l'état de l'art, notamment le SVM à une classe *OC-SVM*, le facteur local aberrant LOF et la forêt d'isolement IF.

La Fig 4.7 montre que tous les modèles évalués par la métrique F1 obtiennent des résultats corrects. Le *OC-SVM* a les résultats les plus bas, obtenant un score d'environ 0,6 qui augmente avec le niveau de bruit. Le LOF obtient un score stable de 0,67 pour tous les niveaux de bruit, tandis que l'IF est d'environ 0,75. La robustesse de la forêt d'isolement est surpassée par les méthodes proposées, l'algorithme OT atteignant des scores entre 0,77 et 0,88. La méthode multibande-OT présentée dans cet article atteint le score le plus élevé, autour de 0,93.

Les mêmes modèles évalués par les métriques Accuracy (voir la Fig 4.8) et AUC-ROC (voir la Fig 4.9) obtiennent des résultats similaires aux résultats de la métrique F1, démontrant ainsi la robustesse et la supériorité de la méthode de classification en transport optimal pour la détection d'anomalies dans les sé-

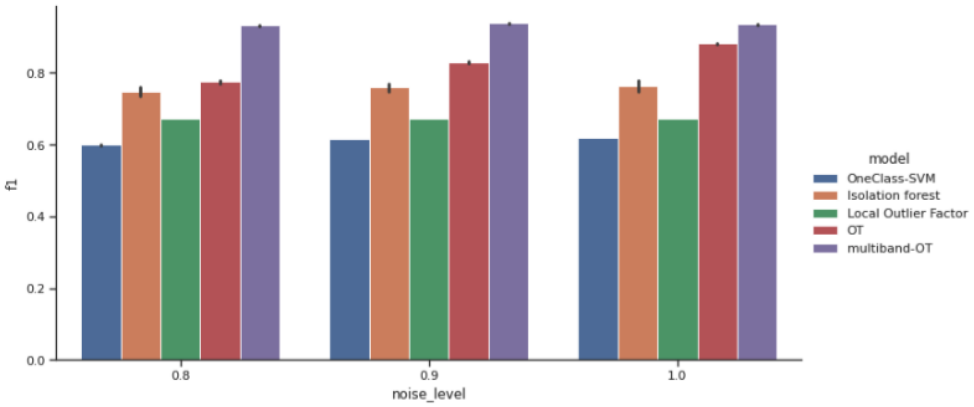


Figure 4.7: Comparaison des algorithmes de détection par la métrique F1 sur l'ensemble de données 1 pour différents niveaux de bruit.

ries temporelles. Le tableau 4.2 résume l'ensemble des scores pour les différents algorithmes évalués sur l'ensemble de données 1.

L'ensemble de données 2 a servi pour le même type d'expérience, pour lequel les courbes ROC de l'algorithme multiband-OT ont été produites pour différents niveaux de bruit. La Fig 4.10 montre une valeur AUC stable autour de 0,87 quelque soit le niveau de bruit.

La même procédure d'évaluation des performances a été effectuée pour l'application des algorithmes sur le deuxième ensemble de données. La Fig 4.10 montre les courbes ROC de l'algorithme multiband-OT avec une valeur AUC stable autour de 0,87 quelque soit le niveau de bruit, prouvant ainsi la capacité de l'algorithme de classification non-paramétrique en transport optimal à détecter efficacement les anomalies dans les séries temporelles quelque soit la nature et niveau de bruit dans lequel est plongé le signal. Les métriques d'évaluation singulières ont été également calculées pour ce deuxième ensemble de données. Les métriques Accuracy et F1 donnent des résultats similaires, la Fig 4.11 montre les résultats les moins performants pour le LOF avec un score de 0,5 suivi du $OC\text{-}SVM$ présentant un score autour de 0,63. Le OT est classé troisième pour cet ensemble de données avec un score autour 0,75 moins bien que l'IF qui a su s'adapter aux deux ensembles de données avec un score stable autour de 0,8. Le $multiband\text{-}OT$ a réussi encore une fois à décrocher la première position en surpassant les quatre autres algorithmes présentant un score croissant allant de 0,93 jusqu'à 0,97 en fonction du niveau de bruit.

4.3 Conclusion

Certes, la détection d'anomalies est un problème complexe sans une solution uniformément meilleure car le choix de la méthode à utiliser dépend largement du contexte, des propriétés des variables et des données observées, et aussi de l'objectif visé. Ce chapitre a présenté de nouvelles méthodes de classification semi-supervisées en transport optimal pour la détection d'anomalies dans les séries temporelles, particulièrement les signaux acoustiques. Le choix s'est porté sur des

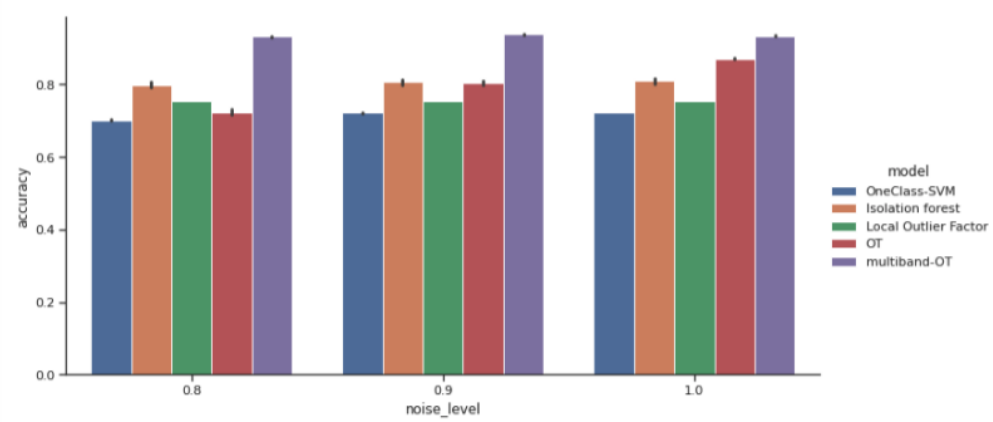


Figure 4.8: Comparaison des algorithmes de détection par la métrique Accuracy sur l'ensemble de données 1 pour différents niveaux de bruit.

signaux acoustiques pour des raisons liées aux contraintes industrielles et sensibilité des produits classifiés mais la méthode pourrait s'appliquer sur d'autres séries temporelles comme les vibrations, les EEG, etc.

Après avoir rappelé les modèles théoriques des méthodes, nous avons fourni une analyse expérimentale détaillée et les résultats correspondants. La robustesse de l'algorithme *OT* et sa version améliorée *multiband-OT* a été démontrée en observant les courbes ROC sur des ensembles de données synthéiques et réels. Les algorithmes se sont distingués des méthodes *OC-SVM*, *LOF* et *IF* avec une concurrence entre l'algorithme *OT* et l'algorithme *IF* mais le *multiband-OT* a marqué les meilleurs résultats en montrant un haut niveau de performance pour différents ensembles de données grâce à son évaluation par les deux métriques F1 et Accuracy.

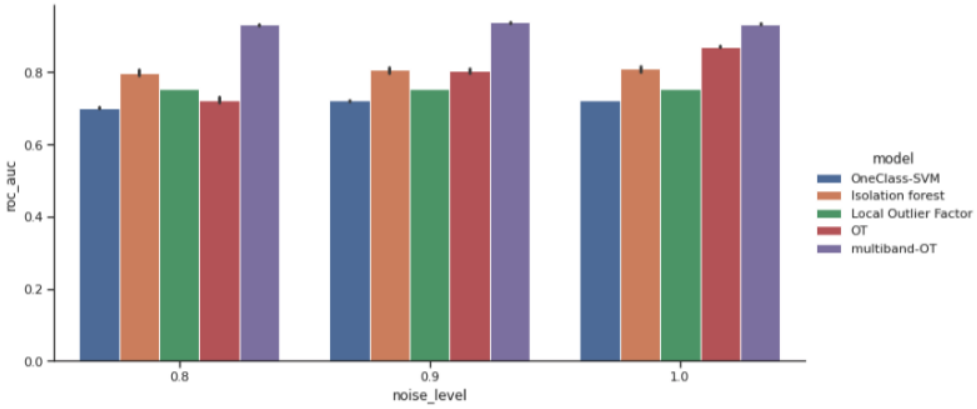


Figure 4.9: Comparaison des algorithmes de détection par la métrique AUC-ROC sur l'ensemble de données 1 pour différents niveaux de bruit.

Niveau de bruit	Modèle	F1	Accuracy	AUC-ROC
0,8	IF	0,75	0,79	0,79
	LOF	0,67	0,75	0,75
	OC-SVM	0,60	0,70	0,71
	OT	0,77	0,72	0,73
	multiband-OT	**0,93**	**0,93**	**0,93**
0,9	IF	0,75	0,80	0,80
	LOF	0,67	0,75	0,75
	OC-SVM	0,61	0,72	0,72
	OT	0,82	0,80	0,80
	multiband-OT	**0,94**	**0,94**	**0,94**
1,0	IF	0,76	0,80	0,80
	LOF	0,67	0,75	0,75
	OC-SVM	0,61	0,72	0,72
	OT	0,88	0,87	0,80
	multiband-OT	**0,93**	**0,93**	**0,93**

Table 4.2: Tableaux des métriques de performances des algorithmes de détection d'anomalies sur l'ensemble de données 1.

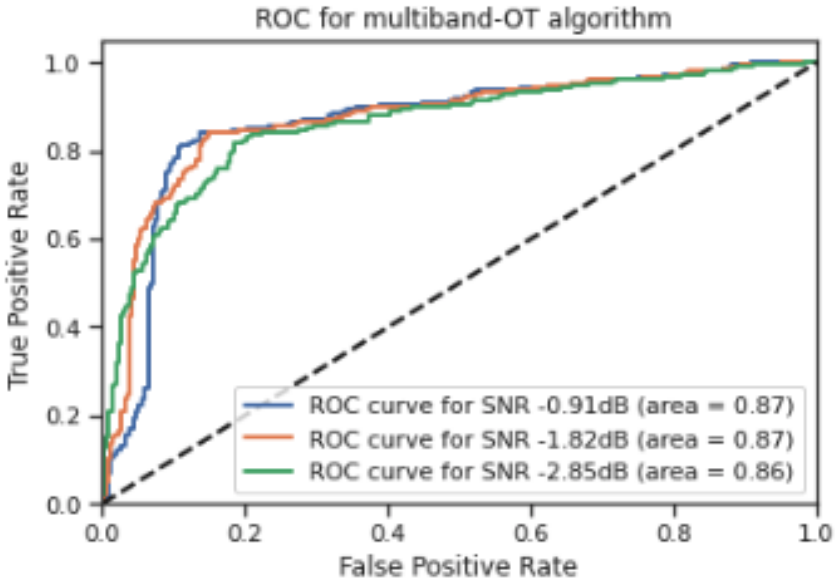


Figure 4.10: Courbe ROC de l'algorithme *multiband-OT* sur le deuxième ensemble de données pour différents niveaux de bruit .

