UNE CONTRIBUTION AU POST-TRAITEMENT DES REGLES D'ASSOCIATION DANS UN CONTEXTE DE DATA MINING

Nora Lounici,
Ahmed Zakane
Laboratoire de Statistique Appliquée (LASAP) **ENSSEA**
noralounici@yahoo.fr

Résumé. La fouille de données est un domaine de recherche de prédilection, à l'origine de nombreuses publications. L'extraction de règles d'association utiles à partir de données volumineuses constitue une des tâches innovantes qui a fait découvrir au monde industriel le domaine du Data Mining. Nous proposons d'abord une présentation synthétique du processus d'ECD et la place qu'occupe le Data Mining dans un tel processus. Nous expliquons, ensuite, brièvement, la démarche globale que nous proposons. Nous nous intéresserons tout particulièrement au problème du nombre de règles d'association prohibitives générées par les algorithmes de type Apriori. Les techniques employées posent deux problèmes majeurs : l'un algorithmique pour la génération de règles, et l'autre qualitatif pour l'évaluation et la validation des règles. Notre proposition s'appuie, en partie, sur l'évaluation objective de l'intérêt des règles à l'aide de mesures de qualité. Afin de classer les règles d'association par intérêt décroissant, une des solutions consiste à utiliser différentes mesures de qualité. Néanmoins, ces mesures renvoient à des résultats différents et parfois conflictuels. Notre approche consiste dans un premier temps à sélectionner un sous-ensemble de mesures complémentaires à partir d'une liste de 25 mesures. Le but est d'effectuer un classement des règles d'association, par intérêt sur la base d'un sous-ensemble de mesures établies. Cette approche est illustrée sur un exemple.

Mots clés : *Fouille de données (Data Mining), ECD, Règles d'association, post-traitement, mesures de qualité.*

1. Introduction

Les avancées technologiques fulgurantes de l'informatique, ayant entrainé des réductions de coûts et l'augmentation des performances, ont permis la collecte de volume de données de plus en plus important[1]. Afin de tirer profit de ces ″gisements d'information″, les techniques d'extraction de connaissances à partir de données (notée ECD), [en anglais Knowledge Discovery in Database (KDD)[2]] se sont développées. L'objectif est de découvrir des connaissances à partir de données brutes. Le concept d'ECD est souvent confondu avec la fouille de données [en anglais Data Mining (que nous notons DM)]. En fait, les deux procédés, fouille de données et ECD ont pour objectif commun : l'extraction des connaissances à partir de volumes importants de données. Classiquement l'ECD est un processus qui se déroule en plusieurs étapes (voir section 2), et au centre de ce processus se situe le DM.

Les méthodes et techniques de DM sont multiples et variées. Parmi les travaux de recherche en DM, l'extraction des règles d'association est indéniablement la méthode la plus novatrice, qui a captivé le monde des chercheurs et pour laquelle beaucoup de travaux ont été consacrés. Cette technique produit des motifs fréquents, à partir desquels est extrait un ensemble de règles. Les règles formulent des associations latentes entre les attributs d'une base de données.

1. *On estime que le volume de données stockées de par le monde double tous les ans.*
2. *Le terme KDD (Knowledge Data Discovery) , en français ECD (l'extraction des Connaissances à partir de données)* a été introduit par Piatetsky-Shapiro lors du premier workshop KDD en 1991.

Plus généralement étudiée dans un cadre d'apprentissage non supervisé, la recherche de règles d'association est probablement la méthode de DM qui a contribué le plus à faire connaître à la communauté des chercheurs et au monde industriel l'importance de cette « nouvelle vision » que l'on pourrait avoir des statistiques. Ce qui rend cette technique attrayante est que les besoins sont énormes, elle est utilisé aussi bien par des entreprises que par des administrations : banques, assurance, commerce, grande distribution, hôpitaux, …etc.

Le but est de chercher à dériver des connaissances, afin de mieux comprendre les liens qui existent entre ses données. Depuis sa première formulation et la proposition des algorithmes AIS et Apriori[1] par [Agra & al 1993], [Agra & Srik 1994], [Agra & al 00], ce problème a suscité beaucoup d'intérêt. Cependant, si les algorithmes *type Apriori* ont donné des résultats intéressants, ils présentent l'inconvénient majeur de générer de façon aveugle un trop grand nombre de règles dont la plupart sont soit redondantes, soit inintéressantes. Quand le nombre de règles d'association extrait est trop grand, allant jusqu'à égaler la quantité des données acquises à l'origine, on se retrouve face à un problème épineux, à savoir comment extraire la connaissance de cette masse de règles inintelligibles. Dans de pareilles situations, une phase post-traitement dotée d'outils et techniques efficaces devient une solution incontournable.

Cet article porte sur l'évaluation de la qualité des règles d'association. L'une des solutions consiste à évaluer et ordonner les règles par des mesures de qualité objectives[2] afin de placer les règles les plus dominantes en tête de classement. Classiquement, les règles sont estimées sur deux mesures de qualités : le Support et la confiance, qui ont été très critiquées car elles ne permettent pas d'évaluer correctement la solidité d'une règle. En complément de ces deux mesures, d'autres mesures alternatives ont été introduites. Nous examinons dans le cadre de cette étude, le comportement d'un certain nombre d'entre-elles. Une analyse statistique est menée sur ces mesures pour retrouver les éventuelles corrélations. Ce qui va nous permettre de nous concentrer sur un sous-ensemble de mesures complémentaires.

L'étude des propriétés du sous-ensemble de mesures nous mène naturellement à opter pour une mise en application d'une analyse d'aide multicritères et faire ainsi participer l'utilisateur métier au processus de choix et lui permettre ainsi d'exprimer ces préférences. L'idée consiste à pouvoir faire ressortir du lot de règles inintelligibles, celles qui présentent le meilleur compromis selon les mesures retenues.

Dans ce contexte, nous proposons une démarche opérationnelle pour tenter d'extraire les règles les plus informatives. Notre approche *est indépendante du contexte applicatif et des connaissances* que l'on peut avoir du domaine. Elle répond à une problématique générale. [On s'intéressera tout particulièrement à certaines propriétés formelles des mesures afin de mener une analyse d'aide à la décision multicritère. L'extraction des règles les plus pertinentes selon cette approche, *le deuxième objectif de nos travaux de recherche*, n'est pas traitée dans cet article et fera l'objet d'une autre publication].

Comme nous le verrons en section 2, un certain nombre de mesures de qualité, pratiques pour filtrer les règles sont présentées. Le but recherché consiste à ne considérer que les mesures "vraiment" complémentaires.

---

1 Il existe plusieurs algorithmes de résolution de fouille de règles d'association, on cite notamment, AIS, APRIORI, APRIORITID [[Agra & Srik 1994], *Max-Miner [Bay1998] CLOSE [Pas & al 1999], etc.*

2 *Les mesures objectives sont axées sur les données, alors que les mesures subjectives, elles intègrent les connaissances du décideur*

Afin de comparer les résultats obtenus et extraites un premier ensemble de règles exploitables, nous nous appuyons sur la mesure de précision. Pour calculer la précision, nous avons défini *un seuil de pertinence* lié à la nature des règles fournies par chaque mesure.

Cette procédure a un double intérêt. Tout d'abord, elle produit des règles générales sur un nombre restreint de mesures qui sont par la suite affinées par cette seconde étude, ce qui nous permettra d'effectuer une analyse multicritère plus simplement, et faire ainsi participer l'utilisateur métier au processus d'extraction final.

Cet article comprend 4 sections :

La **section 2** introduit le processus d'extraction de connaissances à partir de données, ainsi que la place qu'occupe le DM dans un tel processus. La **section 3** est consacrée à la formulation des règles d'association et plus particulièrement à l'évaluation des règles par les mesures de qualité. La **section 4** décrit la démarche adoptée, illustrée sur un exemple. Enfin, nous terminons la **section 5** en ouvrant les perspectifs qu'offre ce travail.

2. Extraction des connaissances :

L'accumulation des données durant cette décennie a motivé l'émergence d'un nouveau domaine de recherche : L'extraction des connaissances à partir des données (ECD). L'ECD est définie, comme « le processus de *découverte non triviale de connaissances inconnues, valides, compréhensibles et potentiellement utiles à partir de données stockées dans les bases de données* » *("The non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.")* [Fay & al 1996].

