REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE MOULOUD MAMMERI DE TIZI-OUZOU
FACULTE DES SCIENCES

DEPARTEMENT DE MATHEMATIQUE

MEMOIRE
En vue de l'obtention du Diplôme de Master
En Recherche Opérationnelle

Thème

DATA MINING ET DATA WAREHOUSE

Présenté par :

M^{me} BELARIF Hayat
M^{elle} BOUZID Tasssadit

Encadré par :

Mr OUKACHA Brahim

Membre du jury :

Mr AIDENE Mohamed	professeur	Président
Mr OUKACHA Brahim	M.C.A	Rapporteur
Mr KASDI Kamel	M A. A	Examinateur
Mr TALEB Youcef	M.A.A	Examinateur

Promotion : 2011/2012

Louange à Allah le miséricordieux de nous avoir donné le courage, la force et la volonté pour la réalisation de ce mémoire.

Nous tenons à exprimer notre gratitude à notre promoteur Monsieur OUKACHA pour avoir accepté d'encadrer notre travail et pour tous ses précieux conseils.

Nous remercions également les membres de jury qui nous ferons l'honneur d'évaluer notre travail.
Nous remercions enfin, toutes les personnes que nous n'avons pas citées qui ont contribué à la réalisation de ce mémoire.

Table des matières

Introduction générale ……...2

Chapitre I : généralité sur le dada mining

1. Introduction……...5
2. Historique……..5
3. Définition du data méninge……..…….……6
4. Pourquoi la naissance du data mining……...……6
5. Statistiques et Data Méninge…….7
   5.1 Les statistiques…….7
   5.2 Relations entre variables……..7
6. Les tâches du data mining …….8
7. Les étapes du processus de data mining ……9
   7.1. Classification…….9
   7.2. Estimation……..9
   7.3. La prédiction……...……..10
   7.4. Le regroupement par similitudes.……....10
   7.5. L'analyse des clusters……..10
   7.6. La description……10
   7.7. L'optimisation……...10
   7.8. Extraction de règle d'association……..11
8. Techniques du data mining…..…….11
   8.1. Les réseaux de neurones…….11
   8.2. Les arbres de décision…….12
   8.3. Les algorithmes génétiques…….12
   8.4. Les règles associatives…….12
   8.5. Réseaux Bayésien……13
9. Les avantages du data mining ……...……...13
10. conclusion……...…….14

Chapitre II : Data Warehouse et OLAP

1. Introduction…….16
2. Data Waterhouse……..16
   2.1. Historique……...16
   2.2. Définition……16
   2.3. Les caractéristiques de data Waterhouse……17
   2.4. Les avantage de data Waterhouse……17

3. Le Data Mart……...18

Table des matières

3.1. Définition……18
3.2. Intérêt et limites……..18
3.3. Exemples de types de DataMarts……19
4. Modèle multidimensionnel……...19
4.1. Le Cube de Données OLAP…….20
4.2. Faits ……..21
4.3. Dimension……..……21
5. Opérateurs OLAP……..22
5.1. Opérateurs d'agrégation……22
5.2. Opérateurs de présentation pour la navigation…….22
5.2.1. Opérations de forage……22
5.2.2. Opérations de sélection……22
6. Rafraichir le data Waterhouse……23
7. Conclusion……..23

Chapitre III : règle d'association

1. Introduction……...…….24
2. Règle d'association ……....24
2.1. Définition……24
2.2. Utilité des règles d'associations……...……...27
2.3. La recherche de règles d'association ……..27
2.3.1. Les étapes d'extraction de règles d'association…….28
3. Algorithme apriori de recherche de règles d'association……..30
3.1. L'étape de jointure……...……30
3.2. L'étape d'élagage……31
4. Exemple de l'algorithme Apriori……..31
5. Avantages et inconvénients des règles d'association ……....34
5.1. Avantages……34
5.2. Inconvénients…….....34
6. Conclusion…….35

Chapitre IV : La Classification

1. Introduction…….…….36
2. Les domaines d'application de la classification……36
3. La définition de la classification…….36
4. La classification supervisée…….36
4.1. Sélection des variables de la classification supervisée……...36
5. La classification non supervisée…….37
5.1. Sélection des variables en classification non supervisée……....37
6. Critère d'agrégation ……38

Table des matières

6.1. Distance moyenne…….38
6.2. Evaluation d'un système de classification……38
6.3. Distance moyenne……...39
6.4. Evaluation d'un système de classification…….39
6.4.1. Corpus de test……...39
6.4.1.1. Cas supervisé……..……39
6.4.1.2. Cas non-supervisé……..40
7. Choix de la méthode de classification…….41
7.1. Classifieur de Bayes Multinomial……41
7.1.1. La Classification Bayesienne……..41
7.1.2. Le Règle de Bayes……...44
7.2. Classifieur par la méthode des Machines à Vecteurs Support……..46
7.3. Classifieur par la méthode des réseaux RBF……..46
7.4. Classifieur par la méthode adaboost sur le classifieur Naive Bayes Multinomial…….46
8. Conclusion……47

Chapitre V : réseau de neurone

1. Les réseaux de neurones……49
1.1. Introduction……49
1.2. Historique……49
1.3. Définition……50
1.4. Applications……50
1.5. Fonctionnement……..50
2. Modèle biologique…….51
2.1. Définition et structure…….51
2.2. Fonctionnement……..52
2.3. Plasticité synaptique……...……53
3. Étude et synthèse d'un réseau de neurone formel……..53
3.1. Neurone formel……...53
3.2. Fonction d'activation……..54
3.3. Les étapes d'un réseau de neurones……55
3.4. Structure des réseaux de neurones……..56
3.4.1. Réseau mono-couches …….56
3.4.2. Réseau muIti-couches……..57
3.5. Fonctionnement d'un réseau……57
3.6. Apprentissage……..58
3.6.1. Apprentissage supervisé……...58
3.6.2. Apprentissage non supervisé……58
3.7. Normalisation des données……..58

Table des matières

4. Développement d'un réseau de neurones……59
   4.1. Collecte des données……59
   4.2. Analyse des données……60
   4.3. Séparation des bases de données ……...……60
   4.4. Choix d'un réseau de neurones……60
   4.5. Mise en forme des données pour un réseau de neurones……61
   4.6. Apprentissage du réseau de neurones……..61
   4.7. Validation……61
5. Exemple de Classification par les réseaux de neurone.
6. Conclusion…….61
Conclusion générale……...64
Références bibliographique……...…66

Dans le monde mouvant des technologies et sciences de l'information, de nouveaux concepts surgissent sans qu'on soit sûr de leur pérennité. Parfois ils expriment des concepts anciens qui n'ont pu se développer faute de technologies ou de maturité. Dans l'univers du décisionnel, plusieurs concepts émergent ou resurgissent grâce à l'évolution des technologies de l'information: Le Data Mining et l'Analyse de données.

Concernant le Data Mining qui est considéré comme un processus non élémentaire de mises à jour des relations, corrélations, dépendances, associations, modèles, structures, tendances, classes, facteurs obtenus en navigant à travers de grands ensembles de données, généralement consignés dans des bases de données, navigation réalisée au moyen de méthodes mathématiques, statistiques ou algorithmique.

On comprend, derrière le concept du Data Mining, l'héritage de l'intelligence artificielle et des systèmes experts. Mais on comprend aussi l'utilisation des méthodes d'analyse des données qui ont pour objet de découvrir des structures, des relations entre faits au moyen de données élémentaires et de techniques mathématiques appropriées. On ne s'étonnera pas donc de trouver au catalogue des méthodes de Data Mining aussi bien les réseaux de neurones, les arbres de décision.

Donc, on peut dire que la tâche principale du Data Mining c'est utilisé des méthodes pour extraire automatiquement l'information utile de ces données et la mettre à disposition des décideurs.

Le présent mémoire est articulé autour de cinq chapitres.

Nous présenterons dans le premier chapitre les généralités sur data mining, les différentes étapes et tâches et les différentes techniques les plus connues.

Dans le deuxième chapitre on définit le concept de data warehouse (entrepôt de données), qui joue un rôle essentiel dans le processus de data mining et les outils OLAP et les différents operateurs OLAP.

Dans le troisième chapitre, nous définirons les règles d'association et les différentes mesures utilisées pour les extraire, comme par exemple le support et la confiance. Nous montrerons par la suite les principales étapes d’extraction d'une règle

d’association. Ce chapitre mettra aussi en évidence l'algorithme APRIORI, qui représente la base des algorithmes d'extraction des règles d'association. Il exposera, entre autres, les différentes étapes du déroulement de cet algorithme. Enfin nous conclurons ce chapitre par la présentation des avantages et des inconvénients liés à l’utilisation des règles d’association.

Au quatrième chapitre nous aborderons la tâche la plus commune du data mining qui est la classification dont on parlera des différents états ou sera supervisé et non supervisé et comment sélectionner ses valeurs et comment choisir la bonne méthode de classification ou on détaillera la méthode bayésienne qui est la plus utilisée en probabilité

Le chapitre cinq est complètement consacré aux réseaux de neurones avec une explication de leur principe de fonctionnement et leurs structures, ainsi qu'une présentation du perceptron monocouche et multicouche qui constitue l'architecture neuronale la plus utilisée dans le domaine de réseaux de neurones.

Enfin, nous achèverons notre travail par une conclusion générale.

1. Introduction

Le data mining ou exploration des données identifie des tendances dans vos données grâce à diverses techniques prédictives. Grâce au data mining, les organisations comme la vôtre obtiennent des informations pertinentes sur les conditions externes, les processus internes, les marchés et les clients. Vous bénéficiez également de fonctionnalités prédictives utiles pour la planification stratégique ainsi que pour les interactions quotidiennes.

Ces connaissances et fonctions prédictives contribuent largement à améliorer la gestion des campagnes marketing de votre société, les ventes à plus haute valeur unitaire et les ventes croisées, ou encore la rétention des clients, l'analyse des risques et la détection des fraudes. Mais il y a de grandes chances que vous puissiez aller encore plus loin. En recourant à d'autres types de données, en associant des méthodes de data mining éprouvées au sein de nouvelles initiatives, ou en utilisant des options de déploiement évoluées, votre société peut obtenir un meilleur retour sur son investissement dans le data mining.

2. Historique [1]

L'expression data mining est apparu vers le début des années 1960, a cette époque les ordinateurs étaient de plus en plus utilisés pour toutes sortes de calculs qui n'était pas envisageable d'effecteur manuellement jusque là.

Certains chercheurs ont commencé à traiter sans a priori statique les tableaux de données relatifs à des enquêtes ou des expériences dont ils disposaient. Comme ils constataient que les résultats obtenus lois d'être aberrants, étaient tout au contraire prometteurs, ils furent incités à systématiser cette approche opportuniste, les statisticiens officiels considéraient toutefois cette démarche peu scientifique et utilisèrent alors les termes data mining ou data fishing pour les critiquer.

Cette attitude opportuniste face aux données coïncida en France avec la diffusion dans le grand public de l'analyse de données dont les promoteurs, comme Jean-Paul Benzecri, ont également du subir dans les premiers temps les critiques venant des membres de la communauté des statisticiens, le succès de cette démarche empirique ne s'est pas démenti malgré tout.

L'analyses des données s'est développée et sont intérêt grandissait en même temps que la taille des bases de données, vers la fin des années 1980, des chercheurs en base de données, tel que Rakesh Agrawal, ont commencé à travailler sur l'exploitation de contenu des bases de données volumineuses, comme par exemple celles des tickets de caisses de grande surface, convaincus de pouvoir valoriser ces masses de données dormantes, ils utilisèrent l'expression database mining mais, celle-ci étant déjà déposée par une entreprise (database Mining Workstation), ce fut data mining qui s'imposa.

En mars 1989, Shapiro Piatetski proposa le terme knowledge discovery à l'occasion d'un atelier sur la découverte des connaissances dans les bases de données.

Actuellement, les termes data mining et knowledge discovery in data bases (KDD, ou ECD en français) sont utilises plus au moins indifféremment.

La communauté de data mining a initie sa première conférence en 1995 a la suite de nombreux workshop sur la KDD entre 1989 et 1994.

En 1998 s'est créé, sous les auspices de l'ACM (Association for Computing Machinery), un chapitre spécial baptisé ACM-SIGKDD (ACM- Spécial Interest Groupe KDD), qui réunit la communauté international du KDD.

La première revue du domaine data mining and knowledge discovery journal publiée par (Kluwers) a été lancé 1990

3. Définition du data mining [2]

Selon le Groupe Gartner, le Data Mining appelé aussi fouille de données est le processus de découverte de nouvelles corrélations, modèles et tendances en analysant une grande quantité de données, en utilisant les technologies de reconnaissance des formes ainsi que d'autres techniques statistiques et mathématiques **[3]**.

Ils existent d'autres définitions :

- Le Data mining est l'analyse de grandes ensembles de données observationnelles pour découvrir des nouvelles relations entre elles et de

les reformuler afin de les rendre plus utilisables de la part de ses propriétaires **[4]**.

- Le Data mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatiques, de reconnaissance des formes, des statistiques, des bases de données et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données **[5]**.
- Le Data Mining est un processus inductif, itératif et interactif dont l'objectif est la découverte de modèles de données valides, nouveaux, utiles et compréhensibles dans de larges Bases de Données **[6]**.

4. Pourquoi la naissance du data mining ? [7]

- Augmentation des capacités de stockage des données (disques durs de giga octets).
- Augmentation des capacités de traitements des données (facilité d'accès aux données : il n'y a plus de bandes magnétiques, accélération des traitements).
- Maturation des principes des bases de données (maturation des bases de données relationnelles).
- Croissance exponentielle de la collecte des données (scanners de supermarché, internet,etc.)
- Croissance exponentielle des bases de données : capacités atteignant le terabits (1012 bits) et émergence des entrepôts de données : data warehouse, rendant impossible l'exploitation manuelle des données.
- Plus grande disponibilité des données grâce aux réseaux (intranet et internet).
- Développement de logiciels de data mining.

5. Statistiques et Data Mining [8]

La théorie voudrait que le data mining soit exploratoire, tandis que les statistique seraient confirmatoires.

On pourrait croire que les techniques de Data Mining viennent en remplacement des statistiques. En fait, il n'en est rien et elles sont omniprésentes. On les utilise **:**

- Pour faire une analyse préalable.
- Pour estimer ou alimenter les valeurs manquantes.
- Pendant le processus pour évaluer la qualité des estimations.
- Après le processus pour mesurer les actions entreprises et faire un bilan.

Par ailleurs, certaines techniques statistiques récentes (travaux de BENZECRI, analyse en composantes principales, analyse factorielle des correspondances, …) peuvent être apparentées aux techniques de Data Mining.

Statistiques et Data Mining sont tout à fait complémentaires.

- **Les statistiques**

Les statistiques sont à la base de tout raisonnement sur les données. Elles permettent de synthétiser un grand nombre de valeurs pour une variable grâce à un nombre très réduit d'informations. Pour chaque variable, on va ainsi rechercher au moins deux indicateurs : un pour mesurer la tendance centrale, un pour mesurer la dispersion.

5.1. Relations entre variables

Les besoins des décideurs ont amené les statisticiens à rechercher des liens entre plusieurs variables ou plusieurs populations. Ils ont donc créé de nouveaux indicateurs comme le khi2, la covariance ou le coefficient de corrélation. La corrélation entre les variables ne recouvre pas que la causalité; elle peut s'expliquer de plusieurs manières :

- La causalité : on observe qu'une variation de A entraîne une variation de B. Il existe un vrai lien entre A et B.
- Le hasard : une variation de A entraîne une variation de B mais celle-ci est uniquement due au hasard.
- La réponse commune : une variation de C entraîne une variation de A et B.
- La confusion : la variation de A et C entraîne la variation de B.

Lorsque le coefficient de corrélation est significatif, il y a souvent confusion entre ces différentes possibilités, surtout entre causalité et hasard.

D'autres techniques : régressions simples ou multiples (linéaires ou non), ajustements vers des lois statistiques (loi normale, binomiale, hypergéométrique, de Poisson, ...) permettent de modéliser les séries, et facilitent les estimations.

Ces techniques statistiques permettent de savoir s'il existe une relation entre plusieurs variables, de faire des prévisions ou estimations.

Le but de ce type d'analyse est souvent de rechercher des liens de causalité.

La recherche de connaissances par l'utilisation de méthodes statistiques est souvent limitée car on ne peut étudier simultanément que quelques variables (une à deux). Les problèmes sont en général plus complexes et mettent en œuvre plusieurs dizaines de variables. Pour répondre à ces besoins, il a fallu créer de nouveaux algorithmes, parfois issus de la recherche opérationnelle, alliant la recherche intelligente et les statistiques.

