Utilisation des règles d'association pour la prédiction de valeurs manquantes

Tao-Yuan Jen*, Dominique Laurent*, Gorgoumack Sambe* **

*ETIS-CNRS, Université Cergy Pontoise,
F-95000 Cergy Pontoise
jen@u-cergy.fr, dlaurent@u-cergy.fr
**Université de Ziguinchor
BP : 523 Ziguinchor Sénégal
gsambe@univ-zig.sn

Résumé. Le traitement des valeurs manquantes est une problématique importante dans le domaine des entrepôts de données. Plusieurs solutions ont été proposées pour la prédiction de valeurs manquantes, présentant les caractéristiques suivantes : (i) la prédiction traite soit des valeurs continues soit des valeurs discrètes, et (ii) la prédiction est approximative (soit elle est associée à une probabilité soit elle concerne un ensemble de valeurs). Récemment, une méthode de prédiction permettant de traiter indépendamment les cas continu et discret a été proposée, en se basant sur les règles d'association. Cette méthode permet de prédire, avec une confiance *toujours égale à* 1, soit un ensemble de valeurs dans le cas discret, soit un intervalle de valeurs dans le cas continu.

Dans cet article, nous reprenons cette approche basée sur l'extraction de règles d'association et nous montrons comment générer des règles de prédictions portant sur une *unique valeur* et dont la confiance est toujours égale à 1. Afin d'obtenir de telles règles, notre méthode suppose que l'on dispose d'une hiérarchie décrivant des concepts généralisant les valeurs qui peuvent être prédites.

1 Introduction

Avec la mondialisation, la croissance et la compétition effrénée, les entreprises ont vu s'accroître le volume de leurs données. Ces données dispersées sur plusieurs sites de l'entreprise, sont regroupées et fédérées sur un seul support de données appelé *entrepôt de données*, à des fins de consultation et d'analyse.

La présence de *valeurs manquantes* dans les entrepôts de données est un problème crucial car les méthodes utilisées pour l'analyse de ces entrepôts produisent généralement des résultats erronés ou incomplets. Dans la mise en place d'un entrepôt de données, la phase de nettoyage des données est estimée entre 30 et 80% du temps de développement. Pour remédier au problème des valeurs manquantes, plusieurs solutions sont proposées (Kamber et Han (2005)) :

- ignorer les données comportant des valeurs manquantes,
- les remplacer manuellement,

- les remplacer automatiquement soit par une valeur fictive, soit par une valeur calculée,
- utiliser la valeur la plus probable prédite à l'aide d'*algorithmes de prédiction* comme la régression, les arbres de décision, et les réseaux de neurones (Kamber et Han (2005)).

La dernière solution qui exploite le plus d'information disponible pour faire la prédiction est la plus utilisée. Les travaux de Ragel et Cremilleux (1999); Shen et al. (2007); Jami et al. (2005) utilisent les *règles d'association* à des fins de prédiction.

Les valeurs prédites par les approches autres que celle de Jami et al. (2005) sont approximatives au sens ou elles sont le plus souvent associées à une probabilité. L'approche proposée dans Jami et al. (2005) est au contraire certaine, car les règles de prédiction ont une confiance de 1. Comme notre approche est également fondée sur des règles de confiance 1 et qu'à notre connaissance, la méthode de Jami et al. (2005) est la seule à traiter des règles de prédiction de confiance 1, nous comparons notre approche seulement à cette méthode. Toutefois, selon l'approche de Jami et al. (2005), l'approximation vient du fait qu'un *ensemble* de valeurs est prédit de manière certaine au sens où on est certain que la valeur manquante appartient à l'ensemble.

Dans cet article, nous proposons une méthode de prédiction selon laquelle une *seule valeur* est prédite de manière *certaine*. L'approche est basée sur les travaux de Jami et al. (2005), initialement introduits dans Jami et al. (1998), qui ont pour but d'extraire des règles permettant la prédiction de valeurs manquantes dans les cas où l'attribut fixé sur lequel la prédiction est faite est soit continu soit discret. Dans cette approche, on considère une table R définie sur un ensemble fixé d'attributs $\{A_1, A_2, \ldots, A_n\}$, parmi lesquels se trouve l'attribut à prédire noté A_{i_0}. Le but de l'approche est d'extraire des règles de la forme $\rho : (A_{i_1} = v_1, A_{i_2} = v_2, \ldots, A_{i_k} = v_k) \rightarrow (A_{i_0} \in E)$, où pour $j = 1 \ldots k, v_j$ appartient au domaine de l'attribut A_{i_j}, E est un sous-ensemble du domaine de A_{i_0} et :

1. dans la partie droite de ρ, notée plus simplement E, E est soit un intervalle, soit un ensemble de valeurs selon que l'attribut prédit A_{i_0} est de type continu ou discret,
2. la partie gauche de ρ, notée plus simplement Γ, est fréquente dans l'ensemble $\overline{R}$ des n-uplets de R ne contenant pas de valeurs manquantes (*i.e.,* la proportion dans $\overline{R}$ de n-uplets dont les valeurs sont égales aux v_i est supérieure à un seuil donné), et
3. la confiance de ρ est 1 (*i.e.,* la règle est satisfaite dans $\overline{R}$).

