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

$\mathcal{C}_1 \Longrightarrow \mathcal{C}_2$ $\quad$ $\mathcal{C}_1, \mathcal{C}_2$ conditions $\texttt{item}_1 \wedge \cdots \wedge \texttt{item}_k$

- $item \equiv produit, objet...$

  Thé $\Longrightarrow$ Biscuits

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

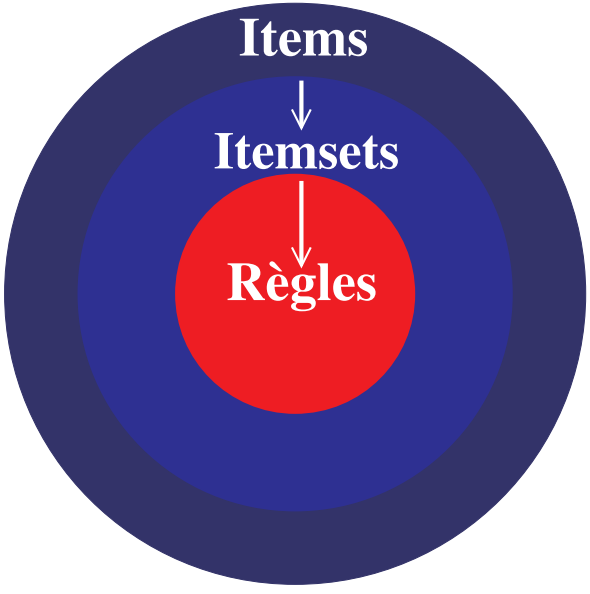
- $item \equiv (attribut = v_i)$ ou $(attribut \in [l_i, u_i])$

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

CADRE CLASSIQUE (AGRAWAL 1993)

- Support$(\mathcal{C}_1 \Rightarrow \mathcal{C}_2)$ = Support $(\mathcal{C}_1 \wedge \mathcal{C}_2)$
- Confiance$(\mathcal{C}_1 \Rightarrow \mathcal{C}_2) = \frac{\text{Support}(\mathcal{C}_1 \wedge \mathcal{C}_2)}{\text{Support}(\mathcal{C}_1)}$
- 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

ÉTAT DE L'ART

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

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

- Rastogi et Shim ICDE 1999
- Auman et Lindell KDD 1999 et Webb KDD 2001 (distributions statistiques)
  Région=Sud $\Longrightarrow$ 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 :

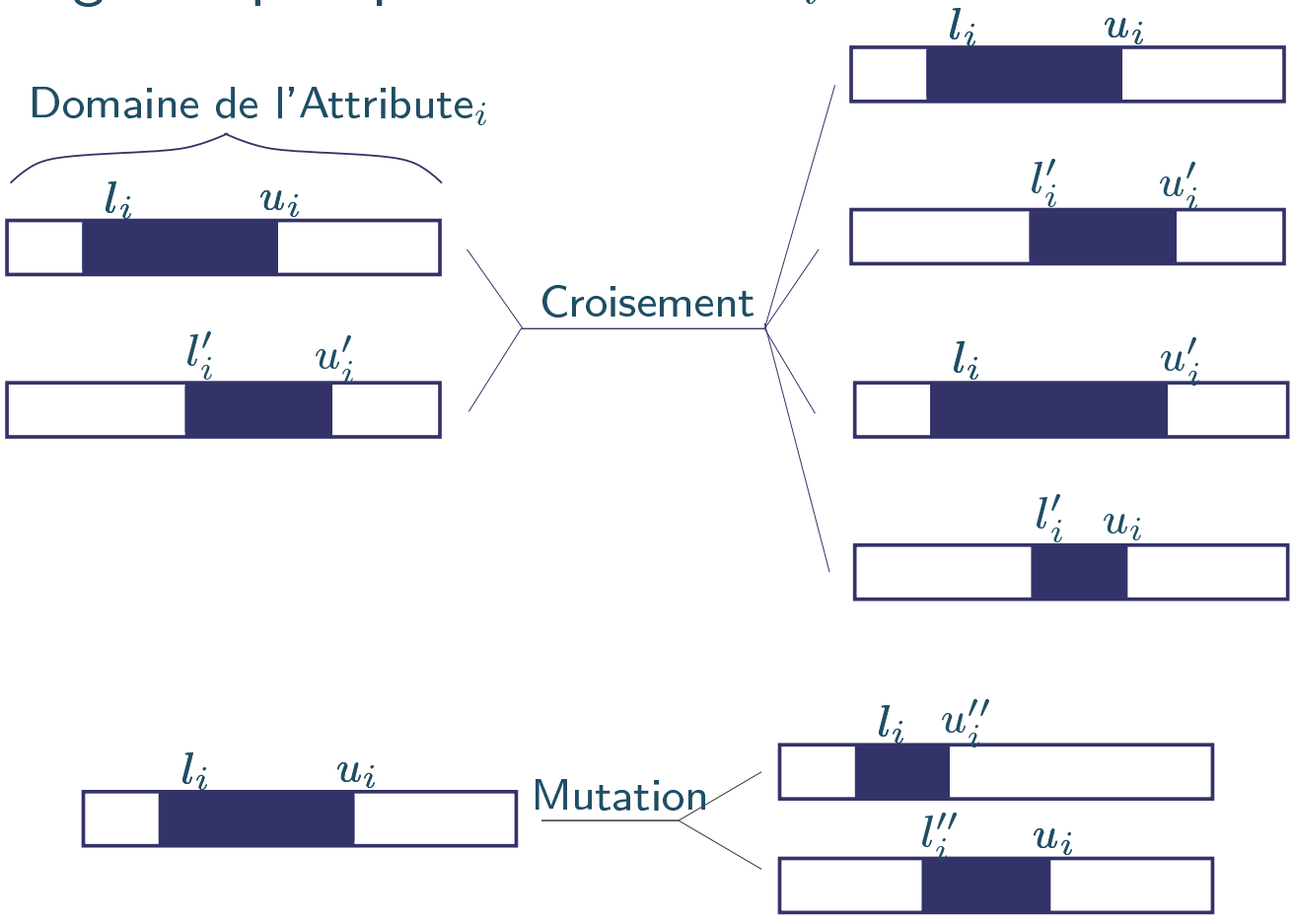
- limitée aux attributs numériques
- seul le support est optimisé
- ne tient pas compte de la position de l'intervalle dans la règle