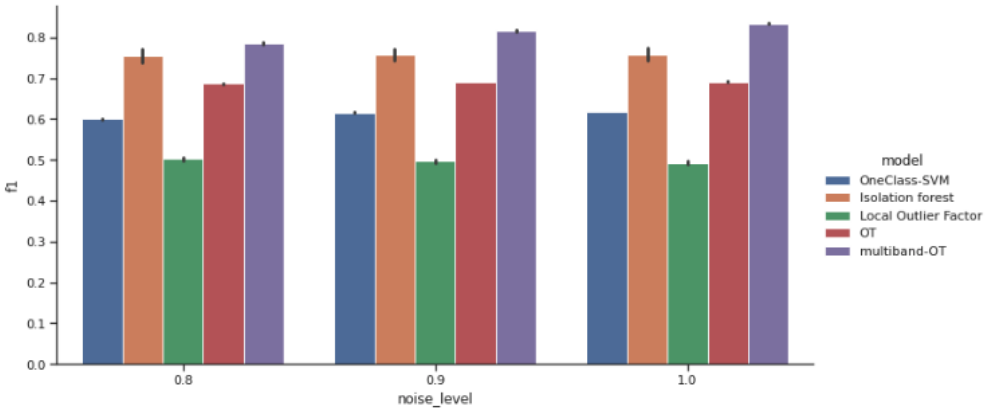


Figure 4.11: Comparaison des algorithmes de détection par la métrique F1 sur l'ensemble de données 2 pour différents niveaux de bruit.

Chapitre 5

Identification des valeurs localement aberrantes reposant sur la distance de Wasserstein

Sommaire

Après avoir abordé le modèle théorique de l'algorithme facteur aberrant local en transport optimal *LOFO*, ce chapitre traite son étude empirique. Il commence par présenter un exemple explicatif montrant les différentes étapes de l'algorithme sur des données synthétiques. Il enchaîne ensuite par une étude expérimentale détaillée et fournit les résultats de cette étude. D'abord, un aperçu des ensembles de données utlisées dans le cadre de cette étude, puis une évaluation des performances

du *LOFO* et enfin une étude méta-analyse pour la comparaison du *LOFO* avec les algorithmes de l'état de l'art.

5.1 Aperçu de la méthode facteur local aberrant en transport optimal : LOFO

Le facteur de valeur aberrante locale en transport optimal (*LOFO*) est un score qui indique la probabilité qu'un certain point de données soit une anomalie. Tout comme l'algorithme *LOFO*, un score *LOFO* inférieur à un paramètre dit offset signifie une valeur aberrante et un score *LOFO* supérieur au paramètre offset signifie que le point de données est normal. Cependant la particularité de l'algorithme *LOFO* réside dans la manière de calcul de la densité d'atteignabilité locale. En effet, celle-ci repose sur des métriques de Wasserstein tel qu'expliqué dans la section 3.3.2.

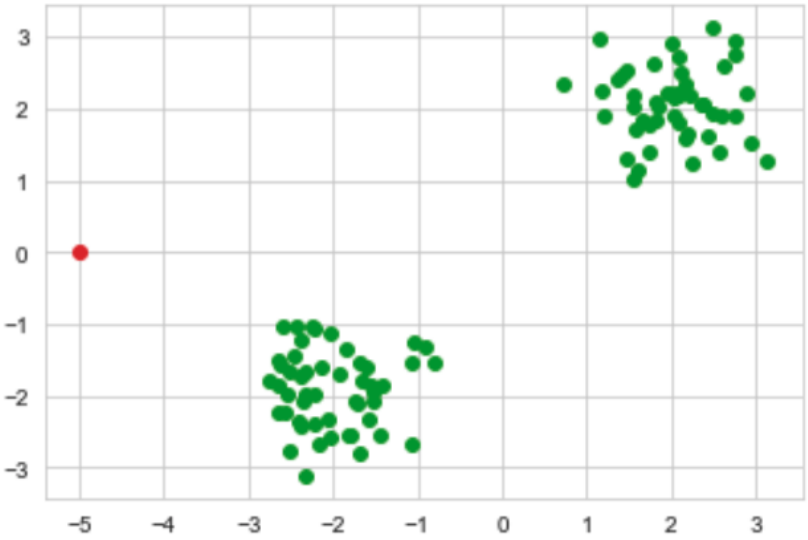


Figure 5.1: Ensemble de données synthétiques. Les nuages de points en vert représentent les données d'entraînement et l'échantillon en rouge est le point à évaluer.

Afin de rappeler les différentes étapes de l'algorithme *LOFO*, nous utilisons dans cette section un ensemble de données synthétique (voir la Fig 5.1). Avec un

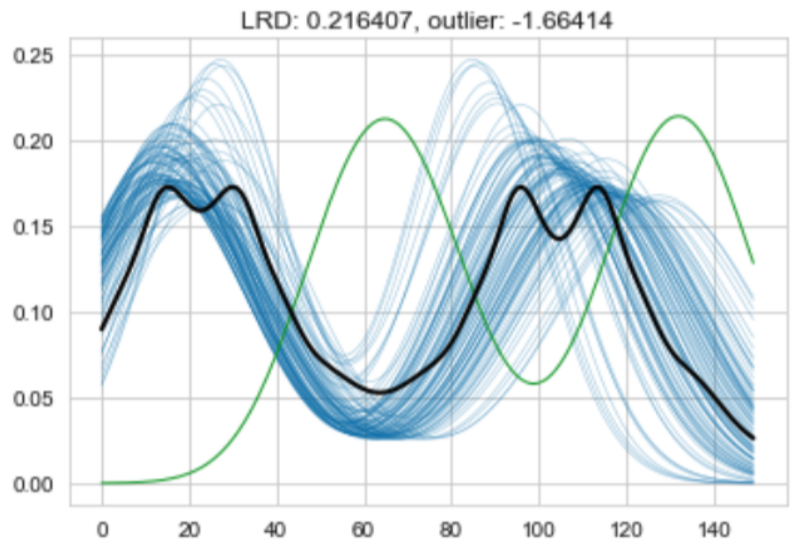


Figure 5.2: Illustration des densités. Les courbes en bleu représentent les densités des k-distances des données d'entraînement. La courbe noire représente la densité moyenne des courbes bleues et la courbe verte représente la densité de la k-distance de l'échantillon rouge de la Fig 5.1

k défini, nous pouvons introduire la k-distance. Pour chaque point de données, la densité locale au voisinage est estimée à l'aide d'un noyau gaussien à partir des valeurs de la k-distance, pour lesquelles un barycentre est calculé en transport optimal (voir la Fig 5.2). La distance d'atteignabilité est ensuite calculée par la métrique de Wasserstein régularisée par l'entropie entre la densité locale des k-distance du point de données et la densité moyenne par le barycentre de Wasserstein. À partir de cette distance d'atteignabilité, la densité d'atteignabilité locale est alors calculée. Le score *LOFO* fait apparaitre dans sa formule le rapport entre la densité d'atteignabilité locale de chaque point de données et la densité d'atteignabilité locale moyenne des points de données au voisinage. Pour les données d'entraînement comme sur la Fig 5.3, le score *LOFO* intervient dans le calcul d'un paramètre offset représenté en noir sur la Fig 5.3 à partir duquel les données de test sont prédites comme normales ou anomalies. Un exemple d'échantillon normal testé par *LOFO* est fourni par la Fig 5.4.

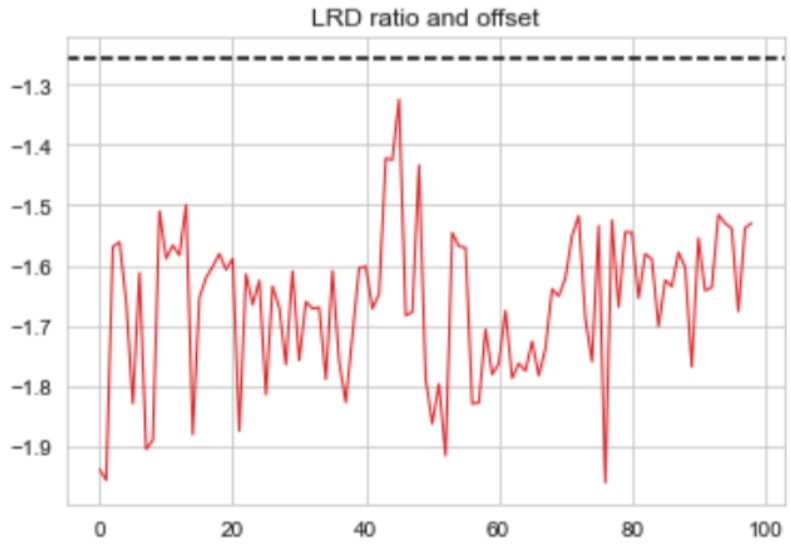


Figure 5.3: Visualisation du rapport des densités d'atteignabilité locales pour l'échantillon de test (courbe rouge).

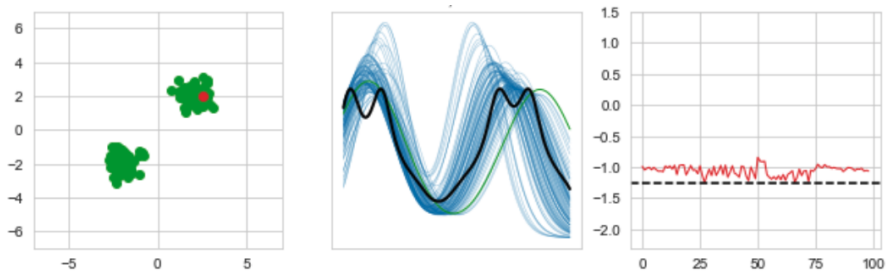


Figure 5.4: Cas d'un échantillon de test normal : La figure à gauche présente l'ensemble d'entraînement en vert et l'échantillon normal à prédire en rouge. La figure du milieu montre les courbes des densités des k-distances des données d'entraînement en bleu. La courbe noire représente la densité moyenne des courbes bleues et la courbe verte représente la densité de la k-distance de l'échantillon rouge. La figure à droite montre le rapport des densités d'atteignabilité locales pour l'échantillon de test (courbe rouge) et le paramètre offset en ligne discontinue noire.

5.2 Analyse expérimentale et résultats

Dans cette partie, nous présentons l'analyse expérimentale et les résultats correspondants de l'algorithme *LOFO* proposé. Cette étude contient une évaluation des performances par les courbes ROC et PR et une méta-analyse pour une comparaison du *LOFO* avec les algorithmes de la bibliothèque *Scikit-learn*, notamment les algorithmes *LOF*, *Robuste Covariance*, *Isolation Forest* et *OneclassSVM*.