De plus, il est montré dans Jami et al. (2005) que si $\rho_1 : \Gamma_1 \rightarrow E_1$ et $\rho_2 : \Gamma_2 \rightarrow E_2$ sont deux règles telles que Γ_2 est une "sous-condition" de Γ_1, alors le support de ρ_2 est supérieur au support de ρ_1 et $E_1 \subseteq E_2$. Ces propriétés ont pour conséquence que, dans Jami et al. (2005), les règles de prédiction peuvent être engendrées selon un algorithme par niveau tel que Apriori (Agrawal et Srikant (1994)).

De plus, afin d'éviter d'engendrer des règles redondantes, la notion de gain de précision, associée à un seuil donné, est introduite dans Jami et al. (2005) comme suit. Si ρ_1 et ρ_2 sont deux règles telles que spécifiées ci-dessus, le gain de précision de ρ_2 par rapport à ρ_1 est défini par $\frac{(|E_2|-|E1|)}{|E_2|}$, où $|E|$ désigne la taille de E.

Une règle $\rho_1 : \Gamma_1 \rightarrow E_1$ n'est alors retenue que si, outre les conditions citées ci-dessus, pour toutes les règles retenues $\rho_2 : \Gamma_2 \rightarrow E_2$ telles que Γ_2 est une "sous-condition" de Γ_1, le gain de précision de ρ_2 par rapport à ρ_1 est supérieur au seuil de gain de précision donné.

On peut donc constater que, bien que l'approche de Jami et al. (2005) permette d'engendrer des règles de prédiction de confiance 1, les prédictions associées ne sont pas réduites à une

valeur unique, et de plus, que cette approche nécessite l'introduction d'un seuil de gain de précision, en plus du seuil de support.

Notre contribution est de fournir une méthode de calcul de règles de prédiction de confiance 1, basée sur Apriori (comme dans Jami et al. (2005)) et permettant de réduire l'ensemble prédit par les règles à une valeur *unique*, selon un *seuil unique*, à savoir le seuil de support minimal.

Pour cela, notre méthode suppose qu'une *hiérarchie* est définie sur l'attribut prédit, permettant ainsi de généraliser des ensembles de valeurs ou des intervalles par de nouveaux concepts, non présents dans les données. Cette méthode est donc très adaptée aux contextes dimensionnels des entrepôts de données et cubes OLAP, dans lesquels des hiérarchies définies sur les dimensions sont considérées.

Nous illustrons la méthode présentée dans Jami et al. (2005), ainsi que celle présentée dans cet article, sur la relation R de la table 1. Cette relation contient des données cliniques codées. Un patient est décrit par son identifiant (ID) avec trois attributs X, Y et Z ayant pour domaines respectifs $\{1, 2, 3, 4, 5\}$, $\{10, 20, 30, 40, 50\}$ et $\{100, 200, 300, 400\}$. De plus, un champ prescription, qui est l'attribut prédit, donne le nom du médicament prescrit au patient selon les symptômes X, Y ou Z qu'il présente.

ID	X	Y	Z	prescription
1	1	10	100	efferalgan
2	1	10	100	aspegique
3	1	20	200	betsenol
4	1	30	100	?
5	1	30	300	efferalgan
6	1	30	300	betsenol
7	3	10	100	aspegique
8	1	30	200	?
9	3	10	100	efferalgan
10	1	30	100	dafalgan
11	3	?	100	?
12	2	30	300	dafalgan
13	1	?	400	?
14	1	40	100	betsenol
15	2	20	300	betsenol
16	2	30	100	aspegique
17	2	20	300	coversyl
18	4	40	400	coversyl
19	5	50	400	coversyl

X	Y	Z	prescription
1	10	100	efferalgan
1	10	100	aspegique
1	20	200	betsenol
1	30	300	efferalgan
1	30	300	betsenol
3	10	100	aspegique
3	10	100	efferalgan
1	30	100	dafalgan
2	30	300	dafalgan
1	40	100	betsenol
2	20	300	betsenol
2	30	100	aspegique
2	20	300	coversyl
4	40	400	coversyl
5	50	400	coversyl

TAB. 1 – *Table des données R et la table extraite $\overline{R}$*

Considérons la table complète $\overline{R}$ issue de la table 1 et contenant tous les n-uplets de R sans valeurs manquantes.

Selon Jami et al. (2005), pour un seuil de support de 0.1 et un seuil de gain de précision de 0.2, la règle $\rho_1 : (X = 1) \rightarrow \{$aspegique, dafalgan, efferalgan, betnesol$\}$ a un support de $7/15$ et un gain de précision (calculé dans ce cas par rapport à toutes les valeurs possibles sur l'attribut de prédiction) de $4/5$. Cette règle est donc retenue et est interprétée comme suit :

pour X, l'observation du symptôme 1 est suffisamment fréquente $(7/15)$ pour conclure avec une confiance de 1 que le médicament prescrit est l'un des quatre cités.

De même, la règle $\rho_2 : (Y = 30) \rightarrow \{$aspegique, dafalgan, efferalgan, betnesol$\}$ est retenue car son support est $5/15$ et son gain de précision est $4/5$.