6. Les étapes du processus de data mining [2]

1- Collecte des données : la combinaison de plusieurs sources de données, souvent hétérogènes, dans une base de données.
2- Nettoyage des données : la normalisation des données : l'élimination du bruit (les attributs ayant des valeurs invalides et les attributs sans valeurs)
3- Sélection des données **:** Sélectionner de la base de données les attributs utiles pour une tâche particulière du data mining**.**
4- Transformation des données **:** le processus de transformation des structures des attributs pour être adéquates à la procédure d'extraction des informations.
5- Extraction des informations (Data mining): l'application de quelques algorithmes du Data Mining sur les données produites par l'étape précédente
6- Visualisation des données : l'utilisation des techniques de visualisation (histogramme, camembert, arbre, visualisation 3D) pour exploration interactive de données (la découverte des modèles de données).
7- Evaluation des modèles : l'identification des modèles strictement intéressants en se basant sur des mesures données.

7. Les tâches du data mining [10] [11]

Il existe plusieurs modèles de la fouille de données, mais aucun modèle n'est meilleur pour tous les domaines d'application. Il faudra faire des compromis selon les besoins dégagés et les caractéristiques connus des outils. Pour une utilisation optimale, une combinaison de méthodes est recommandée. On peut présenter les modèles de datamining en huit catégories

- Classification
- Estimation
- Prédiction
- Le regroupement par similitudes
- Segmentation (ou clustérisassions)
- Description
- Optimisation
- Extraction de règle d'association

7.1. Classification

La classification est la tâche la plus commune du Data Mining et qui semble être une obligation humaine. Afin de comprendre notre vie quotidienne, nous sommes constamment classifiés, catégorisés et évalués.

La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. L'objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées.

7.2. Estimation

L'estimation est similaire à la classification à part que la variable de sortie est numérique plutôt que catégorique. En fonction des autres champs de

l'enregistrement l'estimation consiste à compléter une valeur manquante dans un champ particulier.

7.3. La prédiction

La prédiction est la même que la classification et l'estimation, à part que dans la prédiction les enregistrements sont classés suivant des critères (ou des valeurs) prédites (estimées). La principale raison qui différencie la prédiction de la classification et l'estimation est que dans la création du modèle prédictif on prend en charge la relation temporelle entre les variables d'entrée et les variables de sortie.

7.4. Le regroupement par similitudes [2]

Le groupement par similitude consiste à déterminer quels attributs vont ensemble. La tâche la plus répandue dans le monde du business, où elle est appelée l'analyse d'affinité ou l'analyse du panier du marché, est l'association des recherches pour mesurer la relation entre deux et plusieurs attributs. Les règles d'associations sont de la forme "Si antécédent, alors conséquent.

7.5. L'analyse des clusters [8]

L'analyse des clusters consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous populations ne sont pas préétablis. La technique la plus appropriée à la clusterisation est l'analyse des clusters.

7.6. La description

Parfois le but du Data Mining est simplement de décrire se qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour en premier lieu comprendre le mieux possible les individus, les produit et les processus présents sue cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci.

7.7. L'optimisation [12]

Pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d'y associe une fonction d'évaluation. Le but de l'optimisation est de maximiser ou de minimiser cette fonction. Quelque spécialiste considèrent

que ce type de problème ne relève pas du data mining. La technique la plus appropriée à l'optimisation est le réseau de neurones.

7.8. Extraction de règle d'association [13]

L'extraction de règle d'association est l'un des principaux problèmes de l'ECD. Ce problème fut développé, à l'origine, pour l'analyse de bases de données de transaction de vente. Chaque transaction est constituée d'une liste d'articles achètes, afin d'identifier les groupes d'articles les plus fréquemment vendus ensembles.

La principale application des règles d'association est donc « l'analyse du panier de la ménagère ». Néanmoins, on assiste aujourd'hui à l'application de cette technique à tout domaine cherchant à regrouper des produits ou des services. Le problème d'extraction de règles d'associations s'est étendu au secteur bancaire, médicale, industriel, des nouvelle technologique.

Il faut noter les connaissances ainsi obtenues le sont par induction, à savoir un raisonnement de généralisation. Afin que les résultats acquièrent le statut de connaissances et afin d'éviter une dépendance trop grande par apport a un jeu de données particulier.

8. Techniques du data mininge [18]

Pour effectuer les tâches du Data Mining il existe plusieurs techniques issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) afin de faire apparaître des corrélations cachées dans des gisements de données pour construire des modèles à partir de ces données. Dans ce chapitre, nous présentons les techniques du data mining les plus connues.

8.1. Les réseaux de neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques. Ces unités peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connections. Les réseaux

de neurones sont capables de prédire de nouvelles observations (sur des variables spécifiques) à partir d'autres observations (soit les même ou d'autres variables) après avoir exécuté un processus d'apprentissage sur des données existantes.

La phase d'apprentissage d'un réseau de neurones est un processus itératif permettant de régler les poids du réseau pour optimiser la prédiction des échantillons de données sur lesquelles l'apprentissage été fait. Après la phase d'apprentissage le réseau de neurones devient capable de généraliser.

8.2. Les arbres de décision

Les arbres de décisions sont des outils d'aide à la décision qui permettent selon des variables discriminantes de répartir une population d'individus en groupes homogènes en fonction d'un objectif connu. Les arbres de décision sont des outils puissants et populaires pour la classification et la prédiction. Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions.

Chaque nœud interne d'un arbre de décision permet de répartir les éléments à classifier de façon homogène entre ses différents fils en portant sur une variable discriminante de ces éléments. Les branches qui représentent les liaisons entre un nœud et ses fils sont les valeurs discriminantes de la variable du nœud. Et en fin, les feuilles d'un arbre de décision représentent les résultats de la prédiction des données à classifier.

8.3. Les algorithmes génétiques

Un algorithme génétique se constitue d'une catégorie de programmes dont le principe est la reproduction des mécanismes de la sélection naturelle pour résoudre un problème donné. L'optimisation des problèmes combinatoires et surtout les problèmes dits NP-complets (dont le temps de calcul croit de façon non polynomiale avec la complexité du problème) est l'objectif principale des algorithmes génétiques, ils sont particulièrement adaptés à ce type de problèmes. Ces algorithmes constituent parfois une alternative

intéressante aux réseaux de neurones mais sont le plus souvent complémentaires.

8.4. Les règles associatives

Les règles associatives sont des règles extraites d'une base de données transactionnelles et qui décrivent des associations entre certains éléments. Elles sont fréquemment utilisées dans le secteur de la distribution des produits où la principale application est l'analyse du panier de la ménagère (Market Basket Analysis) dont le principe est l'extraction d'associations entre produits sur les tickets de caisse. Le but de la méthode est l'étude de ce que les clients achètent pour obtenir des informations sur qui sont les clients et pourquoi ils font certains achats. La méthode recherche quels produits tendent à être achetés ensemble. La méthode peut être appliquée à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services : services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical pour la recherche de complications dues à des associations de médicaments ou à la recherche de fraudes en recherchant des associations inhabituelles. Une règle d'association est de la forme : Si condition alors résultat. Dans la pratique, nous nous limitons généralement à des règles où la condition se présente sous la forme d'une conjonction d'apparition d'articles et le résultat se constitue d'un seul article. Par exemple, une règle à trois articles sera de la forme : Si X et Y alors Z ; règle dont la sémantique peu être énoncée : Si les articles X et Y apparaissent simultanément dans un achat alors l'article Z apparaît.

8.5. Réseaux Bayésien [12]

C'est une technique permettant la modélisation de la connaissance sous la forme d'un réseau dont les nœuds correspondent à des événements affectés de leurs probabilités respectives. Des liens de causalité permettent en outre de modeler ces

probabilités selon la connaissance que l'on a de certains autres événement. Dans le cadre de l'extraction de connaissance, cela signifie que ces

applications sont capables d'inférer des connaissances à partir d'indicateurs incomplets. Extrêmement puissants, bâtis sur le formalisme Bayésien.

9. Les avantages du data mining [7]

- Le data mining aide a la prise de décision des dirigeants. Par l'analyse des données, la méthode peut résumer la situation et alors accélérer la prise de décision des dirigeants à un problème donné. Par contre, le data mining ne remplace pa ces dirigeants.
- Le data mining permet de faire des liens pertinents entre des données qui a première vue n'ont aucune relation.
- Cette méthode peut améliorer la satisfaction des clients en analysant leurs besoins et en proposant des améliorations en fonction des évènements passés.
- Permet d'effectuer des profils type : des profils des clients et en proposant des améliorations en fonction des évènements passer.
- Le data mining facilite le développement de nouveaux produits.
- Accélère la gestion des stocks, des inventaires, de la logistique.
- Peut augmenter les revenues tout en diminuent les couts. C'est évident, le data mining a été étudie pour augmenter et optimiser le rendement d'une entreprise ou l'amélioration d'un critère.

10. Conclusion

Le data mining est l'extraction d'informations prédictives cachés dans de grandes base de données. C'est une technologie nouvelle et puissante qui donne la possibilité aux entreprises de se concentrer sur les informations les plus importantes dans leurs data warehouses. Les outils du data mining peuvent prédire les futurs tendances et actions, permettant de prendre les bonnes décisions. C'est ce qui rend le data mining la technologie la plus importantes.

1. Introduction

Les entrepôts de données et les systèmes OLAP (Online Analytical Processing) permettent un accès rapide et synthétique à de gros volumes de données à des fins d'analyse. Afin d'améliorer encore les performances des systèmes décisionnels, une solution consiste en la mise en œuvre d'entrepôts de données sur des systèmes répartis toujours plus puissants.

2. Data warehouse

2.1. Historique [8]

Le concept de **data warehouse** (entrepôt de données ou base de données décisionnelle) a été formalisé pour la première fois en 1990. L'idée de constituer une base de données orientée sujet, intégrée, contenant des informations datées, non volatiles et exclusivement destinées aux processus d'aide à la décision, fut dans un premier temps accueillie avec une certaine perplexité. Beaucoup n'y voyaient que l'habillage d'un concept déjà ancien : l'infocentre. Mais l'économie actuelle en a décidé autrement. Les entreprises sont confrontées à une concurrence de plus en plus forte, des clients de plus en plus exigeants, dans un contexte organisationnel de plus en plus complexe et mouvant.

Pour faire face aux nouveaux enjeux économiques, l'entreprise doit anticiper. L'anticipation ne peut être efficace qu'en s'appuyant sur de l'information pertinente. Cette information est à la portée de toute entreprise qui dispose d'un capital de données gérées par ses systèmes opérationnels et qui peut en acquérir d'autres auprès de fournisseurs externes. Mais actuellement, les données sont surabondantes, non organisées dans un perspectif décisionnel et éparpillées dans de multiples système hétérogènes. Pourtant, les données représentent une mine d'informations. Il devient fondamental de rassembler et d'homogénéiser les données afin de permettre d'analyser les indicateurs pertinents pour faciliter les prises de décisions. Pour répondre à ces besoins, le nouveau rôle de l'informatique est de définir et d'intégrer une architecture qui serve de fondation aux applications décisionnelles : le data warehouse (DW).

2.2. Définition [9]

Le data warehouse est une collection des données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision. Le data warehouse est donc une sorte de point focal stockant en un endroit unique toute l'information utile provenant des systèmes de production et des sources externes. Avant d'être chargée dans le data Waterhouse, l'information donc doit être extraite, nettoyée et préparée.

Le data warhouse est organisé autour des sujets majeurs de l'entreprise. Ainsi les données sont structurées par thèmes contrairement aux données des organisations traditionnelles généralement organisées par processus fonctionnel.

2.3. Les caractéristiques de data Waterhouse [20]

Un entrepôt de données, offre des données intégrées, consolidées et historisées pour faire des analyses. Il s'agit d'une collection de données pour le support d'un processus d'aide à la décision.

Les données de DW possèdent les caractéristiques suivantes :

- **Orientation sujet :** Les données de DW s'organisent par sujets ou thèmes (clients, activités, items…). Cette organisation permet de rassembler toutes les données, pertinentes à un sujet et nécessaires aux besoins d'analyse, qui se trouvent répandues à travers les structures fonctionnelles d'une entreprise.
- **Intégration :** Les données de DW sont le résultat de l'intégration de données en provenance de multiples sources ; ainsi, toutes les données nécessaires pour réaliser une analyse particulière se trouvent dans le DW. L'intégration est le résultat d'un processus qui peut devenir très complexe du à l'hétérogénéité des sources.
- **Histoire :** Les données de DW représentent l'activité d'une entreprise pendant une longue période ou il est important de gérer les différentes valeurs qu'une donnée prises au cours du temps. Cette caractéristique donne la possibilité de suivre une donnée dans le temps pour analyser ses variations.

- **Non-volatilité :** les données chargées dans le DW sont surtout utilisées en interrogation et ne peuvent pas être modifiées, sauf dans certains cas de rafraîchissement.

2.4. Les avantage de data Waterhouse

Le data Waterhouse offre à l'entreprise les avantages suivants

- Il constitue une collection de données centralisée disponible pour l'aide à la décision (OLAP, datamining,...)
- Les évolutions des données de l'entrepôt sont conservées (historisation des données)
- Il contient un ensemble de données consolidées (données homogènes et fiables)
- Il contient des données agrégées permettant une analyse à différents niveaux de détails
- Il permet de développer différents thèmes d'analyse (réorganisation en fonction des sujets à analyser)
- L'accès direct aux données : facile, rapide, sécurisé
- Plus de contrôle sur les données
- Possibilité d'ajouter facilement des annotations

3. Le Data Mart [24]

3.1. Définition

Le DataMart est un ensemble de données ciblées, organisées, regroupées et agrégées pour répondre à un besoin spécifique à un métier ou un domaine donné. Il est donc destiné à être interrogé sur un panel de données restreint à son domaine fonctionnel, selon des paramètres qui auront été définis à l'avance lors de sa conception.

- Le DataMart est issu d'un flux de données provenant du DataWarehouse. Contrairement à ce dernier qui présente le détail des données pour toute l'entreprise, il a vocation à présenter la donnée de manière spécialisée, agrégée et regroupée fonctionnellement.

- Le DataMart est un sous-ensemble du DataWarehouse, constitué de tables au niveau détail et à des niveaux plus agrégés, permettant de restituer tout le spectre d'une activité métier. L'ensemble des DataMarts de l'entreprise constitue le DataWarehouse.

3.2. Intérêt et limites

Les DataMarts étant des extraits simplifiés du détail des données de l'entreprise, ils ne présentent d'intérêt que pour des requêtes identifiées et répétitives ; il est plus facile pour le système d'interroger un DataMart qui ne contient que le nécessaire que d'avoir à cerner et à trier toute la base. Par ailleurs, les DataMarts permettent de classifier et de clarifier l'information, de manière à ce que chaque métier ait accès à des chiffres correspondant à ses attentes fonctionnelles, sans être pollué par des données contigues.

Les choix de simplification qui donnent lieu aux DataMarts rendent ceux-ci naturellement moins flexibles, des demandes d'utilisateurs sortant de leur cadre habituel requièrent fréquemment d'interroger la base à un autre niveau, générant des coûts de développement ou la création de solutions de rechange. Des problèmes peuvent de fait survenir lorsque les DataMarts constituent l'unique moyen d'accès aux données pour l'utilisateur final.

3.3. Exemples de types de DataMarts

Les thèmes suivants se retrouvent dans la plupart des DataMarts d'entreprise :

- DataMart commercial : utilisé pour produire l'information liée au client et à son comportement, avec impact sur le chiffre d'affaires de l'entreprise. Il permet notamment de suivre les succès ou les échecs des produits, ou encore de vérifier l'impact d'une opération commerciale sur la clientèle.
- DataMart financier : il permet de suivre l'activité boursière de l'entreprise.
- DataMart RH : utilisé pour suivre les arrivées et les départs dans l'entreprise.

