Extraction de règles d'association pour la prédiction de valeurs manquantes

Sylvie Jami[1], Tao-Yan Jen[2], Dominique Laurent[3], George Loizou[1], Oumar Sy[3,4]

1. Birkbeck College - University of London
London - United Kingdom
{s.jami, george}@dcs.bbk.ac.uk
2. LI - Université de Tours
Tours - France
jen@univ-tours.fr
3. LICP - Université de Cergy-Pontoise Cergy-Pontoise - France
dominique.laurent@dept-info.u-cergy.fr
4. UFR Sciences Appliquées - Université Gaston Berger
Saint Louis - Sénégal
oumar.sy@ugb.sn

RÉSUMÉ. La présence de valeurs manquantes dans les bases de donnnées a suscité de nombreuses recherches dans le domaine de la découverte des connaissances, notamment en ce qui concerne la prédiction. Cependant, à notre connaissance, aucune de ces approches n'utilise les règles d'association pour la prédiction des valeurs manquantes. Dans cet article, nous montrons comment adapter les différents concepts et algorithmes par niveau liés aux règles d'association afin d'obtenir des règles fréquentes et de confiance 1, permettant la prédiction de valeurs manquantes dans une table relationnelle. La particularité des règles extraites dans notre approche est que leurs conséquents se présentent sous la forme d'intervalles ou d'ensembles de valeurs, selon que le domaine de l'attribut sur lequel les valeurs sont prédites est soit continu soit discret.

ABSTRACT. Missing values in databases has motivated many researches in the field of KDD, specially concerning prediction. However, to the best of our knowledge, no appraoch based on association rules has been proposed so far. In this paper, we show how to adapt the levelwise algorithm for the mining of association rules in order to mine frequent rules with a confidence equal to 1 from a relational table. In our approach, the consequent of extracted rules are either an interval or a set of values, according to the type of the predicted attribute.

MOTS-CLÉS : Bases de données, valeurs manquantes, règles d'association, prédiction.

KEYWORDS : Databases, missing values, association rules, prediction.

1. Introduction

La présence de valeurs manquantes dans les bases de données suscite de nombreuses recherches dans le domaine de la découverte de connaissances. Parmi ces travaux de nombreuses approches ont été proposées dans le cadre de la prédiction (voir [2, 3, 4]). L'approche que nous présentons, qui est la continuation des travaux de [5], permet d'extraire des règles pour la prédiction de valeurs manquantes dans une table relationnelle. Il est important de noter que, dans notre approche, la prédiction est faite sur un attribut de type soit discret soit continu, sans qu'aucune discrétisation du domaine ne soit nécessaire, comme cela est le cas pour les arbres de décision ([8]). Dans cet article, nous présentons cette approche et nous montrons que, contrairement à [5], les règles permettant la prédiction peuvent être extraites en utilisant un algorithme par niveau de type Apriori ([1]). On notera que la problématique des valeurs manquantes dans la découverte de règles d'association est abordée dans [9, 10], mais les auteurs ne traitent pas de la prédiction dans ces travaux.

L'article est organisé comme suit : Dans la section 2, nous donnons les notations et définitions nécessaires ainsi que la forme des règles extraites. Dans la section 3, nous introduisons la mesure de gain de précision et nous en donnons une propriété fondamentale. L'algorithme général de notre méthode est donné en section 4. Enfin, la section 5 présente notre proposition pour effectuer des prédictions à partir des règles calculées par notre algorithme. De plus, dans cette section, nous rappelons brièvement des résultats expérimentaux obtenus antérieurement dans [5] et nous concluons l'article.

2. Règles de prédiction

2.1. Définitions et notations

Notre approche utilise les définitions et notations de base du modèle relationnel ([6]). Nous considérons un ensemble fini d'attributs $U = \{A_1, A_2, \ldots, A_n\}$ et nous supposons qu'à chaque A_i $(i = 1, 2, \ldots, n)$ est associé un *domaine* de A_i noté $dom(A_i)$ qui est soit un ensemble discret soit un ensemble continu. Une *relation* sur U est un ensemble fini de tuples d'arité n pouvant éventuellement contenir des valeurs manquantes, dénotées par le seul symbole '?'. Soit R une telle relation sur U, nous notons $\overline{R}$ la relation sur U obtenue en éliminant de R tous les tuples contenant au moins une valeur manquante.

