Classification des données

2.1 Introduction

Ce chapitre, sera consacré à la présentation des deux méthodes de classification des données. La première méthode est basée sur l'utilisation des techniques de reconnaissance des formes, la seconde méthode utilise les séparateurs à vaste marge (SVM).

La méthode de reconnaissance des formes est une branche de l'intelligence artificielle, son objectif est le classement et l'identification des formes ou des objets à partir de données brutes afin de prendre une décision dépendant de la catégorie attribuée à cette forme **[POISSON 06].** Cette méthode regroupe les techniques informatiques de représentation et de décision qui donnent à la machine la capacité de simuler un comportement sensible.

Les méthodes SVM (Machines à vecteurs de support), est un outil très efficace pour la sélection des paramètres et aussi pour l'optimisation des paramètres d'un classifieur. Nous commençons par une présentation des principes de la méthode SVM où nous mettrons en évidence l'utilisation des SVM dans le cas des données linéairement séparables et le cas des données non linéairement séparables. Ensuite nous présentons des méthodes pour la classification SVM multi-classes.

Nous avons choisi d'utiliser la méthode de reconnaissance des formes floue parmi les autres méthodes, car elle est très performante devant les modèles imprécis (système complexe) et sa robustesse permet de limiter les effets des erreurs et imprécisions associées aux données et aux modèles. Elle nous permet d'identifier l'état actuel du système ainsi que l'évolution vers un mode indésirable. Tandis que La démarche de diagnostic par SVM a été proposée dans le seul but de valider les résultats de classification par reconnaissance des formes floue.

On termine ce chapitre par une conclusion.

Reconnaissance des formes

La reconnaissance des formes interprète toute nouvelle observation (Forme), les observations sont regroupées en classes, constituant des prototypes auxquels la nouvelle observation est comparée pour être identifiée.

Chaque observation est caractérisée par un ensemble de d paramètres, est représentable par un vecteur X_i dans l'espace R^d.

Les différentes étapes d'une procédure de la reconnaissance des formes sont illustrées par la figure 2.1.

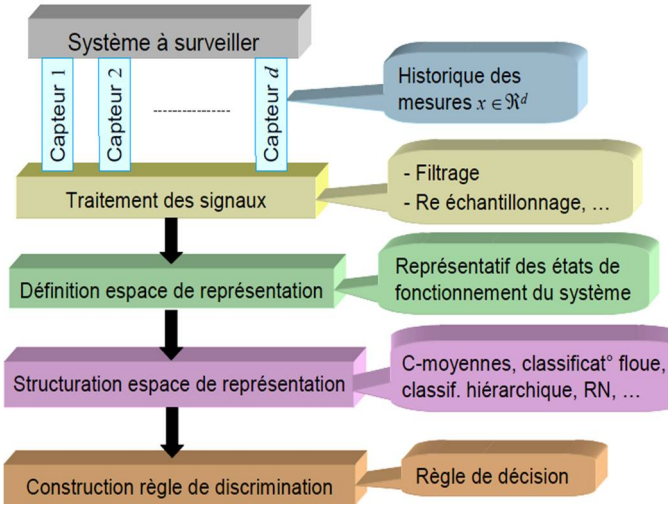

Figure 2.1 : Diagnostic d'un système : approche reconnaissance des formes.

Le diagnostic par reconnaissance des formes est posé comme suite :

Définition de l'espace de représentation, structuration de l'espace de représentation et la construction d'une règle de décision **[HANDIS 12].**

Définition de l'espace de représentation

Cette première phase de prétraitement consiste à construire le vecteur forme, c.à.d. sélectionner les *d* paramètres à extraire à partir des signaux recueillis. Le but est de réduire la quantité de données en retenant que les variables les plus importantes et filtrer les informations redondantes. Cela fait réduire le temps de classification pour les nouveaux observateurs.

L'ensemble de *d* paramètres obtenus après traitement des signaux est représenté par *un vecteur forme* X_i représenté dans l'espace de représentation R^d.

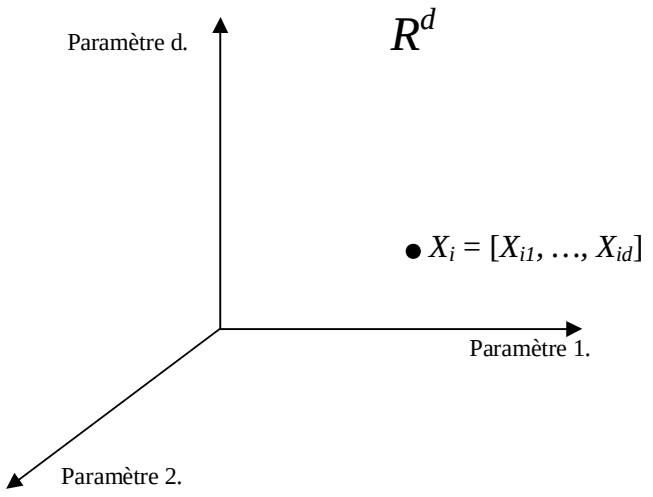

Figure 2.2 : Représentation d'un vecteur forme dans un espace de dimension *d*.

Il est préférable de disposer des paramètres variant, de manière significative, en fonction des différents modes de fonctionnement du système. Le choix des paramètres pertinents à partir des paramètres recueillis fait diminuer le temps de calcul. Des méthodes de réduction des paramètres doivent être utilisées pour extraire les paramètres les plus représentatifs **[GAOUAOUI 12].**

2.2.1.1 Réduction de l'espace de représentation

Les données collectées lors d'une observation d'un phénomène sur un système peuvent être peu significatives, corrélées ou redondantes, aberrantes ou simplement inutiles au problème de classification. L'objectif principal de la réduction de l'espace de représentation est de définir un nouvel espace de représentation de dimension réduite, afin de conserver les informations importantes du signal acquis. Les méthodes d'extraction des variables ainsi que les méthodes de sélection des variables peuvent être utilisées pour la réduction de l'espace de représentation.

2.2.1.1.1 Extraction des variables

L'extraction de variables est une étape importante qui consiste à construire un nouveau ensemble de variables ($d' < d$), qui contient la plus grande part d'information, à partir des variables initiaux. Plusieurs méthodes peuvent être appliquées pour l'extraction des variables, les plus connues sont l'ACP (Analyse en Composantes Principales), LDA (Analyse Discriminante linéaire), PLS (la régression aux moindres carrés), …

a) Analyse en composantes principales

L'analyse en composantes principales (ACP) est une technique statistique d'analyse de données multivariées, capable de comprimer les données pour réduire leur dimensionnalité c.à.d. trouver un sous - espace propre, qui maximise la variance des données après projection de sorte que l'information soit maintenue.

