Base de caractérisation des valeurs manquantes

Leila Ben Othman[*,**] — **François Rioult**[**] — **Sadok Ben Yahia**[*] — **Bruno Crémilleux**[**]

[*] *Département des Sciences Informatiques,*
Faculté des Sciences de Tunis
Campus Universitaire,1060 Tunis, Tunisie.

sadok.benyahia@fst.rnu.tn

[**] *GREYC UMR 6072,*
Université de Caen Basse-Normandie,
Campus 2 Côte de Nacre,
14000 Caen, France.

{lbenothm,François.Rioult,Bruno.Cremilleux}@info.unicaen.fr

RÉSUMÉ. *Les données issues du monde réel ne sont pas toujours complètes, puisque certaines informations ne sont pas disponibles ou sont non renseignées. Dans ce cadre, le traitement des valeurs manquantes a suscité l'intérêt de plusieurs communautés scientifiques. Ainsi, expliquer l'origine des valeurs manquantes permet d'une part de mieux contrôler la qualité des données et d'autre part de proposer des méthodes de traitement adéquates, e.g., pour leur complétion. Dans cet article, nous proposons de caractériser précisément les mécanismes d'apparition des valeurs manquantes dans les données. Dans la littérature, ce mécanisme a été défini par Little et Rubin, selon un point de vue statistique. Nous montrons d'abord que ces modèles sont difficiles à utiliser pour les techniques de fouille de données actuelles, si l'on cherche à les détecter en examinant les données disponibles. Nous présentons ensuite, selon un point de vue complémentaire, comment caractériser le type des valeurs manquantes à l'aide d'une base de règles d'association non redondantes. Cette nouvelle caractérisation permet* a priori *de recommander des stratégies fines de complétion adaptées à des groupes d'objets plutôt que d'imposer le même traitement à toutes les valeurs. Finalement, nous illustrons expérimentalement cette caractérisation sur des données médicales réelles relatives à la maladie de* HODGKIN *d'une part et des* MENINGITES *infantiles d'autre part.*

ABSTRACT. *When tackling real-life datasets, it is common to face the existence of missing values within data. In this respect, handling missing data has grasped the interest of many scientific*

communities, specially that of statistics. Indeed, explaining the origin of the missing values appearance allows to better control the quality of the data, as well as proposing suitable handling methods, e.g., their completion. The abundant literature heavily relies on the missing value appearance models proposed by Little and Rubin. However, a careful scrutiny of these statistic-based models highlights that they constitute an actual hamper towards their use by data mining techniques. The main thrust of this paper is the proposition of a new model for missing values appearance. Such introduced models rely on the use of the proper implication basis. In addition, we show that these new models allow to straightforwardly advise adequate completion strategy. Carried out extensive experiments, on medical datasets, i.e. HODGKIN *and* MENINGITIS*, show how statistically these models are scattered within the considered datasets.*

MOTS-CLÉS : *Fouille de données, Valeurs manquantes, Données incomplètes, Règles d'association.*

KEYWORDS: *Data mining, Missing values, Incomplete data, Association rules.*

1. Introduction

Les données issues du monde réel ne sont pas toujours complètes, notamment parce que certaines informations ne sont pas disponibles ou sont non renseignées. Ceci semble constituer un phénomène inévitable et imprévisible, dû à de multitudes raisons : oubli de la part de l'utilisateur, refus de réponse lors de sondages, impossibilité d'acquisition de valeurs, *etc.* Bien que de nombreuses techniques dédiées à l'extraction des connaissances (réseaux de neurones, k-means, *etc.*) soient maintenant arrivées à maturité, elles demeurent très délicates à mettre en œuvre dans le cas où les données sont incomplètes. Éliminer les données présentant les valeurs manquantes lors du processus d'extraction constitue souvent une perte considérable d'information : leur traitement est nécessaire.

La problématique des valeurs manquantes suscite donc l'intérêt constant de la communauté d'analyse de données, matérialisé par la proposition d'un large éventail de méthodes. Parmi les statisticiens, nous avons les propositions de (Dempster *et al.*, 1977; Ghahramani *et al.*, 1994). Certains ont abordé cette problématique dans le domaine de la bioinformatique (Brock *et al.*, 2008; Ryan *et al.*, 2010), d'autres, dans la communauté de la fouille de données, l'ont abordée par l'exploitation de régularités locales (e.g., règles d'association (Ben Othman *et al.*, 2010; Calders *et al.*, 2007; Shen *et al.*, 2007; Jami *et al.*, 2005; Wu *et al.*, 2004; Ragel *et al.*, 1998), motifs séquentiels (Fiot *et al.*, 2007), mais aussi les représentations concises de motifs (Rioult *et al.*, 2006) ou encore les ensembles approximatifs (Nelwamondo *et al.*, 2007)).

Même si elle n'est pas toujours clairement formulée, l'hypothèse selon laquelle les valeurs manquantes apparaissent selon un modèle uniquement aléatoire, selon la classification proposée dans (Little *et al.*, 2002), est implicitement utilisée par ces méthodes. Dans la réalité, les valeurs manquantes ne sont pas nécessairement aléatoires (Pearson, 2006) : il existe le plus souvent une explication à l'absence de mesure d'une valeur. De plus, nous constatons dans de nombreux cas que le modèle aléatoire est trop restrictif et ne prend pas en considération les spécificités des origines potentielles des valeurs manquantes.

À notre connaissance, il n'existe pas de travaux qui ont étudié la pertinence de l'hypothèse aléatoire, ni précisé sous quelles conditions les valeurs manquantes sont aléatoires. Cependant, une analyse préliminaire sur le modèle d'apparition des valeurs manquantes permettrait de mieux adapter le traitement. Par exemple, lorsqu'une valeur manquante n'est pas aléatoire, elle est dite *informative* car elle permet de caractériser une situation particulière qui apporte une information sur son contexte d'apparition. Si nous considérons que lors d'un sondage, les personnes ayant un sur-poids ont tendance à le cacher, nous saurons que les personnes présentant une valeur manquante sur l'attribut *Poids* cachent potentiellement un sur-poids.

Ce type d'information devrait être exploité par les méthodes de traitement des valeurs manquantes, mais le modèle d'apparition des ces dernières est rarement discuté, souvent réduit au seul modèle aléatoire. Cette limitation et la non prise en compte d'information importantes sur l'origine des valeurs manquantes pénalisent la complé-

tion (Delavallade *et al.*, 2007; Fiot *et al.*, 2007). Nous souhaitons dans ce travail souligner l'importance d'effectuer une analyse préliminaire sur les origines des valeurs manquantes afin de prescrire une méthode de traitement adéquate.

Le recours à l'hypothèse aléatoire des valeurs manquantes est souvent justifié par la difficulté liée à la caractérisation même des modèles existants (Little *et al.*, 2002). En effet, ces modèles (cf. section 2.2) sur les valeurs manquantes utilisent des informations sur le monde *réel* pour expliquer le monde *mesuré*. Or, lors du traitement de données incomplètes, seul le monde mesuré est disponible. Nous avons donc besoin de nouveaux modèles pour caractériser la présence des valeurs manquantes *en fonction des données mesurées.*

Une idée forte de cet article est de proposer une caractérisation plus fine de l'origine des valeurs manquantes afin de mieux y faire face dans le cadre d'un processus d'extraction de connaissances à partir des bases de données ou encore en amont d'un traitement statistique des valeurs manquantes. Ainsi, l'impact du présent travail est double : d'une part, il permet de mieux comprendre les causes des valeurs manquantes et contribue ainsi à l'amélioration de la qualité des données ; d'autre part des méthodes de complétion plus efficaces peuvent être développées, tirant ainsi bénéfice de ces modèles d'apparition des valeurs manquantes. En effet, l'examen des données disponibles peut montrer que les valeurs manquantes présentent des régularités. L'identification de ces régularités permet de proposer une valeur de remplacement plus pertinente que celle issue d'une complétion reposant sur un modèle d'apparition aléatoire. Nous défendons donc l'hypothèse selon laquelle la présence d'une valeur manquante peut en elle-même être une information porteuse de connaissance implicite, qui pourrait s'avérer de grande utilité lors de l'analyse de données incomplètes.

