Discovering Interesting Association Rules by Clustering

Arwa Bezzaouia Aymen Fterich
abezzaouia@uqu.edu.sa
Umm Al-Qura University

Abstract:

The problem of mining association rules has attracted a lot of attention in the research community. Several techniques of efficient discovery of association rules have appeared. Those rules which exceeded a predetermined minimum threshold for support and confidence are considered to be interesting. However, association rule mining does not discover the true correlation relationship, because high minimum support usually generates common sense knowledge, while low minimum support generates huge number of rules, the majority of which are uninformative. Therefore, many metrics of interestingness, such as Convictional Loevinger, Centered Confidence, Lift, etc… have been devised to help find interesting rules while filtering out uninteresting ones. The application of those measures left the user in front of several shortcomings due to the impossibility of filtering out uninteresting rules and respectively removing interesting ones. Those problems are directly linked to the choice of couplets (metric, threshold). The use of metrics is meant to reduce the number of rules. However, after their filtration, the number is still huge which confronts the domain expert with problems during validation.

Our approach consists in attributing to every rule its own vector of metrics, then clustering rules into groups relying on the vectors of measures' value. Besides, every generated cluster is described by the vector of its centre. From this viewpoint, a domain expert can simply validate the centers of those clusters. The result of the expert's decision can be automatically generalized to all members (rules) of the clusters. This approach allows us to reduce significantly the cognitive effort provided in this task.

Keywords: Association Rules, Interestingness, Metrics, Clustering.
DOI Number: 10.48047/nq.2024.22.1.NQ24022 **NeuroQuantology 2024; 22(1):233-245**

1. Introduction

L'extraction des itemsets fréquents à partir des bases de données transactionnelles est une technique permettant la découverte de règles intelligibles et exploitables dans un ensemble de données volumineux. Ces règles exprimant des associations entre items ou attributs dans une base de données. La recherche d'associations est un sujet de recherche attractif et très actif, avec un champ d'applications très large touchant pratiquement tous les domaines: le marketing, l'aide au diagnostic médical, les télécommunications, l'analyse de données spatiales, la téléphonie, etc.

Les algorithmes du type Apriori, Apriori-TID et Apriori-Hybrid [Agrawal et Srikant, 1994], DHP [Park et al., 1995], SON [Savasere et al., 1995], CLOSE [Pasquier et al., 1999], PASCAL [Bastide et al., 2002], fondés 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. Alors, Il faut disposer d'autres mesures de qualité, tels que la

confiance centrée, la conviction, le lift…, pour compléter le support et la confiance afin de réduire le nombre de règles générées [Lallich et Teytaud, 2004]. L'utilisation de ces différentes mesures de qualité présente aussi le défaut de ne pas toujours filtrer correctement les règles et introduit un biais non négligeable dans les règles supprimées (respectivement conservées) lié au choix des couples (mesure de qualité, seuil) et le nombre de ces règles reste important, malgré l'utilisation de ces mesures de qualité.

Nous proposons, dans cet article, de représenter chaque règle d'association par un vecteur de mesures de qualité, puis d'utiliser un algorithme de regroupement (''clustering'') pour regrouper les règles qui semblent les plus proches en se basant sur les valeurs des vecteurs des mesures de qualité étudiées. L'expert intervient uniquement pour valider les différents vecteurs associés aux centres des clusters. La validation réalisée par l'expert pourra être utilisée pour évaluer automatiquement les règles appartenant aux clusters analysés.

2. Définition du problème

Soit L = {i_1, i_2, …, i_m} un ensemble de m littéraux appelés *items*. Soit B = {t_1, t_2,…, t_n} une base de n transactions, chaque transaction t_i étant constituée d'un sous ensemble I $\subseteq$ L d'items et identifiée par un identifiant unique appelé Tid. Un sous ensemble I $\subseteq$ L de taille k est appelé k-Itemset. Une transaction t_i contient un itemset *I* si et seulement si I $\subseteq$ t_i. Soit I_1 et I_2 deux itemsets disjoints tel que $I_2 \neq \emptyset$ et $I_1 \cap I_2 = \emptyset$.

Le support d'un itemset est le nombre de transactions contenant cet itemset. Une règle d'association est une implication de la forme $I_1 \rightarrow I_2$, exprimant le fait que les items dans I_1 tendent à apparaître avec ceux de I_2. Le support de la règle représente la proportion de transactions dans B contenant I_1 et I_2. La confiance est la proportion de transactions parmi celles contenant I_1 qui contiennent aussi I_2.

Le problème d'extraction des règles d'association peut être décomposé en deux sous-problèmes [Agrawal et Srikant, 1994]:

1. Déterminer l'ensemble des itemsets fréquents dans B, ce sont les itemsets dont le support est supérieur ou égale à un seuil *minsupport.*
2. Génération des règles associatives confiantes (>= à un seuil *minconfiance*).

Afin de limiter le nombre de candidats considérés dans chaque itération, en réduisant dynamiquement l'espace de recherche des itemsets fréquents, l'algorithme Apriori se base sur les deux propriétés suivantes :

Propriété 1 : Tous les sous ensemble d'un itemset fréquent sont fréquents.

Cette propriété permet de limiter le nombre des candidats de taille k générés dans la $k^{ème}$ itération en réalisant une jointure conditionnelle des itemsets fréquents lors de l'itération précédente.

Propriété 2 : Tous les sur ensembles d'un itemset infréquents sont infréquents.

Cette propriété permet de supprimer un candidat de taille k lorsqu' au moins un de ses sous ensembles de taille k-1 ne fait pas partie des itemsets fréquents lors de l'itération précédente [Pasquier, 2000].