QUANTMINER

- 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 : $[l_1, u_1] \ \ldots \ [l_k, u_k]$
- Opérateurs génétiques pour un attribut$_i$:



QUANTMINER

- 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)

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

EXEMPLE : LES IRIS

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

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

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

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

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)

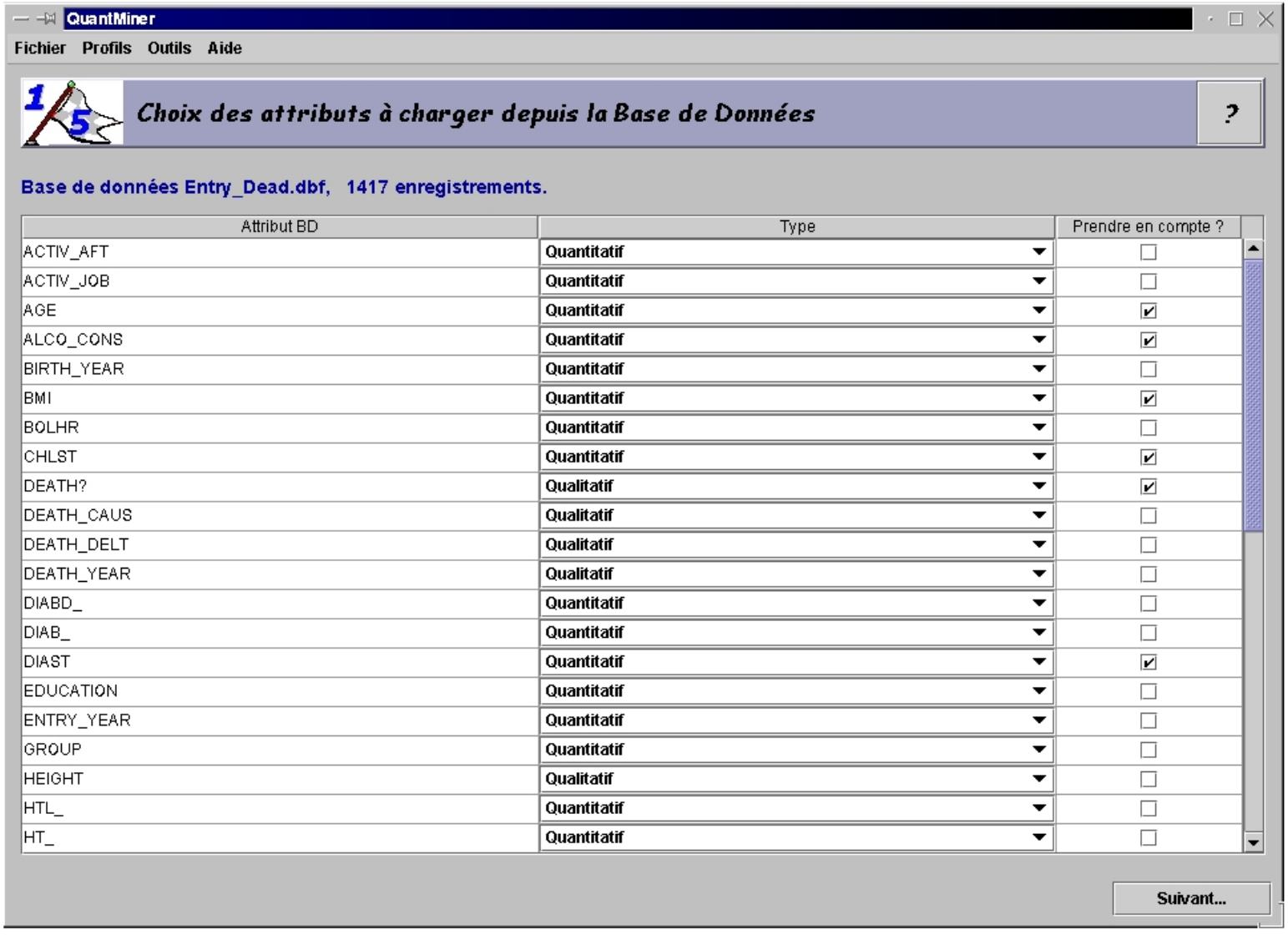
- Étude effectuée pendant 20 ans sur les facteurs de risque de cette maladie