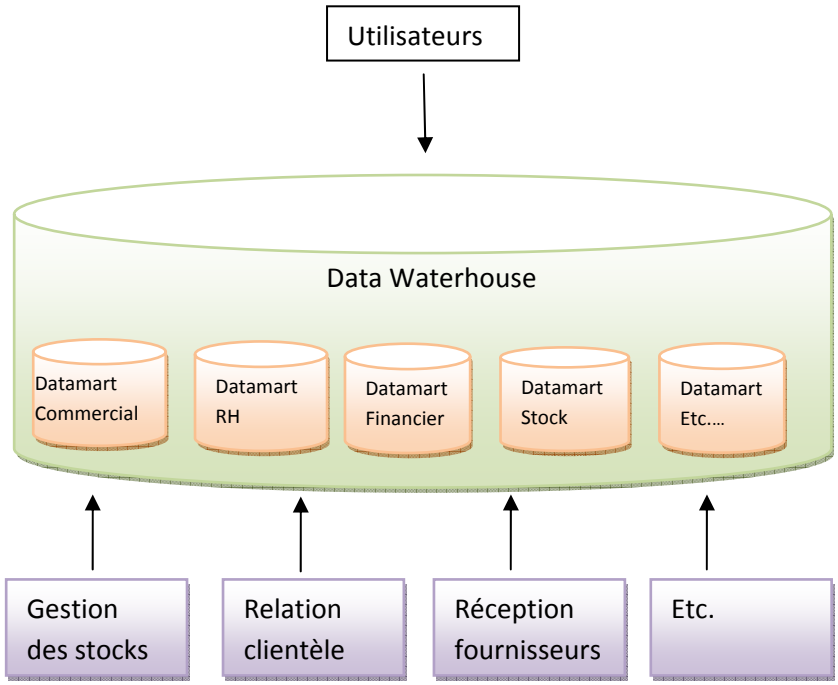


Figure 2.1: Architecture d'un data warehouse.

4. Modèle multidimensionnel [14]

Le modèle conceptuel doit être simplifie au maximum pour permettre au plus grand nombre d'utilisateurs d'appréhender l'organisation des données et de comprendre que le Data Waterhouse mémorise. On parle de modèle multidimensionnel. Ce dernier est utilisé dans les applications dont l'objectif est d'analyser les données plutôt que de procéder à des transactions on-line.

La technologie des bases de données multidimensionnelles est un facteur clé dans l'analyse de larges quantités de données dans l'informatique décisionnelle. En effet, contrairement aux technologies précédentes, les données sont vues comme des cubes particulièrement adaptés à l'analyse de données dans le modèle multidimensionnel.

4.1. Le Cube de Données OLAP [14]

Le cube représente toutes les données qui seront nécessaires à l'analyse, ces données peuvent être stockées ou calculées à la volée, ce qui impose respectivement de l'espace de stockage et du temps de calcul. Afin de conserver des dimensions raisonnables autant dans l'espace de stockage nécessaire que dans les temps de réponse, il faut choisir quelles données

seront calculées d'avance et stockées pour une acquisition immédiate dans le futur, et quelles données seront plutôt calculées à chaque fois qu'elles seront nécessaires.

En effet le cube est produit à partir des données brutes de systèmes d'information, donc des résultats généralement très précis. Ces résultats peuvent être regroupés par familles de différents niveaux de généralité (les niveaux d'agrégation) selon les besoins de précision de l'analyse. Les valeurs correspondant à ces niveaux sont des totaux, des moyennes et autres de toutes les valeurs précises appartenant au niveau. Les niveaux d'agrégation ont une forme d'arbre où toutes les données brutes sont les feuilles.

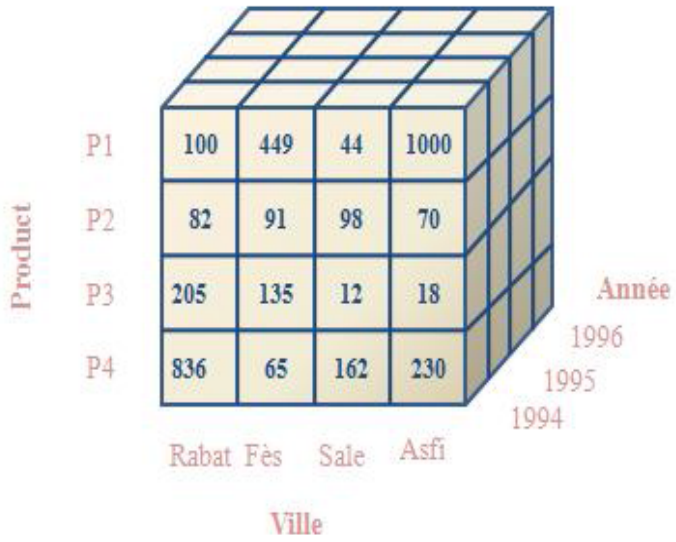


Figure 2.2 : Exemple de cube de données (Ventes par ville)

4.2. Faits [15]

Un fait représente un sujet d'analyse. Il est constitué de plusieurs mesures relatives au sujet traité. Ces mesures sont numériques et généralement valorisées de façon continue.

Dans la plupart des modèles multidimensionnels, les faits sont implicitement représentés par la combinaison des valeurs des dimensions. Un fait n'existe

que si une combinaison particulière des dimensions découle vers une cellule non vide du cube le représentant.

4.3 Dimension [9]

Chaque dimension peut être mesurée par une unité plus ou moins fine. Par exemple, le temps est exprimé en jours, mois, trimestres ou années, la géographie en villes, régions ou pays, les produits en numéros, types, gammes et marques. L'or de l'exploration, les analystes étendent ou réduisent le cube selon certaines dimensions.

Exemple :

«250 000 euros » est un fait qui exprime la valeur de la mesure

« coût des travaux » pour le membre « 2002 » du niveau année de la dimension « temps » et le membre « Versailles » du niveau « ville » de la dimension « découpage administratif »

5. Opérateurs OLAP [16] [9]

L'opérateur OLAP (On-Line Analytical Processing) fait référence à une méthode d'analyse qui peut être représentée par un cube.

Les opérateurs OLAP pour la manipulation des cubes de données sont de deux types : les opérateurs d'agrégation et les opérateurs de présentation pour la navigation.

5.1. Opérateurs d'agrégation

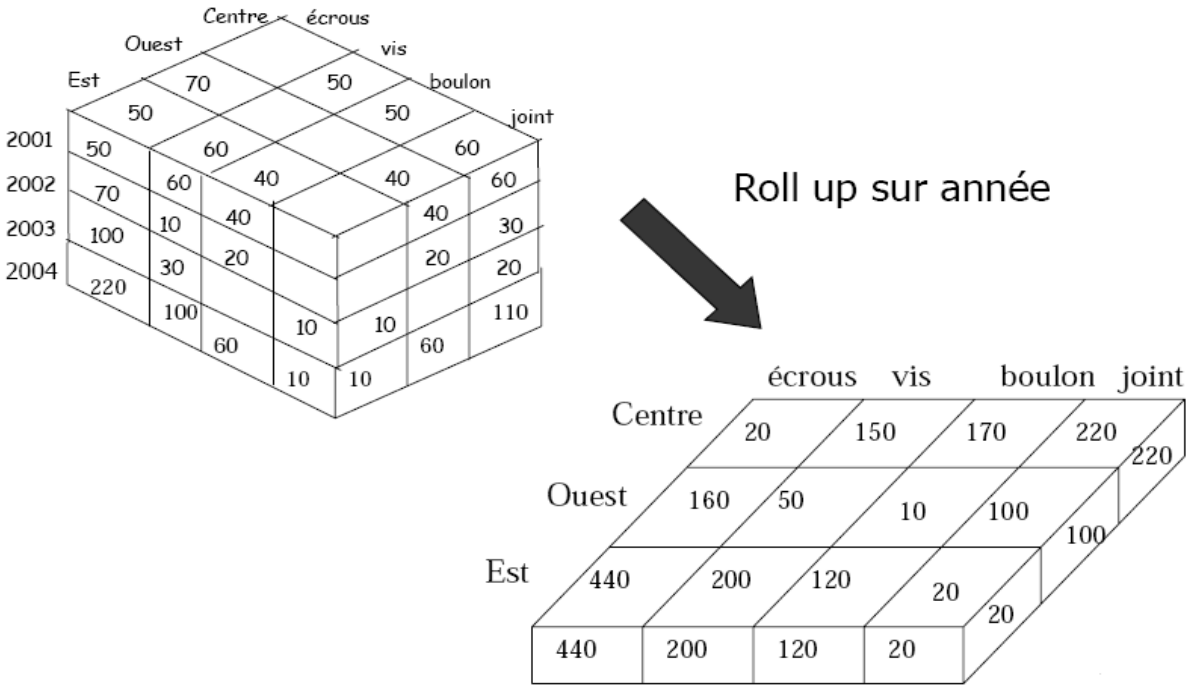
Etant donné le principe de granularité, la navigation dans le cube de données permet a l'utilisateur de passer de données détaillées a des données moins détaillées. Ce genre de manipulation nécessite de résumer les données.

5.2 Opérateurs de présentation pour la navigation [14]

Les outils OLAP utilisent des opérateurs particuliers pour la navigation dans les cubes.

5.2.1. Opérations de forage

Cube Roll-up: aller d'une vue détaillées vers une vue plus globale, par exemple, de l'article, mois, magasin, vers l'article, région.



Figute2.3 Cube Roll-up

Drill-down : inverse de roll-up, aller du global vers le détail, faire un drill-down, c'est avoir un niveau de détails sur les données. Par exemple Supposons qu'on veuille voir le détail des ventes pour le premier trimestre de l'année 1997. On dit qu'on fait un drill-down sur l'axe (ou dimension) temps. C'est-à -dire qu'on ne veut pas voir seulement les données de l'année 1997 mais descendre à un niveau de détail plus bas.

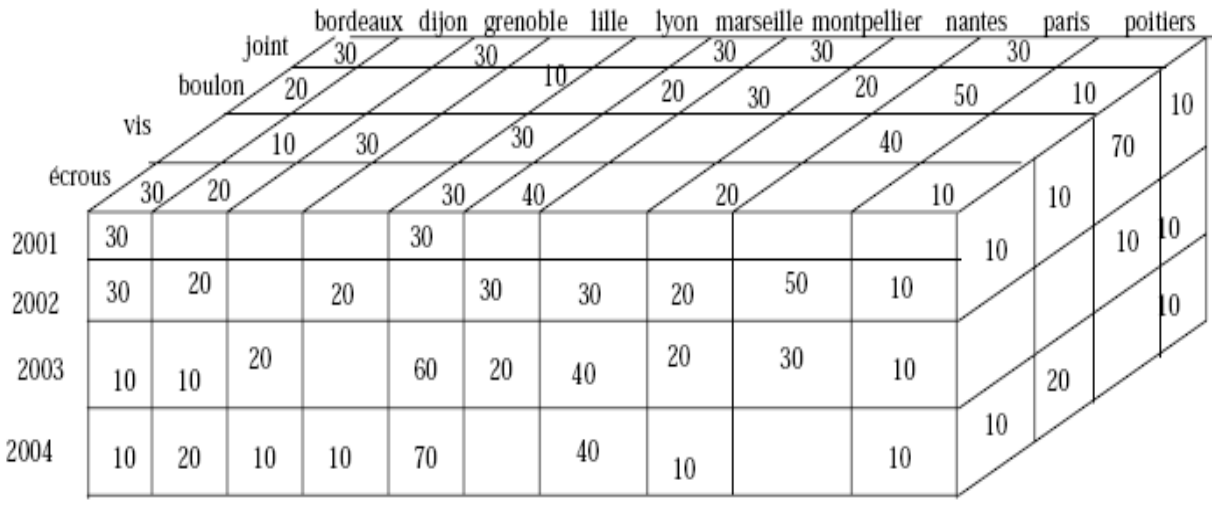
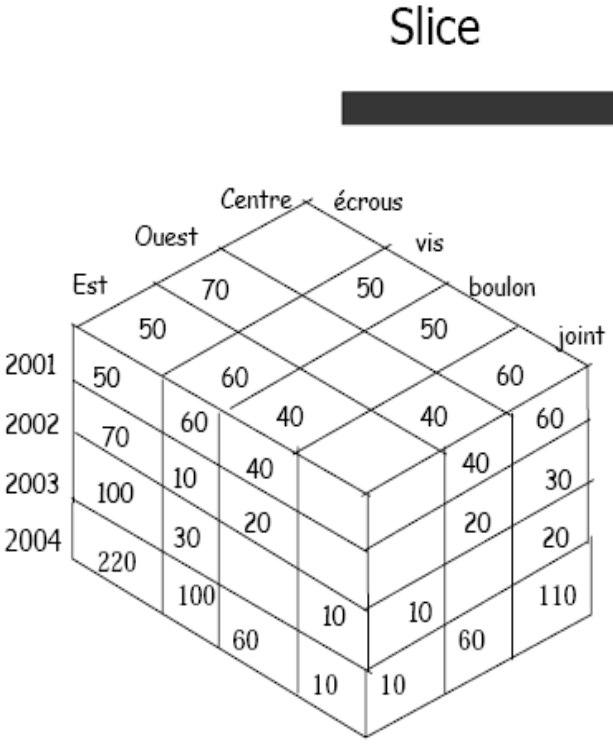


Figure2.4 Drill-down

5.2.2 Opérations de sélection.

Slicing : Extraction d'une tranche d'informations : Sélection d'une dimension pour passer à un sous-cube.

Exemple: Slice (2004) : on ne retient que la partie du cube qui correspond à cette date.



Slice (2004)

Ventes 2004	écrou	vis	boulon	joint
est	220	100	60	10
ouest	160	50	10	60
centre	20	150	170	110

Figure2.5 Slice (2004) : on ne retient que la partie du cube qui correspond à cette date.

Dice : Extraction d'un bloc de données : Sélection de deux ou plusieurs dimensions.

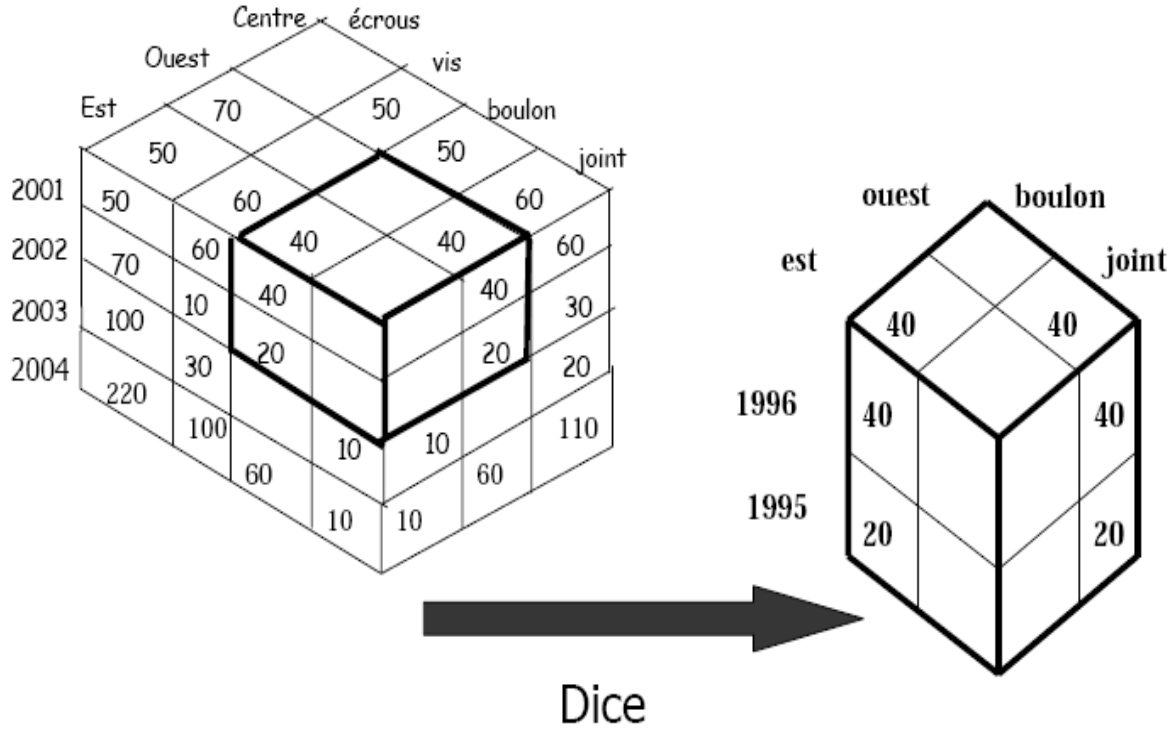


Figure 2.6 Dice : Extraction d'un bloc de données : Sélection de deux ou plusieurs dimensions.

