Étude, proposition et validation d'une méthode de complétion utilisant les modèles d'apparition des valeurs manquantes

Leila Ben Othman

Département Informatique

École doctorale SIMEM
UFR Sciences

Conception et validation d'une méthode de complétion des valeurs manquantes fondée sur leurs modèles d'apparition

THÈSE

présentée et soutenue publiquement le 18 Novembre 2011

pour l'obtention du

Diplôme de Doctorat en Informatique

par

Leila Ben Othman Amroussi

Rapporteurs :	Rafik Bouaziz	Maître de Conférences - H.D.R.	Faculté des Sciences Économiques et de Gestion de Sfax
	Maguelonne Teisseire	Directrice de Recherche	CEMAGREF
Examinateurs :	Gael Dias	Professeur	Université de Caen Basse-Normandie
	Boutheina Ben Yaghlane	Maître de Conférences - H.D.R.	Institut des Hautes Études Commerciales de Carthage
Directeurs :	Sadok Ben Yahia	Maître de Conférences - H.D.R.	Faculté des Sciences de Tunis
	Bruno Crémilleux	Professeur	Université de Caen Basse-Normandie

Unité de Recherche en Programmation, Algorithmique et Heuristique
Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen — CNRS UMR 6072

Mis en page avec la classe thloria.

Remerciements

Je tiens tout d'abord à exprimer ma gratitude envers M. Rafik Bouaziz et Mme Maguelonne Teisseire qui ont accepté de rapporter ce travail. Je les remercie pour leur lecture minutieuse de mon manuscrit ainsi que de leurs remarques qui m'ont été précieuses.

J'exprime ma gratitude envers M. Gael Dias et Mme Bouthaina Ben Yaghlane de me faire l'honneur de participer au jury.

Je remercie François Rioult de m'avoir permis de poursuivre mes travaux de recherche sur la problématique des valeurs manquantes. Il a su diriger ce travail de thèse avec beaucoup de patience et de rigueur scientifique. Il s'est toujours montré chaleureusement disponible et a su guider mes réflexions. Je le remercie particulièrement pour sa présence amicale lors de mes premiers séjours à Caen.

Je suis particulièrement heureuse que ce ce travail ait été mené sous la direction de Sadok Ben Yahia. Il m'a témoigné de beaucoup de confiance et a su guider mes premiers pas dans le domaine de la recherche. Je le remercie pour son soutien tout au long de l'élaboration de ce travail.

Je remercie également chaleureusement Bruno Crémilleux. Auprès de lui, j'ai beaucoup appris en termes d'habileté d'analyse et de finesse scientifique. Je le remercie de la confiance qu'il m'a accordée.

Je remercie également tous les membres de l'équipe Greyc qui m'ont toujours chaleureusement accueilli pendant ces années de thèse.

Je remercie tout spécialement M. Khaled Bsaies directeur de l'URPAH et M. Mohamed Jemal directeur de l'École Doctorale de la Faculté des Sciences de Tunis d'avoir accepté de financer en partie mes séjours à Caen, ainsi que Mme Chiraz Latiri, directrice de l'Institut Supérieur des Arts Multimédia de la Manouba pendant mes années de thèse, qui m'a toujours soutenu et aidé pour la gestion des aspects administratifs.

Une pensée spéciale à mon amie tunisienne de Caen : Lamia Belouaer pour sa présence incontournable à Caen et à Sylvie Huguenin qui m'a toujours accueilli avec beaucoup d'attention au sein de sa famille.

Je remercie spécialement mes amis : Tarek Hamrouni, Sarah Ayouni et Sameh Guezguez pour la relation amicale et scientifique que nous entretenons. Merci à mes amies de toujours : Chaima et Houda. Merci à ma tante Rafla Mrabet pour ses conseils en anglais lors de mes interventions internationales.

Merci également à ma famille : mes parents pour leur soutien qui ne m'a jamais fait défaut, mes beaux-parents, mes frères et leurs femmes, ainsi que mes chers petits neveux pour leur présence quotidienne.

Ces années de thèse auraient été pénibles sans la présence de Mohamed. Je voudrai le remercier pour sa patience. Il a toujours su s'adapter à mes absences, a accepté tant de sacrifices et n'a jamais cessé de m'encourager. J'espère qu'il est fier de moi.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Contexte

La *fouille de données* est fréquemment employée comme synonyme de l'ECBD (Extraction de Connaissances dans les Bases de Données). Elle constitue en réalité la phase d'extraction des connaissances à partir des données dans le processus de l'ECBD. L'ECBD rend exploitable les données collectées et extrait des connaissances, utiles pour la prise de décision [Han et Kamber, 2000]. Cette discipline est classiquement décrite comme un processus semi-automatique, itératif, constituée de plusieurs étapes : sélection et pré-traitement des données, fouille de données (ou data mining) à l'aide d'algorithmes, visualisation et interprétation des résultats [Lefébure et Venturi, 1999]. Cependant, bien que les techniques classiques de fouille de données parviennent maintenant à maturité de développement, les spécialistes se tournent désormais vers l'étude des contextes difficiles. En effet, ces techniques classiques ne restent pas sans faille dans le cas de données complexes, *i.e.,* incomplètes, peu structurées ou redondantes, *etc* [Han et Kamber, 2000]. Ainsi, la fouille de données complexes détermine un axe de recherche en plein essor, puisque la qualité des connaissances extraites va de pair avec la qualité des données collectées.

C'est dans le cadre de l'extraction de connaissances à partir de données *incomplètes* que se situe cette thèse. Ce travail contribue plus particulièrement à l'étape de pré-traitement des données, en proposant une méthode de complétion des valeurs manquantes, afin de rendre les données exploitables par des techniques de fouille de données. Plus précisément, nous proposons d'aborder cette problématique par la définition de *modèles d'apparition* des valeurs manquantes permettant de proposer une méthode fine de complétion.

Motivations

Les données issues du monde réel ne sont pas toujours complètes, lorsque certaines informations ne sont pas disponibles, pas renseignées ou sont aberrantes [Pearson, 2006]. Ceci semble constituer un phénomène aussi imprévisible qu'inévitable, dû à de multiples raisons : oubli de la part de l'utilisateur, refus de réponse lors de sondages, impossibilité d'acquisition de valeurs, *etc.*

Le traitement des valeurs manquantes a suscité l'intérêt de plusieurs communautés scientifiques. La solution la plus simple consiste en la suppression de toute donnée comportant des valeurs manquantes. Cependant, cette solution ne peut-être appliquée que dans le cas où les valeurs manquantes sont peu nombreuses ; la perte des données consécutives serait sinon considérable et lourde de conséquence. D'autres travaux ont été proposés et ont consisté à chercher pour chaque valeur manquante une valeur de remplacement. On parle dans ce cas de *complétion des valeurs manquantes*.

Nous verrons dans le chapitre 1 que la littérature regorge de nombreuses contributions dans ce domaine. Bien qu'elle ne soit pas toujours clairement formulée, l'hypothèse que les valeurs manquantes apparaissent suivant un modèle uniquement aléatoire, selon la classification proposée dans [Little et Rubin, 2002], est implicitement utilisée par la majorité des méthodes de la littérature. 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.

Au meilleur de 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. Pourtant, une analyse préliminaire sur le modèle d'apparition des valeurs manquantes permettrait de mieux adapter le traitement. Lorsqu'une valeur manquante n'est pas aléatoire, elle est dite *informative* car elle permet de caractériser une situation particulière et 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, alors 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. Cependant, le modèle d'apparition de ces dernières est rarement discuté et souvent réduit au seul modèle aléatoire. Cette limitation et la non prise en compte d'informations importantes sur l'origine des valeurs manquantes pénalisent la complétion [Delavallade et Dang, 2007, Fiot *et al.*, 2007]. Nous souhaitons dans cette thèse souligner l'importance d'effectuer une analyse préliminaire sur les origines des valeurs manquantes afin de prescrire une méthode de traitement adéquate.

Contributions

Dans ce manuscrit, nous présentons et évaluons une nouvelle méthode de complétion mettant à profit les modèles d'apparition des valeurs manquantes. Nous procédons pour cela aux quatre étapes suivantes :

1. Caractérisation non redondante des valeurs manquantes et proposition d'une nouvelle ty-

pologie utilisant la base d'implications propres ;

2. Extraction de caractérisation non redondante par traverses minimales ;
3. Conception d'une nouvelle méthode de complétion contextualisée ;
4. Évaluation de cette méthode de complétion.

Caractérisation non redondante des valeurs manquantes et proposition d'une nouvelle typologie

Le premier axe de notre travail consiste en la proposition d'une caractérisation plus fine que celle de la littérature [Little et Rubin, 2002] de l'origine des valeurs manquantes. Cette caractérisation permettra de mieux prendre en compte les valeurs manquantes, dans le cadre d'un processus d'extraction de connaissances à partir des bases de données, ou encore en amont d'un traitement statistique. L'impact de cet axe de travail est double : d'une part, il permet de mieux comprendre les causes des valeurs manquantes et contribue à l'amélioration de la qualité des données ; d'autre part des méthodes de complétion plus efficaces peuvent être développées, tirant bénéfice de ces modèles d'apparition des valeurs manquantes.
En effet, l'examen minutieux 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 d'une grande utilité lors de l'analyse de données incomplètes. Il peut également y avoir plusieurs explications à l'absence d'une valeur selon les 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 ici une nouvelle typologie des valeurs manquantes adaptée à des groupes d'objets spécifiques. En mettant à profit ces caractérisations, ceci permet d'affiner le niveau de granularité de toute méthode de complétion en aval. La contribution de ce premier axe de travail est donc 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 ?

Au chapitre 4, nous avons ainsi explicité différents modèles de valeurs manquantes et proposé une nouvelle typologie (aléatoire, directe, indirecte et hybride) reposant uniquement sur les

données connues, qui différencie les origines de valeurs manquantes selon les groupes d'objets où elles apparaissent. Les points clés de notre typologie peuvent être résumés comme suit :

- Une caractérisation propre pour chaque valeur manquante et non pas une caractérisation globale, comme c'est le cas dans [Little et Rubin, 2002].
- Cette typologie est mise en évidence par une approche *fouille de données*, *i.e.*, nous ne sommes pas partis d'une typologie définie *a priori*, mais nous avons utilisé une démarche analytique pour la faire émerger.
- Nous utilisons une technique efficace de recherche de régularités (base d'implications propres [Taouil et Bastide, 2001]).

Calcul de caractérisation non redondante par traverses minimales

Notre deuxième contribution porte sur le calcul de la base de règles qui permet de caractériser les valeurs manquantes de façon non redondante. Dans ce but, nous avons relié la problématique de l'extraction de la base d'implications propres au cadre des hypergraphes. En effet, la règle $X \rightarrow i$ est une implication propre si X est une traverse minimale [Berge, 1989] des complémentaires des objets contenant l'item i. Ainsi, adapter l'algorithme d'extraction des traverses minimales proposé dans [Hébert *et al.*, 2007] en lui ajoutant une contrainte de fréquence nous permet de calculer efficacement les implications propres.

Complétion contextualisée des valeurs manquantes

Notre troisième contribution porte sur la proposition d'une méthode de complétion qui consiste à tirer profit de l'information supplémentaire obtenue lors de la phase de caractérisation, pour l'intégrer dans la phase de complétion. Cette complétion est *contexualisée* : elle prend différentes formes, suivant le type et le contexte de la valeur manquante.
Les principales caractéristiques de notre méthode de complétion sont comme suit :

- Une complétion caractéristique du type de la valeur manquante ;
- Une complétion *contextualisée* par « valeur spéciale », caractéristique de l'origine d'une valeur manquante, en étendant l'ensemble de définition des différents attributs.

Évaluation des techniques de complétion des valeurs manquantes

Finalement, nous avons été amenée à étudier les techniques d'évaluation et de validation des méthodes de complétion, où nous avons distingué deux approches principales : celles mesurant la proximité des données complétées par rapport aux données de référence et celles mesurant l'impact d'une méthode de complétion sur des techniques d'apprentissage. Pour cela, nous avons mis en place une nouvelle technique d'évaluation fondée sur la stabilité d'un clustering entre les données de référence et les données complétées.

Organisation du mémoire

Ce mémoire de thèse est organisé en deux parties.

La première partie dresse un état de l'art des contributions relatives aux valeurs manquantes : les modèles et les méthodes de traitement des valeurs manquantes sont présentés dans le chapitre 1, tout en montrant l'importance de l'étude de leurs modèles d'apparition. La synthèse de ces travaux permet de positionner notre contribution.

Dans le chapitre 2, les deux techniques d'évaluation des méthodes de complétion des valeurs manquantes sont présentées : celle mesurant la proximité des données complétées avec celles de références, et celle analysant l'impact de la complétion sur des techniques d'apprentissage supervisé. Ensuite, les protocoles expérimentaux associés à ces techniques d'évaluation sont exposés. Finalement, une discussion ainsi qu'une étude de l'adéquation entre les techniques et les protocoles d'évaluation sont menées.

Le chapitre 3 de l'état de l'art porte sur les règles de caractérisation non redondantes minimales. Nous présentons dans ce chapitre la notion de règles redondantes en expliquant l'intérêt de telles règles et en détaillant les différentes couvertures de règles d'association de la littérature.

La deuxième partie de ce manuscrit présente l'ensemble de nos contributions : dans le chapitre 4, nous montrons que les valeurs manquantes ne sont pas forcément aléatoires. Nous proposons une nouvelle typologie des valeurs manquantes que nous caractérisons de manière non redondante en employant la base d'implications propres. Le chapitre 5 présente une nouvelle méthode d'extraction de la base d'implications propres par traverses minimales. Le chapitre 6 montre l'usage pratique de la caractérisation des valeurs manquantes en proposant une nouvelle méthode de complétion. Le chapitre 7 est consacré à la validation de cette méthode, où nous commençons par évaluer la pertinence des techniques mesurant l'impact de la complétion sur les méthodes d'apprentissage supervisé. Ensuite, nous introduisons une nouvelle technique d'évaluation à base de stabilité de méthode d'apprentissage non supervisé.

La conclusion du mémoire et les perspectives de travaux ultérieurs sont présentées dans le dernier chapitre.

Première partie

État de l'art

Chapitre 1

Modèles et méthodes de traitement des valeurs manquantes

Sommaire

La présence des valeurs manquantes est un problème, qui s'est toujours posé depuis l'émergence du domaine de l'analyse des données. Ainsi, nous assistons depuis longtemps à une grande prolifération des méthodes de traitement des valeurs manquantes [Calders *et al.*, 2007]. En effet, un large éventail de méthodes sont proposées dans la littérature : des méthodes les plus simples ou palliatives aux méthodes faisant l'adaptation d'algorithmes existants ou encore celles présentant une technique de complétion [Ragel et Crémilleux, 1999]. Bien que l'objectif de ces méthodes soit souvent le même, la problématique des valeurs manquantes est loin d'être traitée de la même façon.

Dans ce chapitre, nous commençons tout d'abord par présenter le formalisme de notre travail (contexte réel, mesuré, valeur manquante, motif, *etc*). Ensuite, nous présentons les modèles d'apparition des valeurs manquantes, largement connus [Little et Rubin, 2002], tout en discutant cette modélisation, en amont de toute méthode de traitement. Nous décrivons ensuite les méthodes dédiées au traitement des valeurs manquantes. Sans prétendre à l'exhaustivité, cette étude bibliographique passe en revue les principaux domaines dans lesquels la problématique des valeurs manquantes a été abordée : statistique, base de données, apprentissage supervisé, *etc.* Dans la quatrième section de ce chapitre, nous nous intéresserons en particulier aux méthodes de fouille de données, domaine de recherche sur lequel porte notre contribution. Nous y présentons d'abord les méthodes faisant de l'extraction de motifs en présence de valeurs manquantes. Ensuite, nous présentons les méthodes de complétion à base de règles d'association, une technique classique de fouille de données. Finalement, nous mènerons une discussion sur les méthodes de cet état de l'art tout en positionnant notre travail par rapport à ces méthodes.

1.1 Préliminaires

Les données que nous étudions sont initialement au format "attribut/valeur". 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*, par opposition à *contexte mesuré* présentant des valeurs manquantes :

Définition 1 (Contexte réel) *Un* contexte réel *est un triplet* $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$*, décrivant deux ensembles finis* $\mathcal{O}$ *et* $\mathcal{I}$ *et une fonction* $\mathcal{R}$ *sur* $\mathcal{O} \times \mathcal{I}$ *prenant ses valeurs dans* $\{présent, absent\}$*. L'ensemble* $\mathcal{O}$ *est appelé ensemble des objets (ou transactions) et* $\mathcal{I}$ *est appelé ensemble des items. Ainsi,* $\mathcal{R}(o, i) = présent$ *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*.*

La table 1.1 montre un exemple d'un contexte réel, où le symbole « × » indique la présence de l'item en question.

Un motif est un ensemble d'items. Un motif X est fréquent si et seulement si son support, défini par le nombre d'objets supportant X et noté $Supp(X)$, dépasse un certain seuil, noté *minsup* fixé au préalable. Une règle d'association basée sur un motif $Z \neq \emptyset$ est une expression entre deux motifs X et Y de la forme $R : X \rightarrow Y$ telle que $X \subsetneq Z$ et $Y = Z \backslash X$. Les motifs X et Y sont respectivement appelés *prémisse* et *conclusion* de la règle R. Le support de la règle R, noté $Supp(R)$ est égal à celui du motif 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 est dite *exacte* si sa confiance vaut 1, sinon elle est dite *approximative*.

	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		×		×			×		×

TAB. 1.1 – Exemple d'un contexte réel $\mathcal{K}$.

La problématique de découverte des règles d'association valides[1] repose sur l'extraction des motifs fréquents. En effet, l'étape de génération des règles est plutôt facile relativement à l'étape d'extraction des motifs fréquents : pour chaque motif fréquent X, il suffit de générer les règles valides de la forme $X' \rightarrow X \setminus X'$ [Agrawal et Srikant, 1994] dont la confiance dépasse le seuil $minconf$, avec $X' \subset X$.

Dans un contexte réel $\mathcal{K}$, un attribut A_i présente parfois une valeur non renseignée, dite *valeur manquante*, notée par " ?". Ainsi, les données mesurées peuvent être incomplètes. Dans ce cas, la définition d'un contexte nécessite d'être adaptée comme suit :

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 $\{présent, absent, manquant\}$.*

La table 1.2 montre un exemple d'un contexte mesuré associé au contexte réel de la table 1.1. Par exemple, les valeurs manquantes sur les items a et b, dans l'objet o_8 du contexte $mv(\mathcal{K})$, cachent en réalité la présence de l'item b dans $\mathcal{K}$.

[1] Une règle est dite valide si son support et sa confiance sont respectivement supérieurs ou égaux à *minsup* et $minconf$.

	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	?	?		×			×	?	?

TAB. 1.2 – Contexte mesuré $mv(\mathcal{K})$ associé au contexte réel $\mathcal{K}$ donné par la table 1.1.

1.2 Modèles d'apparition des valeurs manquantes

Le modèle d'apparition des valeurs manquantes définit la loi selon laquelle les valeurs manquantes apparaissent dans les données. Plus précisément, un modèle indique la probabilité qu'une valeur manquante dépende de sa propre valeur, d'une ou plusieurs autres valeurs ou qu'elle soit complétement indépendante des données.

L'immense majorité des travaux sur les valeurs manquantes cite la classification basée sur trois modèles statistiques proposés dans [Little et Rubin, 2002]. L'étude menée par LITTLE et RUBIN propose de répartir les modèles d'apparition des valeurs manquantes en modèle complètement aléatoire (MCAR), aléatoire (MAR) et non aléatoire (NMAR). Les acronymes MCAR, MAR et NMAR correspondent respectivement à Missing Completely at random, Missing at Random et Not Missing at Random. Ces trois modèles sont définis comme suit :

- **MCAR** : une valeur manquante est dite *complètement aléatoire* lorsque la non-réponse est totalement indépendante de toute autre valeur. Une valeur manquante complètement aléatoire affecte donc n'importe quel attribut et n'importe quel objet : la probabilité qu'elle soit manquante est la même pour toutes les valeurs. Par exemple, dans la table 1.3 où nous représentons deux attributs *sexe* et *poids*, la colonne *MCAR* indique le cas où les valeurs manquantes sur l'attribut *poids* sont de type *MCAR*. En effet, sur cette colonne, les valeurs manquantes affectent d'une manière équirépartie les personnes de sexe féminin ou masculin. De plus, ces valeurs manquantes ne dépendent pas des valeurs réelles de l'attribut *poids*, qu'on ne possède pas en réalité.
- **MAR** : une valeur manquante est dite *aléatoire* lorsque l'absence d'une valeur dépend des valeurs réelles particulières d'autres attributs. Il est à noter que, dans ce cas, la désignation *aléatoire* prête à confusion, car elle concerne les valeurs réelles de l'attribut présentant la valeur manquante. Si nous reprenons la table 1.3, alors nous remarquons dans la colonne

MAR que toute personne de sexe féminin présente une valeur manquante sur l'attribut *poids*. Ces valeurs manquantes dépendent donc du sexe de la personne : on peut, par exemple, déduire que seules les femmes ont tendance à cacher leurs poids, quelle que soit la valeur réelle du poids. On peut très bien remarquer, à partir de cet exemple, que les valeurs réelles du poids de ces femmes n'appartiennent pas à un intervalle de valeurs précis, d'où le terme « *aléatoire* ».

– **NMAR** : si une valeur est manquante lorsque la valeur réelle de l'attribut correspondant est particulière, alors le modèle est dit *non-aléatoire*, *i.e.,* lorsque le phénomène de « non-réponse » dépend de la valeur manquante elle-même. Les valeurs manquantes sur l'attribut *poids* (dernière colonne) illustrent un exemple de valeurs manquantes de type *NMAR*. En effet, nous remarquons que les personnes concernées par ces valeurs manquantes sont celles dont le *poids* est supérieur à 100, car les personnes ayant un sur-poids ont tendance à le cacher.

sexe	**poids**			
	Valeur réelle	**MCAR**	**MAR**	**NMAR**
M	80	?	80	80
F	67	67	?	67
F	53	?	?	53
M	132	?	132	?
M	62	62	62	62
F	57	57	?	57
M	70	?	70	70
F	62	62	?	62
M	115	115	115	?
F	55	55	?	55
F	145	?	?	?
F	110	110	?	?

TAB. 1.3 – Exemple des trois modèles d'apparition des valeurs manquantes.

Dans le chapitre 4, nous discuterons des limites de la modélisation de LITTLE et RUBIN. Nous expliquerons notamment la difficulté liée à la caractérisation des valeurs manquantes *NMAR*. Cette difficulté explique pourquoi la majorité des travaux ne considère que le modèle *MCAR* [Hawarah, 2008, Fiot, 2007, Ragel, 1999].

En présence de valeurs manquantes, il est primordial de considérer leurs modèles d'apparition. Ceci permet de comprendre pourquoi les données sont manquantes, et facilite par la suite leur traitement [Delavallade, 2007]. En considérant l'exemple des personnes présentant un sur-poids et qui ont tendance à le cacher, il s'avère intéressant de pouvoir caractériser le fait que toutes les

personnes présentant un poids manquant sont des personnes concernées par un sur-poids. Il est alors facile de classer toute personne ayant un poids manquant dans la catégorie des personnes obèses.

Bien que la majorité des travaux dédiés au traitement des valeurs manquantes évoque l'importance de la modélisation des valeurs manquantes, dans la pratique, ils ne traitent que du modèle aléatoire. Dans [Shafer et Graham, 2002], cette hypothèse est justifiée par le fait que cette modélisation est difficile à effectuer. Dans [Song et Shepperd, 2007], les auteurs parlent même du risque encouru lorsqu'on ignore cette modélisation, mais comme tous les autres travaux, cette considération n'a pas été prise en compte lors du protocole expérimental.

1.3 Méthodes de traitement des valeurs manquantes

Il existe une abondante littérature sur les méthodes de traitement des valeurs manquantes. On peut répartir l'ensemble de ces méthodes en des méthodes simples ou palliatives et des méthodes plus élaborées. Ces dernières concernent plusieurs domaines tels que la statistique, l'apprentissage supervisé, les bases de données, la théorie des ensembles approximatifs ou la fouille de données. Dans ce qui suit, nous exposons une panoplie de travaux qui couvrent tant que possible les différents domaines dans lesquels la problématique des valeurs manquantes a été abordée. Nous nous intéresserons spécialement aux méthodes de *complétion* utilisant une technique de fouille de données (*e.g.*, les règles d'association).

1.3.1 Les méthodes palliatives

Comme leur nom l'indique, les méthodes palliatives n'ont pas pour objectif de compléter de façon précise les valeurs manquantes. Elles cherchent plutôt à remplacer les trous engendrés par les valeurs manquantes pour fournir des données exploitables par les techniques d'analyse. Parmi les techniques palliatives, on trouve la technique de *suppression* qui consiste à supprimer toute donnée comportant des valeurs manquantes. En réalité, cette technique peut-être appliquée de deux façons différentes : suppression par ligne (listwise deletion) ou suppression par attribut (pairwise deletion) [Little et Rubin, 2002]. Cependant, ces solutions ne peuvent être appliquées sans conséquence lorsque les valeurs manquantes sont nombreuses, car la perte des données serait considérable. Par ailleurs, ces techniques ne sont envisageables que lorsque les valeurs manquantes sont de type *MCAR*. Dans le cas contraire, les données retenues ne seront pas représentatives et l'analyse sera biaisée.

D'autres travaux ont été proposés, consistant à chercher, pour chaque valeur manquante, une valeur de remplacement. On parle dans ce cas de *complétion* ou d'*imputation* des valeurs manquantes [Pearson, 2006]. Parmi les techniques palliatives qui procèdent à l'imputation, la

moyenne, le *mode* ou la *médiane* sont utilisées. Ces techniques consistent à remplacer chaque valeur manquante respectivement par la *moyenne*, le *mode* ou la *médiane* de l'ensemble des valeurs observées de l'attribut affecté par la valeur manquante. Ces techniques sont connues pour leur impact sur la distribution des données, notamment sur la variance, ainsi que le biais qu'elles introduisent au niveau de la corrélation entre les attributs.

Outre les méthodes palliatives, le premier domaine pour lequel la problématique des valeurs manquantes a suscité beaucoup d'intérêt est celui de la statistique, où la diversité des contributions souligne l'importance du problème pour les statisticiens.

1.3.2 Les méthodes statistiques

L'algorithme EM

L'algorithme Esperance-Maximization (EM) [Dempster *et al.*, 1977] est une technique de référence pour estimer les valeurs manquantes. Le pseudo-code de l'algorithme EM est illustré par l'algorithme 1. EM utilise les notations suivantes :

- Y : la variable explicative présentant les valeurs manquantes ;
- Y_{obs} et Y_{miss} : représentent respectivement la partie observée et la partie manquante de Y.
- θ : le paramètre du modèle ;

L'algorithme est itératif et procède en deux étapes :

- L'étape E (Espérance) est l'étape de calcul de l'espérance et consiste à compléter les données manquantes Y_{miss} à partir de l'estimation de θ.
- L'étape M (Maximisation) consiste à estimer θ en maximisant une fonction de vraissemblance.

Algorithme 1 : L'algorithme EM.

```
1 Algorithme : EM
2 début
3     initialisation au hasard de θ
4     tant que l'algorithme n'a pas encore convergé faire
5         Étape-E (Espérance)
6         compléter les valeurs manquantes Y_miss en se basant sur l'estimation de θ
7         Étape-M (Maximisation)
8         estimer θ en se basant sur Y_obs de manière à accroître la vraissemblance
9 fin
```

Plusieurs contributions portant sur le traitement des valeurs manquantes se sont basées sur l'algorithme EM, *e.g.*, le travail proposé dans [Ghahramani et Jordan, 1994]. D'autres travaux ont

également été proposés dans le domaine de la statistique tel que [Ben Salem, 1999], où l'auteur procède à l'imputation des valeurs manquantes considérées comme des trous dans un tableau de données en employant itérativement la technique des plus proches voisins.

L'imputation multiple

Afin de réduire le biais engendré par les techniques d'imputation simple, la technique de l'imputation multiple [Rubin, 1978] est largement utilisée au sein de la communauté des statisticiens. Son principe se décompose en trois étapes : en premier lieu, chaque valeur manquante est complétée m fois de façon à obtenir m bases complètes. Le but de cette première étape est de créer plusieurs valeurs possibles d'une valeurs manquante. La deuxième étape consiste à faire l'analyse des m bases obtenues. Finalement, les résultats des m analyses sont combinés. L'avantage de l'imputation multiple est qu'elle permet de réduire le biais induit par l'imputation selon une valeur unique. En revanche, cette technique peut s'avérer coûteuse, bien que l'auteur montre dans [Rubin, 1978] que le résultat de 3 à 5 imputations peut s'avérer intéressant.

La technique de régression

La technique de régression a pour objectif d'étudier le lien entre deux variables : l'une est dite variable *expliquée*, la variable qu'on cherche à expliquer ou à prédire, à partir d'autres variables dites *explicatives*. Cette technique est souvent employée pour la prédiction des valeurs manquantes, où une valeur manquante est considérée comme étant la variable *expliquée*. Dans ce cas, une régression linéaire est souvent employée. Le principal inconvénient des techniques de régression est qu'elles se basent sur des modèles non justifiés [Magnani, 2004].