En revanche, la règle $\rho_3 : (X = 2) \rightarrow \{$aspegique, dafalgan, efferalgan, betnesol, coversyl$\}$ n'est pas retenue, car même si son support $(4/15)$ est supérieur à 0.1, son gain de précision est nul. Toutefois, si l'on considère maintenant $\rho_4 : (X = 2, Y = 30) \rightarrow \{$aspegique, dafalgan$\}$, dont le support est $2/15$, ses gains de précison par rapport à ρ_2 et ρ_3 sont respectivement $2/5$ et $2/4$. Cette règle est donc retenue.

La méthode de prédiction proposée dans Jami et al. (2005) consiste à prédire une valeur manquante de l'attribut, prescription dans cet exemple, en utilisant l'intersection des ensembles prédits des règles dont le membre gauche est une sous-condition de la condition représentée par le n-uplet possédant la valeur manquante.

Afin d'obtenir une prédiction ne portant que sur une valeur *unique*, nous proposons d'utiliser une hiérarchie sur l'attribut à prédire. Cette hiérarchie est telle que les feuilles sont les valeurs de l'attribut à prédire. Dans notre exemple, nous considérons la hiérarchie de la figure 1 où les feuilles sont aspegique, dafalgan, efferalgan, betnesol, coversyl, et où all est la racine.

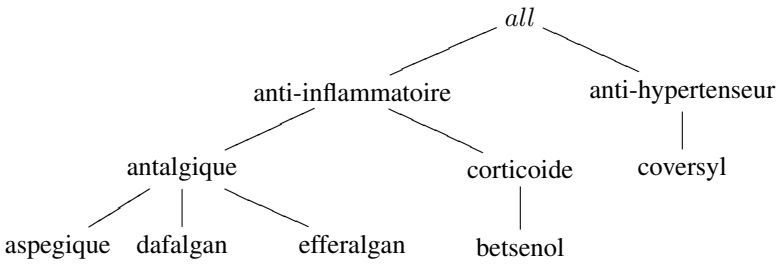

FIG. 1 – *Hiérarchie des médicaments*

Avec cette hiérarchie et le même seuil de support que précédemment, les règles ρ_1, ρ_2, ρ_3 et ρ_4 considérées ci-dessus deviennent alors respectivement :

1. $(X = 1) \rightarrow$ anti-inflammatoire, car anti-inflammatoire est la généralisation la plus spécifique dans la hérarchie de aspegique, dafalgan, efferalgan et betnesol.
2. $(Y = 30) \rightarrow$ anti-inflammatoire, pour la même raison que ci-dessus.
3. $(X = 2) \rightarrow all$, car toutes les valeurs possibles de prescription apparaissent dans l'ensemble prédit.
4. $(X = 2, Y = 30) \rightarrow$ antalgique, car antalgique est la généralisation la plus spécifique dans la hiérarchie de aspegique et dafalgan.

La première de ces règles est interprétée comme suit : pour X, l'observation du symptôme 1 est suffisamment fréquente pour conclure avec une confiance de 1 que le médicament prescrit est un anti-inflammatoire.

Nous remarquons toutefois que toutes les règles obtenues ne présentent pas un intérêt pour l'utilisateur. Ainsi, la règle $(X = 2) \rightarrow all$ n'est d'aucun intérêt puisqu'elle prédit all, c'est à dire tous les médicaments.

D'autre part, $\rho : (X = 1, Y = 30) \rightarrow$ anti-inflammatoire est une règle de prédiction potentielle, mais, puisque la valeur prédite est la même que pour les conditions $(X = 1)$ et $(Y = 30)$ seules, la règle ρ est considérée comme redondante. Cette règle ne sera pas retenue selon notre approche.

Afin de prendre en compte ces remarques, une règle ρ qui prédit la valeur all ne sera pas retenue, et une règle ρ plus spécifique qu'une règle ρ' ne sera retenue que si elle prédit une valeur plus spécifique (donc plus précise) que celle prédite par la règle ρ'.

Dans la section 2, nous définissons les concepts et notations utilisées par notre approche, puis, dans la section 3, nous donnons les algorithmes de génération des règles retenues. La section 4 propose une méthode de prédiction basée sur les règles retenues et donne certains résultats expérimentaux. Nous concluons en section 5 en précisant les directions envisagées à partir de ces travaux.

2 Définitions et notations

Considérons un ensemble fini d'attributs $U = \{A_1, A_2, \ldots, A_n\}$ et supposons qu'à chaque A_i $(i = 1, 2, \ldots, n)$ est associé un ensemble de valeurs, appelé domaine de A_i et noté $dom(A_i)$. Pour tout $i = 1, 2, \ldots, n$, $dom(A_i)$ est soit un ensemble discret soit un ensemble continu. Soit R une relation sur U, on note $\overline{R}$ la relation sur U obtenue en éliminant de R tous les n-uplets contenant au moins une valeur manquante.

