UNIVERSITE DE SAAD DAHLEB DE BLIDA

Faculté des Sciences

Département d'informatique

MEMOIRE DE MAGISTER

Spécialité: Ingénierie des systèmes et de la connaissance

UTILISATION DES TECHNIQUES DE DATA MINING POUR LA MODELISATION DU PARCOURS SCOLAIRE ET LA PREDICTION DU SUCCES ET DU RISQUE D'ECHEC

Par

HAMMOUDA Mohamed

Devant le jury composé de

M. A. GUESSOUM	Professeur, U. de Blida	Président
M. M. KOUDIL	Professeur, INI Oued Smar Alger	Examinateur
Mme H. ABED	Maître de Conférence, U de Blida	Examinateur
Mme S. OUKID	Maître de Conférence, U de Blida	Rapporteur
Mme N. BENBLIDIA	Maître de Conférence, U de Blida	Co-Rapporteur

Blida, 2009

RESUME

Le data mining, ou La fouille de données, constitue le cœur d'un processus d'extraction des connaissances à partir d'un large volume de données. Son spectre d'applications s'élargit de plus en plus, mais il est relativement récent dans le domaine de l'éducation.

Dans le présent travail nous exposons un modèle prédictif du succès et de l'échec d'un parcours scolaire d'une population estudiantine. Pour atteindre cet objectif nous proposons une procédure qui va s'articuler autour des trois étapes suivantes:

- Dans un premier temps nous employons la technique du clustering pour structurer la population en groupes. Pour ce faire, nous proposons dans ce mémoire une version améliorée de l'algorithme k-means que nous baptisons t-means.
- Dans un deuxième temps nous utilisons un réseau bayésien pour modéliser le parcours éventuel de chaque groupe vers une situation de succès ou une situation d'échec.
- Par la suite nous exposons une procédure de caractérisation des groupes dans le but de saisir les facteurs déterminants qui mènent à une situation d'échec.

Enfin, le travail que nous défendons dans ce mémoire présente des intérêts forts intéressants notamment dans l'explication de l'échec scolaire, qui devient de plus en plus un phénomène très préoccupant, et dans la contribution du développement d'un système d'évaluation pédagogique avancé.

Mots clés: extraction de connaissances, fouille de données et éducation, clustering, réseaux bayésiens, caractérisation et description des classes, évaluation pédagogique.

Abstract

The data mining is the heart of a process of knowledge extraction from a large volume of data. Its spectrum of applications is widening more and more, but it is relatively new in the field of education.

This paper presents a predict model of a success and a failure of a school path of a student population. In order to achieve that objective we propose a methodology that is articulated around the three following phases:

- First of all we use the technique of clustering in order to structure a student population into sub-populations (groups). To do this, we propose an improved version of the algorithm k-means that we name t-means.
- In the second phase we use a Bayesian networks to model the school path of each sub-populations (group) that have a situation of success or a situation of failure.
- Thereafter we outline a procedure for the characterization of groups in order to grasp the factors that lead to a situation of failure.

Finally, the work that we defend in this paper presents strong interests in the explanation of school failure, which is increasingly becoming a very worrying phenomenon, and contributes to the development of an advanced educational assessment system.

Keywords: knowledge extraction, data mining and education, educational clustering, Bayesian networks, characterization and description of classes, educational assessment.

ملخص

"داتا ماينغ" أو التنقيب على البيانات تعتبر الأساس في عملية استخراج المعرفه من كم هائل من البيانات. مجالات تطبيقها تتسع يوم بعد يوم ، الا انها جديدة نسبيا في مجال التربية والتعليم.

هذا البحث يعرض نموذج تنبؤي لنجاح أو فشل مسار مدرسي لشريحة طلابية . ومن أجل تحقيق هذا الهدف نقترح منهجية تدور حول المراحل الثلاث التالية :

- المرحلة الاولى نستعمل تقنية التجميع (clustering) من أجل تقسيم شريحة الطلبة إلى مجموعات. ولتحديد ذلك، نقترح صيغة محسنه من الخوارزميه ك – مينز (k-means) والتي لقبناها بـ تي- مينز (t-means).
- اما في المرحلة الثانية نستخدم أسلوب شبكات بايزيه (Bayesian networks) من اجل نمذجة المسار المدرسي لكل مسار مجموعة ناجحة و مسار مجموعة فاشلة.
- بعد ذلك، نعرض طريقة لوصف المجموعات من اجل فهم العوامل التي تؤدي للفشل.

وأخيرا فالبحث الذي ندافع عنه يعتبر هاما جدا لأنه من جهة يقدم تفسيرا للفشل المدرسي الذي أصبح وعلى نحو متزايد ظاهرة مقلقة ومن جهة اخرى يساهم في تطوير نظام التقييم المدرسي.

الكلمات الرئيسية : استخراج المعرفه ، التنقيب على البيانات والتربية والتعليم، المجموعات، شبكات بايزيه ، وصف الفصول والأقسام، تقييم التعليم.

REMERCIEMENTS

*Dans un premier temps je tiens à remercier **Madame Oukid Khouas**, maître de conférence à l'Université Saad Dahlab de Blida (USDB), pour avoir accepté la direction de ce mémoire et pour les entrevues enrichissantes ayant traits à la partie interprétation des résultats.*

*Mes sincères remerciements à **Madame Benblidia Nadjet**, maître de conférence à l'USDB, d'avoir accepté de rapporter ce mémoire, pour sa disponibilité, ses commentaires, ses conseils, son temps consacré à la lecture de ce manuscrit et ses remarques qui m'ont considérablement permis à mener à terme ce travail.*

*Mes vifs remerciements à **Monsieur Hadj Yahia**, maître assistant à l'USDB, pour le temps consacré à la lecture de ce mémoire, et pour les suggestions et les remarques judicieuses qu'il m'a indiquées.*

*Je tiens aussi à remercier vivement tous mes collègues du département informatique de l'université de Saad DAHLAB de Blida pour leurs encouragements et l'intérêt qu'ils ont apportés à mon travail. Je tiens à exprimer tout particulièrement ma reconnaissance à **Mr Ait Akkache** pour m'avoir aidé à appréhender certaines notions de statistiques, Mr **Djamel Bennouar, Madame Ouahrani, Madame Bensititi, Mr Bala Mahfoud** pour tous les entretiens intéressants concernant les aboutissements de ce travail.*

*Mes remerciements vont aussi à **Mr Massiad** Chef du département informatique de l'USDB pour avoir accepté de mettre à ma disposition la base de données du service de scolarité du département informatique.*

*Je voudrais aussi exprimer ma sincère gratitude à **Monsieur Guessoum Abderrazak** professeur à l'USDB, **Mr Koudil Mouloud** professeur à l'Institut National d'Informatique (INI) et **Madame Abed Hafidha** maître de conférence à l'USDB pour l'honneur qu'ils m'ont fait en acceptant d'être membres de mon jury de mémoire.*

Enfin, j'en profite pour saluer toutes les personnes que j'ai côtoyées ainsi que ma famille, mes proches et mes amis pour leur soutien permanent.

TABLE DES MATIERES

LISTE DES FIGURES ET TABLEAUX

Liste des figures

Liste des tableaux

INTRODUCTION

Préambule

Les préoccupations actuelles concernant l'économie du savoir ont permis de mettre en exergue l'importance du capital humain et par-la même de l'éducation. En effet, l'éducation représente un investissement dans les qualifications qui peuvent contribuer à promouvoir la croissance économique et à accroître la productivité [10]. Un système d'éducation médiocre représentera par conséquent une contrainte en terme de production de main-d'œuvre hautement qualifiée.

Le rôle d'un système éducatif est de transmettre des savoirs et des compétences. Pour accomplir cette tâche, les enseignants ont besoin de repères pour aider l'élève dans ses apprentissages, comme l'élève a besoin de repères, de guides, pour apprendre à s'approprier ce qu'il apprend; pour qu'il puisse mesurer ses progrès, anticiper, pointer les domaines précis à retravailler, se connaître afin de progresser [11]. Il y a donc nécessité de développer des mécanismes d'évaluation qui vont permettre d'apprécier dans quelle mesure et jusqu'à quel point les objectifs assignés à un système éducatif ont été atteints.

L'évaluation pédagogique fait partie intégrante du processus éducatif, dont voici quelques définitions données dans [12]:

- l'évaluation est l'estimation par une note d'une modalité ou d'un critère dans un comportement ou un produit. La notion d'évaluation a donc une acceptation plus large que celle de mesure, celle-ci est en effet une simple description quantitative, alors que l'évaluation comporte à la fois la description qualitative

des comportements, mais également des jugements de valeur concernant leur désirabilité.

- L'évaluation pédagogique peut être définie comme le processus systématique visant à déterminer dans quelle mesure des objectifs éducatifs sont atteints par des élèves.

- L'évaluation est le processus qui consiste à décrire, recueillir et fournir des informations utiles pour porter un jugement décisif en fonction de diverses possibilités.

- La démarche de l'évaluation consiste à se donner des objectifs, à opérationnaliser et à définir les moyens appropriés qui permettront de déterminer si les objectifs sont atteints par les élèves. Il s'agira ensuite de procéder à une analyse des résultats; une analyse qui conduira à une prise de décision qui devra être communiquée aux différents intéressés.

L'évaluation pédagogique peut prendre plusieurs formes, à savoir:

- l'évaluation sommative: c'est le contexte le plus visible, elle sert à mesurer les acquis des élèves par des tests et des examens. Il s'agit donc d'un bilan qui permet d'aboutir à une sanction de réussite ou de classement.

- l'évaluation formative: c'est une évaluation au cours de l'enseignement et ayant pour objet d'informer du degré de progrès des élèves et des difficultés rencontrées; en vue d'ajuster les méthodes pédagogiques.

- l'évaluation diagnostique: c'est un bilan des acquis, elle peut être au début d'une formation sur les connaissances pré requises pour tracer par exemple un programme d'enseignement répondant aux besoins; comme elle peut être à la fin d'une étape de formation pour donner par exemple des avis d'orientation.

- l'évaluation pronostique: elle a pour objet de prévoir la réussite d'un élève dans une formation. Elle peut être au début d'une formation pour estimer les chances de réussite d'un élève à partir de ses connaissances pré-requises,

comme elle peut être aussi à la fin pour pronostiquer les chances de réussite dans des formations ultérieures.

Actuellement, l'évaluation pédagogique dans l'université algérienne est principalement sommative. Elle prend la forme d'un bilan à la fin d'une période d'étude qui peut être un semestre ou une année. Ce bilan sert à établir un classement éventuel des étudiants comme il peut servir aussi de décider si un tel étudiant est digne de tel grade ou s'il peut accéder à un niveau supérieur.

En plus de cette évaluation sommative, les bilans sont utilisés pour réaliser une évaluation diagnostique au début ou à la fin d'une formation dont le but principal est d'accepter ou de refuser le choix d'un étudiant de s'inscrire dans une filière donnée.

En terme d'analyse pédagogique, nous remarquons qu'elle se limite essentiellement à de simples descriptions quantitatives et des bilans statistiques.

Travaux de mémoire

L'échec universitaire est devenu un phénomène très préoccupant qui touche de plus en plus un grand nombre d'étudiants [28]. De ce fait, une lutte efficace contre ce fléau devient une priorité. Cela nécessite fondamentalement le développement des mécanismes d'évaluation pédagogique qui vont permettre de cerner les facteurs ayant contribué à son émergence.

Les méthodes ainsi que les outils d'évaluation pédagogique ne font pas l'optique du présent travail que nous l'inscrivons comme étant une contribution dans l'étude du phénomène d'échec. Notre recherche sera axée principalement sur la proposition d'une modélisation du parcours scolaire permettant de faire des prédictions sur les chances de réussite et les risques d'échec. Nous montrons à travers une simulation du modèle prédictif proposé comment on peut mettre en valeur les facteurs influents du cas d'échec.

Pour atteindre cet objectif, nous employons les techniques de data mining dans l'exploration des données d'une population estudiantine; ces données sont

issues de l'évaluation sommative. La démarche globale va s'articuler autour de trois étapes principales.

Dans la première étape, nous utilisons la technique du clustering dans le but de structurer la population d'étude en un certain nombre de classes ou groupes (appelés aussi clusters). Chaque classe représente une tranche de la population d'étude ayant un certain degré d'homogénéité et dont les membres partagent des propriétés intéressantes. Pour réaliser le clustering, nous proposons dans ce mémoire une version améliorée d'un l'algorithme très largement utilisé le k-means.

Le k-means est un algorithme très populaire, simple, rapide et s'adapte parfaitement aux cas des bases de données larges, mais malgré tous ses atouts, il souffre d'un inconvénient majeur qui est la très forte sensibilité aux choix des paramètres initiaux. Ceci a motivé beaucoup de ses utilisateurs à mettre en œuvre un ensemble de variantes, et c'est dans cette suite que s'inscrit notre contribution en introduisant une nouvelle version que nous baptisons t-means.

L'idée générale de t-means consiste en la définition d'un seuil de regroupement comme étant le seuil paramètre initial et laisse le soin à l'algorithme de structurer d'une manière automatique et itérative la population initiale.

La seconde étape, utilise un réseau bayésien construit sur la base des classes obtenues précédemment dans le but de modéliser le parcours scolaire. Chaque chemin dans le réseau exprime, suivant une certaine valeur de probabilité, le parcours éventuel d'une population d'une classe donnée vers une situation de succès ou une situation d'échec.

La troisième étape, est une tâche de caractérisation des classes. Elle consiste en l'étude des caractéristiques générales de chaque classe dans le but de saisir les facteurs déterminants qui mènent à une situation d'échec.

Le résultat final du travail est un modèle prédictif qui peut avoir des intérêts forts intéressants tels que:

- pronostiquer les chances de réussite et les risques d'échec,

- cerner les paramètres influents du succès et de l'échec.
- offrir à l'étudiant la possibilité de s'auto évaluer,
- suivre le comportement pédagogique des étudiants et intervenir au moment opportun et au cas de besoin.

Organisation du document:

Après avoir introduit les éléments nécessaires pour la lecture du document, la suite s'organise de la manière suivante:

Dans le premier chapitre nous introduisons les concepts de base du data mining et de la classification supervisée et non supervisée.

Dans le deuxième chapitre nous montrons par des exemples de quelques travaux l'apport du data mining dans les milieux éducatifs.

Le troisième chapitre est un état de l'art sur les méthodes de clustering couramment utilisées. Nous adoptons dans cet état de l'art une taxonomie classique basée sur le principe de partitionnement.

Les trois premiers chapitres ont permis de prendre connaissance de la discipline du data mining et de sa complexité. Le quatrième chapitre est consacré à la modélisation. Nous commençons dans un premier temps par expliquer la méthodologie adoptée, suivi d'un schéma détaillé du principe de fonctionnement de l'algorithme proposé t-means, que nous appuyons par un test sur un jeu de données synthétiques conçu de telle sorte que la classification optimale est connue et un jeu de données réelles. Pour valider t-means, nous confrontons ses résultats avec ceux de la version standard k-means et un algorithme génétique. Pour la simulation du modèle prédictif nous utilisons un cas réel concernant une population estudiantine du département informatique de l'université SAAD DAHLAB de Blida. Ce cas consiste en l'étude du succès et de l'échec d'un échantillon d'étudiants de la deuxième année. L'interprétation des résultats de la simulation va nous permettre de déduire des connaissances d'ordre pédagogique que nous jugeons très intéressantes.

En guise de conclusion, nous présentons l'intérêt et les limites de notre travail. Ces considérations pourront inspirer les perspectives des travaux futurs dans la continuation de ceux présentés ici.

CHAPITRE 1
CONCEPTS ET GÉNÉRALITÉS SUR LE DATA MINING ET LE CLUSTERING

Ce chapitre est une présentation des concepts fondamentaux, il va nous permettre de prendre connaissance de la discipline et de sa complexité ; il s'articule autour des deux points suivants :

- Le data mining dont la traduction littérale correspond à "forage de données" mais la plus couramment utilisée est "fouille de données",
- et le clustering, connu sous l'appellation de segmentation ou classification automatique non supervisée

1.1 Le data mining

1.1.1 Introduction

Les bases de données d'aujourd'hui sont en croissance continue. C'est une conséquence directe des moyens qu'offre l'informatique en termes de stockage et de puissance avec des coûts de plus en plus faibles, ce qui a motivé les entreprises à accumuler toujours plus de données.

Cette croissance a enrichi les bases de données en terme de données mais les a appauvries en terme d'informations. L'institut EDS (Environnement, développement et société de Québec) estime que la quantité de données collectées dans le monde double tous les 20 mois alors que le volume d'informations fournies aux utilisateurs n'augmente lui que très peu [5].

L'analyse de ces grandes masses de données par les statistiques s'avère être limitée car elle ne permette d'étudier simultanément que quelques variables (une à deux) mais les problèmes sont en réalité plus complexes et mettent en œuvre plusieurs dizaines de variables. Pour répondre à ce besoin, il a fallu mettre en œuvre

de nouveaux algorithmes, parfois issus de la recherche opérationnelle, alliant la recherche intelligente et les statistiques [5].

Le processus d'analyse des bases de données volumineuses est connu dans la littérature sous l'appellation de data mining (DM) dont la traduction la plus couramment utilisée est fouille de données. Ce dernier constitue l'étape essentielle d'un processus d'extraction des connaissances (ECD ou KDD pour dire knowledge discovery in databases) comme le montre la figure 1.1.

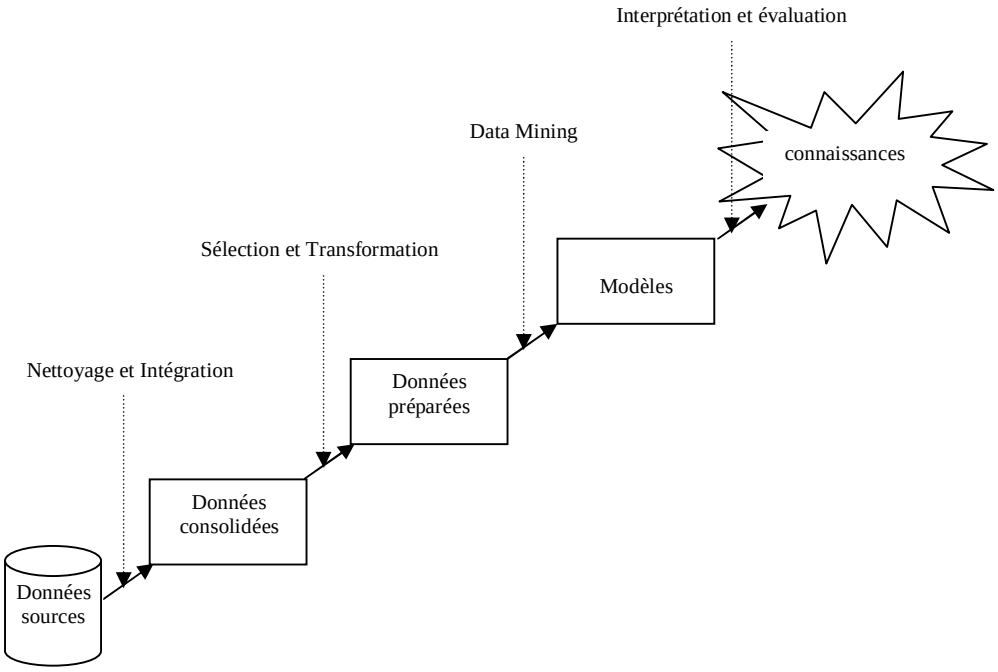


Figure 1.1 : Les étapes d'un processus ECD [6]

1.1.2 Etapes du processus ECD

- le nettoyage des données : les données brutes peuvent contenir des anomalies qui peuvent altérer la suite du processus d'où l'utilité de leur nettoyage : suppression des bruits, correction des données erronées, traitement des valeurs manquantes, …

- l'intégration des données : l'intégration s'impose au cas où les données proviennent de différentes sources et peuvent être aussi avec des formats différents.

- la sélection des données : plusieurs traitements peuvent être utiles dans cette étape, tels que :
    - ignorer certaines variables trop corrélées,
    - ignorer certaines variables absolument non pertinentes ou non discriminantes par rapport à l'objectif à atteindre, au phénomène à détecter,
    - rassembler plusieurs variables en une seule,
    - réduire le nombre de variable au moyen de l'analyse factorielle

- la transformation des données : à cette étape les données sont transformées dans des formes appropriées aux algorithmes de data mining qui vont être employés dans la suite du processus, telles que:

    - remplacer les grandeurs absolues par des pourcentages
    - calculer des ratios qui parfois permettent de diminuer le nombre de variables
    - recoder certaines variables, par exemple "faible, moyen, fort" transformés en "1, 2, 3"
    - remplacer les dates par des durées
    - remplacer les lieux géographiques par des coordonnées
    - discrétiser certaines variables continues (c'est à dire les transformer en tranches de valeurs).

- le data mining : c'est l'étape la plus importante du processus d'ECD, elle consiste en l'application des techniques de data mining dans le but d'extraire les modèles cachés dans les données.

- interprétation et évaluation des modèles : Dans cette étape s'effectue l'interprétation des résultats et la comparaison des modèles. Cette comparaison peut se faire de plusieurs manières entre autres :

-
    - Sélectionner le meilleur modèle sur la base du taux d'erreur.
    - Réitérer le processus en réajustant les données ou en combinant d'autres techniques de DM.

- visualisation des résultats : pour faciliter l'exploitation des résultats ils sont souvent représentés sous des formes graphiques conviviales (histogrammes, courbes, …). Les connaissances inférées par le processus peuvent être enfin intégrées dans un système décisionnel : prise de décisions, contrôle de processus, …

Les différentes étapes d'un processus ECD citées précédemment peuvent être regroupées en trois phases comme le montre la figure 1.2 :

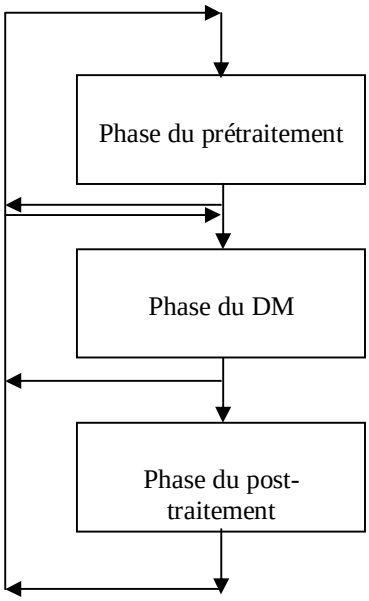


Figure 1.2: résumé du processus ECD

1.1.3 Le data mining (DM)

Le DM est un domaine multidisciplinaire. Il fait appel aux bases de données, l'intelligence artificielle, l'apprentissage automatique, les statistiques, …Il peut être vu comme une suite de l'évolution de la technologie des bases de données ; il est généralement employé pour désigner l'ensemble des outils permettant de générer des informations riches à partir d'un ensemble de données de volume important et

d'en découvrir les modèles qui se cachent. Pour le décrire nous employons la définition donnée dans [8] :

- Le DM est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données (souvent larges) informatiques de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données. [8]

- en bref le DM est l'art d'extraire des informations (ou même des connaissances) à partir des données.

