

Le rôle de l'utilisateur dans un processus d'extraction de règles d'association

Cyril Nortet Ansaf Salleb
Teddy Turmeaux Christel Vrain

`Ansaf.Salleb@irisa.fr`



RÈGLES D'ASSOCIATION

$C_1 \implies C_2$ C_1, C_2 conditions $item_1 \wedge \dots \wedge item_k$

● $item \equiv produit, objet...$

Thé \implies Biscuits

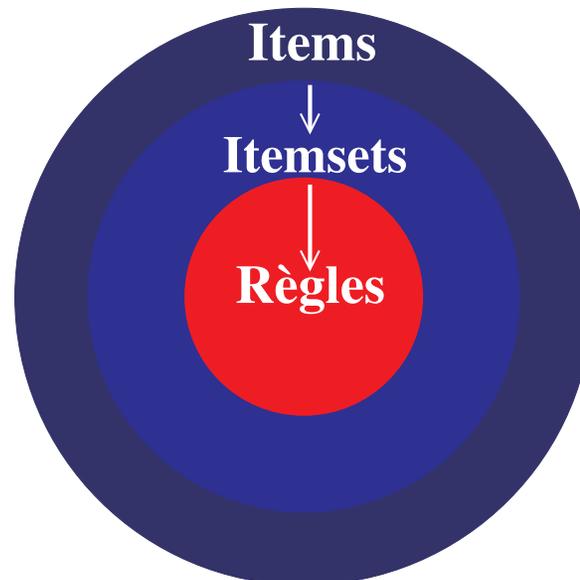
Le fabuleux destin
d'Amélie Poulain \implies Un long dimanche
de fiançailles

● $item \equiv (attribut = v_i) \text{ ou } (attribut \in [l_i, u_i])$

âge $\in [30, 40] \wedge nb_enfants=3 \implies crédit=oui$

CADRE CLASSIQUE (AGRAWAL 1993)

- $\text{Support}(\mathcal{C}_1 \Rightarrow \mathcal{C}_2) = \text{Support}(\mathcal{C}_1 \wedge \mathcal{C}_2)$
- $\text{Confiance}(\mathcal{C}_1 \Rightarrow \mathcal{C}_2) = \frac{\text{Support}(\mathcal{C}_1 \wedge \mathcal{C}_2)}{\text{Support}(\mathcal{C}_1)}$
- Deux étapes :
 - recherche des itemsets fréquents (seuil MinSupp)
 - recherche des règles fortes (seuil MinConf)



CADRE CLASSIQUE : PROBLÈMES

- **Problème 1 :** Trop de règles
- **Problème 2 :** Extension aux règles quantitatives

I - Approches fondées sur une pré-discrétisation (Binning)

- La discrétisation consiste à découper le domaine d'un attribut numérique :
 - en k intervalles de largeur uniforme (equi-width),
 - en k intervalles de fréquences égales (equi-depth),
 - de façon non régulière (connaissances du domaine)
- Équilibre Support-Confiance **Srikant et Agrawal 1996**
 - petits intervalles \longrightarrow faible support
 - grands intervalles \longrightarrow faible confiance
- Ramener la recherche au cadre classique

I - Approches fondées sur une pré-discrétisation (Binning)

- Srikant et agrawal SIGMOD 1996
- Lent et al. ICDE 1997
- Zhang et al. PAKDD 1997
- Miller et Yang SIGMOD 1997
- Wang et al. KDD 1998

Inconvénients :

- Choix de k
- Sensibilités au bruit (outliers)
- Attributs discrétisés indépendamment les uns des autres
- On peut manquer des règles à fort potentiel

II - Approches guidées par des schémas de règles

- le mot clé n'est plus *discrétisation* mais plutôt *optimisation*
- la forme de la règle est spécifiée par l'utilisateur
- instancier les intervalles en optimisant une mesure de la *qualité* de la règle

II - Approches guidées par des schémas de règles

- Fukuda et al. SIGMOD 1996

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

- Rastogi et Shim ICDE 1999

- Auman et Lindell KDD 1999 et Webb KDD 2001 (distributions statistiques)

Région=Sud \implies Salaire : moyenne = 1200 euros / mois

Inconvénients :

- limitées à 2 attributs numériques

III - Approche évolutionnaire (AG)

- Optimisation du support des itemsets par AG

$$\text{Qualité} = \text{Support} - (\alpha * \text{amplitude}) - (\beta * \text{marqué}) + (\delta * \text{nbAttributs})$$

- individu = {(attribut numérique, disjonction d'intervalles)}

- Mata et al. SAC 2002

Inconvénients :

