REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Août 1955 - Skikda

Faculté des Sciences Département d'Informatique

Mémoire de fin d'études pour l'obtention du diplôme de

Master en informatique.

Option : Genie Logiciel Avancé et Applications (GLAA)

Thème

UTILISATION DES REGLES D'ASSOCIATIONS POSITIVES ET NEGATIVES POUR LA PREDICTION

(APPLICATION SUR LES MALADIES URINAIRES)

Réalisé par :

- Guehairia Noujoud
- Guerza Rokaya

Encadré par : A. MANSOUL

Année Universitaire 2021-2022

Remerciement

Avant tout, nous remercions ALLAH qui nous a accordé cette faveur. Toute notre gratitude vas à ceux dont les encouragements, les contributions et les idées nous furent si précieux durant l'élaboration de ce travail.

Parmi ceux qui nous ont patiemment aidé et soutenu dans cette épreuve nous citons notre encadrant Monsieur A.MANSOUL qui a supervisé la réalisation de ce travail dans les meilleures conditions. Nous tenons aussi à remercier les membres du jury d'avoir accepté de juger notre travail et de nous honore de leur présence.

Nous voulons adresser nos vifs remerciements à tous les enseignants du département informatique de l'Université du 20 aout 1955 de Skikda nous remercions également tous ceux qui nous ont aidé de prés ou de loin dans la réalisation de notre mémoire.

Dédicace

Je dédie cet ouvrage

A ma mère « Djamila » qui m'a soutenu et encouragé durant ces années d'études. Qu'elle trouve ici le témoignage de ma profonde reconnaissance.

A mon père « Ibrahime » qui a toujours été pour moi un exemple du père respectueux, honnête, de la personne méticuleuse.

A mon mari pour tout l'encouragement, le respect et l'amour que tu m'as offert, Je te dédis ce travail, qui n'aurait pas pu être achevé sans ton éternel soutien et optimisme. Tu es un modèle d'honnêteté, de loyauté et de force de caractère. J'espère te combler et te rendre toujours heureux.

A ma sœur et mon frère « Imen » et « Mahmoud ». Sans oublier ma belle sœur pour leur soutien et leur amour

A tous les membres de la famille « Guehairia » « Laib » et « Boumedine » : tantes, oncles, cousins, cousines qui m'ont soutenu dans mes efforts au cours de ces longues années d'études.

A mon binôme, ma copine et ma meilleure amie « Guerza Rokaya » d'être là et à mes côtés.

A mes meilleures amies Ines, Sara, Chems, Ilhem, Amina pour leur encouragements et leur soutien.

A tous ceux que j'aime.

Dédicace

Avec tout respect et amour je dédie ce mémoire :

A mes chers parents MAHFOUD et YASMINA qui méritent tout le bonheur du

monde et je leur dis « vous avez tout sacrifié pour votre fille n'épargnant ni santé

ni efforts.. Vous avez donné un magnifique modèle de labeur et de persévérance

merci je vous aime ».

A mes chers frères et sœurs : WAFA, ZEYNEB , MOHAMED ,

DJAMELEDINNE et ABDELMOUIZ ,sans oublier mes belle-sœur

HANIFA , MERIEME pour leur soutien et leur amour.

A tous les membres de la famille GUERZA et KHENCHOUL : tantes, oncles,

cousins, cousines qui m'ont soutenu dans mes efforts au cours de ces longues

années d'études.

A mes chers et meilleurs amis LINA ; HOUDA ; AMIRA ; mon binôme NOUNOU et surtout YASMINE SAYOUD pour leur encouragements et leur soutien !

Je remercie aussi SALAH et MOUMOUH pour les efforts

A tous ceux qui ont été à mes côtés.

Je vous aime....

Sommaire

Résumé

Le data mining, ou la fouille de données, constitue le cœur d'un processus d'extraction des connaissances à partir d'un large volume de données. Son champ d'applications est très vaste.

Dans le présent travail nous exposons un modèle de prédiction pour la recherche d'une donnée ou valeur et qui est en fait une valeur inconnue au départ et dont nous voulons prédire.

Pour atteindre cet objectif nous proposons un système qui va s'articuler autour de quatre modules dont les tâches sont les suivantes :

1. Dans un premier temps nous employons la technique L'extraction de connaissance à partir de données (ECD) pour la collection et la préparation des données

2. Dans un deuxième temps nous utilisons un module que nous avons développé afin de faire la prédiction à partir du modèle construit par les règles d'association (positive et négative).

3. Dans un troisième temps l'objectif est de proposer un nouvel algorithme d'extraction de règles d'association positives et négatives utilisant les optimisations que nous avons proposées dans le module précèdent.

4. Dans une étape finale nous expérimentons notre approche sur des données se rapportant aux maladies urinaires. En effet, dans ce genre de maladie nous nous concentrons surtout sur les symptômes et le développement pendant laquelle la maladie est à son plus haut degré de développement. Les formes aigues de l'inflammation de la vessie et de la néphrite soudaines et de grande intensité, sont des signes alarmants qui guident vers le diagnostic prématurément, ceci nous facilite un traitement d'attaque adéquat. Le patient peut ainsi être sauvé.

Le travail que nous présentons dans ce mémoire est très intéressant notamment dans la recherche de l'information manquante dans les maladies.

Ceci, permettra de contribuer au développement d'un système pour l'étude des maladies urinaires.

Mots clés : Algorithme apriori, Fouille de données, L'ECD, règles d'associations.

Introduction générale

Dans nos jours, on trouve que les hommes qui achètent des couches pour bébé quand ils font leurs courses aux grandes surfaces ou superettes les jeudis et samedi après-midi sont tendance à acheter du lait du café ? Avait entendu parler de cet exemple, c'est que quelqu'un a déjà essayé de vous expliquer à quoi sert la feuille de données ou plus globalement l'extraction de connaissance à partir des données., la chaîne de grande distribution Ardis aurait découvert une corrélation entre l'achat de couches culottes par des hommes et l'achète de café les jeudis et samedi après-midi Ardis aurait donc réorganiser ses rayons en positionnant les packs les packs du café à côté des couches culottes.

Bien que base sur un fait réel, l'histoire a été enjolivée pour mettre en avant que l'analyse des données pourrait augmenter les profits d'analyses donner pouvait augmenter les profits en tirant parti des outils informatiques d'analyses des données. Même enjolive, c'est exemple et lustre parfaitement l'importance de extractions de connaissance à partir de données pour les entreprises et comment elle peut leur permettre d'augmenter leurs ventes et d'obtenir un avantage sur leurs concurrents.

Les travaux de recherche effectué dans ce document s'inscrivent dans le domaine de l'extraction de parler connaissances extrait tu peux prendre différentes formes mais nous allons nous intéresser, Mango connaissance Express ou la forme de règles d'association. Extraction de règles d'association consiste à découvrir relations les variables d'une base de données.

Dans ce mémoire, nous nous intéresserons à l'extraction de règles d'association et plus particulièrement à l'extraction de règles d'association négatives et positives. Permettre Dexter des règles dans lequel la présence ainsi que l'absence de ne variable peuvent être utilisé point-là port de faire règle négative n'ai pas négligeable puisqu'elles peuvent non seulement contenir des informations non présentes dans les règles positives mais permettent également d'élargir la sémantique des connaissances. Ainsi, en médecine, cela peut permettre de trouver les caractéristiques qui empêchent une maladie de se déclarer, en plus de chercher les caractéristiques déclenchant une maladie. En combinant l'extraction des règles positives avec celles des règles négatives, sans le chant de connaissance et par conséquent le champ des possibilités. Cependant cet ajout des règles négatives va être un défi en raison essentiellement des coups de calcul qui vont augmenter exceptionnellement, mais également en raison du temps du nombre prohibitif pour la plupart redondante et intéressante point ces problèmes provient proviennent du fait de l'absence de variable et en général plus important que la présence de ces mêmes variables. Par exemple dans la base de données de la grande distribution, chaque consommateur n'achète qu'un sous-ensemble des milliers d'articles recenses dans le magasin Point l'objectif de synthèse et donc d'extraire de manière optimale l'ensemble des règles d'association positives et négatives intéressantes.

Dans le but de répondre à cette problématique, nous avons structure ce manuscrit en quatre chapitres.

D'abord on commence par le premier chapitre, sur les concepts et les notions de feuilles de données et les différentes phases de processus d'extraction de connaissance à partir de données (ECD). En trouver une définition abstraite de la fouille de données, le processus de l'ECD, le data mining……

Dans ce chapitre on a abordé le la recherche de règles d'association, on a basé sur des concepts et des définitions des item, itemset, itemsetFrequent). Et aussi les règles d'association : des exemples, leur

but principal, le processus de recherche des règles d'association et quelques avantages et inconvénients des règles d'association. A la fin, l'algorithme Apriori.

Le troisièmes chapitre détaille les trois algorithmes majeurs d'extraction de règles d'association positives et négatives évoques dans le premier chapitre. Ces trois algorithmes, possédant chacun ses avantages et ses inconvénients, ont la particularité d'être des dérives d'Apriori et utilisent donc les mesures du support et de la confiance. Cependant chaque algorithme définit de manière différente ce qu'il considère comme une règle valide. Par conséquent les règles extraites seront différentes d'un algorithme `à l'autre comme nous pourrons le voir sur un exemple fil-rouge. Leude approfondie de ces trois méthodes met en avant essentiellement deux failles, à savoir un nombre encore important de règles inintéressantes et un parcours non optimise de recherche des règles.
Dans le dernier chapitre, nous avons mis le point sur l'expérimentation des résultats. Nous avons touché le nécessaire partie qui sont : les outils utilisés pour la construction du système (Weka, React), l'architecture du système, le choix de donner et l'application qui traite des données et on a choisi la maladie urinaire comme titre d'exemple

Chapitre 1 : L'ECD

1. Introduction

Le premier chapitre de notre travail aide à présenter les nécessaires concepts ou les notions de fouilles de données, ou les différentes phases du processus d'extraction de connaissances à partir des données sont décrites. Nous appuyons sur les différentes démarches et les différentes approches de mise en œuvre d'un modèle de fouille de données.

2. L'extraction de connaissance à partir de données (ECD)

L'approche moderne de l'extraction des connaissances à partir des données se veut la plus générale possible. Elle ne privilégie ni une source particulière d'informations (celles-ci peuvent être localement stockées ou distribuées) ni une nature spécifique des données (elles peuvent être structurées en attributs-valeurs, des textes de longueurs variables, des images ou des séquences vidéo). Elle ne se limite pas aux outils d'analyse les plus récents et incorpore explicitement des méthodes pour la préparation des données, pour l'analyse et pour la validation des connaissances produites. Ces méthodes proviennent en majorité de la statistique, de l'analyse des données, de l'apprentissage automatique et de reconnaissance de formes.

L'ECD est un processus anthropocentré, les connaissances extraites doivent être les plus intelligibles possibles à l'usager. Elles doivent être validées, mises en forme et agencées. Nous allons détailler toutes ces notions et les situer dans le processus général de l'ECD. L'introduction de l'ECD dans les entreprises est récente. Le rattachement des activités liées à l'ECD n'est pas toujours clair. Selon les cas, elle peut être intégrée au service ou la direction : informatique, organisation, études, statistique, marketing, etc. Comme le montre le schéma de la figure 1, il convient de distinguer le niveau opérationnel et le niveau «analyse» que nous allons décrire

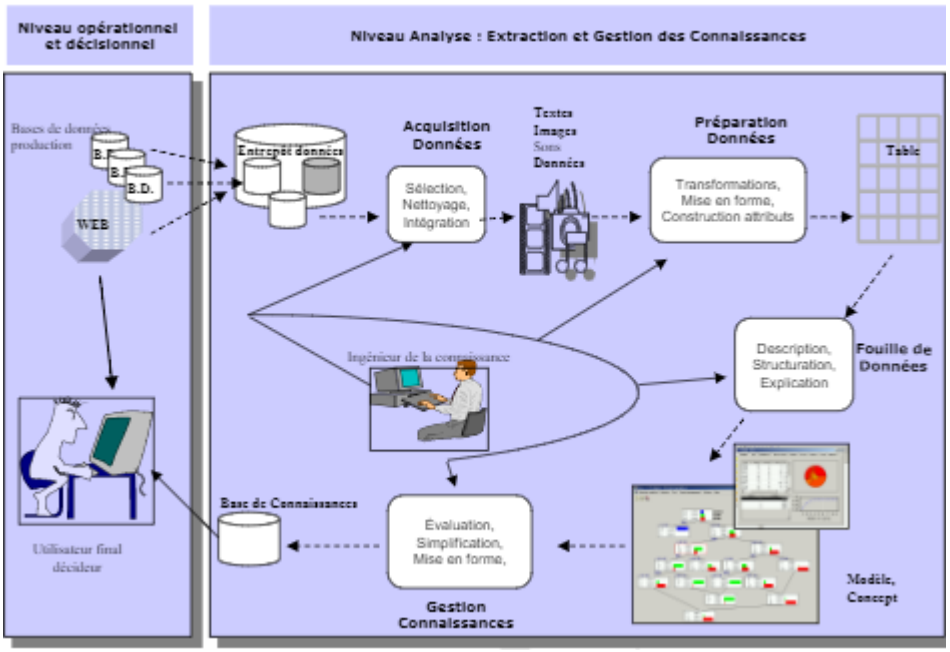


Figure 1 : L'extraction de connaissance à partir de données.[1]

3. La fouille de données

La fouille de données concerne le data mining dans son sens restreint est au cœur du processus d'ECD. Cette phase fait appel à de multiples méthodes issues de la statistique, de l'apprentissage automatique, de la reconnaissance de formes ou de la visualisation. Les méthodes de data mining permettent de découvrir ce que contiennent les données comme informations ou modèles utiles. Si nous essayons de classifier les méthodes de fouille de données utilisées, trois catégories se distinguent :

- Les méthodes de visualisation et de description.
- Les méthodes de classification et de structuration.
- Les méthodes d'explication et de prédiction.

Chacune de ces familles de méthodes comporte plusieurs techniques appropriées aux différents types de tableaux de données. Certaines sont mieux adaptées à des données numériques continues alors que d'autres sont plus généralement dédiées aux traitements de tableaux de données qualitatives. Nous allons donner à présent un aperçu général sur les principales méthodes

4. Le Processus de L'ECD

Comme elle montre la figure 2, l'Extraction de Connaissances à partir de Données est un processus constitué de plusieurs étapes [12] [13] [14] .

Elles sont répétées dans des itérations multiples (des feedbacks et des boucles récursives peuvent être observés durant le processus) et à chaque itération ou étape du ce processus une intégration des connaissances des expertes de domaine est nécessaire (le processus est en perpétuelle interaction avec les utilisateurs) pour découvrir de nouvelles connaissances cachées interprétables et utilisables. De ce fait, le processus de l'ECD est souvent qualifié d'itératif et d'interactif. Les étapes de ce processus consistent principalement en collection des données contenant dans les différentes sources opérationnelles de l'entreprise, la préparation des données nécessaires pour accomplir la ou les tâches de Data Mining souhaitées, l'application des méthodes de Data Mining nécessaires pour résoudre ces tâches et enfin l'évaluation et la validation des résultats obtenus.

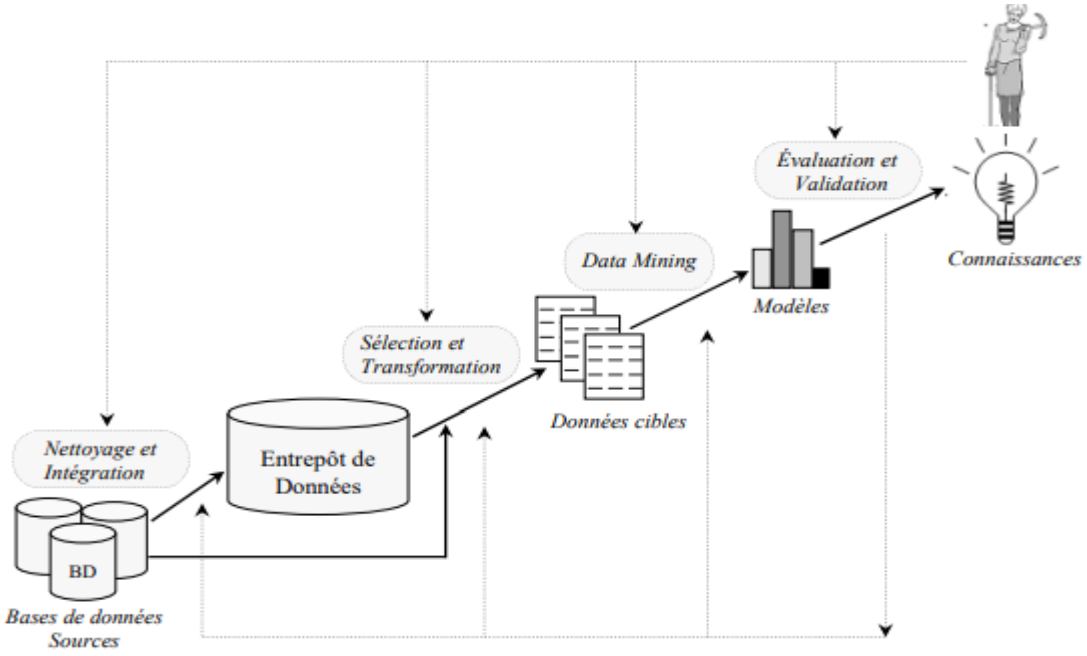


Figure 2 : Le Processus de L'ECD. [6]

4.1 La collection de données

Cette étape consiste généralement à collecter les données contenant dans les différentes sources opérationnelles de l'entreprise, éventuellement réparties et hétérogènes, après nettoyage, prétraitement, transformation et consolidation de ces données, dans une base de données cohérente ou le plus souvent dans un entrepôt de données. La construction d'un entrepôt de données, qui fait l'objet du chapitre 2, est pratiquement réalisée au niveau de cette étape. Dans cette section, nous expliquons seulement les tâches de nettoyage et d'intégration de données, effectuées lors de construction de l'entrepôt de données, puisqu'elles ont un impact important sur les étapes ultérieures du processus de l'ECD et ensuite sur les modèles résultants.