1.3.3 Les bases de données

Dans le domaine des bases de données, l'expression « valeur nulle » est plus couramment utilisée que celle de valeur manquante. Dans [Date, 1995], l'auteur propose de remplacer les valeurs nulles sur chaque colonne par « valeur spéciale » (ou default-value). De cette façon, ces valeurs spéciales seront traitées comme de nouvelles modalités. Dans [Codd, 1990], le fondateur de l'algèbre relationnelle adopte une vision plus pragmatique et indique qu'il existe deux sortes de valeurs manquantes : celle manquante mais applicable et celle manquante mais qui cache en réalité une valeur inapplicable. Sans chercher à trouver les valeurs réelles, il propose deux nouvelles valeurs de remplacement : « A-mark » (missing-but-applicable) et « I - mark » (inapplicable). Ces deux nouvelles valeurs seront par la suite employées pour remplacer les valeurs nulles. Codd a recommandé l'utilisation de ces nouveaux symboles et a présenté un cadre théorique pour leur prise en considération. D'autres travaux ont étudié la problématique des valeurs manquantes dans le cadre des dépendances fonctionnelles, citons à titre d'exemple [Levene et Loizou, 1993].

1.3.4 Les méthodes supervisées

Les méthodes supervisées sont utilisées pour prédire la classe d'appartenance d'un objet. L'idée repose sur l'utilisation d'exemples déjà classés pour apprendre un modèle puis de l'utiliser afin de déterminer la classe de tout nouvel exemple. Ces méthodes opèrent en deux étapes :

- l'étape d'apprentissage, qui consiste à construire le *modèle* des exemples supervisés ;
- l'étape de test, de classement ou de décision, qui valorise ce modèle pour classer les exemples, dont la classe est inconnue.

Pour exhiber un modèle, l'objectif consiste à trouver des relations dans les données permettant d'expliquer leur appartenance à une classe spécifique. Le modèle utilisé pour la prédiction peut se présenter sous différentes formes : règles d'association [B. Liu, 1998, Li *et al.*, 2001], arbres de décision [Quinlan, 1986], réseaux de neurones [Ripley, 1996], *etc.*

Dans ce qui suit, nous présentons les travaux qui ont abordé la problématique des valeurs manquantes dans le cadre de l'apprentissage supervisé. Ces travaux emploient principalement le modèle à base d'arbre de décision.

L'algorithme C4.5

La méthode C4.5 [Quinlan, 1993] se situe dans le cadre de l'apprentissage supervisé à base d'arbre de décision. Un arbre de décision est construit à partir des exemples d'apprentissage, où chaque branche de l'arbre correspond à un test portant sur les valeurs d'un attribut. Un chemin de la racine à un nœud feuille correspond à une règle de classement, qui précise la relation entre une classe spécifique et les valeurs des attributs. La phase de test d'un nouvel exemple de classe inconnue consiste à parcourir l'arbre en choisissant à chaque niveau la branche adéquate. La classe d'appartenance de l'exemple est indiquée par le nœud feuille sur lequel a porté le parcours.

Dans le cadre d'apprentissage supervisé à base d'arbre de décision, le problème des valeurs manquantes se présente sur deux niveaux [Liu *et al.*, 1997] : lors de la construction de l'arbre ainsi que lors du classement de nouveaux exemples. En effet, si la valeur d'un attribut est manquante, il est alors impossible de trancher dans quelle branche de l'arbre il faut envoyer l'exemple.

Dans C4.5, le traitement des valeurs manquantes a été abordé lors de la phase d'apprentissage ainsi que lors de la phase de classement selon une approche probabiliste. Au cours de la phase d'apprentissage, l'idée consiste à envoyer dans différentes branches de l'arbre un exemple présentant une valeur manquante en associant à chaque branche une probabilité. Cette probabilité est calculée à partir des exemples ne présentant pas de valeurs manquantes. Ainsi, un exemple se retrouve fragmenté dans plusieurs sous-branches avec différentes probabilités. Lors du classement d'un nouvel exemple, l'algorithme C4.5 renvoie la valeur de classe ayant la probabilité la plus élevée.

Utilisation de l'entropie pour substituer les valeurs manquantes

Dans [Delavallade et Dang, 2007], la problématique des valeurs manquantes est abordée de façon différente des autres approches de la littérature. L'originalité de ce travail est qu'il importe peu que la valeur de complétion soit proche de la valeur réelle. En effet, au cours de la construction de l'arbre de décision, le choix des attributs pour faire la division en branches est primordial. Ce choix se fait généralement en ayant recours à des mesures telles que l'*entropie* [Shannon, 1948] utilisé dans ID3 et C4.5 [Quinlan, 1993] et l'*indice de Gini* employé dans CART [Breiman *et al.*, 1984]. La division en branches est effectuée selon l'attribut le plus discriminant, c'est-à-dire celui qui sépare le mieux les différents exemples en classes distinctes. Par exemple, dans le cadre de l'apprentissage supervisé, l'*entropie* mesure l'homogénéité des exemples. En choisissant l'attribut dont la valeur d'*entropie* est la plus réduite, la division permet de séparer, autant que possible, les exemples de l'échantillon d'apprentissage. Ainsi, l'intérêt de ce travail est de préserver le pouvoir discriminant d'un attribut lorsqu'il présente des valeurs manquantes. L'inconvénient de cette approche est qu'elle n'est applicable que lors de la phase d'apprentissage puisque l'information de la classe est utilisée [Hawarah, 2008].

Approche probabiliste pour le classement d'objets incomplets dans un arbre de décision

À l'inverse de la méthode décrite précédemment, l'approche proposée dans [Hawarah *et al.*, 2006] s'intéresse au traitement des valeurs manquantes lors la phase de classement. Cette méthode étend l'approche à base d'arbres d'attributs ordonnés (AAO) proposée dans [Lobo et Numao, 1999]. L'approche AAO consiste à construire, pour chaque attribut présentant des valeurs manquantes, un arbre de décision appelé *arbre d'attributs*, dont les nœuds feuilles représentent les valeurs de l'attribut manquant. La construction des arbres se fait selon l'ordre croissant de l'information mutuelle [Shannon, 1948], calculée entre l'attribut en question et l'attribut de classe. L'ordre croissant garantit de commencer par l'attribut le moins dépendant de la classe. L'arbre est donc utilisé pour déterminer les valeurs manquantes sur l'attribut associé en considérant à chaque fois les attributs traités. L'extension de AAO a donné lieu à deux approches : l'approche AAOP (Arbres d'Attributs Ordonnés Probabilistes) [Hawarah *et al.*, 2004], qui déploie la notion d'arbre probabiliste et construit des arbres d'attributs selon la méthode AAO, mais en conservant toutes les valeurs possibles d'un attribut manquant. Le résultat du classement est une distribution probabiliste de classe. La deuxième approche est AAP (Arbres d'attributs Probabilistes) [Hawarah *et al.*, 2006] et consiste à construire un arbre pour chaque attribut en considérant les attributs dont il dépend.

Nous renvoyons le lecteur à un état de l'art détaillé sur les méthodes de traitement des valeurs manquantes à base d'arbre de décision [Hawarah, 2008]. Dans cet état de l'art, d'autres

méthodes sont décrites telles que la méthode CART [Breiman *et al.*, 1984], qui repose sur le principe d'attribut de substitution ou la méthode de Shapiro, décrite dans [Quinlan, 1986], qui consiste à considérer l'attribut manquant comme étant la classe à prédire.

1.3.5 Les ensembles approximatifs

La théorie des ensembles approximatifs est une extension de la théorie des ensembles classiques, où les données sont imprécises. Dans le cadre de cette théorie, une contribution proposée dans [Nayak *et al.*, 2001] a porté sur l'extraction de règles approximatives où l'évaluation de présence des motifs est effectuée de façon probabiliste. Dans [Li et Cercone, 2006], les auteurs ont proposé une approche intitulée *RSFit*, où l'ensemble des attributs les plus représentatifs permet de trouver les valeurs de complétion à partir de l'instance la plus proche à travers un calcul de distance. L'instance la plus proche est déterminée en se basant soit sur la totalité des instances, soit sur celles ayant la même valeur de l'attribut décision. *RSFit* a été combinée à une approche à base de motifs, où les auteurs ont exploité en outre les corrélations induites par les motifs fréquents ; cette combinaison a donné lieu à *ItemRSFit* [Li *et al.*, 2007].
Les travaux les mieux référencés, dans le cadre de cette théorie, sont ceux de Grzymala. Dans [Grzymala-Busse *et al.*, 1999], les auteurs ont introduit l'algorithme *ClosetFit* qui a inspiré *RSFit*. Cet algorithme a été employé dans un cadre de classification d'objets en se basant sur des règles dites *certaines* ou *possibles* issues de la théorie des ensembles approximatifs. Dans [Grzymala-Busse et Hu, 2001], les auteurs comparent plusieurs approches :

- ignorer les objets incomplets ;
- remplacer par la valeur la plus fréquente ;
- remplacer par la valeur moyenne ;
- remplacer par toutes les possibilités : un objet incomplet sera remplacé par un ensemble d'objets dans lequel une valeur manquante est remplacée par les valeurs possibles de l'attribut ;
- créer une valeur particulière « manquante » pour chaque attribut.

Toutes ces approches ont été par ailleurs, appliquées en se restreignant aux objets relatifs à une même classe. Dans [Grzymala-Busse, 2004], Grzymala évoque trois types de valeurs manquantes intrinsèquement liées à leurs origines :

- valeur manquante qui cache une valeur perdue ;
- valeur manquante qui cache une valeur non pertinente ;
- valeur manquante non pertinente mais relative à une classe.

La caractéristique de ce travail est que l'auteur propose un traitement en conséquence qui s'adapte à chaque type de valeur manquante.

La communauté de la fouille de données s'est également intéressée à la problématique des valeurs manquantes. Dans la section qui suit, nous passons en revue les différentes approches proposées dans le cadre de l'extraction des motifs fréquents [Agrawal et Srikant, 1994] en présence de valeurs manquantes ainsi que les approches de complétion à base d'une technique populaire de la fouille de données : les règles d'association.

1.4 Valeurs manquantes en fouille de données

Dans cette section, nous présentons deux grandes familles de travaux de traitement des valeurs manquantes en fouille de données : les travaux faisant l'extraction de motifs en présence de valeurs manquantes et ceux faisant de la complétion à partir de règles d'association.

1.4.1 L'extraction de motifs en présence de valeurs manquantes

Dans le cadre de l'extraction de motifs et de la génération des règles d'association, la présence des valeurs manquantes cause un désagrément. En effet, le problème se pose lors de l'étape de calcul du support d'un motif, étape clé pour le processus d'extraction des règles d'association. Plusieurs travaux ont abordé cette problématique dans le cadre de l'extraction des motifs fréquents [Ragel et Crémilleux, 1998, Kryszkiewicz, 1999, Calders *et al.*, 2007]. En revanche, les travaux qui traitent de l'extraction de représentations concises des motifs fréquents en présence de valeurs manquants sont peu nombreux. Au meilleur de notre connaissance, seuls les travaux [Rioult et Crémilleux, 2003, Rioult et Crémilleux, 2006] en font partie.

La méthode RAR

Dans [Ragel et Crémilleux, 1998], méthode *pionnière* dans le domaine de l'extraction des règles d'association en présence de valeurs manquantes, les auteurs proposent la méthode RAR[2]. Afin de rendre l'extraction des règles d'association robuste aux valeurs manquantes, RAR évalue les motifs dans leurs bases valides. Ainsi, à chaque motif est associé une base valide, qui représente le sous-ensemble d'objets où l'évaluation du support du motif n'est pas sensible à la présence des valeurs manquantes. Par conséquent, la décision de présence d'un motif est uniquement effectuée en fonction de l'information disponible. Pour cela, des modifications des définitions de support et de confiance ont été introduites afin de les adapter aux bases valides. L'inconvénient de cette approche est que l'évaluation du support ne peut se faire que si la base valide d'un motif est un échantillon représentatif de la base totale. Pour cela, les auteurs ont introduit une contrainte de représentativité relative à un motif, où le seuil minimal doit être fixé au préalable. En revanche, la méthode RAR est couplée à un processus MVC (Missing Values Completion) de complétion

[2]L'acronyme RAR désigne Robust Association Rules.

des valeurs manquantes [Ragel et Crémilleux, 1999], où les règles sont filtrées et appliquées selon un vote. Pour cela, différentes mesures ont été utilisées telles que *RI* [Piatetsky-Shapiro, 1991], *J-measure* [Smyth et Goodman, 1992] et *Score-vm* introduite par la méthode MVC.

Nous reviendrons sur cette approche à la sous-section 1.4.2, dédiée aux méthodes de complétion des valeurs manquantes à base de règles d'association, pour décrire en détail le principe de RAR et MVC.

Approche probabiliste de calcul de support

Kryszkiewicz propose de considérer deux stratégies lors de l'évaluation du support d'un motif en présence de valeurs manquantes. Une stratégie dite *optimiste*, pour laquelle le motif est supposé présent et une stratégie dite *pessimiste*, pour laquelle le motif est supposé absent [Kryszkiewicz, 1999]. Lorsqu'il existe un écart important entre la valeur optimiste et la valeur pessimiste, l'auteur suggère d'utiliser la valeur espérée du support, notée *eSup*, et définie en fonction de la valeur optimiste du support d'un motif et de sa négation. Dans [Kryszkiewicz, 2000], une évaluation probabiliste du support est proposée ainsi qu'une redéfinition de la notion du support pour l'adapter à la présence des valeurs manquantes. Pour déterminer le support d'un motif, l'auteur propose de calculer une quantité appelée $probSup_o$ relativement à chaque objet o, où $probSup_o$ est égale à 1 dans le cas où le motif est présent dans l'objet o, elle est égale à 0 si le motif est absent dans l'objet o et égale à $\mu(X_i, v_i)$ dans le cas où X_i présente une valeur manquante. $\mu(X_i, v_i)$ désigne la probabilité d'apparition de la valeur v_i relativement à l'attribut X_i parmi l'ensemble des valeurs observées de l'attribut. Finalement, le support probable du motif X, noté par $probSup(X)$, est égal à $probSup(X) = \sum_{o \in \mathcal{O}} probSup_o(X)$. Cette adaptation du support lui permet de retrouver ses propriétés de monotonie [Kryszkiewicz, 2000].

L'algorithme XMiner

Dans [Calders *et al.*, 2007], les auteurs se sont basés sur la méthode RAR [Ragel et Crémilleux, 1998] pour proposer une solution à l'extraction des motifs fréquents en contournant le problème de l'anti-monotonie du support. Ils ont utilisé les mêmes redéfinitions du support, de la confiance et de la représentativité en ajoutant la notion d'*extensibilité*, afin d'assurer l'extraction des motifs fréquents en présence de valeurs manquantes. Un motif *extensible* est un motif, qui présente au moins un super motif fréquent et représentatif. Cette notion présente deux avantages : elle reste compatible avec le cas classique lorsque les données ne présentent pas de valeurs manquantes, où la représentativité sera égale à 100% et l'extensibilité se ramènera à la notion de fréquence. De plus, l'*extensibilité* est anti-monotone : le super-motif d'un motif inextensible est aussi inextensible, puisque l'extensibilité inclut la contrainte de fréquence. Par conséquent, les motifs inextensibles sont les motifs infréquents ou non représentatifs. L'algorithme XMiner

applique le même principe que l'algorithme ECLAT [Zaki, 2000b], où les motifs extensibles sont extraits en exploitant le critère d'élagage basé sur l'anti-monotonie de l'extensibilité.

Motifs séquentiels et valeurs manquantes

Les motifs séquentiels permettent de découvrir des corrélations entre des motifs respectant une relation temporelle. En présence de valeurs manquantes, l'extraction de ces motifs devient impraticable comme c'est le cas des motifs classiques. Une adaptation de la méthode RAR [Ragel et Crémilleux, 1998] a été proposée dans ce cadre par [Fiot *et al.*, 2007] où la contrainte d'élagage utilisée est la même que celle utilisée dans [Calders *et al.*, 2007] : un motif est élagué s'il est infréquent ou non représentatif. Cette méthode a été également exploitée pour la complétion des valeurs manquantes affectant des données séquentielles, où les motifs extraits sont utilisés pour compléter les séquences manquantes. En cas de conflit lors de la complétion, le choix du motif retenu est effectué selon des indices de pertinence tels que la longueur du motif ou sa fréquence.

Extraction des motifs δ-libres et k-libres en présence de valeurs manquantes

Dans le cadre d'extraction des représentations concises des motifs fréquents en présence de valeurs manquantes, RIOULT et CRÉMILLEUX ont proposé une technique pour les motifs *δ-libres*, qu'ils ont généralisée par la suite aux motifs *k-libres*. La technique proposée dans [Rioult et Crémilleux, 2003] consiste à adapter la notion de presque-fermeture aux bases valides. L'idée clé se matérialise par une stratégie optimiste lors du calcul de la *presque-fermeture* d'un motif X et ceci par une prise en considération des objets désactivés selon X. Les objets désactivés selon un motif X sont les objets qui ont été ignorés temporairement pour obtenir la base valide du motif X. Ces objets sont considérés comme étant des objets contenant X. Il a été démontré qu'avec la nouvelle définition de la presque-fermeture, les motifs *δ-libres* extraits à partir d'un contexte incomplet sont également *δ-libres* dans le contexte complet dont on ne dispose pas en pratique. De même, dans [Rioult, 2005, Rioult et Crémilleux, 2006], un mécanisme de désactivation temporaire des objets incomplets a été proposé. Cette désactivation permet d'extraire les motifs *k-libres* dans une base incomplète. L'originalité de ce travail est la mise en évidence de propriétés caractérisant les données incomplètes mais compatibles avec les propriétés des données complètes. Ainsi, la construction des règles d'association informatives généralisées en présence de valeurs manquantes est rendue praticable en comparant ces propriétés avec celles obtenues dans la base opposée. L'avantage de la base opposée est qu'elle permet d'inverser la présence et l'absence de motifs tout en laissant invariantes les valeurs manquantes.

1.4.2 Complétion des valeurs manquantes à l'aide de règles d'association

Partant du fait que les règles d'association décrivent des relations entre les données, plusieurs approches se proposent d'exploiter ces relations à des fins de complétion, en cherchant les règles concluant sur des valeurs possibles de complétion. Leurs différences concernent essentiellement la manière d'extraire ces règles, la façon dont les valeurs manquantes sont prises en considération et la résolution du problème de conflit lors de la complétion.

La méthode RAR et MVC

Comme nous l'avons présenté dans la sous-section 1.4.1, le principe de RAR et MVC [Ragel et Crémilleux, 1998, Ragel et Crémilleux, 1999] se déroule en deux étapes : une première étape (RAR) consiste à extraire les règles d'association en présence de valeurs manquantes en effectuant un traitement particulier par désactivation temporaire des objets manquants. Ensuite, une deuxième étape (MVC) procède à la complétion des valeurs manquantes en utilisant les règles déjà extraites. Dans ce qui suit, nous décrivons brièvement les idées clés du fonctionnement de RAR et MVC.
La désactivation d'un objet o est définie de la manière suivante :

Définition 3 (Objet désactivé) *[Ragel et Crémilleux, 1998] Un objet o est désactivé pour un motif X, si o contient un item i de X tel que $mv(\mathcal{R})(o,i) = manquant$.*

Le principe de RAR repose sur l'extraction des règles à partir d'une base valide. Une base valide est associée à chaque motif et représente l'ensemble d'objet, où l'évaluation du support est insensible à la présence des valeurs manquantes. Notons par $Des(X, mv(\mathcal{K}))$ l'ensemble des transactions désactivées pour un motif X, la base valide d'un motif X est alors définie par :

Définition 4 (Base valide) $vdb(X) = mv(\mathcal{K}) \backslash Des(X, mv(\mathcal{K}))$.

Ainsi, en considérant la base valide associée à un motif X, le support de X serait égal à :

$$Support(X) = \frac{|mv(\mathcal{K})(X)|}{|vdb(X)|}$$

De même, la confiance d'une règle R : $(X \Rightarrow Y)$ est redéfinie par :

$$Confiance(R) = \frac{|mv(\mathcal{K})(XY)|}{|mv(\mathcal{K})(X)| - |Des(Y, mv(\mathcal{K})) \cap mv(\mathcal{K})(X)|}$$

Cependant, la base valide d'un motif peut parfois ne pas être représentative de la base initiale. Ainsi, afin d'éviter de travailler avec des bases non représentatives, une contrainte de représentativité a été introduite :

Définition 5 (Représentativité) *La représentativité d'une base valide relative à un motif X est :*

$$Représentativité(X) = \frac{|vdb(X)|}{|mv(\mathcal{K})|}$$

De cette façon, l'évaluation du support en présence de valeurs manquantes n'est adéquate que si une base valide est représentative de la base totale. Dans la pratique, les auteurs utilisent un seuil minimal de représentativité (*minRep*) défini par l'utilisateur.

Exemple 1 *Considérons l'exemple d'un contexte mesuré, illustré par la table 1.4. La table 1.5 montre un exemple de calcul du support et de la confiance de la règle $(X_1 = a) \Rightarrow (X_2 = d)$.*
Si nous considérons une valeur $minRep = 25\%$, la base valide relative au motif XY est représentative du contexte initial puisque $Représentativité(XY) = 3/8$ (soit $37,5\%$).

	X_1	X_2	X_3
o_1	a	e	i
o_2	a	c	g
o_3	a	d	h
o_4	?	c	g
o_5	a	?	?
o_6	?	e	i
o_7	b	?	h
o_8	a	?	f

TAB. 1.4 – Contexte mesuré ($mv(\mathcal{K})$).

$R : (X \Rightarrow Y)$	objets désactivés pour X	objets désactivés pour Y	objets désactivés pour XY	support (R)	confiance (R)
$(X_1 = a) \Rightarrow (X_2 = d)$	$\{o_4, o_6\}$	$\{o_5, o_7, o_8\}$	$\{o_4, o_5, o_6, o_7, o_8\}$	1/(8-5)	1/(5-2)

TAB. 1.5 – Exemples de calcul du support et de la confiance d'une règle selon RAR et MVC.

Le principe d'extraction des règles d'association en présence de valeurs manquantes consiste donc à adapter l'algorithme APRIORI [Agrawal et Srikant, 1994], en utilisant les notions de support et confiance redéfinies par RAR. Ensuite, MVC procède à la complétion des valeurs manquantes. Afin de réduire le nombre de règles produites suite à l'extraction des motifs fréquents, *MVC* applique une technique de filtrage selon deux niveaux :

1. au niveau de la production même des règles, en ne conservant que les règles pertinentes, *i.e.*, celles de confiance élevée et celles présentant une corrélation positive selon la mesure *RI* (Rule Interest) [Piatetsky-Shapiro, 1991] sont retenues ;

2. lors de la résolution de conflit en utilisant des métriques telles que *J-measure* [Smyth et Goodman, 1992] et *Score-vm* [Ragel et Crémilleux, 1999], qui permettent d'évaluer la pertinence d'une règle par rapport à une autre lors de la phase de complétion.

Règles d'association pour la prédiction des valeurs manquantes de *Jami et al.*

La méthode proposée dans [Jami *et al.*, 2005] consiste à extraire les règles de prédiction : une règle d'association exacte, dont la conclusion est un intervalle ou un ensemble de valeurs, selon que le domaine de l'attribut sur lequel porte la prédiction est continu ou discret. La complétion s'effectue sur la base d'intersection des conclusions des règles caractérisant l'objet à compléter. Cette approche procède par l'extraction des règles exactes selon l'algorithme APRIORI [Agrawal et Srikant, 1994]. De plus, une nouvelle mesure intitulée *gain de précision* est introduite et permet d'élaguer certaines règles afin de ne retenir que celles qui apportent une précision au niveau des valeurs prédites.

Notons par A_{i0} l'attribut de prédiction et par $mv(\mathcal{K})_{no\_vms}$ le sous-ensemble du contexte $mv(\mathcal{K})$ obtenu en supprimant tout objet présentant au moins une valeur manquante. Une règle de prédiction est définie comme suit :

Définition 6 (Règle de prédiction) *[Jami* et al.*, 2005] On appelle règle de prédiction, toute règle d'association de la forme $T \Rightarrow A_{i0} \in E_T$ où :*
- T est la conjonction de conditions élémentaires de la forme $(A_i = v_i)$, où A_i est un attribut différent de A_{i0} et $v_i \in dom(A_i)$.
- E_T est défini par $E_T = \{v \in dom(A_i) \mid \exists t \in mv(\mathcal{K})_{no\_vms}(t \vDash T \text{ et } t.A_{i0} = v)\}$[3] si $dom(A_i)$ est discret. Si $dom(A_i)$ est continu, alors E_T est défini par $E_T = [\mu_i, \nu_i]$, où $\mu_i = min\{v \in dom(A_{i0}) \mid (\exists t \in mv(\mathcal{K})_{no\_vms})(t \vDash T \text{ et } t.A_{i0} = v)\}$ et $\nu_i = max\{v \in dom(A_{i0}) \mid (\exists t \in mv(\mathcal{K})_{no\_vms})(t \vDash T \text{ et } t.A_{i0} = v)\}$.

Exemple 2 *Reprenons l'exemple du contexte de la table 1.4, supposons que l'attribut X_3 est un attribut de prédiction, la règle $(X_1 = a) \rightarrow X_3 \in \{f, g, h, i\}$ est une règle de prédiction.*

Afin d'éviter d'extraire certaines règles, la mesure *gain de précision* est utilisée et est définie de la manière suivante :

Définition 7 (Gain de précision) *Soient R et R' deux règles de formes respectives $(T \Rightarrow A_{i0} \in E_T)$ et $(T' \Rightarrow A_{i0} \in E_{T'})$, telles que $T \subseteq T'$.*
Le gain de précision de R par rapport à R', noté $gain(R, R')$, est défini comme suit :
- Si $T \neq \emptyset$ alors $gain(R, R') = (|E_T| - |E_{T'}|)/|E_T|$.
- Si $T = \emptyset$ alors $gain(\emptyset, R') = (|dom(A_{i0})| - |E_{T'}|)/|dom(A_{i0})|$.

[3] $t \vDash T$ signifie que le tuple t satisfait les conditions élémentaires de T.

$|E|$ désigne la cardinalité de E si E est un ensemble discret. En revanche, si E est l'intervalle $[\mu, \nu]$, alors $|E|$ désigne la différence $\nu - \mu$. Le *gain de précision* mesure la réduction relative de l'intervalle ou de l'ensemble des valeurs lorsque la prémisse d'une règle est raffinée par l'ajout d'une ou plusieurs conditions de la forme $(A_i = v_i)$. Cette mesure permet de ne retenir une règle que si elle apporte effectivement une réduction de la taille de l'ensemble de valeurs ou de l'intervalle prédit.

Exemple 3 *Reprenons l'exemple de la règle* $R : (X_1 = a) \rightarrow X_3 \in \{f, g, h, i\}$. *Si nous rajoutons la condition* $(X_2 = c)$ *à la prémisse, la règle devient* $R : (X_1 = a) \wedge (X_2 = c) \rightarrow X_3 \in \{g\}$, *nous constatons une réduction de l'ensemble de valeurs prédites qui passe de* $\{f, g, h, i\}$ *à* $\{g\}$. *Ceci indique une réduction de 75% de l'ensemble des valeurs prédites.*

Ainsi, une règle $R : T \Rightarrow A_{i0} \in E_T$ n'est retenue que si pour toutes les règles retenues $R' : T' \Rightarrow A_{i0} \in E_{T'}$ telles que $T' \subseteq T$, le gain de précision de R' par rapport à R est supérieur au seuil de gain fixé au préalable.

Finalement, étant donné un objet présentant une valeur manquante sur un attribut donné, la prédiction s'effectue sur la base de l'intersection des ensembles (ou des intervalles) qui représentent les conclusions des règles de prédiction retenues et vérifiant l'objet en question. Lorsque l'intersection est vide, la complétion est impossible. Lorsque l'intersection est non vide, c'est la règle présentant la prémisse maximale qui sera utilisée pour la complétion, puisque cette approche considère que les règles à prémisse maximale sont celles qui apportent le plus de gain en terme de précision lors de la complétion, en réduisant la taille de l'ensemble ou de l'intervalle des valeurs prédites.

Règles d'association pour la prédiction des valeurs manquantes de *Jen et al.*

L'approche proposée dans [Jen *et al.*, 2009] se base sur l'approche de *Jami et al.* [Jami *et al.*, 2005]. La contribution de *Jen et al.* porte essentiellement sur trois niveaux : (1) la génération des règles de prédiction porte sur une unique valeur ; (2) les valeurs portant sur l'attribut de prédiction sont organisées selon une hiérarchie ; et (3) l'élagage des règles ne se fait plus en exploitant l'anti-monotonie du *gain de précision* mais en considérant les règles à prémisses minimales.

Tout comme *Jami et al.*, *Jen et al.* considèrent un attribut de prédiction fixe, noté A_{i0}. De plus, ils supposent que cet attribut est associé à une hiérarchie, notée $H(A_{i0})$, ayant pour racine *all* et pour feuilles les éléments de $dom(A_{i0})$. Par exemple, la figure 1.1 représente une hiérarchie des médicaments considérée dans [Jen *et al.*, 2009].