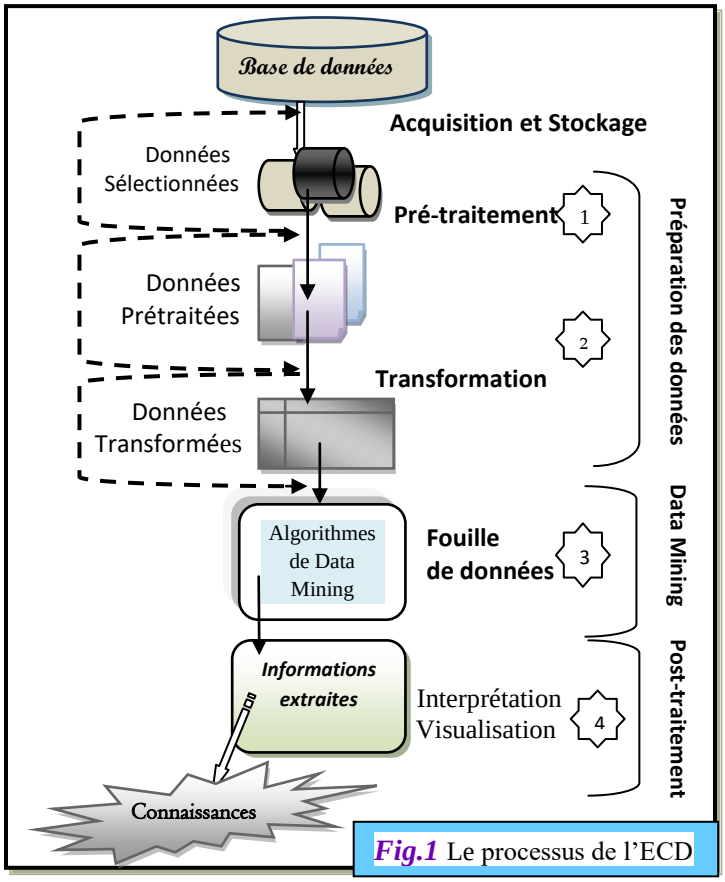


Fig.1 Le processus de l'ECD

La méthodologie d'ECD a été introduite sous l'appellation KDD (Knowledge Discovery in Database) par Piatetsky-Shapiro lors du premier workshop KDD en 1989 [Piat 1991]. L'ECD est un processus complexe ″interactif et itératif ″, et concerne, par nature, *de grandes bases de données*. Il comprend plusieurs phases, dont certaines sont classiques en statistique **:** la préparation des données ou pré-traitement (*pre-processing)*. Le processus aboutit à la construction de modèles, de découverte de nouvelles corrélations et de nouvelles relations pour décrire et /ou expliquer un phénomène ou faire des prédictions, c'est précisément à ce niveau là qu'intervient le DM qui est l'étape motrice de l'ECD.

Classiquement, on divise le processus d'ECD en quatre grandes phases (cf. ***Fig.1*** ci-dessus)
– *Entreposage des données* : Ciblage et Acquisition des données (Data WareHousing)
– *Pré-traitement* : la transformation des données (Pre-Processing)
– *La fouille des données (Data Mining)*
– *Evaluation et déploiement* (Post-Processing)

L'étape de fouille de données consiste à produire des modèles explicatifs ou prédictifs des données ou à rechercher les motifs[1]. A l'issue de cette étape, les résultats produits par les algorithmes de fouille de données ne sont toujours pas exploitables. Il est alors nécessaire de les soumettre à une évaluation dans une perspective d'exploitation.

2.1 Le Data Mining

Dans les années 90, la recrudescence d'informations variables et dynamiques dans les bases de données a occasionné le développement du domaine de l'ECD. Cette situation d'étranglement informationnel mondiale due à la profusion d'informations stockées sur différents supports , internes (BD) ou externes (Internet) a été un véritable casse tête pour les spécialistes, qui traditionnellement étudiés périodiquement les données de l'entreprise pour prendre connaissance des tendances du moment. Ils produisaient ainsi des rapports réguliers au décideur.

Cette forme d'investigation manuelle est devenue impossible pour plusieurs raisons :

- Elle est très coûteuse en temps, dû essentiellement au nombre croissant d'analyse.
- On ne dispose pas forcement toujours d'un spécialiste à portée de main.
- Les spécialistes qui au départ arrivaient à formuler des hypothèses réalisables se trouvent totalement dépassés par le flux de données.

Face à cette surcharge informationnelle et dans un contexte fortement concurrentiel, les techniques anciennes s'effondrent. Pour une gestion plus stratégique de l'entreprise, il est devenu impératif de mettre en œuvre des processus d'analyse nouveaux pour faire face aux défis du moment, en proposant aux décideurs, grâce aux moyens informatiques de plus en plus performants, des outils appropriés leur permettant d'extraire de l'information pertinente de ces masses de données et en temps réel.

Ces données confinées dans les systèmes de production sont considérées comme des « data jails » littéralement « prison de données » par *Meta Group*. Par conséquent, l'idée de traduire ces données en connaissance afin d'en tirer le meilleur profit s'est imposée par elle-même.

[1].*Un Motif (en anglais pattern) ou patron est un ensemble d'attributs (Items).*

Néanmoins, plusieurs facteurs ont contribué à promouvoir cette intuition. En effet, le développement des capacités de stockage, couplé à l'évolution des machines et des logiciels, ainsi que la prolifération des réseaux ont inéluctablement conduit à l'émergence de la fouille de données.

C'est dans la grande distribution que se sont déroulés les premiers essais historiques. L'étude se place dans le contexte de la gestion de la relation client. Pour fidéliser la clientèle, une analyse du comportement du client face aux produits proposés a permis de mettre en évidence des relations cachées entre certains produits. Les distributeurs ont ainsi procédé à un meilleur agencement des rayons, ce qui a fait booster les ventes de manière conséquente.

Le *Data Mining (DM)* traduit fidèlement de l'anglais *par fouille ou forage des données* est un domaine récent, même si ses fondements méthodologiques sont antérieurs. Né aux Etats-Unis en 1989 lors du workshop sur le KDD [Piat 1991], le terme *data mining* a été approuvé officiellement pour la première fois en 1991. Cependant, il faudra attendre 1995 pour que les premières conférences internationales sur le sujet soient tenues et ce n'est qu'en 1997 qu'a eu lieu le premier séminaire Européen.

Même si les premiers résultats étaient loin d'être aberrants, cette discipline a, à ses débuts éveillé beaucoup de curiosité et également des appréhensions, notamment par la communauté des statisticiens qui considèrent la démarche comme peu scientifique. Ce comportement diatribe a froissé les informaticiens, dont l'objectif premier été de valoriser les données dormantes dans les bases de données en offrant des perspectives nouvelles d'extraction de connaissances.

En termes de positionnement, la fouille de données ne représente qu'une étape d'un processus complexe d'ECD, qui englobe, la préparation des données, l'exploration des données et l'interprétation des résultats. Le DM est formé de l'adhésion d'autres disciplines préexistantes, telles que les bases de données, les statistiques, l'apprentissage automatique, les outils de visualisation, l'intelligence artificielle …etc.

La fouille de données ne peut pas être vue comme une agrégation de méthodes hétérogènes et bien connues. C'est un domaine de recherche établit, clairement identifié et confirmé. Il l'a été, notamment quand furent proposées certaines méthodes innovantes adaptées aux bases de données de grandes tailles et qui ne sont rattachées à aucunes techniques déjà existantes, telles que *les règles d'association.*

Les problèmes de recherche des Items fréquents et d'extraction des règles d'association pertinentes seront détaillés dans la suite de cette communication qui s'inscrit dans cette thématique de recherche.

Enfin, l'association entre le *DM* et les statistiques est à notre avis l'ultime recours pour les deux communautés (Statisticiens et Informaticiens), notamment lorsqu'on comprend que la spécificité des applications du *DM* concerne l'utilisation de grandes bases de données, plus en phase avec les réalités actuelles du monde industriel.

2.1 Présentation du Processus de Data Mining

La fouille de données se définit comme un ensemble d'outils regroupés autour d'une problématique, dont les principales tâches sont:

- La recherche de règles d'association.
- les méthodes de Segmentation (clustering)
- les méthodes de classification supervisée (discrimination)

La figure ***Fig.2*** présente les tâches de la fouille de données et quelques algorithmes associés.

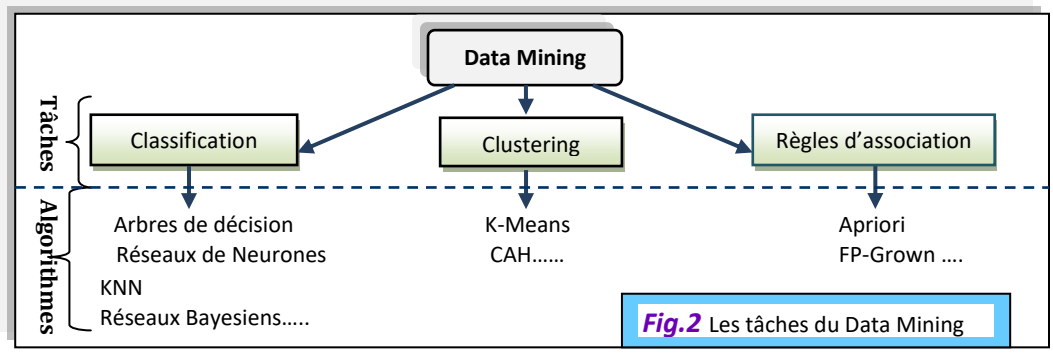


Fig.2 Les tâches du Data Mining

On peut regrouper les objectifs des méthodes de la fouille de données suivant deux approches :

- *approche supervisée :* Etant donné un fichier décrivant des données, associant une description à une classe, on cherche à affecter des individus, qui possèdent des caractéristiques communes, à des classes préétablies.

- *approche non supervisé* : Dans cette approche, les classes ne sont pas obligatoirement connues à priori. L'objectif est de trouver des liens de causalité entre différents attributs dans le but de diviser les données en catégories.

Le choix d'une approche dépendra de la nature du problème, des objectifs visés et du type de données ciblées.

Une des techniques largement étudiée en fouille de données [Brin & al 1997], [Han & Kam 01] et qui rentre dans le cadre de l'apprentissage non supervisé est ***l'extraction de règles d'association***. Originellement introduite par [Agra & al 1993], l'objectif est de trouver des relations surprenantes et/ou des ″régularités ″ intéressantes (comportements similaires), pouvant apporter plus de connaissances au domaine étudié. Nous approfondissons particulièrement ci-dessous le point concernant la technique d'extraction de règles d'association.