Les algorithmes d'extraction liés à l'approche support-confiance parcourent le treillis des itemsets pour rechercher les itemsets fréquents, ceux dont le support dépasse *minsupp*, pour en déduire les règles d'association dont la confiance dépasse *minconf*. Ces algorithmes engendrent un très grand nombre de règles qui sont difficiles à gérer et dont beaucoup n'ont que peu d'intérêt. Afin de compenser cette limite, de nombreuses mesures de qualité complémentaires ont été proposées dans la littérature.

Etant donné que l'intérêt dépend à la fois des préférences de l'utilisateur et des données, les mesures peuvent être dissociées en 2 groupes [Freitas, 1999]: les mesures objectives et les mesures subjectives. Les mesures subjectives dépendent essentiellement des buts, connaissances, croyances de l'utilisateur qui doivent être préalablement recueillis. Elles sont associées à des algorithmes supervisés ad hoc [Padmanabhan et Tuzhilin, 1998 ; Liu et al., 1999] permettant de n'extraire que les règles conformes ou au contraire en contradiction

avec les croyances de l'utilisateur, et ainsi d'orienter la notion d'intérêt vers la nouveauté ou l'inattendu. Les mesures objectives, quant à elles, sont des mesures statistiques s'appuyant sur la structure des données ou plus exactement la fréquence des combinaisons fréquentes d'attributs (itemsets). De nombreux travaux de synthèse, récapitulent et comparent leurs définitions et leurs propriétés [Lenca et al., 2004].

Nous avons étudiés les mesures de qualité les plus intéressantes, pour compléter le support et confiance [Tan et al., 2002]. Les principales sont : Rappel, confiance centrée, Lift, Moindre contradiction, Loevinger, Satisfaction et Spécificité, Sebag-Schœnauer…

Nous avons tenu compte des travaux de [Lallich et Teytaud, 2004], [Guillaume, 2000] et [Azé, 2003] afin de choisir une mesure en tenant compte des critères permettant d'apprécier la qualité d'une mesure *m*.

Nous nous intéressons aux mesures qui ont obtenu les notes les plus élevées selon [Lenca et al., 2004] : la confiance centrée, à laquelle nous rajoutons le lift en raison de son utilisation très répandue et de son interprétation facile [Plasse et al., 2006], la conviction et le Lœvinger. La figure 1 rappelle les propriétés de ces mesures.

Mesure	***Expression***	***Référence***
confiance centrée	P(B\A) – P(B)	[Lallich et Teytaud, 2004]
Lift	$\frac{P(AB)}{P(A)\times P(B)}$	[IBM, 1996]
Lœvinger	$\frac{P(B \setminus A) - P(B)}{\overline{P(B)}}$	[Lœvinger, 1947]
Conviction	$\frac{P(A)P(\overline{B})}{P(A\overline{B})}$	[Brin et al., 1997]

Figure 1 : Les mesures de qualité retenues et leurs définitions.

3. Evaluation des règles d'association en utilisant le Clustering

3.1 Idée de base

Les règles d'association, de la forme A,B→ C où A, B et C représentent des objets d'intérêt dans plusieurs domaines, sont couramment utilisées pour rechercher des corrélations cachées dans les données et ce de manière non supervisée, c'est-à-dire non guidée par un expert. Ces règles sont ensuite soumises à un expert pour être validées et pour déterminer si elles sont utiles, nouvelles et pertinentes pour le domaine étudié.

L'un des problèmes majeurs lié à l'utilisation des règles d'association est le nombre considérable de règles obtenues à partir des données.

L'une des méthodes utilisées pour réduire le volume de règles à analyser est d'utiliser différents couples (mesure de qualité, seuil) pour supprimer des règles a priori non intéressantes.

Cette méthode présente le défaut de ne pas toujours filtrer correctement les règles et introduit un biais non négligeable dans les règles supprimées (respectivement conservées) lié au choix des couples (mesure de qualité, seuil).

L'identification et la validation des règles d'association pertinentes reposent toujours sur l'utilisation de mesures de qualité. Nous proposons, dans notre travail de recherche, de représenter chaque règle d'association par un vecteur de mesures de qualité puis d'utiliser un algorithme de clustering pour regrouper les règles qui semblent les plus proches sur la base des valeurs des mesures de qualité étudiées. Ensuite, chaque groupe (cluster) étant décrit par le centre du groupe en question, l'expert intervient uniquement pour valider les différents vecteurs associés aux clusters. La validation réalisée par l'expert pourra être utilisée pour évaluer automatiquement les règles appartenant aux clusters analysés.

Notre approche peut être décrit par l'algorithme suivant :

1. Utiliser un algorithme d'extraction des itemset fréquent.
2. Générer les règles en se basant sur le couple mesure de qualité et seuil.
3. Assigner à chaque règle un vecteur de mesure de qualité qui la représente.
4. Utiliser un algorithme de clustering qui permet de regrouper les vecteurs de mesures en des clusters homogènes.
5. Validation de l'expert

3.2 How to Choose the Algorithm for Clustering

There are many algorithms for clustering. Which one is suitable for a given application? It should be chosen according to the specific scenario and the requirement of users. Generally speaking, the algorithm should be capable to handle hybrid data. The reason is that most data for association analysis are of hybrid data, so the algorithm for clustering is required to be able to cluster data both of numeric attributes and categorical ones. On the other hand, since the dissimilarity between clusters will be used as the interestingness of rules, the dissimilarity or distance should be easy to judge or compute. There are mainly four categories of algorithms for clustering, namely, partitioning, hierarchical, density-based and grid-based approaches. For density-based and grid-based ones, the clusters are generated by expanding the densely-populated regions or combining dense neighboring cells, so it is difficult to judge the dissimilarity between clusters. Fortunately, for partitioning and hierarchical algorithms, the clusters are usually compact and it is easy to compute the dissimilarity between clusters. For k-means or k-medoids algorithms, the mean or medoid is used to represent a whole cluster, so the dissimilarity can be easily computed as the distance between the means or medoids. For hierarchical approaches, single linkage, average linkage, complete linkage, and mean linkage are main measures for calculating the distances between clusters, and they can be used as the dissimilarity [Zhao et al., 2005]. Therefore, partitioning and hierarchical algorithms can be readily used in our approach, while density-based or grid-based ones are not.