En considère $X_a=(X_{a1}, X_{a2}, \ldots, X_{aN})$ l'ensemble d'apprentissage composé de N vecteurs de dimension d (Matrice des données). L'analyse en composantes principales est une méthode permettant d'identifier un sous-espace $R^{d'}$ de dimension $d'<d$ représentant au mieux l'ensemble X_a **[HANDIS 12]**. Si les composantes de ces vecteurs ne sont pas homogènes les résultats de classification risquent d'être erronés. Pour éviter cette situation les données sont normalisées suivant les étapes qui suivent :

Le centre de gravité $\overline{X}$ de l'ensemble d'apprentissage est défini par :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (2.1)$$

Et les composantes de la variance sont :

$$\sigma_{j=1\ldots d}^{2} = \frac{1}{N}\sum_{i=1}^{N}\left(X_{ij} - \overline{X}_j\right)^2 \qquad (2.2)$$

À partir des deux relations précédentes on en déduit les vecteurs formes normalisés centré réduits (moyenne nulle et écart-type unité).

$$X_{i\,norm} = \Delta^{-1/2}\left(X_i - \overline{X}\right) \tag{2.3}$$

$$\Delta = \begin{bmatrix} 1/\sigma_1^2 & & & 0 \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_d^2 \end{bmatrix} \tag{2.4}$$

Après normalisation des données, les vecteurs propres de la matrice de variance – covariance aide à définir la base du sous-espace :

$$\sum = \frac{1}{N}\sum_{i=1}^{N}\left(X_i \cdot X_i^t\right) \tag{2.5}$$

$\sum$ Une matrice d'inertie du nuage de points (matrice carrée, réelle et symétrique). Ses éléments diagonaux sont les variances de l'ensemble d'apprentissage, et les éléments non diagonaux sont les covariances.
Les vecteurs propres de la matrice de variance-covariance $\sum$ sont les vecteurs v_j tels que $x \times v_j = \lambda_j \times v_j$. Les scalaires λ_j repris dans la matrice diagonale (2.7), sont les vecteurs propres.

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_{d'} & \cdots & v_d \end{bmatrix} \tag{2.6}$$

$$\lambda = \begin{bmatrix} \lambda_1 & & & & & 0 \\ & \lambda_2 & & & & \\ & & \ddots & & & \\ & & & \lambda_{d'} & & \\ & & & & \ddots & \\ 0 & & & & & \lambda_d \end{bmatrix} \tag{2.7}$$

L'ACP garantit que la déformation du nuage de points est minimale, et conduit à minimiser le critère J **[Dubuisson 90]:**

$$J = \frac{1}{2}\sum_{i=1}^{N} d^2\left(X_i, X_i^p\right) \tag{2.8}$$

Les résultats de l'ACP peuvent être visualisés en considérant la projection du nuage de points sur les deux ou trois premiers axes factoriels. Autrement dit, en visualisant la matrice des données dans l'espace défini par les deux ou trois premiers vecteurs propres.

Enfin la qualité de la représentation peut être quantifiée par le calcul du pourcentage d'inertie expliquée sur chaque axe factoriel ou sur le sous-espace de dimension d'obtenu. Ce pourcentage d'inertie est défini en fonction des valeurs propres **[HAFA12]**, **[ROLAND 03].**
Le pourcentage d'inertie expliquée par un axe i est donné par :

$$\Gamma_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{2.9}$$

Et la qualité de la représentation du nuage de points est représentée par :

$$\Gamma = \frac{\sum_{i}^{d'} \lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{2.10}$$

Plus le rapport Γ est proche de 1, la présentation est plus proche de la réalité.

2.2.1.1.2 Sélection des variables

La sélection des variables consiste à chercher un *sous-ensemble* de d' variables les plus pertinentes parmi les *d* variables originaux.

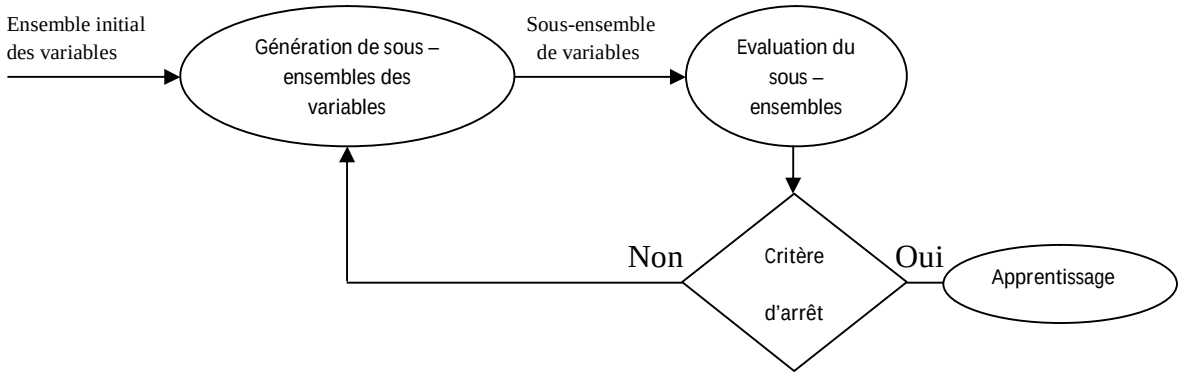

Figure 2.3 : Processus de sélection de variables.

La sélection des variables comporte trois éléments essentiels **[HAFA 12]**, **[PHILIPPE 00]** :

- Un critère d'évaluation des variables, pour qualifier la qualité d'un paramètre, et déterminé un sous-ensemble de variables les plus pertinentes.
- Une procédure de recherche pour explorer l'espace des différentes combinaisons de variables,
- Un critère d'arrêt, pour stopper la procédure de recherche.

a) Critère d'évaluation (mesure de pertinence des variables)

La mesure de pertinence associée aux méthodes de sélection de variables est basée sur des heuristiques calculant l'importance individuelle de chaque variable, ce qui correspond intuitivement à :

- La minimisation de la variance intra-classe (regroupement des points de chaque classe).
- La maximisation de la variance inter-classe (éloignement maximal entre les différentes classes).