3. Etat de l'art des règles d'association

En Intelligence Artificielle, de nombreuses théories de représentation des connaissances sont fondées sur les règles [Kay 1997]. Cette structure implicative est de la forme **:** " Si *prémisse* Alors *Conclusion* " (notée prémisse → *Conclusion)*. En DM, une des principales méthodes produisant des connaissances sous forme de règles est *l'extraction de règles d'association*, qui a pour avantage de représenter les connaissances de manière explicite et donc facilement interprétables par l'utilisateur.

Ce concept initialement introduit par [Agra & al 1993] et testé sur des bases de données (BD) de transactions de ventes, pour mieux comprendre les besoins des clients dans les activités de la grande distribution, s'est étendu par la suite à d'autres secteurs d'activité tels que le management des entreprises, la biologie, le marketing, la robotique, etc.

Une règle d'association est représentée par la relation d'implication probabiliste du type $a \rightarrow b$, *où a et b* représentent respectivement des ensembles d'attributs disjoints. La règle $a \rightarrow b$ exprime une association entre *a* et *b*. En d'autres termes, étant donné un ensemble d'attributs, le but est de repérer dans la base si l'occurrence de l'ensemble *a* pour un individu donné est associée à l'occurrence de l'ensemble *b*.

En général le processus de fouille s'appuie sur la recherche des Items fréquents pour générer *les règles d'association*.

3.1 Notation et concepts de base

On dispose d'un tableau T que nous allons présenter à l'entrée d'un processus de fouille de données. Le tableau est composée de *N* enregistrements ou individus $\{e_1, e_2, e_3 \ldots, e_n\}$ d'un ensemble *E,* décrits par m variables *binaires* $I=\{i_1, i_2, i_3, \ldots, i_m\}$ appelées ***items,*** qui précise la présence (codée 1)/ l'absence (codée 0) de chaque *Item* dans T. La conjonction d'un ensemble d'Items qui est aussi une variable booléenne est appelée ***ItemSet*** ou ***motif***. Ce formalisme peut être appliqué à toute table de type individus/variables préalablement transformée sous une forme disjonctive complète. A partir de cette matrice nous cherchons à extraire des règles d'association.

Définition1 : ***Une règle d'association*** est une paire d'ItemSets de la forme : $a \rightarrow b$ où $a \neq \Phi$ et $b \neq \Phi$, et $a \cap b = \Phi$ (*a* et *b* n'ont pas d'Items en commun) [Agra & al 1993].

Afin d'éviter les règles triviales, on impose ainsi deux contraintes de taille, d'une part la prémisse et la conclusion doivent être non vides et disjoints et, d'autre part, les règles générées sont soumisses à une évaluation selon deux mesures de qualité classiques : le support et de la confiance que nous définissons comme suit :

Définition2 : Soient *a* et *b* deux ItemSets d'un contexte de DM. On appelle :

- *support d'un ItemSet a* le rapport du nombre d'individus contenant *a* par le nombre total d'individus de la base : $\sup(a) = \frac{n_a}{n}$ et
- *Support d'une règle* $a \rightarrow b$ la proportion d'individus qui vérifient la règle dans la base de données ou Pr(*ab*): $\sup(a \rightarrow b) = \frac{n_{ab}}{n}$

Le support peut être calculé pour un ItemSet ou pour une règle d'association.

Définition3 : On définit la confiance d'une règle d'association $a \rightarrow b$ comme étant la proportion d'individus qui possèdent *b* parmi ceux qui possèdent *a* (ie. Elle exprime une probabilité condition--nelle Pr (*b*/*a*)) *:* $Conf(a \rightarrow b) = \frac{n_{ab}}{n_a}$.

La confiance, ne s'applique qu'aux règles d'association

Les différentes méthodes d'extraction et de traitement de règles d'association présentées dans cet article sont illustrées à partir d'un exemple de données fictifs *« jeu d'essai »,* du tableau ***Tab.1***. Cette table est constituée de 6 enregistrements décrits par 5 items $\{i_1,\ldots, i_5\}$. Nous utilisons ce jeu de données pour des raisons de clarté.

D'après notre exemple, les supports respectifs de l'ItemSet $\{i_1, i_2\}$ et de la règle $i_5 \rightarrow i_2$ sont de :
$\boldsymbol{sup}(\{i_1, i_2\}) = \frac{|\{e_1,\ e_3,\ e_5\}|}{6} = \frac{3}{6} = 50\%$ *et* $\boldsymbol{sup}(\{i_5 \rightarrow i_2\}) = \frac{|\{e_1,\ e_2,\ e_3,\ e_5\}|}{6} = \frac{4}{6} \approx 67\ \%$

Tab. 1 Jeu d'essai

Individus	Les Items (attributs)
e_1	i_1 i_2 i_4 i_5
e_2	i_2 i_3 i_5
e_3	i_1 i_2 i_4 i_5
e_4	i_1 i_3 i_5
e_5	i_1 i_2 i_3 i_4 i_5
e_6	i_2 i_4

ItemSet (i_2, i_4)

La confiance de la règle " $i_2 \rightarrow i_5$", est : $\boldsymbol{conf}(\{i_2 \rightarrow i_5\}) = \frac{\sup(\{i_2,\ i_5\})}{sup(\{i_2\})} = \frac{4}{5} = 80\%$

Remarque : Afin de réduire le nombre impressionnant de règles d'association générées, la solution consiste alors, à ne conserver, que les règles dont le support et la confiance sont au moins égaux à des seuils minimaux de support et de confiance préalablement définis par l'utilisateur.

On se fixe donc deux seuils σ_{smin} et $\sigma_{sconf} \in [0,1]$.

Définition4 :Un ItemSet est considéré comme *fréquent* si son support vérifie la condition suivante : $\sup(ItemSet) \geq \sigma_{smin}$. Une règle est *valide* si sa confiance est supérieure ou égale au seuil σ_{sconf} : $\text{conf}(a \rightarrow b) \geq \sigma_{sconf}$.

Ainsi la règle $i_2, i_3 \rightarrow i_5$ de support 67% et de confiance 100% est valide (avec un seuil σ_{smin} = 50%). Cette règle signifie que l'ensemble des individus vérifiant les Items i_2 *et* i_3 ont tendance à vérifier aussi l'Item i_5.

Les règles générées ne sont pas exactes[1] et admettent des contre exemples. Il s'avère donc nécessaire de valider chaque règle en quantifiant la puissance de sa tendance implicative à l'aide de mesures de qualité.

3.2 Extraction de règles

L'extraction des règles d'association consiste à déterminer l'ensemble des règles *potentiellement intéressantes*[2]. La figure ***Fig. 3*** décrit les différentes étapes du processus de fouille de règles d'association.

Les algorithmes d'exploration et d'extraction des règles d'association, dont le plus populaire est l'algorithme Apriori permettent, ainsi, de dégager deux types de connaissances :

- Les ensembles d'Items fréquents (dont le support est au moins égal à σ_{smin}).
- L'ensemble des règles d'association. Ce dernier se construit à partir des ItemSets fréquents obtenus lors de l'étape précédente. Les règles retenues sont celles possédant une confiance suffisante par rapport au seuil de confiance fixé.

L'efficacité de l'algorithme repose sur la propriété *d'antimonotonicité du support* qui permet d'élaguer les ItemSets non fréquents d'une base de données volumineuse.

---

1 *Une règle d'association $X \rightarrow Y$ est dites exacte si $X \subseteq Y$. C'est l'implication à 100%.*
2 *dont le support et la confiance sont au moins égaux, aux seuils σ_{smin} et σ_{sconf} prédéfinis.*

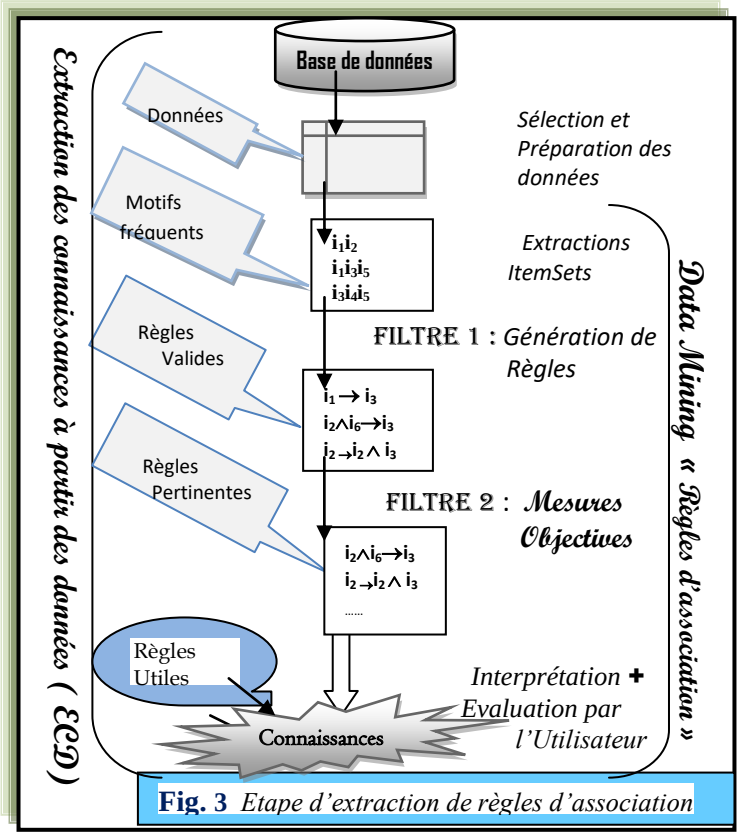


Fig. 3 *Etape d'extraction de règles d'association*

Sans chercher à rentrer dans les détails de cette propriété, si un motif est déclaré non fréquent, alors tous ses sur-ensembles sont écartés. Cette propriété se base sur les ItemSets de l'itération k-1 pour extraire ceux de l'itération k et les motifs potentiellement fréquents sont les motifs dont tous les sous-ensembles ont été désignés fréquents. Les deux étapes de l'algorithme sont très coûteuses en temps et en espace mémoire, notamment lorsque le nombre de variables est grand[1]. Ainsi, par exemple, pour n= 100 (nombre d'Items), le nombre d'itemSets pouvant être généré (pour $\sigma_{smin} = 0$) approche les $2^{100} - 1 \approx 10^{30}$.