In our approach, k-means is used as the algorithm for clustering. Since the k-means algorithm is only for numeric attributes (the values of measures in our approach).

3.3 Clustering Kmeans

Étant donné le nombre ***K*** de clusters recherchés, la méthode de clustering Kmeans [MacQueen, 1967], est la suivante :

1. choisir aléatoirement ***K*** objets de la base qui formeront l'ensemble des centroïdes initiaux représentant les ***K*** clusters recherchés.
2. assigner chaque objet au cluster, dont le centroïde est le plus proche.
3. puis tant qu'au moins un objet change de cluster d'une itération à l'autre :
   - mettre à jour les centroïde des clusters en fonction des objets qui leur sont associés.

$$\overrightarrow{\mu_K} = \frac{1}{N_K} \sum_{i \in D_K} \overrightarrow{x_i}$$

   - mettre à jour les assignations des objets aux clusters en fonction de leur proximité aux nouveaux centroïdes.

$$D_K = \left\{ i \in D \mid \forall j \in [1,K] et\ j \neq K, dist(\overrightarrow{x_i}, \overrightarrow{\mu_K}) < dist(\overrightarrow{x_i}, \overrightarrow{\mu_j}) \right\}$$

On notera *K* le nombre de clusters identifiés, C_k le $K^{ème}$ de ces clusters, D_k l'ensemble des indices des objets associés au cluster C_k, N_k le nombre de ces objets associés, on notera $\overrightarrow{x_i}$, le $i^{ème}$ objet de la base et $\overrightarrow{\mu_K}$ le centre de gravité du cluster(centroïde).

Différentes mesures de distance peuvent être utilisées pour définir la similarité entre objets. La mesure la plus utilisée est indubitablement la distance euclidienne, présentée ci-après. Mais d'autres mesures peuvent également être envisagées.

$$dist(\overrightarrow{x_i}, \overrightarrow{x_j}) = \sqrt{\sum_{d=1}^{M} \left(x_{id} - x_{jd}\right)^2}$$

M : les différentes attributs d'un objet.

3.4 Our Algorithm

Notre algorithme génère, d'abord, les règles en se basant sur le couple mesure de qualité et seuil, en faisant appel à l'algorithme **Apriori()**. Ensuite, il représente chaque règle d'association par un vecteur de mesures de qualité ***(m₁, m₂, m₃, …)*** à l'aide de la procédure **GenRuleVector().** Puis, il regrouper, par la procédure **Kmeans(),** les règles qui semblent les plus proches en se basant sur les valeurs des vecteurs des mesures de qualité étudiées. Après, chaque groupe (``cluster'') généré est décrit par le vecteur de mesure de la règle d'association du centre du groupe en question. Enfin, l'expert intervient uniquement pour valider les différents vecteurs associés aux centres des clusters.

Algorithme 3.1 (Description de notre algorithme)
Entrée :
D : Base de transactions.
minsup : le seuil de support.
minconf : le seuil de confiance.
minconfcen : le seuil de confiance centrée.
minlif : le seuil du lift.
minconv : le seuil de la conviction.
minloe : le seuil du loevinger.
K : le nombre de clusters.
Sortie :
C = {C_1, C_2,.....,C_K} : Les vecteurs de mesures de qualité en des clusters homogène
Algorithme
1. *F* = $\cup_k$ F_k = **Apriori** (D, *minsup*). **//** *F* est l'ensemble des itemsets fréquents.
2. *R* , *V* = **GenRuleVector** (*F, minconf, minconfcen, minlif, minconv, minloe*). **/*** R : est
l'ensemble des règles qui satisfait toutes les mesures.
V : est l'ensemble des vecteurs de mesures qui leurs correspondent.***/**
3. C = {C1, C2,.....,CK}= Kmeans(k, V).

Algorithme 3.2 (Apriori)
Entrée : Base de transactions D, seuil minimal de support γ
Sortie : Ensemble des itemsets fréquents *F*
Algorithme
1. F_1= {1-itemsets fréquents}
2. **Pour** (k=2 ; $F_{k-1} \neq \emptyset$; k++)
3. C_k=Apriori-Gen(F_{k-1}) ;
4. **Pour chaque** transaction t ∈D **Faire**
5. C_t=Subset (C_k, t) ;
6. **Pour chaque** candidat c ∈C_t **Faire**
7. c.support++ ;
8. **Fin Pour**
9. **Fin Pour**
10. F_k={c∈C_k| c.support≥γ} ;
11. Fin Pour
12. **Retourner** *F* = $\cup_k$ F_k

Algorithme 3.3 (GenRuleVector)
Entrée :
F = $\cup_k$ F_k est l'ensemble des itemsets fréquents.
minsup : le seuil de support.
minconf : le seuil de confiance.
minconfcen : le seuil de confiance centrée.
minlif : le seuil du lift.
minconv : le seuil de la conviction.
minloe : le seuil du loevinger.
Sortie :
R : est l'ensemble des règles qui satisfait toutes les mesures.
V : est l'ensemble des vecteurs de mesures qui leurs correspondent (assigner à chaque règle).
Algorithme
1. R = ∅