5.2.1 Description des ensembles de données

Afin d'assurer la reproductibilité des résultats et de garantir l'objectivité dans la construction de l'étude méta-analyse, nous avons appliqué l'algorithme *LOFO* et les autres algorithmes qui font l'objet de l'étude comparative sur des ensembles de données de la bibliothèque ODDS [1] qui les récupère du référentiel de données UCI. Nous avons sélectionné les ensembles de données qui correspondent aux critères suivants :

- Données : multidimentionnelles. Un enregistrement pour chaque instance, et chaque enregistrement contient plusieurs attributs(caractéristiques)
- Tâche : classification binaire pour la détection d'anomalies.
- Instances : Au moins 100, pas de limite supérieure
- Caractéristiques : Pas plus de 300, pas de limite inférieure
- Valeurs : numériques. Les caractéristiques catégorielles sont ignorées si elles sont présentes.

Le tableau 5.1 présente les ensembles de données choisis respectant les critères cités ci-dessus.

5.2.2 Évaluation des performances de l'algorithme LOFO

Les ensembles de données de la bibliothèque ODDS sont tous déséquilibrés. La classe des valeurs aberrantes a une proportion très faible (voir tableau 5.1)

1. Référentiel ODDS

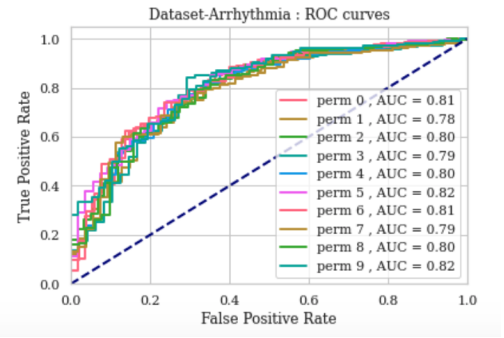
Ensemble de données	Instances	Caractéristiques	Valeurs aberrantes(%)
Wine	129	13	7,7%
WBC	278	30	5,6%
Vowels	1456	12	3,4%
Vertebral	240	6	12,5%
Musk	3062	166	3,2%
Pima	768	8	35%
Lympho	148	18	4,1%
Letter	1600	32	6,25%
Ionosphere	351	33	36%
Glass	214	9	4,2%
Cardio	1831	21	9,6 %
BreastW	683	9	35%
Arrythmia	452	274	15%

Table 5.1: Ensembles de données multidimentionnelles

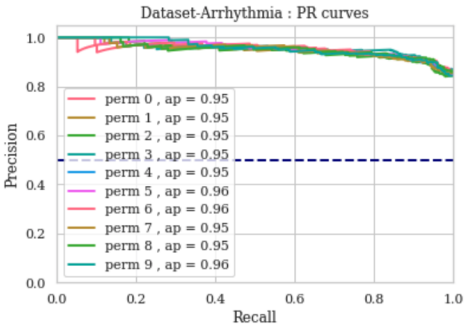
par rapport à la classe majoritaire (normale). Il est donc primordial d'évaluer correctement les performances empiriques de l'algorithme *LOFO* par des métriques d'évaluation de l'apprentissage déséquilibré. En effet, les courbes ROC et PR fournissent des méthodes puissantes pour visualiser les performances empiriques des algorithmes de détection d'anomalies. Elles présentent une forte correspondance : une courbe domine dans l'espace ROC si et seulement si elle domine dans l'espace PR. Bien qu'elles fournissent des résultats analogues, l'objectif des courbes ROC est d'être dans le coin supérieur gauche de l'espace ROC, alors que les courbes PR dominantes résident dans le coin supérieur droit de l'espace PR.

Nous fournissons ici les courbes ROC et PR de l'algorithme *LOFO* sur quelques ensembles de données pour une évaluation qualitative en fonction des nombres d'instances et d'attributs. Une évaluation sur l'ensemble des jeux de données choisis est présentée dans la section de la méta-analyse. Les ensembles de données ne contiennent pas assez de données pour entraîner les modèles, les tester et les valider. Le fait d'en retirer une partie pour la validation poserait un problème de sous-apprentissage. En réduisant les données d'apprentissage, nous risquons de perdre des tendances importantes dans l'ensemble de données, ce qui augmente

l'erreur induite par le biais. Nous utilisons donc la validation croisée *K-Fold*, qui fournit de nombreuses données pour l'entraînement du modèle et laisse également de nombreuses données pour la validation. Nous utilisons également les tests de permutation afin d'évaluer la signification des scores à validation croisée.

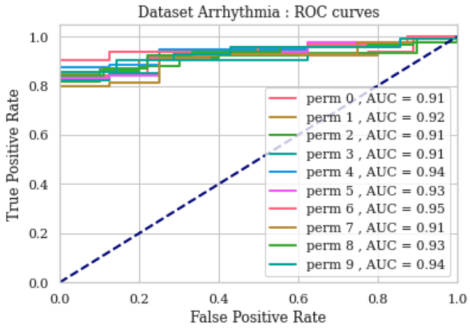


(a) Courbes ROC et valeurs AUC

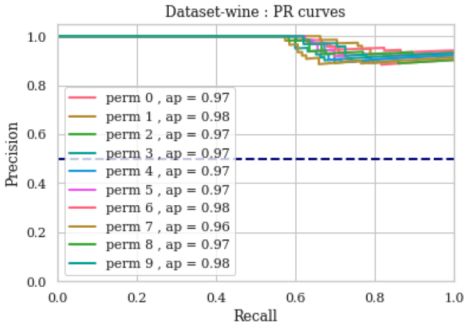


(b) Courbes PR et valeurs AP

Figure 5.5: Courbes ROC et PR du *LOFO* sur les données Arrhythmia



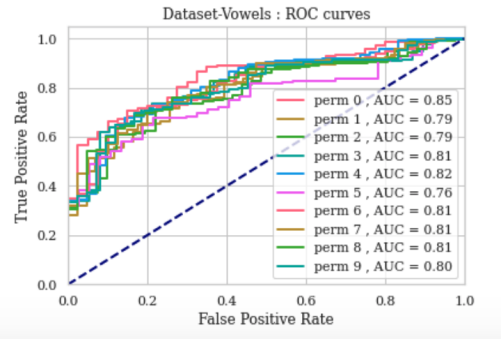
(a) Courbes ROC et valeurs AUC



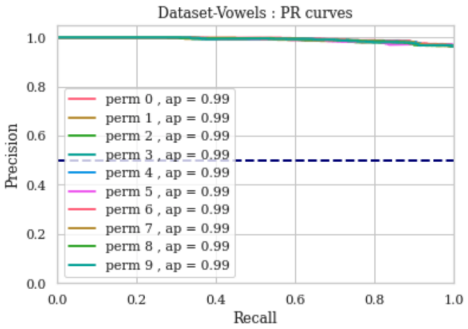
(b) Courbes PR et valeurs AP

Figure 5.6: Courbes ROC et PR du *LOFO* sur les données Wine

Après une recherche par grille des hyper-paramètres du modèle (le nombre des voisins et le coût du transport optimal), nous avons entraîné le modèle dans le cadre d'une validation croisée et un test de dix permutations. La Fig 5.5 présente les courbes ROC et PR de l'ensemble de données Arrhythmia qui contient peu d'instances (452) et un nombre de caractéristiques important (274). La courbe ROC fournit un AUC autour de 0,8 et la courbe PR une AP autour de 0.95. L'ensemble de données Wine pour lequel la Fig 5.6 présente la courbe ROC avec



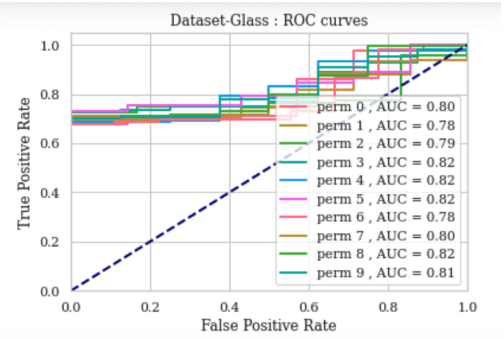
(a) Courbes ROC et valeurs AUC



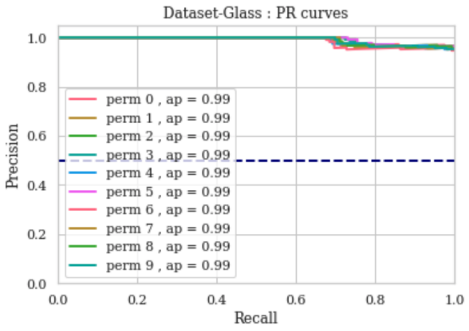
(b) Courbes PR et valeurs AP

Figure 5.7: Courbes ROC et PR du *LOFO* sur les données Vowels

un AUC autour de 0,93 et la courbe PR avec AP autour de 0,97 contient très peu d'instances (129) et peu de caractéristiques (13). La Fig 5.7 montre les courbes ROC pour les différentes permutations avec AUC autour de 0,8 et les courbes PR avec AP égale à 0,99 pour l'ensemble Vowels qui contient 1456 instances et 12 attributs. L'ensemble Glass contenant peu d'instances(214) et peu d'attributs(9) a permis aussi de très bonnes performances du *LOFO* avec une mesure AUC autour de 0,81 et une mesure AP égale à 0,99 pour toutes les permutations comme le montre la Fig 5.8.



(a) Courbes ROC et valeurs AUC



(b) Courbes PR et valeurs AP

Figure 5.8: Courbes ROC et PR du *LOFO* sur les données Glass

5.2.3 Méta-analyse pour la comparaison des algorithmes de détection d'anomalies

Tout d'abord, nous calculons les scores ROC-AUC (AUC) et Average Precision (AP) pour chaque méthode sur chaque référence. Le tableau 5.2 présente les scores AUC qui sont très proches les uns des autres pour chaque référence. En général, aucune méthode ne devrait donner les meilleurs résultats dans tous les scénarios possibles, mais sur les treize jeux de données de référence, *LOFO* fournit les meilleures performances sur six jeux de données. De même pour les scores AP, *LOFO* fournit les meilleures performances sur huit ensembles de données comme le montre le tableau 5.3.

