UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE
Ecole Doctorale Sciences, Technologies, Santé

THÈSE

Présentée à

L'UFR SCIENCES EXACTES ET NATURELLES

Pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE

Spécialité

Génie Informatique, Automatique et Traitement du Signal

Contribution à l'application de la reconnaissance des formes et la théorie des possibilités au diagnostic adaptatif et prédictif des systèmes dynamiques

Par

Mohamed Saïd BOUGUELID

Soutenue le 12 décembre 2007 devant le jury composé de :

Rapporteurs	Stéphane LECOEUCHE Professeur, Ecole des Mines de Douai
	Jean-Christophe POPIEUL Professeur, Université de Valenciennes
Examinateurs	Bernard RIERA Professeur, Université de Reims
Président du jury	Dimitri LEFEBVRE Professeur, Université du Havre
Directeur de thèse	Patrice BILLAUDEL Professeur, Université de Reims
Co-directeur de thèse	Moamar SAYED-MOUCHAWEH Maître de conférences, Université de Reims

Dédicaces

Je dédie ce modeste travail,

A ma mère avec toute mon affection,

A mon père avec toute ma reconnaissance,

A mes sœurs,

A mes frères,

A ma famille,

Et à tous mes amis.

Remerciements

Ces travaux de thèse ont été réalisés à l'Université de Reims Champagne-Ardenne au sein du Centre de Recherche en STIC (CReSTIC). Ils n'auraient jamais été concrétisés sans la collaboration de personnes que je tiens à remercier.

En premier lieu, je tiens à exprimer ma profonde gratitude à Monsieur Janan ZAYTOON, Professeur à l'Université de Reims Champagne-Ardenne et directeur du CReSTIC pour m'avoir accueilli au sein de cette structure de recherche et de m'avoir permis d'enrichir ma réflexion sur différents aspects de la recherche, de ces moyens et de ces enjeux.

Ces travaux de recherche n'auraient pas pu être réalisés sans la confiance dont ma fait preuve mon directeur de thèse Monsieur Patrice BILLAUDEL, Professeur à l'Université de Reims Champagne-Ardenne, sa disponibilité et son dévouement méritent d'être soulignés.

Je tiens aussi à exprimer mes remerciements et sentiments les plus respectueux à Monsieur Moamar SAYED MOUCHAWEH, Maître de conférences à l'Université de Reims Champagne-Ardenne et co-directeur de thèse pour avoir suivi cette thèse tout le long de sa réalisation par sa disponibilité, ses conseils et orientation, qu'il n'a pas cessé de me prodiguer.

Mes remerciements s'adressent également aux professeurs qui m'ont fait l'honneur d'accepter d'être membre du jury de cette thèse. Je voudrais citer particulièrement Monsieur Stéphane LECOEUCHE, Professeur à l'Ecole des Mines de Douai et Monsieur Jean-Christophe POPIEUL, Professeur à l'Université de Valenciennes, pour avoir accepté d'être les rapporteurs de cette thèse. Monsieur Dimitri LEFEBVRE Professeur à l'Université du Havre pour m'avoir fait l'honneur de présider le jury de soutenance, ainsi que Monsieur Bernard RIERA, Professeur à l'Université de Reims Champagne-Ardenne, d'avoir accepté d'examiner mes travaux .

Merci à l'ensemble du personnel de l'Institut de Formation Technique Supérieur de Charleville-Mézières et du laboratoire CReSTIC de Reims, pour leur gentillesse et leur soutien.

Enfin, je remercie toute ma famille, pour m'avoir soutenu dans les moments difficiles et encouragé durant les derniers mois de stress, et en particulier mon père, ma mère, mes frères et sœurs qui m'ont apporté tout au long de mes études leur soutien indéfectible.

Ce travail a été financé par le conseil général des Ardennes de Charleville-Mézières.

Tables des Matières

Chapitre 2 :
Discrimination tenant compte de la forme des classes et de la corrélation des attributs

Chapitre 3 :
Classification adaptative et évolutive pour le diagnostic des systèmes dynamiques

Conclusion générale et perspectives

Introduction générale

Ce mémoire de thèse apporte une contribution au diagnostic des systèmes industriels en utilisant la reconnaissance floue des formes.

Un système industriel est susceptible d'évoluer entre des modes de fonctionnement normaux et défaillants. Un mode de fonctionnement défaillant correspond à un dysfonctionnement du système entraînant une réalisation partielle ou complètement non conforme de la tâche pour laquelle le système a été conçu. Afin d'éviter le fonctionnement sous un mode défaillant, une maintenance périodique, permettant de remplacer des éléments du système après une durée définie, est nécessaire. Toutefois, l'arrêt périodique conduit à la diminution du temps de disponibilité des moyens de production qui cause à son tour une baisse de productivité. De plus, la durée de vie de certains éléments peut ne pas être définie au préalable. Une maintenance conditionnelle, effectuée seulement si le système présente un dysfonctionnement, est donc plus intéressante. En effet, ce type de maintenance permet d'augmenter la productivité et de diminuer le coût d'entretien. Cependant, il nécessite de déterminer à tout instant l'état de fonctionnement, normal ou défaillant, du système. Le but est de détecter l'apparition d'un défaut, d'isoler l'élément responsable et d'identifier éventuellement les causes, afin de proposer à l'opérateur humain des procédures correctives. Cette détection et cette isolation sont réalisables en utilisant un module de diagnostic.

Il existe de nombreuses méthodes de diagnostic dans la littérature. Elles se différencient par rapport au type de connaissances *a priori* disponible sur le fonctionnement du système. En général, ces méthodes peuvent être divisées en deux catégories :

- les méthodes internes ou les méthodes à base de modèle. Ces méthodes nécessitent une information approfondie suffisante pour construire un modèle quantitatif, qualitatif ou mixte du fonctionnement du système. Les équations de parité, l'estimation d'état ou paramétrique, les méthodes à base d'automate ou de réseaux de Petri sont des exemples de ces méthodes,
- les méthodes externes qui regroupent les méthodes à base de connaissances et celles à base de données historiques. Les méthodes à base de connaissances exploitent les compétences et le raisonnement des experts sur le fonctionnement du système et les transforment en règles. En revanche, les méthodes à base de données historiques cherchent à découvrir des informations sous forme de structures des classes ou de tendance des signaux au sein de mesures issues des capteurs et des actionneurs. Ces classes ou ces tendances peuvent identifier le comportement normal ou défaillant du système. Les réseaux de neurones et la reconnaissance des formes sont des exemples très connus de ces méthodes.

Quand le système est complexe ou mal connu et lorsque la connaissance *a priori* est insuffisante pour construire un modèle interne du système, les méthodes à base de Reconnaissance des Formes (RdF) sont particulièrement adaptées pour réaliser le diagnostic. Ces méthodes construisent le modèle du système non pas de façon analytique mais par apprentissage. Les modes de fonctionnement, normaux ou défaillants, sont représentés par des classes. Le diagnostic par RdF est réalisé en associant une nouvelle observation sur le fonctionnement du système, représentée par un point dans l'espace de représentation, à une de ces classes apprises.

L'application de la RdF en diagnostic se heurte souvent au problème de l'insuffisance de la connaissance *a priori* concernant le fonctionnement de ces systèmes. Dans une base de connaissance incomplète, tous les modes de fonctionnement ne sont pas représentés et notamment les modes défaillants ou dangereux qui ne peuvent pas être provoqués pour des raisons de coût ou de sécurité. Un processus de diagnostic par RdF doit donc être réalisé en utilisant une méthode de classification adaptative, c'est-à-dire capable de détecter et d'apprendre les états inconnus, représentés par l'apparition de nouvelles classes dans l'espace de représentation, afin d'enrichir la connaissance *a priori*. De plus, un système dynamique peut évoluer entre plusieurs modes de fonctionnement. Cette évolution se traduit par un ensemble de points formant un pont entre deux classes. Ces points ne présentent pas une structure cohérente et stable contribuant à la formation d'une nouvelle classe. Par contre, il est intéressant de prédire le sens de cette évolution afin d'anticiper la dérive du système d'un mode de fonctionnement normal vers un mode anormal. Cette prédiction permet d'éviter les modes anormaux et leurs conséquences. Le diagnostic pour qu'il soit qualifié de prédictif nécessite donc une méthode de classification prédictive, c'est-à-dire capable d'extraire et d'inclure à la base de connaissance l'information manquante, sur l'évolution du système, portée par chaque nouveau point.

Les méthodes de RdF sont divisées en méthodes paramétriques et non paramétriques. Nous nous sommes intéressés aux méthodes non paramétriques, plus adaptées aux cas industriels parce qu'elles nécessitent aucune connaissance sur la forme de la loi de probabilité des classes. Il existe plusieurs méthodes non paramétriques. Nous avons choisi la méthode de classification Fuzzy Pattern Matching (FPM) parce qu'elle est simple à appliquer et que son temps de classification est faible et indépendant de la taille de l'ensemble d'apprentissage. De plus, elle est capable de sélectionner les sources d'informations les plus pertinentes et de traiter des données qui sont à la fois incertaines et imprécises.

Cependant FPM est une méthode de classification naïve, c'est-à-dire qu'elle classifie un nouveau point par la sélection d'une des décisions partielles. Chaque décision partielle est calculée pour chaque classe et par rapport à chaque attribut de l'espace de représentation. FPM ne tient donc pas compte de la corrélation entre les attributs et considère la forme des classes comme convexe. Ces inconvénients rendent FPM inutilisable pour de nombreuses applications réelles et en particulier pour celles qui nécessitent une discrimination non linéaire entre les classes.

De plus, FPM n'est pas une méthode de classification adaptative ou prédictive. En fait, FPM classifie un nouveau point dans une des classes connues ou le rejette. Les points rejetés portent l'information sur l'apparition d'une nouvelle classe ou l'évolution entre deux classes. FPM ne peut donc pas extraire l'information manquante de ces points rejetés en quantifiant leur représentativité vis a vis des classes de l'ensemble d'apprentissage.

Les travaux de ce mémoire de thèse portent donc sur le développement de la méthode FPM afin qu'elle soit d'une part adaptée pour les classes non-convexes et/ou décrites par des attributs corrélés, et d'autre part une méthode de classification adaptative et prédictive sans connaissance *a priori* de la forme des classes.

Le mémoire est découpé en trois chapitres. Dans le premier chapitre, les théories de représentation et de traitement de l'information imparfaite, à savoir les théories des probabilités, des fonctions de croyance et des possibilités sont étudiées. Une comparaison entre les méthodes de classification basées sur ces trois théories est ensuite réalisée afin de montrer leurs points forts et leurs points faibles, et de justifier notre choix de la théorie des possibilités. Les principales transformations de la théorie des probabilités en possibilités existantes ainsi que la Transformation Variable (TV), proposée dans [SAY06], sont présentées. Une amélioration de la TV, afin de remédier à ses inconvénients, est proposée. La TV améliorée fournie une distribution de possibilité aussi spécifique que celle obtenue par la

transformation optimale de Dubois et Prade tout en respectant la condition de cohérence de Dubois et Prade dans le cas continu. Egalement la TV améliorée est plus simple à mettre en œuvre, surtout dans le cas d'ignorance totale, que celle de Dubois et Prade optimale. Ces transformations sont évaluées selon plusieurs critères afin de montrer l'intérêt de la TV améliorée pour les applications de la reconnaissance des formes. Enfin, plusieurs exemples académiques et réels sont utilisés pour illustrer et tester ces transformations ainsi que les différentes méthodes étudiées dans ce chapitre.

Le deuxième chapitre présente une étude comparative des méthodes de classification utilisées pour la discrimination des classes de forme non convexe et/ou décrites dans un espace de paramètres corrélés. Une amélioration de la méthode FPM est proposée pour être adaptée à cette discrimination. Une évaluation des performances de la méthode améliorée est effectuée et comparée avec les principales méthodes de classification selon deux critères : le taux d'erreur de classification et le temps de calcul. Cette évaluation est réalisée en utilisant plusieurs bases de données académiques et réelles.

Le dernier chapitre traite le problème de diagnostic adaptatif et prédictif pour des classes dynamiques. D'abord, une synthèse de l'état de l'art sur la classification dynamique (cas des données non-stationnaires) et le lien avec la classification statique (cas des données stationnaires) est présentée. Ensuite, une étude des performances de l'algorithme adaptatif et prédictif basé sur la méthode FPM, et proposé dans [SAY02a, SAY02c], a été réalisée afin d'en déduire ses limites. Enfin, une solution basée sur FPM améliorée, développée dans le deuxième chapitre, a été proposée. Cette solution se résume en deux algorithmes. Premièrement, il s'agit de l'intégration de l'apprentissage incrémental à FPM améliorée pour la rendre capable de réaliser le suivi de la déformation locale pour le cas des classes non convexes. Deuxièmement, il s'agit d'une amélioration de cet algorithme pour la détection de l'apparition de nouvelles classes de forme non convexes. Les performances de la solution proposée sont illustrées à travers plusieurs exemples académiques.

Chapitre 1

Diagnostic des défauts : Terminologie, Méthodes, Modélisation et traitement de l'information

1.1. Introduction

Les progrès récents des sciences ont entraîné un changement considérable dans divers secteurs industriels. Les industriels modernes s'équipent de plus en plus avec des systèmes automatiques complexes, afin d'améliorer la productivité et la qualité de leurs produits tout en réduisant leur coût de traitement. Les équipements modernes sont sujets aux défaillances. Ces dernières peuvent réduire considérablement la production et même dans certain cas, mettre en péril la vie des personnes et l'équilibre de l'environnement. Il est alors légitime pour ces industriels d'acquérir une technicité efficace de supervision, dotée d'un outil de diagnostic adapté afin de limiter les conséquences engendrées par les défaillances et d'améliorer la sécurité des personnels assurant ainsi une fiabilité et une disponibilité accrues de leurs outils de production.

Le diagnostic, en exploitant les données recueillies sur un système et sur son environnement, permet de déterminer le mode des défaillances dans lequel se trouve ce système et de localiser les éléments responsables en explicitant les causes qu'ils ont induit. Toutes ces informations, apportées par le diagnostic, sont très utiles pour prendre une décision, qui est soit de maintenir le système sous le même mode de fonctionnement si celui-ci est normal, soit de corriger ce mode ou bien d'arrêter le système s'il est interdit ou dangereux.

L'information manipulée peut être constituée de données statistiques, d'observations directes ou de résultats de traitement sur ces observations, ou encore des connaissances sur les classes de ces observations. Ces connaissances sont exprimées sous forme de règles, de contraintes ou d'avis d'experts. Cette information peut être représentée sous forme numérique ou sous forme symbolique.

Dans le domaine des systèmes d'aide à la décision, les informations manipulées sont souvent imparfaites. Ces imperfections se manifestent sous de multiples formes : incertitude, imprécision, incomplétude, ambiguïté, et conflit [DUB88]. L'incertitude est relative à la vérité d'une information et elle caractérise son degré de conformité à la réalité. L'imprécision concerne le contenu de l'information et mesure donc un défaut quantitatif sur une mesure. Elle concerne le manque d'exactitude en quantité, en taille, en durée, etc. L'incertitude et l'imprécision sont souvent présentes simultanément et l'une peut induire l'autre. L'incomplétude caractérise l'absence d'information apportée par la source sur certaines caractéristiques du problème. Cette incomplétude des informations issues de chaque source est la raison principale qui motive la fusion des informations de plusieurs sources. En effet, l'information fournie par chaque source est en général partielle, elle ne fournit qu'une vision du monde ou du phénomène observé, en n'en mettant en évidence que certaines

caractéristiques [BLO96]. L'ambiguïté exprime la capacité d'une information de conduire à deux interprétations différentes. Elle peut provenir des imperfections précédentes, par exemple de l'imprécision d'une mesure qui ne permet pas de différencier deux situations, ou de l'incomplétude qui induit des confusions possibles entre des objets ou des situations. Ici la fusion est également utile pour lever les ambiguïtés d'une source grâce aux informations apportées par les autres sources ou par les connaissances supplémentaires. Le conflit caractérise deux ou plusieurs informations conduisant à des interprétations contradictoires et donc incompatibles.

La robustesse d'un système conçu pour aider un opérateur ou pour prendre des décisions nécessite l'inclusion de ces imperfections dans la représentation de l'information afin de la mettre sous forme utile pour le raisonnement. Il faut donc des méthodes performantes capables de tenir compte de ces imperfections.

Ce chapitre vise à rappeler la terminologie utilisée dans la littérature et en particulier celle qui est relative à nos travaux. Il introduit en premier lieu, le diagnostic et son intérêt dans le domaine industriel, les différentes modélisations des systèmes du point de vue du diagnostic ainsi que les caractéristiques auxquelles toute procédure de diagnostic doit répondre. Une classification des méthodes de diagnostic en deux catégories est adoptée : les méthodes internes (à base de modèle) et les méthodes externes (sans modèle). Nous nous intéressons à la deuxième catégorie et en particulier aux méthodes de Reconnaissance des Formes (RdF). Ensuite, nous détaillerons les principales théories permettant de représenter les connaissances imparfaites et de raisonner à partir de celles-ci. Ces théories sont : les probabilités, les fonctions de croyance, les possibilités et les ensembles flous. Les méthodes de RdF utilisant ces théories ainsi qu'une comparaison entre elles seront présentées.

En général les informations peuvent être représentées par plusieurs théories afin de tenir compte de différents types d'imperfections. Il faut donc convertir ces informations pour qu'elles soient représentées par une seule théorie. Cette conversion est réalisée par des transformations. Une solution pour remédier à l'inconvénient de la Transformation Variable (TV) de probabilité en possibilité proposée dans [SAY06] sera proposée. Les caractéristiques de cette solution rendent la TV améliorée intéressante pour les applications de la RdF. Les transformations les plus connues de la littérature, ainsi que la transformation proposée, seront développées et comparées en utilisant les critères les plus utilisés de la littérature. Enfin le problème de la fusion de l'information provenant de plusieurs sources ainsi que les modes de combinaison les plus connus de la littérature seront abordés.

1.2. Définition et intérêt du diagnostic

Les systèmes industriels, devenus aujourd'hui très complexes, doivent répondre aux contraintes de sécurité et de fiabilité tout en préservant leurs disponibilités et leurs performances.

La technicité de supervision de ces systèmes doit alors suivre leurs évolutions. La doter d'un outil de diagnostic adapté, à la fois rapide et robuste, peut prémunir au mieux des défaillances qui sont engendrées, la plupart du temps, par une propagation de défauts simples. Le diagnostic peut permettre d'éviter les conséquences catastrophiques sur le plan économique, humain et environnemental [PER84].

On rencontre souvent trois types de défaillance. Le premier est lié à un composant du procédé commandé. Par exemple, l'explosion de la capsule (Apollo 13) en 1970, est causée principalement par une gaine dénudée se trouvant à l'intérieur du réservoir d'oxygène du module de service. Les données recueillies après l'explosion, ont mis en évidence des dommages subis par ce réservoir, qui fut récupéré d'une autre capsule (Apollo 10) lors de

l'opération de démontage. Les tests effectués au sol, avant le décollage, avaient montré une défaillance au niveau de la vidange de ce réservoir, mais celle-ci fut aussitôt contournée par une nouvelle procédure. Cette dernière a contribué à une forte augmentation de la température faisant fondre l'essentiel de l'isolation du câblage électrique à l'intérieur du réservoir. De plus, les systèmes d'alarme et de sécurité n'étaient pas adaptés pour faire face à ce type de problème.

Un autre exemple de défaillance d'un composant du procédé commandé est l'explosion d'un réacteur dans une usine chimique à Seveso, en Italie, en 1976. L'origine de l'accident est une fuite dans un réservoir provoquée par une probable corrosion du serpentin de réchauffage interne qui a libéré un nuage contenant plusieurs produits toxiques. Après analyse, il s'avère que la quantité de produit contenue dans ce réservoir était plus élevée que celle prévue. L'opérateur ne disposait pas d'indicateur de niveau, ni d'autre instrumentation qui aurait permis de détecter l'anomalie.

Un deuxième type est lié au système de commande. On peut citer à titre d'exemple, l'explosion de l'un des sites pétroliers les plus complexes du monde, la raffinerie pétrochimique au Texas qui s'est produite suite, entre autres, à une défaillance d'une vanne.

Un troisième type est lié à de mauvaises manipulations des opérateurs humains de supervision. Ce type de défaillance a été fortement impliqué dans la surchauffe et l'emballement du réacteur de la centrale nucléaire de Tchernobyl en 1986. Suite à une mauvaise interprétation des mesures de température et à des manœuvres indélicates pendant le recuit des barres de contrôle au cœur du réacteur, l'explosion inévitable a engendré des conséquences catastrophiques sur l'environnement.

1.3. Méthodes de diagnostic

Dans la littérature, on retrouve deux catégories de méthodes dédiées au diagnostic [ZWI95], internes et externes. Ces méthodes diffèrent selon la nature et l'étendue de la connaissance accessible du système. Les critères du choix d'une méthode de diagnostic peuvent se résumer comme suit :

- la dynamique du système (discrète, continue ou hybride),
- la structure d'implémentation (comparateur, filtre, référence, ...),
- la nature de l'information (quantitative et/ou qualitative),
- la complexité du système (large ou simple),
- la profondeur de l'information disponible sur le système (structurelle, analytique et/ou heuristique).

1.3.1. Méthodes internes

Les méthodes internes nécessitent une connaissance approfondie du système étudié, afin de le représenter analytiquement [ZWI95] sous forme d'un modèle quantitatif et/ou qualitatif. L'estimation des variations des paramètres du modèle, dans le cas continu, permet de détecter un éventuel défaut ou de déceler un résidu par rapport au système réel.

Parmi les méthodes internes à base de modèles, on peut distinguer les méthodes basées sur des modèles quantitatifs, les méthodes basées sur des modèles qualitatifs et les méthodes basées sur les deux modèles.

1.3.1.1. Méthodes à base de modèle quantitatif

Les modèles quantitatifs sont utilisés pour l'estimation de paramètres d'état ou d'espace de parité à travers des modèles mathématiques et/ou structurels pour représenter l'information disponible du fonctionnement d'un procédé. Un défaut provoque alors des changements dans certains paramètres physiques du procédé. Les modèles mathématiques comparent les différentes valeurs des variables avec des seuils de détection afin de générer un résidu qui sera fourni à l'outil de diagnostic. A partir de toutes les signatures de défauts déterminées par apprentissage, il est possible d'isoler et d'identifier la panne pour prendre une décision [ZWI95]. Les avantages de ces méthodes internes sont tout d'abord la capacité à détecter les variations abruptes et progressives des pannes à travers une analyse de tendance des signaux. De plus, ces méthodes possèdent la capacité de donner une localisation précise du défaut. Par contre, elles nécessitent une information dite "profonde" sur le comportement du système et de ses pannes, rendant les calculs complexes pour le diagnostic en ligne. Elles sont également très sensibles aux erreurs de modélisation.

1.3.1.2. Méthodes à base de modèle qualitatif

Les méthodes à base de modèles qualitatifs permettent de représenter le comportement du procédé avec un certain degré d'abstraction à travers des modèles non plus mathématiques mais des modèles de type symbolique [TRA97]. Les modèles qualitatifs doivent représenter de manière qualitative des systèmes continus, discrets et/ou hybrides pour que le diagnostic soit capable de détecter les déviations du fonctionnement normal, localiser la défaillance et en déterminer la ou les causes. Pour les systèmes continus, les modèles qualitatifs sont fréquemment basés sur des graphes causaux [MON00] ou des graphes causaux temporels [MOS01]. Une abstraction qualitative des comportements continus peut être représentée par des modèles à base d'événements discrets (SED) [SU03], ou par la théorie de supervision [RAM89]. Pour les SED, de nombreuses approches sont proposées utilisant des outils tels que les automates, les équations logiques ou les Réseaux de Petri (RdP) avec observation partielle ou totale du fonctionnement du procédé [PHI06].

1.3.1.3. Méthodes mixtes

Une intégration des modèles discrets et des modèles continus peut être retrouvée également dans les systèmes dynamiques hybrides [ALL98]. Les méthodes à base de modèles quantitatifs et qualitatifs reposent d'une part sur une évaluation quantitative pour la détection d'un défaut et d'autre part sur une analyse qualitative des transitoires pour la localisation et l'identification. Ces méthodes ont l'avantage de combiner les points forts des méthodes à base de modèles quantitatifs et ceux à base de modèles qualitatifs. Cependant, elles sont lourdes à implémenter. On peut citer, comme exemple de ces méthodes, celles développées par [MAN00].

1.3.2. Méthodes externes

Ces méthodes ont été développées pour pouvoir étudier efficacement la dynamique d'un système pour lequel le modèle mathématique est difficile à établir voir même inexistant. Le système est considéré comme étant une boite noire où seules les entrées et les sorties observables peuvent être mesurées. Ces mesures sont appelées signatures externes. La connaissance qualitative et/ou quantitative de ces signatures est précieuse pour l'étude de ces

systèmes. Ces méthodes sont basées sur un retour d'expérience et ont donc l'avantage d'être performantes avec un minimum de connaissance *a priori*. Dans la littérature, de nombreux travaux ont permis leur mise au point et leur utilisation comme par exemple la reconnaissance des formes [DUB90], les systèmes experts [ZWI95] et les réseaux de neurones artificiels [DUB01].

1.3.2.1. Systèmes experts

A partir d'informations heuristiques sur le fonctionnement d'un système, le diagnostic par système expert essaie de reproduire le comportement d'un expert humain dans son domaine en accomplissant une tâche d'association empirique entre les symptômes observés et les causes. D'une façon générale, le système expert [CHA93] est composé d'une base de connaissances et d'un moteur d'inférence. La base de connaissance correspondant d'une part aux règles qui modélisent la connaissance du domaine considéré et d'autre part aux faits contenant les informations concernant le cas observé à traiter (symptôme). Le moteur d'inférence doit être capable de raisonner par un choix de règles (causes) selon les faits observés par deux raisonnements différents. Le premier raisonnement est inductif et infère tous les symptômes qui sont la conséquence d'un symptôme initial (état réel du système), le deuxième est déductif et infère toutes les causes possibles pour expliquer un symptôme.

Les systèmes experts présentent certains avantages quant à leurs utilisations en diagnostic. En effet, ils sont efficaces au niveau du temps de calcul. Le système doit simplement attendre les évènements observables des règles pour "sauter" directement aux conclusions [UNG93] qui sont facilement interprétables pour l'opérateur. De plus, la facilité d'implémentation, en énumérant simplement les règles, contribue d'une certaine façon, à la capitalisation du savoir faire de l'expert du domaine.

Cependant, les systèmes experts présentent aussi certains points faibles trop difficiles à contourner. Par exemple, il est difficile d'extraire les connaissances des experts dans le cas d'une nouvelle installation sujette au manque d'informations concernant les modes de fonctionnements. Aussi, ils sont souvent soumis aux problèmes d'expertise du moment qu'ils sont liés à des applications prédéfinies dans un domaine précis. De plus, ces méthodes demandent un certain temps pour leur apprentissage au début de leurs phases d'exploitation et dans la mise à jour des règles.

1.3.2.2. Réseaux de neurones

Un Réseau de Neurones Artificiels (RNA) est un système informatique, constitué de processeurs élémentaires (ou nœuds) interconnectés entre eux, qui traite de façon dynamique l'information qui lui arrive à partir des signaux extérieurs.

De manière générale, l'utilisation des RNA se fait en deux phases. Tout d'abord, la synthèse du réseau est réalisée et comprend plusieurs étapes : le choix du type de réseau, du type de neurones, du nombre de couches, des méthodes d'apprentissage.

L'apprentissage permet alors, sur la base de l'optimisation d'un critère, de reproduire le comportement du système à modéliser. Il consiste à rechercher un jeu de paramètres (poids) et peut s'effectuer de deux manières : supervisé (le réseau utilise les données d'entrée et de sortie du système à modéliser) et non supervisé (seules les données d'entrée du système sont fournies et l'apprentissage s'effectue par comparaison entre exemples). Quand les résultats d'apprentissage obtenus par le RNA sont satisfaisants, il peut être utilisé pour la généralisation. Il s'agit ici de la deuxième phase.

On présente au RNA, de nouveaux exemples n'ayant pas été utilisés pendant l'apprentissage, pour juger de sa capacité à prédire les comportements du système ainsi modélisé.

Les propriétés rendant l'utilisation des RNA attrayante sont leur faible sensibilité aux bruits de mesures ; leur capacité, dans un premier temps, à résoudre des problèmes non linéaires et multi-variables, puis à stocker les connaissances de manière compacte et enfin à « apprendre » en ligne et en temps réel. Leur emploi peut alors se faire à trois niveaux :

- comme modèle du système à surveiller en état normal et pour générer un résidu d'erreur entre les observations et les prédictions,
- comme système d'évaluation de résidus pour le diagnostic,
- ou comme système de détection en une seule étape (en tant que classificateur), ou en deux étapes (pour la génération de résidus et le diagnostic).

Nous pouvons citer comme avantages, de l'utilisation de ces réseaux, leur réponse presque instantanée et leur applicabilité pour des systèmes dynamiques. Cependant, ils demandent un apprentissage lourd.

1.3.2.3. Reconnaissance des Formes (RdF)

La Reconnaissance des Formes (RdF) est une science qui regroupe l'ensemble des algorithmes ou méthodes permettant la classification d'objets ou de formes en les comparants à des formes-types [BOU97]. On suppose que chaque observation, appelée aussi forme, réalisée sur un système est caractérisée par un certain nombre de paramètres ou d'attributs. Chaque forme peut donc être représentée à l'aide d'un vecteur appelé "vecteur forme" dans un espace appelé "espace de représentation". On suppose aussi que dans cet espace on peut observer des formes de types différents, appelées aussi "prototypes" ou "formes-types".

Dans un cas idéal, où la notion de bruit n'est pas prise en compte, chaque nouvelle forme serait exactement confondue avec l'une des formes-types. Par contre, dans un cas réel, afin de traduire l'influence des perturbations sur le système étudié (bruit de mesure, précision des capteurs,…), une nouvelle observation sera rarement confondue avec l'une des formes-types. Il est alors difficile d'isoler un point unique de l'espace comme représentant de la forme-type. Une zone restreinte, appelée "classe" et notée C_i, est définie autour de chaque forme-type de l'espace de représentation en englobant les formes semblables comme le montre la figure ci-dessous.

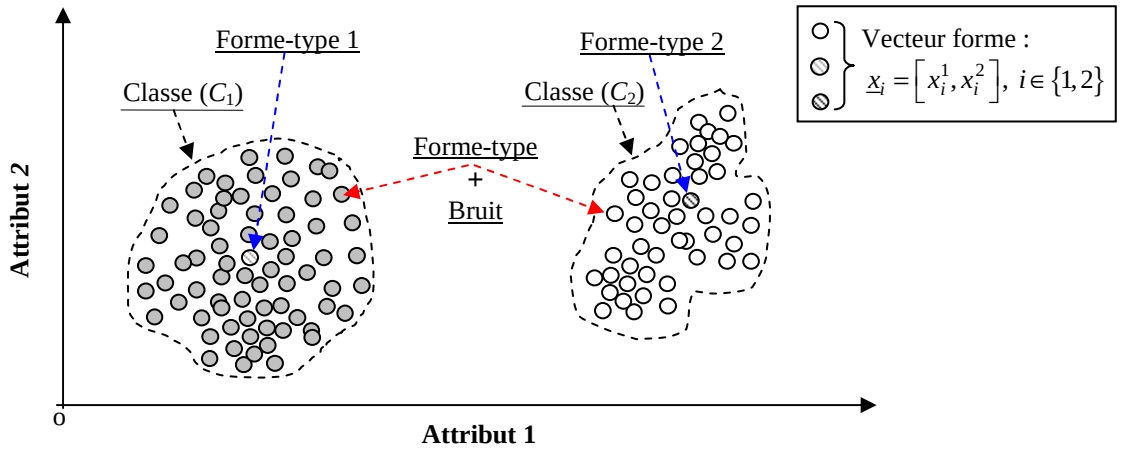


Figure 1.1 Exemple montrant les vecteurs formes dans un espace de dimension 2 et le lien entre les notions de classe et de forme-type en RdF.

De manière générale, on distingue selon la nature de la forme étudiée deux types de RdF. Le premier, appelé RdF statistique, se base sur les propriétés numériques des formes étudiées [FUK72]. Le second type, qui ne sera pas développé par la suite, est qualifié de RdF structurelle, pour lequel les formes sont essentiellement caractérisées par des propriétés grammaticales [FRI99].

Dans le contexte du diagnostic des systèmes industriels, une forme est l'observation de l'état du système étudié et chaque classe de l'espace de représentation représente un mode de fonctionnement. Le problème du diagnostic par RdF est de savoir décider à quelle classe (mode de fonctionnement normal ou défaillant), parmi les classes connues *a priori*, une nouvelle observation recueillie sur le système ressemble le plus. Cela devient alors un problème de classification qui cherche à identifier, selon les règles de décisions utilisées, un des modes de fonctionnement connus *a priori*.

De nombreux travaux, [DUB90, SAY02a, OND06, AMA06], ont permis de montrer l'intérêt du diagnostic par RdF dans différentes applications réelles où la modélisation du procédé est souvent difficile à établir tels que la biométrie, la reconnaissance vocale, la reconnaissance de caractères, l'automatisation industrielle, le diagnostic médical, le diagnostic des machines électriques, le comportement humain, etc.

On peut citer plusieurs propriétés qui rendent l'utilisation de la RdF intéressante comme l'efficacité en termes de temps de calcul pour la classification d'une nouvelle observation, la capacité à traiter des données qui sont à la fois incertaines et imprécises par son association avec la théorie des ensembles flous [BOU97, SAY02a]. Egalement, la possibilité de détecter et de suivre l'évolution des modes de fonctionnement, procurant ainsi au système de diagnostic des moyens d'appréhender la connaissance *a priori* incomplète de ces modes.

1.4. Modélisation de l'information imparfaite

Dans ce qui suit, nous étudions les principales théories permettant, de représenter les connaissances imparfaites et de raisonner à partir de celles-ci. Ces théories sont : les probabilités, les fonctions de croyance, les ensembles flous et les possibilités.

1.4.1. Théorie des probabilités

La théorie des probabilités est l'outil le plus traditionnel qui permet de modéliser une information imparfaite [SAN91]. La notion de probabilité est liée à celle d'expérience aléatoire dont on ne peut prédire avec certitude le résultat. Cette expérience est un événement représentant une proposition logique relative au résultat. Cet événement est un élément de l'ensemble Ω de tous les événements possibles, appelé univers ou référentiel. La probabilité de l'occurrence d'un événement A appartenant à l'ensemble $S(\Omega)$ des sous-ensembles de l'univers est décrite par une mesure de probabilité P qui est une fonction de l'ensemble des parties de Ω dans [0, 1] et qui satisfait les axiomes suivants :

$$\forall A \subseteq \Omega, \qquad (P(\varnothing) = 0) \leq P(A) \leq (P(\Omega) = 1) \tag{1.1}$$

$$\forall A, B \subset \Omega, \qquad \text{si } A \cap B = \varnothing,\ P(A \cup B) = P(A) + P(B) \tag{1.2}$$

L'axiome (1.1) établit que Ω est un évènement certain contrairement à l'ensemble vide $\varnothing$ qui est un évènement impossible. Le nombre non-négatif $P(A)$ quantifie donc dans quelle

mesure l'évènement A est probable. L'axiome d'additivité (1.2) établit le fait que, si deux événements sont mutuellement exclusifs, alors la probabilité qu'au moins l'un d'entre eux ait lieu est la somme de leur probabilité individuelle. On peut déduire de ces axiomes de base que la connaissance de la probabilité d'un événement A détermine complètement celle de son événement contraire $\bar{A}$:

$$\forall A \subseteq \Omega, \qquad P(A) = 1 - P(\bar{A}) \tag{1.3}$$

En lien avec la mesure de probabilité P, et dans le cas d'un univers fini et discret formé d'un ensemble de singletons, une fonction $p : \Omega \to [0,1]$ peut être définie par $p(\omega) = P(\{\omega\})$. Cela veut dire que la probabilité des événements peut être caractérisée aussi par la distribution de probabilité p sur un univers discret Ω.

De l'axiome de l'additivité (1.2) et en vérifiant la condition de normalisation ($\sum_{\omega \in \Omega} p(\omega) = 1$), on déduit que:

$$\forall A \subseteq \Omega, \qquad P(A) = \sum_{\omega \in A} p(\omega) \tag{1.4}$$

Dans le cas où Ω est continu, la fonction p devient une densité de probabilité $p : \Omega \to [0,1]$, telle que $\int_{\Omega} p(\omega) d\omega = 1$ et $P(A) = \int_{A} p(\omega) d\omega$.

On peut distinguer deux interprétations concernant la mesure de probabilité P. La première, appelée fréquentiste, considère une probabilité $P(A)$ comme la limite d'une fréquence de l'occurrence de l'événement A lorsque l'expérience est répétée un grand nombre de fois. Elle est à la base de l'utilisation des informations statistiques. La deuxième interprétation, dite subjectiviste, exclue la notion de répétition d'une expérience aléatoire et suppose qu'on peut attacher à la probabilité $P(A)$ une valeur subjective quantifiant la croyance qu'on accorde, à l'occurrence de A, de façon à ce que P obéisse aux axiomes ci-dessus.

Lorsque certains indices observés permettent d'augmenter la confiance à une hypothèse, l'axiome d'additivité (1.2) nécessite de diminuer d'autant la confiance correspondant à l'hypothèse contraire. Néanmoins, dans certains cas, des informations peuvent très bien favoriser une hypothèse sans pour autant discréditer l'hypothèse contraire. La théorie des probabilités semble donc peu adaptée à des situations où la connaissance d'un événement comme la connaissance de son contraire sont très limitées [ZOU97].

Exemple 1.1

Prenons comme exemple un problème de classification des objets dans une des deux classes C_1 et C_2. Si un objet ω appartient fortement à C_1, il aura une probabilité $p_{\omega}(\{C_1\})$ d'appartenance à cette classe très forte, tandis qu'il aura une valeur faible d'appartenance à C_2. Cette constatation se justifie par la condition de normalisation (1.3) : $p_{\omega}(\{C_1\}) + p_{\omega}(\{C_2\}) = 1$.

Supposons maintenant que la connaissance que l'on a des classes est incomplète et qu'il y a une troisième classe inconnue. Si l'objet ω n'appartient ni à C_1 ni à C_2, à cause de la condition de normalisation cet objet aura certainement une probabilité d'appartenance forte à une des deux classes C_1 ou C_2. La théorie des probabilités est donc inadaptée pour

représenter une information incomplète. De plus, cette théorie est trop rigide pour exprimer le cas de l'ignorance totale. En effet elle modélise ce cas par un ensemble d'événements mutuellement disjoints et équiprobables :

$$\forall \omega \in \Omega, \qquad p(\omega) = \frac{1}{|\Omega|} \tag{1.5}$$

Exemple 1.2

Pour comprendre ce problème de non distinction entre le cas d'ignorance totale et une distribution équiprobable, prenons l'exemple cité en [QUI00]. Cet exemple représente une machine qui tombe en panne si la fréquence d'excitation se trouve dans un intervalle dit dangereux, comme le montre la figure 1.2. Nous n'avons aucune information sur la distribution de probabilité dans cet intervalle ni sur sa largeur Δ. Nous nous trouvons dans le cas d'ignorance totale représenté par une distribution équiprobabilité $1/\Delta$. Puisque nous n'avons aucune connaissance sur la largeur de l'intervalle et afin d'augmenter la sécurité, nous élargissons cet intervalle de Δ à Δ'. Cela va diminuer la probabilité assignée à chaque fréquence possible dans Δ' de $1/\Delta$ à $1/\Delta'$. Cela conduit au problème de sous-estimation de la probabilité de l'occurrence d'une fréquence appartenant à Δ'.

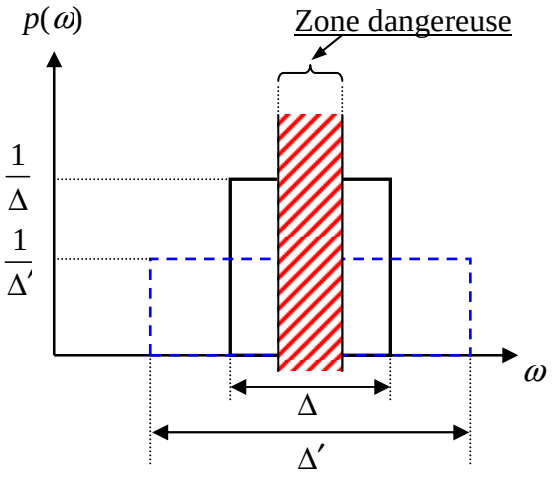


Figure 1.2 Exemple montrant le problème de non distinction entre l'ignorance totale et la distribution équiprobable.

Pour tenir compte de l'imprécision via la théorie de probabilités, une variable donnée se voit affecter un intervalle de valeurs possibles au lieu d'une valeur unique. De plus, on accorde souvent plus d'importance à la valeur médiane de l'intervalle qu'à ses bornes. Le cumul des erreurs, au bout d'un certain nombre de mesure, influence la largeur de l'intervalle de telle façon que celui-ci devient tellement large qu'il engendre une perte de toute signification. La théorie de probabilités est donc peu adaptée pour tenir compte de l'imprécision.

1.4.2. Méthodes de RdF basées sur la théorie des probabilités

1.4.2.1. Méthodes paramétriques

Les méthodes paramétriques supposent la connaissance des lois de probabilité régissant les observations dans les classes. La connaissance de ces lois se porte sur leur expression mathématique ainsi que sur leurs paramètres réels ou estimés. En effet, dans une classe donnée C_i, tout vecteur $\underline{x} \in \Re^a$ suit une loi de probabilité $p(\underline{x} \mid C_i)$ avec $i \in \{1,2,\ldots,c\}$. Les probabilités *a priori* $P(C_i)$ de chaque classe C_i sont connues et vérifient la condition d'orthogonalité :

$$\sum_{i=1}^{i=c} P(C_i) = 1 \tag{1.6}$$

Parmi ces méthodes, la méthode bayésienne [DUB90] est la plus courante. Elle utilise la formule de Bayes afin de calculer la probabilité *a posteriori*. Cette dernière, définit la probabilité qu'une nouvelle observation $\underline{x}$ dans $\Re^a$ provienne de la classe C_i, comme suit :

$$p(C_i \mid \underline{x}) = \frac{p(\underline{x} \mid C_i) \times P(C_i)}{\sum_{j=1}^{c} p(\underline{x} \mid C_j) \times P(C_j)} = \frac{f(\underline{x} \mid C_i)}{f(\underline{x})} \tag{1.7}$$

Cette formule est une conséquence des définitions des lois de probabilités conditionnelles $f(\underline{x} \mid C_i)$ et non conditionnelles ou du mélange $f(\underline{x})$.

La règle de décision de Bayes, concernant la classe C_i d'une nouvelle observation $\underline{x}$, s'écrit comme suit :

$$\underline{x} \in C_i \quad \text{si} \quad p(C_i \mid \underline{x}) = \max_{i=1,\ldots,c} p(C_i \mid \underline{x}) \tag{1.8}$$

Exemple 1.3

Dans le cas d'une classe gaussienne C_1, $\underline{x}$ obéit à une loi de Gauss multidimensionnelle dans $\Re^a$ si sa densité de probabilité s'écrit :

$$p(\underline{x} \mid C_1) = p(\underline{x}) = (2\pi)^{-\frac{a}{2}} \times |\Sigma|^{-\frac{1}{2}} \times e^{\left(-0.5 \times (\underline{x}-m)^T \times \Sigma^{-1} \times (x-m)\right)} \tag{1.9}$$

Les paramètres $\underline{m}$ et Σ sont respectivement le vecteur d'espérance mathématique et la matrice de variance covariance de $\underline{x}$ dans la classe C_1 qui traduit la dispersion des points de l'ensemble d'apprentissage sur chaque dimension. Ces paramètres peuvent être calculés comme suit :

$$E[\underline{x}] = \underline{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_a \end{bmatrix} \quad et \quad \Sigma = E[(\underline{x} - \underline{m}) \times (\underline{x} - \underline{m})^T] \tag{1.10}$$

$|\Sigma|$ représente le déterminant de Σ et Σ^{-1} sa matrice inverse. Cette loi gaussienne est représentée par $\mathcal{N}(\underline{m}, \Sigma)$. La figure 1.3 montre un exemple d'une distribution gaussienne de moyenne 0 et de variance 1.

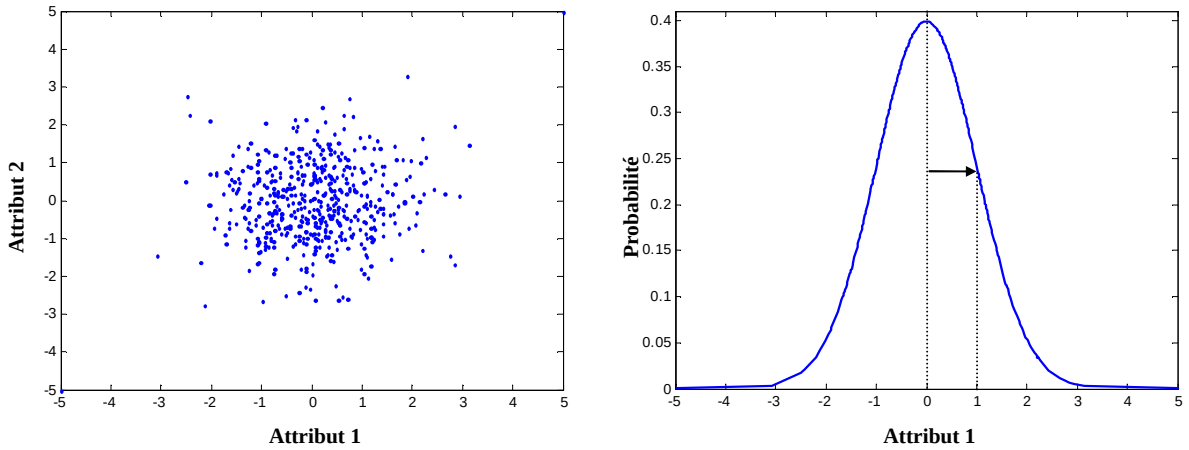


Figure 1.3 Classe gaussienne comportant 500 points représentée à gauche et sa loi gaussienne, selon l'attribut 1, de moyenne 0 et de variance 1, représentée à droite.

1.4.2.2. Méthodes non paramétriques

Dans le cas où la forme des lois de probabilité est inconnue on utilise les méthodes non paramétriques. Il s'agit dans ce cas d'estimer :

- soit directement les probabilités *a posteriori* d'appartenance aux classes, à partir de l'ensemble d'apprentissage, grâce à la méthode des k plus proches voisins au sens des distances de Cover et Hart [COV67],
- soit d'estimer la probabilité conditionnelle par la méthode des noyaux de Parzen [Par62] ou par les Méthodes à base d'histogrammes.

Ces méthodes sont détaillées ci-dessous.

1.4.2.2.1. Méthode des k Plus Proches Voisins

La méthode des k plus proches voisins (kppv) compte parmi les méthodes non paramétriques les plus simples. Elle a été introduite par Fix et Hodges [Fix51]. Elle considère l'estimation de la densité de probabilité comme une fonction purement locale. Un volume $V(\underline{x})$ centré sur un point $\underline{x}$ est déterminé de façon à ce qu'il englobe k observations de l'ensemble d'apprentissage. Une fois ce volume déterminé, le nombre k_i de voisins de $\underline{x}$ appartenant à chaque classe permet de calculer la densité de probabilité de la classe C_i contenant N_i points comme suit :

$$\hat{p}(\underline{x} \mid C_i) = \frac{k_i}{N_i . V(\underline{x})} \tag{1.11}$$

La définition de la notion de plus proches voisins est liée au choix d'une distance. Ce choix influence la forme du volume $V(\underline{x})$. Cette forme est une sphère avec la distance euclidienne, un cube avec la distance de Manhattan et enfin une ellipsoïde avec la distance de Mahanalobis.

Pour la prise de décision concernant une nouvelle observation, celle-ci sera affectée à la classe la plus représentée parmi ces k plus proches voisins. La règle de décision de Bayes (1.8), concernant la classe C_i dans laquelle un nouveau point $\underline{x}$ sera classé, devient alors comme suit :

$$\underline{x} \in C_i \quad \text{si} \quad p(C_i \mid \underline{x}) = \max_{i=1,\ldots,c}(k_i) \tag{1.12}$$

où k_i est le nombre de plus proches voisins appartenant à C_i parmi les k, avec $k = \sum_{i=1}^{c} k_i$.

Cette méthode, en apparence très simple, est robuste. Elle permet de reconnaître les formes même lorsque la distribution des classes n'est pas convexe. Son taux d'erreur de classification tend vers l'erreur minimale obtenue par la méthode bayésienne si k tend vers l'infini et il est majoré par deux fois cette erreur si $k = 1$.

Cependant, cette méthode devient inefficace quand la taille de l'ensemble d'apprentissage est importante. En effet, à chaque nouvelle observation à classer, cette méthode mesure toutes les distances entre cette observation et celles de toutes les observations de l'ensemble d'apprentissage. Elle nécessite, dans ce cas, des ressources mémoires importantes et un temps de calcul élevé qui dépend principalement de la distance utilisée. Par exemple, le choix d'une distance euclidienne dans le cas d'un ensemble d'apprentissage contenant une forme allongée n'est pas adapté. Celle-ci ne prend pas en compte la dispersion des points contenus dans ces classes. La distance de Mahanalobis peut être une solution mais le temps de calcul peut devenir encore plus conséquent avec le choix de cette métrique.

Dans la littérature plusieurs solutions ont été proposées pour accélérer le temps de classification de cette méthode, notamment la réduction des classes à des prototypes [KEL85] mais elles engendrent une perte d'information sur les classes.

Exemple 1.4

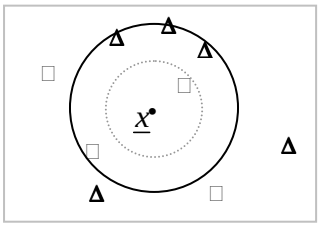


Figure 1.4 Exemple de classification par la méthode des k plus proches voisins.

D'après cet exemple, la classe de la nouvelle donnée $\underline{x}$ dépend du choix de k qui est déterminant pour le résultat final :

- Pour $k = 1$, $\underline{x}$ sera classé "□" (la forme la plus proche de $\underline{x}$),
- Pour $k = 5$, le même $\underline{x}$ sera classé "Δ".

Lorsqu'une nouvelle observation est située au voisinage de la frontière de décision elle peut engendrer une situation d'ambiguïté concernant son affectation à l'une des classes apprises. Cette observation devrait donc être rejetée en ambiguïté.

Rejet en ambiguïté

C'est la réponse à la question : comment une observation $\underline{x}$ peut être rejetée si elle est trop près d'une frontière de décision. L'algorithme ci-dessus, sous cette forme, ne traite pas le cas d'un classement incertain ou erroné de $\underline{x}$. Afin de réduire ce risque, Chow [CHO57] propose d'ajouter une classe C_0, appelée classe de rejet, à laquelle sont affectés les observations ambiguës. En diagnostic industriel ce type de rejet correspond à des régimes de fonctionnement proches de ceux des modes déjà prédéfinis et qu'il est impossible de différencier. Cette règle de rejet s'exprime sous la forme suivante :

$$\underline{x} \in \begin{cases} C_i,\ i = 1,2,\ldots,c \quad \Leftrightarrow \quad \underline{x} \text{ est classé dans } C_i \\ C_0 \quad \Leftrightarrow \quad \underline{x} \text{ est rejeté} \end{cases} \tag{1.13}$$

Hellman [HEL70] et Tomek [TOM76] fixent un seuil minimal, k', de Plus Proches Voisins (PPV) appartenant à la même classe, requis pour que $\underline{x}$ ne soit pas rejeté en ambiguïté. Ils ont obtenu la règle des (k, k')-PPV:

$$\underline{x} \in \begin{cases} C_i \quad \Leftrightarrow \quad k_i = \max_{i=1,\ldots,c}(k_i) \geq k' \\ C_0 \quad \Leftrightarrow \quad k_i = \max_{i=1,\ldots,c}(k_i) < k' \end{cases} \tag{1.14}$$

- Si $k' = 0$, cette règle devient celle des kppv sans rejet.
- Plus k' est proche de k, plus le rejet est important.

La règle (1.14) a été généralisée par Devijver [DEV77] en fixant un nombre minimal de voisins propres à chacune des classes.

Lorsqu'une observation est située loin de toutes les classes son affectation à une classe est très risquée et peut induire des erreurs de classement. Cette observation devrait donc être rejetée en distance.

Rejet en distance

Si le point se trouve loin de toute classe connue, ce point peut être le représentant d'un nouveau mode de fonctionnement pour lequel on ne disposait *a priori* d'aucune information dans l'ensemble d'apprentissage. Donc ce point sera affecté à une classe d'attente appelée classe de rejet en distance C_d.

La règle de décision de kppv doit travailler dans des zones denses et donc il faut introduire la notion de distance minimum acceptable entre le nouveau point et son voisin. La prise en considération de l'éloignement des voisins peut se faire de différentes façons. L'idée la plus simple consiste à effectuer un seuillage de distance. Le point est rejeté si la distance moyenne à ses kppv est supérieure à un seuil T. Ce seuil doit dépendre de la classe et tenir compte de sa géométrie c'est-à-dire être lié au diamètre de la classe ou à des considérations statistiques.

Une autre possibilité est : le point $\underline{x}$ est affecté à la classe C_i si le nombre des plus proches voisins acceptables parmi ses k plus proches voisins est supérieur à un seuil k_i. Ce seuil est déterminé par la recherche du nombre minimum de voisins acceptables dans la classe C_i pour tout point $\underline{x}$ de l'ensemble d'apprentissage.

Exemple 1.5

Reprenons le nuage de points de la figure 1.3. Les figures 1.5 montrent la densité de probabilité jointe et la densité de probabilité marginale, conditionnelle, par rapport à un seul attribut : l'attribut 1, pour différentes valeurs de k. Nous constatons qu'une valeur de $k = 25$ conduit à l'obtention d'une densité de probabilité très hachée, c'est-à-dire que sa dérivée est discontinue et qu'elle est donc loin de la densité théorique représentée dans la figure 1.3. Par contre, une estimation avec une valeur de k plus élevée tend à se rapprocher d'avantage de la densité théorique. Cependant, quelle que soit la valeur de k, la dérivée de l'estimation comportera toujours une discontinuité. Ce problème peut être résolu par la méthode des noyaux de Parzen que nous allons illustrer par la suite.

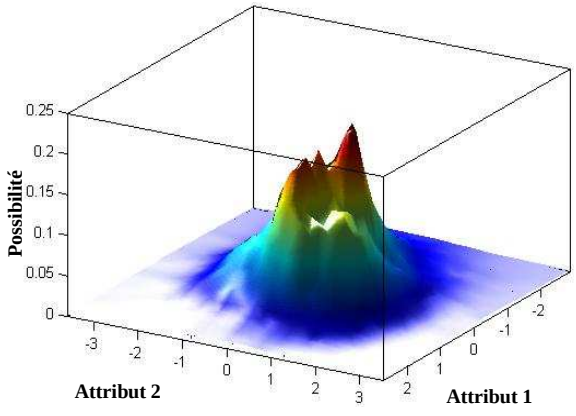


Figure 1.5.a Densité de probabilité jointe (kppv avec k=25) pour l'ensemble de points de la figure 1.3.

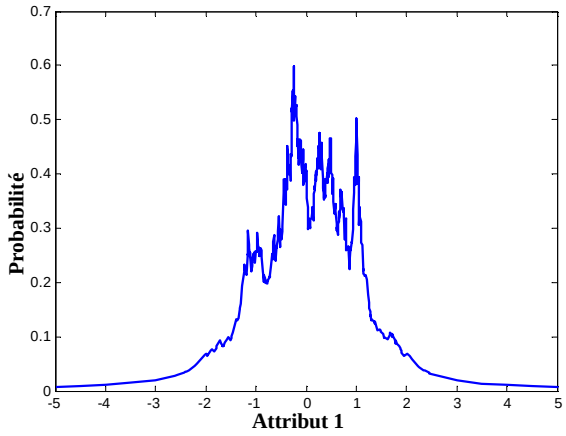


Figure 1.5.b Densité marginale estimée par kppv avec k=25.

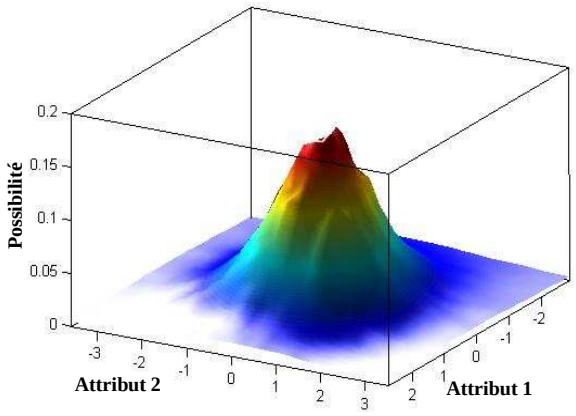


Figure 1.5.c Densité de probabilité jointe (kppv avec k=70) pour l'ensemble de points de la figure 1.3.

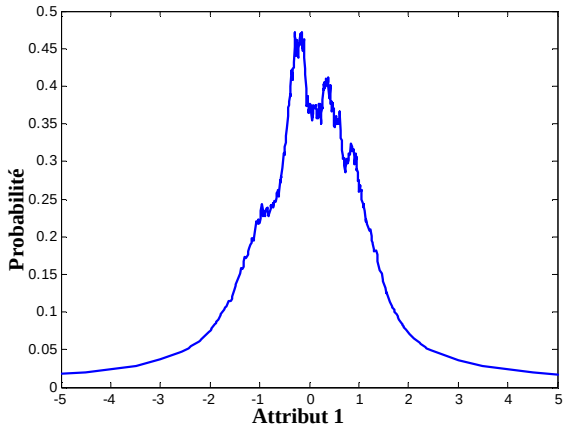


Figure 1.5.d Densité marginale estimée par kppv avec k=70.