Ces deux conditions sont traduites par Les matrices de variance intra–classe et inter–classes.
La matrice de variance intra-classe fait intervenir les échantillons de chaque classe ainsi que les centres gravités :

$$\sum_w = \frac{1}{N}\sum_{k=1}^{M}\sum_{j=1}^{N_k}\left(X_{kj} - \overline{X}_k\right)\left(X_{kj} - \overline{X}_k\right)^t \tag{2.11}$$

Alors que la matrice inter-classes caractérise la dispersion des classes entre elles.

$$\sum_B = \frac{1}{N}\sum_{k=1}^{M}\left(\overline{X}_k - \overline{X}\right)\left(\overline{X}_k - \overline{X}\right)^t \tag{2.12}$$

Le centre de gravité de la classe ω_k est donnée par :

$$\overline{X}_k = \frac{1}{N}\sum_{j=1}^{N_k} X_{kj} \tag{2.13}$$

La matrice variance-covariance qui traduit la dispersion totale de l'ensemble vaut :

$$\sum = \sum_w \sum_B \tag{2.14}$$

Le critère de Fisher permet de quantifier le pouvoir de discrimination pour chacun des paramètres contenus dans le vecteur forme initial de dimension *d*.

Dans le cas d'un problème à *M* classes, le critère s'exprime, pour un paramètre α de la manière suivante :

$$J(\alpha) = \sum_{k=1}^{M}\sum_{j=1}^{M-1}\frac{\left(\overline{X}_k(\alpha) - \overline{X}_j(\alpha)\right)}{\left(N_k\sigma_k^2(\alpha) - N_j\sigma_j^2(\alpha)\right)} \tag{2.15}$$

Avec :

$\overline{X}_k \rightarrow (k = 1, 2)$: Centre de gravité de la classe Ω_k en ne considérant que le paramètre α

$\sigma_k^2(\alpha) \rightarrow (k = 1, 2)$: Variance de la composante α des vecteurs de la classe Ω_k.

$$\overline{X_k}(\alpha) = \frac{1}{N_k}\sum_{i_k}^{N_k} X_{i_k}(\alpha) \tag{2.16}$$

$$\sigma_k^2(\alpha) = \frac{1}{N_k}\sum_{j=1}^{N_k}\left[X_{kj}(\alpha) - \overline{X}_k(\alpha)\right]^2 \tag{2.17}$$

b) Procédure de recherche

Le choix d'une procédure de recherche dépend de la taille de l'espace de représentation des observations. Il est possible de rechercher parmi l'ensemble initial E des données (de *d* paramètres) un sous-ensemble *E'* (de *d*' paramètres) de telle sorte que :

$$J(E') = \underset{i=1,..n_c}{Max}\left[J\left(E_{d'}^{(i)}\right)\right] \tag{2.18}$$

La première solution consiste à calculer le critère *J* sur toutes les combinaisons de *d*' paramètres. Le nombre total de combinaisons est égal à $C_d^{d'}$:

$$C_d^{d'} = d!\Big/(d-d')!\cdot d'! \tag{2.19}$$

La solution devient dissuasive est difficile à assurer quand la taille de l'espace de représentation des données est grande. Il existe aussi des méthodes de recherche efficaces comme le Branch & Bound.

En pratique, les méthodes de sélection sous optimales comme les algorithmes de recherche gloutonne (Forward Selection, Backward Elimination) ou les méthodes de recherche aléatoire sont les plus utilisées **[HAFA 12].**

c) Critère d'arrêt

Pour stopper la procédure de recherche et définir les sous-ensembles des variables les plus pertinents il faut un critère d'arrêt. Le choix de ce dernier reste très difficile à faire, une première solution consiste à arrêter la procédure de recherche lorsqu'un nombre maximum de variables à sélectionner sont atteintes, la fixation de nombre de variables à sélectionner est pratiquement difficile, ce qui rend l'utilisation de ce critère d'arrêt rare. L'idéal est de baser sur le critère d'évaluation en fixant un seuil sur J, dans ce cas la procédure de recherche est arrêtée lorsque ce dernier est atteint.

2.2.1.2 Structuration de l'espace de représentation

L'objectif de la structuration de l'espace de représentation est de regrouper les observations ayant les mêmes paramètres dans la même classe. Cette étape fait appel à des techniques de classification.

La classification se divise en deux familles : classification supervisée, et classification non supervisée. Dans le cas où la classe de chaque observation est connue, l'espace de représentation est parfaitement connu et l'apprentissage est effectué en mode supervisé. Si l'information sur la structuration de l'ensemble d'apprentissage en classe n'est pas disponible, la classification doit être faite en mode non supervisée. La figure suivante représente les différentes méthodes de classification non supervisée, autrement appelée méthode de coalescence.