- limitée aux attributs numériques

- seul le support est optimisé

- ne tient pas compte de la position de l'intervalle dans la règle

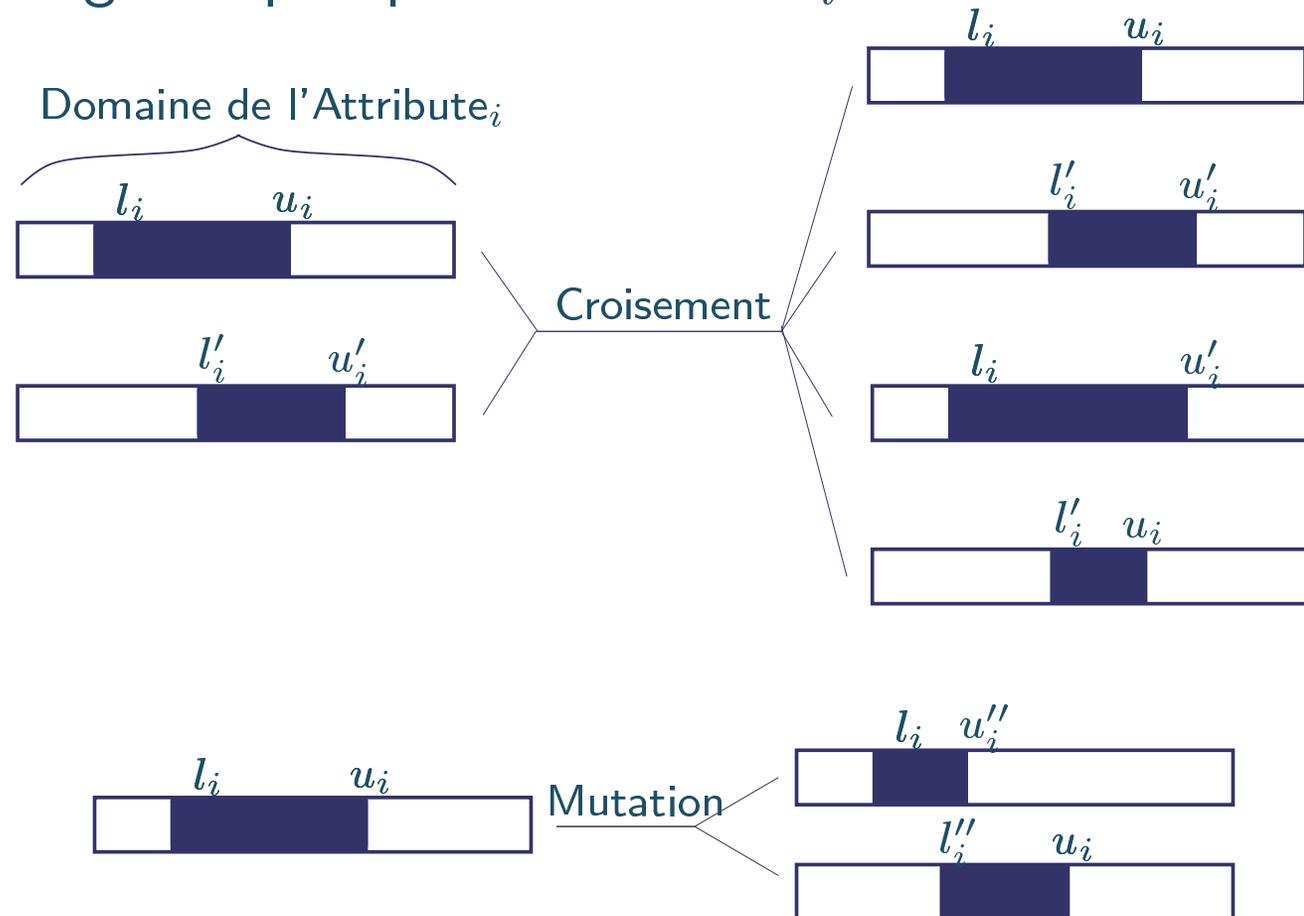


Règles

- Travailler sur des schémas de règles prédéfinis sur des attributs numériques et catégoriques (**Problème 1**)
- Optimiser le Gain de **Fukuda et al. 1996** (support et confiance) par un algorithme génétique (**Problème 2**)

QUANTMINER

- Individu : $[l_1, u_1] \dots [l_k, u_k]$
- Opérateurs génétiques pour un attribut i :



● Fonction de qualité

1. Qualité($A \Rightarrow B$)=
2. QualitéTemporaire = Gain($A \Rightarrow B$);
3. Si QualitéTemporaire ≥ 0
4. Pour tout intervalle I Faire
5. QualitéTemporaire * = (1-Proportion(I))**2
6. Si Support($A \Rightarrow B$) < MinSupp
7. QualitéTemporaire - = NbEnregistrements;
8. 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 items (Attribut=valeur) fixés dans les schémas
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

EXEMPLE : LES IRIS (ANDERSON 1935, FISHER 1936)

- 150 Iris, 3 espèces : Setosa (50), Virginica(50), Versicolor(50)
- 5 attributs :
 - Espèce (Species)
 - Largeur Pétale (PW)
 - Largeur Sépale (SW)
 - Longueur Pétale (PL)
 - Longueur Sépale (SL)

$$\begin{array}{l} \text{Species=} \\ \text{value} \end{array} \Rightarrow \left\{ \begin{array}{ll} \text{PW} \in [l_1, u_1] & \text{SW} \in [l_2, u_2] \\ \text{PL} \in [l_3, u_3] & \text{SL} \in [l_4, u_4] \end{array} \right\} \begin{array}{l} \text{supp}\% \\ \text{conf}\% \end{array}$$

EXEMPLE : LES IRIS

$$\begin{array}{l} \text{Species=} \\ \text{Setosa} \end{array} \Rightarrow \left\{ \begin{array}{l} \text{PW} \in [1, 6] \quad \text{SW} \in [31, 39] \\ \text{PL} \in [10, 19] \quad \text{SL} \in [46, 54] \end{array} \right\} \begin{array}{l} 23\% \\ 70\% \end{array}$$

