Extraction de Règles d'Association Quantitatives Application à des Données Médicales

Cyril Nortet* Ansaf Salleb**

Teddy Turmeaux* Christel Vrain*

*LIFO Rue Léonard de Vinci BP 6759 45067 Orléans Cedex 02
{Cyril.Nortet, Teddy.Turmeaux, Christel.Vrain}@lifo.univ-orleans.fr

**IRISA/INRIA, Projet Dream, Campus de Beaulieu, Rennes
Ansaf.Salleb@irisa.fr

Résumé. L'extraction de règles d'association est devenue aujourd'hui une tâche populaire en fouille de données. Cependant, l'algorithme Apriori et ses variantes restent dédiés aux bases de données renfermant des informations catégoriques.
Nous proposons dans cet article QuantMiner, qui est un outil que nous avons développé dans le but d'extraire des règles d'association gérant variables catégoriques et numériques. L'outil que nous proposons repose sur un algorithme génétique permettant de découvrir de façon dynamique les intervalles des variables numériques apparaissant dans les règles.
Nous présentons également une application réelle de notre outil sur des données médicales relatives à la maladie de l'athérosclérose et donnons des résultats de notre expérience pour la description et la caractérisation de cette maladie.

Mots clé : Fouille de Données, Règle d'Association, Attribut Numérique, Discrétisation, Attribut Catégorique, Algorithme Génétique.

1 Introduction

L'extraction de règles d'association est devenue aujourd'hui une tâche populaire en fouille de données. Elle a pour but de dégager des relations intelligibles entre des attributs dans une base de données. Une règle d'association (Agrawal et al. 1993) est une implication $\mathcal{C}_1 \Rightarrow \mathcal{C}_2$, où $\mathcal{C}_1$ et $\mathcal{C}_2$ expriment des conditions sur les attributs de la base de données. La qualité d'une règle est classiquement évaluée par un couple de mesures *support* et *confiance*, définis comme suit :

- Support($\mathcal{C}$), où $\mathcal{C}$ exprime des conditions sur les attributs, est le nombre de n-uplets (lignes de la table) qui satisfont $\mathcal{C}$.
- Support($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = Support($\mathcal{C}_1 \wedge \mathcal{C}_2$)
- Confiance($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = $\frac{\text{Support}_{(\mathcal{C}_1 \wedge \mathcal{C}_2)}}{\text{Support}_{(\mathcal{C}_1)}}$

Étant donnés deux seuils, MINSUPP et MINCONF, une règle est dite *solide*, si son support et sa confiance dépassent MINSUPP et MINCONF respectivement.

Quoique très utilisés, l'algorithme *Apriori* et ses variantes restent dédiés aux bases de données renfermant des informations catégoriques (dites aussi symboliques ou encore

qualitatives). Or, dans la pratique, les bases de données réelles traitent non seulement de variables catégoriques mais aussi numériques.

Nous proposons dans cet article QuantMiner, outil développé en collaboration avec le BRGM, Bureau de Recherches Géologiques et Minières (service REM/VADO) dans le but d'extraire des règles d'association quantitatives, exprimant des corrélations entre des variables qui peuvent être catégoriques et/ou numériques. L'outil que nous proposons repose sur un algorithme génétique permettant de découvrir de façon dynamique les intervalles des variables numériques qui optimisent le support et la confiance d'une règle d'association. L'idée de base est la suivante : QuantMiner engendre un ensemble de schémas de règles et cherche pour chaque règle les meilleurs intervalles pour les attributs numériques la composant. La discrétisation obtenue varie donc pour chaque schéma et dépend des attributs catégoriques et numériques qui le composent.

Divers travaux ont déjà porté sur l'apprentissage de règles d'association quantitatives. Plusieurs méthodes ont été proposées, reposant sur une discrétisation *a priori* des attributs quantitatifs (voir par exemple (Lent et al. 1997, Srikant et Agrawal 1996)), ou sur une discrétisation pendant le processus d'apprentissage. La discrétisation peut alors être effectuée soit pendant la phase de génération des règles (Fukuda et al. 1996a, Rastogi et Shim 1999), mais dans ce cas le format des règles est souvent très limité, soit pendant la phase de génération des itemsets fréquents (Mata et al. 2002).
À notre connaissance, seul ce dernier travail repose sur l'utilisation d'un algorithme génétique. La problématique est cependant différente dans la mesure où l'algorithme génétique est dans leur cas, utilisé pour recherche des itemsets fréquents optimisant un critère de qualité fondé principalement sur le support. QuantMiner cherche à trouver des intervalles numériques optimisant la qualité (support et confiance) d'une règle.

