UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTÉ DES SCIENCES ET TECHNIQUES FÈS
DÉPARTEMENT D'INFORMATIQUE

FST FES

# PROJET DE FIN D'ETUDES

MASTER SCIENCES ET TECHNIQUES
SYSTÈMES INTELLIGENTS & RÉSEAUX

---

# APPROCHE DE LA CLASSIFICATION SUPERVISÉE À BASE DU GRANULAR COMPUTING FLOU

---

LSIA

LIEU DE STAGE : LABORATOIRE DES SYSTÈMES INTELLIGENTS ET APPLICATIONS

RÉALISÉ PAR : YAHYA KHARBANE

SOUTENU LE 23/06/2015

ENCADRÉ PAR :

MR AZEDDINE ZAHI

DEVANT LE JURY COMPOSÉ DE :

MR ZAHI AZEDDINE
MR ABBAD KHALID
MR BEN ABBOU RACHID
MLLE CHAKER ILHAM

ANNÉE UNIVERSITAIRE 2014-2015

# Remerciements

Avant tout, Je tiens à remercier Allah pour cette grâce d'être en vie et en bonne santé, et pour avoir terminé ce mémoire dans des meilleures conditions, et ce malgré toutes les contraintes que j'avais opposées.

Je remercie mes parents Mohamed et Fatima, mes frères Ayoub et Najib, et tous les autres membres de la famille. Leur soutien dans le meilleur comme dans le pire, Leur sacrifice pour mon propre confort et leur bonne humeur permanente, des facteurs qui m'ont permis d'atteindre cette étape de ma vie. Qu'Allah les récompense pour toutes ces années qu'ils m'ont consacrées.

Ensuite, je formule de sincères remerciements à Monsieur Azeddine Zahi, Enseignant chercheur à la Faculté des Sciences et Techniques de Fès (FSTF) et membre du Laboratoire Systèmes intelligents et Applications (LSIA), pour m'avoir confié ce travail de recherche, ainsi que pour son aide, sa disponibilité, ses conseils pertinents au cours de ce travail.

Une mention spéciale est décernée à Soufiane Ezghari, Tawfiq Khaldi, Imad Elghoubach, Imad Batioua, Ali Ouajjani, Tarik Mouqtassid, Safa Jida, Hajar Mrabti, Yassmine Aghil et Karima Laadnani …, avec leur sincère amitié, leur encouragement et leur soutien inconditionnel, ce travail a pu aboutir.

# Résumé

Dans ce travail, on s'intéresse à la résolution des problématiques traitées par la fouille de données plus précisément celui de la classification automatique. Actuellement les méthodes de classification supervisée doivent être capables de faire face aux différents problèmes à savoir l'adaptation avec le domaine traité, l'interprétabilité de la base des connaissances et le résultat obtenu, l'imprécision dans les données et l'incertitude pendant la résolution du problème. A cet égard, dans ce travail on présente une amélioration de la classification supervisée par l'introduction de la logique floue comme un outil de représentation des données et de calcul, ainsi que le *Granular Computing* qui fournit un concept de représentation des données et du raisonnement structuré.

La *logique floue* a été émergée comme une approche puissante pour modéliser la connaissance qualitative des problèmes de fouille de données. Elle a été utilisée de façon satisfaisante dans diverses applications en particulier lorsque le bruit, l'incertitude et l'imprécision sont inévitables.

***Le** Granular Computing* reflète la façon dont l'humain observe, apercevoit et résout les problèmes complexes. Un modèle basé sur le *Granular Computing* permet la représentation du domaine dans une *structure granulaire* hiérarchique qui capture les différents aspects du domaine. Chaque niveau dans la structure granulaire se forme par des *granules d'informations et* définit un degré d'abstraction du domaine, ces granules représente un ensemble d'objets qui sont regroupés selon un critère comme la similarité, la dépendance, etc. Enfin le calcul avec les granules permet de résoudre un problème dans un niveau d'abstraction adéquat.

L'objectif de travail est de simuler un système de classification qui est composé de deux parties :

1. Un modèle de représentation des connaissances en se basant sur l'approche *granular computing* et la *logique floue***.**
2. Un modèle de classification basé sur la représentation des connaissances obtenues.

Ce système va être validé dans plusieurs domaines comme diagnostic médical, reconnaissance de forme, etc.

# Abstract

In this work, we are interested in solving the problems processed by data mining specifically that of supervised classification. Currently, supervised classification methods must be able to deal with different problems namely the adaptation with the field processed, the interpretability of the knowledge base and the result obtained, the inaccuracy in the data and uncertainty during the resolution. In this regard, this work present a supervised classification improved by the introduction of fuzzy logic as a tool data representation and calculation, as well as the Granular Computing concept that provides a data representation and structured reasoning.

Fuzzy logic has been emerged as a powerful approach to model the qualitative knowledge of data mining problems. It has been satisfactorily used in various applications especially when noise, uncertainty and imprecision are unavoidable.

The Granular Computing reflects how the human observes perceives and solves complex problems. A model based on Granular Computing allows the field's representation in a hierarchical granular structure that captures the various aspects of the field. Each level in the granular structure is formed by information granules of and sets a degree of abstraction of the area, such granules represent a set of objects that are grouped according to a criterion such as the similarity, addiction, etc. Finally the calculation with the granules solves a problem in an adequate level of abstraction.

The work objective is to simulate a classification system which is composed of two parts:

- A knowledge representation model based on granular computing and fuzzy logic approach.
- A classification model based on the representation of the obtained knowledge.

This system will be validated in several fields such as medical diagnosis, pattern recognition, etc.

# Sommaire :

# Liste des figures :

# Liste des tableaux

# Liste des abréviations

AD: arbre de décision

BIC: critère bayésien d'information

CAD : clustering hiérarchique descendant

CAH: clustering hiérarchique ascendant

CART: classification and regression trees

CFCM: Constrained fuzzy c-means

ChAID: Chi-squared Automatic Detection

DC: double clustering

ECD : extraction des connaissances à partir des données

FCM : fuzzy c-means

FD : fouille de données

GrC: Granular Computing

ML-DC: multi-level double clustering

N1: structure granulaire avec un seul niveau

N2: structure granulaire avec deux niveaux

RN: réseau de neurone

SIF: système d'inférence floue

# Introduction

Actuellement, la Fouille de Données est devenue une discipline principale dans plusieurs domaines. En effet, grâce aux avancements technologiques, de nombreux outils sont inventés afin d'acquérir les données dans différents domaines, ainsi ils offrent des opportunités d'automatiser aux d'améliorer plusieurs tâches. A cet égard, la Fouille de Données fournit des multitudes de solutions permettant la représentation des données, la visualisation et la résolution des problèmes.

La Fouille de Données peut réaliser plusieurs tâches comme la recherche des associations entre les données, la détection des anomalies, le regroupement des données ou la prédiction. Dans ce travail on s'intéresse à la tache de prédiction. D'ailleurs, la classification supervisée est une méthode de prédiction qui permet de prédire la nature d'un objet donné. Actuellement la classification supervisée joue un rôle très important dans plusieurs domaines comme le diagnostic médical, la reconnaissance des formes, l'évaluation, etc. Cette variation de domaine d'application révèle d'une part d'autres besoins tels la précision lors du développement d'un modèle de classification supervisée, l'interprétabilité , la rapidité et l'adaptabilité , d'autre part faire face aux problèmes de l'imprécision dans les données et l'incertitude pendant le raisonnement.

Dans ce cadre plusieurs travaux dans la littérature [1]–[3] sont réalisés dans l'objectif de repondre aux nouveaux besoins de la classification supervisée ; les plus connus sont ceux qui se basent sur la *logique floue* ou le *granular computing* [4], [5]. La logique floue, par sa proximité de l'esprit humain, a suscité l'intérêt des chercheurs, des ingénieurs et des industriels. Cet intérêt réside dans la capacité de cette logique à manipuler et à représenter les connaissances, imprécises et incertaines. En effet, les connaissances sont représentées par des variables, appelées variables linguistiques, qui prennent des valeurs dans un ensemble de termes linguistiques tels que, petit, grand, très grand, etc. Chaque terme linguistique est manipulé par une fonction à valeurs dans [0,1], appelée fonction d'appartenance.

Le granular computing est une approche permettant la représentation structurée des données de domaine et la résolution structurée des problèmes. Le granular computing permet une interprétation élevée de domaine traité puisqu'il est similaire à la perception et le stockage d'informations chez l'humain, il stocke les données avec un niveau d'abstraction où

l’ensemble des niveaux représente une hiérarchie. Ainsi la résolution de problème est effectuée dans le niveau d’abstraction adéquat.

Dans ce mémoire, nous nous intéressons à la classification supervisée en se basant sur la logique floue et le granular computing.

Le présent rapport est organisé comme suit :

- Chapitre 1 : ce chapitre situe le contexte général du travail effectué.
- Chapitre 2 : ce chapitre présente un état de l’art su le concept du granular computing, puis il se focalise sur celui basé sur la théorie de logique floue.
- Chapitre 3 : dans cette partie, on a conçu un modèle du granular computing flou qu’on va appliquer à la classification supervisée.
- Chapitre 4 : ce dernier chapitre présente les résultats des expérimentations du modèle conçu appliqué au problème de la classification supervisée des cas de certains domaines réels.

# Chapitre 1 : Contexte du travail

# 1 Introduction

La *fouille de données*[1] est une discipline qui a émergé à l'issue des avancées technologiques réalisées dans les domaines de la collecte, de stockage et de traitement des données (scanners, internet, base de données, entrepôts de données, XML etc.). En effet, ces avancées ont largement contribué à l'accumulation de grands volumes de données que seul un traitement automatique est capable de les gérer, de les analyser et de les explorer.

La fouille de données est considéré comme un procédé d'exploration, et d'analyse en vue, d'une part, elle vise à rendre compréhensible les données accumulées, et d'autre part elle essaie de découvrir les corrélations significatives existantes. Les corrélations trouvées peuvent s'exprimer sous forme de descriptions de connaissances ou des règles de raisonnement capables de traiter de nouvelles situations.

Parmi les objectifs lors du développement d'une solution de *fouille de données*, c'est d'aboutir à un modèle :

- Permettant de capturer les différents aspects du problème.
- Tolérant aux imprécisions dans le sens où il sera capable de traiter des valeurs vagues et floues tels que petit, moyen, chère, etc.
- Capable de gérer les incertitudes où plusieurs solutions sont générées mais avec des degrés de certitudes différents.

Dans ce contexte, on a fait recourt à deux émergentes approches de fouille de données à savoir la *logique floue* et le *Granular Computing* pour répondre à ces objectifs.

Dans ce chapitre, nous allons présenter, une vue générale sur la fouille de données, ainsi que les outils associées et plus particulièrement, la *logique floue* et le *Granular Computing*.

# 2 Processus ECD :

La mise en place d'une solution de fouille de données est inclue dans un processus nommée ECD (Extraction des connaissances à partir des données). Ce processus [6] structure les prétraitements essentiels pour exécuter un système FD convenablement, ainsi l'évaluation et l'interprétation des connaissances découverte, la figure (1) montre les différentes étapes du processus ECD.

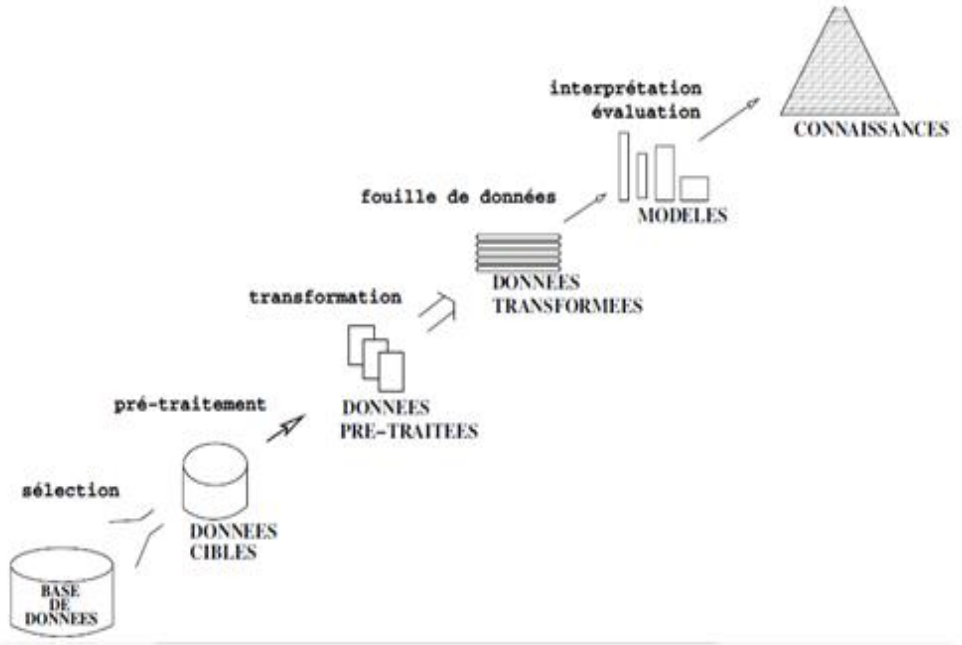



Figure 1 : Processus ECD [6]

On peut distinguer trois grandes parties dans le processus ECD, La première partie permet de récolter et construire les données convenables dans un domaine donné par rapport au contexte de la problématique traitée par la FD. Cette partie est appelé *prétraitement*. Elle est composée de trois éléments :

- *La sélection* : consiste à choisir les données et les propriétés (attributs) pertinentes pour l'objectif étudié.
- *Le traitement* : consiste à appliquer une série de modifications aux données sélectionnées afin de les rendre homogènes. Plusieurs actions sont possibles:
  - Le nettoyage où il s'agit de supprimer les données incomplètes, inconsistantes et incorrectes.
  - La normalisation, la discrétisation, la binarisation, etc.
- *La transformation :* où il s'agit de consolider les données hétérogènes provenant de diverses sources de stockage dans un même fichier. Les données sont généralement présentées sous forme d'un tableau où une colonne représente une *propriété* et une ligne représente un *exemple*.

La deuxième partie est la *fouille de données* qui consiste à extraire les structures cognitives existantes dans les données sélectionnées. Ces structures représentent des tendances, des groupes pertinents, des anomalies, des règles de raisonnement, etc.

La troisième partie est le *post-traitement* qui consiste à :

- Visualiser les connaissances extraites sous forme de schémas, de tableaux, etc. afin d'aider l'utilisateur à mieux comprendre les résultats obtenus.
- Evaluer et interpréter les résultats selon des critères de qualité tels que la précision et l'interprétabilité.

Le processus ECD est itératif ce qui signifie que parfois il peut être nécessaire de refaire certaines étapes. Il est aussi hautement interactif, l'utilisateur y est impliqué à chaque étape pour effectuer des choix.

## 3 Méthodes de Fouille de Données

La fouille de données fournit plusieurs méthodes afin de répondre à plusieurs problématiques, ces derniers peuvent être classées en quatre catégories :

- L'*association* : caractérise les problèmes qui cherchent à déterminer les tendances du concept étudié. Par exemple, le comportement d'un client dans un supermarché lorsqu'il fait ses courses. D'une manière formelle, un problème d'association consiste à découvrir, les relations les plus fréquentes entre les attributs.
- La *prédiction* : caractérise les problèmes qui cherchent à estimer la valeur d'un attribut (variable cible) à l'aide des valeurs des autres attributs (variables prédictives). Par exemple estimer la *dotation* du carburant nécessaire, pour un voyage, en se basant sur le type de véhicule et la distance à parcourir. Un problème de prédiction consiste à trouver la relation qui exprime la variable cible (sortie) en fonction des variables prédictives (entrées). Selon le type de la variable cible, on distingue : la *classification* où il s'agit de prédire des valeurs discrètes et la *régression* où il s'agit de prédire des valeurs continues.
- Le *regroupement* : caractérise les problèmes qui cherchent à trouver les classes pertinentes du concept étudié. Par exemple, un client *fidèle, infidèle*, etc. D'une manière formelle, un problème de regroupement consiste à trouver une *partition,* pertinente des données, constituée par un ensemble de *clusters* où les données d'un même cluster sont très similaires et celles appartenant à des clusters différents présentent une forte dissimilarité.

- La *Détection des anomalies* : caractérise les problèmes qui cherchent à détecter les comportements anormaux d'un concept. Par exemple, la détection des fraudes, la détection des pannes, etc. Cette problématique est souvent utilisée dans les systèmes de surveillance. D'une manière formelle, il s'agit de trouver des valeurs inhabituelles, aberrantes, irrégulières dans des données supposées homogènes.

Pour cela plusieurs méthodes de fouille de données sont développées pour répondre à chacune des problématiques déjà notées. Elles sont classées en trois grandes rubriques :

- Les méthodes de visualisation
- Les méthodes de prédiction
- Les méthodes d'explication

Chacune de ces rubriques contient plusieurs techniques appropriées aux différents types de vecteurs de données. Certaines de ces techniques sont mieux adaptées à des données numériques continues tandis que d'autres sont dédiées aux données qualitatives.