La différence avec les outils d'exploration des données et les outils d'aide à la décision c'est que ces derniers cherchent à présenter les données sous des formes structurées et homogènes (comme les cubes) et laissent l'initiative à l'utilisateur d'observer et d'analyser les éléments qui l'intéressent alors que le DM prend l'initiative des observations et des analyses comme par exemple projeter l'avenir à partir des données historiques. Il reste, par la suite, aux experts d'observer la pertinence ou non des résultats fournis.

1.1.4 Tâches, Modèles et techniques du DM

1.1.4.1 Tâches

Le champ d'application du DM est devenu très large d'où les problèmes abordés sont de plus en plus variés. La formalisation des problèmes permet de les regrouper dans l'une des six tâches suivantes [6] : la description des classes, l'analyse des liens, la classification et la prédiction, la segmentation et enfin l'analyse des tendances.

- Description des classes: la description des classes est l'opération qui permet d'associer à ses individus membres un sens sémantique appelé concept. Par exemple, décrire les clients selon les concepts : dépensier, fidèle, …

La description des classes peut se faire de deux manières : caractérisation et/ou discrimination.

Caractérisation (ou profiling) : le concept d'une classe est défini sur la base des caractéristiques générales de ses membres.

Discrimination : la discrimination cherche à identifier les caractéristiques générales qui différencient une classe d'une autre.

- Analyse des associations : c'est une tâche qui permet de découvrir les rapports de liens qui peuvent exister à l'intérieur d'une base de données. Ces liens sont généralement exprimées sous la forme : $A \Rightarrow B$, qui signifie que la présence de A implique la présence de B mais avec un certaine probabilité exprimée par la notion de support et de confidence [7].

- Classification et prédiction: le but de la classification est de construire le ou les modèles qui permettent de décrire les classes tels que : les arbres de décision, les réseaux de neurones, … La construction d'un modèle se fait d'une manière supervisée par apprentissage automatique à partir d'une base d'exemples étiquetés, c'est-à-dire que la classe de chaque exemple est connue a priori.

Le modèle généré par la classification peut être alors exploité pour réaliser la tâche de prédiction dans le but de [6]:

- prédire ou estimer la valeur d'une donnée manquante ou erronée en utilisant le modèle de classification comme référence,
- ou bien prédire la classe d'une donnée nouvellement arrivée.

- Clustering : contrairement à la classification supervisée le clustering est une tâche de classification non supervisée qui utilise une base d'exemples non étiquetés. L'objectif est d'étiqueter cette base pour générer les classes. Le principe général du clustering repose sur le regroupement des objets de telle manière que les objets d'un même groupe, ou classe, soient fortement similaires entre eux et fortement dissimilaires avec les objets des autres classes.

- Analyse des tendances : cette tâche consiste en l'étude de l'évolution dans le temps d'un modèle de données afin de comprendre son comportement et d'expliquer

les déviations qu'il a subies à un instant donné. Une déviation est tout simplement une évolution anormale significative du modèle.

1.1.4.2 Modèles

Le modèle est un formalisme qui permet de décrire une méthodologie de réalisation d'une ou plusieurs tâches du data mining. Le choix d'un modèle dépend surtout :

- de la nature des données à manipuler
- et du type du problème à résoudre.

A titre indicatif nous présentons quelques exemples de modèles dans le tableau 2.1. Le lecteur intéressé trouvera dans [9] une taxonomie très riche.

1.1.4.3 Techniques (ou algorithmes)

Il existe plusieurs algorithmes différents, chacun propose une manière pour construire le modèle auquel il est associé. Le choix entre eux dépend surtout du contexte de son application. Nous citons à titre d'exemples :

- l'algorithme C5.5, introduit par Quinlan en 1986, pour les arbres de décision
- l'algorithme Apriori, pour les règles d'association, développé par Agrawal et Strikant en 1994 [9].
- L'algorithme SOM (Self Organizing Map), introduit par von der Malsburg en 1973, pour les réseaux de neurones [9]
- L'algorithme MCMC (Markov Chain Monte Carlo), introduit par Gilks, Richardson et Spiegelhalter en 1996, pour les réseaux bayésien [9]
- L'algorithme k-means pour le clustering par partitionnement [9]
- ….

Modèles	**Données étiquetées**	**Données non étiquetées**	**Séries temporelles**	**Prédiction et classification**	**Découverte de modèles, associations et structures**	**Reconnaissance des similarités et dis similarités entre données**
Arbres de décision	×			×	×	×
Règles d'association		×			×	×
Réseaux de neurones	×	×	×	×		×
Réseaux bayésien	×	×	×	×	×	×
Chaîne de Markov	×	×	×	×		×
Plus proche voisin	×		×	×	×	×
Clustering (Partitionnement, Hiérarchique, …)		×			×	×
Analyse de la Composante Principale	×		×	×	×	×

Tableau 1.1: Exemple de modèles de Data Mining

1.1.5 Classification des techniques de DM

Les techniques de DM se répartissent en deux grandes familles : les techniques descriptives et les techniques prédictives [5].

1.1.5.1 Les techniques descriptives

Elles sont de nature exploratoire, elles permettent de réduire, de résumer et de synthétiser les données. Leur but n'est pas de trouver des explications pour des variables mais d'explorer les données pour mettre en évidence les structures cachées. Cette famille regroupe principalement les techniques associées aux tâches de description des classes, d'analyse des associations et de clustering.

1.1.5.2 Les techniques prédictives

Elles sont de nature explicative, elles visent surtout à extrapoler de nouvelles informations à partir des informations présentes. Elles se basent essentiellement sur des modèles qui utilisent des données présentes ou passées pour construire des scénarios futurs. Cette famille regroupe principalement les techniques associées aux tâches de la classification automatique, de la prédiction et de l'analyse des tendances.

En général, le data mining consiste en une description suivie d'une prédiction. Un bon modèle prédictif nécessite une bonne compréhension des données qui s'appuie principalement sur des tâches descriptives comme elle peut aussi se faire par certains indicateurs statistiques (moyenne, écart type, déviation, distribution des données, …) ou par une préparation des données (groupement, transformation, …).

1.1.6 Etapes de mise en œuvre du DM

Dans une première étape il est important de bien définir la finalité attendue du processus de DM car ce sont les objectifs qui vont permettre de choisir les tâches du processus et les variables appropriées. La figure 1.3 montre l'influence de chaque étape sur une autre.

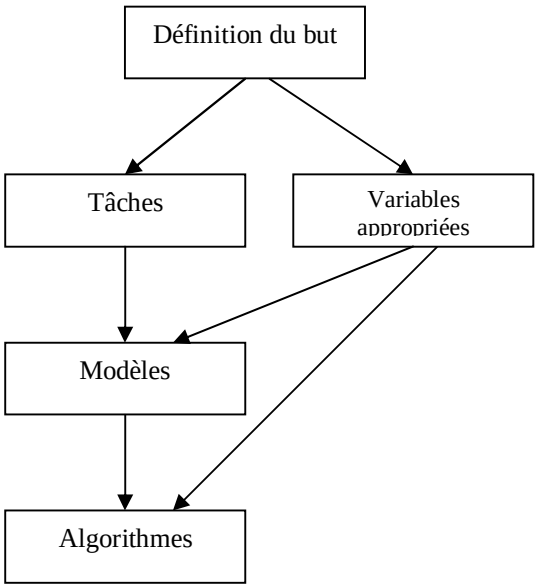


Figure 1.3 : Etapes de mise au œuvre du DM

1.2 Le clustering

1.2.1 Introduction

La classification est une tâche appliquée quotidiennement dans la vie courante. Elle est utilisée pour expliquer les nouveaux phénomènes rencontrés en les comparant avec les concepts et les phénomènes connus en essayant de rapprocher les caractéristiques les plus saillantes.

C'est un sujet de recherche actif qui se place au cœur de l'analyse de grandes masses de données, ce qui justifie pleinement l'intérêt qui lui est porté. Il se trouve au confluent de plusieurs disciplines dont les statistiques, l'apprentissage automatique, le data mining, …

Le problème majeur auquel les méthodes de classification essayent de répondre : Comment associer une classe à un objet ?

Essentiellement, les techniques de classification sont réparties en deux approches : la classification dite supervisée et la classification dite non supervisée appelée aussi clustering [13].

L'approche supervisée tente de construire un modèle de classification à partir d'un ensemble de données étiquetées, c'est-à-dire que la classe de chaque objet est connue à priori, on parle alors de l'apprentissage supervisé ; cet ensemble de données est souvent nommé dans la littérature par ensemble d'apprentissage. Ceci suppose donc une bonne connaissance des classes notamment sur quelle base deux objets sont regroupés ou séparés. Le modèle obtenu doit être capable de prédire la classe la plus vraisemblable sur de nouvelles entrées (ou objets) afin de garder la cohérence avec la structure initiale des classes. Cette cohérence est évaluée sur la base d'une mesure de qualité des classes. La cohérence peut être mise à défaut si le nombre de nouvelles entrées à classer dépasse largement celui de l'ensemble d'apprentissage ; ceci est dû par les nouvelles informations apportées. D'où la nécessité d'actualiser la classification initiale.

Par opposition, dans une approche non supervisée, l'ensemble d'apprentissage est non étiqueté. Le problème de classification devient alors plus compliqué, car on ne dispose d'aucune information à priori sur les classes. L'objectif est de détecter les objets similaires en fonction des variables (attributs) qui les décrivent afin de les regrouper toute en ignorant certains détails de similarité ou dis-similarité. Ce rapprochement entre les objets se base sur une mesure appelée mesure de similarité ou distance. Le résultat obtenu est jugé satisfaisant ou non en fonction du degré d'homogénéité intra classes et d'hétérogénéité inter classes.

L'approche non supervisée est connue aussi sous l'appellation de clustering, segmentation et parfois regroupement; les groupes obtenus sont appelés classes ou clusters.

La modélisation des données par le clustering trouve ses racines dans les mathématiques, les statistiques et l'analyse numérique. Le DM ajoute à cette méthode de modélisation la complexité d'ensembles de données volumineux avec beaucoup d'attributs ou variables de types différents. Cela impose plus de puissance de calculs et nécessite des algorithmes plus performants.

Souvent le clustering est la première tâche à effectuer pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation.

1.2.2 Objectifs du clustering

En résumé, nous pouvons dire que le clustering doit permettre d'atteindre un des objectifs suivants [14] :

- Trouver une vraie typologie.
- Ajustement à un modèle.
- Prédiction basée sur les groupes.
- Tests d'hypothèses.
- Exploration des données.
- Génération d'hypothèses.
- Réduction des données

1.2.3 Domaines d'application

Le clustering est un sujet abordé par plusieurs métiers. Ce fait lui donne l'avantage d'être utilisé par divers domaines d'application qui peuvent être catalogués suivant le but recherché selon trois grands types : [13]

- Extraction des connaissances : dans ce type d'application on cherche généralement à donner une description sémantique aux groupes constitués afin de donner un sens à l'information dont on dispose. Le résultat représente des concepts de données et la littérature parle parfois de clustering conceptuel. Le diagnostic médical qui procède au groupement des patients en se basant sur leurs caractéristiques est un exemple de ce type d'application.

- Réduction des bases de données : l'objectif visé est de réduire l'espace des données sur lequel on travaille en espérant que le problème qu'on cherche à résoudre devient beaucoup plus simple. Ceci est effectué par une segmentation de cet espace en sous espaces homogènes.

- Etude de comportement (profiling) : le but est de détecter des sous populations qui partagent des caractéristiques proches ce qui leurs donne un certain comportement commun. Les systèmes de décision se basent souvent

sur le profiling afin d'entreprendre des actions convenables à chacune des sous populations.

1.2.4 Etapes du processus de clustering

Tout processus de classification, comprend généralement les étapes suivantes [13]:

(1) Représentation et préparation des données
(2) Choix d'une mesure de distance appropriée aux données
(3) Choix d'une méthode (un algorithme) de clustering
(4) Choix d'une forme d'abstraction des données
(5) Validation des résultats

1.2.4.1 Etape 1 : Représentation et préparation des données

Les bases de données sont rarement conçues dans un objectif d'applications de data mining. La préparation des données, bien qu'elle soit optionnelle, est souvent intégrée dans le processus car elle peut contribuer efficacement à sa réussite. En effet, la manière dont les données doivent être présentées diffère d'une technique à une autre et elle a un impact direct sur la performance de l'algorithme adopté et sur les résultats obtenus. Citons à tire d'exemple, que les techniques basées sur les réseaux de neurones manipulent mieux les valeurs numériques mais elles sont fortement sensibles aux valeurs manquantes, alors que les arbres de décision sont beaucoup mieux adaptés pour les valeurs catégoriques et sont faiblement sensibles aux valeurs manquantes. [9]

Le problème de la préparation est comment trouver des sous ensembles réduits qui se comportent plus ou moins de la même manière que la population concernée dans le but de: [9]

- Réduire la dimension des données : en effet plus la dimension des données est élevée plus l'algorithme employé perd de sa performance. Ceci peut être réalisé en ne sélectionnant que les variables ayant une pertinence avec le problème traité et qui ont une certaine indépendance entre elles c'est-à-dire qui ne présentent, dans la mesure du possible, aucune corrélation ou une corrélation très faible.

- Normaliser la distribution des variables numériques

- Diviser l'ensemble des données en trois sous ensembles (échantillons) aléatoires et significatifs : [9]

  - un ensemble d'apprentissage pour la construction du modèle,
  - un ensemble de test pour améliorer la performance du modèle,
  - et un ensemble d'évaluation pour évaluer le comportement du modèle face à différentes situations.

- Effectuer des transformations si nécessaire sur certains types de données : par exemple les valeurs qui expriment la notion de temps et de distances peuvent être groupées par tranche comme les dates naissance.

1.2.4.2 Etape 2 : Choix d'une mesure de distance

Les méthodes de clustering traditionnelles se basent sur une mesure de distance calculée à partir des valeurs prises par les attributs pour effectuer des rapprochements entre les objets [13]. Une mesure de distance (appelée aussi mesure de similarité ou de proximité) exprime une similarité ou une dis similarité : plus deux objets se ressemblent plus leur similarité est grande et plus leur dis similarité est petite. Cette mesure peut être exprimée par une formule qui génère un nombre positif reflétant la proximité/éloignement entre deux points de l'espace des données. Il existe plusieurs sortes de mesures dont le choix dépend surtout [17] :

- du type de données considérées : numériques, nominales, …
- de la nature de groupement cherchée : par exemple pour mettre en avant certaines propriétés dans la similarité ou la dissimilarité entre objets
- de l'algorithme de clustering adopté,
- et de la complexité du calcul

D'une manière générale, le rapprochement ne concerne pas uniquement les objets mais selon les algorithmes il peut désigner trois situations : entre deux objets, entre un objet et un cluster, entre deux clusters [13].

L'estimation d'une distance entre des objets décrits uniquement par des variables numériques n'est pas aussi simple que cela puisse paraître. En effet plusieurs difficultés peuvent être rencontrées à savoir :

- différence d'échelle entre les attributs : Les attributs ayant des valeurs supérieures seront avantagés dans le calcul de la distance par rapport aux restes des attributs. La solution est de les standardiser en les ramenant à une moyenne nulle et un écart-type unitaire pour avoir des valeurs comparables et pour que la distance ait un sens.

- pertinence des attributs : La plupart des mesures considèrent les attributs au même niveau d'importance mais la réalité est que certains peuvent être plus pertinents et par conséquent participent plus dans le rapprochement ou l'éloignement des objets. La solution est d'allouer des poids aux attributs pour que chacun participe selon son degré de pertinence. Ces poids peuvent être estimés en fonction de la variance des attributs. Cette solution devient difficile lorsque le nombre d'attributs est important.

- corrélation des attributs : Même si les attributs sont ramenés à la même échelle la présence d'attributs corrélés peut provoquer une distorsion dans le calcul de la distance. Plusieurs solutions peuvent être utilisées, telle que l'application de l'analyse en composantes principale pour n'utiliser que les composantes factorielles.

1.2.4.3 Etape 3: Choix et application d'une méthode (un algorithme)

Une méthode de clustering a pour objectif principal la découverte des "patterns" cachés dans un ensemble de données brut. Il est extrêmement difficile de dresser un état de l'art complet sur la discipline pour les raisons suivantes :

- le fait que le clustering est abordé par divers métiers a favorisé la diversité des méthodes,

- les méthodes de clustering sont "data driven" c'est-à-dire conduites par les données ; il n'existe pas donc une méthode supérieure à une autre de

manière absolue. Autrement dit, une méthode peut réussir dans un contexte et échouer dans un autre,

- la variété des mesures de distance

Le choix d'une méthode dépend surtout:

- du type des attributs (qualitatif, quantitatif ou mélangé),
- du volume des données, en effet certaines méthodes s'adaptent bien à de grands jeux de données alors que d'autres non,
- de la dimension des données : certaines méthodes perdent leur efficacité si la dimension devient trop élevée,
- de la capacité de traiter les cas (objets) isolés ou exceptionnels,
- de la complexité en terme de temps de calcul,
- de la dépendance de l'ordre des données,
- de la dépendance des paramètres prédéfinis : les connaissances à priori qui vont constituer les paramètres initiaux de la méthode.

Notons qu'en pratique, même si toute l'intention est prise pour choisir la méthode la plus appropriée, ce choix n'est réellement estimé qu'après validation des résultats obtenus.

Dans le chapitre 3, un état de l'art non exhaustif est établi sur les méthodes les plus couramment utilisées.

1.2.4.4 Etape 4 : Choix d'une forme d'abstraction des données

Une fois une méthode de clustering est appliquée sur un ensemble de données, une partition est alors constituée et l'ensemble de départ devient un ensemble de classes. Chaque classe doit avoir un représentant qui est son centre de gravité. Ceci permet de définir une forme d'abstraction des données.

Parmi les formes d'abstraction, nous citons :

- le centroïde : c'est l'objet prédominant parmi les objets d'une classe

- le médoïd : c'est un objet fictif calculé sur la base des objets de la classe.

Notons que le choix du représentant des classes dépend surtout de la méthode de clustering adoptée.

1.2.4.5 Etape 5 : Validation des résultats

Les algorithmes de clustering doivent tenir compte de deux propriétés importantes, appelées inerties, durant le partitionnement : l'homogénéité intra clusters et l'hétérogénéité inter clusters

L'homogénéité intra clusters permet d'assurer une bonne cohésion à l'intérieur des groupes ; elle se base sur un rapprochement objet-cluster. De manière générale, un objet n'est accepté dans un groupe que si la similarité est supérieure à une certaine valeur définie comme étant le seuil d'admission.

Parmi les mesures les plus couramment utilisées citons le rayon moyen défini par :

$$R_q = \frac{1}{n_q} \sum_{i=1}^{n_q} d(X_i, G_q) \qquad [13]$$

où n_q représente le nombre d'objets du cluster C_q de centre de gravité G_q
Plus le rayon R_q diminue plus la cohésion interne du groupe augmente.

Contrairement à l'homogénéité intra clusters, l'hétérogénéité inter clusters permet d'assurer l'éloignement entre les groupes. Son principe revient à utiliser le rapprochement cluster-cluster pour maximiser la dissimilarité entre un cluster et les objets qui lui sont extérieurs. Parmi ces mesures nous citons le lien moyen qui considère la distance entre les centroïdes de deux clusters [13] :

Ajoutons que les deux propriétés évoquées n'empêchent pas l'opération de clustering d'aboutir à des résultats aberrants, d'où la nécessité :

- de vérifier si les classes fournies sont significatives selon des critères souvent subjectifs qui relèvent du métier,

- d'évaluer la qualité de la partition obtenue. Ceci peut se faire de deux manières différentes :

  - comparer le résultat à une partition de référence, ceci suppose qu'on dispose déjà d'un ensemble de données étiqueté comparable à notre ensemble de données,
  - procéder à plusieurs partitionnement et garder la meilleure,
  - utiliser des mesures de qualité (exemple l'erreur quadratique) [13].

1.3 Conclusion

Dans ce chapitre nous avons exposé les principales notions évoquées sur le data mining : une introduction sur les motivations de son émergence, les principales étapes d'un processus d'extraction de connaissances, sa définition ainsi que ses objectifs, ses tâches et ses principaux modèles, un aperçu sur ses techniques et la démarche de sa mise en œuvre.

Nous nous sommes intéressés plus particulièrement à la classification automatique en montrant qu'elle peut être de deux natures, supervisée ou non supervisée, tout en résumant aussi ses grandes étapes de mise en œuvre ainsi que ses concepts les plus importants : la nature des données, la notion de distance et la mesure de qualité.

CHAPITRE 2
LE DATA MINING ET SON APPLICATION DANS L'ÉDUCATION

2.1 Introduction

Le spectre des applications de DM est devenu très large, il envahit de plus en plus de domaines [8] : génomique, astrophysique, e-commerce, gestion de la relation client, gestion des risque comme la détection automatique de fraudes dans la téléphonie mobile, le contrôle de qualité dans l'industrie, les études théoriques dans les sciences humaines, les études biologiques et même dans des domaines de divertissement comme les prévisions d'audiences TV, …

Le DM est surtout utilisé pour appuyer les systèmes d'aide à la décision. En effet, le but n'est plus seulement de modéliser la réalité mais d'orienter la subjectivité humaine dans le processus de décision [8], en voici des exemples :

- Mettre au point certains traitements du cancer sur la base d'analyse de données car le mécanisme biologique de la maladie demeure mal connu du fait de sa complexité.

- Etudier l'efficacité d'un traitement thérapeutique,

- Découvrir les substances qui peuvent être utilisées dans le traitement d'une maladie donnée.

- Gérer toutes les phases du cycle de vie des clients, acquérir de nouveaux clients, fidéliser les clients, retenir les bons clients, déterminer les caractéristiques des bons clients (profiling), …

- Détecter les utilisations frauduleuses dans la télécommunication et les cartes de crédits.

- Réduire les fraudes dans les compagnies d'assurance.

- Etudier les stocks comme les différents éléments affectant le stockage d'un produit.

Dans ce qui suit nous nous intéressons particulièrement aux applications du DM dans le domaine de l'éducation.

2.2 Le Data Mining et l'éducation

L'application des techniques de data mining dans le domaine de l'enseignement est une discipline émergente et relativement récente par rapport aux domaines des affaires et des entreprises [3] tels que l'étude du marché, profil des clients, …

Le souci est de mettre au point des méthodes pour explorer les types de données qui proviennent des milieux éducatifs, et d'utiliser ces méthodes afin d'agir dans le sens qui permet d'améliorer la qualité des enseignements.

Nous présentons dans ce qui suit un bref résumé sur quatre travaux. Le premier travail est un exemple d'utilisation du data mining dans la prédiction des résultats finaux des élèves, le second propose une manière pédagogique pour répartir les élèves dans des groupes de travail, le troisième montre comment on peut construire des modèles pédagogiques pour en inférer des connaissances utiles et intéressantes et le dernier propose une étude sur le comportement d'apprenants dans un environnement interactif.

2.2.1 Prédiction des résultats des élèves [1]

Le système LON-CAPA (Learning Online Network With Computer Assisted Personalized Approch) est un système d'enseignement en ligne développé par MSU (Michigan State University). Il permet de collecter une grande quantité d'informations sur les élèves utilisateurs sur lesquelles le data mining peut être appliqué. Ces

informations constituent par conséquent une base de données qui peut être scindée en deux types :

- les ressources Internet sollicitées par les élèves,
- les travaux et les devoirs réalisés.