De plus, nous appelons le *domaine actif* de l'attribut A_i dans R, noté $adom(A_i)$, soit l'ensemble des valeurs de $dom(A_i)$ présentes dans R, si $dom(A_i)$ est de type discret, soit l'intervalle $[\mu_i, \nu_i]$ où μ_i et ν_i sont respectivement les plus petite et plus grande valeurs de $dom(A_i)$ présentes dans R, si A_i est de type continu.

Dans cet article, nous supposons fixé un attribut A_{i_0} de U, appelé *attribut de prédiction*, sur lequel les prédictions sont effectuées. Les règles sont définies comme suit.

Définition 1 - Règle de prédiction. *On appelle* règle de prédiction *ou simplement* règle, *toute règle de la forme* $(\Gamma \Rightarrow A_{i_0} \in E_\Gamma)$, *notée* $\langle \Gamma, E_\Gamma \rangle$ *ou* $\langle \Gamma \rangle$, *où :*

– Γ *est la conjonction (assimilée à un ensemble) de conditions élémentaires* $A_i = v_i$ *où* A_i *est un attribut de* U *différent de* A_{i_0} *et* $v_i \in dom(A_i)$.

– E_Γ *est défini par* $E_\Gamma = \{v \in dom(A_{i_0}) \mid (\exists t \in \overline{R})(t \models \Gamma \text{ et } t.A_{i_0} = v)\}$, *si* $dom(A_{i_0})$ *est discret, ou, si* $dom(A_{i_0})$ *est continu, par* $E_\Gamma = [\mu_\Gamma, \nu_\Gamma]$ *avec*
$\mu_\Gamma = \min\{v \in dom(A_{i_0}) \mid (\exists t \in \overline{R})(t \models \Gamma \text{ et } t.A_{i_0} = v)\}$ *et*
$\nu_\Gamma = \max\{v \in dom(A_{i_0}) \mid (\exists t \in \overline{R})(t \models \Gamma \text{ et } t.A_{i_0} = v)\}$.

Pour tout t *de* $\overline{R}$, t *satisfait* Γ *(respectivement* E_Γ*), noté* $t \models \Gamma$ *(respectivement* $t \models E_\Gamma$*), si pour tout* $A_i = v_i$ *de* Γ, $t.A_i = v_i$ *(respectivement* $t.A_{i_0} \in E_\Gamma$*).*

Exemple 1 *Dans l'article, nous considérons l'univers* U *constitué des quatre attributs* $VARIETE$, $CATEG$, $PURETE$ *et* $FGERM$. *La relation* $\overline{R}$ *utilisée est donnée ci-dessous et l'attribut de prédiction considéré est* $FGERM$, *qui est supposé de type continu. On suppose de plus que* $dom(FGERM) = [90, 98]$.
En utilisant les définitions ci-dessus, la règle $\rho : (VARIETE = JAYA) \Rightarrow FGERM \in [90, 95]$, *ou* $\langle VARIETE = JAYA, [90, 95] \rangle$, *est une règle de prédiction.*

VARIETE	CATEG	PURETE	FGERM
JAYA	BASE	BONNE	93
JAYA	BASE	BONNE	90
JAYA	PBASE	FAIBLE	95
JAYA	CERT R1	MOYENNE	93
JAYA	CERT R1	MOYENNE	95
JAYA	CERT R2	BONNE	95
IR1529	PBASE	MOYENNE	95
IR1529	PBASE	MOYENNE	98
IR1529	CERT R1	BONNE	90
IKP	BASE	BONNE	90
IKP	BASE	BONNE	93
JAYA	CERT R1	BONNE	92
IR1529	CERT R1	MOYENNE	92
JAYA	BASE	MOYENNE	92

2.2. Support et confiance

Le support et la confiance d'une règle sont définis de manière analogue à [1].

Définition 2 - Support. *Soit* $\overline{R}$ *une relation sur* U *ne contenant aucune valeur manquante,* Γ *une condition et* $\langle \Gamma, E_\Gamma \rangle$ *une règle.*

Le support de Γ dans $\overline{R}$, *noté* $sup(\Gamma, \overline{R})$, *est le rapport :* $|\{t \mid t \models \Gamma\}| \ / \ |\overline{R}|$

Le support de $\langle \Gamma, E_\Gamma \rangle$ dans $\overline{R}$, *noté* $sup(\langle \Gamma, E_\Gamma \rangle, \overline{R})$, *est le rapport :*
$|\{t \mid t \models \Gamma \text{ et } t \models E_\Gamma\}| \ / \ |\overline{R}|$.