## 3.1 Les méthodes de visualisation:

Ces méthodes consistent à fournir une perception visuelle de ce qui n'est pas normalement visible, et à effectuer une représentation compréhensible à l'aide des mots, des dessins ou des tracés géométriques (histogramme, nuages de points, …).

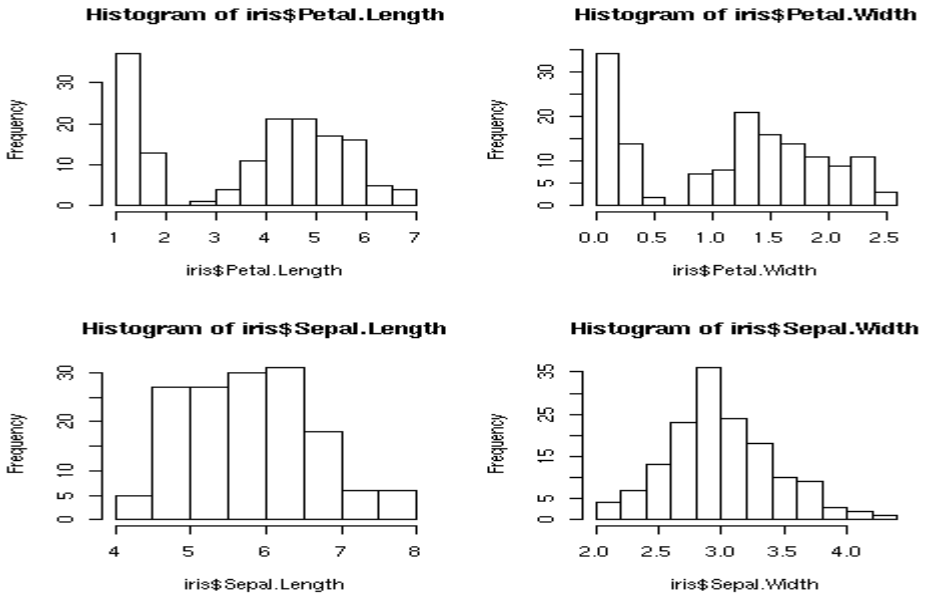



Figure 2: Exemple de représentation graphique par des histogrammes de la base de données Iris[7]

## 3.2 Les méthodes de prédiction :

Ces méthodes ont pour objectif de relier un phénomène à expliquer à un ou plusieurs phénomènes explicatifs. Elles sont issues de la statistique, de la reconnaissance de formes, de l'apprentissage automatique, du connexionnisme ou des bases de données. Ces méthodes sont mises en œuvre pour extraire des modèles de classement ou de prédiction.

Il existe plusieurs méthodes d'explication ou de prédiction développées dans différents contextes parmi lesquelles :

- Les arbres de décision :

L'arbre de décision permet de diviser les données en groupes basés sur les valeurs des attributs. C'est un outil puissant et apprécié pour la prédiction. L'arbre donne des modèles facilement compréhensibles par l'utilisateur. Ainsi les modèles de classifieur comme les arbres de décision s'expriment comme un ensemble de règles de classification de la forme : Si description Alors classe.

Il existe une variété d'algorithmes pour construire les arbres de décision. Par exemple, on trouve CART (classification and régression Trees) [8] qui amène à un arbre binaire, comme on trouve ChAID (Chi-squared Automatic Detection)[9] qui produit un nombre variable d'arcs pour chaque nœud.

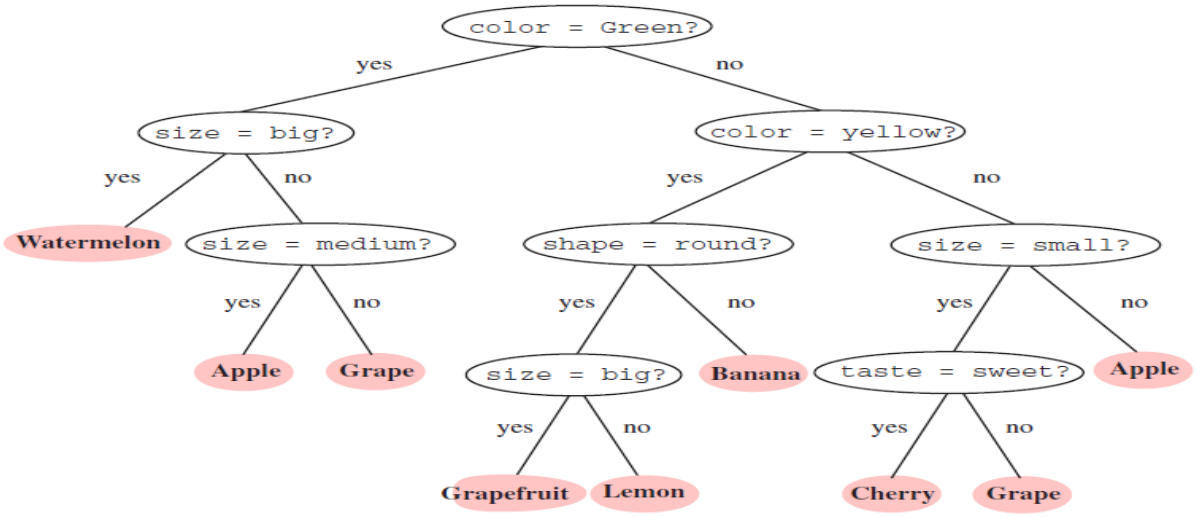



Figure 3: Exemple d'arbre binaire de décision [10]

- Les réseaux de neurones :

Les réseaux de neurones [11] sont des réseaux complexes d'unités de calcul élémentaires interconnectées. Ils sont issus de modèles biologiques, sont constitués d'unités élémentaires (neurones) organisés selon une architecture. Ils se composent de trois parties essentielles : Neurones de la couche d'entrée, ceux de la couche cachée et ceux de la couche de sortie.

Dans les réseaux de neurones la principale difficulté est de faire le bon choix de l'architecture : nombre de couches cachées et nombre de neurones par couche cachée. Les couches d'entrée et de sortie sont déterminées par la nature du problème : le nombre de neurones de la couche d'entrée est le plus souvent égal au nombre des attributs tandis que le nombre de neurones sur la couche de sortie est égal au nombre de classes du problème étudié.

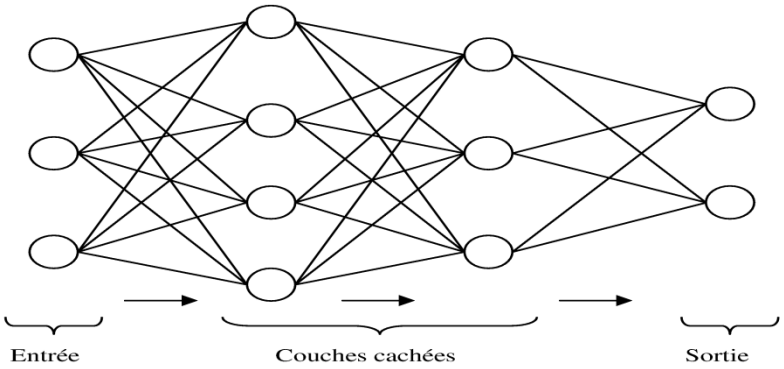



**Figure 4 : structure générale des réseaux de neurones** [12]

- La régression :

En régression [13], il s'agit d'expliciter une relation de type linéaire ou non, entre un ensemble d'attributs et un ensemble de cibles. Dans le cadre de régression, toutes les variables sont considérées continues. Elle possède de nombreux résultats statistiques intéressants permettant d'apprécier la qualité du modèle qu'elle produit.

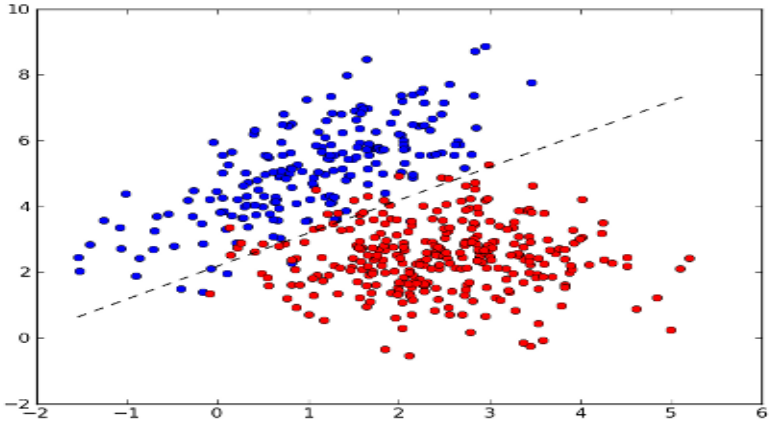



Figure 5 : classification avec régression linéaire[14]

### 3.3 Les méthodes d'explication :

- Les règles d'association :

Les règles d'association [15] permettent de découvrir à partir d'un ensemble de données, un ensemble de règles qui exprime une possibilité d'association entre les différentes observations. Un exemple classique de l'utilité de cette approche est le panier de la ménagère qui décrit un ensemble d'achats effectué au supermarché ; les règles d'association permettent de découvrir de régularités dans l'ensemble de transactions comme par exemple : Si thé alors sucre, etc. Ces règles permettent par exemple au gérant de proposer des bons de réductions significatifs sur les achats futurs des clients.

- Clustering :

Le clustering [9], [16], [17] est une méthode d'analyse de données qui opte à rassembler les objets homogènes dans des groupes appelés clusters en faisant appel à des mesures de similarité. Un bon clustering donc convient à la fois à minimiser l'inertie intra-classe et à maximiser l'inertie inter-classe. Généralement, le clustering est utilisé dans la fouille de données dans le but d'extraire les connaissances pour pouvoir générer des hypothèses ou des modèles prédictifs qui serviront à expliquer des concepts éventuellement impossibles à distinguer naturellement.

## 4 Fouille données : challenges et évolution

Malgré le succès de la fouille de données et sa contribution dans l'évolution de plusieurs domaines, au point de vue des chercheurs, elle souffre toujours de nombreux

désavantages qui sont toujours le sujet de nombreuses recherches. On peut identifier les plus importants challenges de la fouille de données dans trois titres:

- Le coût de la généralisation : la généralisation permet de définir une abstraction des données de domaine traité ce qui engendre une perte d'information, le coût de perte d'information est très important parce que ces informations peuvent être très importantes dans la résolution d'un autre problème.
- L'adaptabilité : l'évolution du domaine traité par la fouille de données peut générer des problèmes un peu différents de ceux résolus lors de la construction du modèle de la fouille de donnée. Dans ce cas, le modèle développé doit doter d'une stratégie d'adaptation pour aborder le problème d'une façon qu'il change ou évolue.
- Tolérance aux imprécisions dans les données : les données récoltées se constituent des mesures d'un ensemble d'attributs, caractéristiques, critères, alternatives ou indices. Ces mesures sont souvent imprécises, par conséquent elles peuvent influencer sur la qualité du modèle.

A cet égard, plusieurs techniques sont introduites pour améliorer l'efficacité des techniques de la fouille de données à savoir les heuristiques, les techniques de boosting, le raisonnement adaptatif, etc [18]. Aujourd'hui la logique floue marque une très grande contribution dans l'amélioration des techniques de la fouille de données grâce à sa capacité dans la tolérance aux imprécisions et la gestion des incertitudes, ainsi elle adopte une méthode de raisonnement très proche de l'être humain basée sur des valeurs linguistiques, l'étendue de de cette technique est le granular computing qui représente un concept complet de représentation et du raisonnement en se basant sur la logique floue. Dans ce travail on s'est intéressé sur le granular computing basé sur la logique floue est son implémentation pour la résolution de la problématique de la classification supervisée.

## 4.1 La logique floue

La logique floue a été fondée par L.Zadeh [19]. Depuis son apparition, elle est devenue un outil prometteur pour l'amélioration des différentes techniques de la fouille de données [1], [5]. L'idée principale consiste à traiter graduellement est partiellement la notion de vérité, en d'autres termes, une variable peut prendre plusieurs valeurs de vérité. Ainsi, la

logique floue est un outil fondamental et indispensable dans la représentation et le traitement des connaissances imprécises et incertaines.

### 4.1.1 Ensemble flou

La logique floue repose sur la théorie des ensembles flous, qui est une extension de la théorie des ensembles classiques. Au contraire des ensembles classiques où la notion d'appartenance est binaire définie par une fonction caractéristique :

$$\mu_A(x) = \begin{cases} 0 \; si \; x \notin A \\ 1 \; si \; x \in A \end{cases} \quad (1)$$

Qui signifie qu'un élément $x$ est soit dans $A$ soit à l'extérieur de $A$.

Dans l'extension de la théorie floue un objet $x$ appartient à un ensemble flou avec un *degré d'appartenance* mesuré par une fonction nommée *fonction d'appartenance.*

Soit $E$ un ensemble. Un sous-ensemble flou $A$ de $E$ est caractérisé par une fonction d'appartenance :

$$\begin{array}{rl} \mu_A: & E \to [0,1] \\ : & x \longmapsto \mu_A(x) \end{array} \quad (2)$$

qui associe à chaque élément $x$ de $E$ une valeur,$\mu_A(x)$, dans$[0,1]$ qui représente le degré d'appartenance de $x$ à $A$.

Un ensemble flou $A$ est caractérisé par plusieurs propriétés :

Soit E un ensemble, A un sous-ensemble flou de E et $\mu_A(x)$ la fonction d'appartenance le caractérisant.

- La hauteur de $A$ correspond à la borne supérieure de l'ensemble d'arrivée de sa fonction d'appartenance : $h(A) = \sup\{\mu_A(x) | \; x \in E\}$
- $A$ est dit normalisé si et seulement si $h(A) = 1$
- Le support de $A$ est l'ensemble des éléments de E appartenant au moins un peu à $A$ : $supp(A) = \{x \in E \mid \mu_A(x) > 0\}$
- Le noyau de A est l'ensemble des éléments de A appartenant totalement à A : $noy(A) = \{x \in E \mid \mu_A(x) = 1\}$
- L' α-coupe de $A$ est l'ensemble des éléments ayant un degré d'appartenance au moins égal à α : $\alpha - coupe(A) = \{x \in E \mid \mu_A(x) \geq \alpha\}$

- Un ensemble flou $A$ est dit convexe si :

$$\forall\, x_1, x_2, x_3 \in X, x_1 \leq x_2 \leq x_3\ alors\ \mu_A(x_2) \geq \min\left(\mu_A(x_1), \mu_A(x_3)\right)$$

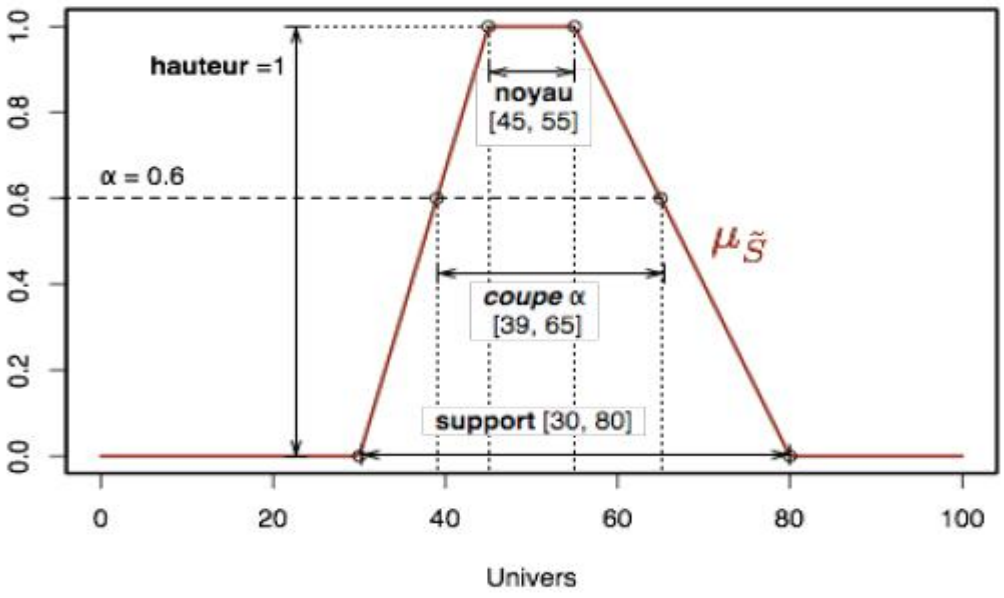


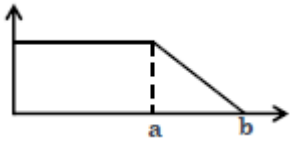

Figure 6 : Propriétés d'un ensemble flou

### 4.1.2 Fonctions d'appartenance :

Les fonctions d'appartenance peuvent théoriquement prendre n'importe quelle forme. Il existe trois formes qui sont souvent utilisées : les triangles [20], les trapézoïdes et les gaussiennes [21], [22] car elles sont simples et simplifient le recueil d'expertise.