1.4.2.2.2. Méthode des Noyaux de Parzen

Cette méthode non paramétrique a vu le jour grâce aux travaux de Rosenblatt [ROS56] et de Parzen [PAR62]. Ces derniers, afin d'estimer la densité de probabilité conditionnelle $p(\underline{x}|C_i)$ à partir de l'ensemble d'apprentissage, ont introduit une fonction φ appelée noyau. Le noyau φ doit être défini positif et normalisé, vérifiant donc respectivement les conditions des densités de probabilités ci-dessous :

$$\forall \underline{x} \in \Re^a, \quad \varphi(\underline{x}) \geq 0 \tag{1.15}$$

$$\int_{\Re^a} \varphi(\underline{x}) d\underline{x} = 1 \tag{1.16}$$

Le principe d'estimation de la densité en un point $\underline{x}$ par la méthode des fenêtres de Parzen consiste à compter le nombre de points de chaque classe C_i qui se trouve dans un volume $V(\underline{x})$, fixé *a priori*. Ce volume est une fenêtre définie par une distance, choisie selon la répartition des points d'apprentissage, et ayant comme centre le point $\underline{x}$. Le plus simple correspond à un hypercube de côté w. Cependant, si le nombre des points comptabilisé est trop faible, cela peut conduire à une estimation très bruitée de $p(\underline{x}|C_i)$. Afin de contourner ce problème, un noyau $\varphi(\frac{\|\underline{x}-\underline{x}_k\|}{w})$ est défini pour chaque point $\underline{x}_k$ appartenant à $V(\underline{x})$ et la somme de ces noyaux au point $\underline{x}$ permet d'évaluer la densité en ce point. De cette manière, chaque observation de la classe C_i intervient dans l'estimation de $p(\underline{x}|C_i)$ et non plus uniquement celles situées à proximité immédiate de $\underline{x}$ comme dans la méthode des k plus proches voisins. La densité de probabilité estimée peut alors s'écrire comme suit :

$$\hat{p}(\underline{x}|C_i) = \frac{1}{N.w^a} \sum_{k=1}^{N} \varphi\left(\frac{\|\underline{x}-\underline{x}_k\|}{w}\right) \tag{1.17}$$

Le plus souvent, le noyau gaussien est le modèle le plus utilisé. Il permet d'introduire une approximation satisfaisante dans beaucoup d'applications. En plus, les éléments mathématiques de ce noyau sont relativement bien maîtrisés [GUI05]. Le paramètre d'ajustement w joue alors le rôle d'écart-type du noyau, plus w augmente et plus l'estimation de la densité sera lissée.

Exemple 1.6

La figure 1.6 présente un nuage comportant 10 points. Nous avons choisi ce nombre afin d'illustrer l'estimation de la densité par la méthode des Noyaux de Parzen. Les figures 1.7 montrent la densité de probabilité jointe des deux attributs 1 et 2, et la densité de probabilité marginale, conditionnelle, par rapport à un seul attribut, l'attribut 1, pour différentes valeurs de w. Nous constatons qu'une valeur faible de w génère des densités de probabilité restreintes aux points d'apprentissage, cf. figures 1.7.a et 1.7.b. Par contre, une valeur de w trop élevée conduit à la perte d'informations sur la distribution des points de l'ensemble d'apprentissage, cf. figures 1.7.e et 1.7.f. Par conséquent, il faut bien choisir la valeur de w afin d'obtenir une estimation qui se rapproche plus au moins de la densité jointe théorique, cf. figures 1.7.c et 1.7.d.

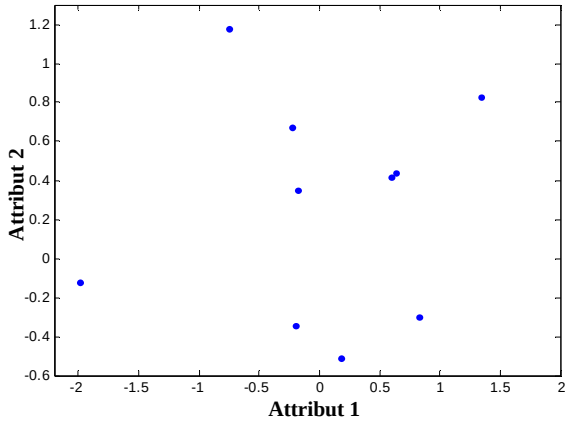


Figure 1.6 Classe gaussienne comportant 10 point, de moyenne 0 et de variance 1.

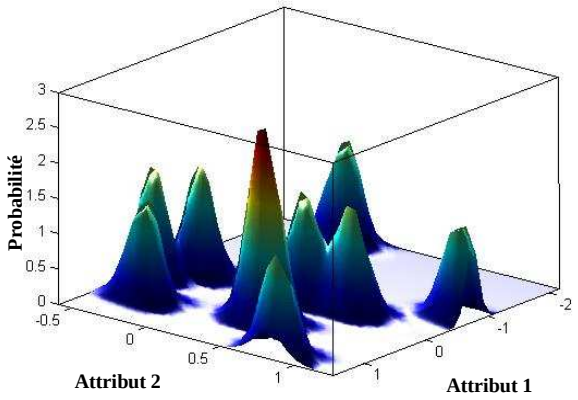


Figure 1.7.a Densité de probabilité jointe (fenêtre de Parzen w =0,1) pour un ensemble de 10 points.

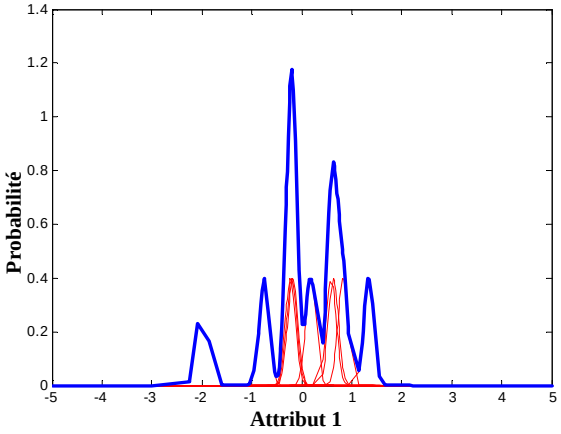


Figure 1.7.b Densité marginale estimée par le noyau gaussien avec w =0,1.

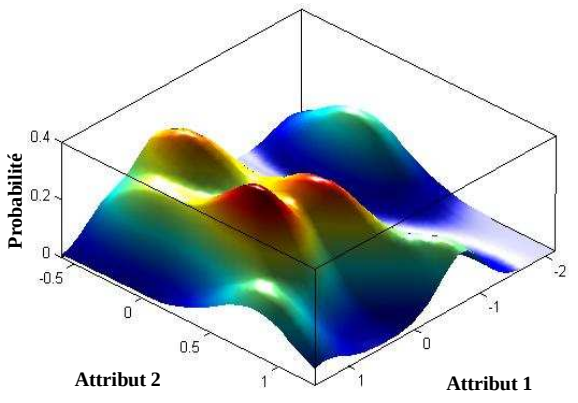


Figure 1.7.c Densité de probabilité jointe (fenêtre de Parzen w =0,3) pour un ensemble de 10 points.

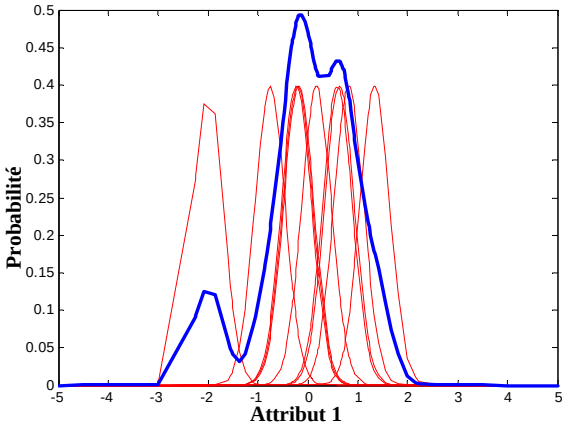


Figure 1.7.d Densité marginale estimée par le noyau gaussien avec w =0,3.

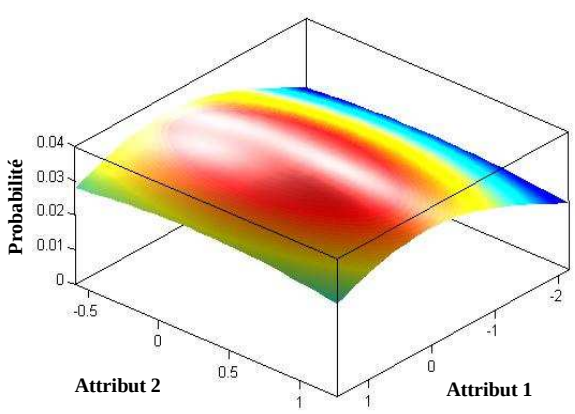


Figure 1.7.e Densité de probabilité jointe (fenêtre de Parzen w =1,9) pour un ensemble de 10 points.

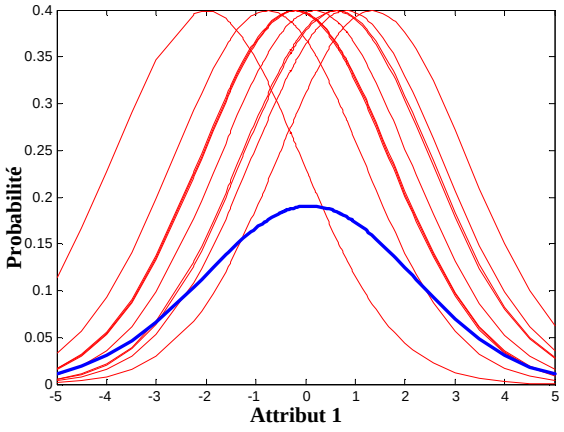


Figure 1.7.f Densité marginale estimée par le noyau gaussien avec w =1,9.

1.4.2.2.3. Méthodes à base d'histogrammes

L'histogramme donne une information sur la distribution d'un ensemble de points, appartenant à une classe, selon un attribut. Il est le moyen le plus ancien utilisé pour estimer une densité de probabilité. En effet, il est un outil d'estimation de l'occurrence des points d'apprentissage, dans une classe, dont découle la densité de probabilité. Les histogrammes se rapprochent de plus en plus de la vraie forme de la densité de probabilité quand la taille N de l'ensemble d'apprentissage grandit.

Chaque histogramme est défini par deux paramètres : le nombre de barres h et sa largeur représentée par un intervalle défini par les bornes inférieure Y_1^j et supérieure Y_2^j selon chaque attribut j. La largeur Δ^j d'une barre selon l'attribut j est calculée par :

$$\Delta^j = \frac{Y_2^j - Y_1^j}{h} \tag{1.18}$$

Une valeur trop élevée de h entraîne une représentation des données avec trop de détails alors qu'une valeur trop faible peut engendrer une perte d'informations. Dans les deux cas, l'histogramme résultant est une fausse estimation de la densité de probabilité. Il faut donc choisir h pour révéler l'essentiel de la structure de la densité de probabilité.

Malgré l'importance de ce paramètre h, peu d'études ont été consacrées à la détermination de sa valeur, surtout dans le cas où la forme de la densité de probabilité est inconnue. En supposant que cette forme est gaussienne, Scott [SCO79] propose de choisir la largeur Δ^j par :

$$\Delta^j = 3,49.\sigma.N^{-1/3} \tag{1.19}$$

où σ est l'écart-type des N points de l'ensemble d'apprentissage.

Freedman et Diaconis proposent une formule similaire mais plus robuste [IZE91, SCO92]:

$$\Delta^j = 2 \,.\, (IQR) \,.\, N^{-1/3} \tag{1.20}$$

où IQR est l'interquartile. Cette formule est plus robuste parce que l'interquartile est moins sensible que l'écart-type aux points aberrants.

Sturges donne une formule théorique pour le calcul de h, dans le cas d'une densité de probabilité normale. Cette formule donne la valeur minimale de h, utilisée comme une valeur par défaut dans les modules statistiques [SCO92] :

$$h = 1 + \log_2 N \tag{1.21}$$

He [HE97] propose de choisir h qui minimise la fonction d'erreur définie par :

$$h = \min\left(\frac{\sum_{k=1}^{h} \frac{p_{b_k}.(1-p_{b_k})}{(N+1)}}{\left\{\frac{(1-\frac{1}{h})}{(N+1)}\right\}^{\{(1+(1-r))\}}}\right), \qquad 5 \le h \le 20 \tag{1.22}$$

où r est l'entropie de Shannon divisée par le nombre de barres h : $r = -\sum_{i=1}^{h} p_{b_i}.\log_2(p_{b_i})/h$ et p_{b_k} est la probabilité de la barre k de l'histogramme. L'entropie de Shannon est utilisée pour mesurer l'incertitude contenue dans l'ensemble d'apprentissage. Cette formule traduit le fait qu'un ensemble d'apprentissage pauvre en informations a besoin d'un histogramme avec moins de barres qu'un ensemble d'apprentissage riche.

En supposant que la densité de probabilité est normale et pour un niveau de coupe $\alpha = 0.05$, Otnes [OTN72] calcule h par :

$$h = 1{,}87 \,.\, (N-1)^{2/5} \tag{1.23}$$

Klein [KLE98] utilise une formule empirique, pour minimiser le taux d'erreur de classification de la méthode Fuzzy Pattern Matching (FPM), en faisant l'analogie avec la formule de détermination du nombre de voisins de la méthode k plus proches voisins (kppv) proposée par Dubuisson [DUB90] :

$$h = \frac{\sqrt{N}}{2} \tag{1.24}$$

Nous proposons également par analogie avec la méthode kppv la formule suivante, utilisée par Enas [ENA86] :

$$h = N^{\left(\frac{2}{8}\right)} \tag{1.25}$$

ou bien :

$$h = N^{\left(\frac{3}{8}\right)} \tag{1.26}$$

Rudemo [RUD82] a proposé de minimiser l'erreur carrée intégrale définie par :

$$ISE = \int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \tag{1.27}$$

où $\hat{f}_h(x)$ est la densité de probabilité estimée de $f(x)$, en faisant intervenir le paramètre h. On utilise les couples d'apprentissage $\{y_k = f(x_k), x_k : k = 1,\ldots,n \text{ et } k \neq i\}$ pour estimer $\hat{f}_{h,-i}(x)$, la densité de probabilité sans le $i^{ème}$ point. Cette estimation est validée en calculant l'erreur d'estimation $y_i - \hat{f}_{h,-i}(x_i)$ définie par :

$$CV(h) = \frac{1}{N^2} \sum_{i=1}^{N} \left\{ y_i - \hat{f}_{h,-i}(x_i) \right\} \tag{1.28}$$

Finalement, h est choisi comme étant la valeur qui minimise l'équation ci-dessus.

Exemple 1.7

On considère la classe gaussienne $\mathcal{N}(0,1)$ de la figure 1.3. Les figures 1.8 comparent les densités de probabilités estimées par les formules précédentes.

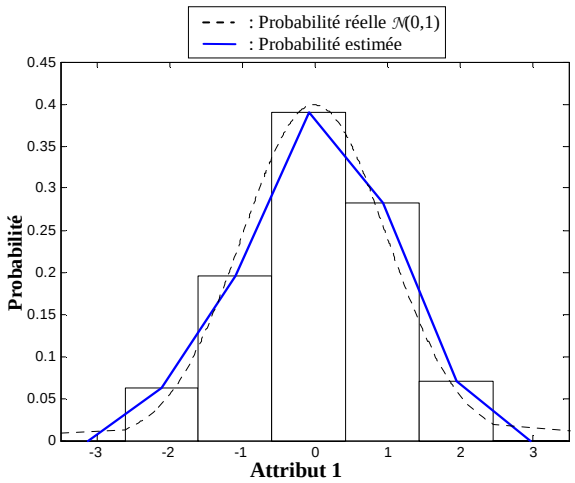


Figure 1.8.a Estimation par un histogramme avec un pas $h = N^{(2/8)}$ ou par minimisation de carrée intégrale de l' équation (1.27).

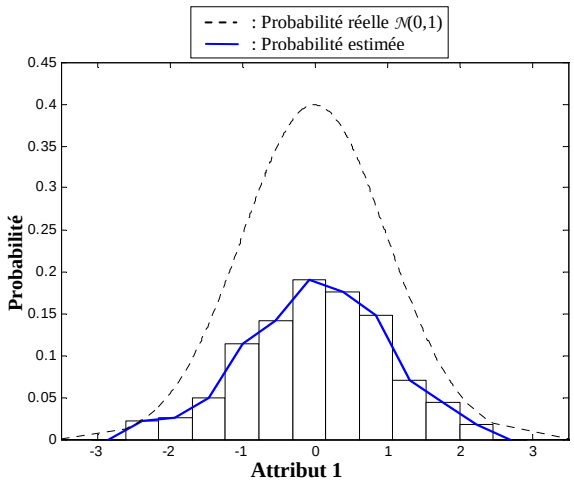


Figure 1.8.b Estimation par un histogramme avec un pas $h = 11$ calculé par $h = \sqrt{N}/2$.

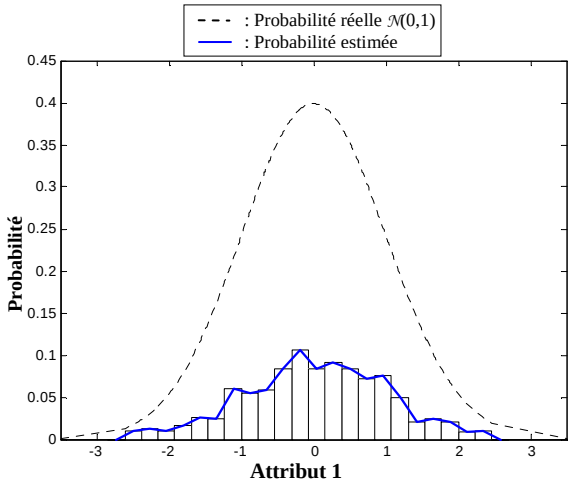


Figure 1.8.c Estimation par un histogramme avec h =1,87×(N-1)$^{2/5}$ =22.

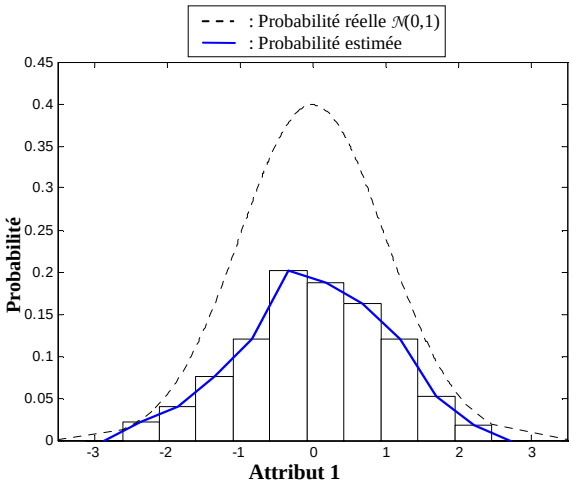


Figure 1.8.d Estimation par un histogramme avec h =1+log2(N) =10.

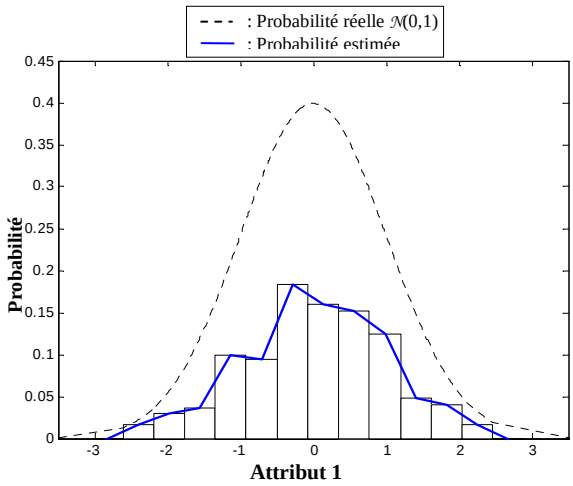


Figure 1.8.e Estimation par un histogramme avec h= 12 et $\Delta^1 = 3,49.\sigma.N^{-1/3}$.

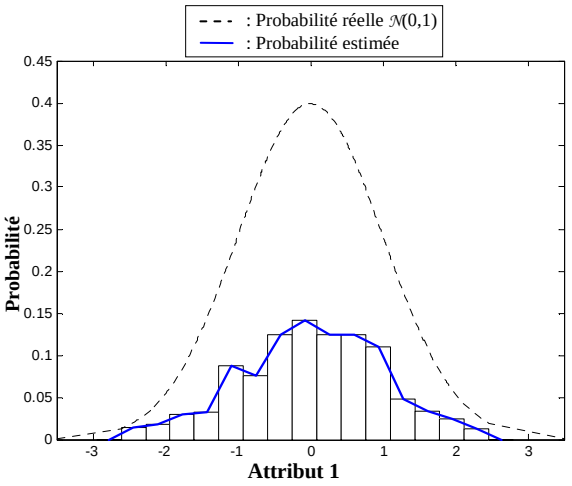


Figure 1.8.f Estimation par un histogramme avec un pas $\Delta^1 = 2\,.\,(IQR)\,.\,N^{-1/3}$ et (h = 15) calculé par la formule (1.22).

On a parlé précédemment de l'importance du choix du paramètre h pour l'estimation de la densité de probabilité à partir d'un ensemble d'apprentissage. Afin d'étudier la robustesse de cette estimation en fonction du nombre de points d'apprentissages disponible, nous avons généré une distribution gaussienne $\mathcal{N}(0,1)$ dont le nombre de points varie de 10 à 1000. La mesure de la qualité de l'estimation de la reconstruction $\hat{f}$ de f est représentée dans la figure 1.9. Elle est calculée par le taux d'Erreur Quadratique Moyen (EQM) :

$$EQM = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(x_i) - f(x_i)\right)^2 \qquad (1.29)$$

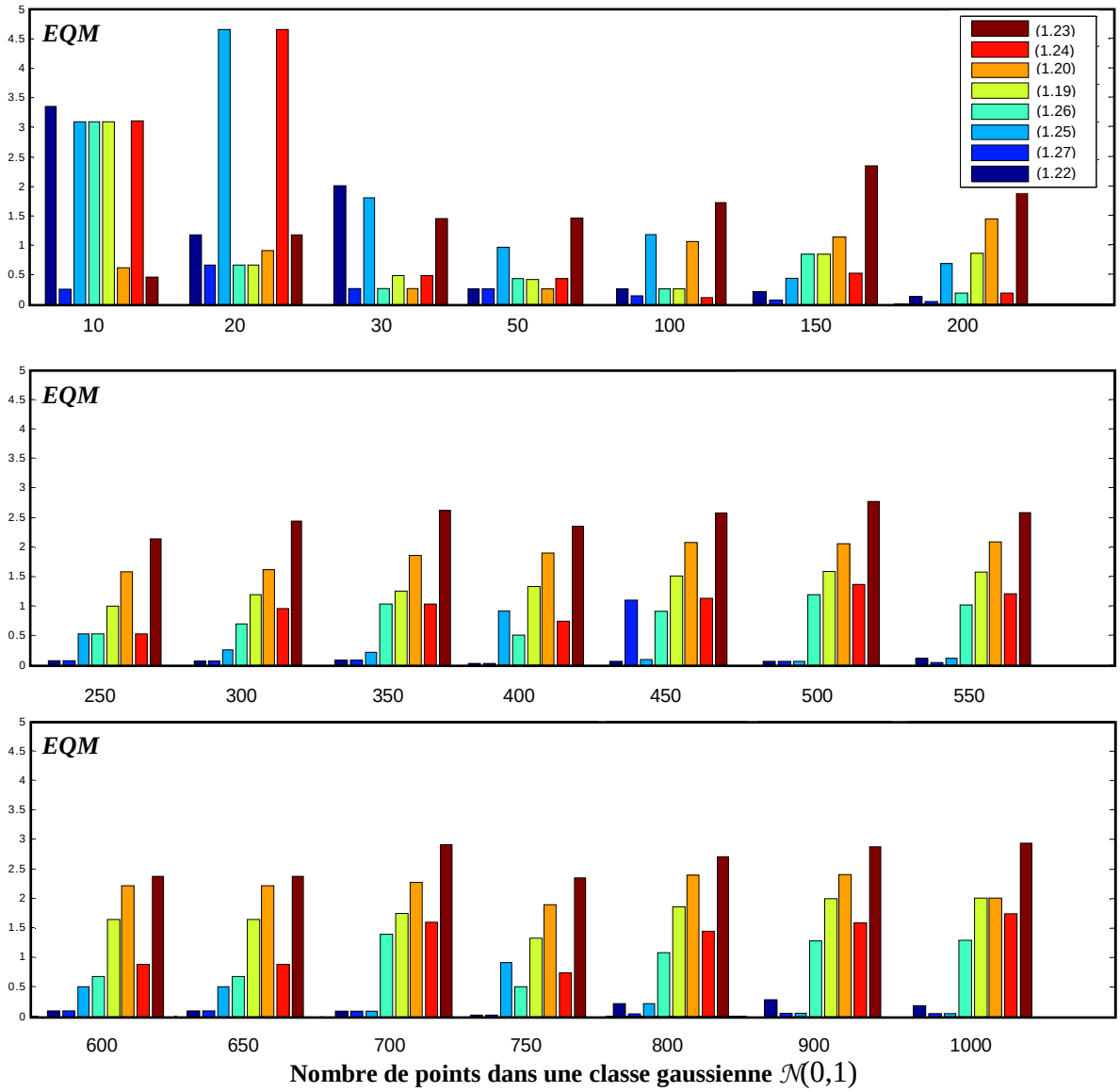


Figure 1.9 Taux d'Erreur Quadratique Moyen (EQM) calculé par un histogramme utilisant différentes formules pour le choix de h et différentes tailles de l'ensemble d'apprentissage.

D'après la figure 1.9, nous pouvons constater que la meilleure estimation est réalisée par les formules définies par (1.25) et (1.26).

Toutes ces méthodes nécessitent soit la connaissance de la forme de la densité de probabilité, soit la minimisation d'une fonction d'erreur. La minimisation d'une fonction d'erreur demande beaucoup de calcul et la complexité de ce calcul augmente considérablement avec la dimension. De plus, le minimum obtenu est très dépendant de l'ensemble d'apprentissage. Cela pose un grand problème si l'ensemble d'apprentissage est incomplet ce qui est souvent le cas dans la plupart des applications industrielles.

La distance entre les bornes inférieure Y_1^j et supérieure Y_2^j exprime l'intervalle de variation des valeurs pour un processus par rapport à l'attribut j. Toutes les observations à l'extérieur de cet intervalle ne sont pas considérées par l'histogramme. Un intervalle de

variation trop petit cause donc une perte d'informations et un intervalle trop large entraîne une sensibilité de l'histogramme aux points aberrants ou au bruit.

Dans la littérature, les bornes d'un histogramme sont déterminées soit comme les valeurs minimum et maximum de l'ensemble d'apprentissage, si on ne connaît pas la fonction de densité de probabilité [HE97]; soit en cherchant ses bornes pour que chaque barre ait au moins deux points d'apprentissage, si on connaît la fonction de densité de probabilité [OTN72].

1.4.3. Théorie des fonctions de croyance

La théorie des fonctions de croyance a été développée par Shafer en 1976 à la suite des travaux de Dempster sur les probabilités inférieures et supérieures. Une connaissance imparfaite sur Ω est représentée par une masse de croyance [SHA76, SME94] définie comme une fonction, nommée fonction d'allocation de masse, de 2^{Ω} dans [0, 1], vérifiant :

$$\sum_{A \subseteq \Omega} m(A) = 1 \qquad (1.30)$$

Les éléments A de Ω tels que $m(A) > 0$ sont appelés des éléments focaux de m. $m(A)$ peut être interprétée comme la fraction de la masse de croyance placée strictement en A et qui ne peut être allouée à aucune hypothèse plus restrictive, compte tenu de l'information disponible.

A la différence des probabilités, on voit qu'il est possible d'allouer de la masse à des sous-ensembles de Ω et non uniquement à des singletons. Cette possibilité fournit au modèle une grande souplesse de représentation. Il est en effet possible de modéliser des connaissances précises ou imprécises, certaines ou incertaines de manière très naturelle. L'ignorance complète correspond à $m(A) = 1$, alors qu'une connaissance précise et sûre correspond à l'attribution de la totalité de la masse à un singleton de Ω. Dans ce cas m est appelée une masse certaine. Une connaissance imprécise et sûre se traduira par l'allocation de la masse unité à un élément focal non singleton. Une connaissance incertaine correspondra à l'allocation de fractions de la masse unité à plusieurs singletons. Enfin une connaissance imprécise et incertaine correspond à l'allocation de fractions de la masse unité à plusieurs éléments focaux non singletons. La théorie des fonctions de croyance, contrairement à la théorie des probabilités, est donc bien adaptée pour représenter des connaissances à la fois imprécises et incertaines.

Un autre intérêt de la modélisation par la théorie des fonctions de croyance réside dans le cas d'une connaissance incomplète, notion de monde ouvert, qui est mal appréhendée par la théorie des probabilités. En fait la masse $m(\varnothing)$ est interprétée comme la part de croyance dans le fait que la vérité se trouve ailleurs que dans Ω. Si $m(\varnothing) = 0$, alors on a une connaissance complète, le monde est fermé.

Exemple 1.8

Reprenons l'exemple 1.1 de classification avec une information incomplète. Si un objet appartient à une classe encore inconnue, sa plausibilité d'appartenance pour les classes C_1 et C_2 aura la valeur 0. Par contre et afin de vérifier (1.30), la masse de croyance sera allouée à l'ensemble vide $\varnothing$ indiquant que la connaissance qu'on a des classes est incomplète et que cet objet appartient à une classe inconnue.

Une masse m peut être représentée par deux mesures non additives : une mesure de crédibilité Bel, de 2^{Ω} dans [0, 1], définie par :

$$\forall A \subseteq \Omega, \qquad Bel(A) = \sum_{B \subseteq A} m(B) \tag{1.31}$$

et une mesure de plausibilité Pl : de 2^{Ω} dans [0, 1], définie par :

$$\forall A \subseteq \Omega, \qquad Pl(A) = \sum_{A \cap B \neq \varnothing} m(B) \tag{1.32}$$

La fonction de croyance $Bel(A)$ prend en compte toutes les hypothèses qui impliquent A, en revanche la fonction de plausibilité $Pl(A)$ prend en compte toutes les hypothèses qui ne discréditent pas A. L'intervalle $[Bel(A), Pl(A)]$ peut être vu comme un intervalle encadrant une probabilité $P(A)$ mal connue [ZOU97] :

$$Bel(A) \leq P(A) \leq Pl(A) \tag{1.33}$$

La différence $Pl(A) - Bel(A)$ mesure l'ignorance relative à A. Les deux mesures de croyance Bel et de plausibilité Pl mettent en rapport un événement A avec l'événement contraire $\overline{A}$:

$$Bel(A) = 1 - Pl(\overline{A}) \tag{1.34}$$

L'équation ci-dessus interprète le fait que, plus on augmente la croyance dans une hypothèse, moins l'hypothèse contraire devient plausible. La théorie des fonctions de croyance de Dempster-Shafer présente l'avantage de permettre à la fois de représenter et de faire la distinction entre le cas de l'ignorance totale et le cas de répartition uniforme de la croyance. En théorie de Dempster-Shafer, on modélise l'ignorance totale par [SAN91] :

$$\forall\ A \subset \Omega : Bel(\Omega) = 1\ et\ Bel(A) = 0 \tag{1.35}$$

L'équi-répartition des croyances est représentée par :

$$\forall\ \omega \in \Omega,\ Pl(\{\omega\}) = \text{constante} \tag{1.36}$$

Il y a une infinité de distributions de plausibilité qui vérifient (1.36), on choisit pour la simplification la distribution $Pl(\{\omega\}) = \frac{1}{|\Omega|}$.

Exemple 1.9

Reprenons l'exemple 1.2, nous pouvons augmenter l'intervalle Δ sans pour autant sous-estimer la croyance que l'on a en l'occurrence d'une fréquence dans cet intervalle, comme le montre la figure ci-dessous.

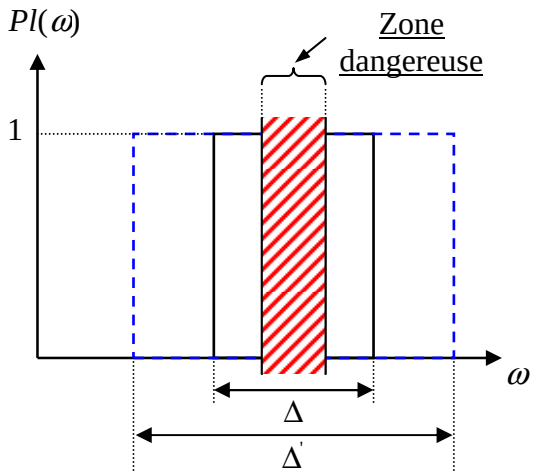


Figure 1.10 Exemple montrant l'intérêt de la théorie de fonctions de croyance comme solution pour le problème de non distinction entre l'ignorance totale et l'équi-répartition pour l'exemple 1.2.

1.4.4. Méthodes de RdF basées sur la théorie des fonctions de croyance

Nous prenons comme exemple de l'utilisation de la théorie des fonctions de croyance en RdF, la méthode des kppv fondée sur la théorie de Dempster et Shafer [DEN06]. Cette méthode considère les points de l'ensemble d'apprentissage comme étant des sources d'informations utilisées pour le classement d'un nouveau point. Ces différentes sources sont représentées par des masses de croyance m et combinées par la règle de Dempster.

Soit l'ensemble d'apprentissage X composé de N points répartis en c classes $\Omega = \{C_1, \ldots, C_i, \ldots, C_c\}$, soit Φ_k l'ensemble des k plus proches voisins d'un nouveau point $\underline{x}$ dans X, l'information apportée par $\underline{x}_q \in \Phi_k$ appartenant à la classe C_i et représentée par la masse de croyance $m_q : 2^\Omega \rightarrow [0,1]$, ayant pour élément focaux les classes $\{C_i\}$ et Ω, est définie comme suit :

$$m_q(\{C_i\}) = \alpha_0 . \varphi_{C_i}(d_i) \quad , \quad 0 < \alpha_0 < 1 \tag{1.37}$$

$$m_q(\Omega) = 1 - \alpha_0 . \varphi_{C_i}(d_i) \tag{1.38}$$

$$m_q(A) = 0 \quad , \quad \forall A \in 2^\Omega \setminus [\Omega, \{C_i\}] \tag{1.39}$$

$\varphi_{C_i}(d_i)$ est une fonction de la distance d_i entre le nouveau point $\underline{x}$ et son plus proche voisin $\underline{x}_q$. Il existe une infinité de fonctions de distance. La fonction inverse utilisée dans [ZOU97] est donnée par :

$$\varphi_{C_i}(d_i) = \frac{1}{1 + \lambda_{C_i} . d_i^2} \tag{1.40}$$

où $\lambda_{C_i} > 0$ est un paramètre représentant la distribution des points de la classe C_i.

La combinaison de toutes les masses de croyance de tous les points de Φ_k est réalisée par la règle de combinaison de Dempster définie par :

$$m_{C_i}(\{C_i\}) = 1 - \prod_{\underline{x}_q \in \Phi_{C_i}} (1 - \alpha_0 . \varphi_{C_i}(d_i)) \tag{1.41}$$

$$m_{C_i}(\Omega) = \prod_{\underline{x}_q \in \Phi_{C_i}} (1 - \alpha_0 . \varphi_{C_i}(d_i)) \tag{1.42}$$

La crédibilité et la plausibilité de la classe C_i est calculée par :

$$bel(\{C_i\}) = \underset{C_i}{m}(\{C_i\}) \tag{1.43}$$

$$pl(\{C_i\}) = \underset{C_i}{m}(\{C_i\}) + \underset{C_i}{m}(\Omega) \tag{1.44}$$

La classification d'un nouveau point $\underline{x}$ est réalisée par son affectation à la classe, $C(\underline{x})$, la plus crédible :

$$C(\underline{x}) = \underset{C_i, i=1,\ldots,c}{\arg\max}\, m(\{C_i\}) \tag{1.45}$$

Le rejet en ambiguïté et le rejet en distance sont deux situations qui se distinguent, selon la répartition de la masse de croyance. Dans le cas d'une répartition uniforme entre les classes, le maximum de plausibilité $pl(\{C(\underline{x})\})$ et le maximum de crédibilité $bel(\{C(\underline{x})\})$ prennent des valeurs relativement faibles au voisinage de la frontière entre les classes. Le rejet en ambiguïté est alors donné par :

$$pl(\{C(\underline{x})\}) < pl_{\min} \tag{1.46}$$

Par contre, dans le cas d'une répartition en grande partie sur l'ensemble Ω, le maximum de crédibilité $bel(\{C(\underline{x})\})$ devient proche de 0, tandis que le maximum de plausibilité $pl(\{C(\underline{x})\})$ devient proche de 1. Cela traduit le fait que le point $\underline{x}$ est éloigné de l'ensemble d'apprentissage. Le rejet en distance est réalisé donc par :

$$bel(\{C(\underline{x})\}) < bel_{\min} \tag{1.47}$$

Le problème de l'utilisation des seuils, $bel_{\min}$ et $pl_{\min}$, pour le rejet est délicat en particulier lorsque l'ensemble d'apprentissage est incomplet. L'inconvénient de l'application de la théorie des fonctions de croyance pour les kppv est qu'elle nécessite une base de données exhaustive contenant toutes les classes. De plus la complexité de calcul est importante.

1.4.5. Théorie des possibilités

La théorie des possibilités, introduite par Zadeh [ZAD78], est étroitement liée à la théorie des sous-ensembles flous [ZAD65]. Elle est présentée comme un cadre alternatif pour

représenter des informations entachées à la fois d'incertitudes et d'imprécisions. En effet, elle permet de formaliser des incertitudes de nature non probabiliste sur des évènements. En d'autres termes, c'est un moyen de dire dans quelle mesure la réalisation d'un événement est possible et dans quelle mesure on en est certain. On peut définir la mesure de possibilité Π d'un événement A comme une fonction $\Pi : S(\Omega) \to [0,1]$ qui associe à A un coefficient (degré de possibilité) compris entre 0 et 1. De plus cette fonction doit vérifier les axiomes suivants [DUB87] :

$$\forall A \subseteq \Omega, \qquad (\Pi(\varnothing) = 0) \le \Pi(A) \le (\Pi(\Omega) = 1) \tag{1.48}$$

$$\forall A, B \subset \Omega, \quad \Pi(A \cup B) = \max(\Pi(A), \Pi(B)) \tag{1.49}$$

Dans l'axiome (1.48), $\Pi(A)$ quantifie dans quelle mesure l'évènement A est possible : c'est-à-dire que A est tout à fait possible si la mesure de sa possibilité est égale à 1, et impossible si celle-ci est nulle. La mesure de possibilité est dite normale si $\Pi(\Omega) = 1$. Cela signifie que l'hypothèse d'un monde fermé est vérifiée.

L'axiome (1.49) traduit le fait qu'une mesure de possibilité n'est pas additive mais a un caractère d'ordre [DUB93]. De plus cette mesure n'exige pas que les deux événements A et B soient disjoints. On peut déduire des axiomes (1.48) et (1.49) que :

$$\forall A \subset \Omega, \qquad \max(\Pi(A), \Pi(\bar{A})) = 1 \tag{1.50}$$

Un événement possible A n'interdit donc pas pour autant l'événement contraire $\bar{A}$ de l'être également, ce qui peut conduire à une situation d'ignorance totale :

$$\forall A \subset \Omega, \qquad \Pi(A) = 1 \tag{1.51}$$

Pour remédier à cette situation, une mesure de nécessité $N(A)$, duale de la mesure de possibilité, a été introduite. Elle indique le degré avec lequel la réalisation d'un événement A est certaine. On dit qu'un événement A est nécessaire si son événement contraire est impossible $\Pi(\bar{A}) = 0$, d'où le lien entre les deux mesures :

$$\forall A \subset \Omega, \qquad N(A) = 1 - \Pi(\bar{A}) \tag{1.52}$$

La mesure de nécessité $N : S(\Omega) \to [0,1]$ vérifie les axiomes suivants :

$$\forall A \subseteq \Omega, \qquad (N(\varnothing) = 0) \le N(A) \le (N(\Omega) = 1) \tag{1.53}$$

$$\forall A, B \subset \Omega, \quad N(A \cap B) = \min(N(A), N(B)) \tag{1.54}$$

Le modèle possibiliste permet alors de modéliser l'imprécision incluse dans une proposition A par un couple de valeurs $(N(A), \Pi(A))$. De façon similaire à la mesure de probabilité P, une mesure de possibilité Π peut être définie à partir d'une distribution de possibilité π sur les singletons de Ω dans [0, 1] :

$$\Pi(\{\omega\}) = \pi(\omega), \quad \Pi(A) = \sup_{\omega \in A} \pi(\omega) \quad \text{avec} \quad \sup_{\omega \in \Omega} \pi(\omega) = 1 \tag{1.55}$$

(1.55) traduit le fait que la possibilité d'un événement formé par une collection d'éléments est égale au plus grand degré de possibilité parmi ses éléments. La mesure de nécessité peut également s'exprimer en fonction de la distribution de possibilité π associée à Π :

$$\forall A \subset \Omega, \qquad N(A) = \inf_{\omega \notin A}(1-\pi(\omega)) \tag{1.56}$$

Avec la théorie des possibilités on est seulement capable de représenter l'ignorance totale par (1.51). Nous ne pouvons pas donc distinguer ce cas d'ignorance totale, où tous les éléments ont la valeur de possibilité 1, du cas déqui-répartition.

1.4.6. Théorie des ensembles flous

La théorie des ensembles flous a été introduite par Zadeh [ZAD65] pour éviter la transition brusque qui existe entre deux sous-ensembles classiques et pour représenter les connaissances imprécises.

Dans la théorie classique, un sous-ensemble A d'un référentiel Ω est défini par une collection d'objets possédant une ou plusieurs propriétés communes caractéristiques de ce sous-ensemble. Chaque objet w de Ω est ainsi caractérisé par son appartenance ou non aux sous-ensembles A. Cela s'exprime par la fonction caractéristique $\mu_A : \Omega \to \{0,1\}$:

$$\mu_A(w) = \begin{cases} 1 & si \quad w \in A \\ 0 & si \quad w \notin A \end{cases} \tag{1.57}$$

Un sous-ensemble flou F d'un univers de discours Ω est défini par une fonction d'appartenance $\mu_F : \Omega \to [0, 1]$ qui associe à chaque élément w de Ω un coefficient $\mu_F(w)$ indiquant son degré d'appartenance à F. Plus w s'approche de la caractérisation typique du sous-ensemble flou F, plus sa valeur d'appartenance $\mu_F(w)$ tend vers 1. Le sous-ensemble flou F est défini par la connaissance des couples $(w, \mu_F(w))$.

Exemple 1.10

Soit Ω l'ensemble des tailles des personnes vivant en France. Le sous-ensemble classique T_{grand} correspondant aux personnes de grandes tailles, représenté figure 1.11, est défini par :

$$\mu_{T_{grand}}(w) = \begin{cases} 1 & si \;\; w \geq 170\ cm \\ 0 & sinon \end{cases} \tag{1.58}$$

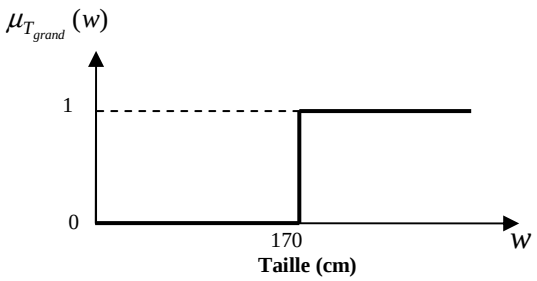


Figure 1.11 Fonction d'appartenance exclusive T_{grand}.

La fonction $\mu_{T_{grand}}(w):\Omega\rightarrow\{0,1\}$ traduit une appartenance tout ou rien de la taille au sous-ensemble T_{grand}. Afin d'éviter la transition brusque entre le 1 et le 0, l'idée d'appartenance graduelle de w au sous-ensemble T_{grand} est exprimée par la théorie des sous ensembles flous. La figure 1.12 montre la fonction d'appartenance pour l'exemple de la taille.

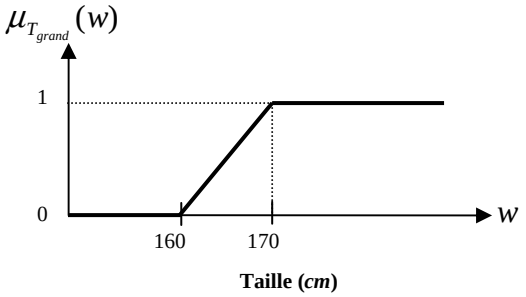


Figure 1.12 Fonction d'appartenance floue T_{grand}.

Un sous-ensemble flou A est caractérisé par :

- le noyau défini comme l'ensemble classique (ou net) des éléments ω appartenant totalement à A (c'est-à-dire pour lesquels $\mu_A(w)=1$),
- le support défini comme l'ensemble net des éléments ayant un degré d'appartenance non nul,
- la hauteur $h(A)$ comme la valeur $\sup_{w\in A}\mu_A(w)$. Si elle est égale à 1 on dit que A est normalisé.

Par ailleurs, on définit l'α-coupe de A (ou la coupe de niveau α), A_α, comme l'ensemble net des éléments ayant une appartenance $\mu_A(w)\geq\alpha$. La figure 1.13 montre ces éléments.

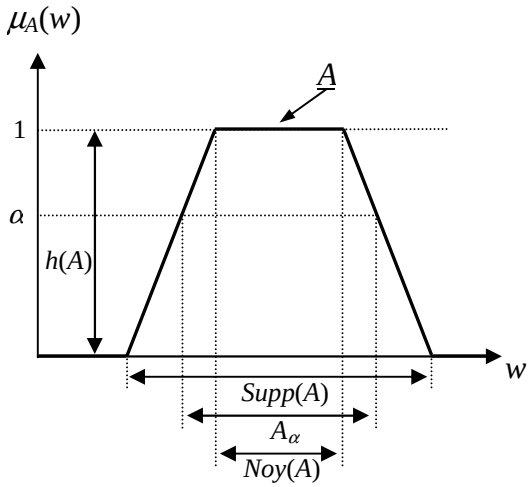


Figure 1.13 Quelques éléments caractéristiques de A.

Pour trouver le lien entre une distribution de possibilité et une fonction d'appartenance, prenons la proposition ''v est A'', A peut être interprété comme l'ensemble des valeurs possibles que peut prendre la variable v. Si certaines valeurs sont davantage possibles que

d'autres, A devient un ensemble flou. Pour un élément w de A, on associe une distribution de possibilité π. $\pi_v(w)$ indique la possibilité que la valeur de v soit w, en revanche $\mu_A(w)$ indique la valeur d'appartenance de w à A. La distribution de possibilité π peut être interprétée comme une fonction d'appartenance d'un ensemble flou [DEV93, ZOU97] :

$$\forall\, w \in A,\ \pi_v(w) = \mu_A(w) \tag{1.59}$$

Exemple 1.11

Soit Ω l'ensemble des entiers positifs et F le sous ensemble flou des petits entiers décrit par les paires $(w,\ \mu_F(w))$ suivantes : {(1, 1), (2, 1), (3, 0,8), (4, 0,6), (5, 0,4), (6, 0,2)}. La proposition ''v est un petit entier'' est associée avec la distribution de possibilité décrite en utilisant les paires $(w,\ \pi_v(w))$ suivantes : $\pi_v(w)$ = (1, 1), (2, 1), (3, 0,8), (4, 0,6), (5 0,4), (6, 0,2)}. La possibilité que l'élément $w \in \Omega$ soit égale à 3, sachant que w est un petit entier, est 0,8. Elle est égale au degré d'appartenance de cet élément w à l'ensemble flou F.

1.4.7. Méthodes de RdF basées sur la théorie des possibilités et des ensembles flous

La théorie des possibilités offre un formalisme pour manipuler l'imprécision et l'incertitude dans un système de RdF. C'est pour cela que plusieurs recherches ont contribué à utiliser cette théorie sur les algorithmes classiques de classification, supervisée ou non supervisée.

1.4.7.1. Méthode des k Plus Proches Voisins Floue

La méthode des k plus proches voisins floue, comme son nom l'indique, introduit la notion de flou dans la règle de décision de la méthode des k plus proches voisins classique [JOS83, KEL85]. Cette approche nécessite d'associer à chaque nouvelle observation $\underline{x}$ un degré d'appartenance $\mu_i(\underline{x}) \in [0,1]$ à chacune des c classes, vérifiant la propriété suivante :

$$\sum_{i=1}^{c} \mu_i(\underline{x}_q) = 1, \quad \text{avec} \quad q \in \{1, 2, \ldots, N\} \tag{1.60}$$

Les degrés d'appartenance $\mu_i(\underline{x})$ peuvent être fixés directement par des experts. Dans le cas où cette connaissance n'est pas disponible alors il faut les calculer. Keller [KEL85] a proposé de calculer le degré $\mu_i(\underline{x})$ en fonction de la distance entre $\underline{x}$ et ses k plus proches voisins, et des degrés d'appartenance de ses voisins aux différentes classes C_i :

$$\mu_i(\underline{x}) = \frac{\sum_{q=1}^{k} \mu_{iq} \cdot \left(\frac{1}{d(\underline{x}, \underline{x}_q)^{\frac{2}{m-1}}} \right)}{\sum_{q=1}^{k} \left(\frac{1}{d(\underline{x} - \underline{x}_q)^{\frac{2}{m-1}}} \right)} \tag{1.61}$$

où $d(\underline{x},\underline{x}_q)$ est la distance entre $\underline{x}$ et son $q^{ième}$ plus proche voisin. Le paramètre m sert à ajuster la valeur de l'appartenance selon l'importance de la distance $d(\underline{x},\underline{x}_q)$. Un choix judicieux de ce paramètre est $m=2$ [KEL85]. μ_{iq} est le degré d'appartenance à la classe C_i du $q^{ième}$ plus proche voisin de $\underline{x}$. La méthode kppv floue, contrairement à kppv classique, nécessite donc une initialisation des degrés d'appartenance de chaque point étiqueté par rapport à toutes les classes. Pour cela, plusieurs méthodes peuvent être utilisées. Keller [KEL85] a proposé d'utiliser :

- soit une initialisation nette (crisp en anglais) qui consiste à assigner à chaque $\underline{x}$ de l'ensemble d'apprentissage un degré d'appartenance égal à 1 pour sa classe et 0 pour toutes les autres classes,
- soit une initialisation selon la méthode des kppv classique comme ci-dessous :

$$\mu_{iq}=\begin{cases} 0,51+0,49\left(\frac{k_i}{k}\right) & \text{si} \quad C_i=C_q \\ 0,49\left(\frac{k_i}{k}\right) & \text{si} \quad C_i\neq C_q \end{cases} \tag{1.62}$$

où k_i est le nombre des voisin de $\underline{x}$ appartenant à la classe C_i parmi ses k plus proches voisins. C_q est la classe de $\underline{x}$ dans l'ensemble d'apprentissage.

La prise de décision, concernant une nouvelle observation $\underline{x}$, consiste à affecter celle-ci à la classe dont la fonction d'appartenance est maximale. La règle de décision de Bayes ou bien celle de la méthode des k plus proches voisins classique devient alors comme suit : $\underline{x}$ sera classé dans la classe C_i si :

$$\mu_i(\underline{x}) = \max_{i=1,\ldots,c}(\mu_i(\underline{x})) \tag{1.63}$$

Comme pour la règle k-PPV classique, on peut introduire l'option de rejet en ambiguïté et en distance définie précédemment, pour réduire le risque d'une mauvaise classification. L'option de rejet en appartenance (en distance) n'est possible que dans le cas d'une approche possibiliste comme dans la méthode décrite ci-dessous.

1.4.7.2. Fuzzy Pattern Matching (FPM)

Soit X un ensemble d'apprentissage contenant N points $\underline{x}$ et c classes. Chaque classe C_i comporte N_i points, dans un espace de représentation, $\Re^a$, de a attributs. La méthode FPM utilise la définition de c profils $\varphi_1,\ldots,\varphi_i,\ldots,\varphi_c$ pour les c classes [GRA92]. Chaque profil φ_i est représenté par une collection de a sous-ensembles flous $\varphi_i^1,\ldots,\varphi_i^j\ldots,\varphi_i^a$ qui expriment la plage des valeurs typiques que prend l'attribut j dans chaque classe C_i. Dans notre cas, il s'agit des densités de possibilité estimées notées Π_i^j. Le problème de classification consiste à affecter un nouveau point $\underline{x}\in\Re^a$, dont les valeurs pour les différents attributs sont $x^1,\ldots,x^a$, à une des classes. Le principe de fonctionnement de FPM nécessite deux phases : la phase d'apprentissage et la phase de classification.

1.4.7.2.1. Phase d'apprentissage

La phase d'apprentissage se préoccupe de la construction des profils φ_i sous forme de fonctions d'appartenance à partir de l'ensemble X. Cette construction est basée d'une part sur l'utilisation des histogrammes pour l'estimation de la probabilité conditionnelle de chaque classe et, d'autre part, sur la théorie des possibilités. Les étapes principales de cette phase sont présentées dans la figure 1.14.

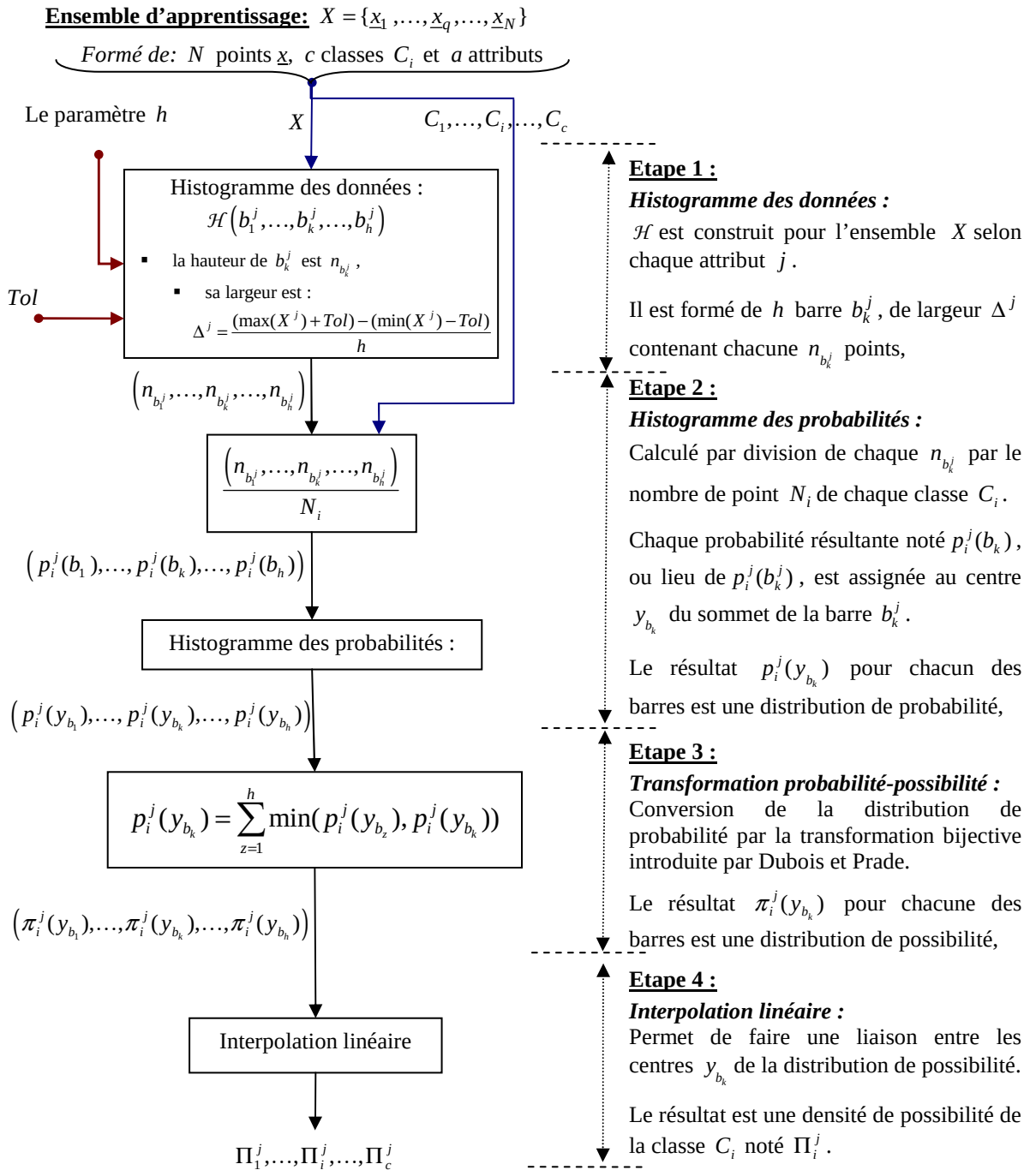


Figure 1.14 Phase d'apprentissage de FPM.

Les histogrammes des données sont établis à partir de l'ensemble X pour chaque classe C_i et par rapport à chaque attribut j. Le nombre h de barres b_k, $k \in \{1, 2, ..., h\}$ des

histogrammes est un paramètre de la méthode et il est déterminé expérimentalement [SAY02a]. Il est bien connu que le choix de h conditionne les performances de FPM de manière critique. Les bornes inférieures et supérieures de chaque histogramme sont généralement les valeurs minimales et maximales des coordonnées des points d'apprentissage suivant l'attribut j. Toutefois, l'ajout d'une marge de tolérance, Tol, déterminée manuellement par l'utilisateur ou par validation croisée permet d'élargir la base des histogrammes. L'intérêt est d'agrandir le champ d'action de FPM afin d'améliorer ses performances. La hauteur de chaque barre est l'effectif des points d'apprentissage, n_{b_k} dans la barre b_k. La division de n_{b_k} par N_i permet de calculer la probabilité $p_i^j(y_{b_k})$ assignée au centre y_{b_k} du sommet de la barre b_k. La distribution de probabilité $\left\{p_i^j(y_{b_k}),\ k \in \{1,2,...,h\}\right\}$ est convertie en distribution de possibilité $\left\{\pi_i^j(y_{b_k}),\ k \in \{1,2,...,h\}\right\}$ en utilisant une transformation de probabilité en possibilité. Nous avons choisi la transformation bijective introduite par Dubois et Prade [DUB93] en raison de ses bons résultats, dans les applications de RdF [SAY02b]. Cette transformation sera détaillée et évaluée par rapport à d'autres transformations dans la suite de ce chapitre. Elle est définie par :

$$\pi_i^j(y_{b_k}) = \sum_{z=1}^{h} \min(p_i^j(y_{b_z}), p_i^j(y_{b_k})) \tag{1.64}$$

Finalement, la densité de possibilité Π_i^j de l'attribut j pour la classe C_i est déduite par l'interpolation linéaire de la distribution de possibilité.