2. V = ∅
3. Pour (k = 2 ; Fk-1 ≠ ∅ ; k++)
4. Pour chaque sous-ensemble S ≠ ∅ de Fk **faire**
5. Si (Conf (S → Fk-S) >= *minconf*) et (Confcen (S → Fk-S) >= *minconfcen*) et (Conv (S → Fk-S) >= *minconv*) et (Lift (S → Fk-S) >= *minlif*) et (Loev (S → Fk-S) >= *minloe*)
.
Alors
6. rule = " S → (Fk-S) "
7. vecteur = (Supp, Conf, Confcen, Lift, Conv, Loev)
8. R = R ∪ {rule}
9. V = V ∪ {vecteur}
10. Fin pour
11. Fin pour.
12. Retourner R, V

Algorithme 3.4 (Kmeans)
Entrée :
V : est l'ensemble des vecteurs de mesures de qualité qui leurs correspondent (assigner à chaque règle).
K : le nombre de clusters.
Sortie :
C = {C_1, C_2,.....,C_K} : Les différents clusters c'est-à-dire les vecteurs de mesures de qualité en des clusters homogènes
Algorithme
1. Initialiser les différentes centres des clusters $\overrightarrow{\mu_1}$, $\overrightarrow{\mu_2}$,....., $\overrightarrow{\mu_K}$ /*Sélection aléatoire de K instances de V*/
2. Répéter
3. Pour chaque objet $\overrightarrow{x_i}$ ∈ V **faire**
4. $C(\overrightarrow{x_i}) = \min_{j=1..k}(dist(\overrightarrow{x_i},\overrightarrow{\mu_j}))$ /* affectation de chaque objet (vecteur de mesures de qualité) à son cluster le plus proche*/
5. Fin pour
6. Pour (i = 1 ; i ≤ K ; i++)
7. $\overrightarrow{\mu_i} = \frac{1}{N_i}\sum_{i \in C_K} \overrightarrow{x_i}$ //recalculer le centre $\overrightarrow{\mu_i}$ de chaque cluster *Ci.*
8. Fin pour
9. Jusqu'à (Convergence) // tous les centres des clusters deviennent stables.
10. Retourner C = {C_1, C_2,.....,C_K}

3.5 Exemple Illustratif
Soit la base de données transactionnelles *D* suivante :

Tid	Items
1	ABCDE
2	ABCD
3	ACD
4	ABE
5	CD
6	CE

Le seuil de support *minsup* est de 30%.
Le seuil de confiance *minconf* est 40%.
Le seuil de confiance centrée *minconfcen* est 10%.
Le seuil du lift *minlif* est 1.
Le seuil de la conviction *minconv* est 1
Le seuil du loevinger *minloe* est 0,2.

eISSN1303-5150

Soit l'ensemble F_4 de 4-itemsets fréquents {A, B, C, D}. Le nombre des règles d'association extraites (générées) à partir de F_4 et qui vérifient les couples (mesure de qualité, seuil) indiqués précédemment est 11. Chaque règle extraite est représentée par un vecteur de mesure de qualité qui a la structure suivante ***(sup, conf, confcen, lift, conv, loe).***

- R1 : A ➔ B C D ======= (0.333 | 0.5 | 0.166 | 1.5 | 1.333 | 0.25)
- R2 : B ➔ A C D ======= (0.333 | 0.666 | 0.166 | 1.333 | 1.499 | 0.333)
- R3 : A C ➔ B D ======= (0.333 | 0.666 | 0.333 | 2 | 1.999 | 0.499)
- R4 : B C ➔ A D ======= (0.333 | 1 | 0.5 | 2 | ∞ | 1)
- R5 : A B C ➔ D ======= (0.333 | 1 | 0.333 | 1.5 | ∞ | 1)
- R6 : D ➔ A B C ======= (0.333 | 0.5 | 0.166 | 1.5 | 1.333 | 0.25)
- R7 : A D ➔ B C ======= (0.333 | 0.666 | 0.333 | 2 | 1.999 | 0.499)
- R8 : B D ➔ A C ======= (0.333 | 1 | 0.5 | 2 | ∞ | 1)
- R9 : A B D ➔ C ======= (0.333 | 1 | 0.166 | 1.2 | ∞ | 1)
- R10 : A C D ➔ B ======= (0.333 | 0.666 | 0.166 | 1.333 | 1.499 | 0.333)
- R11 : B C D ➔ A ======= (0.333 | 1 | 0.333 | 1.5 | ∞ | 1)

En appliquant un algorithme de clustering avec ***K***=2 qui permet de regrouper les règles qui vont ensembles sur la base de leur vecteur de mesure, on obtient les résultats suivants :

Deux clusters C_1 et C_2 avec les caractéristiques suivantes (Figure 2) :

- Cluster C_1 a pour centroïde le vecteur (0.333 | 1 | 0.366 | 1.64 | ∞ | 1) et contient les règles suivante : R4, R5, R8, R9 et R11.
- Cluster C_2 a pour centroïde le vecteur (0.333 | 0.611 | 0.222 | 1.611 | 1.611 | 0.361) et contient les règles suivante : R1, R2, R3, R6, R7 et R10.