Species	Attribut	Min	Max	Moyenne	Écart type
Setosa	PW	1	6	2.46	1.05
	SW	23	44	34.28	3.79
	PL	10	19	14.62	1.74
	SL	43	58	50.6	3.52

EXEMPLE : LES IRIS

$$\begin{array}{l} \text{Species=} \\ \text{Versicolor} \end{array} \Rightarrow \left\{ \begin{array}{ll} \text{PW} \in [10, 15] & \text{SW} \in [22, 30] \\ \text{PL} \in [35, 47] & \text{SL} \in [55, 66] \end{array} \right\} \begin{array}{l} 21\% \\ 64\% \end{array}$$

$$\begin{array}{l} \text{Species=} \\ \text{Virginica} \end{array} \Rightarrow \left\{ \begin{array}{ll} \text{PW} \in [18, 25] & \text{SW} \in [27, 33] \\ \text{PL} \in [48, 60] & \text{SL} \in [58, 72] \end{array} \right\} \begin{array}{l} 20\% \\ 60\% \end{array}$$

Species	Attribut	Min	Max	Moyenne	Écart type
Versicolor	PW	10	18	13.26	1.98
	SW	20	34	27.7	3.14
	PL	30	51	42.6	4.70
	SL	49	70	59.39	5.16
Virginica	PW	14	25	20.26	2.75
	SW	22	38	29.74	3.22
	PL	45	69	55.52	5.52
	SL	49	79	65.88	6.36

IMPLÉMENTATION

- Développé en JAVA par C. Nortet au BRGM
- Processus interactif et itératif (assistant)
- temps ?

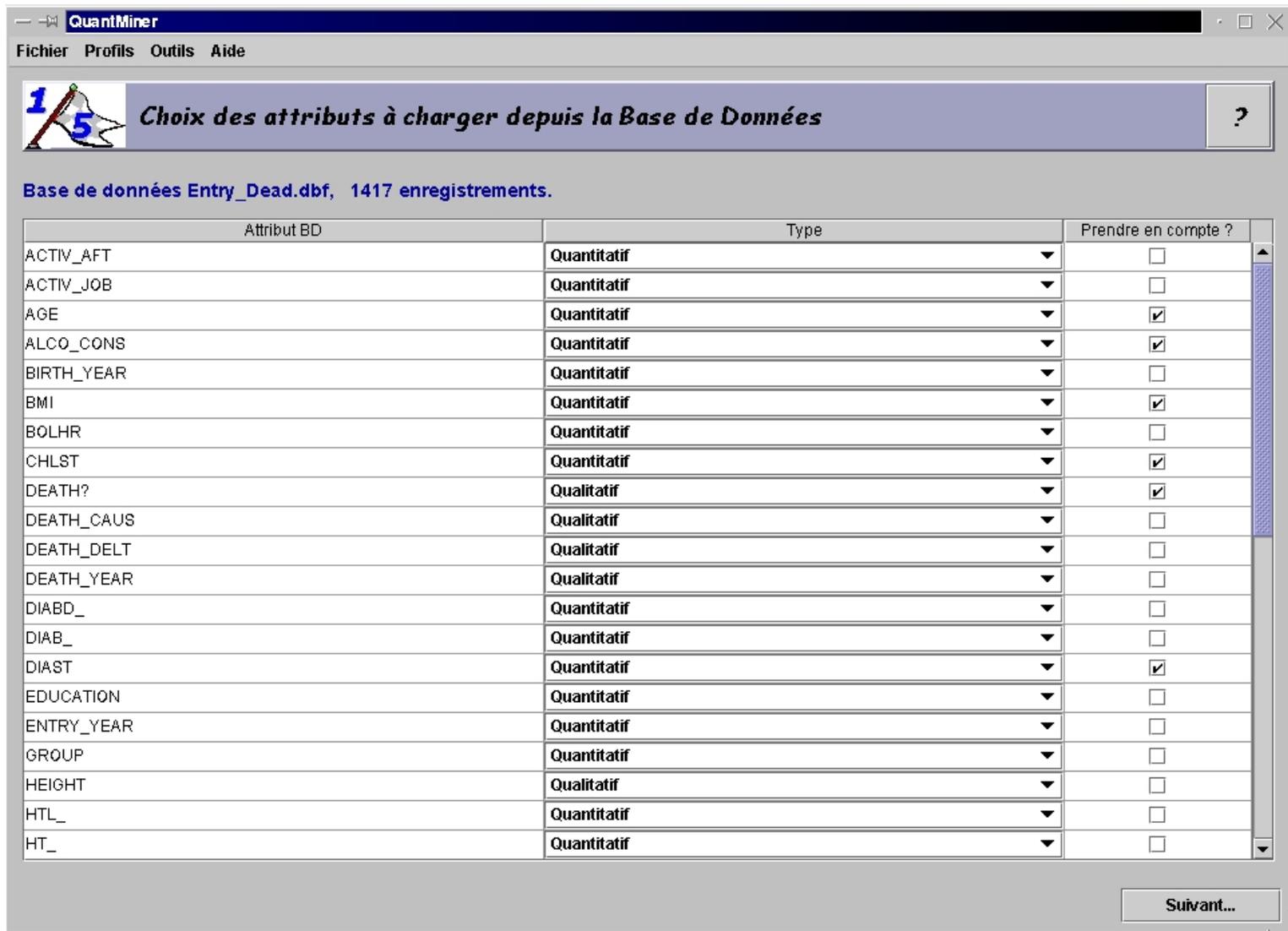
Sur une base de 2 500 n-uplets, population de 250 individus, 150 générations, un schéma de règle/seconde.