La génération de règles d'association est beaucoup moins coûteuse en termes de temps d'exécution que celle des Itemsets fréquents, qui demande de nombreux balayages de la base pour déterminer les supports des ItemSets. Mais, malgré cela, pour des bases de données denses, l'extraction des règles peut conduire à une explosion combinatoire due essentiellement au nombre de conjonctions des items manipulés.

Pour que les algorithmes soient exécutables, on est souvent amené, notamment lorsque la taille du tableau est trop grande à augmenter le seuil du support, ceci conduit à la diminution du nombre de règles au détriment de leur l'intérêt. Cette contrainte supplémentaire rend difficile l'extraction *de Pépites* [2], à savoir les connaissances de forte confiance et de faible support et qui peuvent présenter un réel intérêt pour l'expert. Afin d'améliorer la qualité des règles extraites, il convient de lancer un filtrage ultérieur basé sur d'autres mesures d'intérêt plus adaptées aux données que la confiance et le support.

---

1 *Dans le cas de données réelles, le nombre d'items est en général de l'ordre du millier.*

2 *Les "pépites" sont des exceptions, des règles surprenantes. Il faut remarquer que cette appellation est celle qui est la plus en phase avec la métaphore de la fouille de données où " les pépites d'or " découvertes sont assimilées ici aux règles utiles, de faible support mais de forte confiance.*

L'idée est de procéder à une fouille de données en deux temps. La première concerne l'extraction des règles d'association valides. A l'issu de ce filtrage, toutes les règles découvertes, ne peuvent pas être proclamées les plus intéressantes et doivent subir un second filtrage plus sélectif. Cette étape du processus n'est pas automatisée. Par conséquent une évaluation des règles extraites de façon semi-automatique, où l'expert est sollicité, proclamée étape de *post-traitement* est considérée comme la plus importante.

De nombreux travaux ont eu pour objectif d'assister l'utilisateur dans sa quête à la recherche de la connaissance rare et à forte valeur ajoutée. Nous nous somme intéressés à une des solutions qui est l'évaluation des règles d'association par les mesures de qualité objectives, mais dirigée par l'utilisateur. Dans cette optique, nous avons retenues certaines mesures, *parmi les utilisées*, nous commençons par les énumérer en vue d'un éventuel classement.

3.3. Recensement de certaines mesures d'évaluation

En fouille de règles d'association, plusieurs mesures de qualité ont été proposées pour quantifier leur intérêt. Dans un contexte binaire, les valeurs de ces mesures sont exclusivement déterminées par la table de contingence de la règle $a \rightarrow b$, présentée dans la table ***Tab.2.*** croisant *a* et *b*. En d'autres termes, lorsque les effectifs marginaux *n*, n_a, n_b sont fixés, il suffit de connaître, par exemple "*le nombre d'exemple* n_{ab}" ou *"le nombre de contre exemples* $n_{a\bar{b}}$ " pour retrouver les reste des valeurs.

Tab. 2 Tableau de contingence de a→b

b / a	1	0	Σ
1	n_{ab}	$n_{a\bar{b}}$	n_a
0	$n_{\bar{a}b}$	$n_{\overline{ab}}$	$n_{\bar{a}}$
Σ	n_b	$n_{\bar{b}}$	$\boxed{n}$

Les différentes observations du tableau de contingence ***Tab.2*:**
n_a = le nombre d'enregistrements vérifiant *a*.
n_b = le nombre d'enregistrements vérifiant *b*.
n_{ab} = le nombre d'exemples de la règle.
$n_{a\bar{b}}$ = le nombre de contre-exemples à la règle

Les grandeurs n_a,n_b, n_{ab}, $n_{a\bar{b}}$ sont situées dans l'intervalle [0, n], elles sont souvent utilisées dans les algorithmes de fouille pour caractériser des règles d'association. Elles servent également de base pour le calcul des autres mesures de qualité des connaissances, à condition de supposer que l'estimation de la probabilité d'un ensemble *A de E* est égal à sa fréquence, soit $\frac{n_A}{n}$.

3.4 Définition d'une mesure de qualité

Définition3 : Une mesure de qualité (notée μ) est une fonction de l'ensemble des règles d'association à valeurs dans $\mathbb{R}$, telle que :

$$\mu = \begin{cases} (a \rightarrow b) \rightarrow \mathbb{R} \\ \mu(a \rightarrow b) = f\,(n, n_a, n_b, n_{a\bar{b}}\,(ou\ n_{ab})) \end{cases}$$ pour toute règle d'association $a \rightarrow b$, $\mu(a \rightarrow b)$ dépend des quatre paramètres n, n_a, n_b et une cardinalité de variables conjointes $n_{a\bar{b}}$ *ou* n_{ab}.

Situation de référence : Un point de référence est une valeur constante d'une mesure μ décrivant une situation particulière pour les données. Les trois situations de référence mesurables auquel on se réfèrent pour évaluer une règle sont: le cas d'indépendance[1], d'incompatibilité et d'implication logique.

Selon Piatetsky-Shapiro[Piat 1991], un des pionniers du DM, la valeur d'une mesure μ de la règle $a \rightarrow b$ en cas d'indépendance [2] est une valeur de référence et doit vérifier les principes suivants :

Propriétés objectives

comportement de μ	***Condition***	***$\mu(a \rightarrow b)$***
(1) *Indépendance*	*Pr(ab) = Pr(a).Pr(b*	***= 0***
(2) *Attraction*	*Pr(ab) > Pr(a). Pr(b)*	***> 0***
(3) *Répulsion*	*Pr(ab) < Pr(a). Pr(b*	***< 0***

Ainsi, en cas d'attraction, la valeur d'une mesure μ doit être positive. Lorsque *a* et *b* se trouvent dans une situation d'indépendance, μ est nulle ou proche de zéro et donc peut être éliminée. Enfin, dans le cas de répulsion, les valeurs doivent être négatives. Ces principes énoncés par Piatetsky permettent d'identifier de façon précise où se situer par rapport à l'indépendance.

Cependant, les conditions (1) et (3) ne sont pas vérifiées par la confiance. La confiance permet certes d'identifier l'implication logique (Conf=1) et l'incompatibilité (Conf =0) mais concernant l'indépendance, sa valeur n'est pas constante, elle dépend de Pr(*b*) [Pr(*b*/*a)* = Pr(*b)*], ceci entraine forcement la sélection de règles non pertinentes. Pour pallier à cette insuffisance de nombreux indices de qualité de règles ont été développés, ceci ajoute une difficulté supplémentaire aux contraintes énoncées précédemment, car en plus du problème de la sélection des bonnes règles s'ajoutent celui du choix des bonnes mesures d'intérêt.

3.5 Vers d'autres mesures de qualité

La table ***Tab.3*** liste 25 mesures d'intérêts objectives ([Len & al 04], [Lall & al 04], [Tan & al 02]) utilisées pour quantifier l'intérêt des règles d'association. Les mesures de qualité sont nombreuses, on comptabilise approximativement une cinquantaine. Nous allons examiner un certain nombre d'entre-elles, parmi les plus utilisées. Elles permettent d'attribuer un score aux règles d'association afin de les soumettre à évaluation et procéder à leur classement.

Remarque : Notons à ce propos, qu'à notre connaissance, *les Progiciels actuels,* disponibles sur le marché, *proposent* ***tous*** *un sous-ensemble* réduit de mesures (au maximum 5), utilisées essentiellement pour des raisons pratiques et non à des fins comportementales.

Dans cette table *a* et *b* sont respectivement l'antécédent et le conséquent d'une règle. P(*a*) dénote la probabilité de *a* et P(*b*/*a*) = $\frac{P(ab)}{P(a)}$ représente la probabilité d'observée *b* dans les données indépendamment de *a*.