6. Rafraichir le data Waterhouse [9]

Une fois le système décisionnel alimente la première fois, le processus ne s'arrête pas la et périodiquement, les flux de données vont rafraichir la base décisionnelle. L'idéal dans ce cas est de ne recharger le data warehouse qu'avec les données modifiées ou ajoutées depuis la dernière extraction. Certains outils commencent à fournir ce type de fonctionnement pour optimiser la phase d'extraction, l'idée consiste à réaliser des extractions différentielles en utilisant un mécanisme de marquage des données, souvent par examen de la date de dernière modification associée a la donnée.

7. Conclusion

Les informations des différentes bases de données d'une entreprise sont collectées dans un seul entrepôt de données, ou alors il existe différents entrepôts de données en fonction du sujet ou du métier en rapport avec chaque

information (datamart). Les informations collectées serviront à faire des statistiques, des recherches et des rapports. Les entrepôts de données sont utilisés notamment en informatique décisionnelle. La base de données est le constituant principal du datawarehouse et le cœur de l'intelligence d'affaires : c'est dans celle-ci que l'on va stocker les informations extraites des bases de production

1. Introduction

Dans le domaine du data mining la recherche des Règles d'Association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données. Piatetsky-Shapiro présentent de règles d'Association extrêmement fortes découvertes dans des bases de données en utilisant différentes mesures d'intérêt. En se basant sur le concept de relations fortes.

2. Règle d'association

2.1. Définition [18] [12]

Notation

Soit $I = \{i_1, i_2, \ldots\ldots, i_k\}$

B : la base de données, $|B|$ le nombre de transaction totale de la base.

T : une transaction.

- **Item :**

  Un item est un littéral correspondant à une valeur ou un ensemble de valeurs pour un attribut sélectionné dans la base de données.

- **Itemset** :

  Un itemset est un ensemble non vide et non ordonné d'items note $(i_1, i_2, \ldots\ldots, i_k)$, ou i_j est un etem de I.

- **Support d'un itemsets** :

  Soit X un itemset, son support est le nombre de transaction de la base B contenant X divisé par le nombre total de transaction.

$$\text{supp (X)} = \frac{| \ T \in B \ / \text{X} \subseteq T \ |}{| \ B \ |}$$

- **Itemsets fréquent**

  Un itemset est dit fréquent si son support est supérieur ou égal à un seuil correspondant au support minimum exigé par un utilisateur pour une règle d'association.

- **Règle d'association**

  Les règles d'associations se définissent comme un ensemble de techniques recherchant des relations intéressantes parmi les items contenus dans une série de données particulière

  Une règle d'association est une relation d'implication $X \rightarrow Y$ entre deux ensembles d'articles X et Y, avec $X, Y \neq \emptyset$ et $X \cap Y = \emptyset$.

  Cette règle indique que les transactions qui contient les articles de l'ensemble X ont tendance à contenir les articles de l'ensemble Y.

  L'ensemble X est appelé condition ou prémisse.

  L'ensemble Y est appelé résultat ou conclusion.

- **Support d'une règle d'association**

  Le support d'une règle $(X \rightarrow Y)$ est le rapport entre le nombre de transaction de B contenant $X \cup Y$ et le nombre total de transaction.

$$\text{supp}(X \rightarrow Y) = \frac{| T \in B / X \cup Y \subseteq T |}{| B |}$$

- **Confiance d'une règle d'association :**

  La confiance d'une règle d'association est le rapport entre le nombre de transaction de B contenant $X \cup Y$, et le nombre de transaction de B contenant X.

$$\text{conf}(X \rightarrow Y) = \frac{|T \in B \text{ tq } X \cup Y \subseteq T|}{| T \in B / X \subseteq T |}$$

La confiance se définit aussi par un rapport de support

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}\ (X \cup Y)}{\text{supp}\ (X)}$$

Exemple 1 :

N^0 transaction	**Items achètes**
1	1, 2,5
2	2,4
3	2,3
4	1, 2,4
5	1, 2,3
6	2, 3,5
7	1,3
8	1, 2, 3,5
9	1, 2,3
10	2,3

Tableau 1: Exemple d'une base de données contenant 10 transactions.

I= {1, 2, 3, 4,5}.

Il ya 10 transaction.

Supp (1→2)= 5/10 = 50% car le nombre de transaction de B contenant

$(1 \cup 2) = 5$ et $B = 10$.

Conf (1→2) = 5/6 = 83.3% puisque parmi les 6 transactions où l'item 1 étaient présent, 5 d'entre-elles contenaient également l'item 2.

Exemple 2 :

De manière générale, une règle d'association s'énonce comme suit:

Pain => beurre [support = 2%, confiance = 60 %].

La règle d'association de l'énoncé se lit de la manière suivante: 60 % des clients qui ont acheté du pain ont également acheté du beurre. De plus 2 % des transactions totales enregistrées dans la base de données respectent cette règle. De façon générale, la confiance se définit comme la probabilité d'obtenir dans une transaction l'item résultant de la règle, dans ce cas-ci du beurre, sachant qu'un ou que d'autres items, dans ce cas-ci le pain, s'y retrouvent également. Le support, quant à lui, est donné par la proportion des transactions de la base de données transactionnelles qui contiennent tous les items présents dans la règle. Ce petit exemple démontre l'avantage principal des règles d'associations: leur simplicité. Contrairement à d'autres méthodes d'analyse de données, les règles d'associations fournissent des réponses simples et facilement interprétables quel que soit le degré de complexité de la règle.

2.2. Utilité des règles d'associations [21]

Les règles d'associations sont appliquées dans plusieurs domaines. En marketing, par exemple, elles permettent d'identifier les produits ou services qui sont achetés lors d'une même transaction ou par un même client dans le temps et offrent donc la possibilité d'identifier des opportunités de ventes croisées. En analysant l'ordre dans lequel les internautes accèdent aux pages d'un site WEB, les règles d'associations séquentielles permettent d'entrevoir quelles modifications rendraient le site plus convivial, permettant ainsi aux internautes de trouver rapidement les informations recherchées.

2.3. La recherche de règles d'association [17]

L'extraction des règles d'association est l'un des principaux problèmes de l'ECD (Extraction de Connaissances à partir de Données). Ce problème fut développé à l'origine pour l'analyse de base de données de transactions de ventes. Chaque transaction est constituée d'une liste d'articles achetés, afin d'identifier les groupes d'articles vendus le plus fréquemment ensemble.

Ces règles sont intuitivement faciles à interpréter car elles montrent comment des produits ou des services se situent les uns par rapport aux autres. Ces règles sont particulièrement utiles en marketing. Les règles d'association

produites par la méthode peuvent être facilement utilisées dans le système d'information de l'entreprise.

2.3.1. Les étapes d'extraction de règles d'association

L'extraction des règles d'association peut être décomposée en quatre étapes qu'illustre la Figure 1 Les étapes d'extraction de règles d'association suivante

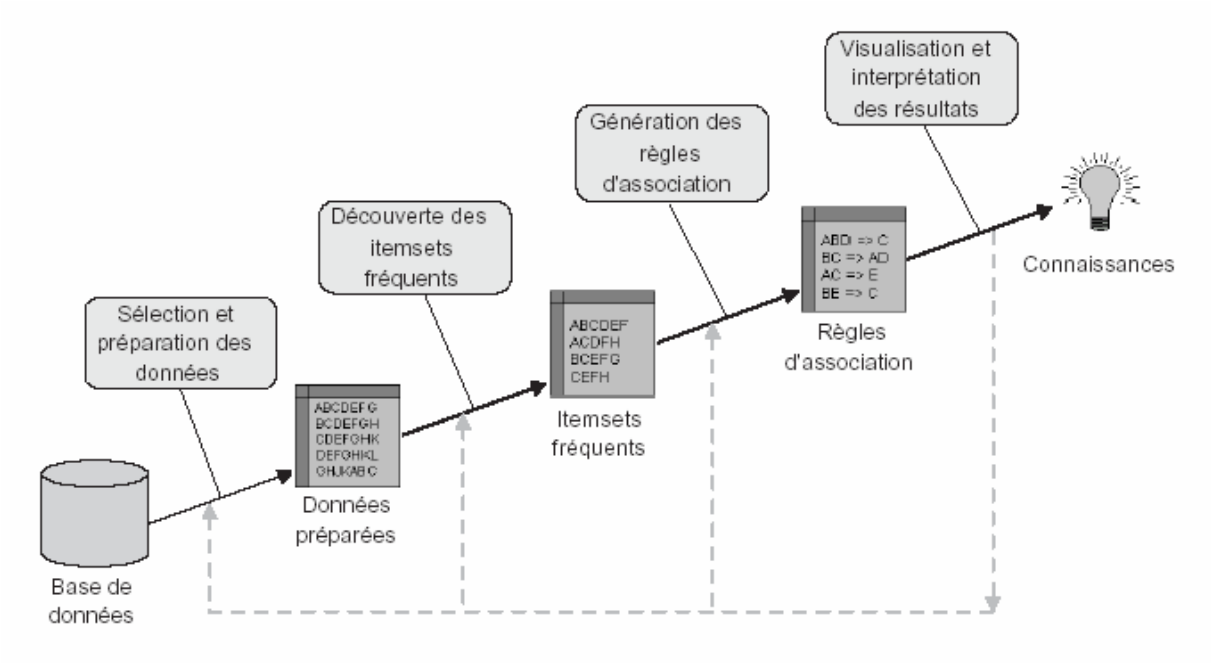


Figure 1.4 : Les étapes d'extraction de règles d'association

a) Préparation des données:

Cette étape permet de préparer les données afin de leur appliquer les algorithm d'extraction des règles d'association. Elle est constituée de deux phases :

- La sélection des données de la base qui permettront d'extraire les informations intéressant l'utilisateur. Ainsi la taille des données traitées est réduite ce qui assure une meilleure efficacité de l'extraction.
- La transformation de ces données en un contexte d'extraction (il s'agit d'un triplet constitué d'un ensemble d'objets, d'un ensemble d'itemsets et d'une relation binaire entre les deux).

La transformation des données sélectionnées en données binaires améliore l'efficacité de l'extraction et la pertinence des règles d'association extraites.

b) Recherche des ItemSets fréquents:

Un Itemset fréquent est un ensemble d'éléments dont le support est supérieur ou égal à un certain support minimal spécifié par l'utilisateur. Cette étape est très coûteuse en temps d'exécution car le nombre d'itemsets fréquents dépend exponentiellement du nombre d'items. Pour un ensemble de *n* items par exemple, le nombre d'ltemsets fréquents qui peut être générés est de 2^n

c) Production des règles d'association:

La génération des règles d'association consiste à déterminer les règles d'association dont le support et la confiance sont supérieurs ou égaux à un certain support et confiance minimaux définis par l'utilisateur. .

d) Visualisation et interprétation des règles d'associations :

Elle met entre les mains de l'utilisateur un ensemble de déductions fiables qui peuvent l'aider à prendre une décision.

3. Algorithme apriori de recherche de règles d'association [21]

L'algorithme Apriori, introduit par Agrawal et al. (1994) est un algorithme clé pour les règles d'association car il est à la base de la majorité des algorithmes servant à découvrir des règles d'associations plus complexes telles que les associations séquentielles et multidimensionnelles. Il tire son nom de son heuristique qui utilise l'information connue a priori sur la fréquence des items. Cette heuristique stipule que si A, un sous-ensemble d'items de l'ensemble I, ne possède pas le support minimal, il ne peut être engagé dans une règle d'association avec tout autre item i de l'ensemble I, $i \notin A$. Ainsi, si A est peu fréquent la règle $A \Rightarrow i$ l'est également et il est donc inutile d'examiner toute règle d'association où A est impliqué. Le problème consistant à identifier des règles d'associations se divise en deux étapes : une étape de jointure et une autre d'élagage.

3.1. L'étape de jointure :

Pour trouver les itemsets fréquents dans la base de données transactionnelle, l'algorithme Apriori effectue plusieurs balayages de la base de données. Le premier balayage sert à identifier les candidats C_k, un ensemble d'itemsets, et

à compter le nombre de fois qu'apparaît chaque item, c'est-à-dire leur support respectif. Tous les items dont le support est plus grand qu'une valeur prédéterminée appelée minsup, sont conservés afin de former L_k, l'ensemble des k-itemsets fréquents. Cet ensemble sert d'amorce pour générer l'ensemble de candidats C_{k+1}. L'ensemble C_{k+1}, qui regroupe les (k+1) itemsets, est généré en liant L_k avec lui-même. Pour que deux k-itemsets puissent être liés, ils doivent posséder k-1 items en commun. Par conséquent la liaison de deux 1-itemsets ne requiert aucun élément en commun, alors que la liaison de deux 3-itemsets requiert 2 éléments en commun. Les deux 1-itemsets {1} et {2} peuvent être liés ensemble pour générer le 2-itemset {1,2}. Le 3-itemset {1,2,3} peut être lié avec {2,3,4} pour générer {1,2,3,4}, mais ne peut pas être lié avec {3,4,5} puisque seul l'items {3} est commun aux itemsets {1,2,3} et {3,4,5}. Il est fort possible qu'un itemset généré ne respecte pas le seuil de support minimal. Si c'est le cas, cet itemset est éliminé lors de l'étape d'élagage.

3.2. L'étape d'élagage :

Une fois l'ensemble des candidats C_{k+1} généré, le support de tous les (k+1)-itemsets est calculé. Tous les (k+1)-itemsets $\in C_{k+1}$ dont le support ne dépasse pas le minsup sont retirés de la liste des candidats. Comme la liste des candidats C_{k+1} est réalisée à partir de la liste antérieure des candidats C_k, tout candidat retiré à l'étape k n'est plus considéré dans l'étape k+1.

4. Exemple de l'algorithme Apriori :

Voici un exemple détaillé des étapes suivies par l'algorithme Apriori. La base de données utilisée pour cet exemple est celle qui se retrouve dans le Tableau 1.Cette base de données contient 10 transactions et pour cet exemple, le support minimal est fixé à 30 % soit un décompte minimal requis de 3 transactions. Les résultats détaillés de chacune des étapes sont illustrés dans le Tableau 2.

No transaction	Items achetés
1	1, 2,5
2	2,4
3	2,3
4	1, 2,4
5	1, 2,3
6	2, 3,5
7	1,3
8	1, 2, 3,5
9	1, 2,3
10	2,3

Tableau 1.1 : Exemple d'une base de données contenant 10 transactions.

1. Lors de la première itération, l'algorithme compte le support de chaque 1-itemset de la base de données. Ces itemsets forment l'ensemble des candidats C_1 qui sert à générer l'ensemble L_1, c'est à dire l'ensemble des 1-itemsets fréquents.
2. Premier élagage : L'algorithme compare ensuite la fréquence de chaque 1-itemset avec la fréquence minimale prédéterminée ici à 3 ou un support de 30 %. Tous les itemsets ayant une fréquence inférieure à 3 sont retirés et ceux dont la fréquence est supérieure ou égale à 3 sont conservés afin de générer l'ensemble L1. Les 1-itemsets fréquents qui forment L sont dans ce cas-ci {1}, {2}, {3} et {5}.
3. Étape de jointure : Les 1-itemsets de l'ensemble L1 sont utilisés pour générer les candidats C2. La génération des candidats est réalisée en liant l'ensemble L1 avec lui-même. Comme il s'agit de 1-itemsets, le nombre de combinaison possible est de n (n-1)/2, n étant le nombre d'itemsets. Dans cet exemple, le nombre d'itemsets étant de 4, six candidats sont formés. Les candidats sont {1,2}, {1,3}, {1,5}, {2,3}, {2,5} et {3,5}.
4. Calcul du support : Lorsque les 2-itemsets candidats ont été générés, l'algorithme effectue un autre balayage de la base de données afin de calculer la fréquence

respective des candidats. Cette fréquence est inscrite dans une table comme illustré dans le Tableau 2 : étapes de l'algorithme Apriori avec un support minimal de 30%.