En outre, une valeur manquante sur un attribut n'admet pas toujours une seule explication. La section 6, relative à un cas d'étude portant sur la maladie de Hodgkin (un cancer du système lymphatique), illustre précisément ce point. Les données correspondantes mesurent l'envahissement par les cellules cancéreuses de ganglions particuliers. Certaines données sont manquantes car leur format a évolué au cours des années, certains ganglions n'étant pas examinés dans les premiers temps de l'étude. Cependant, les mêmes données peuvent aussi être manquantes car le résultat de l'examen n'a pas été transmis. Il y a donc plusieurs explications à l'absence d'une valeur, qui dépendent des objets d'étude et il est illusoire de se contenter d'une unique caractérisation pour cet attribut qui soit valable sur l'ensemble des objets étudiés.

Nous proposons donc une nouvelle typologie des valeurs manquantes adaptée à des groupes d'objets spécifiques. Ceci permet d'affiner le niveau de granularité de toute méthode de complétion introduite en aval, mettant à profit ces caractérisations (*cf.* section 4). Pour cela, nous utiliserons une technique efficace de recherche de régularités (base d'implications propres, (Taouil *et al.*, 2001)) qui confère à la caractérisation obtenue une propriété de minimalité limitant la redondance des explications.

En résumé, la contribution de ce travail est de répondre à des interrogations, que nous jugeons cruciales, lors d'un processus de découverte de connaissances dans des

données incomplètes, et ceci en amont de toute méthode de complétion ou de traitement des valeurs manquantes :

– Quels modèles de valeurs manquantes sont détectables à l'examen des données disponibles ?

– Est-il possible d'expliquer la présence des valeurs manquantes ?

– Comment peut-on caractériser ces valeurs manquantes ?

– Peut-on avoir des caractérisations différentes pour un même attribut, selon des objets spécifiques ?

Le reste de cet article est organisé comme suit : la section 2 détaille le phénomène des valeurs manquantes en rappelant leurs modèles d'apparition proposés dans la littérature. Nous mettons aussi en exergue les limites de ces modèles et précisons le positionnement de notre travail. Nous présentons la base d'implications propres dans la section 3, un outil essentiel pour le calcul d'associations non redondantes et un moyen efficace pour l'extraire. La section 4 propose une nouvelle typologie des valeurs manquantes à partir des données mesurées et indique la forme des règles d'association qui permettent de les caractériser. Dans la section 5, nous montrons l'intérêt de l'utilisation de la base d'implications propres pour la caractérisation des valeurs manquantes. Les résultats d'une étude expérimentale de la nouvelle typologie, sur des données relatives à la maladie de Hodgkin d'une part et des méningites infantiles d'autre part sont enfin présentés dans la section 6. La section 7 conclut et livre quelques perspectives à propos de la complétion des valeurs manquantes.

2. Valeurs manquantes dans les bases de données

Dans cette section, nous commençons par introduire le matériel technique utile à la compréhension de ce travail : contexte booléen, réel ou mesuré. Nous présentons également les modèles de valeurs manquantes recensés dans la littérature.

2.1. *Définitions et notations*

Les données que nous étudions sont initialement sous le format "attribut/valeur". Considérons l'exemple illustré par la table 1 (a). Chaque objet est représenté par quatre attributs A_1, A_2, A_3 et A_4. À chaque attribut est associé un domaine de valeurs, *e.g.,* $dom(A_1) = \{a, b\}$, $dom(A_2) = \{c, d\}$, $dom(A_3) = \{e, f, g\}$ et $dom(A_4) = \{h, i\}$. Un attribut A_i présente parfois une valeur non renseignée, dite *valeur manquante*, notée par " ?". La table 1 (b) représente le même exemple que la table 1 (a) avec des valeurs manquantes.

Lors d'un processus d'extraction de règles d'association, les données sont généralement représentées selon un format binaire ou transactionnel utilisant des *items*, où les attributs quantitatifs sont discrétisés. Nous donnons ci-dessous la définition correspondante d'un contexte réel.

	A_1	A_2	A_3	A_4
o_1	a	c	**e**	h
o_2	**b**	c	e	**i**
o_3	a	c	**f**	h
o_4	a	d	f	**i**
o_5	**a**	c	f	**i**
o_6	b	**c**	f	h
o_7	a	**d**	g	**i**
o_8	**b**	d	g	**i**

	A_1	A_2	A_3	A_4
o_1	a	c	**?**	h
o_2	**?**	c	e	**?**
o_3	a	c	**?**	h
o_4	a	d	f	**?**
o_5	**?**	c	f	**?**
o_6	b	**?**	f	h
o_7	a	**?**	g	**?**
o_8	**?**	d	g	**?**

Tableau 1. *Exemple de données au format attribut-valeur. (a) : contexte réel. (b) : contexte mesuré.*

Définition 1 (Contexte réel) *Un* contexte réel *est un triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, décrivant deux ensembles finis $\mathcal{O}$ d'objets (ou transactions), $\mathcal{I}$ d'items et une fonction $\mathcal{R}$ sur $\mathcal{O} \times \mathcal{I}$ prenant ses valeurs dans $\{présent, absent\}$. $\mathcal{R}(o, i) = présent$ signifie que l'item $i \in \mathcal{I}$ est présent dans l'objet $o \in \mathcal{O}$. Par contre, $\mathcal{R}(o, i) = absent$ indique l'absence de l'item i dans l'objet o.*

Comme nous l'avons vu, les données réelles ne sont pas toujours disponibles, seules les données mesurées le sont et elles peuvent être incomplètes. Nous modélisons la transformation des données réelles en données mesurées grâce à l'opérateur mv :

Définition 2 (Contexte mesuré) *Un opérateur mv (pour $missing\ value$) de modélisation des valeurs manquantes transforme un contexte réel $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ en un* ***contexte mesuré*** *noté $mv(\mathcal{K}) = (\mathcal{O}, \mathcal{I}, mv(\mathcal{R}))$. La nouvelle relation $mv(\mathcal{R})$ prend ses valeurs dans l'ensemble $\{présent, absent, manquant\}$.*

Un contexte réel $\mathcal{K}$ représente les données parfaites (sans valeurs manquantes, mais purement théorique). Les données mesurées indisponibles sont matérialisées par la présence de valeurs manquantes. Le contexte mesuré $mv(\mathcal{K})$ désigne les données que nous détenons en pratique ; l'opérateur $mv()$ modélise un effacement de données.

Même si nous ne faisons pas d'hypothèse sur cet opérateur, il satisfait les contraintes suivantes pour des besoins de cohérence :

– une valeur connue dans $mv(\mathcal{K})$ correspond à celle du contexte réel $\mathcal{K}$;

– une valeur présente ou absente dans $\mathcal{K}$ conservera sa valeur à l'identique ou sera manquante dans $mv(\mathcal{K})$.

La table 2 montre un exemple d'un contexte mesuré associé au contexte de la figure 1 (a), correspondant aux données de la figure 1 (b). Par exemple, les valeurs manquantes sur les items a et b dans l'objet o_8 de $mv(\mathcal{K})$ (table 2) cachent en réalité la présence de l'item b dans $\mathcal{K}$ (table 1 - (a)).

	A_1		A_2		A_3			A_4	
	a	b	c	d	e	f	g	h	i
o_1	×		×		**?**	**?**	**?**	×	
o_2	**?**	**?**	×		×			**?**	**?**
o_3	×		×		**?**	**?**	**?**	×	
o_4	×			×		×		**?**	**?**
o_5	**?**	**?**	×			×		**?**	**?**
o_6		×	**?**	**?**		×		×	
o_7	×		**?**	**?**			×	**?**	**?**
o_8	**?**	**?**		×			×	**?**	**?**

Tableau 2. *Contexte mesuré $mv(\mathcal{K})$ associé au contexte donné par la table 1 (b).*

2.2. *Modèles classiques pour les valeurs manquantes*

La communauté des statisticiens s'est toujours préoccupée des valeurs manquantes. Le modèle de LITTLE et RUBIN (Little *et al.*, 2002), reposant sur l'hypothèse d'apparition aléatoire des valeurs manquantes, est largement répandu. Nous présentons cette modélisation comme formulée dans (Little *et al.*, 2002).