APPLICATION

- Base de données médicale sur l'athérosclérose
<http://euromise.vse.cz/STULONG>
(Salleb et al. PKDD Discovery Challenge 2004)
- Étude effectuée pendant 20 ans sur les facteurs de risque de cette maladie
- 1 419 patients classés en 3 groupes : normaux (N), à risque (R), malades (P).
- décrits par 27 attributs symboliques et 17 numériques (poids, taille, âge, activités physiques, taux de cholestérol, consommation de tabac,...).

ÉTAPE 1 - CHOIX DES ATTRIBUTS

Sélection des Attributs et définition de leur type.



QuantMiner

Fichier Profils Outils Aide

1/5 Choix des attributs à charger depuis la Base de Données ?

Base de données Entry_Dead.dbf, 1417 enregistrements.

Attribut BD	Type	Prendre en compte ?
ACTIV_AFT	Quantitatif	<input type="checkbox"/>
ACTIV_JOB	Quantitatif	<input type="checkbox"/>
AGE	Quantitatif	<input checked="" type="checkbox"/>
ALCO_CONS	Quantitatif	<input checked="" type="checkbox"/>
BIRTH_YEAR	Quantitatif	<input type="checkbox"/>
BMI	Quantitatif	<input checked="" type="checkbox"/>
BOLHR	Quantitatif	<input type="checkbox"/>
CHLST	Quantitatif	<input checked="" type="checkbox"/>
DEATH?	Qualitatif	<input checked="" type="checkbox"/>
DEATH_CAUS	Qualitatif	<input type="checkbox"/>
DEATH_DELT	Qualitatif	<input type="checkbox"/>
DEATH_YEAR	Qualitatif	<input type="checkbox"/>
DIABD_	Quantitatif	<input type="checkbox"/>
DIAB_	Quantitatif	<input type="checkbox"/>
DIAST	Quantitatif	<input checked="" type="checkbox"/>
EDUCATION	Quantitatif	<input type="checkbox"/>
ENTRY_YEAR	Quantitatif	<input type="checkbox"/>
GROUP	Quantitatif	<input type="checkbox"/>
HEIGHT	Qualitatif	<input type="checkbox"/>
HTL_	Quantitatif	<input type="checkbox"/>
HT_	Quantitatif	<input type="checkbox"/>

Suivant...

ÉTAPE 2 - CHOIX DES SCHÉMAS

Répartition fine des attributs dans les règles à extraire.

QuantMiner

Fichier Profils Outils Aide

2/5 Contribution des attributs dans les règles générées ?

Tout sélectionner Ne rien sélectionner

Attribut / Modalité	Informations	Position dans la règle	Présence obligatoire
DEATH?	2 modalités différentes	à gauche (condition)	<input checked="" type="checkbox"/>
SUPER_GROU	4 modalités différentes	nulle part	<input type="checkbox"/>
AGE	[38.0, 53.0], 0 valeurs manquantes.	nulle part	<input type="checkbox"/>
ALCO_CONS	[1.0, 1.69], 0 valeurs manquantes.	à droite (objectif)	<input type="checkbox"/>
BMI	[0.0, 44.96], 7 valeurs manquantes.	à droite (objectif)	<input type="checkbox"/>
CHLST	[112.0, 530.0], 0 valeurs manquantes.	nulle part	<input type="checkbox"/>
DIAST	[50.0, 125.0], 455 valeurs manquantes.	nulle part	<input type="checkbox"/>
SUBSC	[4.0, 70.0], 137 valeurs manquantes.	nulle part	<input type="checkbox"/>
SYST	[80.0, 220.0], 460 valeurs manquantes.	nulle part	<input type="checkbox"/>
TOBA_CONSO	[0.0, 1.25], 17 valeurs manquantes.	à droite (objectif)	<input checked="" type="checkbox"/>
TOBA_DURA	[5.0, 20.0], 383 valeurs manquantes.	à droite (objectif)	<input checked="" type="checkbox"/>
TRIC	[1.0, 35.0], 136 valeurs manquantes.	nulle part	<input type="checkbox"/>
TRIGL	[28.0, 1197.0], 0 valeurs manquantes.	nulle part	<input type="checkbox"/>

Précédent... Suivant...

ÉTAPE 3 - CHOIX DE LA MÉTHODE

Choix de la technique d'optimisation et réglage de ses paramètres.

The screenshot shows the 'QuantMiner' application window. The title bar reads 'QuantMiner' and the menu bar includes 'Fichier', 'Profils', 'Outils', and 'Aide'. The main window title is '3 5 Configuration de la technique d'extraction'. The interface is divided into two main sections: 'Paramétrage des règles à extraire' and 'Paramétrage de la technique'.

Technique à utiliser : ?