Étant donné un seuil de support S, *une condition* Γ *(respectivement une règle* $\langle \Gamma, E_\Gamma \rangle$*) est dite* fréquente dans $\overline{R}$ par rapport à S, *ou simplement* fréquente *si* S *et* $\overline{R}$ *sont fixés, si* $sup(\Gamma, \overline{R}) \geq S$ *(respectivement* $sup(\langle \Gamma, E_\Gamma \rangle, \overline{R}) \geq S$*).*

Définition 3 - Confiance d'une règle. *Soit* $\overline{R}$ *une relation sur* U *ne contenant aucune valeur manquante, et soit* $\langle \Gamma \rangle$ *une règle. La* confiance de $\langle \Gamma \rangle$ dans $\overline{R}$, *notée* $conf(\langle \Gamma \rangle, \overline{R})$, *est le rapport :* $sup(\langle \Gamma \rangle, \overline{R}) / sup(\Gamma, \overline{R})$.

Exemple 2 *Si l'on reprend la règle* $\langle VARIETE = JAYA, [90, 95] \rangle$ *de l'exemple 1 ci-dessus, il est facile de voir que son support dans* $\overline{R}$ *est égal à* $8/14$. *Donc, pour un seuil de support de* $S = 0.14$, *cette règle est fréquente dans* $\overline{R}$ *par rapport à* S. *De plus, comme le support de la condition* $VARIETE = JAYA$ *est égal à* $8/14$, *la confiance de cette règle dans* $\overline{R}$ *est 1.*

Puisque $\overline{R}$ est *fixée*, $sup(\Gamma, \overline{R})$, $sup(\langle \Gamma \rangle, \overline{R})$ et $conf(\langle \Gamma \rangle, \overline{R})$ sont respectivement notés $sup(\Gamma)$, $sup(\langle \Gamma \rangle)$ et $conf(\langle \Gamma \rangle)$. Nous terminons cette section par l'énoncé de propriétés concernant les règles.

Proposition 1 *Soit* Γ *une condition, alors* $sup(\langle \Gamma \rangle) = sup(\Gamma)$ *et donc* $conf(\langle \Gamma \rangle) = 1$. *Si* Γ' *est une condition telle que* $\Gamma \subseteq \Gamma'$, *alors* $E_{\Gamma'} \subseteq E_\Gamma$.

3. Gain de précision

Le gain de précision, introduit dans [5], mesure la réduction relative de l'intervalle (ou de l'ensemble) lorsque la condition d'une règle est raffinée par l'ajout d'une ou plusieurs conditions de la forme $A_i = v_i$. Dans la définition ci-dessous, si E est un ensemble, $|E|$ désigne la cardinalité de E, et si E est l'intervalle $[\mu, \nu]$, $|E|$ désigne la différence $\nu - \mu$.

Définition 4 - Gain de précision. *Soit* $\langle \Gamma \rangle$ *et* $\langle \Gamma' \rangle$ *deux règles telles que* $\Gamma \subseteq \Gamma'$. *Le* gain de précision *de* $\langle \Gamma' \rangle$ *par rapport à* $\langle \Gamma \rangle$, *noté* $gain(\Gamma, \Gamma')$ *est défini par :*

Si Γ *n'est pas vide,* $gain(\Gamma, \Gamma') = (|E_\Gamma| - |E_{\Gamma'}|) \ / \ |E_\Gamma|$

Si Γ *est vide,* $gain(\emptyset, \Gamma') = (|adom(A_m)| - |E_{\Gamma'}|) \ / \ |adom(A_m)|$.

Nous définissons maintenant le critère utilisant la mesure de gain de précision pour sélectionner une règle.

Définition 5 - Règle retenue. *Soit G un seuil de gain et soit $\langle\Gamma\rangle$ une règle telle que $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ où, pour $i = 1, 2, \ldots, k$, γ_i est une condition élémentaire. Γ est* retenue par rapport *à G si :*

– Lorsque $k = 1$ alors $gain(\emptyset, \Gamma) \geq G$

– Lorsque $k > 1$ alors, pour toute permutation θ de $\{1, 2, \ldots, k\}$, et pour tout $j = 1, 2, \ldots, k-1$, alors $gain(\Gamma_j, \Gamma_j \cup \{\gamma_{\theta(j+1)}\}) \geq G$ avec $\Gamma_j = \{\gamma_{\theta(1)}, \gamma_{\theta(2)}, \ldots, \gamma_{\theta(j)}\}$.