QuantMiner est un outil interactif et convivial. Il a été utilisé sur deux applications, l'une concerne la recherche de règles d'association dans un Système d'Information Géographique développé par le BRGM, la seconde concerne des données médicales relatives à la maladie de l'athérosclérose. Nous présentons des résultats obtenus sur cette seconde application.

Le papier est organisé comme suit. Dans la section suivante, nous faisons un état de l'art des divers travaux portant sur l'apprentissage de règles d'association quantitatives. Dans la section 3, nous décrivons l'outil QuantMiner que nous avons développé. La section 4 est dédiée à l'application de QuantMiner sur des données médicales. La section 5 discute des améliorations possibles de notre approche ainsi que des perspectives que nous envisageons dans cette thématique.

2 Règles d'Association Quantitatives

Cette section, inspirée de (Salleb 2003) et de (Nortet), est d'une part une synthèse des principaux travaux d'extraction de règles d'association quantitatives et d'autre part apporte un point de vue critique des approches proposées dans la littérature.

L'algorithme Apriori (Agrawal et Srikant 1994) et ses variantes (voir par exemple (Brin et al. 1997, Bayardo 1998, Zaki 2000, Salleb et al. 2002)) ont été conçus pour des bases de données booléennes, c'est à dire dont les items sont catégoriques et prennent leurs valeurs dans un ensemble discret, de cardinal fini. Mais dans la plupart des cas, les

données sont décrites dans un formalisme attribut-valeur et stockées dans une relation (table). Un attribut peut être de type :

1. catégorique, par exemple *couleur* est un attribut qui prend ses valeurs dans l'ensemble {*bleu, rouge, vert*}; dans ce cas, par exemple, *(couleur=bleu)* est considéré comme un item,
2. numérique discret, *i.e.* ayant peu de valeurs numériques différentes les unes des autres; par exemple, l'attribut *nb_voitures* a pour valeurs possibles {*0, 1, 2, 3*}, on peut former l'item *(nb_voitures=2)*,
3. numérique continu défini sur une large plage de valeurs, comme par exemple *âge* dont le domaine est *[0,120]*, un item possible serait, par exemple, *âge* ∈ *[20,40]*.

Pour les deux premiers types d'attributs, pour passer de la relation aux items, il suffit d'énumérer les modalités de l'attribut et de former pour chaque valeur possible l'égalité (attribut=valeur). En revanche, ceci n'est pas envisageable pour les attributs numériques continus dans le cadre de l'extraction de règles d'association. En effet, c'est rarement sur une valeur particulière, peu fréquente (souvent n'apparaissant qu'une seule fois), que l'on générera des règles d'association de qualité, les valeurs numériques étant souvent toutes différentes d'un n-uplet à un autre. C'est plutôt dans un intervalle de valeurs que des règles intéressantes de support suffisant peuvent être découvertes.

Une règle d'association quantitative est une règle dont l'antécédent et le conséquent sont des conjonctions d'items de la forme $(A_i = v_i)$ pour les attributs catégoriques ou numériques discrets, ou de la forme $A_i \in [l_i, u_i]$ pour les attributs numériques (v_i, l_i et u_i sont des valeurs du domaine de l'attribut A_i). En voici un exemple :

$$\text{âge} \in [27, 38] \text{ et (marié=oui)} \implies \text{(nb\_voitures=2)}$$

Il va sans dire que le choix de la technique de découpage des attributs numériques en intervalles a des conséquences très importantes sur la qualité des règles générées.

Nous pouvons classer les approches proposées dans la littérature pour l'extraction de règles quantitatives en trois grandes catégories :

Approche fondée sur une discrétisation préalable Un attribut numérique continu est souvent discrétisé avant l'étape d'extraction de règles. La discrétisation consiste à découper son domaine en un nombre fini d'intervalles. L'attribut, rendu ainsi catégorique, peut donc être utilisé dans l'extraction de règles d'association au même titre que les attributs non continus. Parmi les méthodes les plus standard de discrétisation, citons :
- la discrétisation en k intervalles de largeur uniforme (equi-width), où chaque intervalle prend la même proportion du domaine,
- la discrétisation en k intervalles de fréquences égales (equi-depth), où à chaque intervalle correspond un même nombre de n-uplets de la relation,
- la discrétisation non régulière qui s'appuiera sur des connaissances du domaine.

La difficulté de ces méthodes de découpage réside dans le choix du paramètre k pour les deux premières et de la disponibilité de connaissances *a priori* pour la troisième.
Le choix d'une pré-discrétisation dans le contexte d'extraction des règles d'association a été soulevé et discuté dans (Srikant et Agrawal 1996). En effet, en se basant sur de trop petits intervalles, on risque d'omettre des règles pour insuffisance de support, alors que si les intervalles sont trop grands, c'est par défaut de confiance que nous sommes

susceptibles de les manquer. Le choix de la discrétisation influe de manière opposée sur les deux paramètres vitaux du problème d'extraction de règles.