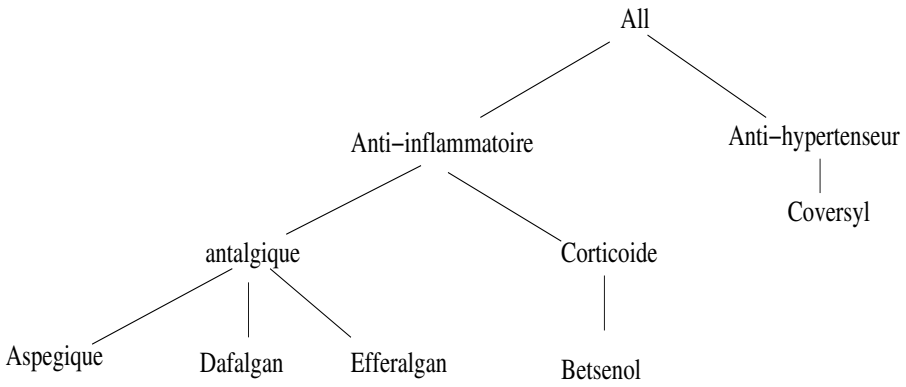


FIG. 1.1 – Exemple d'une hiérarchie des médicaments considérée par *Jen et al.*

Une règle de prédiction selon *Jen et al.* est définie comme suit :

Définition 8 (Règle de prédiction) *[Jen* et al.*, 2009] Une règle de prédiction est une règle de la forme $T \Rightarrow A_{i0} \in E_T$ où T est la conjonction de conditions élémentaires de la forme $(A_i = v_i)$, où A_i est un attribut $\neq A_{i0}$, tels que $v_i \in dom(A_i)$ et $E_T \in H(A_{i0})$ avec $E_T \neq All$.*

Exemple 4 *Si nous reprenons le cas de la règle $(X_1 = a) \rightarrow X_3 \in \{f, g, h, i\}$. Considérons que les valeurs f, g, h et i correspondent respectivement aux valeurs Aspegique, Dafalgan, Efferalgan et Betnesol.*
La règle serait équivalente à $(X_1 = a) \rightarrow X_3 \in \{Aspegique, Dafalgan, Efferalgan, Betnesol\}$. En considérant l'hiérarchie de ces valeurs représentée par la figure 1.1, la règle devient alors $(X_1 = a) \rightarrow X_3 =$anti-inflammatoire.

Cette approche extrait les règles exactes selon l'algorithme APRIORI et applique un élagage selon la notion de *restriction* d'une règle par rapport à une autre. Soient $R : T \Rightarrow A_{i0} \in E_T$ et $R' : T' \Rightarrow A_{i0} \in E'_T$, si toute condition élémentaire de T est également une condition élémentaire de T', on dit que T est une *restriction* de T'. De plus, supposons que $E_T = E'_T$, c'est à dire que les deux règles ont la même conclusion, alors la règle R' est redondante et ne sera donc pas retenue pour la prédiction. Par conséquent, à la différence de l'approche de *Jami et al.*, la redondance des règles est évitée en considérant les règles à prémisse minimales.

Au final, la prédiction s'effectue selon le principe suivant : pour chaque objet, on examine les règles retenues et vérifiant l'objet, en considérant la valeur de prédiction constituant l'ancêtre commun des valeurs dans $H(A_{i0})$. Si l'ancêtre commun est la valeur *All*, alors aucune prédiction n'est possible.

Exemple 5 *Supposons que les deux règles suivantes sont retenues pour la prédiction d'un objet donné : R : $(X_1 = a) \rightarrow X_3 =$Anti-inflammatoire et R' : $(X_2 = e) \rightarrow X_3 = Betsenol$. La valeur de prédiction de l'attribut X_3 sur cet objet serait "Anti-inflamatoire".*

La méthode $\mathcal{GBAR}_{\mathcal{MVC}}$

Dans [Ben Othman et Ben Yahia, 2011], nous avons proposé l'approche $\mathcal{GBAR}_{\mathcal{MVC}}$[4]. Cette approche se démarque des autres approches de complétion des valeurs manquantes par l'utilisation d'une base de générique de règles d'association. Le principe d'extraction de ces règles repose sur la redéfinition de la notion de presque-fermeture proposée dans [Rioult et Crémilleux, 2003], que nous avons décrite à la section 1.4. Ceci a permis de prendre en considération la présence des valeurs manquantes lors de l'extraction des règles par désactivation temporaire des objets incomplets. Par conséquent, les règles employées lors de la complétion ont le mérite d'être les plus fiables possibles vis-à-vis de l'aspect incomplet des données. De plus, ces règles sont exemptes de redondance. En effet, nous avons montré que ces règles sont les plus intéressantes, *i.e.,* présentent le minimum de contraintes à satisfaire lors de la complétion. Ceci est garanti par la minimalité des prémisses constituées de motifs libres (cf. section 3.1.1 - page 42). Ainsi, moyennant l'utilisation d'une base générique, les conflits entre règles sont considérablement réduits, ce qui assure plus d'efficacité lors de la complétion. De plus, une nouvelle métrique intitulée *Robustesse* a été introduite et permettant d'évaluer la pertinence d'une règle par rapport à une autre en cas de conflit.

La fermeture d'un motif X, notée $h(X)$, rassemble les items i à l'intersection de tous les objets contenant X. En présence de valeurs manquantes, nous ne parlons plus en termes de fermeture mais plutôt en termes de pseudo-fermeture, définie comme suit :

Définition 9 (Pseudo-fermeture) *La pseudo-fermeture d'un motif X dans un contexte incomplet $mv(\mathcal{K})$, notée $\mathcal{PF}(X)$, est définie comme suit :*

$$\mathcal{PF}(X) = X \cup \{i \,|\, i \in \mathcal{I} \ \wedge \ Supp(X) - Supp(Xi) = |mv(\mathcal{K})(X) \cap Des(i, mv(\mathcal{K}))| \}.$$

Pour calculer correctement les items i figurant dans la fermeture d'un motif X dans le cas de données incomplètes, il est nécessaire de distinguer les objets qui présentent une valeur manquante sur ces items. Ainsi, il faut prendre en compte les objets désactivés relativement à i pour tester si $i \in \mathcal{PF}(X)$.

Exemple 6 *Considérons l'exemple du contexte de la table 1.4 (page 24), nous avons $Supp(ei) = 2$ mais $Supp(aei) = 1$, donc a priori, $a \notin \mathcal{AC}(ef)$. Toutefois, en considérant la redéfinition de la Pseudo-fermeture (Définition 9), l'égalité suivante est vérifiée :*
$Supp(ei) - Supp(aei) = |mv(\mathcal{K})(ei) \cap Des(a, mv(\mathcal{K}))|$. Par conséquent, $a \in \mathcal{PF}(ei)$.

En s'appuyant sur la redéfinition de la pseudo-fermeture, il devient possible d'extraire les motifs fermés fréquents à partir d'un contexte présentant des valeurs manquantes. De plus, nous

[4] L'acronyme $\mathcal{GBAR}_{\mathcal{MVC}}$ désigne Generic Basis of Association Rules for Missing Values Completion.

avons adapté la notion de base générique des règles d'association exactes introduite dans [Bastide, 2000, Bastide *et al.*, 2000] au cas d'un contexte incomplet. Cette adaptation est définie comme suit :

Définition 10 (Base générique de règles d'association pseudo-exactes) *Soit $\mathcal{PFF}$ l'ensemble des motifs fermés fréquents extraits à partir d'un contexte incomplet $mv(\mathcal{K})$. Pour chaque motif fermé fréquent c, notons par $\mathcal{GM}_c$ l'ensemble des motifs libres associés à c. La base générique des règles d'association pseudo-exactes est définie comme suit :*

$$\mathcal{BG} = \{R : g \Rightarrow (c - g) \mid c \in \mathcal{PFF},\ g \in \mathcal{GM}_c \text{ et } g \neq c^{(5)}\}.$$

Au final, la complétion des valeurs manquantes se fait selon le principe classique des approches de complétion à base de règles d'association : pour chaque valeur manquante, nous examinons les règles concluant sur des valeurs de complétion possibles de l'attribut manquant. Parfois, les règles peuvent indiquer plusieurs valeurs de complétion à la fois. Dans de tels cas, nous appliquons un vote selon une métrique intitulée *Robustesse*. Cette métrique évalue la capacité d'une règle à compléter une valeur manquante. Il s'agit de prendre en considération le degré de corrélation entre la prémisse et la conclusion d'une règle, grâce à la mesure *lift* [Brin *et al.*, 1997]. De plus, cette mesure tient compte du degré de correspondance d'une règle avec une transaction manquante moyennant la notion de *Correspondance*. Les définitions de ces différentes métriques sont comme suit :

Définition 11 *La mesure **Correspondance** d'une règle R : prémisse $\Rightarrow (X_i, v_i)$ avec un objet o présentant une valeur manquante sur l'attribut X_i, est définie comme suit :*

$$Correpondance(R, o, X_i) = \begin{cases} 0 & \text{si } R \text{ ne vérifie pas } o \\ \frac{\sum_{j=1..n} vérifie(X_j, v_j)}{nombre\ d'attributs} telque\ i \neq j & \text{sinon.} \end{cases}$$

où

$$vérifie(X_j, v_j) = \begin{cases} 0 & \text{si } X_j \text{ présente une valeur manquante dans } o \\ 1 & \text{sinon.} \end{cases}$$

Exemple 7 *Considérons un objet o tel que $o : (X_1, ?)(X_2, C)(X_3, F)(X_4, K)$. La règle $R_1(X_2, D) \wedge (X_3, F) \Rightarrow (X_1, A)$ ne vérifie pas l'objet o, puisque la valeur de l'attribut X_2 est C dans o. Par conséquent, $Correpondance(R_1, o, X_1) = 0$. En revanche, si nous considérons une règle $R_2 : (X_2, C) \wedge (X_3, F) \Rightarrow (X_1, B)$, nous avons $Correspondance(R_2, o, X_1) = \frac{1}{2}$.*

La mesure *Correspondance* s'avère utile dans le cas des valeurs manquantes. En effet, une règle peut correspondre à un objet même si cet objet présente des valeurs manquantes sur la prémisse de la règle. Dans ce cas, la pertinence de la règle est incertaine. La *Correspondance*

[5] La condition $g \neq c$ permet de ne pas retenir les règles de la forme $g \Rightarrow \emptyset$.

permettra ainsi de tenir compte des valeurs manquantes dans la prémisse de la règle. L'idée est de privilégier la règle présentant une plus grande correspondance avec l'objet en question. La deuxième mesure employée par $\mathcal{GBAR}_{\mathcal{MVC}}$ est la mesure *lift* introduite dans [Brin *et al.*, 1997]. Elle permet de mesurer la corrélation induite entre la prémisse X et la conclusion Y d'une règle $R : X \Rightarrow Y$ et d'assurer que la présence de X favorise la présence de Y.

Définition 12 *[Brin* et al.*, 1997] Le lift d'une règle $R : (X \Rightarrow Y)$, notée $Lift(R)$, est calculé de la manière suivante :*

$$Lift(R) = \frac{Supp(XY)}{Supp(X) \cdot Supp(Y)}.$$

Si $Lift(R) = 1$, alors X et Y sont dits indépendants. Si $Lift(R) < 1$, X et Y sont dits négativement corrélés. Si $Lift(R) > 1$, alors X et Y sont dits positivement corrélés [Brin et al.*, 1997].*

Ainsi, lors de la complétion des valeurs manquantes, $\mathcal{GBAR}_{\mathcal{MVC}}$ utilise la règle qui maximise à la fois le critère *Correspondance* et la mesure *lift* moyennant la mesure *Robustesse* :

Définition 13 *La Robustesse d'une règle R à compléter un objet o présentant une valeur manquante sur l'attribut X_i est définie comme suit :*

$$Robustesse(R, o, X_i) = Correspondance(R, o, X_i) \times lift(R).$$

D'autres travaux de complétion à base de règles d'association ont été proposés, tels que [Wu *et al.*, 2004, Wu *et al.*, 2008, Bashir *et al.*, 2009]. Par exemple, dans le premier travail, les auteurs font l'extraction de règles à partir des motifs fréquents sans une omission totale des objets incomplets. Cependant, toute valeur manquante est ignorée, *i.e.,* considérée inexistante. Dans le second travail, les auteurs emploient le même principe des approches de complétion à base de règles en appliquant un vote. À la différence du vote classique qui se fait par règle [Ragel et Crémilleux, 1999], le vote selon cette approche se fait par groupe de règles. Dans le dernier travail, les auteurs proposent une méthode hybride : l'idée consiste à combiner une approche à base de règle et la technique des plus proches voisins. Cette dernière technique est employée lorsque les corrélations dans les données sont insuffisantes pour compléter les valeurs qui manquent. Un autre travail de complétion à base de motifs fréquents a été proposé dans [Vreeken et Siebes, 2008]. L'originalité de ce travail est qu'une "*bonne*" complétion ne se mesure pas nécessairement en termes de précision. Les auteurs utilisent un algorithme de compression des données intitulé MDL (Minimum Description Length) [A. Siebes et van Leeuwen, 2006] et considèrent que la meilleure complétion est celle qui compresse au maximum les données, où seul l'écart entre les supports d'un motif dans la base d'origine et celle complétée permet d'évaluer la pertinence de la complétion.

1.5 Discussion et positionnement

À la lumière de cette étude bibliographique, nous distinguons trois grandes catégories de méthodes dédiées au traitement des valeurs manquantes :

- **Suppression** : cette première catégorie regroupe les méthodes, qui procèdent à une suppression "*brutale*" des données incomplètes.
- **Ajustement** : la deuxième catégorie regroupe les méthodes, qui font de l'ajustement d'algorithmes déjà existants pour les adapter aux valeurs manquantes.
- **Complétion** : cette dernière catégorie rassemble les méthodes de pré-traitement des valeurs manquantes moyennant une technique de complétion.

La table 1.6 montre les différentes approches décrites dans ce chapitre, regroupées selon la classification citée précédemment. Nous constatons que seules les approches « listwise » et « pairwise deletion » avec les approches proposées dans [Jami *et al.*, 2005, Jen *et al.*, 2009] font de la suppression *brutale* des données incomplètes, catégorie des méthodes connues pour n'être applicables que lorsque les valeurs manquantes sont peu nombreuses. Sinon, la perte de données serait lourde de conséquence. Cependant, les deux dernières approches de cette première catégorie, appliquent en plus une méthode de complétion.

Nous remarquons aussi que les méthodes de traitement des valeurs manquantes tendent vers l'ajustement de méthodes déjà existantes. Ceci s'illustre par la diversité des méthodes faisant partie de cette catégorie et montre que les chercheurs sont conscients des difficultés liées à la complétion des valeurs manquantes. Bien que les méthodes d'ajustement conçoivent de nouveaux algorithmes adaptés à la présence des valeurs manquantes, ces méthodes présentent certaines limites : (i) elles ne sont adaptées qu'à la tâche pour laquelle elles sont conçues (extraction de règles d'association, de motifs fréquents) ; (ii) elles ne fournissent pas de données complètes qui puissent être exploitables par n'importe quelle autre méthode d'analyse de données ou d'extraction de connaissances. Ce dernier point constitue incontestablement l'intérêt des méthodes de complétion. Finalement, nous estimons que les méthodes procédant à la fois à un ajustement et à une complétion sont à recommander : elles prennent en considération les valeurs manquantes tout en fournissant des données complètes à d'autres méthodes d'analyse. Malgré tout, l'écueil majeur de ces méthodes est l'absence totale de prise en considération des modèles d'apparition des valeurs manquantes.

Méthode	**Suppression**	**Ajustement**	**Complétion**
Technique de suppression (listwise et pairwise)	×		
Moyenne - Mode - Médiane			×
EM [Dempster *et al.*, 1977]			×
Imputation multiple [Rubin, 1978]			×
Technique de régression			×
« valeur spéciale » [Date, 1995]			×
« A-mark » et « I-mark » [Codd, 1990]			×
C4.5 [Quinlan, 1993]		×	×
Delavallde et Dong [Delavallade et Dang, 2007]		×	×
Lobo et Numao [Lobo et Numao, 1999]			×
Hawarah et al. [Hawarah *et al.*, 2004, Hawarah *et al.*, 2006]		×	×
RAR et MVC [Ragel et Crémilleux, 1998, Ragel et Crémilleux, 1999]		×	×
Kryszkiewiez [Kryszkiewicz, 1999, Kryszkiewicz, 2000]		×	
Xminer [Calders *et al.*, 2007]		×	
Fiot et al. [Fiot *et al.*, 2007]		×	×
Rioult et Crémilleux [Rioult et Crémilleux, 2003]		×	
Rioult et Crémilleux [Rioult et Crémilleux, 2006]		×	
Closet Fit [Grzymala-Busse *et al.*, 1999]			×
RSFit [Li et Cercone, 2006]			×
ItemRSFit [Li *et al.*, 2007]			×
Grzymala et Hu [Grzymala-Busse et Hu, 2001]			×
Jami et al. [Jami *et al.*, 2005]	×		×
Jen et al. [Jen *et al.*, 2009]	×		×
Shen et al. [Shen *et al.*, 2007]		×	×
$\mathcal{GBAR}_{\mathcal{MVC}}$ [Ben Othman et Ben Yahia, 2011]		×	×

Tab. 1.6 – Classification des méthodes de la littérature présentées dans ce chapitre.

Pour notre part, nous proposons dans les chapitres suivants une méthode, qui se démarque par :

1. Une prise en considération préalable du modèle d'apparition des valeurs manquantes. Comme nous l'avons mentionné dans la section 1.2, les travaux de la littérature sont conscients de l'importance d'une telle modélisation mais aucune contribution ne s'est orientée vers cet axe. On peut néanmoins citer le travail introduit dans [Grzymala-Busse, 2004] (page 19), qui présente une catégorisation des valeurs manquantes et propose un traitement en conséquence. L'originalité de ce travail est précisément cette adaptation du traitement. À

l'inverse de cette proposition, où la catégorisation est formulée selon une hypothèse, notre contribution portera d'une part sur la catégorisation des valeurs manquantes et d'autre part sur l'adaptation du traitement de complétion en conséquence.

2. Une complétion qui ne cherche pas à trouver nécessairement la valeur réelle d'une valeur manquante. Parmi les travaux de la littérature qui ont adopté cette stratégie, citons les travaux de [Delavallade et Dang, 2007, Vreeken et Siebes, 2008]. En effet, selon l'objectif que l'on souhaite atteindre, la complétion des valeurs manquantes ne se mesure pas nécessairement par la précision, *i.e.*, la capacité d'une complétion à compléter correctement les valeurs manquantes.
3. Une complétion par une « valeur spéciale ». Pareillement, les travaux qui ont adopté une telle complétion en ajoutant « valeur spéciale » par attribut manquant sont [Date, 1995, Grzymala-Busse et Hu, 2001]. De même, le travail de [Codd, 1990] a consisté à ajouter deux catégories de « valeur spéciale » : valeur manquante applicable et valeur manquante non applicable. L'originalité de notre contribution est qu'une « valeur spéciale », renfermant intrinsèquement l'origine de la valeur qui manque, sera employée de façon contextualisée et dépendra de l'objet complété.

1.6 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art des principales méthodes dédiées au traitement des valeurs manquantes. Nous nous sommes d'abord intéressé aux modèles d'apparition des valeurs manquantes. Ensuite, nous avons présenté des travaux faisant partie du domaine de la statistique, des bases de données, de l'apprentissage supervisé ainsi que des ensembles approximatifs. Cette étude bibliographique montre l'importance de cette problématique d'une part et la richesse des contributions d'autre part. Les travaux auxquels nous nous sommes intéressé de près sont ceux de la fouille de données, où nous avons distingué deux grandes approches : les travaux faisant l'extraction de motifs en présence de valeurs manquantes et ceux faisant de la complétion à partir de règles d'association. Finalement, nous avons dressé un tableau comparatif de ces différents travaux, à partir duquel nous avons positionné notre contribution relative à la problématique des valeurs manquantes.

Chapitre 2

Techniques d'évaluation de méthodes de complétion

Sommaire

Dans la littérature, la complétion des valeurs manquantes est généralement évaluée selon deux méthodes : mesurer la proximité des données complétées avec celles de référence ou analyser l'impact du traitement sur des techniques d'apprentissage supervisé ou non supervisé. Dans ce chapitre, nous présentons ces deux techniques d'évaluation. Ensuite, nous nous intéressons aux protocoles expérimentaux, généralement associés à ces techniques d'évaluation. Nous mènerons également une discussion et nous étudierons l'adéquation entre les techniques et les protocoles d'évaluation.

2.1 Techniques d'évaluation

2.1.1 Mesure de la proximité des données de référence

Le principe des techniques d'évaluation mesurant la proximité entre les données de référence et les données complétées est représenté par la figure 2.1. Cette figure montre que les données incomplètes sont issues d'un contexte complet dit de *référence*, indisponible dans les situations réelles. Cependant, lors de protocoles expérimentaux, on peut générer un contexte incomplet à

partir d'un contexte complet en introduisant artificiellement des valeurs manquantes. Considérant que la complétion doit s'approcher des données de *référence*, l'évaluation consiste donc à mesurer une distance entre les deux contextes [Ragel et Crémilleux, 1999, Wu *et al.*, 2004, Jami *et al.*, 2005, Jen *et al.*, 2009, Ben Othman et Ben Yahia, 2011].

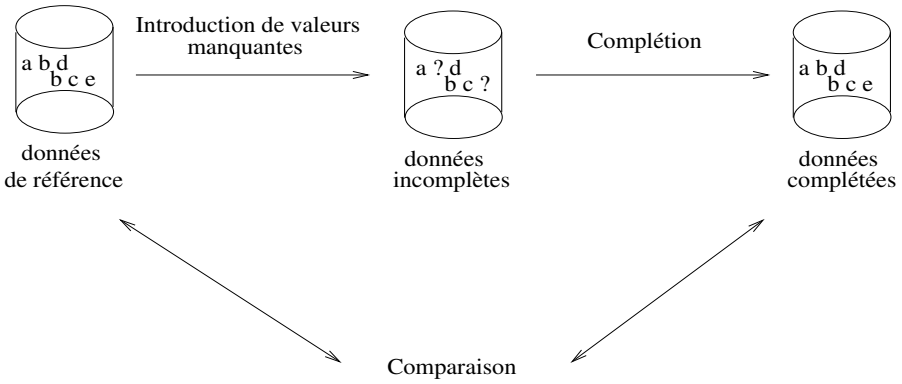


FIG. 2.1 – Évaluation selon la proximité des données de référence.

Parmi les métriques qui permettent de calculer cette distance, on trouve la *précision* d'une complétion qui mesure la proportion de valeurs manquantes correctement complétées par rapport au nombre de valeurs manquantes complétées. Cette métrique est utilisée par la majorité des travaux [Ragel et Crémilleux, 1999, Jami *et al.*, 2005, Ben Othman et Ben Yahia, 2008, Jen *et al.*, 2009]. L'objectif de l'évaluation est alors d'étudier la variation de la *précision* en fonction du pourcentage des valeurs manquantes. Pour compléter cette évaluation, ces travaux emploient également une métrique qui mesure l'exhaustivité de la complétion, *i.e.,* la proportion des valeurs manquantes complétées parmi le nombre total des valeurs manquantes. On considère alors qu'une méthode est efficace si elle arrive à compléter les valeurs manquantes de façon précise et exhaustive. Au final, ces méthodes cherchent à avoir un compromis entre la précision et l'exhaustivité lors de la complétion.

2.1.2 Impact selon des techniques d'apprentissage supervisé

L'autre motivation pour la complétion est de proposer des données complètes aux algorithmes d'apprentissage qui le requièrent. La qualité de la complétion est alors mesurée selon son impact sur les résultats des techniques d'apprentissage supervisé [Bastia et Monard, 2003, A. Farhangfar, 2004, Acuna et Rodriguez, 2004, Delavallade et Dang, 2007]. Dans ce cas, l'évaluation a pour objectif de montrer qu'une "*bonne*" méthode de complétion des valeurs manquantes permet de maximiser les performances d'un classifieur.
À la figure 2.2, nous représentons le schéma général sur lequel se basent les travaux faisant partie d'une telle stratégie d'évaluation.

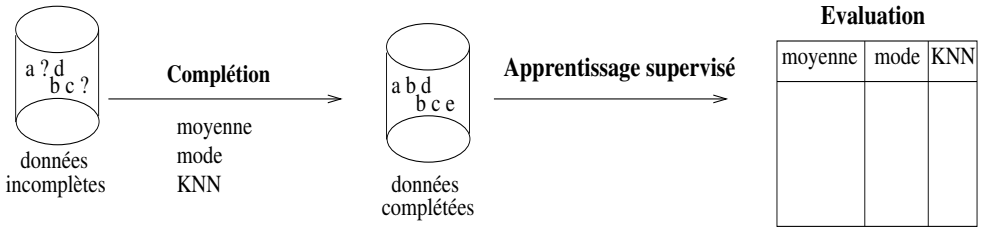


FIG. 2.2 – Évaluation selon des techniques d'apprentissage supervisé.

Cette technique d'évaluation consiste à appliquer une méthode de complétion à partir des données incomplètes. Afin d'évaluer cette dernière, on applique une technique de classification supervisée dont le score, obtenu par validation croisée, permet de discuter la pertinence de la méthode de complétion. L'objectif est d'étudier l'impact d'une méthode de complétion des valeurs manquantes sur la performance d'un classifieur et de tirer des recommandations générales permettant d'évaluer l'efficacité de ce traitement. Par la suite, une discussion est menée en comparant les performances d'une complétion par rapport à d'autres techniques de complétion. En revanche, l'adéquation du classifieur employé lors de l'évaluation n'a jamais été abordée. Bien qu'il existe une littérature très riche qui traite de l'évaluation des méthodes de complétion selon une technique supervisée, il nous semble toujours difficile de dégager une conclusion précise. Ceci est dû à quelques disparités, notamment en ce qui concerne la technique employée pour l'évaluation, la façon dont les données de référence sont obtenues, *etc.* Nous invitons le lecteur à se référer à [Delavallade, 2007] pour un état de l'art concernant l'évaluation par des techniques supervisées, basé sur une taxonomie selon plusieurs critères.

2.2 Protocoles expérimentaux d'évaluation

Après avoir présenté les techniques d'évaluation d'une méthode de complétion des valeurs manquantes, nous détaillons maintenant les protocoles expérimentaux d'évaluation associés. Nous discuterons également l'adéquation de chaque protocole aux techniques d'évaluation présentées ci-dessus. La figure 2.3 montre deux protocoles expérimentaux : le premier s'inscrit dans le cadre des situations réelles, où les données sont réellement incomplètes, *i.e.,* nous ne disposons pas des données de référence. Ce protocole consiste par conséquent à travailler sur les données incomplètes, sur lesquelles on applique une méthode de complétion ensuite une technique d'évaluation. Le deuxième protocole est employé dans un cadre où les valeurs manquantes sont artificiellement introduites dans les données, une méthode de complétion et une technique d'évaluation sont ensuite appliquées successivement.

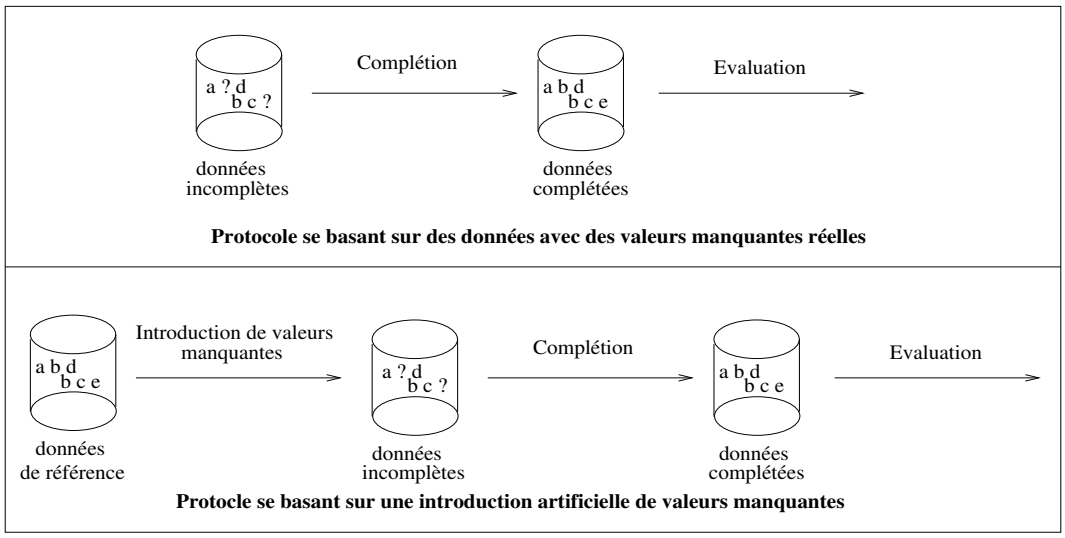


FIG. 2.3 – Protocoles expérimentaux d'évaluation.