– ***MCAR*** *(Missing Completely at Random)* : la présence d'une valeur manquante est complètement aléatoire. Elle affecte n'importe quel objet et n'importe quel attribut et ce phénomène ne dépend d'aucune autre valeur.

Supposons par exemple que les valeurs manquantes sur les attributs A_1 et A_2 de la table 1 (b) soient de type *MCAR*. Ces valeurs manquantes n'ont *a priori* aucune explication particulière.

– ***MAR*** *(Missing at Random)* : quand la présence d'une valeur manquante dépend de valeurs réelles particulières d'autres attributs, le modèle d'apparition est dit *aléatoire*. L'aléatoire concerne les valeurs réelles de l'attribut ayant la valeur manquante.

Les valeurs manquantes sur l'attribut A_3 de la table sont un exemple supposé de valeurs manquantes de type *MAR*. On remarque que toutes les valeurs manquantes sur l'attribut A_3 sont liées à la présence simultanée de a, c et h, qui en constitue une explication.

– ***NMAR*** *(Not Missing at Random)* : si une valeur est manquante lorsque la valeur réelle de l'attribut correspondant est particulière, alors le modèle est dit *non aléatoire*. Supposons que les valeurs manquantes sur l'attribut A_4 illustrent un exemple de valeurs manquantes de type *NMAR*. En effet, si l'item i est impossible à mesurer – ou provoque une défaillance du capteur – il en résultera une valeur manquante chaque fois que i aurait dû être observé.

La section suivante discute des limites de ces modèles et précise le positionnement de notre contribution.

2.3. *Discussion et positionnement*

Comme cela été souligné par (Shafer *et al.*, 2002), l'utilisation et l'interprétation de la modélisation de LITTLE et RUBIN prêtent à confusion. En effet, cette modélisation qualifie des relations entre des valeurs réelles théoriques et des valeurs mesurées. Dans la pratique, seules les données du monde mesuré sont disponibles et ces modèles apportent peu pour la caractérisation des valeurs manquantes. L'ambiguïté de cette modélisation réside essentiellement dans l'emploi du terme *aléatoire* pour le modèle *MAR*. Ce terme ne se justifie vraiment que dans le cas du modèle *MCAR*, où les valeurs manquantes apparaissent au hasard.

Le modèle *NMAR* pose également problème lors de la caractérisation des valeurs manquantes car il relève du contexte réel d'observation, dans lequel l'expert maîtrise l'origine des valeurs manquantes. Si nous reprenons l'exemple introduit dans la section 1 : lors d'un sondage, sachant que les personnes obèses ont tendance à cacher leur poids, nous saurons que les personnes présentant une valeur manquante sur l'attribut *Poids* cachent potentiellement un sur-poids. Mais, cette observation entre dans le cadre de l'expertise du domaine. Une solution envisageable serait de remplacer les valeurs manquantes par "Poids supérieur à 100 kg".

Dans cette configuration, la valeur manquante est dite *informative* car elle cache une certaine valeur que l'expert pourrait *a priori* reconstituer, du moins caractériser précisément et y réserver un traitement particulier. L'expert n'est cependant pas toujours présent et on ne disposera parfois même pas d'expertise. Ce modèle qui caractérise des relations entre la valeur réelle d'un attribut et sa valeur mesurée est difficile à formaliser et à identifier ; une valeur manquante *NMAR* n'est pas reconnaissable au seul examen des données mesurées, faute de connaître le contexte réel. Pour la suite de notre présentation, nous considérons donc que la gestion du type *NMAR* relève du pré-traitement des données, sous la responsabilité de l'expert.

Par ailleurs, nous avons remarqué que lorsqu'une valeur manquante apparaît, d'autres valeurs seront en conséquence indisponibles. Par exemple, lorsqu'un ganglion n'est pas examiné, une valeur manquante est notée. La dimension de l'envahissement est alors manquante, mais c'est pour une bonne raison : le ganglion n'a pas été examiné, ni mesuré.

Ce type d'enchaînement de valeurs manquantes n'est pas pris en compte dans la modélisation classique citée précédemment. Pourtant, l'analyse de ces situations caractéristiques permettrait d'affiner les méthodes de traitement.

Enfin, nous considérons la modélisation classique comme un peu restrictive, car elle explique *de la même façon* toutes les valeurs manquantes affectant un attribut donné. La caractérisation classique est ainsi relative à l'*intégralité* des objets incomplets. Nous proposons une vision plus réaliste en affinant le niveau de granularité, grâce à des caractérisations propres à *un groupe restreint d'objets*. Cette finesse d'analyse repose sur les relations entre les valeurs manquantes et les données, ou impliquent

d'autre valeurs manquantes. Les différents types de relations entre les valeurs manquantes et les données seront présentés à la section 4, qui introduit un cadre théorique pour l'identification de ces relations.

La section suivante présente la *base d'implications propres*, un outil essentiel pour le calcul d'associations permettant une caractérisation non redondante des valeurs manquantes.

3. Base d'implications propres

L'utilisation de la *base d'implications propres* comme base de règles d'association non-redondantes est au cœur de notre travail sur la caractérisation des valeurs manquantes. Ainsi, nous consacrons cette section à une présentation détaillée des résultats connus sur les bases de règles en général, et de la base d'implications propres en particulier.

Nous commençons par rappeler les définitions de motifs et de règle d'association. Nous discutons ensuite des méthodes pour calculer des associations non redondantes et terminons par la présentation de la base d'implications propres et de l'algorithme dédié à son extraction.

3.1. *Définitions*

Un *itemset* (ou *motif*) $X \subset \mathcal{I}$ est un ensemble d'items. Un objet $o \in \mathcal{O}$ supporte l'itemset X si $\forall\, i \in X$, $\mathcal{R}(o,i) = présent$, qu'on notera par $X \subseteq o$. Le support absolu de X, noté $Supp(X)$, est défini par $Supp(X) = |\{o \in \mathcal{O} | X \subseteq o\}|$.

Une *règle d'association* basée sur un itemset $Z \neq \emptyset$ est une expression entre deux itemsets X et Y de la forme $R : X \rightarrow Y$ telle que $X \subsetneq Z$ et $Y = Z \backslash X$. Les itemsets X et Y sont respectivement appelés *prémisse* et *conclusion* de la règle R. Le support de la règle R est égal à celui de l'itemset Z. La confiance de la règle est définie par la probabilité conditionnelle de présence de Y simultanément avec X : $Conf(R) = \frac{Supp(Z)}{Supp(X)}$.

L'extraction des règles d'association est généralement contrainte par un seuil minimal de support, noté $minsup$, et une confiance minimale $minconf$. Une règle d'association $X \rightarrow Y$ est dite *exacte* et nous notons $\models X \rightarrow Y$, si sa confiance vaut 1 sinon elle est dite *approximative*. Dans notre exemple (cf. table 1 - (a)), la règle $fh \rightarrow c$ est exacte.

3.2. *Calcul d'associations non redondantes*

Les méthodes à base de motifs (et donc de règles) produisent des quantités d'information très importantes, souvent très redondantes. Pour pallier cela, on sélectionne les

règles dont la prémisse est minimale. Par exemple, dans nos données (table 1 - (a)), la règle la règle $h \rightarrow c$ est non redondante par rapport à la règle $fh \rightarrow c$ car elle donne la même information à partir d'hypothèse minimale. De plus, la règle $fh \rightarrow c$ peut être déduite de $h \rightarrow c$, au support près. On évite donc beaucoup de redondance en extrayant directement des associations à prémisse minimale : c'est la notion de base ou de couverture de règles (Maier, 1980).

Les règles à prémisse minimales sont décrite dans (Taouil *et al.*, 2001) sous le nom de *base d'implications propres*. Ce n'est pas la plus petite base possible – en terme de nombre de règles – résumant les associations exactes d'un contexte. En effet, il est montré que la *base canonique* dite de Duquennes-Guigues (Guigues *et al.*, 1986) satisfait ce critère. Sans rentrer dans les détails de son calcul, cette base utilise la notion de motif pseudo-fermé, incompatible avec la notion de prémisse minimale. En conséquence, cette base ne permet pas de retrouver simplement les caractérisations utiles pour notre problème. Enfin, la condensation apportée par la base canonique peut être marginale : sur notre exemple il y a 17 implications propres quand la base canonique en contient 16.