- 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,...).

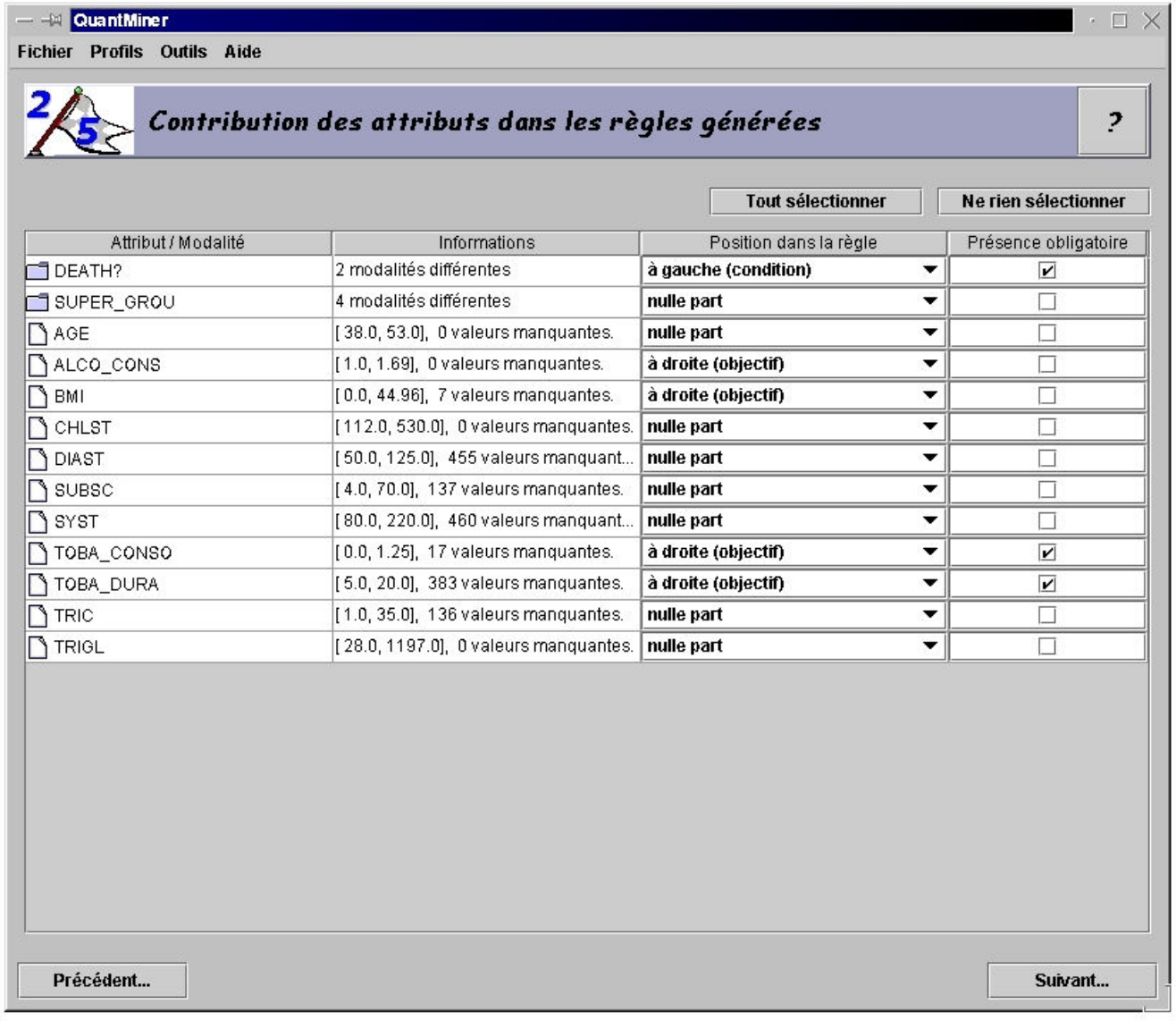
Étape 1 - Choix des attributs

Sélection des Attributs et définition de leur type.