2.3 Discussion et positionnement

L'étude bibliographique, que nous avons menée, nous a permis de séparer les techniques d'évaluation d'une méthode de complétion des valeurs manquantes en deux classes, ainsi que les protocoles expérimentaux associés. Cependant, la question qui se pose nécessairement est : *quelle technique d'évaluation faut-il employer avec quel protocole expérimental ?* La table 2.1 indique l'adéquation (représentée par le symbole $\surd$) entres les deux techniques d'évaluation et les protocoles expérimentaux. Cette table montre que la technique mesurant la proximité des données incomplètes aux données de référence ne peut être employée qu'avec le protocole se basant sur une introduction artificielle de valeurs manquantes, car elle nécessite que les données de référence soient disponibles, ce qui est impossible avec le protocole où les données de référence sont indisponibles. En revanche, la technique qui consiste à étudier l'impact selon l'apprentissage supervisé peut-être employée avec les deux protocoles expérimentaux.

Par ailleurs, il est important de signaler que l'évaluation de la complétion selon la proximité avec les données de référence n'est possible que lorsque les valeurs utilisées par la complétion appartiennent au domaine de définition de l'attribut manquant. Ce n'est pas le cas pour les complétions qui utilisent de nouvelles valeurs : elles ne peuvent être évaluées par proximité des données de référence.

Quant à l'évaluation selon l'impact sur des techniques supervisées, la littérature montre que la majorité des approches emploie exclusivement le modèle *MCAR* lors de l'introduction des valeurs manquantes. Ce choix est justifié par le fait que c'est le modèle le plus simple lorsque l'on veut faire de l'introduction artificielle de valeurs manquantes. Nous montrerons dans le chapitre 4 que ce modèle est en pratique le moins réaliste.

Technique \ Protocole	**Protocole se basant sur des données avec valeurs manquantes réelles**	**Protocole se basant sur une introduction artificielle de valeurs manquantes**
Proximité des données de référence		√
Impact selon une technique d'apprentissage supervisé	√	√

TAB. 2.1 – Adéquation entre les techniques d'évaluation et les protocoles expérimentaux d'évaluation.

2.4 Conclusion

Dans ce chapitre, deux techniques d'évaluation de méthodes de complétion ont été décrites. La première consiste à mesurer la proximité entre les données de référence et les données complétées. Elle s'intéresse aux objectifs suivants :

- la méthode offre-t'elle une complétion fiable des valeurs manquantes ?
- permet-elle toujours la complétion de ces valeurs ?

Quant à la deuxième technique, qui mesure l'impact de la complétion sur le résultat d'une méthode d'apprentissage supervisé, elle cherche à montrer qu'une méthode de complétion des valeurs manquantes doit maximiser les performances d'un classifieur. Bien que cette technique soit largement utilisée dans la communauté « fouille de données », nous présenterons dans le chapitre 7, des expériences montrant que cette question d'amélioration des performances d'un classifieur est loin d'être simple à aborder. Nous montrerons en effet que cette technique d'évaluation doit être employée avec beaucoup de précautions. Cependant, nous envisageons de revoir le protocole expérimental en incorporant les éléments suivants :

- ne pas se limiter aux valeurs aléatoires lors de l'introduction artificielle des valeurs manquantes ;
- par ailleurs, nous mettons en place une nouvelle technique d'évaluation basée sur la stabilité d'une méthode de clustering détaillée dans le chapitre 7.

Chapitre 3

Règles de caractérisation non redondantes

Sommaire

Dans ce chapitre, nous commençons par présenter la notion de représentation condensée de motifs fréquents [Agrawal et Srikant, 1994], ensuite nous rappelons le concept de redondance, et de couverture des règles d'association exactes démunie de redondance, où nous nous intéresserons particulièrement aux règles à prémisses minimales. Dans ce cadre, nous présentons un état de l'art relatif aux couvertures des règles exactes. En effet, dans le chapitre 4, nous utiliserons une caractérisation non redondante des valeurs manquantes en se basant précisément sur l'une de ces couvertures de règles exactes. Nous menons également ici une discussion et une étude expérimentale comparative de ces différentes couvertures de règles.

3.1 Découverte d'association non redondantes

La tendance des algorithmes d'extraction de motifs s'est rapidement orientée vers des représentations condensées de motifs fréquents [Pasquier *et al.*, 1999b, Boulicaut *et al.*, 2000, Bykowski et Rigotti, 2001, Zaki et Hsiao, 2002, Calders et Goethals, 2003]. Dans ce qui suit, nous présentons cette notion.

3.1.1 Représentation condensée de motifs fréquents

Le nombre de motifs fréquents qui peut-être extrait à partir d'un contexte est exponentiel en fonction du nombre d'items de la base [Pasquier *et al.*, 1999b]. La concision ou la condensation de motifs fréquents signifie que l'on s'intéressera seulement à un résumé de ces motifs. Les motifs couramment utilisés pour construire une représentation condensées sont les *motifs libres* et les *motifs fermés*.

Les motifs fermés ont une propriété intéressante, à savoir la structuration de l'espace de recherche en classes d'équivalence des supports. Une classe d'équivalence regroupe un ensemble de motifs ayant un même support et une même fermeture, où la fermeture est définie par l'ensemble des items qui sont toujours présents avec un motif X, qu'on notera par $h(X)$. Une classe d'équivalence regroupe donc des motifs présents dans les mêmes objets. Au sein de chaque classe d'équivalence, les motifs fermés (les éléments maximaux d'une classe d'équivalence) constituent une représentation condensée des motifs fréquents. En effet, le support d'un motif quelconque peut-être déduit grâce au support du motif fermé appartenant à la même classe d'équivalence. Ainsi, en se limitant aux motifs fermés, on obtient une représentation condensée des motifs fréquents qui peuvent être résumés par les motifs fermés.

De façon similaire, les motifs minimaux des classes d'équivalence constituent une représentation condensée des motifs fréquents. Ces motifs minimaux sont appelés motifs *libres* [Boulicaut *et al.*, 2003, Calders et Goethals, 2003], *générateurs minimaux* [Bastide *et al.*, 2000] ou encore motifs *clés* [Pasquier *et al.*, 1999a]. Un motif X est libre si et seulement si pour tout motif Y tel que $Y \subsetneq X, Supp(Y) > Supp(X)$. La liberté d'un motif exprime le fait que son support a strictement décru par rapport à ceux de ses sous-ensembles. Ainsi, la liberté est une contrainte anti-monotone [Boulicaut *et al.*, 2000]. Il est donc très simple d'extraire par niveaux des motifs libres (et fréquents) [Pasquier *et al.*, 1999a, Stumme *et al.*, 2002, Hamrouni *et al.*, 2005]. Tout comme les motifs fermés, les motifs libres permettent de déterminer le support de n'importe quel autre motif : le support d'un motif quelconque est égal au support du plus grand motif libre qu'il contient.
De plus, la minimalité des motifs libres exprime en leur sein l'absence de corrélations. Cette propriété place les motifs libres au cœur du calcul des règles d'association non redondantes.

De nombreuses autres représentations condensées existent en littérature. Citons par exemple les motifs *δ-libres* [Boulicaut *et al.*, 2000], les motifs libres disjonctifs [Bykowski et Rigotti, 2001], les motifs non dérivables [Calders et Goethals, 2002], les motifs *k-libres* [Calders et Goethals, 2003], ou encore les motifs essentiels [Casali *et al.*, 2005, Hamrouni *et al.*, 2007].

Les motifs libres et fermés ont de bonnes propriétés pour le calcul des règles d'association non redondantes. Nous exposons un état de l'art sur cette problématique, qui constitue le cœur de notre travail en parallèle avec le traitement des valeurs manquantes. Le chapitre 5 présentera notre contribution en matière de calcul des règles non redondantes.

La technique de génération des règles d'association à partir des motifs fréquents souffre d'un inconvénient majeur : la quantité de règles produites. Cette quantité est si importante que leur exploitation devient difficile. D'autant plus que parmi cette quantité de règles, de nombreuses sont redondantes, du point de vue de l'information véhiculée. Ainsi, plusieurs travaux se sont intéressés à la sélection d'un sous-ensemble de règles (les plus pertinentes), appelé *couverture* de règles. Les approches proposées dans le cadre des couvertures des règles d'association utilisent deux techniques : réduction orientée utilisateur ou réduction structurelle [Pasquier, 2000]. La première cherche à extraire un sous-ensemble de règles valides du point de vue d'une contrainte spécifiée par l'utilisateur. La deuxième s'intéresse à la sélection d'un sous-ensemble de règles ayant certaines propriétés structurelles. Dans ce chapitre, nous nous intéressons précisément à la réduction de règles par couverture structurelle. Plusieurs définitions de couvertures ont été proposées dans la littérature [Kryszkiewicz, 1998, Bastide *et al.*, 2000, Zaki, 2000a, Gasmi *et al.*, 2006]. Ces couvertures se basent sur l'élimination des règles redondantes.

3.1.2 Notion de redondance

Pour présenter brièvement l'intuition derrière la notion de redondance, reprenons l'exemple du contexte réel (Définition 1 - page 10) représenté par la table 3.1.
Toutes les règles d'association suivantes ont une confiance égale à 1 :

- $R_1 : d \rightarrow i$
- $R_2 : dg \rightarrow i$
- $R_3 : g \rightarrow i$
- $R_4 : g \rightarrow di$

Nous remarquons que les règles R_2 et R_3 sont redondantes. En effet, la règle R_2 présente un item de plus en prémisse (contrainte supplémentaire) par rapport à la règle R_1 pour obtenir une même conclusion. Cependant, l'item g n'est pas nécessaire, l'item d seul permettant de conclure sur l'item i. Quant à la règle R_3, elle est constituée de la même prémisse que la règle R_4, mais conclut sur moins d'items, donc véhicule moins d'information. Ainsi, les règles R_1 et R_4 sont les plus pertinentes, et elles permettent de générer les autres règles (via un mécanisme d'inférence).

	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		×		×			×		×

TAB. 3.1 – Exemple d'un contexte réel $\mathcal{K}$.

3.1.3 Couvertures des règles d'association

L'idée générale des couvertures des règles d'association est de produire un nombre minimal de règles tout en véhiculant le maximum d'information. Les couvertures de règles ont pour objectif, d'une part de rendre l'extraction possible, même dans le cadre de données très corrélées. D'autre part, elles permettent de se focaliser sur les règles les plus intéressantes. Les couvertures de règles les plus connues sont : la base Guigues-Duquenne [Guigues et Duquennes, 1986] pour les implications globales, la base propre de Luxenburger [Luxenburger, 1991] pour les implications partielles, adaptées au contexte des règles d'association exactes et approximatives [Pasquier, 2000]. Parmi les couvertures de règles figurent également les règles représentatives [Kryszkiewicz, 1998], la base représentative [Phan-Luong, 2001], la base générique des règles exactes ($\mathcal{GBE}$) et la base génériques des règles approximatives ($\mathcal{GBA}$) [Bastide *et al.*, 2000], les règles d'association non redondantes [Zaki, 2000a], la base d'implications propres [Taouil et Bastide, 2001] ou encore la base $\mathcal{IGB}$ [Gasmi *et al.*, 2006].

Le calcul de toutes ces couvertures de règles se base sur les motifs minimaux (motifs libres) et sur les motifs maximaux (motifs fermés). Une exception concerne néanmoins la base de Guigues-Duquenne, où la construction est faite à partir des motifs pseudo-fermés (voir section 3.2.3). Le principe de construction fait la différence entre ces couvertures, c'est-à-dire la nature même de la prémisse et de la conclusion. Les mécanismes de *dérivabilité* et d'*informativité* [Gasmi *et al.*, 2006] font également la différence entre toutes ces couvertures.

Ces deux mécanismes constituent des critères permettant d'évaluer une couverture de règles :

1. La *dérivabilité* : exprime la possibilité de générer l'ensemble de toutes les règles si besoin est. Ceci se fait généralement à travers un mécanisme d'inférence tel que celui introduit dans [Kryszkiewicz, 1998]. Deux aspects servent alors à étudier le mécanisme d'inférence : est-il valide ? (dans la mesure où on arrive à ne dériver que les règles valides) et est-il complet ? (dans la mesure où on arrive à dériver toutes les règles valides).

2. L'*informativité* : exprime la possibilité de déterminer avec exactitude les supports et les confiances des règles dérivées.

Une étude comparative et détaillée concernant toutes ces couvertures de règles se trouve dans [Gasmi *et al.*, 2006]. Cette étude montre que seul le couple $\mathcal{GBE}$ (cf. Définition 14), $\mathcal{GBA}$ [Bastide *et al.*, 2000] ainsi que la base $\mathcal{IGB}$ [Gasmi *et al.*, 2006] sont des couvertures de règles permettant la dérivation de toutes les règles valides et sans perte d'information. Il est donc très difficile qu'une couverture de règles satisfasse les deux critères de *dérivabilité* et d'*informativité*. De plus, nous constatons qu'il n'existe pas un unique point de vue sur la redondance, puisque chaque proposition ne garde pas les mêmes règles. Ceci justifie la diversité des mécanismes d'inférences proposés pour la dérivation des règles redondantes.

3.2 Couvertures des règles d'association exactes

3.2.1 La base générique des règles exactes ($\mathcal{GBE}$)

Dans [Pasquier, 2000], l'auteur considère qu'une règle R est non redondante minimale s'il n'existe pas une autre règle R' possédant le même support et la même confiance, dont la prémisse est un sous-ensemble de la prémisse de R et la conclusion est un sur-ensemble de la conclusion de R.

Bastide caractérise les règles d'association exactes non redondantes minimales en utilisant les motifs fermés et leurs générateurs (motifs libres). Cette base de règles est connue sous le nom de *base générique des règles d'association exactes*, ou $\mathcal{GBE}$ [Bastide *et al.*, 2000] :

Définition 14 *[Bastide* et al.*, 2000] Soit $\mathcal{IFF}_{\mathcal{K}}$ l'ensemble des motifs fermés fréquents extrait d'un contexte $\mathcal{K}$. Pour chaque motif fermé fréquent $f \in \mathcal{IFF}_{\mathcal{K}}$, nous désignons par $\mathcal{ML}(f)$ l'ensemble de ses motifs libres. La base générique de règles d'association exactes est donnée par : $\mathcal{GBE} = \{R : g \Rightarrow (f - g) \mid f \in \mathcal{IFF}_{\mathcal{K}}$ et $g \in \mathcal{ML}(f)$ et $g \neq f$[6]$\}$.*

Exemple 8 *La base générique des règles exactes extraite du contexte de la table 3.1 pour un $minsup = 2$ est représentée à la table 3.2.*

L'avantage de cette base de règles est que toutes les règles exactes valides peuvent être dérivées ainsi que leurs supports. Cette base de règle est donc sans perte d'information [Pasquier, 2000].

Cependant, la minimalité de $\mathcal{GBE}$ est relative aux classes d'équivalence, dans le sens où elle regroupe les motifs minimaux par classe d'équivalence. Par exemple, la règle $h \rightarrow c$ est générée à partir du motif libre h de la classe d'équivalence dont la fermeture est hc, tandis que la règle $fh \rightarrow c$ est dérivée à partir du motif libre fh de la classe d'équivalence dont la fermeture est fhc.

[6]La condition $g \neq f$ est nécessaire pour ne pas retenir les règles de la forme $g \Rightarrow \emptyset$ lorsque $g{=}f$.

Libre	Fermé	Règle	Support
d	di	$d \rightarrow i$	3
e	ce	$e \rightarrow c$	2
g	dgi	$g \rightarrow di$	2
h	ch	$h \rightarrow c$	3
ad	adi	$ad \rightarrow i$	2
ah	ach	$ah \rightarrow c$	2
fh	cfh	$fh \rightarrow c$	2
fi	afi	$fi \rightarrow a$	2

TAB. 3.2 – Base générique des règles exactes ($\mathcal{GBE}$) relative au contexte de la table 3.1 pour $minsup = 2$.

Les deux motifs sont donc minimaux, mais vis-à-vis de deux classes d'équivalences distinctes. L'utilisation de $\mathcal{GBE}$ n'évite donc pas toute redondance.

Nous considérons donc que les règles sont non redondantes si elles sont à prémisse minimale :

Définition 15 (Règle à prémisse minimale) *Une règle $X \rightarrow i$ exacte est à prémisse minimale si pour tout Y sous-ensemble propre de X, la règle $Y \rightarrow i$ n'est pas exacte.*

Il est important de noter que si une règle exacte est à prémisse minimale, alors sa prémisse est un motif libre. Nous présentons cette constatation à travers la proposition 1.

Proposition 1 *Si une règle $X \rightarrow i$ est à prémisse minimale, alors X est libre.*

Preuve : Si X est non libre et $X \rightarrow i$ est exacte, alors il existe $X_1 \cup X_2 = X$ tels que $X_1 \rightarrow X_2$ est exacte. Par transitivité $X_1 \rightarrow i$ est exacte, donc $X \rightarrow i$ n'est pas à prémisse minimale. □

Remarquons que la réciproque de la proposition 1 est fausse, car tout motif libre ne donne pas nécessairement lieu à une règle à prémisse minimale. Par exemple, h et fh sont libres et les règles $h \rightarrow c$ et $fh \rightarrow c$ sont exactes. Mais $fh \rightarrow c$ n'est pas à prémisse minimale.

Ainsi, la liberté seule n'assure pas l'obtention de règles non redondantes minimales. Il faudrait associer à la contrainte de liberté une contrainte supplémentaire afin de garantir l'obtention de règles exactes non redondantes et minimales. La base d'implications propres introduite dans [Taouil et Bastide, 2001] permet l'obtention de telles règles.

3.2.2 La base d'implications propres ($\mathcal{BIP}$)

La base d'implications propres ($\mathcal{BIP}$) satisfait les deux critères suivants : elle fournit un ensemble de règles de caractérisation exactes tout en présentant des règles à prémisse minimale.

Les règles de la $\mathcal{BIP}$ constituent un sous-ensemble de la base générique des règles exactes ($\mathcal{GBE}$), en se restreignant aux règles de la forme $X \rightarrow i$ où i ne figure pas dans la fermeture d'un des sous-ensembles de X.
Reprenons notre exemple, où nous avons indiqué que la règle $fh \rightarrow c$ n'est pas à prémisse minimale bien que fh soit un motif libre. Cette règle n'est pas une implication propre puisque l'item c figure déjà dans la fermeture de h un sous-ensemble de fh. Cependant, la $\mathcal{BIP}$ n'est pas informative, mais dans le cadre de notre tâche liée à la caractérisation des valeurs manquantes, nous ne chercherons pas à dériver les autres règles valides. Nous cherchons plutôt à obtenir une caractérisation des valeurs manquantes qui soit non redondante.

La minimalité des règles de la $\mathcal{BIP}$ implique une réduction du nombre de règles. Nous illustrons cette condensation sur des données réelles relatives à la maladie de Hodgkin et celle de la Meningite à la fin de ce chapitre. La table 3.3 montre cette condensation sur le contexte de la table 3.1.

Libre	Fermé	Règle	Support
d	di	$d \rightarrow i$	3
e	ce	$e \rightarrow c$	2
g	dgi	$g \rightarrow d$	2
g	dgi	$g \rightarrow i$	2
h	ch	$h \rightarrow c$	3
fi	afi	$fi \rightarrow a$	2

Tab. 3.3 – Base des implications propres ($\mathcal{BIP}$) relative au contexte de la table 3.1 pour $minsup = 2$.

Dans ce qui suit, nous présentons une autre couverture de règles d'association exactes, connue sous le nom de la *base canonique* dite de Duquennes-Guigues [Guigues et Duquennes, 1986].

3.2.3 La base de Duquenne-Guigues ($\mathcal{GD}$)

La base de Duquenne-Guigues [Guigues et Duquennes, 1986], qu'on notera par $\mathcal{GD}$ est une base d'implications, adaptée au cadre des règles d'association exactes [Pasquier, 2000]. Cette base utilise la notion de motif pseudo-fermé, qui est cependant incompatible avec la notion de prémisse minimale.

Définition 16 (Motif pseudo-fermé fréquent) *[Pasquier, 2000] Un motif X est un pseudo-fermé s'il n'est pas fermé et s'il contient les fermetures de tous ses sous-ensembles qui sont des motifs pseudo-fermés fréquents.*

$$X \text{ pseudo-fermé} \Leftrightarrow h(X) \neq X \wedge \forall Y \subset X \text{ tel que } Y \text{ pseudo-fermé, } h(Y) \subset X.$$

Définition 17 (Base de Duquenne-Guigues) *[Pasquier, 2000] La base de Duquenne-Guigues pour les règles d'association exactes est définie par :*

$$\mathcal{GD}=\{r : X \rightarrow h(X) | \, X \text{ est pseudo-fermé}\}.$$

Exemple 9 *La base $\mathcal{GD}$ extraite du contexte de la table 3.1 pour un minsup = 2 est représentée dans la table 3.4.*

Pseudo-fermé	Fermé	Règle	Support
d	di	$d \rightarrow i$	3
g	gdi	$g \rightarrow di$	2
h	hc	$h \rightarrow c$	2

TAB. 3.4 – Base de Duquennes-Guigues ($\mathcal{GD}$) relative au contexte de la table 3.1 - $minsup = 2$.

Sur notre exemple, les règles de la base canonique $\mathcal{GD}$ constituent un sous-ensemble de la $\mathcal{BIP}$. Cependant, ceci n'est pas toujours le cas. En effet, si nous considérons par exemple un $minsup = 1$, la base canonique contiendra la règle $bce \rightarrow i$, tandis que la $\mathcal{BIP}$ ne contiendra pas cette règle, où on trouve plutôt l'implication propre $be \rightarrow i$. Ainsi, cet exemple nous montre qu'un motif pseudo-fermé, par exemple bce n'est pas nécessairement un motif libre. Les règles de la base $\mathcal{GD}$ ne sont donc pas à prémisses minimales. Cependant, il a été prouvé que la base de Duquennes-Guigues ($\mathcal{GD}$) est la base la plus réduite en termes de règles exactes que peut contenir un contexte, car il ne peut exister un ensemble générateur de règles plus réduit que les motifs pseudo-fermés [Pasquier, 2000].

3.3 Bilan et expériences

3.3.1 Discussion

La table 3.5 donne une synthèse sur la forme des règles des différentes couvertures des associations exactes abordées dans ce chapitre. Nous remarquons que les bases $\mathcal{GBE}$ et la $\mathcal{BIP}$ sont construites à partir des motifs libres, tandis que la base $\mathcal{GD}$ est construite à partir des motifs pseudo-fermés. Bien que ces derniers confèrent à la base canonique la minimalité en termes de nombre de règles, ils ne garantissent pas la minimalité des prémisses de ces règles. De plus, il nous semble difficile de trouver une interprétation sémantique à un motif pseudo-fermé.

Nous préférerons donc l'emploi de la $\mathcal{BIP}$, qui permettra de retrouver toutes les explications possibles des valeurs manquantes lors de leur caractérisation.

Couverture	Forme des règles	Prémisse minimale
$\mathcal{GBE}$	$libre \rightarrow fermé$	non
$\mathcal{BIP}$	$libre \rightarrow fermé$ avec restriction	oui
$\mathcal{GD}$	$pseudo\text{-}fermé \rightarrow fermé$	non

TAB. 3.5 – Synthèse sur les couvertures des règles exactes présentées dans ce chapitre.

3.3.2 Expériences

Nous présentons pour terminer les expériences que nous avons menées dans le cadre de cet état de l'art. Ces expériences ont pour objectif de comparer le nombre des règles extraites pour chaque couverture des règles d'association exactes.

Pour réaliser ces expériences, nous avons utilisé :

- le prototype MVMINER pour extraire la base $\mathcal{GBE}$ et la $\mathcal{BIP}$. MVMINER calcule $\mathcal{GBE}$ mais distingue les implications propres : elles sont obtenues par filtrage de la sortie. Ce prototype a été développé par FRANÇOIS RIOULT et est disponible à l'adresse suivante : https ://forge.greyc.fr/projects/kdariane.
- l'outil SPMF (Sequential Pattern Mining Framework) de PHILIPPE FOURNIER-VIGER, disponible à l'adresse suivante : http ://www.philippe-fournier-viger.com/spmf/.

Toutes ces expériences ont été réalisées avec un processeur Intel 1.66 GHz fonctionnant avec 2 Go de mémoire centrale sous Linux. Les tables 3.6 et 3.7 montrent le nombre des règles de chaque couverture en fonction de la variation de *minsup*, respectivement sur les données de la maladie de HODGKIN ainsi que celle de la MENINGITE données utilisées comme application réelle dans ce travail. Ces tables indiquent le nombre total des règles.

minsup (%)	$\mathcal{GBE}$	$\mathcal{BIP}$	$\mathcal{GD}$
50	3 239	7	5
40	15 926	8	6
30	83 565	10	8
20	625 158	25	12
10	-	-	12
5	-	-	28

TAB. 3.6 – Comparaison entre le nombre de règles des différentes couvertures sur les données de la maladie de HODGKIN.

minsup (%)	$\mathcal{GBE}$	$\mathcal{BIP}$	$\mathcal{GD}$
50	36	1	1
40	102	3	3
30	290	13	3
20	1987	114	4
10	24 580	1 493	13
5	159 708	8 165	13

TAB. 3.7 – Comparaison entre le nombre de règles des différentes couvertures sur les données de la maladie de MENINGITE.

À partir de ces tables, nous constatons que la base $\mathcal{GBE}$ contient le plus de règles extraites. Ensuite, à la deuxième position, nous trouvons la $\mathcal{BIP}$ qui condense énormément par rapport à $\mathcal{GBE}$. Ceci s'explique par la relation d'inclusion, qui existe entre ces couvertures, que nous avons mise en évidence dans ce chapitre.

Notons que sur les données de la maladie de HODGKIN, le symbole « - » indique que l'exécution du programme MVMINER échoue faute de mémoire avec des valeurs de faible support. Ainsi, l'extraction des bases $\mathcal{GBE}$ et $\mathcal{BIP}$ est impraticable. Au chapitre 5, nous proposons un nouvel algorithme pour calculer efficacement la $\mathcal{BIP}$, sans être limité par la quantité de mémoire disponible.

3.4 Conclusion

Dans ce chapitre, nous avons décrit l'intérêt des couvertures des règles d'association exactes et nous avons présenté les différentes couvertures proposées dans la littérature : la base générique des règles exactes, la base d'implications propres et la base de Duquenne-Guigues. Cette étude nous a permis de mettre en évidence deux propriétés : (1) seule la base d'implications propres fournit des règles à prémisses minimales, donc véhicule moins de redondance et (2) il existe une relation d'inclusion entre ces couvertures. La $\mathcal{BIP}$ répond donc à notre besoin de caractériser les valeurs manquantes de façon non redondante, et le chapitre suivant montre sa mise en œuvre.

Deuxième partie

Complétion contextualisée par caractérisation non redondante des valeurs manquantes

Chapitre 4

Caractérisation non redondante des valeurs manquantes et proposition d'une nouvelle typologie

Sommaire

De nombreux travaux dédiés au traitement des valeurs manquantes dans le domaine de la fouille de données ont été développés. Même si elle n'est pas clairement formulée, l'hypothèse selon laquelle les valeurs manquantes apparaissent aléatoirement est implicitement utilisée par

ces méthodes. Or, dans la réalité, les valeurs manquantes ne sont pas nécessairement aléatoires : il existe le plus souvent une explication derrière 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 causes d'apparition. À notre connaissance, il n'existe pas de travaux qui ont étudié la validité de cette hypothèse ni précisé sous quelles conditions les valeurs manquantes sont réellement aléatoires. Nous estimons également que l'impact d'une telle hypothèse a longtemps été passé sous silence par la majorité des méthodes de traitement des valeurs manquantes.

Dans ce chapitre, nous commençons par discuter l'impact de la restriction au modèle aléatoire comme modèle d'apparition des valeurs manquantes. Nous positionnons également notre travail par rapport à la modélisation de Little et Rubin (*cf.*, chapitre 1, page 12). Dans la section 4.2, nous introduisons une nouvelle typologie des valeurs manquantes à partir des données mesurées et indiquons la forme des règles d'association qui permettent de la caractériser [Ben Othman *et al.*, 2008, Ben Othman *et al.*, 2009b]. Dans la section 4.3, 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 aux méningites infantiles d'autre part, sont enfin présentés dans la section 4.4.

4.1 Les valeurs manquantes sont-elles vraiment aléatoires ?

Dans cette section, nous essayons de dégager les caractéristiques des valeurs manquantes aléatoires par rapport à celles qui sont non aléatoires.

4.1.1 L'exceptionnel contre la généralité

Les approches classiques de traitement des valeurs manquantes qualifient souvent l'apparition de ces dernières de phénomène *aléatoire*. En supposant que les données sont représentées sous forme matricielle, les valeurs manquantes seront distribuées aléatoirement dans la matrice. Ceci suppose qu'une valeur manquante affecte n'importe quelle ligne et/ou n'importe quelle colonne. En pratique, cette hypothèse est *naïve* et irréaliste. Par exemple, dans le cadre de sondages, lorsqu'une personne ne fournit pas de réponse à une question, ceci n'est pas le fruit d'un simple hasard. Le plus souvent, les explications derrière l'absence de réponse sont multiples : la personne refuse tout simplement de répondre ou juge qu'elle n'est pas concernée par la question. Dans ce cas, la valeur manquante est dite *non aléatoire*, elle affecte une information bien précise et cache généralement une explication particulière. Évidemment, nous n'écartons pas l'éventuelle occurrence aléatoire des valeurs manquantes : quand une personne oublie simplement de répondre. Dans ce cas, nous qualifions cette situation d'accidentelle et donc d'aléatoire.