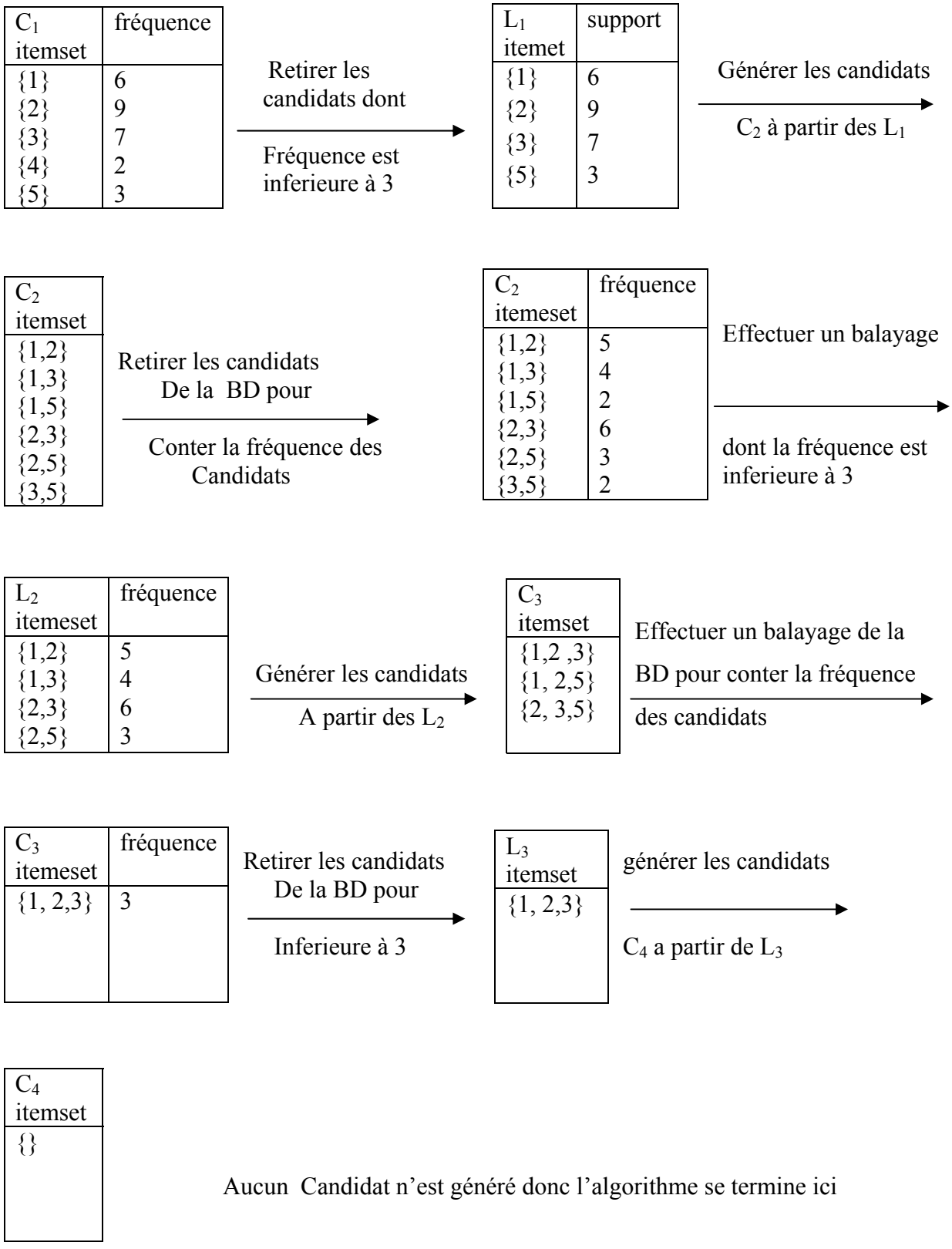


Tableau 1.2 : étapes de l'algorithme Apriori avec un support minimal de 30%.

5. Deuxième élagage. L'algorithme va par la suite parcourir l'ensemble C2 afin d'éliminer tous les 2-itemsets ayant une fréquence inférieure à 3. Dans ce cas-ci, les 2-itemsets {1,5} et {3,5} sont retirés. Les autres 2-itemsets sont conservés et forment l'ensemble L_2, c'est-à-dire l'ensemble des 2-itemsets fréquents.
6. Génération des candidats. La génération des 3-itemsets est réalisée en fusionnant L2 avec lui-même. Comme cette fusion implique des 2-itemsets, les itemsets doivent avoir 1- itemset (2-1) en commun. De plus, tous les sous-ensembles de (k-1)-itemsets formés doivent être fréquents Si un k-itemset généré est composé de (k-1)-itemsets non fréquents, celui-ci est automatiquement éliminé, ce qui évite de calculer son support ou sa fréquence. Dans l'exemple du Tableau 2: étapes de l'algorithme Apriori avec un support minimal de 30%.{1,2} et {2,5} peuvent être fusionnés pour former l'itemset {1, 2,5} Les sous-ensembles de {1, 2,5} sont {1,2}, {1,5} et {2,5} Comme {1,5} ne figure pas parmi l'ensemble L_2 des 2-itemsets fréquents, l'itemset {1, 2,5} ne peut pas avoir une fréquence supérieure à 3. Il est donc retiré des candidats et son support n'est pas comptabilisé. L'itemset {2, 3,5} est également élagué car {3,5} ne figure pas parmi l'ensemble L_2 des 2-itemsets fréquents.

   À la fin de cette étape, seul l'itemset {1, 2,3} est généré.
7. Calcul du support. Un troisième balayage de la base de données sert à calculer la fréquence de l'itemset {1, 2,3} et ce dernier est de 3.
8. Étape d'élagage. L'algorithme compare la fréquence de l'itemset, {1, 2,3} avec la fréquence minimale. Comme {1, 2,3} possède la fréquence minimale, celui-ci est conservé et devient le seul itemset de L_3, l'ensemble des 3-itemsets fréquents.
9. Génération des candidats. Étant donné que l'ensemble d'amorce L_3 ne contient qu'un seul itemset, soit {1, 2,3}, aucun 4-itemset candidat ne peut être généré. L'algorithme se termine ici.

   Les associations identifiées par l'algorithme sont celles formées par les itemsets de l'ensemble L_2 et de l'ensemble L_3 soit {1,2}, {1,3}, {2,3}, {3,5} et {1, 2,3}. Ces itemsets fréquents engendrent les règles d'associations. Ainsi, l'itemset

{1,2} engendre les règles d'associations « 1 => 2 » et « 2 =>1 » alors que l'itemset {1, 2,3} engendre les associations.

« 1 => 2 => 3 », « 1 => 3 => 2 », « 2 => 1 => 3 », « 2 => 3 => 1 », « 3 => 1=> 2 » et « 3 => 2 => 1 ».

5. Avantages et inconvénients des règles d'association [17]

5.1. Avantages

Les règles d'association représentent plusieurs avantages parmi lesquels on peut citer par exemple

- Leur application dans plusieurs domaines de la vie quotidienne, comme l'analyse du panier de la ménagère.
- La découverte de connaissances utiles, cachées dans les grandes bases des données.
- Leur simplicité, efficacité et facilité de compréhension.
- Leur formalisme non supervisé et général.
- Leurs résultats clairs et faciles à interpréter.

5.2. Inconvénients

Malgré les grands avantages que les règles d'association peuvent représenter, elles ont aussi des faiblesses qu'on peut résumer dans :

- Le temps énorme consacré à la recherche des ltemSets fréquents.
- La grande quantité des règles d'association générées.
- La difficulté d'évaluer la qualité des règles d'associations par des indices statiques ou par l'expert du domaine.
- La production des règles triviales et inutiles qui n'apportent pas de nouvelles informations.

6. Conclusion

Les règles d'associations tant traditionnelles que séquentielles sont des outils efficaces pour identifier des relations qui existent parmi les données. En plus d'être aisément interprétables, les règles d'associations peuvent faire découvrir aux analystes des associations inattendues. Ces techniques permettent donc de tirer

profit d'une grande quantité de savoir caché dans les données qui pourrait difficilement être découvert autrement. Ces découvertes peuvent ensuite être utilisées et intégrées dans les processus d'affaires de l'entreprise afin d'en améliorer les performances.

1. Introduction

Dans le domaine du data mining la classification avec ces différents états et différents modèles est la tâche la plus commune qui semble être une obligation humaine. A fin de comprendre notre vie quotidienne, nous somme constamment classifiées, catégorisés et évalués.

2. Les domaines d'application de la classification

La classification s'applique à un grand nombre d'activités humaines et convient en particulier au problème de la prise de décision automatisée. Il s'agit, par exemple, d'établir un diagnostic médical, de donner une réponse à la demande de prêt bancaire d'un client, …. En économie, la classification peut aider les analystes à découvrir des groupes distincts dans leur base clientèle, et à caractériser ces groupes de clients, en se basant sur des habitudes de consommations. En biologie, on peut l'utiliser pour dériver des taxinomies de plantes et d'animaux, pour catégoriser des gènes avec une ou plusieurs fonctionnalités similaires, pour mieux les structures propres aux populations. La classification peut tout aussi bien aider dans l'identification des zones de paysage similaire, utilisée dans l'observation de la terre, et dans l'identification de groupes de détenteurs de police d'assurance automobile ayant un coût moyen d'indemnisation élevé, ou bien dans la reconnaissance de groupes d'habitation dans une ville, selon le type, la valeur et la localisation géographique. Il est possible également de classifier des documents sur le Web, pour obtenir de l'information utile …

3. La définition de la classification [26]

La classification automatique consiste à regrouper divers objets (les individus) en sous-ensembles d'objets (les classes). Elle peut être :

- supervisée : les classes sont connues à priori, elles ont en général une sémantique associée
- non-supervisée (en anglais clustering) : les classes sont fondées sur la structure des objets, la sémantique associée aux classes est plus difficile à déterminer

Dans les deux cas, on a besoin de définir la notion de distance entre deux classes : le critère d'agrégation.

4. La classification supervisée [26]

4.1. Sélection des variables de la classification supervisée

La classification supervisée consiste à regrouper des objets dans des classes prédéfinies en fonction de leurs caractéristiques. Ses applications sont nombreuses : diagnostic médical (évaluation des risques de cancer, détection d'arythmie cardiaque), catégorisation des textes (classification des emails – spam ou non, classification des pages web), reconnaissance de forme (reconnaissance de visages, de chiffres manuscrits, …), etc.

Les tests numériques seront réalisés sur les benchmark problèmes via les bases de données biologiques de très grande dimension dont la sélection des gènes pour le diagnostic médical (cancer, diabète et prostate).

Il est à noter que la sélection des variables implique la norme zéro dans le modèle d'optimisation. La minimisation de la norme zéro est un problème NP-difficile qui attire l'attention de nombreux chercheurs durant ces dernières années.

D= $\{d_1, d_2, \ldots, d_i, \ldots, d_m\}$ un ensemble de documents représentés chacun par une description, $\overrightarrow{d_1}, \overrightarrow{d_2}, \ldots, \overrightarrow{d_i}$ et C= $\{C_1, C_2, \ldots, C_k, \ldots, C_n\}$un ensemble de classes, la classification supervisée suppose connues deux fonctions. La première fait correspondre à tout individu d_i une classe C_k Elle est définit au moyen de couples (d_i,c_k) donnés comme exemples au système. Le deuxième fait correspondre à tout individu d_i sa description di. La classification supervisée consiste alors à déterminer une procédure de classification :

$$C^f : d_i \rightarrow c_k$$

qui à partir de la description de l'élément détermine sa classe avec le plus faible taux d'erreurs. La performance de la classification dépend notamment de l'efficacité de la description. De plus, si l'on veut obtenir un système d'apprentissage, la procédure de classification doit permettre de classer efficacement tout nouvel exemple (pouvoir prédictif).

5. la classification non supervisée [26]

5.1. Sélection des variables en classification non supervisée

Un objet peut être présenté par des variables de différentes natures (quantitatives, qualitatives ou structurées). La nature des variables influe forcement sur la définition de similarité des objets et ce choix est très important. Dans les applications réelles, le nombre de variables représentés un objet est souvent grand. Ce nombre important de variable entraîne un grand coût de calcul (la taille de mémoire utilisée, ...). La question est donc de pouvoir choisir parmi les variables celles qui sont pertinentes et d'éliminer celles qui sont redondantes. Cette question de sélection des variables a été largement étudiée pour la classification supervisée mais reste encore ouverte pour la classification non-supervisée.

La classification non-supervisée est utilisée lorsque que l'on possède des documents qui ne sont pas classés et dont on ne connaît pas de classification. A la fin du processus de classification non-supervisée, les documents doivent appartenir à l'une des classes générées par la classification. On distingue deux catégories de classifications non-supervisées : hiérarchiques et non-hiérarchiques.

Dans la classification hiérarchique(CH), les sous-ensembles créés sont emboîtés de manière hiérarchique les uns dans les autres. On distingue la CH descendante (ou divisive) qui part de l'ensemble de tous les individus et les fractionne en un certain nombre de sous-ensembles, chaque sous-ensemble étant alors fractionné en un certain nombre de sous-ensembles, et ainsi de suite. Et la CH ascendante (ou agglomérative) qui part des individus seuls que l'on regroupe en sous-ensembles, qui sont à leur tour regroupés, et ainsi de suite. Pour déterminer quelles classes on va fusionner, on utilise le critère d'agrégation.

Dans la classification non-hiérarchique, les individus ne sont pas structurés de manière hiérarchique. Si chaque individu ne fait partie que d'un sous-ensemble, on parle de partition. Si chaque individu peut appartenir à plusieurs groupes, avec la probabilité p_i d'appartenir au groupe i, alors on parle de recouvrement.

6. Critère d'agrégation [26]

Le critère d'agrégation permet de comparer les classes deux à deux pour sélectionner les classes les plus similaires suivant un certain critère. Les critères les plus classiques sont le plus proche voisin, le diamètre maximum, la distance moyenne et la distance entre les centres de gravités.

6.1. Plus proche voisin

La distance entre la classe C_p et la classe C_q est la plus petite distance entre un élément de C_p et un élément de C_q

$D(C_p, C_q)=\min \{dist(i, j) ; i \square C_p, j \square C_q\}$

6.2. Diamètre maximum

La distance entre la classe C_p et la classe C_q est la plus grande distance entre un élément de C_p et un élément de C_q.

$D(C_p, C_q)=\max \{dist(i, j) ; i \square C_p, j \square C_q\}$

6.3. Distance moyenne

La distance entre la classe C_p et la classe C_q est la moyenne des distances entre les éléments de G_p et les éléments de G_q

$$D(Cp, Cq) = \frac{\Sigma i, j\{dist(i,j); i \epsilon C_p, j \epsilon C_q\}}{card(C_p) \times (C_q)}$$

Distance entre les centres de gravité Si G_p est le centre de gravité de la classe C_p et si G_q est le centre de gravité de la classe C_q alors la distance entre la classe G_p et la classe G_q est la distance entre leurs centres de gravités.

$D(C_p, C_q)=\min \{dist(G_p, G_q)$

Ce critère n'a de sens que si le calcul du centre de gravité possède lui-même un sens sur les données de l'étude.

6.4. Evaluation d'un système de classification [26]

Nous présentons ici une méthode permettant d'évaluer une classification supervisée, et des techniques classiques pour mesurer et comparer des systèmes de classifications non-supervisées.

6.4.1. Corpus de test

6.4.1.1. Cas supervisé

Pour tester la qualité d'une procédure de classification supervisée, on sépare aléatoirement les éléments classés entre une base de référence(R) et une base de test(T). Ensuite, on détermine la procédure de classification C^f à partir des exemples de la base de référence. Puis, on utilise C^f pour retrouver la classe des éléments de la base de test. Enfin, on estime l'erreur de la procédure de classification.

Pour estimer le taux d'erreur TE d'une procédure de classification C^f, une méthode simple est de calculer le nombre d'éléments mal classés sur le nombre d'éléments à classer :

$$\mathbf{TE(C^f) = \frac{1}{card(T)} \sum_{t=1}^{card(T)} (C^f(d_t) \neq C_{dt})}$$

où C_{dt} *e*st la classe d'origine de dt

Dans les cas de classifications simples, on peut être amené à calculer l'erreur résultant d'une classification purement aléatoire C□ pour la comparer avec l'erreur faite par notre procédure C^f afin de vérifier la performance de notre système.

Soit p_k la fréquence (ou probabilité à priori) de la classe k dans la base de test, on appelle erreur TE□ du système aléatoire :

$$\mathbf{TE_\alpha} - \sum_{k=1}^{c} (p_k)^2 = 1 - \sum_{k=1}^{c} (\frac{card(C_{k\backslash T}}{card(T)})^2$$

où c est le nombre de classes et card(C_k\T) est le nombre d'éléments de T qui sont dans la classeC_k.

L'erreur apparente TE(C^f) est dépendante de l'échantillon considéré. Cependant, plus le nombre d'éléments de l'échantillon est grand, plus l'erreur mesurée tend vers l'erreur réelle de C^f.