La pré-discrétisation est souvent utilisée conjointement avec des techniques de fusion d'intervalles et de regroupement pour améliorer le partitionnement (Lent et al. 1997, Miller et Yang 1997, Srikant et Agrawal 1996, Wang et al. 1998, Zhang et al. 1997). Néanmoins, cette approche n'est pas satisfaisante car elle est sensible au bruit et les attributs sont discrétisés indépendamment les uns des autres. De plus, rien ne prouve que la discrétisation préliminaire reflète bien la distribution des données et ne nous fait pas manquer des règles à fort potentiel.

Approche guidée par des schémas de règles Ici, les attributs numériques ne sont pas, à proprement parler, discrétisés. Une nouveauté dans les travaux de Fukuda et al. présentés dans (Fukuda et al. 1996a, Fukuda et al. 1996b), est que le mot clé n'est plus *discrétisation* mais plutôt *optimisation*. Ces auteurs ont développé des méthodes où l'utilisateur spécifie la forme de l'antécédent et du conséquent de la règle. Les règles d'association sont dans ce cas restreintes à des règles où les attributs du membre gauche et droit de la règle sont instanciés sauf *un ou deux attributs numériques* dans la partie gauche de la règle. Il faut ensuite trouver les intervalles pour ces attributs numériques non instanciés qui maximisent le support ou la confiance ou toute autre mesure d'intérêt. Dans ces travaux, une nouvelle mesure très intéressante nommée *Gain* est introduite. Elle met en relation le support et la confiance et mesure une sorte de compromis entre les deux.

$$\text{Gain}(A \Rightarrow B) = \text{Supp}(AB) - \text{MinConf} * \text{Supp}(A)$$

Les supports sont donnés en nombre d'occurrences et non en fraction. Cette mesure fut par la suite reprise dans (Rastogi et Shim 1999), où deux algorithmes fondés sur l'optimisation du gain sont proposés et permettent d'introduire la disjonction de plusieurs intervalles dans les règles gérant, là aussi, au plus deux attributs numériques.

Une autre vision du problème fondée sur la distribution des attributs continus a été proposée dans (Aumann et Lindell 1999, Webb 2001). Ici, les distributions statistiques (moyenne, variance, écart-type, minimum etc.) des attributs numériques sont autorisées dans la partie droite d'une règle. Plus précisément, deux sortes de règles sont traitées dans ce travail : dans le membre gauche de la règle un ensemble d'attributs catégoriques et à droite la distribution de plusieurs attributs numériques, ou encore un seul attribut numérique à gauche et la distribution d'un seul attribut numérique à droite. Par exemple :

$$\text{Région=Sud} \Longrightarrow \text{Salaire : moyenne} = 1200 \text{ euros / mois}$$

Comme le précisent les auteurs de ces travaux, quoique intéressante, la forme des règles reste néanmoins restreinte.

Approche reposant sur un algorithme génétique L'optimisation est également la voie choisie dans les travaux de (Mata et al. 2002). Seulement, ici il n'est pas question d'optimiser directement des règles mais simplement de trouver les itemsets fréquents, pour ensuite générer les règles selon la méthode traditionnelle. L'algorithme génétique proposé dans leurs travaux est classique. Un individu est codé par une liste de couples (attribut numérique, disjonction d'intervalles). Comme tout algorithme évolutionnaire,

la qualité des individus (qui sont donc des itemsets) est évaluée par une mesure permettant d'optimiser le support des itemsets, tout en veillant à ne pas retenir les domaines entiers des attributs numériques et à favoriser les itemsets les plus spécifiques.
Il est clair, d'après la forme de la représentation d'un individu, que l'algorithme extrait d'un seul coup plusieurs fréquents numériques. Par contre, il ne prend en compte aucun attribut catégorique, ce qui limite l'applicabilité d'une telle approche. Rajoutons à cela que seul le support est optimisé, ce qui est insuffisant.

Cette étude bibliographique, exposant les principaux travaux dans ce domaine, nous a permis de ressentir la difficulté du problème des règles quantitatives mais surtout de réaliser que pour approcher ce problème difficile, il faut faire nécessairement des compromis, comme nous le verrons dans la section suivante dédiée à notre approche.