4.1.1 Nettoyage de données

Puisque l'entrepôt de données est construit pour fin d'extraire des connaissances utiles pour la prise de décisions stratégiques, il est important que les données collectées dans l'entrepôt soient correctes. Cependant, puisque les grands volumes de données des différentes sources opérationnelles de l'entreprise sont impliqués, il y a une haute probabilité d'erreurs et d'anomalies dans les données. Donc, la routine de nettoyage de donnes [15] [4] consiste à éliminer, ou à réduire le maximum possible, les données erronées, aberrantes, inconsistantes, bruyantes et les valeurs manquantes dans l'ensemble de données collectées avant de les chargées dans l'entrepôt de données

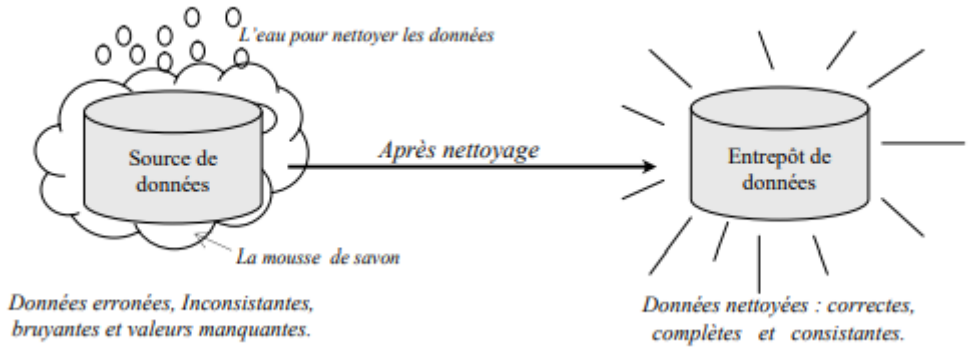


Figure 3 : Nettoyage de données. [2]

Il existe plusieurs méthodes de nettoyage de données. En effet, [4] présentent d'utiliser des techniques comme la régression linéaire et le Clustering pour le traitement des données bruyantes. Quant au traitement des données manquantes, [1] présentent un ensemble d'opérations à savoir :

- Ignorer les enregistrements contenant des valeurs manquantes ;
- Remplir les valeurs manquantes manuellement ;
- Employez une constante globale pour remplir les valeurs manquantes ;
- Utiliser une valeur moyenne pour remplir les valeurs manquantes ;
- Utiliser la valeur la plus probable pour remplir les valeurs manquantes (exemple de techniques : la formule de Bayes ou l'arbre de décision) ;
- Chercher à estimer ces valeurs manquantes par des méthodes d'induction comme la régression, les réseaux de neurones ou les graphes d'induction.

4.1.2 Intégration de données

L'intégration de données [16] implique la combinaison de multiples sources de données, éventuellement réparties et hétérogènes, dans une base de données cohérente qui sera adaptée par la suite pour l'extraction de connaissances. Comme elle montre la figure 1. 3, les sources de données peuvent inclure plusieurs bases de données relationnelles, les bases de données de production (transactionnelles), les bases de données multidimensionnelles, les fichiers plats ou même d'autres sources externes à l'entreprise. La combinaison de telles sources de données construite un entrepôt de masse volumes de données et plus important de données homogènes, résumées, consolidées qui facilite la tâche d'analyse et d'exploration de données. En effet, le but de construction d'un entrepôt de données est la création des vues consolidées pour aider le processus d'extraction de connaissances et ensuite celui de la prise de décision.

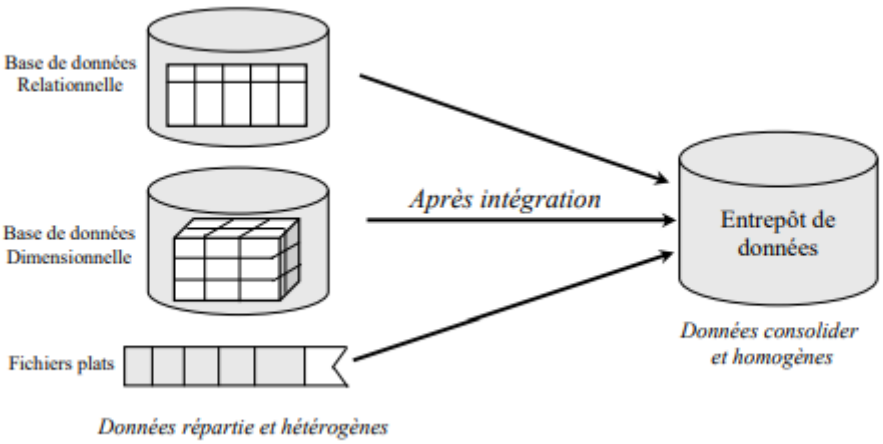


Figure 4 : Intégration de données. [4]

Il existe plusieurs problèmes à considérer durant la phase d'intégration de données. En effet, l'intégration des schémas des sources fait apparaître des conflits, depuis longtemps bien répertoriés dans la littérature. Les principaux conflits pouvant survenir entre deux schémas sont les suivants :

- Les problèmes de terminologie : Un conflit de terminologie survient lorsqu'un même objet du réel est désigné par des noms différents ou au contraire lorsqu'un même nom est utilisé pour deux objets différents. Ces cas peuvent correspondre à des problèmes de synonymie ou d'homonymie, mais sont le plus souvent dus à une différence de niveau de généricité (par exemple : « Personne » et « Étudiant »), ou à des converses (par exemple : « Vente » et « Achat »). En outre, les conflits de terminologie sont souvent à l'origine des problèmes de la redondance des données ; elle constitue un autre problème important à résoudre lors de l'intégration des données et sa détection fait souvent référence à la méthode d'analyse des corrélations [4].

- Les incompatibilités de contraintes : Un conflit de contraintes apparaît lorsque sur deux concepts établis comme équivalents ont des contraintes incompatibles (par exemple : « Un âge supérieur à 18 » et « Un âge inférieur à 17 »).

- Les conflits de structures : Les conflits de structure sont caractérisés par un choix différent de propriétés à stocker pour un même concept du réel. Par exemple, on peut définir une personne dans une vue par son numéro, son nom et son âge, et dans une autre vue par son nom, son prénom et son adresse.

- Les conflits de représentation : On détecte un conflit de représentation lorsque deux représentations différentes sont choisies pour les mêmes propriétés d'un même concept. Par exemple la date de commande peut être incluse dans la commande ou former un objet relié à la commande.

4.2 La Préparation de données

La qualité des résultats d'un processus de l'ECD dépend en grande partie de la qualité des données utilisées, d'où l'importance de l'étape de préparation de ces données [14]. D'après [15], le

prétraitement de données consiste en toute action effectuée sur les données avant l'application d'une technique de Data Mining. La collection des données dans un entrepôt de données facilite considérablement l'étape de préparation de données, puisque les données sont déjà nettoyées, transformées, fusionnées, agrégées, harmonisées et intégrées en vue de les préparées pour un processus d'aide à la décision. Il est important de distinguer entre la préparation de données pour le Data Mining et la préparation de données pour la conception d'un entrepôt de données. En effet, l'entrepôt de données permet de regrouper les données qui seront éventuellement accessibles aux méthodes de Data Mining, mais cet entrepôt ne supporte pas que le Data Mining. Il peut aussi supporter d'autres utilisations, comme l'analyse statistique et des applications de production qui nécessitent des données analytiques (moyennes, médianes, etc.). D'ailleurs, la préparation des données pour un entrepôt de données doit permettre d'optimiser les utilisations pour améliorer les temps de calculs lors de l'analyse. Par contre, la préparation des données pour le Data Mining doit être faite en fonction d'un ou plusieurs buts (ou tâches) définis à l'avance. C'est pourquoi on peut dire que la préparation des données pour le Data Mining est très différente de la préparation des données pour un entrepôt de données [17]. Malgré ces différences, les deux se complémentent très bien, et un entrepôt de données constitue une base solide pour supporter les techniques de Data Mining.

Les principaux traitements, effectués lors de l'étape de préparation de données, peuvent être résumés en sélection des données pertinentes pour accomplir la ou les tâches de Data Mining requises et ensuite, la transformation des données sélectionnées en une forme plus approprie pour réaliser ces tâches.

4.2.1 Sélection de données

L'objectif de cette étape est de sélectionner et d'analyser l'état des données requises pour exécuter la ou les tâches de data Mining souhaitées. Il s'agit d'identifier les attributs (colonnes) et/ ou les enregistrements (lignes) à utiliser pour le Data Mining. Les critères de sélection incluent la pertinence de ces données, la qualité et les contraintes techniques, comme les limites de volume de données ou les types de données à utiliser. Les raisons éventuelles de la sélection de données peuvent être : la réduction du temps de construction du modèle qui décroît avec le nombre d'attributs, l'optimisation de la qualité des modèles obtenus (en suppriment les attributs fortement corrélés) ou les contraintes d'outils utilisés (nombre d'enregistrements supportés). Les étapes ultérieures du processus de l'ECD s'appliquent exclusivement sur l'ensemble de données sélectionné dans cette étape, d'où son importance.

4.2.2 Transformation de données

Cette phase consiste à transformer les données en une forme appropriée pour accomplir la ou les tâches de Data Mining requises. Précisément, les transformations incluent la normalisation des valeurs des champs des enregistrements et ainsi que la réduction de nombre de champs des enregistrements des tables de la base de données, puisque chaque tâche de Data Mining se concentre seulement sur un sous ensemble de champs. Également, certaines d'autres modifications et combinaisons des champs des enregistrements peuvent être faites afin d'arranger les données

originales dans un espace de données plus approprié aux tâches de Data Mining, qui seront exécutées à l'étape suivante (Data Mining).

La technique de normalisation de données consiste à appliquer des opérations de normalisation en vue de représenter les données sous une forme de mesures de la même plage de données. Par exemple, les données « -2, 23, 100, 59, 48 » peuvent être normalisées en « -0.02, 0.23, 1. 00, 0.59, 0.48 ». La technique de réduction de données [4], ou « Data Reduction », peut être appliquée sur un ensemble de données pour obtenir une représentation réduite de celle-ci (voir la figure 1. 4). L'objectif de la réduction de données est d'avoir un espace de travail mois volumineux, au lieu de travail sur l'intégrité des données originale de très large volume. Puisque l'extraction de connaissances à partir d'un ensemble de données réduite doit être plus efficace et produit le même (ou presque le même) résultat analytique, il est important de travail sur des ensembles de données réduits. Parmi les stratégies de réduction de données, nous citons l'agrégation par les cubes de données, la compression de données, la discrétisation et la génération de hiérarchie [4].

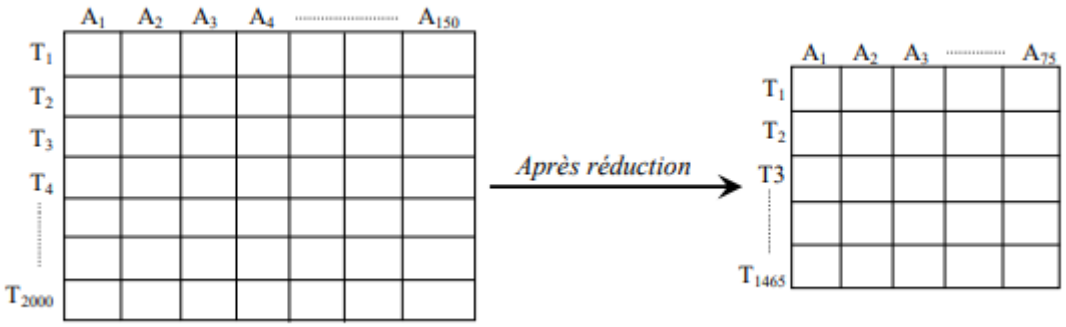


Figure 5 : Transformation de données. [7]

4.3 Le Data Mining

L'étape de Data Mining constitue le cœur du processus d'Extraction de Connaissances à partir de Données. Elle s'agit d'appliquer des méthodes intelligentes, spécifiques au Data Mining, afin d'extraire à partir de données, dite données d'apprentissage, des modèles, des règles ou toutes autres formes compréhensibles et interprétables en connaissances utiles (voir la figure 1. 5). Parmi les méthodes les plus connues du Data Mining, nous citons les méthodes de classification, les méthodes de Clustering et les méthodes de recherche de règles d'association. Dans le chapitre 3, un état de l'art sur ces différentes méthodes est réalisé. En générale, durant cette étape, elle est souhaitable de mettre en œuvre différentes techniques ou méthodes de Data Mining afin de les comparer et d'en retenir une ou plusieurs combinées [7].

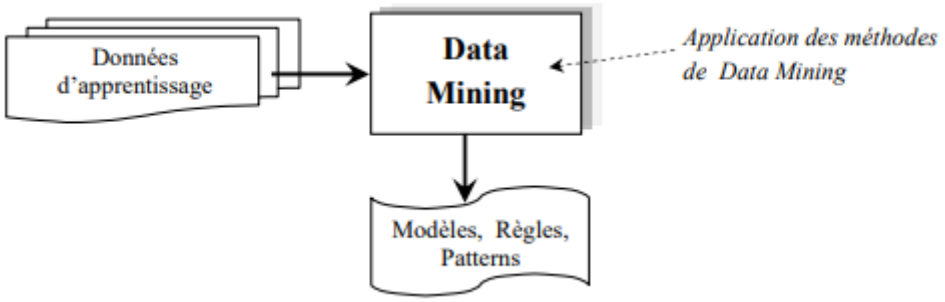


Figure 6 : Data Mining. [6]

4.4 L'évaluation et validation

C'est l'étape finale du processus d'Extraction de Connaissances à partir de Données. Elle consiste à évaluer les résultats obtenus pour déterminer quels modèles peuvent être considérés comme des connaissances nouvelles et intéressantes. Cette étape comporte aussi une interprétation des résultats et une comparaison des modèles. En effet, la pertinence de la connaissance découverte est estimée par des critères de certitude imposés par les utilisateurs ou les experts de domaine [6]. Ensuite, les modèles énumérés comme des connaissances pertinentes seront validés sur d'autres ensembles de données ou sur d'autres systèmes.

Les méthodes de validation vont dépendre de la nature de la tâche et du problème considéré. Par exemple, pour la segmentation et l'association, la validation est essentiellement du ressort de l'expert qui jugera de la pertinence des segments constitués (segmentation) ou de la pertinence des règles (association). Aussi, pour faire de la classification, on décompose les données en trois ensembles disjoints : un ensemble d'apprentissage, un ensemble de test et un ensemble de validation. Au mois deux ensembles sont nécessaires : l'ensemble d'apprentissage permet de générer le modèle et l'ensemble de test permet d'évaluer l'erreur réelle du modèle sur un ensemble indépendant. Ainsi, lorsqu'il s'agit de tester plusieurs modèles et de les comparer, on peut sélectionner le meilleur modèle selon ses performances sur l'ensemble de test et ensuite évaluer son erreur réelle sur l'ensemble de validation [18]. Ce processus de validation est illustré sur la figure ci-dessous.

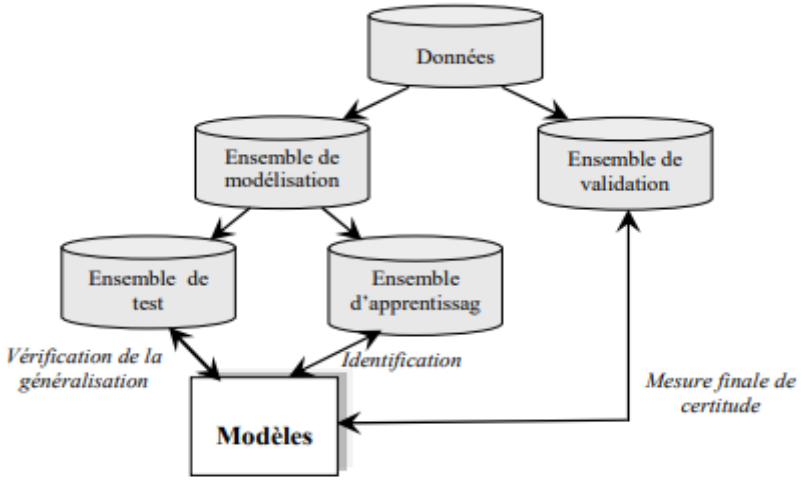


Figure 7 : L'évaluation et validation [6]

5. Les tâches de la fouille de données

Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des six tâches suivantes :

- La classification.
- L'estimation.
- Le groupement par similitude (règles d'association).
- L'analyse des clusters.
- La description.

Les trois premières tâches sont des exemples de la fouille supervisée de données dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Le groupement par similitude et l'analyse des clusters sont des tâches non-supervisées où le but est d'établir un certain rapport entre toutes La description appartient à ces deux catégories de tâche, elle est vue comme une tâche supervisée et non-supervisée en même temps.

- **Classification**

  La classification est la tâche la plus commune de la fouille de données qui semble être une tâche humaine primordiale. Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer. La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie. Les objets à classifier sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jours chaque enregistrement en déterminant la valeur d'un champ de classe. Le fonctionnement de la classification se décompose en deux phases. La première étant la phase d'apprentissage. Dans cette phase, les approches de classification utilisent un jeu d'apprentissage dans lequel tous les objets sont déjà associés aux classes de références connues. L'algorithme de classification apprend du jeu d'apprentissage et construit un modèle. La seconde phase est la phase de classification proprement dite, dans laquelle le modèle appris est employé pour classifier de nouveaux objets.

- **L'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. Par exemple on cherche à estimer la lecture de tension systolique d'un patient dans un hôpital, en se basant sur l'âge du patient, son genre, son indice de masse corporelle et le niveau de sodium dans son sang. La relation entre la tension systolique et les autres données vont fournir un modèle d'estimation. Et par la suite nous pouvons appliquer ce modèle dans d'autres cas.

- **Le groupement par similitude**

  (Analyse des associations et de motifs séquentiels) Le groupement par similitude consiste à déterminer quels attributs "vont ensemble". La tâche la plus répandue dans le monde du business, est celle appelée l'analyse d'affinité ou l'analyse du panier du marché,

elle permet de rechercher des associations pour mesurer la relation entre deux ou plusieurs attributs. Les règles d'associations sont, généralement, de la forme "Si, alors ".

- **L'analyse des clusters**

  Le clustering (ou la segmentation) est le regroupement d'enregistrements ou des observations en classes d'objets similaires. Un cluster est une collection d'enregistrements similaires l'un à l'autre, et différents de ceux existants dans les autres clusters. La différence entre le clustering et la classification est que dans le clustering il n'y a pas de variables sortantes. La tâche de clustering ne classifie pas, n'estime pas, ne prévoit pas la valeur d'une variable sortante. Au lieu de cela, les algorithmes de clustering visent à segmenter la totalité de données en sous-groupes relativement homogènes. Ils maximisent l'homogénéité à l'intérieur de chaque groupe et la minimisent entre les différents groupes

- **La description**

  Parfois le but de la fouille est simplement de décrire ce qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour premier lieu comprendre le mieux possible les individus, les produit et les processus présents dans cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Dans la société Algériennes nous pouvons prendre comme exemple comment une simple description, "les femmes supportent le changement plus que les hommes", peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologues, économistes et les spécialistes en politiques.

6. Conclusion

Nous avons vu dans ce chapitre que des données brutes présentes dans des entrepôts ou des bases de données, peuvent être exploités de manière à en extraire des connaissances ; ces dernières sont obtenues en passant par plusieurs phases, à savoir l'acquisition des données, leurs préparations, leurs traitements pour en extraire des règles, et enfin la validation des connaissances extraites ; toutes ces phases regroupées représentent le processus d'extraction de connaissances « ECD ».

Dans notre travail, nous nous sommes intéressés particulièrement aux deux dernières phases de l'ECD : fouille et validation des données.