AUC	IForest	LOF	OneClassSVM	LOFO	RobustCovariance
Wine	0,8129	0,8397	0,8139	**0,8510**	0,8226
WBC	0,9213	0,9176	0,9089	**0,9277**	0,9228
Vowels	0,7490	0,7717	0,7422	0,7429	0,7502
Vertebral	0,4085	0,3857	0.3943	**0,5933**	0,4034
Musk	**0,8576**	0,8538	0,8531	0,8553	0,8524
Pima	0,6515	**0,6564**	0,6533	0,6548	0,6478
Lympho	0,9288	0,9467	0,9411	**0,9516**	0,9390
Letter	0,6936	0,6822	**0,6956**	0.6912	0,6903
Ionosphere	0,8695	0,8726	0.8760	0.8757	**0,8847**
Glass	0,6870	0,6802	0,6913	**0,7169**	0,6991
Cardio	0,8406	**0,8524**	0,8384	0,8467	0,8389
BreastW	0,8415	0,8342	0.8349	0,8431	**0,8438**
Arrythmia	0,7912	0,7900	0,7811	**0,7930**	0,7849

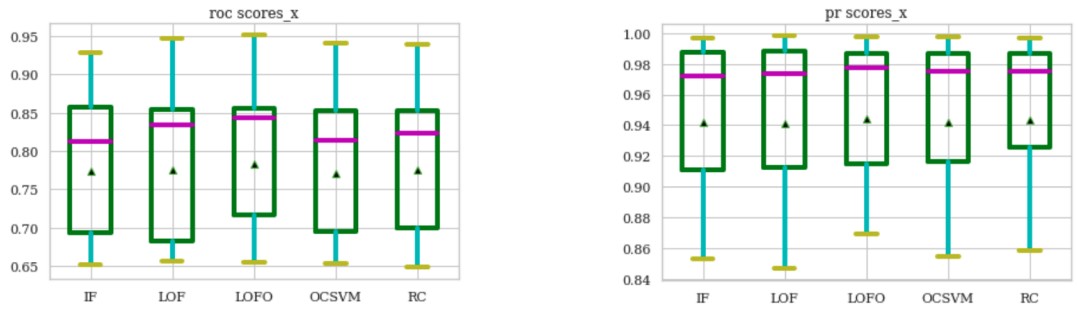
Table 5.2: Performances AUC pour l'ensemble des jeux de données

Nous avons ensuite comparé les distributions globales des scores AUC et AP des cinq méthodes sous forme de boxplots (voir la Fig 5.9). Il semble que toutes les méthodes aient des performances AUC comparables avec une légère supériorité du *LOFO*. Alors que les performances AP favorisent *LOFO* suivi de *RobustCovariance*.

Nous avons en outre vérifié cette affirmation via des tests de rang signé de

AP	IForest	LOF	OneClassSVM	LOFO	RobustCovariance
Wine	0,9759	0,9798	0,9797	**0,9861**	0,9798
WBC	0,9950	0,9946	0,9939	**0,9955**	0,.9954
Vowels	0,9874	0,9868	0,9866	**0,9879**	0,9868
Vertebral	0,8530	0,8464	0.8546	**0,8689**	0,8586
Musk	0,9893	0,9895	0,9894	**0,9906**	0,9900
Pima	0,7751	0,7767	**0,7771**	0,7726	0,7758
Lympho	0,9969	0,9975	0,9973	**0,9981**	0,9968
Letter	0,9682	0,9658	0,9664	0.9660	0,9658
Ionosphere	0,9110	0,9122	0.9161	0.9147	**0,9253**
Glass	0,972	0,9733	0,9747	**0,9772**	0,9748
Cardio	0,9787	**0,9805**	0,9785	0,9797	0,9778
BreastW	0,8924	0,8822	0.8776	0,8906	**0,8949**
Arrythmia	0,9467	0,9480	0,9471	**0,9490**	0,9446

Table 5.3: Performances AP pour l'ensemble des jeux de données



(a) AUC

(b) AP

Figure 5.9: Boxplots pour les scores AUC et AP sur 13 jeux de données.

Wilcoxon par paires entre les méthodes, et les différences moyennes standardisées signées. La Fig 5.10 présente une matrice de signification des mesures AP qui contient les p-value et les différences moyennes standardisées pour une comparaison par paires entre les algorithmes, en fournissant des cases vertes pour les algorithmes significativement meilleurs. Cette matrice montre des différences statistiquement significatives au niveau 0,05 en faveur de *Isolation Forest* et *LOFO*. En effet, ces deux derniers ont de meilleures performances AP que les trois autres algorithmes, *OneClass SVM*, *Robust Covariance* et *Local Outlier Factor*.

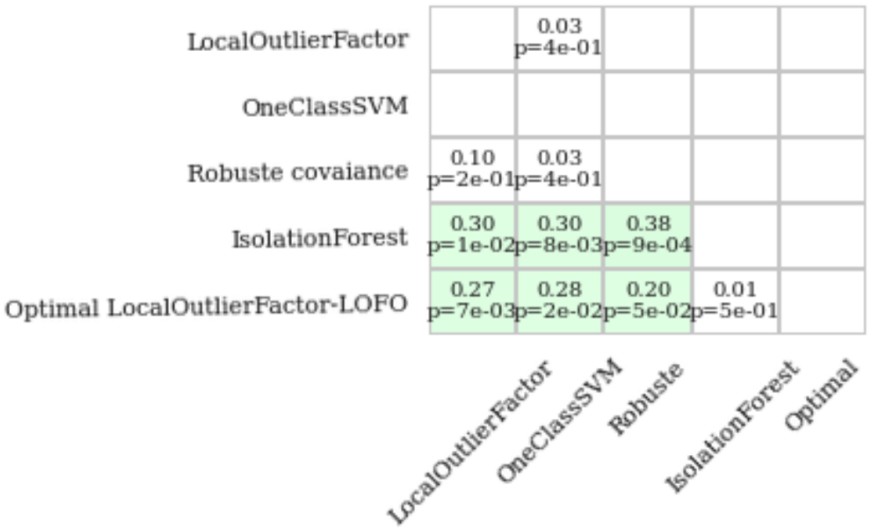


Figure 5.10: Matrice de signification statistique des performances AP

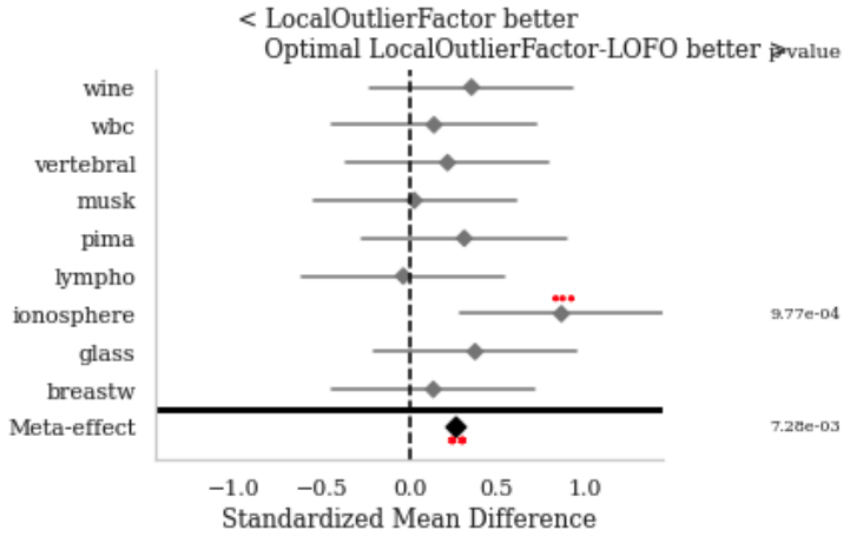


Figure 5.11: Méta-analyse pour les performances AP entre les algorithmes *LOF* et *LOFO*.

Toujours dans le cadre des comparaisons des performances algorithmiques entre les méthodes, la Fig 5.11 présente un graphique de style méta-analyse qui montre l'effet standardisé avec des intervalles de confiance sur tous les ensembles de données pour l'algorithme *Local Outlier Factor* et sa version optimale *LOFO* en iden-

tifiant les aberrations locales dans l'espace de Wasserstein. La figure montre bien un méta-effet en faveur du *LOFO* avec une différence moyenne standardisé de 0,27 et une p-value de 0,007.

5.3 Conclusion

Trouver des valeurs aberrantes est une tâche importante pour de nombreuses applications. La plupart des propositions existantes considèrent une valeur aberrante une propriété binaire. Dans cette contribution, nous montrons empiriquement que pour de nombreuses situations, il est significatif de considérer une valeur aberrante comme le degré auquel l'objet est isolé de son voisinage. Nous avons introduit la méthode *LOFO* qui capture exactement ce degré relatif d'isolement dans l'espace de Wasserstein. Nous avons expliqué les différentes étapes algorithmiques par l'application de la méthode sur des données synthétiques. Nous avons ensuite procédé à l'analyse expérimentale de la méthode sur des ensembles de données multidimensionnels réels. Une première évaluation a montré d'excellentes performances AUC et ROC quelque soit le nombre d'instances composant l'ensemble de données et le nombre de caractéristiques décrivant les instances de données. Un deuxième lot d'expérience a concerné quatre autres algorithmes de détection d'anomalies non supervisée de l'état de l'art. Le facteur local aberrant, la forêt d'isolement, la covariance robuste et le SVM à une classe. L'ensemble de ces algorithmes a été comparé avec l'algorithme *LOFO* dans le cadre d'une étude méta-analyse. Les résultats ont montré une différence statistiquement significative en faveur du *LOFO* en le comparant à *Robust Covariance*, *LOF* et *One Class SVM*. Cependant, nous n'avons pas trouvé de différence statistiquement significative entre *LOFO* et *Isolation Forest*.

Dans l'ensemble, nos expériences ont montré que les méthodes *LOFO* et *Isolation Forest* sont les deux méthodes les plus performantes avec d'excellentes performances globales sur les ensembles de données synthétiques et réelles de faible et moyenne dimensions. Cependant, la déficience de *Isolation Forest* dans les dimensions élevées sera attendue, car les arbres *Isolation Forest* sont générés par par-

titionnement aléatoire le long d'une caractéristique sélectionnée au hasard. Dans des dimensions élevées, il y a une forte probabilité qu'un grand nombre de caractéristiques soient négligées dans ce processus. En revanche, *LOFO* pourrait être plus robuste dans les dimensions élevées, car l'identification des valeurs aberrantes locales qui reposent sur des distances horizontales au sens transport optimal est toujours réalisable mais au détriment du temps de calcul. Nous prévoyons dailleurs, prouver ce propos dans le cadre d'une perspective à court terme de ce manuscrit.

Chapitre 6

Conclusion et perspectives

Sommaire

6.1 Conclusion

Les avancées technologiques émergentes de l'Internet des objets (IoT) ont conduit à une interférence significative des stratégies manufacturières. À cette fin, des concepts tels que "Industrie 4.0", "fabrication intelligente" et "usine numérique" ont vu le jour. Dans ces contextes, la maintenance prédictive joue de plus en plus un rôle crucial dans la réduction des coûts et l'amélioration des performances commerciales car elle utilise des sources de données hétérogènes pour détecter les comportements anormaux des équipements (diagnostic), prédire les modes de défaillance futurs (pronostic) et soutenir les décisions en amont (prise de décision proactive).