1.4.7.2.2. Phase de classification

La classification d'un nouveau point $\underline{x} \in \Re^a$, dont les valeurs pour les différents attributs sont $x^1,\ldots,x^a$, s'effectue, comme illustré par la figure 1.15, en trois étapes :

- détermination de la possibilité d'appartenance π_i^j de $\underline{x}$ à la classe C_i selon l'attribut j. Cette possibilité est calculée par simple projection de x^j sur la densité de possibilité Π_i^j,
- fusion, pour chaque classe C_i, de toutes les valeurs de possibilité d'appartenance $\pi_i^1,\ldots,\pi_i^j\ldots,\pi_i^a$ par un opérateur d'agrégation. Cet opérateur peut être un produit, un minimum, une moyenne, une intégrale floue [GRA94], ou encore l'opérateur Ordered Weighted Averaging (OWA) [YAG88] et son extension basée sur la combinaison de t-norme et l'opérateur OWA (TOWA) [YAG05]. Le résultat de cette fusion représente la possibilité d'appartenance π_i de $\underline{x}$ à la classe C_i. Nous avons choisi l'opérateur "min" comme opérateur d'agrégation. En effet, cet opérateur, dans le contexte d'aide à la décision, a un caractère fortement non tolérant et procure à FPM le caractère sélectif. La possibilité π_i est alors calculée par :
$$\pi_i = \min(\pi_i^1,\ldots,\pi_i^j,\ldots,\pi_i^a) \tag{1.65}$$
- affectation du point $\underline{x}$ à la classe pour laquelle il a la valeur de possibilité d'appartenance la plus élevée.

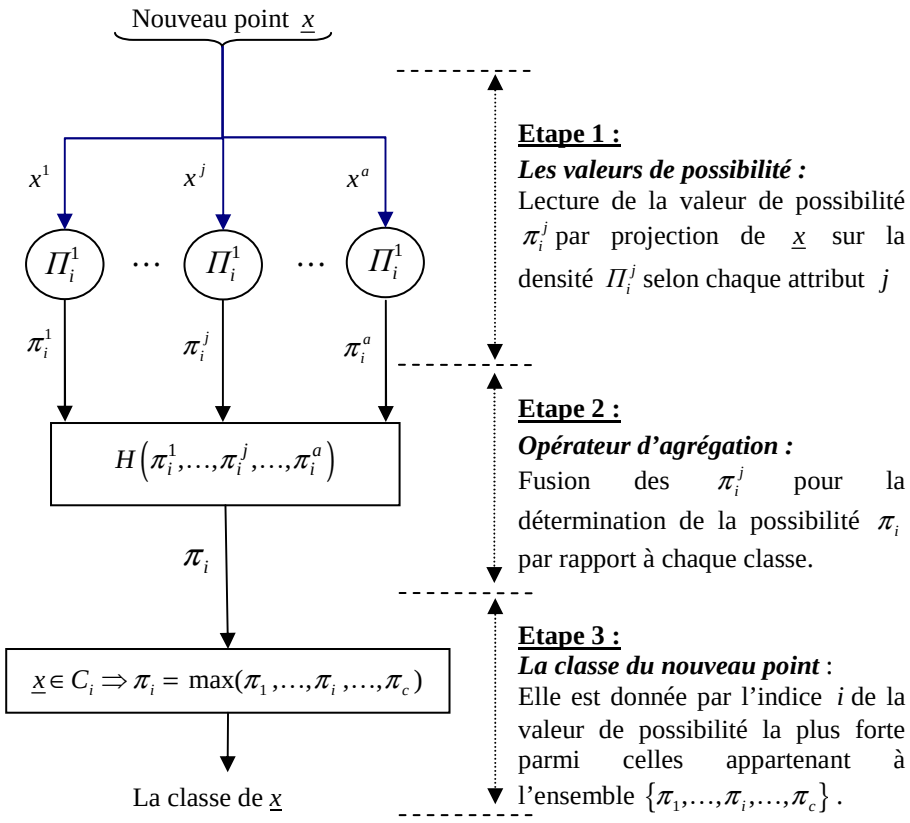


Figure 1.15 Phase de classification par FPM.

1.4.7.2.3. Exemple illustratif

Le polyéthylène téréphtalate PET, le polyéthylène haute densité PEHD et Polychlorure de Vinyle PVC sont les matières utilisées pour la fabrication des bouteilles. Ces bouteilles doivent être triées en vue de leur recyclage. La figure 1.16 présente le nuage correspondant à 120 valeurs de transmission pour deux longueurs d'onde infrarouge. Les 3 classes de 40 points correspondent aux trois polymères [DEV04].

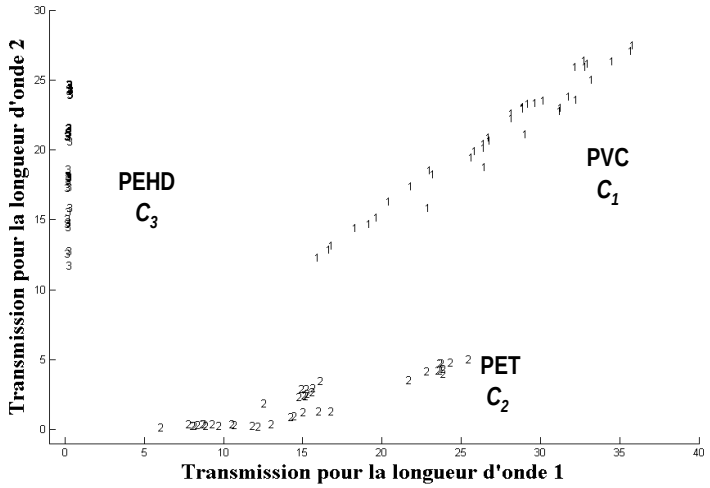


Figure 1.16 Espace de représentation des données de matières plastiques : PVC (C1), PET (C2) et PEHD (C3).

La phase d'apprentissage requiert le calcul de trois histogrammes de probabilité pour chaque attribut. Le paramètre h est fixé à 10. Cette valeur de h permet à FPM de séparer parfaitement les trois classes. Les histogrammes de probabilité, de possibilité et les densités relatives aux classes sont représentés respectivement dans les figures 1.17 et 1.18.

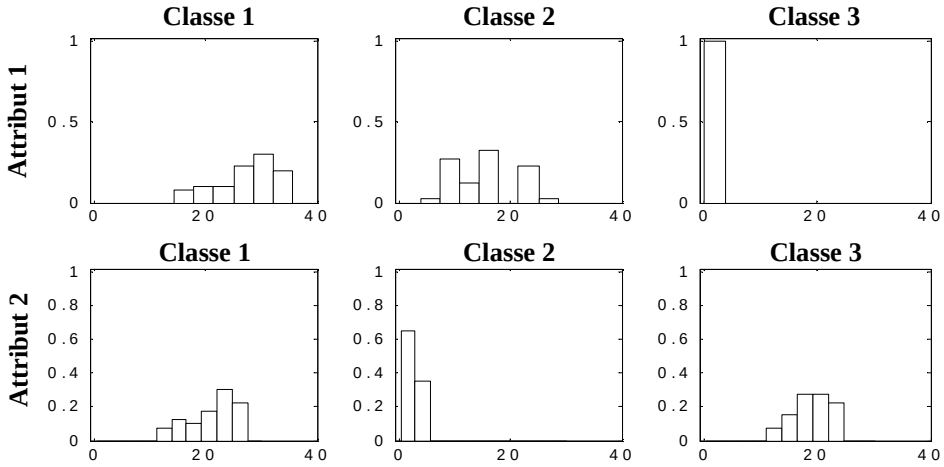


Figure 1.17 Histogrammes de probabilité pour les trois classes de matières plastiques.

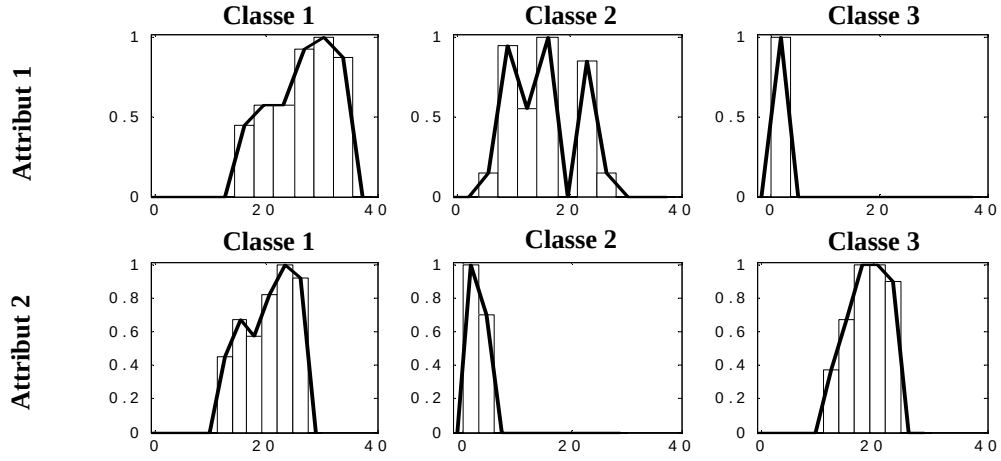


Figure 1.18 Histogrammes et densités de possibilité pour les trois classes de matières plastiques.

La phase de classification d'un nouveau point $\underline{x}$, correspondant à une matière plastique inconnue, est illustrée dans la figure 1.19.

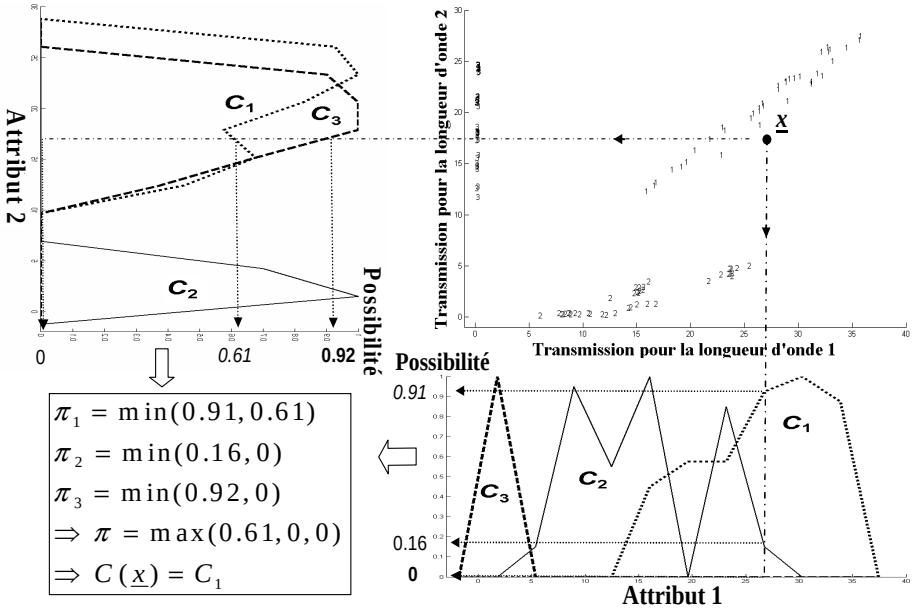


Figure 1.19 Classification d'un nouveau point $\underline{x}$ pour l'exemple de matières plastiques.

Après agrégation, on constate que la possibilité d'appartenance est de 0,61 pour la classe C_1 et 0 pour les deux autres. La nouvelle observation sera donc affectée à la classe C_1.

1.4.7.3. Méthodes de coalescences floues

1.4.7.3.1. Méthode Floue des C Moyennes (FCM)

Les premiers travaux concernant l'introduction d'une fonction d'appartenance dans les algorithmes de coalescence ont été réalisés par Ruspini [RUS69]. Le résultat est vu comme une partition floue sous forme d'une matrice définie par :

$$U = [u_i(\underline{x}_q)]_{(i=1,2,..,c\ ;\ q=1,2,..,N)} \tag{1.66}$$

avec la vérification des propriétés suivantes :

$$u_i(\underline{x}_q) \in [0,1] \quad ,\forall i,q, \tag{1.67}$$

$$\sum_{i=1}^{c} u_i(\underline{x}_q) = 1 \quad ,\forall q, \tag{1.68}$$

$$0 < \sum_{q=1}^{N} u_i(\underline{x}_q) < N \quad ,\forall i, \tag{1.69}$$

La méthode Floue des C Moyennes (FCM) est une extension directe de la méthode des Centres Mobiles (CM). Etudiée essentiellement par Bezdeck [BEZ81], cette technique de classification non supervisée a pour principe de base, la formation, à partir des individus non étiquetés, d'un nombre c de groupes. Ils doivent contenir les individus les plus semblables

possibles, tandis que les individus de groupes différents doivent être les plus dissemblables possibles. Ceci se traduit par la minimisation d'un critère des moindres carrés dont l'expression est la suivante :

$$J(U,G)=\sum_{q=1}^{N}\sum_{i=1}^{c}(u_{iq})^{m}.(d_{iq})^{2} \tag{1.70}$$

La matrice G est la matrice des coordonnées des centroïdes des classes. La distance d_{iq} représente la distance d'un point $\underline{x}_q$ au centroïde g_i. La distance Euclidienne est la plus utilisée, mais dans certain cas l'utilisation de la distance de Mahalanobis donne de meilleurs résultats. La variable m est le coefficient de fuzzyfication qui prend ses valeurs dans l'intervalle $[0,+\infty]$. Lorsque m vaut 1, on retrouve l'algorithme classique et lorsque m tend vers l'infini, on obtient une partition infiniment floue où les coefficients u_{iq} tendent vers 1/c. Les calculs du coefficient d'appartenance d'un nouveau point $\underline{x}_k$ à la classe C_i et du centroïde g_i de la classe C_i sont donnés par les expressions suivantes :

$$u_{ik}=\frac{1}{\sum_{z=1}^{c}\left(\frac{\left\|\underline{x}_k-g_i\right\|^2}{\left\|\underline{x}_k-g_z\right\|^2}\right)^{\frac{1}{m-1}}} \tag{1.71}$$

$$g_i=\frac{\sum_{q=1}^{N}(u_{iq})^{m}.x_q}{\sum_{q=1}^{N}(u_{iq})^{m}} \tag{1.72}$$

L'algorithme de FCM s'arrête lorsque la partition devient stable, c'est-à-dire lorsqu'elle n'évolue plus entre deux itérations successives :

$$\left\|U^{it}-U^{it+1}\right\|<\varepsilon \tag{1.73}$$

ε étant le seuil de convergence.

Il est prouvé que l'algorithme converge toujours mais il faut éviter les minima locaux, en ajustant judicieusement la valeur du seuil de convergence.

L'hypothèse probabiliste de l'algorithme de FCM impose que la somme des coefficients d'appartenance de chaque point doit être unitaire. Le coefficient d'appartenance maximal d'un point ne peut donc être inférieur à la valeur $1/c$. Or les seuils de rejets en appartenance sont souvent inférieurs à cette valeur. Par conséquent, un point éloigné de toutes les classes, tel que le point $\underline{x}_2$ de la figure 1.20, possède un coefficient d'appartenance maximal égal à 0,6 supérieur au seuil de rejet. Il est placé dans une des classes bien qu'il en soit éloigné. Le rejet en appartenance est donc inapplicable.

Lorsque les deux coefficients d'appartenance sont proches, le point est rejeté en ambiguïté. Si le nombre de classes n'est pas connu *a priori*, ou si ces classes ont des formes complexes, la partition obtenue n'est pas conforme à la répartition des points.

1.4.7.3.2. Méthode Possibiliste des C Moyennes (PCM)

Une solution au problème des points isolés a été proposée par Krishnapuram [KRI93]. Elle est basée sur une hypothèse possibiliste pour la définition de la partition. La condition d'orthogonalité, garantissant l'appartenance totale des points au nuage et traduisant l'hypothèse probabiliste n'existe plus. Les coefficients u_{iq} ne reflètent plus l'appartenance d'un point $\underline{x}_q$ à une classe C_i mais son degré de compatibilité. Les hypothèses deviennent :

$$\forall i,q: \ u_{iq} \in [0,1], \tag{1.74}$$

$$\max_{i=1,..,c}(u_{iq}) > 0, \ \forall q \tag{1.75}$$

$$0 < \sum_{q=1}^{N} u_{iq} < N \ , \ \forall i \tag{1.76}$$

Le nouveau critère à minimiser est :

$$J(U,G) = \sum_{q=1}^{N}\sum_{i=1}^{c}(u_{iq})^m.(d_{iq})^2 + \sum_{i=1}^{c}\eta_i.\sum_{q=1}^{N}(1-u_{iq})^m \tag{1.77}$$

Dans cette expression, le premier terme correspond au critère, de FCM, de l'équation (1.70). Le second terme impose des valeurs les plus grandes possibles aux possibilités d'appartenance. La valeur de η_i fixe la distance à partir de laquelle la possibilité d'appartenance est égale à 0,5. Krishnapuram propose de prendre une valeur proportionnelle à la distance moyenne intra-classe. L'algorithme est initialisé avec la matrice de partition obtenue par la méthode FCM. La valeur de η_i est déterminée par :

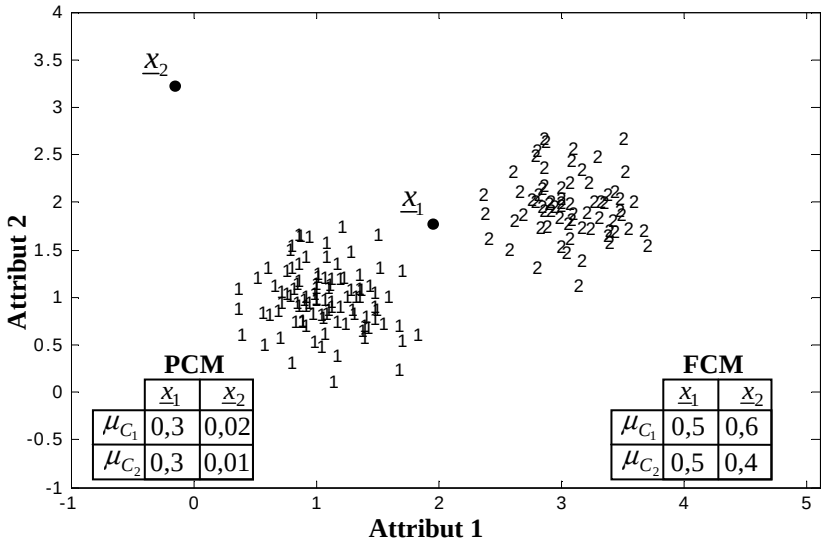
$$\eta_i = \frac{\sum_{q=1}^{N}(u_{iq})^m.(d_{iq})^2}{\sum_{q=1}^{N}(u_{iq})^m} \tag{1.78}$$

Les centroïdes et les possibilités d'appartenance sont calculés de manière itérative par les expressions :

$$g_i = \frac{\sum_{q=1}^{N}(u_{iq})^m.x_q}{\sum_{q=1}^{N}(u_{iq})^m} \tag{1.79}$$

$$u_{iq} = \left[1+\left(\frac{(d_{iq})^2}{\eta_i}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{1.80}$$

Le calcul se termine lorsque la différence entre chaque possibilité d'appartenance et la possibilité de l'itération précédente est inférieure à un seuil ε fixé.



PCM	$\underline{x}_1$	$\underline{x}_2$
μ_{C_1}	0,3	0,02
μ_{C_2}	0,3	0,01

FCM	$\underline{x}_1$	$\underline{x}_2$
μ_{C_1}	0,5	0,6
μ_{C_2}	0,5	0,4

Figure 1.20 La méthode PCM distingue l'éloignement du point $\underline{x}_2$, par rapport au point ambigu $\underline{x}_1$, contrairement à la méthode FCM.

La figure ci-dessus présente un nuage formé de deux classes et les valeurs d'appartenance pour les points $\underline{x}_1$ et $\underline{x}_2$, obtenues par FCM et PCM. Les résultats font bien apparaître que la somme des possibilités d'appartenance n'est pas unitaire dans le cas de PCM. Le point $\underline{x}_2$ possède des valeurs d'appartenance très faibles aux deux classes, cela traduit son éloignement. Un tel point peut donc être facilement rejeté. Par contre, pour FCM les deux points $\underline{x}_1$ et $\underline{x}_2$ sont considérés comme étant ambigus.

La méthode PCM permet de résoudre le problème des points éloignés des classes en leur associant une possibilité d'appartenance plus faible que le coefficient d'appartenance de nature probabiliste défini par la méthode FCM, rendant possible l'introduction des notions de rejet. Cependant, PCM ne solutionne ni le problème des classes de forme complexe, ni celui de la recherche du nombre de classes dans un nuage de points. Son utilisation demeure restreinte à la caractérisation de classes de formes hypersphériques et hyperelliptiques ayant une orientation parallèle aux axes. Dans le chapitre 2, nous verrons des extensions de FCM permettant de tenir compte de la forme et de l'orientation de chaque classe.

1.4.8. Lien entre les différentes théories

La théorie des fonctions de croyance englobe les mesures de probabilité, de possibilité et de nécessité. La fonction de croyance est dite probabiliste si les éléments focaux sont des singletons. La fonction de croyance est une mesure de nécessité et la fonction de plausibilité est une mesure de possibilité si les éléments focaux sont emboîtés comme le montre la figure ci-dessous [ZOU97].

Le choix d'une de ces théories dépend de l'application et principalement de sa complexité, de la connaissance disponible et des types d'imperfections. Cela constitue une motivation

pour des représentations hybrides ou complémentaires permettant de modéliser des informations hétérogènes dont les types d'imperfections sont différents. Le problème dans ce cas est de transformer toutes les informations dans un même modèle ou théorie de telle façon que les caractéristiques principales de ces informations soient préservées. Cette conversion est réalisée en utilisant des transformations permettant de passer d'une théorie à une autre. Quelques-unes de ces transformations seront détaillées par la suite.

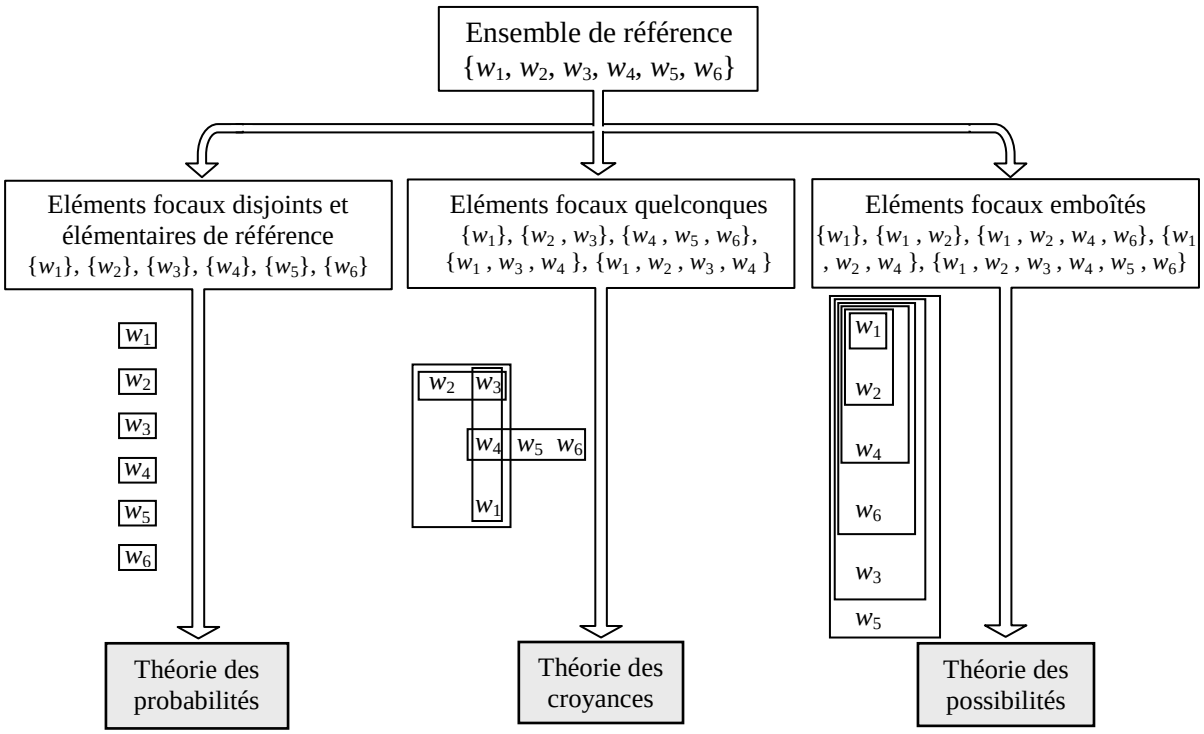


Figure 1.21 Lien entre la théorie de probabilités, la théorie de possibilités, et la théorie des fonctions de croyance.

1.5. Transformations probabilités-possibilités

Dans un système de diagnostic, les informations rencontrées sont à la fois imprécises et incertaines. Cela est dû aux capteurs et différents instruments de mesures, au bruit et aux experts humains. La théorie des possibilités ainsi que la théorie des fonctions de croyance sont bien adaptées pour la représentation et le traitement des connaissances à la fois imprécises et incertaines. Nous traitons des problèmes industriels dont les données sont modélisées par des distributions de probabilité construites d'après des histogrammes. Ces histogrammes ne sont pas exploitables par la théorie de fonctions de croyance. En revanche, il existe de nombreuses transformations permettant de passer des distributions de probabilité à des distributions de possibilité [DUB86; KIL90, LAS00, OUS00]. Chaque transformation doit garder un lien entre la mesure de probabilité et la mesure de possibilité permettant de conserver un certain ensemble de caractéristiques. Ce lien est appelé le principe de cohérence et il en existe plusieurs approches dans la littérature. Nous allons étudier la plus connue, à savoir le principe de cohérence de Dubois et Prade. Les transformations basées sur ce principe seront détaillées et évaluées selon des critères d'évaluation. Ces transformations donnent de bons résultats pour les applications de la RdF.

1.5.1. Condition de cohérence de Dubois et Prade

Le principe de cohérence proposé par Dubois et Prade n'est qu'une traduction formelle du fait que ce qui est probable devrait être possible. Ce principe est le suivant : soit p une distribution de probabilité dont P est la mesure. Toute distribution de possibilité π ou toute mesure de possibilité Π définie sur un domaine $\Omega = \{w_1, \ldots, w_i, \ldots, w_n\}$ satisfait:

$$\begin{cases} \forall A \subseteq \Omega,\ \Pi(A) \geq P(A) & \text{(Cas continu)} \\ \forall w_i \in A,\ \pi(w_i) \geq p(w_i) & \text{(Cas discret)} \end{cases} \tag{1.81}$$

Nous considérons pour la simplicité dans la suite de ce mémoire que $p(w_i) = p_i$ et $\pi(w_i) = \pi_i$

Toute distribution de possibilité satisfaisant la condition de cohérence peut être vue comme un majorant de la distribution de probabilité. Dans ce cas on dit que π domine *p*. De plus, la condition (1.81) s'appuie sur le fait que la représentation possibiliste est moins informative que la représentation probabiliste. En effet, pour le modèle probabiliste, l'incertitude est représentée par une seule valeur, tandis que pour le modèle possibiliste l'incertitude et l'imprécision sont représentées simultanément par un intervalle [DUB93].

Lors d'une transformation de probabilité en possibilité ($p \to \pi$), on perd de l'information, car d'après (1.81), la même distribution de possibilité peut être représentée par plusieurs distributions de probabilités. Alors que pour une transformation de possibilité en probabilité ($\pi \to p$), on ajoute arbitrairement de l'information.

Pour une transformation ($\pi \to p$) donnée, le choix de l'une des distributions de probabilités satisfaisant (1.81) peut se faire à l'aide du principe de la raison insuffisante qui consiste à choisir la distribution de probabilité qui contient le minimum d'informations ou le maximum d'incertitude [DUB93].

1.5.2. Transformations probabilité-possibilité de Dubois et Prade

Les deux transformations $p \to \pi$ et $\pi \to p$ sont définies respectivement par : [DUB93]

$$\pi_i = \sum_{j=i}^{n} \{ p_j / p_j \leq p_i \},\ \forall i = 1, \ldots, n \tag{1.82}$$

$$p_i = \sum_{j=i}^{n} \frac{(\pi_j - \pi_{j+1})}{j} \tag{1.83}$$

Ces deux transformations ne sont pas réciproques. (1.82) est appelée la transformation $p \to \pi$ asymétrique (ou optimale) de Dubois et Prade.

Dubois et Prade ont défini une autre transformation $p \to \pi$ symétrique satisfaisant le principe de cohérence de Dubois et Prade définie par :

$$\pi_i = \sum_{j=1}^{n} \min(p_i, p_j) \tag{1.84}$$

Ainsi, à l'inverse de la transformation (1.82) qui utilise une "somme conditionnelle" et qui présente donc une certaine complexité de calcul, la transformation (1.84) utilise seulement les opérateurs "min" et "somme". Dans cette transformation, on abandonne le fait d'additionner des degrés de possibilités, on les compare seulement [LAS99].

Exemple 1.12

Soit la distribution de probabilité $p = (p_1 = 0,3,\ p_2 = 0,2,\ p_3 = 0,1,\ p_4 = 0,4)$, la distribution de possibilité correspondante selon la transformation asymétrique de Dubois et Prade est : $\pi_{asy} = (\pi_1 = 0,6,\ \pi_2 = 0,3,\ \pi_3 = 0,1,\ \pi_4 = 1)$. Par contre la distribution de possibilité symétrique de Dubois et Prade est $\pi_{sym} = (\pi_1 = 0,9,\ \pi_2 = 0,7,\ \pi_3 = 0,4,\ \pi_4 = 1)$.

1.5.3. Transformation Variable de probabilité en possibilité (TV)

Soit la distribution de probabilité $p = (p_1, ..., p_i, ..., p_n)$ ordonnée comme suit : $p_1 > p_2 > \cdots > p_n$. La distribution de possibilité $\pi = (\pi_1, ..., \pi_i, ..., \pi_n)$ selon la Transformation $p \to \pi$ proposée dans [SAY06] est $\pi_1 > \pi_2 > \cdots > \pi_n$. Chaque possibilité est définie par :

$$\pi_i = \left(\frac{p_i}{p_1} \right)^{k.(1-p_i)}, \ \forall i = 1, 2, \ldots, n \tag{1.85}$$

Cette transformation est appelée la Transformation Variable (TV) car elle permet d'ajuster la spécificité via un paramètre variable k. Afin de vérifier la condition de cohérence de Dubois et Prade définie par (1.81) dans le cas discret, il faut que toutes les inégalités suivantes soient vérifiées :

$$\pi_i = \left(\frac{p_i}{p_1} \right)^{k.(1-p_i)} \geq p_i, \ \forall i = 1, 2, \ldots, n \tag{1.86}$$

Le paramètre *k* doit donc vérifier la condition définie par [SAY06] :

$$0 \leq k \leq \frac{\log(p_n)}{(1-p_n).\log \frac{p_n}{p_1}} \tag{1.87}$$

Pour que la TV soit plus spécifique que la transformation optimale du Dubois et Prade il faut choisir $k = k_{\max} = \max(k_i)$ tel que :

$$k = \max \left(\frac{\log(p_i)}{(1-p_i).\log \frac{p_i}{p_1}} \right), \ \forall i = 1, 2, \ldots, n \tag{1.88}$$

Or dans ce cas la TV viole la condition de cohérence (1.81).

Nous proposons d'améliorer la TV afin qu'elle soit aussi spécifique que la transformation optimale de Dubois et Prade. Cela, tout en vérifiant la condition de cohérence (1.81). Pour cela, au lieu de définir un seul paramètre k, comme dans (1.85), un paramètre k_i est défini pour chaque π_i. L'équation (1.85) s'écrira alors comme suit :

$$\pi_i = \left(\frac{p_i}{p_1}\right)^{k_i(1-p_i)}, \quad \forall\, i = 1,2,\ldots,n \tag{1.89}$$

La possibilité correspondant à la probabilité maximale p_1 est π_1, elle est égale à 1 $\forall k_1 \in IR$. Pour simplifier, nous imposons $k_1 = 1$. Afin de vérifier (1.81) dans le cas continu, il faut donc que toutes les inégalités suivantes soient vérifiées :

$$\pi_i = \left(\frac{p_i}{p_1}\right)^{k_i(1-p_i)} \geq p_i + p_{i+1} + \ldots + p_n, \quad \forall\, i = 2,3,\ldots,n \tag{1.90}$$

Cela signifie que le paramètre k_i doit vérifier la condition suivante :

$$0 \leq k_i \leq \frac{\log(p_i + p_{i+1} + .. + p_n)}{(1-p_i).\log\frac{p_i}{p_1}}, \quad \forall\, i = 2,3,\ldots\, n \tag{1.91}$$

Pour que la TV soit aussi spécifique que la transformation optimale de Dubois et Prade tout en vérifiant la condition de cohérence (1.81), il faut choisir k_i tel que :

$$k_i = \frac{\log(p_i + p_{i+1} + .. + p_n)}{(1-p_i).\log\frac{p_i}{p_1}}, \quad \forall\, i = 2,3,\ldots,n \tag{1.92}$$

Exemple 1.13

Soit la distribution de probabilité $p' = (p'_1 = 0,3,\ p'_2 = 0,2,\ p'_3 = 0,1,\ p'_4 = 0,4)$. Une fois ordonnée de façon descendante : $p = (p_1 = 0,4,\ p_2 = 0,3,\ p_3 = 0,2,\ p_4 = 0,1)$, la distribution de possibilité correspondante selon la TV, proposée dans [SAY06], et calculée en utilisant (1.89) est donnée par $\pi_{TV} = (\pi_1 = 1,\ \pi_2 = 0,69,\ \pi_3 = 0,36,\ \pi_4 = 0,1)$ avec $k = 1,84$. Après avoir été remises dans l'ordre de départ: $\pi_{TV} = (\pi'_1 = 0,69,\ \pi'_2 = 0,36,\ \pi'_3 = 0,1,\ \pi'_4 = 1)$. La valeur $k = 1,84$ est la valeur maximale pour laquelle la condition de cohérence (1.81) est vérifiée que ce soit dans le cas continu ou discret. En effet, dans le cas discret on a $(\forall i \in \{1,\ldots,4\},\ p_i < \pi_i)$ qui sont vérifiées. Dans le cas continu, pour que cette condition soit respectée, il faut que toutes les inégalités (1.90) soient vérifiées. C'est-à-dire : $p_1 + p_2 + p_3 + p_4 \leq \pi_1$, $p_2 + p_3 + p_4 \leq \pi_2$, $p_3 + p_4 \leq \pi_3$ et $p_4 \leq \pi_4$.

Bien que cette transformation respecte la condition de cohérence (1.81), elle n'est pas plus spécifique que la transformation optimale $\pi_{asy} = (\pi'_1 = 0,6,\ \pi'_2 = 0,3,\ \pi'_3 = 0,1,\ \pi'_4 = 1)$ puisque $\pi_{asy} < \pi_{TV}$. Pour que la TV soit plus spécifique que la transformation optimale, il faut choisir $k = k_{\max} = \max(k_i)$. En utilisant l'équation (1.88) on aura $k_{\max}$ = max(5,98 , 2,90 , 1,84

,1) = 5,98. Dans ce cas $\pi_{TV} = (\pi'_1 = 0,3,\ \pi'_2 = 0,036,\ \pi'_3 = 0,\ \pi'_4 = 1)$ et la condition de cohérence (1.81) n'est plus vérifiée. La distribution de possibilité calculée en utilisant (1.89) proposée dans cette thèse est : $\pi_{TV'} = (\pi'_1 = 0,6,\ \pi'_2 = 0,3,\ \pi'_3 = 0,1,\ \pi'_4 = 1)$ avec $k_1 = 2,54,\ k_2 = 2,17,\ k_3 = 1,845,\ k_4 = 1$. Grâce à ces valeurs de *k*, la spécificité de la TV devienne équivalente à celle de la transformation optimale de Dubois et Prade et respecte ainsi la condition de cohérence (1.81).

1.5.4. Evaluation des transformations

Pour comparer ces deux transformations, nous allons utiliser les critères d'évaluation les plus courants. Ces critères définissent les propriétés et les qualités que peut posséder une transformation. Le choix de la transformation dépend du contexte du problème et donc des propriétés à vérifier. Dans la littérature il existe plusieurs critères, nous allons développer ceux qui sont intéressants pour les applications de la RdF.

1.5.4.1. Critère de normalisation

Ce critère vérifie que la distribution obtenue lors de l'utilisation d'une transformation est normalisée quelle que soit la distribution d'origine. Les distributions de possibilité π ou de probabilité p sont dites normalisées si :

$$\exists\, w \in \Omega \text{ tel que } \pi_v(w) = 1 \tag{1.93}$$

$$\sum_{i=1}^{n} p_v(w_i) = 1 \tag{1.94}$$

Ces conditions signifient que l'univers Ω contient toutes les valeurs de l'attribut ou de la variable v. On parle alors de monde fermé. Dans le cas des données incomplètes, l'hypothèse de monde fermé n'est pas satisfaite, on est dans le cas du monde ouvert.

Les transformations, symétrique et asymétrique, de Dubois et Prade ne vérifient pas ce critère. En conséquence, dans le cas d'une distribution de probabilité sous-normalisée, $\sum_{i=1}^{n} p_v(w_i) < 1$, ces deux transformations donnent une distribution de possibilité sous normalisée. Par contre, la TV améliorée vérifie la condition de normalisation définie par (1.94), car en prenant $i = 1$ dans l'équation (1.89), nous trouvons : $\pi_{\max} = \pi_1 = 1$.

1.5.4.2. Critère du maximum de spécificité

Ce critère cherche, pour une distribution de probabilité donnée, et parmi toutes les distributions de possibilité celle qui maximise l'information ajoutée. La transformation la plus spécifique est celle pour laquelle la distribution de possibilité vérifie :

$$\exists\, u \in \Omega :\ \pi(u) = 1, \quad \forall\ w \neq u \in \Omega :\ \pi(w) = 0 \tag{1.95}$$

La mesure de spécificité est évaluée par le degré pour lequel une distribution de possibilité π est proche de la distribution la plus spécifique définie par (1.95). On dit qu'une théorie

d'incertitude capture l'imprécision si elle permet de comparer les différentes distributions selon leur spécificité [SAN91]. La théorie des probabilités ne capture pas l'imprécision du fait que la condition de normalisation conduit à des distributions ayant des spécificités égales dans le sens de (1.95).

Soient π^1 et π^2 , deux distributions de possibilité, on dit que π^1 est plus spécifique que π^2 si et seulement si [DUB93, LAS99] :

$$\forall\ w \in \Omega,\ \pi^2(w) < \pi^1(w) \tag{1.96}$$

Cette inégalité est appelée le principe de maximum de spécificité. On peut déduire de (1.96) qu'entre deux distributions de possibilité, celle qui est la plus spécifique est celle qui est la plus petite dans le sens de la cardinalité de l'ensemble flou, c'est-à-dire celle qui minimise $\sum_{i=1}^{n} \pi(\omega_i)$. Le degré de spécificité peut donc être exprimé par :

$$S_{opt} = \sum_{i=1}^{n} \pi(w_i) \tag{1.97}$$

Le degré de spécificité de la distribution de possibilité qui est la plus spécifique et qui vérifie le principe de cohérence de Dubois et Prade, défini par (1.81) :

$$S_{\max} = 1 + \sum_{j=2}^{n} \left(\sum_{i=j}^{n} p_i \right) \tag{1.98}$$

La transformation optimale de Dubois et Prade, définie par (1.82) , est la transformation la plus spécifique qui vérifie (1.98) [DUB93].

La spécificité de la TV améliorée est équivalente à la transformation optimale de Dubois et Prade pour des valeurs de k_i calculées en utilisant (1.92) afin de respecter la condition de cohérence (1.81).

Exemple 1.14

Soit $p = (p_1 = 0,4,\ p_2 = 0,3,\ p_3 = 0,2,\ p_4 = 0,1)$. La distribution de possibilité selon la transformation optimale de Dubois et Prade, définie par (1.82), est : $\pi_{opt} = (\pi_1 = 1,\ \pi_2 = 0,6,\ \pi_3 = 0,3,\ \pi_4 = 0,1)$. La spécificité de cette distribution, calculée en utilisant (1.97), est $S_{opt} = 2$. La spécificité maximale, définie par (1.98), de cette distribution est égale à $S_{\max} = (1 + (0,6) + (0,3) + (0,1)) = 2$. C'est pourquoi la spécificité de la transformation optimale est toujours maximale.

Trouvons maintenant, la distribution de possibilité selon la TV améliorée. D'abord il faut calculer les paramètres k_i en utilisant (1.92) : $k_1 = 1,\ k_2 = 2,54,\ k_3 = 2,17,\ k_4 = 1,84$. Avec ces paramètres, la TV améliorée donne une distribution de possibilité équivalente à celle optimale.

Exemple 1.15

Afin de constater l'intérêt de la spécificité sur la méthode FPM, nous allons utiliser la base de données issue du simulateur TEP (Tennesse Eastman Process) [DOW93] développé par la

société Eastman chemical company. TEP comporte 12 variables d'entrée et 41 variables de sortie. De plus il est idéal pour étudier les performances des méthodes de contrôle et de surveillance, car la base de données peut répertorier 20 modes de défauts qui sont identifiés. Dans [CHI04, KUL05, VER06], les auteurs s'intéressent uniquement à trois modes défaillants identifiés comme étant les numéros 4, 9 et 11. Ces modes sont représentés par 3 classes qui sont très chevauchées et donc difficilement discernables. La base de données, qu'ils ont utilisée, est divisée en deux parties : l'ensemble d'apprentissage contenant 480 points et l'ensemble de test contenant 800 points. Pour notre application nous nous sommes intéressés uniquement à l'espace de représentation formé par les variables 9 et 51 pour leur représentativité comme le montre la figure 1.22.

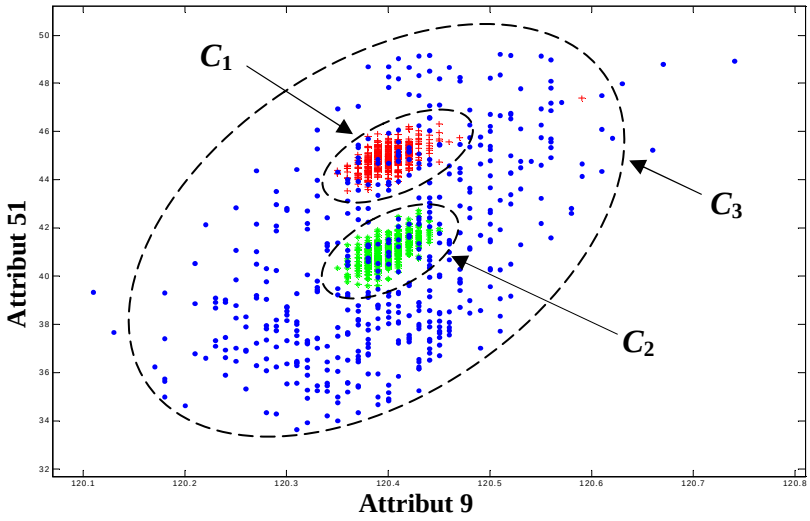


Figure 1.22 Représentation des Classe 1 (+), Classe 2 (*) et Classe 3 (.) pour les données Tennessee Eastman Process.

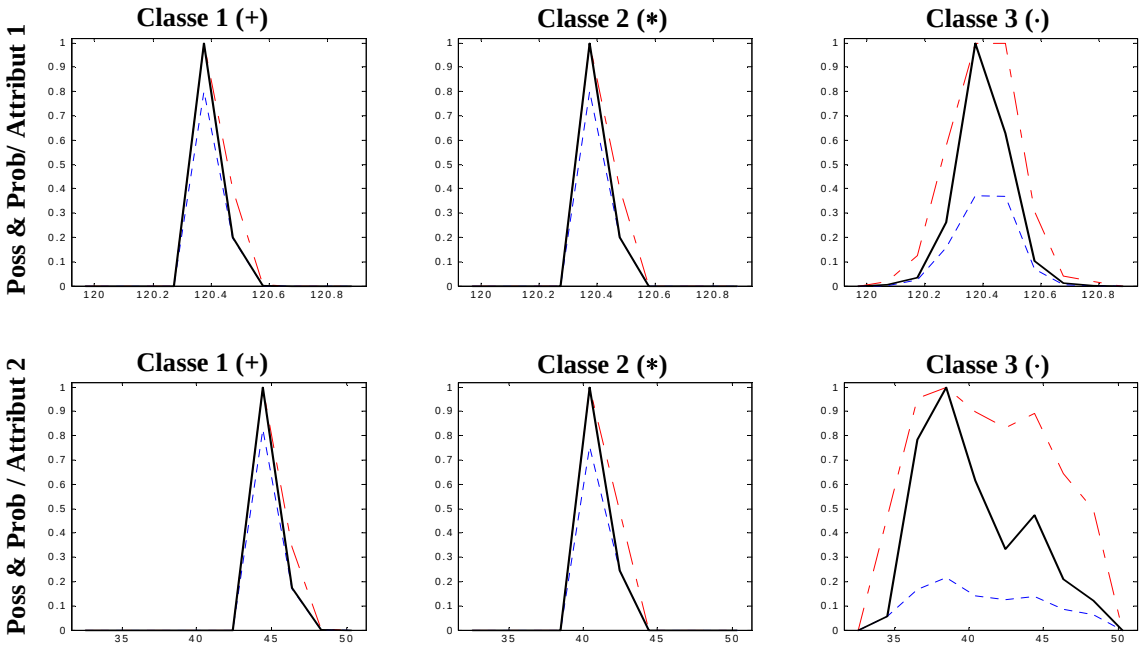


Figure 1.23 (---) probabilité, (-.-) possibilité résultant de la transformation non optimale de Dubois et Prade ; (-) possibilité résultant de la TV améliorée.

Comme on peut le constater sur la figure 1.23, la classe 3 qui recouvre les deux autres, d'après la figure 1.22, nécessite une représentation plus spécifique pour pouvoir la discriminer. Cela se justifie par les Taux d'Erreur de Classification (TEC) obtenus par :

- FPM utilisant la TV améliorée : 8,29%,
- FPM utilisant la transformation non optimale de Dubois et Prade : 53,96%.

Les paramètres de FPM utilisés sont $Tol = 0,09$ et $h = 8$. Ces derniers sont déterminés expérimentalement en utilisant la technique du "leave-one-out", afin d'estimer la borne supérieure du TEC.

1.5.4.3. Critère de conservation de la forme

Ce critère permet de garantir que la distribution de possibilité obtenue a la même forme que la distribution de probabilité de départ. La vérification du critère nécessite la conservation forte de la préférence. Il existe deux types de conservation de la préférence : la conservation faible définie par (1.99) et la conservation forte définie par (1.100) :

$$\forall w,u \in \Omega, \quad p(w) > p(u) \quad \Leftrightarrow \quad \pi(w) > \pi(u) \tag{1.99}$$

$$\begin{cases} \forall\ w, u \in \Omega : \\ p(w) = p(u) \Leftrightarrow \pi(w) = \pi(u), \\ p(w) > p(u) \Leftrightarrow \pi(w) > \pi(u), \\ p(w) < p(u) \Leftrightarrow \pi(w) < \pi(u) \end{cases} \tag{1.100}$$

La préservation de la préférence forte conserve l'ordre des éléments d'une distribution lors de la transformation. Cela veut dire que l'élément le plus probable devient l'élément le plus possible et dans l'autre sens, l'élément le plus possible devient le plus probable. En même temps les éléments qui ont une probabilité égale doivent avoir une possibilité égale ou vice-versa.

La transformation optimale de Dubois et Prade, définie par (1.82), conserve la forme seulement dans le cas où les probabilités sont toutes différentes. Masson [MAS06] a proposé une solution en trois étapes pour le cas où deux probabilités ou plus sont égales. D'abord, il faut construire toutes les permutations possibles de toutes les probabilités égales. Ensuite il faut calculer la possibilité pour chaque permutation. Enfin, les possibilités maximales entre toutes les permutations sont conservées afin de garantir la condition de cohérence de Dubois et Prade. Cette conservation se fera au détriment de la spécificité maximale.

Pour la TV améliorée, la conservation de la forme est plus simple. Si deux probabilités p_i, p_{i+1} sont égales, alors il faut juste imposer que $k_i = k_{i+1}$.

Exemple 1.16

Soit $p = (p_1 = 0,5,\ p_2 = 0,2,\ p_3 = 0,2,\ p_4 = 0,1)$. D'après Masson [MAS06], nous pouvons avoir deux permutations possibles : $p^1 = (p_1 = 0,5,\ p_2 = 0,2,\ p_3 = 0,2,\ p_4 = 0,1)$ et $p^2 = (p_1 = 0,5,\ p_3 = 0,2,\ p_2 = 0,2,\ p_4 = 0,1)$. Les deux distributions de possibilités correspondantes selon la transformation optimale sont : $\pi^1 = (\pi_1 = 1,\ \pi_2 = 0,5,\ \pi_3 = 0,3,\ \pi_4 = 0,1)$ et $\pi^2 = (\pi_1 = 1,\ \pi_3 = 0,5,\ \pi_2 = 0,3,\ \pi_4 = 0,1)$. La distribution finale est obtenue en prenant les valeurs maximales entre les deux distributions : $\pi_{opt} = (\pi_1 = 1,\ \pi_2 = 0,5,\ \pi_3 = 0,5,\ \pi_4 = 0,1)$.

En utilisant la TV améliorée, pour l'exemple ci-dessus, on a $k_1 = 1,\ k_2 = 0,94,\ k_3 = k_2,\ k_4 = 1,59$. Ces paramètres conduisent à l'obtention de la distribution de possibilité suivante $\pi_{TV'} = (\pi_1 = 1,\ \pi_2 = 0,5,\ \pi_3 = 0,5,\ \pi_4 = 0,1)$. Cette simplicité se voit plus clairement dans le cas où il y a plusieurs probabilités égales, comme dans le cas de l'ignorance totale où toutes les probabilités sont égales.

Exemple 1.17

Cas de connexité des classes :

Soit un ensemble d'apprentissage X contenant une seule classe C_1, celle-ci est composée de 4 sous-classes identiques et non connexes, cf. figure 1.25. Chacune de ces sous-classes est générée selon une même loi gaussienne dans $\Re^2$.

L'estimation de la densité de probabilité de C_1 par rapport à chaque attribut est réalisée avec un histogramme de pas h. Cela nous permet de calculer selon la transformation choisie la densité de possibilité de C_1. La figure 1.24 montre les possibilités obtenues en utilisant $h = 5$ puis $h = 20$, par la transformations de Dubois et Prade non optimale et par la TV améliorée.

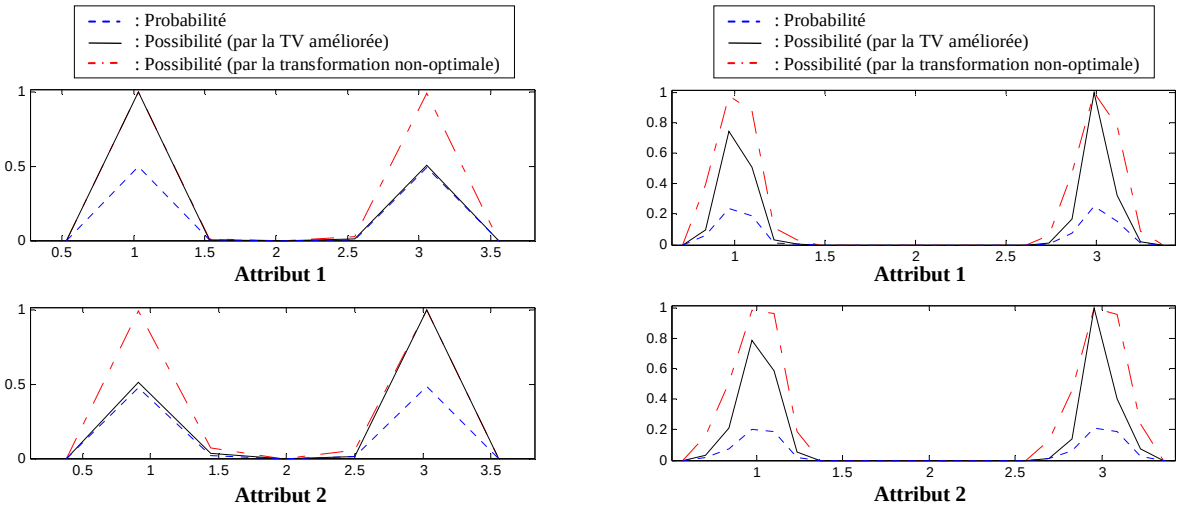


Figure 1.24 Comparaison des densités de possibilités obtenues par la transformation non optimale de Dubois et Prade et par la TV améliorée, représentées à droite pour h=5 et à gauche pour h=20 dans le cas d'une classe connexe.

Comme on peut le constater, la forme de la densité de possibilité obtenue par la transformation non optimale conserve la forme de la densité de probabilité par rapport à chaque sous-classe ; ce n'est pas le cas pour la densité de possibilité obtenue par la TV améliorée. De plus, d'après les figures 1.25 et 1.26, l'intérêt de la conservation de la forme permet une représentation plus réaliste de la distribution des données dans chaque sous classe de C_1. En effet, les courbes de niveaux d'appartenances construites à partir de la transformation non optimale représentent un niveau plus élevé aux centres de chaque sous classes de C_1 quel que soit h, cf. figures 1.25 et 1.26 représentées à gauche. Par contre, la TV améliorée est très sensible au paramètre h. Par exemple, pour une petite valeur de h, $h = 5$, cette transformation n'attribue le niveau le plus élevé, des courbes d'appartenances, qu'à une

seule sous classe de C_1, cf. figure 1.25 à droite. Cela engendre une perte d'informations sur la vraie densité des autres sous-classes. Pour une grande valeur de h, $h = 20$, les courbes de niveaux d'appartenances sont comme celles obtenues par la transformation non optimale à la différence qu'elles se concentrent fortement aux centres des sous classes et ne couvrent pas leur périphérie, cf. figure 1.26 à droite.

Bien que la TV améliorée soit la plus spécifique, son utilisation dans le cas des classes non connexes en RdF n'est pas adaptée.

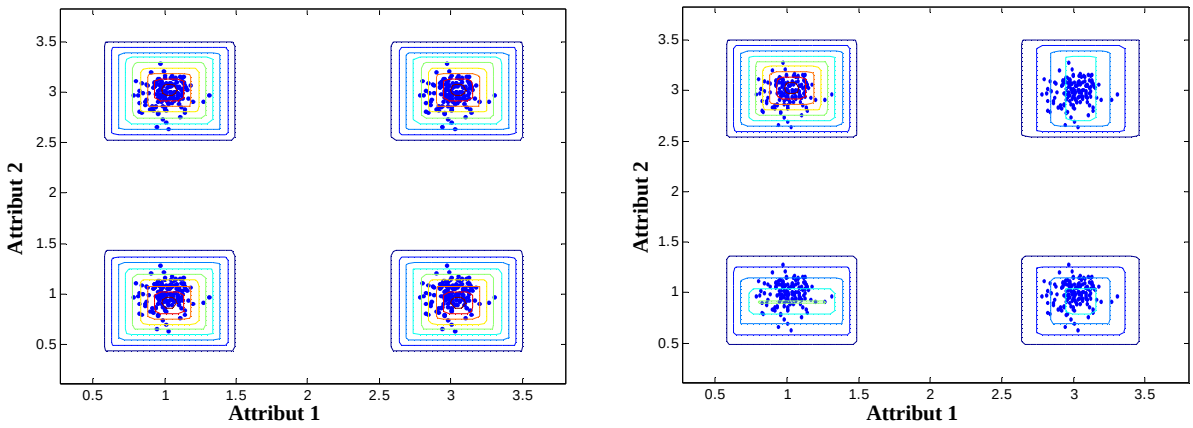


Figure 1.25 Comparaison des courbes de niveaux d'appartenances obtenues par la transformation non optimale du Dubois et Prade à gauche et par la TV améliorée à droite, avec h=5 dans le cas d'une classe connexe.

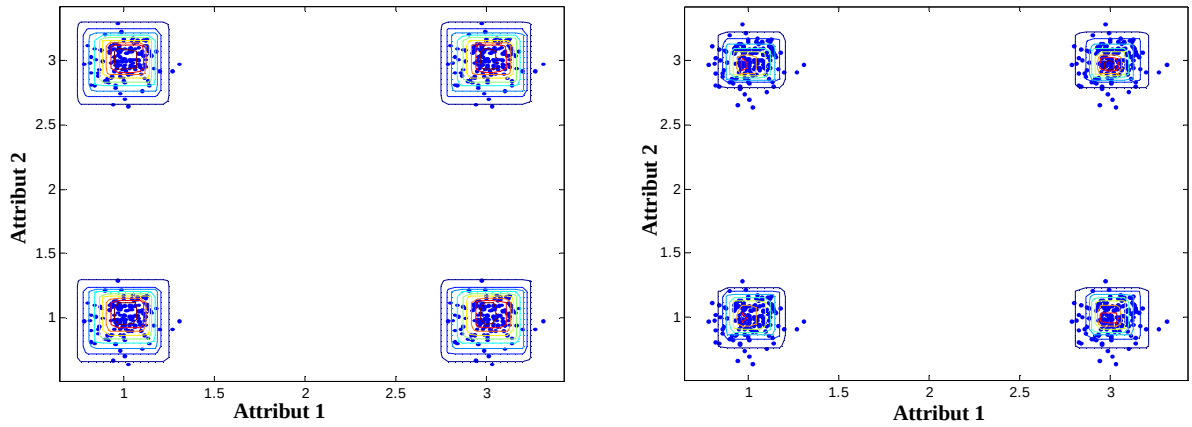


Figure 1.26 Comparaison des courbes de niveaux d'appartenances obtenues par la transformation non optimale du Dubois et Prade à gauche et par la TV améliorée à droite, avec h=20 dans le cas d'une classe connexe.

Exemple 1.18

Cas d'une distribution uniforme :

Soit un ensemble d'apprentissage X contenant une seule classe C_1 générée selon une loi gaussienne dans $\Re^2$ représentée dans la figure 1.28. L'estimation de la densité de probabilité de C_1 par rapport à chaque attribut est réalisée avec un histogramme de paramètre $h = 20$. Cela nous permet de calculer selon la transformation choisie les densités de possibilité de C_1, représentées figure 1.27. Comme on peut le constater, la forme de la densité de possibilité

obtenue par la transformation non optimale est conservée par rapport à la forme de la densité de probabilité. Ce qui n'est pas tout à fait le cas pour la densité de possibilité obtenue par la TV améliorée.

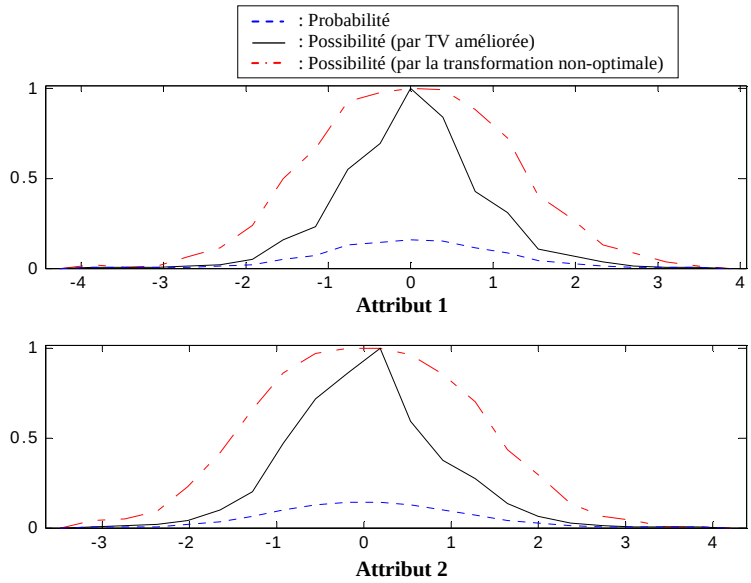


Figure 1.27 Comparaison des densités de possibilités obtenues par la transformation non optimale de Dubois et Prade et par la TV améliorée avec h=20.