II. Chapitre 2 : La recherche de règles d'associations

Chapitre 2 : La recherche de règles d'associations

1. Introduction

Traiter un grand volume de données sauvegardé sur un support informatique ne nous permet pas d'avoir une vue complète sur les informations contenues dans ces données. Souvent, des informations restent cacher, invisibles pour l'utilisateur. Ces informations cachées sont la plus part de temps importantes aux yeux de l'utilisateur. Pour les exploiter, on utilise les méthodes de Data Mining

La méthode de Data Mining la plus intéressante dans ce contexte est l'extraction de règles d'association. Cette méthode met en valeur les relations cachées qui existe dans une grande collection de données. L'exemple de règle d'association le plus connu : '' Dans super marché, un homme qui achète des couches pour bébés, achète aussi 2 packs de bières dans 65% des cas''. Dans cet exemple, cette règle pourra inciter le gérant du super marché à faire des réductions sur des achats des couches pour bébés et des packs de bières. Ce type de connaissances est généralement utile pour la prise de décision (exemple : Choisir les produits à mettre en promotion, organiser l'emplacement des produits, ...etc.).

L'extraction de règles d'association de Data mining sont aussi utilisées dans la fouille de texte pour analyser un texte ou une collection de textes afin extraire des connaissances dissimulées qui existe entre les termes. C'est cette technique de fouille données sera étudier durant toute la suite de notre mémoire.

2. Concepts et définitions (Item, Etemset, ItemsetFrequent)

- **Item :** Un item est tout article, attribut, littéral appartenant à un ensemble fini d'éléments distincts X = {x_1, x_2, ..., x_n}.
  Par exemple, dans les applications de type analyse du panier de la ménagère, les articles en vente dans un magasin sont des items. L'ensemble X peut contenir les items A, B, C et D correspondant aux articles lait, beurre, pain et confiture par exemple. [3]
- **Etemset :** Un itemset est un ensemble d'items d'objets ou d'articles d'une base de données
  Soit R, un ensemble d'items. 1 est un itemset si 1 ç R. Soit i E~, on note I, ['itemset de la Transaction à la position i dans une BDB. On note k-itemset, un itemset ayant une taille de k items

- **ItemsetFrequent** Supposons R un ensemble d'items, r une BOB sur R et Y ç R un itemset. On note par (J un seuil de support minimum défini par un utilisateur. On définit un ensemble d'itemsets (Jfréquents comme suit : Freq(r, (J) = {Y 1 y ç R /\ support(r, Y) (J} [1]

3. Utilité des règles d'associations

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. Les règles d'associations séquentielles sont également utiles dans l'étude des comportements d'achat des consommateurs, car elles permettent de mettre en lumière les comportements d'achats à travers le temps. Par exemple, si les associations séquentielles indiquent que les propriétaires d'une marque et d'un modèle précis de voiture changent leur véhicule aux 18 mois, le concessionnaire pourra fidéliser sa clientèle en faisant parvenir un dépliant sur les nouveaux modèles disponibles quelques semaines avant la date du changement de voiture. [2]

4. Les règles d'associations

Les règles d'association sont une des méthodes de Data Mining les plus répandus dans le domaine du marketing et de la distribution. Les règles d'association générées sont de la forme "Si action1 ou condition alors action2". Elles peuvent se situer dans le temps : "Si action1 ou condition à l'instant t1 alors action2 à l'instant t2" c'est les règles d'association séquentielles.

Leur principale application est « l'analyse du panier de la ménagère », qui consiste, comme l'indique son nom, en la recherche d'associations entre produits sur les tickets de caisse et l'étude de ce que les clients achètent. La méthode recherche quels produits tendent à être achetés ensemble. Elles peuvent être appliquées à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services.

Voici quelques exemples de règles :

- Si un client achète du lait alors il achète du pain (90%)
- Si un client achète une télévision, il achètera un récepteur satellite dans un mois (50%)
- Si maladie X et traitement Y alors guérison (95%)
- Si maladie X et traitement Y alors guérison dans Z années (97%)
- Si présence et travail alors réussite à l'examen (99%)

Ces règles sont intuitivement faciles à interpréter car elles montrent comment des produits ou des services se situent les uns par rapport aux autres. Elles sont particulièrement utiles en marketing et peuvent être facilement utilisées dans le système d'information de l'entreprise.

Le but principal de cette technique est donc descriptif. Dans la mesure où les résultats peuvent être situés dans le temps, cette technique peut être considérée comme prédictive. Cependant, il faut noter que cette méthode, si elle peut produire des règles intéressantes, peut aussi produire des règles triviales ou inutiles (provenant de particularités de l'ensemble d'apprentissage). La recherche de règles d'association est une méthode non supervisée car on ne dispose en entrée que de la description des achats.

5. Le processus de recherche de règles d'association

Le processus d'extraction de règles d'association est constitué de plusieurs phases allant de la sélection et la préparation des données jusqu'à l'interprétation des résultats, en passant par la phase

de recherche des connaissances (extraction des ensembles fréquents d'attributs et génération des règles d'association). Ci-dessous une description de différentes phases de ce processus. [3]

5.1.Sélection et préparation des données (nettoyage)

Cette phase consiste à sélectionner les données (attributs et objets) de la base de données utiles à l'extraction des règles d'association et transformer ces données en un contexte d'extraction. L'extraction de règles d'association peut être effectuée à partir des bases de données de divers types, comme des données spatiales, temporelles, orientées objets, multimédia, etc. Cette première phase est très importante car à partir de la qualité des données en entrées dépend la qualité des résultats. [3]

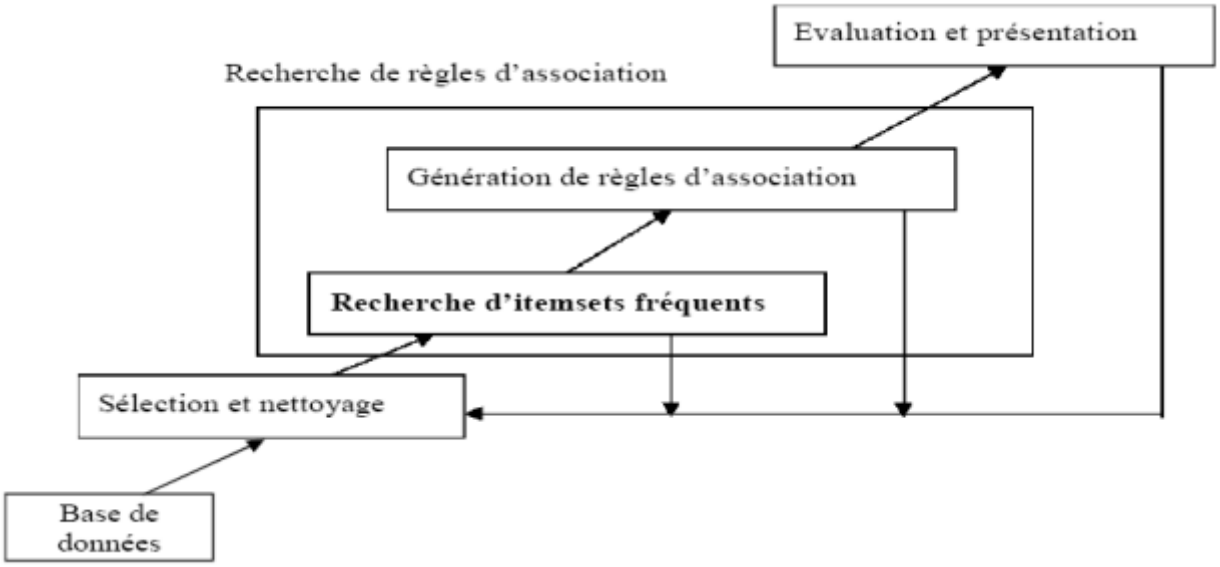


Figure 8 : Processus d'ECD adapté à la recherche de règles d'association [3]

Cette phase est nécessaire pour pouvoir appliquer les algorithmes d'extraction des règles sur des données de natures différentes provenant de sources différentes, de concentrer la recherche sur les données utiles pour l'application et de minimiser le temps d'extraction. A noter que le problème des données incomplètes (valeurs manquantes), et les données erronées ou incertaines et la taille du jeu de données doivent être pris en considération dans cette phase.

Par exemple, le tableau ci-dessous suivant représente un contexte d'extraction D constitué de 6 objets, chacun représenté par son identifiant et de quatre items. Ce contexte sera utilisé comme exemple dans tout le reste de ce chapitre. [3]

ITEM ID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E
6	B C E

Tableau 1 : Contexte d'extraction de règles d'association D.

II. Chapitre 2 : La recherche de règles d'associations

5.2.Recherche d'itemsets fréquents

Cette phase consiste à extraire du contexte D tous les itemsets qui sont fréquents. La recherche des itemsets fréquents est un problème non trivial car le nombre d'itemsets fréquents potentiels est exponentiel en fonction du nombre d'items du contexte D.

Dans le cas d'un ensemble d'items I de taille m, le nombre d'itemsets potentiels est de 2m-1.

Ces items forment le treillis des itemsets de I, dont la hauteur est de m+1. Les balayages du contexte doivent être réalisés lors de cette phase et il est donc nécessaire de développer des méthodes efficaces d'exploration de cet espace de recherche exponentiel.

La phase découverte des items fréquents constitue la phase la plus coûteuse en temps d'exécution et en espace. L'espace de recherche est de taille exponentielle par rapport au nombre d'items. Plusieurs méthodes ont été proposées dans le but de réduire l'espace de recherche de cette phase ainsi que le nombre de balayages du contexte réalisé. [3]

Voici un exemple d'un treillis des itemsets du contexte D donné dans le tableau 1 précédent

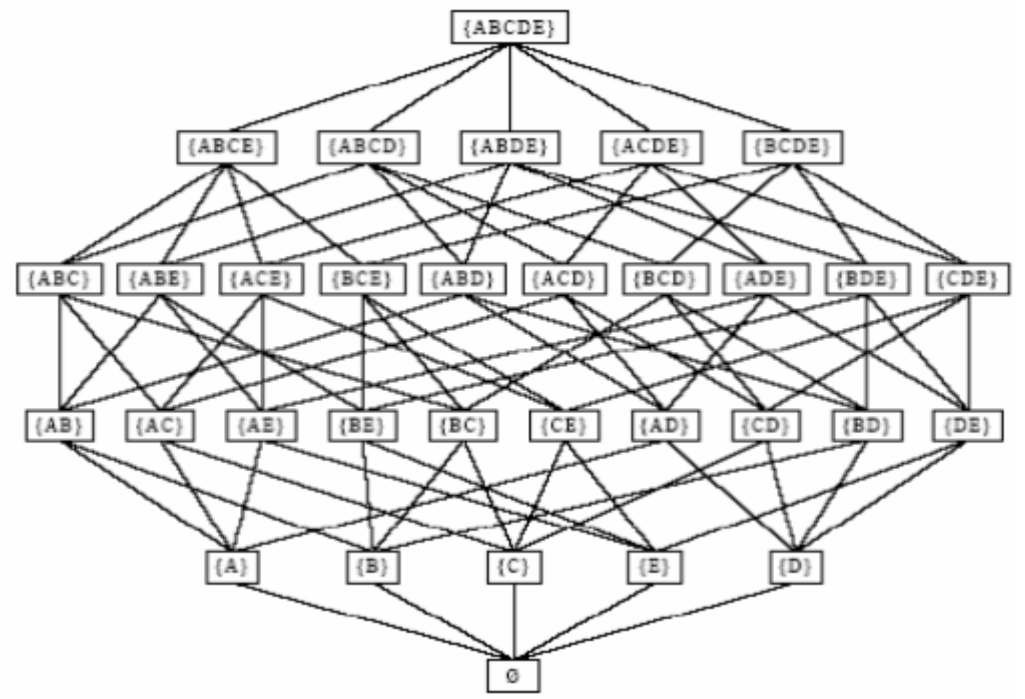


Figure 9 : Représentation sous forme de treillis d'itemsets fréquents du contexte D [3]

5.3.Génération des règles d'association

La génération des règles d'association s'effectue à partir des itemsets fréquents générés précédemment.

II. Chapitre 2 : La recherche de règles d'associations

En général, la génération des règles d'association est réalisée de manière directe, sans accéder au contexte d'extraction, et le coût de cette phase en temps d'exécution est donc faible par rapport au coût de l'extraction des itemsets fréquents.

La génération de règles d'association est assez simple suivant le principe donné dans la définition donnée pour l'ensemble de règles d'association générées.

Pour chaque itemset fréquent l1 dans F, tous les sous-ensembles l2 de l1 sont déterminés et la valeur de la confiance(r) est calculée. Si cette valeur est supérieure ou égale au seuil minimal de confiance alors la règle d'association l2 ◊ (l1 - l2) est générée. [3]

5.4.Visualisation et interprétation

C'est la phase finale du processus d'ECD. Cette phase consiste en la visualisation par l'utilisateur des règles d'association extraites du contexte et leur interprétation afin d'en déduire des connaissances utiles pour l'amélioration de l'activité concernée. Ainsi l'expert du domaine peut juger de leurs pertinences et utilités. Mais le nombre important des règles d'association extraites impose le développement d'outils de classification de règles selon leurs propriétés, de sélection de sous-ensembles de règles selon des critères définis par l'utilisateur, et de visualisation de ces règles sous une forme intelligible. Cette nouvelle problématique est également appelée « Knowledge Mining ».

La forme de présentation de règles peut être textuelle, graphique ou bien une combinaison de ces deux formes intelligibles. Ceci va donner naissance à un nouveau domaine de recherche : la fouille visuelle de données « Visual Data Mining » afin d'améliorer le processus d'extraction de connaissances en proposant des outils de visualisation adaptés à différentes problématiques.

Les connaissances de l'utilisateur concernant le domaine d'application sont nécessaires lors des phases de pré-traitement afin d'assister la sélection et la préparation des données et de post-traitement, pour l'interprétation et l'évaluation des règles extraites. En fonction de l'évaluation des règles extraites, les paramètres utilisés lors des précédentes phases (critères de sélection et préparation des données et seuils minimaux de support et de confiance) peuvent être modifiés avant d'effectuer à nouveau l'extraction des règles d'association, ceci afin d'améliorer la qualité du résultat.

Il ressort de la grande majorité de ces applications qu'au final, beaucoup de règles sont générées par les algorithmes et qu'il est parfois difficile aux experts du domaine de les exploiter dans leur intégralité, car cela engendre un travail cognitif très important. Devant cette tâche, leur premier souhait est souvent de réduire cet ensemble pour ainsi diminuer le temps d'expertise correspondant. En effet, dans le domaine industriel, les experts n'ont pas forcément beaucoup de temps à consacrer à l'analyse des résultats. [3]

II. Chapitre 2 : La recherche de règles d'associations

6. Avantages / inconvénients des règles d'associations

L'avantage principal des règles d'association réside dans sa clarté et sa simplicité d'implantation. Cependant cette technique souffre d'inconvénient majeur qui la rend n'appropriée dans le cas de bases de données volumineuses (VLDB

7. Evaluation et validation des règles d'association

Résume. La recherche de règles d'association intéressantes est un thème privilégie de l'extraction des connaissances à partir des données. Les algorithmes du type Apriori fondes sur le support et la confiance des règles ont apporté une solution élégante au problème de l'extraction de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes. Il faut disposer d'autre mesures venant compléter le support et la confiance. Dans cet article, nous passons en revue les principales mesures proposées dans la littérature et nous proposons des critères pour les évaluer. Nous suggérons ensuite une méthode de validation qui utilise les outils de la théorie de l'apprentissage statistique, notamment la VC-dimension. Face au grand nombre de mesures et a la multitude de règles candidates, l'intérêt de ces outils est de permettre la construction de bornes uniformes non asymptotiques pour toutes les règles et toutes les mesures simultanément

8. L'algorithme APRIORI

C'est l'inévitable algorithme d'extraction des règles d'association proposé par Agrawal et al. En 1994. Cet algorithme se base essentiellement sur les deux propriétés suivantes :

- **Propriété sur les sous-ensembles (antimonotonicité) :**

  «Tous les sous-ensembles d'un itemset fréquent sont fréquents ». Cette propriété permet de limiter le nombre de candidats de taille k générés lors de la 4 -ème itération (parmi les 2k -1 itemsets de taille k existants) en réalisant une jointure conditionnelle des itemsets fréquents de taille k-1 découverts lors de l'itération précédente.

- **Propriété sur les sur-ensembles :**

  « Tous les sur-ensembles d'un itemset non fréquent sont non fréquents ». Cette propriété permet de supprimer un candidat de taille k lorsque au moins un de ses sous-ensembles de taille k-1 ne fait pas partie des itemsets fréquents découverts lors de l'itération précédente.

8.1. La recherche des Itemsets Fréquents

L'algorithme APRIORI utilise une approche itérative par niveaux pour générer les itemsets fréquents. Pour cela, le treillis des itemsets est exploré en largeur d'abord. APRIORI effectue à chaque itération k, un passage dans la base de transactions afin de calculer le support de chaque kitemset.

II. Chapitre 2 : La recherche de règles d'associations

Soit Ck l'ensemble des k-itemsets candidats, et Fk l'ensemble des k-itemsets fréquents de taille k.

- Considérons le treillis des itemsets. C1 est l'ensemble des 1-itemsets candidats du treillis

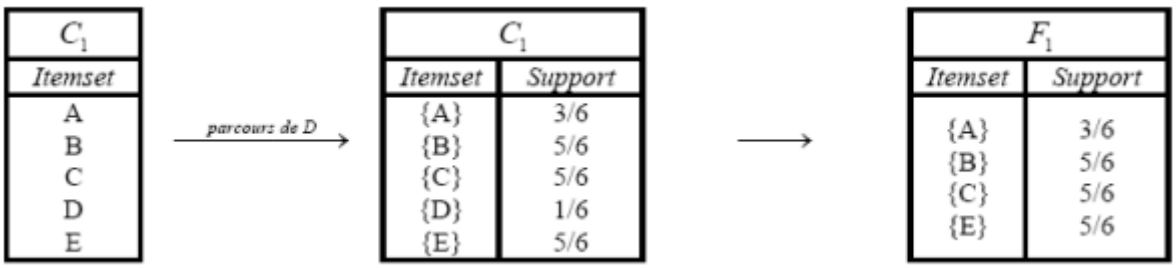
- Le contexte (base de données) D est parcouru afin de trouver F1 (l'ensemble des 1-itemsets fréquents dans D). L'algorithme calcul le support pour chaque 1-itemset, s'il est supérieur au support minimum, cet 1-itemset est ajouté à F1. Une fois la liste F1des 1-itemsets fréquents est construite, alors l'ensemble C2 des 2-itemsets candidats est généré par jointure entre F1 et F1 elle-même, en combinant tous les cas possibles des 1-itemsets pour avoir les 2-itemsets sans répétition d'un item. Le processus se répète ensuite pour C2

- Pour l'itération k, l'ensemble Fk-1 des (k-1)-itemsets fréquents correspondant aux itemsets de niveau (k - 1) du treillis (calculé à l'étape précédente), est utilisé pour générer l'ensemble Ck des k-itemsets candidats. Pour cela, une jointure est effectuée entre Fk-1 et Fk-1 pour former tous les k-itemsets possibles (deux itemsets p et q de Fk-1 forment un kitemset c si et seulement s'ils ont k-2 itemsets en commun, et c est de taille k). Donc, le nouveau candidat ne contient pas un sous-ensemble non fréquent (selon la propriété d'antimonotonicité). Une fois l'ensemble Ck, des candidats de taille k calculé, la base de transactions est parcourue transaction par transaction afin de déterminer le support de chaque candidat. Parmi les candidats seuls les candidats fréquents, i.e. de support suffisant sont gardés dans l'ensemble Fk. C'est cet ensemble qui est retourné à la fin du processus de génération des itemsets fréquents.