6.4.1.2. Cas non-supervisé

Dans le cas non-supervisé, on peut évaluer la classification par rapport à certaines de ces caractéristiques. On distingue d'une part, les caractéristiques numériques : le nombre de classes obtenues, le nombre d'éléments par classe, le nombre moyen d'éléments par classe, l'écart-type des classes obtenues, et d'autre part, les caractéristiques sémantiques. Par exemple, si à un document est associé un ensemble de mots clés, la sémantique associée à une classe pourra se composer des mots les plus fréquents dans la classe.

Pour évaluer l'homogénéité du nombre d'images par classe, on peut utiliser la variance :

V=σ2=Σ (card(C_k)moy)^2

$$V = \sigma^2 = \frac{1}{c}\sum_{k=1}^{c}(\text{card}(C_k) - \text{moy})^2$$

$$\text{moy} = \frac{1}{c}\sum_{k=1}^{c}\text{card}(C_k)$$

est le nombre moyen d'éléments par classe et C est le nombre de classes obtenues. L'écart-type σ=√V permet d'exprimer la dispersion dans la même unité que la moyenne.

7. Choix de la méthode de classification [27]

Une fois l'espace vectoriel réduit nous procédons au calcul du modèle de classification. Ce modèle sera ensuite utilisé pour l'évaluation des textes du jeu de test.

Nous avons utilisé plusieurs méthodes de classification. Elles sont fondées sur quatre méthodes principales.

Nous avons également testé d'autres procédures de classification dont les performances se sont révélées moins intéressantes.

Le choix de la procédure de classification s'est fait sur chaque ensemble d'apprentissage ou corpus. La sélection fut très simple, nous avons conservé la méthode de classification la plus performante pour un corpus donné. Les mesures de performances sont décrites ci après.

Nous décrivons brièvement ci-après les trois méthodes de classification.

En voici la liste :

- La classification probabiliste utilisant la combinaison de la loi de Bayes et de la loi multinomiale,
- La classification par les machines à vecteurs support S.V.M type SMO.
- La classification par les machines à vecteurs support S.V.M type Libsvm.
- La classification par la méthode des réseaux RBF (Radial Basis Function)
- La classification par boosting sur le classifieur de Bayes

7.1. Classifieur de Bayes Multinomial

Cette technique est classique pour la catégorisation de textes. Elle combine l'utilisation de la loi de Bayes bien connue en probabilités et la loi multinomiale. Nous avons simplement précisé le calcul de la loi à priori en utilisant l'estimateur de Laplace pour éviter les biais dus à l'absence de certains mots dans un texte.

7.1.1 La Classification Bayesienne [28]

La technique Bayesienne de Classification repose sur une fonction de vérité probabiliste et le règle de Bayes.

Dans un système de vérité probabilisé, la valeur de vérité de la proposition une probabilité :

$$p(w_k) = p(E \square w_k) = \text{vérité}(E \square w_k)$$

Par convention on écrit $p(w_k)$ pour $p(E \square w_k)$.

Le critère de décision est de minimiser le nombre d'erreur. Dans un système probabiliste, ça revient de minimiser la probabilité d'erreur. Ceci est équivalent à choisir la classe le plus probable.

$W\hat{}_k$ = Decider (E □ w_k) = arg-max $_{wk}${p(w_k | X)}

Pour estimer la probabilité nous utilisons les caractéristiques, $\vec{X}$ de l'événement.

Considère le cas D = 1 et K = 2. Dans ce cas, le domaine d'X est un axe.

La classification est équivalente à un découpage du domaine d'X en deux zones : Z_1 et Z_2.

$w\hat{}_1$ si X □ Z_1 et w^2 si X □ Z_2

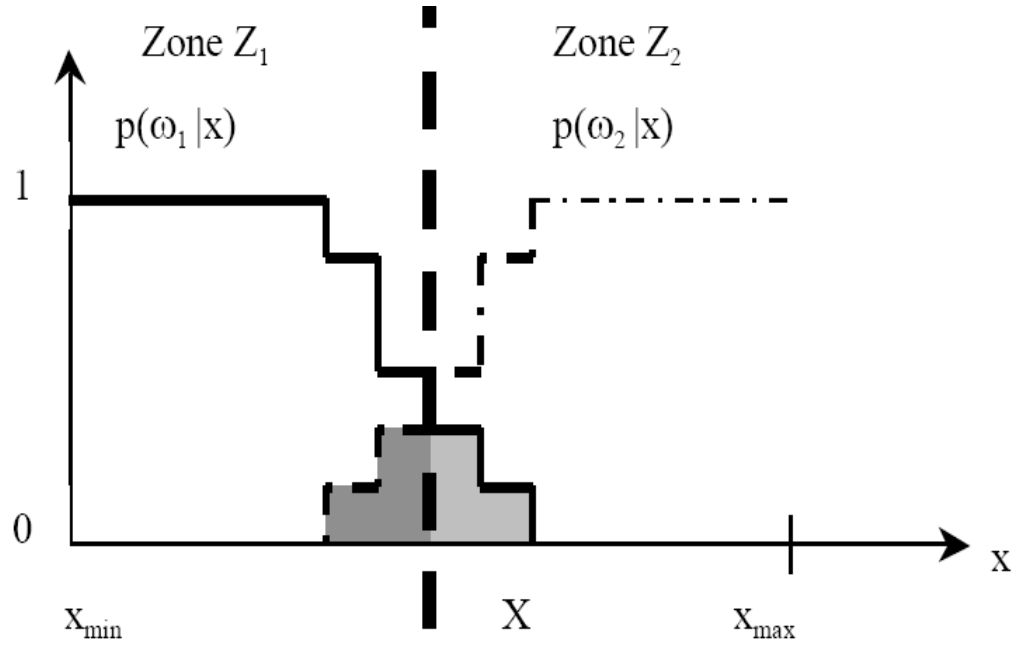


Figure 4.1 : découpage de X en deux zones Z1et Z2

La probabilité d'erreur est la somme des probabilités de $p(w_2)$ en Z_1 et la somme de probabilité de $p(w_1 |X)$ en zone 2.

p(erreur) = □ $_{Z1}$p(w2 | X) + □$_{Z2}$p(w1 | X)

La minimum est atteint quand :

Donc $p(g_k(X))$ = arg-max w_k {p(w_k | X)}

Dans ce cas, nous avons utilisé arg-max w_k{p(w_k | X)} en tant que fonction de décision et $g_k(X) = p(w_k | X)$ comme la fonction de discrimination

Probabilité :

La probabilité peut être assimilée à la fréquence d'occurrence.

Par exemple, si nous observons M événements et M_k de ces événements sont issus de la classe w_k , on dit que la probabilité qu'un événement E soit issu de la classe w_k est :

$$P(E \in W_k) = P(W_k) = \lim_{M \to \infty} \{\frac{M_k}{M}\}$$

$$\text{pour M finit on écrit} \quad P(W_k \approx \frac{M_k}{M})$$

En règle générale, il faut observer $M_k \geq 10$ pour que cette approximation soit raisonnable.

Cette approche se généralise pour la probabilité des caractéristiques.

Soit une caractéristique de valeur entier, X, tel que x □ $[X_{min}, X_{max}]$

On peut représenter p(X=x) avec une table de fréquence, h(x), composé de $N = X_{max} - X_{min} + 1$ cellules.

Pour estimer p(X), on observe un ensemble aléatoire de M événements avec leurs caractéristiques. Cet ensemble est dit l'ensemble d'entrainement (training set) $\{X_m\}$

□ Xm □ $\{X_m\}$: $h(X_m) := h(X_m) + 1$

Ensuite :

La probabilité à priori que X=x :

$$P(X = x) \approx \frac{1}{M} h(x)$$

par convention on écrit P(X=x) comme

$$p(X) \approx \frac{1}{M} h(x)$$

On peut voir chaque cellule comme une sorte de classe de la caractéristique X.

Donc, en règle générale, il faut plus que 10 exemples par cellule de h(x).

Pour N cellules, il faut un ensemble d'entrainement de M > 10 N exemples.

Dans le cas de K classes, la probabilité conditionnelle de X:

$$p(X = x \backslash W_k) \approx \frac{1}{M_k} h_k(x)$$

par convention on écrit p(X=x | $w_{k)}$ comme p(X | w_k)

Et il faut M_k > 10 N pour chaque classe.

7.1.2 Le Règle de Bayes

Dans la dernière séance nous avons vu que pour deux classes indépendantes d'événement , A et B. .

p(E □ A □ E □ B) = p(E □ B | E □ A) p(E □ A) = p(E □ A | E □ B) p(E □ B)

ou bien

p(A □ B) = p(B | A) p(A) = p(A | B) p(B)

Ceci peut se généraliser pour X et w_k. Soit A ≡ (X=x) et B ≡ (E □ w_k)

p(X=x □ E □ w_k) = p(E □ w_k| X=x) p(X=x) = p(X=x | E □ w_k) p(E □ w_k)

ou bien

p(X □ w_k) = p(w_k | X) p(X) = p(X | w_k) p(w_k)

Donc :

$$P(W_k \backslash X) = \frac{P(X \backslash W_k) P(W_k)}{P(X)}$$

En utilisant une table de fréquence pour les probabilités, on obtient une résulte intéressante :

$$p(W_k \backslash X) = \frac{P(X \backslash W_k) P(W_k)}{P(X)} \approx \frac{\frac{M_k}{M} \frac{1}{M_k} h_k(X)}{\frac{1}{M} h(X)} = \frac{h_k(X)}{h(X)}$$

Cette technique marche également pour les vecteurs de caractéristiques XL' histogramme est une table a D dimensions : h(X)

ce qui donne :

probabilité conditionnelle de X :

$$P(\vec{X} \backslash W_k) \approx \frac{1}{M_k} h_k(\vec{X})$$

Probabilité à priori de X :

$$P(\vec{X}) \approx \frac{1}{M} h_k(\vec{X})$$

$$P(W_k \backslash \vec{X}) = \frac{P(\vec{X} \backslash W_k) P(W_k)}{P(\vec{X})} \approx \frac{\frac{M_k}{M} \frac{1}{M_k} h_k(\vec{X})}{\frac{1}{M} h(\vec{X})} = \frac{h_k(\vec{X})}{h(\vec{X})}$$

Cette technique s'avère très utile dans les cas où il y a suffisamment d'échantillons pour faire un histogramme valable. Par exemple quand on traite des images ou les signaux. Mais il faut toujours M > 10 N exemples. Pour un tableau de D dimensions, avec B, cellules pers dimensions, celui-ci devient

$M > 10\ B^D$

M accroitre exponentiellement avec le nombre de dimensions.

Un détaille :

Que faire si la masse d'exemple est insuffisante : M < 10 N ?

Que faire si x n'est pas entier ?

Dans ces cas, on peut faire appel à une fonction paramétrique pour p(X).

La fonction paramétrique la plus utilisée est la loi Normale.

7.2. Classifieur par la méthode des Machines à Vecteurs Support (S.V.M.)[27]

Cette technique consiste à délimiter par la frontière la plus large possible les différentes catégories des échantillons de l'espace vectoriel du corpus d'apprentissage. Les vecteurs supports constituent les éléments délimitant cette frontière.

Plusieurs méthodes de calcul des vecteurs supports peuvent être utilisées :

- une méthode linéaire
- une méthode polynomiale

- une méthode fondée sur la loi gaussienne normale
- une méthode fondée sur la fonction sigmoïde

Nous avons essentiellement utilisé la méthode linéaire et celle fondée sur la loi.

7.3. Classifieur par la méthode des réseaux RBF (Radial Basis Function) [27]

Cette technique implémente un réseau de neurones à fonctions radiales de base. Elle utilise un algorithme de « clustering » de type « k-means » et utilise au dessus de cet algorithme une régression linéaire. Les gaussiennes multivariables symétriques sont adaptées aux données de chaque « cluster ». Toutes les données numériques sont normalisées (moyenne à zéro, variance unitaire).

Cette technique est présentée dans (Parks & Sandberg, 1991).

7.4 Classifieur par la méthode adaboost sur le classifieur Naive Bayes Multinomial [27]

Ce classifieur a pour objectif de doper les performances d'un classifieur associé par l'utilisation de la méthode Adaboost .Cet algorithme améliore souvent de façon importante les résultats d'un classifieur mais quelquefois déprécie les résultats. Dans le cas du classifieur de Bayes nous avons constaté que les résultats de Adaboost étaient souvent légèrement meilleurs.

8. Conclusion

La difficulté majeure de classification réside dans la non convexité du modèle d'optimisation associé d'une part, et la taille très grande de ce modèle d'autre part, vu la dimension et le volume de masse de données considérées. Avec les différentes mesures de distance, tous les modèles d'optimisation en classification sont de la forme DC ou peuvent être transformés en une programmation DC par les techniques de reformulation.

Dès lors DCA peut être développée pour la résolution de ces problèmes, en particulier pour les problèmes de très grande dimension.

1. Les réseaux de neurones

1.1. Introduction

On peut dire que parmi les buts essentiels de la recherche scientifique est de développer des machines intelligentes qui peuvent exécuter toute tâche pénible et encombrante. Parmi les technologies qui sont consacrées à ce type de recherche: l'intelligence artificielle et les systèmes de neurones artificiels. Ces derniers sont basés essentiellement sur le mécanisme de transmission nerveuse d'un être humain.

L'élément fonctionnel essentiel du système nerveux est la cellule nerveuse ou neurone qui a pour rôle d'élaborer l'information reçue et transmettre les résultats à d'autres neurones. Le cerveau humain développe mieux les solutions intelligentes qu'un ordinateur, cependant ce dernier est rapide dans l'exécution des opérations.

1.2. Historique [25]

Pour faire un bref historique, les réseaux de neurone ont connu leurs débuts dans les années 1943 avec les travaux de Warren Mc Culloch & Walter Pitt sur le « neurone formel ». En 1949, D. Hebb présente dans son ouvrage « The Organization of Behavior » une règle d'apprentissage. De nombreux modèles de réseaux aujourd'hui s'inspirent encore de la règle de Hebb En 1958 les travaux de Franck Rosenblatt sur « le perceptron » proposent au Cornell Aeronautical Laboratory le premier algorithme d'apprentissage permettant d'ajuster les paramètres d'un neurone. Il est à présent communément admis que le perceptron, comme classifieur linéaire, est le réseau de neurones le plus simple.

En 1969, Minsky et Papert publient le livre Perceptrons dans lequel ils utilisent une solide argumentation mathématique pour démontrer les limitations des réseaux de neurones à une seule couche. Ce livre aura une influence telle que la plupart des chercheurs quitteront le champ de recherche sur les réseaux de neurones. En 1982, Hopfield propose des réseaux de neurones associatifs et l'intérêt pour les réseaux de neurones renaît chez les

scientifiques. En 1986, Rumelhart, Hinton et Williams publient l'algorithme de la rétropropagation de l'erreur, qui permet d'optimiser les paramètres d'un réseau de neurones à plusieurs couches. À partir de ce moment, la recherche sur les réseaux de neurones connaît un essor fulgurant et les applications commerciales de ce succès académique suivent au cours des années 90.

1.3. Définition [22]

Un réseau de neurones est un ensemble de méthodes d'analyse et de traitements des données permettant de construire un modèle de comportement à partir de données qui sont des exemples de ce comportement. Un réseau de neurones est constitué d'un graphe pondéré orienté dont les nœuds symbolisent les neurones. Ces neurones possèdent une fonction d'activation qui permet d'influencer les autres neurones du réseau. Les connexions entre les neurones, que l'on nomme liens synaptiques, propagent l'activité des neurones avec une pondération caractéristique de la connexion. On appelle poids synaptique la pondération des liens synaptiques.
Les neurones peuvent être organisés de différentes manières, c'est ce qui définit 1'architecture et le modèle du réseau. L'architecture la plus courante est celle dite du perceptron multicouche.

1.4. Applications

Les réseaux de neurones sont essentiellement utilisés pour faire de la classification. Construit à partir d'exemples de chaque classe qu'il a appris, un réseau de neurones est normalement capable de déterminer à quelle classe appartient un nouvel élément qui lui est soumis.

1.5. Fonctionnement

- La construction de la structure du réseau.
- La constitution d'une base de données de vecteurs représentant au mieux le domaine à modéliser. Celle-ci est partagée en deux parties: une partie servant à l'apprentissage du réseau (on parle de base d'apprentissage) et une autre partie aux tests de cet apprentissage (on parle de base de test).
- Le paramétrage du réseau par apprentissage. Au cours de 1'apprentissage, les vecteurs de données de la base d'apprentissage sont présentés

séquentiellement et plusieurs fois au réseau. Un algorithme d'apprentissage ajuste le poids du réseau afin que les vecteurs soient correctement appris. L'apprentissage se termine lorsque l'algorithme atteint un état stable.