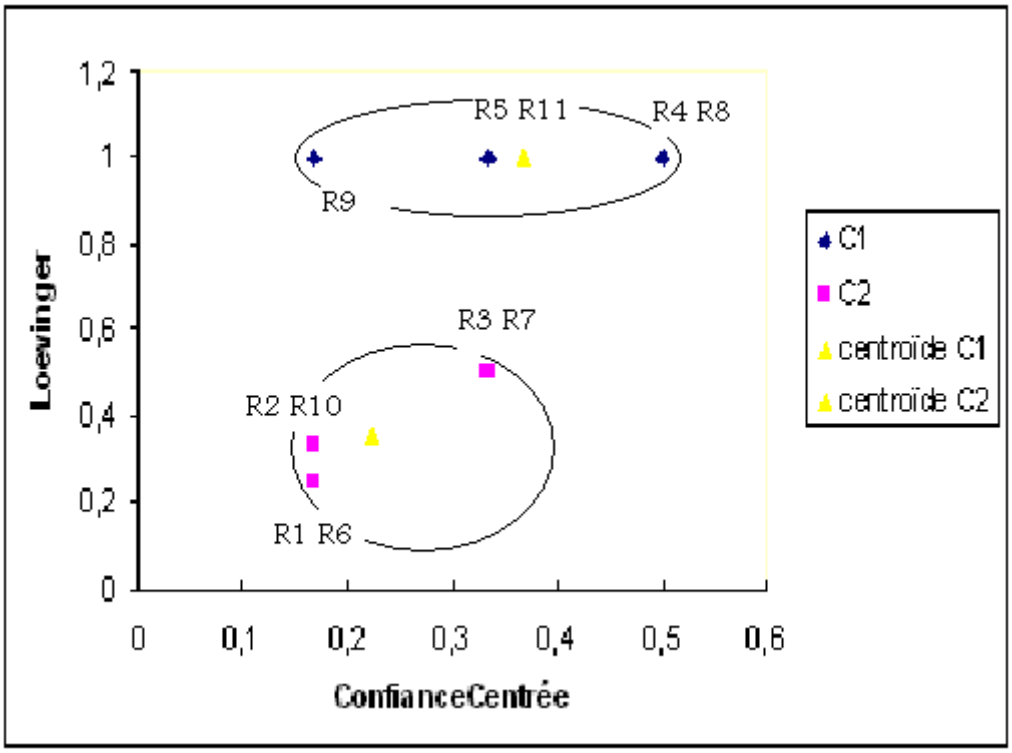

Figure 2 : Représentation des différentes clusters.

On constate que l'interprétation, ainsi que la validation des centres des clusters, peuvent être utilisées pour valider automatiquement les différentes règles appartenant aux clusters analysés. En effet, en examinant uniquement le centre de cluster C1, on constate que les différentes règles qu'il contient, ont une confiance =1 et par la suite une conviction qui tend vers l'infini. C'est qui constitue un avantage pour l'expert lors de la tâche de validation, en réduisant le coût cognitif de cette tâche.

4. Expérimentations

4.1. Résultats des expérimentations

Dans le cadre de notre travail, nous avons implémenté notre approche qui se base sur l'algorithme Apriori qui permet de rechercher l'ensemble des itemsets fréquents, à partir de ces derniers, nous avons implémenté un programme qui permet de générer les règles qui satisfait différents couples (mesure de qualité, seuil) et assigner pour chaque règle son vecteur de mesures de qualité. Enfin nous avons implémenté l'algorithme k-means qui permet de regrouper les règles qui semblent les plus proches sur la base des valeurs des vecteurs des mesures de qualité.

Le comportement des algorithmes d'extraction des itemsets fréquents varie selon les caractéristiques de la base de données à utiliser. Il existe deux types de base de données. Les bases de données faiblement corrélées et les bases de données corrélées. Les bases de données faiblement corrélées constituent des cas faciles pour l'extraction car peu des motifs sont fréquents. Donc, l'espace de recherche est très limité. Les données corrélées constituent des cas bien plus difficiles du fait de l'importante proportion de motifs fréquents [Bastide et al., 2000]. Dans notre étude, nous allons travailler sur deux bases de types différents et c'est pour discuter plus l'effet de ce facteur sur la performance de notre approche.

Les deux bases de données que nous allons utiliser, sont téléchargées à partir du UC Irvine Machine Learning Database Repository **(http://www.ics.uci.edu/~mlearn/)** :

La première base de données D_1, est corrélée. Elle est constituée de 6.483 enregistrements, avec une taille moyenne des objets de 41 items.

La deuxième base de données D_2, est faiblement corrélée. Elle est constituée de 3.196 enregistrements, avec une taille moyenne des objets de 37 items.

Les expériences sont effectuées sous la plateforme Windows, sur un PC équipé d'un processeur AMD Turion 64 ML à 1,6 GHz avec 1Go de mémoire DDR.

On a appliqué notre approche, qui est déjà implémentée sur la base D_1, en fixant les valeurs des seuils des mesures comme suit:
Le seuil de support *minsup* est de 86%.
Le seuil de confiance *minconf* est 70%.
Le seuil de confiance centrée *minconfcen* est 2%.
Le seuil du lift *minlif* est 1.
Le seuil de la conviction *minconv* est 1.
Le seuil du loevinger *minloe* est 0,1.

Le nombre des règles d'association extraite, en se limitant à l'utilisation du support et confiance est égale à 12.282, qui un nombre important et en utilisant les différentes couples (mesure, seuil) indiqués précédemment, le nombre des règles extraites est 4.847, ce dernier reste important malgré l'utilisation de ces mesures ! Ce qui constitue un vrai problème pour l'expert.

En appliquant notre approche qui consiste à représenter chaque règle extraite par son vecteur de mesure de qualité et d'appliquer un algorithme de clustering, avec ***K***=10, permettant de regrouper les règles qui semblent homogènes sur la base de leurs vecteurs de mesures.

Les résultats sont présentés par les figures suivantes :

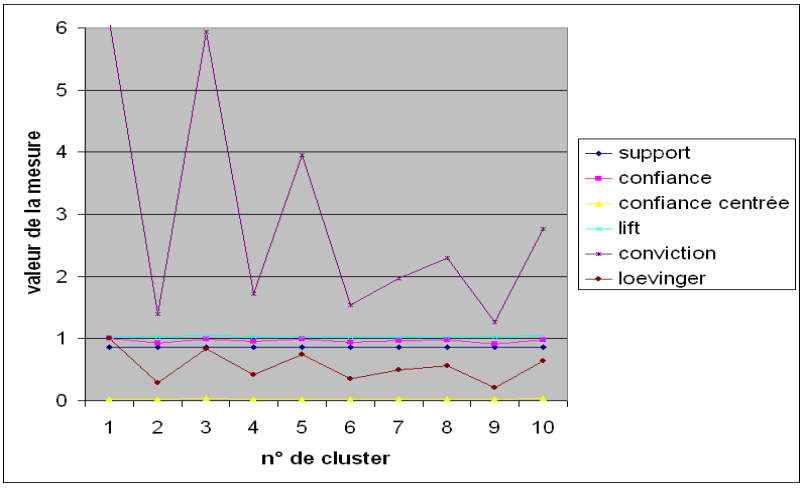

Figure 3 : Représentation des différentes centres des cluster en fonction des valeurs de ces mesures (D1).