  Exemple:

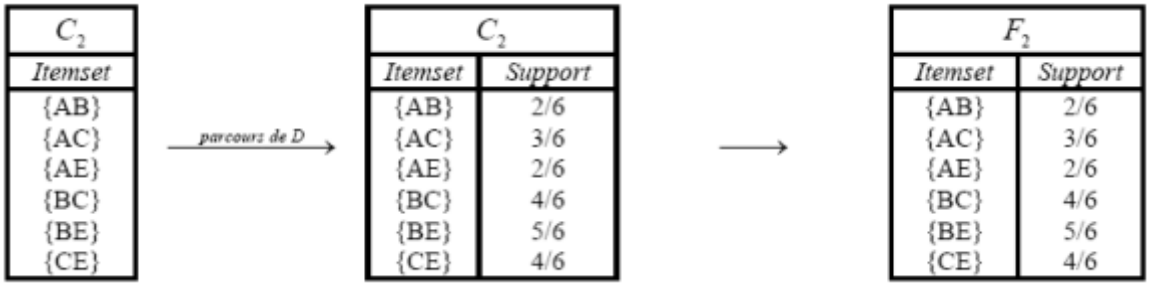
  L'exemple ci-dessous montre le processus d'extraction des itemsets fréquents sur la base de transactions D de notre exemple (tableau 1) pour un support minimum=2/6.

  - Á la première itération, chaque item X est un 1-itemset de C1. Un 1er parcours de D permet de trouver le support de chaque 1-itemset. Tous les 1-itemsets fréquents, i.e. de support supérieur ou égal à 2/6 seront gardés dans F1.

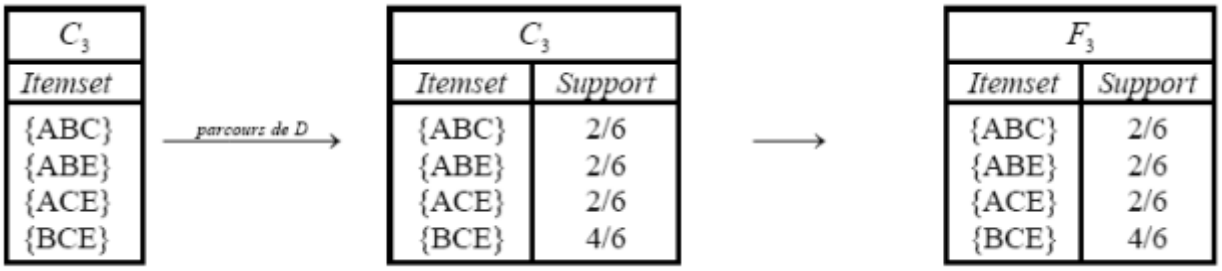


  - Afin de découvrir les 2-itemsets fréquents, Apriori effectue dans la 2ème itération une jointure de F1 X F1 pour trouver l'ensemble C2 des candidats de

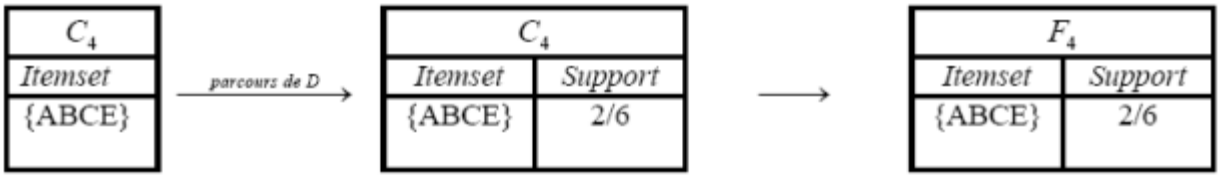
taille 2. Seuls les 2-candidats n'ayant pas de sous-ensembles non fréquents sont gardés. Un 2ème parcours de D est alors effectué pour déterminer le support de chacun des 2-itemsets candidats, seuls les 2-itemsets fréquents sont gardés dans F2.



- Les 3-itemsets sont obtenus en combinant les itemsets de F2 deux à deux, i.e. par jointure F2 X F2. Seuls les 2-itemsets ayant le même préfixe de taille 1 sont générés. Par exemple les 2-itemsets AB et AC forment le candidat ABC. On s'assure également que les candidats générés n'ont pas de sous-ensembles non fréquents. Un 3ème parcours de D est alors effectué pour déterminer les 3-itemsets fréquents.



- Les 4-itemsets sont obtenus en combinant les itemsets de F3 deux à deux, i.e. par jointureF3 X F3. Seuls les 3-itemsets ayant le même préfixe de taille 2 sont générés. Par exemple les 2-itemsets ABC et ABE forment le candidat ABCE. On s'assure également que les candidats générés n'ont pas de sous-ensembles peu fréquents. Un 4ème parcours de D est alors effectué pour déterminer les 4-itemsets fréquents.



- A la fin, 15 itemsets sont fréquents, l'itemset ∅ n'étant pas considéré et le plus grand des itemsets fréquents est {ABCE} dont la taille est 4.

II. Chapitre 2 : La recherche de règles d'associations

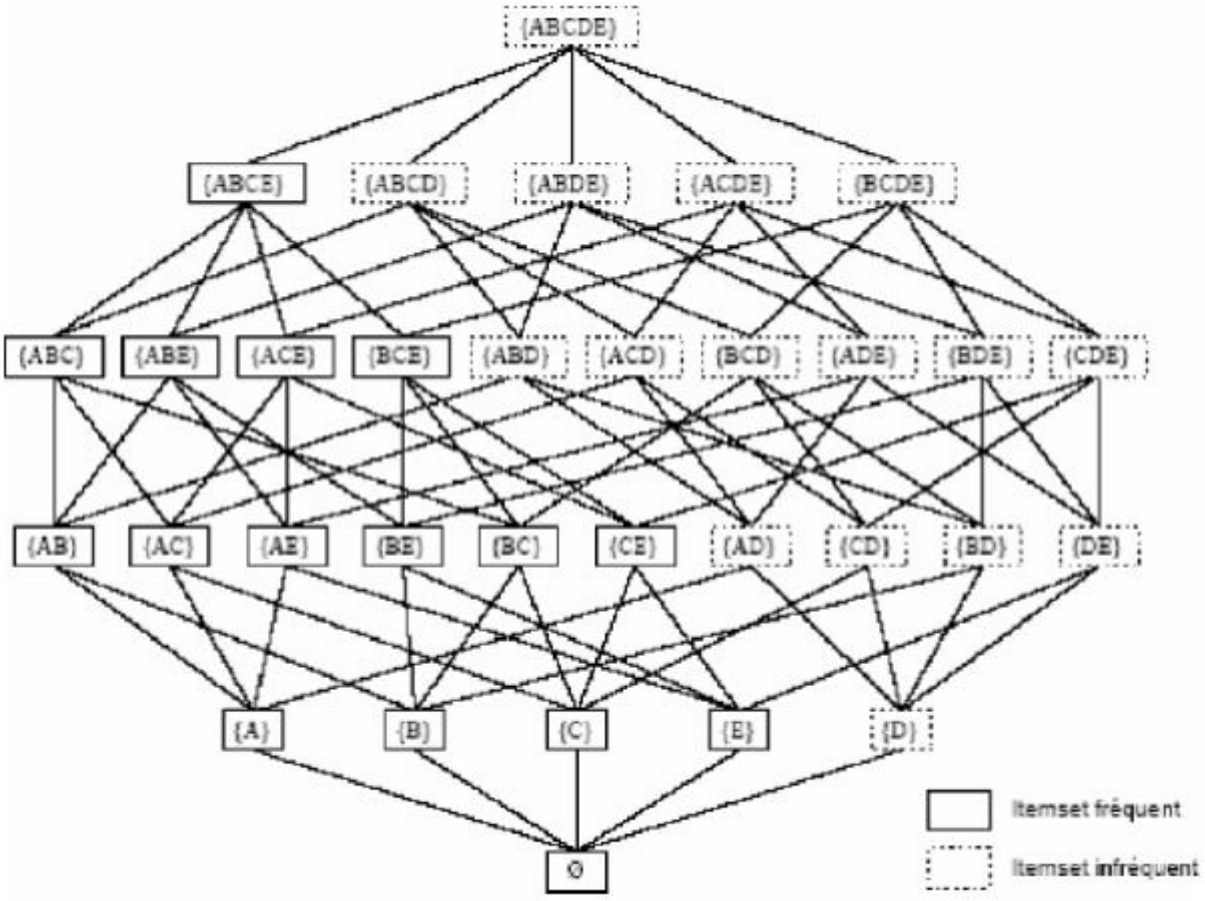


Figure 10 : Représentation des items ets fréquents dans le treillis des itemsets [3]

8.2.Génération des règles d'association

Cette étape est plus simple et beaucoup moins coûteuse que la génération des itemsets fréquents car il n'est plus nécessaire de faire des parcours coûteux de la base de transactions. Toutefois, cette phase reste tout de même exponentielle dans la taille des itemsets fréquents car le nombre de règles pouvant être générées à partir d'un k-itemset de taille supérieure à 1 est égal 2k -2. Agrawal et Srikant ont proposé une optimisation à la génération des règles d'association. Cette optimisation est basée sur la propriété suivante : « Si une règle ayant une conséquence C est valide, toutes les règles ayant pour conséquence des sous-ensembles de C sont aussi valides ». Ceci peut être formalisé comme suit :

Soit l un itemset fréquent, alors :

$$\forall C \subset l, C \neq \phi, \left[(l - C) \rightarrow C\right] est\ solide \Rightarrow \forall \tilde{C} \subset C, \tilde{C} \neq \phi, \left[(l - \tilde{C}) \rightarrow \tilde{C}\right] est\ solide$$

Cette propriété permet de ne pas considérer tous les sous-ensembles possibles des itemsets fréquents. Le principe de génération des règles d'association est le suivant :

II. Chapitre 2 : La recherche de règles d'associations

- Soit F l'ensemble des itemsets fréquents trouvés en phase précédente. Pour chaque itemset fréquent X appartenant à F, on considère tous ses sous-ensembles (qui sont aussi fréquents d'après la propriété d'antimonotonicité).

- Á partir de ces sous-ensembles fréquents, on génère toutes les règles de la forme générale suivante : (X -C) ◊ C Cette règle est valide si sa confiance calculée est supérieure à la confiance minimum donnée. Le calcul de la confiance se fait par la formule donnée en 2.3.1 (Définition de la confiance). A partir des 3-itemsets, on utilise la propriété d'optimisation lors de la génération des règles pour éliminer dans les calculs inutiles.

  Exemple :

  Reprenons notre exemple, les tableaux ci-dessous 2, 3, 4, et 5 montrent les règles d'association générées pour une confiance minimumφ = 60%. Les règles d'association sont générées en considérant d'abord les itemsets fréquents de taille 2, puis ceux de taille 3, 4..., etc.

  Les itemsets fréquents de taille 2, F2 ont permis de générer les règles d'association du tableau 2. Dans ce cas la propriété d'optimisation n'est pas employée vu qu'on ne peut avoir plus d'un item en conséquence des règles d'association.

Itemset	N° Règle	Règle	Confiance	Support	Valide ?
AB	1	$A \rightarrow B$	2/3 = 66,66%	2/6 = 33,33%	Yes
	2	$B \rightarrow A$	2/5 = 40%	2/6 = 33,33%	No
AC	3	$A \rightarrow C$	3/3 =100%	3/6 = 50,00%	Yes
	4	$C \rightarrow A$	3/5 = 60%	3/6 = 50,00%	Yes
AE	5	$A \rightarrow E$	2/3=66,66%	2/6=33,33%	Yes
	6	$E \rightarrow A$	2/5=40%	2/6=33,33%	No
BC	7	$B \rightarrow C$	4/5 = 80%	4/6 = 66,66%	Yes
	8	$C \rightarrow B$	4/5 = 80%	4/6 = 66,66%	Yes
BE	9	$B \rightarrow E$	5/5 = 100%	5/6 = 83,33%	Yes
	10	$E \rightarrow B$	5/5 = 100%	5/6 = 83,33%	Yes
CE	11	$C \rightarrow E$	4/5 = 80%	4/6 = 66,66%	Yes
	12	$E \rightarrow C$	4/5 = 80%	4/6 = 66,66%	Yes

Tableau 2: Règles d'association à deux items et à un item comme conséquence

Les itemsets fréquents de taille 3, F3 à savoir les itemsets{ABC}{ABE}{ACE}{BCE} ont permis de générer les règles, d'abord avec un conséquent à un item figurant dans le tableau 3

II. Chapitre 2 : La recherche de règles d'associations

Itemset	N° Règle	Règle	Confiance	Support	Valide ?
ABC	13	AB→C	2/2=100%	2/6=33,33%	Yes
	14	AC→B	2/3=66,66%	2/6=33,33%	Yes
	15	BC→A	2/4=50%	2/6=33,33%	No
ABE	16	AB→E	2/2=100%	2/6=33,33%	Yes
	17	AE→B	2/2=100%	2/6=33,33%	Yes
	18	BE→A	2/5=40%	2/6=33,33%	No
ACE	19	AC→E	2/3=66,66%	2/6=33,33%	Yes
	20	AE→C	2/2=100%	2/6=33,33%	Yes
	21	CE→A	2/4=50%	2/6=33,33%	No
BCE	22	BC→E	4/4=100%	4/6=66,66%	Yes
	23	BE→C	4/5=80%	4/6=66,66%	Yes
	24	CE→B	4/4=100%	4/6=66,66%	Yes

Tableau 3 : Règles d'association à trois items et à un item comme conséquence.

Les conséquences de taille d'un item trouvées ayant constitué des règles valides seront ensuite combinées pour constituer des conséquences de taille 2 pour les règles d'association du tableau 4.

Itemset	N° Règle	Règle	Confiance	Support	Valide ?
ABC	25	A→BC	2/3=66,66%	2/6=33,33%	Yes
	26	C→AB	2/5=33,33%	2/6=33,33%	No
	27	B→AC	2/5=33,33%	2/6=33,33%	No
ABE	28	A→BE	2/3=66,66%	2/6=33,33%	Yes
	29	E→AB	2/5=33,33%	2/6=33,33%	No
	30	B→EA	2/5=33,33%	2/6=33,33%	No
ACE	31	A→EC	2/3=66,66%	2/6=33,33%	Yes
	32	E→AC	2/5=33,33%	2/6=33,33%	No
	33	C→EA	2/5=33,33%	2/6=33,33%	No
BCE	34	B→EC	4/5=80%	4/6=66,66%	Yes
	35	E→BC	4/5=80%	4/6=66,66%	Yes
	36	C→EB	4/5=80%	4/6=66,66%	Yes

Tableau 4 : Règles d'association à trois items et à deux items comme conséquence.

L'itemset fréquent de taille 4 {ABCE} a permis de générer les règles d'association figurant dans le tableau 5.

Itemset	N° Règle	Règle	Confiance	Support	Valide?
ABCE	37	A→BCE	2/3=66,66%	2/6=33,33%	Yes
	38	C→ABE	2/5=40%	2/6=33,33%	No
	39	B→ACE	2/5=40%	2/6=33,33%	No
	40	E→ABC	2/5=40%	2/6=33,33%	No
ABCE	41	AB→CE	2/2=100%	2/6=33,33%	Yes
	42	AC→BE	2/3=66,66%	2/6=33,33%	Yes
	43	AE→BC	2/2=100%	2/6=33,33%	Yes
	44	BC→AE	2/4=50%	2/6=33,33%	No
	45	BE→AC	2/5=40%	2/6=33,33%	No
	46	CE→AB	2/4=50%	2/6=33,33%	No
ABCE	47	ABC→E	2/2=100%	2/6=33,33%	Yes
	48	ACE→B	2/2=100%	2/6=33,33%	Yes
	49	BCE→A	2/4=50%	2/6=33,33%	No
	50	ABE→C	2/2=100%	2/6=33,33%	Yes

Tableau 5 : Règles d'association à quatre items

9. Conclusion

Les règles d'associations en générale sont des outils efficaces pour identifier des relations entre attributs dans les bases de données. De même, ils peuvent faire découvrir aux analystes des associations inattendues. Ces relations peuvent ensuite être utilisées et intégrées dans les processus d'affaires de l'entreprise afin d'en améliorer les performances.

Sachant que les bases de données volumineuses demandent beaucoup de temps de calculs, qui devient un problème remarquable, les algorithmes discutés réussissent tout de même à identifier les règles d'associations présentes dans les données.

III. Chapitre 3 : La prédiction par les règles d'associations

Chapitre 3 : La prédiction par les règles d'associations

1. Introduction

L'objectif de ce chapitre est de proposer un nouvel algorithme d'extraction de règles d'association positives et négatives utilisant les optimisations que nous avons proposées dans le chapitre précèdent. Ces travaux ont été publiés dans [8] et [9]. La première partie de ce chapitre présente les différentes contraintes que doivent respecter les règles pour être considérées comme valides et être extraites par notre algorithme. Nous présentons ensuite les différents algorithmes que nous utilisons au cours du processus d'extraction. Le processus d'extraction se divise en trois parties, à savoir la recherche des motifs raisonnablement fréquents, la recherche des motifs négatifs minimaux raisonnablement fréquents et enfin l'extraction des règles valides à partir des motifs raisonnablement fréquents précédemment trouves.

2. Règles d'association positives et négatives valides

Chaque règle, quel que soit son type, est générée à partir des motifs raisonnablement fréquents XY. Il faut donc que le support du motif XY vérifie min_{sup}, max_{sup} et que le support du motif X¨Y vérifie min_{sup}. La règle est ensuite générée si son support, sa confiance et sa valeur de MG sont valides. Pour les règles négatives X ⇒ Y, il faut vérifier que les prémisses et/ou les conclusions négatives sont constituées de motifs négatifs minimaux raisonnablement fréquents dont nous donnons la définition :

Définition - Motif négatif minimal raisonnablement fréquent :

Un motif XY est motif négatif minimal raisonnablement fréquent si min_{sup} 6 sup(XY) 6 max_{sup} et sup(X) < min_{sup} et sup(Y) < min_{sup}. Une dernière contrainte est appliquée au nouveau type de règle ¨X ⇒ ¨ Y, ou le motif X¨Y doit avoir une taille strictement supérieure a` 2. L'omission de cette contrainte amenée a généré des règles en double. En effet, pour un 2 -motif i1i2, les règles i1 ⇒ i2 et ¨i 1 ⇒ ¨i2 sont identiques. Afin d'éviter cette redondance, nous pouvons appliquer cette dernière contrainte sur les règles ¨X ⇒ ¨ Y ou sur les règles X ⇒ Y. Notre approche étant la seule à notre connaissance à générer les règles composées de conjonctions de motifs négatifs, il est plus pertinent de garder les règles X ⇒ Y afin de pouvoir les comparer avec celles extraites par les autres approches. C'est pourquoi, nous choisissons d'appliquer cette dernière contrainte sur les règles ¨X ⇒ Y¨ (i.e. taille (X¨Y)> 2).