[1] *On précise cependant que ces principes ont été énoncés pour des mesures de qualité telles que la confiance ou le support*
[2] *Les variables binaires a et b sont indépendantes si et seulement si P(a /b) = P(a)×P(b).*

Tab.3 ***Liste des mesures retenues***

	Nom de la mesure	Formules de $\mu(a \rightarrow b)$	
1	Confiance	**Conf** = $\frac{n_{ab}}{n_a}$	**Conf** = *P(b/a)*
2	Confiance Centrée	**ConfC** = $\frac{n.n_{ab}-n_b}{n.n_a}$	**ConfC** = Conf – P(b)
3	Conviction	**Conv** = $\frac{n_a.n_{\bar{b}}}{n.n_{a\bar{b}}}$	**Conv** = $\frac{\boldsymbol{P(a).\ P(\bar{b})}}{\boldsymbol{P(a\bar{b})}}$
4	Cosinus	***Cos*** = $\frac{n_{ab}}{\sqrt{n_a.n_b}}$	***Cos*** = $\frac{P(ab)}{\sqrt{P(a)\ .\ P(b)}}$
5	Dépendance	***Dep*** = $\left\lvert\frac{n_{ab}}{n_a} - \frac{n_a}{n}\right\rvert$	***Dep*** = $\lvert P(b/a) - P(a)\rvert$
6	Jaccard	***Jac*** = $\frac{n_{ab}}{n_{a\bar{b}}+ n_b}$	***Jac*** = *P(ab) / (P(a)+P(b)–P(ab))*
7	Laplace	***Lap*** = $\frac{n_{ab}+1}{n_a+2}$	***Lap*** = $\frac{n_{ab}+1}{n_a+2}$
8	Lift ou Intérêt	***Lift*** = $\frac{n.n_{ab}}{n_a.n_b}$	***Lift*** = $\frac{\boldsymbol{P(ab)}}{\boldsymbol{P(a)\ .\ P(b)}}$
9	Loevinger	***Loe***= $1 - \frac{n.n_{a\bar{b}}}{n_a.n_{\bar{b}}}$	***Loe*** = $1 - \frac{P(a)\ .\ P(\bar{b})}{P(\ a\bar{b}\ )}$
10	Moindre Contradiction	***MoCo*** = $\frac{n_{ab}-n_{a\bar{b}}}{n_b}$	***MoCo*** = $\frac{\boldsymbol{P(ab)-\ P(a\bar{b})}}{\boldsymbol{P(b)}}$
11	Nouveauté (leverage)	***Nouv*** = $\frac{n_{ab}}{n} - \frac{n_a n_b}{n^2}$	***Nouv*** = *P(b/a) – P(a).P(b)*
12	Rule-Interest (PS)	$\boldsymbol{M_{PS}}$ = $n_{ab} - \frac{n_a n_b}{n}$	$\boldsymbol{M_{ps}}$ = *P(ab) – P(a).P(b)*
13	Satisfaction	***Sat*** = $1 - \frac{n_{a\bar{b}}}{n_{\bar{b}}}$	***Sat*** = $\frac{\boldsymbol{P(\bar{b})\ -P(\bar{b}/a)}}{\boldsymbol{P(\bar{b})}}$
14	Surprise	***Surp*** = $\frac{n_{ab}-n_{a\bar{b}}}{n_a}$	***Surp*** = $\frac{\boldsymbol{P(ab)-\ P(a\bar{b})}}{\boldsymbol{P(b)}}$
15	Support	***Sup*** = $\frac{\boldsymbol{n_{ab}}}{\boldsymbol{n}}$	***Sup*** = *P(ab)*
16	Taux des Exemples et des Contre-exemples	***Tec*** = $\frac{n_{ab}-n_{a\bar{b}}}{n_{ab}}$	***Tec*** = $1 - \frac{P(\ a\bar{b})}{P(ab)}$
17	Coefficient de Corrélation	***R(a,b)***= $\frac{n.n_{ab}-n_a.n_b}{\sqrt{n_a.n_b.n_{\bar{a}}.n_{\bar{b}}}}$	**R(a,b)** = $\frac{P(ab)-P(a)\ .\ P(b)}{\sqrt{P(a).P(b).P(\bar{a}).P(\bar{b})}}$
18	*Khi2* (χ^2)	$\boldsymbol{Khi^2} = \frac{n(n.n_{ab} - n_a n_b)^2}{n_a.n_b\ .n_{\bar{a}}.n_{\bar{b}}}$	$\boldsymbol{Khi^2} = \frac{n(P(ab) - P(a).P(b))^2}{P(a).P(b).P(\bar{a}).P(\bar{b})}$
19	Zhang	***Zhang*** = $\frac{n.n_{ab}- n_b n_b}{\max\{n_{ab}.n_{\bar{b}}\ ,\ n_b.n_{a\bar{b}}\}}$	***Zhang*** = $\frac{P(ab)- P(a).P(b)}{\max\{P(ab).P(\bar{b})\ ,p(b)\ .\ P(\ a\bar{b})\}}$
20	Gain d'information	***GI*** = $\log\left(\frac{n\ .\ n_{ab}}{n_a\ .\ n_b}\right)$	***GI*** = $\log\left(\frac{p(ab)}{P(a)\ .\ P(b)}\right)$
21	Spécificité	***Spec***= $\frac{n_{\bar{a}\bar{b}}}{n_{\bar{b}}}$	***Spec*** = $P(\bar{b}/\bar{a})$
22	Pearl	***Pearl*** = $\frac{n_a}{n}.\left\lvert\frac{n_{ab}}{n.n_a} - \frac{n_b}{n}\right\rvert$	***Pearl*** = *P(a)* * $\lvert P(b/a) - P(b)\rvert$
23	Fiabilité négative	***FiNeg*** = $\frac{n_{\bar{a}\bar{b}}}{n- n_a}$	***FiNeg*** = $P(\bar{b}/\bar{a})$
24	Rappel	***Rap*** = $\frac{n_{ab}}{n_b}$	***Rap*** = *P(a/b)*
25	Indice de Ganascia	***Gan*** = $\frac{n_{ab}- n_{\bar{a}\bar{b}}}{n_a}$	***Gan*** = $\frac{\boldsymbol{P(ab)-P(\ \bar{a}\bar{b})}}{\boldsymbol{P(a)}}$

4. Description de notre approche

Nous avons vu dans la section précédente que, pour construire les règles d'association, la *confiance* n'était pas une mesure inéluctable et qu'elle pouvait très bien être remplacée par d'autres mesures. Les mesures de qualité sont caractérisées par des propriétés intrinsèques qui contribuent à ordonner les règles d'association. Néanmoins, ces mesures retournent des résultats diffèrent et parfois conflictuels sur la qualité des règles. Il est souvent difficile de déterminer la meilleure mesure.

Afin d'aider l'analyste dans sa quête, nous proposons une approche *semi-automatique*, qui consiste à repérer et à conserver les règles les plus informatives. Nous procédons en deux temps. Nous calculons d'abord une classification sur l'ensemble des 25 mesures, sans savoir à priori le nombre de classes produites. Nous présentons la classification non supervisée selon deux méthodes statistiques : l'analyse en composante principale (ACP) et une association regroupant la méthode des K-means à la classification hiérarchique (CAH). Il s'agit de classer les 25 mesures en catégories de manière à dériver un sous-ensemble de mesures d'intérêt que nous appelons $\mathcal{Smc}$. Les différents rangements obtenus par ce sous-ensemble de mesures d'accompagnement, que nous dévoilons ci-après, représentent des « *points de vue* » complémentaires sur l'ensemble des règles.

[*Ensuite, nous exposons un ensemble de critères à travers la formulation des propriétés des mesures de $\mathcal{Smc}$, dont le but de mener une analyse d'aide multicritères et faire ainsi participer l'analyste dans le processus d'extraction de règles intéressantes*]. L'objectif recherché par ce processus de sélection est d'apporter de nouveaux outils capables d'identifier un sous-ensemble de règles d'associations pertinentes, réduites, de les placer en tête de liste avant de les présenter à l'analyste.

Etapes de la démarche :

1. Extraction de l'ensemble des règles d'association.
2. Classement des règles suivant 25 mesures de qualité de la table ***Tab.3***
3. Par une ACP, former un assortiment de mesures similaires.
4. Application d'une méthode mixte (pour renforcer les résultats de l'ACP).
5. Confrontation des résultats obtenus par L'ACP aux résultats de la méthode mixte.
6. Construction d'un ensemble $\mathcal{Smc}$, formé d'un représentant de chaque classe.
7. *[Adopter une méthodologie multicritères pour extraire les règles les plus pertinentes après une étude formelle de certaines propriétés des mesures de l'ensemble $\mathcal{Smc}$]*.

L'objet de cette section est de sélectionner un sous-ensemble d'indicateurs de qualité en étudiant les éventuelles corrélations entre les mesures retenues, afin de réduire l'espace de recherche. Nous commençons par énumérer *toutes les règles d'association* avec les mesures classiques : *le support et la confiance***.** Ensuite, nous évaluons chacune de ces règles sur les autres mesures. Les résultats sont conservés dans un tableau récapitulatif, dont lequel chaque ligne matérialise une règle d'association et chaque colonne la qualité de la règle par rapport à l'une des mesures. Dans le cadre de notre étude, nous avons utilisé le *logiciel Weka*[1] pour extraire les règles d'association.

[1] ***Weka*** *signifie Waikato Environment for Knowledge Analysis (Environnement Waikato pour l'analyse de connaissances). Le logiciel ainsi que son code source sont disponible à l'adresse : http //www.cs.waikato.ac.nz/ml/weka/.*

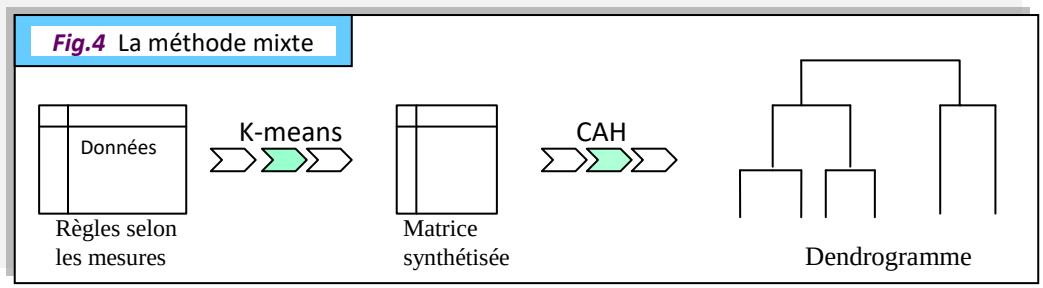
Les logiciels disponibles, les plus exploités en DM, ne disposant que d'un nombre limité de mesures *(Weka supporte les mesures suivantes : le support, la confiance, le lift, la nouveauté et la conviction)* pour l'évaluation des règles. A cet effet, nous avons développé un algorithme pour calculer l'intérêt des règles selon les 25 mesures retenues.