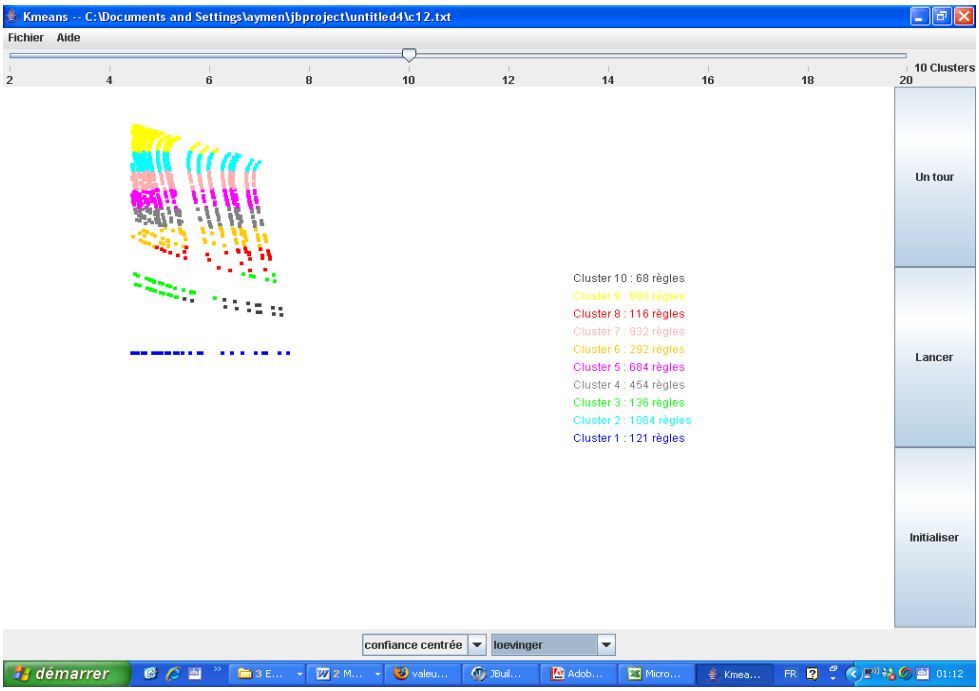

Figure 4 : Interface graphique représentant la distribution des différentes clusters en fonction des mesures de qualité utilisées(D1).

D'après les deux figures, en examinant juste les 10 vecteurs de mesures de qualité, qui représentent les centres des clusters, on remarque que les valeur des certaines mesures tel que : Le support, la confiance, la confiance centrée et le lift possèdent la même tendance et ont des valeurs très proche. Par contre, les valeurs de la conviction et le loevinger sont remarquables et varient d'un centre à l'autre. Ce qui peut être généralisé sur les autres règles appartenants aux différents clusters (Figure 3).

Le centre du cluster C_1 possède une confiance maximale égale à 100%, cette information peut être propagée sur les différentes règles appartenant à ce cluster, c'est-à-dire les 121 règles sont des règles exactes (Figure 4).

Dans le deuxième volet de cette section, nous allons considérer la base de données D_2. Cette base est faiblement corrélée. Le nombre

des règles générées, en utilisant l'approche support-confiance est de 1.778. En utilisant les mêmes seuils des mesures que D_1 (citer dans le volet précédent), le nombre des règles se réduit à 434 règles. En effet, même avec ce type de base (faiblement corrélée), le nombre des règles reste important.

En utilisant notre approche qui consiste à représenter chaque règle extraite par son vecteur de mesure de qualité et d'appliquer un algorithme de clustering, avec ***K***=10.

Les résultats sont présentés par les figures suivantes :

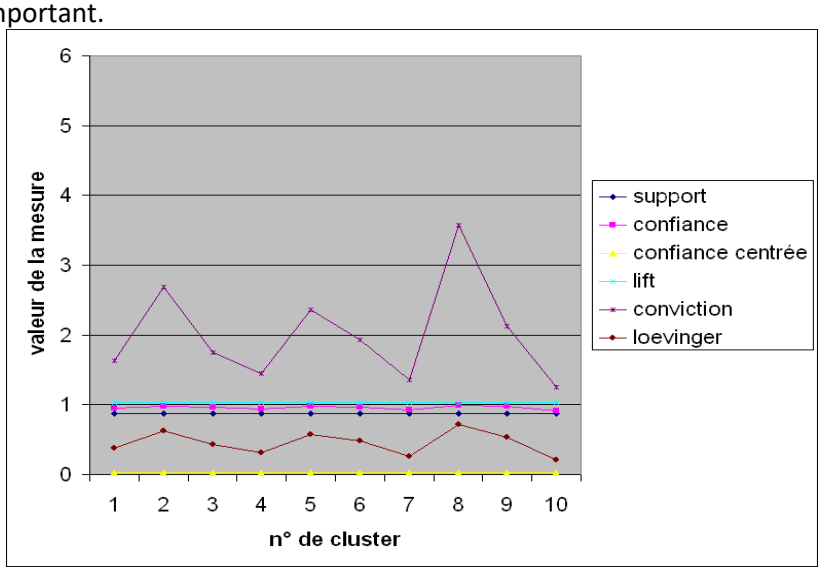

Figure 5 : Représentation des différentes centres des cluster en fonction des valeurs de ces mesures (D2).