Il nous paraît donc raisonnable de préférer l'emploi de la base d'implications propres plutôt que la base canonique.

3.3. *Calcul efficace des implications propres*

Pour présenter notre méthode de calcul des implications propres, nous utilisons le concept de *complémentaire* d'un objet. Considérant qu'un objet $o \in \mathcal{O}$ est un ensemble d'items, soit $o \subseteq \mathcal{A}$, nous notons $\overline{o} = \mathcal{A} \backslash o$. Nous utilisons également le concept de *traverse* : X traverse o s'ils partagent un item, *i.e.,* $X \cap o \neq \emptyset$.

Nous utilisons le lemme suivant :

Lemme 1 *Pour tout motif X et tout objet o :* $\quad X \subseteq o \iff X \cap \overline{o} = \emptyset$.

Corollaire 1 *La règle $X \rightarrow i$ est valide si seulement si X est une traverse des complémentaires des objets qui contiennent i.*

> Preuve : Soit X une traverse des complémentaires des objets qui contiennent i. $\forall o \mid i \in \overline{o}, X \cap \overline{o} \neq \emptyset$ donc $\forall o \mid i \notin o, \neg(X \subseteq o)$ et par contraposition : $X \subseteq o \implies i \in o$ donc la règle $X \rightarrow i$ est valide. □

Pour extraire les implications propres $X \rightarrow i$, il nous faut donc extraire les traverses **minimales** des complémentaires des objets contenant i. Le calcul des traverses minimales est bien identifié dans la communauté algorithmique comme candidat à la séparation des classes P et NP et a fait l'objet de nombreuses études et propositions d'algorithmes (Hagen, 2008). Son lien avec l'extraction de motifs fréquents est bien connu puisque la génération de candidats utilise également les traverses minimales

des complémentaires pour calculer la bordure négative (Mannila *et al.*, 1997; Gunopulos *et al.*, 1997). On pourrait donc résoudre notre problème en considérant que la prémisse X d'une implication propre pour un item i appartient à la bordure négative associée à la contrainte anti-monotone $\models X \rightarrow i$.

Pour le calcul des traverses minimales, nous disposons d'un vaste choix d'algorithmes fournis par la communauté : en largeur, en profondeur, avec ou sans consommation de mémoire. Disposant de bases très denses, nous avons finalement privilégié l'algorithme de Stavropoulos et Kavvadias (Kavvadias *et al.*, 1999), qui s'exécute en profondeur et fournit les traverses incrémentalement, donc ne requiert pas de mémoire.

La section suivante présente notre contribution portant sur les modèles de valeurs manquantes (Ben Othman *et al.*, 2009b). Dans le travail présent, nous montrons essentiellement comment ces modèles peuvent être caractérisés de façon non redondante à l'aide de la base d'implications propres.

4. Nouvelle typologie des valeurs manquantes

Notre intuition est que des régularités de valeurs peuvent expliquer la présence de valeurs manquantes. Ces explications concernent des valeurs particulières d'attributs ou d'autres valeurs manquantes. Ainsi, nous proposons une nouvelle typologie des valeurs manquantes :

– **Valeur manquante directe** : une valeur manquante est dite directe quand il existe une relation entre cette valeur manquante et *des données mesurées*.

– **Valeur manquante indirecte** : une valeur manquante est dite indirecte quand il existe une relation entre cette valeur manquante et *d'autres valeurs manquantes*.

– **Valeur manquante hybride** : une valeur manquante est dite hybride quand il existe à la fois une relation entre cette valeur manquante, *des données mesurées et d'autres valeurs manquantes*.

– **Valeur manquante *aléatoire*** : une valeur manquante est dite *aléatoire* quand il n'existe *aucune relation* entre cette valeur manquante avec les données mesurées, ni avec d'autres valeurs manquantes.

Nous définissons maintenant formellement cette typologie des valeurs manquantes et nous présentons la structure des règles d'association permettant de caractériser les types de valeurs manquantes de cette typologie.

4.1. *Règles de caractérisation des valeurs manquantes*

La définition des règles de caractérisation des valeurs manquantes requiert des précisions sur la capacité de *mesure* d'un itemset dans un contexte $mv(\mathcal{K})$, selon les cas suivants :

Définition 3 *Soient $X \subseteq \mathcal{I}$ un itemset et $o \in \mathcal{O}$ un objet.*

En o, X est dit :	*si et*	*seulement si*
présent	$\forall\, x \in X$	$mv(\mathcal{R})(o,x) = présent$
manquant	$\forall\, x \in X$	$mv(\mathcal{R})(o,x) = manquant$
partiellement présent	$\forall\, x \in X$	$mv(\mathcal{R})(o,x) \neq absent \wedge$ $\exists\, x_1 \in X, mv(\mathcal{R})(o,x_1) = présent \wedge$ $\exists\, x_2 \in X, mv(\mathcal{R})(o,x_2) = manquant$
absent	$\exists x \in X$	$mv(\mathcal{R})(o,x) = absent$

On notera respectivement $Présent(X,o)$, $Manquant(X,o)$ $PartPrésent(X,o)$ et $Absent(X,o)$.

Exemple 1 *Dans le contexte $mv(\mathcal{K})$ illustré par la table 2, nous avons $Présent(adf,o_4)$, $Manquant(ah,o_8)$, $PartPrésent(bdg,o_8)$ et $Absent(bd,o_1)$.*

Remarquons qu'une partition est ainsi définie : un itemset est exclusivement présent, absent, manquant ou partiellement présent.

Nous pensons que les régularités qui permettent de caractériser les différents types de valeurs manquantes peuvent être décelées grâce à des règles d'association. Dans la pratique, leur extraction est paramétrée par un support minimal $minsup$, sur le choix duquel nous reviendrons.

Nous proposons maintenant une formalisation de cette nouvelle typologie des valeurs manquantes :

Définition 4 *Soit $\mathcal{T} \subseteq \mathcal{O}$ (avec $|\mathcal{T}| \geq minsup$). Une valeur manquante sur $i \in \mathcal{I}$*

est dite	*en $\mathcal{T}$*	*si et*	*seulement si*
directe	$\exists X \subseteq \mathcal{I} \setminus \{i\}$	$\forall\, o \in \mathcal{T}$	$Présent(X,o) \Rightarrow Manquant(i,o)$
indirecte	$\exists X \subseteq \mathcal{I} \setminus \{i\}$	$\forall\, o \in \mathcal{T}$	$Manquant(X,o) \Rightarrow Manquant(i,o)$
hybride	$\exists X \subseteq \mathcal{I} \setminus \{i\}$	$\forall\, o \in \mathcal{T}$	$PartPrésent(X,o) \Rightarrow Manquant(i,o)$
aléatoire	$\forall X \subseteq \mathcal{I} \setminus \{i\}$	$\exists\, o \in \mathcal{T}$	$Manquant(i,o) \wedge Absent(X,o)$.

Remarquons là encore que les quatre types sont exclusifs. De fait, le type aléatoire peut être considéré par défaut, lorsqu'aucune caractérisation n'est découverte, du moins au seuil de support considéré.

4.2. *Relation avec la typologie classique*

Une valeur manquante est de type *direct* lorsque l'absence de valeur s'explique par des valeurs observées sur d'autres attributs. On retrouve dans ce type le modèle *MAR*.

Elle est de type *indirect* lorsqu'elle s'explique par la présence d'autres valeurs manquantes sur d'autres attributs. Parfois, une valeur manquante i s'explique par la présence simultanée de valeurs observées et d'autres valeurs manquantes. Dans ce cas, elle est de type *hybride*. Ces types n'ont pas d'équivalent dans la typologie classique.

Finalement, lorsqu'il n'existe aucune explication à la présence d'une valeur manquante, elle est dite *aléatoire*. Dans ce dernier cas, on retrouve le modèle *MCAR*.

De façon attendue, notons que le modèle *NMAR* ne relève pas de notre typologie. En effet, sa définition est relative aux relations entre les valeurs réelles et mesurées et notre hypothèse de travail est de disposer de modèles détectables au seul examen des données disponibles (mesurées). Il s'agit d'un avantage majeur de notre typologie.

Les caractérisations proposées par la définition 4 peuvent être calculées à l'aide de règles d'association. En effet, **si les valeurs manquantes sont codées comme des items, il suffit de calculer les règles d'association qui concluent sur ces items**.