Bien que l'hypothèse aléatoire ne soit pas à écarter, nous sommes persuadés qu'en pratique elle est assez rare. En effet, nous pensons que les valeurs manquantes révèlent le plus souvent des situations particulières. Cette hypothèse erronée constitue l'écueil des approches classiques de traitement des valeurs manquantes.

Pour approfondir l'idée présentée précédemment, nous allons nous baser sur un exemple de données illustré par la table 4.1. Cette table représente deux catégories de valeurs manquantes : les aléatoires et les non aléatoires. Une valeur manquante non aléatoire peut-être caractérisée par une règle exacte concluant sur une valeur manquante. Une règle de caractérisation se présente sous la forme $(A_1 = a) \rightarrow (A_2 =?)$ et indique qu'étant donnée une prémisse $(A_1 = a)$, alors une valeur manquante est notée sur l'attribut A_2. Remarquons que les valeurs manquantes sur les objets o_1, o_2 et o_3 rentrent dans ce cas, puisque la règle $(A_1 = a) \rightarrow (A_2 =?)$ est exacte. En revanche, les valeurs manquantes, portant sur les objets o_6 et o_7, sont des valeurs manquantes aléatoires puisque chacun de ces objets caractérise un cas distinct. Ainsi, nous pouvons considérer que les valeurs manquantes aléatoires sont la conséquence de circonstances exceptionnelles pouvant affecter n'importe quel objet. À l'inverse, les valeurs manquantes non aléatoires caractérisent une situation unique. Cette caractérisation permet d'identifier un contexte qui régulièrement conduit à des valeurs manquantes. Sur l'exemple de la table 4.1, ce contexte est le motif $(A_1 = a)$. Les valeurs manquantes non aléatoires sont donc caractérisées par une généralité, tandis que les aléatoires constituent une exception.

	A_1	A_2
o_1	a	**?**
o_2	a	**?**
o_3	a	**?**
o_4	b	d
o_5	c	e
o_6	b	**?**
o_7	c	**?**
o_8	b	d
o_9	c	e

TAB. 4.1 – Données avec des valeurs manquantes aléatoires et non aléatoires

4.1.2 Hypothèse *aléatoire* : quel impact ?

Comme nous l'avons précédemment évoqué, le type aléatoire représente la principale catégorie de valeurs manquantes traitée dans la littérature [Shafer et Graham, 2002, Pearson, 2006]. Cette considération suppose des hypothèses, souvent négligées par la grande majorité des méthodes

de traitement des valeurs manquantes. Dans cette section, nous montrons l'impact d'une telle considération, distingué en termes d'informativité, de biais et de représentativité.

Impact en termes d'informativité

Étant donné qu'elle permet de caractériser une situation particulière, une valeur manquante non aléatoire est informative, car elle apporte une information sur son contexte d'apparition. Si nous reprenons l'exemple (page 13) des personnes ayant un sur-poids et qui ont tendance à le cacher, nous saurons que les personnes présentant une valeur manquante sur l'attribut *Poids* cachent potentiellement un sur-poids. Une telle informativité devrait être exploitée par les méthodes de traitement des valeurs manquantes. Malheureusement, le modèle d'apparition des valeurs manquantes étant ignoré, voire réduit au seul modèle aléatoire, la perte d'informativité est conséquente.

Impact en termes de représentativité

Un échantillon est dit *représentatif* lorsqu'il possède les mêmes caractéristiques que l'échantillon que l'on souhaite étudier. Ici, on s'intéresse à la représentativité des données observées (ne présentant pas des valeurs manquantes) par rapport aux données non observées (présentant des valeurs manquantes), puisque c'est à partir des données observées que l'on se base pour compléter les valeurs manquantes. Ainsi, la question qui se pose nécessairement est : "*les données observées sont-elles représentatives des données avec valeurs manquantes ?*" La figure 4.1 illustre les échantillons de données relatifs au contexte de la table 4.1, où nous distinguons l'échantillon de données non observées selon qu'il s'agisse de valeurs manquantes aléatoires ou non aléatoires.

	A_1	A_2
o_4	b	d
o_5	c	e
o_8	b	d
o_9	c	e

	A_1	A_2
o_1	a	**?**
o_2	a	**?**
o_3	a	**?**

	A_1	A_2
o_6	b	**?**
o_7	c	**?**

FIG. 4.1 – **(a)** : Échantillon de données observées. **(b)** : Échantillon de données non observées avec valeurs manquantes non aléatoires. **(c)** : Échantillon de données non observées avec valeurs manquantes aléatoires - Contexte de la table 4.1.

La figure 4.1 met en exergue que seul l'échantillon de données non observées avec valeurs manquantes aléatoires **(c)** est représentatif de l'ensemble des données observées **(a)** : les instances présentant une valeur manquante ne peuvent pas être distinguées des instances ayant une valeur mesurée. Ceci n'est pas le cas pour l'échantillon de données non observées avec valeurs

manquantes non aléatoires **(b)**, où cette distinction est évidente.

Impact en termes de biais

Faute de représentativité, le résultat du traitement effectué sur les données non observées sera certainement biaisé. Pour montrer cela, plaçons-nous dans le cadre de l'exemple concernant le surpoids. En faisant l'hypothèse d'un modèle aléatoire, un moyen de traiter les valeurs manquantes sur l'attribut *Poids* consisterait à s'appuyer sur les habitudes alimentaires des personnes ayant indiqué leurs poids (l'échantillon des données observées). Il est certain que le résultat sera biaisé puisque l'échantillon de personnes ayant indiqué leurs poids n'est pas représentatif de l'échantillon de personnes n'ayant pas indiqué leurs poids. Le biais se présentera sous forme d'écart entre la vraie valeur de la variable non observée et la valeur estimée. Ce biais est dû au fait que l'hypothèse de départ est fausse.

Récaptitulatif

Le tableau 4.2 montre les différences entre les valeurs manquantes aléatoires et non aléatoires, en termes de contexte d'apparition, d'informativité, de représentativité et de biais.

Nous souhaitons donc orienter notre travail vers une prise en considération préalable du modèle d'apparition des valeurs manquantes puis vers une adaptation du traitement en fonction du modèle. Cette orientation constitue le point clef de notre travail, développé dans la section suivante.

	valeur manquante aléatoire	**valeur manquante non aléatoire**
Contexte d'apparition	exception	généralité
Informativité		×
Représentativité	×	
Biais		×

TAB. 4.2 – Caractéristiques des valeurs manquantes aléatoires contre celles non aléatoires.

4.1.3 Positionnement par rapport à LITTLE et RUBIN

Comme cela a été souligné par [Shafer et Graham, 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, où l'expert maîtrise l'origine des valeurs manquantes. Ceci est le cas de l'exemple des personnes ayant un sur-poids, qui ont tendance à cacher leur 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 donc 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 ce chapitre, nous considérons donc que la gestion du type *NMAR* relève du pré-traitement des données, sous la responsabilité de l'expert.

En outre, nous trouvons que la modélisation classique est 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. En revanche, nous pensons qu'une valeur manquante sur un attribut n'admet pas toujours une seule explication. La section 4.4, 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 une vision plus réaliste en affinant le niveau de granularité, grâce à des caractérisations propres à *un groupe restreint d'objets*.

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 été ni examiné, ni mesuré. Cette finesse d'analyse repose sur les relations entre les valeurs manquantes et les données, ou impliquent d'autres valeurs manquantes. Ce type de valeurs manquantes n'est pas pris en compte dans la modélisation de Little et Rubin. Pourtant, l'analyse de ces situations caractéristiques permettrait d'affiner les

méthodes de traitement.

4.2 Nouvelle typologie des valeurs manquantes

Nous commençons cette section par préciser la relation qui lie un contexte réel (Définition 1 - page 10) à un contexte mesuré (Définition 2 - page 11).

Propriété 1 *Étant donnés un contexte réel $\mathcal{K}$, un opérateur $mv()$ qui transforme $\mathcal{K}$ en un contexte mesuré $mv(\mathcal{K})$.*
Le contexte $mv(\mathcal{K})$ vérifie alors les propriétés suivantes pour $valeur \in \{présent, absent\}$:

1. $mv(\mathcal{R})(o,i) = valeur \Rightarrow \mathcal{R}(o,i) = valeur$
2. $\mathcal{R}(o,i) = valeur \Rightarrow mv(\mathcal{R})(o,i) \in \{valeur, manquant\}$

Un contexte réel $\mathcal{K}$ représente des données parfaites (sans valeurs manquantes). Cependant, les données mesurées correspondantes ne sont pas toutes disponibles et donneront lieu éventuellement à l'apparition de valeurs manquantes. Un contexte mesuré $mv(\mathcal{K})$ désigne les données que nous détenons en pratique ; l'opérateur $mv()$ modélise un effacement de données. Quand une valeur est manquante dans $mv(\mathcal{K})$, il est impossible de connaître sa vraie valeur dans $\mathcal{K}$. En revanche, quand la valeur est connue dans $mv(\mathcal{K})$, elle correspond à celle du contexte réel $\mathcal{K}$ (première relation de la propriété 1). La deuxième relation assure qu'une valeur présente ou absente dans $\mathcal{K}$ conservera sa valeur à l'identique ou sera manquante dans $mv(\mathcal{K})$.

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, qui se présente comme suit :

- **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 et 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.

4.2.1 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 motif dans un contexte $mv(\mathcal{K})$, selon les cas suivants :

Définition 18 *Soient $X \subseteq \mathcal{I}$ un motif et $o \in \mathcal{O}$ un objet.*

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

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

Exemple 10 *Dans le contexte $mv(\mathcal{K})$, illustré par la table 1.2 (page 12), nous avons $Présent(adf, o_4)$, $Absent(bc, o_1)$, $Manquant(ah, o_8)$ et $PartPrésent(bdg, o_8)$ et $Absent(bd, o_1)$.*

Remarquons qu'une partition est ainsi définie : un motif 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 19 *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.

4.2.2 Relation avec la typologie classique de Little et Rubin

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 19 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. La section suivante présente notre contribution concernant l'utilisation de la base d'implications propres pour mettre en évidence la nouvelle typologie des de valeurs manquantes.

4.3 Caractérisation des valeurs manquantes à l'aide d'implications propres

Nous expliquons maintenant comment utiliser la base d'implications propres 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.

4.3.1 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 4.3 compare les règles de la base $\mathcal{GBE}$ (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 $vm(A_i)$.*

	Règle	Support
R_1	$vm(A_1) \rightarrow vm(A_4)$	3
R_2	$d \rightarrow vm(A_4)$	2
R_3	$g \rightarrow vm(A_4)$	2
R_4	$c \wedge vm(A_1) \rightarrow vm(A_4)$	2

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

TAB. 4.3 – Règles concluant sur $vm(A_4)$. (a) : les règles de la base $\mathcal{GBE}$. (b) : les implications propres.

Nous remarquons que la règle R_4 n'a pas été générée par la base d'implications propres ($\mathcal{BIP}$) qui contient déjà R'_1 à prémisse minimale. Grâce à la propriété 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 de la base $\mathcal{GBE}$, 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 ($\mathcal{BIP}$), elles seront simplement caractérisées comme indirectes.

Pour la suite de ce chapitre, nous caractérisons les valeurs manquantes grâce aux implications propres qui modélisent la définition 19. 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 prémisses des règles de caractérisation. En effet, nous cherchons seulement à avoir une caractérisation non redondante. Dans cet esprit, l'utilisation de la base d'implications propres présente, 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. Dans ce qui suit, nous examinons plus particulièrement le type hybride.

4.3.2 Caractérisation du type hybride

Selon la définition 19, une valeur manquante hybride s'explique par la présence simultanée de valeurs observées et de valeurs manquantes. Comment alors considérer une valeur manquante caractérisée à la fois directe et indirecte ?

Sur notre exemple, la valeur manquante affectant l'attribut A_4 de l'objet o_8 est caractérisée par deux règles : $vm(A_1) \rightarrow vm(A_4)$ et $d \rightarrow vm(A_4)$. La première règle caractérise un type indirect, la deuxième un type direct. Cependant, la règle $d \wedge vm(A_1) \rightarrow vm(A_4)$ est valide et caractérise un type hybride, mais ce n'est pas une implication propre car elle n'est pas à prémisse minimale.

Dans la pratique, 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 4.4 indique les implications propres concluant sur des valeurs manquantes relatives au contexte de la table 1.2 (page 12). Ces implications permettent de déduire la caractérisation du type des valeurs manquantes sur les objets qui supportent la règle. Cette caractérisation est présentée à la table 4.5.

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

TAB. 4.4 – Implications propres concluant sur une valeur manquante relatives au contexte de la table 1.2 ; $minsup = 2$.

	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}

TAB. 4.5 – Typologie des valeurs manquantes du contexte de la table 1.2.

4.3.3 Comparaison des typologies de valeurs manquantes

Dans ce qui suit, nous résumons les principales caractéristiques (table 4.6) de notre nouvelle typologie en la comparant à la typologie classique de Little et Rubin [Little et Rubin, 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 la portée de travail de Little et Rubin de purement théorique. En effet, il est impossible de mettre en pratique un modèle de valeurs manquantes concernant des données indisponibles.

En outre, la caractérisation de Little et Rubin est globale car toutes les valeurs manquantes d'un même attribut sont caractérisées de la même façon. Notre typologie propose une vision locale, portant sur 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

Tab. 4.6 – Caractéristiques des typologies des valeurs manquantes.

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 : celui de Little et Rubin est global, tandis que notre approche est locale : la correspondance entre les types est localisé sur un groupe d'objets.

4.3.4 Discussion

La définition 19 des nouveaux types de valeur manquante dépend du paramètre de support relatif aux règles permettant la caractérisation. 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 items en prémisse ?

D'autre part, l'utilisation de la base d'implications propres nous limite à des caractérisations par des règles valides, de confiance égale à 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. 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.

4.4 Expérimentations

Cette section rapporte les résultats de nos expériences sur la caractérisation des différents types de valeurs manquantes sur 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 : l'une relative à la maladie de Hodgkin, un cancer des ganglions du système lymphatique, 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).

4.4.1 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.

Résultats : L'extraction des règles a été effectuée avec un support absolu minimum égal à 700. Les 15 règles découvertes sont reportées à la table 4.7. Le faible nombre de règles est

caractéristique d'une réduction drastique du nombre de règles à calculer, par rapport aux règles de la base $\mathcal{GBE}$ (*cf.* table 4.8), qui évite leur redondance et facilite ainsi fortement leur étude.

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

TAB. 4.7 – Implications propres extraites à partir de la base HODGKIN pour une valeur de minsup=700. $vm(attribut)$ indique que *attribut* est manquant.

		Nombre de règles	Nombre de règles concluant sur une valeur manquante
Base	$\mathcal{BIP}$	**49**	**15**
de HODGKIN	$\mathcal{GBE}$	2 923 070	2 681 045
Base	$\mathcal{BIP}$	**17 508**	**152**
de la MENINGITE	$\mathcal{GBE}$	182 317	7659

TAB. 4.8 – Comparaison entre le nombre d'implications propres et de règles de la base $\mathcal{GBE}$.

Par exemple, la règle R_4 (table 4.7) indique que tous les objets contenant l'item $plaq <= 600$ et une valeur manquante sur l'attribut *chd* (ganglion cervical haut droit) contiennent également une valeur manquante sur l'attribut *chg*. C'est une caractérisation de valeur manquante de type *hybride*.

Les règles R_1 et R_2 concluent sur un envahissement du ganglion cervical haut droit *chd* ou gauche *chg* 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 caractérisées 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 cherchent à compléter ces valeurs soit par la moyenne, soit par une valeur aléatoirement choisie, soit encore 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*. En revanche, la règle R_{15} caractérise des valeurs manquantes de type *hybride*.

La table 4.9 présente la répartition des caractérisations pour chaque attribut manquant, en considérant comme hybride la combinaison direct-indirect. Remarquons 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*. Notons 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%

TAB. 4.9 – 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 : cette caractérisation est réaliste et utile en pratique pour appréhender les multiples causes pouvant expliquer 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%).

4.4.2 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 de 1% à 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 à la table 4.10.

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

TAB. 4.10 – Implications propres extraites à partir de la base MENINGITE pour une valeur de *minsup*=10.

La règle R_1 conclut sur l'attribut *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 a été 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 4.11 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	Nombre de 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%

TAB. 4.11 – Caractérisation des valeurs manquantes dans la base MENINGITE.

4.5 Conclusion

Dans ce chapitre, nous avons montré que les valeurs manquantes ne doivent pas être exclusivement 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 : les 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.

De plus, 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. Nous avons ainsi distingué les valeurs manquantes aléatoires qui sont caractérisées par l'absence de relations dans les données sur leur contexte d'apparition. Dans ce cas, nous proposerons une complétion à l'aide d'un modèle calculé sur les données, par exemple à l'aide de règles d'association [Ben Othman et Ben Yahia, 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. Dans cette situation, la complétion par une valeur du domaine de définition n'aurait pas de sens. Notre but est d'aller vers une complétion des valeurs manquantes qui prenne différentes formes, selon les différents types mis en évidence par notre caractérisation puis valorisés lors de la phase de complétion.

Chapitre 5

Calcul de la base d'implications propres par traverses minimales

Sommaire

Dans ce chapitre, nous proposons un nouvel algorithme de calcul des implications propres [Ben Othman *et al.*, 2011]. En effet, filtrer un algorithme classique d'extraction de règles à prémisse minimale n'est pas efficace, puisque l'on n'a besoin que de certaines règles, celles concluant sur des valeurs manquantes. Nous formalisons notre contribution par un calcul de traverses minimales extraites à partir du contexte complémentaire. Pour cela, nous présentons brièvement le formalisme autour de la théorie des hypergraphes, matériel technique utile à la compréhension de notre contribution. Ensuite, nous montrons le principe de notre méthode de calcul de la base d'implications propres, ainsi que l'algorithme MTBipminer associé et nous illustrons sa mise en œuvre à travers un exemple.

5.1 Préliminaires

5.1.1 Définition du problème

Nous avons montré dans le chapitre 4 que la caractérisation des valeurs manquantes se fait grâce à la base d'implications propres concluant sur ces valeurs. Dans ce chapitre, nous proposons précisément un nouvel algorithme de calcul de cette base de règles. Une solution *naïve* pour extraire cette base serait de filtrer le résultat d'un algorithme classique de calcul de règles. Cependant, cette solution n'est pas efficace car ces algorithmes calculent toutes les implications propres, tandis que pour faire la caractérisation, on a seulement besoin de celles concluant sur des valeurs manquantes.

La solution que nous proposons consiste à extraire les traverses minimales à partir du contexte complémentaire (*cf.*, proposition 2, propriété 2). Dans ce qui suit, nous présentons brièvement la notion de traverse minimale.

5.1.2 Traverses minimales d'un hypergraphe

Un hypergraphe est une structure plus générale que celle d'un graphe, où on parle d'hyperarête au lieu d'arête. Une hyperarête relie un nombre quelconque de sommets. Un hypergraphe est défini comme suit [Berge, 1973] :

Définition 20 (Hypergraphe) *Un hypergraphe $\mathcal{H}$ est un couple ($\mathcal{V}$,$\mathcal{E}$) où $\mathcal{V}$=$\{v_1, v_2, ...v_n\}$ est un ensemble de sommets de $\mathcal{H}$ et $\mathcal{E}$=$\{e_1, e_2, ..., e_m\}$ est un ensemble de parties non vides de $\mathcal{V}$. Les éléments de $\mathcal{E}$ sont appelés les hyperarêtes de $\mathcal{H}$.*

Dans le domaine de la fouille de données, les items peuvent représenter les sommets d'un hypergraphe, tandis que les objets peuvent matérialiser les hyperarêtes. Par exemple, la figure 5.1 montre l'hypergraphe constitué par l'ensemble des objets $\{o_1, o_2, o_3\}$ du contexte de la table 1.1 - page 11.

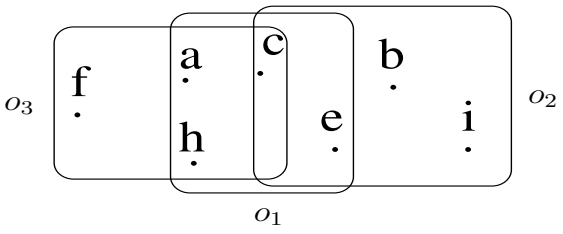


FIG. 5.1 – Hypergraphe associé aux objets $\{o_1, o_2, o_3\}$ du contexte de la table 1.1.

Définition 21 (Traverse - Traverse minimale) *Un ensemble $T \subseteq V$ est appelé transversal (ou traverse) de $\mathcal{H}$, si T intersecte toutes les hyperarêtes de $\mathcal{H}$, i.e., $\forall\ e \in \mathcal{E}$, $e \cap T \neq \emptyset$. Une*

transverse T est dite minimale *si aucun de ses sous-ensembles n'est une traverse. L'ensemble des traverses minimales de $\mathcal{H}$ est noté MinTr($\mathcal{H}$).*

Le problème de calcul des traverses minimales est un problème difficile, car le nombre de traverses minimales dans un hypergraphe $\mathcal{H}$ est potentiellement exponentiel en la taille de $\mathcal{H}$. Un premier algorithme faisant ce calcul a été proposé par BERGE dans [Berge, 1989]. Ensuite, plusieurs autres algorithmes ont suivi, tels que ceux décrits dans [Kavvadias et Stavropoulos, 2005, Hébert, 2007]. Bien qu'il soit simple, l'algorithme de BERGE a été critiqué dans [Kavvadias et Stavropoulos, 2005]. En effet, cet algorithme nécessite le stockage des traverses intermédiaires dont le nombre peut-être exponentiel et reste inapplicable même dans le cas d'hypergraphes constitués d'une dizaine de sommets. Les auteurs [Kavvadias et Stavropoulos, 2005] ont donc proposé un nouvel algorithme basé sur celui de BERGE, qui procède de la même façon incrémentale, mais emploie un parcours en profondeur. L'algorithme proposé dans [Hébert, 2007] emploie quant à lui un parcours par niveaux et exploite des critères d'élagage, bien connus en fouille de données. Nous invitons le lecteur à se référer à [Hagen, 2008] pour un état de l'art concernant les algorithmes de calcul des traverses minimales.

Pour le calcul des traverses minimales, nous disposons donc d'un vaste choix d'algorithmes fournis par la communauté : en largeur, en profondeur, avec ou sans consommation de mémoire. Dans la suite de ce chapitre, nous montrons que c'est l'algorithme *MTminer* [Hébert, 2007] qui est le mieux adapté à notre problématique.

5.2 Extraction de la base d'implications propres par traverses minimales

5.2.1 Principe de notre méthode

L'extraction de la $\mathcal{BIP}$ par traverses minimales peut-être décrite de la façon suivante : la règle $X \rightarrow i$ est une implication propre si X est une traverse minimale (cf. propriété 2) des complémentaires des objets contenant l'item i.

L'idée générale de notre approche est illustrée par la figure 5.2. L'approche est itérative selon les attributs manquants du contexte. Pour chaque item i, nous considérons le contexte initial, divisé en deux sous-ensembles : les objets contenant i et les objets ne contenant pas i. Nous considérons ensuite le contexte complémentaire. Un hypergraphe est alors associé aux objets complémentaires contenant i, à partir duquel nous faisons l'extraction des traverses minimales. Nous obtenons ainsi les prémisses des implications propres concluant sur i.

Dans ce qui suit, nous présentons la formalisation de notre méthode ainsi que l'algorithme MTBIPMINER associé.

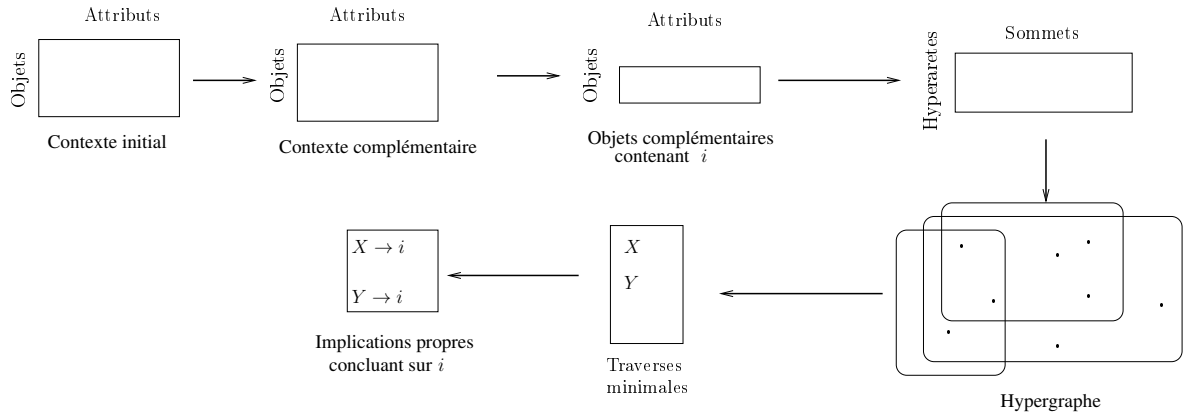


FIG. 5.2 – Principe général de notre approche pour l'extraction des implications propres concluant sur un item i par traverses minimales.

5.2.2 Définitions

Considérant qu'un objet $o \in \mathcal{O}$ est un ensemble d'items, soit $o \subseteq \mathcal{I}$, nous notons $\overline{o} = \mathcal{I} \backslash o$ le complémentaire de o.
Nous utiliserons la proposition suivante :

Proposition 2 *Pour tout motif X et tout objet o :*

$$X \subseteq o \iff X \cap \overline{o} = \emptyset$$

Preuve : $X \subseteq o \iff \forall i \in X, i \in o \iff \neg(\exists i \in X, i \in \overline{o})$ □

Notre méthode d'extraction de la $\mathcal{BIP}$ par traverses minimales repose sur la propriété suivante :

Propriété 2 *La règle exacte $X \to i$ 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. Cela veut dire : $\forall o \mid i \in \overline{o}, X \cap \overline{o} \neq \emptyset$ donc $\forall o \mid i \notin o, \neg(X \subseteq o)$ et par contraposition : $X \subseteq o \implies i \in o$ donc la règle $X \to i$ exacte est valide. □

Pour extraire les implications propres $X \to i$, il nous faut donc extraire les traverses **minimales** des complémentaires des objets contenant i. Cependant, la majorité des traverses obtenues dans le complémentaire ont un support nul dans le contexte initial. Les tables 5.1, 5.2 et 5.3 montrent le déroulement de notre méthode pour l'extraction des implications propres concluant sur $vm(A_3)$: calcul du contexte complémentaire contenant $mv(A_3)$ et extraction des traverses

minimales à partir de ce dernier contexte. La table 5.4 montre ces traverses minimales. Remarquons que parmi ces 30 traverses, seulement trois : ac, ah et ch ont un support non nul et constituent donc des prémisses d'implications propres concluant sur $mv(A_3)$.

	A_1			A_2			A_3				A_4		
	a	b	$vm(A_1)$	c	d	$vm(A_2)$	e	f	g	$vm(A_3)$	h	i	$vm(A_4)$
o_1	×			×						×	×		
o_2			×	×			×						×
o_3	×			×						×	×		
o_4	×				×			×					×
o_5			×	×				×					×
o_6		×				×		×			×		
o_7	×					×			×				×
o_8			×		×				×				×

TAB. 5.1 – Contexte initial.

	A_1			A_2			A_3				A_4		
	a	b	$vm(A_1)$	c	d	$vm(A_2)$	e	f	g	$vm(A_3)$	h	i	$vm(A_4)$
o_1		×	×		×	×	×	×	×			×	×
o_2	×	×			×	×		×	×	×	×	×	
o_3		×	×		×	×	×	×	×			×	×
o_4		×	×	×		×	×		×	×	×	×	
o_5	×	×			×	×	×		×	×	×	×	
o_6	×		×	×	×		×		×	×		×	×
o_7		×	×	×	×		×	×		×	×	×	
o_8	×	×		×		×	×	×		×	×	×	

TAB. 5.2 – Contexte complémentaire.

	A_1			A_2			A_3			A_4		
	a	b	$vm(A_1)$	c	d	$vm(A_2)$	e	f	g	h	i	$vm(A_4)$
o_2	×	×			×	×		×	×	×	×	
o_4		×	×	×		×	×		×	×	×	
o_5	×	×			×	×	×		×	×	×	
o_6	×		×	×	×		×		×		×	×
o_7		×	×	×	×		×	×		×	×	
o_8	×	×		×		×	×	×		×	×	

TAB. 5.3 – Restriction à $vm(A_3)$ du contexte complémentaire.