Ce manuscrit a donné une vue d'ensemble des architectures de maintenance prédictive composées principalement de trois grands piliers : acquisition des données, traitement des données et prise de décision qui repose sur des processus de

diagnostic et de pronostic. Le chapitre 1 a abordé les supports méthodologique et technologique permettant d'adopter une stratégie de maintenance prédictive et de mettre en place des solutions techniques afin de réussir le déploiement de la stratégie et atteindre des niveaux d'efficacité opérationnelle considérables.

L'efficacité opérationnelle d'une stratégie de maintenance prédictive est conditionnée par plusieurs critères cruciaux, entre autres la partie prise de décision. Cette partie qui se compose de deux processus déterminants (diagnostic et pronostic) constitue un volet de recherche technologique et opérationnelle très vaste dont ce manuscrit fait partie. En effet, la détection d'anomalies qui est l'étape fondamentale du processus de diagnostic et l'étape transitoire entre le processus de perception et le reste de l'architecture de maintenance prédictive constitue la question de recherche ouverte de ce manuscrit à laquelle nous avons essayé d'apporter des éléments de réponse.

Afin de bien mener cette question de recherche, le chapitre 2 a présenté les concepts fondamentaux pour la prédiction de situation anormale pour une procédure de diagnostic des équipements industriels. Il a dégagé dans un premier volet une compréhension globale du principe de la détection d'anomalies par apprentissage automatique et a présenté les aspects à prendre en compte lors de la conception ou l'utilisation d'une technique de détection d'anomalies, notamment la nature des données, les mécanismes d'apprentissage automatique, les méthodes d'analyse statistique et probabiliste appropriées et enfin les techniques d'évaluation des performances des méthodes de détection d'anomalies par apprentissage automatique.

Des réflexions approfondies sur les contributions de cette thèse ont conduit à l'exploration de la théorie mathématique du transport optimal comme outil solide d'analyse et d'exploitation des distributions des données de perception pour la proposition de nouvelles méthodes de détection d'anomalies. Afin de fonder nos méthodes sur des bases solides, le deuxième volet du chapitre 2 a présenté une synthèse haut-niveau sur la théorie du transport optimal en abordant le problème de transport Monge-Kantorovich et sa formulation dans la géométrie discrète qui a mené à la définition de l'espace de Wasserstein et ses multiples métriques. Ces

métriques de traitement des mesures de probabilité sont caractérisées par un plan de transport d'un espace de probabilité à un autre selon une matrice de coût. Les approches ont été appliquées à une grande variété de tâches et se sont avérées efficaces pour l'étude de similarité des distributions de données qui est un problème courant dans les applications d'apprentissage automatique, mais nécessitant souvent une grande quantité de ressources de calcul et donc limitées par des défis informatiques. La régularisation entropique a permis l'atténuation des problèmes de coût de calcul et a rendu les problèmes de transport optimal réalisables sur le plan informatique et donc applicables. Nous avons en outre présenté dans ce chapitre une liste non exhaustive des applications en apprentissage automatique pour des modèles supervisés, semi-supervisés et non-supervisés qui reposent sur la théorie du transport optimal.

Théoriquement, la détection d'anomalies est l'ensemble des processus et méthodes mis en place pour reconnaître automatiquement les modèles anormaux. Elle concerne un grand nombre d'applications de surveillance de l'état de santé des machines. Cet état de santé est décrit par des données suivant un modèle temporel régulier ou présentant des perturbations qui mènent à la dérivation du modèle temporel régulier. La détection d'anomalies dans ces cas repose alors sur l'analyse de ces modèles temporels appelés aussi séries temporelles. Le chapitre 3 dans sa première section a détaillé le contexte de la détection d'anomalies dans les séries temporelles, et a présenté les méthodes de détection utilisées dans ce contexte. Dans sa deuxième section, le chapitre 3 a introduit la première contribution de ce manuscrit, une nouvelle méthode semi-supervisée de détection d'anomalies dans les séries temporelles qui repose sur des métriques en transport optimal pour l'étude de similarité des séries temporelles. Elle se distingue par rapport aux méthodes existantes par l'utilisation d'une distance spectrale horizontale significative au sens transport optimal et sensible aux variations de puissance des séries temporelles. A l'issue de cette approche, nous avons proposé deux algorithmes de classification des séries temporelles. Les premiers développements ont mené à un algorithme paramétrique qui étudie la similarité des séries temporelles en calculant la distance de Wasserstein entre leurs densités spectrales de puissance. Cette

version paramétrique a présenté des limites, notamment lorsque les anomalies à détecter sont restreintes à des bandes de fréquences spécifiques. D'ailleurs ce premier algorithme a fait l'objet d'une publication avec comité de relecture dans un congrès international (Springer ICANN 2019). Une version non-paramétrique de la méthode de détection d'anomalies a été proposée pour remédier aux limites de la version paramétrique en s'appuyant sur des statistiques non-paramétriques calculées sur une décomposition de signal en banque de filtres, et la densité spectrale de puissance de la série temporelle est analysée indépendamment pour chaque bande de fréquence. Ce deuxième algorithme a donné lieu à une deuxième publication avec comité de relecture dans un congrès international (IEEE ICASSP 2020).

L'énorme potentiel de la théorie de transport optimal et les résultats positifs de la première contribution nous ont poussés à explorer d'autres méthodes de détection d'anomalies et d'en apporter des améliorations en impliquant les dernières avancées de cette théorie. En effet, nous nous sommes intéressés aux méthodes du plus proche voisin pour la détection d'anomalies car l'efficacité de celles-ci dépendent principalement de la métrique de distance utilisée pour la mesure de similarité entre les instances de données. Nous avons donc mis à contribution une nouvelle méthode LOFO permettant l'identification des valeurs aberrantes locales dans l'espace topologique de Wasserstein. Le détail de cette nouvelle méthode qui a montré se preuves est fourni dans la troisième section du chapitre 3.

Cette thèse s'est inscrite dans le cadre d'un contrat CIFRE (Conventions Industrielles de Formation par la Recherche). Afin de répondre aux exigences industrielles, nous avons appliqué la première contribution (méthodes de classification semi-supervisées pour la détection d'anomalies dans les séries temporelles)sur des données acoustiques. Pour des raisons liées aux contraintes du cadre applicatif et de la confidentialité des données industrielles, nous avons publié les expérimentations et les résultats correspondants sur des données similaires aux données industrielles privées. Le chapitre 4 a présenté l'analyse expérimentale et les résultats de ces méthodes en démontrant leur efficacité et en comparant leurs performances aux performances des méthodes de l'état de l'art.

Après avoir défini le cadre théorique de la deuxième contribution de ce ma-

nuscrit LOFO (facteur local aberrant en transport optimal) dans le chapitre 3. Le chapitre 5 a présenté le cadre applicatif de l'algorithme LOFO, d'abord l'application de la méthode sur des données synthétiques pour expliquer les différentes étapes de l'algorithme et puis une analyse expérimentale complète de l'application de l'algorithme sur des données multidimensionnelles de petite et moyenne dimensions. Cette analyse a contenu l'évaluation des performances AUC et PR de l'algorithme LOFO qui a montré de très bons résultats quelque soit le nombre d'instances de l'ensemble de données et quelque soit le nombre de dimensions. Une étude méta-analyse a complété le chapitre pour comparer LOFO aux algorithmes de l'état de l'art prouvant une différence statistiquement significative en faveur du LOFO.

6.2 Perspectives

Les travaux expérimentaux de ce manuscrit sur la méthode de classification semi-supervisée en transport optimal pour la détection d'anomalies dans les séries temporelles ont fourni des résultats très positifs sur des données acoustiques. Cette même approche sera adaptée pour être appliquée sur des données vibratoires comme première perspective à court terme de ce manuscrit.

La construction d'un système de détection d'anomalies en temps réel pour les séries temporelles constitue le second axe de poursuite de ce travail. Pour ce, nous devons répondre à certaines exigences liées aux mécanismes de perception comme la minimisation des faux positifs car ils entraînent une diminution des performances et des incidents potentiellement manqués à l'avenir. La robustesse aux anomalies est aussi une exigence importante car lorsqu'une anomalie se produit, le système ne doit pas incorporer ces points de données dans l'estimation du comportement normal. Nous devons donc utiliser des statistiques très robustes et de longues fenêtres de données historiques. Sans oublier de tenir compte de l'actionnabilité car certaines anomalies sont bien plus importantes et exploitables que d'autres. Nous devons définir donc des règles qui peuvent aider à filtrer les anomalies exploitables du bruit. Une autre question capitale et inévitable c'est comment mettre à jour

notre modèle en temps réel ? Si nous voulons trouver des anomalies en temps réel, s'entraîner une seule fois n'est sans doute pas suffisant, nous devons constamment mettre à jour notre modèle pour nous adapter au dernier comportement de l'équipement à surveiller. Ainsi, une approche de mise à jour des paramètres du modèle au fil du temps est donc nécessaire. La littérature fait référence à quatre approches de mises à jour automatique : Mises à jour par force brute qui consiste simplement à recalculer les paramètres du modèle sur la fenêtre de données la plus récente à chaque fois qu'un nouveau point de données arrive. Cependant, cela peut être irréalisable si l'ajustement du modèle à la fenêtre est trop complexe en termes de calcul. Mises à jour programmées où il est possible de mettre en cache les paramètres de modèle pendant une période donnée, et le réentraîner sur les nouveaux points de données à la fin de chaque période. Cependant, un nombre excessif des faux positifs peut se produire si le comportement du système à surveiller change avant la mise à jour programmée. Mises à jour événementielles, c'est le cas de l'apparition d'une erreur de prédiction élevée pour l'ensemble récent de points de données, il faut recalculer les paramètres du modèle. Ce genre de mises à jour est imprévisible, ce qui peut entraîner des défis opérationnels. La dernière approche est la mise à jour en ligne qui consiste à faire des lectures en continu de nouveaux points de données et mise à jour efficace des paramètres avec chaque point de données mais cette forme de mise à jour est très coûteuse.

En ce qui concerne la deuxième contribution du manuscrit sur l'identification des valeurs localement aberrantes dans l'espace de Wasserstein, les travaux expérimentaux ont été effectués sur des données multidimensionnelles de faible et moyenne dimensions et qui ont fourni de très bons résultats. Nous voulons compléter l'étude de cette contribution par l'application de la méthode LOFO sur des données de grande dimension. D'ailleurs, les métriques de Wasserstein régularisées par l'entropie ont plusieurs propriétés intéressantes car il s'agit d'une formulation non euclidienne sans échelle qui est moins sujette à la malédiction de la dimensionnalité. Nous prévoyons donc des résultats très positifs dans le cadre d'application du LOFO sur des ensembles de données de grande dimension. Après l'affirmation des performances du LOFO indépendamment de la dimensionnalité, son implé-

mentation dans la bibliothèque *Scikit-learn* est prévue comme perspective de ce travail.