3 QuantMiner

À la lumière de l'étude bibliographique exposée dans la section précédente, nous avons d'abord écarté l'idée d'utiliser une pré-discrétisation des variables numériques. En fait, elle rigidifie tout le reste du processus d'extraction; c'est en définitive une mauvaise façon d'appréhender le problème. En revanche, nous avons retenu l'idée de travailler directement sur **l'optimisation de règles quantitatives** par algorithme génétique. Notre choix de l'*optimisation*, comme quelques travaux cités plus haut, se justifie par le fait qu'on travaille sur des variables numériques, continues qui donnent au problème une complexité telle qu'on ne peut songer à un algorithme exact.

Il convient ici de montrer clairement ce qui démarque notre approche des travaux de Mata et al. D'une part, nous prenons en compte les attributs numériques et catégoriques, et d'autre part, optimiser seulement le support des itemsets nous semble insuffisant. C'est pourquoi, nous proposons de travailler directement sur les règles en utilisant le support et la confiance.

Ainsi, nous considérons des **schémas de règles**. Un schéma de règle est une règle présentant dans chacun de ses membres gauche et droit des items catégoriques aux valeurs fixées et des items numériques dont les intervalles correspondant ne sont pas encore instanciés. Puis par optimisation nous cherchons les bornes les plus adaptées pour chacun de ses intervalles, en prenant en compte la mesure du Gain (Fukuda et al. 1996a). L'algorithme génétique que nous proposons suit une organisation tout à fait traditionnelle (donnée ci-dessous), dont voici les mécanismes adaptés au problème :

- **Représentation d'un individu** : Un individu est représenté par un ensemble de couples $(attribut_i, [l_i, u_i])$, où $attribut_i$ est le i$^{\text{ème}}$ attribut numérique en partant de la gauche du schéma de règle à optimiser. Lors de la génération de la population initiale, on fabrique des individus dont les intervalles sont de supports décroissants, pour garantir une bonne diversité et exprimer des règles aussi bien générales que spécifiques.
- **Fonction de croisement** : Pour ajouter un peu de variété à ces règles initiales, nous procédons par cette opération à l'échange aléatoire entre parents. Pour chaque attribut numérique, l'intervalle correspondant chez l'enfant sera,

avec une certaine probabilité, ou bien celui du père, ou bien celui de la mère, ou alors un mélange des deux (Figure 1).

- **Fonction de mutation** : La mutation a pour rôle d'affiner les règles. Il s'agit de déplacer chaque borne, par augmentation ou réduction (Figure 1). Le déplacement se fait dans le domaine trié de l'attribut numérique et ne doit pas permettre d'englober ou d'écarter plus de 10% de n-uplets couverts par l'intervalle.

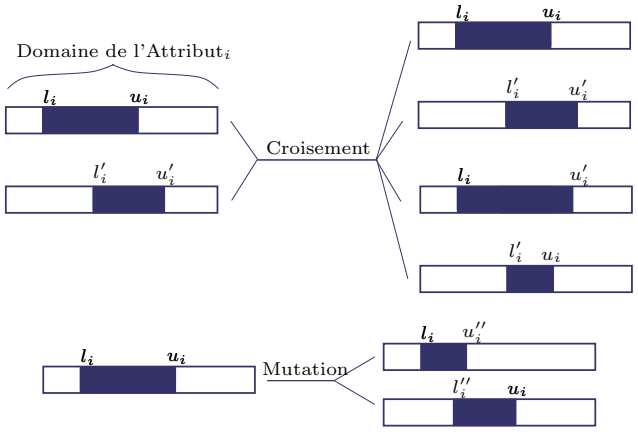

FIG. 1 – Opérateurs génétiques utilisés dans QUANTMINER

- **Fonction de qualité**
  La qualité repose avant tout sur la valeur du gain. Si celui-ci est positif, on va tenter d'en affiner les intervalles en favorisant ceux d'amplitudes faibles. Pour accentuer la pénalité sur un intervalle d'amplitude large, on élève au carré sa proportion (rapport de son amplitude sur l'amplitude du domaine). Une pénalité supplémentaire est attribuée aux règles qui ne dépassent pas le seuil de support : on diminue la qualité du nombre de n-uplets (valeur qui est en somme dans la même unité que le gain), pour bien s'assurer qu'un tel individu sera vite écarté de la population.

```
Qualité(A=>B)=
   {   QualitéTemporaire = Gain(A=>B);
       Si QualitéTemporaire >= 0
       Alors   Pour tout intervalle I
               Faire QualitéTemporaire * = (1-Proportion(I))**2;
       Si Support(A=>B) < MinSupp
       Alors QualitéTemporaire - = NbEnregistrements;
       Renvoyer QualitéTemporaire; }
```

```
Algorithme QuantMiner

Entrée:  une relation (table), POP_SIZE, GEN_NB, MINSUPP, MINCONF
Sortie:  Règles d'association quantitatives
 1.  Choisir un ensemble d'attributs de la relation
 2.  Choisir un ensemble de schémas de règles sur ces attributs
 3.  Calculer les itemsets fréquents sur les attributs catégoriques
     apparaissant dans les schémas[1]
 4.  Pour chaque schéma de règle
 5.      Générer une population aléatoire de taille POP_SIZE
 6.      Itérer les étapes suivantes GEN_NB fois
 7.          Former une nouvelle génération de population
             par croisements et mutations
 8.          Garder les règles de bonne qualité
```