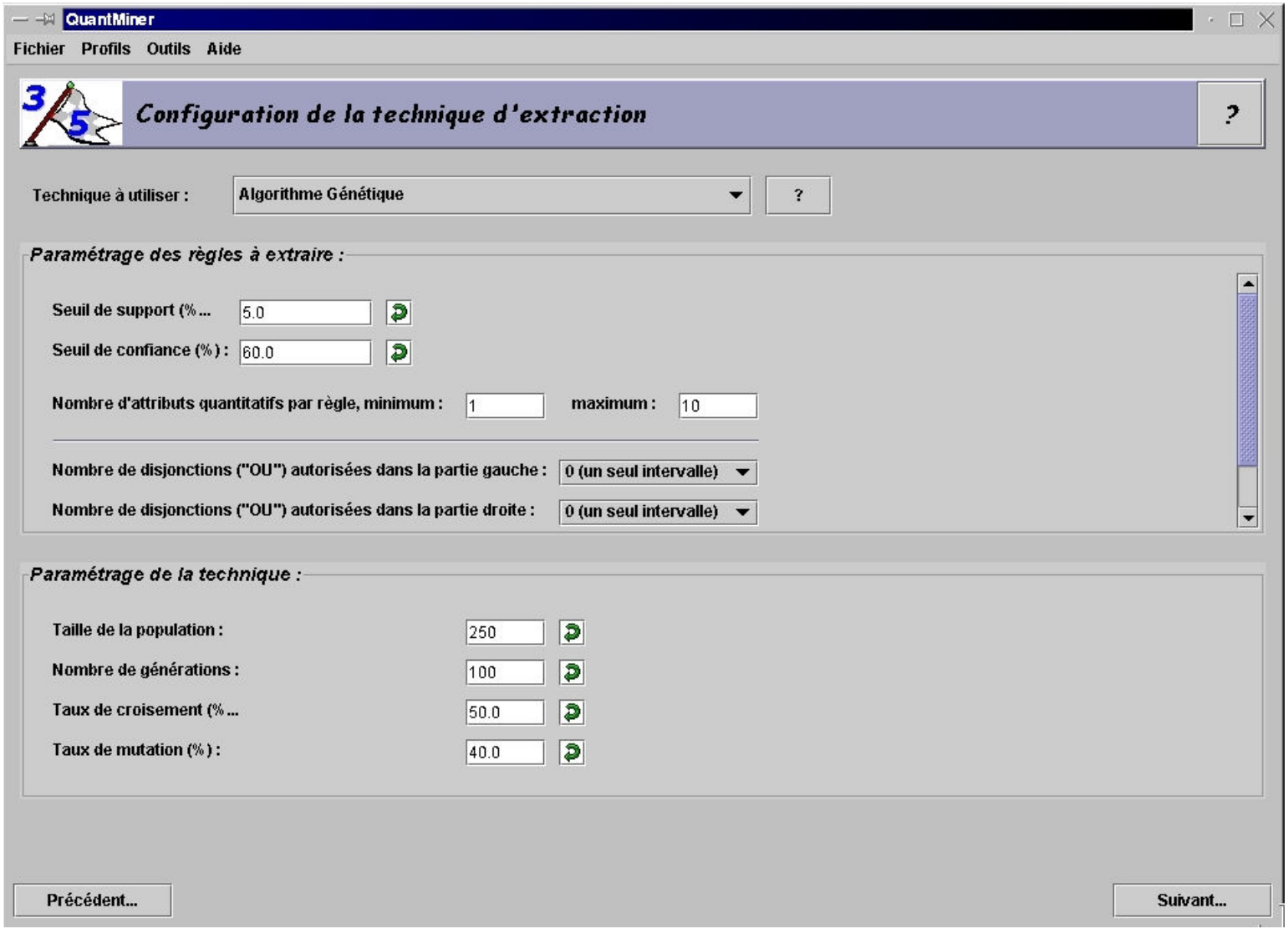
ÉTAPE 2 - CHOIX DES SCHÉMAS

Répartition fine des attributs dans les règles à extraire.