4.1 *Classification des mesures par une méthode mixte*

La complexité algorithmique de la CAH est de l'ordre de $O(n^3)$ dans le pire des cas, ce qui restreint son application à des tableaux de taille raisonnable. Afin de réduire le temps d'exécution de l'algorithme et pouvoir l'appliquer à des données de grande taille, nous proposons de mettre en œuvre une méthode mixte de classification qui combine à la fois la méthode des *k-means* et la *CAH*.

Il est utile de rappeler que la méthode des *K-means* est une technique de classification non supervisée (clustering), qui vise à créer K groupes (appelés aussi clusters) de valeurs les plus compacts possibles. Une des limites de cette méthode, concerne le choix du nombre *K* de classes à retenir. Pour surmonter cette difficulté et tenter de se faire une idée du nombre de clusters à fixer, nous l'avons fait précéder d'une *ACP*.

Nous commençons par utiliser la méthode des *k-means* pour construire 10 classes, qui seront ensuite réduites par la classification hiérarchique (voir *la figure* ***Fig. 4)*****.** Le choix des 10 classes est empirique et non arbitraire, on s'est référer d'une part au découpage résultant de l'ACP et d'autre part au critère de [Won 1999] qui suggère $n^{0.3}$ classe intermédiaires, que l'on arrondi à la dizaine supérieure.



Une fois cette catégorisation effectuée, on applique la CAH non pas à l'ensemble des individus de la population initiale mais aux *centres finaux* de la classification émanant des *k-means*. Le but est de consolider les résultats obtenus par les deux méthodes et donner un plus large pouvoir de généralisation à nos résultats.

Dans l'attente d'une expérimentation sur des données réelles, Nous illustrons cette approche, sur un exemple : "*Jeu d'essai* " (***Tab.1***). Nous avons généré toutes les règles avec un seuil de support $\sigma_{smin} = 0.1$ et un seuil de confiance $\sigma_{sConf} = 0.2$. Les deux seuils fixés correspondent au niveau de significativité minimale requis pour conserver une règle sous *Weka*. Le nombre de règles générées au lancement de l'algorithme était de 221(énumération exhaustive). Les règles extraites n'étaient pas intelligibles et beaucoup trop nombreuses. Après éliminations des règles invalides, le nombre de règles potentiellement intéressantes est passé à 126. Afin de se limiter à un nombre maîtrisable de règles, nous nous sommes restreints à celles dont la conclusion est formée d'un seul Item Fréquent. Ce qui réduit le nombre à 89 (soit 40% des règles obtenues au départ) : un extrait de ces règles est présenté dans le tableau ***Tab.4.***

Algorithme : « Apriori » de *Weka*

Relation: " jeu "
Minimum support: 0.17 (3 instances)
Minimum metric <confidence>: 0.2
Number of cycles performed: 10

Tab.4 *exemple de règles générées*

N°	Règles	Sup	Conf	N°	Règles	Sup	Conf
1	i1 ==> i5	0.67	1	16	i4, i5 ==> i2	0.5	1
2	i4 ==> i2	0.67	1	17	i4, i5 ==> i1	0.5	1
3	i2 ==> i4	0.67	0.8	18	i2, i4 ==> i1	0.5	1
4	i3 ==> i5	0.5	1	19	i2, i5 ==> i1	0.5	1
5	i1 ==> i4	0.5	0.75	20	i1, i5 ==> i4	0.5	0.75
6	i5 ==> i1	0.67	0.8	21	i2, i5 ==> i4	0.5	0.75
7	i5 ==> i3	0.5	0.6	22	i1, i4 ==> i2	0.17	1
8	i1, i2 ==> i4	0.5	1	23	i2, i4, i5 ==> i1	0.5	1
9	i1, i2 ==> i5	0.5	1	24	i1, i3, i4 ==> i2	0.17	1
10	i3, i4 ==> i1	0.17	1	25	i2, i3, i4 ==> i1	0.17	1
11	i1, i3 ==> i5	0.33	1	26	i3, i4, i5 ==> i1	0.17	1
12	i2, i3 ==> i5	0.33	1	27	i1, i2, i3 ==> i5	0.17	1
13	i3, i4 ==> i2	0.17	1	28	i1, i3, i4 ==> i5	0.17	1
14	i3, i4 ==> i5	0.17	1	29	i3, i4, i5 ==> i2	0.17	1
15	i1, i4 ==> i5	0.5	1	30	i1, i3, i4, i5 ==> i2	0.17	1
	.....			31	i1, i2, i3, i5 ==> i4	0.17	1
					.....		

Best rules found:

Nous avons confronté ces règles aux 25 mesures de l'étude. Ce qui nous donne un tableau de 89 lignes et de 25 colonnes dans lequel chaque ligne représente une règle d'association et chaque colonne la qualité de la règle par rapport à l'un des mesures. La table ***Tab. 5,*** donne un aperçu des règles obtenues.

On constate dans cet ensemble que *les règles 1 et 2 sont* « *relativement » les mieux* classées par l'ensemble des mesures et peuvent être considérées comme pertinentes. A l'inverse *les règles : de 63* à *77*, se trouvent être parmi celles qui ont reçu les plus mauvaises scores pour quasiment toutes les mesures, elles sont par conséquent rejetées. Les règles mitoyennes, qui, selon le tableau, sont trouve entre [3-61] et [78–89] et qui sont mauvaises pour certaines mesures et bonnes pour d'autres, sont précisément celles qu'il faut départager.

Tab.5 *Classement de quelques règles par les mesures*

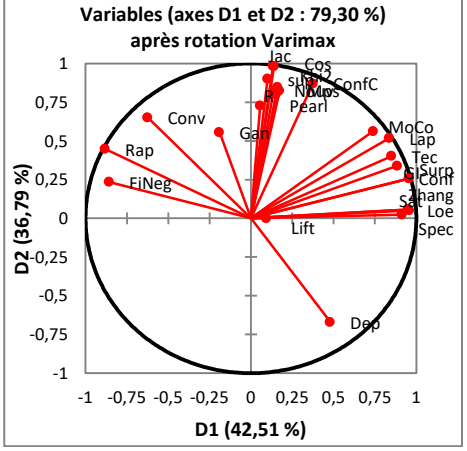
N°	sup	Conf	Lift	Nouv	Conv	Loe	ConfC	Cos	MoCo	Mps	Sat	Jac	Lap	Surp	Tec	Dep	R	Khi2	Zhang	G I	Spec	Pearl	FiNeg	Rap	G a n
1	0,50	1,00	2,00	0,25	1,50	1,00	0,83	1,00	1,00	1,50	1,00	1,00	0,80	1,00	1,00	0,50	1,00	6,00	1,00	2,89	1,00	0,17	1,00	1,00	0,00
2	0,50	1,00	2,00	0,25	1,50	1,00	0,83	1,00	1,00	1,50	1,00	1,00	0,80	1,00	1,00	0,50	1,00	6,00	1,00	2,89	1,00	0,17	1,00	1,00	0,00
3	0,17	1,00	2,00	0,08	0,50	1,00	0,50	0,58	0,33	0,50	1,00	0,33	0,67	1,00	1,00	0,83	0,45	1,20	1,00	2,89	1,00	0,06	0,60	0,33	-2,00
.....			.....			.....				.....				.....					.....			.....			
60	0,33	1,00	1,20	0,06	0,33	1,00	0,58	0,63	0,40	0,33	1,00	0,40	0,75	1,00	1,00	0,67	0,32	0,60	1,00	3,40	1,00	0,22	0,25	0,40	0,50
62	0,17	0,20	1,20	0,03	0,83	0,04	0,17	0,45	-3,00	0,17	0,20	0,20	0,29	-0,60	-3,00	0,63	0,20	0,24	0,20	0,18	0,20	0,11	1,00	1,00	0,00
63	0,17	1,00	1,20	0,03	0,17	1,00	0,17	0,45	0,20	0,17	1,00	0,20	0,67	1,00	1,00	0,83	0,20	0,24	1,00	3,40	1,00	0,11	0,20	0,20	0,00
64	0,17	1,00	1,20	0,03	0,17	1,00	0,17	0,45	0,20	0,17	1,00	0,20	0,67	1,00	1,00	0,83	0,20	0,24	1,00	3,40	1,00	0,11	0,20	0,20	0,00
.....			.....			... ...				... ...				... ...					... ...			... ...			
77	0,17	0,20	1,20	0,06	0,83	0,04	0,17	0,45	-3,00	0,17	0,20	0,20	0,29	-0,60	-3,00	0,63	0,20	0,24	0,20	0,18	0,20	0,11	1,00	1,00	0,00
78	0,50	0,75	1,13	0,06	0,67	0,26	0,58	0,75	0,50	0,35	0,50	0,60	0,67	0,50	0,67	0,08	0,26	0,40	0,34	2,89	0,50	0,36	0,51	0,75	0,50
.....			.....			... ...				... ...				... ...					... ...			... ...			
89	0,50	0,75	1,13	0,06	0,67	0,26	0,58	0,75	0,50	0,35	0,50	0,60	0,67	0,50	0,67	0,08	0,26	0,40	0,34	2,89	0,50	0,36	0,51	0,75	0,50