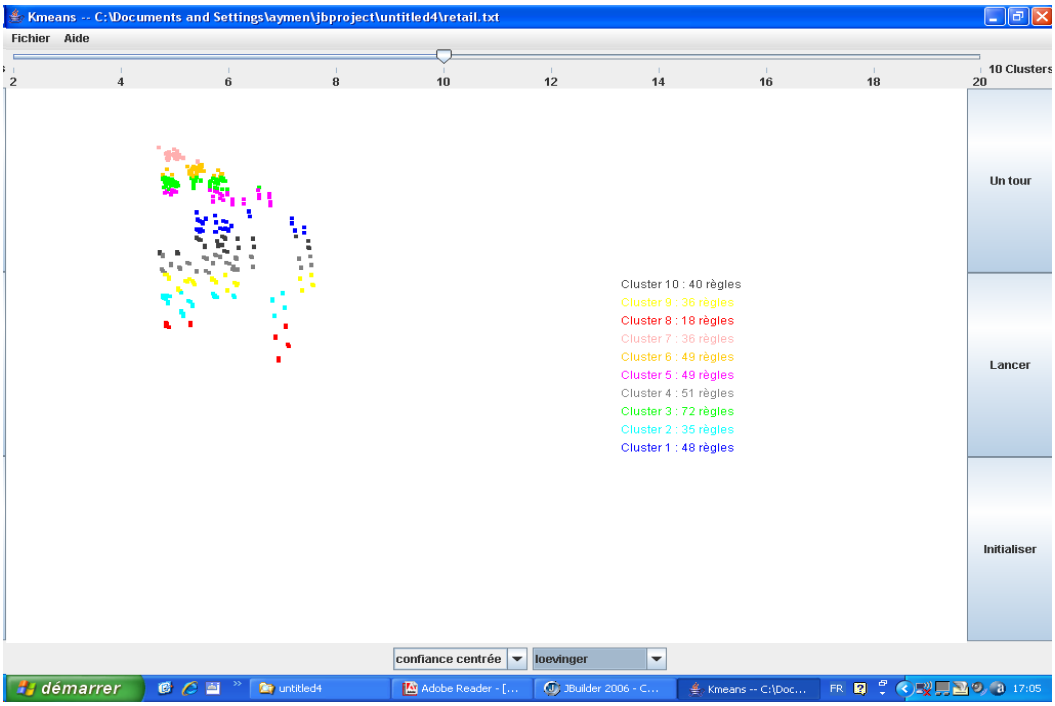

Figure 6 : Interface graphique représentant la distribution des différentes clusters en fonction des mesures de qualité utilisées(D2).

La figure 5 montre que le centre du cluster C_8 possède une conviction importante, cela signifie que la force de l'implication des règles qui appartiennent à ce cluster est meilleure.

En effet, une conviction élevée indique que le nombre de contre-exemples vérifiant chaque règle est inférieur à celui attendu sous l'hypothèse d'indépendance entre les attributs de la règle. D'une autre manière, la prémisse et la conclusion des règles tendent à apparaître ensemble. Ceci peut être généralisé sur les 18 règles du cluster C_8 (Figure 6).

Donc, en utilisant cette nouvelle approche, l'expert n'a que validé uniquement les différents centres des clusters et par suite, son interprétation est utilisée pour valider automatiquement les différentes règles appartenants aux clusters en question. Ce qui a permet de réduire le coût de l'évaluation et la validation des règles d'associations par l'expert en terme de temps.

4.2. Synthèse

D'après les expérimentations effectuées, nous pouvons déduire que le nombre des règles générées, en se limitant à l'utilisation du support et de la confiance est très important quelque soit le type de la base, ce nombre devient de plus en plus important avec les bases de données denses. Il en découle que les données de ventes sont éparses et faiblement corrélées alors que les données statistiques, spatiales ainsi que les données concernant l'historique d'accès Internet sont des données corrélées ou denses. Ces dernières représentent des données réelles [Pasquier, 2000]. Le recourt à d'autres mesures de qualité pour filtrer les règles a priori non intéressantes et donc réduire le volume de règles à analyser.

Malheureusement, même avec l'utilisation de ces mesures de qualité, l'utilisateur (expert des données ou analyste) se trouve confronté à deux problèmes majeurs : la quantité de règles reste important. De plus, cette méthode présente le défaut de ne pas toujours filtrer correctement les règles et introduit un biais non négligeable dans les règles supprimées (respectivement conservées) lié au choix des couples (mesure de qualité, seuil). Avec ces deux problèmes, la tâche de validation de l'expert, devient difficile.

En appliquant notre approche qui consiste à représenter chaque règle d'association par son vecteur de mesures de qualité. Ces derniers sont regroupés, à l'aide d'une méthode de clustering. Ensuite, une fois que les clusters sont formés, l'expert intervient uniquement pour valider les différents centres (vecteurs de mesures de qualité) associés aux clusters.

La validation réalisée par l'expert pourra être utilisée pour évaluer automatiquement les règles appartenant aux clusters analysés. Ce qui facilite la tâche d'évaluation et de validation des règles par l'expert du domaine et par la suite, réduire le coût cognitif de cette tâche.

5. Conclusion et Perspectives

Dans le cadre de notre travail de recherche, nous nous sommes intéressés à l'extraction des itemsets fréquents à partir de bases de données transactionnelles. Ce thème de recherche fait partie du domaine du data mining et plus généralement de l'ECD.