Paramétrage des règles à extraire :

- Seuil de support (% ...): ?
- Seuil de confiance (%): ?
- Nombre d'attributs quantitatifs par règle, minimum: maximum:
- Nombre de disjonctions ("OU") autorisées dans la partie gauche:
- Nombre de disjonctions ("OU") autorisées dans la partie droite:

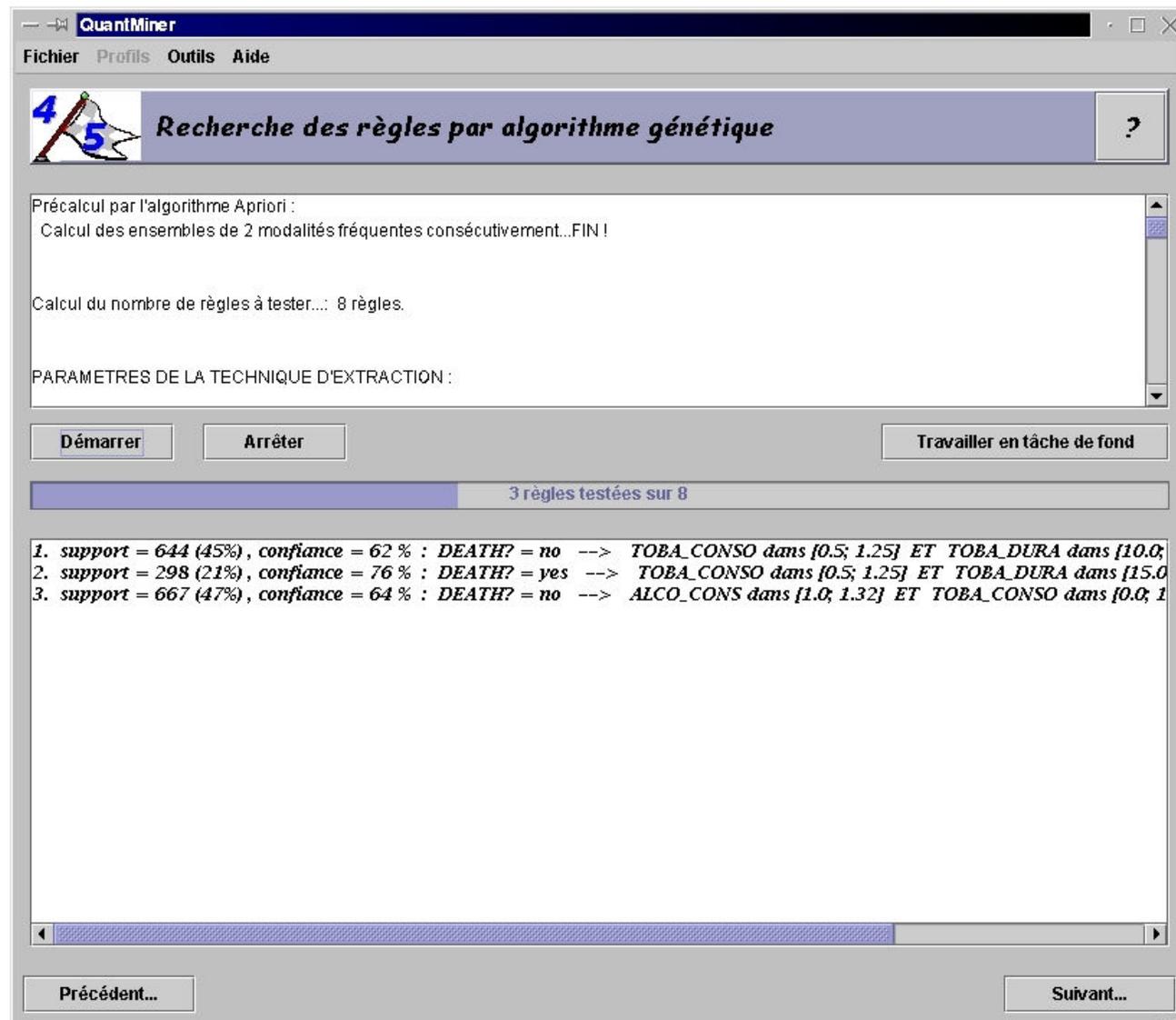
Paramétrage de la technique :

- Taille de la population: ?
- Nombre de générations: ?
- Taux de croisement (% ...): ?
- Taux de mutation (%): ?

Buttons:

ÉTAPE 4 - OPTIMISATION

Exécution de l'algorithme d'optimisation.



The screenshot shows the QuantMiner application window. The title bar reads "QuantMiner". The menu bar includes "Fichier", "Profils", "Outils", and "Aide". The main window has a header bar with a "4/5" icon and the text "Recherche des règles par algorithme génétique". Below this, a text area displays the progress of the algorithm:

Précalcul par l'algorithme Apriori :
Calcul des ensembles de 2 modalités fréquentes consécutivement...FIN !

Calcul du nombre de règles à tester...: 8 règles.

PARAMETRES DE LA TECHNIQUE D'EXTRACTION :

Buttons: Démarrer, Arrêter, Travailler en tâche de fond

Progress bar: 3 règles testées sur 8

Results list:

1. support = 644 (45%), confiance = 62 % : DEATH? = no --> TOBA_CONSO dans [0.5; 1.25] ET TOBA_DURA dans [10.0;
2. support = 298 (21%), confiance = 76 % : DEATH? = yes --> TOBA_CONSO dans [0.5; 1.25] ET TOBA_DURA dans [15.0
3. support = 667 (47%), confiance = 64 % : DEATH? = no --> ALCO_CONS dans [1.0; 1.32] ET TOBA_CONSO dans [0.0; 1

Buttons: Précédent..., Suivant...

ÉTAPE 5 - VISUALISATION

Affichage des résultats, avec tri sélectif des règles produites.