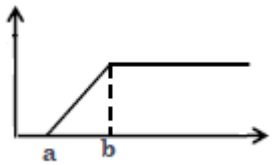
- Forme demi-trapézoïde gauche :

$$f(x \mid a, b) = \max\left(\min\left(\frac{x-a}{b-a}, 1\right), 0\right)$$



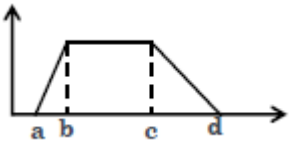

- Forme demi-trapézoïde droit :

$$f(x \mid a, b) = \max\left(\min\left(\frac{b-x}{b-a}, 1\right), 0\right)$$



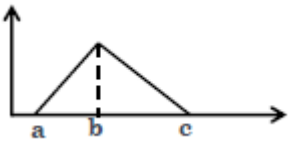

- Forme trapézoïde symétrique ou asymétrique:

$$f(x \mid a,b,c,d\ ) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\ \right)$$



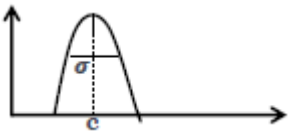

- Forme triangulaire symétrique ou asymétrique :

$$f(x \mid a,b,c\ ) = \max\left(\min\left(\frac{x-a}{b-a},,\frac{c-x}{c-b}\right), 0\ \right)$$




- Forme gaussienne :

$$f(x \mid \sigma,c\ ) = e^{\frac{-(x-c)^2}{2\sigma^2}}$$




### 4.1.3 Variable linguistique

Une Variable linguistique est définie par le triplet $(V; X_V\ ;\ T_V)$ avec :

- V est le nom de la variable linguistique, par exemple Age, taille, masse, etc.;
- $X_V$ est l'ensemble des valeurs (termes linguistiques) pouvant être, prises par V;

- $T_V$est une partition floue, où chaque sous-ensemble est associé à une valeur dans$X_V$ .

Par exemple la variable *taille* définie sur $[0\,;250]$ peut-être représentée par la variable linguistique :

- $V$ : taille,
- $X_V = \{\text{très petit, petit, moyen, grand et très grand}\}$ ,
- La partition floue représenté dans la figure 7.

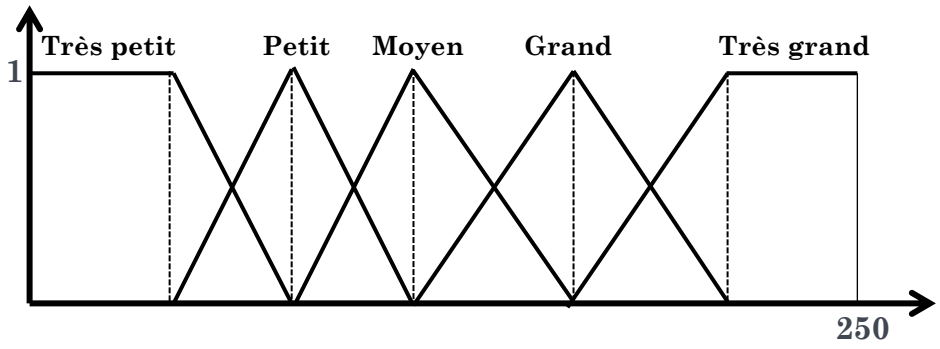



Figure 7: Représentation graphique de la variable linguistique Taille

### 4.1.4 Opérateurs

Afin de pouvoir manipuler aisément les ensembles flous, nous redéfinissons les opérateurs de la théorie des ensembles classiques afin de les adapter aux fonctions d'appartenance propres à la logique floue permettant des valeurs strictement entre 0 et 1.

Contrairement aux définitions des propriétés des ensembles flous qui sont toujours les mêmes, la définition des opérateurs sur les ensembles flous est choisie de la même manière que les fonctions d'appartenance. Voici les deux ensembles d'opérateurs les plus connus pour le complément (NON), l'intersection (ET) et l'union (OU) utilisés le plus couramment :

Tableau 1 : quelques opérateurs flous

| Dénomination | Intersection | Union | Complément |
|---|---|---|---|
| Opérateurs de Zadeh | $\min\big(\mu_A(x), \mu_B(x)\big)$ | $\max\big(\mu_A(x), \mu_A(x)\big)$ | $1 - \mu_A(x)$ |
| Probabiliste | $\mu_A(x) \times \mu_B(x)$ | $\mu_A(x) + \mu_B(x) - \mu_A(x) \times \mu_B(x)$ | $1 - \mu_A(x)$ |

Avec les définitions usuelles des opérateurs flous, nous retrouvons toujours les propriétés de commutativité, distributivité et associativité des opérateurs classiques. Cependant, relevons deux exceptions notables :

- En logique floue, le principe du tiers exclu est contredit : $A \cup \bar{A} \neq E$.
- En logique floue, un élément peut appartenir à A et non A en même temps : $A \cap \bar{A} \neq E$

### 4.1.5 Modélisation floue d'un modèle de fouille de données

Généralement un modèle de fouille de données est défini par trois partie la base de connaissance ou les entrées de modèle, le mécanisme de raisonnement et la sortie. L'introduction de la logique floue va permettre au modèle de fouille de données la tolérance aux imprécisions qui réside dans la base de connaissance et la gestion d'incertitude durant le raisonnement. En fin, un modèle flou est défini par trois parties :

- Les variables d'entrées : sont une collection des variables linguistiques définies par des termes linguistique.
- L'inférence floue : est un raisonnement basé sur la représentation de la base de connaissance par des règles floues SI….ALORS en utilisant les termes linguistiques associés à leur fonction d'appartenance et les opérateurs logiques.
- La sortie : le modèle fournit en premier temps une sortie vague sous forme d'un ensemble floue qui contient l'ensemble des solutions possibles, ensuite il peut calculer une solution précise et optimale à partir de l'ensemble des solutions proposé.

La figure 8 montre un exemple d'un modèle floue qui détermine si un client actif ou non.

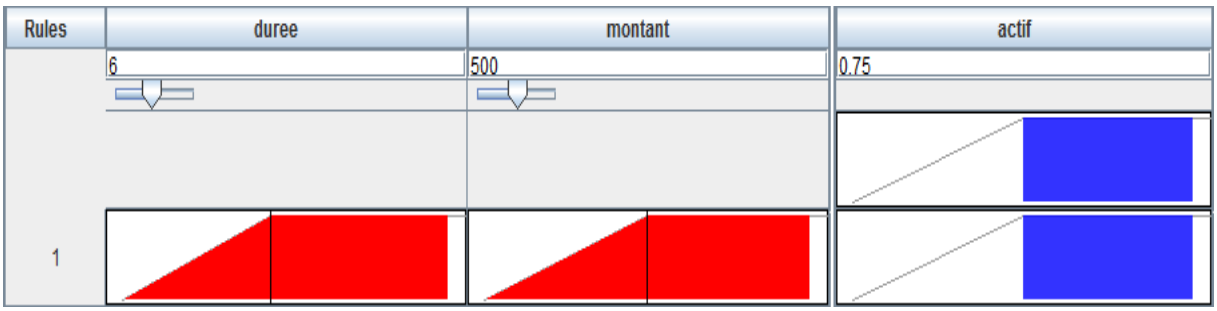



Figure 8 : Modèle flou décisionnel construit par le logiciel FISpro

## 4.2 Granular Computing

Le Granular Computing (GrC) [23]–[26] est une approche de résolution structurée de problème en se basant sur la représentation structurée des connaissances. En effet, le GrC représente les informations dans plusieurs niveaux de granularité (détail) où chaque niveau est composé de plusieurs granules d'information, qui définit un niveau d'abstraction du domaine traité. Ensuite l'ensemble des niveaux sont insérés dans une structure granulaire qui capture les différents aspects du problème traité (figure 9). Enfin, La structure granulaire permet le raisonnement dans les niveaux d'abstraction défini, et d'explorer des niveaux variables avec des solutions approximatives.

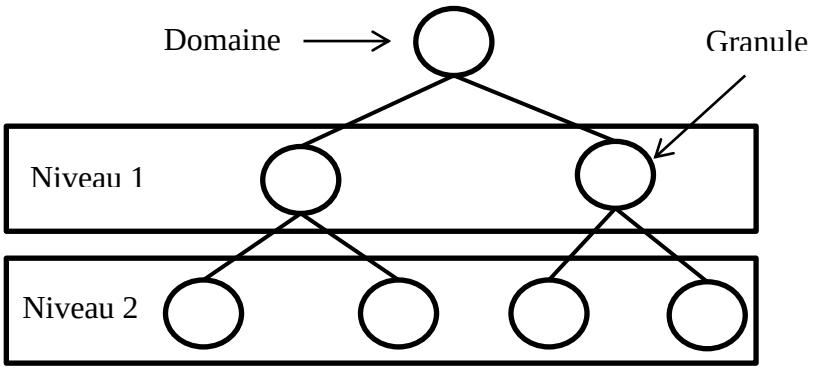



Figure 9: Exemple d'une structure granulaire à deux niveaux

Les éléments de base d'un modèle basé sur le GrC[26] peuvent être énumérés en trois éléments :

- Granule : est une abstraction du domaine traité, il est défini par un ensemble d'objets rassemblés selon un critère de similarité, dépendance, fonctionnalité, etc. on dit qu'un granule a une abstraction élevé s'il contient un grand nombre d'éléments.
- Structure granulaire : est une description structuré du domaine, il permet de représenter le domaine traité dans un schéma hiérarchique où chaque niveau représente un niveau d'abstraction.
- Calcul avec les granules : est un mécanisme qui permet la résolution du problème par l'exploration de la structure granulaire. Dans un premier temps un niveau d'abstraction supérieur doit être capable de résoudre le problème qui n'a pas besoin de trop de détails, si c'est le cas, il faut explorer un niveau inférieur pour le résoudre. Dans la figure 10 on montre une structure granulaire

d’une entreprise. Dans cet exemple, les problèmes concernant la comptabilité requiert beaucoup de détails et sont résolus dans les niveaux inférieurs au service de comptabilité, par contre la validation d’un modèle d’affaire est faite dans un niveau supérieur parce qu’il s’intéresse seulement aux résultats finaux du modèle.

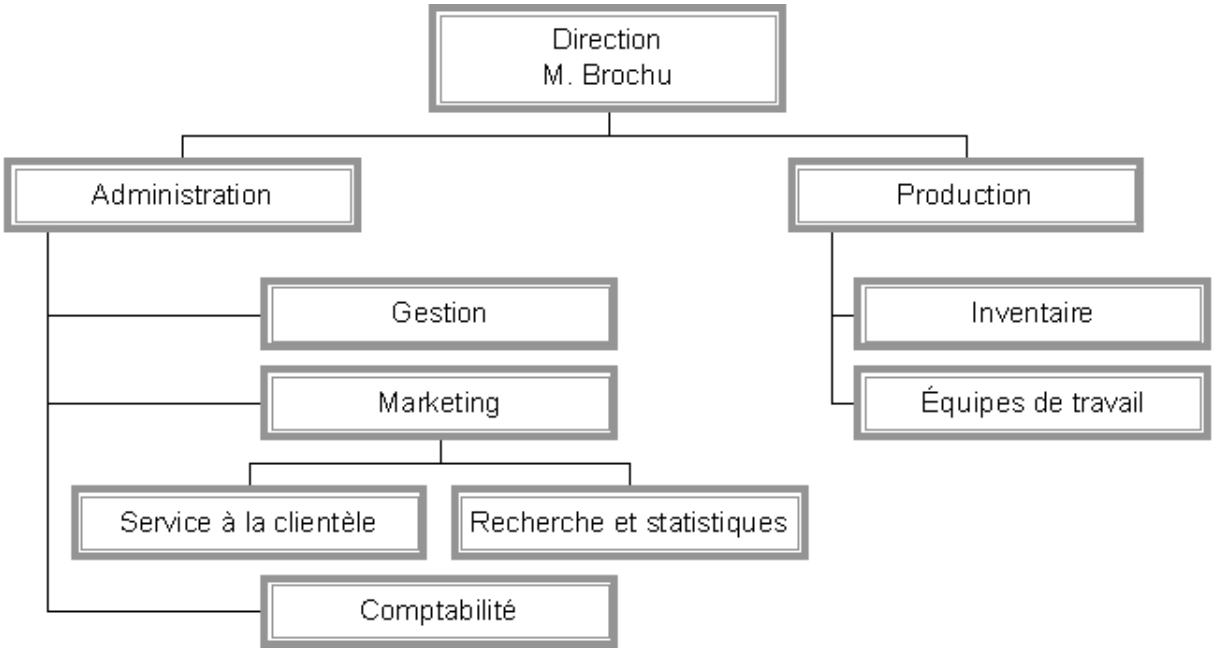



Figure 10: structure granulaire d'une entreprise

# 5 Conclusion

La fouille de donnée essaie de résoudre plusieurs problématiques qu’n avait classé en 4 catégories. Notre travail s’intéressera essentiellement aux problèmes de la classification supervisée, c’est pourquoi on essayera de concevoir un modèle basé sur le granular computing flou qui est un outil très fort dans la perception et la structuration du domaine et dans le traitement de l’information.

# Chapitre 2 : Granular Computing

# 1 Introduction

Le GrC est un outil informatique émergent pour le traitement de l'information. Il s'intéresse à la découverte, à la représentation et au traitement des entités d'information complexes appelées *granules d'information*. On peut le considérer comme une façon de penser qui repose sur la capacité humaine à comprendre des problèmes réels dans de différents niveaux de granularité. En se focalisant sur les différents niveaux de granularité, on peut obtenir plusieurs niveaux de connaissance d'où plusieurs résolutions du problème.

Le GrC est donc une méthode prometteuse de résolution de problème car il se base sur la représentation de domaine de problème dans un schéma multi-niveaux, chaque niveau définit une vue différente de résolution avec un niveau de détail élevé.

Dans ce chapitre on va présenter l'architecture générale d'un modèle de fouille de donnée basé sur le GrC, après on va définir les différents éléments ainsi comment le concevoir.

# 2 Architecture générale

Dans la figure 11 on montre les différentes parties d'un modèle de fouille de donnée basé sur le GrC:

- base de données : un ensemble d'échantillons qui capture les différents cas possibles dans un domaine donné. Un cas est composé de deux parties description de problème et sa solution.
- Granulation d'information : est un processus qui permet de construire une structure granulaire en se basant sur les données, la granulation d'information est réalisé en trois étapes :
    1. La construction des granules : permet de rechercher des granules d'informations par le rassemblement des objets selon un critère de **similarité, dépendance, fonctionnalité, etc**.
    2. Représentation des granules : permet la définition des granules trouvés par des motifs comme fonction, règles, etc.

3. Structure granulaire : permet de définir la hiérarchie des granules, pour cela il faut placer les granules d'une abstraction élevé dans un niveau élevé et le contraire pour les granules ayant une abstraction faible.

- Calcul avec les granules : est un mécanisme qui permet à partir de la description d'un nouveau cas de dériver une solution, deux mécanismes existent:
    1. le mécanisme de raisonnement dans un niveau donné, il utilise les granules d'information pour dériver une solution dans le niveau courant.
    2. Le mécanisme d'exploration des niveaux dans la structure granulaire, dans ce cas il faut prendre en considération les résultats obtenus dans un niveau supérieur.

Dans la suite une description détaillée de la partie de granulation de l''information, ensuite dans le chapitre suivant, une implémentation sur la classification supervisée est effectuée où on va expliquer le calcul avec les granules.