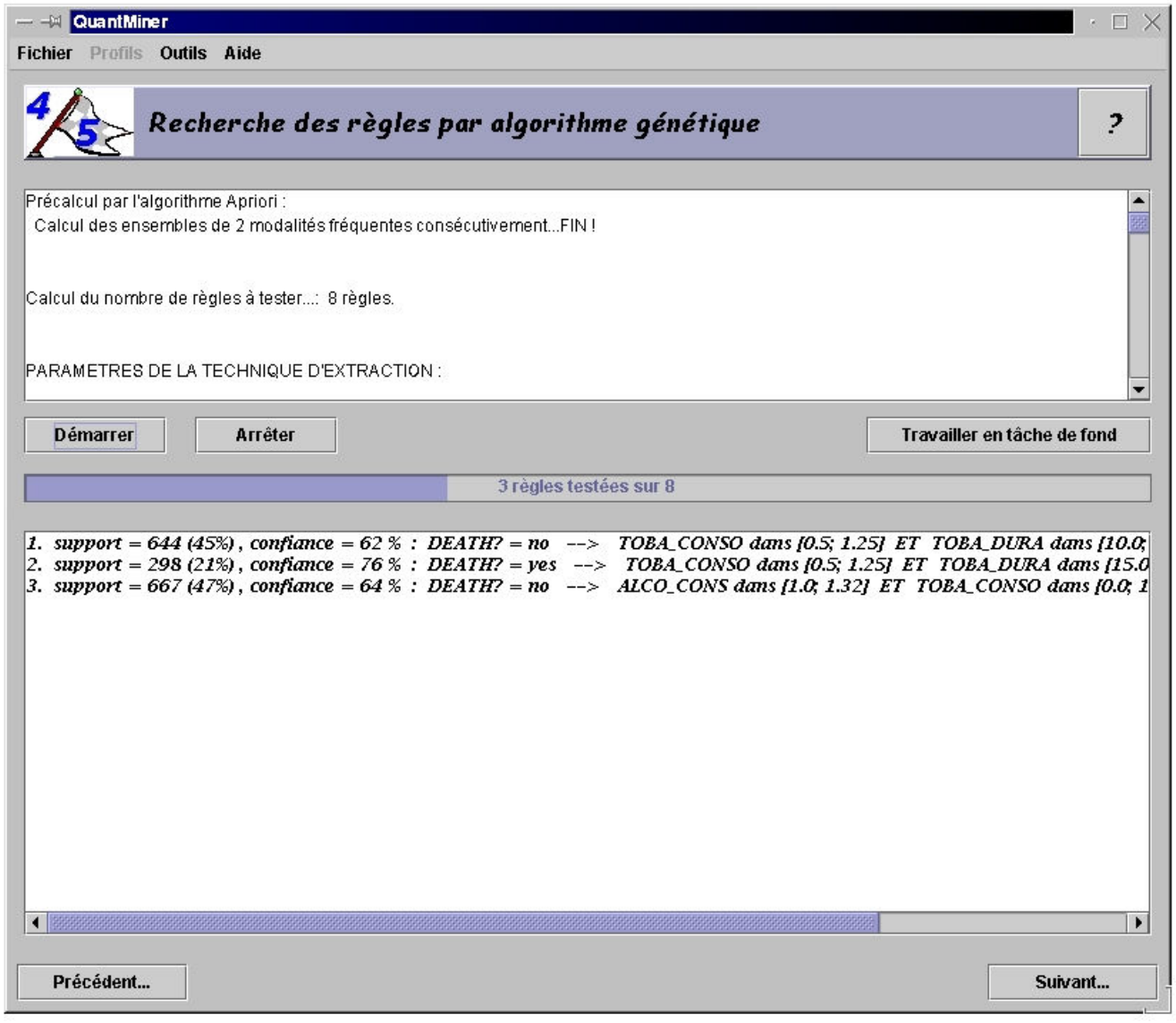
ÉTAPE 3 - CHOIX DE LA MÉTHODE

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



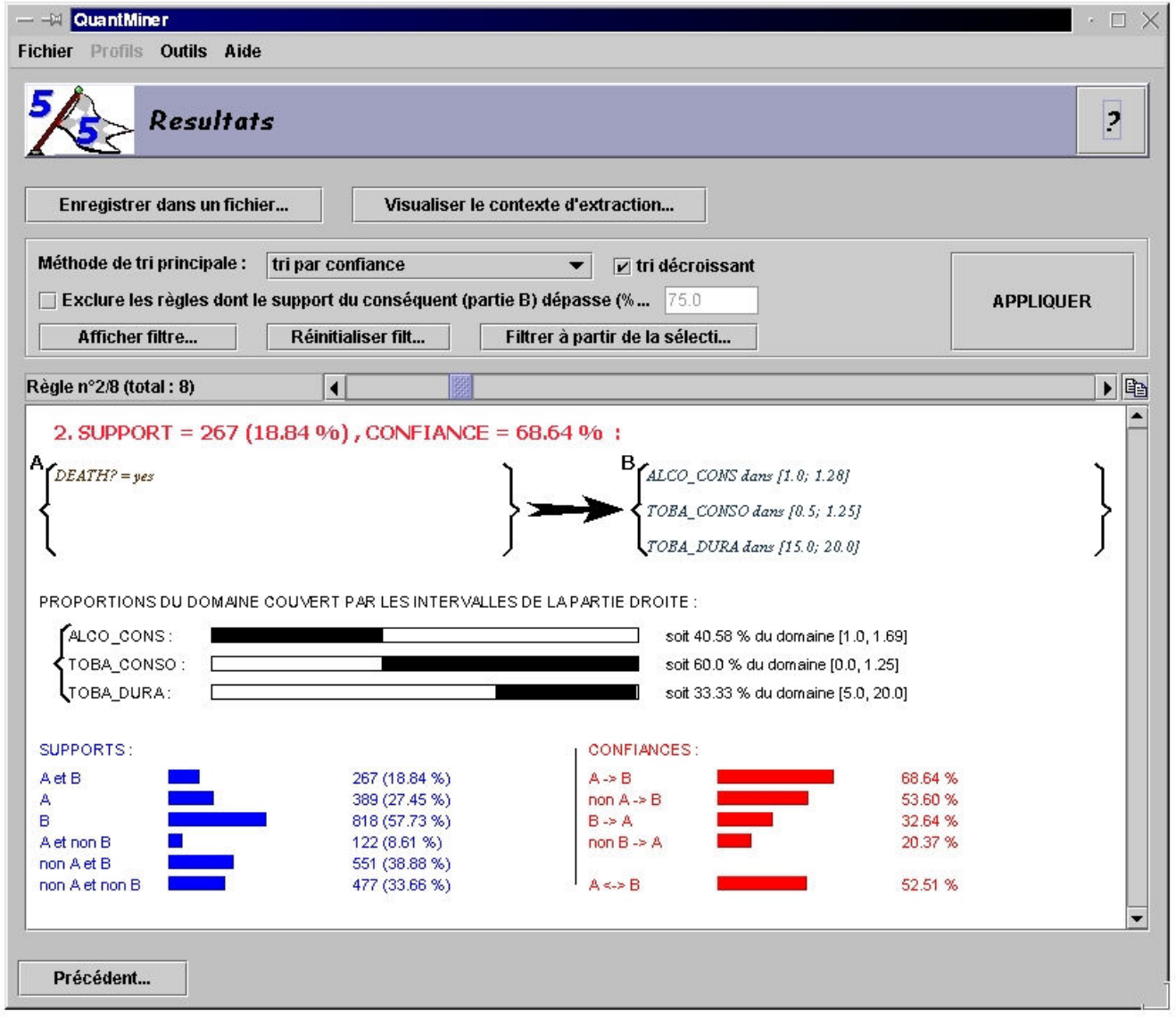
ÉTAPE 4 - OPTIMISATION

Exécution de l'algorithme d'optimisation.



ÉTAPE 5 - VISUALISATION

Affichage des résultats, avec tri sélectif des règles produites.



EXEMPLES DE RÈGLES

$$\text{DEATH?