De plus, d'après la figure 1.28, les courbes de niveaux d'appartenance obtenues par la transformation non optimale couvrent entièrement le nuage de points de la classe C_1. Tandis que celles obtenues par la TV améliorée se concentrent fortement au centre de la classe et ne couvrent pas la périphérie. Encore une fois, l'intérêt de la conservation de la forme permet une représentation plus réaliste de la distribution des données dans C_1. Bien que la TV améliorée soit la plus spécifique, son utilisation dans le cas des classes uniformes en RdF n'est pas adaptée.

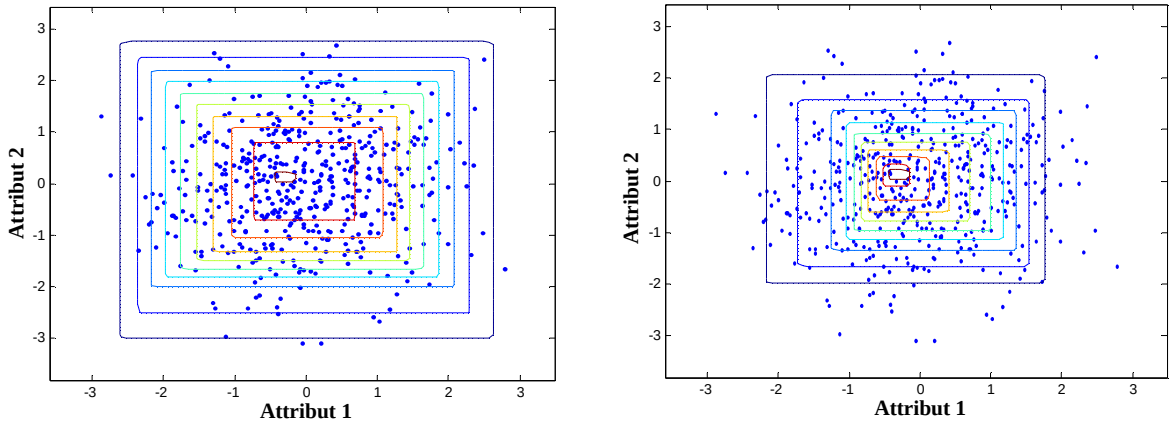


Figure 1.28 Comparaison des courbes de niveaux d'appartenance obtenues par la transformation non optimale du Dubois et Prade à gauche et par la TV améliorée à droite), avec h=20 pour le cas d'une distribution uniforme.

1.5.4.4. Critère de conservation de l'ignorance

Ce critère permet de garantir la conservation du cas de l'ignorance totale lors de la transformation. Le cas de l'ignorance totale est indésirable parce qu'il correspond au cas où on ne peut pas prendre de décision.

D'après la solution proposée par Masson [MAS06], présentée dans la section précédente, pour une distribution de probabilité contenant n éléments égaux, il faut calculer $n!$ permutations ; ensuite les transformer en possibilité en utilisant la transformation optimale de Dubois et Prade ; enfin, garder les valeurs maximales de possibilités entre toutes ses distributions. La TV améliorée, quant à elle, ne nécessite aucune permutation ou aucun calcul supplémentaire. A chaque fois qu'on rencontre deux probabilités égales p_i, p_{i+1}, il suffit de garder la même valeur du paramètre k calculé auparavant $k_i = k_{i+1}$.

1.6. Combinaison des sources d'information

Dans le domaine des systèmes d'aide à la décision, les informations manipulées sont souvent imparfaites. Ces imperfections se manifestent sous de multiples formes : incertitude, imprécision, incomplétude, ambiguïté, et conflit [DUB88]. La fusion est donc utile pour lever les ambiguïtés d'une source grâce aux informations apportées par les autres sources ou par les connaissances supplémentaires. Dans ce qui suit, le problème de la fusion de l'information provenant de plusieurs sources ainsi que les modes de combinaisons les plus connus de la littérature seront abordés.

1.6.1. Modes de combinaison

L'agrégation floue intervient dans le domaine de la décision multi-critère. Elle permet de combiner les données fournies par des capteurs ou des sources d'informations. Cette combinaison doit être adaptée à la situation conflictuelle ou non des données. Il existe quatre grands modes d'agrégation qui sont la conjonction ou s-norme, le compromis, la disjonction ou t-norme et des opérateurs prenant en compte des mesures de conflit ou encore de fiabilité des sources.

1.6.1.1. t-normes

On appelle "t-norme", toute fonction t vérifiant les contraintes suivantes [ZIM91] :

$$t:\left[\begin{array}{c}[0,1]\times[0,1]\to[0,1]\\ (x,y)\to t(x,y)\end{array}\right] \tag{1.101}$$

$$t(0,0)=0,\quad t(x,1)=t(1,x)=x \tag{1.102}$$

$$\text{Monotonie : si } x_1\le x_2 \text{ et } y_1\le y_2 \text{ : } \quad t(x_1,y_1)\le t(x_2,y_2) \tag{1.103}$$

$$\text{Commutativité : } t(x,y)=t(y,x) \tag{1.104}$$

$$\text{Associativité : } t(x,t(y,z))=t(t(x,y),z) \tag{1.105}$$

Les t-normes définissent, pour les ensembles flous, une classe générale d'opérateurs d'intersection. Quelques exemples de t-normes sont présentés dans le tableau 1.1. Ces opérateurs ont un caractère non tolérant, cela veut dire que le résultat de l'agrégation dépend fortement de la plus faible des valeurs à agréger. Une propriété importante est que l'opérateur "min" est une borne supérieure pour toutes les t-normes, autrement dit le plus tolérant des t-normes.

1.6.1.2. s-normes

On appelle " s-normes ", toute fonction *s* vérifiant les contraintes suivantes [ZIM91] :

$$s:\left[\begin{array}{c}[0,1]\times[0,1]\to[0,1]\\ (x,y)\to s(x,y)\end{array}\right] \tag{1.106}$$

$$s(1,1)=1,\quad s(x,0)=s(0,x)=x \tag{1.107}$$

$$\text{Monotonie : si } x_1\le x_2 \text{ et } y_1\le y_2 \text{ :} \quad s(x_1,y_1)\le s(x_2,y_2) \tag{1.108}$$

$$\text{Commutativité : } s(x,y)=s(y,x) \tag{1.109}$$

$$\text{Associativité : } s(x,s(y,z))=s(s(x,y),z) \tag{1.110}$$

Les s-normes définissent, pour les ensembles flous, une classe générale d'opérateurs d'union. Quelques exemples de s-normes sont présentés dans le tableau 1.1. Ces opérateurs ont un caractère tolérant ou compensatoire, cela veut dire que le résultat de l'agrégation dépend fortement de la plus grande des valeurs à agréger. Une propriété importante est que l'opérateur "max" est une borne inférieure pour toutes les s-normes, autrement dit le moins tolérant des s-normes.

Tableau 1.1 Exemple de t-normes et de s-normes

	t-normes	*s-normes*
Brute (Weber)	$t_W(x,y)=\begin{cases}\min(x,y)\ si\ \min(x,y)=1\\ 0 \quad sinon\end{cases}$	$s_W(x,y)=\begin{cases}\max(x,y)\ \ si\ \ \min(x,y)=0\\ 1 \quad sinon\end{cases}$
Bornée (Lukasiewicz)	$t(x,y)=\max(0,x+y-1)$	$s(x,y)=\min(1,x+y)$
Hamacher ($\gamma=0$), Algébrique ($\gamma=1$), Einstein ($\gamma=2$)	$t(x,y)=\dfrac{x.y}{\gamma+(1-\gamma).(x+y-x.y)}$	$s(x,y)=\dfrac{x+y-x.y-(1-\gamma).x.y}{1-(1-\gamma).x.y}$
Propriétés importantes	$t(x,y)=\min(x,y)$ tel que : $t_W(x,y)\le t(x,y)\le\min(x,y)$	$s(x,y)=\max(x,y)$ tel que : $\max(x,y)\le s(x,y)\le s_W(x,y)$

1.6.1.3. Opérateurs de compromis

On appelle opérateur de compromis toute fonction d'agrégation se situant entre l'opérateur de conjonction "min" et celui de disjonction "max". Les figures 1.29 et 1.30 illustrent respectivement le comportement de l'opérateur moyenne "m", ainsi que les courbes de niveaux en fonction de la valeur de α, donnée par :

$$m(x,y)=\left(\frac{x^{\alpha}+y^{\alpha}}{2}\right)^{1/\alpha} \qquad (1.111)$$

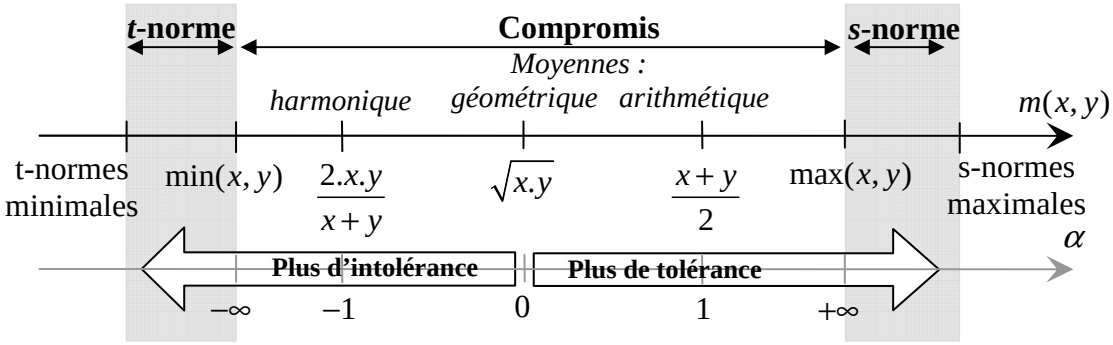


Figure 1.29 le comportement de l'opérateur de compromis moyenne $m(x,y)$ en fonction de la valeur de α.

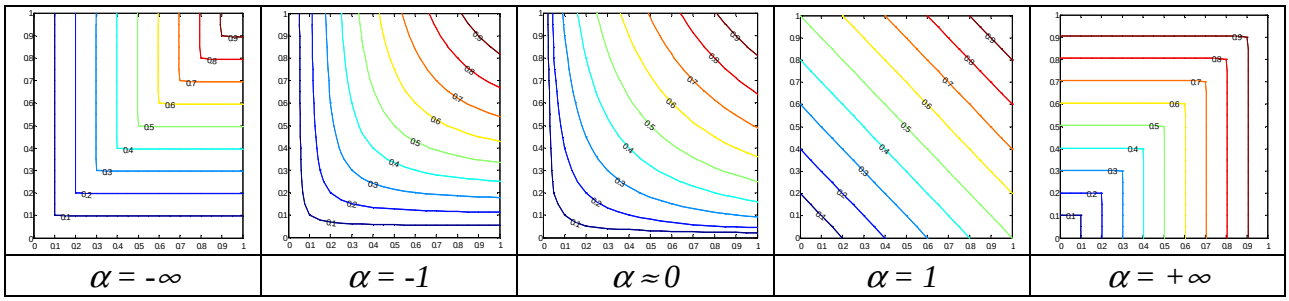


Figure 1.30 Courbes de niveaux obtenues par l'opérateur de compromis moyenne " $m(x,y)$ " en fonction de la valeur de α.

Il arrive que lorsqu'on combine deux distributions de possibilités $\pi_1(\omega)$ et $\pi_2(\omega)$, la distribution résultante $\pi(\omega)$ peut devenir sous-normale ($\sup_{\omega\in\Omega}\pi(\omega)<1$), notamment lorsqu'un mode conjonctif ou de compromis est appliqué. Il convient alors de procéder à un processus de normalisation.

1.6.1.4. Opérateurs dépendant du contexte

Le résultat de la combinaison obtenu par les opérateurs précédents ne dépend que des valeurs à combiner. Pour les opérateurs dépendant du contexte, le résultat des combinaisons dépend d'autres informations telles que la fiabilité des sources d'informations ou capteurs, de

l'interaction ou du conflit entre ces sources. Citons comme exemples de cette famille les opérateurs adaptatifs et l'intégrale floue.
Les opérateurs adaptatifs, utilisés en théorie de possibilités, adaptent leur mode de combinaison en fonction du degré du conflit entre les sources d'informations ainsi que des fiabilités. Ils se comportent comme un opérateur conjonctif si les distributions de possibilités sont consonantes (faible conflit). Dans ce cas les deux sources sont fiables et le mode de combinaison peut être sévère. Par contre si les sources sont dissonantes (conflit fort), les opérateurs auront un comportement disjonctif favorisant l'ensemble de solutions données par les deux sources. Enfin, ils se comportent comme un opérateur compromis dans le cas du conflit partiel, en adoptant un comportement prudent. Plusieurs exemples de cette famille d'opérateurs de combinaison peuvent se trouver dans [DUB92a]. A titre illustratif, l'équation (1.112) représente un opérateur combinant deux distributions de possibilités π_1 et π_2, définies sur D, représentant par exemple l'imprécision sur une variable estimée de deux manières différentes. La fonction $\lambda(\pi_1,\pi_2)$ est une mesure de leurs conflits, calculée par l'équation (1.113).

$$\pi'(s) = t\left[\pi_1(s),\pi_2(s)\right] - \lambda(\pi_1,\pi_2) \tag{1.112}$$

$$\lambda(\pi_1,\pi_2) = 1 - \max_{s\in D} \min(\pi_1(s),\pi_2(s)) \tag{1.113}$$

Les intégrales floues permettent de modéliser de nombreux comportements décisifs. Elles intègrent les moyennes généralisées, les minima et maxima pondérés et encore bien d'autres opérateurs. Leur principal intérêt est de représenter une interaction entre les critères. Ceci est dû au fait qu'un poids d'importance peut être attribué à chaque sous-ensemble de critère. Elles ont tout d'abord été introduites par Sugeno en 1974 avant d'être généralisées par Weber, puis par Murofushi et Sugeno [MUR91]. Cependant et pour a sous-ensembles de critères, ces intégrales nécessitent la détermination de $2^a - 2$ paramètres. Ces derniers peuvent être déterminés par l'expérience humaine ou automatiquement [GRA93, GRA95].

1.7. Conclusion

Dans un système de diagnostic, les informations recueillies sont à la fois imprécises et incertaines. Cela est dû aux capteurs et différents instruments de mesures, au bruit et aux experts humains. La robustesse d'un système conçu dans le but d'aider un opérateur ou de prendre des décisions nécessite des théories capables d'inclure ces imperfections dans la représentation de l'information et le raisonnement. Dans ce chapitre, nous avons étudié les principales théories permettant de représenter les connaissances imparfaites et de raisonner à partir de celles-ci. Ces théories sont : les probabilités, les fonctions de croyance, les ensembles flous et les possibilités. Nous avons montré que la théorie des probabilités est incapable de représenter des informations à la fois incertaines et imprécises. Ensuite, puisque la théorie des fonctions de croyance a une complexité importante pour la construction des fonctions de masses de croyance, nous avons opté pour l'utilisation de la théorie des possibilités. En effet la construction d'une mesure de possibilité étant intuitive, nous préférons déterminer une mesure probabiliste, pauvre et imprécise par nature, et en déduire la mesure de possibilité par une transformation bijective. C'est pourquoi nous avons choisi la méthode de classification Fuzzy Pattern Matching (FPM) qui est basée sur la transformation des histogrammes des probabilités en ceux des possibilités.

Il existe plusieurs transformations probabilité-possibilité dans la littérature. Nous avons étudié trois transformations, à savoir optimale et non optimale de Dubois et Prade et la Transformation Variable (TV), pour les bons résultats obtenus dans les applications de la Reconnaissance des Formes (RdF). Nous avons proposé une solution pour rendre la Transformation Variable (TV) proposée dans [SAY06] aussi spécifique que la transformation optimale de Dubois et Prade tout en respectant la condition de cohérence de Dubois et Prade. Les trois transformations sont ensuite comparées selon plusieurs critères d'évaluation. Nous avons démontré que la TV améliorée est meilleure que les deux autres transformations selon ces critères. Enfin le problème de la fusion de l'information provenant de plusieurs sources ainsi que les modes de combinaisons les plus connus de la littérature ont été brièvement abordés.

Dans le chapitre suivant, nous allons étudier les performances de FPM afin de constater ses limites. En effet, FPM est inopérante pour la discrimination des classes de forme non-convexe et/ou décrites par des attributs corrélés. Il existe dans la littérature quatre solutions qui sont basées sur FPM. Ces dernières sont FPM Multi-prototype (FPMM) et sa version améliorée : FPM utilisant une fonction Exponentielle (FPME), développées dans [DEV99] ; FPM Corrélée (FPMC) introduite dans [SAY02a] et FPM utilisant la méthode des fenêtres de Parzen présentée dans [CAD04]. Une étude détaillée a permis de révéler les avantages et les inconvénients de chacune de ces méthodes. Nous avons constaté que FPMC peut résoudre plusieurs inconvénients dont souffrent les autres méthodes basées sur FPM. Par contre FPMC manque de formalisation. C'est pourquoi nous allons d'abord la formaliser afin d'en déduire ses limites et ensuite l'améliorer. Nous appellerons la nouvelle version de FPMC, FPM Améliorée (FPMA). Enfin, nous allons comparer les performances de FPM et FPMA à celles de la méthode k plus proches voisins et la méthode des noyaux de Parzen en utilisant des données académiques et réelles.

Chapitre 2

Discrimination tenant compte de la forme des classes et de la corrélation des attributs

2.1. Introduction

Nous avons choisi la méthode Fuzzy Pattern Matching (FPM) comme méthode de classification non paramètrique pour réaliser le diagnostic. Nous avons justifié ce choix à travers le chapitre 1 en la comparant à d'autres méthodes de classification. Cependant, FPM est une méthode de classification naïve, c'est-à-dire que sa décision globale est basée sur la sélection d'une des décisions partielles. Chaque décision partielle est calculée en utilisant une fonction d'appartenance construite pour chaque classe et par rapport à chaque attribut. FPM ne tient donc pas compte de la corrélation entre les attributs et considère la forme des classes comme convexe. Ces inconvénients rendent FPM inutilisable pour de nombreuses applications réelles et en particulier pour celles qui demandent une discrimination non linéaire entre les classes. Ce chapitre propose d'améliorer FPM afin de remédier à ces inconvénients tout en conservant ses avantages. Ce chapitre est structuré comme suit. Premièrement, le problème de la discrimination des classes non-convexes ou décrites par des paramètres corrélés est étudié. Les performances de FPM dans ce cas sont évaluées à travers des exemples. Ensuite, un état de l'art des solutions existantes pour la discrimination des classes non-convexes est présenté. Cet état de l'art est divisé en deux parties. La première partie présente les méthodes de classification les plus connues de la littérature, autre que FPM, qui sont opérantes pour la discrimination non linéaire des classes. La deuxième partie, quant à elle, traite les solutions basées sur FPM pour discriminer des classes non-convexes. Cet état de l'art est suivi par une nouvelle solution pour remédier aux limites de FPM. Enfin les performances de cette solution sont évaluées à travers plusieurs exemples académiques et réels et comparées par rapport à FPM classique, à la méthode k plus proches voisins (kppv) et à la méthode des noyaux Parzen.

2.2. Limites de la méthode Fuzzy Pattern Matching (FPM)

FPM s'apparente à un discriminateur Bayesien naïf qui suppose que les attributs sont statistiquement indépendants. Pour représenter les classes, ce discriminateur utilise les fonctions de densités de probabilités marginales [GRA00]. La probabilité d'appartenance *a posteriori* $P(C_i|\underline{x})$ de $\underline{x}$ à la classe C_i s'écrit alors comme suit :

$$P(C_i|\underline{x}) = \prod_{j=1}^{a} p(x^j|C_i)P(C_i) \qquad (2.1)$$

où $p(x^j|C_i)$ est la densité conditionnelle de x^j, selon l'attribut *j*, sachant la classe C_i. Si les probabilités *a priori* $P(C_i)$, $i \in \{1,2,...,c\}$ des classes sont équiprobables alors le produit peut être remplacé par le minimum (min) [GRA00].

Rappelons que l'opérateur "min" procure à FPM le caractère sélectif des attributs pour la décision globale. L'équation (2.1) devient, dans ce cas, équivalente à l'équation (1.65). Chaque densité de possibilité, de chaque attribut, représente alors une densité marginale. Cette marginalité de FPM induit des limites, pour une discrimination non linéaire entre les classes, comme dans le cas des classes non convexes ou bien des classes représentées par des attributs corrélés.

2.2.1. Forme des classes et corrélation des attributs

Il a été démontré dans plusieurs applications réelles que la classification par FPM donne de bons résultats à condition que les classes soient parfaitement séparables et de formes convexes orientées parallèlement aux axes de représentation [DEV04, GRA92, SAY02a], comme le cas (a) de la figure 2.1. Par contre, les résultats sont nettement moins satisfaisants si les classes sont de formes non convexes, obliques ou décrites par des attributs corrélés comme dans les cas (d), (b), et (c), respectivement. Le cas (c) présente le problème OU-exclusif (XOR). Ce problème compte parmi les exemples les plus adaptés pour l'étude de la corrélation des attributs.

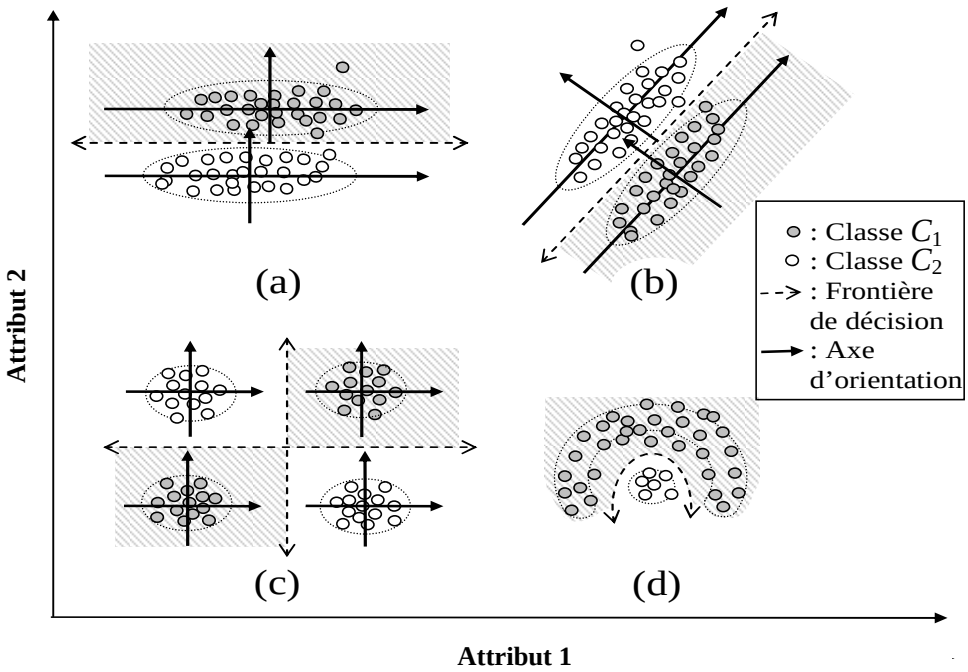


Figure 2.1 Exemple de classification dans un espace de deux attributs. La séparabilité des classes dans le cas (a) et (b) est linéaire, contrairement au cas (c) et (d). Les formes des classes dans les cas (a), (b) et (c) sont convexes contrairement au cas (d). L'orientation des classes (a) et (c) est parallèle aux axes de représentation contrairement au cas (b).

La marginalité de FPM impose une forme rectangulaire ou circulaire des courbes de niveaux d'appartenance aux classes indépendamment de leurs formes réelles qui peuvent être convexes ou non.

2.2.2. Performances de FPM

Exemple des matières plastiques

L'intérêt de respecter la forme de la classe est illustré par la figure 2.2 qui présente les courbes de niveaux d'appartenance obtenues par FPM en utilisant comme opérateur d'agrégation le minimum. Nous constatons que la forme rectangulaire de ces courbes ne respecte pas la forme oblique de la classe C_1. Cela traduit le fait qu'un point se trouvant à la périphérie de celle-ci peut avoir la même valeur d'appartenance qu'un point se trouvant en son centre. De plus, un point peut avoir une valeur d'appartenance à cette classe alors qu'il ne lui appartient pas, comme pour le point $\underline{x}_2$. En effet, le point $\underline{x}_2$ qui est éloigné de C_1, sera affecté avec la même possibilité d'appartenance que $\underline{x}_1$, alors qu'il ne le devrait pas. Cela est une conséquence directe de l'opérateur utilisé qui synthétise le degré de possibilité global π_i en se basant sur la sélection d'une seule possibilité marginale π_i^j par rapport à un seul attribut j.

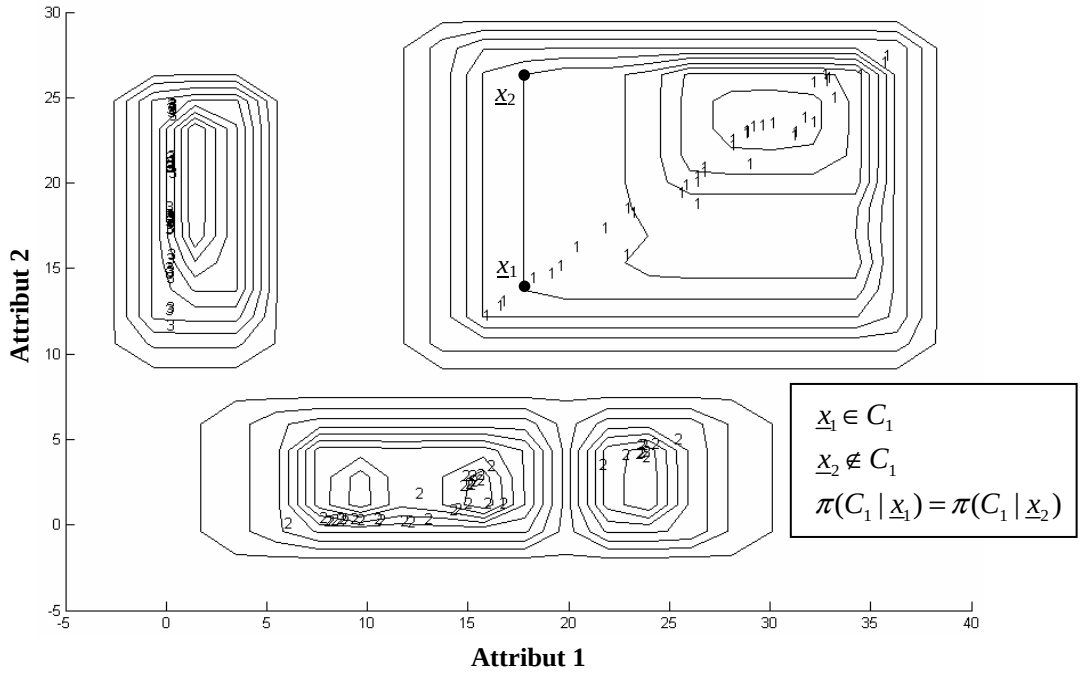


Figure 2.2 Courbes de niveaux d'appartenance, obtenues par FPM classique, pour l'exemple des matières plastiques.

Problème double XOR

Les courbes de niveaux d'appartenance, obtenues par FPM, pour la classe C_1 sont décrites dans la figure 2.3.a. Comme on peut le constater, FPM respecte la forme convexe de la classe

C_1, par contre elle ne distingue pas ses points de ceux de la classe C_2 bien qu'ils soient éloignés les uns des autres; puisque leurs densités de possibilité, représentées dans la figure 2.3.b, se superposent par rapport à l'attribut 1 et à l'attribut 2. Le même problème se répète lorsqu'il s'agit de discriminer la classe C_2. On obtient, dans ce cas, des courbes similaires à celles des figures 2.3.a et 2.3.b.

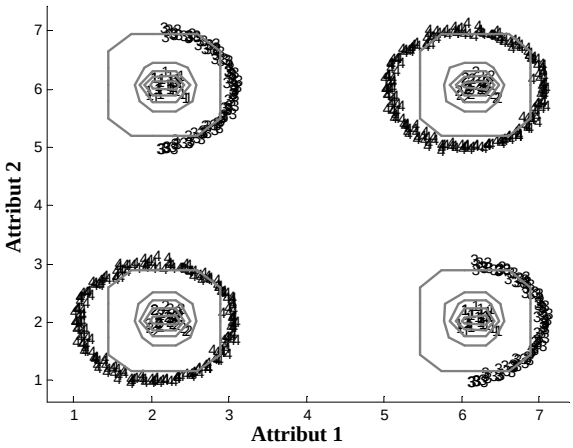


Figure 2.3.a Courbes de niveaux d'appartenance à C_1.

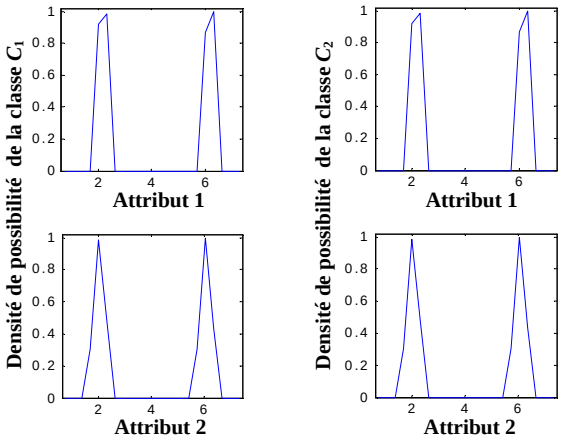


Figure 2.3.b Densités de possibilité de C_1 et C_2 par rapport aux attributs 1 et 2.

Les courbes de niveaux d'appartenance, obtenues par FPM, pour la classe C_3 sont décrites dans la figure 2.4.a. Comme on peut le constater, FPM considère que la forme de cette classe est convexe et par conséquent elle ne distingue pas les points de la classe C_1 de ceux de la classe C_3. Le fait de ne pas tenir compte de la corrélation, FPM confond les points de la classe C_4 et ceux de la classe C_2 et les considère comme des points de la classe C_3.

La même remarque peut être faite pour les courbes d'appartenance à la classe C_4 comme le montre la figure 2.4.b.

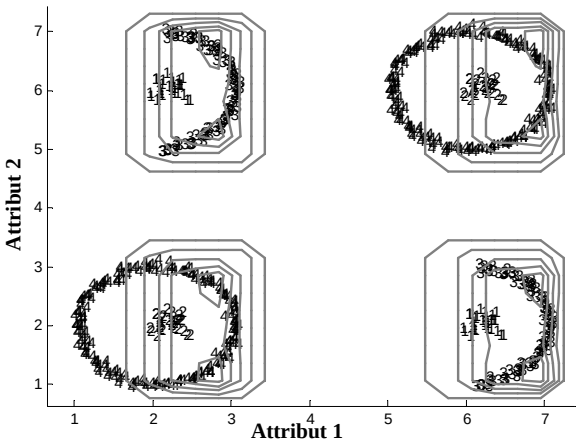


Figure 2.4.a Courbes de niveaux d'appartenance à C_3.

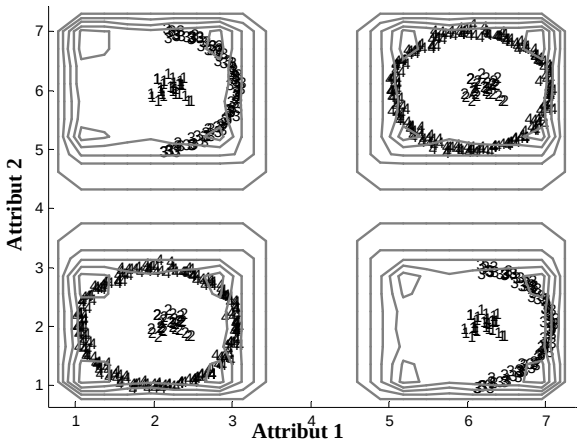


Figure 2.4.b Courbes de niveaux d'appartenance à C_4.

2.3. Solutions basées sur des méthodes de RdF autre que FPM

2.3.1. Classification par Support Vector Machines (SVM)

La méthode Machines à Vecteur de Support, Support Vector Machines (SVM), connue également sous le nom de séparateur à vaste marge, maximum margin classifiers, est une méthode de classification supervisée binaire [VAP95, BUR98, CRI00, AMA06]. Elle repose sur l'existence d'un séparateur linéaire dans un espace de représentation approprié. Elle est basée sur l'utilisation de fonctions dites noyaux (Kernels) qui permettent une séparation optimale entre deux classes.

La figure 2.5 montre un exemple de séparation linéaire entre deux classes. Les points de l'ensemble d'apprentissage les plus proches de l'hyperplan qui sépare les deux classes sont appelés les vecteurs de support. SVM cherche l'hyperplan optimal qui maximise la distance, ou la marge, entre les points de deux classes et cet hyperplan.

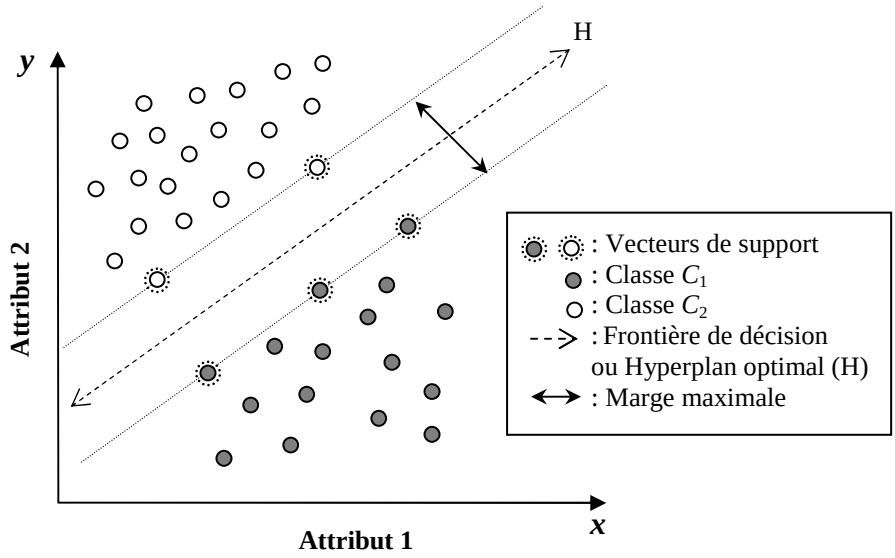


Figure 2.5 Principe de fonctionnement de la méthode SVM dans le cas de classes linéairement séparables. Les vecteurs de support désignent les points les plus proches de l'hyperplan séparateur H.

La méthode SVM, tout comme la méthode des k plus proches voisins ou la méthode de Parzen, est capable de discriminer des classes non-convexes sans *a priori* sur la distribution des données. Par contre elle réalise cette discrimination par la transformation de l'espace de représentation initial, dans lequel la séparation est non linéaire, en un espace de dimension plus élevée. Grâce à cette transformation la séparation des données devient linéaire. Elle est réalisée via une fonction noyau (polynomiale, gaussienne, etc.) qui répond au critère de Mercer [VAP95].

Pour illustrer ce cas de transformation, prenons le problème XOR dans un espace de représentation de 2 attributs, cf. figure 2.6. La séparation de la classe C_1, représentée par les points (0,1) et (1,0), de la classe C_2, formée par les points (0,0) et (1,1), nécessite une séparation non linéaire comme le montre la frontière de décision de la figure 2.6 à gauche. Si on prend la fonction polynomiale suivante $(x, y) \mapsto (x, y, x \times y)$ qui fait passer d'un espace de dimension 2 à un espace de dimension 3, on obtient un problème en trois dimensions

linéairement séparable : $(0,0) \mapsto (0,0,0)$, $(0,1) \mapsto (0,1,0)$, $(1,0) \mapsto (1,0,0)$, $(1,1) \mapsto (1,1,1)$, comme le montre la figure 2.6 à droite. Nous constatons, à travers cette figure, que la séparation entre les deux classes est linéaire par rapport à chaque plan de l'espace de représentation.

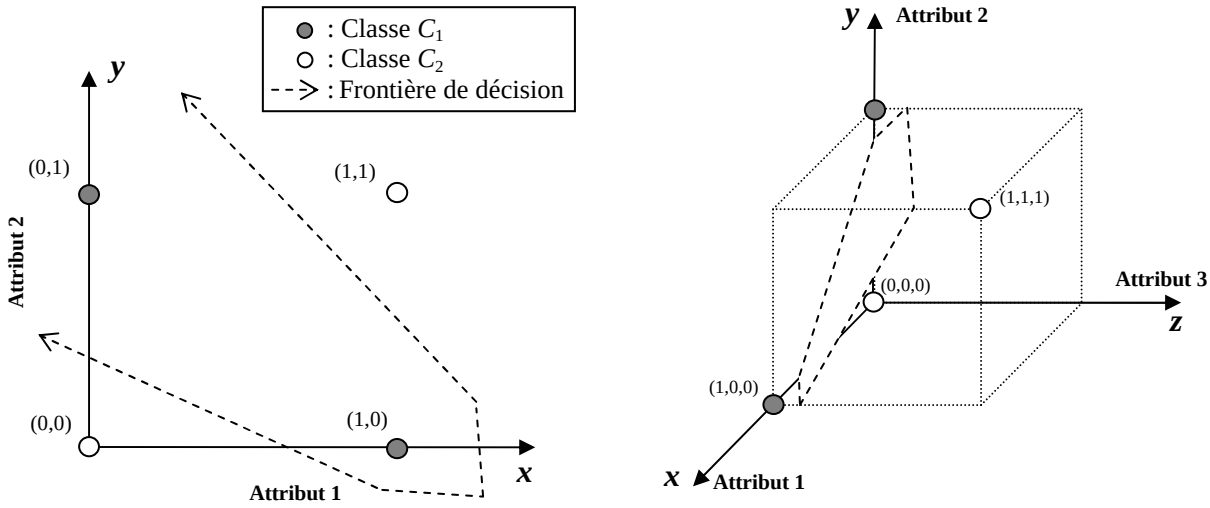


Figure 2.6 Séparation non linéaire pour le problème XOR, à gauche. Séparation linéaire du problème XOR grâce à la transformation de l'espace de représentation par une fonction à noyau polynomiale, à droite.

SVM a montré son efficacité pour la discrimination non linéaire des classes [LUR03, AMA06]. L'estimation de la densité de probabilité d'une façon non paramétrique à l'aide de la méthode SVM (de type One-Class SVM) est définie uniquement avec les vecteurs de support qui ne sont qu'un sous-ensemble restreint de la base d'apprentissage, contrairement aux estimateurs classiques, cf. chapitre 1, qui utilisent toutes les données d'apprentissage. La méthode SVM est adaptée pour la classification des données évolutives, comme nous allons le voir dans le chapitre 3. Comme limites de cette méthode, on peut citer l'impossibilité d'intégrer un mécanisme pour la scission des classes ou encore le temps de calcul qui peut exploser quand il y a un grand nombre de points dans l'ensemble d'apprentissage.

2.3.2. Classification floue à base de règles graduelles

Les méthodes de classification à base de règles graduelles [DUB92b, DAR06] définissent des quadrilatères pour chaque classe. Ces quadrilatères s'adaptent mieux à la forme géométrique des classes contrairement aux règles conjonctives qui ont une forme rectangulaire aux côtés parallèles aux axes de l'espace de représentation. La forme générale des règles graduelles est de type « *Plus x est* A_i*, plus y est* B_i ». Cette règle correspond à la définition d'une relation implicative ou une contrainte liant les attributs x et y. L'intégration d'une règle graduelle dans une règle floue de classification, de type « Si vecteur d'attributs vérifie condition z alors la classe est C_i », conduit à la forme suivante :

$$\begin{aligned} &Si(x\ est\ A_1 \rightarrow y\ est\ B_1) \\ &et\ (x\ est\ A_2 \rightarrow y\ est\ B_2)\ alors\ classe = C_i \end{aligned} \qquad (2.2)$$

A_i et B_i sont des termes linguistiques qui décrivent les attributs x et y. Le symbole « $\rightarrow$ » est un opérateur d'implication. Un exemple de ce dernier est l'implication de Rescher-Gaines définie par :

$$V \rightarrow T = \begin{cases} 1 & V \leq T \\ 0 & V > T \end{cases} \qquad (2.3)$$

L'exemple de la figure 2.7 illustre un quadrilatère généré par une règle graduelle de la forme (2.2) et l'implication (2.3). Un point aura un degré d'appartenance de 1 à la classe C_i s'il se trouve dans le quadrilatère. Sinon ce point aura la valeur d'appartenance 0 à cette classe. Le résultat de classification est donc net et pas flou. Obtenir des résultats flous de classification est un point actuel de recherche.

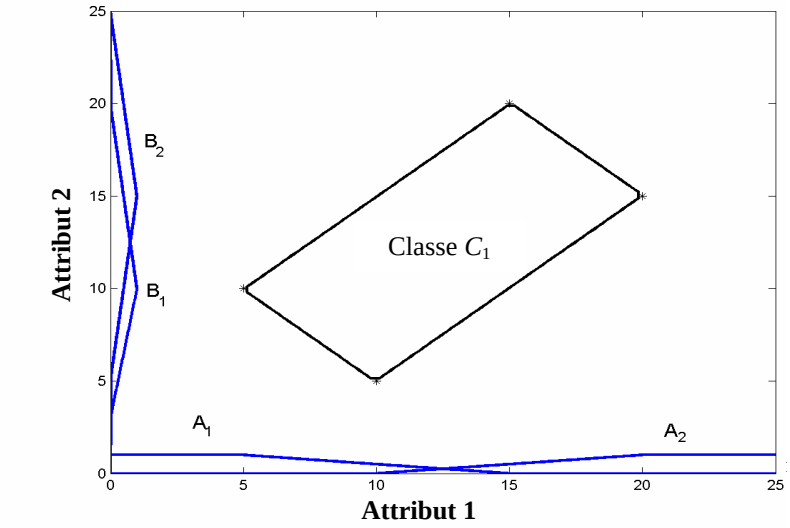


Figure 2.7 Quadrilatère généré par une règle graduelle.

L'augmentation du nombre de règles, ou contraintes, conduit à une surface plus restreinte et donc plus adaptée à des formes de classes complexes. Ces règles définissent alors des quadrilatères assez flexibles en respectant quelques contraintes. Toutefois, la construction de ces règles est un problème complexe dans un espace à dimension élevée.

2.3.3. Méthodes de coalescence floue

La méthode FCM et son extension PCM, étudiées au chapitre 1, imposent un modèle unique pour la forme des classes. Cette forme est hypersphérique ou hyperelliptique, ayant la même orientation, selon la métrique de la distance Euclidienne ou Mahalanobis. Elles ne permettent pas donc de définir un modèle propre à chaque classe en fonction de sa forme non-convexe.

[BEZ81] propose une solution qui ne considère plus les centres des classes comme des points mais comme des variétés linéaires d'ordre r. Cette solution est appelée Fuzzy C Varieties (FCV). r = 1 définit la variété comme une droite, r = 2 comme un plan, r = 3 comme un hyperplan, …. L'inconvénient majeur de FCV est le cas de colinéarité des classes. Ce cas est caractérisé par deux classes différentes ayant la même variété. Afin de remédier à cet inconvénient, une solution basée sur la combinaison de FCM et FCV est proposée [LES04]. Cette solution consiste à minimiser, d'une part, la distance entre un point et le centre de la classe et d'autre part entre ce même point et la variété linéaire caractérisant la classe.

Toutefois cette solution nécessite plus de paramètres à régler et un temps de calcul relativement important.

[DEV99] propose une méthode permettant de trouver les partitions des classes de forme non-convexe dans un nuage de points bruités. Cette solution est appelée Unsupervised Fuzzy Graph Clustering (UFGC). Cette méthode est basée sur la division du nuage de points en un nombre surestimé de sous-classes. Ensuite la fusion des sous-classes est réalisée en utilisant un indice de similarité floue, ou d'ambiguïté, pour chaque couple de sous-classes. Cet indice est la moyenne de la somme des différences de valeurs d'appartenance des points à ce couple de sous-classes. Plus cet indice est proche de 1, plus les deux sous-classes sont proches l'une de l'autre. Ces deux sous-classes sont fusionnées si leur indice de similarité floue est supérieur ou égale à un seuil. Toutefois les performances de cette méthode sont très sensibles au choix du seuil de fusion et le temps de calcul long empêche l'utilisation en ligne.

2.4. Solutions basées sur FPM

Il existe dans la littérature quatre solutions qui sont basées sur FPM. Ces dernières sont FPM Multi-prototype (FPMM) et sa version améliorée : FPM utilisant une fonction Exponentielle (FPME), développées dans [DEV99] ; FPM utilisant la méthode des fenêtres de Parzen présentée dans [CAD04] et FPM Corrélée (FPMC) introduite dans [SAY02a]. Une étude détaillée a permis de révéler les avantages et les inconvénients de chacune de ces méthodes.

2.4.1. Fuzzy Pattern Matching Multi-prototype (FPMM)

FPM dans sa version classique est basée sur l'approche mono-prototype. C'est-à-dire qu'elle considère qu'une classe de forme simple, représentant un mode de fonctionnement, est modélisée par un unique prototype. Par contre, afin de conférer à FPM la capacité de modéliser une classe de forme non-convexe ou complexe, il convient de définir plusieurs prototypes, dont chacun doit s'adapter localement à la distribution de données, et de les combiner. Suivant cette description, qu'on appelle approche multi-prototype, chaque classe sera représentée par un ensemble de sous-classes. Etablir un diagnostic basé sur FPM en utilisant l'approche Multi-prototype (FPMM) [DEV99] nécessite deux phases décrites ci-dessous.

2.4.1.1. Phase d'apprentissage

L'enrichissement de l'apprentissage de FPM en utilisant l'approche multi-prototype comporte deux étapes, illustrées dans la figure 2.8. La première étape se base sur la méthode de coalescence, C Moyennes Floues ou Fuzzy C Means (FCM) [BEZ81], afin de décomposer chaque classe de l'ensemble d'apprentissage en un nombre fini de sous-classes de cardinalités différentes. Dans la deuxième étape, la phase d'apprentissage de FPM est appliquée sur chaque sous-classe de l'ensemble de toutes les classes, afin de calculer leurs densités de possibilité par rapport à chaque attribut. La phase d'apprentissage de FPMM nécessite donc un paramètre supplémentaire par rapport à FPM. C'est le nombre s_i de sous-classes, par rapport à chaque classe C_i. Ce paramètre est nécessaire pour FCM afin qu'elle puisse diviser chaque classe C_i en plusieurs partitions SC_i^k, $k \in \{1,2,\ldots,s_i\}$, appartenant à s_i sous-classes.

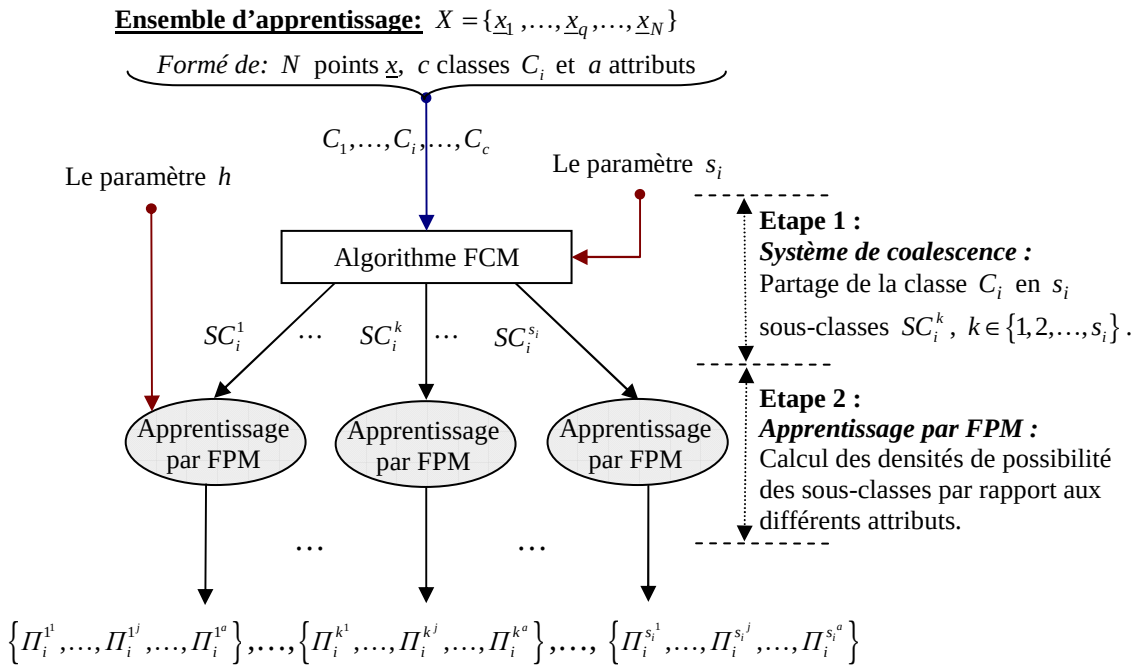


Figure 2.8 Phase d'apprentissage de FPMM. Illustration pour une classe C_i.

2.4.1.2. Phase de classification

La classification d'un nouveau point $\underline{x} \in \Re^a$, dont les valeurs pour les différents attributs sont $x^1, \ldots, x^a$, s'effectue en quatre étapes :

- détermination de la valeur de possibilité d'appartenance $\pi_i^{k^j}$ du point $\underline{x}$ à la sous-classes SC_i^k de C_i selon l'attribut j. Cette possibilité est calculée par simple projection de x^j sur la densité de possibilité $\Pi_i^{k^j}$,
- fusion, pour chaque sous-classe SC_i^k, de toutes les valeurs de possibilité d'appartenance $\pi_i^{k^1}, \pi_i^{k^2}, \ldots, \pi_i^{k^a}$ par un opérateur d'agrégation comme dans la phase de classification par FPM. Le résultat de cette fusion représente la possibilité d'appartenance π_i^k à chaque sous-classe SC_i^k de la classe C_i. La possibilité π_i^k est alors calculée par :

$$\pi_i^k = H\left(\pi_i^{k^1}, \pi_i^{k^2}, \ldots, \pi_i^{k^a}\right) \qquad (2.4)$$

- détermination de la valeur d'appartenance du point $\underline{x}$ à la classe C_i. Cette valeur est donnée par fusion des s_i possibilités d'appartenance aux sous-classes SC_i^k, $k \in \{1,2,\ldots,s_i\}$, de C_i par un opérateur d'agrégation. L'opérateur choisi est une somme bornée comme défini dans [MAS96]. Il permet entre autres d'atténuer les pics de forte appartenance qui se traduisent par des îlots d'appartenance sur les courbes de niveaux [DEV99]. Le résultat de cette fusion est donné par :

$$\pi_i(\underline{x}) = \pi_i^1(\underline{x}) \cup \ldots \cup \pi_i^{s_i}(\underline{x}) = \min\left[1, \sum_{k=1}^{s_i} \pi_i^k(\underline{x})\right] \qquad (2.5)$$

- affectation du point $\underline{x}$ à la classe pour laquelle il a la valeur de possibilité d'appartenance la plus forte.

La figure 2.9 résume, les étapes principales du déroulement de la phase de classification par FPMM.

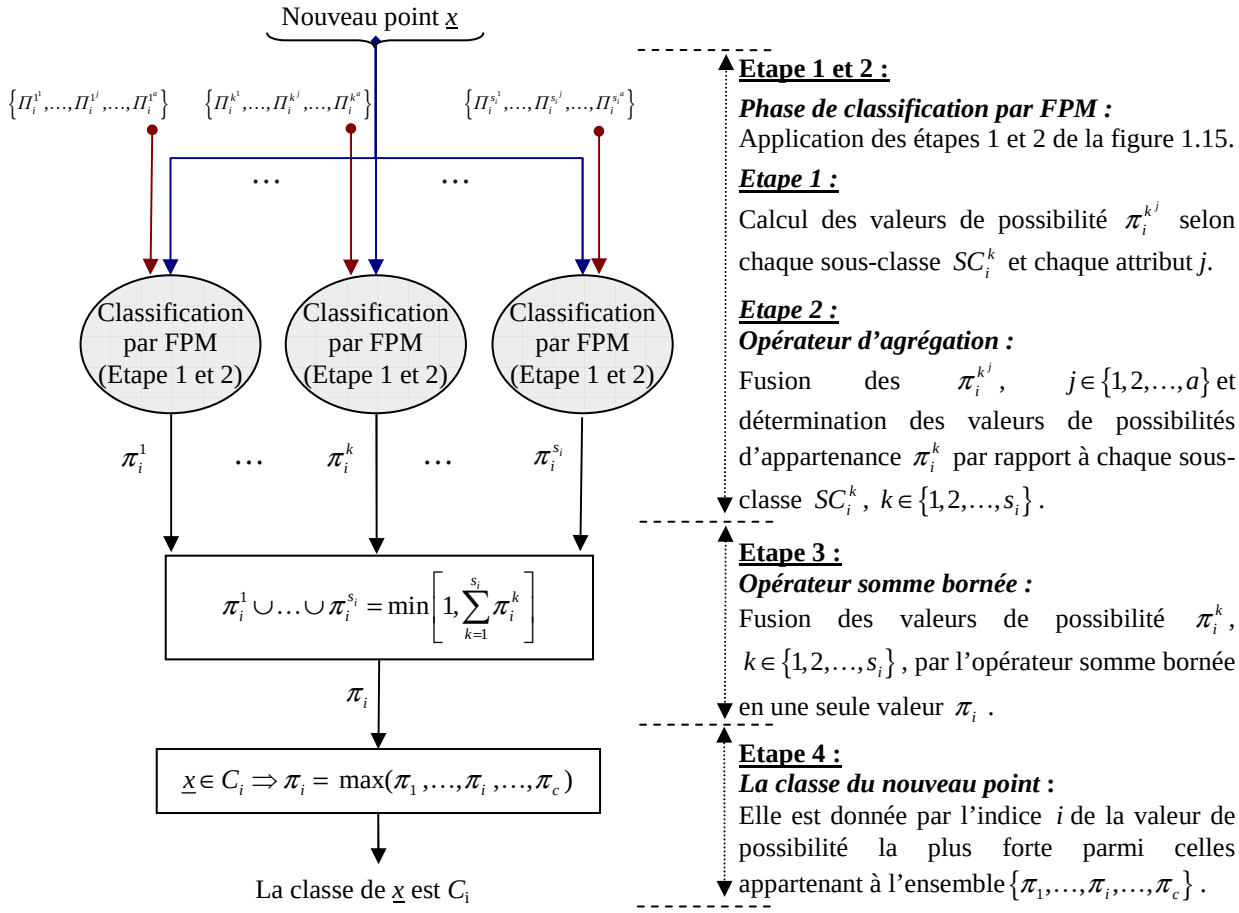


Figure 2.9 Phase de classification de FPMM.

2.4.1.3. Performances de FPMM

La figure 2.10 représente une classe ayant une forme non convexe. La forme de cette classe est prise en compte dans la méthode FPMM suite à sa décomposition en sous-classes convexes et grâce aux calculs de leurs densités marginales.

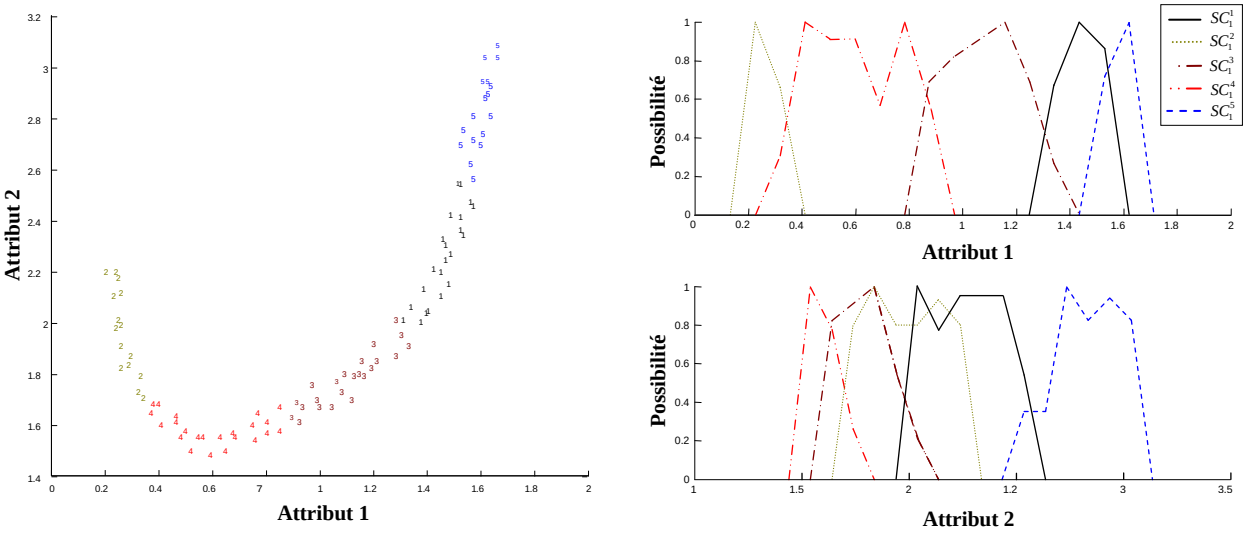


Figure 2.10 Décomposition de la classe de forme non convexe en 5 sous-classes convexes. Les points marqués avec la même étiquette appartiennent à la même sous-classe, à gauche. Les densités de possibilités des 5 sous-classes par rapport à chaque attribut sont représentées à droite.

Dans la figure 2.11, on peut constater que la précision de ces courbes en terme du respect de la forme de la classe est proportionnelle au nombre de sous-classes créées dans cette classe. Plus ce nombre est important, plus ces courbes respectent la forme de la classe.