Le tableau 6 recapitule les contraintes pour qu'une règle soit valide pour notre approche.

En conclusion, l'ensemble des règles X ⇒ Y doit respecter le support minimum, le support maximum, le support du motif négatif X¨Y, la mesure M_G et la confiance.

Dans les prochaines sections, nous présentons les différentes étapes de l'algorithme. L'algorithme va se dérouler en trois étapes :

- Rechercher l'ensemble RF des motifs raisonnablement fréquents.
- Rechercher l'ensemble NMRF des motifs négatifs minimaux raisonnablement fréquents.
- Générer les règles valides.

La prochaine section se focalise sur la première étape de l'algorithme qui correspond à la recherche des motifs raisonnablement fréquents.

$X \Rightarrow Y$	$\overline{X} \Rightarrow Y$
$sup(XY) \geq min_{sup}$ $sup(\ddot{X}\ddot{Y}) \geq min_{\ddot{sup}}$ $sup(XY) \leq max_{sup}$ $M_G(X \Rightarrow Y) \geq min_{M_G}$ $conf(X \Rightarrow Y) \geq min_{conf}$	$sup(XY) \geq min_{sup}$ $sup(\ddot{X}\ddot{Y}) \geq min_{\ddot{sup}}$ $sup(XY) \leq max_{sup}$ $sup(\overline{X}Y) \geq min_{sup}$ $sup(\overline{X}Y) \leq max_{sup}$ $M_G(\overline{X} \Rightarrow Y) \geq min_{M_G}$ $conf(\overline{X} \Rightarrow Y) \geq min_{conf}$ $\overline{X}$ *minimal*
$X \Rightarrow \overline{Y}$	$\overline{X} \Rightarrow \overline{Y}$
$sup(XY) \geq min_{sup}$ $sup(\ddot{X}\ddot{Y}) \geq min_{\ddot{sup}}$ $sup(XY) \leq max_{sup}$ $sup(X\overline{Y}) \geq min_{sup}$ $sup(X\overline{Y}) \leq max_{sup}$ $M_G(X \Rightarrow \overline{Y}) \geq min_{M_G}$ $conf(X \Rightarrow \overline{Y}) \geq min_{conf}$ $\overline{Y}$ *minimal*	$sup(XY) \geq min_{sup}$ $sup(\ddot{X}\ddot{Y}) \geq min_{\ddot{sup}}$ $sup(XY) \leq max_{sup}$ $sup(\overline{X}\,\overline{Y}) \geq min_{sup}$ $sup(\overline{X}\,\overline{Y}) \leq max_{sup}$ $M_G(\overline{X} \Rightarrow \overline{Y}) \geq min_{M_G}$ $conf(\overline{X} \Rightarrow \overline{Y}) \geq min_{conf}$ $\overline{X}$ et $\overline{Y}$ *minimaux*
$\ddot{X} \Rightarrow \ddot{Y}$	
$sup(XY) \geq min_{sup}$ $sup(\ddot{X}\ddot{Y}) \geq min_{\ddot{sup}}$ $sup(XY) \leq max_{sup}$	$M_G(\ddot{X} \Rightarrow \ddot{Y}) \geq min_{M_G}$ $conf(\ddot{X} \Rightarrow \ddot{Y}) \geq min_{conf}$ taille$(\ddot{X}\ddot{Y}) > 2$

Tableau 6 : Règles valides

3. Génération des règles d'association positives et négatives

Avant de présenter l'algorithme d'extraction des règles d'association positives et négatives, nous allons modéliser la propriété de la confiance présentée dans la propreté

1, sous forme de méta-règle :

M R_9 : ∀ (X, Y, Z) tel que Z ⊄ Y ⊄ X.

Si $conf(Z \Rightarrow X \backslash Z) < min_{conf}$ alors $conf(Y \Rightarrow X \backslash Y) < min_{conf}$

Cette méta-règle sera utilisée pour l'extraction des règles du nouveau type et permettra d´éviter l'étude de règles dont nous savons par avance que la valeur de la confiance est insuffisante. Nous aurions également pu utiliser cette méta-règle pour les règles positives, seulement nous devons calculer la confiance de la règle positive pour déterminer la zone d'appartenance de la règle. Par conséquent, nous sommes obligés de calculer la confiance de la règle positive et la méta-règle n'est donc plus utile pour règles positives

III. Chapitre 3 : La prédiction par les règles d'associations

L'algorithme utilise ´également une méta-règle que nous avons présentée au chapitre précédent. Nous la rappelons ci-dessous. La méta-règle dont nous allons nous servir est MR_4.

M R_4 : $\forall X \Rightarrow Y$ avec($\frac{1}{2} < sup(X) < \sup(Y)\ ou\ (\sup(X) < \frac{1}{2} < \sup(\sup(Y))$

Si $M_G(X \Rightarrow Y) < min_{M_G}\ alors\ M_G(\bar{X} \Rightarrow \bar{Y}) < min_{M_G}$

Cette méta-règle M R_4 nous relève, que si la règle $X \Rightarrow Y$ est invalide pour la mesure M_G dans le cas où $(\frac{1}{2} < \sup(X) < \sup(Y) < \frac{1}{2} < \sup(Y))$ ou $(\sup(X) < \frac{1}{2} < \sup(Y))$, alors la règle $\bar{X} \Rightarrow \bar{Y}$ sera également invalide. Cette méta-règle sera utilisé dans la zone attractive mais également dans la zone répulsive. Dans la zone0 répulsive l'invalidité de la règle $\bar{X} \Rightarrow Y$ sera déduite à partir de l'intérêt de la règle $X \Rightarrow \bar{Y}$.

La génération des règles d'association positive et négative, à partir des motifs raisonnablement fréquents préalablement trouvés, et effectuée par la fonction RAPN (algorithme1). Cet algorithme étant assez long, nous l'avons divisé en plusieurs fonctions. Chaque type de règle va être étudié par une fonction différente : l'algorithme 2 l'algorithme de pour les règles du type $X \Rightarrow Y$ l'algorithme 3 pour les règles du type $\bar{X} \Rightarrow \bar{Y}$ l'algorithme 4 pour les règles du type $X \Rightarrow \bar{Y}$ l'algorithme 5 pour les règles du type $\bar{X} \Rightarrow Y$ et l'algorithme 6 pour les règles du type $\ddot{X} \Rightarrow \ddot{Y}$.

Après avoir initialisé l'ensemble R des règles valides à l'ensemble nul (ligne 1), le processus va extraire l'ensemble des règles valides (lignes 2 à 14). Ainsi pour chaque motif raisonnablement fréquent $X \in RF$ avec une taille strictement supérieure à 1 (ligne 2) (puisqu'on une règle ne peut pas générer une règle comportant qu'un seul item) et pour chaque motif conclusion $Y \subsetneq X$ (ligne 3) ordonné par taille croissante (comme pour Apriori), nous commençons par déterminer la zone d'appartenance de la règle $X \backslash Y \Rightarrow Y$ en comparant la confiance de cette règle au support de la conclusion Y. Pour ce faire, nous devons d'abord calculer la confiance de la règle $X \backslash Y \Rightarrow Y$ (ligne 4).

Si la règle est dans la zone attractive $i.e. conf(X \backslash Y \Rightarrow Y) > \max(\frac{1}{2}, sup(Y)))$ (ligne 5), alors nous étudiants uniquement deux règles : les règles$XY \Rightarrow Y$ et $\overline{XY} \Rightarrow \bar{Y}$en effet comme nous pouvons le voir dans la figure renseignant sur les règles potentiellement intéressantes et les règles inintéressantes en fonction de l'intérêt de la règle positive, nous pouvons voir que lorsque la confiance de la règle est supérieure au support de la conclusion la règle se situe dans la zone potentiellement intéressante, et seules les règles du type peuvent être intéressantes La règle du type $X \backslash Y \Rightarrow Y, Y \Rightarrow X \backslash Y, \overline{X \backslash Y} \Rightarrow \bar{Y}\ et\ \bar{Y} \Rightarrow \overline{X \backslash Y}$ peuvent être intéressantes.

Algorithme 1: RAPN – Extraction des règles d'association Positives et Négatives

III. Chapitre 3 : La prédiction par les règles d'associations

Entrées : base de données D, ensemble des motifs Raisonnablement Fréquents RF, ensemble des motifs Négatifs Minimaux Raisonnablement Fréquents NMRF, support minimum min_{sup}, support maximum max_{sup}, confiance minimum min_{cof}, M_G minimum min_{M_G}.

Sortie : ensemble des règles valides R

```
1  R = ∅
2  pour tout k-motif X ∈ RF tel que k > 1 faire
3      pour toute conclusion Y ⊊ X telle que taille(Y) ↑ faire
4          c = conf(X\Y ⇒ Y)
5          si c > max(1/2, sup(Y)) alors
6              Étude de X\Y ⇒ Y
7              si X\Y ∈ NMRF et Ȳ ∈ NMRF alors
8                  [¬MR4] Étude de X\Y ⇒ Ȳ
9          sinon si c < min(1/2, sup(Y)) alors
10             si Ȳ ∈ NMRF alors
11                 Étude de X\Y ⇒ Ȳ
12             si X\Y ∈ NMRF alors
13                 [¬MR4] Étude de X\Y ⇒ Y
14         [¬MR9] Étude de X¨\Y ⇒ Ÿ
15 retourner R
```

La zone attractive de la mesure M_G correspond à la zone potentielle intéressante à laquelle nous ajoutons la prise en compte de l'équilibre, c'est-à-dire qu'il faut également vérifier que la confiance de la règle est supérieure à $\frac{1}{2}$.

La zone dans la zone attractive seul c'est 4 règles peuvent être intéressantes à condition qu'elle s'éloigne suffisamment du point d'équilibre. Par ailleurs par ailleurs, vu que notre algorithme parcourt l'ensemble des conclusions Y possible, nous allons uniquement étudier les règles $X\backslash Y \Rightarrow Y$ *et* $\overline{XY} \Rightarrow \bar{Y}$ dans la zone attractive. Réciproquement, nous étudiants uniquement les règles $X\backslash Y \Rightarrow \bar{Y}$ et $\overline{X\backslash Y} \Rightarrow Y$ dans la zone répulsive.

Début no débutant donc par l'étude de $X\backslash Y \Rightarrow Y$ (ligne 6) avec l'algorithme 2.

Algorithme 2 : : Etude des règles du type . $X\backslash Y \Rightarrow Y$

```
1 si c ⩾ min_conf alors
2     si M_G(X\Y ⇒ Y) ⩾ min_MG alors
3         R = R ∪ {X\Y ⇒ Y}
```

Dans l'algorithme 2, nous vérifions que la valeur de la confiance (déjà calculée dans l'algorithme principal (algorithme 1 ligne 4) est supérieur au seuil (ligne 1 algorithme 2). Puis, si la confiance de la règle est valide alors on vérifie la validité de la règle pour la mesure M_G (ligne 2

algorithme 2). Si M_G vérifie le seuil alors la règle$X\backslash Y \Rightarrow Y$ est ajoutée à l'ensemble R des règles valides (ligne 3 algorithme 2).

Nous retournons ensuite à l'algorithme principal (algorithme 1) pour étudier la règle négative $\overline{X\backslash Y} \Rightarrow \bar{Y}$. Si $\overline{X\backslash Y}$et $\bar{Y}$ sont des motifs négatifs minimaux raisonnablement fréquent, c'est-à-dire si $\overline{X\backslash Y}$ et $\bar{Y}$ sont contenus dans la NMRF (ligne 7), alors nous étudions la règle $\overline{X\backslash Y} \Rightarrow \bar{Y}$ (ligne 8) si la méta-règle ne peut être appliquée. L'étude de cette règle est réalisée avec l'algorithme 3.

Algorithme 3: Etude des règles du type $X\backslash Y \Rightarrow Y$.

1 **si** $min_{sup} \leqslant sup(X\overline{Y}) \leqslant max_{sup}$ **alors**
2 **si** $conf(X\backslash Y \Rightarrow \overline{Y}) \geqslant min_{conf}$ **alors**
3 **si** $M_G(X\backslash Y \Rightarrow \overline{Y}) \geqslant min_{M_G}$ **alors**
4 $R = R \cup \{X\backslash Y \Rightarrow \overline{Y}\}$

Cet algorithme 3 commence par vérifier si le support de la règle $\overline{X\backslash Y} \Rightarrow \bar{Y}$est valide (ligne 1 algorithme 16). Pour vérifier le support, nous utilisons la formule suivante :

$$\sup(\bar{X}\bar{Y}) = 1 - \sup(X) - \sup(Y) + \sup(XY).$$

Cette formule nous évite d'interroger la base de données D. Si le support est vérifié, nous étudiants ensuite la confiance (ligne 2 algorithme 3). Si la confiance est valide nous calculons la règle de ligne 3 algorithme 3. Si la règle est valide pour la mesure M_G, la règle $\overline{X\backslash Y} \Rightarrow \bar{Y}$ est ajoutée à l'ensemble des règles valides R puis nous retournons à l'algorithme principal (algorithme 1).

Si la règle $X\backslash Y \Rightarrow Y$ est dans la zone répulsive, ($i.e\ conf\left(\frac{X}{Y} \Rightarrow Y\right) < \min\left(\frac{1}{2}, \sup(Y)\right)$ (ligne 9), alors nous étudions les règles : $X/Y \Rightarrow \bar{Y}$et $\overline{X\backslash Y} \Rightarrow Y$. Si $\bar{Y}$est minimale (ligne 10), nous étudions la règle (ligne 11). SI $\overline{X\backslash Y}$est un motif minimal (ligne12) et que M R_4 n'est pas vérifiée alors la règle $X\backslash Y \Rightarrow \bar{Y}$ est ensuite étudier (ligne 13). La vérification des contraintes (support, confiance et M_G) se fait avec l'algorithme 4 pour les règles du type $X \Rightarrow \bar{Y}$ et algorithme 5 pour les règles du type $\bar{X} \Rightarrow Y$. Expliquant maintenant ces deux algorithmes très semblables aux algorithmes 2 et 3.

1 **si** $min_{sup} \leqslant sup(\overline{X}\,\overline{Y}) \leqslant max_{sup}$ **alors**
2 **si** $conf(\overline{X\backslash Y} \Rightarrow \overline{Y}) \geqslant min_{conf}$ **alors**
3 **si** $M_G(\overline{X\backslash Y} \Rightarrow \overline{Y}) \geqslant min_{M_G}$ **alors**
4 $R = R \cup \{\overline{X\backslash Y} \Rightarrow \overline{Y}\}$

III. Chapitre 3 : La prédiction par les règles d'associations

L'algorithme 3 commence par vérifier si le support de la règle $X \backslash Y \Rightarrow \bar{Y}$ est valide (ligne 1 algorithme 3). Le support peut être calculé avec la formule $\sup(X\bar{Y}) = \sup(X) - \sup(XY)$ et évite d'interroger la base de données D. Nous vérifions ensuite, si le support est valide, que la contrainte de la confiance (ligne 2 algorithme 3) et respectée. Si c'est le cas, nous vérifions enfin la valeur de M_G (ligne 3 algorithmes 3). Si la valeur de M_G est suffisante, la règle $X \Rightarrow \bar{Y}$ est ajoutée à l'ensemble des règles valides R (ligne 4 algorithmes 3) et nous retournant à l'algorithme principal (algorithme 1). L'algorithme 5 effectue les mêmes opérations mais utilise une formule différente pour calculer le support.

Algorithme 4: : Etude des règles du type $\overline{X \backslash Y} \Rightarrow Y$.

1 **si** $taille(\ddot{X}) > 2$ **alors**
2 **si** $conf(X\ddot{\backslash}Y \Rightarrow \ddot{Y}) \geqslant min_{conf}$ **alors**
3 **si** $M_G(X\ddot{\backslash}Y \Rightarrow \ddot{Y}) \geqslant min_{M_G}$ **alors**
4 $R = R \cup \{X\ddot{\backslash}Y \Rightarrow \ddot{Y}\}$

En effet dans l'algorithme 5 la formule que nous allons utiliser est la suivante :

$(\bar{X}Y + \sup(Y) - \sup(XY)$.

Les règles du type $\bar{X} \Rightarrow Y$ sont ajoutées à l'ensemble des règles valides R si les trois contraintes successives sont respectées. Nous retournons ensuite à l'algorithme principal (algorithme 1) pour étudier le dernier type de règles.

Le nouveau type de règles $\bar{x}_1..\bar{x}_p \Rightarrow \bar{y}_1 ..\bar{y}_q$ est ensuite étudié à l'aide de l'algorithme 5 si la métarègle M R_9 nous l'autorise. En effet, la contrainte de la confiance fonctionne sur les motifs $\ddot{X}$ puisque le support pour ce motif possède une propriété anti-monotone.

Algorithme 5: : Etude des règles du type $X\ddot{\backslash}Y \Rightarrow \ddot{Y}$.

III. Chapitre 3 : La prédiction par les règles d'associations

1 **si** $min_{sup} \leqslant sup(\overline{X}Y) \leqslant max_{sup}$ **alors**
2 **si** $conf(\overline{X \backslash Y} \Rightarrow Y) \geqslant min_{conf}$ **alors**
3 **si** $M_G(\overline{X \backslash Y} \Rightarrow Y) \geqslant min_{M_G}$ **alors**
4 $R = R \cup \{\overline{X \backslash Y} \Rightarrow Y\}$

Cet algorithme 6 commence par vérifier que le motif générant la règle et la taille strictement supérieur à 2 (ligne 1 algorithme 6) enfin de ne pas extraire des règles en double comme nous l'avons expliqué précédemment. Si la taille du motif est suffisante alors nous vérifions ensuite la contrainte de la confiance (ligne 2 algorithme 6). Si la confiance est valide, nous vérifions la dernière contrainte qui est la mesure M_G (ligne 3 algorithme 6). Si la règle $\ddot{X} \Rightarrow \ddot{Y}$ possède une valeur de M_G suffisante, alors elle est ajoutée à l'ensemble des règles valides R. Dans cette étude, nous n'avons pas besoin de vérifier que le support de la règle est valide puisque cette contrainte est déjà vérifiée lors de l'extraction des motifs raisonnablement fréquents. En effet, durant la recherche des motifs raisonnablement fréquents, nous avons ajouté la contrainte $min_{süp}$ afin de vérifier le support des motifs $\ddot{X}$.