Definition 1 - Condition de prédiction. *Une* condition élémentaire *est une expression de la forme $A_i = v_i, A_i \in U, v_i \in dom(A_i)$. Un n-uplet t sur U satisfait la condition élémentaire $A_i = v_i$, noté $t \models (A_i = v_i)$, si la restriction de t à A_i, notée $t.A_i$, est égale à v_i.*

Une condition de prédiction *(ou simplement condition) Γ est soit une condition élémentaire soit une conjonction de conditions élémentaires de la forme $(A_{i_1} = v_1 \wedge A_{i_2} = v_2 \wedge \ldots \wedge A_{i_k} = v_k)$ telle que, si j et j' sont deux entiers distincts de $\{1, \ldots, k\}$, alors $A_{i_j} \neq A_{i_{j'}}$. L'ensemble $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$ est appelé le* schéma *de Γ et noté $sch(\Gamma)$.*

Un n-uplet t satisfait Γ, noté $t \models \Gamma$, si t satisfait chaque condition élémentaire de Γ.

Toute condition de prédiction de la forme $(A_{i_1} = v_1 \wedge A_{i_2} = v_2 \wedge \ldots \wedge A_{i_k} = v_k)$ sera plus simplement notée $(A_{i_1} = v_1, A_{i_2} = v_2, \ldots, A_{i_k} = v_k)$.

De plus, si Γ et Γ' sont deux conditions de prédiction telles que toute condition élémentaire de Γ est également une condition élémentaire de Γ', on dit que Γ est une *restriction* de Γ'.

Dans l'exemple cité en introduction, $(X = 2)$ et $(X = 2, Y = 30)$ sont des conditions de prédiction, la première étant élémentaire de schéma $\{X\}$, et la seconde étant composée, de schéma $\{X, Y\}$. De plus, $(X = 2)$ est une restriction de $(X = 2, Y = 30)$.

Comme dans Jami et al. (2005), nous supposons que l'attribut sur lequel les ensembles prédits sont calculés est fixé, et nous notons A_{i_0} cet attribut. De plus, nous supposons que l'attribut prédit est associé à une *hiérarchie*, notée $H(A_{i_0})$, ayant pour racine all et pour feuilles les éléments de $dom(A_{i_0})$ dans le cas discret, et dans le cas continu des intervalles deux à deux disjoints dont l'union est égale à $dom(A_{i_0})$.

Ainsi, dans le cas où A_{i_0} est continu, $H(A_{i_0})$ permet de discrétiser de manière hiérarchique l'ensemble $dom(A_{i_0})$. Par conséquent, dans notre approche, les cas continu et discret concernant l'attribut prédit, sont traités de manière analogue.

Pour un *nœud* donné X de $H(A_{i_0})$, tout nœud parent (direct ou indirect) de X est un *ancêtre* de X, et tout nœud fils (direct ou indirect) de X est un *descendant* de X.

Un nœud X est dit *plus spécifique* qu'un nœud Y, noté $Y \preceq X$, si $X = Y$ ou si, dans $H(A_{i_0})$, X est un descendant de Y.

On remarque que, puisque $H(A_{i_0})$ est un arbre, pour tout ensemble H de nœuds de $H(A_{i_0})$, il existe un unique élément de $H(A_{i_0})$, noté $\min_{\preceq}(H)$, qui est maximal par rapport à $\preceq$ et tel que, pour tout X de H, $\min_{\preceq}(H) \preceq X$.

Par exemple, si l'on considère la hiérarchie de la figure 1, pour $H = \{$antalgique, efferalgan, betsenol$\}$, on a $\min_{\preceq}(H) =$ anti-inflammatoire, et pour $H = \{$antalgique, coversyl$\}$, on a $\min_{\preceq}(H) = all$.

De plus, si X est un nœud de $H(A_{i_0})$ et t un n-uplet sur U, on note $t \models X$ le fait que $t.A_{i_0}$ est soit égale à X soit égale à un nœud descendant de X. En d'autres termes, $t \models X$ si $X \preceq t.A_{i_0}$.

On remarque que, comme pour tout X de $H(A_{i_0})$ on a $all \preceq X$, tout n-uplet t satisfait $all \preceq t.A_{i_0}$. Par conséquent, toute règle dont la partie droite est all n'apporte aucune information. La forme des règles extraites dans notre approche est donc définie comme suit.

Definition 2 - Règle de prédiction. *Nous appelons* règle de prédiction *toute règle de la forme* $\Gamma \rightarrow X_\Gamma$*, où* Γ *est une condition de prédiction telle que* $A_{i_0} \notin sch(\Gamma)$ *et* $X_\Gamma \in H(A_{i_0})$ *avec* $X_\Gamma \neq all$.

Le support et la confiance d'une règle de prédiction sont définis comme suit.