	Traverses minimales		Traverses minimales
tr_1	$vm(A_3)$	tr_{16}	$c\ vm(A_2)$
tr_2	i	tr_{17}	de
tr_3	ab	tr_{18}	dh
tr_4	$\boxed{\text{ac}}$	tr_{19}	$d\ vm(A_2)$
tr_5	ae	tr_{20}	ef
tr_6	$\boxed{\text{ah}}$	tr_{21}	eg
tr_7	$a\ vm(A_1)$	tr_{22}	eh
tr_8	bc	tr_{23}	$e\ vm(A_2)$
tr_9	bd	tr_{24}	fg
tr_{10}	be	tr_{25}	gh
tr_{11}	bg	tr_{26}	$h\ vm(A_1)$
tr_{12}	$b\ vm(A_1)$	tr_{27}	$vm(A_1)\ vm(A_2)$
tr_{13}	cd	tr_{28}	adg
tr_{14}	cg	tr_{29}	$af\ vm(A_2)$
tr_{15}	$\boxed{\text{ch}}$	tr_{30}	$df\ vm(A_1)$

TAB. 5.4 – Ensemble des traverses minimales extraites à partir du contexte complémentaire contenant $mv(A_3)$. Seules les traverses encadrés ont un support non nul dans le contexte initial.

Dans le cas de données réelles, l'écart entre le nombre de traverses minimales dans le complémentaire et le nombre de traverses minimales ayant un support non nul dans le contexte initial est encore plus important. Les tables 5.6 et 5.5 montrent d'une part, le nombre de traverses extraites par attribut manquant et d'autre part, le nombre de traverses de support non nul, sur les données de la maladie de HODGKIN et de la MENINGITE. On constate que parmi les millions de traverses minimales, seules quelques dizaines ont un support non nul. Ces nombres montrent la nécessité de concevoir un algorithme de calcul de traverses minimales utilisant une contrainte de support afin d'extraire directement les traverses minimales utiles.

La contrainte de support étant anti-monotone, un algorithme de calcul de traverses minimales par niveaux serait dans ce cas adéquat. Nous réutilisons donc l'algorithme MTMINER, auquel nous ajoutons une contrainte de fréquence. De plus, MTMINER est très efficace sur des données denses [Hébert, 2007]. Ceci est le cas ici puisqu'il est appliqué sur des complémentaires d'objets. Avec une contrainte de support supplémentaire, il est donc encore plus efficace. La sous-section suivante présente l'algorithme MTBIPMINER, une adaptation de l'algorithme MTMINER avec contrainte de support.

Attribut manquant	Nombre de traverses minimales dans le complémentaire	Nombre de traverses minimales de support non nul dans le contexte initial
tonus	214 296	36 825
polysang	228 877	3 3145
polyns	228 718	33 716
gluc	424 362	6 902
vs	287 111	50 630
leuco	422 665	4 532
dfievre	432 254	595
sneuro	431 762	305
prot	430 853	1 062

Tab. 5.5 – Nombres de traverses minimales par attribut manquant sur les données de la Meningite.

5.3 L'algorithme MTBipminer

5.3.1 Présentation

Nous détaillons dans cette sous-section le fonctionnement de notre algorithme MTBipminer.

Définition 22 (Contexte complémentaire) *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. Le contexte complémentaire de $\mathcal{K}$, noté $\overline{\mathcal{K}}$, est le contexte défini sur $\mathcal{K}$ par la relation opposée $\overline{R}$:*

$$\overline{\mathcal{K}} = (\mathcal{O}, \mathcal{I}, \overline{\mathcal{R}})$$

avec $\forall\ i \in \mathcal{I}$, $o \in \mathcal{O}$,

$$\overline{\mathcal{R}}(i, o) = \textit{présent} \iff \mathcal{R}(i, o) = \textit{absent}$$

(et vice-versa).

Définition 23 (contexte projeté) *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte et i un item. Le* contexte projeté *de $\mathcal{K}$ selon i, noté par $\mathcal{K}_i$, est la restriction de $\mathcal{K}$ aux objets contenant i :*

$$\mathcal{K}_i = (\mathcal{O}_i, \mathcal{I}, \mathcal{R})$$

avec

$$\mathcal{O}_i = \{o \in \mathcal{O} \mid \mathcal{R}(o, i) = \textit{présent}\}$$

.

Attribut manquant	Nombre de traverses minimales dans le complémentaire	Nombre de traverses minimales de support non nul dans le contexte initial
chd	7 453 135	81
chddim	1 050 017	167
chg	7448 754	80
chgdim	1 019 500	401
cbddim	3 072 860	690
cbgdim	3 456 431	1 017
axddim	1 367 206	1497
axgdim	1 564 900	2 026
mtr	5 249 326	1 895
ldh	9 550 812	410
polyn2	8 580 356	1 029
cbg	10 266 888	27
axg	10 224 757	49
axd	10 193 813	26
med	10 259 803	19
hild	10 005 456	259
hilg	10 021 572	269
gb	9 885 615	362
cbd	10 226 183	10
vs	10 091 973	16
ext	10 116 561	18
sg	10 284 404	13
age	10 311 217	1
histo2	10 296 212	15

TAB. 5.6 – Nombres de traverses minimales par attribut manquant sur les données de la maladie de HODGKIN.

Notation 2 *Le contexte complémentaire projeté sur l'item i est noté $\overline{\mathcal{K}}_i$.*

Définition 24 (anti-fréquence) *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte et $X \subset \mathcal{I}$ un motif. L'*anti-frequence *est définie par :*

$$\overline{\mathcal{F}}(X, \mathcal{K}) = |\{o \in \mathcal{O} \mid X \cap o = \emptyset\}|.$$

Notons que l'anti-fréquence, comme la fréquence, est décroissante selon la taille du motif. Elle peut donc être utilisée comme contrainte anti-monotone. De plus, l'anti-fréquence caractérise les traverses d'un hypergraphe : ce sont les motifs d'anti-fréquence nulle. En effet, lorsque $\overline{\mathcal{F}}(X, \mathcal{K}) = 0$, cela veut dire que X intersecte tous les objets de $\mathcal{K}$ et par conséquent X est une traverse de $\mathcal{H}_\mathcal{K}$ (l'hypergraphe formé à partir du contexte $\mathcal{K}$).

MTBIPMINER nécessite le calcul des traverses minimales à partir des complémentaires des objets qui contiennent l'item à caractériser, c'est-à-dire à partir du contexte complémentaire projeté sur i ($\overline{\mathcal{K}}_i$). De plus, ces traverses minimales doivent être fréquentes dans le contexte initial ($\mathcal{K}$) selon un seuil de support γ. MTMINER extrait les traverses minimales par niveau, à la façon d'Apriori [Agrawal et Srikant, 1994]. Pour cela, il considère que X est une traverse de $\mathcal{H}_\mathcal{K}$ si elle satisfait la contrainte $\overline{\mathcal{F}}(X, \mathcal{K}) = 0$. Pour obtenir des traverses qui sont minimales, MTMINER ne génère un candidat que si son anti-fréquence décroit strictement.

Pour MTBIPMINER, les candidats générés doivent en outre respecter une contrainte anti-monotone, la fréquence de la traverse. Lors de la procédure de génération *apriori_gen_minimal*, on génère $Y = Xi_1i_2$ à partir de Xi_1 et Xi_2 si tous les sous-ensembles Y' de Y :

1. satisfont la contrainte de fréquence $\mathcal{F}(Y', \mathcal{K}) >= \gamma$;
2. ont une anti-fréquence strictement supérieure à celle de Y : $\overline{\mathcal{F}}(Y', \overline{\mathcal{K}}_i) > \overline{\mathcal{F}}(Y, \overline{\mathcal{K}}_i)$.

Le pseudo-code de l'algorithme MTBIPMINER est donné par l'algorithme 2. Il permet de calculer les implications propres fréquentes concluant sur un item i. L'ensemble $\mathcal{BIP}_k$ sert à stocker les prémisses des implications propres de taille k, concluant sur l'item i, tandis que l'ensemble $\mathcal{C}and_k$ contient les motifs pour générer les prémisses candidates au niveau $k+1$.

Algorithme 2 : Calcul des implications propres fréquentes concluant sur l'item i.

Données : un contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, un item i et un seuil de support γ.

Résultats : l'ensemble $\mathcal{BIP}$ des implications propres fréquentes concluant sur i.

```
1  début
2      k ← 1;
3      Cand_1 ← {j ∈ I | j ≠ i};
4      répéter
5          BIP_k ← {X ∈ Cand_k | F̄(X, K̄_i) = 0 ∧ F(X, K) >= γ};
6          Cand_k ← {X ∈ Cand_k | F̄(X, K̄_i) ≠ 0 ∧ F(X, K) >= γ};
7          Cand_{k+1} ← apriori_gen_minimal(Cand_k);
8          k++;
9      jusqu'à Cand_k = ∅;
10     retourner (BIP = ⋃_{n=1..k} BIP_n);
11 fin
```

Exemple 11 *Dans cet exemple, nous montrons une trace d'exécution de l'algorithme* MTBIPMINER *en considérant une valeur de* $\gamma = 2$ *et* $i = vm(A_3)$.

1. *Initialisation de l'ensemble* $\mathcal{C}and_1$=$\{a, b, c, d, e, f, g, h, i, mv(A_1), vm(A_2), vm(A_4)\}$;
2. *Itération 1 :*
   - $\mathcal{BIP}_1$=$\emptyset$ *(voir table 5.7) ;*
   - $\mathcal{C}and_1$ *devient* $\{a, c, d, f, g, h, vm(A_1), vm(A_2), vm(A_4)\}$, *puisque b, e et i sont infréquents ;*
   - *Suite à l'application de apriori_gen_minimal*($\mathcal{C}and_1$), $\mathcal{C}and_2$=$\{ac, ah, ch\}$.
3. *Itération 2 :*
   - $\mathcal{BIP}_2$=$\{ac, ah, ch\}$;
   - $\mathcal{C}and_2$ *devient égal à* $\emptyset$;
   - $\mathcal{C}and_3$=$\emptyset$;
4. *L'algorithme s'arrête. Ainsi,* $\mathcal{BIP}_{vm(A_3)} = \{ac, ah, ch\}$.

X	$\mathcal{F}(X, \mathcal{K})$	$\overline{\mathcal{F}}(X, \overline{\mathcal{K}}_{vm(A_3)})$
a	4	2
b	1	1
c	4	2
d	2	2
e	1	1
f	3	2
g	2	2
h	3	1
i	0	0
$vm(A_1)$	3	3
$vm(A_2)$	2	1
$vm(A_4)$	5	5

TAB. 5.7 – Calcul effectué par MTBIPMINER pour l'initialisation de $\mathcal{BIP}_1$.

Pour la caractérisation des valeurs manquantes, nous calculons les implications propres concluant sur des valeurs manquantes. Notons qu'il serait intéressant de limiter le nombre d'items des prémisses des implications propres afin d'obtenir des explications pertinentes des valeurs manquantes. En effet, une caractérisation faisant appel à la conjonction de plus de cinq items par exemple n'est pas réaliste. Si besoin est, il est aisé de modifier l'algorithme 2 pour ne retenir que les traverses d'une longueur k fixée au préalable.

Dans la sous-section suivante, nous analysons la performance de notre algorithme MTBIPMINER, en termes de temps d'exécution.

5.3.2 Étude de performance

Les tables 5.8 et 5.9 montrent le temps d'exécution nécessaire au calcul des implications propres par filtrage des traverses minimales fréquentes de l'algorithme proposé dans [Kavvadias et Stavropoulos, 2005] et de notre algorithme MTBIPMINER, avec $\gamma = 1$, sur les données de la MENINGITE ainsi que celles de HODGKIN. Les résultats sont détaillés selon la valeur manquante sur laquelle concluent les implications propres. Tous les temps d'exécution sont donnés en secondes. La dernière colonne des tableaux indique le rapport entre les temps d'exécution des deux algorithmes.

Nous constatons que sur les données de la MENINGITE, notre algorithme MTBIPMINER est beaucoup plus rapide que le filtrage de celui de [Kavvadias et Stavropoulos, 2005] : entre 20 et 30 fois. Sur les données de HODGKIN, c'est plus disparate, notre algorithme est de 4 à plusieurs centaines de milliers plus rapide. Ceci s'explique par le fait que l'algorithme STAVRO calcule toutes les traverses minimales du contexte complémentaire, même si leur support est nul dans le contexte initial. Tandis que le nôtre se focalise sur les traverses fréquentes. Comme nous l'avons montré dans la sous-section 5.2.2, l'écart entre le nombre de traverses minimales dans le complémentaire et le nombre de traverses minimales ayant un support non nul dans le contexte initial est très important, et explique le gain de performance de notre méthode.

Ces expériences prouvent l'intérêt pratique de MTBIPMINER : l'extraction des implications propres avec de très faibles valeurs de support est rendue praticable. MTBIPMINER doit son succès en termes de temps d'exécution à l'ajout d'une contrainte de fréquence qui garantit l'obtention des traverses minimales fréquentes. Il est ainsi plus performant que le filtrage de l'algorithme classique [Kavvadias et Stavropoulos, 2005] de calcul de traverses minimales, d'une part. D'autre part, il est plus performant qu'une solution *naïve* de filtrage du résultat d'un algorithme classique de calcul de règles, puisque MTBIPMINER permet directement d'extraire les implications propres concluant sur des valeurs manquantes.

5.4 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle méthode de calcul de la base d'implications propres basée sur les traverses minimales d'hypergraphe. L'algorithme MTBIPMINER adapte MTMINER, qui calcule les traverses minimales, en lui ajoutant une contrainte sur le support des traverses calculées. Ceci permet de calculer efficacement la base des implications propres nécessaire pour la caractérisation des valeurs manquantes.

Attribut manquant	STAVRO	MTBIPMINER	Rapport
chd	69275	3585	19
chddim	1139	275	4
chg	68456	3771	18
chgdim	1128	413	3
cbddim	12185	2569	5
cbgdim	16370	2974	6
axddim	1239	280	4
axgdim	1816	430	4
mtr	32258	4183	8
ldh	103682	3243	32
polyn2	89529	2260	40
cbg	114921	192	599
axg	69275	42	1 649
axd	113615	34	3 342
med	114735	93	1 234
hild	110995	125	888
hilg	111111	136	817
gb	110840	114	972
cbd	114004	100	1 140
vs	113412	750	151
ext	112235	229	490
sg	115392	0,44	262 255
age	118640	0,16	741 500
histo2	119159	205	581

TAB. 5.8 – Temps de calcul du filtrage des traverses minimales de l'algorithme de [Kavvadias et Stavropoulos, 2005] et celui de MTBIPMINER sur les données de HODGKIN.

Attribut manquant	STAVRO	MTBIPMINER	Rapport
tonus	36	1,76	20
polysang	2	1,53	14
polyns	38	1,05	36
gluc	15	1,11	14
vs	36	1,29	28
leuco	36	1,31	27
dfievre	34	1,87	18

TAB. 5.9 – Temps de calcul du filtrage des traverses minimales de l'algorithme de [Kavvadias et Stavropoulos, 2005] et celui de notre algorithme MTBIPMINER sur les données de la MENINGITE.

Chapitre 6

Complétion contextualisée des valeurs manquantes

Sommaire

Dans ce chapitre, nous utilisons la typologie des valeurs manquantes définie dans le chapitre 4 afin de proposer une technique de complétion *contextualisée*, dont les actions diffèrent selon le type de la valeur manquante (aléatoire, directe, indirecte, hybride) ainsi que sur l'objet portant la valeur manquante. Notons que nous nous intéressons ici à la complétion des valeurs manquantes non aléatoires [Ben Othman *et al.*, 2009a]. Pour la complétion des valeurs manquantes aléatoires, nous préconisons la méthode $\mathcal{GBAR}_{\mathcal{MVC}}$ [Ben Othman et Ben Yahia, 2008] décrite dans le chapitre 1.

Les caractéristiques de notre complétion des valeurs manquantes non aléatoires se présentent comme suit :

1. Une complétion caractéristique du type de la valeur manquante ;
2. Une complétion *contextualisée* par « valeur spéciale », caractéristique de l'origine d'une valeur manquante, en étendant l'ensemble de définition des attributs.

6.1 Contextualisation de la complétion

Notre méthode consiste à tirer profit, pour la phase de complétion, de l'information supplémentaire obtenue lors de la phase de caractérisation. La complétion que nous proposons est une complétion *contextualisée* : elle prend différentes formes, suivant le type et le contexte de la valeur manquante.

6.1.1 Contextualisation selon le type aléatoire/non aléatoire

Notre caractérisation distingue les valeurs manquantes aléatoires de celles qui ne sont pas aléatoires. Quand il s'agit de valeur manquante non aléatoire, nous disposons, sous la forme d'une règle, d'une information relative à son contexte d'apparition, symbolisée par la prémisse de cette règle. Nous proposons dans ce cas de fournir une valeur de remplacement qui symbolise ce contexte d'apparition ou *origine*. Le remplacement d'une valeur manquante par son origine se justifie par le fait que, dans certaines situations, la complétion par une valeur du domaine de définition n'a pas vraiment de sens. Par exemple, les valeurs manquantes qui ne sont pas aléatoires cachent parfois des valeurs inapplicables, des valeurs particulières ou restreintes, *etc.* Dans ce cas, il est inapproprié de compléter par des valeurs du domaine de définition.

6.1.2 Contextualisation selon l'objet

La complétion que nous proposons est également contextualisée selon l'objet où apparaît une valeur manquante.

Le principe est le suivant : bien que deux valeurs manquantes sur un même attribut soient caractérisées par un même type, ceci ne conduit pas à leur complétion par une même valeur. Reprenons notre exemple de la typologie (table 6.1) constituée des règles de la table 6.2. Les valeurs manquantes sur les objets o_4 et o_7, affectant l'attribut A_4, ont toutes les deux été caractérisées en tant que directes. Cependant, ces deux valeurs manquantes sont expliquées par deux règles différentes : $d \rightarrow vm(A_4)$ et $g \rightarrow vm(A_4)$. Par conséquent, elles n'ont pas la même origine et elles seront donc remplacées par deux valeurs différentes. Ainsi, la contextualisation traite différemment les valeurs manquantes selon l'objet où elles apparaissent et suivant leurs origines respectives.

Dans la section suivante, nous montrons comment combiner les règles de caractérisation afin d'effectuer la complétion.

	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}

TAB. 6.1 – Typologie des valeurs manquantes.

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

TAB. 6.2 – Implications propres concluant sur une valeur manquante ; $minsup = 2$.

6.2 Schémas de caractérisation

Nous assimilons notre ensemble de règles de caractérisation à un ensemble de graphes orientés, que nous appelons *schémas de caractérisation*. Pour chaque objet, un *schéma de caractérisation* synthétise les règles de caractérisation supportées par l'objet. Les nœuds sources des *schémas de caractérisation* sont des valeurs connues ou des valeurs manquantes et constituent les prémisses des règles de caractérisation, tandis que les nœuds cibles sont exclusivement des valeurs manquantes et constituent les conclusions de ces règles.

La figure 6.1 montre les différentes représentations des différents types (aléatoire, direct, indirect et hybride) où les valeurs manquantes sont illustrées par des ellipses, tandis que les valeurs connues sont indiquées par des rectangles. Dans cette figure, la caractérisation porte sur la valeur manquante de l'attribut A (notée $vm(A)$).

Les schémas de caractérisation relatifs à notre typologie de la table 6.1 sont donnés par la figure 6.2. Notons que certains objets peuvent présenter le même *schéma de caractérisation*, lorsqu'ils vérifient les mêmes règles de caractérisation. Par exemple, les objets $\{o_1, o_3\}$ partagent le même schéma ainsi que les objets $\{o_2, o_5\}$.

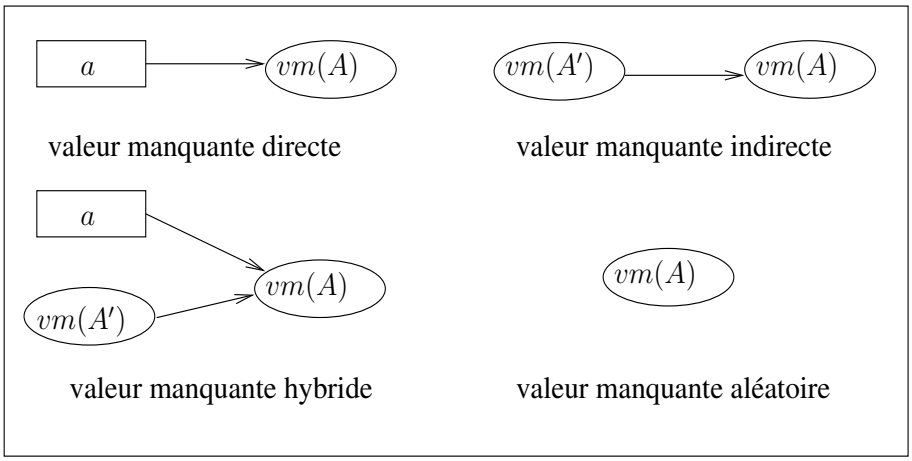


FIG. 6.1 – Les représentations des différents types des valeurs manquantes.

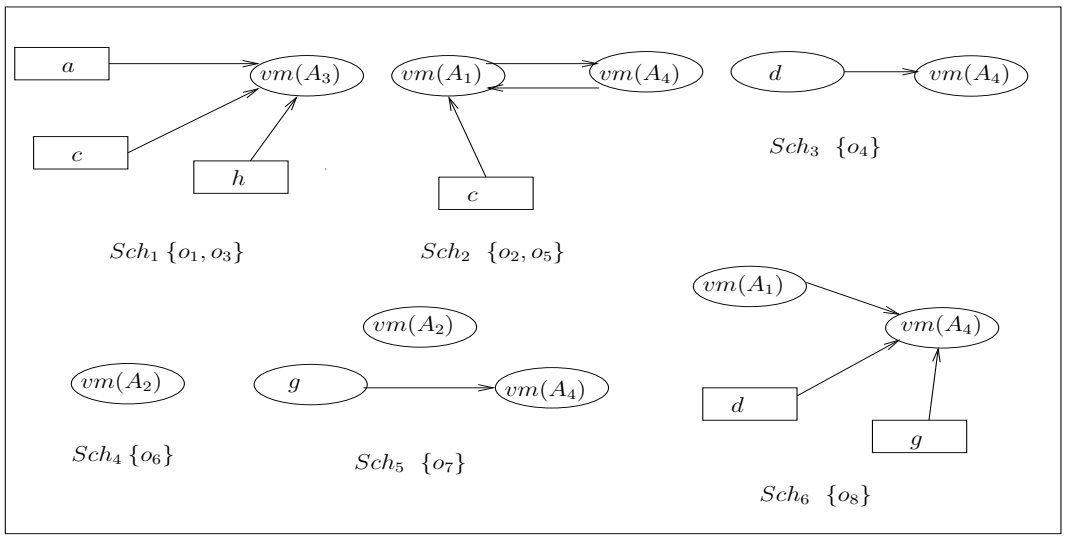


FIG. 6.2 – Schémas de caractérisation relatifs à la typologie de la table 6.1.

Représenter les règles par des *schémas de caractérisation* permet d'exhiber toutes les relations entre les valeurs manquantes sur un objet donné et de remonter à leurs origines. Ces origines constitueront par la suite les valeurs de complétion. De plus, un *schéma de caractérisation* permet de considérer les caractérisations multiples et d'identifier les situations particulières telles que les caractérisations cycliques. L'utilisation de graphes orientés est utile puisqu'elle nous permettra par la suite de lever les ambiguïtés sur les origines des valeurs manquantes indirectes et hybrides en appliquant une réduction des caractérisations cycliques.

Afin de retrouver les origines des valeurs manquantes, nous procéderons également à une propagation transitive de ces origines.

6.2.1 Propagation transitive des origines d'une valeur manquante

L'examen d'une règle de caractérisation nous renseigne le plus souvent sur l'origine d'une valeur manquante. Ceci semble bien fonctionner dans le cas d'une valeur manquante directe, où la prémisse de la règle de caractérisation constitue incontestablement son origine. La valeur manquante sur l'attribut A_4 et affectant l'objet o_4 (table 6.1) a pour origine l'item d. En revanche, si nous considérons une valeur manquante indirecte ou hybride, quel est son origine ? Par exemple, quel serait l'origine de la valeur manquante $vm(A_4)$ affectant l'objet o_2 ? D'après la caractérisation, cette valeur présente à son tour comme origine la valeur manquante $vm(A_1)$. Ainsi, l'observation d'une règle de caractérisation dans le cas de valeurs manquantes indirectes ou hybrides ne nous conduit pas directement à l'origine de la valeur manquante. Il faudra dans ce cas procéder à une propagation transitive de l'origine de la valeur manquante, où l'item c se propagera comme origine de la valeur manquante $vm(A_4)$ par transitivité de $c \rightarrow vm(A_1) \rightarrow vm(A_4)$.

La difficulté de la propagation de l'origine d'une valeur manquante dépend de la complexité du *schéma de caractérisation*. Dans ce qui suit, nous nous intéresserons particulièrement aux caractérisations cycliques.

6.2.2 Réduction cyclique de la caractérisation

Une configuration cyclique se présente lorsqu'au moins deux valeurs manquantes s'expliquent mutuellement. Cette configuration pose problème au niveau de la propagation de l'origine. Par exemple, dans la figure 6.2, on peut voir un exemple de cycle entre les valeurs manquantes des attributs A_1 et A_4 du *schéma de caractérisation* Sch_2, associé aux objets $\{o_2, o_5\}$. Afin de remonter aux origines respectives de ces deux valeurs manquantes, nous sommes confronté à une recherche de cycles. Pour détecter un cycle de longueur minimale dans un *schéma de caractérisation* :

- on calcule la *girth*, longueur du plus petit cycle, en calculant les puissances successives de la matrice d'adjacence M : dès qu'un 1 apparaît sur la diagonale de M^n, c'est qu'un cycle de longueur n existe à partir du sommet S concerné.
- avec l'algorithme Dijkstra, on recherche le plus court chemin entre S et l'un de ses prédécesseurs. Si ce chemin est de longueur n, c'est un des plus petits cycles dans le graphe.

Une fois qu'un cycle de longueur minimale est détecté, les nœuds du cycle sont fusionnés. Par exemple, le schéma associé aux objets $\{o_2, o_5\}$ sera réduit tel que présenté dans la figure 6.3 (Gauche). En pratique, les configurations cycliques sont beaucoup plus complexes. Nous illustrons à la figure 6.4 (Gauche) un *schéma de caractérisation* cyclique pour les données de la Meningite. La même figure (Droite) montre le même schéma après réduction des cycles.

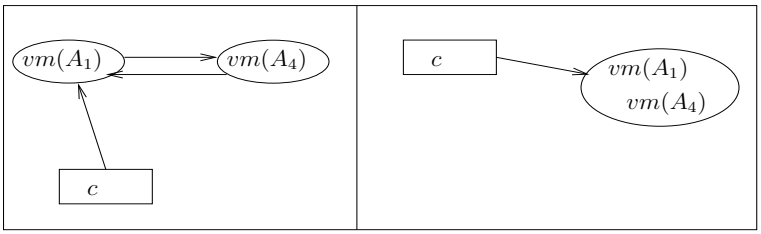


FIG. 6.3 – Schéma de caractérisation Sch_2 relatif aux objets $\{o_2, o_5\}$. **Gauche** : avant réduction du cycle. **Droite** : après réduction du cycle.

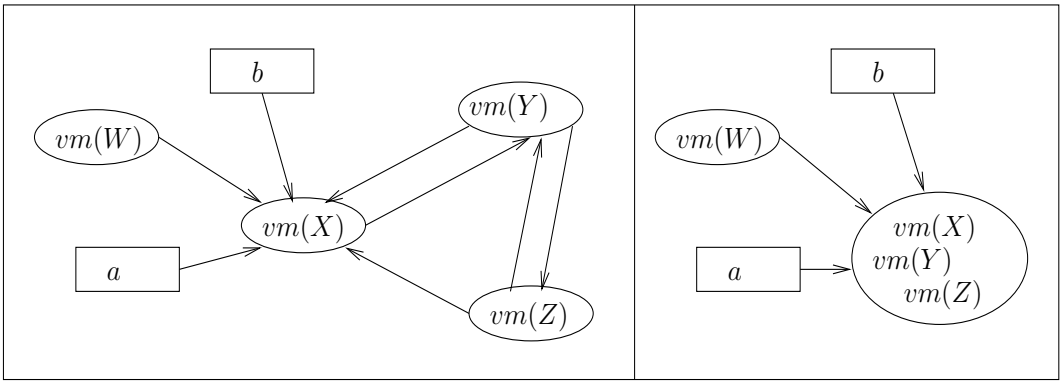


FIG. 6.4 – Schéma de caractérisation sur les données de la MENINGITE. **Gauche** : avant réduction du cycle. **Droite** : après réduction du cycle.

6.3 Méthode de complétion contextualisée des valeurs manquantes

La méthode de complétion suit les étapes suivantes :

1. Réduction des cycles et propagation transitive des origines des valeurs manquantes, dans les *schémas de caractérisation* ;

2. Tri topologique sur les *schémas de caractérisation*, pour définir l'ordre de remplacement des valeurs manquantes concernées.

3. Au final, deux situations peuvent se présenter : soit nous trouvons une origine à la valeur manquante, qui constituera la valeur de complétion ; soit il n'y a pas d'origine, nous dirons que la valeur manquante est aléatoire. Dans ce cas, elle sera complétée par exemple par la méthode $\mathcal{GBAR}_{\mathcal{MVC}}$ [Ben Othman et Ben Yahia, 2008].