Dans ce travail une méthode de classification des d'élèves utilisant le système LON-CAPA est développée dans le but de prédire les résultats finaux en se basant sur un ensemble d'informations (variables) extraites d'une base de données qui décrit le comportement des utilisateurs (élèves) durant le processus de formation.

La méthode utilise des données d'archives et se déroule en deux étapes :

- une étape non supervisée : dans laquelle on constitue des classes d'élèves selon leurs résultats définitifs,

- une étape supervisée : où une série d'algorithmes de classification (classificateur) du système LON-CAPA sont comparés, a savoir :

- classification Bayesien
- 1-plus proche voisin (1-NN)
- k-plus proche voisin (k-NN)
- estimation de densité noyau (Parzen-window)
- les réseaux de neurones multicouches (MLP)
- et les arbres de décision.

L'étude comparative consiste d'abord à estimer le taux d'erreur individuel de chaque algorithme ensuite un algorithme génétique est utilisé pour calculer la solution optimale en considérant que les classes ayant obtenu le maximum de votes par les classificateurs individuellement.

Une mesure de statistique basée sur l'entropie a été employée par la suite pour montrer le degré d'importance de chaque variable dans la constitution des classes.

Une simulation est faite sur un ensemble de test de 227 élèves ayant suivi un cours d'introduction sur la physique le long d'un semestre de l'année 2002. Le cours contient un ensemble de 12 devoirs avec un total de 184 problèmes.

Pour mener l'étude, six variables ont été considérées :

1. le nombre total des réponses correctes,
2. le nombre total de problèmes corrects au premier essai,
3. le nombre total d'essai pour effectuer un devoir,
4. le temps dépensé pour résoudre un problème,
5. le temps total dépensé sur un problème sans tenir compte s'il est résolu ou non,
6. la participation c'est-à-dire l'interaction avec les autres élèves et l'enseignant.

Le tableau 2.1 montre un exemple de 3 classes obtenues par la première étape. Les résultats définitifs sont exprimés en notation américaine :

Niveau de Classe	**Résultat final**	**Nombre d'élèves**	**pourcentage**
Elevé	≥ 3.5	69	30.40 %
Moyen	2.0 < < 3.5	95	41.80 %
faible	≤ 2.0	63	27.80 %

Tableau 2.1 : Partitionnement en trois classes

Le tableau 2.2 résume le degré d'importance en pourcentage de chaque variable :

variables	Degré d'importance (%)
1	100.00
2	58.61
3	27.70
4	24.60
5	24.47
6	9.21

Tableau 2.2 : Degré d'importance des variables dans la classification

Les résultats montrent clairement que les deux premières variables ont la plus grande influence dans la classification. Ce résultat permettra à l'enseignant d'identifier les élèves qui présentent un risque d'échec et de prendre les mesures appropriées au moment opportun.

2.2.2 Distributions des élèves dans des classes [2]

Dans le souci d'améliorer la qualité des enseignements, le département de l'ingénierie industriel de l'université Parahyangan Catholic s'est posé les questions suivantes :

- Doit-on changer notre méthode habituelle dans la distribution des élèves dans les classes?
- Quelles sont les raisons si le changement est nécessaire?
- Comment changer s'il s'avère nécessaire?
- Quels sont les facteurs à considérer si l'université décide ce changement?

L'activité d'enseignement fait intervenir deux parties : l'enseignant et l'élève. L'idéal dans cette activité est d'avoir un seul enseignant pour un seul élève dans une seule matière pour atteindre le maximum de qualité chose qui est impossible à appliquer du point de vue technique et financier.

Dans le souci de s'approcher de cet idéal, le département de l'ingénierie industriel de l'université Parahyangan Catholic a pensé à un compromis qui consiste à prêter attention sur la distribution des élèves dans les classes.

Habituellement la distribution des élèves est faite selon un certain ordre comme l'ordre alphabétique ou l'ordre des matricules d'identification ou peut être même de manière complètement aléatoire sans qu'aucun critère pédagogique ne soit considéré.

L'idée proposée consiste en une distribution sélective basée sur la similarité des élèves. Le processus de distribution devient donc un problème de clustering et les classes des clusters. Chaque classe regroupe les élèves qui partagent une certaine similarité par rapport aux connaissances près requises afin d'assurer une certaine homogénéité du niveau.

Une étude a été menée sur 180 élèves ayant suivi un cours intitulé "Recherche Opérationnelle II" durant le sixième semestre de leur cursus. Les connaissances pré-requises concernent cinq matières comme le montre le tableau2.3 :

Connaissances pré requises	**Les élèves**				
Recherche opérationnelle I	1	4	2	3	…….
Calcul multi variables	1	2	2	3	……
Matrices et espace vectorielle	2	4	2	3	…….
Calcul des intégrales	2	4	2	2	…….
Calcul différentiel	2	2	2	2	……..

Tableau 2.3 : Les connaissances pré requises

Chaque élève est défini par un vecteur de 5 variables représentant respectivement les 5 matières. Les variables représentent les notes obtenues selon la notation américaine A=4, B=3, C=2, D=1 et E=0.

La technique du clustering flou, introduit par James C.Bezdek en 1973, a été adoptée pour attribuer l'ensemble des élèves dans les trois classes. Cette technique repose sur le calcul du degré d'adhésion dans chaque classe. L'élève est alors assigné à la classe ayant le degré plus élevé comme le montre le tableau 2.4 :

	Degré d'adhésion		
Elèves	**Classe 1**	**Classe 2**	**Classe3**
1	**0.840**	0.122	0.038
2	0.200	0.391	**0.409**
3	**0.852**	0.127	0.021
4	0.139	**0.760**	0.101
…..	……	…..	……

Tableau 2.4 : Les degrés d'adhésion de chaque élève dans les classes

Le tableau 2.5 décrit la partition obtenue ou chaque classe est représentée par un vecteur défini comme étant son centre de gravité.

Matières	**Classe 1**	**Classe 2**	**Classe3**
1	1.755	2.904	3.655
2	1.568	2.492	3.842
3	2.226	2.795	3.500
4	1.913	2.220	3.498
5	1.849	2.639	3.472

Tableau 2.5 : Partition finale retenue

Comme exemples d'utilités de ce travail :

- installer des enseignants adaptés aux classes
- regrouper les élèves selon leur rythme d'assimilation.

2.2.3 Extraction de modèles pédagogiques [3]

Ce travail montre comment les algorithmes de data mining peuvent aider à découvrir des connaissances pédagogiques utiles, pour l'enseignant et pour

l'apprenant, à partir d'une base de données obtenue à partir d'un système tutorial ou d'apprentissage basé web.

Les systèmes web ont la particularité de collecter une grande quantité de données sur les élèves à partir desquelles on peut extraire des modèles pédagogiques intéressants et qui peuvent servir à :

- offrir à l'enseignant un support de suivi à distance comme par exemple constater la progression d'un élève, ses difficultés, ses lacunes, sa capacité d'assimilation,
- adapter le cours selon le profil de l'élève,
- prédire la performance des apprenants et en prendre en charge ceux qui présentent le risque d'échec,
- …

Une simulation a été faite sur une base de données obtenue à partir du système Logic-ITA.

Logic-ITA est outil tutorial basé web utilisé par l'université de Sydney depuis 2001. Il est utilisé comme un tuteur secondaire pour aider les étudiants à maîtriser les preuves de la logique formelle d'une part et d'assister les enseignants dans le suivi de leurs classes d'autre part.

La simulation a été faite sur 860 individus (élèves) ayant utilisé l'outil pendant quatre ans. La base de données obtenue regroupe :

- l'identification des élèves,
- les exercices résolus correctement ou incorrectement, le nombre d'essais pour chaque exercice, …
- les erreurs commises,
- …

Différents outils de data mining ont été combinés, tels que :

- le clustering (l'algorithme k-means combiné avec le clustering hiéarchique) pour construire des groupes d'individus homogènes,

- la classification (au moyen des arbres de décision) pour prédire les résultats finaux des individus,

- les règles d'associations pour construire les liens éventuels entres les erreurs comme par exemple : si un élève répond incorrectement à X alors il répondra incorrectement à Y.

Cette étude a conduit à l'extraction d'un ensemble de connaissances ayant été d'une aide précieuse dans l'installation d'une politique pédagogique efficace. Nous citons à titre d'exemples :

- les élèves qui ont tenté de faire au moins deux exercices sont plus enclins à faire plus d'exercices et de les achever. De ce fait, il a été exigé de l'élève de faire au moins deux exercices dans le cadre de son évaluation,

- les règles d'associations sur les erreurs ont montré que les élèves ont des difficultés pour comprendre le concept de la preuve logique. Ce fait a motivé la décision de re-formaliser le concept d'une autre manière ce qui a permis d'obtenir une amélioration conséquente dans le taux du succès,

2.2.4 Etude du comportement dans un espace de collaboration [4]

Le travail proposé est une étude du comportement d'un ensemble d'individus utilisant un espace de collaboration commun d'un système de gestion d'apprentissage.

Les systèmes de gestion d'apprentissage (ou de formation) LMS (Learning Management Systems) mettent l'ordinateur et l'Internet au service de l'éducation. Le principe général de ces systèmes repose sur un espace de travail (ou de communication) invitant ainsi les élèves participants à suivre un cours d'une manière interactive et à utiliser cet espace pour communiquer, collaborer, partager des services d'internet comme les forums, … Les participants sont souvent connus sous

l'appellation de communauté d'apprentissage et le cours est dirigé par un instructeur à qui revient la tâche d'évaluer la progression des participants à travers la dynamique de chacun dans l'espace de travail.

Les LMS permettent de recueillir des informations concernant toutes les interactions produites dans l'espace de communication. Les informations ainsi que leur agrégation ont un degré générique de connaissances faible et ne permettent qu'une analyse quantitative telles que : les emails envoyés, la fréquence des emails,…

Par contre, une analyse qualitative doit permettre d'attribuer des valeurs sémantiques aux évènements qui se produisent dans l'espace d'interaction. Pour ce faire, le problème est formulé comme étant une tâche de data mining et plus précisément le clustering.

La finalité de cette recherche est d'examiner l'application des techniques de data mining sur une base de données obtenue par le LMS pour construire des modèles analytiques résumant les modes d'interactions.

Les modèles obtenus permettront au tuteur d'avoir un aperçu plus concis sur les profils des comportements des élèves participants.

Une étude a été menée sur un cours concernant l'utilisation de l'Internet en éducation où un groupe d'élèves suivi par des instructeurs ont été installés avec un ensemble de services mis à leur disposition : e-mails, forums, chats, espaces de stockage, espaces personnels pour la navigation, ….

Les élèves peuvent s'entraider ou contacter un instructeur en cas de difficulté aux moyens des services proposés par la plateforme. Le système se charge en parallèle d'enregistrer tous les évènements qui se produisent dans l'espace de communications. Plusieurs variables (ou évènements) ont été sélectionnées, préparées et rassemblées dans une table pour les adapter aux exigences des algorithmes choisis. Parmi ces variables on cite :

- nombre d'entrées en chat ou forum du cours
- nombre d'entrées en chat ou forum externe du cours
- nombre de messages envoyés
- nombre de fichiers enregistrés
- nombre de pages visitées
- nombre de travaux lancés
- nombre de travaux terminés
- …

Certaines variables sont avantagées par rapport à d'autres en raison de leurs expressivités comme les travaux lancés qui expriment un facteur important qui est le degré d'initiative et de contribution de l'élève.

L'objectif est de construire à partir de ces variables les modes d'interactions, de combiner ces modes pour inférer des connaissances cachées.

Construction du modèle :

La construction du modèle est faite en utilisant une approche du clustering basé sur des modèles de probabilité. Cette approche suppose que chaque cluster k a une probabilité λ_k où chaque variable i a une distribution θ_i. Ainsi le modèle peut être défini comme étant un ensemble de paramètres noté $\theta = \{\lambda_k, \theta_i\}$ où le modèle naïf de Bayes de classification a été utilisé pour définir la distribution θ_i.

L'algorithme EM (Expectation Maximization) a été adopté comme solution. Cet algorithme se base essentiellement sur les principes probabilistes et permet de transformer le problème de clustering en une estimation de probabilité maximum.

L'expérimentation de l'algorithme a permis de construire un modèle de six clusters. La figure 2.1 illustre graphiquement les quatre variables les plus discriminantes dans la formation des clusters à savoir les travaux lancés, les travaux terminés, la participation au chat et au forum.

Figure 2.1 : Profiles des clusters [4]

Où l'aire de chaque cercle représente la probabilité conditionnelle de la variable dans le cluster.

Une interprétation subjective des résultats obtenus permet de déduire des connaissances qui peuvent être d'une grande utilité pour les instructeurs, telles que:

- le premier graphe montre que les groupes 1 et 4 présentent les mêmes comportements du point de vue initiative mais le quatrième graphe montre que les élèves du groupe 4 présentent une interaction faible. Une des explications possible est peut être la non maîtrise du cours,

- les élèves du groupe 2 essayent de participer mais leurs contributions semblent être peu intéressantes. Cette situation montre que ces élèves ont besoin d'être encouragés à travailler davantage dans le fond du cours.

CHAPITRE 3
ETAT DE L'ART SUR LES TECHNIQUES DE CLUSTERING

3.1 Introduction

Le clustering connu également sous l'appellation de classification automatique non supervisée est l'une des plus importantes tâches du DM. Il recouvre un ensemble de méthodes ayant un objectif précis : simplifier la représentation des données initiales en les organisant en classes homogènes. L'analyse de ces classes permettra alors de dériver un ensemble de connaissances [18].

Il est utilisé dans divers domaines tels la biologie, la médecine, l'ingénierie, … D'ailleurs, à l'appellation de classification automatique non supervisée peut être substitué un ensemble d'autres termes selon le contexte d'utilisation [18] : apprentissage non supervisé (en reconnaissance de formes), taxonomie (en science de la vie), typologie (en sciences humaines).

La démarche générale du clustering est semblable à celle de l'ECD (Extraction de Connaissances à partir de Données) [18] : c'est un enchaînement de plusieurs tâches qui peuvent être intégrées dans un processus itératif. C'est un enchaînement qui se résume comme suit et comme le montre aussi la figure 3.1.

1. sélection des variables pertinentes permettant d'exprimer au mieux le jeu de données initial,
2. choix d'une distance ou d'une mesure de similarité / dis-similarité,
3. choix d'une structure de classification (partitionnement, hiérarchique, …),
4. application d'un algorithme de classification (étape centrale)
5. validation des résultats,
6. interprétation des résultats.

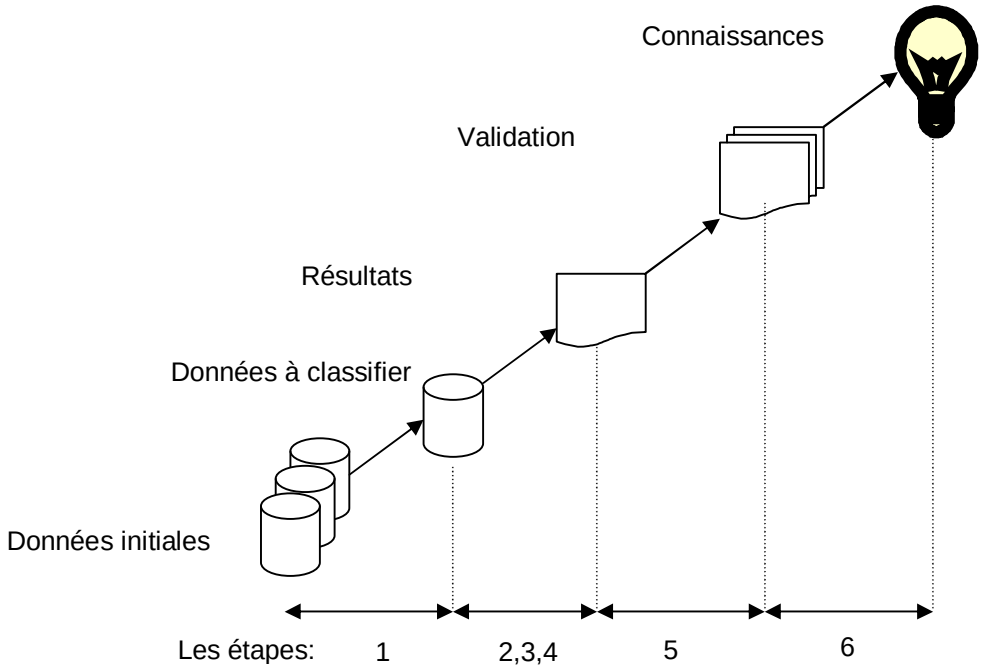


Figure 3.1 : Méthodologie générale d'un processus de clustering [18].

Le choix d'une méthode de classification est fortement dépendant de l'ensemble de données initial, on dit alors que la classification est "data driven"[13]. La diversité des données, notamment celle du web, et l'accroissement constant des volumes des bases de données sont les premières causes de la multiplicité des algorithmes et de leur évolution permanente d'où la difficulté ou peut être même la quasi impossibilité de dresser une taxonomie complète.

La taxonomie des méthodes de classification non supervisée a fait l'objet de plusieurs discussions dans la littérature. Les différences entre les différentes taxonomies dépendent surtout des critères de classification des méthodes. En général, les trois critères suivants sont adoptés dans le classement [18] :

- selon le type des données manipulées,
- selon la mesure de similarité employée,
- selon la stratégie ou le principe global de partitionnement.

On peut également rencontrer d'autres critères de classification tels que :

- les méthodes floues à l'opposé des méthodes classiques : les méthodes floues utilisent la théorie des ensembles flous pour intégrer l'incertitude, ainsi un objet peut appartenir à plusieurs classes avec un certain degré d'appartenance, on parle alors de partition moue à l'opposé de partition dure générée par les méthodes classiques (figure 3.2 et figure 3.3).

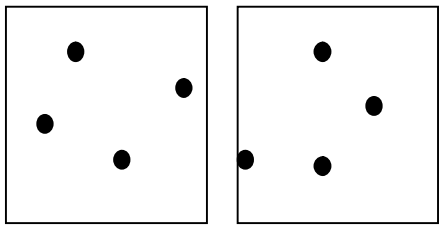


Figure 3.2 : Partition dure

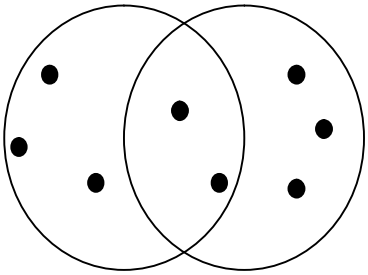


Figure 3.3 : Partition moue

- Les méthodes évolutionnaires basées sur les algorithmes génétiques.

- Les méthodes déterministes ou stochastiques,

- Les méthodes incrémentales ou non incrémentales,

- Les méthodes monothétiques ou polythétiques,

Dans cet état de l'art nous allons adopter la classification qui se base sur la stratégie globale de partitionnement, chacune des classes permet de construire une partition selon une certaine structure. Bien que cette classification soit qualifiée de traditionnelle ou classique [14], elle a pour avantages :

- d'être la plus populaire dans la littérature,
- de couvrir une très grande partie des algorithmes,
- et enfin, nous permettre de bien s'émerger dans la discipline et de construire une idée concise sur le domaine.

Selon cette classification traditionnelle, les méthodes de clustering peuvent être réparties suivant les quatre groupes suivants [6], [18] et [19] :

- Méthodes de partitionnement : elles adoptent une recherche itérative de la partition jusqu'à optimisation d'un critère d'arrêt.

- Méthodes hiérarchiques : elles peuvent être descendante en partant d'une seule classe et par segmentation elles cherchent à atteindre une certaine partition, ou ascendante par fusion de classes jusqu'à la satisfaction d'un critère d'arrêt.

- Méthodes basées sur la densité : le principe du groupement est selon le voisinage des objets et le niveau de densité de chaque objet.

- Méthodes a base de grilles : ces méthodes procèdent à une discrétisation de l'espace de données en un certain nombre de cellules. Elles sont essentiellement utilisées pour des données spatiales [18].

Chaque groupe rassemble un ensemble d'algorithmes. Ces derniers partagent le principe général de construction des partitions suivant une certaine structure.

Dans les prochaines sections, nous développons chaque groupe de méthodes ainsi que les algorithmes les plus utilisés qui lui sont associés.

D'autres méthodes qui s'échappent du cadre de ce regroupement seront brièvement présentées à la fin de ce chapitre.

Nous terminons enfin par une conclusion dans laquelle nous mettons en évidence d'autres critères importants de différenciation entre les différents algorithmes.

3.2 Les méthodes de partitionnement

Son principe général est de démarrer à partir d'un seul cluster qui est partitionné d'une manière itérative en effectuant une redistribution des objets ou en essayant d'identifier les clusters comme étant des régions très peuplées jusqu'à la rencontre d'un critère d'arrêt. L'objectif de ces méthodes est de diviser de manière optimale l'ensemble des objets en un nombre fixe de groupes. Les clusters identifiés ont généralement une forme sphérique.

Les algorithmes de partitionnement les plus utilisés sont *K-means* et *K-medoids* [19]. Les algorithmes récents sont généralement de type medoids mais adoptent une recherche aléatoire au lieu d'un k fixé à priori. Ajoutons que fréquemment c'est l'erreur carrée (squared error) qui est utilisée comme critère d'arrêt.

3.2.1 L'algorithme k-moyennes (k-means)

La méthode des k-means est la plus simple, la plus rapide et la plus utilisée dans les applications scientifiques et industrielles. Elle a été introduite par J.MacQuenn en 1967 et mise en œuvre sous sa forme actuelle par E.Forgy [6].

Dans cet algorithme, chaque classe est représentée par une valeur moyenne pondérée appelée centroid. Au début, les points sont partitionnés dans k classes avec k un paramètre défini à priori par l'utilisateur. Ensuite, itérativement l'algorithme tente de raffiner la solution initiale en procédant à chaque itération à une redistribution des points dans les clusters et en mettant à jour les centroids.

Sous sa forme la plus simple l'algorithme se résume comme suit [6] :

Pour un échantillon X de N objets $x_1, x_2, , x_N$

(1) choisir aléatoirement k objets initiaux (appelés centroids ou noyaux) formant ainsi k clusters C_i de centre initiaux m_i i=1,k

(2) Affecter chaque objet x_j à un cluster C_i de centre m_i telle que la distance x_j et m_i soit minimale

(3) Calculer le centre m_i pour chaque groupe

(4) Répéter les étapes 2 et 3 jusqu'à la stabilité des centres m_i.

La figure 3.4 est un exemple d'exécution de k-means sur un nuage de points bidimensionnels avec un k=3.