- La phase de reconnaissance qui consiste à présenter au réseau chacun des vecteurs de la base de test. La sortie correspondante est calculée en propageant les vecteurs à travers le réseau. La réponse du réseau est lue directement sur les unités de sortie et comparée à la réponse attendue. Une fois que le réseau présent des performances acceptables, il peut être utilisé pour répondre au besoin qui a été à l'origine de sa construction.

2. Modèle biologique [25]

2.1. Définition et structure

Le bloc principal du système nerveux est le neurone. Il transmet l'information reçue vers les diverses parties du corps. Il est constitué,

- D'un corps cellulaire nommé Noyau.
- Des plusieurs épines semblables propagées dans le corps cellulaires nommées dendrites. Leur rôle est de capter les signaux qui proviennent du neurone.
- D'une seule fibre nerveuse nommé axone, qui sert à connecter le corps cellulaire aux autres neurones. L'axone est un moyen de transport pour les signaux émis par le neurone.
- Les connexions entre les neurones se font par l'intermédiaire du corps cellulaire ou les dendrites en jonctions nommées synapses. Les synapses servent à limiter plus ou moins l'amplitude des signaux qui passent d'un neurone à un autre, comme est illustré dans la figure 5.1

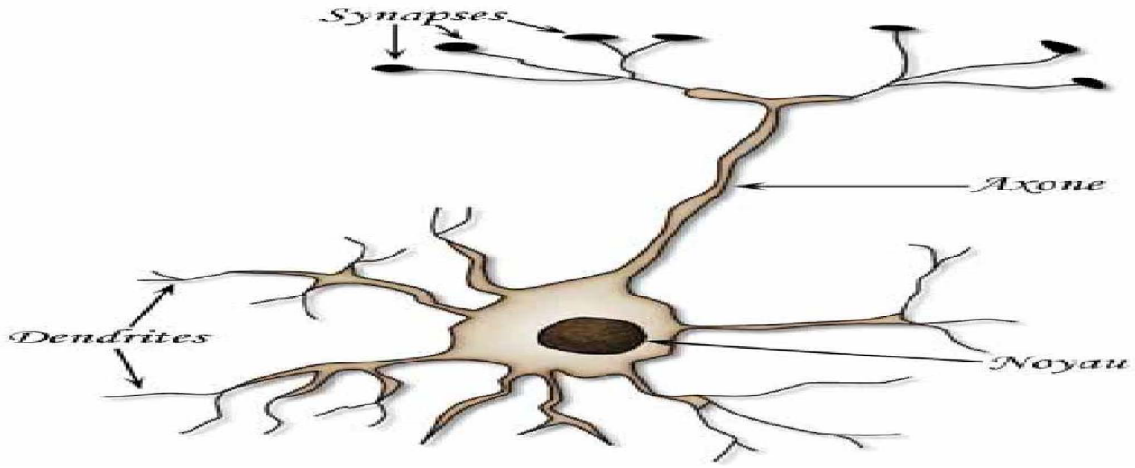


Figure 5. 1 : Représentation simplifiée de neurone.

Le modèle neuronique dont la forme la plus simple est représentée par la figure 5.2

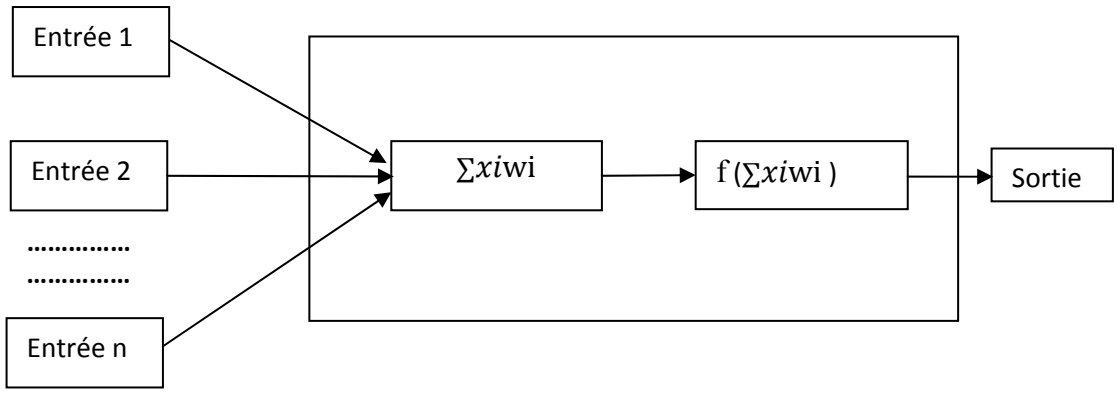


Figure 5.2 : Modèle d'un neurone artificiel

La figure 5.2 représente l'architecture générale d'un neurone artificiel (formel), y compris les différentes couches qui le constitue, la couche d'entrée, la couche cachée et la couche de sortie.

Dans cette figure, on utilise les notations suivantes:

- $(x_i)_{1 <= i < k}$ est les *k* informations parvenant au neurone,
- w_i est le poids associé à la connexion entre le nœud i et le nœud observé, la i -ème information qui parviendra au neurone sera donc en fait ($w_i * x_i$). Il y a toutefois un "poids" supplémentaire, qui va représenter ce que l'on appelle le coefficient de biais également appelé seuil
- f est la fonction de transfert associée au nœud observé.

2.2. Fonctionnement

Le mécanisme de fonctionnement d'un neurone est de recevoir, grâce à ces dendrites, les signaux émis par les autres neurones, puis décider, à partir des données reçues, d'émettre ou non un signal à ses semblables le long de son axone.

Plus précisément, le soma recueille l'ensemble des informations reçues par les dendrites et effectue la sommation. En raison de sa dimension, l'intégration somatique est aussi temporelle. Si le potentiel somatique dépasse un certain seuil, il y a émission d'un potentiel d'action. Le signal, très bref, est transmis sans atténuation le long de l'axone et réparti sur le neurone cible.

2.3. Plasticité synaptique

La notion de plasticité synaptique, c'est à dire le mécanisme de modification progressive des couplages entres neurones, chaque neurone présente deux états (actif ou inactif).

L'efficacité synaptique augmente seulement si les deux éléments sont actifs simultanément, donc elle prévoit exclusivement le renforcement des efficacités synaptiques, c'est à dire que le poids de la synapse ne peut qu'augmenter, chose qui conduit à une fatale saturation du réseau. Nous sommes donc, obligés de préciser un certain intervalle de coïncidence.

3. Étude et synthèse d'un réseau de neurone formel (artificiel) [22]

Un réseau de neurone est une structure de traitement parallèle et distribué d'informations comportant plusieurs éléments de traitement Neurone, qui peut posséder des mémoires locales et exécuter les opérations de traitements sur des informations locales.

Ils sont interconnectés les uns aux autres avec des canaux des signaux unidirectionnels. La synthèse d'un réseau de neurone formel est basée sur des caractéristiques similaires à celle d'un réseau de neurone biologique. Ces caractères sont:

- Il est composé d'un nombre très grand d'éléments de traitement simple.
- Chaque élément de traitement est connecté à plusieurs éléments voisins.
- Le fonctionnement d'un réseau est basé sur le mécanisme de modification de poids de connexion pendant la phase d'apprentissage.

3.1. Neurone formel

Un neurone formel est un petit automate qui réalise la somme pondérée des poids $W_1, W_2, \ldots\ldots, W_n$ des entrées $X_1, X_2, \ldots\ldots, X_n$ dont les valeurs sont estimées dans la phase d'apprentissage. Ils constituent "la mémoire" ou "connaissance répartie" du réseau.

La figure 5.3 suivant montre la structure d'un neurone formel

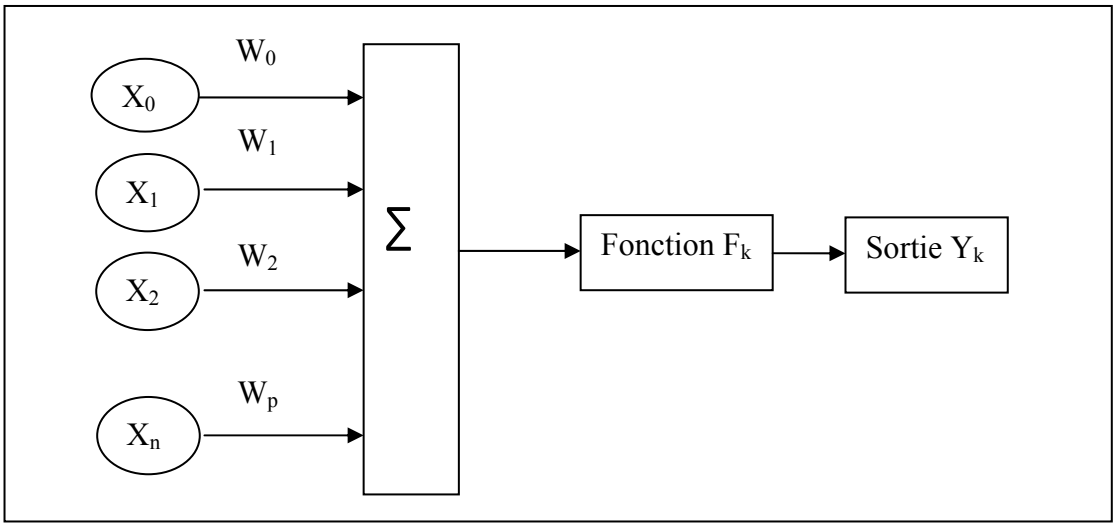


Figure 5.3 : Structure du neurone formel.

- x_n : l'entrée n du k-ème neurone.
- w_i : poids associé au i-ème entrée du neurone k.
- w_0 : seuil , du i -ème neurone
- F_k : fonction d'activation.
- y_k : sortie du neurone k.

Par analogie, le neurone formel est un modèle qui se caractérise par un état interne, des signaux d'entrée x_1, ……., x_n et une fonction d'activation

$$\forall \ 1 \leq i \leq p, \ y = f\left(\sum_{i=0}^{k} xi * w_i\right) = f\left(\left(\sum_{i=1}^{k} xi * w_i\right) - w_0\right)$$

Si $\sum_{i=0}^{k} xi * w_i$ > seuil → activation du neurone de sortie

3.2. Fonction d'activation

Afin de déterminer une valeur en sortie, une fonction appelée fonction d'activation (ou de transfert), est appliquée à cette valeur.

La fonction d'activation la plus généralement rencontrée est une fonction sigmoïde telle que « si la somme des entrées est supérieure à un seuil, alors le neurone de sortie est activé; sinon, rien »

Exemples de fonctions d'activation

Sigmoïde ou logistique

$$f(x) = \frac{1}{1 + e^{-x}}$$

Tangente hyperbolique

$$f(x) = \frac{2e^{x}}{1 + e^{x}} - 1$$

Linéaire

$$f(x) = x$$

3.3. Les étapes d'un réseau de neurones [23]

Les étapes dans la mise en œuvre d'un réseau de neurones pour la prédiction ou la classification sont:

1- L'identification des données en entrée et en sortie
2- La normalisation de ces données
3- La constitution d'un réseau avec une structure adaptée
4- L'apprentissage du réseau
5- Le test du réseau
6- L'application du modèle généré par l'apprentissage
7- La démoralisation des données en sortie

3.4. Structure des réseaux de neurones

La structure du réseau de neurones, encore appelée « architecture» ou « topologie» du réseau de neurones, est le nombre de couches et de nœuds, la façon dont sont interconnectés les différents nœuds (choix des fonctions de combinaison et de transfert) et le mécanisme d'ajustement des poids.

3.4.1. Réseau mono-couches

Dans ce type de réseau, il y a une seule couche cachée, qui relie les cellules d'association *(*couche d'entrée*)* aux cellules de décision (couche de sortie). C'est la seule couche de connexion modifiable.

Un réseau de neurones monocouche, est caractérisé de la manière suivante.

- Il possède n informations en entrée.
- Il est composé de k neurones, que l'on représente généralement alignés verticalement. Chacun peut en théorie avoir une fonction d'activation différente.
- Chacun des k neurones est connecté aux n informations d'entrée.

Le réseau de neurones possède ainsi n informations en entrée et k sorties, chaque neurone renvoyant sa sortie.

Chaque neurone de la couche donnera donc une sortie. Une utilisation courante est que chaque neurone de la couche représente une classe. Pour un exemple X donné, on obtient la classe de cet exemple en prenant la plus grande des k sorties.

La structure d'un réseau monocouche est telle que des neurones organisés en entrée soient entièrement connectés à d'autres neurones organisés en sortie par une couche modifiable de poids.

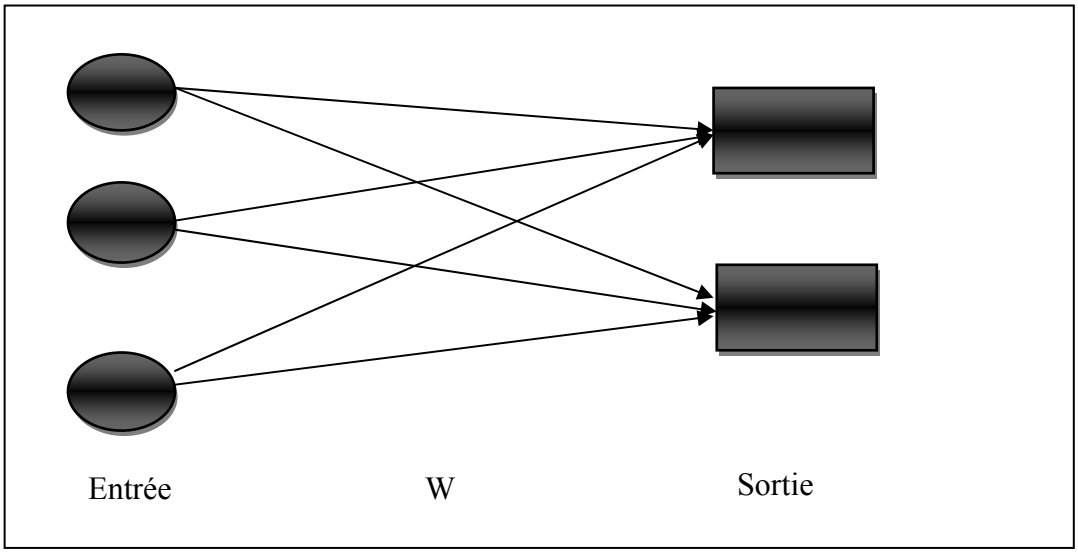


Figure.5.4: Réseau Monocouche.

3.4.2. Réseau muIti-couches

Pour maîtriser les limitations d'un réseau mono-couche, on utilise un réseau multi-couche, où la sortie n'est connectée à l'entrée qu'après quelques couches de neurones intermédiaires apportant une richesse à la structure pour accroître la capacité de réseau. Notons que les couches internes n'ont aucune connexion prédéfinie, elles servent seulement à contribuer à l'obtention de résultats souhaités à la sortie.

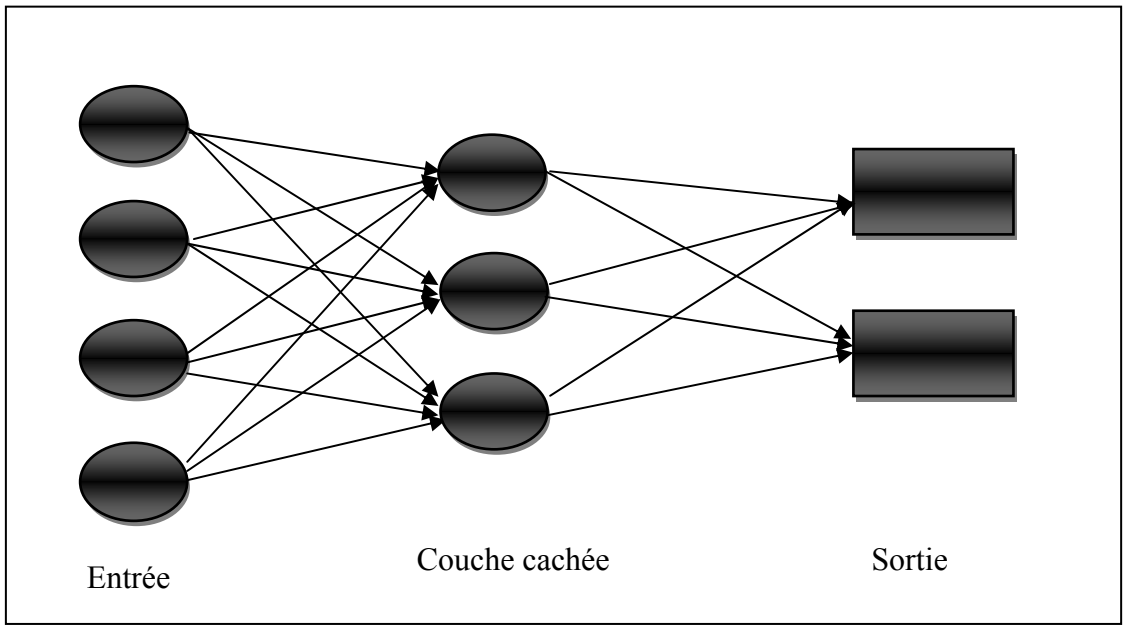


Figure.5.5 : Réseau Multicouche

3.5. Fonctionnement d'un réseau

Un réseau de neurone peut fonctionner en deux modes, parallèle ou séquentiel.