Cependant, comme nous l'avons détaillé à la section précédente, ce calcul peut fournir des associations très redondantes et il est intéressant d'obtenir une représentation condensée de ces règles. La section suivante présente notre contribution concernant l'utilisation la base d'implications propres pour mettre en évidence les modèles de valeurs manquantes.

5. Caractérisation des valeurs manquantes à l'aide d'implications propres

Nous expliquons maintenant comment utiliser la base d'implications propres (*cf.* section 3) pour nos besoins en caractérisation des valeurs manquantes. Cette base est essentielle pour obtenir des associations non redondantes (à prémisse minimale) et nous montrons l'intérêt de cette propriété pour la caractérisation.

5.1. *Caractérisation non redondante des valeurs manquantes*

Pour montrer l'intérêt de l'utilisation de la base d'implications propres pour la caractérisation des valeurs manquantes, la table 3 compare les règles classiques (a) et les implications propres (b) concluant sur des valeurs manquantes de l'attribut A_4. Pour cela, nous employons la notation suivante :

Notation 1 *Une valeur manquante sur l'attribut A_i est notée par $MV(A_i)$.*

Nous remarquons que la règle R'_4 n'a pas été générée par la base d'implications propres qui contient déjà $R"_1$ de prémisse minimale. Grâce aux propriétés de minimalité des prémisses, la base d'implications propres est un atout pour caractériser le type des valeurs manquantes de façon non redondante. En utilisant les règles classiques,

	Règle	Support
R'_1	$MV(A_1) \rightarrow MV(A_4)$	3
R'_2	$d \rightarrow MV(A_4)$	2
R'_3	$g \rightarrow MV(A_4)$	2
R'_4	$c \wedge MV(A_1) \rightarrow MV(A_4)$	2

	Règle	Support
$R"_1$	$MV(A_1) \rightarrow MV(A_4)$	3
$R"_2$	$d \rightarrow MV(A_4)$	2
$R"_3$	$g \rightarrow MV(A_4)$	2

Tableau 3. *Règles concluant sur $MV(A_4)$. (a) : les règles classiques. (b) : les implications propres.*

les valeurs manquantes sur l'attribut A_4 seraient caractérisées comme indirectes et hybrides par les règles R'_1 et R'_4. En revanche, avec la base d'implications propres, elles seront simplement caractérisées comme indirectes.

Pour la suite de cet article, nous caractérisons les valeurs manquantes grâce aux implications propres qui modélisent la définition 4. Il s'agit certes d'une restriction par rapport aux définitions des types de valeurs manquantes, qui ne font pas intervenir de contrainte de minimalité sur les règles de caractérisation. Cependant, nous ne souhaitons pas obtenir l'intégralité des caractérisation possibles, juste une représentation condensée non redondante. Dans cet esprit, l'utilisation de la base d'implications propres possède pour la caractérisation des valeurs manquantes les avantages suivants :

1) l'utilisation d'une représentation condensée réduit drastiquement le nombre de règles générées. Il s'agit là d'un point essentiel pour obtenir une caractérisation concise, entre autres pour interagir avec l'expert responsable des données.

2) ces règles permettent de caractériser le type des valeurs manquantes en minimisant les caractérisations multiples grâce aux propriétés de minimalité des prémisses. On parle de caractérisation multiple ou de conflit lorsqu'une même valeur manquante peut satisfaire deux types différents, selon que les objets diffèrent ou au sein d'un même objet.

Bien que les conflits sur la caractérisation d'une valeur manquante soient minimisés, il est cependant possible d'obtenir plusieurs types simultanément. La section suivante examine plus particulièrement le type hybride.

5.2. *Caractérisation du type hybride*

Selon la définition 4, une valeur manquante hybride s'explique par la présence simultanée de valeurs observées et de valeurs manquantes. Comment alors considérer une valeur manquante à la fois directe et indirecte ?

La base d'implication propres concentre notre attention sur les caractérisations minimales. Sur notre exemple, la valeur manquante affectant l'attribut A_4 de l'objet o_8 est caractérisée par deux règles : $MV(A_1) \rightarrow MV(A_4)$ et $d \rightarrow MV(A_4)$. La première règle caractérise un type indirect, la deuxième un type direct. Cependant, la

règle $d \wedge MV(A_1) \rightarrow MV(A_4)$ est valide et caractérise un type hybride, mais ce n'est pas une implication propre car elle n'est pas à prémisse minimale : nous ne retenons pas ici la caractérisation hybride.

Dans la pratique, il est donc inutile de différencier les types simultanés direct et indirect. Lors des expériences sur des données réelles comme celles reportées à la section suivante, nous qualifions d'hybride les valeurs manquantes à la fois directes et indirectes.

La table 4 indique les implications propres concluant sur des valeurs manquantes. Ces implications permettent de déduire la caractérisation du type de valeur manquante sur les objets qui supportent la règle. Cette caractérisation est présentée à la table 5.

	Règle	Objets supportant la règle
R_1	$a \wedge c \rightarrow MV(A_3)$	$\{o_1, o_3\}$
R_2	$MV(A_1) \rightarrow MV(A_4)$	$\{o_2, o_5, o_8\}$
R_3	$a \wedge h \rightarrow MV(A_3)$	$\{o_1, o_3\}$
R_4	$c \wedge MV(A_4) \rightarrow MV(A_1)$	$\{o_2, o_5\}$
R_5	$c \wedge h \rightarrow MV(A_3)$	$\{o_1, o_3\}$
R_6	$d \rightarrow MV(A_4)$	$\{o_4, o_8\}$
R_7	$g \rightarrow MV(A_4)$	$\{o_7, o_8\}$

Tableau 4. *Implications propres concluant sur une valeur manquante ;* $minsup = 2$.

	A_1	A_2	A_3	A_4
o_1	-		{direct}	-
o_2	{hybride}	-	-	{indirect}
o_3	-	-	{direct}	-
o_4	-	-	-	{direct}
o_5	{hybride}	-	-	{indirect}
o_6	-	{aléatoire}	-	-
o_7	-	{aléatoire}	-	{direct}
o_8	{aléatoire}	-	-	{hybride}

Tableau 5. *Typologie des valeurs manquantes.*

5.3. *Comparaison des typologies de valeurs manquantes*

Pour finir, nous résumons les principales caractéristiques (table 6) de notre nouvelle typologie en la comparant à la typologie classique de LITTLE et RUBIN (Little *et al.*, 2002).

LITTLE et RUBIN se basent sur des données disponibles mais également sur les données non disponibles, car le modèle NMAR utilise la valeur réelle d'un attribut pour qualifier le modèle d'apparition de ses valeurs manquantes. Notre proposition

détecte les modèles présents dans les données disponibles, sous forme d'implications propres, pour construire notre typologie.

Nous qualifions donc le cadre de travail de LITTLE et RUBIN de purement théorique, car il est impossible de mettre en pratique un modèle de valeurs manquantes concernant des données indisponibles.

De fait, la caractérisation de Little et Rubin est globale car toutes les valeurs manquantes d'un même attribut sont caractérisées d'une façon unique. Notre typologie propose une vision locale, éventuellement réduite à un petit groupe d'objets.

	Typologie classique	**Nouvelle typologie**
Données utilisées	disponibles + expertise	disponibles
Cadre de travail	théorique	opérationnel
Focus	global	local
Types	MCAR	aléatoire
	MAR	direct
	NMAR	→ relève de l'expertise
	-	indirect
	-	hybride

Tableau 6. *Caractéristiques des typologies des valeurs manquantes.*

Pour finir, nous tentons d'établir une correspondance entre les types de Little et Rubin et les nôtres. Le type NMAR ne figure pas dans notre typologie puisqu'il se base sur les données *a priori* indisponibles ; leur traitement relève de l'expertise. En revanche, les types indirect et hybride ne figurent pas dans celle de Little et Rubin. Il y a néanmoins une correspondance des types MCAR/aléatoire et MAR/direct, même si lors d'une utilisation pratique ces caractérisations diffèrent notablement par leur focus : celle de LITTLE et RUBIN est globale, tandis que notre approche est locale. La correspondance entre les types est localisée sur un groupe d'objets.