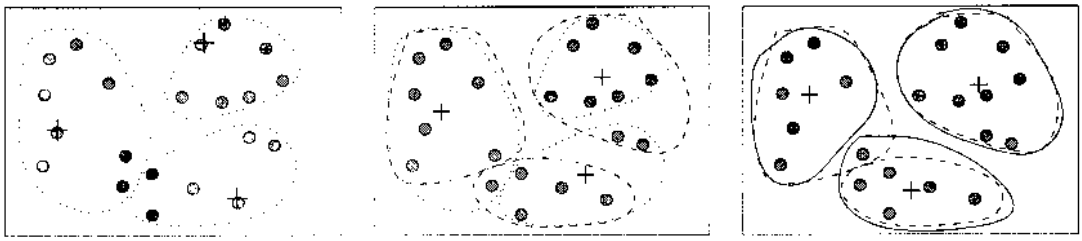


Figure 3.4 : Partitionnement basé sur k-means [6]

Où :

a) choix des trois centres initiaux marqués par + et affectation de chaque objet au centre le plus proche,
b) recalcule des centres pour chaque cluster et redistribution des objets selon les nouveaux centres,
c) le processus est répété jusqu'à stabilisation des clusters.

k-means vise donc à minimiser la variance intra classe qui se traduit par la minimisation de l'énergie suivante :

$$E = \frac{1}{2}\sum_{x \in X} \min \left\| x - m_i \right\|^2 \qquad [13]$$

L'algorithme ainsi défini converge vers un minimum local de l'énergie, appelée aussi fonction objective, qui se traduit par une partition des données en des classes séparées.

Nous faisons remarquer que la fonction objective se base essentiellement sur la variance intra cluster. En d'autres termes, c'est un modèle de distribution normale largement utilisé en statistique, ce qui qualifie la méthode des k-means comme étant une méthode qui dérive d'un cadre probabiliste.

La qualité de la solution trouvée dépend fortement du choix de la valeur de k de départ. Ce choix est généralement empirique de façon à minimiser l'énergie.

Comme avantages de cette méthode, on cite :

- relativement extensible pour traiter des ensembles de données de taille importante,
- indépendante de l'ordre d'arrivée des données,
- complexité linéaire,
- relativement efficace, si n est le nombre d'objet, k le nombre de cluster et t le nombre d'itérations. Généralement l'algorithme converge avec un k et un t inférieur à n,
- produit généralement un optimum local, c'est-à-dire que la solution trouvée n'est pas forcément la meilleure solution, un optimum global peut être obtenu en utilisant d'autres techniques telles que les algorithmes génétiques.

Cette méthode souffre de quelques inconvénients :

- valable seulement dans le cas où la moyenne des objets est définie,
- le nombre de cluster (k) doit être spécifié à priori,
- sensible au choix des centroïdes de départ,
- incapable de traiter les données bruitées,
- les clusters formés sont de forme sphérique et elle est non adaptée pour découvrir des clusters avec des structures convexes,
- les points isolés sont mal gérés.

Pour étendre les capacités de classification de la méthode des k-means de nombreuses variantes se sont succédées [20] :

- k-means itérative optimization : cette version propose d'effectuer une analyse plus détaillée des effets de la fonction objective si un point est déplacé de sa classe à une classe potentielle. Si ce déplacement apporte une amélioration

dans la qualité du partitionnement alors le point est réaffecté et les deux centroids recalculés.

- ISODATA : proposé pour pallier au critère du choix de k. cette version propose de déterminer automatiquement la meilleure valeur de k pour obtenir le meilleur résultat possible.

- k-modes et k-prototypes : développés par Z.Huang. Ces deux versions emploient une mesure de dis-similarité permettant de traiter les données catégoriques pour k-modes et les données catégoriques et numériques pour k-prototypes.

- single pass k-means (scalable k-means) : développé par Badley, Fayyad et Reina. Il vise à augmenter l'extensibilité de k-means pour de gros jeux de données.

- accélération de k-means : proposé par Elkan dans le but de diminuer le coût de calcul des distances en se basant sur l'inégalité triangulaire.

- continuous k-means : cet algorithme diffère par le choix des points de référence initiaux et la sélection des points pour la mise à jour des classes. En effet, les points sont choisis comme un échantillon aléatoire, et pour mettre à jour les classes l'algorithme n'aura à examiner que l'échantillon aléatoire au lieu d'examiner séquentiellement tous les points comme le fait la version standard.

3.2.2 L'algorithme k-medoids

Il est introduit par Kaufman et Rousseew en 1987 [6]. Contrairement aux méthodes k-means ou le cluster est représenté par une valeur moyenne appelée centroïde, dans les méthodes k-medoids un cluster est représenté par un de ses points prédominants appelé medoid. Une telle représentation a deux avantages :

- ces méthodes s'adaptent à n'importe quel type de données
- elles ne sont pas sensibles aux points isolés

k-medoids utilise une fonction objective qui définit la distance moyenne entre un point et le medoid, soit :

k le nombre de clusters défini comme un paramètre à priori

$P = \{p_1, p_2, \ldots, p_n\}$ population ou ensemble des n points à segmenter

$M = \{m_1, m_2, \ldots, m_k\}$ l'ensemble des medoids des k clusters

$C = \{C_1, C_2, \ldots, C_k\}$ l'ensemble des k clusters

Chaque cluster C_i est représenté par son medoid m_i

$d(p_{ij},m_i)$ distance d'un point p_{ij} à son medoid mi avec $(p_{ij},m_i) \in C_i$ pour $j = 1,|C_i|$

$|C_i|$ étant le nombre d'objets de la classe i

La fonction objective peut être définie de la manière suivante :

$$J(M) = J(m_1,...,m_k) = \sum_{i=1}^{k} \sum_{p_{ij} \in C_i} d(p_{ij}, m_i), j = 1,|C_i| \qquad [6]$$

Le principe général des méthodes k-medoids revient à chercher les meilleurs medoids permettant de minimiser la fonction objective *J*.

La figure 3.5 est une illustration sur un nuage de points bidimensionnels avec un k=3:

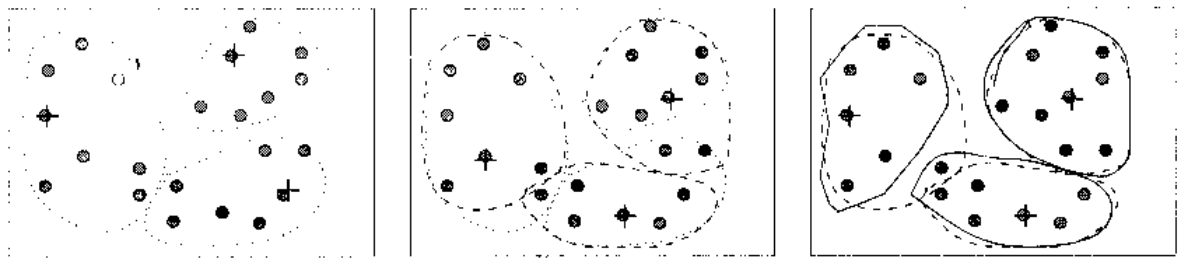


Figure3.5 : Partitionnement basé sur k-medoid [6]

Où :

a) choix des trois centres initiaux marqués par + et affectation de chaque objet au centre le plus proche,

b) recalcul des médoid pour chaque cluster et redistribution des objets selon les nouveaux centres,
c) le processus est répété jusqu'à stabilisation des clusters.

Il existe plusieurs algorithmes adoptant ce principe, les plus connus sont : PAM, CLARA et CLARANS [19]

3.2.3 L'algorithme PAM (Partitioning Around Medoids)

Il est développé par Kaufman et Rousseeuw [20], son principe général se résume ainsi :

(1) sélectionner aléatoirement l'ensemble M des k medoids parmi les n points de données
(2) itérativement pour chaque paire (m_i,p_j) ou p_j est point non medoid des (n-k) points restants :

- calculer la fonction J(M)
- si J(M) diminue alors m_i et p_j changent de rôle

(3) retourner les clusters ainsi que leurs medoids.

L'inconvénient de cet algorithme est le coût total de calcul car il est d'une complexité quadratique de l'ordre de $O(k.(n-k)^2)$ pour chaque itération, ceci le rend non adaptable pour une population importante.

3.2.4 L'algorithme CLARA (Clustering LARge Application)

Il est introduit aussi par Kaufman et Rousseeuw en 1990 [6] et [20], CLARA applique PAM sur un ensemble d'échantillons au lieu de l'ensemble de données en entier. La qualité du partitionnement est améliorée à chaque exécution. Ceci a pour avantage la réduction du coût de calcul, par conséquent, des populations beaucoup plus importantes peuvent être analysées, mais l'algorithme souffre de certains inconvénients, tels que :

- dans certains cas, la définition des paramètres d'échantillonnage peut être très compliquée,

- si un point qui pourrait être le meilleur medoid n'apparaissait dans aucun échantillon alors l'algorithme ne trouvera jamais la meilleure solution.

3.2.5 L'algorithme CLARANS (Clustering Large Application based on RANdomized Search)

Développé par Raymon Ng et Jiawei Han [6] et [20], l'algorithme combine PAM et CLARA. CLARANS ramène le problème de clustering à un problème de recherche sur un graphe abstrait où chaque nœud représente une partition et possède $k\times(n-k)$ nœuds voisins. Un nœud voisin est obtenu par un changement de rôle entre un medoid et un non medoid.

La différence avec CLARA c'est que ce dernier opère sur l'ensemble des données alors que PAM ne s'intéresse qu'à un sous ensemble de nœuds appelés nœuds adjacents.

Deux nœuds M_1 et M_2 tels que $M_1=\{m_{11}, m_{12},\ldots, m_{1k}\}$ et $M_2=\{ m_{21}, m_{22},\ldots, m_{2k}\}$ sont dits adjacents si $|M_1\cap M_2|=k-1$ c'est-à-dire qu'ils ne diffèrent que d'un seul point.
L'algorithme démarre d'abord d'un sommet tiré aléatoirement et explore ensuite le graphe de nœuds voisins en nœuds voisins à la recherche du meilleur nœud.

Son principe peut être résumé de la manière suivante :

(1) choisir aléatoirement un nœud M_0
(2) itérativement se déplacer du nœud M_t à un nœud adjacent M_{t+1} tel que
$J(M_{t+1})<J(M_t)$
avec $M_{t+1} = M_t\cup\{p\}-\{m\}$
où m est un medoid de la partition M_t qui a changé de rôle avec le point p.
(3) s'arrêter lorsque aucun nœud adjacent ne satisfait la relation $J(M_{t+1})<J(M_t)$.

Cet algorithme de clustering a tous les avantages de PAM et de CLARA avec une complexité linéaire par rapport au nombre de points.

L'expérimentation des auteurs a montré qu'il fournit toujours les meilleurs résultats dans tous les cas [20].

3.3 Les méthodes hiérarchiques

Le principe de construction des clusters est graduel comme le développement des cristaux. L'approche peut être ascendante ou descendante, le principe repose soit sur une subdivision dans le cas descendant soit sur une agglomération dans le cas ascendant de manière successive jusqu'à la satisfaction d'un critère d'arrêt [13]. Le résultat obtenu est une hiérarchie de clusters formant un arbre appelé *dendrogramme* (figure 3.6). Le dendrogramme peut être vu comme des niveaux de clustering.

Les algorithmes hiérarchiques sont de deux types : agglomératif et divisive

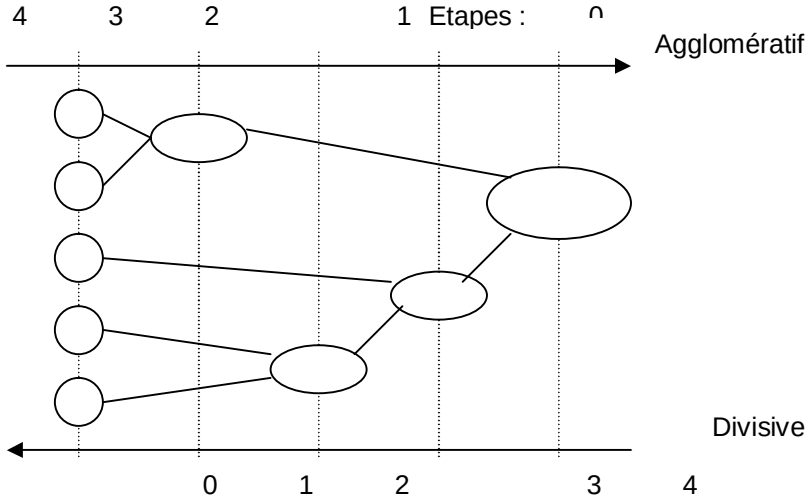


Figure 3.6 : Dendrogramme

La fusion ou l'éclatement des objets se font suivant un critère de distance choisi.
Soient C_1 et C_2 deux clusters :

- Deux objets x_1 et x_2 sont fusionnés avec $x_1 \in C_1$ et $x_2 \in C_2$ si la distance $d(x_1,x_2) \leq$ seuil choisi,
- deux clusters C_1 et C_2 sont fusionnés si la distance $d(C_1,C_2) \leq$ seuil choisi [6].

Les algorithmes hiérarchiques ont l'avantage de ne pas exiger le paramètre k comme entrée, mais juste une condition d'arrêt (par exemple une distance inter clusters est atteinte).

Ces algorithmes sont basés sur les distances inter objets et la recherche des plus proches voisins, ce qui donne une complexité maximale de $o(n^2)$ si toutes les distances inter objets pour un objet sont calculées.

3.3.1 Le groupement agglomératif

La démarche est ascendante (bottom-up), elle démarre d'un seul objet auquel on associe les autres objets un à un.

Au départ, chaque objet constitue un groupe de taille 1. A chaque étape, les deux groupes les plus proches sont fusionnés jusqu'à ce que tous les objets appartiennent à un seul groupe.

La figure 3.7 est un exemple de groupement agglomératif [6] :

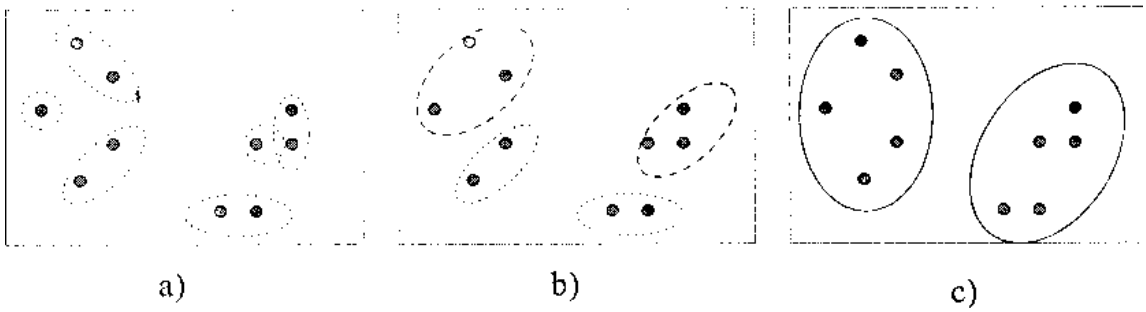


Figure 3.7 : Groupement agglomératif

Où :

a) Au départ chaque objet est considéré comme étant un cluster
b) Dans a et b à chaque itération deux clusters sont groupés
c) En c formation de deux clusters

Caractéristiques de ces méthodes :

- les méthodes hiérarchiques sont les plus utilisées

- les algorithmes associés sont déterministes
- ils se basent sur des matrices de distances
- ils nécessitent la définition d'une distance inter groupes
- et que plus on est haut dans l'arbre, moins est bonne la représentation des données.

Parmi les algorithmes les plus populaires nous distinguons ceux de BIRCH, CURE et CHAMELEON [6] et [19].

3.3.3.1 L'algorithme BIRCH (Balanced Itérative Reducing and Clustering using Hierarchies)

Il est introduit par Zhang, Ramakrishan et Livny en 1996 [6] et [20], la communité du domaine de classification trouve que BIRCH est l'un des meilleurs algorithmes qui peut traiter de gros jeux de données [20]. L'idée principale est que la classification est effectuée sur des données compactées et organisées en une structure d'arbre équilibré appelé CF_tree (Clustering Feature) de taille limitée proche de B_Tree, sa structure se présente comme suit [20] :

- Arbre CF_Tree

CF_Tree est une structure hiérarchique (figure 3.8) où chaque niveau représente une phase de clustering :

- chaque nœud non feuille est une classe contenant au plus B sous classes organisées sous forme d'entrées dans ce nœud,

- l'entrée d'un nœud non feuille est composée d'un pointeur vers un nœud fils et d'un vecteur CF = (N,LS,SS) qui est la somme des CFs de toutes les entrées du noeud fils avec :

  N : nombre d'objets à classer

  $LS = \sum_{i=1,N} \vec{X}_i$ somme linéaire des objets

  $SS = \sum_{i=1,N} \vec{X}_i^2$ somme des carrés

- le vecteur CF est suffisant pour le calcul des informations sur les sous clusters comme le centroid et le diamètre ; il constitue une manière de stockage efficace résumant ainsi les sous clusters au lieu de stocker tous les points,

- une feuille contient au plus L entrées. Chaque entrée stocke un vecteur CF qui représente une sous classe dont le diamètre doit satisfaire un seuil d'absorption T,

- les feuilles représentent les clusters.

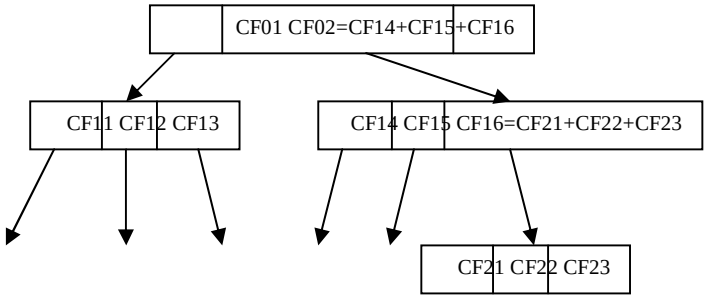


Figure 3.8 : Structure de l'arbre CF_Tree

- Les phases de BIRCH :

BIRCH contient trois phases, la première est obligatoire les deux autres sont optionnelles.

Phase 1 :

- l'algorithme scanne linéairement l'ensemble des points et construit l'arbre d'une manière incrémentale, c'est-à-dire en une seule passe,

- les points sont affectés aux feuilles suivant une similarité définie et sans violation du seuil d'absorption T,

- si le seuil est atteint, une nouvelle entrée dans la feuille est créée,

- si le seuil de L entrées dans la feuille est atteint la feuille est alors décomposée,

Phase 2 :

- l'arbre CF_Tree peut être réduit jusqu'à ce qu'un nombre de feuilles soit atteint.

Phase 3 :

- un algorithme simple de clustering (comme k-means, …) peut être appliqué sur les CF des feuilles pour améliorer la qualité des classes obtenues.

Parmi les avantages de BIRCH, nous citons :

- ✓ il peut traiter de grands volumes de données,
- ✓ il trouve généralement un bon clustering en une seule passe qui peut être amélioré dans une autre étape,
- ✓ la structure CF_Tree peut être ajustée à l'espace mémoire disponible,
- ✓ enfin les expériences ont montré que l'utilisation de BIRCH combiné avec CLARANS donne souvent une très bonne qualité de clustering.

Mais certaines faiblesses sont constatées telles que :

- ✓ la version originale de l'algorithme ne considère que les données numériques,
- ✓ la très forte sensibilité à l'ordre des données.

3.3.3.2 L'algorithme CURE (Clustering Using REpresentatives)

Il est proposé par Guha, Rastagi et Shim en 1998 [6]. Dans CURE la classe est représentée par un nombre fixe de points appelés représentants. Ces derniers sont choisis parmi les points distants du centre qui s'éparpillent bien autour d'une classe afin qu'ils puissent capturer sa forme et son ampleur. Une fois ces points sélectionnés, ils sont alors rétrécis (ou déplacés) vers le centroid par un facteur quelconque. Ceci permet de diminuer l'effet des bruits et des points isolés. Ce n'est

qu'après rétrécissement que les points sélectionnés deviennent les représentants de la classe. La distance entre deux classes est définie comme étant le minimum de distances entre deux représentants des deux classes.
La démarche de l'algorithme se résume comme suit [20] :

- sélectionner aléatoirement un échantillon de points
- partitionner l'échantillon en un certain nombre de partitions partielles
- pour chaque partition procéder au clustering :
    - chaque point d'entrée est considéré comme une classe et par agglomération les classes les plus proches sont fusionnées
    - calcul des représentants de chaque nouvelle classe obtenue
    - les représentants sont stockés dans une structure kd_Tree.

La figure 3.9 est un exemple de déroulement de CURE [6] :

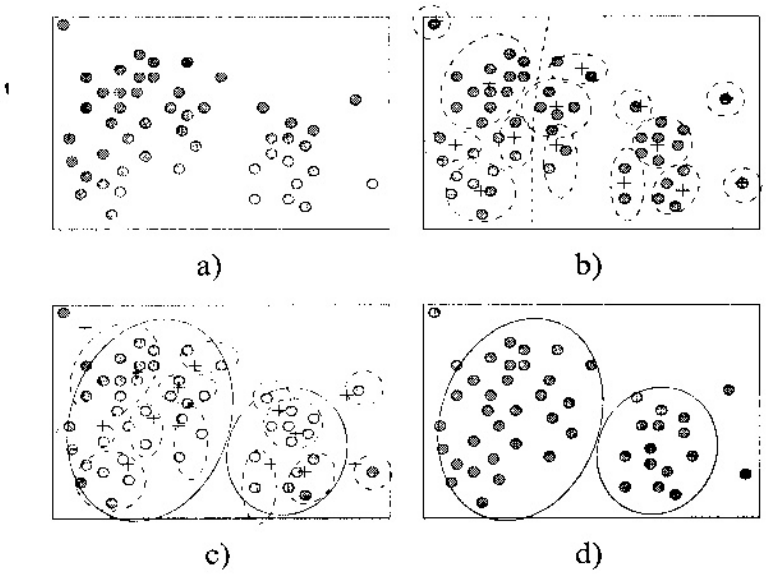


Figure 3.9 : Exemple de partitionnement basé sur CURE [6]

a) l'objectif est d'obtenir une partition de deux clusters d'un nuage de 50 points,
b) partitionnement en 10 classes potentielles,
c) déplacements des représentants vers les centres suivant un facteur,
d) capture de la forme des clusters et formation de deux clusters et suppressions des exceptions.

La qualité des classes formées dépend des paramètres de l'algorithme :

- le facteur de rétrécissement,
- nombre de représentants,
- nombre de partitions,
- et la taille de l'échantillon.

Parmi les avantages de CURE on souligne :

- il est conçu spécialement pour traiter de gros jeux de données,
- il manipule mieux les clusters de formes et de tailles aléatoires,
- il n'est pas sensible aux points isolés,
- il est d'une complexité de temps, au pire des cas, de l'ordre de $O(n^2 \log n)$.

3.3.3.3 L'algorithme CHAMELEON (Hierarchical clustering using dynamic modeling)

Il est introduit par G. Karypis, E. H. Han et V. Kumar en 1999 [6] ; il se déroule en deux étapes :

1. étape de prétraitement : elle consiste à construire un graphe des k plus proches voisins (K-Nearest-Neighbor ou K-NN) sur l'ensemble des points. Ce graphe va servir à capturer les relations entre chaque point et ses K plus proches voisins.

2. étape de clustering : cette étape contient deux phases

Phase 1 : dans cette phase l'algorithme cherche à trouver des sous clusters initiaux en utilisant un algorithme de partitionnement du graphe KNN, précédemment construit, dans le but de trouver le maximum de clusters « solides » dont les sommets sont bien connectés, c'est-à-dire que la distance entre les points est minimisée.