Implémentation QuantMiner a été développé en Java par Cyril Nortet au BRGM. Le logiciel prend en charge des fichiers de données au format DBF et prend la forme d'un assistant (wizard), limité à 5 étapes :
1) Sélection des attributs et définition de leur type
2) Répartition fine des items ($A_i = v_i$ ou $A_i \in [l_i, u_i]$, les valeurs l_i, u_i non instanciées, v_i instanciée ou non) aux places où on souhaite les voir apparaître dans les règles.
3) Choix de la technique d'optimisation et réglage des paramètres de celle-ci.
4) Exécution de l'algorithme d'optimisation.
5) Affichage des résultats, avec tri sélectif des règles produites.

Dans QuantMiner, le processus de fouille de données est interactif; l'utilisateur qui est dans ce cas tout *près des règles*, précisant celles qui l'intéressent. Le processus est aussi itératif dans la mesure où l'utilisateur peut sauvegarder son contexte d'extraction (schémas, méthode et paramètres) et recommencer le processus ultérieurement.

En plus du support minimum et de la confiance minimum, les paramètres de l'algorithme génétique sont la taille de la population, le nombre de générations, les taux de mutations et de croisements. Nous avons fixé ces valeurs par défaut à 250 individus, 150 générations, 40% de mutations et 50% de croisements. Ce choix expérimental permet d'obtenir des intervalles optimisés relativement stables d'une exécution à une autre de l'application. De nombreuses optimisations dans l'évaluation des règles ont été intégrées dans le cœur du logiciel. Nous ne pouvons donner tous les détails d'implémentation faute de place, mais à titre indicatif, avec les paramètres par défaut, le logiciel parvient à traiter en moyenne un schéma de règle par seconde sur une base de 2 500 n-uplets.

Exemple Nous avons effectué à titre expérimental des tests sur un jeu de données classique : les Iris de Fisher (Fisher 1936, Murphy et Aha 1995). Il est composé de 150 iris répartis en trois familles de proportions identiques : *Setosa, Versicolor et Virginica*. Chaque fleur est décrite par un attribut catégorique `ESPÈCE` indiquant la famille de

[1] Cette étape, semblable à Apriori, permet d'instancier les attributs catégoriques dans les schémas.

la fleur, ainsi que par 4 attributs numériques mesurant la longueur et largeur de son sépale, longueur et largeur de son pétale. Considérons le schéma suivant :

ESPÈCE = valeur ⇒ Larg._Pét.$\in [l_1, u_1]$ Larg._Sép.$\in [l_2, u_2]$
Long._Pét.$\in [l_3, u_3]$ Long._Sép.$\in [l_4, u_4]$ supp%-conf%

QuantMiner a généré les règles suivantes (voir aussi Figure 2). Elles sont cohérentes avec les statistiques descriptives de chacune des espèces données en Table 1.

ESPÈCE = Setosa ⇒ Larg._Pét.∈[1; 6] Larg._Sép.∈[31; 39]
Long._Pét.∈[10; 19] Long._Sép.∈[46; 54] 23%-70%
ESPÈCE = Versicolor ⇒ Larg._Pét.∈[10; 15] Larg._Sép.∈[22; 30]
Long._Pét.∈[35; 47] Long._Sép.∈[55; 66] 21%-64%
ESPÈCE = Virginica ⇒ Larg._Pét.∈[18; 25] Larg._Sép.∈[27; 33]
Long._Pét.∈[48; 60] Long._Sép.∈[58; 72] 20%-60%