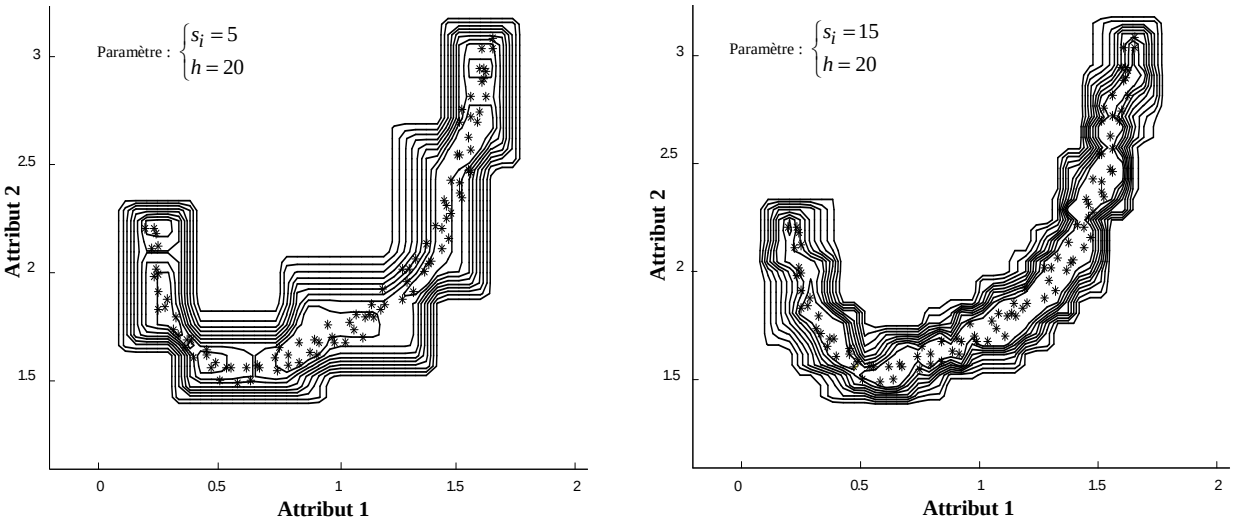


Figure 2.11 Courbes de niveaux d'appartenance obtenues par FPMM pour l'exemple de la figure 2.10. Le paramètre s_i est fixé à 5 sous-classes, à gauche, et à 15 sous-classes, à droite.

Bien que cette solution permette l'application de FPM aux classes de forme complexe, son fondement théorique la rend très dépendante de l'opérateur d'agrégation qui doit être utilisé [DEV99]. De plus, elle nécessite un paramètre supplémentaire à ajuster. Ce dernier, le nombre de sous–classes dans chaque classe, n'est pas toujours facile à déterminer. En effet, la recherche de la précision des courbes par accroissement des valeurs des paramètres de FPMM, favorise l'apparition d'îlots d'appartenance [DEV99]. Cela provoque également l'augmentation du temps d'apprentissage et de classification, ce qui peut compromettre l'utilisation en temps réel. Enfin, puisqu'elle est basée sur l'approche multi-prototype cette méthode n'est pas capable de distinguer les zones de haute densité quand elles existent

[SAY02a]. La forme de la classe obtenue par FPMM n'est pas rigoureusement respectée. Cela est dû d'une part à la sensibilité de cette méthode à la qualité de l'ensemble d'apprentissage. D'autre part, cela est dû à la construction de la fonction d'appartenance, à partir des données locales relatives à chaque sous-classe, d'une façon marginale sans tenir compte de leurs densités jointes. Une solution pour ce problème, décrite ci-dessous, est proposée dans [DEV99].

2.4.2. Fuzzy Pattern Matching utilisant une fonction Exponentielle (FPME)

Dans la littérature, il existe une grande variété de fonctions pour caractériser l'appartenance d'un point par rapport à un prototype. Ces fonctions se différencient par leur forme et la rapidité de leur décroissance. La fonction la plus simple et la plus employée est la fonction exponentielle. L'enrichissement de l'apprentissage et de la classification par FPM est réalisé par l'introduction d'une fonction exponentielle relative à la densité de probabilité jointe des attributs. Cela constitut le principe de la solution développée par [DEV99]. Cette solution est appelée « Fuzzy Pattern Matching using Exponential function » (FPME). FPME nécessite deux phases décrites ci-dessous.

2.4.2.1. Phase d'apprentissage

Les étapes décrivant la phase d'apprentissage sont illustrées dans la figure ci-dessous.

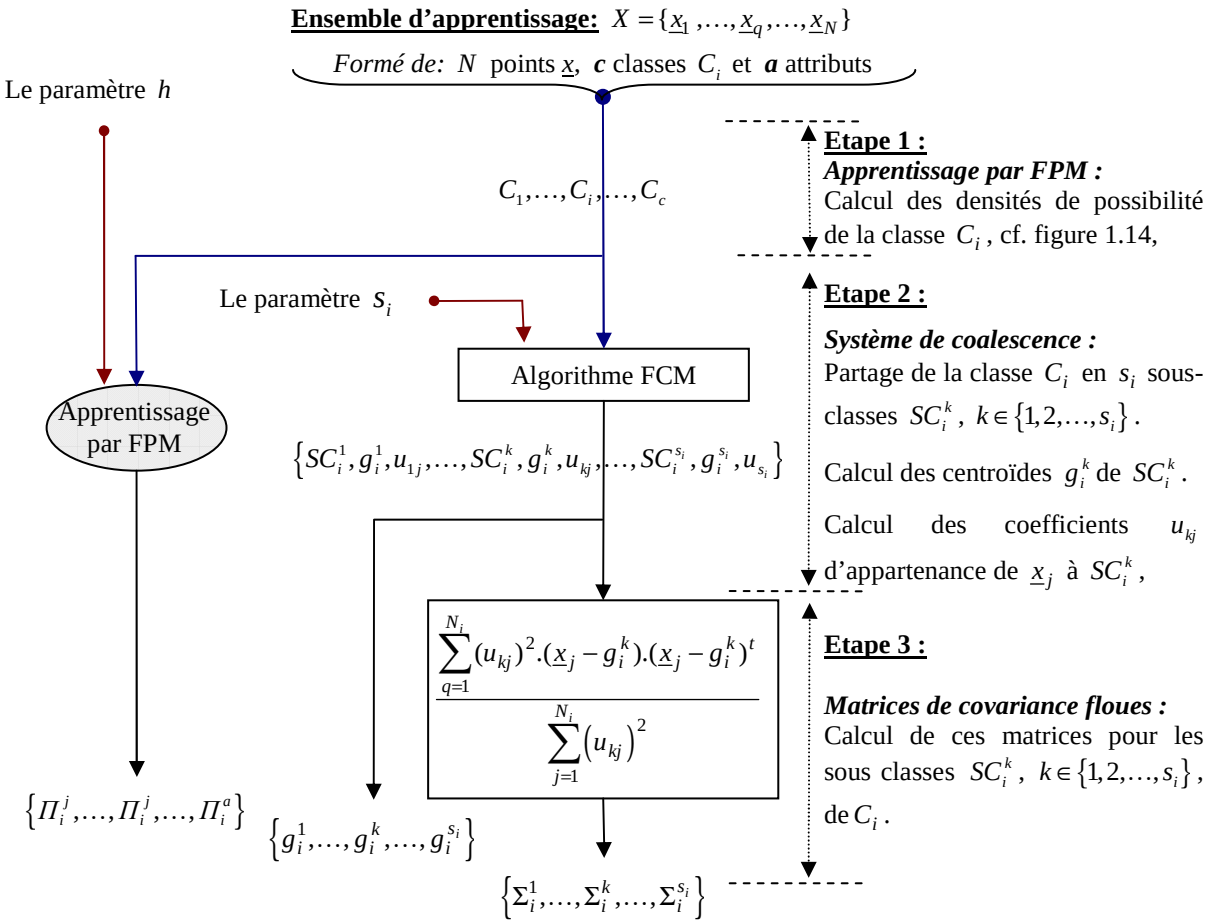


Figure 2.12 Phase d'apprentissage par FPME. . Illustration pour une classe C_i.

La phase d'apprentissage considère les données comme un mélange de distributions. Chaque distribution est modélisée par une fonction exponentielle. Ces données suivent alors une loi dont la fonction de densité est une densité de mélange représentée par une exponentielle englobante. Dans ce cas de modèle de mélange, la spécification des paramètres inconnus de la fonction exponentielle est réalisée par l'algorithme FCM. Chaque classe C_i, de l'ensemble d'apprentissage, est partagée en s_i sous-classes par FCM [BEZ81]. Chacune de ces sous-classes est représentée par un prototype de son centroïde g_i^k. La matrice d'appartenance U est composée des éléments u_{kq} représentant le coefficient d'appartenance du point $\underline{x}_q$ à la sous-classe SC_i^k. Ce qui donne une matrice d'appartenance pour chaque classe.

2.4.2.2. Phase de classification

Une fonction exponentielle est introduite pour caractériser l'éloignement ou le rapprochement de tout nouveau point $\underline{x}$ par rapport aux sous-classes. La valeur de cette fonction doit être de 1 pour les observations qui appartiennent de façon certaine au prototype de la sous-classe, et décroître vers 0 au fur et à mesure que les observations s'en éloignent. La fonction exponentielle est calculée pour chaque sous-classe SC_i^k selon l'expression suivante [MAS96] :

$$\mu_i^k(\underline{x}, g_i^k, \lambda_i^k) = \exp\left[-\lambda_i^k . d(\underline{x}, g_i^k)\right] \tag{2.6}$$

où λ_i^k est un paramètre utilisé pour ajuster la raideur de la courbe exponentielle et $d(\underline{x}, g_i^k)$ représente la distance de Mahalanobis entre le point $\underline{x}$ et le centroïde g_i^k de la sous-classe SC_i^k. Elle utilise la matrice de covariance floue et est définie par l'expression suivante [GUS79] :

$$d(x, y) = (x - y)^t . (\Sigma_i^k)^{-1} . (x - y) \tag{2.7}$$

Connaissant les coefficients d'appartenance des points d'apprentissage, la matrice de covariance floue de chaque sous-classe est définie par :

$$\Sigma_i^k = \frac{S_i^k}{\sum_{j=1}^{N_i} \left(u_{kj}\right)^2} \tag{2.8}$$

où N_i est le nombre de points dans la classe C_i et S_i^k étant la matrice de dispersion floue définie par :

$$S_i^k = \sum_{j=1}^{N_i} (u_{kj})^2 . (x_j - g_i^k) . (x_j - g_i^k)^t \tag{2.9}$$

L'introduction de cette matrice dans les calculs de distance permet de respecter la forme elliptique des sous-classes en tenant compte de la répartition des niveaux d'appartenance.

Une fonction exponentielle globale est calculée pour chaque classe C_i par l'union des s_i fonctions exponentielles définies sur les s_i sous-classes [MAS96] :

$$\mu_i(\underline{x}) = \mu_i^1 \cup ..... \cup \mu_i^{s_i} = \min\left[1, \sum_{k=1}^{s_i} \mu_i^k(\underline{x}, g_i^k, \lambda_i^k)\right] \qquad (2.10)$$

Les étapes principales de déroulement de la phase de classification par FPME sont décrites dans la figure 2.13.

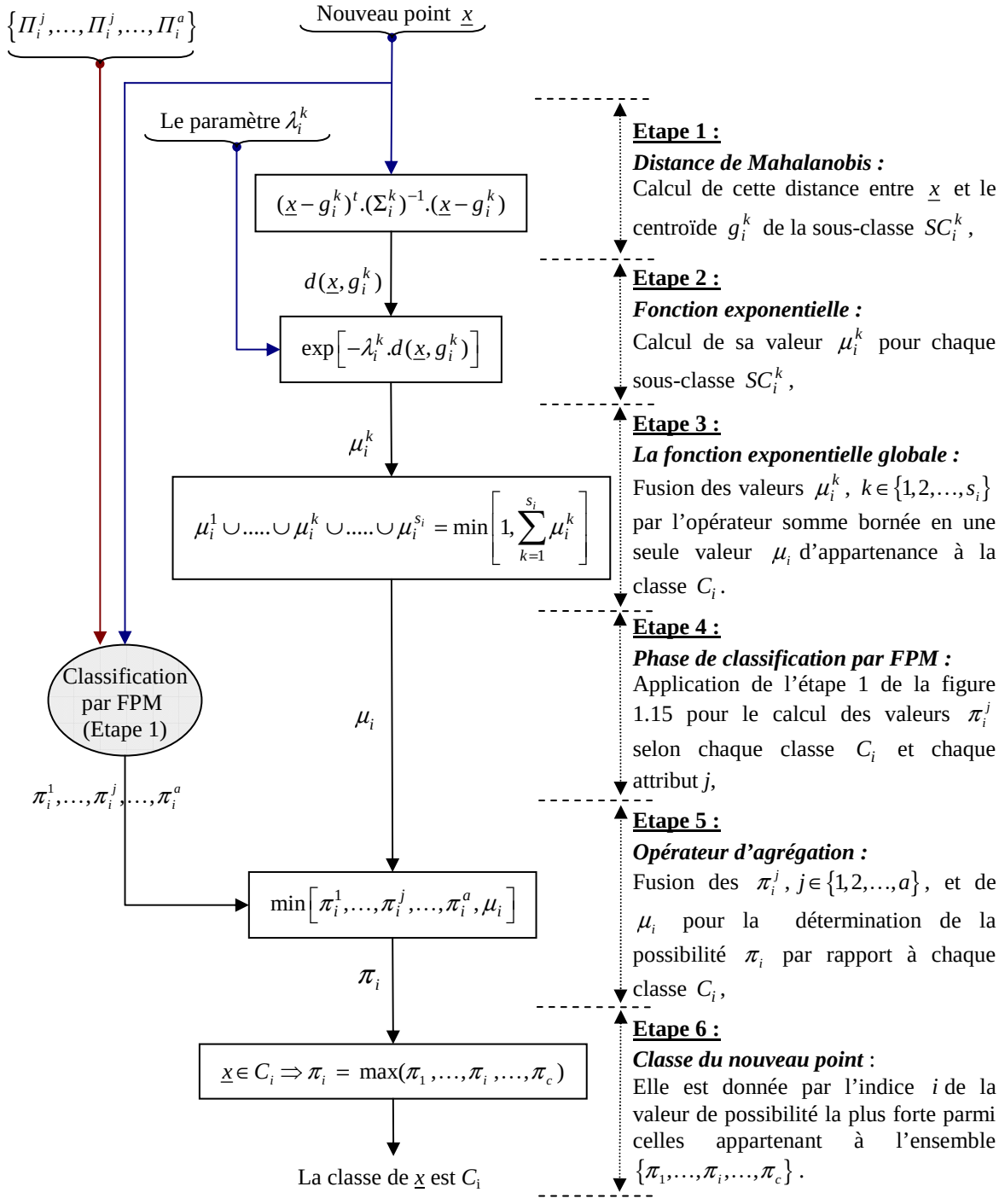


Figure 2.13 Phase de classification par FPME.

Le degré d'appartenance pour chaque classe est obtenu par agrégation de la valeur de la fonction exponentielle globale et des valeurs de possibilités sur chaque attribut. Cette agrégation est réalisée par l'opérateur minimum qui privilégie la valeur d'appartenance qui respecte la forme de la classe [DEV99] :

$$\pi_i(x) = \min\left[\pi_i^1(x^1), \pi_i^2(x^2), \pi_i^3(x^3), ..., \pi_i^a(x^a), \mu_i(\underline{x})\right] \qquad (2.11)$$

où $\pi_i^j\left(x^j\right)$ est la possibilité d'appartenance du point $\underline{x}$ pour la classe C_i selon l'attribut j.

2.4.2.3. Performances de FPME

La forme de la classe de la figure 2.10 est prise en compte dans la méthode FPME suite à sa décomposition en sous-classes convexes et grâce à l'utilisation d'une fonction exponentielle globale relative à la densité de probabilité jointe des attributs et aux calculs de la densité de possibilité marginale de la classe, cf. figure 2.14.

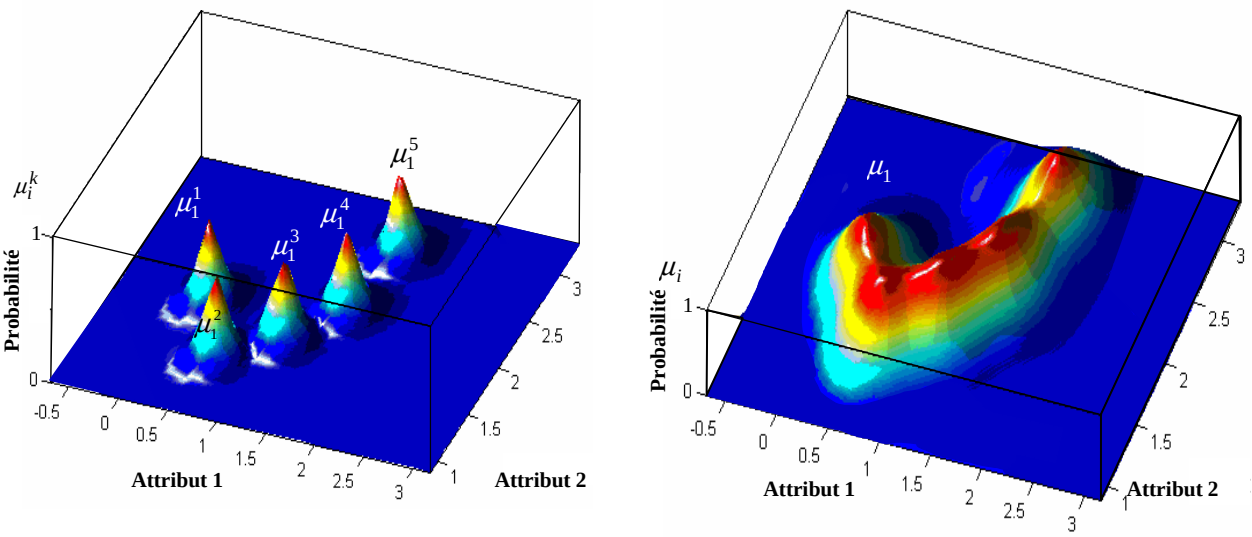


Figure 2.14 Décomposition de la classe ayant une forme non convexe en 5 sous-classes convexes ou prototypes et leur fonction exponentielle d'appartenance, à gauche. La fonction exponentielle globale de la classe est représentée à droite.

On peut constater dans la figure 2.15 que la qualité des courbes de niveaux d'appartenance, en terme du respect de la forme de la classe, est meilleure que celle obtenue par FPMM à condition d'avoir un ajustement adéquat du paramètre supplémentaire λ_i^j. En effet ce paramètre permet d'étaler la courbe de la fonction exponentielle par rapport à son centroïde. La valeur 0,1 est satisfaisante à l'inverse de la valeur 0,4 qui est trop forte, dans ce cas les courbes sont très proches des points, cf. figure 2.15 à droite.

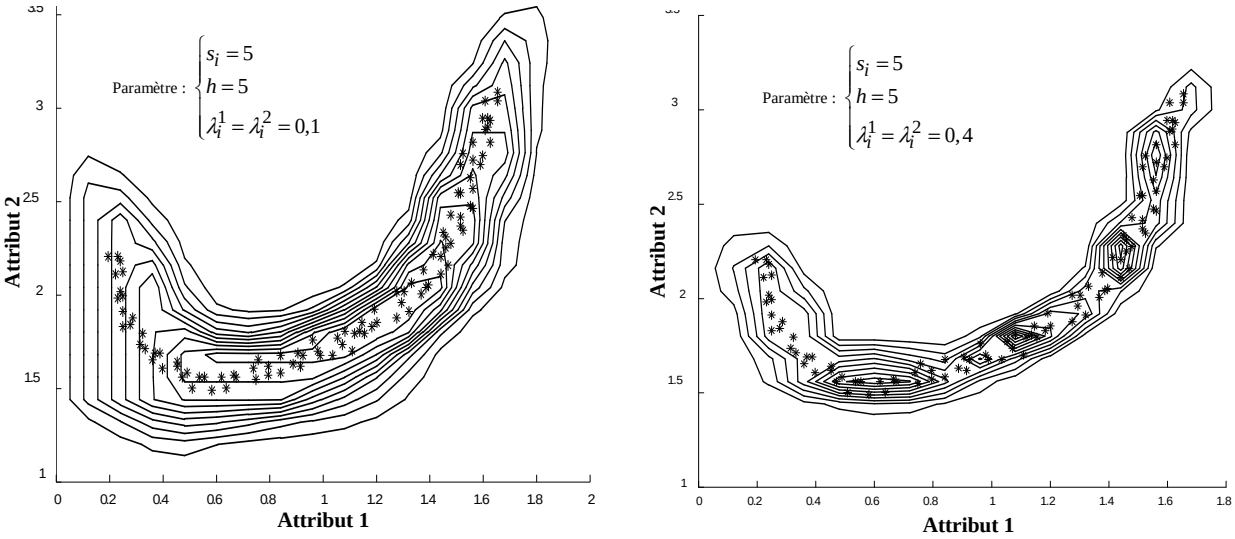


Figure 2.15 Courbes de niveaux d'appartenance obtenues par FPME pour l'exemple de la figure 2.10. Les paramètres h, s_1 et λ_1^j sont fixés respectivement à 5, 5 et 0,1, à gauche, et à 5, 5 et 0,4, à droite.

La méthode FPME utilise, d'une part FPM pour calculer les densités de possibilité pour chaque classe et chaque attribut et d'autre part FCM pour calculer les centroïdes et les matrices de covariance floues des sous-classes. La classification d'un nouveau point, conformément à FPM, nécessite en plus le calcul d'une fonction d'appartenance qui contient l'information sur la forme de la classe. Cette solution permet d'une part l'application de FPM aux classes de forme complexe et d'autre part de remédier au problème de FPMM qui est, rappelons le, le manque de finesse dans le respect de la forme de la classe.

Cependant, FPME souffre de certains inconvénients. En plus des inconvénients de FPMM, vus précédemment, elle nécessite deux paramètres supplémentaires à ajuster par rapport à FPM. Ces derniers, ne sont pas toujours faciles à déterminer.

2.4.3. Fuzzy Pattern Matching utilisant la méthode des fenêtres de Parzen

Une solution pour rendre FPM opérante dans le cas des données décrites par des attributs corrélés a été présentée dans [CAD04]. Cette solution se base sur l'utilisation de la méthode des fenêtres de Parzen, décrite dans le chapitre 1, pour estimer la probabilité conditionnelle de chaque classe selon chaque attribut. Cette estimation est basée sur la recherche de tous les points d'une classe qui sont à l'intérieur d'une fenêtre définie par un centre et une largeur *w* (qui dépend de l'ensemble d'apprentissage). La probabilité d'appartenance d'un nouveau point à une classe C_i sera calculée en considérant ce point comme le centre de la fenêtre. Cette probabilité est égale au nombre de points appartenant à cette classe (à l'intérieur de la fenêtre *w*) divisé par le nombre de tous les points (à l'intérieur de la même fenêtre). Cette solution nécessite, comme FPM, deux phases : la phase d'apprentissage et la phase de classification.

2.4.3.1. Phase d'apprentissage

Cette méthode utilise les fenêtres de Parzen, à la place des histogrammes de possibilités pour estimer la densité de probabilité de chaque classe selon un attribut principal. Chaque point de l'ensemble d'apprentissage est classifié selon cet attribut principal. Si un point est

mal classé, il faut faire appel à un autre attribut auxiliaire afin de bien le classifier. Les points qui nécessitent un attribut auxiliaire sont ceux qui évidemment sont corrélés par rapport à ces deux attributs. Pour chacun de ces points, le numéro de l'attribut auxiliaire j' nécessaire est mémorisé dans une cellule mémoire associée à ce point.

Les étapes principales du déroulement de cette phase sont décrites dans la figure 2.16. Cette phase se résume donc à chercher les points qui sont corrélés dans l'ensemble d'apprentissage. Dans ce cas, il faut trouver un autre attribut, en plus de l'attribut principal, qui pourra mieux classer ces points. Pour les points qui ne sont pas corrélés, seul l'attribut principal est pris en compte.

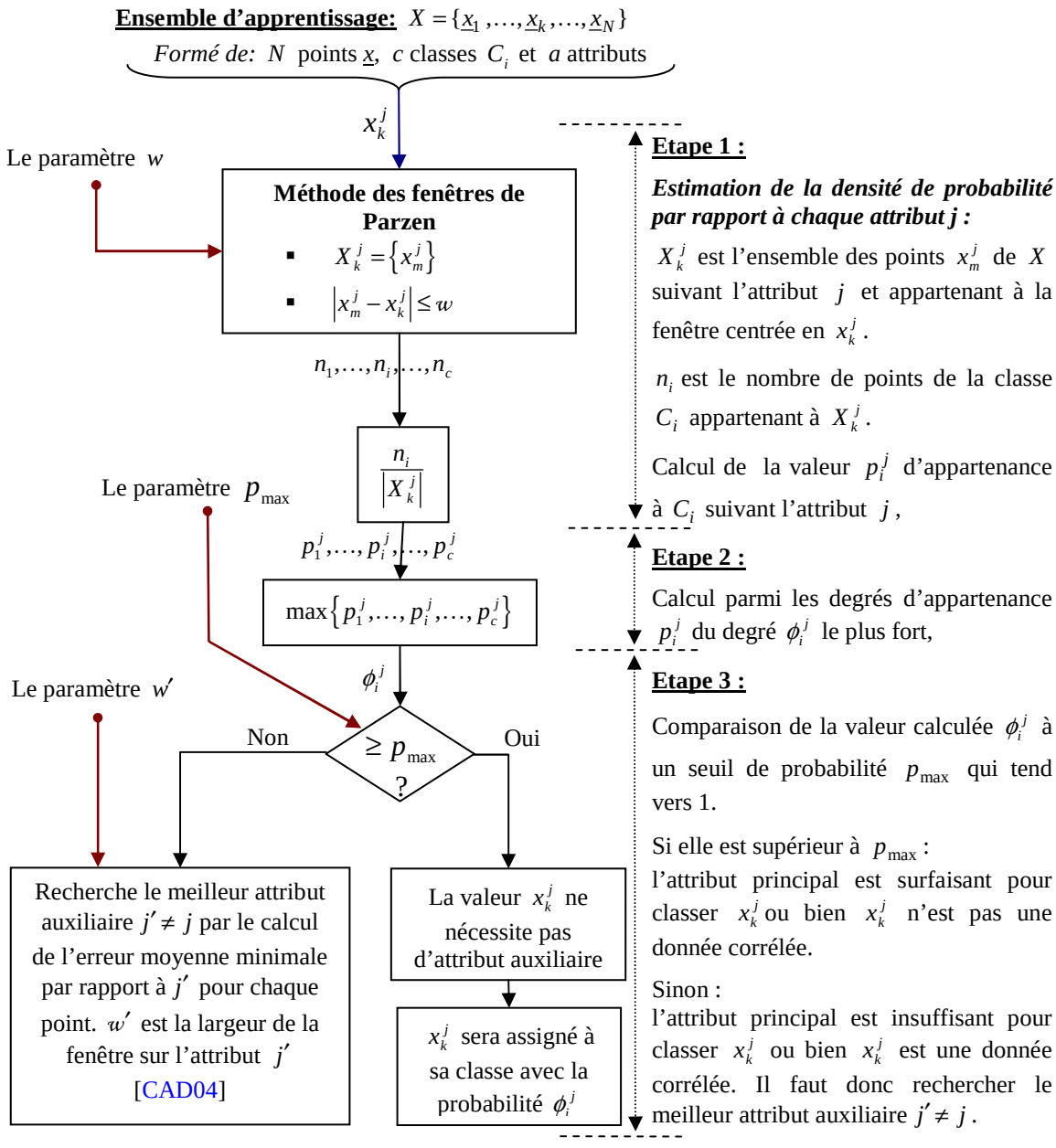


Figure 2.16 Phase d'apprentissage par FPM utilisant les fenêtres de Parzen.

2.4.3.2. Phase de classification

La phase de classification, illustrée par la figure 2.17, commence par la recherche, dans l'ensemble d'apprentissage, du plus proche voisin x_k^j du nouveau point $\underline{x}$ par rapport à l'attribut principal j. Si la cellule mémoire associée à x_k^j est vide (pas de besoin d'un attribut auxiliaire j'), alors la classification du nouveau point $\underline{x}$ sera réalisée en utilisant uniquement l'attribut principal j.

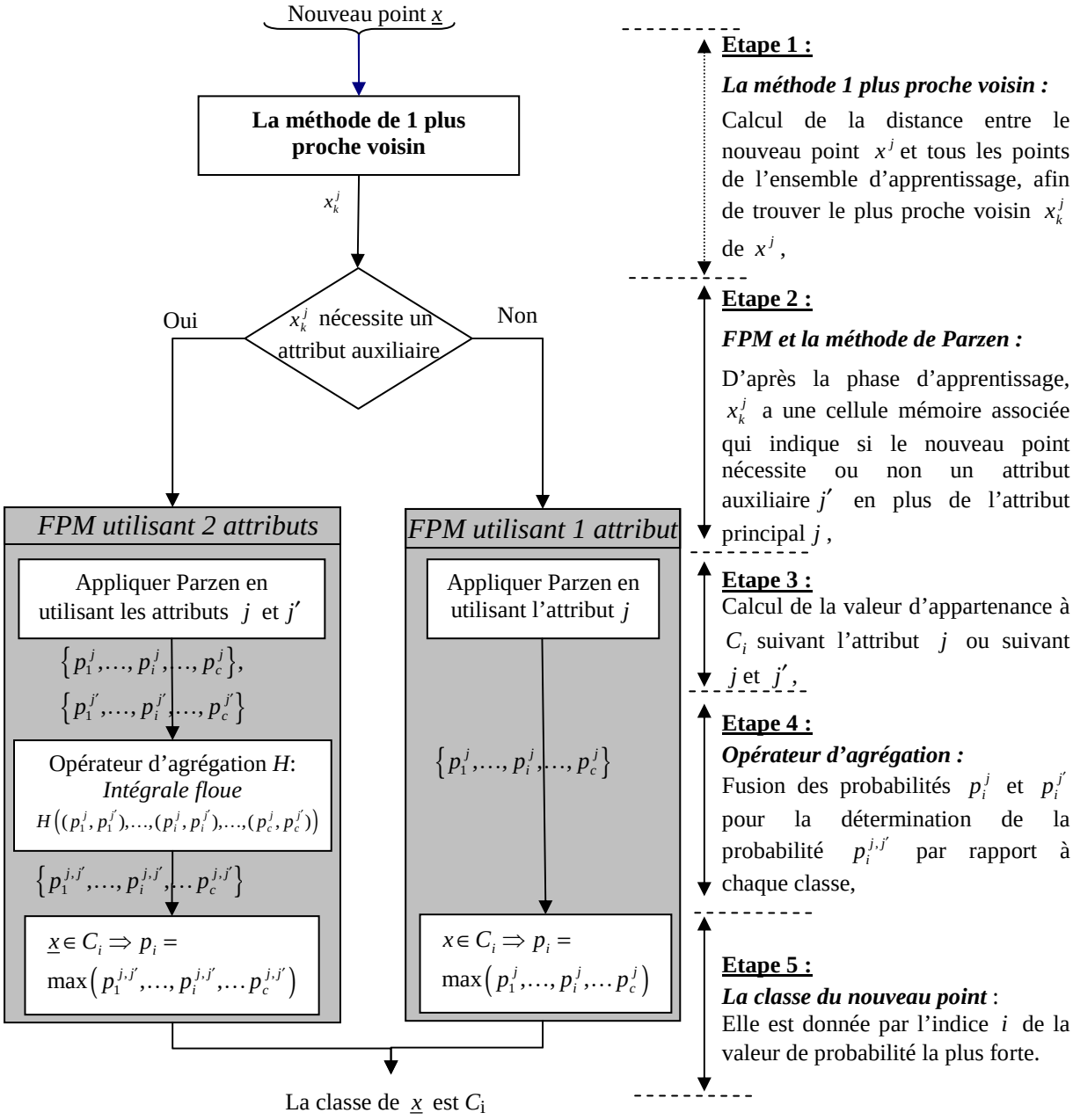


Figure 2.17 Phase de classification de la méthode FPM utilisant les fenêtres de Parzen.

L'estimation de la probabilité d'appartenance de $\underline{x}$ à chaque classe est réalisée en utilisant la méthode de Parzen uniquement sur cet attribut principal. La classe de $\underline{x}$ sera celle pour laquelle la probabilité d'appartenance est maximale. Par contre si un attribut auxiliaire est

nécessaire alors la méthode de Parzen sera appliquée sur ces deux attributs j et j' afin de calculer la probabilité d'appartenance de $\underline{x}$ à chaque classe selon chacun de ces deux attributs. La probabilité d'appartenance de $\underline{x}$ à chaque classe sera déterminée en fusionnant les deux probabilités d'appartenance par rapport aux attributs principal et auxiliaire. Cette fusion est réalisée en utilisant l'opérateur d'agrégation intégrale floue. Enfin, la classe de $\underline{x}$ est celle qui a la probabilité d'appartenance maximale.

2.4.3.3. Performances de FPM utilisant la méthode des fenêtres de Parzen

Cette solution permet de prendre en compte la corrélation entre deux attributs et de respecter la forme non-convexe des classes vue qu'elle est basée sur la méthode de Parzen. Cependant, cette solution souffre de certains inconvénients. En plus des inconvénients de la méthode de Parzen, vus dans le chapitre 1, elle nécessite deux paramètres supplémentaires à ajuster. D'une part, le paramètre $p_{\max}$ qui est le seuil de probabilité d'appartenance à partir duquel la décision ne nécessite pas un attribut auxiliaire. D'autre part le paramètre w' qui est la taille de la fenêtre selon l'attribut auxiliaire. Ces derniers, ne sont pas toujours faciles à déterminer et compliquent d'avantage la méthode. De plus, la recherche de l'attribut auxiliaire en minimisant l'erreur moyenne est une procédure qui alourdi d'avantage l'apprentissage et induit une complexité exponentielle en fonction du nombre d'attributs. Cette solution est également consommatrice en temps, car elle utilise la règle de 1 plus proche voisin dans la phase de classification ainsi que l'intégrale floue en tant qu'opérateur d'agrégation. Enfin, cette solution est inopérante pour certains problèmes en dimension supérieure à 2, comme celui de la figure 2.18.

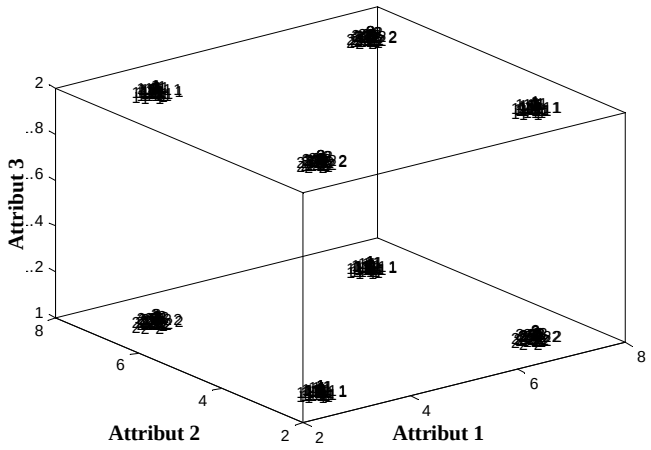


Figure 2.18 Exemple d'un problème nécessitant la prise en compte simultanée de la corrélation entre tous les attributs.

En effet, cette figure montre un problème double XOR dans $\Re^3$ (XOR3). Les données constituant les deux classes de ce problème sont générées à partir d'une même distribution gaussienne, utilisant une moyenne de $(2, 2, 0)^T$, $(6, 6, 0)^T$, $(2, 6, 1)^T$ et $(6, 2, 1)^T$ pour la classe 1 et de $(2, 2, 1)^T$, $(6, 6, 1)^T$, $(2, 6, 0)^T$ et $(6, 2, 0)^T$ pour la classe 2. La variance de chacune de ces classes est égale à 0,02. Chaque classe contient 60 points. Pour ce problème les classes sont confondues par rapport à toutes les combinaisons de 2 attributs. La discrimination de ces classes nécessite donc la prise en compte simultanément de la corrélation entre les 3 attributs.

2.4.4. Fuzzy Pattern Matching Corrélée (FPMC)

Sayed Mouchaweh [SAY02a] a proposé d'intégrer une information sur la corrélation des données en se basant sur la combinaison des barres des histogrammes construits par rapport à chaque attribut. Plus précisément en prenant en compte toutes les combinaisons entre les barres du premier attribut avec les autres barres des autres attributs. L'avantage de cette approche est de pouvoir intégrer l'information sur la corrélation des attributs sans l'ajout de paramètre supplémentaire à la méthode FPM, d'où l'appellation de FPM Corrélée (FPMC). Cependant, FPMC manque de formalisation et de plus elle a été testée uniquement dans le cas des classes convexes. Sa complexité en fonction du nombre de paramètres de l'espace de représentation n'était pas calculée et ses performances ne sont pas comparées aux autres méthodes de RdF et en utilisant une base de données variée. C'est pourquoi nous allons d'abord la formaliser afin d'en déduire ses limites et ensuite l'améliorer. Nous appellerons la nouvelle version de FPMC, FPM Améliorée (FPMA) détaillée ci-dessous.

2.5. Solutions proposées pour FPM

2.5.1. FPM Améliorée (FPMA) utilisant un apprentissage binaire

Nous proposons une solution pour remédier aux limites de FPM [BOU07a, BOU07b]. Cette solution est une version formalisée de FPMC qui tient compte de l'importance (cardinalité) de chaque classe. Nous appelons FPM après l'intégration de cette solution, FPM Améliorée (FPMA). FPMA enrichit l'information donnée par les distributions de possibilités marginales, par une information relative à la distribution conjointe. Cette dernière décrit la corrélation entre les attributs. Elle est représentée par une matrice binaire indiquant la corrélation entre les barres de l'histogramme du premier attribut et toutes les autres barres des autres attributs pour chaque classe en utilisant l'ensemble d'apprentissage. Cette solution ne nécessite la détermination d'aucun paramètre supplémentaire. Son fonctionnement comporte également, comme FPM, deux phases : les phases d'apprentissage et de classification qui sont détaillées ci-dessous.

2.5.1.1. Phase d'apprentissage

La phase d'apprentissage de FPMA est similaire à celle de FPM mais elle intègre en plus une information sur le positionnement des points de l'ensemble d'apprentissage X dans $\Re^a$ en calculant leurs barres par rapport à tous les attributs. Rappelons que chaque histogramme pour chaque attribut *j*, contient un nombre *h* de barres numérotées $b_{k_j}^j$, $k_j \in \{1,...,h\}$. La matrice de corrélation *B* pour l'ensemble X est définie comme suit :

$$B = [B^1,...,B^i,...,B^c] \tag{2.12}$$

où B^i est la matrice de corrélation pour la classe C_i, calculée comme suit :

$$B^i = [\alpha_b^i \in \{1,0\}] \tag{2.13}$$

où α_b^i est le facteur de corrélation pour l'hypercube $b = [b_{k_1}^1,\ldots,b_{k_j}^j,...,b_{k_a}^a]$ formé par l'intersections de toutes les barres $b_{k_j}^j$ de tous les attributs $j \in [1,...,a]$ suivant la classe C_i. Ce facteur de corrélation et calculé par :

$$\forall \underline{x} \in X,\ Classe(\underline{x}) = C_i :$$
$$\alpha_b^i = \begin{cases} 1 & \text{si } \underline{x} \in b_{k_1}^1 \cap \ldots \cap b_{k_1}^j \cap \ldots \cap b_{k_a}^a \\ 0 & \text{sinon} \end{cases} \qquad (2.14)$$

Ce facteur traduit donc la présence ou l'absence de points $\underline{x}$ d'apprentissage, appartenant à la classe C_i, dans chaque hypercube *b*. '$\cap$' est l'opérateur d'intersection. La dimension de la matrice de corrélation B, dans ce cas, est de $h^a \times c$. Afin, de diminuer la complexité de la solution proposée, les hypercubes sont définis par l'intersection des barres du premier attribut et les barres de chacun des autres attributs. Plus précisément chaque hypercube est formé par l'intersection de la barre $b_{k_1}^1$ de l'attribut 1 et chacune des barres des autres attributs suivant la classe C_i, ce qui réduit la dimension de la matrice B à $h^2 \times (a-1) \times c$ hypercubes à déterminer. Le facteur de corrélation, dans ce cas, est calculé par :

$$\forall \underline{x} \in X,\ Classe(\underline{x}) = C_i :$$
$$\alpha_b^i = \begin{cases} 1 & \text{si } \underline{x} \in \left\{ b_{k_1}^1 \cap b_{k_2}^2 \wedge \ldots \wedge b_{k_1}^1 \cap b_{k_j}^j \wedge \ldots \wedge b_{k_1}^1 \cap b_{k_a}^a \right\} \\ 0 & \text{sinon} \end{cases} \qquad (2.15)$$

où '$\wedge$' est l'opérateur logique ET.

La figure 2.19 présente un exemple académique illustrant le calcul de la matrice de corrélation B pour le cas d'un espace de représentation contenant deux classes C_1 et C_2 décrites par deux attributs.

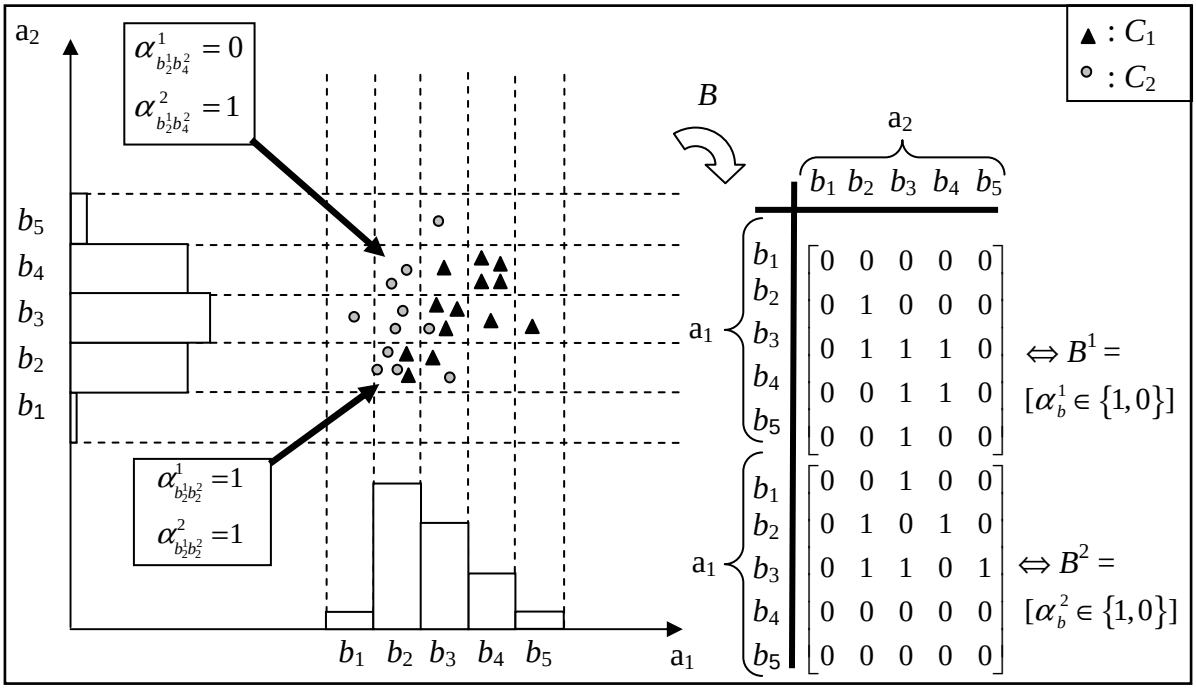


Figure 2.19 Matrice de corrélation B obtenue par l'apprentissage binaire de FPMA pour le nuage de points contenant deux classes C_1 et C_2 dans $\Re^2$.

Les barres b_2^1 et b_4^2 sont corrélés uniquement selon la classe C_2. En effet, l'hypercube formé par leur intersection ne contient que deux points d'apprentissage appartenant à la classe C_2 et aucun point appartenant à la classe C_1. Les facteurs de corrélation $\alpha^1_{b_2^1 b_4^2}$ par rapport à la classe 1 et $\alpha^2_{b_2^1 b_4^2}$ par rapport à la classe 2, correspondants à cet hypercube, auront respectivement, dans la matrice B, les valeurs 0 et 1.

La figure 2.20 représente les étapes nécessaires pour calculer la matrice de corrélation B.

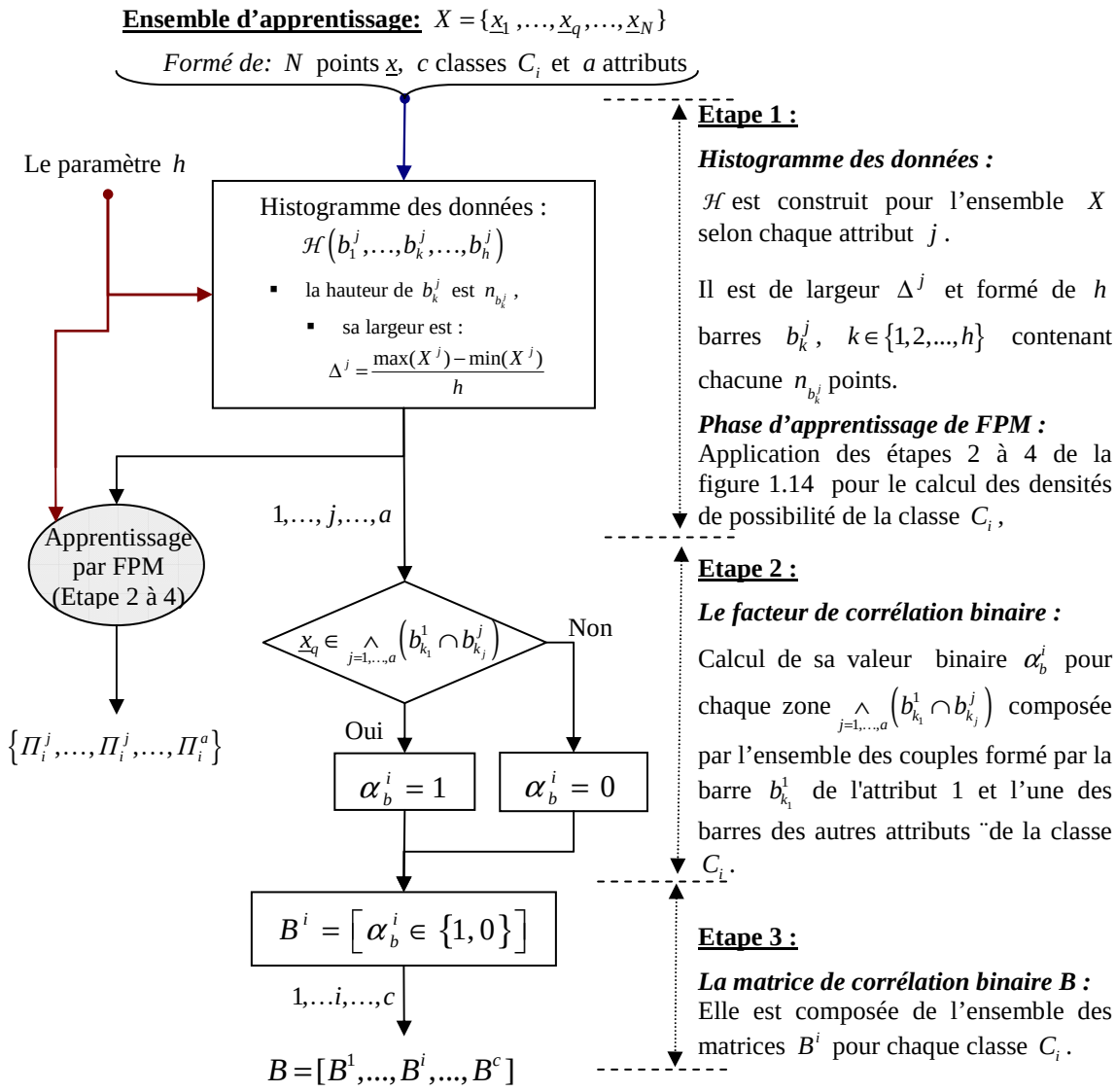


Figure 2.20 Phase d'apprentissage de FPMA avec un apprentissage binaire.

2.5.1.2. Phase de classification

Les étapes nécessaires pour la classification d'un nouveau point $\underline{x}$ sont représentées dans la figure 2.21. Premièrement, le repérage de $\underline{x}$ dans l'espace de représentation est réalisé par le calcul de l'ensemble des barres $\left\{b_{k_1}^1,\ldots,b_{k_j}^j,\ldots,b_{k_a}^a\right\}$ auxquelles il appartient par rapport à

chaque attribut. Cet ensemble délimite un hypercube $b_{\underline{x}} = [b^1_{k_1}, \ldots, b^j_{k_j}, ..., b^a_{k_a}]$ défini par l'intersection de ces barres. Ensuite, le facteur de corrélation $\alpha^i_{b_{\underline{x}}}$ relatif à l'hypercube $b_{\underline{x}}$ est recherché dans la matrice de corrélation B résultante de la phase d'apprentissage. Enfin, la valeur de possibilité d'appartenance à chaque classe est calculée comme dans le cas de FPM si et seulement si $\alpha_{b_{\underline{x}}} = 1$ par rapport à cette classe. Sinon le point $\underline{x}$ sera rejeté selon la classe C_i, c'est-à dire $\pi_i = 0$.

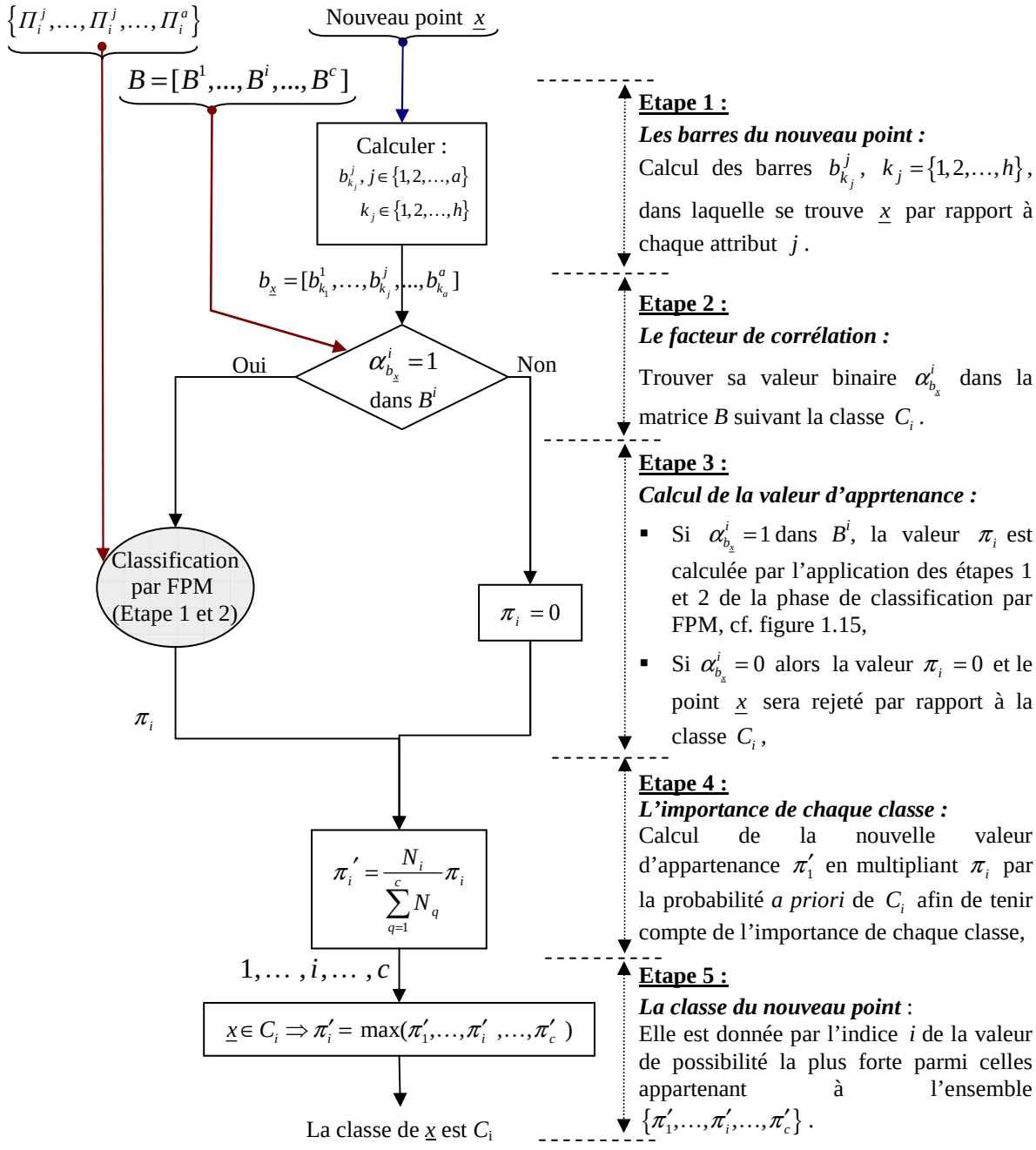


Figure 2.21 Phase de classification de FPMA avec un apprentissage binaire.

Afin de tenir compte de l'importance de chaque classe, la possibilité d'appartenance π_i de $\underline{x}$ à la classe C_i, calculée par FPM, est multipliée par la probabilité *a priori* de C_i :

$$\pi_i' = \frac{N_i}{\sum_{j=1}^{c} N_j} \pi_i \qquad (2.16)$$

Enfin le point $\underline{x}$ sera assigné à la classe pour laquelle il a la valeur de possibilité d'appartenance la plus élevée.

2.5.1.3. Evaluation des performances de FPMA avec un apprentissage binaire

Exemple des matières plastiques

La figure 2.22 présente les courbes de niveaux d'appartenance issues de l'application de FPMA avec un apprentissage binaire pour la classification des matières plastiques. Nous pouvons constater que ces courbes respectent la forme oblique de la classe C_1.

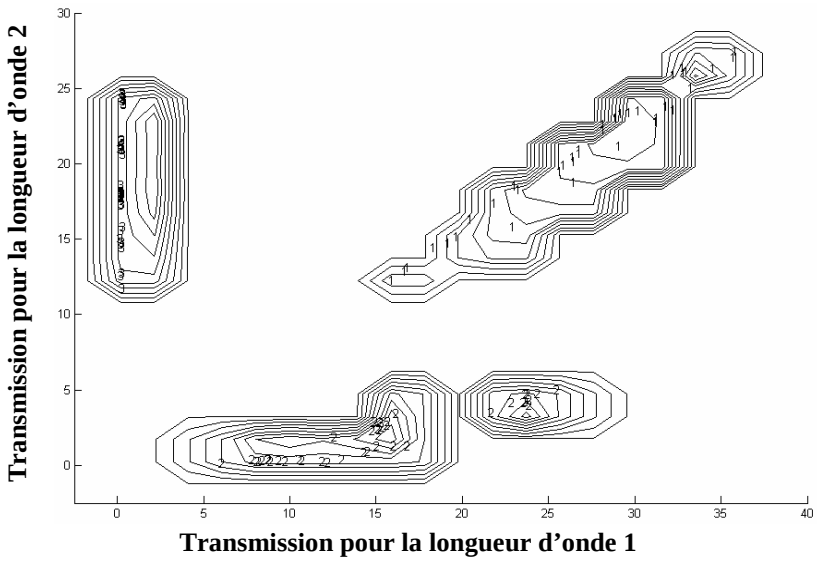
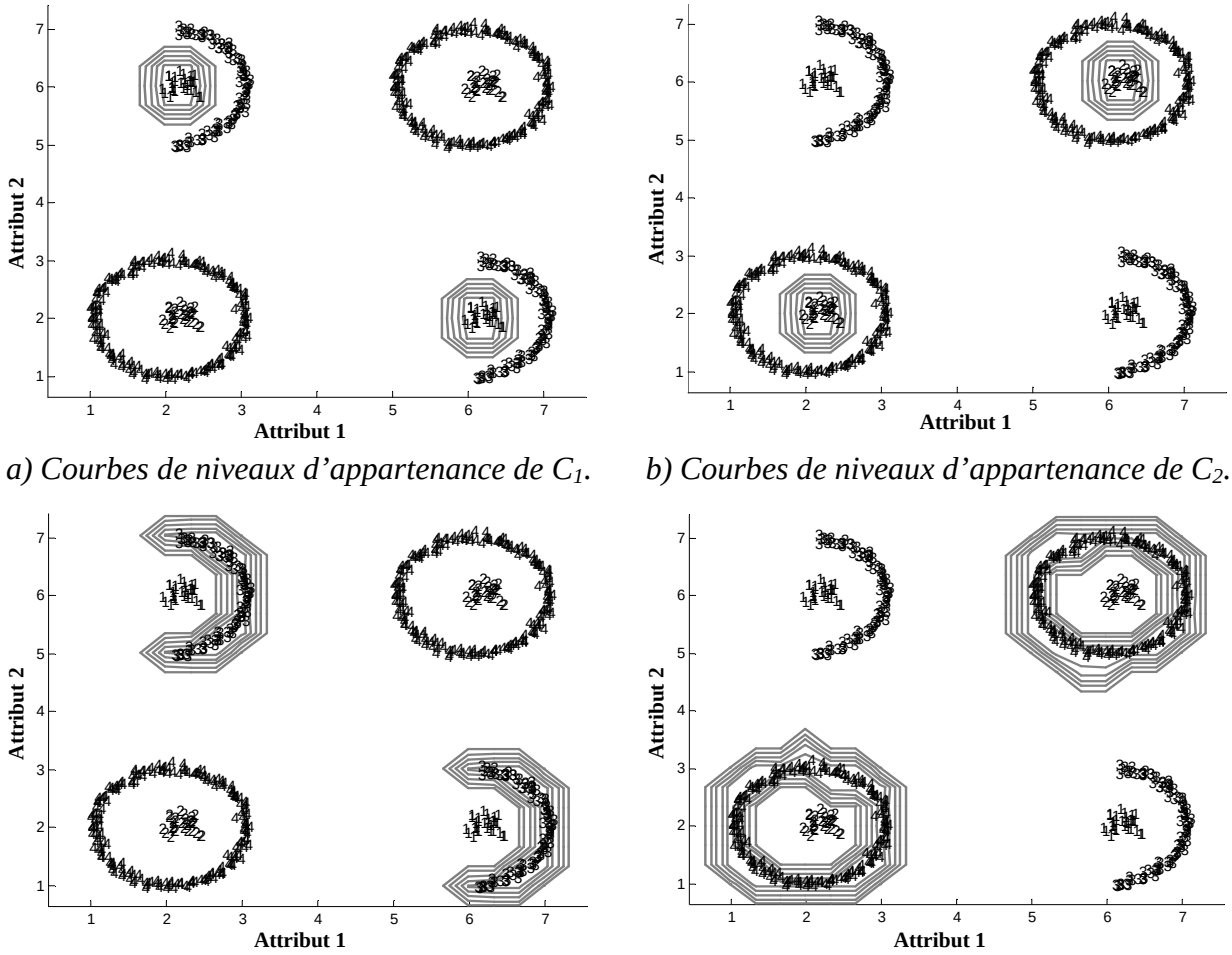


Figure 2.22 Courbes de niveaux d'appartenance, obtenues par FPMA, pour l'exemple de matières plastiques avec h=10.