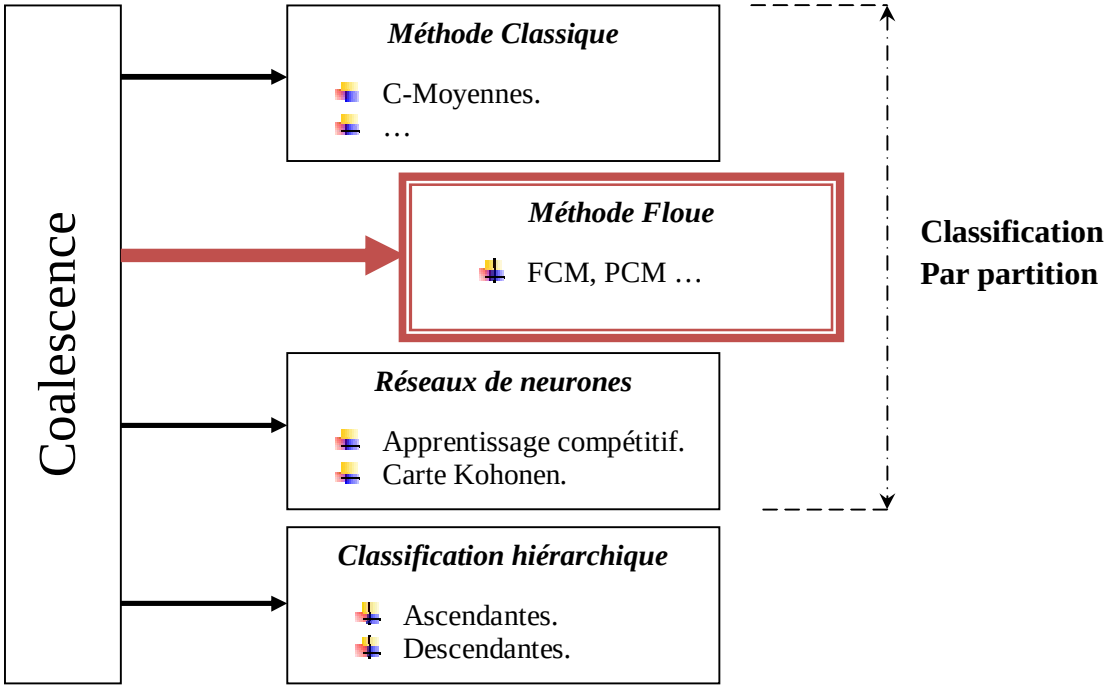

Figure 2.4 : Méthodes de classification non supervisée – Méthode de coalescence –

2.2.1.3 Détermination d'une méthode de décision

À l'issue de la phase d'analyse, l'ensemble d'apprentissage a été structurées en classes distinctes, les données (observations) sont parfaitement connues et étiquetées de telle sorte que chacune des observations appartient à une des *M* classes connues.
Soit $X_a = (X_{a1}, X_{a2}, \cdots, X_{aN})$ l'ensemble d'apprentissage et $\Omega = (\Omega_1, \Omega_2, \cdots, \Omega_N)$ l'ensemble des M classes (modes de fonctionnement) connues. La question qui se pose :

À quelle classe affecter une nouvelle observation X recueillie à un instant donné sur le système ?

Pour répondre à cette question, on doit construire une règle de décision en utilisant une approche statistique ou une approche analytique pour la classification des nouvelles observations aux différentes classes de l'ensemble d'apprentissage. L'algorithme de décision doit donc bénéficier des solutions fournies par les options de rejets d'ambiguïté et de distance afin de diminuer le risque de mauvaise classification **[ROLAND 03].**

Le rejet d'ambiguïté est appliqué aux observations situant aux frontières de décision des classes connues comme montre la figure 2.5.

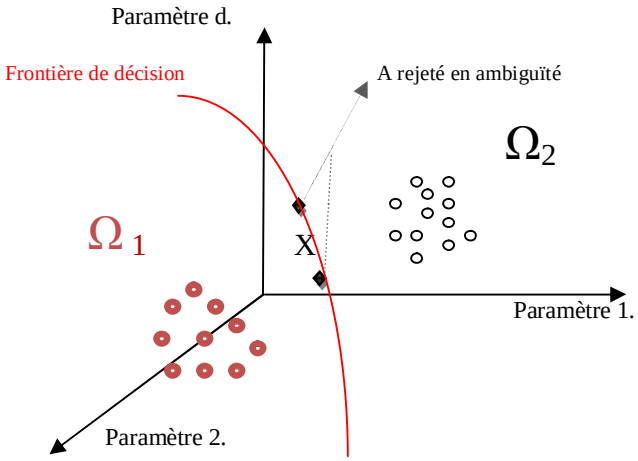

Figure 2.5 : Rejet d'ambiguïté. Observation située près de la frontière de décision entre deux classes.

Les observations se trouvant dans cette situation sont affectés à une nouvelle classe fictive Ω_0 appelée classe de rejet d'ambiguïté.

Si un/ou plusieurs observations sont trop loin des classes connues, en applique un rejet de distance comme montre la figure 2.6.

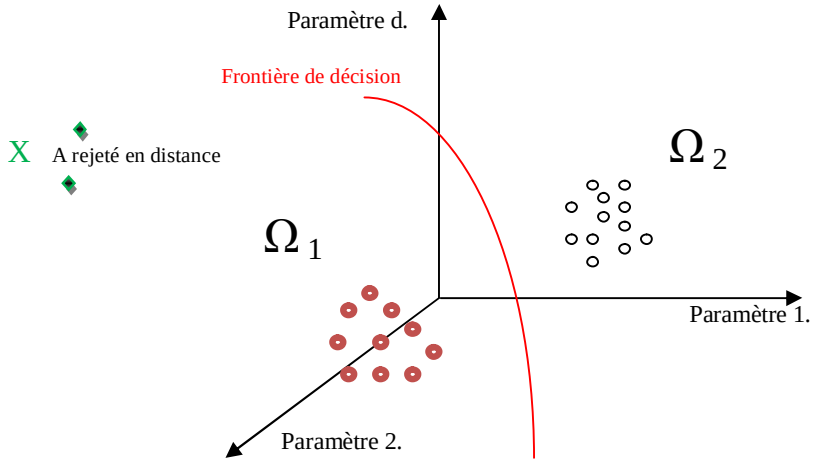

Figure 2.6 : Rejet de distance. Observation ne correspond à aucune des classes connues.

Les observations qui ne correspondent à aucune des classes connues seront affectées à une nouvelle classe fictive Ω_d appelée classe de rejet en distance. Le système se trouve dans un mode de fonctionnement inconnu.

Deux types d'approches peuvent être utilisé pour la mise au point d'une règle de décision. Le premier type consiste à utiliser des méthodes statistiques, paramétriques ou non paramétriques. La deuxième approche, qui est analytique, consiste à privilégier le calcul des frontières de décision entre classes **[ROLAND 03]**, **[HAMOU MAMAR 08].**

Dans une approche statistique, on associe à chaque vecteur forme *X* une densité de probabilité conditionnelle $f(X/\Omega_j)$ par rapport à chaque classe de Ω. En fonction de la connaissance, complète ou incomplète sur la distribution de $f(X/\Omega_j)$ on trouve différentes approches.

Si la connaissance est complète, modèle et paramètres de $f(X/\Omega_j)$ étant supposé connus, la règle de décision optimale est basée sur la théorie Bayésienne de la décision.

Dans le cas d'une connaissance incomplète, le modèle $f(X/\Omega_j)$ inconnu, différentes stratégies peuvent être utilisées :

- Si la forme de $f(X/\Omega_j)$ est supposée connue, on estime les paramètres de $f(x/\omega_j)$,
- Si la forme de $f(X/\Omega_j)$ est inconnue, on utilise des méthodes non paramétriques pour l'estimation de $f(X/\Omega_j)$ tels que les noyaux de Perzen. Sinon on estime les probabilités a posteriori en utilisant la méthode des k plus proches voisins, les réseaux neuromimetriques, … etc.