5.4. *Discussion*

La définition 4 des nouveaux types dépend du paramètre de support. Ainsi, lorsque le support minimum d'extraction varie, on obtient différentes caractérisations, potentiellement conflictuelles.

Le choix de ce paramètre a donc une incidence profonde sur la typologie obtenue. On peut également discuter de la longueur d'une caractérisation (le nombre d'items de la prémisse de la règle). En effet, est-il raisonnable de prendre en compte une caractérisation impliquant par exemple 12 item ?

D'autre part, l'utilisation de la base d'implication propres nous limite à des caractérisations par des règles valides, de confiance 1. Pour obtenir des règles de plus petite confiance, on peut utiliser comme prémisse les sous-ensembles des prémisses des règles valides, mais cette possibilité introduit un nouveau paramètre.

Finalement, le calcul de notre typologie est paramétrable par un support minimal, une longueur de caractérisation et une confiance. Une discussion plus précise sur l'impact de leur choix serait à mener lors de la validation d'une méthode de complétion utilisant cette typologie. Dans la section suivante qui relate nos expériences sur des données médicales, ces paramètres ont été choisis en accord avec les attentes des experts sur la finesse de caractérisation.

6. Expérimentations

Cette section rapporte les résultats de nos expériences sur la caractérisation des différents types de valeurs manquantes dans des données réelles. Le but de ces expériences est de montrer la pertinence des caractérisations introduites. Les bases considérées sont deux bases de données médicales : une relative à la maladie de Hodgkin[1] et l'autre portant sur les méningites infantiles. La raison de ces choix est double : d'une part, les valeurs manquantes de ces deux bases sont réelles (*i.e.,* aucune simulation n'a été menée pour introduire artificiellement des valeurs manquantes) et, d'autre part, nous disposons d'une expertise médicale pour ces données (le Dr M. HENRY-AMAR, responsable de l'unité clinique du centre de lutte contre le cancer FRANÇOIS BACLESSE à Caen pour les données sur les Hodgkin et le Dr P. FRANÇOIS du CHU de Grenoble pour celles sur les méningites infantiles).

6.1. *Données sur la maladie de Hodgkin*

La base HODGKIN regroupe 3904 patients concernant trois *essais thérapeutiques* (H7, H8 et H9) réalisés pendant des périodes temporelles successives. Chaque patient est décrit par 36 attributs, dont 29 présentent des valeurs manquantes avec un taux variant entre 2% et 88%. Outre les attributs concernant certaines caractéristiques sanguines ou histologiques, les données indiquent si les ganglions lymphatiques cervicaux, auxiliaires, hile et médiastin sont envahis par le cancer et la dimension de cet envahissement le cas échéant.

L'extraction des règles a été effectuée avec un support absolu minimum de 700. Les 15 règles découvertes sont reportées à la table 8, ce faible nombre est caractéristique d'une réduction drastique du nombre de règles à calculer, par rapport aux règles classiques (*cf.* table 7), qui évite leur redondance et facilite ainsi fortement leur étude.

Par exemple, la règle R_4 (table 8) signifie que tous les objets contenant l'item $plaq <= 600$ et une valeur manquante sur l'attribut chd (ganglion cervical haut droit)

1. La maladie de Hodgkin est un cancer des ganglions du système lymphatique

		Nombre de règles	Nombre de règles concluant sur une valeur manquante
Base de HODGKIN	Implications propres	**49**	**15**
	Règles classiques	2 923 070	2 681 045
Base de la MENINGITE	Base d'implications propres	**17 508**	**152**
	Règles classiques	182 317	7659

Tableau 7. *Comparaison entre le nombre d'implications propres et de règles classiques.*

contiennent également une valeur manquante sur l'attribut chg. C'est une caractérisation de valeur manquante de type *hybride*.

	prémisse	conclusion	support absolu
R_1	essai H7	MV(chd)	816
R_2	essai H7	MV(chg)	816
R_3	MV(axddim)$\wedge$ MV(chd)	MV(chg)	811
R_4	plaq<=600 $\wedge$ MV(chd)	MV(chg)	778
R_5	chd non envahi	MV(chddim)	2449
R_6	chg non envahi	MV(chgdim)	2407
R_7	cbd non envahi	MV(cbddim)	1969
R_8	cbg non envahi	MV(cbgdim)	1690
R_9	axd non envahi	MV(axddim)	3295
R_{10}	axg non envahi	MV(axgdim)	3185
R_{11}	MV(chd)	MV(chddim)	908
R_{12}	MV(chg)	MV(chgdim)	910
R_{13}	med non envahi $\wedge$ vs <=30	MV(mtr)	920
R_{14}	med non envahi $\wedge$ rechute= non	MV(mtr)	1042
R_{15}	med non envahi $\wedge$ MV(cbgdim)	MV(mtr)	717

Tableau 8. *Exemples de règles exactes extraites à partir de la base* HODGKIN *pour une valeur de minsup=700.* $MV(attribut)$ *indique que* $attribut$ *est manquant.*

Les règles R_1 et R_2 concluent sur un envahissement du ganglion cervical haut gauche *chg* ou droit *chd* manquant. Elles contiennent en prémisse *l'essai H7*. Ces valeurs manquantes sont de type *direct*. Il s'avère que pour l'essai thérapeutique $H7$, le premier chronologiquement, les ganglions cervicaux hauts et bas n'étaient pas différenciés et cette valeur n'existe pas pour les données correspondantes. La valeur *H7* pour le numéro d'essai explique donc la présence de valeurs manquantes sur ces ganglions. Il s'agit là d'un problème classique de fusion de données. Notre méthode permet ainsi de mettre en évidence les problèmes causés par la fusion des données, puisqu'il s'agit de détecter de potentielles anomalies dans les données. Par conséquent, nous arrivons à mieux connaître et à mieux contrôler la qualité des données. Nous avons également constaté que les valeurs manquantes sur l'attribut *chg* ont été ca-

ractérisées par d'autres règles donnant lieu à des valeurs manquantes de type *indirect* (R_3) et *hybride* (R_4). C'est un exemple de caractérisation multiple.

Lorsque qu'un ganglion n'est pas envahi, sa dimension n'est pas mesurée par les médecins, induisant des valeurs manquantes sur l'attribut *dimension*. Cette connaissance sur les valeurs manquantes est retrouvée (règles R_5 à R_{10}). En effet, les prémisses de toutes ces règles présentent des ganglions non envahis, ces valeurs manquantes sont donc *directes* et elles révèlent une relation avec l'envahissement du ganglion. En général, les principales méthodes de traitement des données manquantes chercheront à compléter ces valeurs soit par la moyenne, soit par une valeur aléatoirement choisie ou soit en utilisant l'ensemble des valeurs possibles. En réalité, l'absence de ces dimensions représente des valeurs non applicables, puisque les ganglions ne sont pas envahis. Un des intérêts de notre caractérisation est de mettre en évidence cette relation et de suggérer qu'il ne faut pas chercher à compléter "aveuglement" ces valeurs manquantes : une solution possible est d'ajouter une valeur spéciale sur l'attribut *dimension* indiquant que le ganglion n'est pas envahi.

Nous avons également mis en évidence des valeurs manquantes de type *indirect* sur les dimensions $chgdim$ et $chddim$ des ganglions cervicaux hauts gauche et droite (les règles R_{11} et R_{12}). Des valeurs manquantes sur ces dimensions s'expliquent par des valeurs manquantes sur l'attribut indiquant si le ganglion est envahi ou pas : quand on ne sait pas si un ganglion est envahi – il n'a pas été examiné ou le résultat de cet examen n'a pas été transmis – une valeur manquante affectera nécessairement sa dimension.

Enfin, les valeurs manquantes sur l'attribut *mtr* (*rapport dimension ganglion médiastin / thorax*) ont été caractérisées par trois règles (R_{13} à R_{15}). Les deux premières règles mettent en évidence des valeurs de type *direct*. Par contre, la règle R_{15} caractérise des valeurs manquantes de type *hybride*.

La table 9 présente la répartition des caractérisations pour chaque attribut manquant, en considérant comme hybride la combinaison direct-indirect. Notons que d'après cette table, il existe le plus souvent une explication à la présence des valeurs manquantes, *i.e.,* le pourcentage des valeurs manquantes de type *aléatoire* est relativement faible.