Phase 2 : l'algorithme applique un clustering hiérarchique agglomératif pour fusionner les clusters construits par la première phase selon une similarité définie.

La figure 3.10 montre les étapes de partitionnement basé sur l'algorithme CHAMELEON [6] :

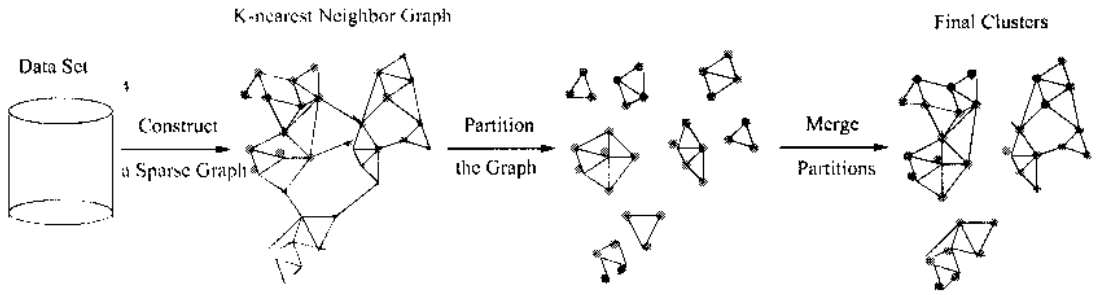


Figure 3.10 : Etape de l'algorithme CHAMELON [6]

Le CHAMELEON est d'une complexité de $O(n^3)$, son point fort c'est qu'il permet de découvrir des classes non connexes. C'est un algorithme très puissant beaucoup plus adapté pour traiter des données spatiales puisque le regroupement repose sur le voisinage d'où la découverte de régions peuplées ou denses [20].

3.3.2 Le groupement par division

Contrairement au groupement agglomératif, le groupement par division est une démarche descendante (top-down), elle démarre de l'ensemble des objets que l'on divise successivement en sous ensembles [13].

Au départ, tous les objets appartiennent à un seul et unique groupe. Un algorithme de partitionnement est ensuite utilisé pour diviser un groupe en deux sous groupes. Cet algorithme est alors appliqué de manière récursive jusqu'à ce que tous les groupes aient une taille de 1.

DIANA est un algorithme très simple qui applique une division récursive des objets selon un critère de dispersion des objets.
Caractéristiques :

- ce sont des algorithmes non déterministes
- plus on est en bas dans l'arbre, moins est bonne la représentation de la structure des données.

Conclusion sur les deux approches :

- l'approche récursive rend les algorithmes de division plus rapides que les méthodes agglomératives,

- les méthodes par division sont largement dépendantes du choix de l'algorithme de partitionnement,

- une approche par division est à préférer lorsque l'on souhaite identifier un faible nombre de groupes.

3.4 Les méthodes basées sur la densité

Ces algorithmes considèrent les clusters comme étant des régions denses d'objets dans l'espace de données. Un point de l'espace est dense si le nombre de ses voisins dépasse un certain seuil. Ils essayent alors d'identifier les clusters en se basant sur la densité des points de données dans une région. Les objets sont alors groupés non pas sur la base d'une distance mais tant que la densité de voisinage excède une certaine limite.

Parmi les algorithmes les plus connus, citons ceux de DBSCAN, OPTICS et DENCLUE [6] et [19].

3.4.1 L'algorithme DBSCAN (Density Based Spatial Clustering of Application with Noise).

Il est introduit par Ester, Kriegel, Sander et Xu en 1996 [6], cet algorithme, très populaire en classification utilise itérativement deux étapes pour découvrir les groupes :

1) choisit aléatoirement un point dense
2) forme un groupe à partir de tous les points accessibles depuis ce point selon un certain seuil de densité.

Pour ce faire, deux paramètres doivent êtres définis :

1) le paramètre de voisinage (ε) qui définit la distance maximale du groupe

2) le seuil de densité (MinPts) qui définit le nombre minimal de points dans un voisinage.

A partir de ces deux paramètres, l'algorithme manipule les notions suivantes :

- voisinage d'un point p par rapport à ε :

Soit : X l'ensemble des points

$V_\varepsilon(p)$ l'ensemble voisinage de p par rapport à ε

$V_\varepsilon(p) = \{q \in X / d(p,q) \leq \varepsilon\}$

Si $|V_\varepsilon(p)| > MinPts$ alors p est un noyau

- point *directement accessible* (directly reachable) : un point p est directement accessible depuis un autre point q relativement aux paramètres ε et MinPts si :

- $p \in V_\varepsilon(q)$
- et q un noyau

- point *d_accessible* (density reachable) : un point p est d_accessible depuis un autre point q s'il existe une série de points $p_1,\ldots,p_n$ telle que :

- $p_1=q$ et $p_n=q$
- p_{i+1} est directement accessible depuis p_i

- point *d_connecté* (density connected) : S est d_connecté à R s'il existe un point O tel que S et R soit d_accessibles depuis O.

- définition d'une classe : si p est un point tel que $|V_\varepsilon(p)| > MinPts$ alors l'ensemble des points d_accessibles depuis p est une classe.

- définition du bruit : tout point n'appartenant à aucune classe est un bruit.

La figure 3.11 est une illustration de l'algorithme DBSCAN :

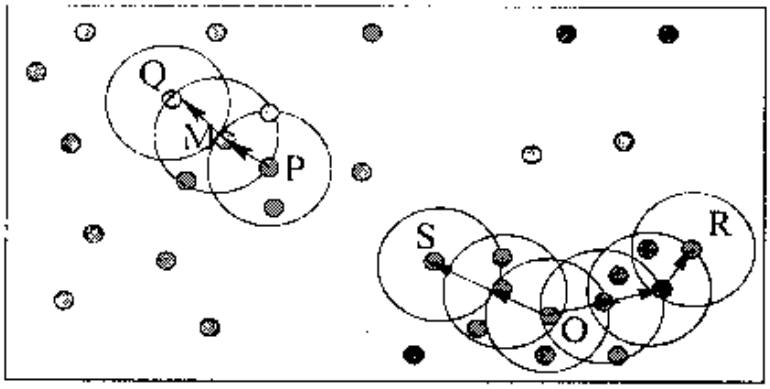


Figure 3.11 : Illustration de DBSCAN

La démarche générale se résume comme suit :

1. sélectionner aléatoirement un point p
2. déterminer l'ensemble E des points d-accessible depuis p
3. si p est un noyau, alors E est une classe
4. sinon sélectionner un autre point et aller en (2) s'arrêter lorsque tous les points ont été sélectionnés

L'algorithme est incrémental et d'une complexité d'ordre $O(n^2)$, ses principaux avantages :

- il peut trouver des classes de forme arbitraires
- il est insensible à l'ordre d'arrivé des objets.

Mais souffre de quelques inconvénients :

- difficulté de fixer les paramètres initiaux, c'est-à-dire, le rayon et le seuil de voisinage
- sa performance peut diminuer pour les données de haute dimension.

3.4.2 L'algorithme OPTICS (Ordering Points to Identify the Clustering Structure)

Il est introduit par Ankerst, Breunig, Kriegel et Sandar en 1999 [6], c'est une extension de DBSCAN qui effectue un clustering pour plusieurs valeurs du rayon de voisinage de manière simultanée. L'algorithme procède ainsi :

1) il commence par créer une liste ordonnée de différents rayons
2) applique ensuite DBSCAN en parallèle à cette liste
3) choisit la plus petite distance qui assurent la plus forte densité.

3.4.3 L'algorithme DENCLUE (DENsity-based CLUstEring)

Il est introduit par Hinneburg et Keim en 1998 [6], cet algorithme adopte une démarche basée sur des principes mathématiques où une modélisation analytique est employée pour calculer l'influence d'un point ainsi que l'espace des points.

DENCLUE formalise la densité globale d'une région dense par une somme des fonctions d'influence de densité associées à chaque point, comme par exemple la fonction gaussienne la plus connue $k(x) = e^{\frac{-d(x,y)^2}{2\sigma^2}}$ [6].

L'algorithme a deux étapes essentielles :

1) étape de pré-traitement : elle consiste à découper l'espace des points par une grille de cellules hyper rectangles de longueur d'arête égale à 2σ,

2) étape de classification : l'algorithme ne considère que les cellules denses et calcule alors les fonctions d'influence correspondantes, ensuite il fusionne les cellules qui peuvent être jointes selon un critère prédéfini.

Notons que DENCLUE est plus efficace que DBSCAN, et il peut même être paramétré pour se comporter comme DBSCAN. Il peut également se comporter comme k-means en jouant sur le choix de σ et en omettant la fusion des classes. En outre, en répétant la classification pour différentes valeurs de σ, on peut obtenir une classification hiérarchique [20].

3.5 Les méthodes basées sur la grille

Le principe de ces méthodes se base sur les algorithmes hiérarchiques ou de partitionnement pour diviser l'espace de données en cellules formant ainsi une grille. Une cellule peut être un cube, une région, un hyper rectangle selon l'intervalle

d’attributs des données. En plus de la densité de la cellule des informations statistiques caractérisant les points de cellules sont aussi stockées. Les cellules voisines sont alors groupées en terme de distance selon la stratégie de l’algorithme adopté. Parmi les algorithmes les plus connus dans ce contexte on cite : STING (STatistical Information Grid introduit par Wang, Yang et Muntz 1997), WaveCluster (développé par Sheikholeslami, Chatterjee et Zhang en 1998) et CLIQUE (développé par Agrawal et Al. En 1998) [19] et [20].

Le principe général est comme suit :

1) fusion des zones denses selon la valeur d'un seuil fixé,
2) zones peu denses permettant d'établir des frontières.

La difficulté majeure réside dans le choix de la taille des cellules :

- cellules de très grande taille : possibilité d'avoir dans la même cellule des objets non forcément proches. Ceci peut générer des cellules peu homogènes, d’où on obtient ce qu'on appelle un sous partitionnement

- cellules de petite taille : elles auront la tendance d'être toujours dense et des frontières qui n'ont aucune raison d'être peuvent être détectées. C'est le cas du sur partitionnement.

Les méthodes basées sur la densité et sur la grille ont pour principe commun de travailler dans un espace métrique et identifient les régions comme étant des régions denses. Les différents algorithmes de ces méthodes sont surtout adaptés pour des contextes de données spatiales [18].

3.6 Autres méthodes

3.6.1 Les méthodes statistiques

Ces méthodes démarrent du principe que les données de départ ont été générées suivant une certaine loi de distribution. Le principe général est d'utiliser le modèle statistique permettant d'approcher au mieux cette distribution pour réaliser le

regroupement. Ensuite, d'une manière itérative, le modèle est amélioré jusqu'à la satisfaction d'un critère d'arrêt.

La limite principale de ces méthodes c'est qu'elles partent du principe que les attributs décrivant les objets sont totalement indépendants entre eux, mais ceci n'est pas toujours vrai, car ils existent souvent des corrélations [6].

Parmi les systèmes de classification basés sur des modèles statistiques : COBWEB (Fisher 1987), CLASSIT (Gennari, Langley et Fisher 1989) et AutoCass (Cheeseman et Stutz 1996) [6].

3.6.2 Les méthodes évolutives

Souvent ces méthodes sont classées dans la famille des algorithmes heuristiques ou génétiques. Ces derniers trouvent leur inspiration dans l'évolution naturelle et le comportement social de certaines espèces comme la colonie des abeilles ou des fourmis. Le principe général se base sur l'évolution d'une population de partitions au lieu d'une seule comme le font les autres méthodes [13].

Les méthodes évolutives adoptent souvent trois opérateurs d'évolution : la sélection, la mutation et la recombinaison. Le schéma type peut être vu ainsi [21] :

1. choisir aléatoirement une population de solutions. Chaque solution est une partition munie d'une certaine qualité (sélection)
2. générer de nouvelles solutions sur la base des anciennes (mutation et recombinaison)
3. refaire (2) jusqu'à atteindre un critère d'arrêt

L'exploration de l'espace des solutions possibles donne l'avantage à ces méthodes d'atteindre toujours la solution optimale mais les inconvénients majeurs restent le coût élevé et la mauvaise adaptation pour des jeux de données larges [19].

3.6.3 Les méthodes hybrides

Dans la pratique un algorithme n'est jugé satisfaisant qu'après avoir évalué le résultat qu'il fournit. Donc pour raffiner le résultat obtenu souvent plusieurs

algorithmes différents sont combinés pour constituer ce qu'on appelle une méthode hybride. A titre d'exemple [13] :

- Utiliser la méthode des k-means pour générer volontairement un nombre k de clusters le plus large possible ensuite appliquer une méthode hiérarchique ascendante pour réaliser un regroupement à base des centroïdes obtenus par k-means. L'avantage est que l'algorithme hiérarchique ne considère que les représentants au lieu de l'ensemble de données en entier et ainsi sa complexité est réduite.

- Utiliser le k-means pour construire une partition, donc une solution, ensuite appliquer un algorithme génétique pour converger vers la solution optimale. Ceci permet de réduire l'espace des solutions (Lee et Antonsson 2000) [19].

3.7 Conclusion

Rappelons que le clustering est "data driven", par conséquent un état de l'art sur les méthodes de clustering ne pourra jamais être exhaustif car beaucoup de méthodes ont été développées pour qu'elles puissent être adaptées à des contextes bien particuliers. Notons aussi que les algorithmes traditionnels restent dans la plus part des cas une origine d'inspiration des nouvelles techniques.

Nous concluons dans un premier temps que les différents algorithmes de clustering se distinguent principalement par les propriétés suivantes :

- le type des attributs manipulés,

- le volume et la dimension des données,
- la robustesse aux bruits : certains algorithmes ne perdent pas leur performance si le jeu de données présentent des anomalies telles que : absence de certaines valeurs, données erronées, …alors que d'autres algorithmes sont très sensibles aux anomalies et nécessitent tout un travail de préparation des données,

- la facilité ou non de l'interprétation des résultats : l'interprétation des résultats permet d'attribuer facilement un sens sémantique aux clusters obtenus,

- la capacité de traiter les observations isolées,

- la complexité de calcul : un algorithme de grande complexité a un coût de calcul élevé et s'adapte mal aux cas de jeux de données importants,

- la dépendance ou non des paramètres prédéfinis : certains algorithmes exigent la définition de certains paramètres initiaux comme le nombre k de clusters de k-means. Cette propriété rend l'algorithme sensible au choix des paramètres de départ,

- la méthode ou la stratégie de regroupement : elle définit le principe général du regroupement (partitionnement, hiérarchique, …) et la structure de la partition fournie,

- déterministe ou non : l'algorithme est déterministe si son résultat reste inchangé dans le cas de plusieurs exécutions sur le même jeu de données sinon il est non déterministe,

- incrémental ou non : un algorithme non incrémental est très dépendant de l'ordre d'arrivée des données, par conséquent, si les données évoluent dans le temps, il devient nécessaire d'adapter le modèle à la nouvelle situation ce qui constitue un grand handicap dans le cas des bases de données larges. Contrairement à cette propriété, un algorithme incrémental est insensible à l'ordre d'arrivée des données. Ceci a pour avantages :

- ✓ de balayer une seule fois les données
- ✓ d'assurer l'interruptibilité et l'incrémentalité c'est-à-dire la possibilité de suspendre l'algorithme et sauvegarder le contexte pour continuer plus tard avec de nouvelle données
- ✓ de mieux gérer l'espace mémoire
- ✓ de traiter des bases de données larges ou les données qui arrivent en flot.

- hard ou fuzzy (dure ou floue) : dans une approche hard les clusters formés sont mutuellement exclusifs alors que dans une approche fuzzy le regroupement se fait sur la base d'une fonction qui exprime un degré d'appartenance. De ce fait, un objet peut appartenir à plus d'un groupe mais avec des degrés différents. Un regroupement non exclusif peut être très expressif dans un contexte subjectif,

- le critère d'arrêt : certains algorithmes exigent un critère de qualité pour l'arrêt, comme le cas des algorithmes de partitionnement, alors que d'autres exigent un critère d'agrégation comme c'est le cas des algorithmes hiérarchiques.

Nous soulignons que beaucoup d'algorithmes, bien qu'ils soient différents, peuvent partager certaines propriétés mais la propriété qui définit la structure de regroupement reste la plus importante dans la différentiation des algorithmes.

Nous remarquons aussi que la majorité des algorithmes se basent soit sur la structure par partitionnement soit sur la structure hiérarchique mais avec des principes différents.

Nous concluons enfin, que le clustering est un processus très sensible au choix de l'algorithme. Ce choix est une étape très difficile surtout que l'expérience des experts du domaine a montré que l'efficacité d'un algorithme ne peut être jugée qu'après son expérimentation.

CHAPITRE 4
CONSTRUCTION D'UN MODELE PREDICTIF DU PARCOURS SCOLAIRE A L'AIDE DU CLUSTERING

4.1 Introduction

Nous commençons, dans un premier temps, par rappeler que dans ce travail nous utilisons les techniques de data mining pour la modélisation du parcours scolaire, dans le but de prédire le succès et l'échec d'une population d'apprenants et de donner une explication sur les facteurs les plus déterminants du cas d'échec.

Nous présentons dans ce chapitre une simulation sur un ensemble de données réelles obtenues à partir de la base de données du service de scolarité du département informatique de la faculté des sciences de l'université SAAD DAHLAB de Blida. Notre objectif est de construire un modèle prédictif de l'échec et du succès des étudiants de la deuxième année vers la troisième année du système classique. Nous envisageons par ce modèle répondre à la question suivante : étant donné un étudiant en deuxième année, quelle est sa probabilité de réussir et de passer en troisième année ainsi que sa probabilité d'échouer et de refaire l'année, et quels sont les facteurs les plus saillants favorisant le cas d'échec.

La démarche que nous adoptons se résume globalement dans les étapes suivantes (figure 4.1) :

- définition des données et sélection des variables pertinentes,
- partition des données sélectionnées en deux échantillons, un échantillon d'apprentissage pour la construction du modèle et un échantillon d'évaluation pour estimer le pouvoir de prédiction du modèle,
- application du clustering sur le premier échantillon en utilisant l'algorithme proposé t-means ; les données sont alors partitionnées en un ensemble de classes que nous appelons classes de références,

- utilisation de la technique des réseaux bayésien pour calculer les probabilités de succès et d'échec de chaque classe de référence,
- évaluation du modèle obtenu (les classes de références et le réseau bayésien) par les données du deuxième échantillon,
- et enfin l'interprétation des résultats.

Ces différentes étapes, décrites dans la figure 4.1, seront développées dans la suite du chapitre.

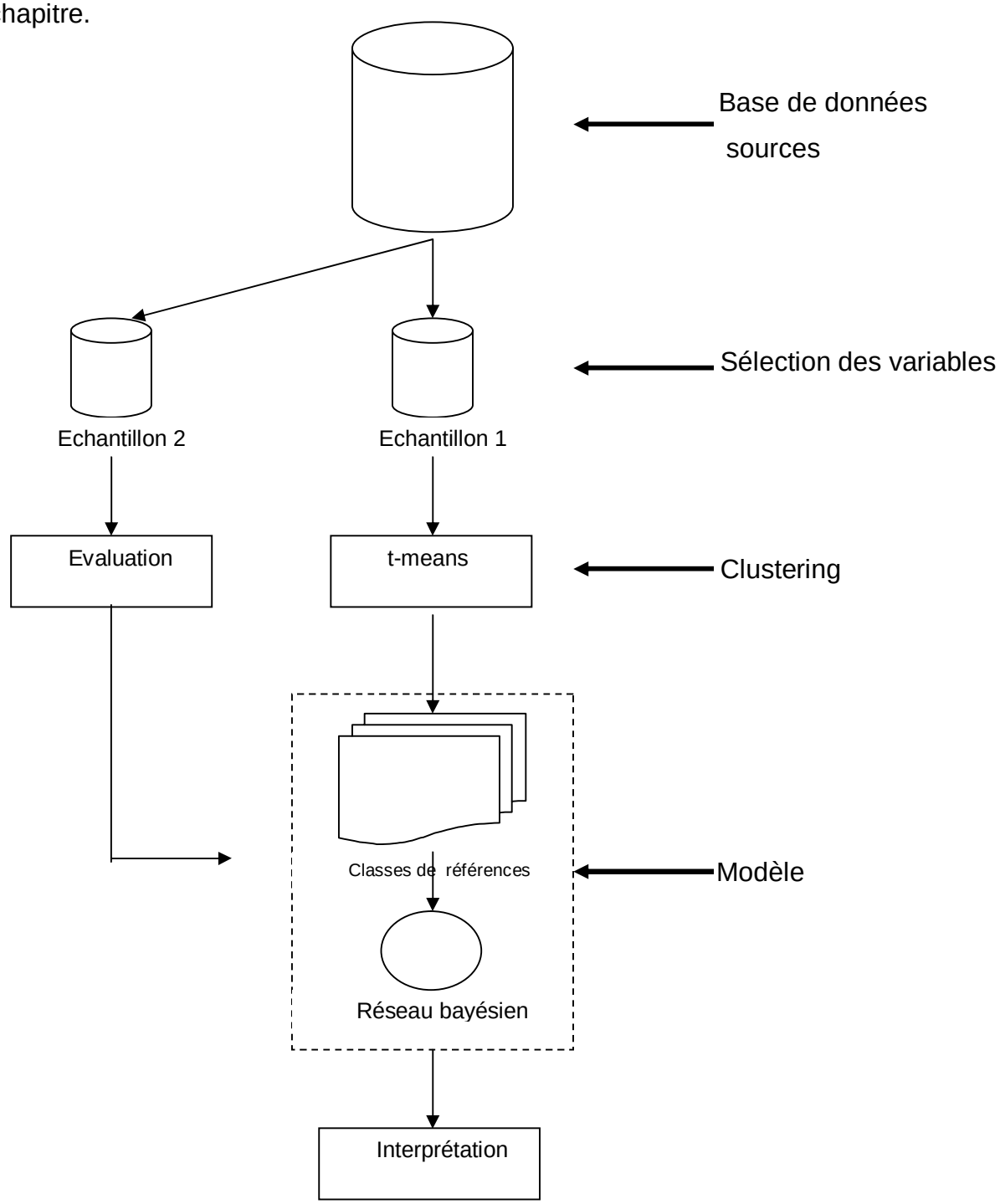


Figure 4.1 : Etapes de modélisation et de simulation du parcours scolaire

4.2 Définition des données et sélection des variables

Nous disposons, au départ, d'une base de données sources de type relationnel qui décrit par ses différents champs (appelés aussi variables ou attributs) une population estudiantine. Ces attributs peuvent êtres de deux types : statique et dynamique.

Les attributs statiques décrivent les individus comme le nom, le prénom, la date de naissance, le type et la mention du bac, …

Les attributs d'évènements décrivent les comportements à travers les différentes notes d'évaluation obtenues comme les examens de l'année, les examens de synthèse ou de rattrapage, les travaux pratiques, les travaux dirigés, …

L'année universitaire du système classique est sanctionnée par trois sessions :

- la session normale : elle se passe généralement au mois de juin où une moyenne est calculée pour chaque module enseigné sur la base de tout les contrôles effectués durant l'année,

- la session de synthèse : elle s'effectue généralement au mois de juillet et ne concerne que les étudiants recalés dans la session normale,

- la session de rattrapage : elle s'effectue au mois de septembre pour les étudiants qui n'ont pas encore réussi à obtenir le passage vers l'année supérieure.