QuantMiner

Fichier Profils Outils Aide

5/5 Resultats

Enregistrer dans un fichier... Visualiser le contexte d'extraction...

Méthode de tri principale : tri par confiance tri décroissant

Exclure les règles dont le support du conséquent (partie B) dépasse (%... 75.0)

APPLIQUER

Afficher filtre... Réinitialiser filt... Filtrer à partir de la sélecti...

Règle n°2/8 (total : 8)

2. SUPPORT = 267 (18.84 %) , CONFIANCE = 68.64 % :

A { DEATH? = yes } → B { ALCO_CONS dans [1.0; 1.28]
TOBA_CONSO dans [0.5; 1.25]
TOBA_DURA dans [15.0; 20.0] }

PROPORTIONS DU DOMAINE COUVERT PAR LES INTERVALLES DE LA PARTIE DROITE :

ALCO_CONS : [bar chart] soit 40.58 % du domaine [1.0, 1.69]
TOBA_CONSO : [bar chart] soit 60.0 % du domaine [0.0, 1.25]
TOBA_DURA : [bar chart] soit 33.33 % du domaine [5.0, 20.0]

SUPPORTS :

A et B	[bar chart]	267 (18.84 %)
A	[bar chart]	389 (27.45 %)
B	[bar chart]	818 (57.73 %)
A et non B	[bar chart]	122 (8.61 %)
non A et B	[bar chart]	551 (38.88 %)
non A et non B	[bar chart]	477 (33.66 %)

CONFIANCES :

A → B	[bar chart]	68.64 %
non A → B	[bar chart]	53.60 %
B → A	[bar chart]	32.64 %
non B → A	[bar chart]	20.37 %
A ↔ B	[bar chart]	52.51 %

Précédent...

EXEMPLES DE RÈGLES

$$\begin{array}{l} \text{DEATH?=} \\ \text{YES} \end{array} \Rightarrow \left\{ \begin{array}{l} \text{ALCO_CONS} \in [1.0, 1.28] \\ \text{TOBA_CONSO} \in [0.5, 1.25] \\ \text{TOBA_DURA} \in [15, 20] \end{array} \right\} \begin{array}{l} 18\% \\ 68\% \end{array}$$

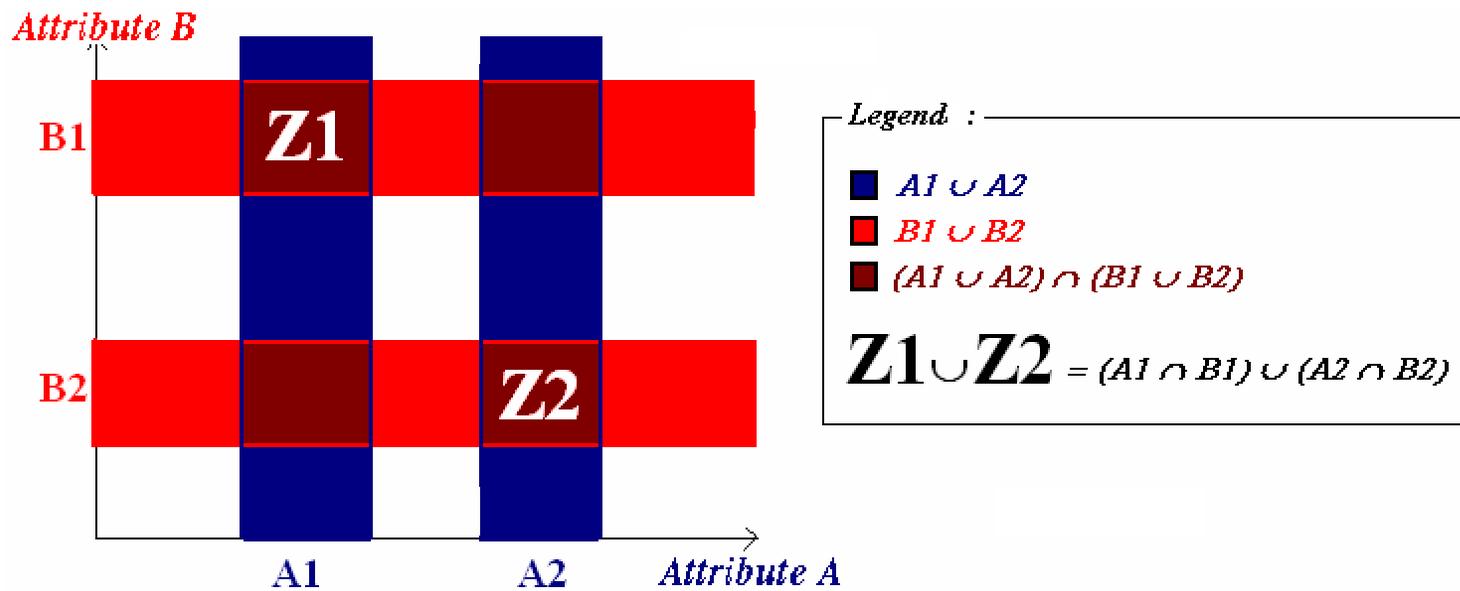
$$\text{GROOUP=N} \Rightarrow \left\{ \begin{array}{l} \text{ALCO_CONS} \in [1.0, 1.2] \\ \text{BMI} \in [19.73, 27.77] \\ \text{TOBA_CONSO} \in [0.0, 0.5] \end{array} \right\} \begin{array}{l} 13\% \\ 69\% \end{array}$$

$$\left\{ \begin{array}{l} \text{ALCO_CONS} \in [1.1, 1.2] \\ \text{BMI} \in [23.18, 26.15] \\ \text{TOBA_CONSO} \in [0.0, 0.5] \end{array} \right\} \Rightarrow \begin{array}{ll} \text{DEATH?=} & 9.5\% \\ \text{NO} & 90\% \end{array}$$