La table 6.3 montre le contexte augmenté à l'aide de ces différentes valeurs de complétion, indiquées par un cadre. Par exemple, cette table indique que les valeurs manquantes sur les objets o_2, o_5 et affectant l'attribut A_1 ont comme origine la présence de l'item [c].

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

TAB. 6.3 – Contexte augmenté des nouvelles valeurs de complétion des valeurs manquantes non aléatoires.

6.4 Conclusion

Les méthodes classiques de complétion des valeurs manquantes ont toujours été limitées aux valeurs manquantes aléatoires. Notre principale contribution pour ce problème est la proposition d'une complétion, qui prend en considération le type de la valeur manquante et fournit un traitement en conséquence. Dans ce chapitre, nous avons exploité la caractérisation des valeurs manquantes afin de proposer une nouvelle méthode fine de complétion, que nous avons qualifiée de *contextualisée*. Cette contextualisation est liée au type (aléatoire/non aléatoire) de la valeur manquante d'une part, et à la finesse de la caractérisation, d'autre part, spécialisée selon les objets. Nous avons montré qu'il est possible de retrouver l'origine d'une valeur manquante à travers la construction des schémas de caractérisation dont on réduit les cycles. Nous avons exploité ces origines afin de proposer des valeurs de complétion aux manquantes non aléatoires. Dans le chapitre suivant, nous montrons l'intérêt pratique de notre complétion.

Chapitre 7

Évaluation

Sommaire

Dans ce chapitre, nous commençons d'abord par mener des expériences pour évaluer la pertinence d'une méthode de classification supervisée pour mesurer l'intérêt d'une complétion de données. Pour cela, nous considérons disposer de bases incomplètes et de leur complétion idéale, et nous comparons les performances en classification des modèles correspondants. Nous montrons qu'il est difficile d'utiliser un tel schéma de validation, car les performances des données incomplètes sont souvent meilleures que celles de la complétion idéale. Dans la deuxième partie de ce chapitre, nous mettons en place une nouvelle technique d'évaluation de méthode de complétion, qui se base sur la stabilité d'une méthode de classification non supervisée.

7.1 Comment introduire des valeurs manquantes non aléatoires ?

Dans cette section, nous expliquons le protocole implémenté pour introduire des valeurs manquantes non aléatoires dans les données. Ce protocole sera employé lors de nos expériences pour mesurer l'impact de notre méthode de complétion.

Nous voulons introduire des valeurs manquantes selon la typologie mise en évidence dans ce travail. Plus précisément, nous voulons insérer des valeurs manquantes non aléatoires (directe, indirecte et hybride).

L'insertion de valeurs manquantes artificielles utilise des modèles à base de règles (figure 7.1), plus précisément, des implications propres. Les paramètres employés lors de ce protocole d'insertion des valeurs manquantes sont les suivants :

- le support minimum *minsup* des implications propres utilisées ;
- la proportion de valeurs manquantes à insérer.

Disposant d'une liste d'implications propres de support compris entre $\frac{minsup}{2}$ et $minsup \times 2$, classées par supports décroissants, la règle la plus fréquente est utilisée pour insérer des valeurs manquantes dans les objets concernés. Ce processus est répété tant que l'on n'a pas obtenu la proportion de valeurs manquantes désirée. Notons que cette proportion de valeurs manquantes peut ne pas être atteinte, lorsqu'il y a peu d'implications propres dans la gamme de support, ou lorsque les règles supportent les mêmes objets.

Dans nos expériences, *minsup* prend les valeurs de 2, 4, 8 et 16%, la proportion attendue est fixée à 2, 4, 8 et 16% également. Selon les bases de données, nous obtenons ainsi des taux maximaux de valeurs manquantes jusqu'à 8 à 18%. L'utilisation de ces deux paramètres permet cependant de varier à la fois la quantité de valeurs manquantes introduites et la taille des effacements réalisés pour y parvenir. Nous présentons les résultats en fonction de la proportion réelle de valeurs manquantes présentes dans les données générées.

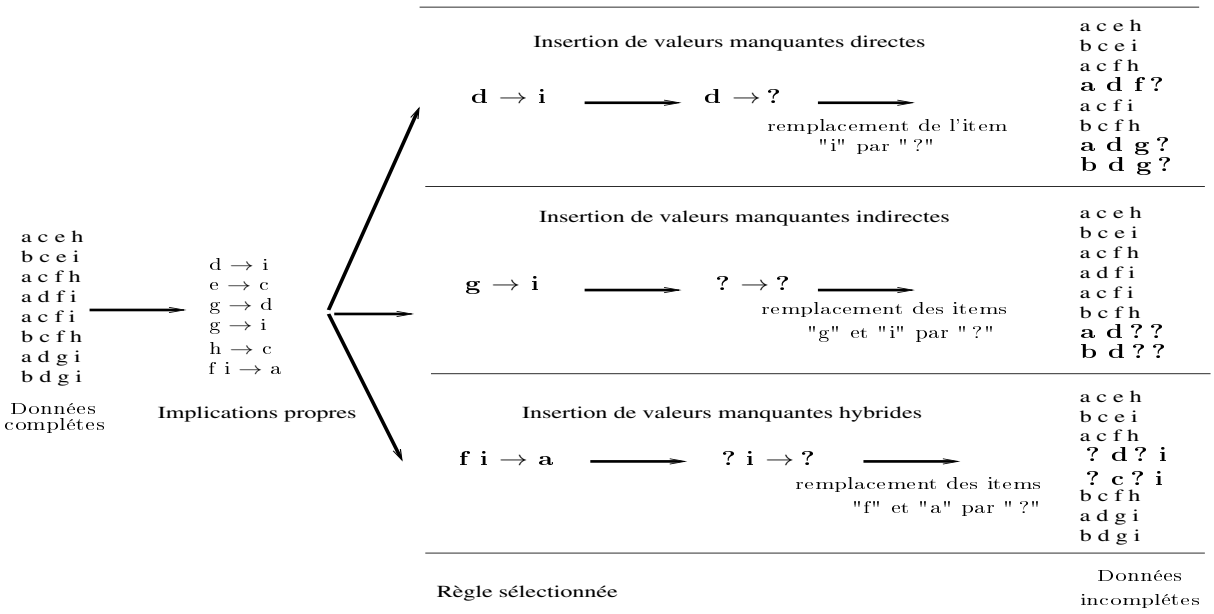


FIG. 7.1 – Protocole d'introduction artificielle de valeurs manquantes non aléatoires.

Dans la section suivante, nous menons des expériences afin d'évaluer la pertinence des méthodes d'apprentissage supervisé comme technique d'évaluation de méthodes de complétion.

7.2 Évaluation selon les techniques supervisées

Les expériences relatées ici se fondent sur le concept de *complétion idéale*, détaillé ci-dessous.

7.2.1 Complétion idéale

D'après les techniques d'évaluation de l'état de l'art présentées dans le chapitre 2, on attend naturellement de la complétion des valeurs manquantes qu'elle satisfasse deux qualités : proximité avec les données de référence et amélioration d'une méthode d'apprentissage. Afin de focaliser l'attention sur les performances de la complétion sur les techniques d'apprentissage, nous supposons disposer d'une complétion idéale, qui satisfait le premier critère (proximité avec les données de référence). L'originalité de notre approche est représentée par la figure 7.2. Plutôt que de partir d'une base complète, en effacer des valeurs puis les compléter, nous partons d'un contexte incomplet (généré par effacement de valeurs sur un contexte $\mathcal{K}$) et prétendons qu'un *oracle* nous a fourni sa complétion idéale $\mathcal{K}$ (les données de référence). Ce modeste point de vue optimise la proximité de la complétion, et permet de focaliser l'attention sur les performances de la complétion idéale en classification.

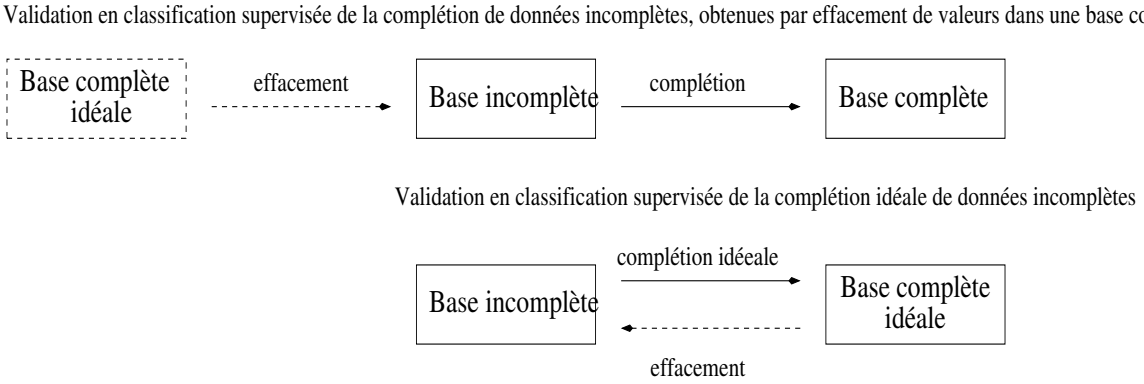


FIG. 7.2 – En haut, la validation classique d'une complétion. En bas, la validation de la complétion idéale.

Nous formalisons la complétion *idéale* comme suit :

Definition 1 (complétion idéale) *Soit $mv()$ un opérateur d'effacement et $\mathcal{K}$ un contexte complet. La* complétion idéale *du contexte incomplet $mv(\mathcal{K})$ est un opérateur de complétion $ideal()$ tel que $ideal(mv(\mathcal{K})) = \mathcal{K}$. Nous dirons également que la complétion idéale de $mv(\mathcal{K})$ est $\mathcal{K}$.*

7.2.2 Protocole de mesure de l'impact de la complétion

Pour mesurer l'impact de la complétion idéale sur la classification supervisée, nous disposons d'un contexte incomplet $mv(\mathcal{K})$ et de sa complétion idéale $\mathcal{K}$. Pour chaque partition de ces deux contextes en échantillon d'apprentissage et échantillon de test, nous comparons les expériences suivantes :

1. calcul d'un score de référence en classification supervisée, le modèle est obtenu sur l'échantillon d'apprentissage de $\mathcal{K}$, et appliqué sur l'échantillon de test de $\mathcal{K}$;
2. mesure des performances du modèle de $\mathcal{K}$ appliqué sur l'échantillon de test de $mv(\mathcal{K})$;
3. calcul du score obtenu par le modèle calculé sur l'échantillon d'apprentissage de $mv(\mathcal{K})$, appliqué sur l'échantillon de test de $\mathcal{K}$ puis de $mv(\mathcal{K})$.

Quatre expériences sont schématisées par la figure 7.3, avec les notations de la table 7.1. Dans la suite, nous examinons séparément les résultats avec deux classifieurs : l'un à base d'arbres de décision et l'autre à base de règles d'association.

	Apprentissage	
Test	$\mathcal{K}$	$mv(\mathcal{K})$
$\mathcal{K}$	$ideal(ideal)$	$mv(ideal)$
$mv(\mathcal{K})$	$ideal(mv)$	$mv(mv)$

TAB. 7.1 – Notation des expériences.

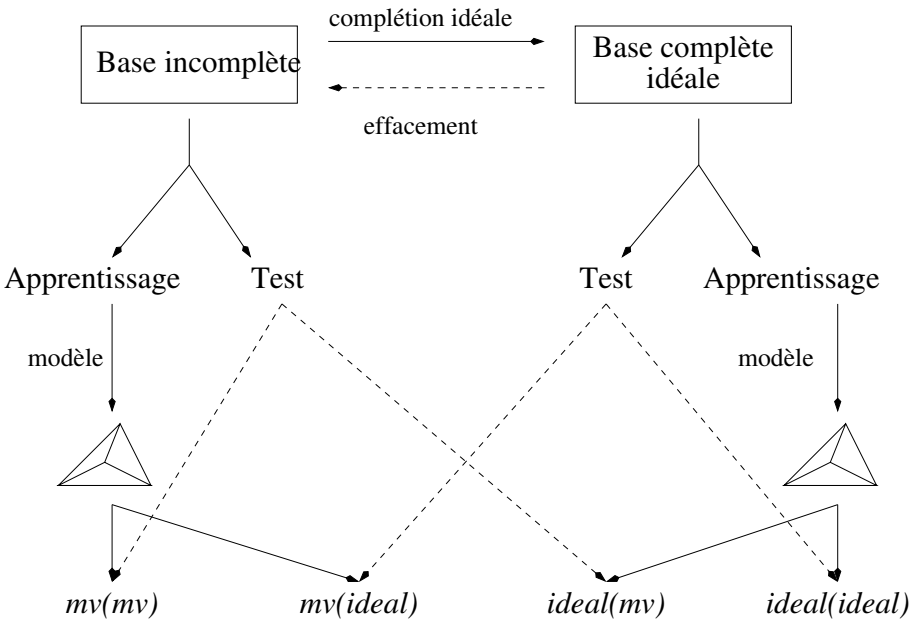


FIG. 7.3 – Protocole expérimental.

7.2.3 Discussion

Nous avons mené les expériences décrites à la section précédente sur des données complètes de l'UCI [Asuncion et Newman, 2007]. Ces expériences comparent les performances obtenues en classification supervisée par les modèles issus de données incomplètes et de leur complétion idéale, avec des classifieurs à base d'arbres de décision ou de règles d'association.

Les données incomplètes résultent de l'effacement de valeurs (sauf sur l'attribut de classe) selon une probabilité uniforme variant de 1 à 20%. Les scores de classification sont obtenus en 10-validation croisée. Les trois courbes $mv(ideal)$, $ideal(mv)$ et $mv(mv)$ coïncident sur l'axe des ordonnées avec le score de référence $ideal(ideal)$, quand le taux de valeurs manquantes est nul. La courbe en trait plein indique le score $ideal(mv)$ obtenu par le modèle des données idéales.

La première série d'expériences (figure 7.4) indique les résultats pour les arbres de décision (méthode C4.5, implémentation de Weka[7] [Witten et Frank, 2005]). Les valeurs manquantes sont prises en charge selon la technique décrite dans le chapitre 1 (page 17). La deuxième série d'expériences (figure 7.5) permet une discussion sur les modèles à base de règles d'association [Li *et al.*, 2001]. Les valeurs manquantes y sont ignorées, c'est-à-dire qu'elles sont considérées comme absentes.

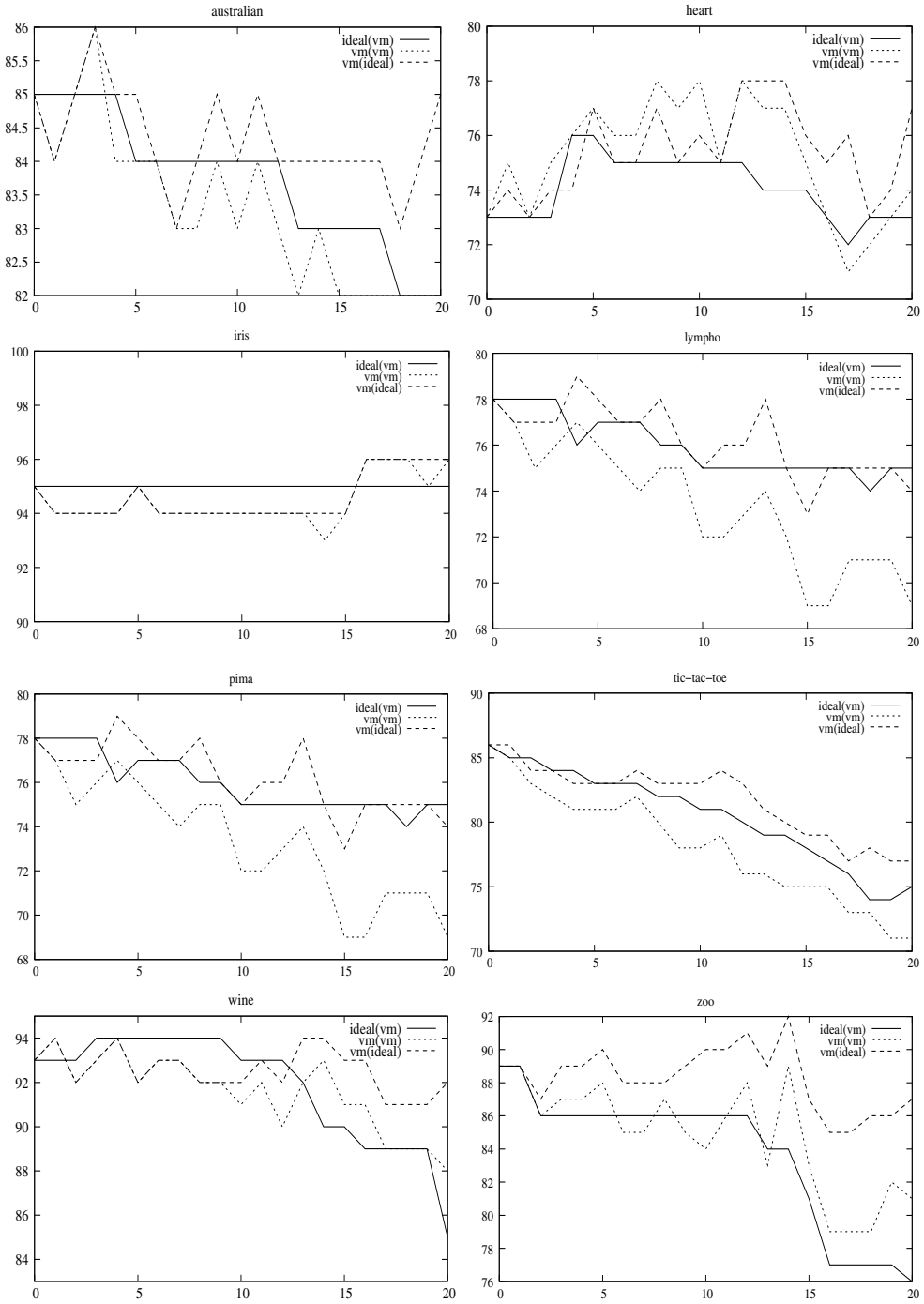
Impact des valeurs manquantes sur un modèle à base d'arbre de décision

En examinant le résultat obtenu avec l'algorithme C4.5 (figure 7.4), nous remarquons que $ideal(vm)$ est souvent au dessus de la ligne en fin pointillés $vm(vm)$. L'allure moyenne est celle observée sur la base tic-tac-toe. Dans ce cas, on peut dire que le modèle idéal fournit un meilleur score de classification. Ce premier résultat confirme, dans le cas des arbres de décision, que les performances d'un classifieur se dégradent en présence de valeurs manquantes.

Dans quelques cas peu significatifs (sauf sur la base heart), la performance de référence est dépassée. Cependant, il est difficile d'en tirer des conclusions opérationnelles. Nous verrons au paragraphe suivant, concernant les modèles à base de règles, que ce résultat y est plus tangible que pour les arbres de décision.

On remarque enfin que le modèle calculé sur les données incomplètes et appliqué sur les données complètes (courbes $vm(ideal)$ en pointillés larges) est plus performant que le modèle idéal appliqué aux données incomplètes. Pour les arbres de décision, l'impact des valeurs manquantes est plus sensible sur les objets à classer que sur le calcul du modèle. Bien-sûr, ce résultat dépend de la méthode utilisée pour prendre en compte les valeurs manquantes dans l'algorithme de décision. Nous estimons cependant qu'une bonne complétion à destination des arbres de décision devrait être évaluée selon l'amélioration apportée sur les objets de test et non pas sur la qualité du modèle calculé.

[7] Disponible à l'adresse : http ://www.cs.waikato.ac.nz/ml/weka/



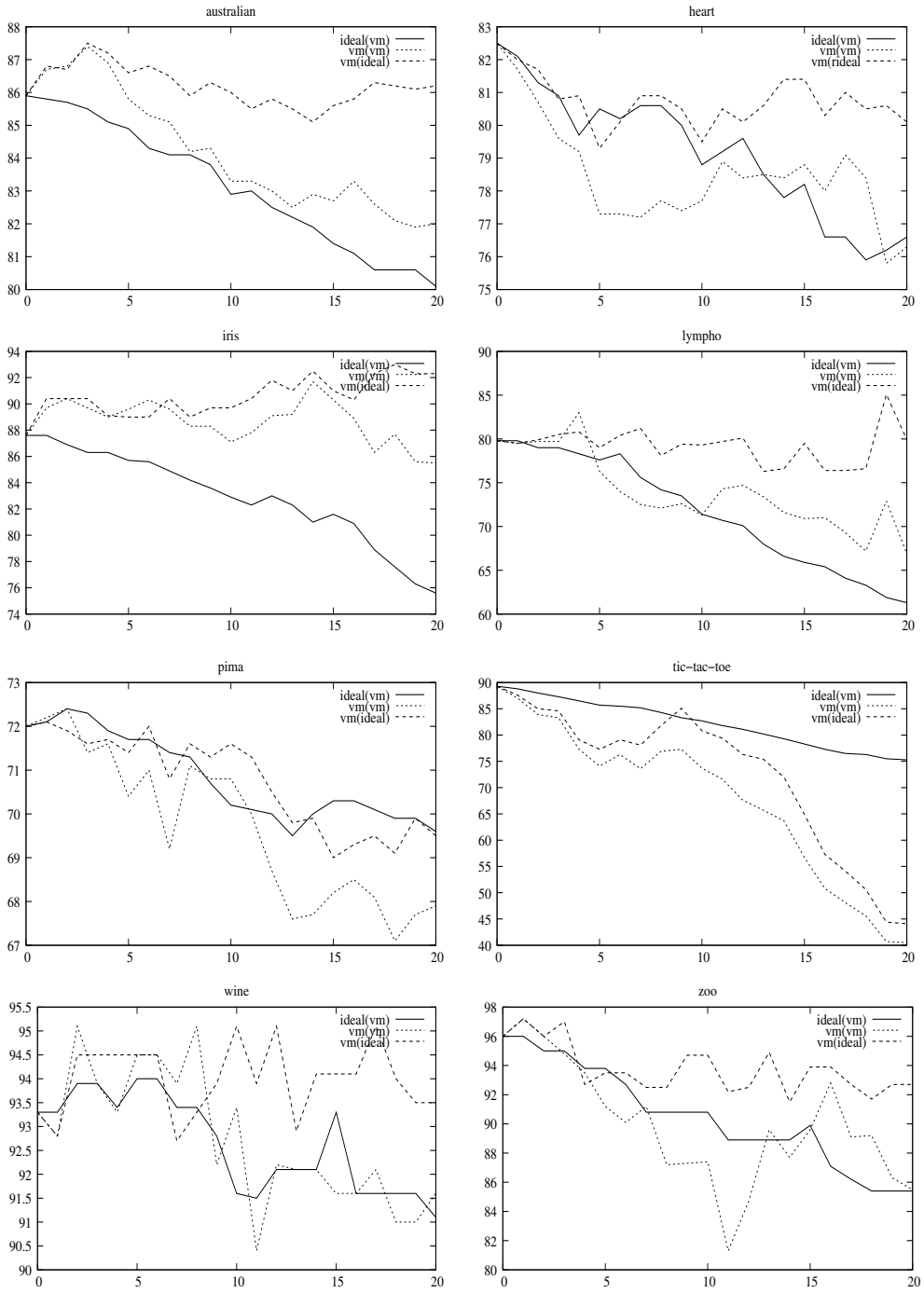
Légende : L'axe des abscisses indique le pourcentage des valeurs manquantes .

FIG. 7.4 – Résultat de la classification supervisée à l'aide de C4.5.

Impact des valeurs manquantes sur le modèle de classification à base de règles d'association

La figure 7.5 montre les résultats de nos expériences avec un classifieur à base de règles. Comme précédemment, on constate que le modèle idéal ($ideal(vm)$, ligne pleine) ne fournit pas les

meilleurs performances ; le résultat est meilleur sur des données complètes en appliquant le modèle obtenu sur les données incomplètes ($vm(ideal)$, pointillés larges). Il semble là encore difficile de prétendre qu'une méthode de complétion idéale peut être évaluée en examinant l'apport du modèle à base de règles correspondant. À la différence des expériences sur les arbres de décision,



Légende : L'axe des abscisses indique le pourcentage des valeurs manquantes.

FIG. 7.5 – Résultats de la classification supervisée à l'aide de règles d'association.

le score obtenu sur les données incomplètes sans traitement ($vm(vm)$) est souvent meilleur que celui fourni par le modèle idéal. Cette constatation réduit encore la pertinence d'une évaluation par la classification à base de règles, car il semble difficile de valoriser un modèle obtenu sur les données complétées, même idéalement.

On observe en outre qu'une faible quantité de valeurs manquantes améliore sensiblement les scores de classification. Pour la classification supervisée à base de règles d'association, l'effacement de quelques données améliore la pertinence des modèles et la classification des échantillons de test. Là encore nous évaluons un phénomène lié à la prise en compte *transactionnelle* des valeurs manquantes pour le calcul des règles et la décision sur les objets de tests : ce qui est manquant est considéré absent. Cette considération induit en elle-même une méthode de complétion, certes triviale. Elle montrera cependant souvent de meilleures performances qu'une méthode complexe de complétion.

La discussion peut également être menée sur l'intérêt d'une méthode de complétion pour premièrement le calcul du modèle et deuxièmement son application aux objets de test incomplets. La section précédente sur les arbres de décision a plutôt montré la pertinence d'une complétion pour le classement des objets que pour le calcul du modèle, qui fonctionne bien avec la stratégie de prise en compte des valeurs manquantes qu'utilise *C4.5*. Pour les règles d'association, l'application du modèle peut être optimisée, mais fournira un gain de performances moindre que dans le cas des arbres de décision.

Au final, toutes ces expériences montrent qu'il reste difficile d'évaluer une *bonne* méthode de complétion à l'aide de technique d'apprentissage supervisé. En effet, à l'aide d'expérimentations menées sur des données de l'UCI, nous avons montré que, contrairement à l'intuition, le modèle obtenu à partir d'une complétion *idéale* n'obtient pas les performances optimales et les améliorations obtenues s'avèrent marginales. En outre, l'introduction de valeurs manquantes aléatoires fournit fréquemment un modèle de meilleure qualité que celui calculé sur les données complètes. Ceci ne remet pas en cause l'utilité des méthodes de complétion, dont les motivations restent la nécessité de fournir des données complètes aux autres méthodes d'exploration des données (k-means, réseaux de neurones, SVM, *etc.*). Ces expériences permettent donc de prendre conscience des précautions qu'il faudra employer lorsqu'une méthode de classification supervisée est utilisée comme technique de validation d'une méthode de complétion.

Dans la section suivante, nous allons mettre en œuvre une nouvelle technique d'évaluation se basant sur des techniques de classification non supervisée.

7.3 Évaluation selon la stabilité d'une méthode d'apprentissage non supervisé

L'objectif d'une tâche de classification non supervisée consiste à proposer une partition des objets en k sous-ensembles, où k est le nombre de regroupements[8] souhaité. Une variation de cette tâche est de ne pas utiliser le nombre souhaité de regroupements comme un paramètre. L'algorithme construit alors plusieurs partitions candidates et choisit la meilleure. La meilleure partition est celle qui optimise un critère de qualité des partitions [Halkidi et Vazirgiannis, 2001, Deborah *et al.*, 2010].

7.3.1 Principe

Une méthode de classification non supervisée est classiquement validée en mesurant sa stabilité, *i.e.,* sa sensibilité à une perturbation des données. La mesure de stabilité se traduit généralement par un calcul de similarité entre deux partitions : celle issue des données originales et l'autre construite à partir de données perturbées. Cette technique de validation a longtemps intéressé les chercheurs [Raghavan et Ip, 1982]. Plus récemment, d'autres travaux se sont intéressés à cette technique, citons par exemple [Pascual *et al.*, 2010].

Dans le cadre de l'évaluation de méthodes de complétion, nous emploierons cette technique en considérant que les données originales sont les données de référence, tandis que les données perturbées sont celles issues d'une complétion. Nous supposons donc que la complétion est le mécanisme qui engendre la perturbation des données, l'objectif étant d'évaluer la méthode de complétion.

7.3.2 Indice de comparaison de partitions

Pour évaluer la stabilité de la méthode, nous comparons deux partitions. Pour cela, nous emploierons l'indice de $Rand(\mathcal{P}, \mathcal{Q})$ [Rand, 1971], qui permet de mesurer la similarité de deux partitions $\mathcal{P}$ et $\mathcal{Q}$ en comptant la proportion de paires d'objets qui sont dans le même cluster d'une partition à l'autre. Lorsque les deux objets sont placés dans le même cluster ou lorsqu'ils sont placés dans des clusters différents dans les deux partitionnements, alors ces deux objets sont considérés comme *similairement placés*. En revanche, les deux objets sont considérés comme *différemment placés* lorsque ils sont placés dans un même cluster dans un partitionnement et dans deux clusters différents dans l'autre partitionnement. L'indice de *Rand* est donc la proportion de paires d'objets similairement placés.

[8] On parle de regroupement, de partition ou de cluster en anglais.

La figure 7.6 illustre le déroulement de la technique d'évaluation à base de stabilité de clustering, employant l'indice de *Rand*.