Les données que nous sélectionnons pour les besoins de la simulation se limitent aux données de la session normale que nous jugeons plus pertinentes et les moins biaisées. En effet, ces données sont les premiers indicateurs qui vont permettre de porter un jugement sur un étudiant et par conséquent de faire la meilleure description possible d'une population estudiantine. Nous voulons aussi par ce choix construire un modèle prédictif avec le minimum de données possible.

Les données sélectionnées sont les moyennes modulaires de la session normale concernant une sous-population à savoir les étudiants de la deuxième année. Ces modules constituent les variables du travail et ils sont au nombre de neuf que nous présentons dans le tableau 4.1 dans l'ordre décroissant des coefficients.

code	Intitulé des modules	Coefficient
V1	Algorithmique	5
V2	Architecture des ordinateurs	5
V3	Composants de base des calculateurs	3
V4	Logique mathématique	3
V5	Système d'informations	3
V6	Maths pour informatique	2
V7	Traitement du signal	2
V8	Probabilités et statistiques	2
V9	Anglais	1

Tableau 4.1 : Modules de la deuxième année informatique

Chaque individu (ou étudiant) dans notre ensemble de données sera représenté par un vecteur de neuf variables. Les données sont alors extraites de la base de données sources, contrôlées, nettoyées et regroupées en deux échantillons sous forme de fichier plat.

4.3 Définition des échantillons

Rappelons que nous utilisons deux échantillons de données :

- le premier est utilisé pour l'apprentissage et la construction du modèle, c'est un échantillon de 404 individus obtenu par regroupement de trois promotions, la promotion 2001, 2002 et 2003 respectivement de 150, 128 et 126 individus.

- le second est utilisé pour l'évaluation du modèle, il est formé de 134 individus obtenus à partir de la promotion 2004.

4.4 Le clustering

Le clustering (ou classification non supervisée) est une étape importante dans la démarche et dont dépend fortement la qualité du modèle. L'objectif de cette étape est de synthétiser les données de départ en un ensemble restreint de classes que nous appelons classes de références et qui vont nous permettre par la suite d'appliquer une classification supervisée sur l'échantillon d'évaluation. La difficulté majeure dans cette étape réside dans le choix de l'algorithme. En effet, comme nous l'avons déjà montré précédemment dans le chapitre 3, la littérature propose toute une diversité d'algorithmes, chacun a ses avantages et ses inconvénients sans que nous puissions dire qu'une version est meilleure qu'une autre. De ce fait, le choix d'une version est fortement lié au contexte de l'application.

Pour réaliser cette étape, nous aurons besoin d'un algorithme de clustering qui doit répondre aux critères suivants :

- ✓ le cas d'un jeu de données large,
- ✓ le cas d'une dimension importante des données,
- ✓ le cas des attributs numériques,
- ✓ la complexité doit être linéaire,
- ✓ et enfin l'algorithme doit être incrémental et déterministe.

Parmi les nombreux algorithmes étudiés, nous remarquons que les critères que nous venons de définir donnent un grand avantage à l'algorithme des k-means. Il est considéré comme un standard dans le domaine du clustering et répond parfaitement aux quatre premiers critères mais ceci ne l'empêche pas de souffrir de quelques insuffisances dont voici les principales :

- le choix des paramètres initiaux : le nombre k des clusters doit être défini à priori; de même les k centres initiaux doivent être choisis d'une manière aléatoire,

- la qualité du partitionnement est fortement dépendante du choix de la valeur de k et des centres initiaux des différents clusters.

- les points isolés, ou les exceptions, sont mal pris en charge et peuvent affecter négativement la qualité du résultat de partitionnement,

- l'arrivée de nouvelles données peut remettre en cause l'incrémentalité de l'algorithme et par conséquent la structure de la partition déjà construite précédemment et par conséquent son amélioartion.

- L'algorithme n'est pas déterministe car il peut générer une nouvelle solution à chaque nouvelle exécution.

Pour pallier à ces inconvénients, nous proposons dans le cadre de ce mémoire une version améliorée du k-means que nous baptisons t-means, t pour dire *Threshold* en anglais ou seuil en français.

4.4.1 Présentation de t-means

Le principe général de la version proposée, t-means, se base sur la définition à priori d'un seuil de regroupement, ensuite on laisse à l'algorithme le soin de découvrir automatiquement les k clusters. L'idée est une inspiration des méthodes basées sur la densité et de ceux basées sur le partitionnement hiérarchique, dont voici le principe de fonctionnement.

Partant du fait que le centre de gravité d'un nuage de points a la tendance d'être attiré vers la plus grande concentration des points, l'algorithme cherche parmi les points du nuage le point le plus proche au centre de gravité et le considère comme étant le premier centroïde initial.

Par la suite, l'algorithme procède à un regroupement de tous les points dont la distance est inférieure à un certain seuil prédéfini et recalcule ensuite le centroïde.

Ce regroupement est réitéré jusqu'à la stabilité du centroïde pour donner enfin naissance à un cluster.
L'algorithme reprend d'une manière itérative l'étape précédente jusqu'à ce que tous les points soient classés.

Dans une seconde étape, il procède à une redistribution éventuelle qui peut toucher surtout les points des frontières. Ceci, bien sur, dans le cas ou un point frontière est plus proche à un centre voisin plutôt qu'au centre auquel il a été assigné. Cette étape est réalisée grâce au k-means standard.

Pour ne garder dans la partition que l'information pertinente, l'algorithme procède dans une phase finale à l'élimination des points isolés et des clusters non significatifs.

Dans le cas où le nombre de points isolés devient significatif, on peut penser soit à créer de nouvelles classes si cela est nécessaire, soit à les traiter comme étant une classe des exceptions.

Le schéma de la figure 4.2 est une illustration du principe de fonctionnement de l'algorithme t-means :

Début de l'algorithme

Initialisation :

a) définir un seuil t,
b) k=1 // k représente le numéro du cluster courant

étape 1 : construction de la partition

a) calculer le centre de gravité G de tous les points non encore groupés
b) prendre le point C_k le plus proche à G comme étant un centre initial
c) grouper tous les points p telle que la distance $d(p,C_k) \leq t$
d) recalculer le centre C_k
e) répéter (c) et (d) jusqu'à la stabilité du centre C_k
f) k=k+1
g) reprendre de (a) jusqu'à ce que tous les points soient affectés à un cluster

étape 2 : élagage et amélioration de la partition

1. appliquer k-means standard pour améliorer la partition,
2. éliminer les points isolés et les clusters non significatifs.

Fin de l'algorithme.

Figure 4.2 : Schéma illustratif de t-means

Pour vérifier la validité de t-means nous procédons à son test sur un ensemble de données bidimensionnelles et synthétiques conçu de telle sorte que la classification optimale soit connue.

Prenons un nuage de soixante trois points répartis aléatoirement autour de six centres selon un seuil t=2.5 ; formant ainsi une partition de six clusters (k=6) avec la présence de trois points isolés (figure 4.3)

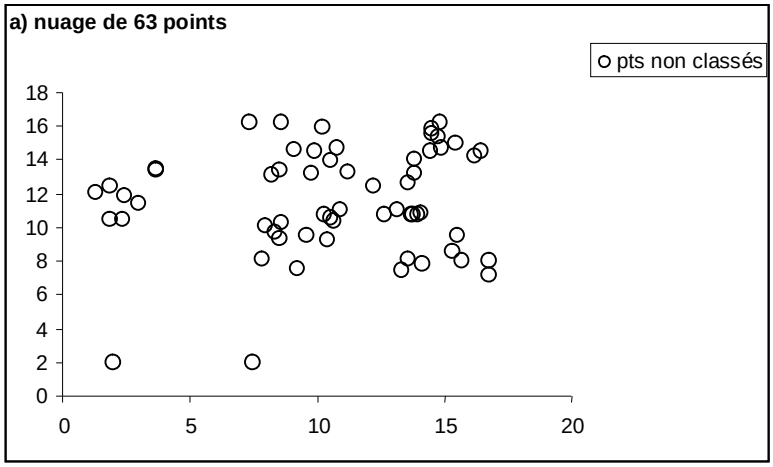


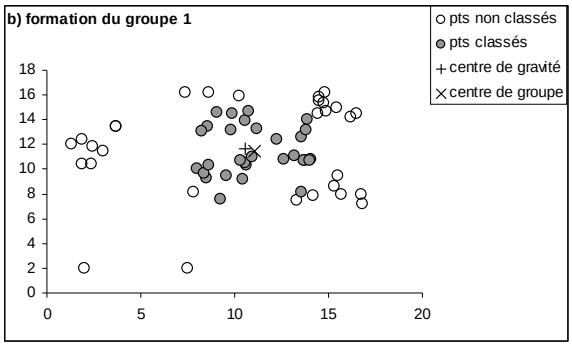
Figure 4.3 : Nuage de test de 63 points

Les graphiques 1 à 8 de la figure 4.4 montrent les itérations de la première étape de l'algorithme en considérant le même seuil et en prenant comme mesure la distance euclidienne définie par :

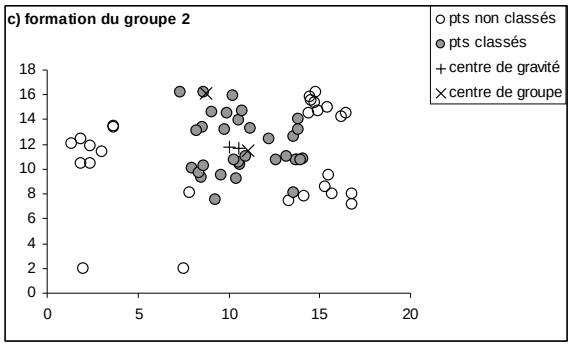
$$d(X,Y) = \sqrt{\sum_{i=1}^{M} \left| x_i - y_i \right|^2} \qquad (13)$$

Avec X et Y deux points de l'espace de dimension M

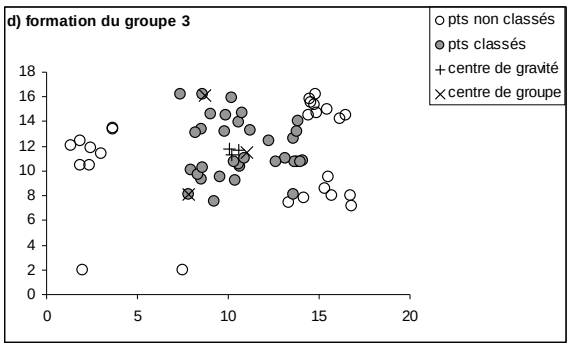
A chaque itération, il y a apparition d'un groupe potentiel.



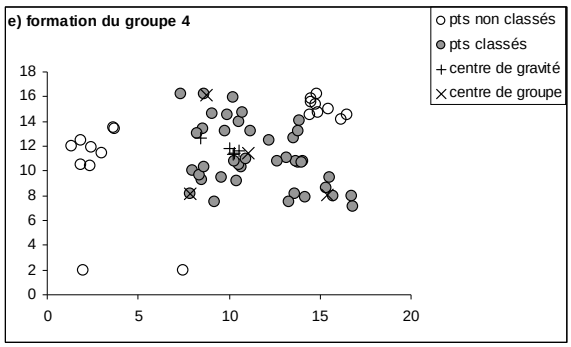
(1)



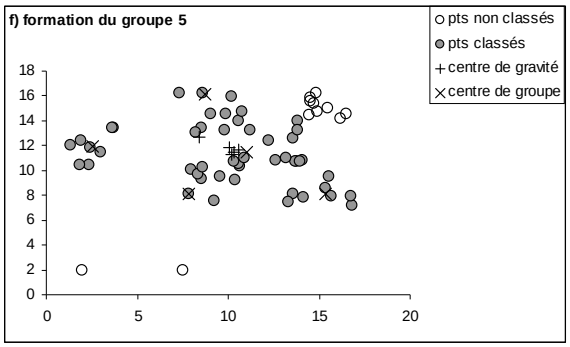
(2)



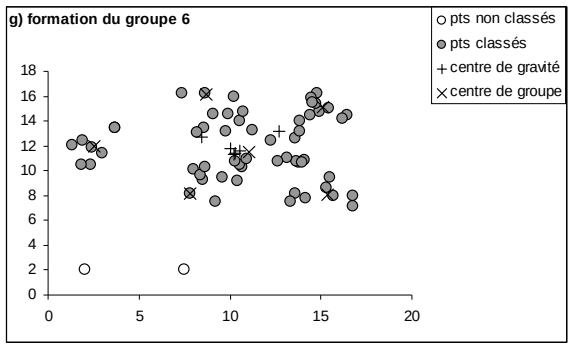
(3)



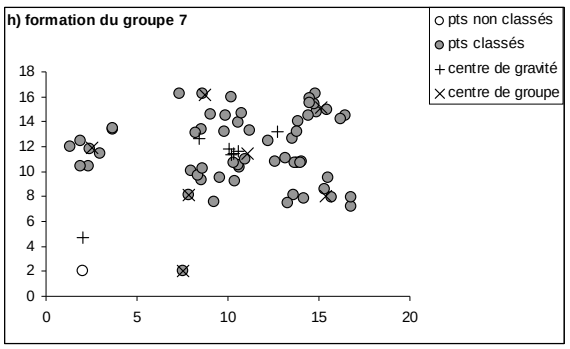
(4)



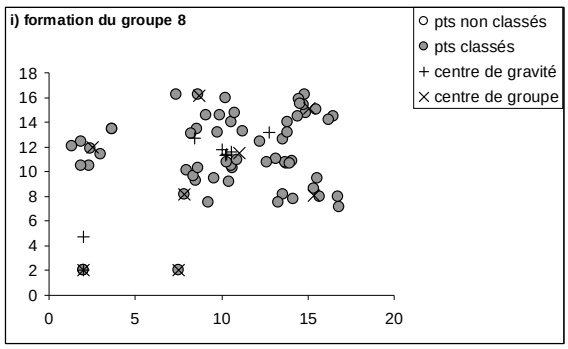
(5)



(6)



(7)



(8)

Figure 4.4 : Déroulement de la première étape de t-means (t=2.5)

La figure 4.5 montre le résultat de la deuxième et dernière étape ; elle a pour but de raffiner les groupes construits précédemment. Nous remarquons que t-means a réussi à reconstituer et à améliorer le partitionnement initial du nuage de points de départ.

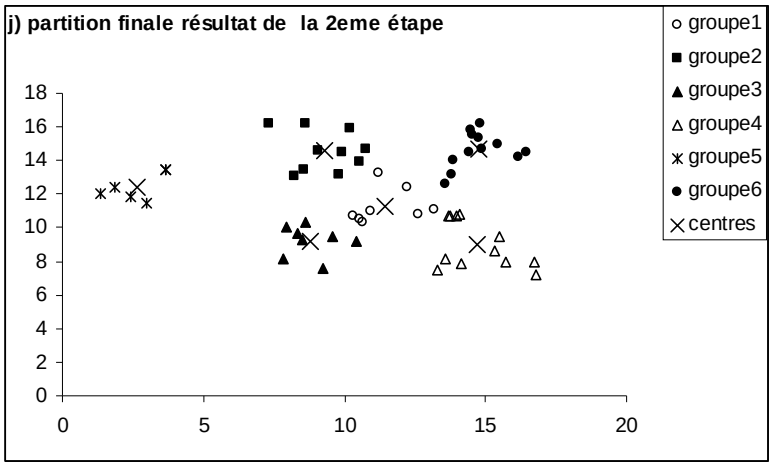


Figure 4.5 : Résultat de la deuxième étape de t-means (t=2.5)

La figure 4.6 montre que la courbe de tendance sur les centres des groupes suit parfaitement la courbe du nuage des points. L'écart remarqué au début des deux courbes est causé par les points isolés que t-means a négligés .

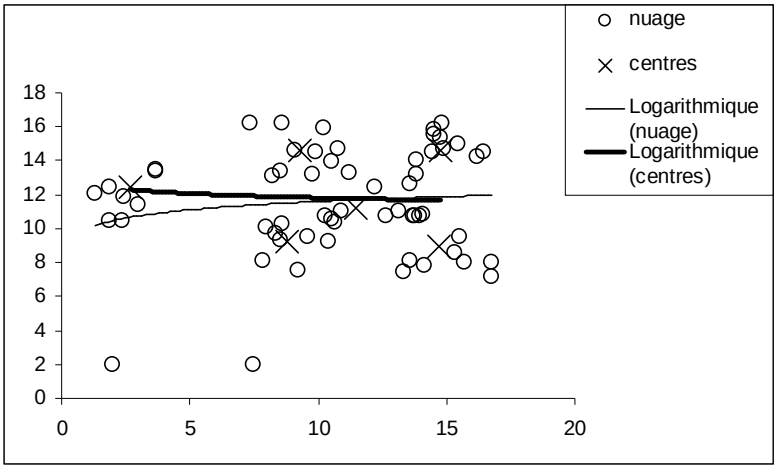


Figure 4.6 : Graphe comparatif par courbe de tendance logarithmique

4.4.2 Discussion

- la complexité : t-means est rapide et d'une complexité linéaire car dans sa première phase, contrairement à k-means, à chaque itération les points déjà classés sont ignorés. La deuxième étape est juste une amélioration du résultat obtenu par la première étape qui touche surtout les points frontières,

- l'incrémentalité : elle est vérifiée par le fait que les nouvelles données qui arrivent sont affectées aux classes suivant le seuil de la partition, et si aucune classe ne peut les accueillir elles sont alors considérés comme étant des points isolés,

- Le déterminisme : t-means est déterministe car chaque nouvelle exécution, avec le même seuil, aboutit toujours au même résultat,

- Les paramètres initiaux : le seuil t est le seul paramètre prédéfini ; il définit le critère d'admission d'un objet dans une classe. Nous pouvons interpréter son effet comme suit :

    - Plus t augmente plus k diminue, l'homogénéité des classes se dégrade et le nombre de points isolés diminue (sur partitionnement) .
    - Plus t diminue plus k augmente, l'homogénéité des classes s'améliore et le nombre de points isolés augmente (sous partitionnement) .

Par conséquent, une recherche empirique de la meilleure valeur de t permettra d'éviter d'aboutir soit à un sur partitionnement soit à un sous partitionnement.

Pour trouver un compromis entre t et k, nous faisons varier le seuil t dans un intervalle et nous retenons celui dont la partition répond le mieux aux quatre critères de qualités suivants :

- Le maximum d'homogénéité à l'intérieur des classes (distance intra-classe) .
- Le maximum d'hétérogénéité entre les classes ((distance inter-classes) .
- Le minimum de classes possibles.
- Le minimum de points isolés possibles.
- Le maximum d'absorption des individus dans les classes.

4.4.3 Validation de t-means par des données réelles

Nous allons maintenant procéder à une validation de l'algorithme proposé sur un ensemble de données réelles de 713 individus de dimension 10 que nous avons construit à partir de la base de données du tronc commun de la promotion 2005. Ceci se fera essentiellement en utilisant l'approche hybride évoqué dans le chapitre 3 et qui consiste à faire converger par un algorithme génétique des solutions obtenues par l'algorithme k-means vers une solution optimale. Le principe que nous adoptons se résume dans les trois étapes suivantes :

- Nous appliquons dans un premier temps l'algorithme t-means sur l'ensemble de données avec un seuil t=20 en utilisant comme mesure la distance euclidienne ; le résultat est une partition P_t de k_t classes avec k_t=6.

- Dans un deuxième temps, nous appliquons le k-means standard plusieurs fois (disons N) avec un k fixé à k_t, obtenu précédemment, pour construire une population de partitions que nous appelons P telle que P = {P_1,P_2, … P_N} et nous retenons parmi les différentes partitions P_i (i=1,N) de P la partition P_k comme étant la meilleure solution dont la valeur de l'intra classe est la plus faible.

- et enfin, dans un troisième temps nous appliquons un algorithme génétique sur la population P, obtenue précédemment, pour construire à partir des différentes solutions P_i, représentant des optimums locaux, la meilleure partition possible que nous appelons P_g, et qui tend vers la solution la plus optimale possible.

Nous résumons à la fin les résultats dans un tableau comparatif entre les trois partitions produites (P_t, P_k, P_g) sur la base du temps d'exécution et des inerties à savoir l'inertie intra classe basée sur le rayon moyen et l'inertie inter classes basée sur le lien moyen défini précédemment dans le premier chapitre.

4.4.3.1 Mise en œuvre de l'algorithme génétique

Les algorithmes génétiques sont dans la famille des algorithmes méta heuristiques, leur le but est d'obtenir une solution convenable dans un temps acceptable et de concevoir des systèmes artificiels possédant des propriétés similaires aux systèmes naturels [29].

Contrairement aux méthodes traditionnelles qui cherchent à trouver une solution analytique exacte ou une bonne approximation numérique, un algorithme génétique tente de trouver un optimum en faisant évoluer d'une manière itérative une population de solutions initiales où chaque itération donne naissance à une nouvelle génération qui optimise au mieux une certaine fonction d'évaluation. La création d'une nouvelle génération se fait par application d'opérateurs génétiques stochastiques et qui sont : la sélection, le croisement et la mutation (figure 4.7)

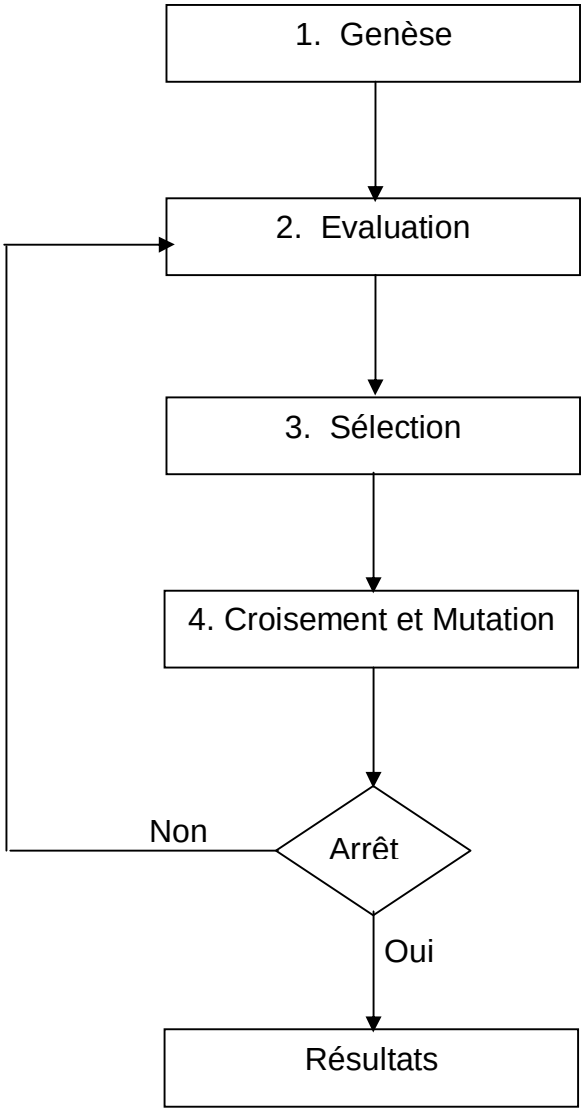


Figure 4.7 : Principe général des algorithmes génétiques [22]

Pour mettre en œuvre cet algorithme nous avons besoin de définir essentiellement :

- le principe du codage des solutions appelées aussi individus,
- les méthodes d'applications des opérateurs génétiques,
- la fonction d'évaluation (fitness function) appelée aussi fonction d'adaptation,
- et enfin le critère d'arrêt des itérations.