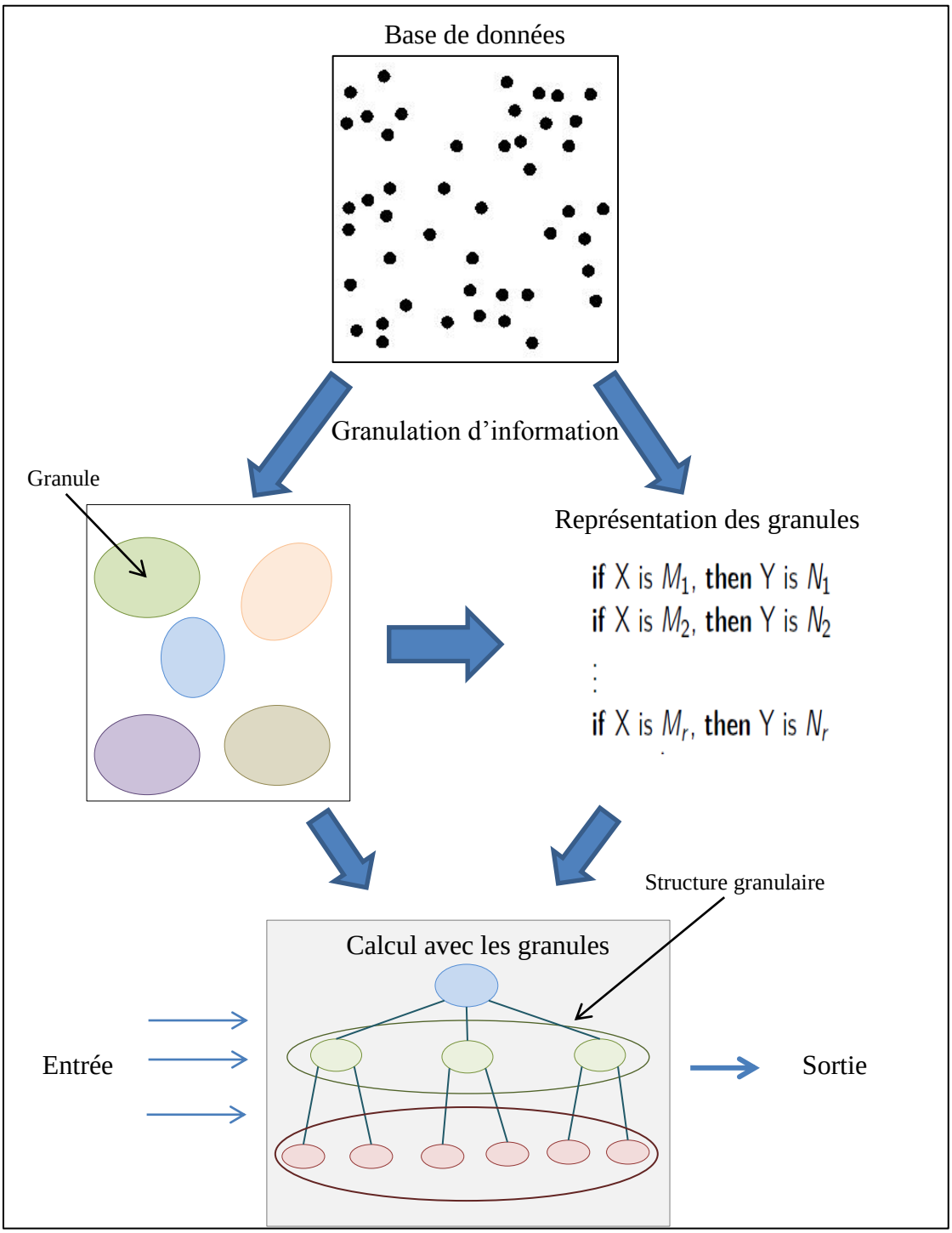



Figure 11: Architecture générale d'un modèle de fouille de données basé sur le Granular Computing

## 3 Granulation d'information :

La granulation d'information est un algorithme qui permet de rechercher des granules d'informations dans les données. Comme les granules sont définis par un ensemble d'objet rassemblés par un critère de ressemblance, fonctionnalité ou dépendance, intuitivement les algorithmes de Clustering sont très adéquats pour la recherche des granules informations.

Les méthodes de Clustering [27], [16] sont des méthodes non-supervisées qui ont pour objectif de déterminer les classes auxquelles appartiennent les objets en se basant sur certaines caractéristiques. Cependant le grand problème des algorithmes de clustering et la définition du nombre de cluster, pour cela des indices de validité sont employés pour définir le nombre de cluster optimal.

Par l'utilisation des algorithmes de clustering dans la granulation d'information, un granule donc sera défini par un cluster.

Dans ce qui suit, on présente les différentes techniques connues de clustering existantes dans la littérature, ainsi que quelques indices de validité.

## 3.1 Méthodes de clustering :

### 3.1.1 K-means :

L'algorithme de K-means[28] est un outil de fouille de données qui vise à partitionner en différentes classes l'ensemble des individus. On cherche à regrouper autant que possible les individus les plus semblables tout en séparant les classes les unes des autres.

La méthode des centres mobiles s'applique lorsqu'on connait déjà le nombre de classes K que l'on veut avoir.

L'algorithme est le suivant :

- Etape 0 : On initialise les centres en tirant aléatoirement K individus qui appartiennent à la population.
- Etape 1 : On répartit l'ensemble des individus en K classes en regroupant autour de chaque centre l'ensemble des individus qui lui sont plus proches que les autres centres.
- Etape 2 : On détermine les centres de gravité des K classes obtenues et on les désigne comme de nouveaux centres.
- Etape 3 : on répète les étapes 1 et 2 jusqu'à ce que l'algorithme se stabilise.

Généralement, cet algorithme est efficace mais présente quelques faiblesses comme l'initialisation des centres des clusters qui conditionne le résultat final (des initialisations différentes mènent à des clusters différents). Il peut arriver qu'un cluster ne contienne que son centre.

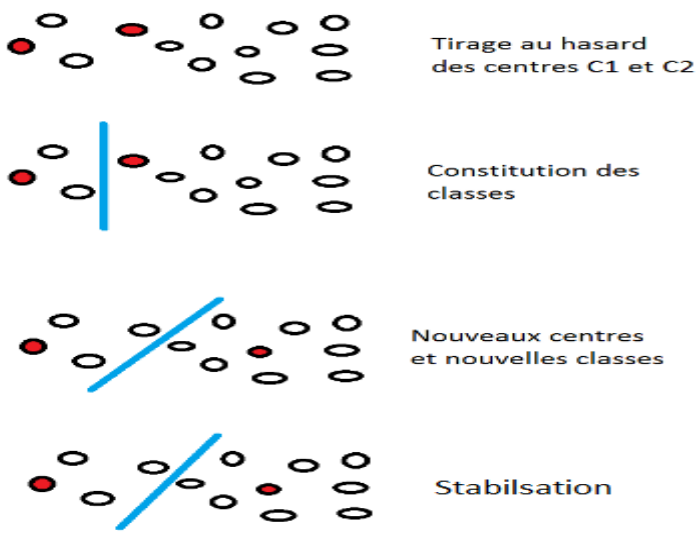



Figure 12: illustration de l'algorithme K-means

### 3.1.2 Clustering hiérarchique :

La méthode hiérarchique [17] génère une partition de l'espace des données, mais aussi une succession de partitions de l'espace des données. Celles-ci sont souvent représentées sous la forme d'un dendrogramme. Un dendrogramme est un arbre de partitions successives de l'espace des données. Selon la direction que l'on parcourt le dendrogramme (de haut en bas ou de bas en haut), la méthode sera appelée descendante (division) ou ascendante (agglomération).

A chaque étape de ces algorithmes, on cherche à fusionner (CAH) ou à diviser (CHD) les clusters un-par-un, la raison pour laquelle on calcule la matrice de distances/proximités puis après on utilise un critère d'agrégation/de division qui nous aidera à faire le bon choix.

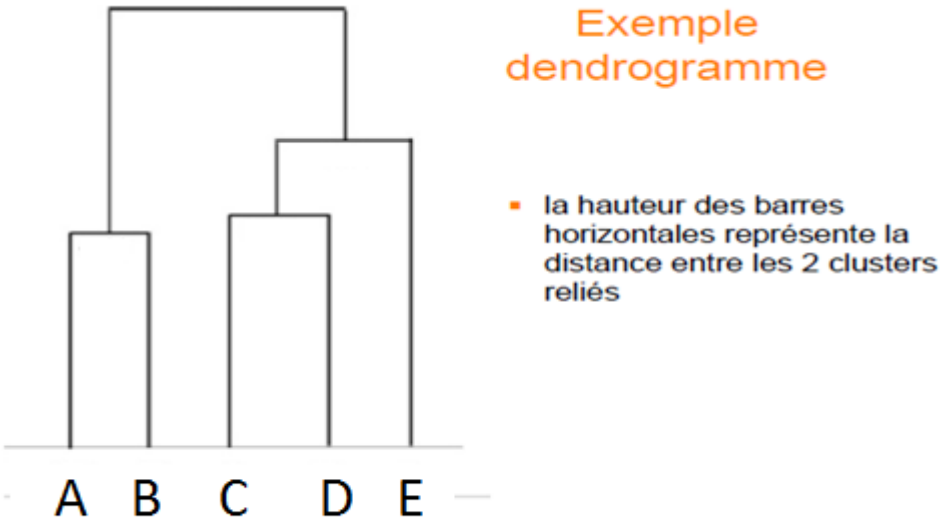



Figure 13: Exemple de dendrogramme de 5 éléments

### *3.1.2.1 Clustering Ascendant Hiérarchique (CAH) :*

Les principaux étapes d'une méthode ascendante quelconque est comme suit :

Algorithme :

Soient :

- Une population X de n individus
- Une fonction de distance *dis* s'appliquant aux paires de sous-ensembles de X.

$$dis: P(X) \times P(X) \rightarrow [\, dis_{min}, dis_{max}]$$

- Etape 1 : Initialiser n clusters chacun réduit à un individu de X.
- Etape 2 : fusionner la paire de cluster ( $C_{n1}, C_{n1}$) de distance minimale.

$$(\, C_{n1}, C_{n1}) = \underset{\substack{(i,j)\in([1,card(C)]\times[1,card(C)]) \\ i\neq j}}{\operatorname{argmin}} dis(\, C_i, C_j) \qquad (3)$$

- Etape 3 : Si $(C) \neq 1$ , revenir à l'étape 2.

Les différentes méthodes se distinguent par le choix de la fonction de distance. Leur nom dépend du choix fait. Dans la littérature on trouve une grande variété de méthodes. Nous allons présenter le comportement de certaines méthodes car la manière de fusion de deux clusters diffère d'une méthode à une autre.

- Saut minimal (single linkage) : cette méthode a tendance de regrouper les deux clusters les plus similaires, et tend à agréer un pont à un cluster déjà existant plutôt qu'à donner naissance à un nouveau cluster. Elle crée des clusters allongés car un seul membre proche suffit pour effectuer le regroupement.

$$Dis(C_i, C_j) = \min_{x \in C_i, y \in C_j}(dis(x, y)) \quad (4)$$

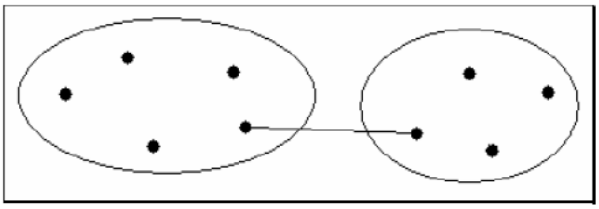


Figure 14 : distance entre deux clusters avec saut minimal

- Saut maximal (complete-linkage) : dans cette méthode, les distances entres classes sont déterminées par la plus grande distance existant entre deux objets de classes différentes. Au contraire du saut minimal, cette approche crée des clusters ramassés

$$Dis(C_i, C_j) = \max_{x \in C_i, y \in C_j}(dis(x, y)) \quad (5)$$

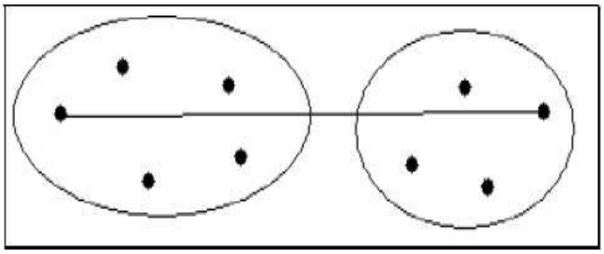


Figure 15 : distance entre deux clusters avec saut maximal

- lien moyen (average-linkage) : Il définit la distance entre deux clusters en faisant intervenir tous les objets présents dans ces clusters.

$$Dis(C_i, C_j) = \text{moyenne}_{x \in C_i, y \in C_j}(dis(x, y)) \quad (6)$$

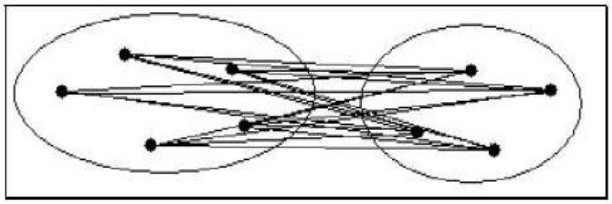


Figure 16: distance entre deux clusters avec lien moyen

- Méthode des centroïdes : dans cette méthode, la distances entres les classes est déterminée par la distance entre leur centres.

$$Dis(C_i, C_j) = \text{dis}(\frac{1}{\text{card}(C_i)}\sum_{k=1}^{k=card(C_i)} x_k , \frac{1}{card(C_j)}\sum_{l=1}^{l=card(C_j)} y_l) \qquad (7)$$

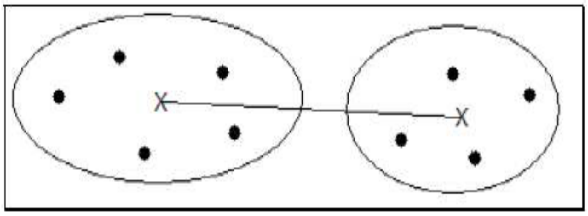


Figure 17 :distance entre les cenroïdes de deux clusters

- Méthode de Ward : Cette méthode se distingue de toutes les autres en ce sens qu'elle utilise une analyse de la variance intra-classe afin d'évaluer les distances entre classes. En résumé, cette méthode tente de minimiser la somme de carrés de toutes les paires de classes pouvant être formées à chaque étape. En d'autre terme, l'agrégation entraîne la diminution minimale de l'inertie du nuage.

  Les indices d'agrégation sont recalculés à chaque étape à l'aide de la règle suivante :

$$dis(C_{ij}, C) = \frac{(N_{C_i}+N_C)dis(C_i,C)+(N_{C_j}+N_C)dis(C_j,C)-N_C dist(C_j,C_j)}{N_{C_i}+N_{C_j}+N_C} \qquad (8)$$

  Où $C_{ij}$ est une classe obtenue en regroupant les classes $C_i$ $et$ $C_j$, désigne le cardinal de la classe.

### *3.1.2.2 Clustering Descendant Hiérarchique*

Contrairement aux méthodes ascendantes, les méthodes du clustering descendant hiérarchique commencent par l'ensemble entier comme étant un seul cluster, et le sépare ensuite en deux clusters dans le sens de minimiser un critère d'optimisation c. Après cette séparation, les clusters résultants à leur rôle seront divisés, et ainsi de suite jusqu'à ce que la division ne soit plus possible.

Il existe essentiellement deux alternatives pour diviser les classes :

- Les méthodes polythétiques qui utilisent tous les variables pour les divisions successives.

- Les méthodes monothétiques qui utilisent une seule variable pour les divisions successives.

Généralement, les méthodes du clustering hiérarchique descendant procèdent comme suit :

Algorithme :

Soient :

- Une population X de n individus
- Un critère $c$: [$ensemble\ des\ clusters\ C$] → [ $dis_{min}, dis_{max}$]

➢ Etape 1 : Initialiser C à l'ensemble vide.

➢ Etape 2 :

pour chaque individu x de X

pour chaque k cluster $C_i$ de C, recalculer $c(C)$ en considérant que x est dans $C_i$.

on construit un nouveau cluster {x} et on le met dans C et on recalcule $c(C)$.

parmi les k+1 possibilités on choisit celle qui la plus grande valeur de $c(C)$

### 3.1.3 Groupage flou:

L'algorithme FCM [29] est parmi les algorithmes non-supervisés les plus connus dans le domaine de la reconnaissance de forme. FCM vise à minimiser la fonction objective suivante :

$$J_m(U,V) = \sum_{j=1}^{c} \sum_{i=1}^{n} {\mu_{ij}}^m \left\| x_i - v_j \right\|^2 \quad (9)$$

Tout en respectant la contrainte suivante :

$$\sum_{j=1}^{c} \mu_{ij} = 1\,, \forall i = 1, \ldots\ldots, n \quad (10)$$

Où :

$x_i$ désigne la donnée i.

$c$ : le nombre de clusters que l'on souhaite crée.

$m$ : l'indice de fuzzification.

$v_j$ : le centre du cluster.

$\mu_{ij}$ : Le degré d'appartenance de la donnée i au cluster j.

Pour effectuer le bon partitionnement, l'algorithme suit les étapes suivantes :

- Etape 1 : tout d'abords on fixe le nombre de clusters désiré, l'indice m et le critère d'arrêt (le choix entre un nombre d'itération maximale ou un taux d'erreur). Ensuite, on initialise aléatoirement la matrice de partition en respectant la contrainte précédente.
- Etape 2 : On calcule les centres des clusters en utilisant la formule suivante

$$v_j = \frac{\sum_{i=1}^{n} {\mu^m}_{ij} x_i}{\sum_{i=1}^{n} {\mu^m}_{ij}} \quad (11)$$

- Etape 3 : On calcule la nouvelle matrice de partition, suivant les nouvelles positions des centres

$$\mu_{ij} = \frac{1}{1+\sum_{k=1}^{c} (\frac{d_{ik}}{d_{ij}})^{\frac{2}{m-1}}} \quad (12)$$

- Etape 4 : Refaire les étapes 2 et 3 si jamais le critère d'arrêt n'est pas vérifié.