La dernière étape de l'algorithme 1 consiste à retourner l'ensemble R des règles valides (ligne 15).

Exemple

Le tableau rappelle les données que nous allons utiliser pour notre exemple. Nous prenons les mêmes valeurs pour le support et la confiance que celle utilisée dans les exemples précédents. Les paramètres sont donc les suivants : 0,25 pour le support minimum et pour le support minimum du motif négatif, 0,90 pour le support maximum, 0,80 pour la confiance minimum et 0,50 pour la valeur de la mesure M_G.

A	B	C	D	E
1	0	1	1	0
0	1	1	0	1
1	1	1	0	0
0	1	0	0	1

Tableau 7: Exemple fil-rouge.

Nous rappelons que l'algorithme se déroule en trois étapes. La première étape consiste à rechercher les motifs raisonnablement fréquents. Nous cherchons ensuite les motifs négatifs

minimaux raisonnablement fréquent. La dernière étape concerne la génération des règles valides. Nous commençons donc par chercher les motifs raisonnablement fréquents.

- **Étape 1 : recherche des motifs raisonnablement fréquents**

Nos débutants en récupérant les items I contenus dans la base de données dont le support est inférieur ou égal au support minimum et dont le support de l'items négatifs i et également inférieur ou égal au support maximum. Le tableau fournit le support des items de la base ainsi que le support d'item négatif.

Item	Support	Support item négatif	Item	Support	Support item négatif
A	0,50	0,50	*B*	0,75	0,25
C	0,75	0,25	*D*	0,25	0,75
E	0,50	0,50			

Tableau 8: Items de taille 1.

L'ensemble L des items contenus dans le tableau respecte bien la contrainte du support maximum (0,90) et la contrainte du suppose maximum pour le motif négatif (0,90) . Par ailleurs, l'ensemble des items respecte la contrainte du support minimum (0,25) et la contrainte du support minimum pour le motif négatif (0,25), ils sont par conséquence ajouter à l'ensemble RF des motifs raisonnablement fréquents. Les items sont ensuite combinés à l'aide de la fonction candidat d'Apriori pour générer les candidats de taille 2. Le support ainsi que le support de items négatif. Le support ainsi que le support de l'item négatif sont ensuite vérifiés. Le résultat est visible dans le tableau.

2-Motif	Support	Support motif négatif	2-Motif	Support	Support motif négatif
AB	0,25	0	**AC**	0,50	0,25
AD	0,25	0,50	*AE*	0	0
BC	0,50	0	*BD*	0	0
BE	0,50	0,25	**CD**	0,25	0,25
CE	0,25	0	*DE*	0	0,25

Tableau 9: Candidats de taille 2.

Les motifs AC, AD, BE, et CD vérifient les deux contraintes suivantes : celle du support ainsi que celle du support du motif négatif. Ils sont donc ajoutés à l'ensemble des motifs raisonnablement fréquents. La génération des candidats se poursuit et engendrer uniquement le motif ACD c'est dans un support à 0,25 et un support négatif à 0,25. Le motif

ACD est fréquent. Le tableau récapitule l'ensemble des motifs raisonnablement fréquents trouver dans cet exemple.

Motif	Support	Support motif négatif	Motif	Support	Support motif négatif
A	0,50	0,50	AC	0,50	0,25
ACD	0,25	0,25	AD	0,25	0,50
B	0,75	0,25	BE	0,50	0,25
C	0,75	0,25	CD	0,25	0,25
D	0,25	0,75	E	0,50	0,50

Tableau 10: Ensemble RF des motifs raisonnablement fréquents extraits.

- **Étape 2 : recherche des motifs négatifs minimaux raisonnablement fréquents**

La seconde étape de l'algorithme consiste à rechercher l'ensemble des motifs négatifs minimaux raisonnablement fréquent point dans cet exemple, la recherche est rapide puisque des initialisations des motifs négatifs minimaux raisonnablement fréquents, nous obtenus l'ensemble des motifs négatifs minimaux raisonnablement fréquents. En effet, lors de l'initialisation, nous cherchons les négations d'items tel que le support de cette négation soit supérieure ou égal au seuil du support minimum et inférieur ou égal au seuil du support maximum. Les résultats sont visibles dans le tableau.

Motif	Support	Motif	Support	Motif	Support
$\overline{A}$	0,50	$\overline{B}$	0,25	$\overline{C}$	0,25
$\overline{D}$	0,75	$\overline{E}$	0,50		

Tableau 11: Ensemble NMRF des motifs négatifs minimaux raisonnablement

- **Étape 3 : génération des règles valides**

Une fois les motifs raisonnablement fréquents et les motifs négatifs minimaux raisonnablement fréquents trouver, la dernière phase de l'algorithme peut commencer. En effet, la recherche des règles se fait à partir des motifs raisonnablement fréquents, et la contrainte de minimalité pour les règles négatives et vérifier grâce au motif négatif minimaux raisonnablement fréquent. Prenons le motif raisonnablement fréquent ACD et cherchons les règles valides. La première étape de la recherche des règles va être de rechercher les règles possédant un seul item en conclusion. Tout d'abord, nous commençons par calculer la confiance afin de connaître les types de règles à étudier :

$$conf(CD \Rightarrow A) = \frac{\sup(ACD)}{\sup(CD)} = \frac{0,25}{0,25} = 1$$

$$conf(AD \Rightarrow C) = \frac{\sup(ACD)}{\sup(AD)} = \frac{0{,}25}{0{,}25} = 1$$

$$conf(AC \Rightarrow D) = \frac{\sup(ACD)}{\sup(AD)} = \frac{0{,}25}{0{,}50} = 0{,}50$$

Les règles $CD \Rightarrow A$ et $AD \Rightarrow C$ et $AC \Rightarrow D$ possèdent une confiance égale à 1. De plus, la confiance de ses règles est bien supérieure au minimum entre 1 / 2 et le support de leur conclusion (sup(A) pour la première et sup(C) pour la seconde). Par conséquence, nous étudions les règles de type $X \backslash Y \Rightarrow Y$ et $\overline{X \backslash Y} \Rightarrow \bar{Y}$ pour ces deux combinaisons. Quant à la confiance de la règle $AC \Rightarrow D$, elle n'est pas supérieure au minimum entre 1 / 2 et le support de leur conclusion et elle n'est pas non plus inférieur au minimum entre 1 / 2 et le support de sa conclusion. Par conséquent, pour cette combinaison ne nous étudions ni les règles du type $X \backslash Y \Rightarrow Y$ et $\overline{X \backslash Y} \Rightarrow \bar{Y}$ ni les règles du type $\overline{X \backslash Y} \Rightarrow Y$ et $X \backslash Y \Rightarrow \bar{Y}$ puisqu'elle se situe dans la zone unique intéressante.

Pour la combinaison $CD \cup A$, nous commençons par étudier la règle positive $CD \Rightarrow A$ en vérifiant si la valeur de la confiance est valide. La confiance a déjà été calculé pour déterminer la zone d'appartenance de la règle, nous la comparant simplement à la valeur de min_{conf}. Comme le seuil est égal à 0,80, la règle est valide pour la confiance. Nous continuons en étudiant la valeur M_G de la règle. La règle est en dans la zone attractive, nous prenons le calcul associé pour M_G :

$$M_{G_a}(CD \Rightarrow A) = \frac{conf(CD \Rightarrow A) - \max\left(\frac{1}{2}, \sup(A)\right)}{1 - \max\left(\frac{1}{2}, \sup(A)\right)}$$

La valeur de M_G pour la règle $CD \Rightarrow A$ est supérieur au seuil minimum de M_G (0,50). La règle est donc valide et sera conservée. L'étude se poursuit avec la règle négative $\overline{CD} \Rightarrow \bar{A}$. Nous vérifions tout d'abord que les motifs $\overline{CD}$ et $\bar{A}$ sont des motifs négatifs minimum. Or $\overline{CD}$ n'en est pas un (absent du tableau) et par conséquent il est inutile de générer la règle négative point nous cherchons ensuite le nouveau type de règle. Nous devant d'abord vérifier que les motifs ont une taille strictement supérieure à 2. Cette vérification permet de s'assurer de ne pas générer des règles redondantes., Comme nous l'avons dit précédemment, pour un motif X de taille 2, la combinaison $X \backslash Y \cup \ddot{Y}$ est similaire à la combinaison $\overline{X \backslash Y} \cup \bar{Y}$. ACD est un motif de taille 3 donc nous pouvons étudier cette nouvelle règle point le support a déjà été vérifié lors de l'extraction des motifs raisonnablement fréquents. Nos étudiants donc la valeur de la confiance. Si cette dernière est valide alors nous calculerons la valeur de M_G :

$$conf(CD \Rightarrow A) = \frac{conf(CD \Rightarrow A) - MAX\left(\frac{1}{2}, \sup(A)\right)}{1 - \max\left(\frac{1}{2}, \sup(A)\right)} = \frac{1 - \max\left(\frac{1}{2}, 0{,}50\right)}{1 - \max\left(\frac{1}{2}, 0{,}50\right)} = \frac{0{,}50}{0{,}50} = 1$$

La confiance de la règle $\ddot{C}D \Rightarrow \ddot{A}$ est valide et est supérieur au support de la conclusion (0,5), par conséquence au calcul M_G :

III. Chapitre 3 : La prédiction par les règles d'associations

$$M_G(\ddot{C}D \Rightarrow \ddot{A}) = \frac{conf(\ddot{C}D \Rightarrow \ddot{A}) - MAX(\frac{1}{2}, \sup(\ddot{A}))}{1 - \max(\frac{1}{2}, \sup(\ddot{A}))} = \frac{1 - \max(\frac{1}{2}, 0{,}50)}{1 - \max(\frac{1}{2}, 0{,}50)} = \frac{0{,}50}{0{,}50} = 1$$

La valeur de M_G est valide : la règle $\ddot{C}D \Rightarrow \ddot{A}$ ou encore $\bar{C}\bar{D} \Rightarrow \bar{A}$ et donc conservée et ajoutée à l'ensemble R.

Pour la combinaison $AD \cup C$, nous étudiants tout d'abord la règle $AD \Rightarrow A$. Comme la confiance de la règle $AD \Rightarrow C$ est valide, et que celle-ci est supérieur au support de la conclusion, nous calculons la mesure M_G :

$$M_G(AD \Rightarrow C) = \frac{conf(AD \Rightarrow C) - MAX(\frac{1}{2}, \sup(C))}{1 - \max(\frac{1}{2}, \sup(C))} = \frac{1 - \max(\frac{1}{2}, 0{,}70)}{1 - \max(\frac{1}{2}, 0{,}70)} = \frac{0{,}25}{0{,}25} = 1$$

La valeur de M_G est valide et la règle $AD \Rightarrow C$ est également conservée et ajoutée à l'ensemble R. Nous passons ensuite à l'étude de la règle $\overline{AD} \Rightarrow \bar{C}$, cependant le motif $\overline{AD}$ n'est pas minimal point par conséquent l'étude de cette règle s'arrête.

Nous cherchons ensuite le nouveau type de règle : $X\ddot{\setminus}Y \Rightarrow \ddot{Y}$. Le motif ACD est de taille 3 donc nous pouvons étudier la règle :

$$conf(\ddot{A}D \Rightarrow \ddot{C}) = \frac{\sup(A\ddot{C}D)}{\sup(A\ddot{D})} = \frac{0{,}25}{0{,}50} = 0{,}50$$

La confiance n'est pas assez élevée donc la règle ne sera pas conservée.

Pour la combinaison $AC \cup D$, nous étudions directement les règles du type $\ddot{X} \cup \ddot{Y}$ car la règle positive tombe dans la zone intéressante et par conséquence M_G=0. Le motif ACD étant de taille 3, nous pouvons donc étudier la valeur de la confiance :

$$conf(\ddot{A}\ddot{C} \Rightarrow \ddot{D}) = \frac{\sup(A\ddot{C}D)}{\sup(A\ddot{C})} = \frac{0{,}25}{0{,}25} = 1$$

Cette dernière est valide, par conséquent nous calculons la valeur de M_G. Comme la confiance de la règle $\ddot{A}\ddot{C} \Rightarrow \ddot{D}$ est supérieur au support de la conclusion (0,75) la règle est dans la zone attractive. Nous utilisons donc la formule :

$$M_G(\ddot{A}\ddot{C} \Rightarrow \ddot{D}) = \frac{conf(\ddot{A}\ddot{C} \Rightarrow \ddot{D}) - MAX(\frac{1}{2}, \sup(\ddot{D}))}{1 - \max(\frac{1}{2}, \sup(\ddot{D}))} = \frac{1 - \max(\frac{1}{2}, 0{,}75)}{1 - \max(\frac{1}{2}, 0{,}75)} = \frac{0{,}50}{0{,}50} = 1$$

La valeur pour M_G étant valide, la règle $\ddot{A}\ddot{C} \Rightarrow \ddot{D}$ est donc conservée et ajoutée à l'ensemble R.

III. Chapitre 3 : La prédiction par les règles d'associations

La prochaine étape va être la recherche des règles possède plusieurs items en conclusion. Nous ajoutons item à la conclusion puis calculons la confiance des règles afin de connaître les types de règles à étudier :

$$conf(D \Rightarrow AC) = \frac{\sup(ACD)}{\sup(D)} = \frac{0{,}25}{0{,}25} = 1$$

$$conf(C \Rightarrow AD) = \frac{\sup(ACD)}{\sup(C)} = \frac{0{,}25}{0{,}75} = \frac{1}{3} \simeq 0.33$$

$$conf(A \Rightarrow CD) = \frac{\sup(ACD)}{\sup(A)} = \frac{0{,}25}{0{,}50} = 0{,}50$$

Seule la $D \Rightarrow AC$ règle possède une confiance supérieure au minimum entre 1 / 2 et le support de sa conclusion D. Les autres règles se situent dans la zone inintéressante :

$$\min\left(\tfrac{1}{2}, \sup(AD)\right) = 0{,}25 \leq conf(C \Rightarrow AD) \simeq 0{,}33 \leq \max\left(\tfrac{1}{2}, \sup(AD)\right) = 0{,}50 \text{ et}$$
$$\min\left(\tfrac{1}{2}, \sup(CD)\right) = 0{,}25 \leq conf(A \Rightarrow CD) = 0{,}50 \leq \max\left(\tfrac{1}{2}, \sup(CD)\right) = 0{,}50 .$$

Par conséquent, nous étudions uniquement les règles $D \Rightarrow AC$ *et* $\overline{D} \Rightarrow \overline{AC}$. Nous commençons par l'étude de la règle $D \Rightarrow AC$. La règle vérifier la contrainte de la confiance. Nous étudions donc la valeur de M_G :

$$M_{Ga}(D \Rightarrow AC) = \frac{conf(D \Rightarrow AC) - MAX(\frac{1}{2}, \sup(AC))}{1 - \max(\frac{1}{2}, \sup(AC))} = \frac{1 - \max(\frac{1}{2}, 0{,}50)}{1 - \max(\frac{1}{2}, 0{,}50)} = \frac{0{,}50}{0{,}50} = 1$$

La règle $D \Rightarrow AC$ est conservée car la valeur de M_G est valide. Nous passons ensuite à l'étude de la règle $\overline{D} \Rightarrow \overline{AC}$ l'étude de cette règle s'arrête car $\overline{AC}$ n'est pas un motif négatif minimal.

Étudiant maintenant le nouveau type de règle pour les trois combinaisons. Commençons par l'étude de la règle $\ddot{D} \Rightarrow \ddot{A}\ddot{C}$. La méta-règle de la confiance MR_9 nous empêche de l'étudier car la règle $\ddot{A}D \Rightarrow \ddot{C}$ n'est pas valide pour la confiance. Comme $\ddot{A}D \Rightarrow \ddot{C}$ ne possède pas une confiance suffisante et que l'inégalité suivante existe :

$conf(\ddot{A}D \Rightarrow \ddot{C}) \geq conf(\ddot{D} \Rightarrow \ddot{A}\ddot{C})$. La confiance de la règle ne peut pas être valide.

Pour la règle $\ddot{C} \Rightarrow \ddot{A}D$, la méta-règle nous autorise à l'étudier car les règles $\ddot{C}D \Rightarrow \ddot{A}$ et $\ddot{A}\ddot{C} \Rightarrow \ddot{D}$ ont une confiance valide. Calculons maintenant la confiance de cette règle :

$$conf(\ddot{C} \Rightarrow \ddot{A}D) = \frac{\sup(\ddot{\ })}{\ddot{\ }} = \frac{0{,}25}{0{,}25} = 1$$

La confiance de la règle est valide. La règle est également dans la zone attractive nous continuons donc en calculant sa valeur pour M_G avec la formule :

$$M_{Ga}(\ddot{C} \Rightarrow \ddot{A}D) = \frac{conf(\ddot{C} \Rightarrow \ddot{A}D) - \max(\frac{1}{2}, \sup(\ddot{C}))}{1 - \max(\frac{1}{2}, \sup(\ddot{C}))} = \frac{1 - \max(\frac{1}{2}, 0{,}25)}{1 - \max(\frac{1}{2}, 0{,}25)} = \frac{0{,}50}{0{,}50} = 1$$

M_G étant valide, la règle $\ddot{C} \Rightarrow \ddot{A}\ddot{D}$ est ajoutée à l'ensemble R des règles valides.

Et enfin pour la règle $\ddot{A} \Rightarrow \ddot{C}\ddot{D}$, notre algorithme s'arrête à la vérification de la méta-règle MR_9 car la confiance de la règle $\ddot{A} \Rightarrow \ddot{C}\ddot{D}$ n'est pas valide.

Le tableau expose les règles extrêmes par notre algorithme.

Règle	Confiance	M_G	Règle	Confiance	M_G
$A \Rightarrow C$	1	1	$AD \Rightarrow C$	1	1
$CD \Rightarrow A$	1	1	$D \Rightarrow A$	1	1
$D \Rightarrow AC$	1	1	$D \Rightarrow C$	1	1
$E \Rightarrow B$	1	1			
$\overline{B} \Rightarrow \overline{E}$	1	1	$\overline{C} \Rightarrow \overline{A}$	1	1
$\overline{A}\,\overline{C} \Rightarrow \overline{D}$	1	1	$\overline{C} \Rightarrow \overline{A}\,\overline{D}$	1	1
$\overline{C}\,\overline{D} \Rightarrow \overline{A}$	1	1			

Tableau 12 Règles extraites sur la base d'exemple par notre algorithme classées par type de règles .

En conclusion, notre algorithme générée 12 règles au total : 7 règles positives $X\backslash Y \Rightarrow Y$, 0 règle négative du type $\overline{X\backslash Y} \Rightarrow \bar{Y}$, 0 règle négative du type $X\backslash Y \Rightarrow \bar{Y}$, 2 règles négatives du type $\overline{X\backslash Y} \Rightarrow \bar{Y}$ et 3 règles du nouveau type $X\ddot{\backslash}Y \Rightarrow \ddot{Y}$. Le chapitre 4 contient une analyse sur les règles extraite afin de comparer cet algorithme aux algorithmes d'Apriori.de [10], de [11] et de [12].