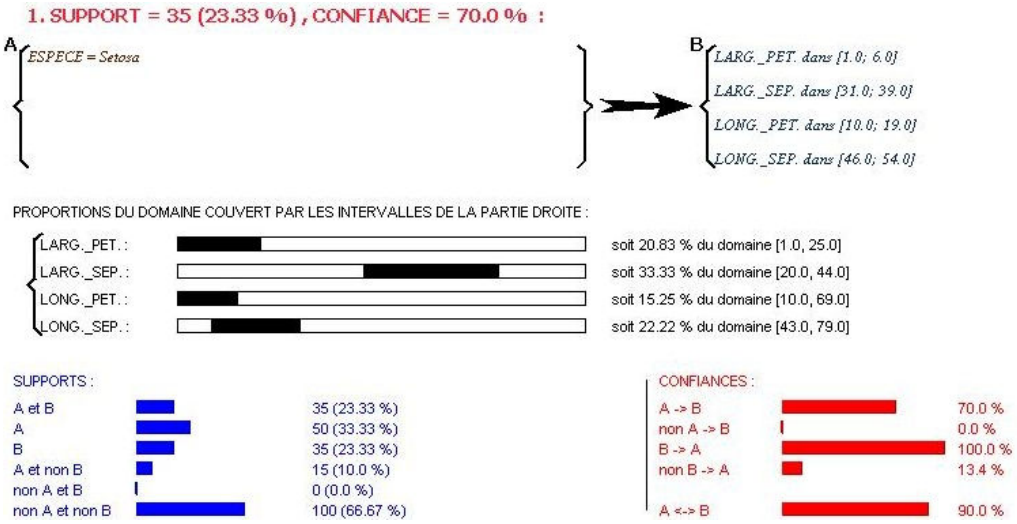

Fig. 2 – Exemple de règle telle que produite par QuantMiner

Espèce	Mesure	Min	Max	Moyenne	Écart-type
Setosa	Larg._Pét.	1	6	2.46	1.05
	Larg._Sép.	23	44	34.28	3.79
	Long._Pét	10	19	14.62	1.74
	Long._Sép	43	58	50.6	3.52
Versicolor	Larg._Pét.	10	18	13.26	1.98
	Larg._Sép.	20	34	27.7	3.14
	Long._Pét	30	51	42.6	4.70
	Long._Sép	49	70	59.39	5.16
Virginica	Larg._Pét.	14	25	20.26	2.75
	Larg._Sép.	22	38	29.74	3.22
	Long._Pét	45	69	55.52	5.52
	Long._Sép	49	79	65.88	6.36

Tab. 1 – Description statistique des trois espèces d'iris

4 Application à des données médicales

L'athérosclérose est une maladie répandue et grave des artères dont les parois se durcissent provoquant une gêne considérable à la circulation du sang.
Le projet STULONG[2] porte sur une étude médicale effectuée pendant 20 ans sur les facteurs de risque de cette maladie. Elle concerne une population de patients composée de 1 419 hommes adultes qui ont été classés en trois groupes : le groupe des patients normaux N, le groupe des patients à risque R et enfin le groupe P des patients présentant la pathologie. Nous nous sommes intéressés à l'extraction de règles descriptives et caractéristiques de cette maladie dans le cadre de ce projet.

Afin de décrire les patients selon leur groupes et selon qu'ils soient décédés ou non, nous nous sommes concentrés sur des règles de la forme[3] :

$$\texttt{Ensemble de patients} \Longrightarrow \texttt{descriptions}$$

Par exemple :

$$\texttt{DEATH?} = \texttt{valeur} \Longrightarrow \texttt{descriptions}$$
$$\texttt{GROUP} = \texttt{valeur} \Longrightarrow \texttt{descriptions}$$

Nous avons ainsi cherché à décrire des patients par rapport à leur consommation d'alcool, de tabac, à leur poids, à leur situation sociale, leurs activités physiques ainsi qu'aux examens cliniques, biochimiques et physiques effectués lors de leur entrée dans l'étude. La Table 2 donne un échantillon de règles générées par QuantMiner.

Règle ($A \Rightarrow B$)	**Supp**($A \Rightarrow B$)	**Conf**($A \Rightarrow B$)	**Conf**($\neg A \Rightarrow B$)
GROUP = N $\Longrightarrow$ ALCO_CONS $\in$ [1.0; 1.2] $\wedge$ BMI $\in$ [19.73; 27.77] $\wedge$ TOBA_CONSO $\in$ [0.0; 0.5]	13%	69%	19%
GROUP = R $\Longrightarrow$ ALCO_CONS $\in$ [1.0; 1.29] $\wedge$ BMI $\in$ [22.28; 30.72] $\wedge$ TOBA_CONSO $\in$ [0.5; 1.25] $\wedge$ TOBA_DURA $\in$ [15.0; 20.0]	39%	64%	38%
GROUP = P $\Longrightarrow$ ALCO_CONS $\in$ [1.0; 1.53] $\wedge$ BMI $\in$ [21.98; 33.14] $\wedge$ TOBA_CONSO $\in$ [0.5; 1.25]	5%	64%	60%
DEATH? = yes $\Longrightarrow$ ALCO_CONS $\in$ [1.0; 1.28] $\wedge$ TOBA_CONSO $\in$ [0.5; 1.25] $\wedge$ TOBA_DURA $\in$ [15.0; 20.0]	18%	68%	53%
DEATH? = no $\Longrightarrow$ ALCO_CONS $\in$ [1.0; 1.23] $\wedge$ TOBA_CONSO $\in$ [0.0; 0.85]	49%	67%	55%