Les algorithmes de type FCM sont sensibles au bruit et aux point aberrants. Les faibles valeurs associées aux points bruits peuvent exprimer la contamination du bruit, cependant, comme on peut voir dans (12), les degrés d'appartenance générés en respectant la contrainte (10), sont des nombres relatifs. Cela veut dire que les point bruits et les points aberrants vont avoir au moins la valeur de 1/c pour tous les clusters. Chaque augmentation de la valeur d'appartenance à un cluster implique la diminution des degrés d'appartenance aux autres clusters. Ce dernier fait, conduit à l'idée que les points bruits peuvent avoir des degrés d'appartenances élevés, ce qui va influencer sur l'estimation des paramètres des prototypes.

La deuxième principale déficience de FCM, est dû à la contrainte (10), les degrés d'appartenance sont interprétés comme des degrés de partage, mais pas comme des degrés de

possibilité d'appartenance d'un point à une classe. Un degré de possibilité d'appartenance peut mieux être adapté pour la théorie des ensembles flous théoriques.

## 3.2 Indices de validité :

Les indices de validité [30] sont des techniques qui nous aide à qualifier un clustering effectué : une bonne partition sera formée de clusters bien séparés les uns les autres (séparation) et dont les éléments sont rassemblés autour du centre de gravité (compacité).

Ils existent deux types d'indice de validation :

- Indices de validité interne : basés sur les informations tirées des données seulement.
- Indices de validité externes : basée sur la connaissance préalable des données.

### 3.2.1 Indices de validité interne :

#### *3.2.1.1 Indice de BIC :*

Le critère bayésien d'information est conçu d'une manière à éviter le surajustement, est défini ainsi :

$$BIC = -\ln(L) + v * \ln(n) \tag{13}$$

Où n est le nombre d'objets, L est la vraisemblance des paramètres que génèrent les données dans le modèle, et v est le nombre de paramètres libres dans le modèle gaussien. L'indice de BIC prend en considération l'ajustement du modèle aux données et sa complexité. Le modèle dont l'indice de BIC est petit, est le meilleur à retenir.

#### *3.2.1.2 Indice de Calinski-Harabasz :*

L'indice est calculé par la formule suivante :

$$CH = \frac{trace(S_B)}{trace(S_w)} . \frac{n_p - 1}{n_p - k} \tag{14}$$

Où $(S_B)$ est la matrice de covariance inter-cluster, $(S_w)$ est la matrice de covariance intra-cluster, $n_p$ le nombre des exemples de l'ensemble de données, et p est le nombre de clusters.

### 3.2.1.3 Indice de Davies-Bouldin (DB) :

Cet indice essaie d'identifier des ensembles de clusters qui sont compacts et bien séparés. Il est défini ainsi :

$$BD = \frac{1}{c}\sum_{i=1}^{c} Max_{i\neq j}\{\frac{d(X_i)+d(X_j)}{d(c_i,c_j)}\} \quad (15)$$

Où c désigne le nombre de clusters, $d(X_i)$ $et$ $d(X_j)$ sont les données appartenant aux clusters i et j respectivement, et $d(c_i, c_j)$ est la distance entre les centroïdes de ces clusters. La petite valeur de DB indique la bonne solution du clustering.

### 3.2.1.4 Indice de Silhouette :

Pour un cluster donné, $X_j(j = 1, \dots c)$, la technique de silhouette attribue au i-ème exemple une mesure de qualité, $s(i) = (i = 1, \dots m)$, connu comme largeur de la silhouette. Cette valeur est un indicateur de confiance du degré d'appartenance de l'élément i au cluster $X_j$ et elle est définie ainsi :

$$s(i) = \frac{b(i)-a(i)}{mac(a(i),b(i))} \quad (16)$$

Où a(i) est la distance moyenne entre le i-ème élément et les élément du cluster $X_j$, et b(i) est la distance moyenne minimale entre le i-ème élément et les éléments des autres clusters.

### 3.2.1.5 Indice de Dunn :

L'indice de Dunn est défini comme suit :

$$Dunn = min_{1\leq i\leq c}\{\min\left\{\frac{d(c_i,c_j)}{max_{1\leq k\leq c}(d(X_k))}\right\}\} \quad (17)$$

Où $d(c_i, c_j)$ désigne la distance inter- cluster entre le cluster $X_i$ et $X_j$, et $d(X_k)$ représente la distance intra-cluster du cluster $X_k$ et c est le nombre de clusters. La grande valeur de l'indice Dunn correspond à la bonne solution du clustering.

### 3.2.1.6 Indice de Xie-beni :

Cet indice [31] est utilisé surtout dans les algorithmes du clustering flou. Il est défini par l'équation suivante :

$$XB = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ki}{}^2.d(x_k,c_i)}{n.min_{i\neq j}d(c_i,c_i)} \quad (18)$$

Où $\mu_{ki}$ désigne le degré d'appartenance flou de l'élément $x_k$ au cluster $c_j$, et n est le taille de l'ensemble des données.

### 3.2.2 Indices de validité externes :

#### *3.2.2.1 F-Measure :*

Il combine entre le concept du rappel et celui de la précision issus de la recherche d'information. On calcule la précision et le rappel de chaque cluster pour chaque classe comme suit :

$$Rappel(i,j) = \frac{n_{ij}}{n_i} \tag{20}$$

Et

$$Précision(i,j) = \frac{n_{ij}}{n_j} \tag{21}$$

Où $n_{ij}$ est le nombre des objets de la classe i qui appartiennent au cluster j, $n_j$ est le nombre des objets du cluster j, et $n_i$ est le nombre des objets du cluster i. L'indice de F-Measure du cluster j et de la classe i est donné par la formule suivante :

$$F(i,j) = \frac{2Rappel(i,j)Precision(i,j)}{Precision(i,j)+Rappel(i,j)} \tag{22}$$

Les valeurs de F-Measure sont comprises entre [0,1] et la valeur la plus grande indique le clustering de plus haute qualité.

#### *3.2.2.2 Pureté :*

La pureté est similaire à l'entropie. Pour calculer la pureté de l'ensemble des clusters, on calcule la pureté de chaque cluster :

$$P_j = \frac{1}{n_j} Max_i(n_j^i) \tag{23}$$

Où $n_j^i$ désigne le nombre des objets du cluster j dont le label de la classe est i. La pureté globale de la solution du clustering est donée par :

$$Pureté = \sum_{j=1}^{m} \frac{n_j}{n} P_j \tag{24}$$

Où $n_j$ est la taille du cluster j, m le nombre de clusters, et n est le nombre total d'objets.

### *3.2.2.3 L'entropie :*

L'entropie mesure la pureté des labels de la classe des clusters. Par conséquent, si chaque cluster contient des objets qui ont tous le même label de la classe, l'entropie sera égale à 0. Pour calculer l'entropie d'un ensemble de donnée, nous avons besoin de calculer la distribution des classes des objets dans chaque cluster comme suit :

$$E_j = \sum_i p_{ij} \log(p_{ij}) \quad (25)$$

L'entropie totale pour un ensemble de clusters est calculée par l'équation suivante :

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \quad (26)$$

Où $n_j$ est la taille du cluster j, m le nombre de clusters, et n est le nombre total d'objets.

## 4 Représentation des granules

La représentation d'une granule doit être effectuée par un concept qui permet de définir la notion d'abstraction, pour cela la théorie des ensembles flous est très adéquate puisqu'elle permet la représentation des granules par un ensemble flou. Ce dernier est défini par une règle floue qui capture d'une manière abstraite la description et la solution des cas inclus dans le granule.

A cet égard, un granule est représenté par une règle floue (figure 18) contenant en entrée les variables linguistiques qui décrivent un cas et en sortie le terme linguistique de la solution.

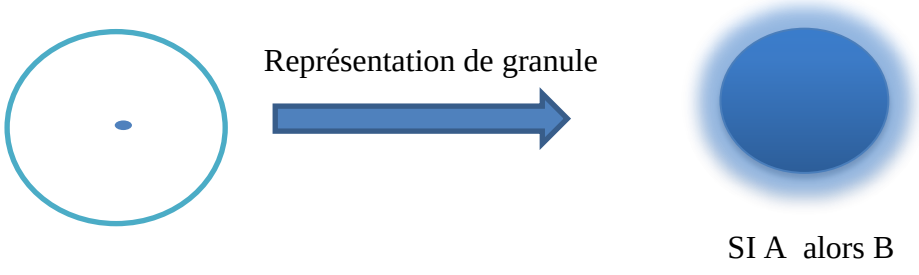



Figure 18 : représentation de granule par une règle floue

## 5 Structure granulaire

Une structure granulaire permet de définir le nombre de niveau de détail dans un domaine, cependant il n'est pas toujours nécessaire de construire plusieurs niveaux pour améliorer le résultat du système, en effet parfois un seul niveau est suffisant, dans ce sens, il existe deux types de structure granulaire qu'on peut définir :

- Structure simple : elle contient un seul niveau, on peut obtenir cette structure par l'application de clustering et la représentation des granules obtenus.
- Structure multi-niveaux : elle contient plusieurs niveaux où chaque niveau définit un degré de détail. Pour obtenir une structure multi-niveaux, l'approche de clustering successive [4] est la plus utilisée. Cette approche permet d'effectuer la granulation d'information dans chaque niveau en prenant en considération le contexte du niveau précédent.

## 6 Conclusion :

Dans cette section, on décrit d'une manière détaillée les éléments de base du GrC, ainsi que les éléments d'un modèle à base du GrC flou, lequel on va exploiter pour résoudre des problèmes de la classification supervisée du monde réel.

# Chapitre 3 : Application à la classification supervisée

# 1 Introduction :

Parmi les domaines d'intérêt de la fouille de donnée, on trouve la classification supervisée dans laquelle on chercher à prédire la classe d'un nouvel élément en se basant sur une base d'exemples. A partie de cette base, on essaie de dériver des classifieurs qui vont nous aider à effectuer cette tâche, l'ensemble de ces classifieurs constitue ce qu'on appelle un système de classification. Dans ce chapitre, on va voir comment appliquer un modèle de classification supervisée basé sur le granular computing

# 2 Problématique de la classification supervisée :

Le problème de classification est implicitement défini par un ensemble de N paires de données :

$$D = \{(X_i, c_i) \in \mathbb{R}^n \times C, i = 1, \dots, N\}$$

Où $X_i = (x_{i1}, \dots, x_{in})$ sont des données de l'Univers du Discours et C est l'ensemble des labels des classes du problème en étude.

L'objectif est donc de prédire la classe c d'un nouveau élément x à partir de l'ensemble de données D en se trompant rarement que possible, où (x,c) est une paire de l'ensemble $\mathbb{R}^n \times C$, indépendante des paires appartenant à D. On parle de ***classification*** car il s'agit d'attribuer une étiquette c à un nouvel objet. Quant au terme ***supervisée***, il est dû au fait que tous les échantillons sont étiquetées, autrement dit on connait au préalable la classe de chaque élément. Il existe une multitude de problèmes dans lesquels la classification supervisée entre en action : reconnaissance et identification de caractères manuscrits, reconnaissance du locuteur, aide au diagnostic médical,…

La résolution d'un problème de classification consiste à trouver un ensemble de classifieurs appelé modèle de classification à partir du jeu de donnée, définis comme suit :

$$\mathrm{S} = \{S_i : \mathrm{x} \rightarrow \mathrm{c}\,, \mathrm{i} = 1, \dots, \mathrm{r}\}$$

Où r est le nombre des classifieurs ou règles de classification constituant le modèle.

Etant donné un nouvel élément $x_j$, le classifieur S lui attribue un label de classe $S(x_j)$, que l'on espère coïncider avec $c_i$. La valeur $S(x_j)$ est déduite d'après les sorties $S_i(x_j)$ de tous les classifieurs constituant le modèle selon un critère d'agrégation dépendant de la méthode utilisée pour la classification. Donc un bon modèle S sera celui dont le taux de

classification est le plus grand, autrement dit dont le risque d'attribuer une mauvaise classe à un nouvel élément est très petit.

Dans la littérature [2], [3], il existe plusieurs méthodes de classification supervisée (Arbres de décision, SVM, …). Dans notre travail on a choisi de résoudre le problème de classification par les systèmes d'inférence flous (système à base de règles floues) où les règles seront construites à partir d'une approche du granular computing qui sera détaillée ensuite.

## 3 Système d'inférence flou :

Les systèmes d'inférence flous (SIF) [32] connus encore sous le nom de systèmes à base de règles, se composent fondamentalement de quatre blocs fonctionnels :

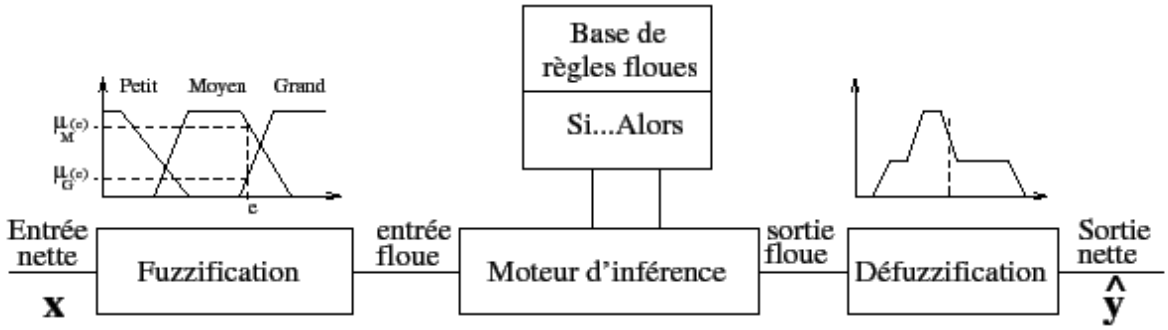



Figure 19: schéma général d'un SIF

- Une base de règles contenant un nombre de règles de la forme suivante: $Si\ (x_1\ est\ A_1\ ET\ \ldots\ x_k\ est\ A_k)\ Alors\ P(y = s_1) = \pi_1 ET\ \ldots\ P(y = s_c) = \pi_c$ où $x_1 \ldots x_n$ sont les variables d'entrée et $A_1 \ldots A_k$ sont les différentes ensembles d'entrée. La zone d'influence d'une règle floue est appelée le prototype de cette règle. Les prototypes des d'inférence sont calculés à l'aide d'algorithmes de clustering. La conclusion d'une règle floue diffère selon le type de système utilisé, elle peut être soit des ensembles flous (FIS Mamdani) ou des combinaisons linéaires des variables d'entrée (FIS Takagi and Sugeno).

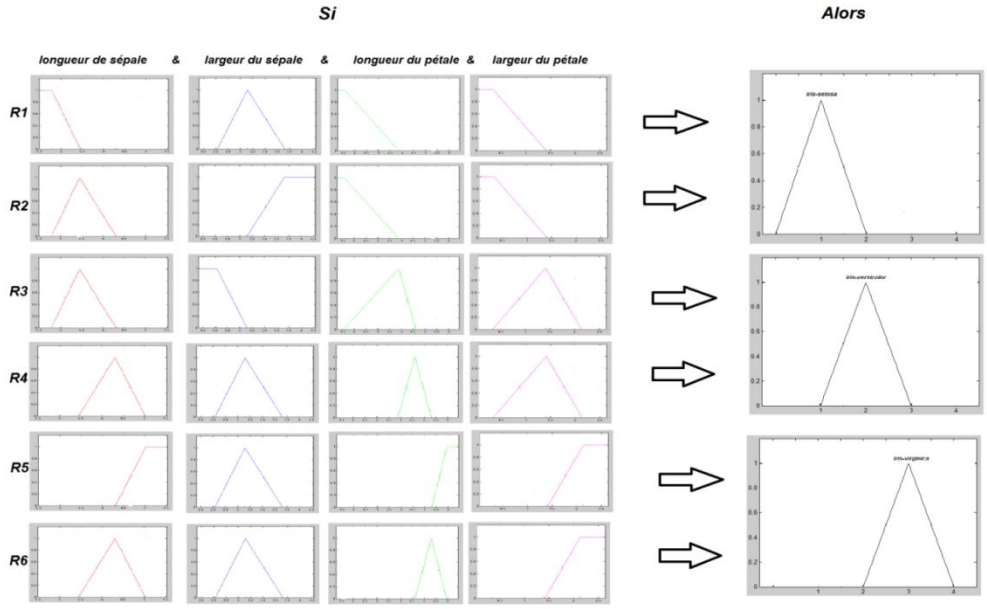



Figure 20: Exemple de règles floues utilisées pour le problème de classification de la base de données Iris

- Une interface de fuzzification qui transforme les entrées nettes à des degrés de correspondance avec des valeurs linguistiques, en d'autre terme elle affecte à chaque attribut d'une certaine entrée, un degré d'appartenance aux ensembles flous correspondants.