Definition 3 - Support et confiance. *Soit* $\overline{R}$ *une table sur* U *et* $\rho : \Gamma \rightarrow X_\Gamma$ *une règle de prédiction.*

- *Le support de* Γ *est défini par :* $sup(\Gamma) = \frac{|\{t \mid t \models \Gamma)\}|}{|\overline{R}|}$
- *Le support de* X_Γ *est défini par :* $sup(X_\Gamma) = \frac{|\{t \mid t \models X_\Gamma)\}|}{|\overline{R}|}$
- *Le support de* ρ *est défini par :* $sup(\rho) = \frac{|\{t \mid t \models \Gamma, t \models X_\Gamma)\}|}{|\overline{R}|}$
- *La confiance de* $\rho : \Gamma \rightarrow X_\Gamma$ *est définie par :* $conf(\rho) = \frac{sup(\rho)}{sup(\Gamma)}$

Soit s *un seuil de support minimum, la condition de prédiction* Γ *(respectivement la règle de prédiction* ρ*) est dite* fréquente *si* $sup(\Gamma) \geq s$ *(respectivement* $sup(\rho) \geq s$*).*

On notera que si Γ et Γ' sont deux conditions de prédiction telles que Γ est une restriction de Γ', alors $sup(\Gamma) \geq sup(\Gamma')$.

Cette remarque, qui exprime la monotonie du support des conditions de prédiction, sera utilisée dans la section suivante concernant les algorithmes d'extraction des règles de prédiction retenues dans notre approche, qui sont définies ci-après.

Il est également important de noter que si $\rho : \Gamma \rightarrow X_\Gamma$ a une confiance de 1, alors $sup(\rho) = sup(\Gamma)$.

Definition 4 - Règle retenue. *Soit* s *un seuil de support, une règle de prédiction* $\rho : \Gamma \rightarrow X_\Gamma$ *est* retenue par rapport à s *si :*

- ρ *est fréquente (*i.e., $sup(\rho) \geq s$*),*
- $conf(\rho) = 1$,
- *Pour toute restriction* Γ' *de* Γ, $conf(\Gamma' \rightarrow X_\Gamma) \neq 1$.

La première condition de la définition ci-dessus exprime que nous ne nous intéressons qu'aux règles de prédiction qui s'appliquent dans une proportion suffisante de n-uplets de $\overline{R}$, et la deuxième condition exprime que seules les règles de prédiction *exactes* sont retenues dans notre approche.

La troisième condition de la définition ci-dessus exprime le fait que l'on ne retient que les règles de prédiction les plus générales, parmi celles ayant le même membre droit.

En effet, soit $\rho : \Gamma \rightarrow X_\Gamma$ et $\rho' : \Gamma' \rightarrow X_{\Gamma'}$ deux règles de prédiction de confiance 1 et telles que ρ est fréquente, $X_\Gamma = X_{\Gamma'}$, et Γ' est une restriction de Γ. Alors, puisque $conf(\rho) = conf(\rho') = 1$, on a $sup(\rho) = sup(\Gamma)$ et $sup(\rho') = sup(\Gamma')$. Donc, $sup(\rho') \geq sup(\rho)$ (car $sup(\Gamma') \geq sup(\Gamma)$, du fait que Γ' est une restriction de Γ). Par conséquent, ρ' est fréquente, et donc devrait être retenue selon les deux premiers critères.

Ces deux règles de prédiction ayant le même membre droit, ρ' est redondante, et ne doit donc pas être retenue. Par conséquent, la troisième condition de la définition ci-dessus permet d'éviter les redondances dans les règles retenues.

Ainsi, si l'on considère $\rho : (Y = 30) \rightarrow$ anti-inflammatoire et $\rho' : (Y = 30, Z = 300) \rightarrow$ anti-inflammatoire, ρ' est une règle de prédiction fréquente, mais pas une règle retenue.

3 Algorithme d'extraction des règles

Il est important de noter en premier lieu que la propriété de monotonie du support d'une condition de prédiction et le fait que l'on ne s'intéresse qu'à des règles de confiance égale à 1 ont pour conséquence qu'il est possible de calculer les règles retenues en utilisant un algorithme par niveau parcourant le treillis des conditions de prédiction (selon l'ordre partiel induit par la notion de restriction liée aux conditions de prédiction).

Ainsi, les algorithmes présentés dans cette section sont basés sur l'algorithme Apriori (Agrawal et Srikant (1994)), avec la particularité, par rapport à l'algorithme classique, que les règles retenues sont engendrées en même temps que les conditions de prédiction fréquentes.

Dans les algorithmes des figures 2 et 3, pour chaque niveau k du treillis des conditions de prédiction (*i.e.,* l'ensemble des conditions de prédiction constituées de k conditions élémentaires), on note respectivement C_k et L_k les ensembles de conditions de prédiction candidates et fréquentes, et R_k dénote l'ensemble des règles retenues dont le membre gauche est dans L_k.

Il est facile de voir que notre méthode est correcte et complète, à savoir que seules, toutes les règles retenues (voir la définition 4) sont effectivement calculées.

En effet, l'algorithme Apriori étant lui-même correct et complet, toutes les conditions de prédiction fréquentes sont calculées par l'algorithme 1. De plus :

- Les règles produites sont des règles de prédiction car aucune règle dont le membre droit est all n'est retournée.
- Les règles produites sont des règles de confiance 1 car, pour toute règle produite $\Gamma \rightarrow X_\Gamma$, la valeur X_Γ est la généralisation la plus spécifique selon $H(A_{i_0})$ de toutes les

valeurs α sur A_{i_0} de $\overline{R}$ pour lesquelles il existe t dans $\overline{R}$ tel que $t \models \Gamma$ et $t.A_{i_0} = \alpha$, c'est à dire $X_\Gamma = \min_{\preceq}(\{\alpha | (\exists t \in \overline{R})(t \models \Gamma, t.A_{i_0} = \alpha)\})$
- Par conséquent, les règles produites sont fréquentes, puisque leurs membres gauches le sont.
- Enfin, la dernière boucle de l'algorithme principal (figure 2) garantit qu'aucune règle redondante, selon la troisième condition de la définition 4, n'est produite.