L'estimation de la densité conditionnelle $f(X/\Omega_j)$ se fait toujours sur un ensemble d'apprentissage.

Dans l'approche analytique, la fonction discriminante est déterminée en estimant les paramètres d'une fonction mathématique de telle sorte à minimiser la probabilité de mauvaise classification pour mieux séparer les classes. La complexité de la frontière de décision impose le choix du modèle de la fonction mathématique. En pratique, on commence par des fonctions linéaires puis si la qualité de classification n'est pas satisfaisante, on passe à des modèles plus complexes, quadratique, réseaux neuromimétiques …etc.

De plus, il y a d'autres méthodes tel que les réseaux de neurones basés sur la notion de distance, les réseaux perceptron ou bien les réseaux Learning Vector Quantization, tous ses méthodes utilisant des indicateurs pour la division directe de l'espace de représentation en des zones de décision **[HAMOU MAMAR 08].**

2.2.1.4 Evaluation du système de diagnostic

Le système de diagnostic par reconnaissance de formes permet de déterminer à quel mode de fonctionnement correspond chaque nouvelle observation recueillie sur le système. L'algorithme de décision ne pourra que confirmer un mode existant ou rejeter une observation différente des autres. En cas de rejet, seul un travail d'expertise peut fournir une analyse des causes de ce changement d'état **[ROLAND 03]**.

Dès que le nombre des observations rejetées devient significatif, la méthode de diagnostic doit permettre de rechercher une nouvelle structure en classe, un nouvel apprentissage est à relancer afin de prendre en compte cette nouvelle classe.

2.3 Reconnaissance des formes floue

Dans la théorie ensembliste classique, l'appartenance d'un élément *x* à une des classes connues est exprimée par une fonction caractéristique (fonction indicatrice) $u_{E_k}(x)$.

$$x \xrightarrow{u_{E_k}(x) \to \{0,1\}} u_{E_k}(x) = \begin{cases} 1 \quad Si & x \in E_k \\ 0 & Sinon \end{cases} \qquad (2.20)$$

Le concept d'appartenance en logique ensembliste classique est binaire (*x* appartient ou n'appartient pas à un ensemble), qui ne permet pas de représenter les données subjectives à cause des incertitudes liées aux informations et aux connaissances.
La notion floue associée à la reconnaissance des formes permet de situer une forme par rapport à toutes les classes connues. Un degré d'appartenance compris dans l'intervalle [0 1] est fourni à la forme étudiée pour chaque classe. Si ce degré est proche de 1, cela signifie une forte ressemblance entre la forme et les individus de la classe. À l'inverse, si ce degré est proche de 0, la forme est différente des individus de la classe **[JEAN-HUGH 99].**
Donc, un ensemble flou F étend la notion d'ensemble classique en généralisant la fonction caractéristique $u_F : E \to [0,1]$. Cela permet de rendre compte de l'appartenance partielle d'un sous-ensemble flou. Les ensembles flous autorisant des états intermédiaires entre l'appartenance et non appartenance, qui permettent une représentation simple des imprécisions et incertitudes.

La figure 2.7 représente une comparaison entre le concept d'appartenance classique et le concept d'appartenance floue.

(a) Ensemble Classique (b) Ensemble floue

Figure 2.7 : Représentation d'une fonction d'appartenance au sens classique et au sens flou.

2.3.1 Caractéristiques des fonctions d'appartenances

Selon **[ALOUANE 08]**, les fonctions d'appartenances possèdent les caractéristiques suivantes :

Le type : les fonctions d'appartenances peuvent avoir différentes formes (triangulaires, trapézoïdales, monotones, gaussiennes).

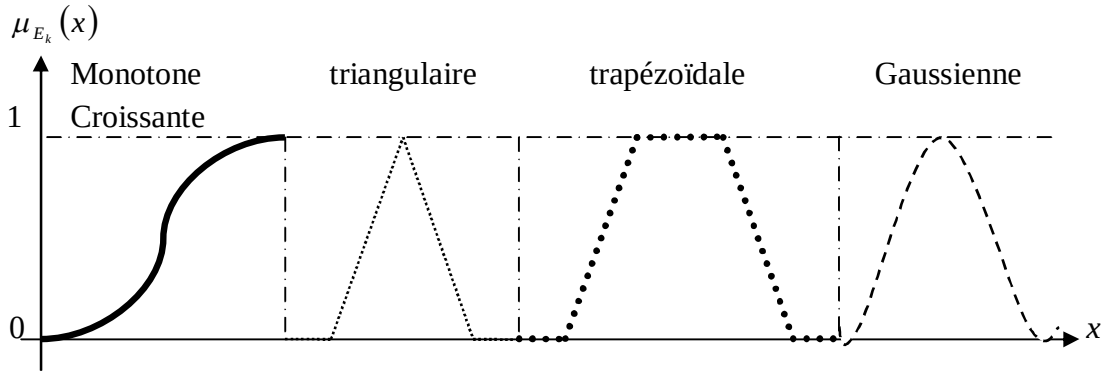

Figure 2.8 : Exemple de fonction d'appartenance

Le noyau $(Noy(F))$**:** C'est l'ensemble des éléments de E dont le degré d'appartenance à F vaut 1 :

$$Noy(F) = \{x \in E \ / \ \mu_F(x) = 1\} \tag{2.21}$$

Le support $(Supp(F))$ **:** C'est l'ensemble des éléments de *E* dont le degré d'appartenance à *F* est non nul.

$$Supp(F) = \{x \in E \ / \ \mu_F(x) \rangle 0\} \tag{2.22}$$

La hauteur $(h(F))$**:** C'est la valeur maximale de la fonction d'appartenance.

$$h(F) = \sup_{x \in E} \ \mu_F(x) \tag{2.23}$$

Si $h(F) = 1 \ \rightarrow$ *F* est normalisé.

Les coupes de niveau α (α-coupe) : C'est l'ensemble des éléments de *E* qui appartiennent à *F* avec un degré au moins égal à α.

$$F_\alpha = \{x \in E \ / \ \mu_F \geq \alpha\} \tag{2.24}$$

L'α-coupe permet à partir de sous-ensembles flous de revenir à la notion d'ensemble classique.