Règles utiles Mauvaises règles

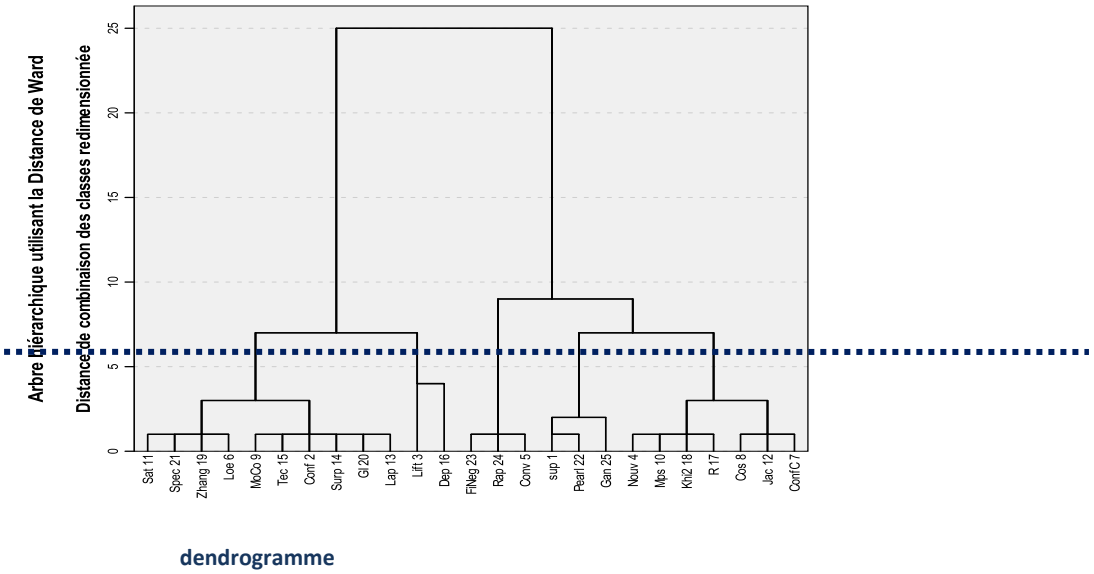
Application des méthodes énoncées : Comme indiqué plus haut, on cherche à découvrir des mesures ayant des comportements similaires à partir de la matrice des règles de la table ***Tab.5bis*** donnée en annexe. Pour cela, nous avons soumis ce tableau de données à l'analyse en composantes principales (ACP) normée disponible sous SPSS18. Nous obtenons en sortie une matrice des corrélations entre les différentes mesures, sur les 89 règles sélectionnées. Cette matrice est présentée en annexe dans la table ***Tab.6*****.**

Ces résultats font apparaitre de fortes corrélations entre de nombreuses mesures. Ainsi la première classe concerne 10 mesures fortement corrélées à savoir *{ Moco, Conf, Lap, Surp, Tec, Gi, Sat, Spec, Loe, Zhang }*. *Le lift et Dep* semble assez marginalisés. Le cercle des corrélations, donne en ***Fig.5*** les pourcentages d'inertie : 42,51% (axe1 horizontal) et 36,79% (axe2 vertical), confirmant ainsi les tendances.

Fig.5 Le cercle des corrélations



Le dendrogramme résultant de la classification ascendante hiérarchique de variables en appliquant l'agrégation selon le critère de Ward et le carré de la distance Euclidienne prescrivent une partition des variables en 3 ou 8 classes. Pour éviter une forte perte d'inertie interclasses, on coupe l'arbre représenté en ***Fig.6,*** à une hauteur de 5 (ie. cette valeur est également jugée optimale en référence à la matrice des corrélations de l'ACP), on obtient 5 classes selon les partitions décrites dans ***Tab.7*****.** Ce résultat est conforme à 90% avec celui obtenu par l'ACP.



dendrogramme

Tab.7 Les classes préconisées par la méthode mixte

Les Classes	*Nombre de mesures par classe*	*Les mesures*
2	10	Moco, Conf, Lap, Surp, Tec, Gi, Sat, Spec, Loe, Zhang
1	3	Sup, Pearl, Gan
3	2	Lift, dep
4	3	Conv, FiNeg, Rap
5	7	Nouv, ConfC, Cos, Mps, Jac, R, Khi2

Chacune des classes regroupe les mesures les plus fortement corrélées. Il suffit à présent de désigner un représentant de chaque classe. Le sous-ensemble choisi est : $Smc = \{ Sup, Conf, ConfC, Lift, Rap \}$.

A partir de l'ensemble (Smc), on range les règles en fonction des valeurs de chaque mesure de qualité. Ensuite, on calcule la précision des évaluations, en considérant les *nr* % meilleures règles fournies par chaque mesure.

La précision est définie par la formule: $Précision = \frac{Nombre\ de\ règles\ potentiellement\ pertinentes}{Nombre\ de\ règles\ total}$

Pour déterminer *nr*, nous avons fixé un seuil de pertinence σi_{pert}, pour chaque mesure μ_i. Nous considérons qu'une règle est pertinente si son évaluation par une mesure donnée est supérieure au seuil de pertinence fixé pour cette même mesure. La fixation des seuils de pertinence dépend de la dispersion des valeurs dans chaque colonne. On note $R_i(\mu_i, \sigma_{ipert})$, l'ensemble des règles potentiellement pertinentes pour une mesure μ_i et un seuil σ_{ipert}.

L'ensemble des meilleures solutions est défini par $\mathcal{E}_{m\mathcal{R}} = [R_1(\mu_1, \sigma_{1pert}), R_2(\mu_2, \sigma_{2pert}), \ldots. R_5(\mu_5, \sigma_{5pert})]$. On décide de conserver une règle de l'ensemble $\mathcal{E}_{m\mathcal{R}}$, lorsqu'au moins deux mesures l'ont classé au même rang dans l'intersection de $\mathcal{E}_{m\mathcal{R}}$ ($\cap\ \mathcal{E}_{m\mathcal{R}}$).

Tab.7 pré-ordres résultant des 5 mesures

Rang/μ	*N°Règl*	*sup*	*N°Règle*	*Conf*	*N°Règle*	*Lift*	*N°Règle*	*Nouv*	*N°Règle*	*Conv*
1	40	0,667	1	1,000	1	2,000	1	0,250	1	1,500
2	41	0,667	2	1,000	2	2,000	2	0,250	2	1,500
3	42	0,667	3	1,000	3	2,000	7	0,170	7	1,000
4	43	0,667	4	1,000	4	2,000	8	0,170	8	1,000
5	1	0,500	5	1,000	5	2,000	9	0,170	9	1,000
6	2	0,500	6	1,000	6	2,000	10	0,170	10	1,000
7	7	0,500	7	1,000	12	2,000	11	0,170	11	1,000
8	8	0,500	8	1,000	13	2,000	16	0,170	16	1,000
9	9	0,500	9	1,000	14	2,000	17	0,170	17	1,000
10	10	0,500	10	1,000	15	2,000	18	0,170	18	1,000
11	11	0,500	11	1,000	7	1,500	19	0,170	19	1,000
12	16	0,500	21	1,000	8	1,500	20	0,170	20	1,000
13	17	0,500	24	1,000	9	1,500	40	0,110	12	0,830
14	18	0,500	26	1,000	10	1,500	41	0,110	13	0,830
15	19	0,500	28	1,000	11	1,500	42	0,110	14	0,830
16	20	0,500	30	1,000	16	1,500	43	0,110	15	0,830
17	44	0,500	31	1,000	17	1,500	3	0,080	22	0,830
18	45	0,500	35	1,000	18	1,500	4	0,080	23	0,830
19	46	0,500	36	1,000	19	1,500	5	0,080	25	0,830
20	47	0,500	39	1,000	20	1,500	6	0,080	27	0,830
21	48	0,500	40	1,000	21	1,500	12	0,080	29	0,830
22	49	0,500	43	1,000	22	1,500	13	0,080	32	0,830
23	50	0,500	44	1,000	23	1,500	14	0,080	33	0,830
24	51	0,500	47	1,000	24	1,500	15	0,080	34	0,830
25	52	0,500	48	1,000	25	1,500	44	0,080	37	0,830
26	53	0,500	50	1,000	26	1,500	45	0,080	38	0,830
27	54	0,500	53	1,000	27	1,500	46	0,080	41	0,830
28	55	0,500	55	1,000	28	1,500	47	0,080	42	0,830
29	56	0,500	57	1,000	29	1,500	48	0,080	45	0,830
30	57	0,500	58	1,000	30	1,500	49	0,080	46	0,830
31	78	0,500	60	1,000	31	1,500	50	0,080	49	0,830

Le tableau ***Tab. 8*** présente un extrait des règles de $\mathcal{E}_{m\mathcal{R}}$, (39 au total). On constate, en générale qu'aucune règle mal classée par les 25 mesures de départ ne figure dans $\mathcal{E}_{m\mathcal{R}_c}$. Les règles ayant reçu les meilleurs notes sont : de 1 à 11, de 16 à 20 et la règle 55. Ce qui conforte notre précédente analyse.