Les algorithmes d'extraction des itemsets fréquents se basent sur deux propriétés : le support et la confiance introduit par l'algorithme Apriori dans [Agrawal et Srikant, 1994]. Le nombre des règles générées ne permet pas aux utilisateurs eux-mêmes de sélectionner les règles pertinentes. Une manière de réduire le coût cognitif de cette tâche consiste à les guider à l'aide de mesures de qualité adaptées à ses préférences. Nous avons constaté que l'utilisation des différents mesures de qualité, présente le défaut de ne pas toujours filtrer correctement les règles pertinentes et introduit un biais non négligeable dans les règles supprimées (respectivement conservées) lié au choix des couples (mesure de qualité, seuil).et que même avec l'utilisation des ces mesures de qualité le nombre des règles reste important. Ces problèmes constituent le cadre de notre nouvelle approche, qui consiste à : représenter chaque règle d'association par un vecteur de mesures de qualité. Puis d'utiliser un algorithme de clustering pour regrouper les règles suivant les valeurs des mesures de qualité. Ensuite, chaque groupe (cluster) étant décrit par son centre, l'expert intervient pour valider les différents vecteurs (centres) associés aux clusters. La validation réalisée par l'expert pourra être utilisée pour évaluer automatiquement les règles appartenant aux clusters analysés. Et comme travaux futurs, il est intéressant d'étendre cette étude en

incluant diverses stratégies de propagation des informations issues des expertises, d'essayer de combiner les différentes mesures de qualité existantes avec les différentes mesures de clustering. Un autre élément important à prendre en compte lors de la mise en oeuvre d'une méthode de clustering est la possibilité que les valeurs de certaines données sur certains attributs ne soient pas renseignées, ainsi que la possible présence de données bruitées, c'est-à-dire d'exemples qui ne suivent pas la distribution générale des exemples sur leur espace de description.

6. Bibliographie

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. (1994). Fast algorithms for mining association rules. In 20th VLDB Conference, Santiago, Chile, Sept, 1994.

[Azé, 2003] J. Azé. (2003). Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. RSTI série RIA-ECA, 17(1-2-3):171-182.

[Bastide et al., 2000] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme et L. Lakhal. (2000). Mining Frequent Patterns with Counting Inference. SIGKDD Explorations, vol 2, n°2, 66-75, ACM Computer, december 2000.

[Bastide et al., 2002] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme et L. Lakhal. (2002). Pascal : un algorithme d'extraction des motifs fréquents. Techniques et Science Informatiques, 21(1):65-95, Hermès Science, march 2002.

[Brin et al., 1997] S. Brin, R. Motwani, et C. Silverstein. (1997). Beyond market baskets : Generalizing association rules to correlations. Dans Proceedings of ACM SIGMOD'97, pages:265-276.

[Freitas, 1999] A. A. Freitas. (1999).On rule interestingness measures, Knowledge-Based Systems 12(5-6), pp 309-315, 1999.

[Guillaume, 2000] S. Guillaume. (2000). Traitement des données volumineuses Mesures et algorithmes d'extraction de règles d'association et règles ordinales. PhD thesis, Université de Nantes.

[IBM, 1996] IBM. (1996). IBM Intelligent Miner User's Guide.Version 1 Release 1, SH12-6213-0o, edition.

[Lallich et Teytaud, 2004] S. Lallich et O. Teytaud. (2004). Evaluation et validation de l'interêt des règles d'association. Revue des Nouvelles Technologies de l'Information, 2004.

[Lenca et al., 2004] P. Lenca, P. Meyer, B. Vaillant, P. Picouet et S. Lallich. (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. Mesures de qualité pour la fouille de données, n° spécial RNTI Revue des Nouvelles Technologies de l'Information, Cepaduès(2004).

[Liu et al., 1999] B. Liu, W. Hsu, L. Mun et H. Lee. (1999). Finding interestingness patterns using user expectations, IEEE.Trans. on Knowl,. and Data Mining (11), pp 817-832, 1999.

[Loevinger, 1947] J. Loevinger. (1947). A. systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 61:1-49.

[MacQueen, 1967] J. MacQueen. (1965). Some methods for classification and analysis for multivariate observations. Dans Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281-297, Berkeley, CA. University of California Press.

[Padmanabhan et Tuzhilin, 1998] B. Padmanabhan et A. Tuzhilin. (1998). A belief-driven method for discovering unexpected patterns, Proc. Of KDD'98, pp 94-100, 1998.

[Park et al., 1995] J. S. Park, M. Chen, et P. S. Yu. (1995). An effective hash based algorithm for mining association rules. In ACM SIGMOD International Conference on Management of Data, May 1995.

[Pasquier, 2000] N. Pasquier. (2000). Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données. Thèse de Doctorat, université Blaise Pascal - Clermont-Ferrand II, LIMOS, Janvier 2000. 223 pages.

[Pasquier et al., 1999] N. Pasquier, Y. Bastide, R. Taouil et L. Lakhal. (1999). Pruning Closed Itemset Lattices for Association Rules. Proc. BDA'98 conference, pp 177-196, october 1998.

[Plasse et al., 2006] M. Plasse, N. Niang, G. Saporta et L. Leblond. (2006). Une comparaison de certains indices de pertinence des règles d'association . In EGC06, Lille, 18-20 janvier, pp. 561-568, Cepadues, 2006.

[Savasere et al., 1995] A. Savasere, E. Omiecinski et S. B. Navathe. (1995). An

efficient algorithm for mining association rules in large databases. In 21st VLDB Conference, 1995.

[Tan et al., 2002] P. Tan, V. Kumar et J. Srivastava. (2002). Selecting the right interestingness measure for association patterns. In Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, pp. 32-41.

[Zhao et al., 2005] Y. Zhao, C. Zhang et S. Zhang. (2005). Discovering interesting association rules by clustering. Australian joint conference on artificial intelligence No17, Cairns, AUSTRALIE. 2004, vol. 3339, pp. 1055-1061.