La cardinalité $|F|$ **:** la cardinalité d'un ensemble floue *F* est la quantité d'éléments de *E* qui appartiennent à *F*.

$$|F| = \sum_{x \in E} \mu_F(x) \tag{2.25}$$

Les principaux éléments caractéristiques d'un sous ensemble flou sont illustré dans la figure2.9

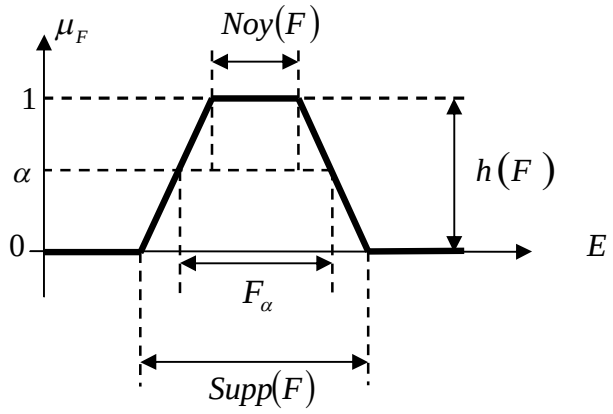

Figure 2.9 : Principaux éléments caractéristiques d'un sous ensemble flou F [ALOUANE08].

2.3.2 Opérations sur les sous-ensembles flous

Le concept de sous-ensemble flou étant une généralisation du concept d'ensemble classique. On présente ici, les opérations couramment utilisées pour la manipulation de deux sous-ensembles floue *F* et *G* **[AMBAPOUR 09]**:

Egalité : Deux sous-ensembles flous *F* et *G* sont égaux si leurs fonctions d'appartenance prennent la même valeur pour tous les éléments *x* de *E*. Formellement *F*=*G* Si seulement si :

$$\forall\ x \in E,\ \mu_F(x) = \mu_G(x) \tag{2.26}$$

Complément : le complément d'un sous-ensemble flou *F* de *E* noté $\overline{F}$ est défini par :

$$\forall\ x \in E,\ \mu_{\overline{F}}(x) = 1 - \mu_F(x) \tag{2.27}$$

Inclusion : Soit *F* et *G* deux ensembles flous de *E*. Si pour n'importe quel élément *x* de *E*, *x* appartient toujours moins à *F* qu'à *G*, alors on dit que *F* est inclus dans *G*. Formellement :

$$F \subseteq G \Leftrightarrow \forall\ x \in E,\ \mu_F(x) \leq \mu_G(x) \tag{2.28}$$

Union : c'est le sous-ensemble flou constitué des éléments de *E* affectés du plus grand des degrés avec lesquels ils appartiennent à *F* et *G*. Formellement :

$$F \cup G \Leftrightarrow \forall x \in E,\ \mu_{F\cup G} = \max\left(\mu_F(x),\ \mu_G(x)\right) \tag{2.29}$$

Intersection : l'intersection de deux sous-ensembles flous *F*et *G* de *E* est le sous ensemble flou constitué des éléments de *E* affectés du plus petit des degrés avec lesquels ils appartiennent à *F* et G, $F \cap G$ est donné par :

$$\forall\ x \in E,\ \mu_{F\cap G}(x) = \min\left(\mu_F(x),\ \mu_G(x)\right) \tag{2.30}$$

2.3.3 Classification floue supervisée

La classification par méthodes supervisées, nécessite la connaissance a priori de nombre de classes. Avant d'effectuer une identification des données, il faut passer par les deux phases suivantes **[CLAUDIA 07] :**

d) Une phase d'apprentissage pour apprendre les caractéristiques de chaque classe.
e) Une phase de reconnaissance pour décider l'appartenance des nouvelles données à telle ou telle classe, les données sont classées en fonction du classifieur obtenu lors de la phase d'apprentissage.

Plusieurs algorithmes peuvent être utilisés en mode supervisé tel que les k plus proches voisins, algorithme de Bayes…

2.3.3.1 K plus proches voisins (*K-ppv*)

L'algorithme de *K* plus proches voisins *K-ppv* est une méthode non paramétrique et supervisé qui ne nécessite pas une étape d'apprentissage, elle est généralement utilisée en classification et en segmentation d'image en particulier. Le principe de base est d'affecter un individu à la classe qui est majoritaire parmi ses *K* plus proches voisins. La version floue de cet algorithme apporte deux modifications :

- Introduction de la notion de distance dans le "vote" : un individu (en fait sa classe) aura d'autant plus de poids qu'il sera proche de l'individu à classer.
- Comme pour l'algorithme des C-moyennes, introduction d'une matrice d'appartenance de chaque individu à chaque classe.

L'algorithme se déroule comme suit :

Étape 1 : on fixe une matrice d'appartenance $\mu_k(X)$ initiale.

Étape 2 : pour chaque échantillon X à classifier, on calcule son degré d'appartenance $\mu_k(X)$ dans la classe Ω_k par une somme pondérée des degrés d'appartenance à la classe *k* de ses *Kppv.*

Plusieurs règles sont proposées dans la littérature pour introduire la notion floue dans la méthode des k plus proche voisin, comme par exemple l'algorithme proposé par Béreau, ou l'algorithme présenté par Keller qui exprime le coefficient d'appartenance d'un nouveau vecteur forme *X* à la classe Ω_k par **[MOCHAWECH 03]** :

$$\mu_k(X) = \frac{\sum_{i \in \phi_k(X)} \mu_{i,k} \cdot \left(\frac{1}{\|X - X_i\|^{(2/m-1)}} \right)}{\sum_{i \in \phi_k(X)} \cdot \left(\frac{1}{\|X - X_i\|^{(2/m-1)}} \right)} \qquad (2.31)$$

La règle floue des K-ppv la plus utilisée est la règle du maximum d'appartenance qui consiste à affecter *X* à la classe dont la fonction d'appartenance est maximale :

$$X \in \Omega_k \; si \; \mu_k(X) = \max_{i=1,\cdots M} \mu_i(X) \tag{2.32}$$

L'algorithme flou des k plus proches voisins est simple à implémenter et donne de bons résultats quand le nombre des données n'est pas important.