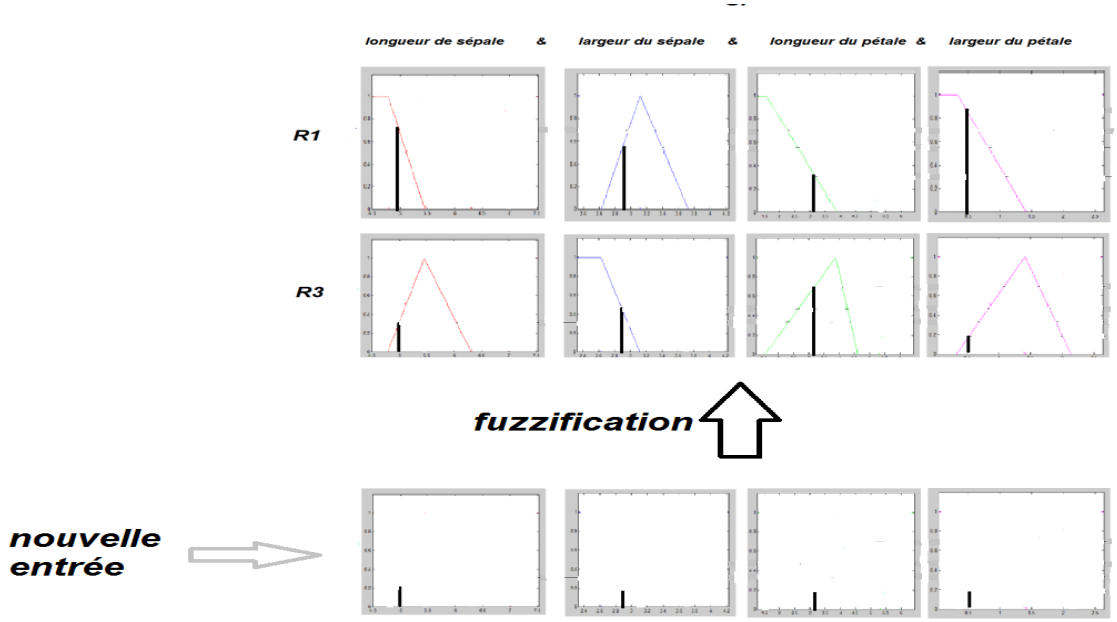



Figure 21: fuzzifcation d'un nouvelle entrée d'iris par l'interface de la figure 19

- Un moteur d'inférence flou qui effectue des opérations d'inférence sur les règles.Elle effectue effectue la conjonction des degrés d'appartenances via des opérateurs T-normes (voir tableau) spécifiques pour obtenir la sortie de chaque règle. Enfin, on combine les sorties par un opérateur T-conorme pour obtenir la sortie totale.

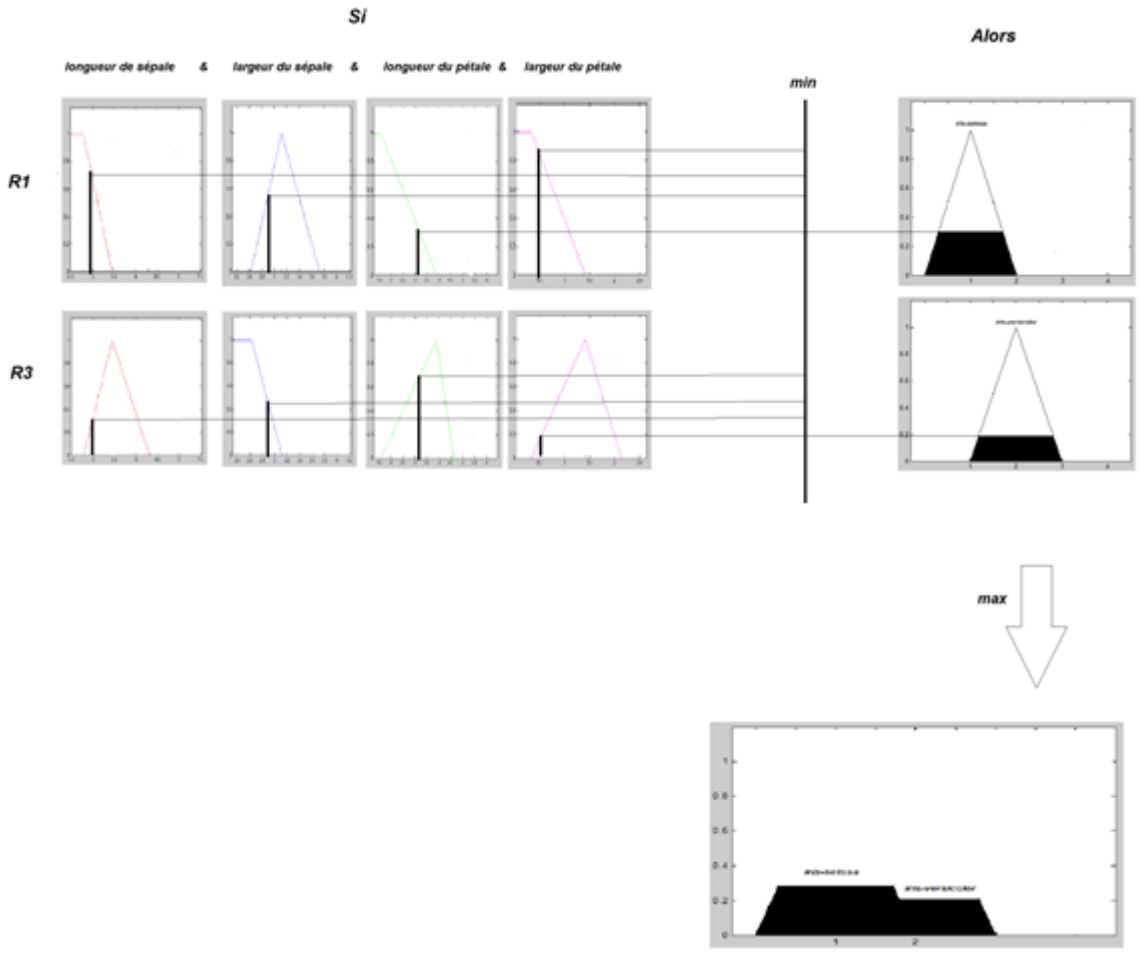



Figure 22: inférence de l'entrée précédente à l'aide d'un FIS Mamdani en utilisant « min » comme T-norme et « max » comme T-conorme

- Une interface de défuzzification qui transforme les résultats d'inférence à une sortie nette, en utilisant certaine techniques. Parmi celle les plus connues : la technique du centroïde de la surface qui consiste à calculer le centre de gravité de la surface de la sortie totale, et celle de la moyenne des maximas dont laquelle la sortie nette est la moyenne des valeurs de la sortie totale qui atteignent la valeur maximale.

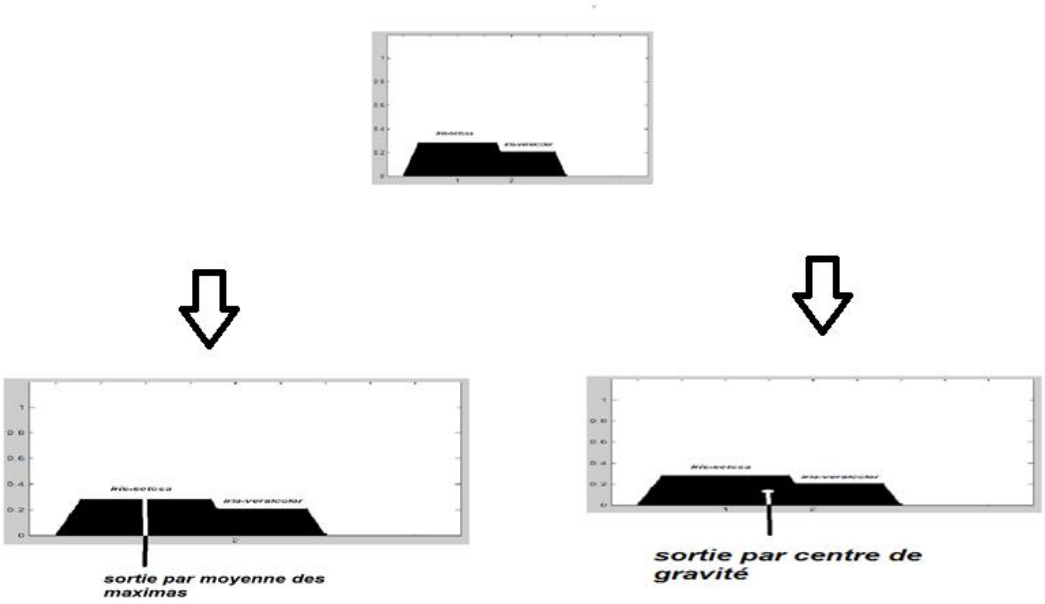



Figure 23: défuzzification

Plusieurs types du raisonnement flou ont été proposés dans la littérature. En se basant sur les types du raisonnement et les règles floues employées, on peut classifier les SIF en trois catégories :

- 1er type : La sortie globale est la moyenne des sorties nettes de chaque règle induites par fonctions d'appartenance de sortie pondérées par les poids des règles. Les fonctions d'appartenance de sortie utilisées dans ce schéma doivent être monotones.
- 2ème type : La sortie globale floue est calculée en appliquant l'opérateur « max » aux sorties floues qualifiées (chaque sortie est égale au minimum du poids de la règle et la fonction d'appartenance de la sortie). Plusieurs schémas ont été proposés à ce stade pour choisir la sortie nette basée sur la sortie globale floue, parmi lesquels on trouve : critère du maximum, la moyenne des maximas, …
- 3ème type : Ce type utilise les règles floues de Takagi et Sugeno. La sortie de chaque règle est une combinaison linéaire des variables des entrées plus une constante, et la sortie finale est la moyenne pondérée des sorties de chaque règle.

La différence principale entre ces types provient de la spécification de la conséquence, en d'autres termes les schémas de défuzzification.

# 4 Modèle de classification supervisée basé sur le GrC

## 4.1 Granulation d'information :

La granulation d'information est effectuée en se basent sur la méthode du Double clustering [33], en effet cette technique permet d'extraire les granules d'information flous représentable par des termes de labels linguistiques qualitatifs. La technique essaie d'exploiter les caractéristiques du Clustering multidimensionnel et celui monodimensionnel. Le Clustering multidimensionnel capture la granularité des données dans l'espace multidimensionnel, mais la fuzzification des granules résultants peut mener à des ensembles flous qu'on ne peut pas associer à des labels linguistiques qualitatifs. Inversement, le clustering monodimensionnel assure des ensembles flous interprétables mais il peut perdre de l'information sur la granularité des données multidimensionnelles. L'intégration du clustering multidimensionnel et celui monodimensionnel permet une granulation interprétable de l'information.

Le DC s'effectue principalement en trois étapes principales :

1) Clustering des données :

Le Clustering est effectué dans l'espace multidimensionnel des données numériques pour regrouper les données similaires dans des granules. Ici, les granules d'informations sont décrits par des prototypes de cluster multidimensionnels. Les prototypes là sont conçus comme des éléments qui caractérisent les relations cachés découvertes via le Clustering.

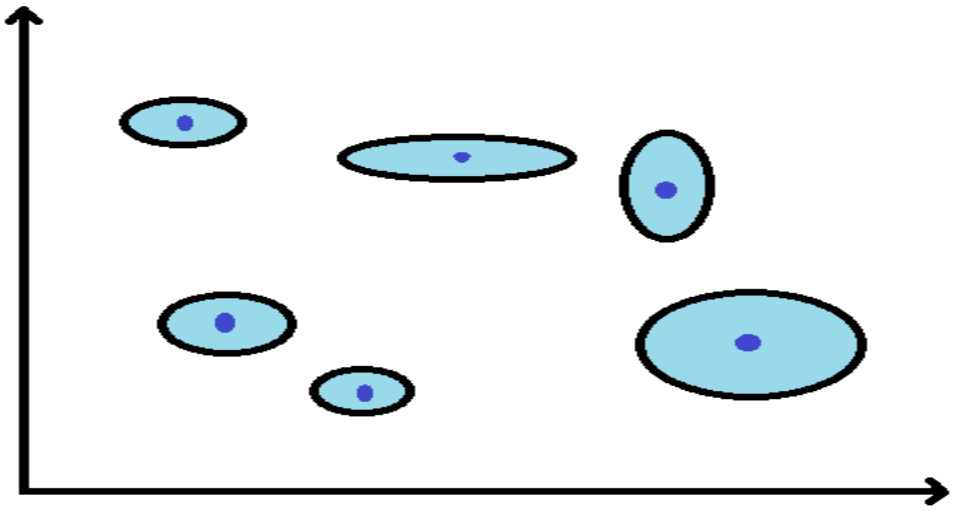


Figure 24: La 1ère étape du DC; clustering multidimensionnel des données

2) Clustering des prototypes :

Les prototypes multidimensionnels obtenus par la première étape sont projetés sur chaque dimension d'attribut. On effectue ensuite un Clustering sur ces projections pour obtenir le nombre de prototypes monodimensionnel pour chaque attribut.

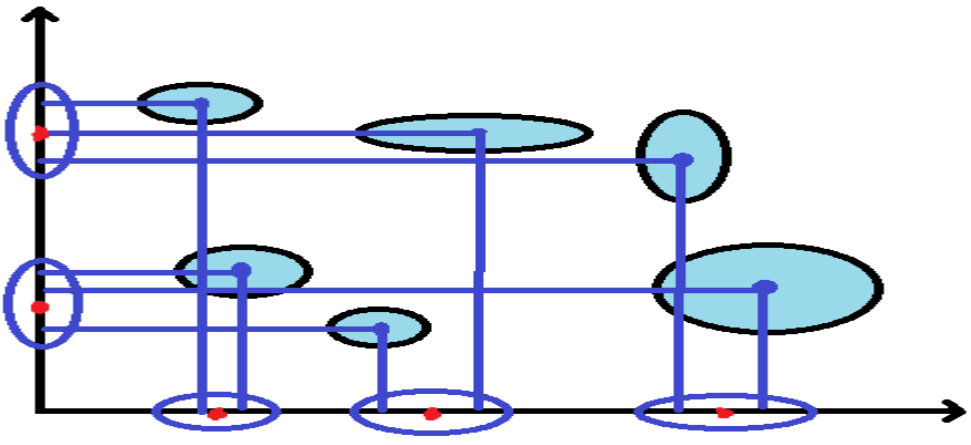


Figure 25: La 2ème étape du DC; clustering des projections des prototypes sur chaque dimension

3) Fuzzification des granules :

Les prototypes multidimensionnel et monodimensionnel fournissent une information utile pour extraire les granules d'informations qu'on peut représenter par des

ensembles flous. De plus, ces ensembles flous sont construits d'une façon répondante à la contrainte d'interprétabilité qui permet une description qualitative des granules de l'information.

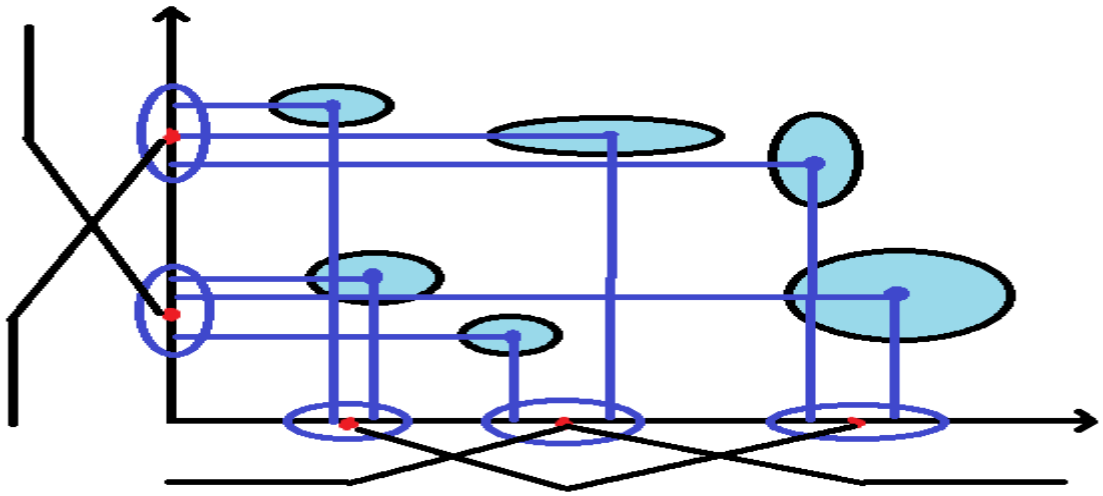


Figure 26: 3ème étape; dérivation des ensembles flous pour chaque dimension

Formellement, Le DC peut être décrit comme suit :

Soit $X \subseteq \mathbb{R}^n$ l'univers n-dimensionnel du discours dans lequel un ensemble de données numériques $D = \{x_i \in X : i = 1 \dots N\}$ est valable.