Exemple 1. Nous illustrons les algorithmes des figures 2 et 3 dans le cadre de la table $\overline{R}$ donnée en introduction de l'article, et avec un seuil de support de 0.1.

Au premier niveau, les conditions élémentaires fréquentes de L_1 et les règles retenues de R_1 sont calculées. On obtient alors les résultats suivants : $L_1 = \{(X = 1), (X = 2), (X = 3), (Y = 10), (Y = 20), (Y = 30), (Y = 40), (Z = 100), (Z = 300), (Z = 400)\}$ et les règles retenues sont données ci-dessous, avec leur support.

- $\rho_1 : (X = 1) \rightarrow$ anti-inflammatoire $s = 7/15$
- $\rho_2 : (X = 3) \rightarrow$ antalgique $s = 2/15$
- $\rho_3 : (Y = 10) \rightarrow$ antalgique $s = 4/15$
- $\rho_4 : (Y = 30) \rightarrow$ anti-inflammatoire $s = 5/15$
- $\rho_5 : (Z = 100) \rightarrow$ anti-inflammatoire $s = 7/15$
- $\rho_6 : (Z = 400) \rightarrow$ coversyl $s = 2/15$

Les conditions $(X = 2)$, $(Y = 20)$, $(Y = 40)$ et $(Z = 300)$ ne correspondent à aucune règle, car une telle règle aurait son membre droit égal à all. En effet, par exemple pour $(X = 2)$, les valeurs à considérer sur l'attribut prescription sont successivement betsenol, coversyl et aspegique et ont all comme unique ancêtre commun dans $H(A_{i_0})$.

Au niveau 2, les conditions élémentaires de L_1 sont combinées dans l'ensemble C_2 et on obtient $L_2 = \{(X = 1, Y = 10), (X = 1, Y = 30), (X = 1, Z = 100), (X = 1, Z = 300), (X = 2, Y = 20), (X = 2, Y = 30), (X = 2, Z = 300), (X = 3, Y = 10), (X = 3, Z = 100), (Y = 10, Z = 100), (Y = 20, Z = 300), (Y = 30, Z = 100), (Y = 30, Z = 300)\}$. Après la dernière boucle de l'algorithme principal, l'ensemble R_2 est alors réduit aux règles suivantes :

- $\rho_7 : (X = 2, Y = 30) \rightarrow$ antalgique $s = 2/15$
- $\rho_8 : (Y = 30, Z = 100) \rightarrow$ antalgique $s = 2/15$

En effet, pour $(Y = 20, Z = 300)$, les valeurs à considérer sont betsenol et coversyl, dont le seul ancêtre commun dans $H(A_{i_0})$ est all. De plus, par exemple pour $(X = 1, Y = 10)$ les valeurs à considérer sont efferalgan et aspegique, ce qui donnerait la règle $(X = 1, Y = 10) \rightarrow$ antalgique, qui est retirée à cause de ρ_3.

Au niveau 3, on a $L_3 = \{(X = 1, Y = 10, Z = 100), (X = 1, Y = 30, Z = 300), (X = 2, Y = 20, Z = 300), (X = 3, Y = 10, Z = 100)\}$, mais on peut voir que finalement, aucune règle prédite n'est retenue à ce niveau car toutes les règles de prédiction potentielles sont redondantes. Par suite, l'ensemble des règles produites dans notre exemple est constitué des règles $\rho_1, \ldots, \rho_8$ mentionnées précédemment.

Algorithme 1 : algorithme principal
Entrée : la table $\overline{R}$, un seuil de support s, la hiérarchie $H(A_{i_0})$
Sortie : l'ensemble des règles retenues.
Méthode
//Calcul des conditions fréquentes élémentaires et des règles retenues
//Nécessite une passe sur la table de données
$C_1 := \emptyset$
Pour tout t de $\overline{R}$ faire
Pour tout $A \neq A_{i_0}$ faire
Si $(A = t.A) \in C_1$ alors
$sup(A = t.A) := sup(A = t.A) + 1$
$X_{A=t.A} := \min_{\preceq}(X_{A=t.A}, t.A_{i_0})$
Sinon
$C_1 := C_1 \cup \{(A = t.A)\}$
$sup(A = t.A) := 1$
$X_{A=t.A} := t.A_{i_0}$
Fin Si
Fin Pourtout
Fin Pourtout
$L_1 := \{\Gamma_1 \in C_1 \mid sup(\Gamma_1) \geq s\}$
$R_1 := \{(\Gamma_1 \rightarrow X_{\Gamma_1}) \mid \Gamma_1 \in L_1 \wedge X_{\Gamma_1} \neq all\}$
$k := 2$
$R_2 := \emptyset$
Tant que $L_{k-1} \neq \emptyset$ faire
//Génération et élagage des candidats selon la méthode de Agrawal and Srikant (1994)
$C_k := join(L_{k-1})$
$C_k := C_k - \{\Gamma_k \in C_k \mid (\exists \gamma \in \Gamma_k)(\Gamma_k - \gamma \notin L_{k-1})\}$
//Calculs de L_k et R_k. Nécessite une passe sur la table de données
Calculer L_k et R_k
$k := k + 1$
$R_k := \emptyset$
Fin Tantque
Pour tout $\Gamma_i \rightarrow X_{\Gamma_i}$ et $\Gamma_j \rightarrow X_{\Gamma_j}$ de $\bigcup_k R_k$ faire
Si Γ_i est une restriction de Γ_j et $X_{\Gamma_i} = X_{\Gamma_j}$ alors
Retirer $\Gamma_j \rightarrow X_{\Gamma_j}$ de $\bigcup_k R_k$
Fin Si
Fin Pourtout
Retourner $\bigcup_k R_k$