Ces résultats confirment, à partir de données réelles, que les valeurs manquantes d'un même attribut ne s'expliquent pas nécessairement de la même façon et par conséquent, ne suivent pas un même type global. C'est le cas des attributs *chg*, *chddim*, *chgdim* et *mtr*. Remarquons que les valeurs manquantes *chddim* et *chgdim* sont caractérisées par deux types de règles distincts. Dans le premier cas, les ganglions ne sont pas envahis (R_5 et R_6) et le type est *direct*, tandis que dans le deuxième cas aucune connaissance ne permet de conclure quant à leur envahissement (R_{11} et R_{12}), *i.e.,* le type est *indirect*. Comme il est impossible d'avoir un même patient vérifiant les deux cas, il s'agit de prémisses mutuellement exclusives. Une valeur manquante sur ces attributs sera *directe* ou *indirecte*, mais pas les deux à la fois.

attribut	valeurs manquantes	directes	indirectes	hybrides	aléatoires
chd	908	90%	0	0	10%
chg	910	10,7%	10%	79%	0,3%
chddim	3435	71%	3%	24%	2%
chgdim	3398	71%	3%	24%	2%
cbddim	2274	87%	0	0	13%
cbgdim	2027	83%	0	0	17%
axddim	3444	96%	0	0	4%
axgdim	3360	95%	0	0	5%
mtr	1512	32%	0	47%	21%

Tableau 9. *Caractérisation des valeurs manquantes dans la base* HODGKIN *selon leurs types.*

Il n'en va pas de même pour les attributs *chg* et *mtr*. En examinant attentivement les règles concluant sur *chg* manquant (R_2 à R_4) et *mtr* manquant (R_{13} à R_{15}), nous remarquons que les prémisses des règles ne sont pas mutuellement exclusives. Une valeur manquante peut donc être expliquée de façon multiple. Nous pensons que ce point constitue un des intérêts de notre typologie : notre caractérisation est réaliste et utile en pratique pour appréhender les multiples causes pouvant expliquer la présence d'une valeur manquante.

D'autres attributs possèdent des valeurs manquantes, mais en faible proportion (entre 2% et 9%). De façon évidente, nous n'avons pas trouvé de règles les caractérisant sous nos conditions expérimentales (le support absolu de 700 objets correspond à $minsup = 18\%$ et est donc supérieur à 9%). Clairement, le choix du support minimum influe profondément sur la caractérisation des valeurs manquantes. Une faible valeur de ce seuil pourrait fournir une quantité importante de caractérisations rendant difficile leur analyse. C'est pourquoi nous avons choisi un seuil relativement important, dans le but de privilégier des tendances générales qui peuvent amener à des préconisations opérationnelles. Une étude plus fine de cette typologie figure dans nos perspectives, en particulier avec un support minimal plus faible et une relaxation de la confiance des règles.

6.2. *Données sur la méningite*

La méningite est une infection des méninges – les enveloppes de la moelle épinière et du cerveau – dans lesquelles circule le liquide céphalorachidien. Une méningite est due à un virus ou à une bactérie. La base MENINGITE que nous avons étudiée décrit 329 enfants atteints de méningite, c'est-à-dire tous les enfants (sur une période de 4 ans) ayant été admis aux urgences pédiatriques du CHU de Grenoble suite à une

méningite. Chaque enfant est décrit par 23 attributs, dont 9 présentent des valeurs manquantes d'un taux variant entre 1% et 23%.

Avec $minsup = 10$ (3%), nous avons caractérisé les valeurs manquantes pour les attributs *tonus*, *polysang* et *polyns*. Une partie des règles produites est donnée dans la table 10.

	prémisse	conclusion	support
R_1	MV(polysang)	MV(polyns)	71
R_2	gram=0 $\wedge$ MV(polyns)	MV(polysang)	61
R_3	leuco<=11.5 $\wedge$ MV(polyns)	MV(polysang)	20
R_4	age<=2.75 $\wedge$ aerien<=0 $\wedge$ gluc<=2.66 $\wedge$ polyns<=38	MV(tonus)	11
R_5	age<=2.75 $\wedge$ aerien<=0 $\wedge$ poly_LCR<=73	MV(tonus)	14
R_6	age<=2.75 $\wedge$ comport<=2 $\wedge$ aerien<=0	MV(tonus)	10
R_7	age<=2.75 $\wedge$ comport<=2 $\wedge$ cytol<=12280$\wedge$ gram=0	MV(tonus)	10
R_8	age<=2.75 $\wedge$ aerien<=0 $\wedge$ prot<=8 MV(vs)	MV(tonus)	10
R_9	age<=2.75 $\wedge$ sneuro<=0 $\wedge$ gram=0 $\wedge$ polysang<=53 MV(vs)	MV(tonus)	10

Tableau 10. *Exemples de règles exactes extraites à partir de la base* MENINGITE *pour une valeur de* minsup=*10.*

La règle R_1 conclut sur *polyns* manquant. Elle met en évidence des valeurs manquantes de type *indirect* : toutes les valeurs manquantes sur l'attribut *polyns* s'expliquent par la présence de valeurs manquantes sur l'attribut *polysang*. De façon symétrique, les valeurs manquantes sur l'attribut *polysang* s'expliquent par des valeurs manquantes sur l'attribut *polyns* (règles R_2 et R_3), sauf que ces deux dernières règles sont de type *hybride*. Ces valeurs manquantes affectent plutôt des méningites d'origine virale (examen bactériologique *direct* négatif (*gram = 0*), faible taux de leucocytes). On met ici en évidence que le recueil des données est effectué avec moins de soin (*polysang* manquant) pour les cas bénins de méningites, c'est-à-dire ceux d'origine virale.

La caractérisation de l'attribut *tonus* est plus complexe, plusieurs règles ont été produites. Elles indiquent des valeurs manquantes de type *direct* (R_4, R_5, R_6 et R_7) et *hybride* (R_8 et R_9) et reposent sur des données de natures différentes : clinique (*âge*, *aérien*), du liquide cephalo-rachidien (*pourcentage de polynucléaires*, *protéinorachie*), de la biologie du sang (*polyns*, *polysang*).

La table 11 donne la répartition des caractérisations pour chaque attribut entaché de valeurs manquantes. Remarquons qu'aucune règle de caractérisation n'a été produite sur l'attribut *vs* et ces valeurs manquantes sont classées comme aléatoires.

Similairement à ce que nous avons constaté sur la base HODGKIN, le support minimal influe sur la caractérisation : les valeurs manquantes de faible proportion (attributs *dfievre*, *sneuro*, *prot*, *gluc* et *leuco*) ne sont pas caractérisées. Nous pensons qu'il serait artificiel de chercher des relations vérifiées par moins de 10 patients, celles-ci ne seraient pas fiables.

attribut	valeurs manquantes	directes	indirectes	hybrides	aléatoires
tonus	88	57%	0	39%	4%
polysang	71	0	0	100%	0
polyns	72	0	98%	0	2%
vs	76	0	0	0	100%
dfievre	1	0	0	0	100%
sneuro	1	0	0	0	100%
prot	1	0	0	0	100%
gluc	1	0	0	0	100%
leuco	6	0	0	0	100%

Tableau 11. *Caractérisation des valeurs manquantes dans la base* MENINGITE.

7. Conclusions et Perspectives

Dans cet article, nous avons montré que les valeurs manquantes ne doivent pas être considérées comme aléatoires ni être caractérisées par un seul type valable pour toutes les données. Nous avons explicité différents modèles des valeurs manquantes et nous avons proposé une nouvelle typologie reposant uniquement sur les données connues, qui différencie les origines de valeurs manquantes selon les groupes d'objets où elles apparaissent. Nous avons alors montré comment il est possible de caractériser ces différents types, correspondant aux modèles d'apparition des valeurs manquantes, en utilisant une base de règles non redondantes constituée des implications propres.

Des expériences sur des bases de données médicales réelles montrent, d'un point de vue pratique, que cette méthode permet de mieux comprendre les causes des valeurs manquantes et qu'elle contribue ainsi à améliorer la qualité des données, en détectant par exemple des incohérences dues à la fusion de données ou des explications de l'origine de valeurs manquantes suggérant un recueil plus fin des données. Nous pensons que des méthodes de complétion plus efficaces peuvent être développées, puisque celles-ci peuvent tirer bénéfice de ces modèles d'apparition des valeurs manquantes. Nous avons récemment fait une proposition en ce sens (Ben Othman *et al.*, 2009a).