Néanmoins pour apporter une certaine crédibilité à nos résultats, il serait intéressant, comme première solution suggérée, de valider les règles obtenues en les comparants à des règles de décision apprises grâce à un algorithme du type C4.5 par exemple. Pour compléter cette approche, vue l'importance du rôle de l'expert dans un tel processus, une seconde solution serait de faire participer ce dernier au processus de fouille de règles en lui donnant la possibilité d'exprimer ses préférences, dans le but de lui recommander la(s) bonne(s) mesure(s) en fonction de ses besoins.

5. Conclusion et travaux futurs:

Dans cet article, nous avons présenté un survol de l'état de l'art de l'extraction de l'ECD et du DM. Nous nous sommes intéressés, tout particulièrement à une des techniques du DM : la découverte de règles d'association pertinentes. Le constat général est que pratiquement tous les logiciels, du moins ceux que nous avons testés (Tanagra, Weka, Orange et Clémentine) ont concentré leur objectifs essentiellement sur la performance, en cherchant à tout prix à baisser les seuils des mesures classiques afin de conserver les pépites rares (les exceptions). Ce qui a eu pour conséquence un nombre de règles exorbitants et inexploitable par l'analyste.

Nous avons à travers cette analyse tenté d'apprécier la qualité des règles d'association sur d'autres mesures de qualité, dont le but d'offrir à l'utilisateur le moyen de choisir d'autres indicateurs selon la nature de son application. Nous avons proposé une approche opérationnelle, permettant d'une part, de s'affranchir de la nécessité de fixer un seuil minimal pour filtrer les règles et d'autre part, de contrôler de manière plus efficace le nombre de règles à soumettre à l'utilisateur. Par ailleurs, le classement final d'une règle a été renforcé par la fixation d'un seuil de pertinence, que nous avons calculé en fonction de la distribution de chaque mesure et arbitrée par la présence, en moyenne, d'un nombre m de mesures, évaluant positivement la règle. Cette approche a été expérimentée sur un exemple didactique. Afin de gagner en généralité, il est nécessaire d'effectuer des expérimentations sur des bases de données réelles.

Le travail est de longue haleine*. Nous comptons apporter diverses améliorations :*
1) Réaliser des tests sur des bases de données réelles, afin d'évaluer l'approche.
1) Donner la possibilité à l'utilisateur de choisir un nombre réduit de mesures, en adéquation avec ses préférences.
2) Classement par dominance des règles d'après les propriétés des mesures soustraites de l'ensemble $\mathcal{S}mc$, mais cette fois-ci sous le contrôle de l'expert.

Les perspectives futures seraient l'implémentation d'un prototype qui intègre toutes ces fonctionnalités.

NB : *les phrases écrites en claire indiquent les axes de recherche futurs.*

Annexes:

Tab.6 Matrice de corrélation

	sup	Conf	Lift	Nouv	Conv	Loe	ConfC	Cos	MoCo	Mps	Sat	Jac	Lap	Surp	Tec	Dep	R	Khi2	Zhang	GI	Spec	Pearl	FiNeg	Rap	Gan
sup	1,000																								
Conf	,362	1,000																							
Lift	-,361	,011	1,000																						
Nouv	,539	,309	,414	1,000																					
Conv	,369	-,454	,242	,651	1,000																				
Loe	,073	,910	,197	,281	-,482	1,000																			
ConfC	,821	,589	-,027	,693	,309	,392	1,000																		
Cos	,863	,394	-,051	,763	,512	,158	,937	1,000																	
MoCo	,646	,875	-,012	,456	-,192	,646	,756	,683	1,000																
Mps	,571	,336	,432	,989	,639	,296	,730	,797	,503	1,000															
Sat	,053	,890	,291	,295	-,472	,978	,391	,162	,659	,316	1,000														
Jac	,804	,392	-,061	,778	,528	,157	,926	,996	,669	,806	,158	1,000													
Lap	,622	,950	-,068	,456	-,242	,794	,794	,645	,940	,491	,785	,641	1,000												
Surp	,361	1,000	,012	,309	-,454	,910	,589	,394	,876	,336	,891	,392	,950	1,000											
Tec	,507	,940	,001	,350	-,357	,752	,658	,538	,980	,393	,759	,525	,947	,940	1,000										
Dep	-,669	,275	,177	-,293	-,564	,562	-,422	-,633	-,173	-,320	,483	-,625	-,001	,275	,004	1,000									
R	,560	,333	,452	,959	,602	,324	,714	,778	,489	,978	,335	,780	,481	,333	,384	-,281	1,000								
Khi2	,517	,316	,432	,973	,634	,315	,664	,739	,438	,978	,317	,759	,452	,316	,338	-,232	,974	1,000							
Zhang	-,013	,886	,341	,309	-,443	,976	,329	,112	,585	,328	,986	,106	,717	,836	,689	,549	,369	,345	1,000						
GI	,502	,970	-,156	,268	-,438	,794	,636	,484	,935	,300	,772	,479	,963	,970	,973	,101	,289	,262	,700	1,000					
Spec	,053	,890	,291	,295	-,472	,978	,391	,162	,659	,316	1,000	,158	,785	,891	,759	,483	,335	,317	,986	,772	1,000				
Pearl	,950	,294	-,586	,252	,187	-,027	,750	,797	,562	,285	-,059	,789	,539	,294	,442	,659	,280	,230	-,140	,472	-,059	1,000			
FiNeg	-,026	-,781	,265	,263	,807	-,741	-,182	,073	-,538	,249	-,744	,074	-,665	-,781	,666	,427	,268	,266	,665	-,762	,744	-,125	1,000		
Rap	,285	-,746	,0.13	,280	,866	-,792	,054	,309	-,431	,273	-,777	,311	-,522	,746	-,600	-,691	,271	,268	-,742	-,667	-,777	,229	,889	1,000	
Gan	,826	,013	-,730	,160	,276	,201	,437	,553	,187	,157	-,232	,564	,226	-,014	,074	-,570	,127	,146	-,275	,149	-,232	,788	-,002	,398	1,000

Références

[Agra & al 1993] Agrawal R., Imielinski T., Swami A. " Mining Associations between Sets of Items in Ma -ssive Databases. " in Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data, Washington D.C., pp. 207-216. 1993.

[Agra & al 1994] Agrawal R., Srikant R. " Fast algorithms for mining association rules ", in Proceedings of the twentieth international conference on very large data bases (VLDB 1994) (J. B. Bocca, M. Jarke & C. Zaniolo, éds.), Morgan Kaufmann, p. 487–499. 1994

[Agra & al 1993] Agrawal R., Mannila, R.H. Srikant, R. Toivonen H., Verkamo, A.I. " Fast Discovry of association rules, in, Advances in Knowledge Discovery and Data Mining", AAAI, MIT/Press, Menio park,CA,307-328. 1996.

[Bay 98] Bayardo. R. J. "*Efficiently mining long patterns from databases"* . Proc. SIGMOD conf., pp 85-93, June 1998.

[Brin & al 1997] Brin, S., Motwani, R. , Silverstein C. "Beyond market baskets: generalizing association rules to correlations." In ACM SIGMOD/PODS ' Joint Conference, pages 265–276. 1997.

[Fay & al 1996] Fayyad U.M., Piatesky- Shapiro G. Smyth P., Uthurusamy R. " *From Data Mining to Knowledge Discovery: An Overview Advances in Knowledge Discovery and Data Mining"*, AAAI Press MIT Press, 1996.

[Han & al 01] Han j., Kamber M. "*Mining Association Rules in Large Databases*, (Chap. 6) dans Data Mi--ning Concepts and Techniques ". Morgan Kauffmann Publishers, San Francisco, CA. 2001.

[Kay 1997] Kayser. D, " La représentation des Connaissances ", Université Paris nord. Hermés Science pub--lication, Broché 1997.

[Lall & al 04] Lallich S., Teytaud O. " Evaluation et validation de l'intérêt des règles d'association ". RN--TI-E-1, pp 193- 217. 2004.

[Len & al 04] Lenca P., Meyer P., Vaillant B. , Picouet P. , Lallich S. " Evaluation des mesures de qualité des règles d'association". Mesures de Qualité pour la Fouille de Données, (RNTI-E-1) :219–246. 2004.

[Pas & al 99] Pasquier N., Bastide Y., Taouil R., Lakhal L., " Discovering Frequent Closed Itemsets for Asso--ciation Rules », *Proc. of the 7th Int'l Conf. on Database Theory (ICDT)*, no 1540 Lectures Notes in Com--puter Sciences, Springer, p. 398-416. 1999.

[Piat 1991] Piatetsky-Shapiro. G " Discovery, analysis, and presentation of strong rules ", in Knowledge Dis--covery in Databases (G. Piatetsky-Shapiro & W. J. Frawley, éds.), AAI/MIT Press, p. 229–248. 1991.

[Tan & al 02] Tan P. N., Kumar V., Srivastava, J. " Selecting the right interestingness measure for associa--tion patterns " In Proceedings of the Eighth ACM Sig KDD International Conference on KDD, page 32-41. 2002.

[Won 1999] Wong P.C., Whitney P., Thomas J. "Visualizing Association Rules for Text Mining". In *Proc--eedings of the IEEE Symposium on Information Visualization (INFOVIS'99)*, Salt Lake City, Utah, USA, 120-128. 1999