FIG. 2 – *Algorithme d'extraction des règles retenues*

Algorithme 2 : Calculs de L_k et R_k
Entrée : ensemble des candidats C_k
Sortie : les ensembles L_k et R_k
Méthode
  Pour tout Γ_k de C_k faire
    $X_{\Gamma_k} := NIL$
    $sup(\Gamma_k) := 0$
  Fin Pourtout
  Pour tout t de $\overline{R}$ faire
    Pour tout Γ_k de C_k faire
      Si $t \models \Gamma_k$ alors
        $sup(\Gamma_k) := sup(\Gamma_k) + 1$
        Si $X_{\Gamma_k} = NIL$ alors $X_{\Gamma_k} := t.A_{i_0}$ Sinon $X_{\Gamma_k} := \min_{\preceq}(X_{\Gamma_k}, t.A_{i_0})$
      Fin Si
    Fin Pourtout
  Fin Pourtout
  $L_k := \{\Gamma_k \in C_k \mid sup(\Gamma_k) \geq s\}$
  $R_k := \{\Gamma_k \rightarrow X_{\Gamma_k} \mid (\Gamma_k \in L_k) \wedge (X_{\Gamma_k} \neq all)\}$
  Retourner L_k et R_k

FIG. 3 – *Algorithme de calcul de L_k et de R_k*

4 Méthode de prédiction et résultats expérimentaux

4.1 Méthode de prédiction

Si nous considérons un cas de prédiction comme une condition de prédiction, le but est alors d'appliquer les règles retenues pour associer une valeur *unique* de prédiction, quand cela est possible.

Si P est une condition de prédiction et $\rho : \Gamma \rightarrow X_\Gamma$ une règle retenue, on dit que ρ est *applicable pour* P si Γ est une restriction de P. On note alors $Pred(P)$ l'ensemble des règles applicables pour P.

Notre méthode de prédiction est la suivante : Si $Pred(P) = \emptyset$, alors aucune prédiction n'est possible, puisque aucune règle retenue ne peut être appliquée. Sinon, soit $\mu = \min_{\preceq}(\{X_\Gamma \mid (\Gamma \rightarrow X_\Gamma) \in Pred(P)\})$,

- si $\mu \neq all$, alors μ est la valeur prédite,
- sinon ($\mu = all$), aucune prédiction *fiable* n'est possible et l'intervention de l'utilisateur est alors requise.

On notera à propos du dernier cas ci-dessus, qu'un critère de choix peut être de déterminer la valeur prédite en ne considérant que la ou les règles de $Pred(P)$ dont le support est maximum.

Exemple 2. Nous illustrons cette méthode sur l'exemple de la table 1, à partir des règles retenues calculées dans l'exemple 1, et en considérant différents cas de prédiction.

- Pour le n-uplet d'identifiant 4 de la table R, *i.e.,* pour $P = (X = 1, Y = 30, Z = 100)$, on a $Pred(P) = \{\rho_1, \rho_4, \rho_5, \rho_8\}$. Dans ce cas, on a $\mu = \min_{\preceq}(\{$antalgique, anti-inflammatoire$\})$ = anti-inflammatoire, qui est ainsi la valeur prédite.

- Pour le n-uplet d'identifiant 11, *i.e.,* pour $P = (X = 3, Z = 100)$, on a $Pred(P) = \{\rho_2, \rho_5\}$ et donc ici, $\mu = \min_{\preceq}(\{$antalgique, anti-inflammatoire$\})$ = anti-inflammatoire. Donc la valeur prédite est de nouveau anti-inflammatoire.
- Pour $P = (X = 1, Z = 400)$, alors on a $Pred(P) = \{\rho_1, \rho_6\}$ et donc $\mu = \min_{\preceq}(\{$anti-inflammatoire, coversyl$\})$, soit $\mu = all$. Dans ce cas aucune prédiction fiable n'est possible, mais, en se basant sur le support, on peut néanmoins prédire la valeur anti-inflammatoire, puisque $sup(\rho_1) = 7/15$ et $sup(\rho_6) = 2/15$.