Problème double XOR

Les figures ci-dessous présentent les courbes de niveaux d'appartenance obtenues par l'application de FPMA avec un apprentissage binaire sur le problème double XOR. Nous constatons, d'après la figure 2.23.a, que FPMA distingue les points de la classe C_1 de ceux des autres classes. Nous constatons la même chose, dans les figures 2.23.b, 2.23.c et 2.23.d respectivement, pour les classes 2, 3 et 4. De plus, ces courbes de niveaux, contrairement à FPM, respectent la forme non convexe des classes 3 et 4.



a) Courbes de niveaux d'appartenance de C_1. *b) Courbes de niveaux d'appartenance de C_2.*

c) Courbes de niveaux d'appartenance de C_3. *d) Courbes de niveaux d'appartenance de C_4.*

Figure 2.23 Courbes de niveaux d'appartenance obtenues par FPMA avec apprentissage binaire pour l'exemple XOR avec des classes de forme convexe et non-convexe.

Résultats expérimentaux

Les performances de FPMA sont comparées avec celles de FPM en appliquant la technique du "leave-one-out", sur trois ensembles de données: le problème XOR, IRIS [NEW98] et les données du cancer du sein, Wisconsin Brest Cancer (WBC) [NEW98]. Cette technique donne une évaluation pessimiste sur les performances des méthodes de classification. Le tableau 2.1 décrit les caractéristiques des bases de données utilisées. Le problème XOR se compose de 2 classes, constituées chacune de 400 points dans un espace de dimension 5. Nous avons choisi la dimension 5 pour montrer les performances de FPMA dans le cas d'un espace de représentation de dimension supérieure à 2 attributs corrélés. Les données IRIS se composent de 3 classes de fleurs : Iris Sétosa, Versicolor, et Virginica. Chacune de ces classes est constituée de 50 échantillons caractérisés par 4 attributs : longueur et largeur des sépales, longueur et la largeur des pétales. La classe Iris Sétosa est bien séparée des deux autres qui se recouvrent partiellement. La troisième base de données, issue du domaine médical, concerne le diagnostic du cancer du sein. Ces données sont fournies par l'Université de Wisconsin. La base de données que nous avons utilisée est constituée de 683 échantillons. Chaque point est décrit par 9 attributs caractérisant des échantillons cytologiques du sein, classés comme bénins ou malins. Ces deux classes n'ont pas la même importance : 443 points étant bénins et 240 points étant malins.

Tableau 2.1 Les bases de données utilisées.

Données	**Classes**	**DIM**	**Pt. /Class**
XOR	2	5	{400, 400}
Iris	3	4	{50, 50, 50}
WBC	2	9	{443, 240}

Les performances de FPM et de FPMA sont évaluées en utilisant le Taux d'Erreur de Classification (TEC), le Temps de Classification (TC), et le Taux de Rejet (TR) pour indiquer le nombre de points qui ne sont assignés à aucune classe. En effet, il vaut mieux rejeter que mal classifier.

Le tableau 2.2 compare les résultats obtenus, pour les deux méthodes, en utilisant un processeur Pentium 2,8 GHz. Le nombre de barres *h*, ainsi que la tolérance *Tol*, sont déterminés expérimentalement, pour chacune des bases de donnée, afin de maximiser les performances de FPM et FPMA.

Tableau 2.2 Comparaison de FPM et FPMA, en utilisant la technique leave-one-out, suivant le Taux de Rejet (TR), le Taux d'Erreur de Classification (TEC) et le Temps de Classification d'un point (TC).

X	XOR		IRIS		WBC	
(h, Tol)	(9, 0,31)		(6, 0,31)		(2, 0)	
Méthode	FPM	FPMA	FPM	FPMA	FPM	FPMA
TR (%)	0	0	0,67	6,67	0	0
TEC (%)	44,12	0	3,33	1,33	5,12	3,66
TC $\times 10^{-3}$ (Sec)	2	2,2	2,7	2,1	3,3	4,4

Les résultats obtenus montrent que les performances de FPMA sont meilleures que celles de FPM. Nous constatons que l'augmentation du taux de rejet est très dépendante de la qualité de l'ensemble d'apprentissage et on ne peut le considérer comme une faiblesse de cette méthode car FPMA nous donne une plus grande certitude sur l'affectation d'un point à une classe que FPM. Concernant le temps de classification, FPM et FPMA ont un temps comparable pour les trois bases de données.

La solution que nous avons proposée permet à FPM, de tenir compte de la corrélation entre les attributs et de respecter la forme non convexe des classes. Cependant, cette méthode comporte plusieurs inconvénients :

- La taille de la matrice de corrélation *B* estimée à partir des données d'apprentissage a été diminuée de $h^a \times c$ à $h^2 \times (a-1) \times c$ pour moins de complexité. En effet, pour un ensemble d'apprentissage contenant une classe avec 2 attributs et *h* barres par histogramme, la taille est de h^2. Ce sera donc $2 \times h^2$ au lieu de h^3 pour 3 attributs. Cependant, elle peut devenir très vite gigantesque si le nombre d'attributs et le nombre de barres deviennent importants,
- Pour estimer correctement la matrice de corrélation *B*, il faut une base d'apprentissage riche. En effet, si un des nouveaux points à classer se trouve dans

un hypercube où aucun point d'apprentissage ne se trouvait ($\alpha^i_{b_x} = 0$), ce nouveau point sera rejeté. Il faut alors que la base d'apprentissage couvre correctement tout l'espace de représentation. Sinon un nombre important de points sera rejeté,

- FPMA se comporte exactement comme FPM dans les zones de chevauchement des classes. En effet, l'attribution d'une valeur binaire à la matrice de corrélation la rend très sensible à la qualité de l'ensemble d'apprentissage. Il suffit qu'un point de l'ensemble d'apprentissage soit présent pour que l'on affecte la valeur 1. Il n'y a aucune pondération par le nombre de points présents dans un hypercube.

Nous proposons par la suite une amélioration de FPMA afin de remédier à ces inconvénients.

2.5.2. FPMA utilisant un apprentissage flou

2.5.2.1. Phase d'apprentissage

Nous proposons une amélioration à FPMA [BOU07c] pour résoudre les problèmes liés :

- à l'utilisation d'une matrice binaire, qui ne tient pas compte du nombre de points de la même classe dans un hypercube,
- au taux de rejet sensible à la richesse de l'ensemble d'apprentissage.

La matrice de corrélation binaire devient alors un ensemble des facteurs de corrélation flous. Cet ensemble est défini pour toutes les classes comme suit :

$$B = \left\{ B^1,\ B^2, ..., B^i, ..., B^c \right\} \tag{2.17}$$

Telque B^i est l'ensemble des facteurs de corrélation flous pour la classe C_i, calculée comme suit :

$$B^i = \left\{ \alpha^i_b \in [0,1] \right\} \tag{2.18}$$

où α^i_b est le facteur de corrélation flou calculé pour l'hypercube formé par l'intersections des barres $b = [b^1_{k_1}, \ldots, b^j_{k_j}, ..., b^a_{k_a}]$. Ces facteurs sont calculés uniquement pour les hypercubes, de l'espace de représentation, contenant des points de X comme suit :

$$\forall \underline{x} \in X, C(\underline{x}) = C_i,\ \underline{x} \in b^1_{k_1} \cap ... \cap b^j_{k_j} \cap ... \cap b^a_{k_a} :$$
$$\alpha^i_b = \left(\frac{n^i_b}{n_b} \right)^{\beta} \tag{2.19}$$

où n^i_b est le nombre de points de X appartenant à la classe C_i et se trouvant dans l'hypercube constitué par les barres b. n_b est le nombre total de tous les points de X appartenant à cet hypercube. Ce facteur tient donc compte du nombre de points d'une même classe présents dans un hypercube. Plus ce nombre est petit, plus la valeur de ce facteur s'approche de zéro pour indiquer une faible corrélation entre les attributs dans cet hypercube. β est un paramètre qui permet d'ajuster le degré de flou de α^i_b. Par exemple dans la figure 2.24, l'augmentation

de la valeur de β de 1 à 5 provoque l'diminution de la valeur de α_b^i pour un nombre de points n_b^i donné.

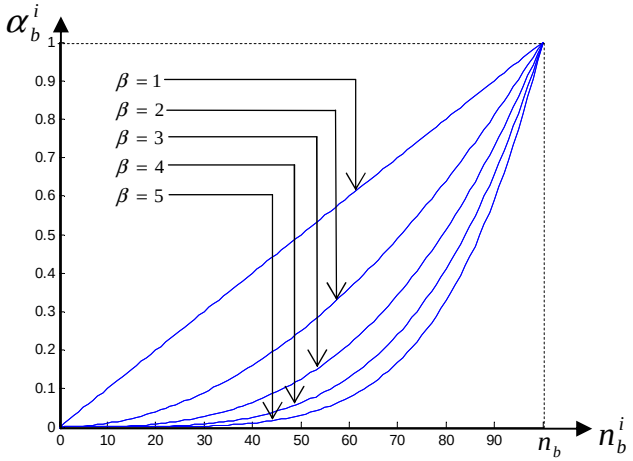


Figure 2.24 Contrôle de l'influence du nombre de points n_b^i de la même classe C_i pour le calcul du facteur de corrélation flou α_b^i via le paramètre β.

Reprenons l'exemple de la figure 2.19, cf. figure 2.25, afin d'illustrer le calcul de l'ensemble des facteurs de corrélation flous *B*.

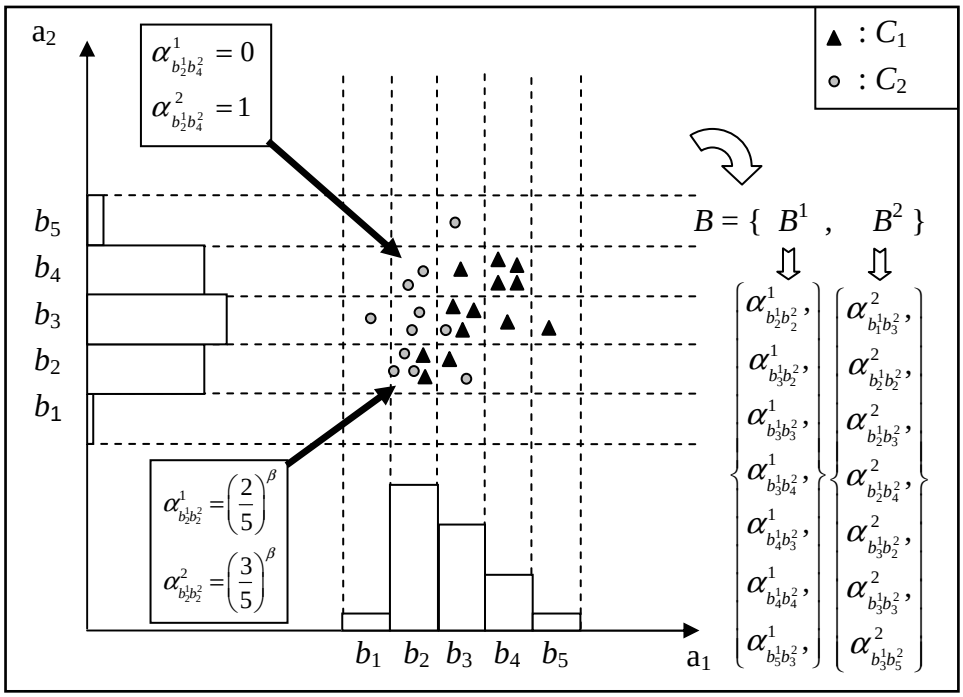


Figure 2.25 Ensemble des facteurs de corrélation flous B obtenu par l'apprentissage flou de FPMA pour le nuage de points contenant deux classes C_1 et C_2 dans $\Re^2$.

La figure 2.26 représente les étapes nécessaires pour calculer l'ensemble des facteurs de corrélation flous B pour toutes les classes.

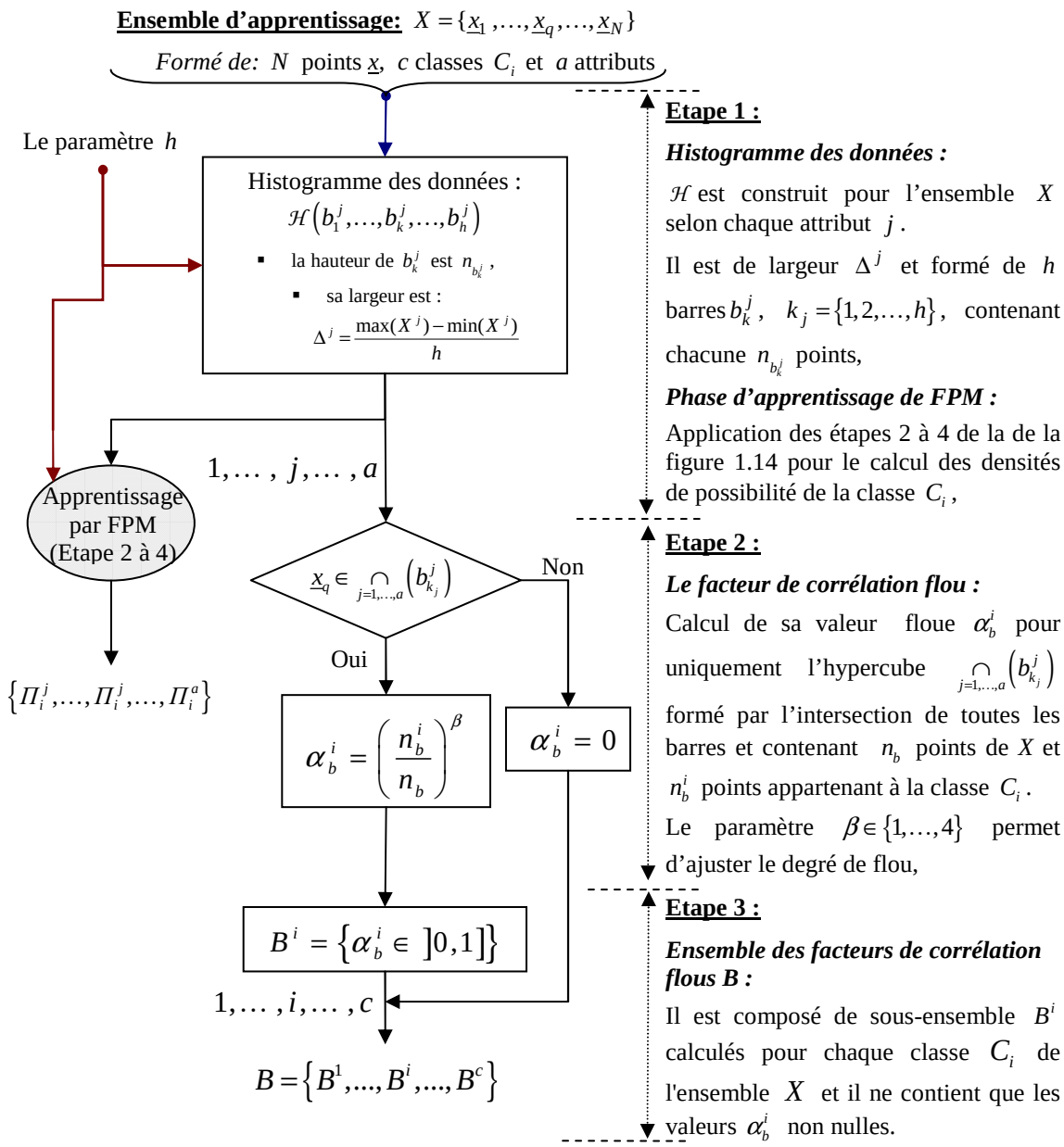


Figure 2.26 Phase d'apprentissage de FPMA avec un apprentissage flou.

Dans la figure 2.25, les barres b_2^1 et b_2^2, par exemple, sont corrélées selon les classes C_1 et C_2. En effet, la zone formée par l'intersection de ces deux barres contient deux points d'apprentissage appartenant à la classe C_1 et 3 points appartenant à la classe C_2. Les facteurs de corrélations correspondants à cette zone ou hypercube $\alpha^1_{b_2^1 b_2^2}$ et $\alpha^2_{b_2^1 b_2^2}$ auront donc, dans l'ensemble B, les valeurs $(2/5)^2=0{,}16$ et $(3/5)^2=0{,}36$ respectivement pour une valeur de $\beta=2$. Cela veut dire que ces barres sont corrélées selon toutes les classes. De plus le facteur de corrélation flou, contrairement au facteur binaire, renseigne sur le degré de corrélation des attributs 1 et 2 par rapport à chaque classe. En effet, le degré de corrélation dans cet hypercube, formé par l'intersection des barres b_2^1 et b_2^2, est plus important par rapport à C_2 que par rapport à C_1. Pour les barres b_2^1 et b_4^2, il n'y a que deux points de la classe C_2 localisés dans la zone de leur intersection et les facteurs de corrélation correspondant à cette

zone auront donc les valeurs 0 et 1 pour les classes C_1 et C_2 respectivement. Seule la valeur non nulle du facteur de corrélation de la classe C_2 est mémorisée dans B. Finalement l'ensemble des facteurs de corrélation flous B contient 14 facteurs de corrélation à mémoriser au lieu de 50, ou 25 facteurs pour chaque classe. En effet, il y'a 25 hypercubes dans l'espace de représentation, cf. figure 2.19.

2.5.2.2. Phase de classification

Les étapes nécessaires pour la classification d'un nouveau point $\underline{x}$ sont représentées dans la figure 2.27.

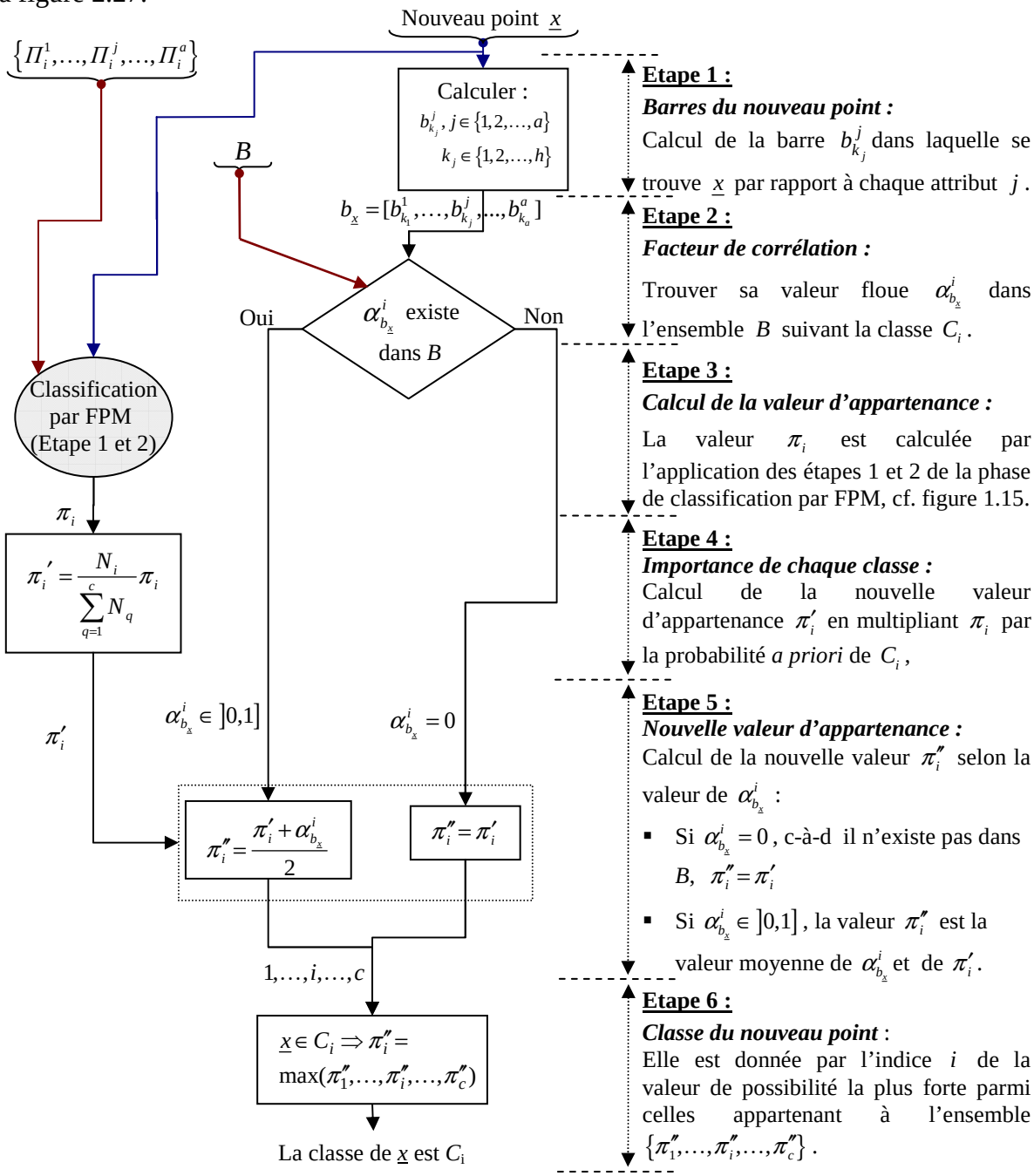


Figure 2.27 Phase de classification de FPMA avec un apprentissage flou.

La classification d'un nouveau point $\underline{x}$, commence par la détermination de ses barres $b_{\underline{x}} = [b^1_{k_1}, b^2_{k_2}, ..., b^a_{k_a}]$. Ensuite, le facteur $\alpha^i_{b_{\underline{x}}}$ pour ces barres $b_{\underline{x}}$ est recherché dans l'ensemble B^i des facteurs de corrélation de la classe C_i. Si l'hypercube constitué par ces barres n'existe pas dans B^i alors $\alpha^i_{b_{\underline{x}}} = 0$. La valeur d'appartenance de ce point $\underline{x}$ par rapport à la classe C_i est calculée donc par :

$$(\pi_i)_{FPMA} = \pi''_i = \begin{cases} \dfrac{\pi'_i + \alpha^i_{b_{\underline{x}}}}{2} & \text{si } \alpha^i_{b_{\underline{x}}} \neq 0 \\ \pi'_i & \text{sinon} \end{cases} \qquad (2.20)$$

La moyenne arithmétique permet de tenir compte de l'information marginale, possibiliste, et de l'information conjointe, facteur de corrélation, avec la même importance. L'information possibiliste est plus riche puisqu'elle est calculée en utilisant tout l'ensemble d'apprentissage, contrairement à l'information conjointe qui est calculée en utilisant les points d'apprentissage localisés dans un hypercube. Afin de tenir compte de cette différence pendant l'agrégation, une solution consiste à diminuer l'influence du facteur de corrélation en augmentant la valeur de β dans l'équation (2.19). L'étude sur plusieurs bases de données avec des valeurs de β variant entre 1 et 4, a montré que la valeur 2 donne en général les meilleurs résultats.

2.5.2.3. Performances de FPMA avec un apprentissage flou

Le taux d'erreur de classification de FPMA est comparé avec celui de FPM ainsi qu'avec les méthodes de k plus proches voisins (kppv), de kppv Floue (kppvF) et de la méthode des noyaux de Parzen, en appliquant la technique du "leave-one-out", sur trois ensembles de données :

- le problème XOR3 représenté dans la figure 2.18,
- les données Vibration [KNI72] issues du domaine industriel qui sont caractérisées par deux classes très voisines comme le montre la figure 2.28,

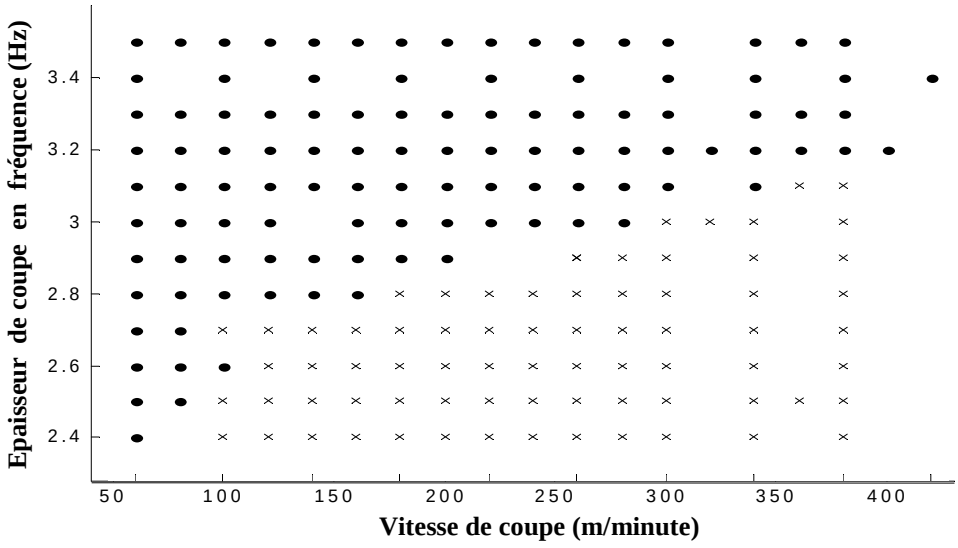


Figure 2.28 Données Vibration dans $\Re^2$. " • " et " × " indiquent les points des classe C_1 et C_2 respectivement.

- les données Iris en tenant compte des attributs 3 et 4 (Iris34). Les résultats obtenus pour Iris en utilisant ces deux paramètres sont meilleurs que ceux obtenus en utilisant toutes les autres combinaisons possibles. En effet, il est démontré que ces deux attributs sont les plus informatifs pour discriminer les trois classes [GRA00] comme le montre la figure 2.29. Nous pouvons constater que les densités de possibilités les mieux séparées par rapport à chaque attributs sont celles des attributs 3 et 4.

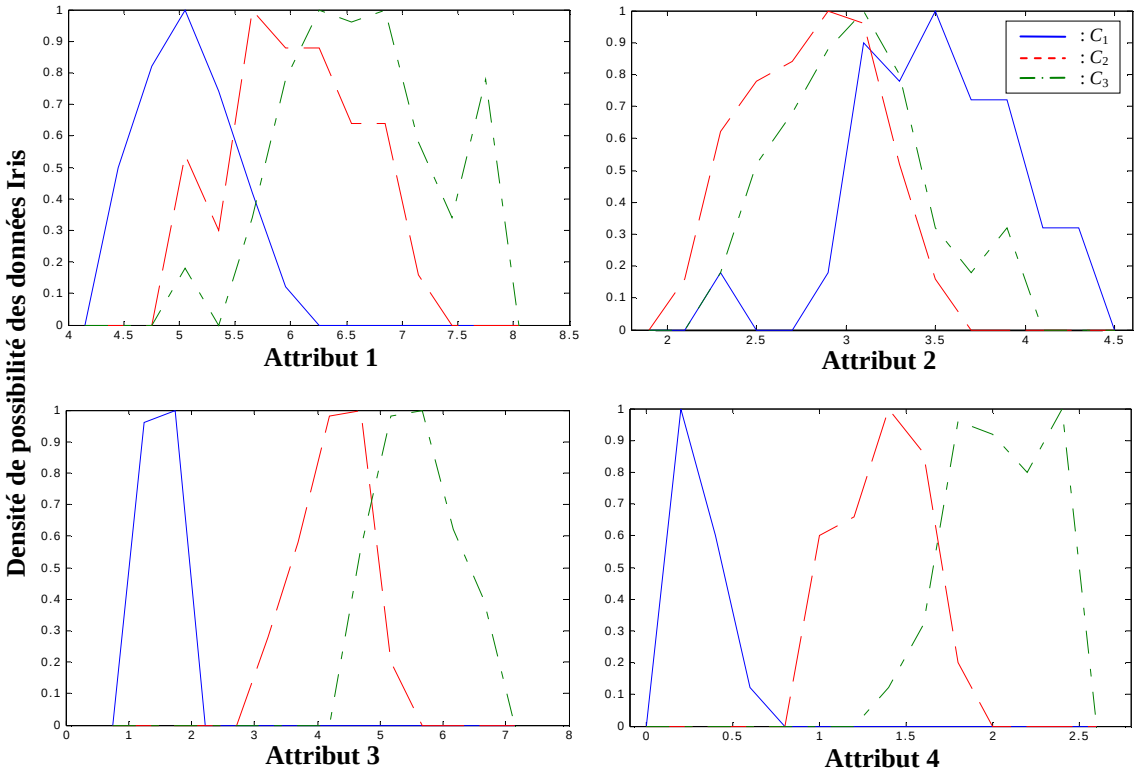


Figure 2.29 Densités de possibilité pour les trois classes de données Iris par rapport à chaque attribut.

La dimension de l'espace de représentation, le nombre de classes de chaque base de données ainsi que le nombre de points de chaque classe qui caractérise chaque base de données sont représentées dans le tableau 2.3. Le tableau 2.4 compare les résultats obtenus, par les cinq méthodes.

Tableau 2.3 Les bases de données utilisées.

Données	**Classes**	**DIM**	**Pt. /Class**
Iris34	3	2	{50, 50, 50}
Vibration	2	2	{107, 73}
XOR3	2	3	{60, 60}

Tableau 2.4 Taux d'erreur de classification pour les trois bases de données utilisées.

	TEC(%)		
Méthode	**Iris34**	**Vibration**	**XOR3**
kppv (k)	3,33 (6)	8,89 (3)	0 (1)
kppvF (k)	4,67 (1)	13,21 (1)	0 (1)
Parzen (w)	3,33 (0,2)	7,78 (0,05)	0 (0,05)
FPM (h, Tol)	2,67 (12, 0,03)	6,67 (5, 0,29)	100 (10, 0)
FPMA (h, Tol)	1,33 (12, 0,03)	2,78 (9, 0,03)	0 (10, 0,03)

Les plages de variation du nombre de barres h et du nombre de plus proches voisins k sont respectivement [2, 15] et [1, 15]. La largeur de la fenêtre de Parzen w et de la tolérance Tol varient dans l'intervalle [0, 0,4] avec un pas de 0,01. L'objectif est de trouver expérimentalement les valeurs optimales de ces paramètres qui minimisent le Taux d'Erreur de Classification (TEC). Le noyau de Parzen est hypercubique de type gaussien. En effet, le noyau gaussien est le noyau le plus utilisé en RdF. Le choix de ce type de noyau est justifié principalement par les très bons résultats qu'il donne en général pour différentes bases de données [GUI05]. A travers le tableau 4, nous constatons que FPMA donne le plus faible taux d'erreur de classification par rapport aux autres méthodes et pour toutes les bases de données utilisées.

2.6. Conclusion

Nous avons proposé une solution pour que la méthode Fuzzy Pattern Matching (FPM), soit opérante dans le cas des classes de forme non-convexe et/ou décrites par des attributs corrélés. Cette solution est une version formalisée de la méthode FPM Corrélée (FPMC) présentée dans [SAY02a]. Elle recherche la corrélation entre les barres des différents histogrammes par rapport à tous les attributs selon chaque classe en utilisant l'ensemble d'apprentissage. Nous avons amélioré FPMC en tenant compte de l'importance (cardinalité) de chaque classe. Nous l'avons appelée FPM Améliorée (FPMA).

Cependant, FPMA se comporte exactement comme FPM dans les zones de chevauchement des classes. Cela revient à l'utilisation d'une matrice de corrélation binaire, qui ne tient pas compte du nombre de points de la même classe dans un hypercube. De plus, la taille de cette matrice peut devenir très vite gigantesque si le nombre d'attributs et le nombre de barres deviennent importants. Pour pouvoir l'estimer correctement, il faut que la base d'apprentissage couvre correctement tout l'espace de représentation. Sinon un nombre important de points sera rejeté.

Nous avons alors proposé une amélioration de FPMA en remplaçons la matrice de corrélation binaire par un ensemble des facteurs de corrélation flous. Ces facteurs tiennent compte du nombre de points appartenant à la même classe dans chaque hypercube. De plus ces facteurs sont calculés uniquement dans les zones, ou hypercubes, contenant des points

d'apprentissage. Le taux d'erreur de classification de FPMA est comparé à celui de FPM, k plus proches voisins classique et floue, et la méthode des noyaux de Parzen en utilisant plusieurs ensembles de données académiques et réelles. Le taux d'erreur de classification de FPMA est le plus faible par rapport aux autres méthodes utilisées.

FPM ne peut pas détecter l'apparition de nouvelles classes ni de suivre l'évolution d'une classe vers une autre. En effet les nouveaux points, qui portent l'information sur l'apparition de nouvelles classes ou leur évolution, sont rejetés par FPM. Egalement, FPM ne peut pas être appliquée pour la classification des données non-stationnaires. Dans le chapitre 3, nous allons étudier le problème de classification évolutive et dynamique ainsi que les performances de FPM de ce cas. Ensuite, la solution proposée dans [SAY02a, SAY02c] pour rendre FPM capable de détecter l'apparition de nouvelles classes ou de leurs évolutions est présentée et évaluée afin d'en déduire ses limites. Enfin, une solution basée sur FPM Améliorée (FPMA), développée dans le deuxième chapitre, a été proposée. Cette solution se résume en deux algorithmes. Premièrement, il s'agit de l'intégration de l'apprentissage incrémental à FPMA pour la rendre capable de réaliser le suivi de la déformation local pour le cas des classes non convexes. Deuxièmement, il s'agit d'une amélioration de cet algorithme pour la détection de l'apparition de nouvelles classes pour le cas des classes non convexes. Les performances des algorithmes proposés sont illustrées à travers plusieurs exemples académiques.

Chapitre 3

Classification adaptative et évolutive pour le diagnostic des systèmes dynamiques

3.1. Introduction

L'application de la Reconnaissance des formes (RdF) en diagnostic des systèmes dynamiques se heurte souvent au problème de l'insuffisance de la connaissance *a priori* concernant le fonctionnement de ces systèmes. Rappelons que, le diagnostic par RdF est réalisé en associant une nouvelle observation sur le fonctionnement du système, représentée par un point dans l'espace de représentation, à une de ces classes représentant un mode de fonctionnement. Dans une base de connaissance incomplète, tous les modes de fonctionnement ne sont pas représentés. En effet, pour des raisons de coût ou de sécurité, certains états défaillants ou dangereux ne peuvent pas être provoqués et seront donc absents de la base. Un processus de diagnostic par RdF doit donc être réalisé en utilisant une méthode de classification adaptative, c'est-à-dire qu'il doit être capable de détecter les états inconnus, représentés par l'apparition de nouvelles classes dans l'espace de représentation, et de les apprendre. De plus, un système dynamique peut évoluer entre plusieurs modes de fonctionnement. Cette évolution se traduit par un ensemble de points formant un pont entre deux classes. Ces points ne présentent pas une structure cohérente et stable pour contribuer à la formation d'une nouvelle classe. Par contre, il est intéressant de prédire le sens de cette évolution afin d'anticiper la dérive du système d'un mode de fonctionnement normal vers un mode anormal. Cette prédiction permet d'éviter les modes anormaux et leurs conséquences.

La base de connaissances ne peut pas contenir *a priori* toutes les évolutions possibles entre deux classes. L'extraction de l'information manquante sur l'évolution du système peut être réalisée par un diagnostic prédictif. Le diagnostic pour qu'il soit qualifié de prédictif nécessite une méthode de classification capable d'extraire l'information portée par chaque nouveau point et de l'inclure à la base de connaissance. Ainsi, l'information manquante est acquise grâce aux nouveaux points, ou données, reçus en ligne de façon séquentielle. Comme ces données ne sont pas connues *a priori* dans la base initiale. Cela nécessite un apprentissage continu conduisant à la mise à jour du modèle de classification chaque fois qu'une nouvelle donnée se présente. Cet apprentissage est réalisé soit hors-ligne soit en-ligne. L'apprentissage hors-ligne recalcule complètement le modèle de classification à chaque acquisition à partir de l'ensemble des données déjà connues auquel on ajoute la nouvelle donnée. Cette technique s'avère très gourmande en temps de calcul puisque l'ensemble d'apprentissage croit d'une façon permanente et ne tire aucun avantage des connaissances extraites au cours des apprentissages précédents. De plus, cette adaptation hors-ligne est peu efficace lorsque le système dérive lentement vers un mode inconnu. Enfin, la détection tardive peut être catastrophique si le nouveau mode de fonctionnement est dangereux. L'apprentissage en-ligne utilise les techniques d'apprentissage incrémental développées avec des règles de mise à jour récursives. Grâce à ces techniques, les informations portées par les nouveaux points sont

incorporées séquentiellement dans la mise à jour du modèle de classification sans réutiliser les anciens points.

Les classes basées sur des données stationnaires sont des classes statiques. Ce qui signifie que les caractéristiques du modèle de classification demeurent inchangées au cours du temps. Pour ce type de données, l'enrichissement de la base de connaissance se traduit par une déformation locale du contour ou bien de la forme des classes sans mettre en cause les informations acquises précédemment. Toutefois, comme l'environnement est en perpétuelle évolution, la majeure partie des données issues du monde réel est non-stationnaire. Ce type de données entraîne une variation au cours du temps des caractéristiques du modèle de classification. La classification dans ce cas est appelée classification dynamique. Cette dynamique ce traduit par le déplacement, l'élimination, la fusion ou la scission des classes. Parmi les applications réelles nécessitant une classification dynamique, on peut citer le diagnostic industriel (évolution des modes de fonctionnement), le diagnostic médical (expansion de cancer), la surveillance vidéo (mouvement des cibles), etc. [AMA06].

Dans la littérature, de nombreux auteurs se sont intéressés au problème du diagnostic adaptatif et prédictif et ont apportés une multitude de solutions pour le cas des données stationnaires. Par contre, peu de solutions concernent la classification dynamique. Dans le cas du diagnostic prédictif pour des données stationnaires, une première solution suppose la connaissance de la loi de densité de probabilité [GRE84, GAN87, FRE92, PEL93, ZIE95, OND06]. Elle utilise donc des méthodes statistiques paramétriques de détection d'évolution comme la méthode CUSUM (CUmulative SUM) ou le filtre de Kalman. Une seconde solution cherche soit à apprendre deux fonctions d'appartenance pour chaque couple de classes, ces fonctions épousent la forme de la trajectoire supposée au sein de ce couple [BOU96, COR02]. Soit cette solution réalise une interpolation linéaire entre les différents centres de gravités des différentes classes [OND06]. Le résultat de cette interpolation est une fonction mathématique polynomiale pour chaque dimension de l'espace de représentation. Cette fonction mathématique représente une trajectoire moyenne permettant d'indiquer la position d'un nouveau point par rapport aux différentes classes. Cette solution nécessite la connaissance *a priori* du chemin d'évolution entre chaque couple de classes, la détermination d'un grand nombre de paramètres pour construire la bonne forme de la trajectoire entre les classes et un temps de calcul relativement élevé. De plus, la fonction mathématique est une relation algébrique qui ne tient pas compte de la dynamique d'évolution des modes de fonctionnement.

Le diagnostic adaptatif peut être réalisé en utilisant : les critères de voisinage ou d'activation [BOU97], le filtre de Kalman [FRE92, OND06], ou la méthode potentielle [GRE84, PEL93]. L'utilisation des critères de voisinage et d'activation nécessite la détermination de plusieurs paramètres et donne des performances et une robustesse assez médiocre. L'utilisation du filtre de Kalman se base sur l'hypothèse de la connaissance de la loi de densité de probabilité. Cette hypothèse n'est pas toujours vérifiée. Enfin l'utilisation de la méthode potentielle, qui découpe l'espace de représentation en régions d'équi-appartenance aux classes, nécessite une connaissance complète de l'espace de décision et un temps de calcul exorbitant. Ces méthodes de classification sont développées pour la construction et l'adaptation locale des classes au cours du temps. Elles ne sont pas appropriées à la classification dynamique.

Dans [LUR03, AMA06, LEC06] deux algorithmes sont développés pour la classification des données non-stationnaires. Ces derniers sont des réseaux de neurones à architecture évolutive constitués de trois couches : d'entrée, cachée et de sortie. Grâce à la structure dynamique de ces réseaux, le nombre de neurones et de connexions peut varier. La création de classe se traduit par l'insertion de neurones dans la couche cachée et dans la couche de sortie : c'est la phase constructive de l'architecture. L'adaptation du modèle de classification entraîne

la mise à jour des poids des neurones cachés et de sortie. Au cours de la procédure de fusion des classes, les neurones correspondant aux classes fusionnées sont remplacés par un neurone unique dans la couche de sortie, la couche cachée est mise à jour par la même occasion. Pendant la scission de classe, l'architecture s'adapte inversement. L'élimination de classes parasites et obsolètes entraîne la suppression de neurones dans la couche cachée et dans celle de sortie, c'est la phase d'élimination ou d'élagage de l'architecture. Ces deux algorithmes, sont des réseaux à prototypes qui utilisent des prototypes comme des unités regroupant les données de caractéristiques typiques. Chaque classe complexe est modélisée par la combinaison de plusieurs prototypes hyper-elliptiques dont chacun s'adapte localement à la distribution de données : approche multi-prototype. Le premier algorithme est appelé l'AUDyC (AUto-adaptive and Dynamical Clustering). Cet algorithme est basé sur une technique de modélisation inspirée du modèle de mélange gaussien. Chaque classe est décrite par un ensemble de sous-classes gaussiennes. Le deuxième algorithme appelé SAKM (Self-Adaptive Kernel Machine) modélise les classes dynamiques en utilisant des fonctions d'apprentissage à noyau. Ces deux algorithmes de classification dynamique sont critiquables pour leur nombre important de paramètres de réglage. L'algorithme AUDyC utilise 9 paramètres et la détermination de certains est critique comme la taille de la fenêtre de définition des prototypes et leur nombre ou encore le choix de la matrice de covariance initiale. Les autres paramètres sont fixés par expérience ou choisis en fonction du bruit dans les données et des objectifs du problème à traiter. Les paramètres du SAKM sont également nombreux (7 paramètres au total). Le choix du paramètre du noyau gaussien peut être délicat lorsque la distribution des données n'est pas connue. Le nombre de paramètres à déterminer pour SAKM est moins élevé que celui de AUDyC mais SAKM ne comporte pas de mécanisme de scission.

Dans notre cas, on ne dispose d'aucune connaissance sur la loi de densité de probabilité. De plus, on cherche à réaliser la prédiction et l'adaptation en ligne, sans connaissance *a priori* de la forme des classes, pour des systèmes disposant de bases de données incomplètes. Dans [SAY02a] une solution, basée sur Fuzzy Pattern Matching (FPM), capable de détecter l'apparition de nouvelles classes, de forme convexe, et de suivre leurs évolutions est proposée. Cependant, cette solution souffre des inconvénients de FPM abordés dans le deuxième chapitre. Dans ce chapitre nous allons développer FPM Améliorée (FPMA), proposée dans le deuxième chapitre, pour qu'elle soit apte à réaliser le diagnostic adaptatif pour des systèmes dynamiques quelque soit la forme des classes, convexes ou non. Ce développement est basé sur l'intégration de deux algorithmes à la méthode FPMA. Le premier algorithme concerne l'intégration d'une approche incrémentale pour la mise à jour des densités de possibilité après la classification de chaque nouveau point. Le deuxième algorithme, quant à lui, est un mécanisme permettant la détection en ligne de l'apparition de nouvelles classes en analysant la similarité entre les points rejetés.

Ce chapitre est structuré comme suit. Premièrement, le problème de la classification adaptative et prédictive, dans le cas des données stationnaires est abordé en évaluant les performances de FPM afin d'en déduire ses limites. Cela est réalisé à travers des exemples concernant le cas d'évolution entre deux classes, le cas d'apparition de nouvelles classes et le cas de modification locale, grossissement, d'une classe. Deuxièmement, la solution développée en [SAY02a], pour la classification adaptative et prédictive basée sur FPM, est présentée. Nous allons démontrer, à travers des exemples que cette solution n'est applicable que dans le cas des classes de forme convexe. De plus, qu'elle est inopérante pour la classification des données non-stationnaires décrivant des situations de rotation, de déplacement, de fusion et de scission des classes.

Pour remédier aux limites de la solution existante pour le cas stationnaire, une amélioration de FPMA, développée dans le chapitre 2, en lui intégrant l'apprentissage

incrémental est proposée. Suivi d'une proposition d'un algorithme basé sur FPMA pour la détection de l'apparition de nouvelles classes pour le cas des classes de forme convexe ou non. Enfin, nous allons illustrer les performances de notre développement en utilisant plusieurs exemples académiques tout en montrant ses limites. Ces dernières induiront les perspectives de cette thèse.

3.2. Limites de FPM en classification adaptative et évolutive

3.2.1. Cas des données stationnaires

Les données stationnaires sont des données qui conservent leurs caractéristiques statistiques au cours du temps. La base de données d'un système complexe est souvent incomplète, tous les modes de fonctionnement et notamment les modes défaillants ou dangereux ne peuvent pas être provoqués. De plus les données disponibles sur un mode de fonctionnement sont souvent insuffisantes pour estimer correctement le modèle de classification. Des modifications locales au niveau des contours des classes de l'ensemble d'apprentissage ou encore au niveau de leurs frontières de décision, sont dues à l'intégration de nouvelles données assignées dans ces classes. Ces modifications ne sont pas causées par la non-stationnarité des données mais parce que la base ne contient pas suffisamment de données pour tracer correctement la structure de ces classes en termes de géométrie, de contour, ou de frontière de décision.

De plus, l'incomplétude de la base de données se traduit par l'absence d'information sur les trajectoires d'évolution entre les différentes classes. En effet, il est impossible d'inclure *a priori* les données de tous les chemins d'évolutions, entre toutes les classes, dans l'ensemble d'apprentissage. Ces évolutions sont dues à la nature dynamique du système qui entraîne un changement de son mode de fonctionnement au cours du temps.

3.2.1.1. Evolution entre deux classes

Un système évolue d'un état normal vers un autre anormal soit d'une façon brutale soit d'une façon successive. Une évolution est brutale si le point apparaissant à l'instant $t+1$ n'est pas affecté à la même classe que le point apparaissant à l'instant t. Dans ce cas, la prédiction de l'évolution est impossible. L'évolution successive se caractérise par l'apparition de plusieurs points intermédiaires entre un point affecté à la classe de départ et un point affecté à la classe d'arrivée. Ces points intermédiaires sont généralement rejetés et ne représentent pas une structure stable et cohérente pour former une nouvelle classe. Dans ce cas d'évolution successive, il est préférable de prédire l'évolution plutôt que d'attendre d'arriver à un mode anormal. Cette prédiction va permettre de prendre d'une façon immédiate les démarches correctives et d'éviter ainsi les conséquences du dysfonctionnement.

Une évolution désigne un état de transition entre deux modes de fonctionnement. Prévoir la classe, ou le mode de fonctionnement, d'arrivée revient à effectuer un pronostic, ou un diagnostic prédictif [PEL93]. La simple connaissance de l'appartenance à une classe ne suffit pas pour la prise en compte du phénomène d'évolution d'un mode de fonctionnement vers un autre. Il faut pouvoir quantifier la proximité ou l'éloignement d'une observation à une classe. Cette quantification traduit la représentativité de l'observation vis a vis des classes.

L'évolution s'appuie sur une suite chronologique d'observations décrivant le comportement d'un système sur un intervalle de temps. Cette évolution est caractérisée par un chemin entre les deux états comme le montre l'exemple de la figure 3.1. Ces points ont été générés aléatoirement selon une loi normale avec une moyenne qui change successivement à

partir de la moyenne de la première classe vers celle de la deuxième classe. Les classes C_a et C_b suivent respectivement les lois normales $\mathcal{N}_a(\underline{m}_a = (3\,,\,3)^T, \Sigma_a = (0{,}5\,,\,0{,}5)^T)$, $\mathcal{N}_b(\underline{m}_b = (15\,,\,15)^T, \Sigma_b = (0{,}5\,,\,0{,}5)^T)$. Le nombre de points nécessaire pour prédire l'évolution du système dépend de son type : lent ou rapide.

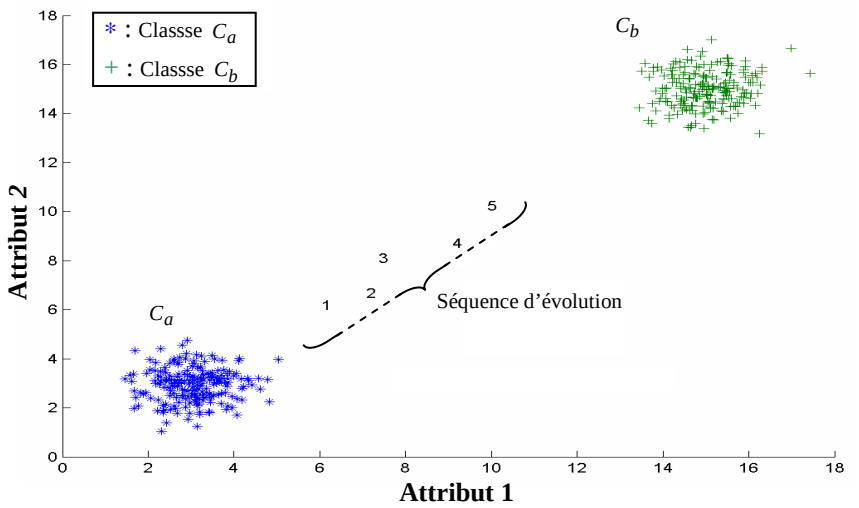


Figure 3.1 Evolution de la classe C_a vers la classe C_b. Le numéro du point indique son ordre d'arrivée

FPM classifie un nouveau point dans une des classes connues ou le rejette. La figure 3.2 présente les courbes de niveaux d'appartenance pour l'exemple de la figure 3.1. FPM divise donc l'espace de représentation en deux zones, la première est celle où les points seront affectés à la classe C_a ou à la classe C_b et la zone 2 est celle où les points seront rejetés.

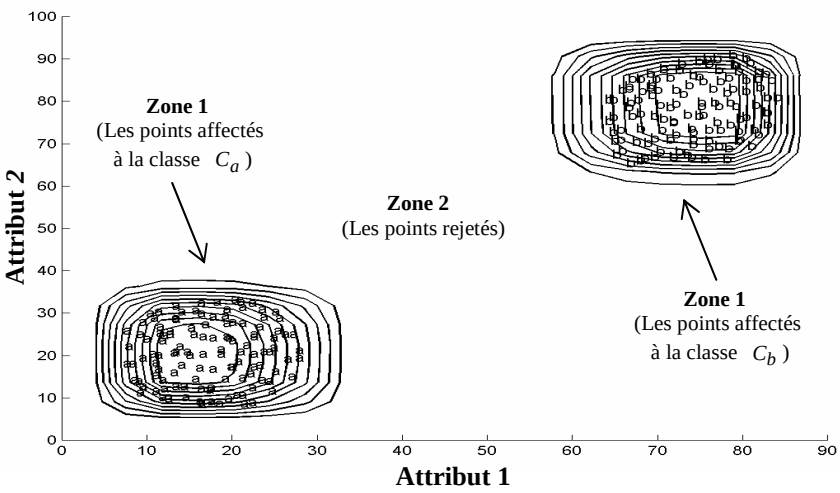


Figure 3.2 Division par FPM de l'espace de représentation en deux zones.

La zone contenant le chemin d'évolution s'appelle la zone d'évolution. Puisque, lorsque l'on utilise FPM, celle-ci correspond à la zone 2 (des points rejetés), FPM ne fournit aucune

information sur la représentativité des points vis à vis des classes. FPM n'est donc pas capable de prédire l'évolution entre les classes [SAY02a].

3.2.1.2. Apparition de nouvelles classes

Un problème de diagnostic par RdF s'accompagne de l'incomplétude de la base de connaissance où tous les modes de fonctionnement ne sont pas représentés. Un système de diagnostic adaptatif doit être capable de détecter et d'inclure à sa base de connaissances les états inconnus. Le rejet confère au système de décision son caractère adaptatif. En effet, détecter un état inconnu nécessite un nombre minimum de points rejetés et géographiquement proches les uns des autres. Une nouvelle classe ne peut donc apparaître que dans la zone d'évolution. FPM ne donne pas d'information sur la situation géographique d'un point dans cette zone. La figure 3.3 montre un exemple d'apparition d'une nouvelle classe représentée par le rejet en appartenance d'un ensemble de 10 points qui suivent une loi normale $\mathcal{N}(\underline{m} = (12\ ,\ 7)^T, \Sigma = (0{,}5\ ,\ 0{,}5)^T)$.

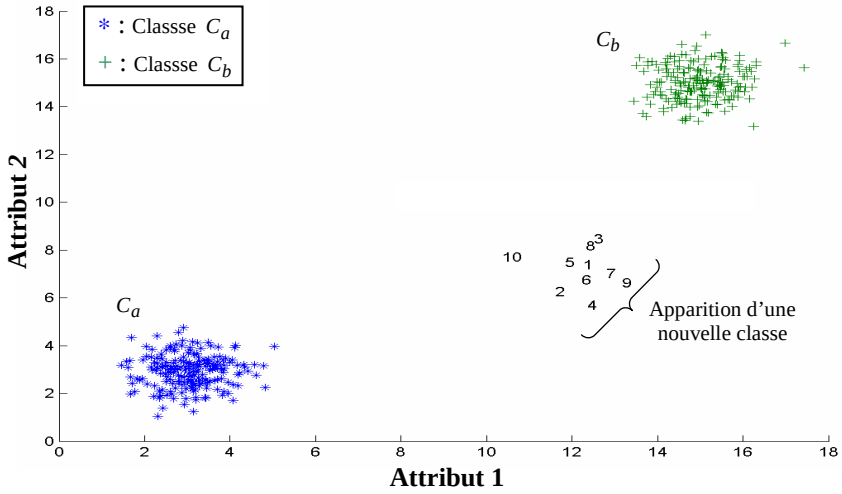


Figure 3.3 Apparition d'une nouvelle classe représentée par des points rejetés en appartenance.

Cependant il faut distinguer l'apparition d'une nouvelle classe représentative de l'apparition d'une classe parasite, créée à cause du bruit et des perturbations extérieures. Dans la littérature, la détection de ce type de classe est basée sur un critère de cardinalité de points ayant un niveau acceptable de similarité entre eux.

Pour inclure un état inconnu dans la base de connaissance, FPM doit arrêter la phase de classification et effectuer un apprentissage hors-ligne [SAY02a]. Le système de diagnostic doit alors utiliser une méthode de coalescence floue pour introduire la nouvelle classe dans l'ensemble des classes. Ensuite les fonctions d'appartenance de FPM sont reconstruites ou bien mises à jour pendant la phase d'apprentissage. Cela est fait en tenant compte de l'intégration de cette nouvelle classe. La figure 3.4 montre une adaptation hors-ligne d'un système de diagnostic.

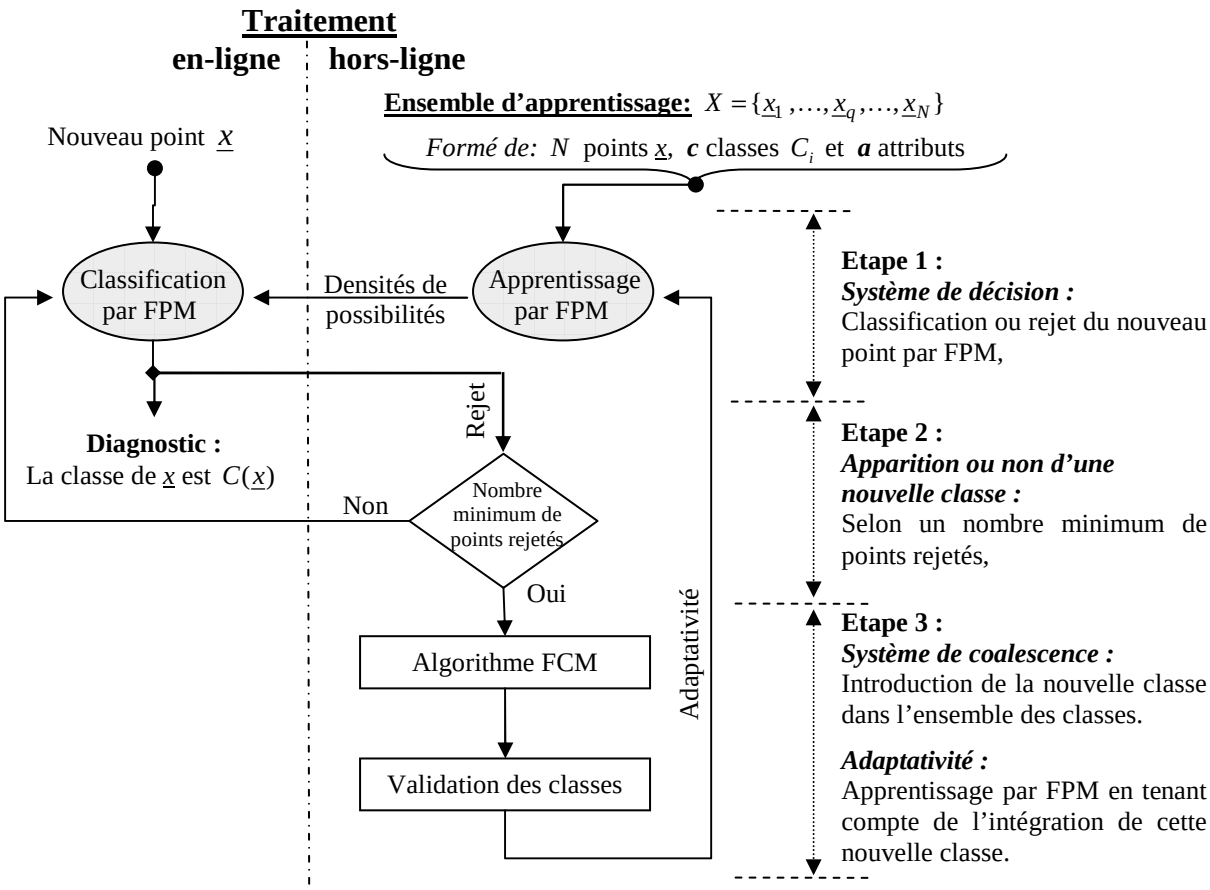


Figure 3.4 Adaptativité du système de diagnostic hors-ligne .

Cette procédure lourde empêche l'utilisation du diagnostic adaptatif, basé sur FPM, en temps réel [SAY02a]. De plus, le traitement hors-ligne des points rejetés est critique lorsque le nouveau mode de fonctionnement est dangereux et doit être détecté aussi tôt que possible.

3.2.1.3. Modification locale d'une classe

Quand les points de l'ensemble d'apprentissage ne sont pas suffisants pour correctement tracer le contour de la classe, ou de bien estimer ses paramètres comme la moyenne et la variance pour une classe gaussienne, la mise à jour du module de classification après l'affectation de chaque nouveau point conduira à une modification locale de la classe comme le montre la figure 3.5. Dans cette figure, la classe suit une loi normale $\mathcal{N}(\underline{m} = (7,15)^T, \Sigma = (1,1)^T)$. L'ensemble d'apprentissage de départ contient 20 points, les étoiles dans la figure 3.5. L'estimation des paramètres de cette classe ainsi que son contour ne sont pas corrects. La classification de 180 nouveaux points, les signes " + " dans la figure 3.5, est réalisée par l'algorithme proposé dans [SAY02c]. Ce dernier, permet à FPM de réaliser un apprentissage incrémental, grâce à la mise à jour des densités de possibilité. Cela, conduit au changement : grossissement, du contour de cette classe permettant ainsi d'estimer correctement ses paramètres. Le contour de la classe est représenté, dans la figure 3.5, par la courbe de niveaux d'appartenance ayant la valeur 0,1.