2.3.3.2 Algorithme de Bayes

L'algorithme de Bayes repose sur la détermination de la valeur de la probabilité a posteriori qu'une observation *X* provienne de la classe Ω_k :

$$P_r(\Omega_k / X_i) = f(X/\Omega_k) \cdot P_r(\Omega_k) \Big/ \sum_{i=1}^{M} f(X/\Omega_i) \cdot P_r(\Omega_i) \tag{2.33}$$

La règle de Bayes détermine l'affectation du vecteur *X* à l'une des classes par :

$$X \in \Omega_k \; si \; P_r(\Omega_k / X) = \max_{i=1,\cdots M} [P_r(\Omega_i / X)] \tag{2.34}$$

Alors l'algorithme cherche à déterminer pour chaque élément X_i la classe Ω_i qui maximise la probabilité de contenir cet élément.

2.3.4 Classification floue non supervisée

2.3.4.1 Algorithme FCM (Fuzzy C-means)

C moyennes floues (FCM) est un algorithme de classification non supervisée floue. Il effectue une optimisation itérative en évaluant de façon approximative les minimums d'une fonction d'erreur.

Les FCM sont un cas particulier d'algorithmes basés sur la minimisation d'un critère ou d'une fonction objectif **[ALOUANE 08]**, [**KHODJA 97]**, **[YANG 02]**, **[TREERATTANAPITAK13]**.

Soit l'ensemble d'apprentissage $X_a = \{X_{a1}, X_{a2}, ..., X_{aN}\}$ défini dans $\Re^d$ que l'on souhaite partitionner en *M* classe. Soit $V = \{V_1, V_2, ..., V_M\}$ qui représente l'ensemble des centres de prototype.

L'objectif de l'algorithme est de minimiser la fonction objective J_m qui représente la distance moyenne intra-groupe floue au sens des moindres carrés, J_m est définie par :

$$J_m(U,V)=\sum_{k=1}^{M}\sum_{i=1}^{N}(u_{ik})^{m}\left\|X_i-V_k\right\|_A^2 \qquad (2.35)$$

Pour tout $k(1\le k\le M)$, V_k est un vecteur à d composantes qui représente le prototype de la $k^{ème}$ (Centre classe de la classe k), $m\rangle 1$ est l'indice flou.

Où :

$\left\|X_i-V_k\right\|_A^2$ Est la distance entre l'échantillon X_i *et le centroïde* du cluster V_k définie par *:*

$$\left\|X_i-V_k\right\|_A^2=(X_i-V_k)^t A(X_i-V_k) \qquad (2.36)$$

Si A est la matrice identité, le cluster V_k sera sphérique, sinon il sera ellipsoïdal **[GLORENNEC 99]**. Dans la version habituelle des FCM, A est la matrice identité quelque soit V_k.

La mise à jour itérative des prototypes et des degrés d'appartenance est donnée par les deux équations suivantes :

La détermination des centroïdes V_k **[BERZEK 81]** : on calcule les centroïdes des classes.

$$V_k=\frac{\sum_{i=1}^{N}u_{i,k}^{m}X_i}{\sum_{i=i}^{N}u_{i,k}^{m}} \qquad (2.37)$$

Détermination des fonctions d'appartenance :

$$u_{i,k}=\left[\sum_{i=1}^{M}\left(\frac{\left\|X_i-V_k\right\|_A}{\left\|X_i-V_l\right\|}\right)^{\frac{2}{(m-1)}}\right]^{-1} \qquad (2.38)$$

L'algorithme FCM est décrit de manière détaillée dans le tableau (2.1) :

Algorithme : C-moyennes floues [FCM]
o **Données** N vecteurs non étiquetés.
o **Paramétrage :** Nombre de classe M. nombre maximal d'itérations T Fixer le degré de flou m. Norme pour calcul des distances $\|.\|_A$. Seuil d'arrêt ε.

- **Initialisations :**
  Initialiser le compteur de boucle $t \leftarrow 0$.
  Tirage au hasard des prototypes $V_0 = \left(V_1^{(0)},...,V_M^{(0)} \right)$.

- **Itérations :**
  Répéter
  $t \leftarrow t+1$
  1. **Calcul de la matrice de partition U_t par :**

$$u_{i,k} = \begin{cases} 1 & \left\|X_i - V_k\right\|_A \le \left\|X_i - V_l\right\|_A \quad \forall l \ne k \\ 0 & \textit{Sinon} \end{cases} \quad \forall\, i,k$$

Ou par : $u_{i,k} = \left[\sum_{i=1}^{M} \left(\frac{\left\|X_i - V_k\right\|_A}{\left\|X_i - V_l\right\|} \right)^{\frac{2}{(m-1)}} \right]^{-1}$

  2. **Calcul des prototypes V_t par**

$$V_k = \frac{\sum_{i=1}^{N} u_{i,k}^m X_i}{\sum_{i=i}^{N} u_{i,k}^m}$$

  Tant que $\left(t \le T\right)$ et $\left\|V_t - V_{t-1}\right\| \ge \varepsilon$

Tableau 2.1 : Algorithme FCM.

2.3.4.1.1 Avantages et inconvénients de l'algorithme FCM

L'avantage principal de l'algorithme FCM et le fait d'obtenir des regroupements dans lesquels l'appartenance d'une observation à un regroupement devient une notion graduelle. Les degrés d'appartenance μ_{jk} rendent le processus itératif beaucoup plus robuste qui permet d'obtenir des partitions plus pertinentes, et plus proches de la réalité.
Cependant, l'algorithme FCM n'est pas robuste face aux bruits introduits par l'imprécision des attributs, ce problème de la sensibilité à l'initialisation des centres des classes peut facilement diminuer l'efficacité de l'algorithme et fournir une solution localement optimale.
De plus, il ne permet que la détection des classes de forme sphérique, puisqu'il est basé sur la distance euclidienne pour mesurer la similarité entre une observation et le centre d'une classe.
Pour résoudre ces inconvénients, plusieurs modifications ont été apportées à l'algorithme de base et qui sont présentées comme des dérivés de l'algorithme FCM.

2.3.4.2 Les algorithmes dérivés de l'algorithme FCM

Pour pouvoir éviter les problèmes liés à l'algorithme FCM standard, de nombreuses variantes se fondent essentiellement sur une modification de la métrique utilisée ont été proposées.

2.3.4.2.1 Algorithme de Gustafson et Kessel GK

Cette méthode est basée sur l'utilisation de la distance Mahalanobis à la place de la distance euclidienne, cela permet de représenter chaque classe k par son centre de gravité et sa matrice de covariance. On parle alors de distance adaptative qui permet de détecter des classes de forme ellipsoïdale.