=}\ \text{YES} \Rightarrow \left\{ \begin{array}{l} \text{ALCO\_CONS} \in [1.0, 1.28] \\ \text{TOBA\_CONSO} \in [0.5, 1.25] \\ \text{TOBA\_DURA} \in [15, 20] \end{array} \right\} \begin{array}{l} 18\% \\ 68\% \end{array}$$

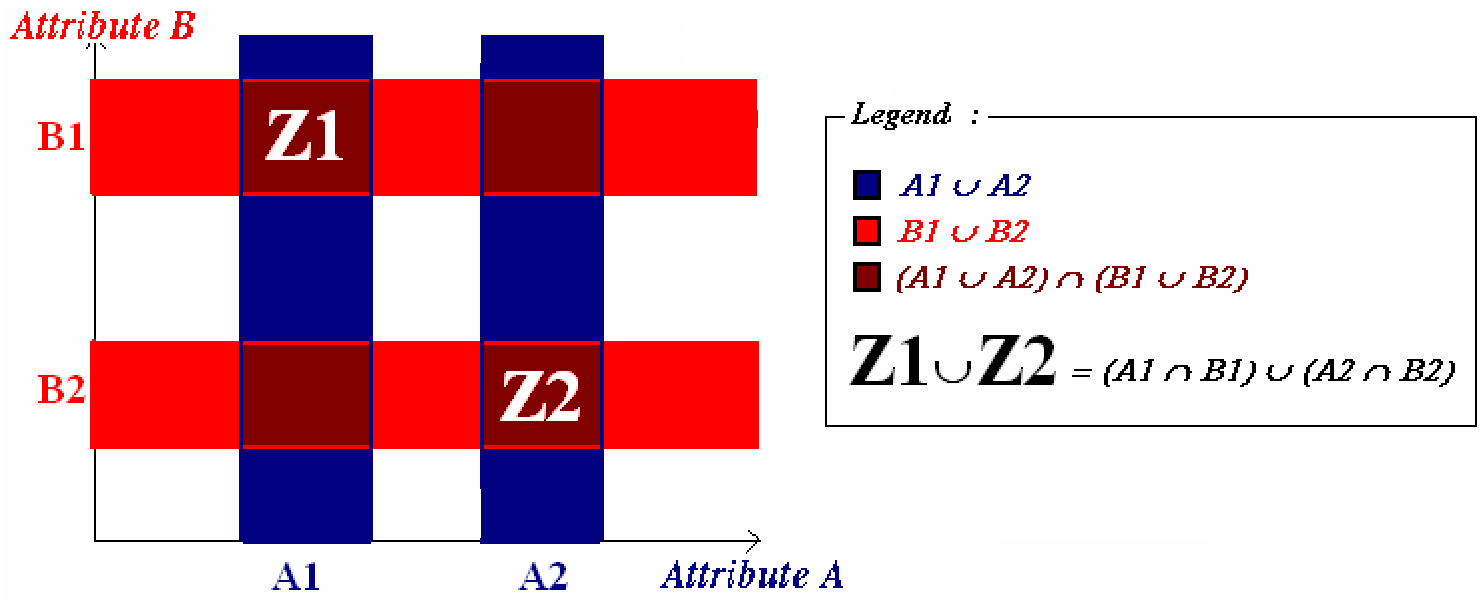
$$\text{GROOUP=N} \Rightarrow \left\{ \begin{array}{l} \text{ALCO\_CONS} \in [1.0, 1.2] \\ \text{BMI} \in [19.73, 27.77] \\ \text{TOBA\_CONSO} \in [0.0, 0.5] \end{array} \right\} \begin{array}{l} 13\% \\ 69\% \end{array}$$