La première étape du DC donne comme résultat une collection de p prototypes multidimensionnels :

$$c_1, c_2, \dots, c_p \in X$$

Etant $c_i = \{c_{11}, c_{12}, \dots, c_{1n}\}$ pour $i = 1, \dots, p$. Les prototypes multidimensionnels sont projetés sur chaque dimension, menant à n ensembles :

$$C(j) = \{c_{1j}, c_{2j}, \dots, c_{pj}\}$$

pour $= 1, \dots, n$ . Dans la deuxième étape, les points de chaque C(j) subit un Clustering monodimensionnel, menant à n ensembles de prototypes monodimensionnels : $P(j) = \{p_{1j}, p_{2j}, \dots, p_{K_j j}\}$ tel que $K_j$ est le nombre de clusters de la dimension j, $j = 1, \dots, n$. Généralement, on utilise le clustering hiérarchique ascendant car le nombre des centres est petit.

La dernière étape du DC nécessite la représentation des granules floues d'informations. Cela est atteint premièrement par la fuzzification des granules monodimensionnel définis par les prototypes pour chaque P(j), et deuxièmement par l'agrégation des ensembles flous monodimensionnels pour former les granules flous d'information multidimensionnels.

La fuzzification des granules d'informations est atteinte en définissant des ensembles floues (gaussiennes, triangulaires, …) qui vérifient la contrainte d'interprétabilité. Précisément pour chaque dimension $j = 1, \dots, n$ , les $K_j$ ensembles flous sont définis par des fonctions d'appartenance $\mu^j{}_{A_k}(x)$ pour chaque $k = 1, \dots, K_j$

Donc, les granules flous d'information multidimensionnels peuvent être formés en combinant les ensembles flous monodimensionnels, un pour chaque dimension. Parmi toutes les combinaisons possibles des ensembles flous monodimensionnels, seulement ceux qui représentent mieux les clusters découverts dans la première étape son sélectionnés. La sélection de tel granules est accomplie pour chaque dimension, en considérant pour chaque cluster , $i = 1, \dots, p$ , l'ensemble flou de la jième dimension pour lequel l'attribut j du cluster i atteint la valeur maximale du degré d'appartenance. La représentation linguistique finale du granule d'information dérivé est une conjonction des contraintes comme suit :

$$G = attribut\_1\ est\ petit\ ET\ attribut\_2\ est\ moyen \dots\dots ET\ attribut_n est\ grand$$

Un fois le processus de la granulation est terminé, On peut construire un modèle basé sur les règles floues à base des granules flous construits, qui nous seront très utiles dans les problèmes de classification.

Pour extraire une base de règle, chaque granule extrait sera divisé en deux parties : prémisse et conséquence. La partie de la prémisse sera définie par la représentation linguistique des granules d'information, tandis que la conclusion sera présentée par la conjonction des fréquences relatives des observations de chaque classe appartenant à ce granule-ci.

Donc, les règles floues construites à partir des granules d'information seront sous cette forme-ci :

$$SI\ x\ est\ G_k\ ALORS\ P(Classe = 1) = \pi_{k,1}\ AND\ \dots AND\ P(Classe = C) = \pi_{k,C}$$

Où,

$$G_k = \wedge_{j=1,\dots,n}\ max_{i=1,\dots,K_j}\ \mu^{i}{}_{A_i}(c_k) \tag{27}$$

Et,

$$\pi_{k,c} = \frac{\sum_{i=1,c_i=c}^{N} \mu_{G_k}(x_i)}{\sum_{i=1}^{N} \mu_{G_k}(x_i)} \tag{28}$$

Ainsi, étant donné une entrée x, les sorties du classifieur seront calculées selon la formule suivante :

$$\pi_c\,(x) = \frac{\sum_{i=1}^{p} \mu_{G_i}(x) * \pi_{i,c}}{\sum_{i=1}^{p} \mu_{G_i}(x)} \tag{29}$$

Pour $c = 1,..,C$. Si on doit attribuer une seule classe à x, on choisit la classe dont la valeur $\pi_c\,(x)$ est grande.

## 4.2 Structure granulaire multiniveaux

L'approche multi-niveaux[4] pour la granulation floue de l'information est effectué à l'aide du clustering successif . Ce dernier se base sur la technique du double clustering, qui est un outil servant à créer des granules interprétables à partir des données, facile à être étiqueté par des labels linguistiques. Cette approche est une extension de la technique du double clustering, elle a été mise en œuvre dans le but d'obtenir une description précise des données.

Le double clustering multi-niveaux exploite la structure extraite par le double clustering, pour donner une vue multiple des données, donc le résultat du premier niveau, sera considéré comme le contexte du deuxième niveau.

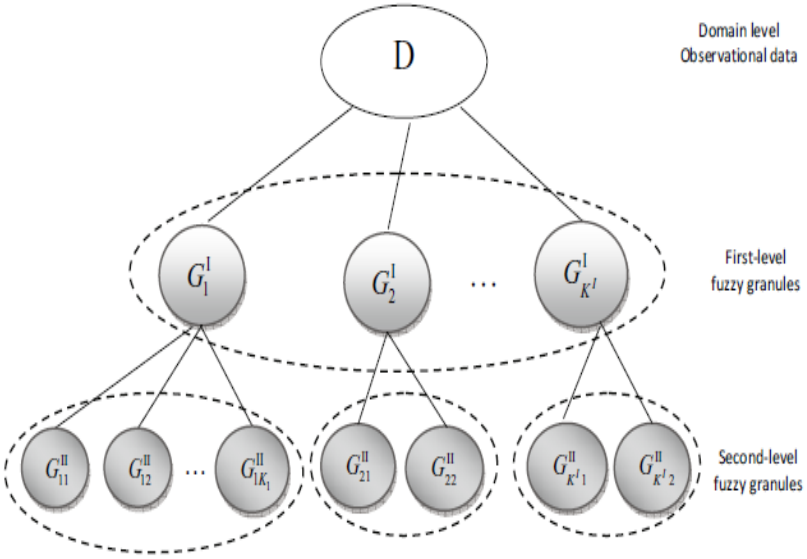



Figure 27: granulation multi-niveaux obtenue par le ML-DC

Le processus peut être répété pour plusieurs niveaux. Cependant une granulation à deux niveaux est adéquate pour obtenir deux vues du problème (une vue qualitative et autre quantitative), de manière à obtenir un compromis équilibré entre l'exactitude et l'interprétabilité des données. Les granules du premier niveau sont utilisés pour décrire les données par des labels linguistiques qualitatifs, tandis que les granules du deuxième niveau décrivent les granules du premier niveau.

La granulation du 2ème niveau est effectuée avec le même schéma du DC, mais on doit prendre en considération le contexte généré par chaque granule d'information du premier niveau. Sinon, si on ignore ce contexte la granulation du 2ème niveau sera identique à celle du premier niveau. Pour assurer cela, on utilise l'algorithme du CFCM (FCM conditionnel) qui est une extension de l'algorithme connu FCM, les deux algorithmes minimisent la même fonction objective et calculent les centres et la matrice d'appartenance de la même manière. La seule différence qui existe dans la contrainte :

$$\forall i = 1, \dots, N \ \sum_{j=1}^{c} \mu_{ij} = f_i \qquad (30)$$

Pour la granulation du deuxième niveau, le contexte est défini par chaque granule flou d'information par

$$\forall i = 1, \dots, N \ \ f_i = {\mu_k}^I(x_i) \qquad (31)$$

Où ${\mu_k}^I(x_i)$ désigne le degré d'appartenance de la i-ème observation au k-ème granule flou d'information découvert lors du processus de la granulation du premier niveau.

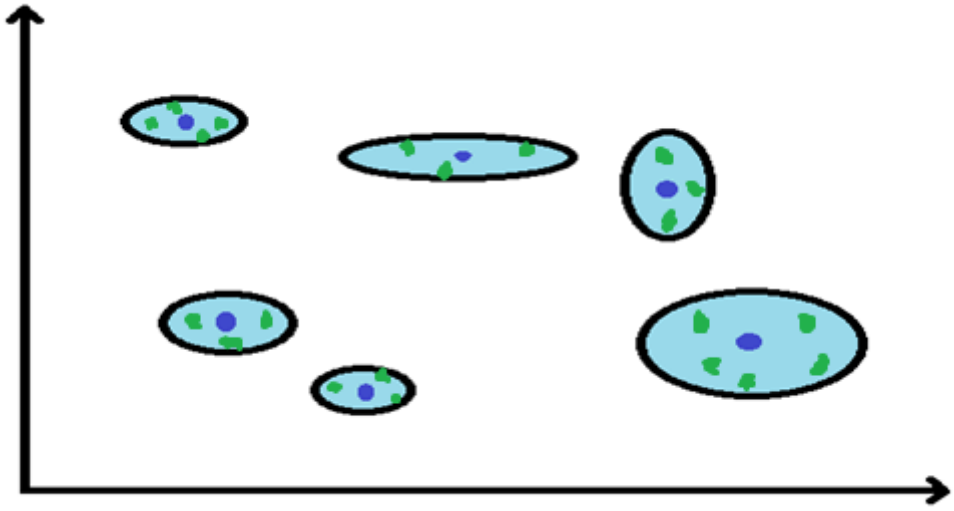


Figure 28 : prototypes obtenus de la granulation dus 2ème niveau selon le contexte de la granulation du premier niveau

Le classifieur SIF conçu via la granulation de l'information du premier niveau est très compact, mais il n'est pas précis. Les granules d'informations du deuxième niveau peuvent être exploités pour améliorer la précision de la classification. Ainsi, pour chaque granule d'information du premier niveau, ML-DC donne un ensemble de granules d'informations du deuxième niveau qu'on va utiliser pour générer un SIF correspondant qui possède le même schéma de celui du premier niveau, donc on aura comme résultat un ensemble de p SIF générés. Ces SIF sont interconnectés pour former un comité hiérarchique à partir duquel la correspondance entre les entrées et les sorties est déduite.

Les sorties du ce FIS hiérarchique sera définie comme la somme pondérée des sorties de chaque FIS appartenant à ce comité.

Formellement, étant donnée une entrée $x\ in\ \mathbb{R}^n$, La sortie du comité des FIS est calculée comme suit :

$$\pi_c^{II}(x) = \frac{\sum_{i=1}^{p} \mu_{G_i}(x) * \pi_{i,c}^{II}(x)}{\sum_{i=1}^{p} \mu_{G_i}(x)} \qquad (32)$$

Où $\pi_{i,c}^{II}(x)$ désigne la sortie relative à la classe c du FIS appartenant au comité dont le contexte est le granule $G_i$ . Cette sortie est calculée selon la formule (29). Pendant que $\mu_{G_i}(x)$est le poids affecté à l'ième FIS (c.à.d. dont le contexte est le granule $G_i$), il correspond au degré d'appartenance de l'entrée x au granule $G_i$ du premier niveau.

## 5 Conclusion :

L’approche multiniveaux décrite précédemment se promet d’être efficace, car elle permet la représentation du domaine, ainsi, elle donne la possibilité de se contenter d’un niveau de détail selon le besoin de la problématique. Pour déterminer la puissance d’un modèle basé sur cette approche, on va l’expérimenter sur des problèmes de la classification supervisée et le comparer avec des modèles connus

# Chapitre 4 : Expérimentations

# 1 Introduction

L'objectif de ce chapitre est d'analyser la classification supervisée à base du GrC. Nous allons expérimenter les deux types de structure granulaire; à un seul niveau et multi-niveaux, ainsi les expérimentations seront réalisées dans plusieurs bases de données issues de différents domaines. Les résultats obtenus sont comparés avec d'autres modèles de classifications. Dans la suite, on présente la configuration des expérimentations et l'analyse des résultats.

# 2 Configuration des expérimentations

## 2.1 Base de données

Nous avons utilisé un ensemble des données standards tiré à partir du répertoire de données KEEL [34], les données sélectionnées représente plusieurs domaines, donc différentes difficultés de classification. Le tableau 2 représente les propriétés principales de chaque data base : le nombre d'attribut (#Attribut), le nombre d'exemples (#Exemple) et le nombre de classe (#classe).

Chaque base et divisé en deux parties : 70% pour la phase d'apprentissage, pendant que 30% sera utilisé pour la phase de test.

Tableau 2 : Propriétés des bases de données utilisées dans la phase d'expérimentation

| Nom | #attribut | #Exemple | #classe |
|---|---|---|---|
| Appendicitis | 7 | 106 | 2 |
| Balance | 4 | 625 | 3 |
| Glass | 9 | 214 | 7 |
| Iris | 4 | 150 | 3 |
| Pima | 8 | 768 | 2 |

## 2.2 Algorithmes de comparaison

Pour bien situer notre travail, nous allons comparer le résultat de la classification à base du GrC avec d'autres algorithmes de classification supervisées. Pour cela on a sélectionné deux algorithmes très connus déjà appliqués dans plusieurs domaines :

- Arbre de décision (AD) [35]
- Réseau de neurone (RN) [36], [37]

## 2.3 Paramètre du modèle de classification

On a expérimenté deux structures granulaires, la première avec un seul niveau (N1) de granulation, quant à la deuxième, elle est multiniveaux (N2), ainsi les deux approches sont réalisées par la méthode de double clustering. Les deux étapes du clustering sont réalisées par :

- L'algorithme CFCM,
- Clustering hiérarchique.

Les deux clustering sont exécutés avec divers nombres de clusters dans l'intervalle [2,9], après le résultat est validé par l'indice Xie-Beni.

# 3 Résultats

## 3.1 Résultat de la granulation d'information

Les figures (29) à (33) représentent les prototypes découverts pour les deux structures granulaires, associés à chaque base de données utilisée dans l'expérimentation, les figures montrent que la méthode de granulation préserve la structure de la base et couvre aussi tout l'espace de données.

Le tableau (3) est un exemple de la représentation des granules découverte lors de la granulation du premier niveau, par des règles floues.

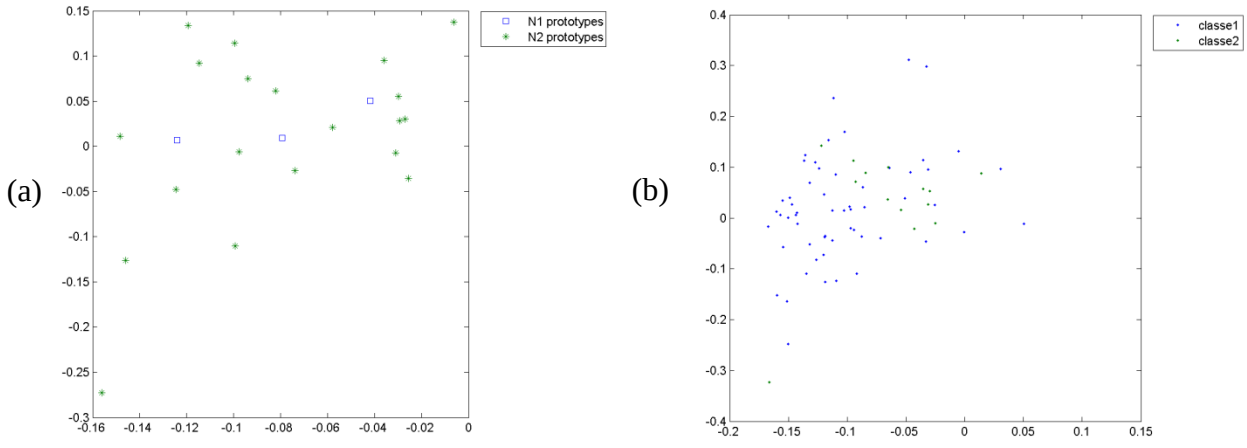



Figure 29: (a) résultat de la granulation d'information pour les deux niveaux (b) structure originale de la base Appendicites