Nous nous limitons dans ce qui suit à la méthode d'implémentation que nous avons adopté. Le lecteur intéressé trouvera dans [22] une présentation détaillée sur les principes de codage et les différentes manières d'application des opérateurs génétiques appuyée par des exemples et des comparaisons entre une résolution déterministe et une résolution avec des algorithmes génétiques.

4.4.3.2 Implémentation de l'algorithme génétique

- population initiale : comme nous l'avons déjà montré précédemment, nous appliquons d'une manière itérative le k-means standard pour construire une population initiale de 20 individus tous différents, où chaque individu représente une partition de six centres. Pour l'algorithme génétique, un individu est un ensemble de six chromosomes dont chacun est formé par un seul gène qui est la valeur réelle du centre.

- codage : un individu P_i est défini par : $\{C_{i1}, C_{i2}, C_{i3}, C_{i4}, C_{i5}, C_{i6}\}$ ou chaque C_{ij} est une valeur réelle qui représente le $j^{ième}$ centre de la partition i.

- fonction d'évaluation : nous utilisons l'inertie intra classe définie dans le chapitre 1 comme fonction pour mesurer l'adaptabilité d'un individu. Plus l'inertie d'un individu est faible plus il est adapté et plus il a de chance de rester dans le circuit de génération.

- sélection : le rôle de cet opérateur est de choisir parmi les N individus de P les individus les mieux adaptés qui vont participer dans le croisement pour être dupliqués dans la nouvelle génération. Dans notre implémentation nous gardons toujours une population de N individus qui va rester dans le cycle de reproduction.

- croisement : Les individus sélectionnés précédemment vont être regroupés maintenant en couples pour les faire reproduire par recombinaison des informations présentes dans le patrimoine génétique. Nous utilisons dans notre cas un croisement en un point dont le principe consiste à choisir au hasard un point de découpe des parents d'un couple en deux segments (figure 4.8), ensuite de faire un échange du segment du premier parent avec son homologue du deuxième parent. Ce processus va permettre à chaque couple d'avoir deux descendants et ainsi la taille de la population P va doubler et passe à 2*N individus.

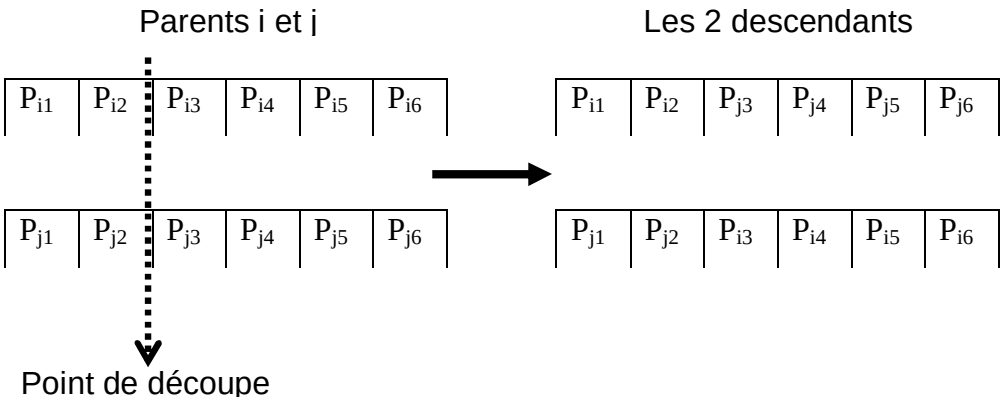


Figure 4.8 : Croisement en un point

- remplacement : le rôle de cet opérateur consiste simplement à faire introduire les descendants dans le cycle de reproduction. Pour ce faire, nous appliquons un remplacement dit élitiste qui consiste à faire remplacer les parents les moins adaptés par les descendants les plus adaptés, et ainsi la population de reproduction repasse à N individus.

- critère d'arrêt : nous limitons le cycle de reproduction à 20 générations

Le tableau 4.2 dresse une comparaison entre les trois partitions P_k, P_g et P_t sur la base de trois paramètres : l'inertie intra classes, l'inertie inter classe et le temps d'exécution. Rappelons seulement que la meilleure partition est celle qui minimise la première inertie et maximise la seconde.

Paramètres	P_k	P_g	P_t
L'inertie intra classes	92,1548	90,992	90,145
L'inertie inter classes	112,914	114,077	114,923
Temps d'exécution (sec)	1,391"	120,922"	0,891"

Tableau 4.2 : Tableau comparatif entre P_k, P_g et P_t

Le tableau 4.2 montre clairement que le programme génétique a réussi à apporter des améliorations nettes aux solutions fournies par l'algorithme k-means mais son inconvénient c'est le temps d'exécution qui est à l'ordre de 87 fois plus.

Il montre aussi que la meilleure solution est la partition P_t obtenue par l'algorithme proposé t-means avec un temps record nettement inférieur même à celui réalisé par k-means.

Au cours de nos expérimentations nous avons constaté qu'en augmentant la taille de la population initiale du programme génétique la partition P_g peut devenir meilleure que la partition P_t fournie par t-means au prix d'un coût de calcul très élevé.

Les résultats obtenus démontrent une performance intéressante de t-means et que ce dernier possède tous les atouts pour être adapté aux cas des bases de données larges et que surtout il apporte des solutions aux inconvénients de k-means cités précédemment tels que la non incrémentalité, le non déterminisme et la gestion des points isolés.

4.4.4 Partitionnement des échantillons

Nous rappelons qu'à ce niveau de la simulation nous appliquons t-means pour partitionner le premier échantillon de données (échantillon d'apprentissage) d'une manière non supervisée pour construire une partition de référence.

Pour déterminer le meilleur seuil nous le faisons varier dans un intervalle de la plus petite distance vers la plus grande distance possible soit [0,…,60], où la valeur 60 représente la distance euclidienne entre les deux vecteurs extrêmes de données, le premier ne contient que des notes nulles (0/20) le second que des notes de 20/20.

Nous retenons après plusieurs itérations une partition de 6 classes (tableau 4.3) fournie par un t=12 qui répond au mieux aux critères de qualité cités dans le paragraphe 4.4.2.

classes	v1	v2	v3	v4	v5	v6	v7	v8	v9
1	12.81	11.19	11.99	11.07	10.46	11.49	11.09	10.06	10.52
2	7.88	10.73	12.21	4.62	10.87	11.76	9.51	8.60	9.25
3	6.14	9.04	11.02	11.41	9.45	9.90	9.69	7.93	9.12
4	7.72	9.06	9.83	8.37	9.72	7.30	5.36	6.15	10.42
5	6.21	9.01	10.19	2.77	9.57	9.43	5.84	5.81	8.55
6	1.94	4.15	5.92	1.63	4.54	4.19	2.78	3.00	6.95

Tableau 4.3 : Partition de référence de six classes

Le tableau 4.4 montre pour chaque classe la moyenne calculée sur la base des coefficients des variables et la densité de sa population.

classes	moyenne	Densité (%)
1	10.62	14,76
2	8.77	17,58
3	8.46	20,36
4	7.82	17,30
5	7.03	24,17
6	3.37	5,85

Tableau 4.4 : Moyennes et densités des classes

Sur le second échantillon (échantillon d'évaluation) nous appliquons une classification supervisée tout en gardant le même seuil et en utilisant comme étiquettes la partition de référence précédente (tableau 4.5)

classes	Densité (%)
1	14,18
2	19,40
3	6,72
4	20,15
5	29,10
6	10,45

Tableau 4.5 : Partition de l'échantillon d'évaluation

4.5 Modélisation du parcours scolaire par un réseau bayésien

La modélisation du parcours scolaire en vue de prédire les résultats des étudiants est un problème de gestion de l'incertain. Ce genre de problème est un très vaste domaine de recherche en intelligence artificielle.

Plusieurs méthodologies ont été proposées pour répondre à ce genre de questions, mais seules les approches probabilistes s'adaptent mieux non seulement au raisonnement avec la connaissance et la croyance incertaine, mais aussi à la structure de représentation de la connaissance. Ces approches probabilistes sont appelées "réseaux bayésiens" [23].

Les réseaux bayésiens sont la combinaison des approches probabilistes et la théorie des graphes ; ils permettent de représenter des situations de raisonnement probabiliste à partir de connaissance incertaines [23]. Comparés aux autres outils tels que les réseaux de neurones, la logique floue, … les réseaux bayésiens ont l'avantage d'offrir un formalisme plus intuitif de représentation des connaissances. Ce formalisme consiste simplement à relier des causes et des effets par des arcs orientés, comme par exemple un arc est orienté de A vers B si A est une cause de B.

Un réseau bayésien peut être formellement défini par deux composantes [6] :

- la première est un graphe orienté acyclique où chaque nœud représente une variable aléatoire, et chaque arc représente une dépendance conditionnelle,

- la seconde consiste en une table de probabilités conditionnelles qui exprime la distribution des probabilités conditionnelles des variables.

La construction d'un réseau bayésien consiste alors à [27] :

- définir dans un premier temps les tables de probabilités des variables : ceci relève généralement des observations historiques,
- définir la structure du réseau en établissant les différents liens de causalité entre les variables. Pour ce faire deux approches sont possibles et peuvent se combiner : le recueil d'expertise et l'apprentissage automatique.

La propagation de l'information dans le réseau ainsi que sa quantification par la formule de Bayes permet de faire des inférences de nature incertaine ; ceci se fait par un calcul de la probabilité marginale à posteriori de quelques variables sachant la valeur des variables observées [24] et [25].

Dans le cadre de la simulation nous utilisons le réseau illustré par la figure 4.9 où les six classes obtenues précédemment par clustering deviennent des nœuds ; le problème d'inférence qu'on cherche à résoudre est ramené à un calcul de la probabilité à posteriori de chaque classe. En d'autres termes, nous voulons répondre à la question suivante : Quelle est la probabilité de réussite et de l'échec de chaque classe?

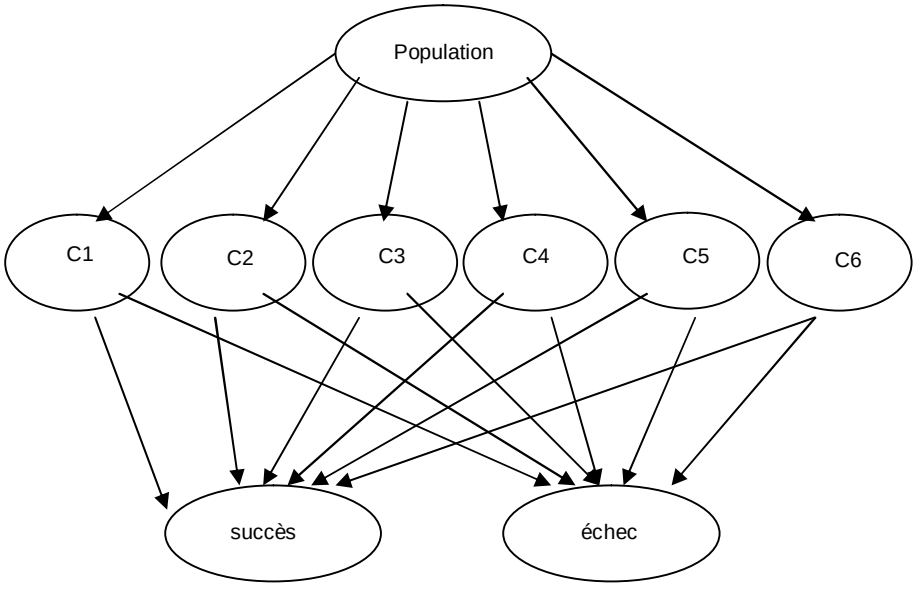


Figure 4.9 : Réseau bayésien du parcours scolaire

Le calcul des probabilités est obtenu par les équations de bayes définies de la manière suivante [26] :

- probabilité de l'échec :

$$P(C_i|Echec) = \frac{P(Echec|C_i).P(C_i)}{P(Echec)} \quad \text{Pour i=1, …,6}$$

Avec $$P(Echec) = \sum_{i=1}^{6} P(Echec|C_i).P(C_i)$$

- probabilité du succès :

$$P(C_i|Succès) = \frac{P(Succès|C_i).P(C_i)}{P(Succès)} \quad \text{Pour i=1, …,6}$$

Avec $$P(Succès) = \sum_{i=1}^{6} P(Succès|C_i).P(C_i)$$

Où

C_i (i=1…6) les six classes de la partition de la population,
P(Succès) et P(Echec) sont calculées par la loi de probabilités totales.

4.5.1 Calcul des probabilités conditionnelles

Nous procédons, à ce niveau de la simulation, aux calculs des probabilités conditionnelles des deux échantillons de données (échantillon d'apprentissage et échantillon d'évaluation).
Notation :
E : Echec
S : Succès
C : Classe de 1 à 6
Nous remarquons que nous faisons abstraction des cas qui quittent le système.

- distribution des effectifs dans les classes (tableau 4.6 et 4.7)

Classes	Effectif initial	Effectif de l'échec	Effectif du succès
1	58	0	54
2	69	11	50
3	80	17	50
4	68	28	30
5	95	64	20
6	23	17	3
Total	393	137	207

Tableau 4.6 : Distribution des effectifs de l'échantillon d'apprentissage

Classes	Effectif initial	Effectif de l'échec	Effectif du succès
1	19	0	17
2	26	5	20
3	9	3	6
4	27	12	9
5	39	36	2
6	14	12	0
Total	134	68	54

Tableau 4.7 : Distribution des effectifs de l'échantillon d'évaluation

- Calcul des probabilités marginales (Tableaux 4.8 et 4.9)

C	P(C)	$P(E \mid C) \times P(C)$	$P(S \mid C) \times P(C)$
1	0,1476	0,0000	0,1374
2	0,1756	0,0280	0,1272
3	0,2036	0,0433	0,1272
4	0,1730	0,0712	0,0763
5	0,2417	0,1628	0,0509
6	0,0585	0,0433	0,0076
Total	393	137	207

Tableau 4.8 : Distribution des probabilités marginales de l'échantillon d'apprentissage

C	P(C)	P(E| C)	P(S| C)
1	0,1418	0,0000	0,1269
2	0,1940	0,0373	0,1493
3	0,0672	0,0224	0,0448
4	0,2015	0,0896	0,0672
5	0,2910	0,2687	0,0149
6	0,1045	0,0896	0,0000
Total	134	68	54

Tableau 4.9 : Distribution des probabilités marginales de l'échantillon d'évaluation

- Calcul des probabilités totales (tableaux 4.10 et 4.11) :

C	P(E| C)×P(C)	P(S| C)×P(C)
1	0,0000	0,0203
2	0,0049	0,0223
3	0,0088	0,0259
4	0,0123	0,0132
5	0,0393	0,0123
6	0,0025	0,0004
Somme	**0,0679**	**0,0945**

Tableau 4.10 : Probabilités totales de l'échantillon d'apprentissage

Nous obtenons P(E) = 0,0679 et P(S) = 0,0945

C	P(E C)×P(C)	P(S C)×P(C)
1	0,0000	0,0180
2	0,0072	0,0290
3	0,0015	0,0030
4	0,0181	0,0135
5	0,0782	0,0043
6	0,0094	0,0000
Somme	**0,1144**	**0,0678**

Tableau 4.11 : Probabilités totales de l'échantillon d'évaluation

Nous obtenons P(E) = 0,1144 et P(S) = 0,0678

- Calcul des probabilités d'inférences (tableaux 4.12 et 4.13)

C	**P(C E)**	**P(C S)**
1	0,0000	0,2146
2	0,0724	0,2364
3	0,1298	0,2741
4	0,1814	0,1397
5	0,5795	0,1302
6	0,0373	0,0047

Tableau 4.12 : Probabilités d'inférences de l'échantillon d'apprentissage

C	**P(C E)**	**P(C S)**
1	0,0000	0,2654
2	0,0633	0,4272
3	0,0132	0,0444
4	0,1578	0,1997
5	0,6835	0,0640
6	0,0818	0,0000

Tableau 4.13 : Probabilités d'inférences de l'échantillon d'évaluation

- Illustration graphique des probabilités d'inférences (figures 4.10 et 4.11)

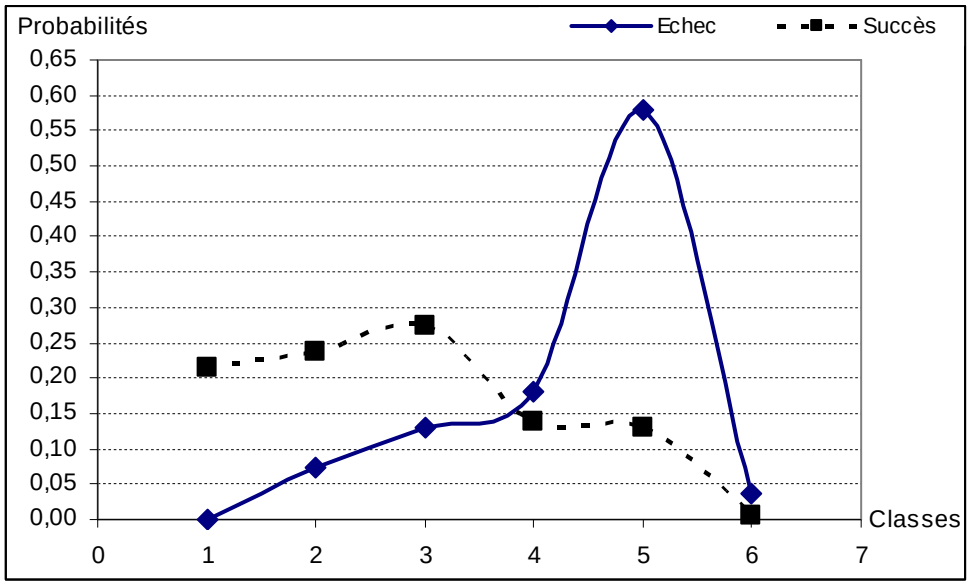


Figure 4.10 : Courbe succès versus échec de l'échantillon d'apprentissage

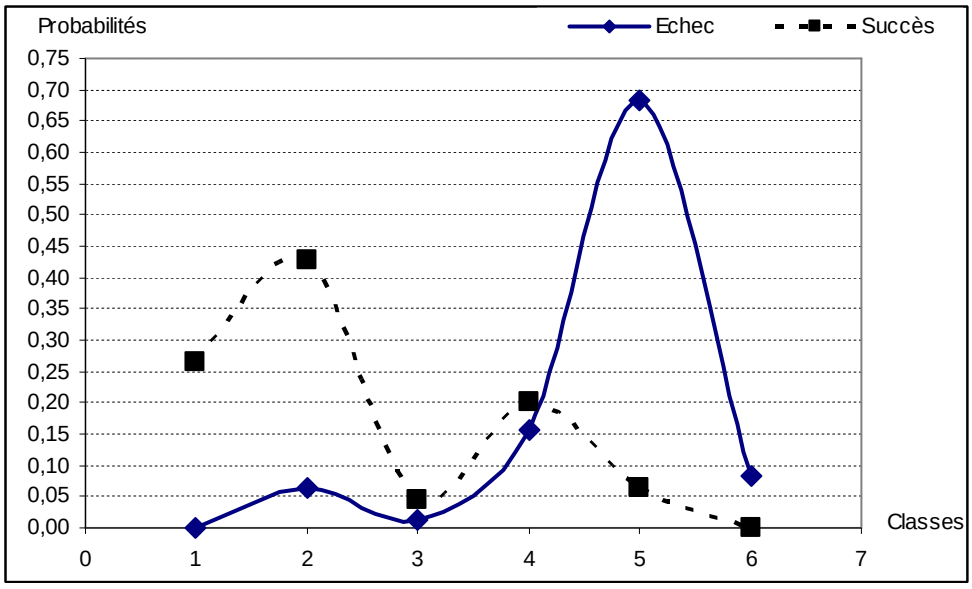


Figure 4.11 : Courbe succès versus échec de l'échantillon d'évaluation

Nous constatons que les figures 4.10 et 4.11 partagent beaucoup de similitudes sauf au niveau de la classe 3 où nous enregistrons un net décalage. Elles montrent aussi, que la classe 4 constitue un seuil où le risque d'échec devient plus important que la chance du succès.

Nous déduisons de cette similitude que les classes de références expriment un niveau de représentativité acceptable des échantillons de données, et par conséquent que l'objectif de l'étape du clustering est atteint.

4.6 Analyse des facteurs prédéterminants de l'échec

Il s'agit maintenant d'étudier pour chacune des classes ses caractéristiques générales dans le but de mettre en évidence les variables déterminantes de l'échec. Pour cela nous adoptons la méthodologie suivante :

Nous prenons comme référence les deux classes symétriques C1 et C6, la première représente le succès et la seconde l'échec et nous cherchons à définir pour les quatre classes restantes (C2, C3, C4 et C5) les cinq espaces suivants (figure 4.12) :

- Espace échec (E6) : l'ensemble des variables communes avec C6.
- Espace frontière échec (F6) : l'ensemble des variables formant une frontière avec C6.
- Espace propre (P) : l'ensemble des variables propres à une classe.
- Espace frontière succès (F1) : l'ensemble des variables formant une frontière avec C1.
- Espace succès (E1) : l'ensemble des variables communes avec C1.

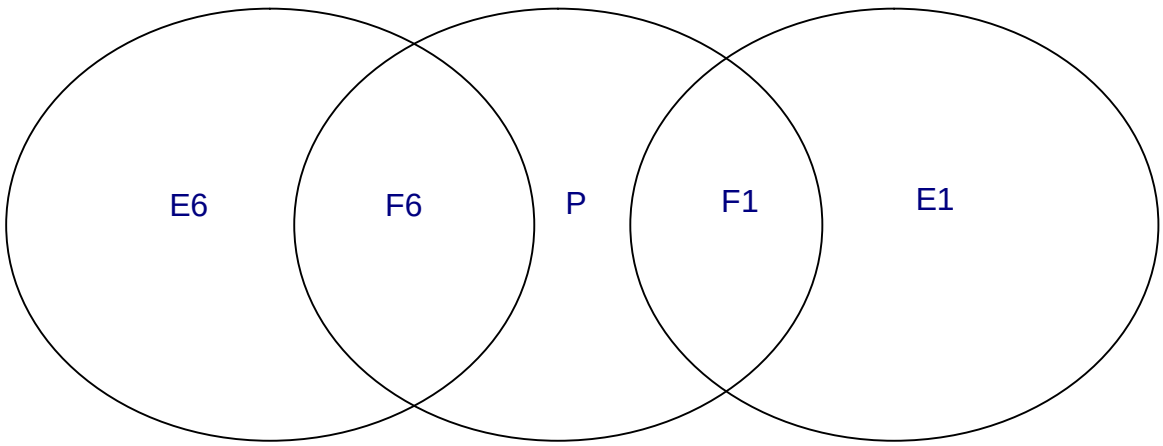


Figure 4.12 : Schéma illustratif des cinq espaces

Le problème que nous voulons résoudre est amené à une distribution des variables de chacune des quatre classes dans les cinq espaces définis suivant le principe de l'algorithme que nous proposons dans la figure 4.13.