Dans le mode parallèle tous les neurones calculent leurs nouvelles activations et leurs sorties, et les transmettent aux neurones auxquels ils sont connectés, à chaque top d'horloge. Contrairement, au mode séquentiel, un seul neurone calcule sa nouvelle activation et sa sortie puis les transmet aux neurones auxquels il est connecté, à chaque top d'horloge. Donc, le calcule est fait en fonction des entrées des neurones au top d'horloge précédent.

3.6. Apprentissage

On peut définir l'apprentissage par la modification des interactions entre neurones, l'apprentissage consiste donc à ajouter les poids synaptiques de telle façon que le réseau présente un certain comportement désiré.

Les procédures d'apprentissage peuvent se subdiviser, en deux grandes catégories:

Apprentissage supervisé ou apprentissage non supervisé.

3.6.1. Apprentissage supervisé

Dans ce type d'apprentissage, le réseau s'adapte par comparaison entre le résultat qu'il a calculé, en fonction des entrées fournies, et la réponse attendue en sortie. Ainsi, le réseau va se modifier jusqu'a ce qu'il trouve la bonne sortie, c'est-à-dire celle attendue, correspondant à une entrée donnée.

Le renforcement est en fait une sorte d'apprentissage supervisé et certains auteurs le classe d'ailleurs, dans la catégorie des modes supervisés. Dans cette approche le réseau doit apprendre la corrélation entrée/sortie via une estimation de son erreur, c'est-à-dire du rapport échec/succès. Le réseau va donc tendre à maximiser un index de performance qui lui est fourni, appelé signal de renforcement. Le système étant capable ici, de savoir si la réponse qu'il fournit est correcte ou non, mais il ne connait pas la bonne réponse.

3.6.2. Apprentissage non supervisé

Les réseaux, utilisant l'apprentissage non supervisé, sont souvent appelés auto-organisateurs, ou encore à apprentissage compétitif. Dans ce type d'apprentissage la connaissance de la sortie désirée n'est pas nécessaire c'est à dire que le réseau s'auto-organise et organise les entrées qui sont présentées de façons à optimiser un critère de coût donné.

3.6.3. Règles d'apprentissage

Règle de correction d'erreurs, Si on considère y comme étant la sortie calculée par le réseau, et d la sortie désirée, le principe de cette règle est d'utiliser l'erreur (d-y), afin de modifier les connexions et de diminuer ainsi l'erreur globale

du système. Le réseau ajuste ensuite les poids des différents nœuds.

3.7. Normalisation des données

Les données utilisées dans un réseau de neurones doivent être numériques et leurs modalités comprises dans l'intervalle [0,1], ce qui implique, quand ce n'est pas le cas, une normalisation des données. Pour que le travail de normalisation soit correct, il faut, bien entendu, que le jeu de données d'apprentissage couvre toutes les valeurs rencontrées dans la population tout entière, et, en particulier, les valeurs extrêmes des variables continues.

- **Variables continues**

  Même en les normalisant, les variables continues peuvent connaître le problème d'écrasement des valeurs normales les valeurs extrêmes. Plusieurs moyens existent pour bien normaliser ce type de variable. On peut discrétiser la variable et la remplacer, par exemple, par ses quartiles. On peut normaliser, non pas la variables, mais le logarithme de cette variable, qui « distend» le début de l'échelle. On peut normaliser la variable linéairement, pour ses valeurs comprises entre -3 et +3 fois l'écart -type σ autour de la moyenne μ et envoyer les valeurs à $\mu - 3\sigma$ sur 0, et les valeurs supérieures à $\mu + 3\sigma$ sur 1.

4. Développement d'un réseau de neurones [25]

Procédure de développement d'un réseau de neurones.

Le cycle classique de développement peut être séparé en sept étapes :

1- la collecte des données.

2- l'analyse des données.

3- la séparation des bases de données.

4- le choix d'un réseau de neurones.

5- la mise en forme des données.

6- l'apprentissage.

7-la validation.

4.1. Collecte des données

L'objectif de cette étape est de recueillir des données, à la fois pour développer le réseau de neurones et pour le tester. Dans le cas d'applications sur des données réelles, l'objectif est de rassembler un nombre de données suffisant pour constituer une base représentative des données susceptibles d'intervenir en phase d'utilisation du système neuronal. La fonction réalisée résultant d'un calcul statistique, le modèle qu'il constitue n'a de validité que dans le domaine où on l'a ajusté. En d'autres termes, la présentation de données très différentes de celles qui ont été utilisées lors de l'apprentissage peut entraîner une sortie totalement imprévisible.

4.2. Analyse des données

Il est souvent préférable d'effectuer une analyse des données de manière à déterminer les caractéristiques discriminantes pour détecter ou différencier ces données. Ces caractéristiques constituent l'entrée du réseau de neurones. Notons que cette étude n'est pas spécifique aux réseaux de neurones, quelque soit la méthode de détection ou de classification utilisée, il est généralement nécessaire de présenter des caractéristiques représentatives Cette détermination des caractéristiques a des conséquences à la fois sur la taille du réseau (et donc le temps de simulation), sur les performances du système (pouvoir de séparation, taux de détection), et sur le temps de développement (temps d'apprentissage). Une étude statistique sur les données peut permettre d'écarter celles qui sont aberrantes et redondantes. Dans le cas d'un problème de classification, il appartient à l'expérimentateur de déterminer le nombre de classes auxquelles ses données appartiennent et de déterminer pour chaque donnée la classe à laquelle elle appartient.

4.3. Séparation des bases de données

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données : une base pour effectuer l'apprentissage et une autre pour tester le réseau obtenu et déterminer ses performances. Afin de contrôler la phase d'apprentissage, il est souvent

préférable de posséder une troisième base de données appelée « base de validation croisée ».

Les avantages liés à l'utilisation de cette troisième base de données seront exposés dans les sections suivantes. Il n'y a pas de règle pour déterminer ce partage de manière quantitative. Il résulte souvent d'un compromis tenant compte du nombre de données dont on dispose et du temps imparti pour effectuer l'apprentissage. Chaque base doit cependant satisfaire aux contraintes de représentativité de chaque classe de données et doit généralement refléter la distribution réelle, c'est à dire la probabilité d'occurrence des diverses classes.

4.4. Choix d'un réseau de neurones

Il existe un grand nombre de types de réseaux de neurones, avec pour chacun des avantages et des inconvénients. Le choix d'un réseau peut dépendre :

- de la tâche à effectuer (classification, association, contrôle de processus,)
- de la nature des données
- d'éventuelles contraintes d'utilisation temps-réel
- des différents types de réseaux de neurones disponibles dans le logiciel de simulation que l'on compte utiliser.

  Ce choix est aussi fonction de la maîtrise ou de la connaissance que l'on a de certains réseaux, ou encore du temps dont on dispose pour tester une architecture prétendue plus performante.

4.5. Mise en forme des données pour un réseau de neurones

De manière générale, les bases de données doivent subir un prétraitement afin d'être adaptées aux entrées et sorties du réseau de neurones. Un pré traitement courant consiste à effectuer une normalisation appropriée, qui tienne compte de l'amplitude des valeurs acceptées par le réseau.

4.6. Apprentissage du réseau de neurones

Tous les modèles de réseaux de neurones demandent un apprentissage. Plusieurs types d'apprentissages peuvent être adaptés à un même type de réseau de neurones. Les critères de choix sont souvent la rapidité de

convergence ou les performances de généralisation. Le critère d'arrêt de l'apprentissage est souvent calculé à partir d'une fonction de coût, caractérisant l'écart entre les valeurs de sortie obtenues et les valeurs de références (réponses souhaitées pour chaque exemple présenté).

4.7. Validation

Une fois le réseau de neurones entraîné (après apprentissage), il est nécessaire de le tester sur une base de données différente de celles utilisées pour l'apprentissage.

Ce test permet à la fois d'apprécier les performances du système neuronal et de détecter le type de données qui pose problème. Si les performances ne sont pas satisfaisantes, il faudra soit modifier l'architecture du réseau, soit modifier la base d'apprentissage.

5. Exemple de Classification par les réseaux de neurone.

Supposons que l'on désire classer des formes en deux catégories, A ou B, en fonction de certaines caractéristiques de ces formes ; on peut définir une fonction *f* qui vaut +1 pour toutes les formes de la classe A et -1 pour toutes les formes de la classe B. Les réseaux de neurones sont de bons candidats pour réaliser une approximation de cette fonction F.

Le perceptron est la forme la plus simple de réseau de neurones, et permet de classifier correctement des objets appartenant à deux classes. Il consiste en un seul neurone qui possède un seuil ainsi qu'un vecteur de poids synaptiques.

Le perceptron associe à chaque classe une fonction discriminante linéaire qui s'exprime:

$$F(x) = W^t * X$$

telle que

$W = [w_0, w_1, \ldots w_n]$ est un vecteur de coefficients de pondération.

$X = [-1, x, \ldots x_n]$ le vecteur de caractéristiques augmenté d'un objet à classifier.

Dans le cas d'un problème à deux classes, la règle de classification s'écrit:

$X \epsilon A$ ssi $F_1(x) \geq F_2(x)$, Sinon $X \epsilon B$

Ce qui peut également s'exprimer selon:

$X \epsilon A$ ssi $F_{12}(x) = F_1(x) - F_2(x) \geq 0$, Sinon $X \epsilon B$

C'est cette règle de décision qui est réalisée dans le perceptron, grâce à l'utilisation de la fonction de signe, définie par

$$signe\ (y) = \begin{cases} 1 & si\ wi * xi > 0 \\ -1 & sinon \end{cases}$$

Comme fonction d'activation du neurone. La sortie du neurone vaut alors +1 si l'objet *X* appartient à la classe A, et -1 dans le cas contraire.

6. Conclusion

Le grand avantage des réseaux de neurones réside dans leur capacité d' apprentissage automatique, ce qui permet de résoudre des problèmes sans nécessiter l'écriture de règles complexes, tout en étant tolérant aux erreurs. Cependant, ce sont de véritables boites noires qui ne permettent pas d'interpréter les modèles construits. En cas, d'erreurs du système, il est quasiment impossible d'en déterminer la cause.

Il serait pertinent de faire une auto évaluation de ce modeste travail, une critique positive nous permettra sûrement de repérer nos difficultés et de nous situer par rapport aux objectifs qu'on s'était assignés au départ.

Data Mining (fouille de données) est un domaine très vaste qui est connu sous l'appellation Extraction de Connaissances à partir de Données (ECD), le data mining peut s'entendre de l' application des techniques statistiques, d'analyse des données et d'intelligence artificielle à l'exploration et l'analyse sans à priori de grandes bases de données informatiques qui sont préparées pour les utilisées dans un data warehouse , en vue d'en extraire des informations nouvelles et utiles pour le détenteur de ces données.

Pour aboutir à ces nouvelles informations, le Data mining utilise plusieurs techniques dont on a étudié en premier lieu la règle d'association qui consiste à déterminer les valeurs associées parmi les données en second la classification qui consiste à étudier les caractéristiques des données pour leurres attribuer une classe prédéfinie et en dernier les réseaux de neurones qui sont utilisés comme méthodes pour la classification.

Il est clair que notre étude est loin d'être complète. Il nous semble néanmoins avoir acquis les connaissances essentielles, lesquelles conjuguées avec beaucoup de persévérance.

Références bibliographique

[1] Clowers,première revue du domaine data mining and knowledge discovery journal 1977.

[2] Melle. CHAMI Djazia Mémoire En vue de l'obtention du diplôme de Magister en informatique, Une plate forme orientée agent pour le data mining (2009-2010).

[3] The Gartner Group, www.gartner.com.

[4] D. HAND, H. MANNILA et P. SMYTH, Principles of Data Mining,MIT Press, Cambridge, MA, 2001.

[5] P. CABENA, P. HADJINIAN, R. STADLER, J. VERHEES et A. ZANASI, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.

[6] E-G. TALBI, Fouille de données (Data Mining) : Un tour d'horizon, Laboratoire d'Informatique Fondamentale de Lille.

[7] Dr Henning Bay-Nielsen, Dr Birthe Frimodt-Moller, Développement d'outils d'aide à la décision, Rapport final octobre (2000).

[8] Georges El Helou et Charbel Abou khalil, Data,technique d'extraction de Connaissances. Lien de téléchargement format DOC.

www.lri.fr/~mdr/DataMining.doc

[9] Jef Wijsen Data Mining et Data warehousing, Université de Mons-Hainaut Février 2001.

[10] D.T. LAROSE, Discovering Knowledge In Data: An Introduction to Data Mining, Central Connecticut State University, 2005.

[11] R. Rahmani, Découverte d'association sémantique dans les bases de données relationnelles par les méthodes de Data Méninge, Mémoire En vue de l'obtention du diplôme de Magister en informatique, UMMTO (2009).

[12] Haddadou siad kamel et Haddar mohamed said, Extraction de règle d'association a partir des données ordinales imprécises, Mémoire D'ingénieur D'état En Informatique, UMMTO (2009-2010).

Références bibliographique

[13] R. Agrawa, T. Imielinski, A.Swami mining association rules between sets of items in large data base.in proc .SIGMOD 93,pp207-216, ACM Press (1993).

[14] Andreas Meier, Le CRM analytique Les outils d'analyse OLAP et le Data Méninge,

Dans le cadre du séminaire « Customer Relationship Management, Fribourg, le 26 avril 2008.

[15] MARHOUMI Fatima Ezzahra, Entrepôts de données: Développement d'un outil Extraction Transformation Load (ETL), Mémoire de fin d'études en vue de l'obtention du grade d'Ingénieur Civil Informaticien en Sciences Appliquées, Université Libre de Bruxelles (2006).

[16] Marlyse Dieungang – Khaoula Ghilani Data warehouse: Cubes OLAP.

[17] M. R. Allia, T. Bouadi, S. Almoutaoukil, k. Mamadou, Fouille de données : Règles séquentielles, rapport, Université Montpellier 2 (2010).

[18] ACIL Souhila et Makour Lamia. Extraction de règle d'association application d'une méthode basée sur les fermés fréquents. Mémoire D'ingénieur D'état En Informatique, UMMTO (2009-2010).

[19] Hassane Hilali, Application de la classification textuelle pour l'extraction des règles d'association maximales, comme exigence partielle de la maitrise en mathématique et informatique appliquées, Université du Québec à Trois - Rivières, Avril 2009.

[20] Elhoussaine Ziyati, Optimisation de requêtes OLAP en Entrepôts de Données Approche basée sur la fragmentation génétique, en vue de l'obtention du grade de DOCTORAT, Université Mohamed V –Agdal (2010).

[21] Marcos D'Urbano, Les règles d'associations séquentielles et leurs applications à des données d'achats et de ventes de fonds de placement, Mémoire présenté en vue de l'obtention du grade de Maîtrise (M.Sc.), Écoles des Hautes Études Commerciales (juin 2003)

Références bibliographique

[22] Benamar Houmadi, Etude Exploratoire D'outils pour le Data Mining, Exigence Partielle de la métrise en Mathématiques et Informatiques Appliquées, mémoire présenter a l'université du Québec a Trois-Rivières (Avril 2007).

[23] Arbatni Khaled, Réseaux de neurones appliqués à l'analyse et à la modélisation non linéaire du signal ECG, memoire de magister, Universite Mentouri Constantine (2007).

[24] fr.wikipedia.org/wiki/Datamart

[25] www.ryounes.net/cours/chapitre%203%20RN.pdf

[26] Mémoire la classification Tollari Sabrina 2003-06-10

[27] Classification de textes en genre et enthème : Votons utile ! Michel Plantié1, Mathieu Roche2, Gérard Dray1