Soit l'ensemble $A = \{A_k\}_{k=1}^{M}$, la fonction objectif donnée par Gustafson et Kassel est:

$$J_m^{GK} = (U,V) = \sum_{k=1}^{M}\sum_{i=1}^{N} \mu_{i,k}^m \left\| X_i - X_k \right\|_{A_K}^2 \tag{2.39}$$

La distance entre l'échantillon X_i et le centroïde du cluster V_k prend la forme suivante :

$$\left\| X_i - V_k \right\|_{A_k}^2 = \left(X_i - V_k \right)^t A_k \left(X_i - V_k \right) \tag{2.40}$$

Où :

$$A_k = \left[\rho_k \det\left(\tilde{\sum_k} \right) \right]^{1/n} \tilde{\sum_k}^{-1} \quad \forall\ k \tag{2.41}$$

$\tilde{\sum_k}$ Représente la matrice covariance floue associée à la classe k, celle-ci est calculée à chaque itération par l'équation (2.42):

$$\tilde{\sum_k} = \frac{1}{\sum_{i=1}^{N} \mu_{ik}^m} \sum_{i=1}^{N} \left(X_i - V_k \right)\left(X_i - V_k \right)^t \tag{2.42}$$

L'algorithme Gustafson et Kessel (GK) est décrit de manière détaillée dans le tableau 2.2 :

Algorithme : Gustafson et Kessel (GK)
o **Données** N vecteurs non étiquetés.
o **Paramétrage :** Nombre de classe M. nombre maximal d'itérations T Fixer le degré de flou m. Norme pour calcul des distances $\|.\|_A$. Seuil d'arrêt ε.

- **Initialisations :**
  Initialiser le compteur de boucle $t \leftarrow 0$.
  Tirage au hasard des prototypes $V_0 = \left(V_1^{(0)},...,V_M^{(0)} \right)$.
  Initialiser l'ensemble des matrices de covariance $A^{(0)}$ respectant $\det(A_k) = \rho_k \rangle 0$

- **Itérations :**
  Répéter $t \leftarrow t+1$
  1. **Calcul de la matrice de partition U_t par :**

$$u_{i,k} = \left[\sum_{l=1}^{M} \left(\frac{\left\| X_i - V_k \right\|_A}{\left\| X_i - V_l \right\|} \right)^{\frac{2}{(m-1)}} \right]^{-1}$$

  2. **Calcul des prototypes V_t par**

$$V_k = \frac{\sum_{i=1}^{N} u_{i,k}^m X_i}{\sum_{i=i}^{N} u_{i,k}^m}$$

$$A_k = \left[\rho_k \det\left(\tilde{\sum_k} \right) \right]^{1/n} \tilde{\sum_k}^{-1} \quad \forall\ k$$

Où : $\tilde{\sum_k} = \frac{1}{\sum_{i=1}^{N} \mu_{ik}^m} \sum_{i=1}^{N} (X_i - V_k)(X_i - V_k)^t$

Tant que $(t \le T)$ et $\left\| V_t - V_{t-1} \right\| \ge \varepsilon$

Tableau 2.2 : Algorithme Gustafson et Kessel

2.3.4.2.2 Algorithme C- moyennes possibiliste PCM

L'algorithme c-moyennes possibiliste a été proposé par Krishnapuram et Keller **[KRISHNAPURAM 05]**, dans ce dernier, le degré d'appartenance d'une observation à une classe est considéré comme une possibilité d'appartenir à une classe **[SBILI 15]**.
La fonction objectif modifiée est :

$$J_m^{PCM}(U,V) = \sum_{k=1}^{M} \sum_{i=1}^{N} \mu_{i,k}^m \left\| X_i - V_K \right\|_A^2 + \sum_{k=1}^{M} w_k \sum_{i=1}^{N} \left(1 - \mu_{i,k}\right)^m \qquad (2.43)$$

Dans cette expression, le premier terme correspond au critère, de FCM, le second terme impose des valeurs les plus grandes possibles aux possibilités d'appartenance. La valeur de ω_k fixe la distance à partir de laquelle la possibilité d'appartenance est égale à 0.5 **[ASSAS13]**. L'algorithme est initialisé avec la matrice de partition obtenue par la méthode FCM.

$$w_k = \frac{\sum_{i=1}^{N} \mu_{ik}^m \left\| X_i - V_k \right\|_A^2}{\sum_{i}^{N} \mu_{ik}^m} \tag{2.44}$$

Les centres de classes sont mis à jour comme dans l'algorithme FCM. Les possibilités d'appartenance sont calculées de manière itérative par les expressions :

$$V_k^{(t)} = \frac{\sum_{i=1}^{N} \mu_{i,k}^m X_i}{\sum_{i=1}^{N} \mu_{i,k}^m} \tag{2.45}$$

$$\mu_{i,k} = \left[1 + \left(\frac{\left\| X_i - V_k \right\|_A^2}{w_k} \right)^{\frac{1}{(m-1)}} \right]^{-1} \qquad \forall i,k \tag{2.46}$$

Le calcule se termine dès que la différence entre chaque possibilité d'appartenance et la possibilité de l'itération précédente est inférieure à un seuil fixé. La confusion des centres est l'inconvénient de cet algorithme.

2.3.4.2.3 Algorithme de C-moyennes possibilistes modifiée MPCM

L'algorithme de c-moyennes possibiliste modifié proposé par KHODJA Lotfi [**KHODJA 97]**, a pour but d'éliminer le problème de confusion des centres, les MPCM limitent la contribution des points les plus éloignés d'un centre donné dans le calcul des coordonnées du centre via la modification de la formule des degrés d'appartenance :

$$\mu_{i,k} = \left[1 + \left(\frac{\left\| X_i - V_k \right\|_A^2}{w_k} \right)^{\frac{m}{(m-1)}} \right]^{-1} \qquad \forall i,k \tag{2.47}$$

La fonction objectif proposée par KHODJA est :

$$J_m^{MPCM}(U,V) = \sum_{k=1}^{M} \sum_{i=1}^{N} \mu_{i,k}^m \left\| X_i - X_K \right\|_A^{2m} + \sum_{k=1}^{M} w_k \sum_{i=1}^{N} \left(1 - \mu_{i,k}\right)^m \tag{2.48}$$

Cet algorithme présente l'avantage de sauvegarder les apports espérés de l'approche possibiliste, tout en éliminant l'inconvénient de confusion des centres.
L'algorithme PCM et MPCM sont décrit de manière détaillée dans le tableau (2.3) :