Début Algorithme :

1- Pour chaque classe Ci i=2,5

2- E1={ }, E6={ }, F1={ }, F6={ }, P={ }

3- Calculer d1=distance(Ci,C1) et d6=distance(Ci,C6)

4- Faire varier j=1,9

5- Enlever la variable Vj

6- Calculer xd1=distance(Ci,C1) et xd6=distance(Ci,C6)

7- Si |d1-xd1|=0 alors ajouter Vj à E1

8- Si |d6-xd6|=0 alors ajouter Vj à E6

9- Si |d1-xd1|<1 et |d1-xd1|<|d6-xd6| alors ajouter Vj à F1

10-Si |d6-xd6|<1 et |d6-xd6|<|d1-xd1| alors ajouter Vj à F2

11- Autre ajouter Vj à P

12- Remettre la variable Vj

13- Reprendre de (4)

14- Rendre le résultat d'espacement de la classe Ci

15- Reprendre de (1)

Fin Algorithme

Figure 4.13 : Algorithme de distribution des variables dans les espaces

L'application de l'algorithme a permis d'aboutir aux situations suivantes :

- cas de C2 (figure 4.14)

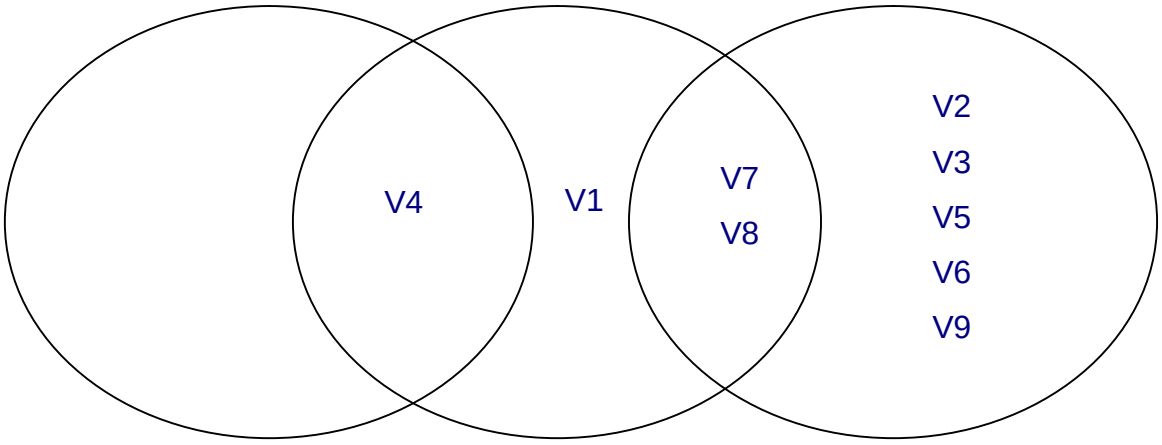


Figure 4.14 : Distribution des variables de la classe C2

La classe C2 est la classe la plus proche à C1 avec une probabilité de succès autour de 0.2364 contre une probabilité d'échec de 0.0724, soit un succès d'environ 3.2 fois plus que l'échec (voir tableau 4.12). La figure 4.14 montre clairement que les variables V1 (module d'algorithmique) et V4 (module de la logique mathématique) constituent les deux premières difficultés qui font défaut aux étudiants

- cas de C3 (figure 4.15)

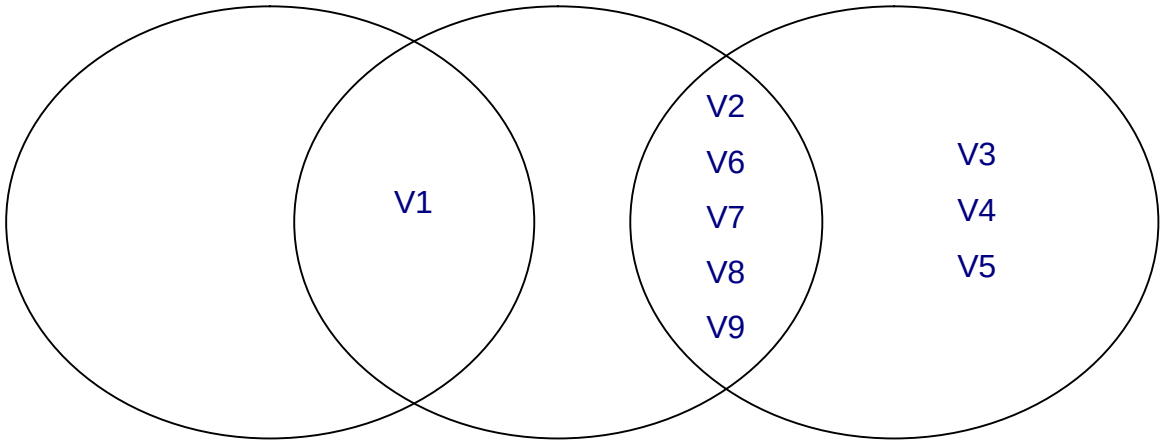


Figure 4.15 : Distribution des variables de la classe C3

La classe C3 (figure 4.15) dispose d'une probabilité de succès de 0.2741 contre une probabilité d'échec de 0.1298, soit un succès d'environ 2.1 fois plus que l'échec, son handicap majeur se situe dans V1. En comparaison avec C2 elle partage moins d'espace avec C1 avec l'avantage d'avoir V4 dans l'espace E1.

- cas de C4 (figure 4.16) :

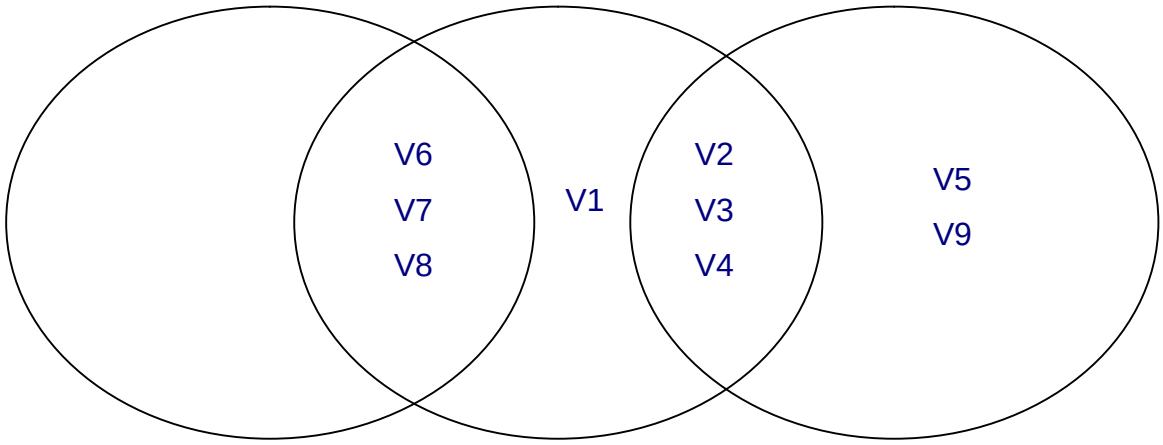


Figure 4.16 : Distribution des variables de la classe C4

La classe C4 (figure 4.16), dispose d'une probabilité de succès de 0.1397 contre une probabilité d'échec de 0.1814, et c'est à partir de cette classe que l'échec devient plus important soit environ 1.3 fois plus que le succès. Nous remarquons qu'en plus de la faible performance en algorithmique elle partage moins d'espace avec C1 et possède trois variables (V6, V7 et V8) dans la frontière de l'échec.

- cas de C5 (figure 4.17) :

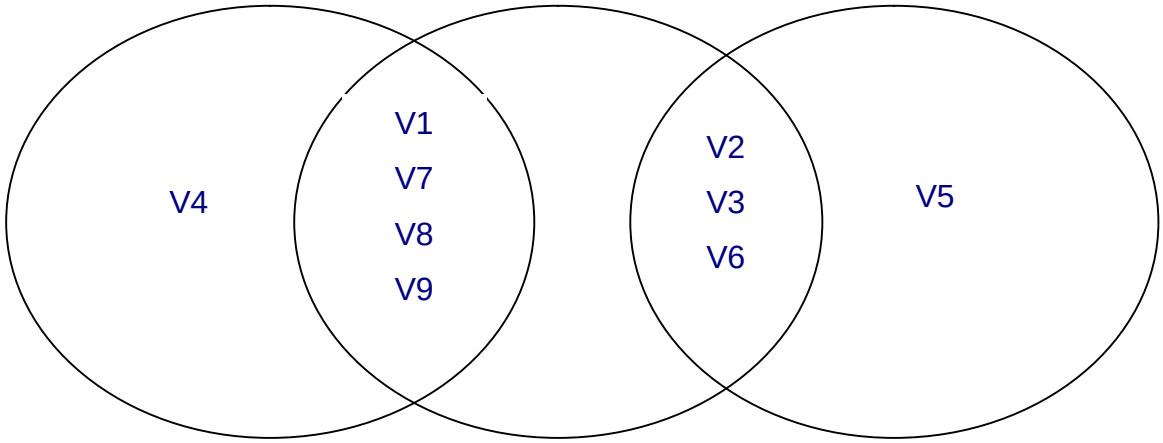


Figure 4.17 : Distribution des variables de la classe C5

La classe C5 (figue 4.17) a une probabilité de succès de 0.1302 contre une probabilité d'échec de 0.5795, soit un échec de 4.4 fois plus que le succès. Elle souffre de beaucoup de lacunes qui se manifestent par un espace très réduit partagé avec C1 et une frontière beaucoup plus importante avec C6 en plus de V4 dans l'espace échec.

4.7 Lecture et interprétation des résultats

- Les classes C2 et C3 montrent que les premiers facteurs déterminants de l'échec sont le module d'algorithmique ensuite le module de logique mathématique. Partant du principe que chaque discipline a ses exigences propres ; l'algorithmique dans une discipline informatique constitue le premier greffon dans la formation d'un informaticien. Il constitue une connaissance de base qui permet de façonner l'esprit d'analyse et un acquis préalable prépondérant pour la réussite dans cette discipline. Nombreux sont les enseignants qui partagent l'idée que les difficultés que

rencontrent beaucoup d'étudiants dans la suite de leur cursus reviennent en grande partie à la non maîtrise de cet acquis et que les acquis préalables doivent constituer des objectifs de court et moyenne durée qu'il faut atteindre en fin de chaque étape d'une formation.

- Globalement la performance des populations des deux classes C2 et C3 est convenable, mais ceci ne les a pas empêché d'avoir une probabilité d'échec significative. Ce constat annonce que le problème de l'échec est beaucoup plus relié à des facteurs pédagogiques tels que :

- ✓ La formulation des concepts des cours.
- ✓ Le manque d'apprentissage.
- ✓ L'absence d'enseignement de certains concepts fondamentaux.
- ✓ Les objectifs de court et moyenne durée ne sont pas explicités.
- ✓ Le manque d'harmonisation des différents enseignements dans la concrétisation des objectifs

- Contrairement à C2 et C3, l'échec dans les classes C4 et C5 semble être beaucoup plus lié à des facteurs reliés aux étudiants. En effet, les figurent 4.14 et 4.15 montrent qu'en sus des difficultés de C2 et C3 les classes C4 et C5 possèdent un espace F6 beaucoup plus important. Dans cet espace nous remarquons la présence de certains modules à savoir V6 (Maths pour informatique), V7 (traitement de signal) et V8 (probabilités et statistiques) pour C4 et la présence de V7, V8 et V9 (anglais) pour C5. Ces modules pénalisants que nous venons d'énumérer reposent en premier lieu sur la nature des acquis dont disposent les étudiants. Ceci démontre que les connaissances dont l'étudiant est doté au début d'un cursus jouent un rôle prépondérant tout au long du déroulement de l'action didactique et que les acquis préalables sont indispensables pour l'atteinte des résultats escomptés. Cette situation accablante est derrière un taux d'échec important ; elle laisse penser d'une part à une déficience chronique du système d'orientation et que ce dernier doit être doté de mécanismes plus élaborés pour qu'il soit en mesure de prendre en charge la diversité et l'hétérogénéité des étudiants, et d'autre part à la nécessité d'avoir des mesures permettant la mise à niveau des connaissances afin de favoriser la réussite.

- Pour C5, la présence de V9 (Anglais) dans la l'espace F6, indique que cette population traîne en plus un handicap linguistique très sérieux.

- Nous remarquons à titre indicatif que les différentes figures montrent une parfaite corrélation :

- ✓ entre V2 (Architecture des ordinateurs) et V3 (Composants de base des calculateurs)
- ✓ et entre V7 (Traitement du signal) et V8 (probabilités et statistiques)

4.8 Conclusion

Au cours de ce chapitre nous avons montré qu'il est possible, en utilisant les techniques de data mining, d'extraire à partir des résultats de l'évaluation sommative un modèle prédictif du succès et de l'échec et dont l'interprétation a montré que les causes d'échec se résument essentiellement en deux sortes de facteurs :

- des facteurs qui peuvent être reliés à des problèmes pédagogiques,
- et d'autres facteurs qui sont reliés directement aux acquis préalables des étudiants.

L'interprétation fait révéler que le premier facteur déterminant de l'échec est le module d'algorithmique. Ce dernier constitue un préalable primordial pour ceux qui débutent dans une spécialité informatique. Cela suppose la présence de problèmes d'ordre pédagogique qui méritent une attention particulière.

L'illustration du modèle par les figures 4.12, 4.13, 4.14 et 4.15 Constitue un portrait qui peut être utile pour réaliser d'autres formes d'évaluation. Parmi ses multiples utilités nous citons les possibilités :

- de raisonner en termes d'objectifs plus qu'en termes de notes.

- d'offrir à l'étudiant un moyen d'autoévaluation lui permettant de se positionner afin qu'il puisse apprécier sa performance et estimer ses chances de réussite et ses risques d'échec.

- d'offrir aussi à l'enseignant un moyen pour mieux connaître ses étudiants et même de juger si ses évaluations sommatives sont en harmonie avec les objectifs explicités.

- de permettre à l'enseignant de suivre la progression de ses étudiants et de les orienter sur les domaines qu'il faut retravailler.

CONCLUSION

Le travail que nous venons de présenter dans ce mémoire est une proposition d'une modélisation d'un parcours scolaire dans le but de prédire les chances de succès et les risques d'échec. Après avoir introduit dans le premier et deuxième chapitre les concepts fondamentaux relatifs au contexte du travail, nous avons exposé dans le troisième chapitre un état de l'art sur les différentes techniques de clustering ce qui nous a permis d'émerger dans cette discipline. Dans le quatrième et dernier chapitre nous avons exposé notre contribution articulée autour de trois axes principaux :

- l'emploi de la technique du clustering pour partitionner les données représentant une population estudiantine en un ensemble de classes (clusters) représentatives.

- L'emploi d'un réseau bayésien pour modéliser le parcours scolaire que nous avons défini comme étant un chemin dans le réseau qui montre l'évolution de chaque classes vers une situation de succès et une situation d'échec. La prédiction dans le réseau est exprimée par la probabilité d'inférence que peut prendre cette évolution.

- La caractérisation des classes dans le but de mettre en évidence les éléments influents sur l'évolution vers la situation d'échec

Pour réaliser le clustering, nous avons proposé dans le cadre de ce mémoire une nouvelle variante de l'algorithme k-means que nous avons baptisé t-means dans laquelle nous avons essayé de remédier aux inconvénients de la version originale tout en préservant ses avantages. La modification principale apportée se situe essentiellement dans la définition d'un seuil de regroupement comme étant le seul paramètre initial qu'il faut définir. Une confrontation des résultats de t-means avec ceux obtenus par la version standard k-means et un algorithme génétique a donné

l'avantage à notre proposition sur le plan complexité du calcul et qualité du partitionnement.

Pour mettre à l'épreuve le modèle prédictif proposé, nous avons réalisé une simulation sur une base de données réelle des étudiants du département informatique de l'université de SAAD DAHLAB de Blida. Elle consistait en l'étude du succès et de l'échec des étudiants de la deuxième année cycle ingénieur vers la troisième année. L'interprétation des résultats obtenus nous a révélé que les facteurs d'échec peuvent être scindées en deux grandes sortes :

- des facteurs liés à des problèmes pédagogiques. En effet, les résultats ont montré que le premier facteur qui fait défaut même pour des étudiants ayant une performance convenable est bien le module d'algorithmique qui constitue un préalable primordial pour continuer dans une filière informatique,

- des facteurs liés aux connaissances préalables des étudiants. Assurément l'échec dans certains modules (les maths, les probabilités, l'anglais, ..) ne peut s'expliquer que par une déficience sur la nature et la qualité des connaissances préalables. Ces facteurs confirment aussi l'urgence de travailler davantage sur un système d'orientation qui prend en charge les différences des étudiants.

Sous un autre angle nous venons de proposer dans ce mémoire un outil qui peut être d'un grand apport aux objectifs assignés à la réforme universitaire, le LMD, en permettant :

- d'effectuer des analyses plus objectives sur la réussite et l'échec scolaire,

- d'offrir aux apprenants un moyen pour se situer, mesurer le chemin parcouru et estimer les chances de succès et les risques d'échec,

- d'exploiter les résultats des épreuves pour réaliser d'autres formes d'évaluation,

- de déceler les causes probables des difficultés rencontrées par les apprenants et d'intervenir au moment opportun.

Les différentes consultations menées avec plusieurs enseignants du département informatique démontrent la pertinence des résultats obtenus. Ceci encourage la continuation des recherches dans l'optique de ce mémoire et d'envisager comme perspective le développement des axes suivants :

- étendre le modèle pour intégrer d'autres facteurs et étudier leur influence tels que les acquis préalables dont disposent les étudiants au début de la formation ainsi que les données de type sociodémographiques et socioéconomiques.

- étendre aussi le modèle sur toutes les étapes de la formation afin de réaliser un tracé prédictif de son début jusqu'à sa fin.

REFERENCES

1. Behrouz Minaei-Bidgoli, Deborah A.Kashy, Gerd Kortemeyer, William F.Punch, "*Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA*", papier, 2003, Boulder, CO.

2. Sani Susanto, Ign.Suharto, et Paulus Sukapto, "*Using fuzzy clustering algorithm for allocation of students*", document "World Transaction on Engeineering and Technology Education" Volume 1 page 245-248, 2002.

3. Agathe Merceron et Kalina Yacef, "*Educational Data Mining: a Case Study*", Article, Procedings of the 12th conference on artificial intelligence in education, Amsterdam, the Netherlands, 2005, IOS Press.

4. Luis Talavera et Elena Gaudioso, "Mining Student Data to Characterize Similar Behavior Groups In Unstructured Collaboration Spaces", Article, Workshop on artificial intelligence in CSCL, 16th European Conference on artificial intelligence, ECAI 2004.

5. Georges EL Helou et Charbel Abou Khalil, " *Data Mining Techniques d'extraction des connaissances*", Rapport de projet, Laboratoire de Recherche en Informatique (LRI) de l'université Paris-Sud et du CNRS, 2004.

6. Jiawei Han and Micheline Kamber, "*Data Mining Concept and Techniques*", Ouvrage, Edition Morgan Kaufmann Publishers, 2000

7. Data Mining '99: Technology Report, "*Introduction to Data Mining and Knowledge Discovery",* by Two Cows Corporation, 1999.

8. Stéphane Tufféry, "*Data Mining et Statistiques décisionnelles l'intelligence dans les bases de données*", livre, édition TECHNIP, 2005.

9. Nong Ye, "*The Handbook of Data Mining*", Arizona State University, IEA, 2003.

10. Claude Montmarquette, Muriel Meunier, Jérôme Schaeffer, Laure Thomas, "*Etude Comparée sur la Réussite Universitaire Québec-Ontario pour la période 1994-1996*", Rapport de Projet, Montréal, Février 2002.

11. Fraçoise Casado, "*L'évaluation Pédagogique au CDI*", Réunion du GTL de Châtelleraut, Juin 2005.

12. Babou Sène, "*L'évaluation pédagogique*", Rapport, Structure de Formation Continuée (SFC), Université de vacances de la SFC, Avril 2003.

13. Nicola BECK, "*application de méthodes de clustering traditionnelles et extension au cadre multicritère* ", Mémoire d'ingénieur, Université Libre de Bruxelles faculté des sciences appliquées, 2006.

14. Thierry Gafner, "*Analyse critique des méthodes classiques et nouvelle approche par la programmation mathématique en classification automatique*", thèse doctorat, Université de Neuchâtel, Faculté de droit et des sciences économiques, 1991.

15. Olivia Parr Rud,"*Data Mining Cookbook, Modeling Data for Marketing, Risk, and Customer Relationship Management*", Wiley Computer Publishing, 2001.

16. Maurice Roux, "*Algorithmes de classification*", Université Paul Cézanne, Marseille, France, 2006.

17. Fabrice Muhlenbach, "*Evaluation de la qualité de la représentation en fouille de données*", thèse doctorat, Université Lumière Lyon II, 2002.

18. Pierre Emmanuel JOUVE, "*Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données*", Thèse Doctorat en Informatique, Université Lumière Lyon2, 2003.

19. Pavel Berkhin, "*Survey of Clustering Data Mining Techniques*", Article, Compagnie Accrue Software, Inc, 2002.

20. LE Anh Tuan, IFI Hanoi, "*Réduction de base de données parla classification automatique*", Rapport de stage, Institut de la francophonie pour l'informatique, 2004.

21. Fabien Moutarde, "*Algorithmes Evolutionnistes*", Cours, Ecole des mines de Paris, 2005.

22. Souket Amédée, Radet François-Gérard, "*Algorithmes génétiques*", rapport de recherche, Université de Nice Sophia Antipolis, département informatique, 2004.

23. D. Cram, M.May, R.Guelton, S.Touch, "*Résumé: Réseaux bayésiens*", Rapport de recherche, Novembre 2005.

24. Eduardo SANCHEZ SOTO, "*Réseaux Bayésiens Dynamiques pour la Vérification du Locuteur*", Thèse doctorat, Spécialité Signal et Images, Ecole Doctorale d'Informatique, Télécommunications et Electronique de Paris, Mai 2005.

25. Salem BENFERHAT, "*Introduction aux réseaux bayésiens*", Support de cours, CRIL université d'Artois

26. Claude Bélisle, "*Processus Aléatoire: Probabilité Conditionnelle*", Support de cours, Université Laval, 2004.

27. Mathieu Hibou, "*Réseaux Bayésiens pour la modélisation de l'apprenant en EIAH: modèles multiples versus modèle unique*", Article, Université René Descartes Paris 5, 2006.

28. Djeknoun Abdelhamid Recteur Université Mentouri Constantine - ALGERIE, "*Vers un espace universitaire euro – maghrébin solidaire: La réforme LMD en*

Algérie état des lieux et perspectives'', 2ème rencontre des Recteurs et des Présidents d'Universités des pays du Maghreb et des Conférences Francophones de l'Union Européenne, Tunis, 1 – 2 Décembre 2006.

29. "*New genetic crossover operator*", Article, site:jedai.afia-france.org