4. Conclusion

Dans ce chapitre nous avons introduit un nouvel algorithme pour extraire plus efficacement les règles d'association positives et négatives d'une base de données. Bien que reposant sur l'algorithme fondateur Apriori, notre approche est différente de celle présentes dans la littérature. Notre algorithme essaye de répondre aux deux problématiques présentes les autres approches, à savoir l'extraction d'un trop grand nombre de règles un inintéressantes, ainsi qu'un parcours non optimisé de recherche des règles.

Notre extraction repose sur un nouveau type de motifs, à savoir les motifs raisonnablement fréquents. L'avantage de ces motifs par rapport aux motifs fréquents repose notamment sur l'élimination des motifs omniprésent. Ces derniers entraînent soit la génération

de règles dans la confiance et invalide, soit la génération de règles dans l'écran à l'indépendance est juger trop faible. Par conséquent, les motifs raisonnablement fréquents permettent d'éliminer dès la première étape de l'algorithme ces deux types de règles inintéressante sans avoir à les étudier. L'utilisation de la mesure M_G, plus sélective que les mesures utilisées par d'autres algorithmes, a également permis d'éliminer un autre type de règles non pertinentes. En effet, les règles possédant un écart trop faible par rapport à l'équilibre sont également éliminées.

Notre algorithme va également cibler les règles potentiellement intéressantes en fonction de la zone d'appartenance de la règle positive $X \Rightarrow Y$ et va donc permettre d'étudier uniquement la moitié des règles. L'utilisation des méta-règle dégage dans le chapitre 3 permet d'insérer la non validité des règles $Y \Rightarrow X$ $\bar{X} \Rightarrow \bar{Y}$ à partir de l'intérêt de la règle $X \Rightarrow Y$? mais également des règles$\bar{Y} \Rightarrow X$ et $\bar{X} \Rightarrow Y$ à partir de l'intérêt de la règle $X \Rightarrow \bar{Y}$.

Le prochain chapitre va présenter les expérimentations que nous avons réalisée sur différentes bases de données afin d'évaluer notre algorithme et le comparer avec les autres approches de la littérature

IV. Chapitre 4 : Expérimentation des résultats

Chapitre 4 : Expérimentation des résultats

1. Introduction

Dans ce chapitre nous présentons la dernière étape, l'étape de réalisation, ainsi que le choix technique utilisé pour le développement de notre application.

2. Les outils utilisés pour la construction du système

2.1.WEKA « environnement Waikato pour l'analyse des connaissances »

C'est l'outil utilisé dans notre expérimentation pour la création de modèle, est une suite de logiciels d'apprentissage automatique écrite en Java et développée à l'université de Waikato en Nouvelle-Zélande. C'est un logiciel libre disponible sous la Licence publique générale GNU (GPL). WEKA est un ensemble de classes et d'algorithmes en Java implémentant les principaux algorithmes de Fouille de données. Il est disponible gratuitement à l'adresse [20]

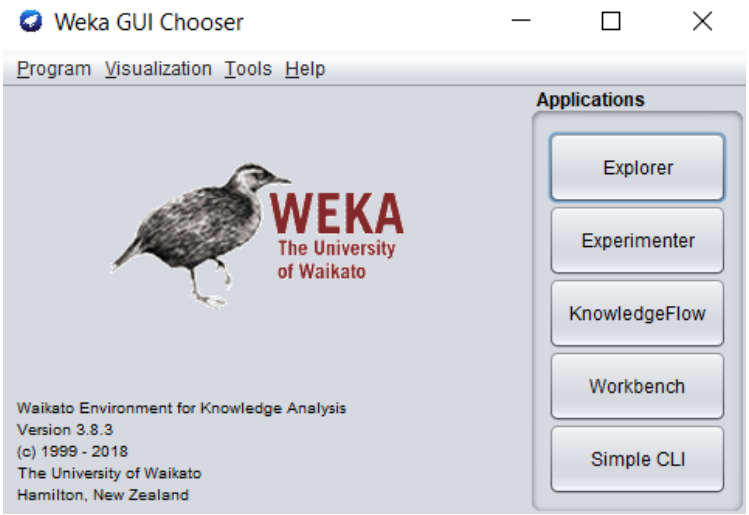


Figure 11: acronyme pour Waikato environment for knowledge analysis (weka).

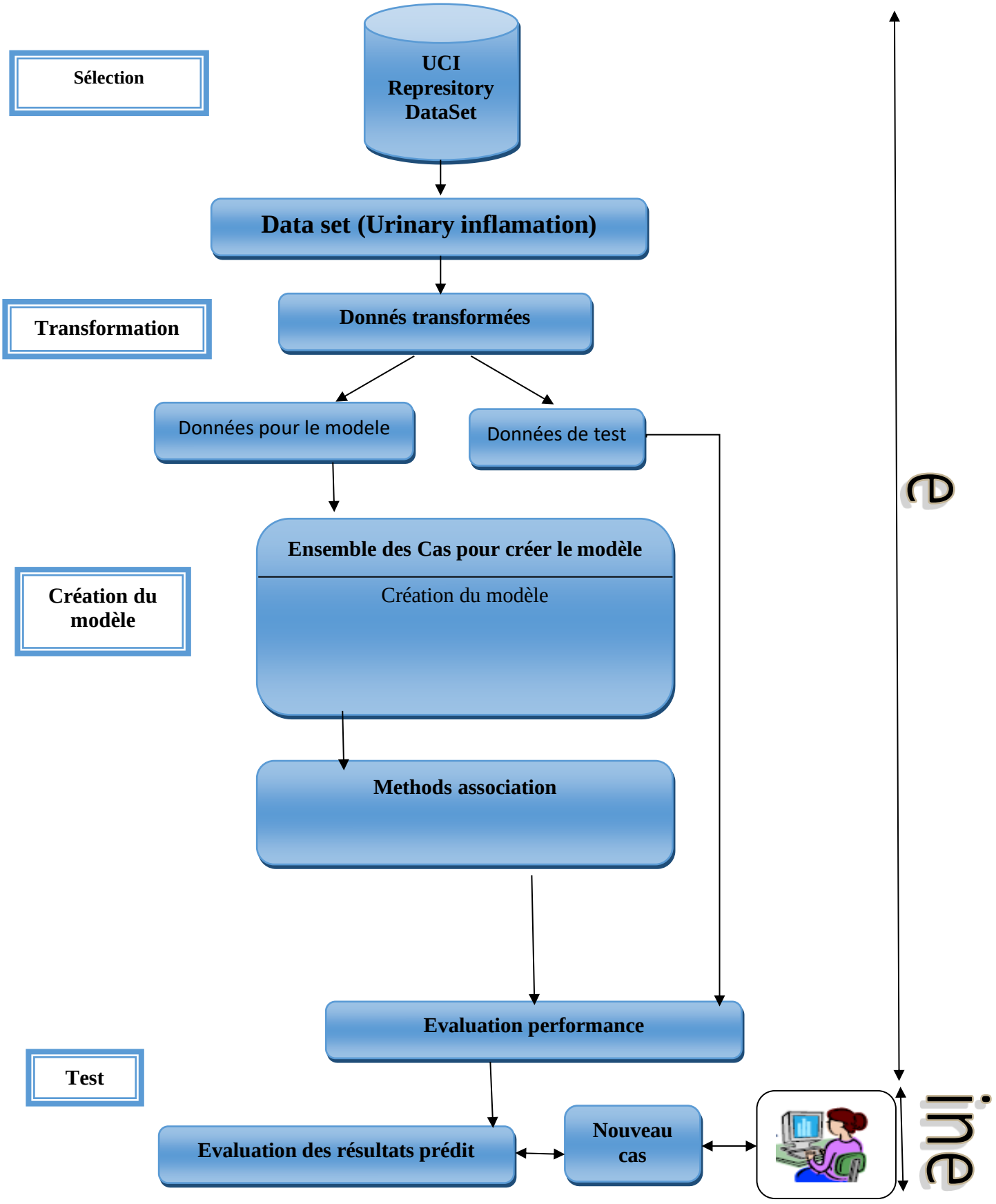
2.1. React

est une bibliothèque JavaScript pour la construction d'interfaces utilisateur (UI). Pour découvrir à quoi sert React. [21]

Figure 12 : logo de react

IV. Chapitre 4 : Expérimentation des résultats

3. Architecture du système



4. Choix de donnée

L'idée principale de cet ensemble de données est de préparer l'algorithme du système expert, qui effectuera le diagnostic présomptif de deux maladies du système urinaire. Ce sera l'exemple du diagnostic de l'aigu inflammations de la vessie et néphrites aiguës. Pour mieux comprendre le problème, considérons les définitions des deux maladies données par les médecins. L'inflammation aiguë de la vessie est caractérisée par l'apparition soudaine de douleurs dans la région de l'abdomen et la miction sous forme de poussée d'urine constante, de douleurs mictionnelles et parfois d'un manque de rétention d'urine. La température du corps augmente, mais le plus souvent pas au-dessus de 38°C. L'urine excrétée est trouble et parfois sanglante. Avec un traitement approprié, les symptômes disparaissent généralement en quelques jours. Cependant, il y a une tendance aux retours. Chez les personnes atteintes d'une inflammation aiguë de la vessie, il faut s'attendre à ce que la maladie se transforme en forme prolongée. La néphrite aiguë d'origine du bassinet du rein survient beaucoup plus souvent à femmes que chez les hommes. Elle débute par une fièvre soudaine, qui atteint, et dépasse parfois 40C. La fièvre s'accompagne de frissons et d'un ou douleurs lombaires bilatérales, parfois très fortes. Symptômes de inflammation aiguë de la vessie apparaissent très souvent. Pas du tout rarement il y a des nausées et des vomissements et des douleurs généralisées abdomen.

Les données ont été créées par un expert médical en tant qu'ensemble de données pour tester le système expert, qui effectuera le diagnostic présomptif de deux maladies du système urinaire. La base de la détection des règles était Rough Sets La théorie. Chaque instance représente un patient potentiel. Les données sont dans un fichier ASCII. Les attributs sont séparés par TAB. Chaque ligne du fichier de données commence par un chiffre indiquant la température du patient.

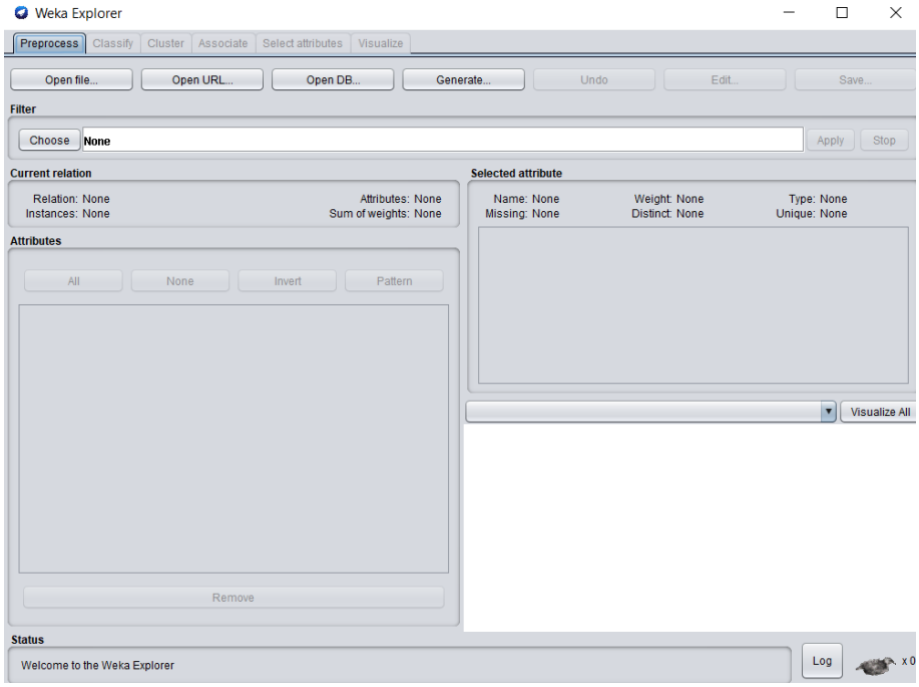
Les données :

1	35.5	no	yes	no	no	no	no	nc
2	35.9	no	no	yes	yes	yes	yes	nc
3	35.9	no	yes	no	no	no	no	nc
4	36.0	no	no	yes	yes	yes	yes	nc
5	36.0	no	yes	no	no	no	no	nc
6	36.0	no	yes	no	no	no	no	nc
7	36.2	no	no	yes	yes	yes	yes	nc
8	36.2	no	yes	no	no	no	no	nc
9	36.3	no	no	yes	yes	yes	yes	nc
10	36.6	no	no	yes	yes	yes	yes	nc
11	36.6	no	no	yes	yes	yes	yes	nc
12	36.6	no	yes	no	no	no	no	nc
13	36.6	no	yes	no	no	no	no	nc
14	36.7	no	no	yes	yes	yes	yes	nc
15	36.7	no	yes	no	no	no	no	nc
16	36.7	no	yes	no	no	no	no	nc

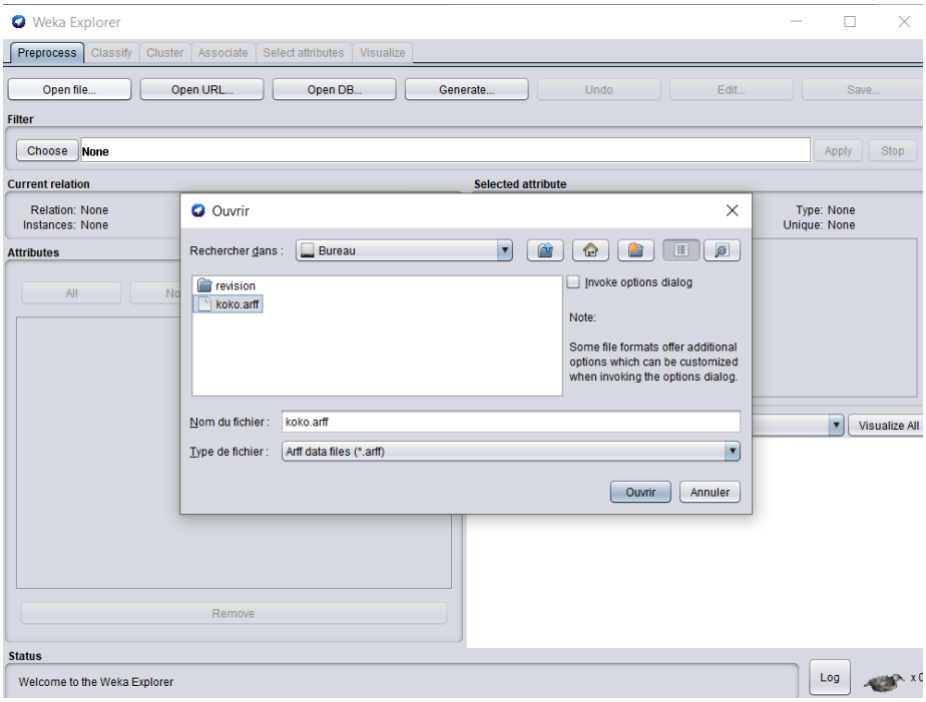
Figure 13: Echantillon des données de fichier data Set.

Ces données vont être importées au WEKA avec ses 8 attributs :

- Ouvrir le fichier (open file) :



- Sélectionner le fichier



➢ L'importation :

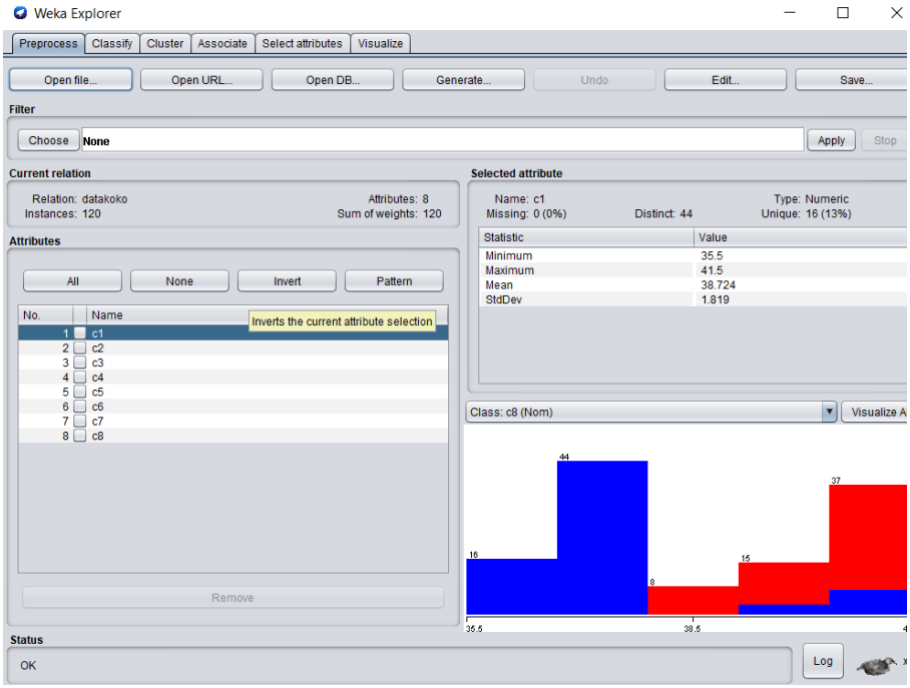
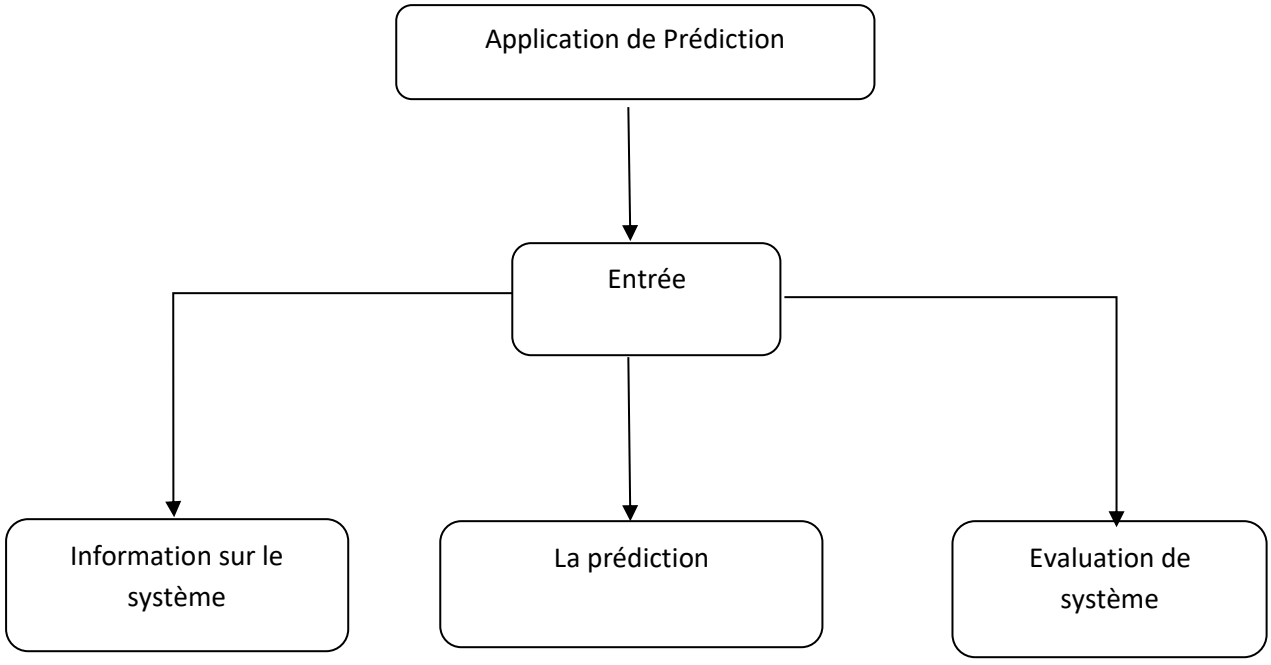


Figure 14 : importation de fichier Data Set. CSV au WEKA .