Bibliographie

Aalst, Wil MP Van der et al. (2010). « Process mining : a two-step approach to balance between underfitting and overfitting ». In : *Software & Systems Modeling* 9.1, p. 87-111.

Abdul-Aziz, Ali et al. (2012). « Rotor health monitoring combining spin tests and data-driven anomaly detection methods ». In : *Structural Health Monitoring* 11.1, p. 3-12.

Abe, Naoki, Bianca Zadrozny et John Langford (2004). « An iterative method for multi-class cost-sensitive learning ». In : *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 3-11.

Abu Alfeilat, Haneen Arafat et al. (2019). « Effects of distance measure choice on k-nearest neighbor classifier performance : a review ». In : *Big data* 7.4, p. 221-248.

Adewumi, Aderemi O et Andronicus A Akinyelu (2017). « A survey of machine-learning and nature-inspired based credit card fraud detection techniques ». In : *International Journal of System Assurance Engineering and Management* 8.2, p. 937-953.

Agueh, Martial et Guillaume Carlier (2011). « Barycenters in the Wasserstein space ». In : *SIAM Journal on Mathematical Analysis* 43.2, p. 904-924.

Ahmad, Najma (2004). *The geometry of shape recognition via the Monge-Kantorovich optimal transport problem*. Brown University.

Alaverdyan, Zaruhi (2019). « Unsupervised representation learning for anomaly detection on neuroimaging. Application to epilepsy lesion detection on brain MRI ». Thèse de doct. Université de Lyon.

Almseidin, Mohammad et al. (2017). « Evaluation of machine learning algorithms for intrusion detection system ». In : *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, p. 000277-000282.

Ambrogioni, Luca et al. (2018). « Wasserstein variational inference ». In : *arXiv preprint arXiv :1805.11284*.

Arlot, Sylvain, Alain Celisse et al. (2010). « A survey of cross-validation procedures for model selection ». In : *Statistics surveys* 4, p. 40-79.

Awoyemi, John O, Adebayo O Adetunmbi et Samuel A Oluwadare (2017). « Credit card fraud detection using machine learning techniques : A comparative analysis ». In : *2017 International Conference on Computing Networking and Informatics (ICCNI)*. IEEE, p. 1-9.

Bagavathiappan, Subramaniam et al. (2013). « Infrared thermography for condition monitoring–A review ». In : *Infrared Physics & Technology* 60, p. 35-55.

Balageas, Daniel L et al. (2015). « The thermographic signal reconstruction method : a powerful tool for the enhancement of transient thermographic images ». In : *Biocybernetics and biomedical engineering* 35.1, p. 1-9.

Bay, Stephen D et Mark Schwabacher (2003). « Mining distance-based outliers in near linear time with randomization and a simple pruning rule ». In : *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 29-38.

Bergmeir, Christoph et José M Benítez (2012). « On the use of cross-validation for time series predictor evaluation ». In : *Information Sciences* 191, p. 192-213.

Bergstra, James et Yoshua Bengio (2012). « Random search for hyper-parameter optimization. » In : *Journal of machine learning research* 13.2.

Bielza, Concha et Pedro Larranaga (2014). « Discrete Bayesian network classifiers : A survey ». In : *ACM Computing Surveys (CSUR)* 47.1, p. 1-43.

Birant, Derya et Alp Kut (2007). « ST-DBSCAN : An algorithm for clustering spatial–temporal data ». In : *Data & knowledge engineering* 60.1, p. 208-221.

Bousdekis, Alexandros et al. (2019). « Decision Making in Predictive Maintenance : Literature Review and Research Agenda for Industry 4.0 ». In : *IFAC-PapersOnLine* 52.13, p. 607-612.

Breunig, Markus M et al. (2000). « LOF : identifying density-based local outliers ». In : *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, p. 93-104.

Bunkhumpornpat, Chumphol, Krung Sinapiromsaran et Chidchanok Lursinsap (2012). « DBSMOTE : density-based synthetic minority over-sampling technique ». In : *Applied Intelligence* 36.3, p. 664-684.

Camerini, Valerio et al. (2019). « Development of a vibration monitoring strategy based on cyclostationary analysis for the predictive maintenance of helicopter gearbox bearings ». In : *Surveillance, Vishno and AVE conferences*.

Castellani, Francesco et al. (2017). « Wind turbine loads induced by terrain and wakes : an experimental study through vibration analysis and computational fluid dynamics ». In : *Energies* 10.11, p. 1839.

Cerrada, Mariela et al. (2016). « Fault diagnosis in spur gears based on genetic algorithm and random forest ». In : *Mechanical Systems and Signal Processing* 70, p. 87-103.

Chaitra, PC et R Saravana Kumar (2018). « A review of multi-class classification algorithms ». In : *International Journal of Pure and Applied Mathematics* 118.14, p. 17-26.

Chandola, Varun, Arindam Banerjee et Vipin Kumar (2010). « Anomaly detection for discrete sequences : A survey ». In : *IEEE transactions on knowledge and data engineering* 24.5, p. 823-839.

Chanthery, Elodie et Pauline Ribot (2013). « An integrated framework for diagnosis and prognosis of hybrid systems ». In : *arXiv preprint arXiv :1308.5332*.

Chazal, Frédéric, David Cohen-Steiner et Quentin Mérigot (2011). « Geometric inference for probability measures ». In : *Foundations of Computational Mathematics* 11.6, p. 733-751.

Chen, Ken, Bao-Liang Lu et James T Kwok (2006). « Efficient classification of multi-label and imbalanced data using min-max modular classifiers ». In : *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, p. 1770-1775.

Cheng, Kevin C et al. (2020). « Optimal transport based change point detection and time series segment clustering ». In : *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, p. 6034-6038.

Chobola, Tomáš, Daniel Vašata et Pavel Kordík (2021). « Transfer learning based few-shot classification using optimal transport mapping from preprocessed latent space of backbone neural network ». In : *arXiv preprint arXiv :2102.05176*.

Choudhuri, Nidhan, Subhashis Ghosal et Anindya Roy (2004). « Bayesian estimation of the spectral density of a time series ». In : *Journal of the American Statistical Association* 99.468, p. 1050-1059.

Courty, Nicolas et al. (2016). « Optimal transport for data fusion in remote sensing ». In : *2016 IEEE international geoscience and remote sensing symposium (IGARSS)*. IEEE, p. 3571-3574.

Cuturi, Marco (2013). « Sinkhorn distances : Lightspeed computation of optimal transportation distances ». In : *arXiv preprint arXiv :1306.0895*.

Dai, Juying et al. (2019). « Fault diagnosis of rolling bearing based on multiscale intrinsic mode function permutation entropy and a stacked sparse denoising autoencoder ». In : *Applied Sciences* 9.13, p. 2743.

Das, Santanu, Bryan L Matthews et Robert Lawrence (2011). « Fleet level anomaly detection of aviation safety data ». In : *2011 IEEE Conference on Prognostics and Health Management*. IEEE, p. 1-10.

Davis, Jesse et Mark Goadrich (2006). « The relationship between Precision-Recall and ROC curves ». In : *Proceedings of the 23rd international conference on Machine learning*, p. 233-240.

Djenouri, Youcef et al. (2019). « A survey on urban traffic anomalies detection algorithms ». In : *IEEE Access* 7, p. 12192-12205.

Evans, Lawrence C et Wilfrid Gangbo (1999). *Differential equations methods for the Monge-Kantorovich mass transfer problem*. 653. American Mathematical Soc.

Fedele, Lorenzo (2011). *Methodologies and techniques for advanced maintenance*. Springer Science & Business Media.

Feng, P et al. (2019). « Monitoring gear surface degradation using cyclostationarity of acoustic emission ». In : *Mechanical Systems and Signal Processing* 131, p. 199-221.

Fernandes, Gilberto et al. (2019). « A comprehensive survey on network anomaly detection ». In : *Telecommunication Systems* 70.3, p. 447-489.

Flamary, Rémi (2019). « Transport optimal pour l'apprentissage statistique ». Habilitation à diriger des recherches. Université Côte d'Azur.

Flamary, Rémi et al. (2016). « Optimal spectral transportation with application to music transcription ». In : *arXiv preprint arXiv :1609.09799*.

Galagedarage Don, Mihiran et Faisal Khan (2019). « Process Fault Prognosis Using Hidden Markov Model–Bayesian Networks Hybrid Model ». In : *Industrial & Engineering Chemistry Research* 58.27, p. 12041-12053.

Galar, Diego et al. (2015). « Context awareness for maintenance decision making : A diagnosis and prognosis approach ». In : *Measurement* 67, p. 137-150.

Gangbo, Wilfrid et Robert J McCann (2000). « Shape recognition via Wasserstein distance ». In : *Quarterly of Applied Mathematics*, p. 705-737.

Gao, Jun et al. (2011). « RKOF : robust kernel-based local outlier detection ». In : *Pacific-Asia conference on knowledge discovery and data mining*. Springer, p. 270-283.

Georgiadis, Anthimos, Xiaoyun Gong et Nicolas Meier (2018). « Vibration analysis based on the spectrum kurtosis for adjustment and monitoring of ball bearing radial clearance ». In : *MATEC Web of Conferences*. T. 211. EDP Sciences, p. 06006.

Gonçalves, Mateus A et al. (2021). « A review of breast cancer aspects and its diagnosis by MRI ». In : *Authorea Preprints*.

Goodfellow, Ian et al. (2016). *Deep learning*. T. 1. 2. MIT press Cambridge.

Gorecky, Dominic et al. (2014). « Human-machine-interaction in the industry 4.0 era ». In : *2014 12th IEEE international conference on industrial informatics (INDIN)*. Ieee, p. 289-294.

Govindaraj, Vishnuvarthanan et al. (2020). « Automated unsupervised learning-based clustering approach for effective anomaly detection in brain magnetic resonance imaging (MRI) ». In : *IET Image Processing* 14.14, p. 3516-3526.

Guha, Sudipto, Rajeev Rastogi et Kyuseok Shim (2000). « ROCK : A robust clustering algorithm for categorical attributes ». In : *Information systems* 25.5, p. 345-366.

Haker, Steven et Allen Tannenbaum (2003). « On the Monge-Kantorovich problem and image warping ». In : *IMA Volumes in Mathematics and its Applications* 133, p. 65-86.

Haker, Steven et al. (2004). « Optimal mass transport for registration and warping ». In : *International Journal of computer vision* 60.3, p. 225-240.

Halim, MH Abdul et al. (2018). « An overview of data-driven and model-driven based prognostics techniques for power modules ». In : *2018 4th International Conference on Electrical, Electronics and System Engineering (ICEESE)*. IEEE, p. 34-39.