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

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)

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

- Règles

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

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

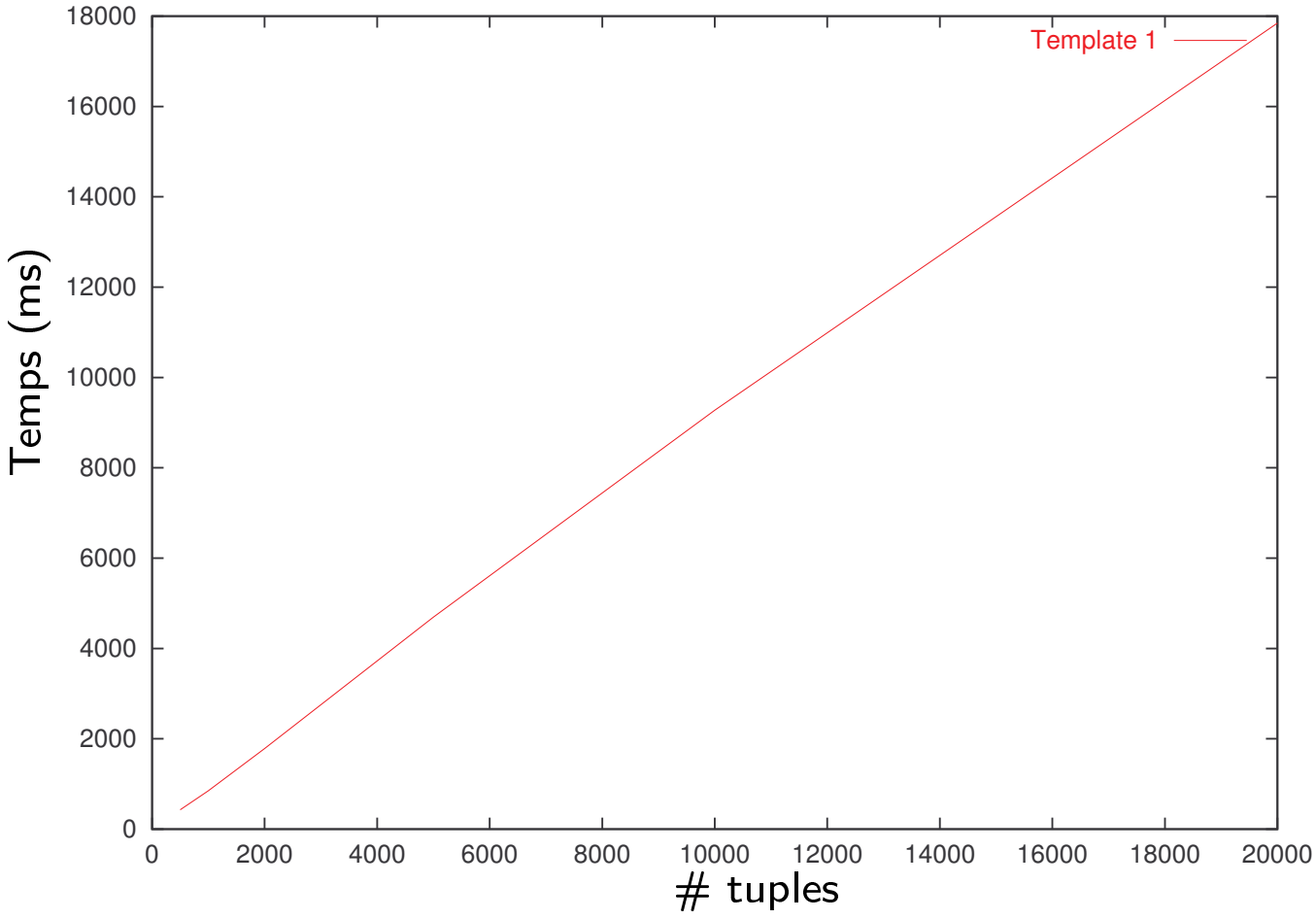
ANNEXE 2 : RÈGLES *versus* ITEMSETS

- Règles générées par QUANMINER