CONCLUSION

- La recherche de règles quantitatives **n'est pas une simple extension** du cadre classique
- QUANTMINER : optimisation dynamique de règles d'association quantitatives par algorithme génétique
- Règles fortes, intéressantes, peu sensibles au bruit
- Recherche guidée par l'utilisateur
- Perspectives :
 - Introduction de la disjonction
 - Extension aux règles de classification, d'exception,...

ANNEXE 1 : VERS LA DISJONCTION



ANNEXE 2 : RÈGLES *versus* ITEMSETS

- Itemset optimisé par GAR (Mata et al 2002)

$$\left\{ \begin{array}{l} \text{Assist} \in [0.0721, 0.2529] \\ \text{Height} \in [179, 198] \\ \text{Age} \in [22, 32] \end{array} \right\} 39.45\%$$

- Règles

$$\begin{array}{l} \text{Assist} \in \\ \mathbf{[0.0721, 0.2529]} \end{array} \Rightarrow \left\{ \begin{array}{l} \text{Height} \in [179, 198] \\ \text{Age} \in [22, 32] \end{array} \right\} \begin{array}{l} 39.45\% \\ 39.08\% \end{array}$$

$$\begin{array}{l} \text{Age} \in \\ [22, 32] \end{array} \Rightarrow \left\{ \begin{array}{l} \text{Assist} \in \mathbf{[0.0721, 0.2529]} \\ \text{Height} \in [179, 198] \end{array} \right\} \begin{array}{l} 39.45\% \\ 38.60\% \end{array}$$