Hayes, Michael A et Miriam AM Capretz (2015). « Contextual anomaly detection framework for big sensor data ». In : *Journal of Big Data* 2.1, p. 1-22.

He, Haibo et Edwardo A Garcia (2009). « Learning from imbalanced data ». In : *IEEE Transactions on knowledge and data engineering* 21.9, p. 1263-1284.

He, Zengyou, Xiaofei Xu et Shengchun Deng (2003). « Discovering cluster-based local outliers ». In : *Pattern Recognition Letters* 24.9-10, p. 1641-1650.

Henriquez, Patricia et al. (2013). « Review of automatic fault diagnosis systems using audio and vibration signals ». In : *IEEE Transactions on Systems, Man, and Cybernetics : Systems* 44.5, p. 642-652.

Hernandez, Julio, Jesús Ariel Carrasco-Ochoa et José Francisco Martínez-Trinidad (2013). « An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets ». In : *Iberoamerican Congress on Pattern Recognition*. Springer, p. 262-269.

Higashi, Y et al. (2018). « SNR enhancement of a spin-MEMS microphone by optimum bias magnetic field and demonstration of operation sound monitoring of rotating equipment ». In : *2018 IEEE Micro Electro Mechanical Systems (MEMS)*. IEEE, p. 1060-1063.

Hu, Li-Yu et al. (2016). « The distance function effect on k-nearest neighbor classification for medical datasets ». In : *SpringerPlus* 5.1, p. 1-9.

Hubert, Mia, Michiel Debruyne et Peter J Rousseeuw (2018). « Minimum covariance determinant and extensions ». In : *Wiley Interdisciplinary Reviews : Computational Statistics* 10.3, e1421.

Huda, AS Nazmul et Soib Taib (2013a). « Application of infrared thermography for predictive/preventive maintenance of thermal defect in electrical equipment ». In : *Applied Thermal Engineering* 61.2, p. 220-227.

Huda, ASN et S Taib (2013b). « Suitable features selection for monitoring thermal condition of electrical equipment using infrared thermography ». In : *Infrared Physics & Technology* 61, p. 184-191.

Iverson, David L (2004). « Inductive System Health Monitoring. » In : *IC-AI*, p. 605-611.

Jeannerod, Marc (1983). *Le Cerveau Machine*. Sous la dir. de Fayard.

Jin, Wen et al. (2006). « Ranking outliers using symmetric neighborhood relationship ». In : *Pacific-Asia conference on knowledge discovery and data mining*. Springer, p. 577-593.

Jin, Xiaohang et Tommy WS Chow (2013). « Anomaly detection of cooling fan and fault classification of induction motor using Mahalanobis–Taguchi system ». In : *Expert Systems with Applications* 40.15, p. 5787-5795.

Jin, Xiaohang et al. (2016). « Anomaly detection and fault prognosis for bearings ». In : *IEEE Transactions on Instrumentation and Measurement* 65.9, p. 2046-2054.

Jin, Xiaohang et al. (2019). « A data-driven approach for bearing fault prognostics ». In : *IEEE Transactions on Industry Applications* 55.4, p. 3394-3401.

Jo, Taeho et Nathalie Japkowicz (2004). « Class imbalances versus small disjuncts ». In : *ACM Sigkdd Explorations Newsletter* 6.1, p. 40-49.

Junsomboon, Nutthaporn et Tanasanee Phienthrakul (2017). « Combining over-sampling and under-sampling techniques for imbalance dataset ». In : *Proceedings of the 9th International Conference on Machine Learning and Computing*, p. 243-247.

Kale, Satyen, Ravi Kumar et Sergei Vassilvitskii (2011). « Cross-validation and mean-square stability ». In : *In Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011*. Citeseer.

Khaled, Ouni, Dhouibi Hedi et Nabli Lotfi (2011). « A New Fault Diagnosis Method Using Fault Directions in Partial Least Square. » In : *International Journal of Computer Science Engineering & Technology* 1.4.

Kishawy, HA et al. (2018). « Application of acoustic emissions in machining processes : analysis and critical review ». In : *The International Journal of Advanced Manufacturing Technology* 98.5, p. 1391-1407.

Koehrsen, Will (2018). « Overfitting vs. Underfitting : A Complete Example ». In : *Towards Data Science*.

Kotsiantis, Sotiris B (2013). « Decision trees : a recent overview ». In : *Artificial Intelligence Review* 39.4, p. 261-283.

Krawczyk, Bartosz (2016). « Learning from imbalanced data : open challenges and future directions ». In : *Progress in Artificial Intelligence* 5.4, p. 221-232.

Krishnakumar, Nathina et Tamer Abdou (2020). « Detection and Diagnosis of Breast Cancer Using a Bayesian Approach ». In : *Canadian Conference on Artificial Intelligence*. Springer, p. 335-341.

Kroll, Björn et al. (2014). « System modeling based on machine learning for anomaly detection and predictive maintenance in industrial plants ». In : *Procee-*

dings of the 2014 IEEE emerging technology and factory automation (ETFA). IEEE, p. 1-7.

Krstajic, Damjan et al. (2014). « Cross-validation pitfalls when selecting and assessing regression and classification models ». In : *Journal of cheminformatics* 6.1, p. 1-15.

Kukita, Yoji et al. (2013). « Quantitative identification of mutant alleles derived from lung cancer in plasma cell-free DNA via anomaly detection using deep sequencing data ». In : *PloS one* 8.11, e81468.

Kurniabudi, Kurniabudi et al. (2019). « Network anomaly detection research : a survey ». In : *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* 7.1, p. 37-50.

Kyritsis, Konstantinos et al. (2017). « Automated analysis of in meal eating behavior using a commercial wristband IMU sensor ». In : *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, p. 2843-2846.

Laclau, Charlotte et al. (2017). « Co-clustering through optimal transport ». In : *International Conference on Machine Learning*. PMLR, p. 1955-1964.

Latecki, Longin Jan, Aleksandar Lazarevic et Dragoljub Pokrajac (2007). « Outlier detection with kernel density functions ». In : *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, p. 61-75.

Lei, Yaguo et al. (2018). « Machinery health prognostics : A systematic review from data acquisition to RUL prediction ». In : *Mechanical systems and signal processing* 104, p. 799-834.

Lemaître, Guillaume, Fernando Nogueira et Christos K Aridas (2017). « Imbalanced-learn : A python toolbox to tackle the curse of imbalanced datasets in machine learning ». In : *The Journal of Machine Learning Research* 18.1, p. 559-563.

Lever, Jake, Martin Krzywinski et Naomi Altman (2017). *Points of significance : Principal component analysis*.

Liu, Fei Tony, Kai Ming Ting et Zhi-Hua Zhou (2008). « Isolation forest ». In : *2008 eighth ieee international conference on data mining*. IEEE, p. 413-422.

— (2012). « Isolation-based anomaly detection ». In : *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1, p. 1-39.

Lu, Jiantao et al. (2021). « Enhanced K-nearest neighbor for intelligent fault diagnosis of rotating machinery ». In : *Applied Sciences* 11.3, p. 919.

Lu, Yanfei, Qing Li et Steven Y Liang (2018). « Physics-based intelligent prognosis for rolling bearing with fault feature extraction ». In : *The International Journal of Advanced Manufacturing Technology* 97.1, p. 611-620.

Mahapatra, Amogh, Nisheeth Srivastava et Jaideep Srivastava (2012). « Contextual anomaly detection in text data ». In : *Algorithms* 5.4, p. 469-489.

Malla, Chandrabhanu et Isham Panigrahi (2019). « Review of condition monitoring of rolling element bearing using vibration analysis and other techniques ». In : *Journal of Vibration Engineering & Technologies* 7.4, p. 407-414.

Marhas, Manmeet Kaur, Anup Bhange et Piyush Ajankar (2012). « Anomaly detection in network traffic : A statistical approach ». In : *International Journal of IT, Engineering and Applied Sciences Research (IJIEASR)* 1.3, p. 16-20.

Marti, Gautier et al. (2016). « Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series ». In : *2016 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, p. 1-5.

Mehta, Parikshit, Andrew Werner et Laine Mears (2015). « Condition based maintenance-systems integration and intelligence using Bayesian classification and sensor fusion ». In : *Journal of Intelligent Manufacturing* 26.2, p. 331-346.

Minkina, Waldemar et Sebastian Dudzik (2009). *Infrared thermography : errors and uncertainties*. John Wiley & Sons.

Mishra, Sitesh Kumar et al. (2021). « An approach to improve high-frequency resonance technique for bearing fault diagnosis ». In : *Measurement* 178, p. 109318.

Mohd, Mohd Rizman Sultan, Sukreen Hana Herman et Zaiton Sharif (2017). « Application of K-Means clustering in hot spot detection for thermal infrared images ». In : *2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE, p. 107-110.

Monge, Gaspard (1781). *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale.

Moretti, Nicola et al. (2020). « Maintenance service optimization in smart buildings through ultrasonic sensors network ». In : *Intelligent Buildings International*, p. 1-13.

Mulak, Punam et Nitin Talhar (2015). « Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset ». In : *International Journal of Science and Research* 4.7, p. 2101-2104.

Murphy, Chris (2020). « Choosing the Most Suitable Predictive Maintenance Sensor ». In : *Analog Devices, Inc.*

Nasr, Saifeddine Ben et al. (2021). « CNN model applied on SNP protein sequences for intestinal cancer early detection ». In : *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, p. 255-263.

Ono, Kanji (2018). « Review on structural health evaluation with acoustic emission ». In : *Applied Sciences* 8.6, p. 958.

Pang, Chee Khiang et al. (2017). *Intelligent Diagnosis and Prognosis of Industrial Networked Systems : Automation and Control Engineering Series*. CRC press.

Pang, Guansong et al. (2018). « Learning representations of ultrahigh-dimensional data for random distance-based outlier detection ». In : *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 2041-2050.

Patcha, Animesh et Jung-Min Park (2007). « An overview of anomaly detection techniques : Existing solutions and latest technological trends ». In : *Computer networks* 51.12, p. 3448-3470.

Peng, Ying et Ming Dong (2011). « A prognosis method using age-dependent hidden semi-Markov model for equipment health prediction ». In : *Mechanical Systems and Signal Processing* 25.1, p. 237-252.

Perez, Deyban et al. (2017). « Intrusion detection in computer networks using hybrid machine learning techniques ». In : *2017 XLIII Latin American Computer Conference (CLEI)*. IEEE, p. 1-10.

Peyré, Gabriel, Marco Cuturi et al. (2019). « Computational optimal transport : With applications to data science ». In : *Foundations and Trends® in Machine Learning* 11.5-6, p. 355-607.