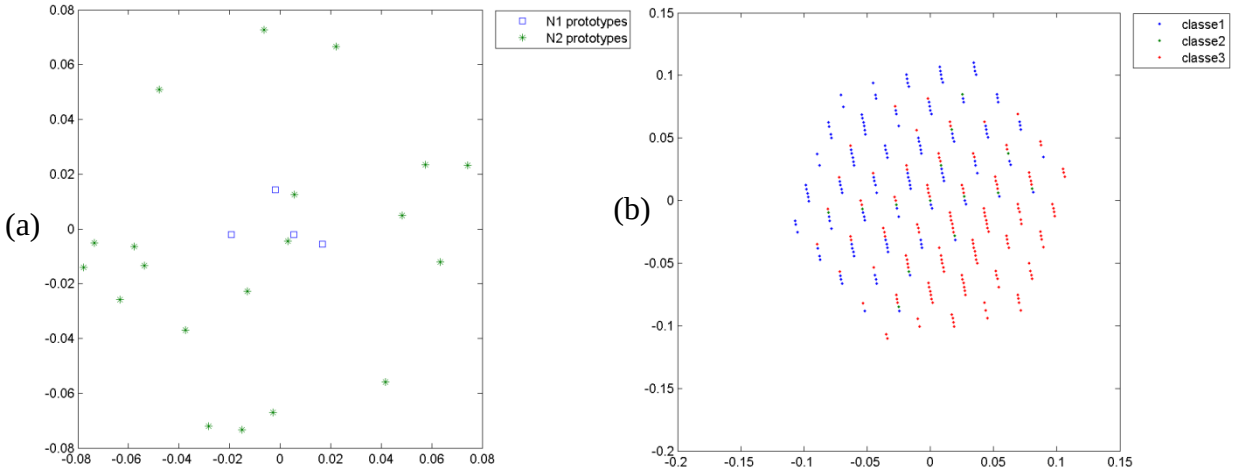



Figure 30: (a) résultat de la granulation d'information pour les deux niveaux (b) structure originale de la base Balance

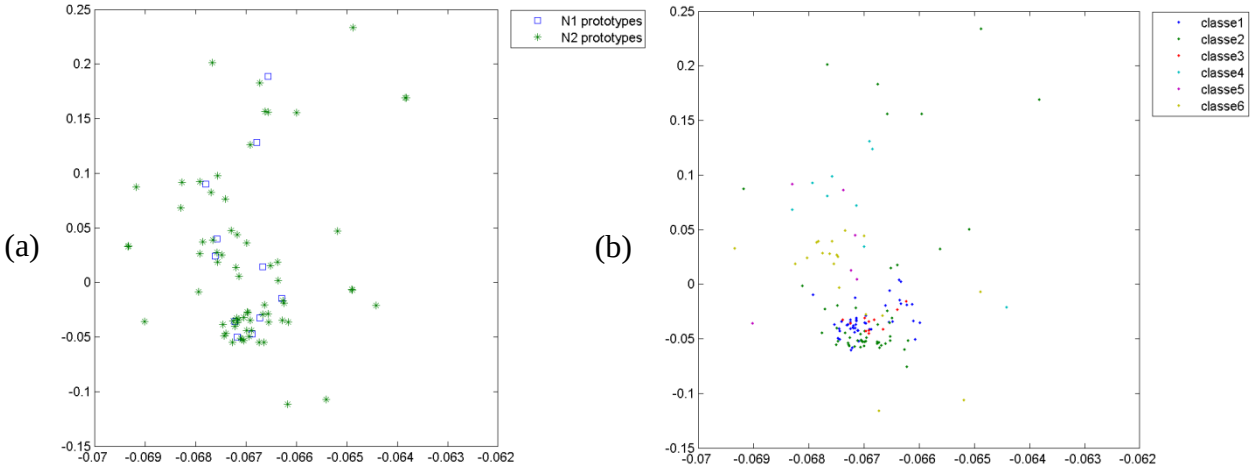



Figure 31: (a) résultat de la granulation d'information pour les deux niveaux (b) structure originale de la base Glass

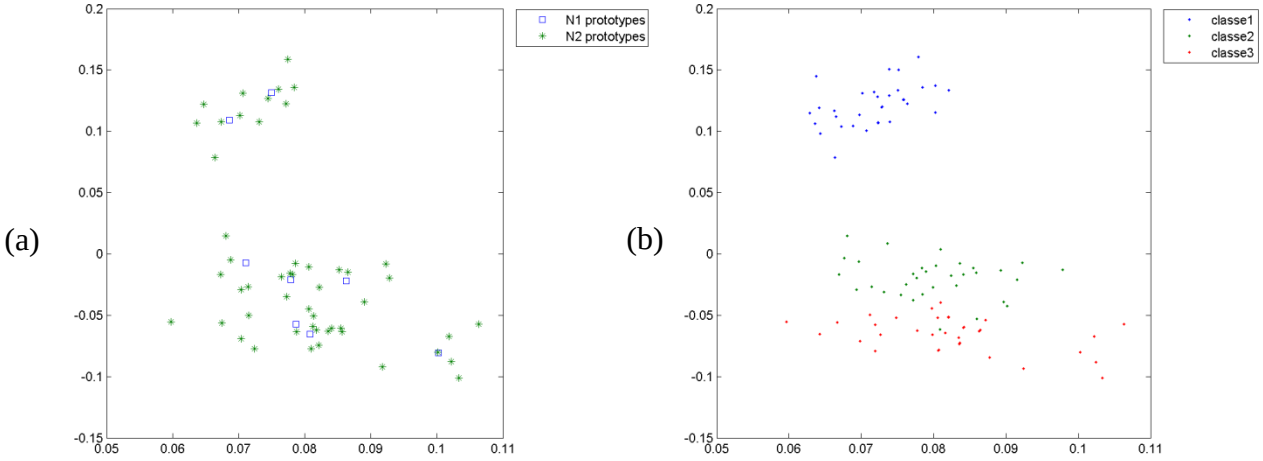



Figure 32:(a) résultat de la granulation d'information pour les deux niveaux (b) structure originale de la base IRIS

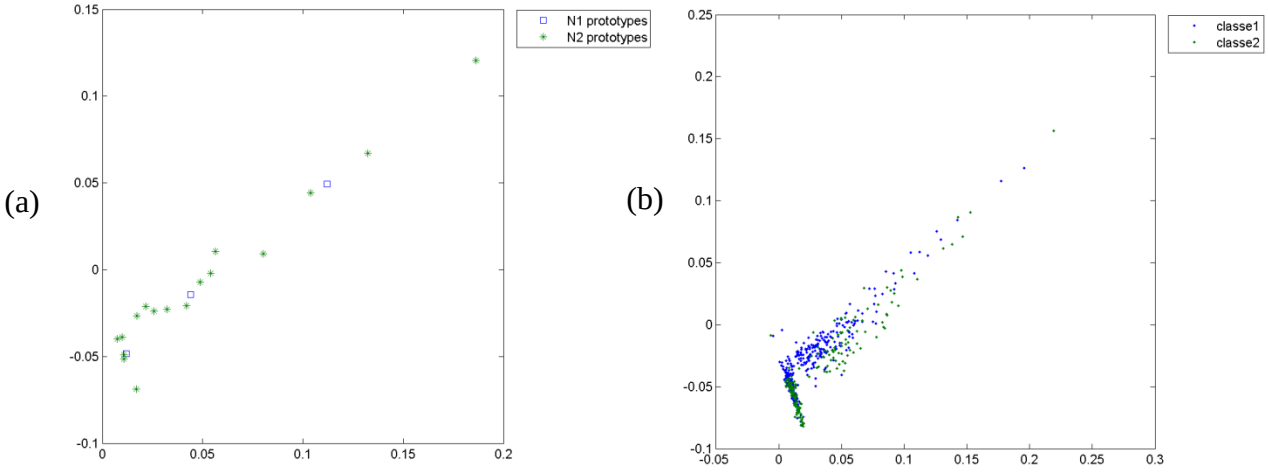



Figure 33: résultat de la granulation d'information pour les deux niveaux (b) structure originale de la base Pima

Tableau 3: représentation linguistiques des granules obtenus dans le niveau 1 pour la base d'Iris

| Si | | | | Alors | | |
|---|---|---|---|---|---|---|
| Longueur du sépale | Largeur du sépale | Longueur du pétale | Largeur du pétale | P(Iris Setosa) = | P(Iris-versicolor) = | P(iris-virginica) = |
| Très petite | moyenne | Très petite | Petite | 100% | 0% | 0% |
| petite | grande | Très petite | Petite | 99.16% | 0.84% | 0% |
| petite | petite | petite | Moyenne | 1.55%% | 96.72% | 1.73%% |
| moyenne | moyenne | moyenne | Moyenne | 0.13%% | 75.42% | 24.45%% |
| grande | moyenne | Très grande | Grande | 0% | 8.16% | 91.84% |
| moyenne | moyenne | grande | Grande | 0% | 25.13% | 74.87% |

## 3.2 Résultat de la classification à base GrC

Le tableau (4) représente la précision de la classification obtenue dans les deux structures granulaires (N1 et N2), les résultats montrent que N2 donne un taux de

classification plus élevé que celui de N1 dans les bases Appendicitis, Balance, Glass et iris. Par contre dans le cas de la base Pima N1 donne le meilleur taux, ce qui signifie qu'il n'est pas toujours évident de descendre vers un niveau bas d'abstraction pour obtenir un résultat optimale.

Tableau 4: Taux de précision de classification pour le modèle d'un seul niveau et multi niveaux

| | N1 (%) | N2(%) |
|---|---|---|
| **Appendicitis** | 88,88 | 91,66 |
| **Balance** | 55,57 | 74,03 |
| **Glass** | 50 | 61,11 |
| **Iris** | 94 | 98 |
| **Pima** | 66,01 | 65,23 |

## 3.3 Comparaison

Les résultats obtenus par la classification à base de GrC sont comparés avec celle du RN et du AD (Tableau 5). La comparaison des moyennes des taux de précision montre que la méthode à base de GrC est très performante que la classification à base de l'AD, tandis que la classification à base de RN donne une meilleur performance que les deux. En effet dans la littérature [11], le RN est utilisé dans les domaines où l'interprétabilité est négligeable, alors que l'AD et utilisé essentiellement dans le but de concevoir un modèle interprétable. D'ailleurs la classification à base de GrC peut être une meilleur alternative que l'AD, ainsi il peut être plus performant que la RN comme le montre le tableau (5), le GrC est efficient que RN dans les bases Appendicitis et Iris.

Tableau 5: Comparaison des taux de précision de classification par les modèles : MLDC, AD et RN

| | RN(%) | AD(%) | GrC(%) |
|---|---|---|---|
| **Appendicitis** | 88,89 | 80,56 | 91,66 |
| **Balance** | 90,38 | 75,48 | 74,03 |
| **Glass** | 69,01 | 59,15 | 61,11 |
| **Iris** | 98 | 94 | 98 |
| **Pima** | 73,83 | 72,66 | 66.01 |
| **Moyenne** | 83,62 | 77,17 | 78,16 |

# 4 Conclusion

Les résultats expérimentaux ont montré que le GrC flou est un outil puissant en terme de représentation des connaissances, ainsi qu'il peut être un bon concurrent pour les modèles qui visé seulement la précision, qu'on peut améliorer si on arrive à définir le niveau d'abstraction convenable pour la résolution des problèmes.

# Conclusion

Dans ce travail, on a présenté les différents problèmes que rencontrent la fouille de données dans divers domaines, on a mis l'accent sur l'adaptation avec le domaine, l'interpérabilité, l'imprécision et l'incertitude. A cet égard, on a développé un modèle basé sur la logique floue pour permettre la tolérance aux imprécisions et la gestion des incertitudes pendant le raisonnement, aussi l'approche GrC qui permet la représentation du domaine dans une structure granulaire interprétable et fournit une méthode de résolution de problème avec une multitude de niveau d'abstraction.

Le modèle développé a été implémenté dans la problématique de classification supervisée, dans cette implémentation nous avons profité des méthodes de clustering floue pour la construction d'un modèle de classification supervisée à base de GrC, ainsi le modèle résultant a été expérimenté dans plusieurs domaines réel et a été comparé avec deux algorithmes très utilisés dans la fouille de données à savoir les arbres de décision et les réseaux de neurones. Le résultat d'expérimentation montre que:

- La granulation d'information et la représentation du domaine dans une structure granulaire représentent les variations dans le domaine traité, ainsi elles fournissent une structure hiérarchique interprétable représentée par des règles floues.
- Les taux de classification obtenus montrent que le modèle développé peut être une alternative de l'algorithme des arbres de décision et aussi du réseau de neurones dans certains domaines.

Ce travail révèle plusieurs perspectives à savoir :

- Création des méthodes de clustering adaptés au GrC.
- Recherche d'une méthodologie pour définir le niveau d'abstraction convenable pour résoudre un problème.

# Références


[1] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2011.

[2] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques." 2007.

[3] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[4] G. Castellano, A. M. Fanelli, and C. Mencar, "Fuzzy Information Granulation with Multiple Levels of Granularity," in *Granular Computing and Intelligent Systems*, Springer, 2011, pp. 185–202.

[5] E. Hüllermeier, "Fuzzy methods in machine learning and data mining: Status and prospects," *Fuzzy sets Syst.*, vol. 156, no. 3, pp. 387–406, 2005.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[7] "Exploration visuelle de données en R." [Online]. Available: http://www.grappa.univ-lille3.fr/~ppreux/ensg/miashs/fouilleDeDonneesII/tp/exploration-visuelle/.

[8] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[9] J. A. Hartigan, *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[10] "Ruminating on Decision Trees." [Online]. Available: http://www.narendranaidu.com/2014/02/ruminating-on-decision-trees.html.

[11] S. P. Curram and J. Mingers, "Neural networks, decision tree induction and discriminant analysis: An empirical comparison," *J. Oper. Res. Soc.*, pp. 440–450, 1994.

[12] "Les réseaux de neurones." [Online]. Available: http://mp.cpgedupuydelome.fr/document.php?doc=Article - Les r?seaux de neurones.txt.

[13] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[14] "Statistics/Data Mining Applications." [Online]. Available: http://stat-mzhong.blogspot.com/2012/09/python-linear-discriminant-analysis.html.

[15] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast Discovery of Association Rules.," *Adv. Knowl. Discov. data Min.*, vol. 12, no. 1, pp. 307–328, 1996.

[16] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *Neural Networks, IEEE Trans.*, vol. 16, no. 3, pp. 645–678, 2005.

[17] J. Grabmeier and A. Rudolph, “Techniques of cluster algorithms in data mining,” *Data Min. Knowl. Discov.*, vol. 6, no. 4, pp. 303–360, 2002.

[18] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[19] L. A. Zadeh, “Fuzzy sets,” *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[20] N. K. Kasabov and Q. Song, “DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction,” *Fuzzy Syst. IEEE Trans.*, vol. 10, no. 2, pp. 144–154, 2002.

[21] E. D. Lughofer, “FLEXFIS: a robust incremental learning approach for evolving Takagi–Sugeno fuzzy models,” *Fuzzy Syst. IEEE Trans.*, vol. 16, no. 6, pp. 1393–1410, 2008.

[22] P. P. Angelov and D. P. Filev, “An approach to online identification of Takagi-Sugeno fuzzy models,” *Syst. Man, Cybern. Part B Cybern. IEEE Trans.*, vol. 34, no. 1, pp. 484–498, 2004.

[23] W. Pedrycz, *Granular computing: an emerging paradigm*, vol. 70. Springer Science & Business Media, 2001.

[24] A. Bargiela and W. Pedrycz, *Granular computing: an introduction*, vol. 717. Springer Science & Business Media, 2012.

[25] L. A. Zadeh, *Computing with words in Information/Intelligent systems 1: Foundations*, vol. 33. Physica, 2013.

[26] Y. Yao and N. Zhong, “Granular computing,” *Wiley Encycl. Comput. Sci. Eng.*, 2008.

[27] P. Berkhin, “A survey of clustering data mining techniques,” in *Grouping multidimensional data*, Springer, 2006, pp. 25–71.

[28] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Appl. Stat.*, pp. 100–108, 1979.

[29] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Comput. Geosci.*, vol. 10, no. 2, pp. 191–203, 1984.

[30] E. Rendón, I. M. Abundez, C. Gutierrez, S. D. Zagal, A. Arizmendi, E. M. Quiroz, and H. E. Arzate, “A comparison of internal and external cluster validation indexes,” in *Proceedings of the 2011 American Conference, San Francisco, CA, USA*, 2011, vol. 29.

[31] X. L. Xie and G. Beni, “A validity measure for fuzzy clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 8, pp. 841–847, 1991.

[32] J.-S. R. Jang, “ANFIS: Adaptive-Network-Based Fuzzy Inference System Jyh-Shing Roger Jang Department of Electrical Engineering and Computer Science University of California, Berkeley, CA 94720.”

[33] G. Castellano, A. M. Fanelli, and C. Mencar, “DCf: a double clustering framework for fuzzy information granulation,” in *Granular Computing, 2005 IEEE International Conference on*, 2005, vol. 2, pp. 397–400.

[34] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, “Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2010.

[35] W. Du and Z. Zhan, “Building decision tree classifier on private data,” in *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, 2002, pp. 1–8.

[36] M. F. Møller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural networks*, vol. 6, no. 4, pp. 525–533, 1993.

[37] A. Hart, “Using Neural Networks for Classification Tasks--Some Experiments on Datasets and Practical Advice,” *J. Oper. Res. Soc.*, pp. 215–226, 1992.