On constate ainsi que pour s'assurer que la règle apporte effectivement une réduction de la taille de l'ensemble prédit, tous les ordres d'écriture des conditions élémentaires de Γ doivent être considérés. L'exemple suivant illustre cette définition.

Exemple 3 *Dans le cadre de l'exemple 1, considérons les règles ρ_1 : $\langle(VARIETE = JAYA), [90, 95]\rangle$, ρ_2 : $\langle(CATEG = BASE), [90, 93]\rangle$ et ρ_3 : $\langle(VARIETE = JAYA, CATEG = BASE), [90, 93]\rangle$. Pour un seuil de gain de précision de 0.15, l'application de la définition 5 ci-dessus nécessite tout d'abord le calcul de $gain(\emptyset, \rho_1)$. Ce gain étant de 0.375, la règle est retenue. Ensuite, $gain(\rho_1, \rho_3)$ est calculé. Puisqu'il est égal à 0.4, on considère tous les ordres possibles d'écriture de la condition de ρ_3 soit $gain(\emptyset, \rho_2) = 0.4$ et $gain(\rho_2, \rho_3) = 0$. Le second gain étant inférieur à 0.15, la règle n'est pas retenue.*

Nous montrons qu'il est possible de tester si une règle doit être retenue sans considérer tous les ordres d'écriture de la condition de cette règle.

Proposition 2 *Soit G un seuil de gain de précision et $\langle\Gamma\rangle$ une règle telle que Γ contient au moins deux conditions élémentaires. $\langle\Gamma\rangle$ est retenue par rapport à G si et seulement si pour tout γ de Γ : $\langle\Gamma \setminus \{\gamma\}\rangle$ est retenue par rapport à G, et $gain(\Gamma \setminus \{\gamma\}, \Gamma) \geq G$.*

Exemple 4 *Si l'on reprend les règles ρ_1, ρ_2 et ρ_3 de l'exemple 3, et le seuil de gain de précision de 0.15, l'application à ρ_3 de la proposition 2 ci-dessus nécessite de savoir si ρ_1 et ρ_2 sont retenues. Puisque c'est le cas, on calcule $gain(\rho_1, \rho_3)$ et $gain(\rho_2, \rho_3)$. Comme le second gain est inférieur à 0.15, la règle n'est pas retenue.*

Il est important de noter par rapport aux calculs de l'exemple 3, que si l'on sait que ρ_1 ou ρ_2 n'est pas retenue, alors en appliquant la proposition 2 ci-dessus, aucun calcul n'est nécessaire pour savoir que ρ_3 n'est pas retenue.

La proposition 2 ci-dessus est utilisée de façon à faire des coupures, comme dans [1] : si au moins l'une des règles $\langle\Gamma \setminus \{\gamma\}\rangle$ (avec $\gamma \in \Gamma$) n'est pas retenue, alors $\langle\Gamma\rangle$ n'est pas retenue.

4. Algorithme d'extraction des règles de prédiction

Dans cette section, nous présentons un algorithme d'extraction des règles de prédiction basé sur Apriori. Il est important de noter les deux points suivants :

1) Comme Apriori, notre algorithme effectue un parcours par niveau du treillis et utilise l'anti-monotonie du support et du gain de précision pour retenir une règle.

2) Contrairement à Apriori, les règles sont engendrées dans la même phase que l'extraction des fréquents, car les règles calculées ont une confiance de 1 (proposition 1).

Algorithme 1

Entrée : *une table de données complètes $\overline{R}$, un seuil de support S, et un seuil de gain de précision G*
Sortie : *l'ensemble des règles de prédiction fréquentes et retenues*
Méthode :