Tab. 2 – Exemples de règles découvertes par QuantMiner

[2] « The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boud ik, MD, ScD, with collaboration of M. Tomeckova, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvarova, DrSc). The data resource is on the web pages http://euromise.vse.cz/STULONG. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107. »

[3] N'avoir qu'une condition dans le membre gauche n'est pas une limitation de QuantMiner

Nos expérimentations ont montré que l'excès de poids, la consommation excessive de tabac pendant une longue durée, des taux de cholestérol, de triglycérides et des pressions artérielles systoliques et diastoliques élevés, ainsi que la présence de plis graisseux sous certains muscles sont des facteurs de risques importants. Les résultats, publiés dans (Salleb et al. 2004), ont été jugés intéressants lors du *Discovery Challenge* organisé autour de ces données et dans lequel des médecins participent et scrutent les connaissances apprises par les diverses contributions.

5 Discussion et Conclusion

QuantMiner ne se contente pas de donner la confiance d'une règle de la forme $A \Rightarrow B$, mais il donne aussi la confiance de $\neg A \Rightarrow B$, $B \Rightarrow A$, $\neg B \Rightarrow A$ (Figure 2). Cependant, lors de l'application aux données médicales, il nous a semblé que ces critères devaient être affinés dans le cas de la caractérisation de plusieurs classes. Dans le cas où A représente des caractéristiques des patients à risques, $\neg A$ regroupe les patients bien portants et malades; il serait plus intéressant de séparer ces deux classes.

Par ailleurs, dans QuantMiner, nous cherchons le meilleur intervalle pour chaque attribut, mais que se passe-t-il s'il n'y a pas vraiment de meilleur, mais un bon intervalle et d'autres intéressants ? Notre algorithme limité à un intervalle ne va pas considérer les autres alors qu'ils auraient pu révéler des informations intéressantes. On peut penser qu'il suffit d'introduire la disjonction d'intervalles dans la représentation des individus pour résoudre le problème, mais cela le rend plus complexe. Ce problème est illustré dans la Figure 3 et sera l'objet de nos travaux futurs dans cette thématique.

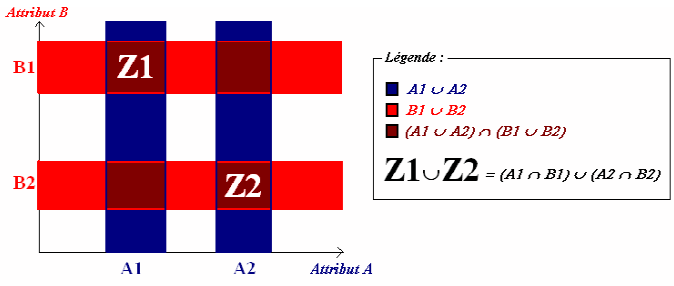

FIG. 3 – Soient 2 concentrations de points « Z1 » et « Z2 » à mettre en évidence. Il est bien évident que l'on ne générera jamais d'itemset les représentant exactement avec la notation disjonctive car « (A, A1∪A2) ∩ (B, B1∪B2) » exprime 4 zones, dont 2 parasites.

Le problème d'extraction de règles quantitatives ne peut pas être une simple extension de celui des règles catégoriques. La principale raison est que dans le cas catégorique, on teste finalement *tous les cas possibles* de règles ce qui est clairement impossible dans le cas numérique. Le problème est donc du ressort de l'optimisation : la recherche de règles d'association perd alors de son caractère non supervisé et est plutôt guidée par la forme des règles.

Nous avons présenté dans cet article QuantMiner, qui repose sur un algorithme génétique permettant de découvrir de façon dynamique les intervalles des variables numériques qui optimisent le support et la confiance d'une règle d'association. QuantMiner est un outil convivial et interactif, où la présence de l'expert reste indispensable pour aller à l'essentiel et visionner les règles au plus fort potentiel.

Références

Agrawal R., Imielinski T. et Swami A. N. (1993). Mining association rules between sets of items in large databases. Dans Buneman P. et Jajodia S., éditeurs, *Proceedings of the 1993 ACM SIGMOD*, pages 207–216, Washington, D.C.