5. Application

5.1.Présentation fonctionnelle



5.2.Les différentes classes

Page d'Accueil : C'est la fenêtre principale qui réunit toutes les classes du système, elle contient 4 fenêtres et 2 boutons pour la validation et l'Ajout d'un Fichier

Dans la grande fenêtre s'affichent les attributs et data importés de WEKA .

Figure 15 : fenêtre principale du système .

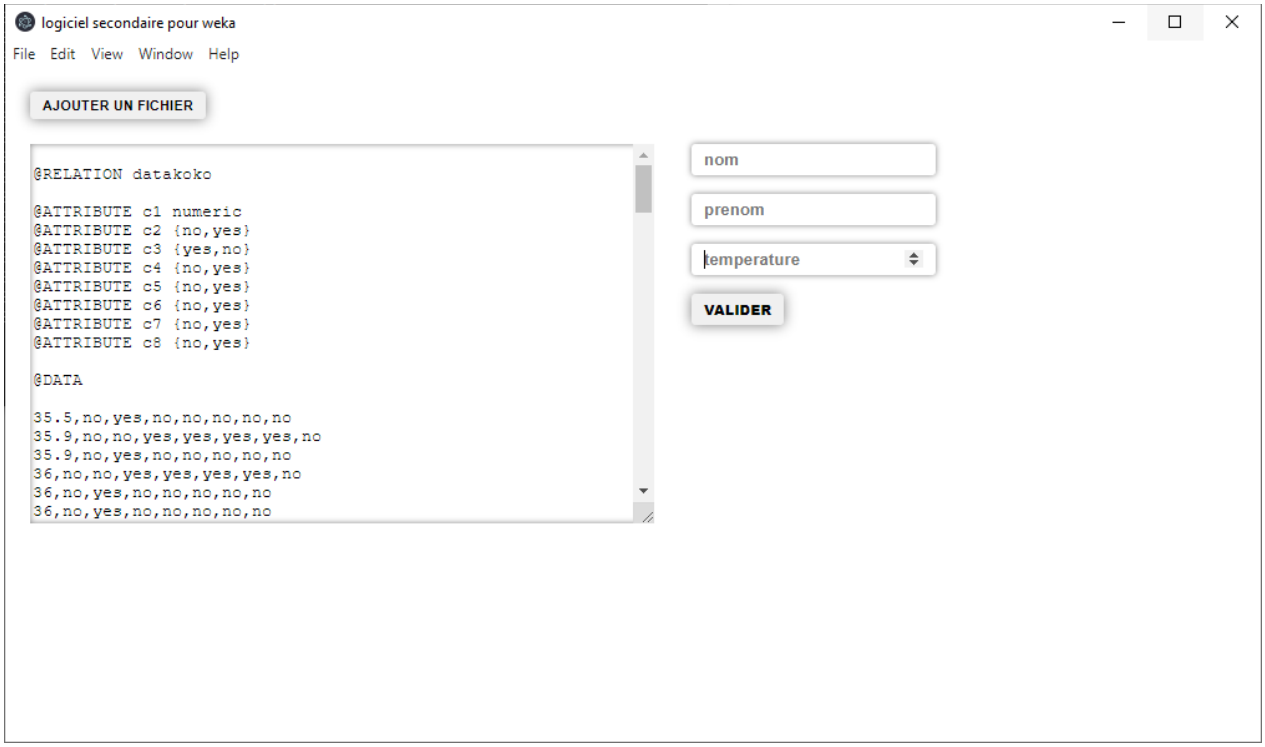


Figure 16 : ajout d'un fichier .

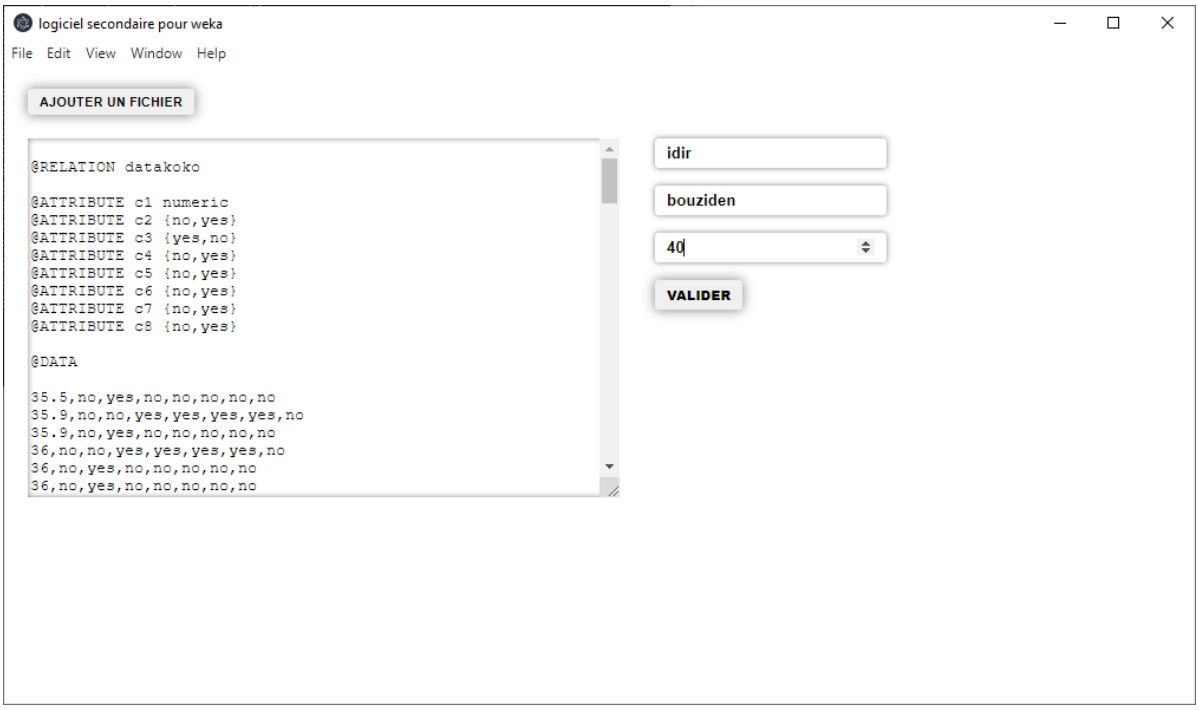


Figure 17 : les information de malade.

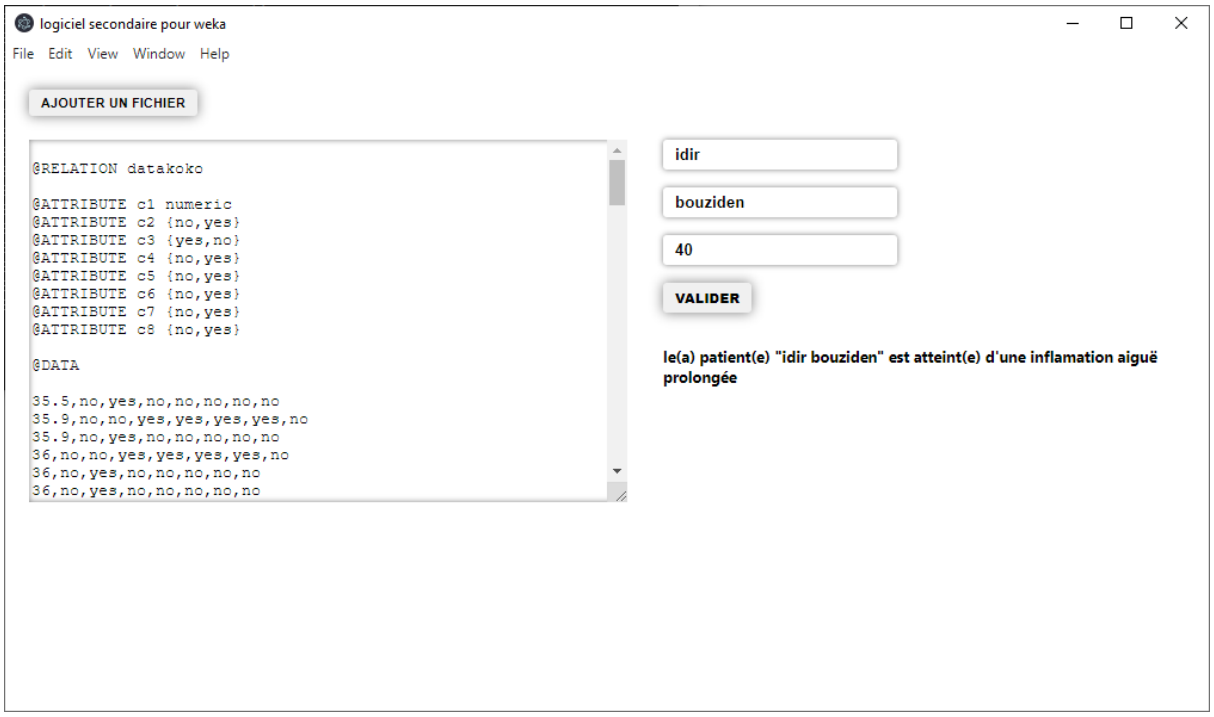


Figure 18: validation et résultat .

6. Conclusion

Dans ce chapitre, nous avons défini un système qui permet de prédire les symptômes d'une maladie urinaire qui est souvent bénin tant qu'il n'y a pas de fièvre ,de douleurs intenses, de nausées, et de vomissements. Une température élevée est mauvais signe car la vessie et les reins pourraient être touchés. Dans ce cas la maladie peut se transformer de la forme aigue à la forme prolongée

Notre résultat d'expérimentation basé sur les techniques de Fouille de données montre que notre système est performant.

Conclusion générale

L'extraction de connaissances à partir des données est le processus de découverte de connaissances utiles à partir d'un jeu de données. Ce processus se décompose en plusieurs étapes mais nous nous sommes intéressés uniquement, dans ce manuscrit, a l'étape qui consiste à extraire les connaissances : la fouille de données. Les connaissances extraites peuvent prendre plusieurs formes, mais nous nous sommes concentrés sur les règles d'association positives et négatives.

Les algorithmes d'extraction de règles d'association positives et négatives de la littérature génèrent un nombre important de règles inintéressantes en utilisant un parcours de recherche des règles non optimise. Afin de combler ces deux failles, nous avons proposé dans ce mémoire une nouvelle approche basée sur l'algorithme fondateur Apriori pour extraire des règles d'association positives et n´ergatives intéressantes. Notre algorithme a également la particularité d'extraire un nouveau type de règles négatives que les autres algorithmes ne recherchent pas : les règles possédant des conjonctions d'items négatifs en prémisse et en conclusion. Par conséquent, en plus de rechercher les règles $X \Rightarrow Y$, $Y \Rightarrow X$ comme le font les méthodes classiques, nous extrayons également les règles : $x_1..x_P \Rightarrow y_1..y_P$.

Les apports de notre approche

Afin de réduire le nombre de règles inintéressantes, nous avons choisi de ne plus baser notre extraction sur les motifs fréquents comme le font les autres méthodes que nous avons pu étudier dans ce manuscrit. Ce choix se justifie par la possible présence de motifs omniprésents dans les bases de données à analyser. Ces motifs omniprésents vont conduire à des règles non valides pour la confiance, a des règles trop proches de l'indépendance ou encore à des règles redondantes. Par conséquent, ces motifs omniprésents vont amener, dans le cas où l'extraction est possible, à extraire des règles inintéressantes. Notre premier objectif fut donc de supprimer ces motifs omniprésents de l´étude. Pour se faire, nous avons proposé de baser notre extraction sur les motifs raisonnablement fréquents qui permettent d'écarter les motifs omniprésents d`es la première phase de l'extraction.

Nous avons également ajoute une autre contrainte lors de la recherche des motifs raisonnablement fréquents : le support minimum pour les conjonctions d'items négatifs. Cette contrainte supplémentaire renforce la contrainte du support maximum et permet d'extraire des motifs XY plus intéressants. C'est ´également l'ajout de cette contrainte qui nous a permis de découvrir les règles intéressantes du nouveau type.

Une fois les motifs raisonnablement fréquents extraits, la seconde étape a consisté `à générer les règles à partir de ces motifs. Afin de connaitre les types de règles à étudier, nous avons utilisé la mesure MG. La mesure MG commence par déterminer la zone d'appartenance de la règle positive en comparant sa confiance au support de sa conclusion. Cette étape nous a indiqué si nous devions étudier les règles $X \Rightarrow Y$ et $Y \Longrightarrow X$ou bien si nous devions étudier les règles $\bar{X} \Rightarrow \bar{Y}$ et $\bar{Y} \Rightarrow \bar{X}$. Cette mesure nous a permis donc d'optimiser le parcours de recherche puisque comme nous l'avons démontré seule la moitié des règles peuvent être intéressantes. Cette mesure possède un autre avantage puisqu'elle permet d´élaguer un autre type de règles inintéressantes : les règles trop proches de l´équilibre.

Afin d'optimiser le parcours de recherche des règles, nous avons utilisé la mesure M_G comme nous l'avons expliqué précédemment mais nous avons également eu recours `a deux métarègles afin de d´enduire l'intérêt d'une règle à partir d'une autre. La première métarègle MR_9 correspond à la

propriété de la confiance. Cette propriété abandonnée par les autres méthodes, nous a permis d'élaguer certaines règles du nouveau type. La deuxième métarègle MR4 permet d'inférer la non validité des règles $X \Rightarrow Y$ et $Y \Rightarrow X$ à partir des règles $\bar{X} \Rightarrow \bar{Y}$ et $\bar{Y} \Rightarrow \bar{X}$respectivement.

Lors de la recherche des règles négatives, nous avons ajoute une contrainte afin de s'assurer de l'intérêt de la règle extraite. En effet, nous avons vérifié que la règle est minimale. Pour se faire, après avoir extrait les motifs raisonnablement fréquents, nous avons recherche les motifs négatifs minimaux raisonnablement fréquents. Cette étape intermédiaire, nous est utile lors de la recherche des règles ou nous vérifions que chaque prémisse et conclusion négatives sont bien présentes dans la liste des motifs minimaux raisonnablement fréquents.

Bibliographie

[1] J.Han, M.Kamber. « Data Mining: Concepts and Techniques », Morgan Kaufmann Publishers, 2000.

[2] A.Famili, M.Shen, R.Weber, E. Simoudis. « Data Preprocessing and Intelligent Data Analysis ». Intelligent Data Analysis, Vol. 1, P. 3-23, 1997.

[3] A. Silberschatz, A.Tuzhilin. « On Subjective Measures of Interestingness in Knowledge Discovery ». In proceedings of KDD-95. First International Conference on KDDM, PP. 275-281, Monlo Park, CA: AAAI Press, 1995.

[4] M.Antonie, and Zaiane, O. R. (2004). Mining positive and négative association rules : an approach for confined rules. In Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD, pages 27–38.

[5] J.Bykowski, Boulicaut, & A.Rigotti. «Free-Sets: A Condensed Representation of Boolean Data for the Approximation of Frequency Queries. DMKD 7, (1), 5-22.» 2003.

[6] C.Yan, Cornelis, P.Zhang, and G.Chen. (2006). Mining positive and negative association rules from large databases. In Cybernetics and Intelligent Systems, CIS, pages 613–618.

[7] D.Calvanese, G.De Giacomo, M.Lenzerini, D.Nardi et R.Rosat1. « Data Integration in Data Warehousing ». Int. Journal. Of Cooperative Information Systems, Vol.10, P. 237-271, 2001. s.d.

[8] D.Pyle. « Data Preparation for Data Mining ». Morgan Kaufmann Publishers, 1999.

[9] M.Zaoui «Fouille des règles d'association guidée par des ontologies et des schémas de règles.» 2011.

[10] G.Gardrin. « Internet / Intranet et Bases de Données: Data Web, Data Media, Data Warehouse - Data Mining », Eyrolles, 2000.

[11] S.Guillaume and Papon, P.-A. (2013a). ´ Etude comparative d'extraction de règles d'association positives et négatives et optimisations. In Apprentissage Artificiel et Fouille de Données, Revue des Nouvelles Technologies de l'Information, vol. RNTI-A.6,.

[12] S.Guillaume and Papon, P.-A. (2013b). Extraction optimisée de règles d'association positives et négatives (RAPN). In Actes de la 13`eme Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, Revue des Nouvelles Technolog.

[14] R.Brachman, T. Anand. « The Process of Knowledge Discovery in Databases: A Human-Centered Approach ». In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, (eds.), Adv. in KDDM, MIT Press, Cambridge, 1996.

[15] R.Gilleron, M. Tommas1. « Découverte de Connaissances à partir de Données. Notes de cours », IUP MIAGE, Lille, Octobre 2002.

[16] U.Fayyad, G. Piatetsky-Shapiro, P. Smyth. « From Data Mining to Knowledge Discovery: An Overview ». In U. Fayyad, G. P. Shapiro, Amith, P. SmythR. Uthurusamy, (eds.), Adv. in KDDM, MIT Press, 1-36, Cambridge, 1996.

[17] U.Fayyad, G. Piatetsky-Shapiro, P. Smyth. « The KDD Process for Extracting Useful Knowledge from Volumes of Data ». Communications of the ACM,5. 39, no. 11, pp. 27-34, November 1996.

[18] U.Fayyad. « Data Mining and Knowledge Discovery: Making Sense Out of Data ». IEEE Expert, 5. 11, no. 5, pp. 20-25, October 1996.

[19] W.Zhang, C.Zhang, S. (2004). Efficient mining of both positive and négative association rules. Transactions on Information Systems (TOIS), 22(3) :381–405.

[20] Logicel REACT *https://fr.reactjs.org/docs/getting-started.html*.

[21] Logiciel WEKA *www.cs.waikato.ac.nz/ml/weka*.

Listes des Tableaux

Listes des figures