Calculer $RC_1 = \{(\langle \Gamma_1 \rangle, sup(\langle \Gamma_1 \rangle)) \mid |\Gamma_1| = 1\}$
//**Nécessite un balayage de la table de données**
$L_1 = \{\langle \Gamma_1 \rangle \mid sup(\langle \Gamma_1 \rangle) \geq S$ *et* $gain(\emptyset, \Gamma_1) \geq G\}$
$k = 2$
Tant que $L_{k-1} <> \emptyset$ **faire**
//**Phase de génération des conditions de cardinalité** k **à partir de** L_{k-1}
$C_k = \{\Gamma_k \mid |\Gamma_k| = k\}$ *à partir des* Γ_{k-1} *dans* C_{k-1}
//**Phase d'élagage de** C_k
$C_k = \{\Gamma_k \in C_k \mid (\forall \gamma \in \Gamma_k)(\langle \Gamma_k \setminus \{\gamma\} \rangle \in L_{k-1})\}$
//**Calcul du support et construction de l'ensemble prédit**
//**Nécessite un balayage de la table de données**
$RC_k = \{(\langle \Gamma_k \rangle, sup(\langle \Gamma_k \rangle)) \mid \Gamma_k \in C_k\}$
//**Sélection des règles fréquentes et retenues**
$L_k = \{\langle \Gamma_k \rangle \mid \Gamma_k \in C_k$ *et* $(sup(\langle \Gamma_k \rangle) \geq S)$ *et* $(\forall \gamma \in \Gamma_k)(gain(\Gamma_k \setminus \{\gamma\}, \Gamma_k) \geq G)\}$
$k = k + 1$
Fin Tantque
Retourner $\bigcup_{i=1..k} L_i$

Nous donnons un exemple d'application de cet algorithme sur les données présentées dans la table de l'exemple 1, pour un seuil de support $S = 0.14$ et un seuil de gain de précision $G = 0.14$.

Exemple 5 *Pour $k = 1$, on peut voir que les conditions $CATEG = CERTR2$ et $PURETE = FAIBLE$ ne sont pas fréquentes, et que $\langle VARIETE = IR1529, [90, 98]\rangle$ a un gain de précision nul. Ces règles ne sont donc pas retenues. Après exécution on obtient :*

$$L_1 = \{ \begin{array}{l} \langle VARIETE = JAYA, [90, 95]\rangle, \langle VARIETE = IKP, [90, 93]\rangle, \\ \langle CATEG = BASE, [90, 93]\rangle, \langle CATEG = PBASE, [95, 98]\rangle, \\ \langle CATEG = CERTR1, [90, 95]\rangle, \langle PURETE = BONNE, [90, 93]\rangle, \\ \langle PURETE = MOYENNE, [92, 98]\rangle \ \} \end{array}$$

Les candidats de niveau 2 sont ensuite engendrés et leurs intervalles calculés. Pour le candidat $\langle (VARIETE = JAYA, CATEG = BASE), [90, 93]\rangle$, le premier calcul effectué est : $gain(\langle CATEG = BASE\rangle, (VARIETE = JAYA, CATEG = BASE))$. Ce gain étant nul, la règle n'est pas retenue.

On considère ensuite $\langle (VARIETE = JAYA, CATEG = CERTR1), [92, 95] \rangle$, *et on calcule* $gain(\langle (VARIETE = JAYA), (VARIETE = JAYA, CATEG = CERTR1) \rangle)$ *et* $gain(\langle (CATEG = CERTR1), (VARIETE = JAYA, CATEG = CERTR1) \rangle)$. *Les deux gains étant supérieurs à 0,15, la règle est retenue.*

Au troisième niveau nous n'avons qu'un seul candidat $\langle VARIETE = JAYA, CATEG = CERTR1, PURETE = MOYENNE \rangle$ *et on évalue les gains ci-dessous. Leurs valeurs étant 0.33, la règle est retenue.*

$$
\begin{array}{ll}
gain(& (CATEG = CERTR1, PURETE = MOYENNE), \\
 & (VARIETE = JAYA, CATEG = CERTR1, PURETE = MOYENNE)) \\
gain(& (VARIETE = JAYA, PURETE = MOYENNE), \\
 & (VARIETE = JAYA, CATEG = CERTR1, PURETE = MOYENNE)) \\
gain(& (VARIETE = JAYA, CATEG = CERTR1), \\
 & (VARIETE = JAYA, CATEG = CERTR1, PURETE = MOYENNE))
\end{array}
$$

5. Prédiction et résultats expérimentaux

Étant donné un n-uplet t de R dont la valeur sur A_{i_0} est inconnue, on peut prédire une approximation de $t.A_{i_0}$ en utilisant les règles de prédiction sur A_{i_0} extraites selon la méthode vue précédemment. La méthode que nous proposons ici est la suivante.

Soit $\overline{t} = \langle v_{i_1}, v_{i_2}, \ldots, v_{i_k} \rangle$ le tuple sur le schéma $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ obtenu à partir de t en ne considérant que les attributs où t est défini. Si l'on assimile $\overline{t}$ à la condition $\Gamma_t = \langle A_{i_1} = v_{i_1}, A_{i_2} = v_{i_2}, \ldots, A_{i_k} = v_{i_k} \rangle$, alors l'intervalle (ou l'ensemble) prédit sur A_{i_0} pour t est l'intersection de tous les intervalles (ou ensembles) E_i tels que $\langle \Gamma_i, E_i \rangle$ est une règle extraite avec $\Gamma_t \subseteq \Gamma_i$. Dans le cas où cette intersection est vide, alors aucune prédiction n'est possible, ce qui signifie intuitivement que les données de la table $\overline{R}$ ne permettent pas de traiter le cas.