Agrawal R. et Srikant R. (1994). Fast algorithms for mining association rules. Dans Bocca J. B., Jarke M. et Zaniolo C., éditeurs, *Proc. 20th Int. Conf. Very Large Data Bases.*, pages 487–499. Morgan Kaufmann.

Aumann Y. et Lindell Y. (1999). A statistical theory for quantitative association rules. Dans *Knowledge Discovery and Data Mining*, pages 261–270.

Bayardo R. J. (1998). Efficiently mining long patterns from databases. Dans *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data*, pages 85–93, Seattle.

Brin S., Motwani R. et Silverstein C. (1997). Beyond market baskets: generalizing association rules to correlations. Dans *Proc. of ACM SIGMOD*, pages 265–276.

Fisher R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*.

Fukuda T., Morimoto Y., Morishita S. et Tokuyama T. (1996a). Data mining using two-dimensional optimized association rules: Scheme, algorithms and visualization. Dans *Proc. of the Int'l Conf. ACM SIGMOD*, pages 12–23.

Fukuda T., Morimoto Y., Morishita S. et Tokuyama T. (1996b). Mining optimized association rules for numeric attributes. Dans *Proc. of the fteenth ACM SIGACT-SIGMOD -SIGART PODS'96*, pages 182–191. ACM Press.

Lent B., Swami A. N. et Widom J. (1997). Clustering Association Rules. Dans Gray A. et Larson P.-Å., éditeurs, *Proceedings of the 13th International Conference on Data Engineering, April 7-11, 1997 Birmingham U.K*, pages 220–231.

Mata J., Alvarez J. L. et Riquelme J. C. (2002). An evolutionary algorithm to discover numeric association rules. Dans *Proceedings of the 2002 ACM symposium on Applied computing SAC'2002*, pages 590–594.

Miller R. J. et Yang Y. (1997). Association rules over interval data. *SIGMOD Rec.*, 26(2):452–461.

Murphy P. M. et Aha D. W. (1995). *UCI Repository of Machine Learning Databases.* Machine-readable collection, Dept of Information and Computer Science, University of California, Irvine. [Available by anonymous ftp from `ics.uci.edu` in directory `pub/machine-learning-databases`].

Nortet C. Extraction de Règles d'Association Quantitatives. Mémoire de Master, BRGM et LIFO Université d'Orléans.

Rastogi R. et Shim K. (1999). Mining optimized support rules for numeric attributes. Dans *Proceedings of the 15th International Conference on Data Engineering, 23-26 March 1999, Sydney, Austrialia*, pages 206–215. IEEE Computer Society.

Salleb A. (2003). *Recherche de Motifs Fréquents pour l'Extraction de Règles d'Association et de Caractérisation*. Thèse de doctorat, Université d'Orléans, France.

Salleb A., Maazouzi Z. et Vrain C. (2002). Mining Maximal Frequent Itemsets by a Boolean Based Approach. Dans Harmelen F., éditeur, *15th European Conference on Artificial Intelligence Ecai*, pages 385–389, Lyon, France. IOS Press Amsterdam.

Salleb A., Turmeaux T., Vrain C. et Nortet C. (2004). Mining quantitative association rules in a atherosclerosis dataset. Dans *Proceedings of the PKDD Discovery Challenge 2004 (co-located with the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases)*, pages 98–103, Pisa, Italy.

Srikant R. et Agrawal R. (1996). Mining quantitative association rules in large relational tables. Dans Jagadish H. V. et Mumick I. S., éditeurs, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 1–12.

Wang K., Tay S. H. W. et Liu B. (1998). Interestingness-based interval merger for numeric association rules. Dans Agrawal R., Stolorz P. E. et Piatetsky-Shapiro G., éditeurs, *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 121–128. AAAI Press.

Webb G. I. (2001). Discovering associations with numeric variables. Dans *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 383–388. ACM Press.

Zaki M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.

Zhang Z., Lu Y. et Zhang B. (1997). An effective partitioning-combining algorithm for discovering quantitative association rules. Dans *Proceedings of the of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Summary

Mining association rules in databases has long been studied. However, most researches have focused on mining efficiently such rules in databases composed of boolean or categorical attributes, when in practice many tables contain also numeric attributes. In this paper, we propose QuantMiner, a system for mining multi-dimensional quantitative association rules. QuantMiner looks for the best intervals for numeric attributes relying on a genetic-based algorithm. Basically, in order to get high quality rules, both the support and confidence are optimized during the mining process.
We conducted an intensive experimental evaluation of our algorithm on real datasets. Our experiments showed the usefulness of QuantMiner as an interactive descriptive data mining tool.

Keywords: data mining, association rule, numeric attribute, discretization, categorical attribute, evolutionary algorithm.