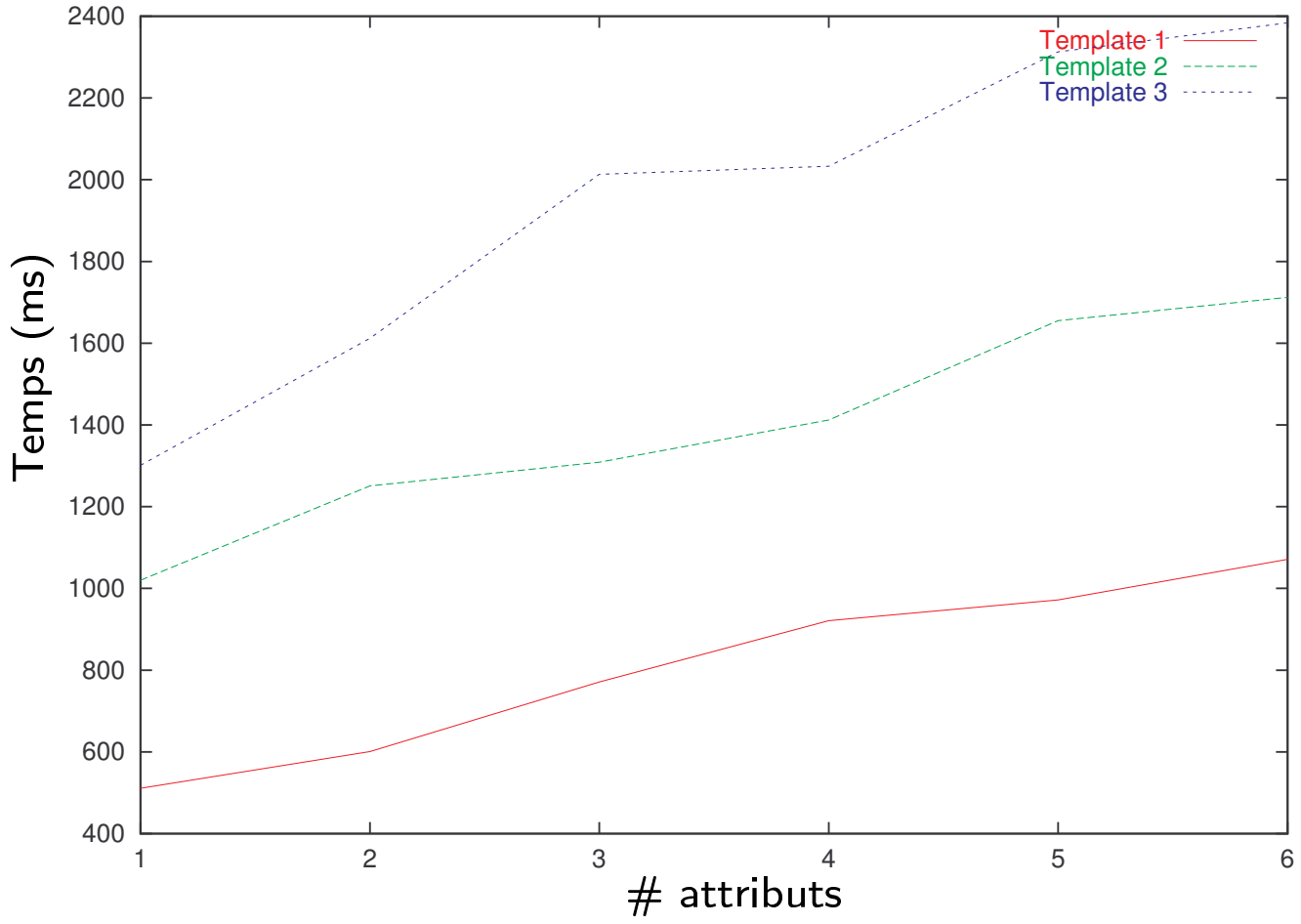
$$\texttt{Assist} \in \textbf{[0.1554, 0.2521]} \Rightarrow \left\{ \begin{array}{l} \texttt{Height} \in [180,193] \\ \texttt{Age} \in [23,30] \end{array} \right\} \begin{array}{l} 35\% \\ 72\% \end{array}$$

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

ANNEXE 4 : PERFORMANCES



Annexe 4 : Performances



Annexe 4 : Performances