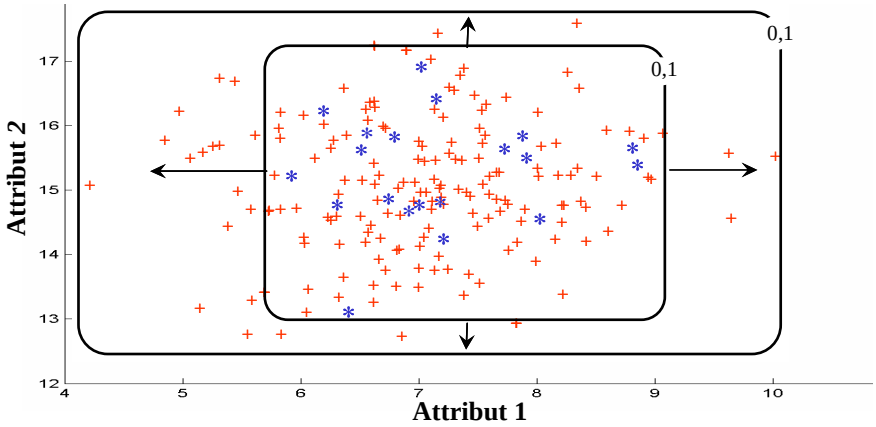


*Figure 3.5 Modification locale, grossissement, de la classe due à la classification par FPM de nouveaux points. " * " indique les points initiaux de l'ensemble d'apprentissage et " + " les nouveaux points affectés à la classe. " → " indique le sens de grossissement de la classe.*

L'algorithme proposé dans [SAY02c] basé sur FPM n'est applicable que quand les classes ont des formes convexes ; par contre dans le cas contraire, comme le montre la figure 3.6, cet algorithme n'est pas opérant. On propose d'utiliser FPMA avec apprentissage flou développée dans le chapitre 2 puisqu'elle tient compte de la forme non-convexe des classes, cf. figure 3.19.

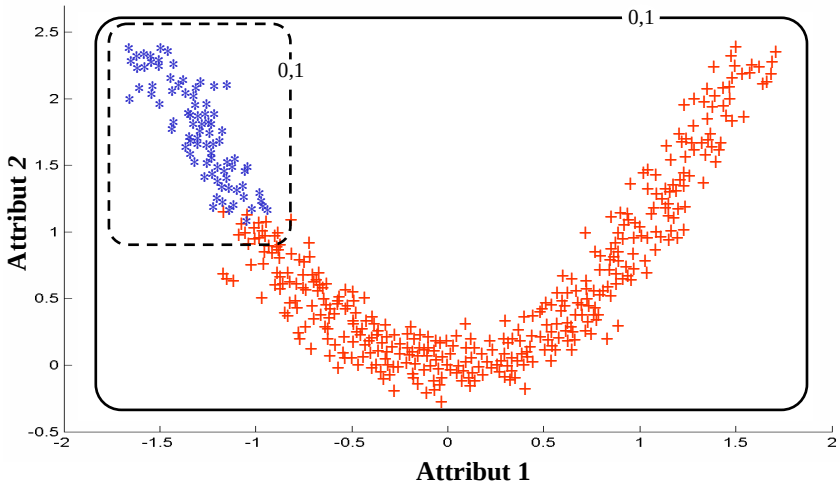


*Figure 3.6 Modification locale du contour de la classe due à la classification par FPM de nouveaux points. " * " indique les points initiaux de l'ensemble d'apprentissage et " + " les nouveaux points affectés à la classe.*

3.2.2. Cas des données non-stationnaires

Les données non-stationnaires sont des données pour lesquelles au moins une caractéristique statistique varie au cours du temps. Les données issues du monde réel sont non-stationnaires parce que le monde est en perpétuelle évolution. Cette évolution se traduit

par une rotation, un déplacement, une fusion ou par une scission des classes, comme le montrent respectivement les figures 3.7, 3.8, 3.9 et 3.10. Dans ces figures les courbes de niveau d'appartenance sont représentées pour le niveau 0,1. Ces courbes sont obtenues par FPM, en supposant que ces données sont stationnaires. Par exemple pour la rotation de la classe, les courbes sont tracées en considérant toutes les données avant la rotation et ensuite toutes les données après la rotation. Cela, afin de représenter les courbes d'appartenances telles qu'elles devraient être obtenues par une méthode de classification adaptée pour le cas des données non-stationnaires.

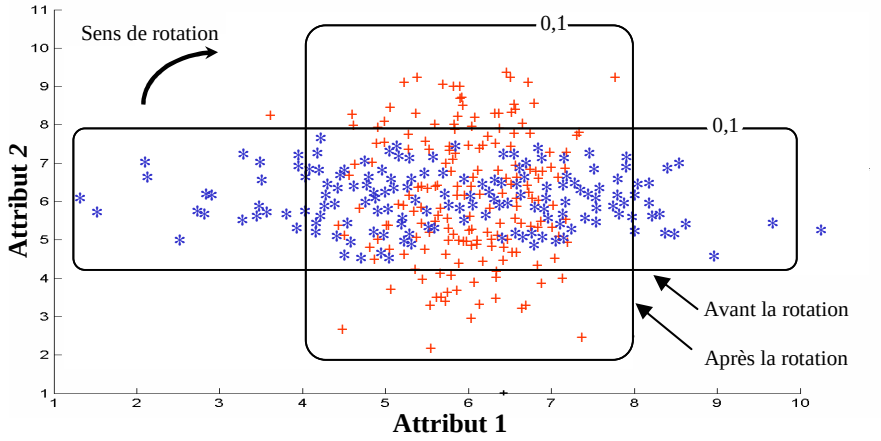


*Figure 3.7 Rotation de la classe due à la non stationnarité des données. " * " les points de départ de la classe et " + " les points après la rotation. Cette rotation est réalisée par la rotation des valeurs de la variance par rapport aux axes de l'espace de représentation.*

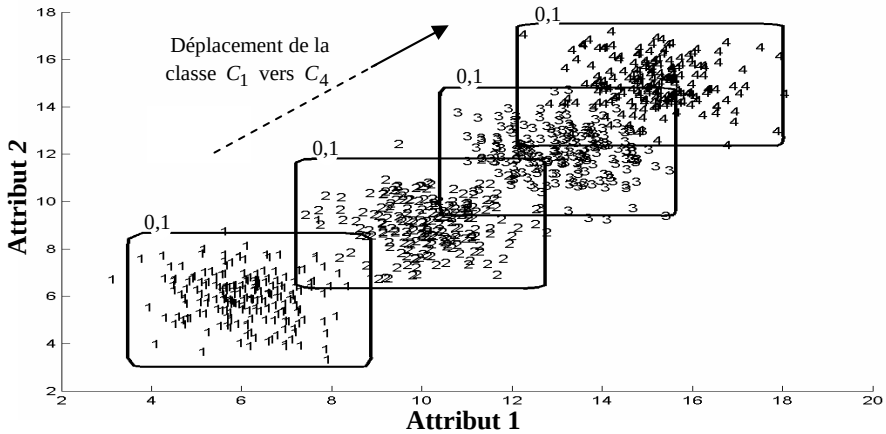


Figure 3.8 Déplacement de classe dû à la non-stationnarité des données. Ce déplacement est réalisé par le changement successif de la moyenne de la classe.

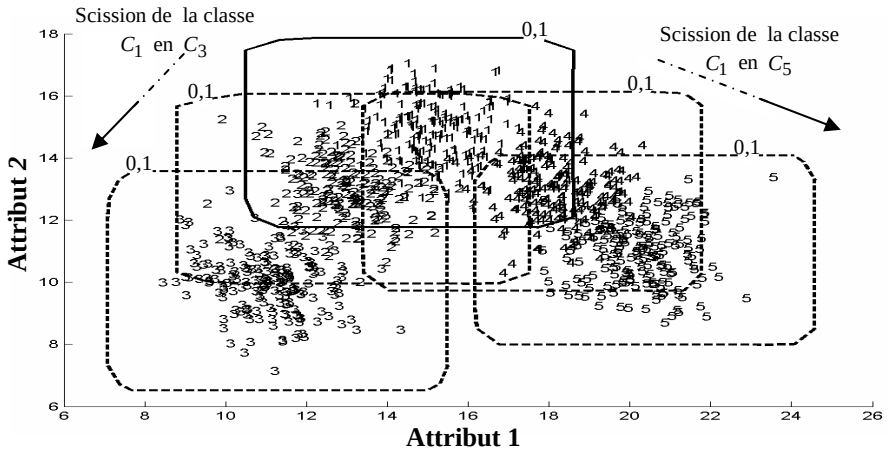


Figure 3.9 Scission de la classe C_1 en deux classes C_3 et C_5, due à la non-stationnarité des données.

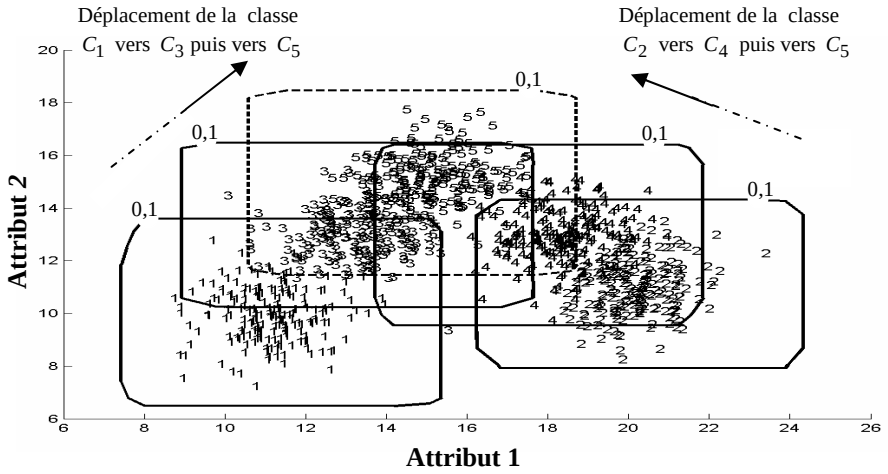


Figure 3.10 Fusion des classes C_1 et C_2 en une seule classe C_5 due à la non-stationnarité des données.

La classification dynamique de ce type de données nécessite la prise en compte des informations récentes utiles apportées par les nouveaux points affectés aux classes et l'élimination des informations qui ne sont plus représentatives à l'instant courant et apportées par les plus anciens points de ces classes. Ces informations, qui ne sont plus valides à l'instant courant, nécessitent un critère d'oubli qui est très subjectif et très dépendant du contexte de l'application. La figure 3.11 montre la conséquence de la non prise en compte d'un facteur d'oubli dans la méthode de classification pour le cas de déplacement de la classe de la figure 3.8. Nous constatons simplement une modification locale du contour, déformation, empêchant tout suivi du déplacement de la classe. La même remarque peut être constatée pour les cas de fusion ou de scission des classes.

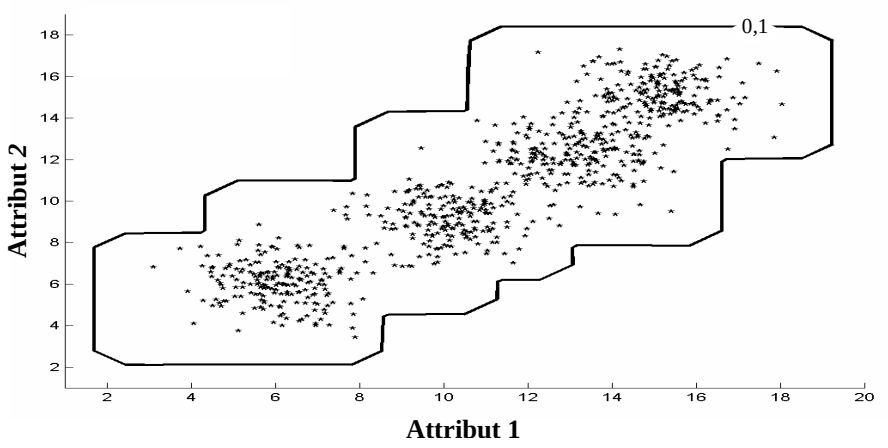


Figure 3.11 La courbe 0,1 de niveau d'appartenance obtenue par FPMA pour le cas de déplacement des classes sans tenir compte d'un facteur d'oubli.

L'intégration d'un critère d'oubli dans une méthode de classification nécessite en général des informations supplémentaires permettant de déterminer l'espace de temps à partir duquel une information n'est plus valide. Par exemple un expert du système, peut donner un indicateur sur la similarité nécessaire pour un point, affecté précédemment à une classe, pour qu'il soit toujours accepté au sein de cette classe.

3.2.2.1. Détection de la rotation des classes

FPM ne détecte pas la rotation de la classe contenant des données non-stationnaires comme le montre la figure 3.12.

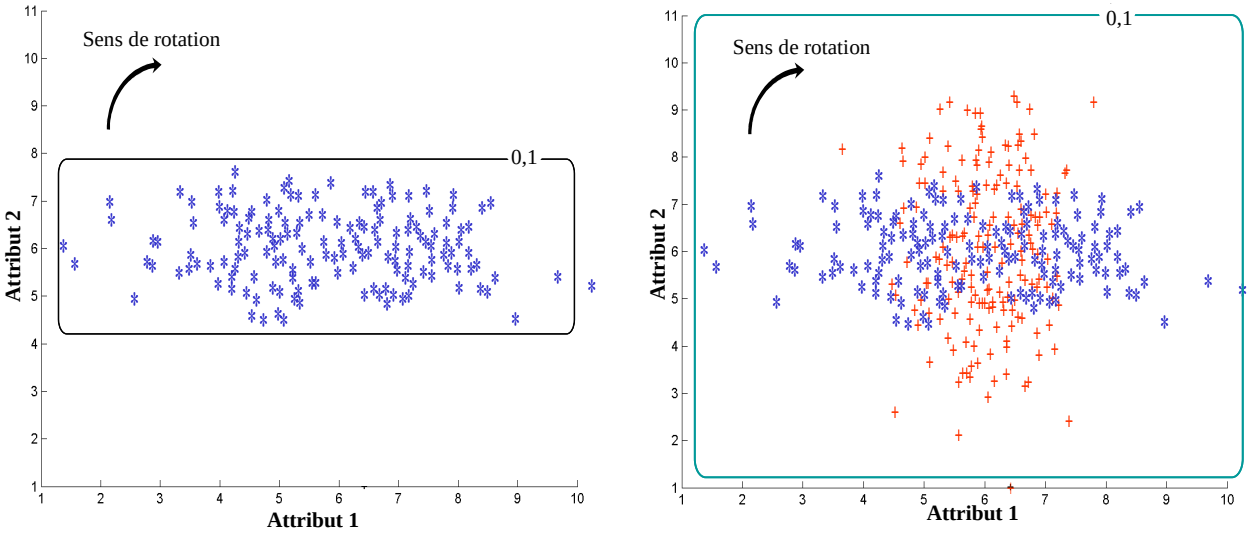


*Figure 3.12 La non détection de la rotation d'une classe, contenants des données non stationnaire, par FPM. Les points de la classe avant la rotation sont représentés par " * ", à gauche. Après la rotation ces points sont représentés par " + ", à droite.*

La flèche, dans cette figure, indique le sens de rotation. Les points de la classe avant la rotation et les points de cette même classe après la rotation sont représentés respectivement par " * " et " + ". Cette rotation est réalisée par le changement des valeurs de la variance par rapport aux axes de l'espace de représentation. On peut constater, d'après la courbe de niveau 0,1 d'appartenance obtenue par FPM avant et après cette rotation, que FPM considère la rotation comme un grossissement de la classe, cf. figure 3.5.

3.2.2.2. Détection de déplacement des classes

Le déplacement des classes, de la figure 3.13, est dû à la non-stationnarité des données. Il est réalisé par le changement successif de la moyenne de la classe C_1 ainsi que de sa variance. FPM ne détecte pas ce déplacement d'après la courbe d'appartenance de niveau 0,1. Cette dernière considère plutôt le déplacement comme un grossissement de la classe C_1.

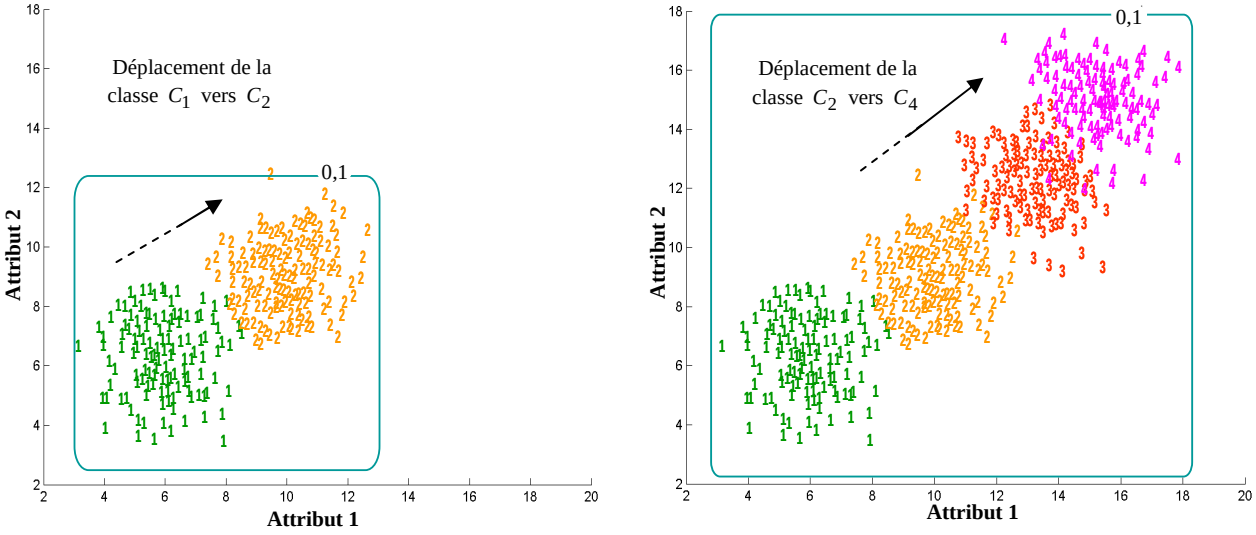


Figure 3.13 La non détection de déplacement d'une classe, contenants des données non stationnaire, par FPM. Le déplacement de classe C_1 vers la classe C_2 est représenté à gauche et le déplacement de la classe C_2 vers la classe C_3 puis vers C_4 est représenté à droite.

3.2.2.3. Détection de la fusion des classes

La fusion des classes C_1 et C_2, de la figure 3.14, en une seule classe C_5 ne peut pas être détectée par FPM comme le montre les courbes d'appartenance de niveau 0,1. Les flèches, dans cette figure, indiquent le sens de déplacement des classes. FPM détecte plutôt deux classes séparables avant la fusion, cf. figure 3.14 à gauche, et deux classes chevauchantes après la fusion, cf. figure 3.14 à droite.

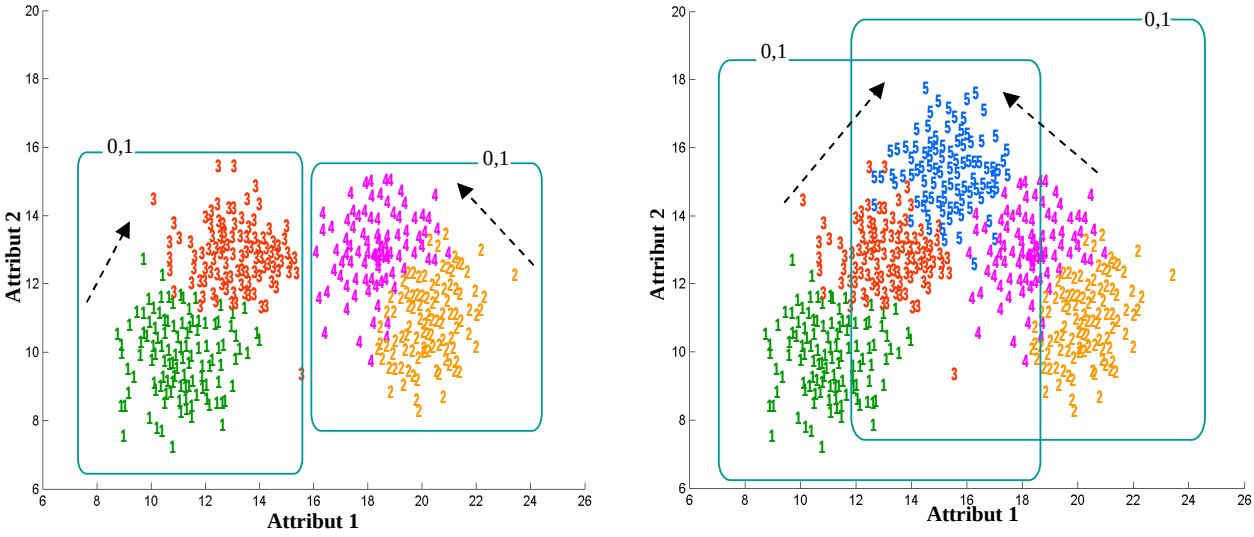


Figure 3.14 Déplacement de la classes C_1 vers la classe C_3 et la classe C_2 vers la classe C_4 due à la non-stationnarité des données, représentée à gauche. Suivie d'une fusion en une seule classe C_5, représentée à droite. La flèche indique le sens de déplacement.

3.2.2.4. Détection de la Scission des classes

La Scission de la classe C_1 de la figure 3.15, en deux classes C_2 et C_4, ne peut pas être détectée par FPM comme le montre les courbes d'appartenance de niveau 0,1. Les flèches, dans cette figure, indiquent le sens de déplacement des classes pendant et après la scission. FPM détecte plutôt une classe avant la scission, la classe C_1, et considère la scission de cette classe comme un grossissement, cf. figure 3.15 à gauche. FPM considère aussi le déplacement après la scission comme un grossissement de la classe C_1, cf. figure 3.15 à droite.

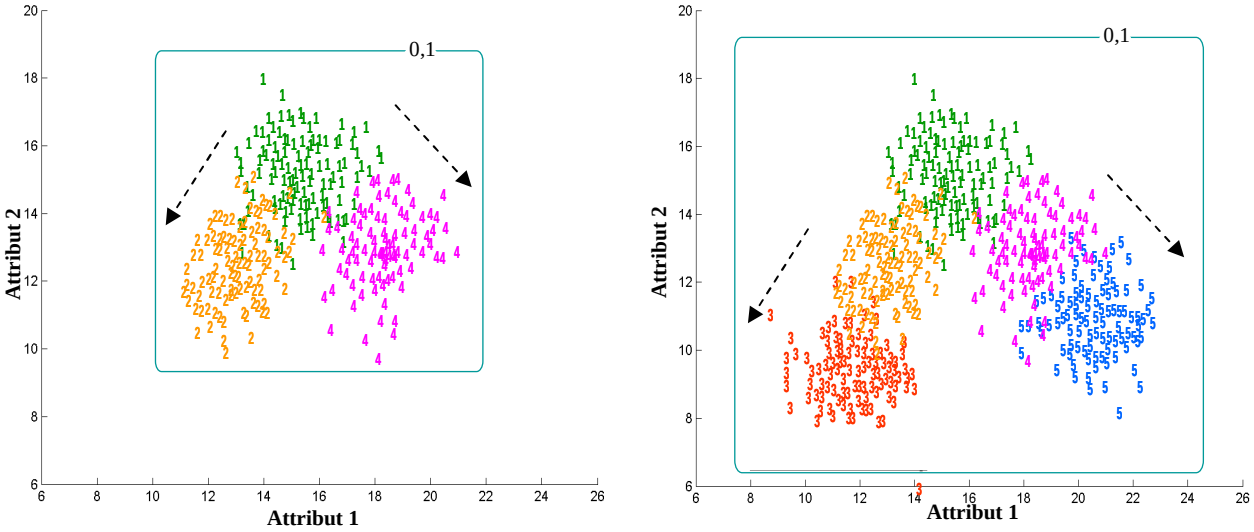


Figure 3.15 Scission de la classe C_1 en deux classes C_2 et C_4 due à la non-stationnarité des données, représentée à gauche. Suivie d'un déplacement de la classes C_2 vers la classe C_3 et de la classe C_4 vers la classe C_5, représentée à droite. La flèche indique le sens de déplacement.

3.3. Solution existante pour le cas stationnaire basée sur FPM

Soit $Z^a \subset IR^a$ un sous-espace de l'espace de représentation définissant la zone d'évolution et dans lequel FPM rejette tout nouveau point. Ce sous-espace contient les informations requises sur le comportement du système en termes d'évolution (éloignement ou rapprochement) entre les classes ou de création de nouvelles classes. Soit $XR = (\underline{x}_1, \ldots, \underline{x}_r, \ldots, \underline{x}_{n_e})$ une série chronologique composée de n_e points ou observations rejetées, par FPM. Cette dernière n'assigne aucune valeur d'appartenance aux points qui sont situés loin ou entre les classes connues *a priori*. Ces points rejetés peuvent apporter des informations sur l'apparition d'une nouvelle classe ou bien l'évolution d'une classe vers une autre. Quand un système évolue progressivement d'une classe C_1 vers une autre classe C_2, la possibilité d'appartenance, de la suite d'observations, à la classe C_1 devrait diminuer au fur et à mesure du rapprochement de C_2. Par contre, la possibilité d'appartenance à la classe C_2 devrait augmenter.

Dans [SAY02a], une approche capable de quantifier l'éloignement ou le rapprochement de chacun des points rejetés par rapport à chacune des classes connues de l'espace de représentation est proposée. Cette approche considère chaque attribut, comme une source d'information. De plus, la représentativité d'une observation rejetée dans la zone d'évolution est quantifiée en utilisant deux informations : la distance entre cette observation et son plus proche voisin ainsi que la valeur d'appartenance du plus proche voisin de cette observation. En utilisant ces deux informations, une fonction d'évolution, quantifiant la représentativité de l'observation rejetée, pour chaque attribut est donc construite [BOU07e]. La construction de cette fonction est détaillée dans ce qui suit

3.3.1. Construction de la fonction d'évolution

La fonction d'évolution $v_i^j(\underline{x}_r) : Z \to [0,1]$, est définie pour un point $\underline{x}_r \in XR$, par rapport à chacune des classes connues C_i et selon chaque attribut j. Elle quantifie non plus la valeur d'appartenance d'un point à une classe puisque ce point a été rejeté, mais la représentativité du point rejeté $\underline{x}_r$ ou bien sa situation géographique par rapport à une classe C_i ou encore l'influence de C_i sur ce point. Cette fonction a la forme suivante :

$$v_i^j(\underline{x}_r) = (1 - d_n^j(\underline{x}_r, \underline{x}_{ppv}))^{\frac{c}{\pi_i^j(x_{ppv})}} \tag{3.1}$$

où $d_n^j(\underline{x}_r, \underline{x}_{ppv}) \in [0,1]$ est la distance normalisée selon l'attribut j entre la point rejeté $\underline{x}_r$ et son plus proche voisin $\underline{x}_{ppv}$ appartenant à C_i. Cette distance est calculée par :

$$d_n^j(\underline{x}_r, \underline{x}_{ppv}) = \frac{d^j(x_r, \underline{x}_{ppv})}{x_{\max}^j - x_{\min}^j} \tag{3.2}$$

où $x_{\max}^j$ et $x_{\min}^j$ sont respectivement les bornes supérieurs et inférieurs de l'histogramme correspondant selon l'attribut j. Ces bornes sont considérées comme étant les coordonnées

maximales et minimales de l'ensemble d'apprentissage selon l'attribut j comme le montre la figure 3.16. Cette figure illustre aussi le calcul de $d^j(\underline{x}_r, \underline{x}_{ppv})$: la distance entre $\underline{x}_r$ et $\underline{x}_{ppv}$ selon l'attribut j, ainsi que la détermination de $\underline{x}_{ppv}$ et sa possibilité $\pi_i^j(\underline{x}_{ppv})$. Puisque le point $\underline{x}_r$ est rejeté de la classe C_i selon l'attribut j, son plus proche voisin $\underline{x}_{ppv}$ devrait alors se trouver soit dans la première barre soit dans la dernière barre de l'histogramme correspondant. Afin de pouvoir réaliser la prédiction ou l'adaptation en temps réel, la distance doit être calculée soit entre $\underline{x}_r$ et la borne supérieure, dans le cas où $\underline{x}_{ppv}$ se trouve en dernière barre, soit entre $\underline{x}_r$ et la borne inférieure, dans le cas où $\underline{x}_{ppv}$ se trouve en première barre, de l'histogramme de la classe C_i par rapport à l'attribut j. Cela permet d'éviter le calcul de toutes les distances entre $\underline{x}_r$ et tous les points de l'ensemble d'apprentissage pour trouver $\underline{x}_{ppv}$. La distance $d^j(\underline{x}_r, \underline{x}_{ppv})$ est normalisée par rapport à $x_{\max}^j$ et $x_{\min}^j$ de l'ensemble d'apprentissage selon chaque attribut j, pour qu'elle prenne des valeurs comprises dans [0, 1] quelque soit le chemin d'évolution. La décroissance de la fonction d'évolution est proportionnelle à l'éloignement du point rejeté par rapport à la classe C_i. Cette fonction doit donc diminuer quand la distance normalisée augmente.

$\pi_i^j(\underline{x}_{ppv})$ est la possibilité d'appartenance du plus proche voisin $\underline{x}_{ppv}$ de $\underline{x}_r$ à la classe C_i par rapport à l'attribut j. Si $\underline{x}_{ppv}$ se trouve en dernière barre de l'histogramme de possibilité correspondant à la classe C_i et l'attribut j, sa possibilité sera égale à celle de cette barre. Par contre si $\underline{x}_{ppv}$ se trouve en première barre, $\pi_i^j(\underline{x}_{ppv})$ prendra la valeur de cette barre.

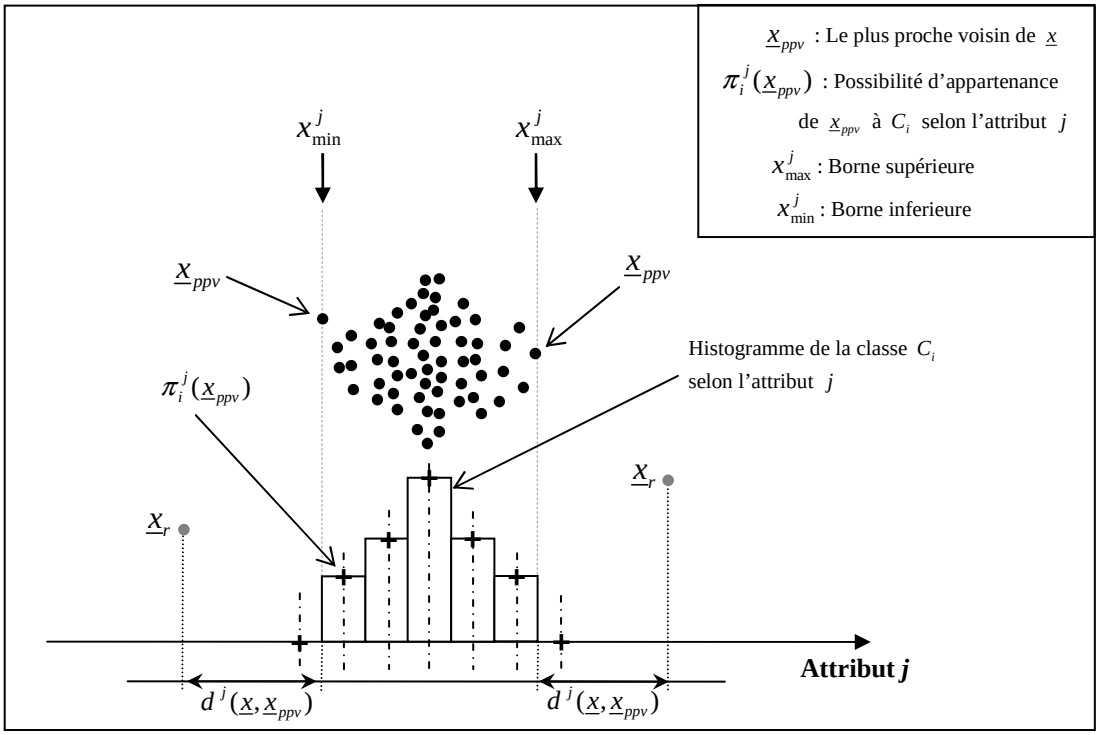


Figure 3.16 Calcul de distance entre le point rejeté $\underline{x}_r$ et son plus proche voisin $\underline{x}_{ppv}$ ainsi que la possibilité $\pi_i^j(\underline{x}_{ppv})$.

3.3.2. Performances de la solution existante

Un point rejeté peut correspondre à un des trois cas suivants :

- c'est un point de bruit,
- c'est un point caractérisant un état d'évolution,
- c'est un point appartenant à une nouvelle classe.

Les performances de l'algorithme proposé dans [SAY02a] ont été testées sur plusieurs exemples simulés et réels. Il donne de bons résultats tant que les classes ont des formes simples. Dans le cas contraire, nous allons démontrer que cet algorithme ne peut pas être appliqué.

3.3.2.1. Détection de nouvelle classe dans le cas des formes convexes

Dans une série chronologique d'évolution représentant l'apparition d'une nouvelle classe, les observations consécutives doivent traduire des comportements voisins. Autrement dit, plus ces observations sont proches les unes des autres dans la série chronologique, plus elles sont susceptibles de traduire le même comportement. Ces observations doivent donc se localiser dans une même zone restreinte de l'espace de représentation formant ainsi un même état. Cela peut être traduit par la notion de stabilité. Un état stable peut être vu comme un ou plusieurs paliers dans les signaux chronologiques VC_i avec des oscillations autour de la valeur moyenne de ce palier. Ces oscillations traduisent la dynamique du système au sein du nouvel état. Reprenons l'exemple de la section 3.2.1.2. qui montre le cas d'apparition d'une nouvelle classe. La série chronologique d'évolution XR contient 10 points qui sont rejetés par rapport aux deux classes. Les vecteurs chronologiques qui guident la détection de cette nouvelle classe son représentés pour les classes C_a et C_b dans la figure 3.17. Comme on peut le constater, ces signaux oscillent autour de la valeur moyenne des deux paliers 1 et 2. De plus, les vitesses moyennes d'évolution calculées sont $VE_a = 0,008$ et $VE_b = -0,002$. Ces valeurs sont très faibles. Cela signifie que les points sont très proches les uns des autres par rapport aux classes connues et qu'il n'y a pas d'évolution [SAY02a]. Enfin, les seuils d'évolutions $npsp_a = 0,55$, $npsn_b = 0,55$ confirment la présence d'un état de stabilité et donc la détection d'une nouvelle classe.

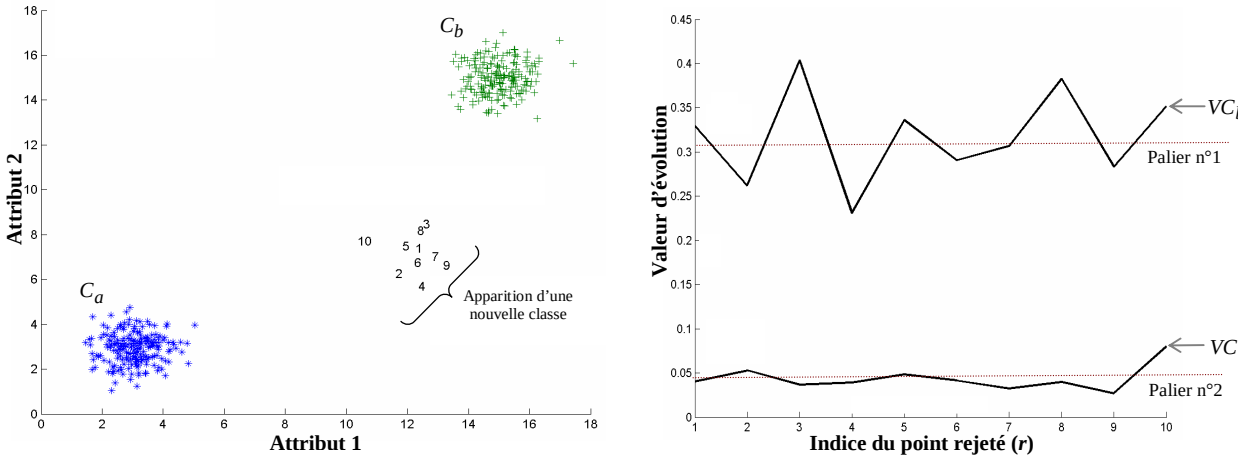


Figure 3.17 Apparition d'une nouvelle classe, à gauche, et les signaux Vecteurs Chronologiques VC_a et VC_b correspondants, à droite. Ces signaux oscillent autour de la valeur moyenne des deux paliers 1 et 2.

3.3.2.2. Limites de la solution existante dans le cas des classes non convexes

Puisque cet algorithme est basé sur la méthode FPM, il n'est pas donc applicable pour détecter et prédire l'évolution entre des classes de forme non convexe. En fait l'algorithme proposé ci-dessus quantifie la représentativité des points rejetés par FPM. Si des nouvelles classes apparaissent entre les classes connues de forme non convexe, leurs points seront classifiés dans ces classes connues et aucun point n'est rejeté. Cela est dû au non respect de la forme des classes connues. La figure 3.18 illustre ce cas pour une classe connue de forme non-convexe. L'apparition d'une nouvelle classe, indiquée par la lettre N, ne peut pas être détectée par l'algorithme existant, présenté ci-dessus, parce que tous les points de cette nouvelle classe sont classifiés dans la classe connue.

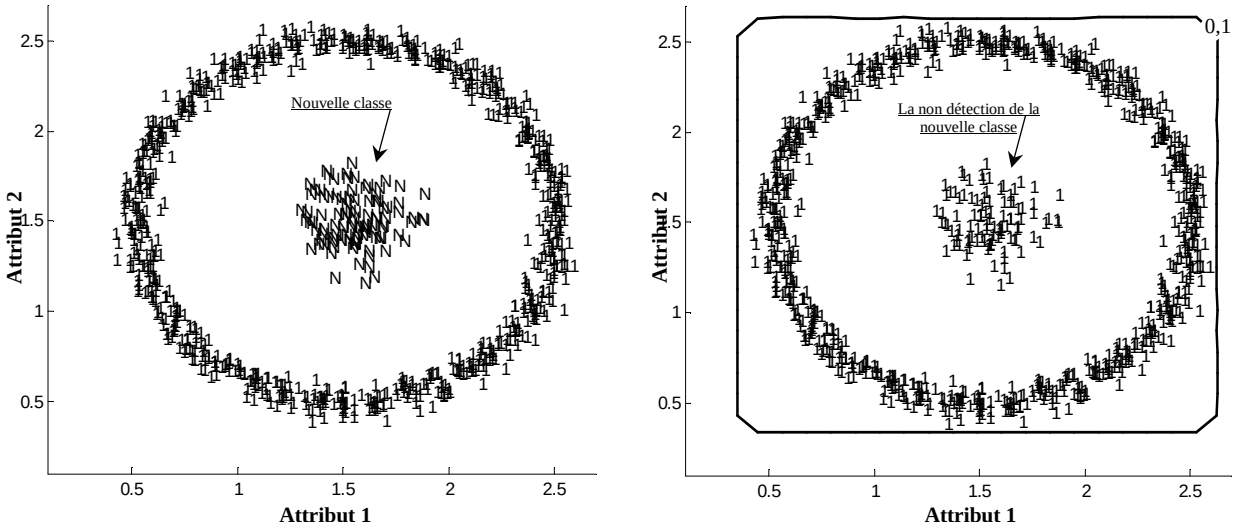


Figure 3.18 Apparition d'une nouvelle classe, indiquée par la lettre N, dans le cas d'une classe connue de forme non convexe, à gauche. Les points de cette nouvelle classe sont affectés à la classe connue par FPM, à droite. Donc la détection ne peut pas être réalisée par l'algorithme présenté ci-dessus.

3.3.2.3. Limites de la solution existante dans le cas non-stationnaire

Les performances de l'algorithme proposé dans [SAY02a] pour le cas des données non-stationnaires sont similaires à celles de FPM. Cette remarque est justifiée par le fait que les courbes d'appartenance de niveau 0,1 présentées dans les figure 3.12, 3.13, 3.14 et 3.15 pour les cas de rotation, de déplacement, de fusion et de scission respectivement, obtenues par FPM seront les mêmes que celles obtenues en utilisant la solution existante proposée dans [SAY02a]. Cela est dû à la non intégration d'un facteur d'oubli permettant d'oublier les données qui ne sont plus représentatives à l'instant courant. Cet algorithme ne peut donc pas être appliqué dans le cas des données non-stationnaires.

3.3.3. Solution proposée pour le cas stationnaire en présence des classes non convexes

3.3.3.1. Intégration de l'apprentissage incrémental à FPMA

La méthode FPMA, développée dans le chapitre 2, n'intègre aucun mécanisme de mise à jour des densités de possibilités après la classification de chaque nouveau point. Nous

proposons une solution à ce problème par la mise à jour des ensembles de corrélation ainsi que des densités de possibilité et cela de façon incrémentale.

Soit un nouveau point $\underline{x} \in C_i$ et se trouvant dans un hypercube non vide ayant un facteur de corrélation $\alpha_b^i = \left(\frac{n_b^i}{n_b}\right)^\beta$. $\left(\alpha_b^i\right)^+$ est la nouvelle valeur du facteur de corrélation de cet hypercube, après la classification de $\underline{x}$ dans la classe C_i, tel que :

$$\left(\alpha_b^i\right)^+ = \left(\frac{n_b^i + 1}{n_b + 1}\right)^\beta \Leftrightarrow \left(\alpha_b^i\right)^+ = \left((\alpha_b^i)^{\frac{1}{\beta}} . \frac{n_b}{n_b + 1} + \frac{1}{n_b + 1}\right)^\beta \qquad (3.3)$$

Les facteurs de corrélation $\alpha_b^j = \left(\frac{n_b^j}{n_b}\right)^\beta$ correspondant aux autres classes C_j tel que $i \neq j$ est :

$$\left(\alpha_b^j\right)^+ = \left((\alpha_b^j)^{\frac{1}{\beta}} . \frac{n_b}{n_b + 1}\right)^\beta \qquad (3.4)$$

Dans le cas où le nouveau point $\underline{x} \in C_i$ se trouve dans un hypercube vide *b*, c'est-à-dire ayant un facteur de corrélation $\alpha_b^i = 0$, la nouvelle valeur du facteur de corrélation de cet hypercube après la classification de $\underline{x}$ est $(\alpha_b^i)^+ = 1$.

La figure 3.19 présente le résultat de l'application de FPMA utilisant un apprentissage incrémental sur l'exemple de la figure 3.6. Puisque la classe a une forme non convexe, la nouvelle courbe de niveau d'appartenance, ayant la valeur 0,1, reflète la vraie forme de la classe.

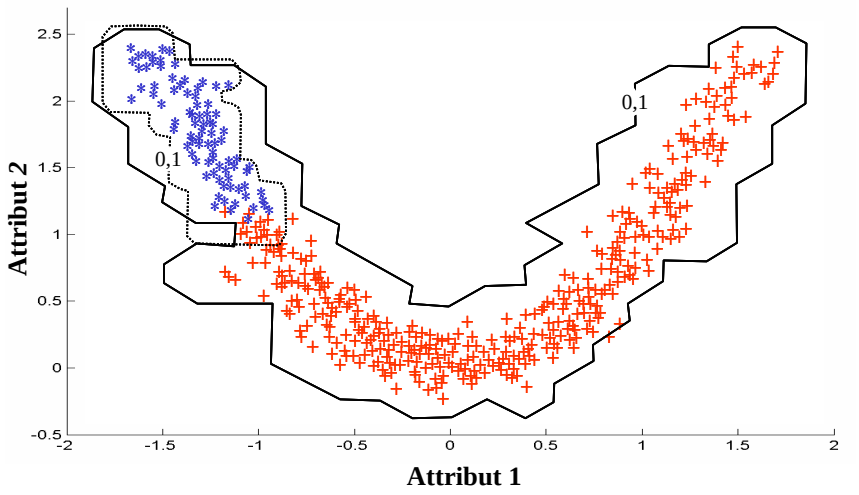


*Figure 3.19 Modification locale du contour de la classe due à la classification par FPMA de nouveaux points. " * " indique les points initiaux de l'ensemble d'apprentissage et " + " les nouveaux points affectés à la classe.*

3.3.3.2. Algorithme de détection basée sur FPMA et FPM non supervisée

La méthode FPMA permet de tenir compte de la forme non convexe des classes comme le montre la figure 3.20. Les points formant la nouvelle classe, indiquée par la lettre N, seront donc rejetés, par rapport à la classe non convexe, par FPMA.

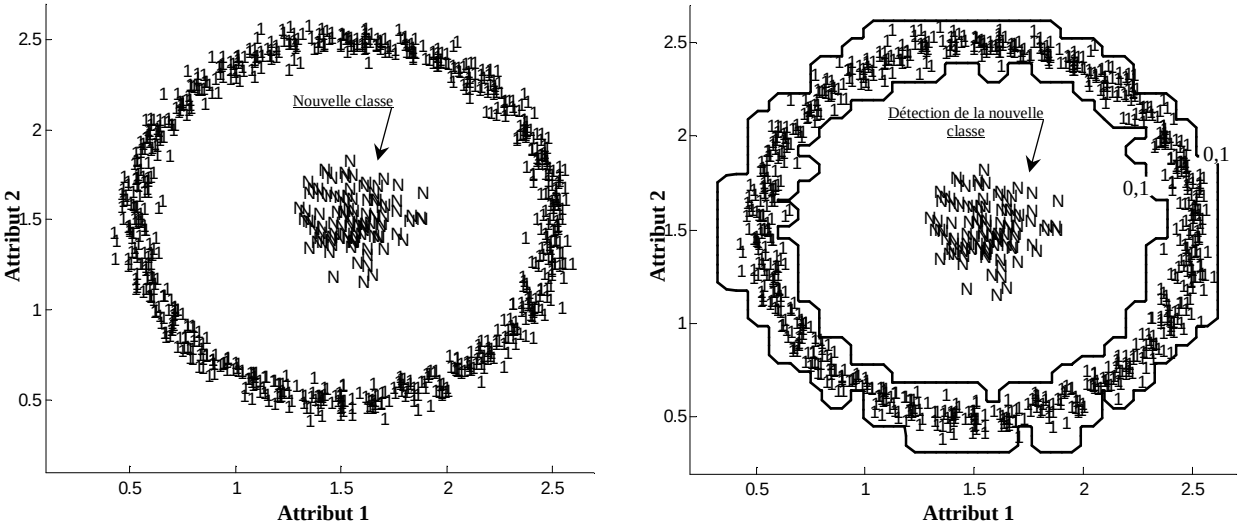


Figure 3.20 Apparition d'une nouvelle classe, indiquée par la lettre N, dans le cas d'une classe connue de forme non convexe, à gauche. Les points de cette nouvelle classe ne sont pas affectés à la classe connue par FPMA, à droite. La détection peut donc être réalisée.

3.3.3.2.1. Détection d'apparition d'une nouvelle classe

Afin de quantifier le rapprochement ou l'éloignement entre les points rejetés, nous proposons un algorithme utilisant la méthode FPM avec un apprentissage non supervisée. Cet algorithme considère chaque nouveau point rejeté par FPMA comme une nouvelle classe comportant un seul point d'apprentissage comme le montre la figure 3.21. L'histogramme de probabilité sera donc construit et transformé en histogramme de possibilité.

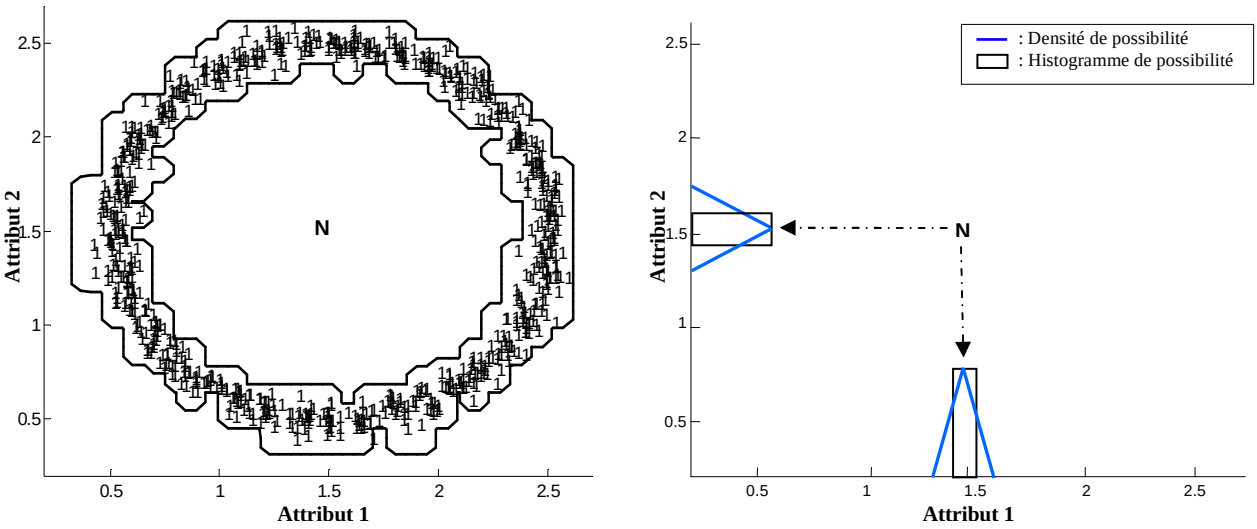


Figure 3.21 Apparition d'un nouveau point indiqué par la lettre N, dans la classe 1 de forme non convexe. Ce point est rejeté par FPMA, à gauche. Initialisation de l'algorithme de détection de l'apparition de nouvelles classes utilisant FPM non supervisée, à droite.

Ensuite la représentativité du deuxième nouveau point rejeté sera déterminée en calculant sa valeur de possibilité d'appartenance. Ce calcul est réalisé par une projection directe sur les densités de possibilité de cette nouvelle classe comme le montre la figure 3.22. Si le deuxième point est classifié dans cette nouvelle classe, cela indique qu'il est proche et donc les histogrammes de probabilités seront mis à jour d'une façon incrémentale.

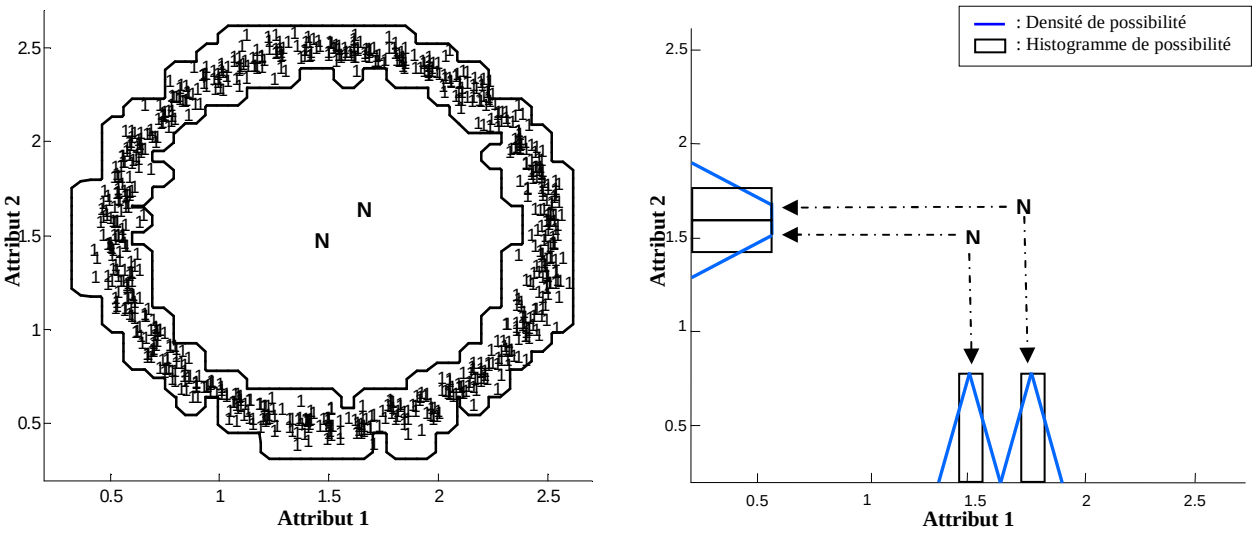


Figure 3.22 Apparition d'un nouveau point en plus de celui de la figure 3.21. Ce point est rejeté par FPMA par rapport à la classe connue, à gauche. Classification de ce nouveau point dans la nouvelle classe en utilisant FPM et mise à jour des densités de possibilité de la figure 3.21, à droite.

Ce processus sera répété pour chaque nouveau point rejeté, par rapport aux classes connues, et classifié dans la nouvelle classe comme le montre la figure 3.23. Par contre si un nouveau point est rejeté par cette nouvelle classe, une deuxième **nouvelle** classe sera créée en considérant que ce nouveau point est le seul point de l'ensemble d'apprentissage.

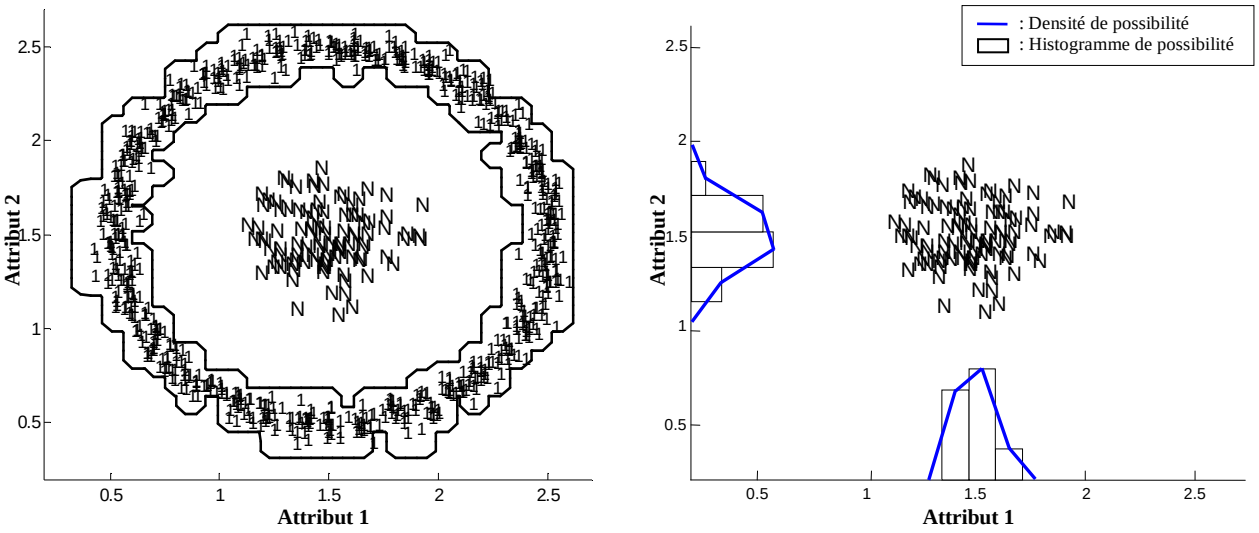


Figure 3.23 Apparition d'un certain nombre de nouveaux points, en plus de ceux de la figure 3.22. Ces points sont rejetés par FPMA par rapport à la classe connue, à gauche. Classification par FPM non supervisée et mise à jour des densités de possibilité de la figure 3.22, et détection de l'apparition d'une nouvelle classe, à droite.

Les étapes principales décrivant cet algorithme en utilisant FPMA et FPM non supervisée, sont illustrées dans la figure ci-dessous.

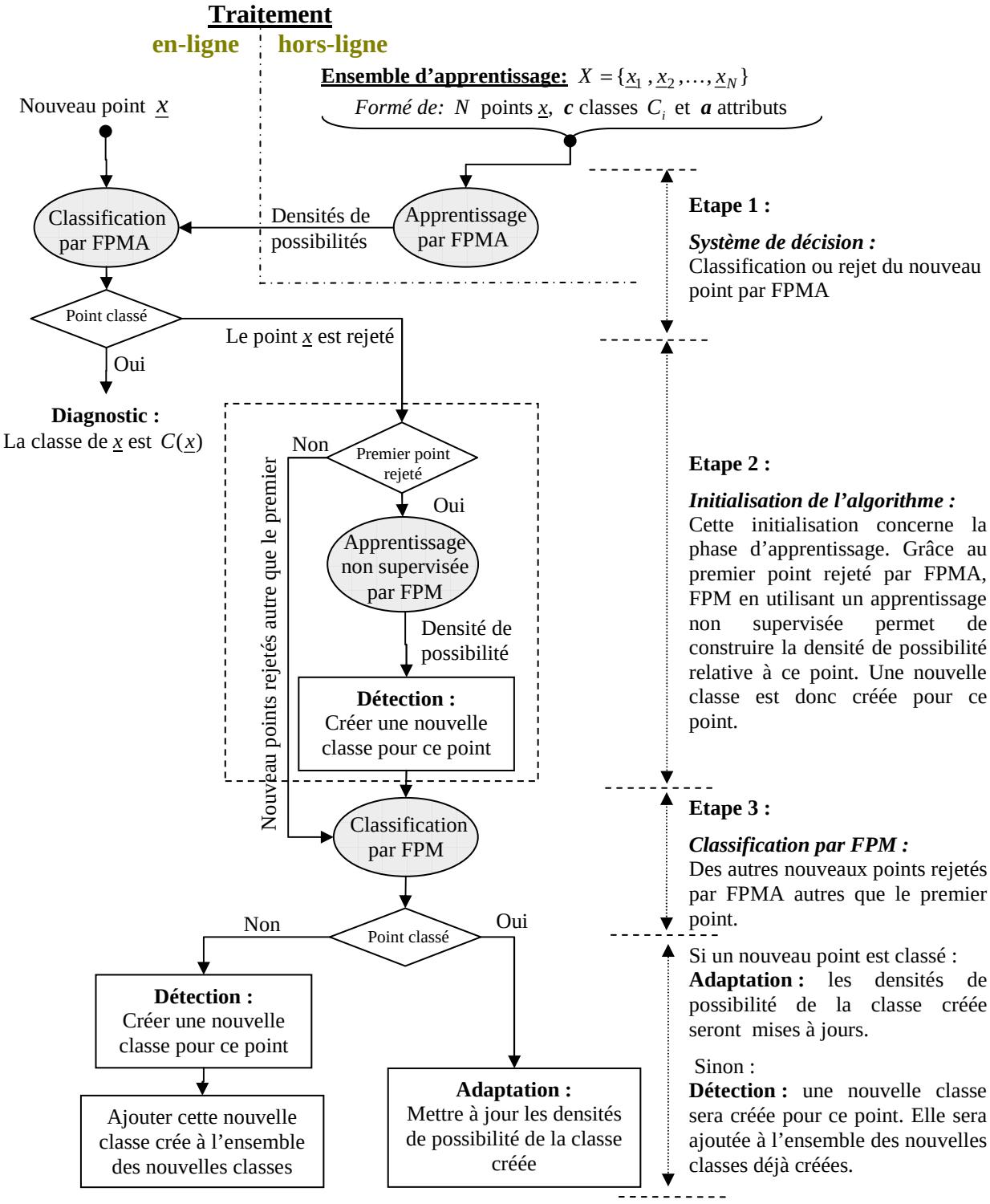


Figure 3.24 Algorithme de détection de nouvelles classes enligne dans le cas des classes connues de forme non convexe en utilisant FPMA et FPM avec apprentissage non supervisé.