4.2 Résultats expérimentaux

L'algorithme de génération des règles fréquentes a été implémenté en C++ et testé sur un système Linux et un ordinateur de 2Ghz de CPU et 2 GO de mémoire centrale.

Pour effectuer les tests, nous avons utilisé deux bases de données provenant de UCI Machine Learning repository : El nino contenant des données océanographiques et météorologiques avec 12 attributs et 178080 n-uplets, et Abalone contenant des données sur une population d'haliotis (mollusque marin appelé abalone en anglais) avec 8 attributs et 4177 n-uplets.

Nous avons introduit de manière aléatoire, au niveau de la dernière colonne de type numérique dans les deux cas, des valeurs manquantes (5%, 10%, 25%, 40%, 50%), et nous avons construit une hiérarchie sur chacun de ces attributs.

Avec un support de 3%, le temps d'exécution est inférieur à 7s, le taux de prédiction, qui est égal au rapport du nombre de prédictions effectuées sur le nombre total de n-uplets ayant une valeur manquante, dépasse 75%, et plus de 99% des règles extraites prédisent un nœud parent de la valeur réelle qui a été remplacée par une valeur manquante.

Nous avons également comparé notre méthode avec celle de Jami et al. (2005) sur la base de la précision définie par $(1 - \frac{|E|}{|dom(A_{i_0})|})100$, où $|E|$ est la longueur ou la cardinalité de l'ensemble prédit.

Nous adaptons cette définition au contexte de l'utilisation des hiérarchies comme suit : la précision est l'expression $(1 - \frac{|X|}{|dom(A_{i_0})|})100$, où $|X|$ est le nombre de feuilles descendant du nœud X dans le cas discret, et la somme des longueurs des intervalles qui sont les feuilles descendant de X dans le cas continu.

Pour la base El nino, le taux de précision est de 30,89% avec la méthode de Jami et al. (2005) et de 55% avec notre méthode. Pour la base Abalone, la précision est de 44% avec la méthode de Jami et al. (2005) et de 37% avec notre méthode.

Ces résultats s'expliquent d'une part par le choix de la hiérarchie et d'autre part par la présence de valeurs isolées dans l'attribut à prédire. En effet, lorsqu'un n-uplet a une valeur isolée sur l'attribut à prédire, la partie droite de toute règle basée sur ce n-uplet est alors généralisée, impliquant une diminution du taux de précision. Toutefois, l'avantage de notre méthode par rapport à Jami et al. (2005),est de toujours fournir une seule valeur plus facilement exploitable qu'un intervalle ou un ensemble.

5 Conclusion

Dans cet article, nous avons proposé une méthode de prédiction de valeurs manquantes sur un attribut discret ou continu en utilisant des règles d'association. Cette méthode, qui utilise en

outre une hiérarchie définie sur l'attribut à prédire, permet d'extraire des règles fréquentes de confiance 1 et ayant une conséquence unique, en ne nécessitant qu'un seuil (de support). Nous projetons dans nos travaux futurs :

- d'étudier la possibilité de déterminer une hiérarchie automatiquement sur la base des ensembles prédits par la méthode de Jami et al. (2005), et
- d'étendre la méthode proposée par Jami et al. (2005) et celle proposée dans cet article, en exploitant l'approche de Ragel et Cremilleux (1998, 1999) pour obtenir une méthode de prédiction sur plusieurs attributs.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *VLDB'94*, pp. 487–499. Morgan Kaufmann.

Jami, S., T.-Y. Jen, D. Laurent, G. Loizou, et O. Sy (2005). Extraction de règles d'association pour la prédiction de valeurs manquantes. *Revue Africaine de la Recherche en Informatique et Mathématique Appliquée ARIMA Spécial CARI04*, 103–124.

Jami, S., X. Liu, et G. Loizou (1998). Learning from an incomplete and uncertain data set : the identification of variant haemoglobins. In *Workshop on IDAMP, ECAI'98*.

Kamber, M. et J. Han (2005). *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

Ragel, A. et B. Cremilleux (1998). Treatment of missing values for association rules. In *PAKDD '98*, Volume 1394 of *Lecture Notes in Computer Science*, pp. 258–270. Springer-Verlag.

Ragel, A. et B. Cremilleux (1999). Mvc - a preprocessing method to deal with missing values. *Knowledge-Based Systems 12*, 285–291.

Shen, J.-J., C.-C. Chang, et Y.-C. Li (2007). Combined association rules for dealing with missing values. *J. Inf. Sci. 33*(4), 468–480.

Summary

Dealing with missing values is an important issue in the field of data warehousing. Several solutions have been proposed in the literature for the prediction of missing values. In general, these approaches are such that: (i) the values to be predicted are either continue or discrete, and (ii) the prediction is not exact (as associated with a probability or as outputing a set of possible values). Recently, an approach for the prediction of missing values based on association rule mining has been introduced. The important features of this approach are that it generates prediction rules with confidence 1, on any kind of values (numeric or discrete). In this paper, we enhance this approach based on association rule mining by generating prediction rules for *single values*, continuous or discrete, with confidence 1. To this end, our method relies on the assumption that a hierarchy modeling concepts that generalize the values to be predicted, is available.