Pittino, Federico et al. (2020). « Automatic anomaly detection on in-production manufacturing machines using statistical learning methods ». In : *Sensors* 20.8, p. 2344.

Pramanik, Sourav, Debotosh Bhattacharjee et Mita Nasipuri (2016). « Texture analysis of breast thermogram for differentiation of malignant and benign breast ». In : *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, p. 8-14.

Purarjomandlangrudi, Afrooz, Amir Hossein Ghapanchi et Mohammad Esmalifalak (2014). « A data mining approach for fault diagnosis : An application of anomaly detection algorithm ». In : *Measurement* 55, p. 343-352.

Qiu, Guangqi, Yingkui Gu et Junjie Chen (2020). « Selective health indicator for bearings ensemble remaining useful life prediction with genetic algorithm and Weibull proportional hazards model ». In : *Measurement* 150, p. 107097.

Ramdas, Aaditya, Nicolás García Trillos et Marco Cuturi (2017). « On wasserstein two-sample testing and related families of nonparametric tests ». In : *Entropy* 19.2, p. 47.

Roberts, David R et al. (2017). « Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure ». In : *Ecography* 40.8, p. 913-929.

Rousseeuw, Peter J et Katrien Van Driessen (1999). « A fast algorithm for the minimum covariance determinant estimator ». In : *Technometrics* 41.3, p. 212-223.

Rubner, Yossi, Carlo Tomasi et Leonidas J Guibas (1998). « A metric for distributions with applications to image databases ». In : *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, p. 59-66.

Sandler, Roman et Michael Lindenbaum (2011). « Nonnegative matrix factorization with earth mover's distance metric for image analysis ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8, p. 1590-1602.

Santambrogio, Filippo (2010). « Introduction to optimal transport theory ». In : *arXiv preprint arXiv :1009.3856*.

Schubert, Erich, Arthur Zimek et Hans-Peter Kriegel (2014). « Generalized outlier detection with flexible kernel density estimates ». In : *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, p. 542-550.

Segla, Mawuena, Shaoping Wang et Fang Wang (2012). « Bearing fault diagnosis with an improved high frequency resonance technique ». In : *IEEE 10th International Conference on Industrial Informatics*. IEEE, p. 580-585.

Selcuk, Sule (2017). « Predictive maintenance, its implementation and latest trends ». In : *Proceedings of the Institution of Mechanical Engineers, Part B : Journal of Engineering Manufacture* 231.9, p. 1670-1679.

Sharma, Nikita, Mahendra Mishra et Manish Shrivastava (2012). « Colour image segmentation techniques and issues : an approach ». In : *International Journal of Scientific & Technology Research* 1.4, p. 9-12.

Sinar, Dogan et George K Knopf (2020). « Disposable piezoelectric vibration sensors with PDMS/ZnO transducers on printed graphene-cellulose electrodes ». In : *Sensors and Actuators A : Physical* 302, p. 111800.

Sollai, Sara et al. (2016). « Performance of a non-contact infrared thermometer in healthy newborns ». In : *BMJ open* 6.3.

Solomon, Justin et al. (2014). « Wasserstein propagation for semi-supervised learning ». In : *International Conference on Machine Learning*. PMLR, p. 306-314.

Stupnikov, Sergey A. et Leonid A. Kalinichenko (2018). « FAIR Data Based on Extensible Unifying Data Model Development ». In : *Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018), Moscow, Russia, October 9-12, 2018*. Sous la dir. de Leonid A. Kalinichenko et al. T. 2277. CEUR Workshop Proceedings. CEUR-WS.org, p. 9-13. URL : http://ceur-ws.org/Vol-2277/paper04.pdf.

Sun, Yanmin, Mohamed S Kamel et Yang Wang (2006). « Boosting for learning multiple classes with imbalanced class distribution ». In : *Sixth international conference on data mining (ICDM'06)*. IEEE, p. 592-602.

Taha, Ayman et Ali S Hadi (2019). « Anomaly detection methods for categorical data : A review ». In : *ACM Computing Surveys (CSUR)* 52.2, p. 1-35.

Tang, Jian et al. (2001). « A robust outlier detection scheme for large data sets ». In : *In 6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*. Citeseer.

Tian, Jing, Michael H Azarian et Michael Pecht (2014). « Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm ». In : *Proceedings of the European Conference of the Prognostics and Health Management Society*. Citeseer, p. 1-9.

Tobar, Felipe (2018). « Bayesian nonparametric spectral estimation ». In : *arXiv preprint arXiv :1809.02196*.

Turner, Richard E et Maneesh Sahani (2014). « Time-frequency analysis as probabilistic inference ». In : *IEEE Transactions on Signal Processing* 62.23, p. 6171-6183.

Vanhoy, Garrett et Noel Teku (2017). « Feature selection for cyclostationary-based signal classification ». In : International Foundation for Telemetering.

Varmedja, Dejan et al. (2019). « Credit card fraud detection-machine learning methods ». In : *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, p. 1-5.

Viharos, Zs J et Krisztián Balázs Kis (2015). « Survey on neuro-fuzzy systems and their applications in technical diagnostics and measurement ». In : *Measurement* 67, p. 126-136.

Villani, Cédric (2003). *Topics in optimal transportation*. 58. American Mathematical Soc.

Vinayakumar, Ravi et al. (2019). « Deep learning approach for intelligent intrusion detection system ». In : *IEEE Access* 7, p. 41525-41550.

Wang, Limin et al. (2019). « Optimizing the topology of Bayesian network classifiers by applying conditional entropy to mine causal relationships between attributes ». In : *IEEE Access* 7, p. 134271-134279.

Wang, Ling et al. (2018). « Application of relative entropy and gradient boosting decision tree to fault prognosis in electronic circuits ». In : *Symmetry* 10.10, p. 495.

Wang, Xi et Chen Wang (2019). « Time series data cleaning : A survey ». In : *IEEE Access* 8, p. 1866-1881.

Wang, Yuyang, Roni Khardon et Pavlos Protopapas (2012). « Nonparametric Bayesian estimation of periodic light curves ». In : *The Astrophysical Journal* 756.1, p. 67.

Xing, Yang et Chen Lv (2019). « Dynamic state estimation for the advanced brake system of electric vehicles by using deep recurrent neural networks ». In : *IEEE Transactions on Industrial Electronics* 67.11, p. 9536-9547.

YAGHOOTKAR, Bahareh, Soheil AZIMI et Behraad BAHREYNI (2017). « A high-performance piezoelectric vibration sensor ». In : *IEEE Sensors Journal* 17.13, p. 4005-4012.

ZHANG, Hongcai et al. (2019a). « A low noise capacitive MEMS accelerometer with anti-spring structure ». In : *Sensors and Actuators A : Physical* 296, p. 79-86.

ZHANG, Ke, Marcus HUTTER et Huidong JIN (2009). « A new local distance-based outlier detection approach for scattered real-world data ». In : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, p. 813-822.

ZHANG, Yuchen et al. (2019b). « Probabilistic anomaly detection approach for data-driven wind turbine condition monitoring ». In : *CSEE Journal of Power and Energy Systems* 5.2, p. 149-158.

ZHANG, Zheng, Ping TANG et Thomas CORPETTI (2020). « Time Adaptive Optimal Transport : A Framework of Time Series Similarity Measure ». In : *IEEE Access* 8, p. 149764-149774.

ZHAO, Chun et al. (2019). « A Resonant MEMS Accelerometer With 56ng Bias Stability and 98ng/Hz 1/2 Noise Floor ». In : *Journal of Microelectromechanical Systems* 28.3, p. 324-326.

ZHAO, Fu-Qiong et al. (2017). « Integrated equipment health prognosis considering crack initiation time uncertainty and random shock ». In : *Chinese Journal of Mechanical Engineering* 30.6, p. 1383-1395.

ZHONG, Jia-Xing et al. (2019). « Graph convolutional label noise cleaner : Train a plug-and-play action classifier for anomaly detection ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 1237-1246.

ZHOU, Zhi-Hua et Xu-Ying LIU (2010). « On multi-class cost-sensitive learning ». In : *Computational Intelligence* 26.3, p. 232-257.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Prédiction de situations anormales par apprentissage automatique pour la maintenance prédictive : approches en transport optimal pour la détection d'anomalies

Mots clés : Maintenance prédictive, Détection d'anomalies, Transport optimal

Résumé : L'émergence de l'Industrie 4.0 et des systèmes intelligents entraîne une attention croissante pour les stratégies de maintenance prédictive qui peuvent réduire le coût et les temps d'arrêt et augmenter la disponibilité des équipements indus-triels. Dans cette thèse, nous présentons une vue d'ensemble des architectures de maintenance prédictive et nous nous intéressons à un pilier capital de ces architectures, la détection d'anomalies comme première étape de prise de décision dans une architecture de maintenance prédictive. Nous apportons deux contributions à cette question de recherche. Une première méthode de classification semi-supervisée en transport optimal dans deux versions(paramétrique et non-paramétrique) pour la détection d'anomalies dans les séries temporelles. Les travaux expérimentaux de l'application de cette méthode sur des ensembles de données acoustiques synthétiques et réels prouvent la robustesse des métriques au sens transport optimal et démontrent en outre la supériorité desperformances de la méthode par rapport aux algorithmes de l'état-de-l'art. La deuxième contribution concerne une méthode non-supervisée de détection d'anomalies dans des données multidimensionnelles. Elle identifie les valeurs aberrantes locales dans un espace topologique non-euclidien en utilisant des métriques en transport optimal. Les résultats expérimentaux montrent l'efficacité de la mé-thode à remédier au problème de la malédiction de dimensionnalité et témoignent de la différence statistiquement significative de la méthode proposée par rapport aux méthodes évaluées de l'état de l'art.

Title : Prediction of abnormal situations by machine learning in a predictive maintenance context : Optimal transport theory for anomaly detection

Keywords : Predictive maintenance, Anomaly detection, Optimal Transport

Abstract : The emergence of Industry 4.0 and smart systems is leading to increasing attention to predictive maintenance strategies that can decrease the cost of downtime and increase the availability of industrial equipment. In this thesis, we present an overview of predictive maintenance architectures and we are interested in a capital pillar of these architectures, the anomaly detection as a first step of decision-making in a predictive maintenance architecture. We provide two contributions to this research question. A first method of semi-supervised classification in optimal transport in two versions (parametric and non-parametric) for the detection of anomalies in time series. The experimental results of acoustic data sets prove the robustness of the metrics derived from optimal transport and further demonstrate the superiority of the method performance over state-of-the-art algorithms. The second contribution concerns an unsupervised anomaly detection method in multidimensional data. It identifies local outliers in a non-Euclidean topological space using optimal transport metrics. The experimental results revealed the effectiveness of the method in solving the dimensionality problem and testify to the statistically significant difference of the proposed method compared the evaluated methods of the state of the art.