Le travail en cours consiste à intégrer la typologie proposée dans un processus de complétion. Une méthode de complétion se doit de prendre en considération les causes et les origines des valeurs manquantes. Ces informations supplémentaires permettent d'adapter à chaque type la stratégie de complétion. Cette typologie nous permet par exemple de distinguer les cas où il est pertinent de compléter une valeur manquante de façon automatique des cas où il est nécessaire de consulter le propriétaire des données et avec quel objectif. Nous avons ainsi distingué le cas des valeurs manquantes aléatoires qui sont caractérisées par l'absence de relations dans les données sur leur contexte d'apparition. Dans ce cas précis, nous proposons une complétion à l'aide d'un modèle calculé sur les données, par exemple à l'aide de règles d'association (Ben Othman *et al.*, 2008). En revanche, quand il s'agit de valeur manquante directe,

indirecte ou hybride, nous détenons une certaine information concernant le contexte d'apparition. Nous proposons dans ce cas de fournir une valeur de remplacement qui symbolise les conditions d'apparition de cette valeur manquante. Nous avons identifié des situations réelles où la complétion par une valeur du domaine de définition n'a pas de sens dans le cas de valeurs manquantes directe, indirecte ou hybride. Notre but est d'aller vers une complétion des valeurs manquantes qui prenne différentes formes, selon les différents types, avec l'objectif de valoriser ces types lors de la phase de complétion.

Remerciements

Les auteurs remercient le Centre Anti-Cancéreux François Baclesse de Caen et le Docteur Michel HENRY-AMAR pour la mise à disposition des données sur le lymphome de Hodgkin et le CHU de Grenoble et le Docteur Patrice FRANÇOIS pour celles sur les méningites infantiles. Ce travail est partiellement financé par le projet Franco-Tunisien *CMCU 05G1412*.

8. Bibliographie

Ben Othman L., Ben Yahia S., « Yet another approach for completing missing values », *Post-Proceedings of the 4th International Conference on Concept Lattices and their Applications (CLA 2006), LNAI Vol. 4923*, Springer Verlag, p. 154-168, 2008.

Ben Othman L., Ben Yahia S., « GBAR$_{MVC}$: Generic Basis of Association Rules based approach for Missing Values Completion », *International Journal of Computing and Information Sciences (IJCIS)*, 2010.

Ben Othman L., Rioult F., Ben Yahia S., Crémilleux B., « Completing non-random missing values. », *Proceedings of the 4th International Conference on Intelligent Systems and Knowledge Engineering (ISKE 2009)*, Hasselt, Belgium, 2009a.

Ben Othman L., Rioult F., Ben Yahia S., Crémilleux B., « Missing Values : Proposition of a Typology and Characterization with an Association Rule-Based Model », *Proceedings of 11th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2009), Springer-Verlag, LNCS 569*, p. 441-452, 2009b.

Brock G., Shaffer J., Blakesley R., Lotz M., Tseng G., « Which missing value imputation method to use in expression profiles : a comparative study and two selection schemes », *BMC Bioinformatics*, vol. 9, n° 1, p. 12, 2008.

Calders T., Goethals B., Mampaey M., « Mining itemsets in the presence of missing values », *Proceedings of the ACM Symposium on Applied Computing*, ACM, Seoul, Korea, p. 404-408, 2007.

Delavallade T., Dang T., « Using Entropy to Impute Missing Data in a Classification Task », *Proceedings of the IEEE International Conference of Fuzzy Systems (FUZZ-IEEE'07)*, London, UK, p. 577-582, 23–26 July, 2007.

Dempster A., Laird N., Rubin D., « Maximum likelihood from incomplete data via the EM algorithm », *Journal of the Royal Statistical Society*, vol. 39, n° 1, p. 1-38, 1977.

Fiot C., Laurent A., Teisseire M., « SPoID : Extraction de motifs séquentiels pour les bases de données incomplétes », *Actes des 7èmes journés d'Extraction et Gestion des Connaissances (EGC'07)*, Namur, Belgium, p. 715-726, Janvier, 2007.

Ghahramani Z., Jordan M. I., « Supervised learning from incomplete data via an EM approach », *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann, p. 120-127, 1994.

Guigues J., Duquennes V., « Familles minimales d'implications informatives résultant d'un tableau de données binaires », *Mathématiques et Sciences Humaines*, vol. 95, p. 5-18, 1986.

Gunopulos D., Mannila H., Khardon R., Toivonen H., « Data mining, hypergraph transversals, and machine learning », *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'97), Tucson, USA*, 1997.

Hagen M., Algorithmic and computational complexity issues of MONET, PhD thesis, Firedrich-Schiller-Universität Jena, Germany, 2008.

Jami S., Jen T., Laurent D., Loizou G., Sy O., « Extraction de règles d'association pour la prédiction de valeurs manquantes », *ARIMA journal, Numéro spécial CARI'04*, vol. , p. 103-124, Novembre, 2005.

Kavvadias D. J., Stavropoulos E. C., A New Algorithm for the Transversal Hypergraph Problem, Technical Report n° Technical Report CTI TR990303, Computer Technology Institute, Patras, Greece, 1999.

Little R., Rubin D., *Statistical Analysis with Missing Data, second edition*, John Wiley, New York, 2002.

Maier D., « Minimum covers in the relational database model », *JACM*, vol. 27, p. 664-674, 1980.

Mannila H., Toivonen H., « Levelwise search and borders of theories in knowledge discovery », *Data Mining and Knowledge Discovery*, vol. 1, n° 3, p. 241-258, 1997.

Nelwamondo F., Marwala T., « Rough Set Theory for the Treatment of Incompltete Data », *Proceedings of the IEEE International Conference of Fuzzy Systems (FUZZ-IEEE'07)*, London, UK, p. 338-343, 23–26 july, 2007.

Pearson R. K., « The problem of disguised missing data », *SIGKDD Explorations*, vol. 8, n° 1, p. 83-92, 2006.

Ragel A., Crémilleux B., « Treatment of Missing Values for Association Rules », *Proceedings of the International Pacific-Asia Conference on Knowlege Discovery and Data Mining (PAKDD'98), LNCS Vol. 1394*, Springer, Melbourne, Australia, p. 258-270, April 15-17, 1998.

Rioult F., Crémilleux B., « Extraction de propriétés correctes dans les bases de données incomplètes », *Actes de la Conférence Francophone sur l'Apprentissage Automatique (CAp'06)*, Trégastel, France, p. 347-362, 2006.

Ryan C., Greene D., Cagney G., Cunningham P., « Missing Value Imputation for Epistatic MAPs », *BMC Bioinformatics*, 2010.

Shafer J. L., Graham J. W., « Mising data : Our view of the state of the art », *Psychological Methods*, vol. 7, n° 2, p. 147-177, 2002.

Shen J. J., Chang C. C., Li Y. C., « Combined Association Rules for Dealing with Missing Values », *Journal of Information Science*, vol. 33, n° 4, p. 468-480, 2007.

Taouil R., Bastide Y., « Computing Proper Implications », *Proceedings of the 9th International Conference on Conceptual Structures (ICCS'2001)*, Stanford, CA, p. 49-61, 2001.

Wu C., Wun C., Chou H., « Using Association Rules for Completing Missing Data. », *Proceedings of the 4th International Conference on Hybrid Intelligent Systems, (HIS'04), IEEE Computer Society Press*, Kitakyushu, Japan, p. 236-241, 5-8 December, 2004.

Article reçu le 9 Avril 2010
Accepté après révisions le 17 mars 2011

Leila Ben Othman est doctorante en Informatique en cotutelle entre la Faculté des Sciences de Tunis et l'Université de Caen, Basse-Normandie. Elle est également enseignante d'informatique au département informatique à l'Institut Supérieur des Arts Multimédia de la Manouba. Ses travaux portent sur le traitement des valeurs manquantes en fouille de données. Elle s'intéresse plus précisément aux modèles d'apparition des valeurs manquantes afin d'améliorer l'étape de complétion.

François Rioult

Sadok Ben Yahia

Bruno Crémilleux