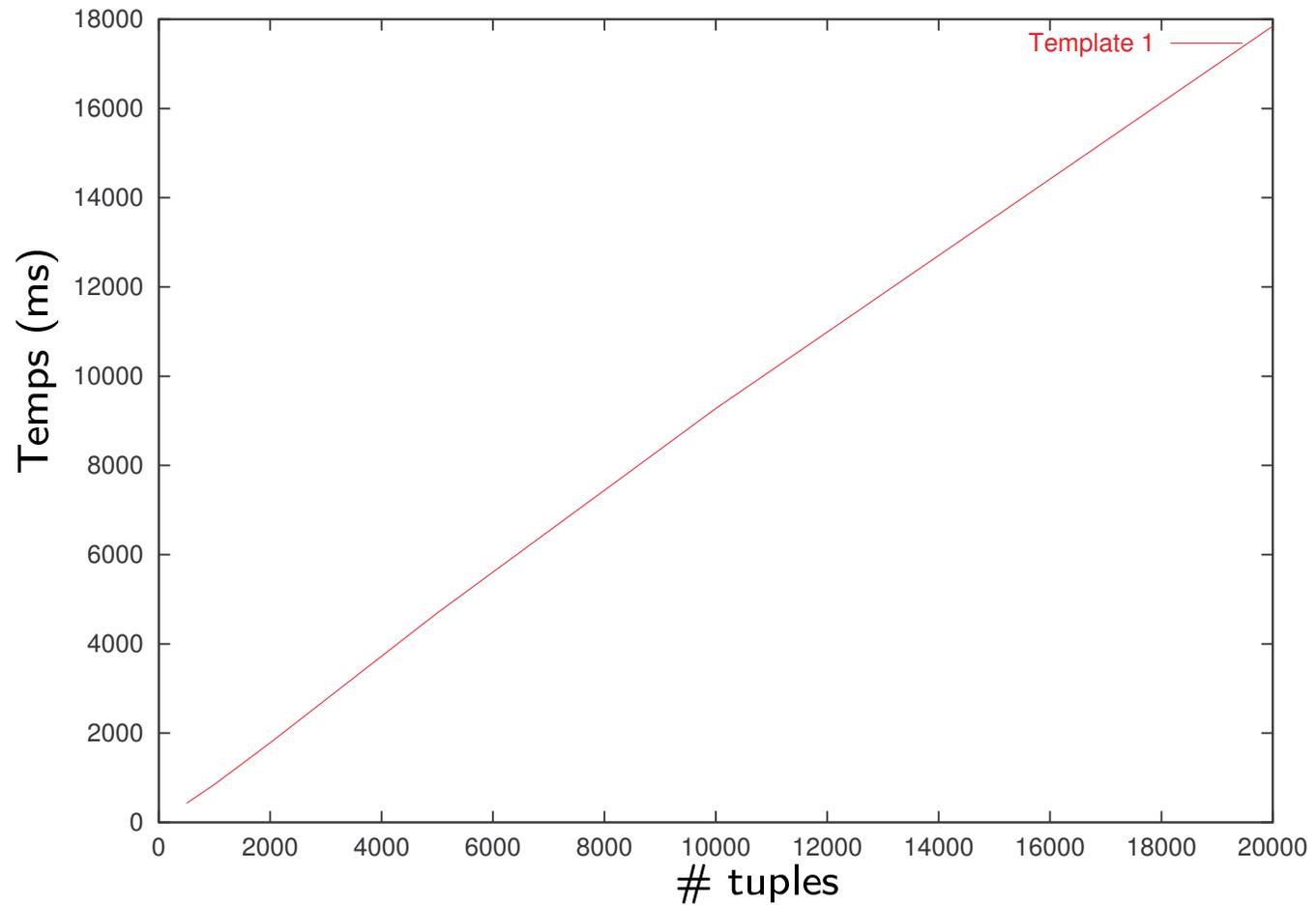
ANNEXE 2 : RÈGLES *versus* ITEMSETS

- Règles générées par QUANMINER

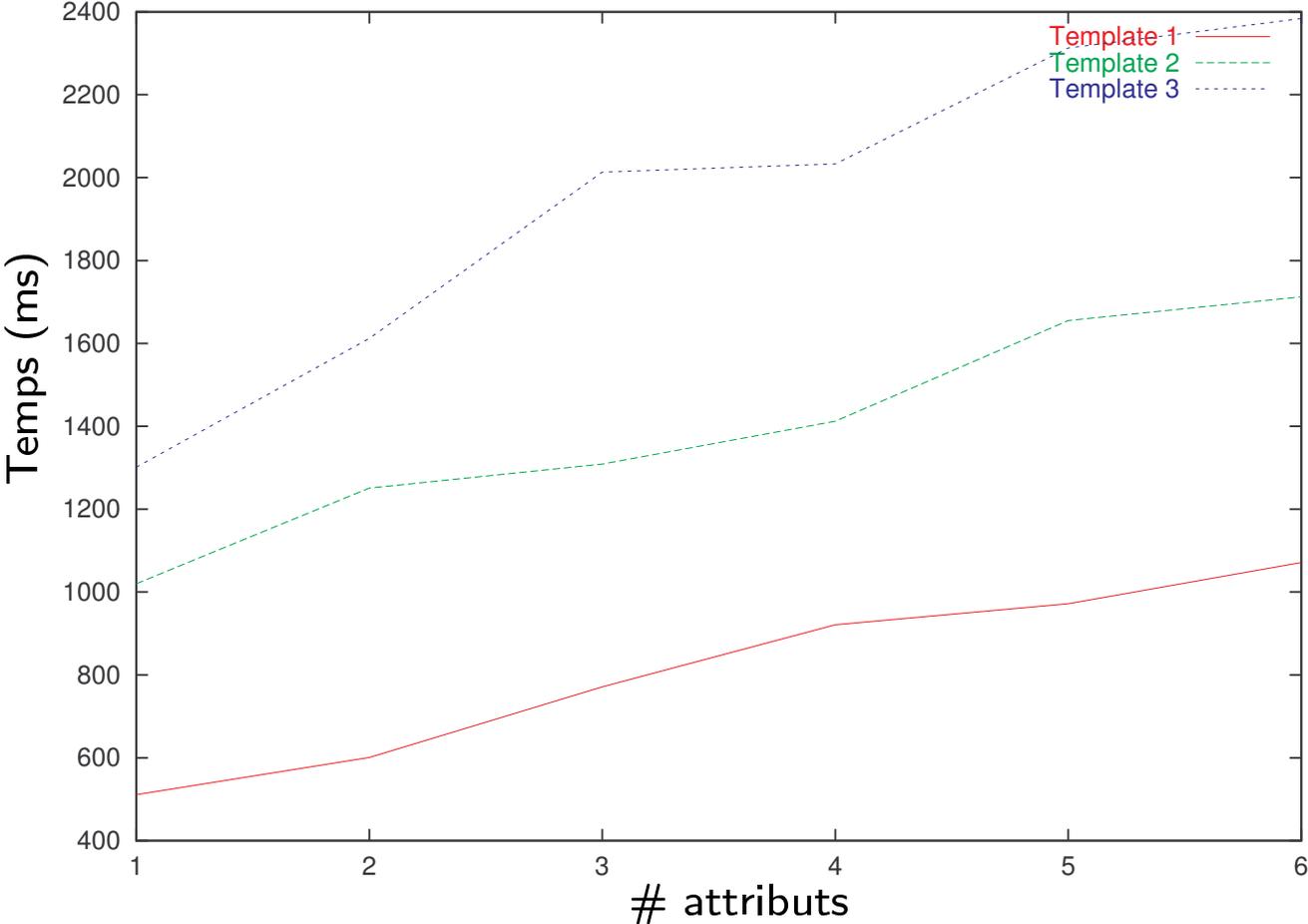
$$\begin{array}{l} \text{Assist} \in \\ \mathbf{[0.1554, 0.2521]} \end{array} \Rightarrow \left\{ \begin{array}{l} \text{Height} \in [180, 193] \\ \text{Age} \in [23, 30] \end{array} \right\} \begin{array}{l} 35\% \\ 72\% \end{array}$$

$$\begin{array}{l} \text{Age} \in \\ [24, 28] \end{array} \Rightarrow \left\{ \begin{array}{l} \text{Assist} \in \mathbf{[0.1058, 0.2495]} \\ \text{Height} \in [185, 196] \end{array} \right\} \begin{array}{l} 35\% \\ 65\% \end{array}$$

ANNEXE 4 : PERFORMANCES



ANNEXE 4 : PERFORMANCES



ANNEXE 4 : PERFORMANCES