L'algorithme décrit ci-dessous considère le premier nouveau point rejeté par FPMA comme le seul point de l'ensemble d'apprentissage d'une nouvelle classe. Ensuite le prochain

nouveau point sera soit classifié par rapport à cette nouvelle classe soit rejeté. Dans le premier cas les densités de possibilité seront mises à jour en ligne en utilisant l'apprentissage incrémental. Par contre dans le deuxième cas une deuxième classe sera créée pour le point rejeté. Les nouvelles classes seront validées par rapport à leur cardinalité, cf. section 3.3.3.2.3. Des nouvelles classes peuvent être fusionnées si un certain nombre de points sont affectés à plusieurs classes, cf. section 3.3.3.2.4. Toutefois cette solution ne permet pas le suivi de l'évolution et la prédiction de son sens parce que cette solution quantifie l'éloignement ou le rapprochement entre les points rejetés et non pas par rapport aux différentes classes.

3.3.3.2.2. Détection d'apparition de trois nouvelles classes

L'application de cet algorithme sur l'exemple de la figure 3.25 conduit à la détection de trois classes de forme convexe dans la classe 1 qui est non convexe.

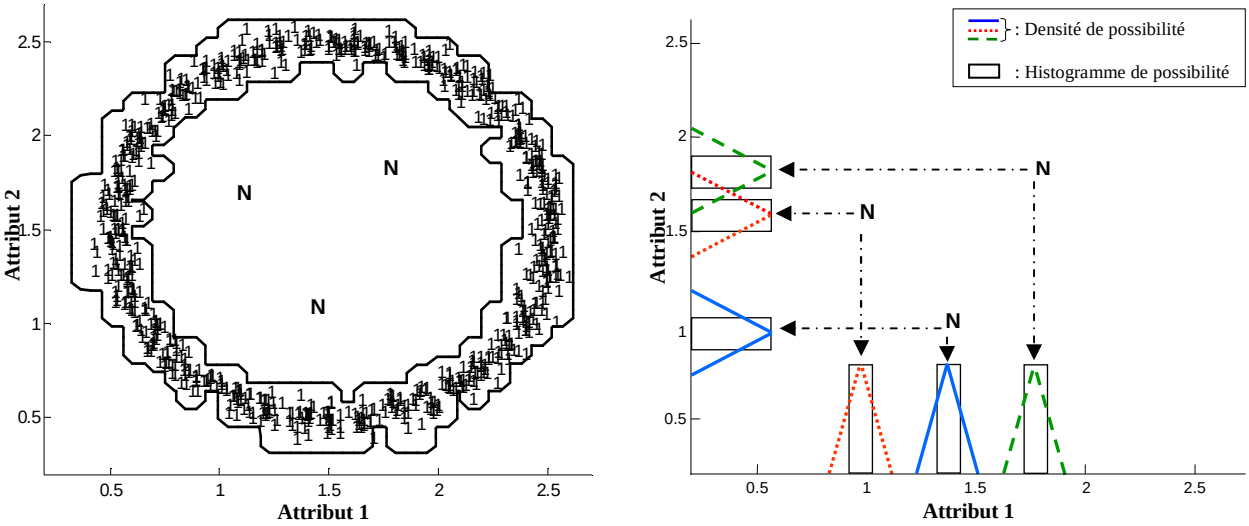


Figure 3.25 Apparition de trois nouveaux points indiqués par la lettre N, dans le cas d'une classe de forme non convexe. Ces points ne sont pas affectés à la classe connue par FPMA, à gauche. Initialisation de l'algorithme de détection (création de trois nouvelles classes par FPM non supervisée), à droite.

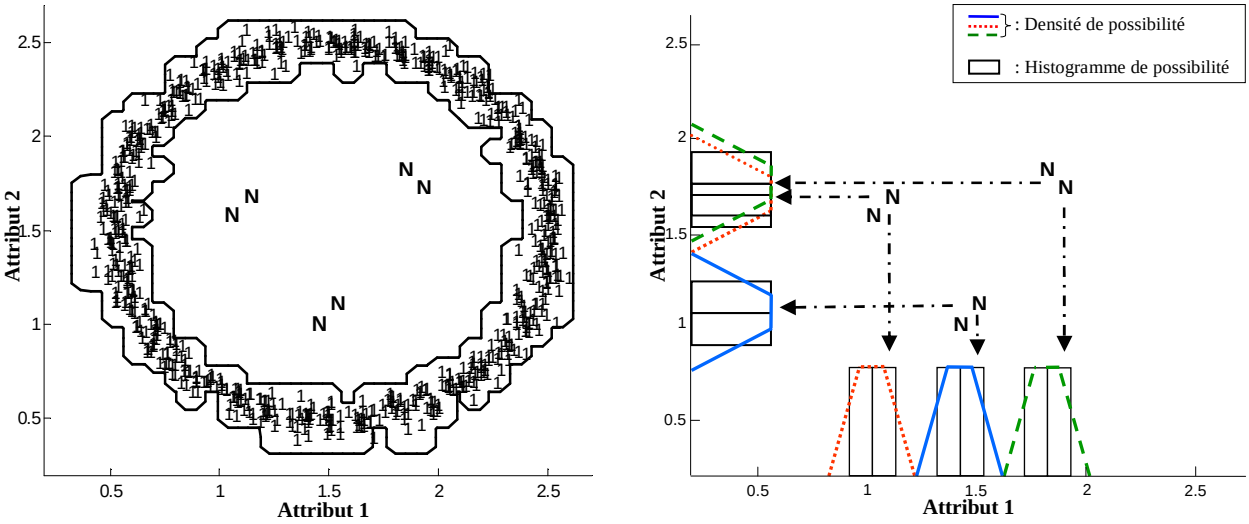


Figure 3.26 Apparition de trois nouveaux points en plus de ceux de la figure 3.25. Ces points ne sont pas affectés à la classe connue par FPMA, à gauche. Mise à jour des densités de possibilité de la figure 3.25, à droite.

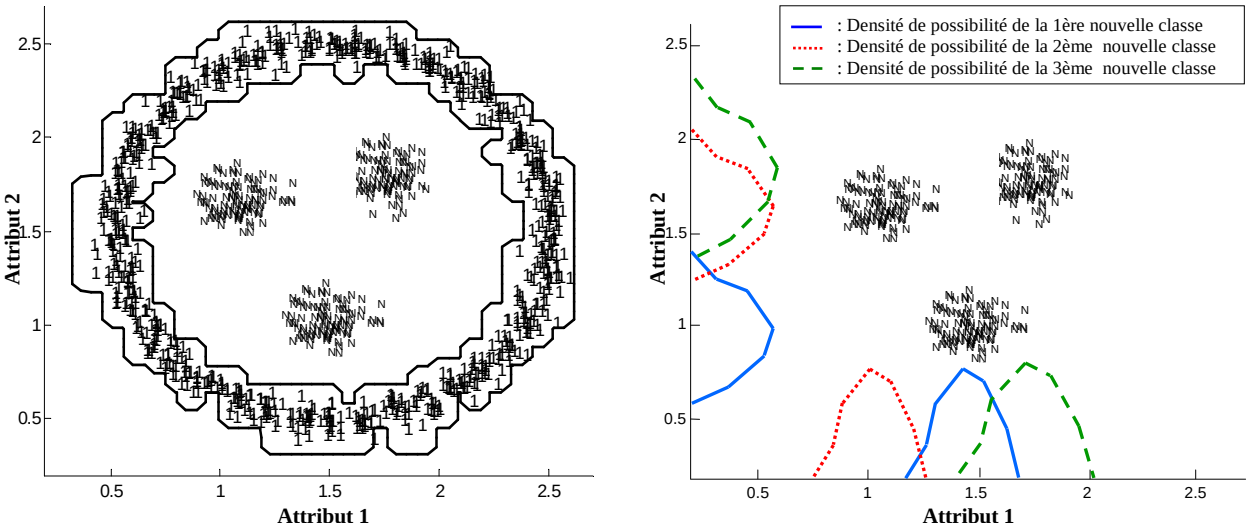


Figure 3.27 Apparition d'un certain nombre de nouveaux points, indiqués par la lettre N, en plus de ceux de la figure 3.26. Ces points ne sont pas affectés à la classe connue par FPMA, à gauche. Mise à jour des densités de possibilité de la figure 3.26 et détection de l'apparition de trois nouvelles classe à droite.

3.3.3.2.3. Détection des points aberrants

Les classes qui sont créées doivent ensuite être validées. Cette validation peut être basée soit sur l'expert soit sur un critère de cardinalité de points dans les nouvelles classes. Toute classe qui ne contient pas un certain nombre de points sera supprimée et ses points sont considérés comme des points aberrants comme ceux de la figure 3.28.

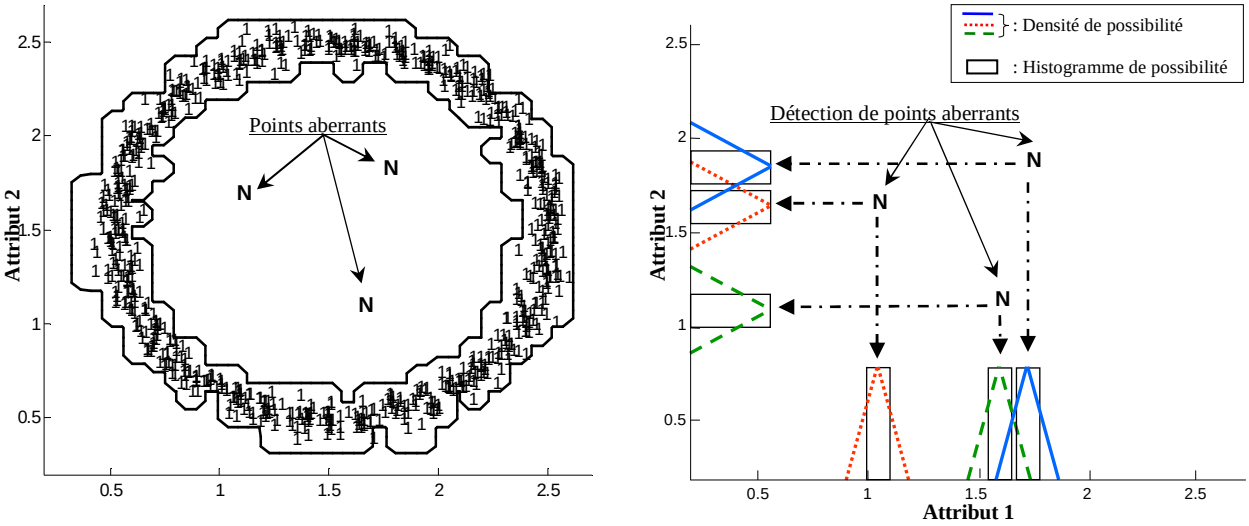


Figure 3.28 Apparition de trois nouveaux points indiqués par la lettre N, dans le cas d'une classe connue de forme non convexe. Ces points ne sont pas affectés à la classe connue par FPMA, à gauche. Initialisation de l'algorithme de détection de l'apparition de nouvelles classes (création de trois nouvelles classes par FPM non supervisée), à droite. Ces trois classes ne sont pas validées parce qu'elles contiennent un nombre très faible de points, un seul point dans chaque nouvelle classe.

3.3.3.2.4. Fusion des classes

Des classes peuvent être fusionnées soit en se basant sur l'expert soit sur un critère de nombre de points en commun entre ces classes. Les points en communs sont les points affectés en même temps à plusieurs classes. Par exemple dans la figure 3.29, les numéros des points indiquent leurs ordres ou instants d'arrivées. A l'instant t = 11, cf. figure 3.29 à gauche, l'algorithme proposé ci-dessus détecte l'apparition de deux nouvelles classes séparables C_1 et C_2. Cela grâce aux densités de possibilités marginales de ces deux classes, à cet instant, qui sont aussi séparables. Entre les instants t = 12 et t = 30, au fur et à mesure que de nouveaux points apparaissent, les densités de possibilités des deux classes précédemment détectées sont misent à jour. Cela, conduit au grossissement de leurs contours. Ce grossissement conduit à l'existence de points ambigus entre ces deux nouvelles classes, comme le montre la figure 3.29.

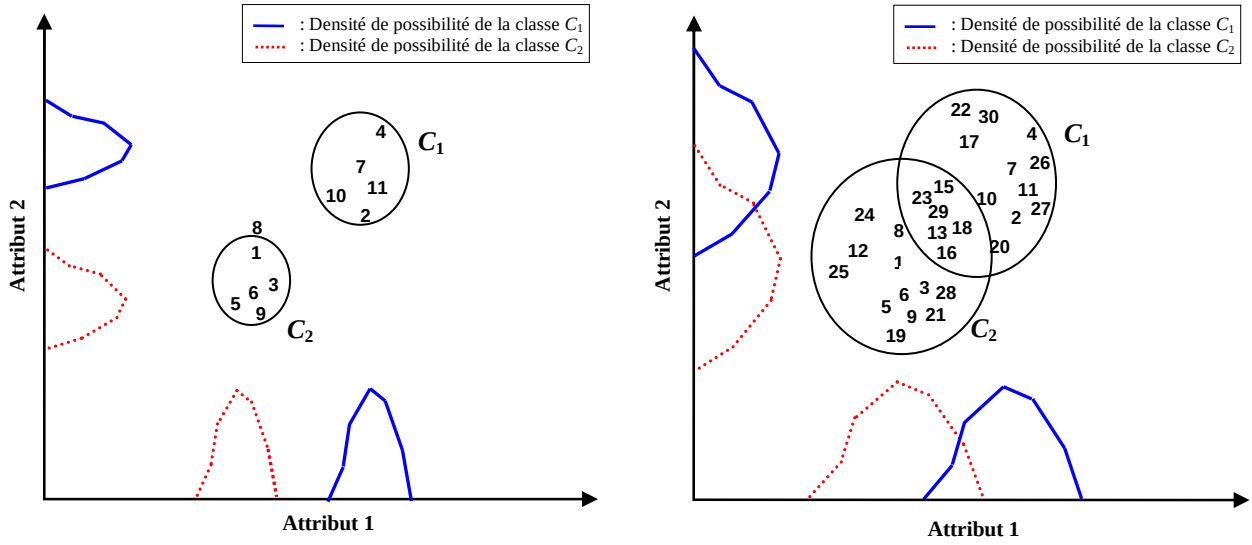


Figure 3.29 Les numéros des points indiquent leurs ordres ou instants d'arrivés. Détection d'apparition de deux nouvelles classes disjointes C_1 et C_2 à l'instant t = 11 et leurs densités de possibilités marginales mises à jour à cet instant, à gauche. Grossissement, entre t =12 et t=30, des classes détectées et leurs densités de possibilités marginales mises à jour, à droite.

Les points observés aux instant t = {13, 15, 16, 18, 23, 29} sont des points ambigus puisqu'ils ont deux valeurs de possibilité d'appartenance non nulles aux deux classes. Selon un critère de cardinalité fixé à 7 points, cf. figure 3.30 à gauche, les deux classes C_1 et C_2 doivent être fusionnées en une seule classe C_1 à t = 31. Dans ce cas, les densités de possibilité marginales de ces deux classes doivent aussi être fusionnées en une seule densité marginale, cf. figure 3.30 à droite.

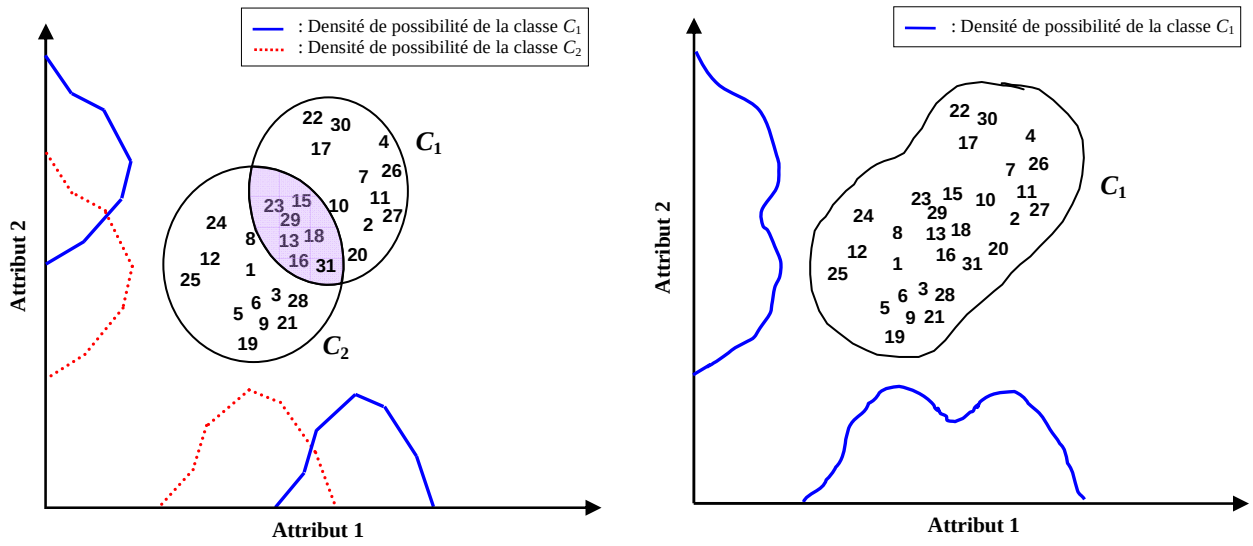


Figure 3.30 Détection de points ayant des possibilités d'appartenances aux deux classes C_1 et C_2, selon un critère de cardinalité égal à 7 points, à gauche. Les deux classes C_1 et C_2 sont fusionnées en une seule classe C_1 ainsi que leurs densités de possibilité marginales en une seule densité, à droite.

3.4. Conclusion

Le diagnostic des systèmes dynamiques par Reconnaissance des Formes (RdF) nécessite une méthode de classification à la fois adaptative et prédictive. L'adaptativité permet à la méthode de classification de détecter les états inconnus et de les apprendre, cela, afin d'enrichir sa connaissance. Le caractère prédictif, quant à lui, permet de suivre l'évolution du système entre plusieurs modes de fonctionnement. Il est intéressant de prédire le sens de cette évolution afin d'anticiper la dérive du système d'un mode de fonctionnement normal vers un mode anormal. L'enrichissement de la base de données est basé sur l'extraction de l'information manquante, sur l'évolution du système, de chaque nouveau point et l'inclure à la base de connaissances.

Les classes regroupant des données stationnaires sont appelées classes statiques. Pour ces données, les caractéristiques du modèle de classification ne changent pas au cours du temps. De plus, l'enrichissement de la base de connaissances se traduit par une déformation locale de la forme des classes sans mettre en cause les informations acquises précédemment.

Toutefois, la plupart des données issues du monde réel, dans lequel l'environnement est en perpétuelle évolution, sont des données non-stationnaires. Les classes regroupant ce type de données sont appelées classes dynamiques. Pour ces classes, les caractéristiques du modèle de classification varient au cours du temps.

Dans la littérature, de nombreux auteurs se sont intéressés au problème du diagnostic adaptatif et prédictif et ont apportés une multitude de solutions pour le cas des données stationnaires. Par contre, peu de solutions concernant la classification dynamique. Concernant celles qui sont basées sur FPM, il existe une seule solution qui a été proposée dans [SAY02a].

Dans ce chapitre, nous avons étudié les performances de FPM et de la solution existante afin d'en déduire leurs limites. Nous avons démontré, à travers des exemples, que la solution existante n'est applicable que dans le cas des classes de forme convexe. En effet, cette solution quantifie la représentativité de chaque nouveau point rejeté par rapport à toutes les classes, de formes convexes, connues. Cette quantification est basée sur la distance entre le point rejeté et son plus proche voisin ainsi que sur la valeur d'appartenance de ce plus proche

voisin. De plus, cette solution est inopérante pour la classification des données non-stationnaires décrivant des situations de rotation, de déplacement, de fusion et de scission des classes.

Pour remédier aux limites de la solution existante pour le cas stationnaire et dans le cas des classes de forme non-convexe, nous avons proposée une amélioration de FPM Améliorée (FPMA), développée dans le chapitre 2, en lui intégrant d'abord l'apprentissage incrémental. Rappelons que FPMA respecte la forme non convexe des classes. L'apprentissage incrémental permet à FPMA de suivre la déformation des contours des classes par la mise à jour des densités de possibilités après la classification de chaque nouveau point. Ensuite, nous avons proposé un algorithme basé sur FPMA et FPM non supervisée qui permet de détecter en ligne l'apparition de nouvelles classes pour le cas des classes de formes convexes ou non-convexes. Enfin, nous avons illustré les performances de cet algorithme dans le cas d'apparition d'une nouvelle classe, de plusieurs nouvelles classes, des points aberrants et enfin le cas d'apparition de deux nouvelles classes et leurs fusions en une seule classe. Ces cas sont illustrés à travers des exemples académiques. Cependant, cet algorithme sous cette version ne permet pas de suivre l'évolution entre les classes car il quantifie l'éloignement ou le rapprochement entre les points rejetés et non pas par rapport aux différentes classes.

L'algorithme proposé dans ce chapitre ne permet pas de suivre l'évolution des classes dynamiques. En effet les densités de possibilités ne sont en aucun moment mises en cause et aucun facteur d'oubli n'est utilisé. Le développement de FPM pour qu'elle puisse suivre l'évolution des classes dynamiques nécessite donc l'intégration du temps pour chaque point de l'ensemble d'apprentissage afin de pouvoir oublier les points les plus anciens au profit des points les plus récents. Ces limites induiront les perspectives de cette thèse.

Conclusion générale et perspectives

Conclusion générale

Le diagnostic consiste à comparer l'information instantanée issue du système à la connaissance *a priori* disponible sur le fonctionnement de ce système. Si cette connaissance est numérique, issue des capteurs, et si elle n'est pas suffisante pour construire un modèle du comportement normal et/ou défaillant du système, les méthodes de Reconnaissance des Formes (RdF) sont particulièrement adaptées pour réaliser le diagnostic. Les modes de fonctionnement, normaux ou défaillants, sont représentés par des classes. Le diagnostic par RdF consiste à classifier chaque nouvelle observation sur le fonctionnement du système, représentée par un point dans l'espace de représentation, dans une des classes apprises. La connaissance de la classe de la nouvelle observation permet de déterminer le mode actuel de fonctionnement du système.

Il existe plusieurs méthodes de RdF. Nous avons choisi la méthode Fuzzy Pattern Matching (FPM) pour sa simplicité, son temps de classification constant et faible, sa capacité à traiter des données à la fois incertaines et imprécises et pour sa sélection des sources d'informations les plus pertinentes. Toutefois FPM est une méthode naïve, c'est-à-dire qu'elle considère les paramètres comme statistiquement indépendants. Aucune information sur la forme non convexe des classes ou la corrélation entre les paramètres n'est intégrée pendant la phase d'apprentissage. Cela rend FPM inutilisable pour la discrimination non linéaire des classes ou pour les classes ayant des axes non parallèles à ceux de l'espace de représentation. De plus, FPM ne peut détecter l'apparition d'aucune nouvelle classe ni l'évolution entre les classes. FPM n'est donc pas une méthode de classification adaptative et prédictive. Elle ne peut pas non plus être intégrée dans un module de diagnostic adaptatif et prédictif.

Les travaux effectués au cours de cette thèse nous ont permis d'apporter trois contributions. La première contribution consiste d'abord en une comparaison entre les méthodes de classification basées sur les théories de représentation et de traitement de l'information imparfaite, à savoir les théories des probabilités, des fonctions de croyance et des possibilités, afin de montrer leurs points forts et leurs points faibles, et de justifier notre choix de la théorie des possibilités. Ensuite, Une amélioration de la Transformation Variable (TV), permettant de passer de la théorie des probabilités en possibilités et présenté dans [SAY06], est proposée. La TV améliorée fournie une distribution de possibilité aussi spécifique que celle de Dubois et Prade optimale tout en respectant la condition de cohérence de Dubois et Prade dans le cas continu. Egalement la TV améliorée est plus simple à mettre en œuvre, surtout dans le cas d'ignorance totale, que celle de Dubois et Prade optimale. Les transformations optimale et non optimale de Dubois et Prade ainsi que la TV améliorée ont été évaluées selon plusieurs critères afin de montrer l'intérêt de la TV améliorée pour les applications de la reconnaissance des formes. Plusieurs exemples académiques et réels sont utilisés pour illustrer et tester ces transformations.

La deuxième contribution propose une solution pour que FPM soit opérante dans le cas des classes de forme non-convexe et/ou décrites dans un espace de représentation formé de paramètres corrélés. Nous avons appelé FPM après l'intégration de cette solution, qui préserve les avantages de FPM, FPM Améliorée (FPMA). Cette solution enrichit l'information donnée par les distributions de possibilités marginales par une information relative à la distribution conjointe. Cet enrichissement repose sur le calcul d'un ensemble de

facteurs de corrélation flous entre les attributs de l'espace de représentation pour chaque classe. Chaque facteur de corrélation est défini pour un hypercube formé de l'intersection des barres des histogrammes construits marginalement à partir des données d'apprentissage. Ces facteurs sont calculés uniquement pour les hypercubes de l'espace de représentation contenant des points d'apprentissage. Ce facteur tient compte du nombre de points d'une même classe présents dans un hypercube. Cette solution nécessite la détermination d'aucun autre paramètre supplémentaire. Les performances de FPMA sont comparées à celles de FPM ainsi qu'à la méthode k plus proches voisins (kppv) et à la méthode des noyaux de Parzen selon deux critères d'évaluation, le taux d'erreur et le temps de classification. Cette comparaison est réalisée en utilisant plusieurs bases de données académiques et réelles issues de différents champs d'application. Les résultats obtenus montrent que les performances de FPMA sont meilleures que celles de FPM, de kppv et de Parzen. Cependant, les performances de FPMA se dégradent quand l'ensemble d'apprentissage est assez pauvre.

La troisième contribution concerne le développement de FPM pour qu'elle soit une méthode de classification adaptative et prédictive, sans connaissance *a priori* de la forme de la classe. Une méthode de classification adaptative permet de détecter les états inconnus, représentés par l'apparition de nouvelles classes dans l'espace de représentation, et de les apprendre, afin d'enrichir sa connaissance. Le caractère prédictif, quant à lui, permet de suivre l'évolution du système entre plusieurs modes de fonctionnement. Cette détection et cette prédiction sont basées sur l'extraction de l'information manquante et sur l'évolution du système à partir de chaque nouveau point. FPM ne peut pas extraire cette information puisqu'elle rejette ces points.

Pour remédier à ce problème, une solution a été proposée dans [SAY02a]. Cette solution quantifiée la représentativité de chaque nouveau point rejeté par rapport à toutes les classes connues. Cette quantification est basée sur la distance entre le point rejeté et son plus proche voisin ainsi que sur la valeur d'appartenance de ce plus proche voisin. Les performances de cet algorithme ont été testées sur plusieurs exemples simulés et réels. Toutefois cet algorithme ne peut pas être appliqué pour les classes de forme non convexe.

Nous avons donc proposé une solution afin de remédier à cet inconvénient. D'abord, nous avons proposée de développer FPM Améliorée (FPMA) en lui intégrant l'apprentissage incrémental. Rappelons que FPMA a l'avantage d'être adaptée pour les classes de formes non convexe. L'apprentissage incrémental permet à FPMA de suivre la déformation des contours des classes par la mise à jour des densités de possibilités après la classification de chaque nouveau point. Ensuite, nous avons proposé un algorithme basé sur FPMA et FPM non supervisée qui permet de détecter en ligne l'apparition de nouvelles classes pour le cas des classes de forme convexe ou non-convexe. En effet, cet algorithme quantifie les points qui sont rejetés par la méthode FPMA. Pour ces points rejetés une nouvelle classe sera créée. Ensuite le prochain nouveau point sera soit classifié par rapport à cette nouvelle classe soit rejeté. Dans le premier cas les densités de possibilité seront mises à jour en ligne en utilisant l'apprentissage incrémental. En revanche, dans le deuxième cas, une deuxième classe sera créée pour le point rejeté. Les nouvelles classes seront validées par rapport à leur cardinalité. Des nouvelles classes peuvent être fusionnées si un certain nombre de points sont affectés à plusieurs classes. Toutefois cette solution ne permet pas le suivi de l'évolution et la prédiction de son sens parce que cette solution quantifié l'éloignement ou le rapprochement entre les points rejetés et non pas par rapport aux différentes classes.

Perspectives

Extension du champ d'application

Nous avons testé les contributions proposées dans ce mémoire de thèse sur des bases de données issues de systèmes industriels et biomédicaux. Il est envisageable et intéressant de tester et de développer ces contributions sur d'autres applications issues du domaine de la reconnaissance de la parole, de signatures, d'images, etc. Nous n'avons pas pu également implanter l'algorithme adaptatif et prédictif dans des systèmes réels et nous n'avons donc pas pu tester ses performances en temps réel.

Association de la Reconnaissance statistique et de la Reconnaissance structurelle

On distingue deux catégories de méthodes de RdF : les méthodes statistiques et les méthodes structurelles. Les méthodes statistiques s'appuient sur une représentation purement quantitative des paramètres de l'espace de représentation. Un objet à classer sera, pour ce type de méthode, un vecteur de taille fixe égale au nombre de paramètres de l'espace de représentation. Les avantages principaux de ces méthodes sont leur indépendance du domaine d'application, leur facilité de mise en œuvre et leur temps de calcul relativement faible. Toutefois la nature quantitative des paramètres empêche ces méthodes d'être applicables pour des données corrélées ou interconnectées représentant une structure. Dans ce cas, Il est intéressant d'utiliser les méthodes de RdF structurelles, qui sont basées sur l'utilisation de grammaires. Ces dernières modélisent les relations entre les composants élémentaires d'une forme. Toutefois, ces grammaires sont construites manuellement et dépendent du domaine de l'application. Il est donc intéressant de combiner les avantages de ces deux catégories de méthodes en réalisant l'extraction des caractéristiques informatives du système par une méthode de Reconnaissance des Formes structurelle (méthodes de segmentation), et la classification par une méthode de Reconnaissance des Formes statistiques (Fuzzy Pattern Matching par exemple).

Combinaison de la connaissance subjective et objective : Méthode hybride

Nous avons commencé à développer une approche combinant la méthode FPM statistique et la connaissance de l'expert représentée par un ensemble de règles de type « Si-Alors » [BOU07d]. La connaissance de l'expert est utile afin de fournir une information sur la forme non-convexe des classes et/ou la corrélation entre les attributs. Toutefois cette connaissance est nette et elle ne renseigne pas sur la graduation de l'appartenance d'un point à une classe. La combinaison entre les règles fournies par l'expert et l'évaluation issue de la méthode FPM statistique est donc intéressante pour, d'une part, tenir compte de la corrélation entre les attributs et d'autre part renseigner sur le degré d'appartenance à une classe. Toutefois la génération des règles reste encore un sujet à développer.

La figure ci-dessous montre un exemple d'une classe de forme oblique linéaire. L'application de FPM statistique sur les données constituant cette classe fournie des courbes de niveau d'appartenance rectangulaires qui ne respectent pas la forme oblique de la classe, cf. figure ci-dessous à gauche. La combinaison entre FPM statistique et une règle fournie par un expert, renseignant sur la corrélation entre les deux attributs ou la forme oblique de la classe, permet d'obtenir des courbes de niveau d'appartenance riches et respectant la forme de la classe, cf. figure ci-dessous à droite [BOU07d].

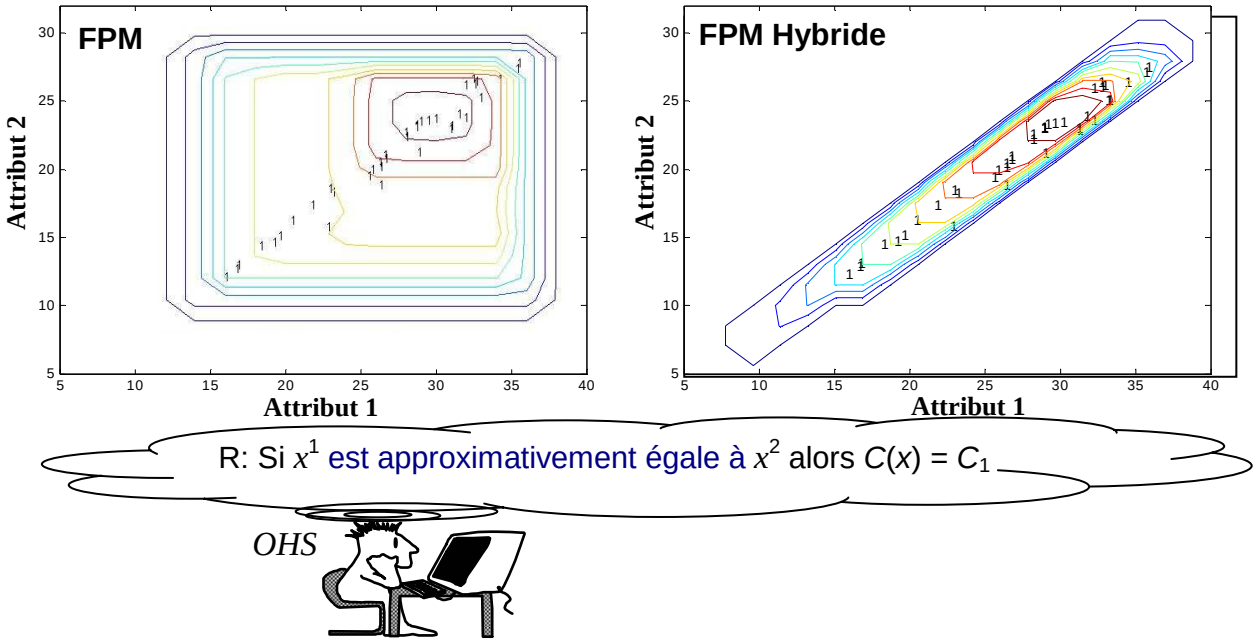


Figure i.i Combinaison entre FPM statistique et une règle fournie par un expert

Classification des données non-stationnaires

Les classes sont basées soit sur des données stationnaires, et dans ce cas elles sont statiques, soit non-stationnaires et là elles sont dynamiques. Les caractéristiques du modèle de classification pour les classes statiques ne changent pas au cours du temps. Dans ce cas, l'enrichissement de la base de connaissance se traduit par une déformation locale de la forme des classes sans mettre en cause des informations acquises précédemment. Toutefois la plupart des données issues du monde réel sont des données non-stationnaires parce que l'environnement est en perpétuelle évolution. Ce type de données conduit à la variation des caractéristiques du module de classification au cours du temps.

L'algorithme proposé dans ce mémoire de thèse ne permet pas de suivre l'évolution des classes dynamiques. En effet les densités de possibilité sont à aucun moment mises en cause et aucun facteur d'oubli n'est utilisé. Le développement de FPM pour le suivi de l'évolution des classes dynamiques nécessite d'une part l'intégration du temps d'arrivée de chaque point de l'ensemble d'apprentissage et d'autre part la définition d'un critère permettant l'évaluation de la validité des points affectés précédemment, cela afin de pouvoir oublier les points les plus anciens au profit des points les plus récents.

Bibliographie

[ALL98] ALLA H., DAVID R., "*Continuous and hybrid Petri nets*", Journal of Circuits, Systems and Computer, 8 (1), pp.159-188, 1998.

[AMA06] AMADOU-BOUBACAR H., "*Classification Dynamique de Données non-stationnaires: Apprentissage séquentiel des Classes évolutives*". Thèse de Doctorat, Université des Sciences et Technologies de Lille (USTL), France, 2003, 2006.

[BEZ81] BEZDEK J. C., "*Pattern recognition with fuzzy objective function algorithms*", New-York, Plenum Press, 1981.

[BLO96] BLOCH I., "*Information Combinaison Operators for Data Fusion : A comparative Review with Classification*", IEEE Trans on SMC, Vol 26 N°1, 1996.

[BOU97] BOUDAOUD N., "*Conception d'un système de diagnostic adaptatif en ligne pour la surveillance des systèmes évolutifs*0, Thèse de Doctorat, Université de Technologie de Compiègne, France, 1997.

[BOU96] BOUTLEUX E., "*Diagnostic et suivi d'évolution de l'état d'un système par reconnaissance des formes floues. Application au modèle du réseau téléphonique français*", Thèse de Doctorat, Université de Technologie de Compiègne, France, 1996.

[BOU07a] BOUGUELID M. S., SAYED MOUCHAWEH M., BILLAUDEL P., "*New results on diagnosis by pattern recognition*", 4th ICINCO/IFAC International Conference Informatics in Control, Automation and Robotic, pp. 167-172, Angers, France, 2007.

[BOU07b] BOUGUELID M. S., Sayed MOUCHAWEH M., BILLAUDEL P., "*Fuzzy pattern Matching améliorée pour la discrimination non linéaire des classes*", 2èmes Journées Doctorales MACS (JDMACS 2007), Reims, France, 2007.

[BOU07c] BOUGUELID M. S., SAYED MOUCHAWEH M., BILLAUDEL P., "*Classification par apprentissage floue pour le diagnostic à base de reconnaissance des formes*", Rencontres francophones sur la Logique Floue et ses Applications, LFA 2007, Nîmes, 22 & 23 Novembre 2007.

[BOU07d] BOUGUELID M. S., SAYED MOUCHAWEH M. BILLAUDEL P., RIERA B., "*Hybrid pattern recognition method to diagnose dynamic systems*", 10th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (IFAC-HMS 2007), September 4-6, Seoul, Korea, 2007.

[BOU07e] BOUGUELID M. S., SAYED MOUCHAWEH M. and BILLAUDEL P. "*Adaptative and predictive diagnosis based on Pattern Recognition*", 11th IEEE International Conference of intelligent Engineering and Systems, pp.139-144. Budapest, Hungary, 2007.

[BUR98] BURGES C. "*A tutorial on support vector machines for pattern recognition*", Data Mining and Knowledge Discovery, 2-2, 1998.

[CAD04] CADENAS J.M., GARRIDO M.C. and HERNANDEZ J.J., "*Improving fuzzy pattern matching techniques to deal with non discrimination ability features*, IEEE International Conference on Systems, Man and Cybernetics, pp. 5708-5713, 2004.

[CHA93] CHATAIN JEAN-NOEL, "*Diagnostic par systèmes experts*", Traité des Nouvelles Technologies, Hermès, Paris, 1993.

[CHI04] CHIANG L.H., KOTANCHEK M.E. and KORDON A.K. "*Fault diagnosis based on fisher discriminant analysis and support vector machines*", Computers and Chemical Engineering, 28(8), pp.1389–1401, 2004.

[CHO57] CHOW C. K., "*An optimum character recognition system using decision*", R.E. Transaction Electronic Compiuter, pp. 247-254, 1957.

[COR02] CORN G., DUBUISSON B., "*Pattern characteristics of an evolution between two classes*", Fuzzy Sets and Systems 126, pp.293-310, 2002.

[COV67] COVER T. M. and HART P.E., "*Nearest neighbor pattern classification*", IEEE Trans. Information Theory, Vol. IT-13, pp. 21-27, Januray 1967.

[CRI00] CRISTIANINI N., SHAWE-TAYLOR J., "*Introduction to Support Vector Machines*", Cambridge University Press, 2000.

[DAR06] DARLEA L., GALICHET S. et Valet L., "*Analyse et prétraitement d'un ensemble d'apprentissage dans le cadre d'un système coopératif*", Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA), pp. 53-60, 2006.

[DEN06] DENOEUX T. and SMETS P., "*Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches*", IEEE Transactions on Systems, Man and Cybernetics B , Vol. 36, Issue 6, pp. 1395-1406, 2006.

[DEV77] DEVIJVER P. A., "*Reconnaissance des formes par la méthode des plus proches voisins*", Thèse de 3ème cycle, Université Paris 6, juin 1977.

[DEV93] DEVEUGHELE S., "*Etude d'une méthode de combinaison adaptative d'informations incertaines dans un cadre possibiliste*", Thèse de Doctorat, Université de Technologie de Compiègne, France, 1993.

[DEV99] DEVILLEZ A., "*Contribution à la classification floue de données comportant des classes de forme quelconque. Application au développement d'un module d'aide à la décision*", Thèse de Doctorat, Université de Reims, France, 1999.

[DEV04] DEVILLEZ A., "*Four Fuzzy supervised classification methods for discriminating classes of non convex shape*", Fuzzy sets and systems, Vol. 141, pp. 219-240, 2004.

[DOW93] DOWNS J.J. and VOGEL E.F., "*Plant-wide industrial process control problem*", Computers and Chemical Engineering 17(3), pp.245–255, 1993.

[DUB86] DUBOIS D. and PRADE H., "*A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets*", International Journal of General Systems, 12 , pp.193–226, 1986.

[DUB87] DUBOIS D. et PRADE H., "*Théorie des possibilités Application à la représentation des connaissances en informatique*", Deuxième édition, Masson, 1987.

[DUB88] DUBOIS D. and PRADE H., "*Possibility Theory*" Plenum Press, New-York, 1988.

[DUB90] DUBUISSON B., "*Diagnostic et reconnaissance des formes*", Traité des Nouvelles Technologies, série Diagnostic et Maintenance, HERMES, 1990.

[DUB92a] DUBOIS D. et PRADE H., "*Combination of information in the framework of possibility*". Dans M. Al ABIDI, Ed, Data fusion in robotics and machine intelligence. Academic Press edition, 1992.

[DUB92b] DUBOIS D. and PRADE H., "*Gradual inference rules in approximate reasoning*", Information Sciences, 61: pp.103–122, 1992.

[DUB93] DUBOIS D. and PRADE H., "*On possibility/probability transformations*", Fuzzy Logic, pp. 103-112, 1993.

[DUB01] DUBUISSON B., "*Automatique et statistiques pour le diagnostic*", Traité IC2 Information, commande, communication. Hermès Sciences, 2001.

[ENA86] ENAS G. and CHOI S., "*Choice of the smoothing parameter and efficiency of k-nearest neighbor classification*", Computers and Mathematics with Applcation 2, 12 A, pp.235-244, 1986.

[Fix51] FIX, E., and HODGES J. L. "*Discriminatory analysis: Non-parametric discrimination: Consistency properties*", pp. 261–279. USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[FRE92] FRELICOT C., "*Un système adaptatif de diagnostic prédictif par reconnaissance des formes floue*". Thèse présentée devant l'Université de Technologie de Compiègne, France, 1992.

[FRI99] FRIEDMAN M. and KANDEL A., "*Introduction to pattern recognition – statistical, structural, neural and fuzzy logic approaches*". Imperial College Press, London 1999.

[FUK72] FUKUNAGA K., "*Introduction to statistical pattern recognition*". Academic Press, New York, 1972.

[GAN87] GANA K. "*Suivi d'évolution et aide au pronostic en maintenance de système industriel*", Thèse de l'Université de Valenciennes et du Hainaut Cambrésis, France, 1987.

[GRA92] GRABISH M. and SUGENO M., "*Multi-attribute classification using fuzzy integral*", Proc. of fuzzy IEEE, pp. 47-54, 1992.

[GRA93] GRABISCH M., "*On the use of fuzzy integral as a fuzzy connective*". Proc. Of the 2nd IEEE Inter. Con. On Fuzzy Systems (FUZZ-IEEE'93), San Fransisco, March,1993.

[GRA94] GRABISCH M. and NICOLAS J.M., "*Classification by fuzzy integral: Performance and tests*", Fuzzy Sets Syst., vol. 65, pp. 255–271, 1994.

[GRA95] GRABISCH M., "*Fuzzy integral in multidecision making*", Fuzzy sets and systems 69, pp.279-298, 1995.

[GRA00] GRABISCH M., "*Fuzzy integral for classification and feature extraction*", In Fuzzy Measures and Integrals, Theory and Applications, M. Grabisch, T. Murofushi, and M. SUGENO (eds), Physica Verlag, 348-374, 2000.

[GRE84] GRENIER M. "*Méthode de détection d'évolution*", Thèse de Doctorat, Université de Technologie de Compiègne, France, 1984.

[GUI05] GUIGUE V., "*Méthodes à noyaux pour la représentation et la discrimination de signaux non-stationnaires*", Thèse de doctorat, INSA de Rouen, 2005.

[GUS79] GUSTAFSON D.E., KESSEL W.C., "*Fuzzy clustering with a fuzzy covariance matrix*", Proc. of IEEE-CDC, San Diego, 1979.

[HEL70] HELLMAN M. E., "*The nearest neighbor classification rule with a reject option*", IEEE. Trans on System, Man and Cybernetics, Vol 6, no 3, pp 179-185, 1970.

[HE97] HE K. and MEEDEN G., "*Selecting the number of bins in a histogram: A decision theoretic approach*", Journal of Statistical Planning and inference (61), pp.1-14, 1997.

[IZE91] IZENMAN A.J., "*Recent developments in nonparametric density estimation*", Journal of the American Statistical Association 86 (413), pp. 205-224.,1991.

[JOS83] JOSWICK A., "*A learning scheme for a fuzzy k-NN rule*", pattern Recognition Letters, vol.1, pp.287-289, 1983.

[KEL85] KELLER J.M.,. GRAY M.R, GIVENS J.A. JR, "*A fuzzy k-nearest neighbor algorithm*", IEEE Transactions on systems, man and cybernetics, Vol. 15, no. 4, pp.580-585, 1985.

[KIL90] KILIR G.J., "*A principle of uncertainty and information invariance*". Int. J. of General Systems, 17(2-3), pp. 249–275, 1990.

[KRI93] KRISHNAPURAM R. and KELLER J., "*A possibilistic approach to clustering*", IEEE Trans. Fuzzy Syst., vol. 1, pp. 98-110, May 1993.

[KLE98] KLEIN F., "*Etude comparative de méthodes de classification floues*", Thèse de Doctorat, Université de Reims, France, 1998.

[KNI72] KNIGHT W.A., "*Chatter in turning: some effect of tool geometry and cutting conditions*", Int. J. Mach. Tool Des. Res. Vol.12, pp. 201–220, 1972.

[KUL05] KULKARNI A., JAYARAMAN V.K. and KULKARNI B.D., "*Knowledge incorporated support vector machines to detect faults in Tennessee eastman process*", Computers and Chemical Engineering 29(10), pp. 2128–2133, 2005.

[LAS99] LASSERRE V., "Modélisation floues des incertitudes de mesures de capteurs", Thèse de Doctorat, Université de Savoie, France, 1999.

[LAS00] LASSERRE V., MAURIS G., and FOULLOY L., "*A simple possibilistic modelisation of measurement uncertainty*". In L.A. Zadeh Eds. B. Bouchon-Meunier, R.R. Yager, editor, Uncertainty in Intelligent and Information Systems, pp.58–69. World Scientific, 2000.

[LEC06] LECOEUCHE S., MERCERE G., AMADOU-BOUBACAR H., "*Modelling of non stationary systems based on a dynamical decision space*", 14th IFAC Symposium on System Identification, Newcastle, Australia, 2006.

[LES04] LESKI J.M., "*Fuzzy c-varieties/elliptotypes clustering in reproducing kernel Hilbert space*", Fuzzy Sets and Systems, Volume 141, Number 2, pp. 259-28, 2004.

[LUR03] LURETTE C., "*Développement d'une technique neuronale auto-adaptative pour la classification dynamique de données évolutives. Application a la supervision d'une presse hydraulique*", Thèse de Doctorat, Université des Sciences et Technologies de Lille, France, 2003.

[MAN00] MANDERS E.J., BISWAS G., NARASIMAH S., MOSTERMAN P.J., "*Combined Qualitative, Quantitative approach for fault isolation in continuous dynamic systems*", In 4th Symposium on Fault Detection Supervision and Safety of Technical Processes, Budapest, Hungary, 2000.

[MAS96] MASSON M. H., DUBUISSON B., FRELICOT C., "*Conception d'un module de reconnaissance de formes floues pour le diagnostic*", RAIRO-APII-JESA 30 (2,3), pp. 319-341, 1996.

[MAS06] MASSON M. et DENOEUX T., "*Inferring a possibility distribution from empirical data*", Fuzzy Sets and Systems, 157(3), 319-340, 2006.

[MON00] MONTMAIN J., GENTIL S., "*Dynamic causal model diagnostic reasoning for online technical process supervision*", Automatica, 36, pp.1137-1152, 2000.

[MOS01] MOSTERMAN J., "*Diagnosis of Physical Systems With Hybrid Models Using Parameterised Causality*", Proceeding of Hybrid Systems: Computation and Control, 4th International Workshop (HSCC01), Rome, Italia, pp.447-458, 2001.

[MUR91] MUROFUSHI T.and SUGENO M., "*Fuzzy t-conorms integrals with respect to fuzzy measures : generalization ef Sugeno integral and Choquet integral*". Fuzzy Sets and Systems, 42, pp.57-71, 1991.

[NEW98] NEWMAN D.J., HETTICH, S., BLAKE C.L. and MERZ, C.J. "*UCI Repository of machine learning databases*", Dept. of Information and Computer Science, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, 1998.

[OND06] ONDEL O., "*Diagnostic par reconnaissance des formes: Application à un ensemble Convertisseur-Machine asynchrone*", Thèse de Doctorat, Ecole centrale de Lyon, France, 2006.

[OTN72] OTNES R.K., ENOCHSON L., "*Digital time series analysis*", Wiley-Interscience Publication, New York, 1972.

[OUS00] OUSSALAH M., "*On the probability/possibility transformations: a comparative Analysis*", Int. J. General Systems 29 (5), 671-718, 2000.

[PAR62] PARZEN E., "*On the estimation of a probability density function and mode*". Annals of Mathematical. Statistics, Vol 33, pp. 1065-1076, 1962.

[PEL93] PELTIER M. A., "*Un système adaptatif sur la reconnaissance des formes floues. Application au diagnostic du comportement d'un conducteur automobile*", Thèse de Doctorat, Université de Technologie de Compiègne, France, 1993.

[PER84] PERROW C., "*Normal Accidents: Living with High Risk Technologies*", Basic Books, Inc., New York, 1984.

[PHI06] PHILIPPOT A., SAYED MOUCHAWEH M., CARRÉ-MÉNÉTRIER V., RIERA B., "*Decentralized Approach to Diagnose Manufacturing Systems*", In Computational Engineering in Systems Applications CESA'06, Beijing, China, 2006.

[QUI00] QINGHONG CHEN S., "*Comparing Probabilistic and Fuzzy Set Approaches*" for Design in the Presence of Uncertainty, A dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Aerospace Engineering, 2000.

[RAM89] RAMADGE P.J.G., WONHAM W., "*Supervisory control of a class of discrete events systems*", SIAM journal of Control and Optimization, Vol. 25, N°1, 1989.

[ROS56] ROSENBLATT M., "*Remarks on some non-parametric estimates of a density function*", Annals of Mathematical Statistics, Vol. 27, pp. 642-669, 1956.

[RUD82] RUDEMO M., "*Empirical choice of histograms and kernel density estimates*", Scandinavian Journal of Statistics 9, pp. 65-78, 1982.

[RUS69] RUSPINI E. H., "*A new approach to clustering*", Information and Control 15, pp.22-32, 1969.

[SAN91] SANDRI S., "*La combinaison de l'information incertaine et ses aspects algorithmiques*", Thèse de Doctorat, Université Paul Sabatier, Toulouse, 1991.

[SAY02a] SAYED MOUCHAWEH M., "*Conception d'un système de diagnostic adaptatif et prédictif basé sur la méthode Fuzzy pattern Matching pour la surveillance en ligne des systèmes évolutifs : Application à la supervision et au diagnostic d'une ligne de peinture au trempé*", Thèse de Doctorat, Université de Reims, France, 2002.

[SAY02b] SAYED MOUCHAWEH M. and BILLAUDEL P., "*Influence of the choice of histogram parameters at Fuzzy Pattern Matching performance*", WSEAS Transactions on Systems, Vol. 1, Issue 2, pp. 260-266, 2002.

[SAY02c] SAYED MOUCHAWEH M., BILLAUDEL P. et VILLERMAIN LECOLIER V. G., "*Diagnostic prédictif et adaptatif pour la supervision en temps réel d'un système évolutif en utilisant Fuzzy Pattern Matching*", Rencontres Francophones sur la Logique Floue et ses Applications, Montpelier, France, 2002.

[SAY06] SAYED MOUCHAWEH M., BOUGUELID M. S., BILLAUDEL P., RIERA B., "*Variable Probability-Possibility Transformation*", 25th European Annual Conference on Human Decision-Making and Manual Control (EAM'06), September 27-29,Valenciennes, France, 2006.

[SCO79] SCOTT D.W., "*On optimal and data-based histograms*", Biometrika (66), pp.605-610, 1979.

[SCO92] SCOTT D.W., "*Multivariate density estimation*", Wiley, New York, 1992.

[SME94] SMETS P. et KENNES R, "*The transferable belief model*", Artificial Intelligence, vol. 66, n°2, pp. 191-234, 1994.

[SHA76] SHAFER G., "*A mathematical theory of evidence*", Princeton, University press, 1976.

[SU03] SU R., ABDELWAHED S., KARSAI G., G. BISWAS, "*Discrete Abstraction and Supervisory Control of Switching Systems*", IEEE International Conference on Systems, Man and Cybernetics, Vol. 1, pp.415-421, 2003.

[TOM76] TOMEK, "*A generalisation of the kNN rule*", IEEE Trans on System, Man and Cybernetics, Vol. 6, no 2, pp 121-126, 1976.

[TRA97] TRAVE-MASSUYES L., DAGUE P., GUERRIN F., "*Le raisonnement qualitatif*", Hermès, France, 1997.

[UNG93] UNGAUER C., "*Problématique d'utilisation de techniques de supervision à base de connaissances profondes : l'exemple de la supervision du réseau TRANSPAC*", Technical report, France Telecom R&D, 1993.

[VAP95] VAPNIK V. "*The nature of statistical learning theory*", Spring-Verlag, New York, 1995.

[VER06] VERRON S., TIPLICA T. and KOBI A., "*bayesian networks and mutual information for fault diagnosis of industrial systems*", Workshop on Advanced Control and Diagnosis, ACD 2006. November 16-17, Nancy, France, 2006.

[YAG88] YAGER R. R., "*On ordered weighted averaging aggregation operators in multi-criteria decision making*". IEEE Trans Syst Man Cyber,Vol.18, pp.183–190, 1988.

[YAG05] YAGER, R. R., "*Extending multicriteria decision making by mixing t-norms and OWA operators*", International Journal of Intelligent Systems 20, pp.453-474, 2005.

[ZAD65] ZADEH L. A, "*Fuzzy Sets*", Information and Control, Vol. 8, pp. 338-353, 1965.

[ZAD78] ZADEH L. A., "*Fuzzy sets as a basis for a theory of possibility*", Fuzzy sets and systems 1, pp. 3-28,1978.

[ZIE95] ZIEBA S., "*Une méthode de suivi d'un phénomène évolutif. Application au diagnostic de la qualité d'usinage*", Thèse de Doctorat, Université de Technologie de Compiègne, France, 1995.

[ZIM91] ZIMMERMANN, H.J., "*Fuzzy Set Theory and its Applications*", Kluwer Academic, Boston ,1991.

[ZOU97] ZOUHAL L., "*Contribution à l'application de la théorie des fonctions de croyance en reconnaissance des formes*", Thèse de Doctorat, Université de Technologie de Compiègne, France, 1997.

[ZWI95] ZWINGELSTEIN G., "*Diagnostic des défaillances*", Traité des Nouvelles Technologies, série Diagnostic et Maintenance, Hermès, 1995.

Résumé : Les méthodes de Reconnaissance est l'ensemble des méthodes permettant de classifier des formes dans des classes. Nous avons choisi, parmi les méthodes de classifications existantes, la méthode possibiliste Fuzzy Pattern Matching (FPM) pour réaliser le diagnostic des systèmes dynamiques. FPM est simple et son temps de classification est constant et faible. De plus, elle est capable de sélectionner les sources d'informations les plus pertinentes et de traiter des données qui sont à la fois incertaines et imprécises. Cependant, FPM est une méthode de classification naïve, c'est-à-dire qu'elle classifie un nouveau point par la sélection d'une des décisions partielles. Chaque décision partielle est calculée pour chaque classe et par rapport à chaque attribut. FPM ne tient donc pas compte de la corrélation entre les attributs et considère la forme des classes comme convexe. Ces inconvénients rendent FPM inutilisable pour les applications réelles qui souvent nécessitent une discrimination non linéaire entre les classes. De plus, FPM n'est pas une méthode de classification adaptative ou prédictive. Elle ne peut pas extraire l'information manquante des points rejetés en quantifiant leur représentativité vis a vis des classes connues. Ces points, portent l'information sur l'apparition d'une nouvelle classe ou l'évolution entre deux classes. Les travaux de ce mémoire de thèse portent donc sur l'amélioration de la méthode FPM afin de remédier à ses limites. Les performances des solutions proposées sont illustrées à travers plusieurs exemples académiques et réels.

Mots–clés : Reconnaissance des formes, Théorie des possibilités, Classification, Corrélation des attributs, Diagnostic, Fuzzy pattern matching, Apprentissage incrémental, Données stationnaires.

Abstract : The problem of diagnosis by Pattern Recognition can be posed as a problem of classification, i.e., the actual functioning mode can be determined by knowing the class of the actual pattern. We use the method Fuzzy Pattern Matching (FPM) to realize the diagnosis because it is a simple method based on a feature selection. In addition it has a small and constant classification time, and it takes into account both the imprecision and uncertainty. However FPM is marginal, i.e., its global decision is based on the selection of one of the intermediate decisions. Each intermediate decision is based on one attribute. Thus, FPM does not take into account the correlation between attributes. Additionally, FPM considers the shape of classes as convex one. Also, FPM cannot realize the adaptive and predictive diagnosis because it rejects all the points which carry the information about the class evolution or the creation of a new class. These drawbacks make FPM unusable for many real world applications. In this thesis, we propose to improve FPM to solve these drawbacks. Several synthetic and real data sets are used to show the performances of the improved FPM with respect to classical one.

Key-words : Pattern recognition, Possibility theory, Classification, Attribute correlation, Diagnosis, Fuzzy pattern matching, Incremental learning, Stationary data.