Lorsque l'intersection est non vide, la proposition 1 montre que seules les règles $\langle \Gamma \rangle$ telles que Γ est maximale (au sens de l'inclusion) sont à considérer dans le calcul de la prédiction. On notera de plus que notre méthode est *cohérente* avec le contenu de $\overline{R}$, puisque si $\overline{t}$ est un sous-tuple d'un tuple t' de $\overline{R}$ et si au moins un des v_{i_j} de $\overline{t}$ apparaît dans une règle extraite, alors $t'.A_{i_0}$ appartient à l'intervalle (ou ensemble) prédit.

Par exemple, si R contient le tuple $t = \langle IKP, BASE, MOYENNE, ? \rangle$, alors $\overline{t} = \langle IKP, BASE, MOYENNE \rangle$, et grâce aux règles $\langle VARIETE = IKP, [90, 93] \rangle$ et $\langle PURETE = MOYENNE, [92, 98] \rangle$, l'intervalle prédit sur $FGERM$ est $[92, 93]$.

On trouvera dans [5] la description de tests réalisés sur des données synthétiques et réelles. Les résultats sont comparés aux résultats obtenus par l'algorithme *C4.5* ([8]). Ces tests, effectués en utilisant une méthode de prédiction basée sur la mesure de Piatetsky–Shapiro ([7]) qui est plus fine de celle décrite ci-dessus, montrent la pertinence de notre approche. En particulier, des expériences concernant la reconnaissance de caractères, ont

porté sur un ensemble de données de 20 000 lignes. Pour un seuil de support de 003 et un seuil de gain de précision de 0.1, 410 règles ont été extraites. Les résultats montrent que 99.62% des valeurs manquantes ont pu donner lieu à une prédiction et que 99.55% des règles extraites étaient correctes.

En conclusion, nous rappelons que l'approche présentée dans cet article permet l'extraction de règles de prédiction de confiance 1 et dont la partie droite est un ensemble de valeurs. Outre les mesures de support et de confiance, la notion de gain de précision a été introduite afin d'améliorer la qualité des règles extraites. Nous avons montré que, malgré ce critère supplémentaire, il est possible d'appliquer un algorithme par niveau de type Apriori pour extraire les règles de prédiction. Enfin, les résultats expérimentaux brièvement rappelés ci-dessus montrent la pertinence de l'approche. L'implémentation de l'algorithme présenté dans cet article est en cours, afin d'améliorer les performances du prototype de [5].

6. Bibliographie

1. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo. *Fast Discovery of Association Rules.* In Advances in Knowledge Discovery and Data Mining. AAAI-MIT Press, 1996.
2. M. Ingunn, E. Stenrud, U. Olsson. *Analysing Data Sets with Missing Data : An empirical Evaluation of Imputation Methods and Likelihood-Based Methods*. IEEE Transactions on Software Engineering, 27(11), 2001
3. Y. Fujikawa. *Efficient Algorithms for Dealing with Missing values in Knowledge Discovery*. Ph.D Thesis, School of Knowledge Science, Advanced Institute of Science and Technology, Japan, 2001
4. J. Han and M. Kamber. *Data Mining : Concepts and Techniques*, Morgan Kaufmann, 2000.
5. S. Jami. *Learning Quality Rules from Sparse and Uncertain Data*. Ph.D Thesis, University of London, Birkbeck College, 2000.
6. M. Levene, G. Loizou. *A Guided Tour of Relational Databases and Beyond*. Springer-Verlag, 1999.
7. G. Piatetsky-Shapiro. *Discovery, Analysis and Presentation of Strong Rules*. In Knowledge Discovery in Databases, AAAI Press, 1991.
8. R. Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann, 1993.
9. A. Ragel and B. Crémilleux. *Treatment of Missing Values for Association Rules*. Pacific-Asia Conference on Knowledege Discovery and Data Mining, 1998.
10. A. Ragel, B. Crémilleux. *MVC - A Preprocessing Method to Deal with Missing Values*. Knowledge-based Systems, Elsevier, 1999.