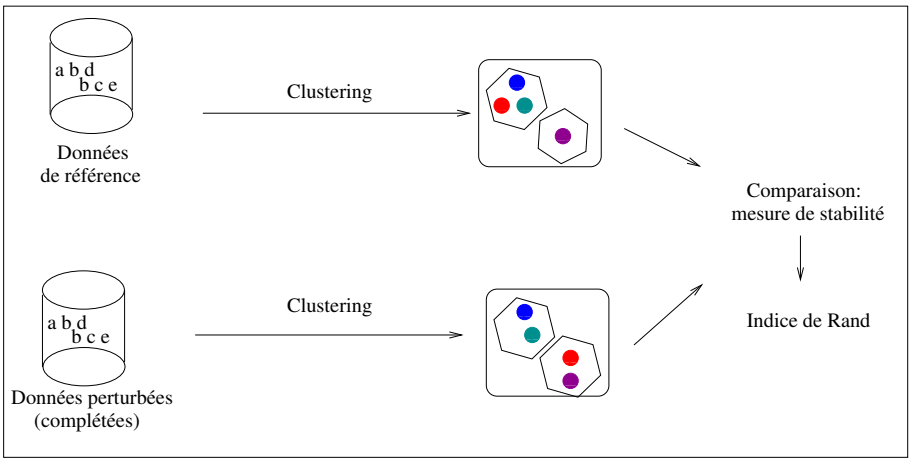


FIG. 7.6 – Évaluation d'une méthode de complétion selon la technique de mesure de stabilité de clustering.

7.3.3 Discussion

Nous réalisons des expériences en employant la méthode *KMeans* [MacQueen, 1967], une méthode classique de partitionnement. La figure 7.7 montre le déroulement du protocole associé à ces expériences. Rappelons que notre méthode complète les valeurs manquantes non aléatoires. Elle doit donc être suivie d'une méthode de complétion des valeurs manquantes aléatoires. Nous évaluons la pertinence de notre méthode de complétion sur les données de l'UCI quand elle est appliquée avant d'autres méthodes de complétion : complétion par *nouvelle valeur* (*new*), par le *mode* (*mode*) ou par une *valeur aléatoire* (*rand*). Nous examinons les écarts de performance entre l'application ou non de notre méthode.

Les données incomplètes résultent de l'effacement de valeurs de façon non aléatoires, selon le protocole décrit dans la section 7.1. La figure 7.8 pour *KMeans* (implémentation de Weka) montre les écarts entre la racine carré[9] des indices de *Rand* avec ou sans notre méthode de complétion. Ceci permet d'évaluer l'impact de notre complétion des valeurs manquantes non aléatoires, en amont d'une autre méthode de complétion de valeurs manquantes aléatoires plus classique.

Nous remarquons qu'à avec la méthode *KMeans*, l'écart avec la méthode qui remplace par une *nouvelle valeur* est peu significatif. En revanche, avec les deux autres méthodes l'écart est souvent tangible. Nous apercevons un bon comportement de notre complétion lors qu'elle est appliquée en amont, essentiellement sur les bases *vehicle* et *wine*.

[9] La racine carré permet de normaliser la distribution.

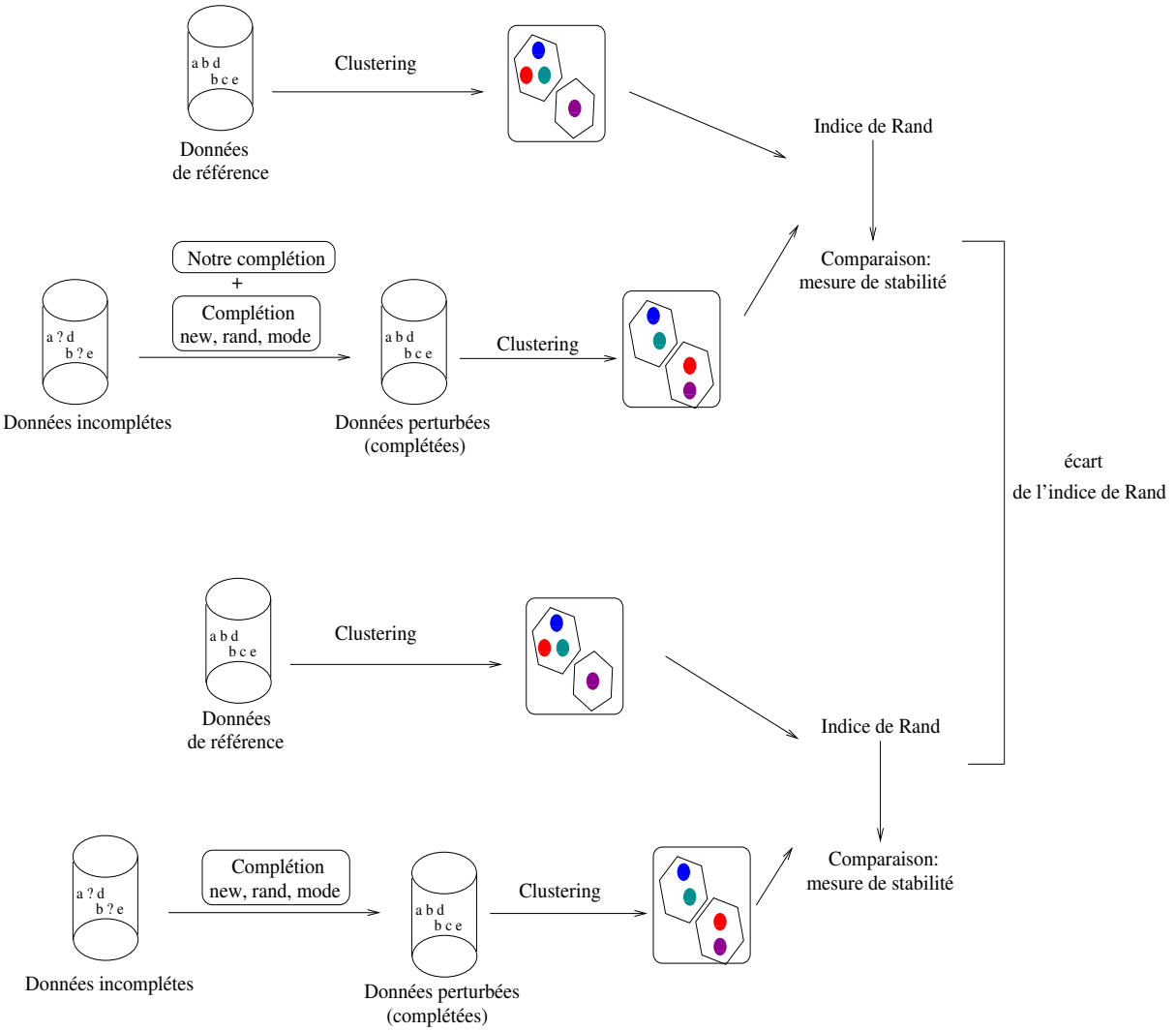


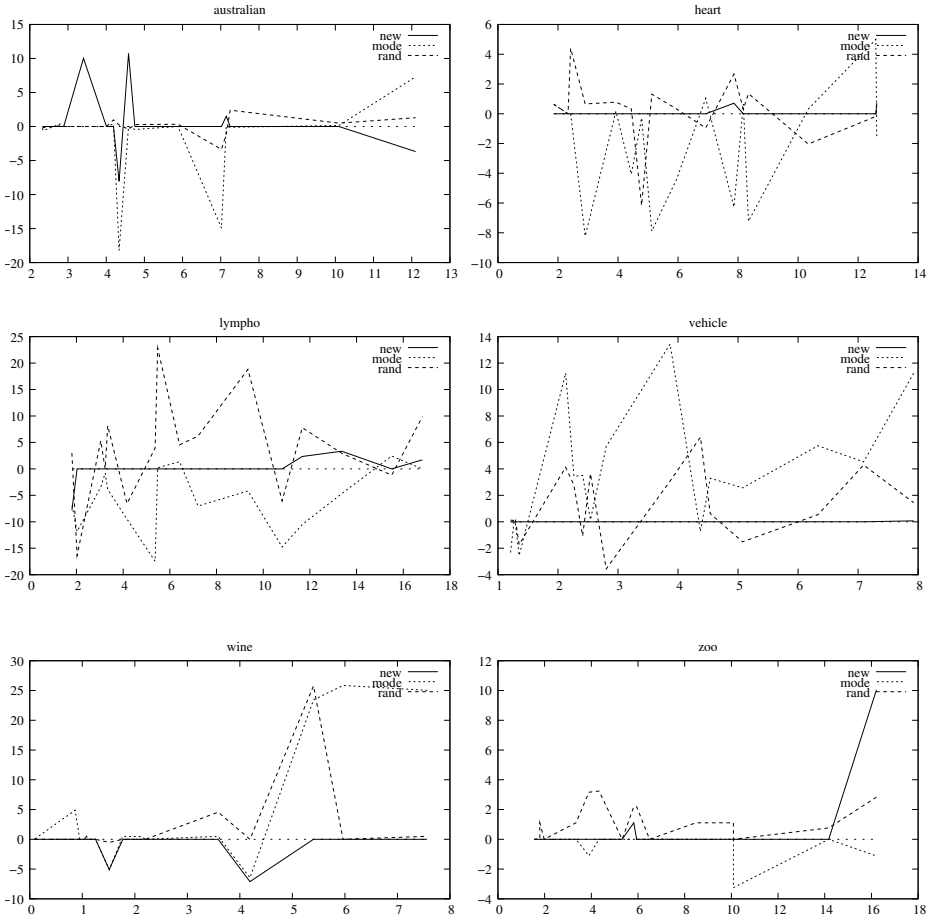
FIG. 7.7 – Évaluation de notre méthode de complétion selon la technique de mesure de stabilité de clustering.

Il apparaît que notre complétion permet de préserver la stabilité de la technique *KMeans* entre les partitions issues des données de référence et de celles complétées. Dans tous les cas, c'est la complétion avec une valeur aléatoire qui cause l'écart le plus significatif. Ceci n'est guère étonnant puisque cette dernière méthode de complétion est la moins satisfaisante.

En revanche, la complétion avec le *mode* fournit un écart moindre. Ce dernier point confirme la précision de notre technique d'évaluation : elle fournit un plus grand écart avec la complétion la moins fiable et inversement.

Par ailleurs, à la différence de l'évaluation selon des techniques supervisées, nous constatons que la technique à base de stabilité de clustering a montré une moindre sensibilité vis-à-vis de l'augmentation du taux des valeurs manquantes. Ce dernier point indique que cette technique

est robuste en présence de valeurs manquantes et s'avère donc plus appropriée à évaluer une méthode de complétion. De plus, nous estimons que cette technique d'évaluation est intéressante puisqu'elle ne présente pas la contrainte sur l'emplacement des valeurs manquantes : *faut-il les introduire dans l'échantillon de test, celui d'apprentissage ou les deux à la fois* ? Ainsi, l'impact des valeurs manquantes sur les objets à classer ou bien sur le calcul du modèle ne se pose pas.



Légende: L'axe des abscisses indique le pourcentage des valeurs manquantes, l'axe des ordonnées indique l'écart de l'indice de Rand entre les données de référence et l'utilisation ou non de notre méthode, préalablement a celles des méthodes new, mode et rand.

FIG. 7.8 – Résultat de la classification non supervisée à l'aide de *KMeans*.

7.4 Conclusion

Dans ce chapitre, nous avons commencé par présenter le protocole qui nous a permis d'introduire les valeurs manquantes de façon non aléatoires selon la typologie mise en évidence dans cette thèse. Ensuite, nous avons mené des expériences dans le but d'évaluer deux techniques de classification d'apprentissage supervisé, l'une à base d'arbre de décision, l'autre à base de règles d'association, lorsqu'elles sont appliquées pour évaluer une méthode de complétion idéale. Nous

avons montré qu'il est très difficile d'employer un tel schéma d'évaluation. Nous avons alors introduit une nouvelle technique d'évaluation qui consiste à mesurer l'impact de la complétion en mesurant la stabilité d'une méthode de classification non supervisée. Nous estimons que cette nouvelle technique présente un intérêt majeur, puisqu'elle n'est pas sensible à l'augmentation des valeurs manquantes. La complétion que nous avons mise en place dans le chapitre 6 a montré un bon comportement vis-à-vis de la stabilité de la technique *KMeans*.

Bilan et perspectives

Bilan

La problématique, à laquelle nous nous sommes intéressée dans cette thèse porte sur le prétraitement des valeurs manquantes dans le cadre d'un processus d'exploration de données. Au cours de ce travail, nous avons abordé la problématique des valeurs manquantes par une définition préalable d'une typologie des valeurs manquantes permettant de proposer une méthode de complétion caractéristique du type de la valeur manquante.

Caractérisation des valeurs manquantes

Le premier axe de notre étude a porté sur la proposition d'une nouvelle typologie des valeurs manquantes. Cette typologie consiste à exploiter les corrélations au sein des données mesurées afin de caractériser la présence des valeurs manquantes. La motivation principale de cette contribution est que la manière dont nous complétons une valeur manquante devrait dépendre de son type (manquante accidentellement, valeur inapplicable, *etc*). Nous avons constaté que dans la littérature, seules les valeurs manquantes aléatoires sont traitées.

En se basant sur la modélisation proposée par Little et Rubin, nous avons proposé une nouvelle typologie qui offre une caractérisation propre pour chaque valeur manquante. L'intérêt de cette caractérisation est qu'elle est mise en évidence selon une approche fouille de données grâce à des implications propres. Ces implications, étant à prémisses minimales, n'induisent pas de redondance.

Calcul de la base d'implications propres

Nous avons aussi proposé une nouvelle méthode de calcul de la base d'implications propres grâce au calcul des traverses minimales du contexte complémentaire. En se basant sur la propriété qui stipule que la règle $X \rightarrow i$ est une implication propre si X est une traverse minimale des objets complémentaires contenant i, nous avons pu extraire les implications propres nécessaires à la caractérisation non redondante des valeurs manquantes. Pour cela, nous avons adapté un algorithme de calcul des traverses minimales procédant par niveaux auquel nous avons ajouté

une contrainte sur la fréquence des traverses minimales afin d'obtenir des implications propres fréquentes.

Complétion des valeurs manquantes

Le troisième axe de notre travail a porté sur la proposition d'une nouvelle méthode de complétion contextualisée. Cette méthode consiste à traiter les valeurs manquantes en fonction du type mis en évidence lors de la caractérisation. Sur la base des règles de caractérisation, nous avons construit des schémas de caractérisation qui nous ont permis d'exhiber toutes les relations entre les valeurs manquantes et les données. En appliquant une propagation transitive et une réduction des caractérisations cycliques, nous avons retrouvé les origines des valeurs manquantes. Ces origines ont finalement constitué les valeurs de complétion relatives aux valeurs manquantes non aléatoires.

Évaluation des méthodes de complétion

Nous avons enfin étudié les techniques d'évaluation des méthodes de complétion des valeurs manquantes. À travers des expériences, nous avons montré qu'il est très difficile de valider une méthode de complétion en mesurant son impact sur des techniques supervisées. Nous avons alors introduit une nouvelle technique d'évaluation consistant à qualifier la pertinence de la complétion selon la stabilité d'une technique d'apprentissage non supervisé entre des données de références et des données complétées. Cette technique a souligné la relative stabilité de notre méthode de complétion.

Perspectives

Les résultats des travaux réalisés dans cette thèse offrent plusieurs perspectives de recherches :

Valeurs manquantes

Tout d'abord, en termes de caractérisation des valeurs manquantes, la définition de la typologie des valeurs manquantes, que nous avons proposée, est indépendante de toute base de règles. Cependant, cette définition dépend du seuil minimal des règles de caractérisation. En pratique, nous avons employé les implications propres pour la propriété de non redondance de la caractérisation qu'elle permet d'assurer. Notre première investigation a montré qu'il est très difficile d'écarter la notion de seuil minimal puisqu'il s'agit d'un paramètre clé lors de la recherche de régularités dans les données. Nous pensons néanmoins qu'il serait intéressant d'obtenir des critères permettant d'évaluer la pertinence d'une caractérisation par rapport à une autre lorsque l'on fait varier le seuil minimal.

Il serait également intéressant de mettre notre complétion à l'épreuve dans divers contextes applicatifs. Par exemple, dans le domaine bio-informatique, les puces à ADN, appelées aussi puces à gènes ou biopuces permettent, aux biologistes de réaliser des expériences sur plusieurs milliers de gènes où il manque souvent certaines mesures.

Une autre piste intéressante concerne la technique d'évaluation proposée dans le cadre de cette thèse qui raisonne sur la stabilité d'un clustering. Par exemple, il conviendrait de mener avec l'expert une discussion sur la pertinence de cette technique : quelle signification ont les résultats pour l'expert, notamment avec les nouvelles valeurs de complétion ? En effet, la qualité d'une complétion dépend de ce qu'on veut en faire. La question de l'objectif à atteindre en réalisant une complétion mérite d'être étudiée plus profondément.

Perspectives générales

Une autre perspective qui nous semble intéressante à creuser est celle de l'étude des propriétés de la base d'implications propres. Nous avons proposé une méthode d'extraction des implications propres concluant sur un item à travers le calcul des traverses minimales extraites à partir du contexte complémentaire contenant cet item. Nous pensons que ces règles peuvent être appliquées dans un autre contexte, autre que la caractérisation des valeurs manquantes, par exemple la caractérisation de classes dans un problème de classification supervisée. De plus, il nous semble intéressant d'exploiter cette méthode à base de traverses minimales afin d'extraire toutes les implications propres indépendamment de leur support. Ceci permettra par exemple d'obtenir les implications propres rares, utiles dans certaines applications.

Bibliographie

[A. Farhangfar, 2004] A. Farhangfar, L. Kurgan, W. P. (2004). Experimental analysis of methods for imputation of missing values in databases. *Intelligent Computing : Theory and Applications II Conference, held in conjunction with the SPIE Defense and Security Symposium, Orlando, FL, 2004*, pages 172–182.

[A. Siebes et van Leeuwen, 2006] A. Siebes, J. V. et van Leeuwen, M. (2006). Item sets that compress. *In Proceedings of the Sixth SIAM International Conference on Data Mining, April 20-22, 2006, Bethesda, MD, USA*, pages 393–404.

[Acuna et Rodriguez, 2004] Acuna, E. et Rodriguez, C. (2004). The treatment of missing values and its effect in the classifier accuracy. *In D. Banks, L. House, F.R. McMorris,P. Arabie, W. Gaul (Eds).Classification, Clustering and Data Mining Applications. Springer-Verlag Berlin-Heidelberg*, pages 639–648.

[Agrawal et Srikant, 1994] Agrawal, R. et Srikant, R. (1994). Fast algorithms for mining association rules. *In* Bocca, J. B., Jarke, M. et Zaniolo, C., éditeurs : *Proceedings of the 20th Intl. Conference on Very Large Databases, Santiago, Chile*, pages 478–499.

[Asuncion et Newman, 2007] Asuncion, A. et Newman, D. (2007). UCI machine learning repository [http ://www.ics.uci.edu/∼mlearn/mlrepository.html]. Irvine, CA : University of California, School of Information and Computer Science.

[B. Liu, 1998] B. Liu, W. H. e. Y. M. (1998). Integrating Classification and association Rule Mining. *In Proceedings of the International Conference in Knowldge Discovery in Databases (KDD'98), New-york, USA*, pages 80–86.

[Bashir *et al.*, 2009] Bashir, S., Razzaq, S., Maqbool, U., Tahir, S. et Baig, A. R. (2009). Using association rules for better treatment of missing values. *CoRR*, abs/0904.3320.

[Bastia et Monard, 2003] Bastia, G. E. A. P. et Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519–533.

[Bastide, 2000] Bastide, Y. (2000). Data mining : algorithmes par niveau, techniques d'implantation et applications. Thèse de doctorat, École Doctorale Sciences pour l'Ingénieur de Clermont-Ferrand, Université Blaise Pascal, France.

[Bastide *et al.*, 2000] Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L. et Stumme, G. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. *In Proceedings of the International Conference DOOD'2000, LNAI, volume 1861, Springer-Verlag, London, UK*, pages 972–986.

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

[Ben Othman et Ben Yahia, 2011] Ben Othman, L. et Ben Yahia, S. (2011). $GBAR_{MVC}$: Generic Basis of Association Rules based approach for Missing Values Completion. *The International Journal of Computing and Information Sciences (IJCIS)*, Volume 9, Number 1, April 2011:pages 19–36.

[Ben Othman *et al.*, 2008] Ben Othman, L., Rioult, F., Ben Yahia, S. et Crémilleux, B. (2008). Typologie des valeurs manquantes : proposition et caractérisation à l'aide de règles d'association. *In Actes des 24ièmes journées des "Bases de Données Avancées" (BDA 2008), Ardèche, France.*

[Ben Othman *et al.*, 2009a] Ben Othman, L., Rioult, F., Ben Yahia, S. et Crémilleux., B. (2009a). Completing non-random missing values. *In Proceedings of the 4th International Conference on Intelligent Systems and Knowledge Engineering (ISKE 2009)*, pages 227–232, Hasselt, Belgium.

[Ben Othman *et al.*, 2009b] Ben Othman, L., Rioult, F., Ben Yahia, S. et Crémilleux, B. (2009b). Missing values : Proposition of a typology and characterization with an association rule-based model. *In Proceedings of 11th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2009), Linz, Autsria, Springer-Verlag, LNCS 569*, pages 441–452.

[Ben Othman *et al.*, 2011] Ben Othman, L., Rioult, F., Ben Yahia, S. et Crémilleux, B. (2011). Base de caractérisation des valeurs manquantes. *Technique et Science Informatiques*, page (à paraître courant 2011).

[Ben Salem, 1999] Ben Salem, K. (1999). Design and analysis of an iterative algorithm for incomplte data estimation. *International Journal of Computer Mathematics*, 71(1):71–82.

[Berge, 1973] Berge, C. (1973). *Graphs and hypergraphs.* American Elsevier Publishing Company, INC, New-York.

[Berge, 1989] Berge, C. (1989). *Hypergraphs : Combinatorics of Finite Sets.* North-Holland.

[Boulicaut *et al.*, 2000] Boulicaut, J.-F., Bykowski, A. et Rigotti, C. (2000). Approximation of frequency queries by means of free-sets. *In Proceedings of the International Confe-*

rence on Principles and Practice of Data Mining and Knowledge Discovery in Databases (PKDD'2000), Lyon, France, pages 75–85.

[Boulicaut *et al.*, 2003] Boulicaut, J.-F., Bykowski, A. et Rigotti, C. (2003). Free-sets : A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery. Kluwer Academics Publishers*, 7:5–22.

[Breiman *et al.*, 1984] Breiman, L., Friedman, J., Stone, C. J. et Olshen, R. (1984). *Classification and Regression Trees.* Wadsworth Publishing Company.

[Brin *et al.*, 1997] Brin, S., Motwani, R. et Silverstein, C. (1997). Beyond market baskets : Generalizing association rules to correlations. *In* Peckham, J., éditeur : *Proceedings of the International Conference on Management of Data (ACM SIGMOD), Tucson, Arizona, USA*, pages 265–276. ACM Press.

[Bykowski et Rigotti, 2001] Bykowski, A. et Rigotti, C. (2001). A condensed representation to find frequent patterns. *In Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium of Principles of Database Systems, Santa Barbara, USA*, pages 267–273.

[Calders et Goethals, 2002] Calders, T. et Goethals, B. (2002). Mining all non-derivable frequent itemsets. *In Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland*, pages 74–85.

[Calders et Goethals, 2003] Calders, T. et Goethals, B. (2003). Minimal k-free representation of frequent sets. *In Proceedings of the International Conference on Principles ans Practice of Knowledge Discovery in Databases (PKDD'03), Cavtat-Dubrovnik, Croatia*, pages 71–82.

[Calders *et al.*, 2007] Calders, T., Goethals, B. et Mampaey, M. (2007). Mining itemsets in the presence of missing values. *In Proceedings of the ACM Symposium on Applied Computing*, pages 404–408, Seoul, Korea. ACM.

[Casali *et al.*, 2005] Casali, A., Cicchetti, R. et Lakhal, L. (2005). Essential patterns : A perfect cover of frequent patterns. *In Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Copenhagen, Denmark, August 22-26, 2005*, pages 428–437.

[Codd, 1990] Codd, E. F. (1990). *The Relational Model for Database Management, version 2*, chapitre 8 : Missing Information. Addison Wesley Publishing Company (April 1990).

[Date, 1995] Date, C. J. (1995). *Relational Database - Writings (1991-1994).* Addison Wesley ; Facsimile edition (1995).

[Deborah *et al.*, 2010] Deborah, L. J., Baskaran, R. et A.Kannan (2010). A survey on internal validity measure for cluster validation. *International Journal of Computer science and engineering Survey (IJCSES)*, 1(2):85 – 102.

[Delavallade, 2007] Delavallade, T. (2007). *Evaluation des risques de crise, appliquée à la détection des conflits armés intra-étatiques.* Thèse de doctorat, Université Pierre et Marie Curie - Paris VI.

[Delavallade et Dang, 2007] Delavallade, T. et Dang, T. (2007). Using entropy to impute missing data in a classification task. *In Proceedings of the IEEE International Conference of Fuzzy Systems (FUZZ-IEEE'07)*, pages 577–582, London, UK.

[Dempster *et al.*, 1977] Dempster, A., Laird, N. et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[Fiot, 2007] Fiot, C. (2007). *Extraction de séquences fréquentes : des données numériques aux valeurs manquantes.* Thèse de doctorat, Université de Montpellier II.

[Fiot *et al.*, 2007] Fiot, C., Laurent, A. et Teisseire, M. (2007). SPoID : Extraction de motifs séquentiels pour les bases de données incomplétes. *In Actes des 7èmes journés d'Extraction et Gestion des Connaissances (EGC'07)*, pages 715–726, Namur, Belgium.

[Gasmi *et al.*, 2006] Gasmi, G., Ben Yahia, S., Nguifo, E. M. et Slimani, Y. (2006). IGB : une nouvelle base générique informative des règles d'association. *Information-Interaction-Intelligence(Revue I3)*, 6(1):31–67.

[Ghahramani et Jordan, 1994] Ghahramani, Z. et Jordan, M. I. (1994). Supervised learning from incomplete data via an em approach. *Advances in Neural Information Processing Systems 6*, pages 120–127.

[Grzymala-Busse, 2004] Grzymala-Busse, J. W. (2004). Three approaches to missing attribute values - a rough set perspective. *In Proceedings of the Workshop on Foundations of Data Mining, the fourth IEEE International Conference on Data Mining, Brighton, UK, November 2004*, pages 139–152.

[Grzymala-Busse *et al.*, 1999] Grzymala-Busse, J. W., Grzymala-Busse, W. J. et Goodwin, L. K. (1999). A closest fit approach to missing attribute values in preterm birth data. *In Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, (RSFDGrC'99)*, pages 405–413, London, UK. Springer-Verlag.

[Grzymala-Busse et Hu, 2001] Grzymala-Busse, J. W. et Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining. *Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing*, pages 378–385.

[Guigues et Duquennes, 1986] Guigues, J. et Duquennes, V. (1986). Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, 95:5–18.

[Hagen, 2008] Hagen, M. (2008). *Algorithmic and Computational Complexity Issues of MONET*. Phd dissertation, Institut für Informatik, Friedrich-Schiller-Universität Jena.

[Halkidi et Vazirgiannis, 2001] Halkidi, M. et Vazirgiannis, M. (2001). Clustering validity assessment : Finding the optimal partitioning of a data set. *In Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM'01, pages 187–194, Washington, DC, USA. IEEE Computer Society.

[Hamrouni *et al.*, 2005] Hamrouni, T., Ben Yahia, S. et Slimani, Y. (2005). Prince : An algorithm for generating rule bases without closure computations. *In Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Copenhagen, Denmark, August 22-26, 2005*, pages 346–355.

[Hamrouni *et al.*, 2007] Hamrouni, T., Denden, I., Yahia, S. B., Nguifo, E. M. et Slimani, Y. (2007). Les itemsets essentiels fermés : une nouvelle représentation concise. *In Actes des 7èmes journés d'Extraction et Gestion des Connaissances (EGC'07)*, pages 241–252, Namur, Belgium.

[Han et Kamber, 2000] Han, J. et Kamber, M. (2000). *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers.

[Hawarah, 2008] Hawarah, L. (2008). *Une approche probabiliste pour le classement d'objets incomplètement connus dans un arbre de décision*. Thèse de doctorat, Université Joseph Fourier - Grenoble I.

[Hawarah *et al.*, 2004] Hawarah, L., Simonet, A. et Simonet, M. (2004). A probabilistic approach to classify incomplete objects using decision trees. *In Proceedings of the 15th International Conference on Database and Expert Systems Applications (DEXA 2004), Lecture Notes in Computer Science 3180 Springer 2004, Zaragoza, Spain*, pages 549–558.

[Hawarah *et al.*, 2006] Hawarah, L., Simonet, A. et Simonet, M. (2006). Evaluation of a probabilistic approach to classify incomplete objects using decision trees. *In Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006), Krakow, Poland*, pages 193–202.

[Hébert, 2007] Hébert, C. (2007). *Extraction et usage de motifs minimaux en fouille de données, contribution au domaine des hypergraphes*. Thèse de doctorat, Université de Caen, Basse-Normandie.

[Hébert *et al.*, 2007] Hébert, C., Bretto, A. et Crémilleux, B. (2007). A data mining formalization to improve hypergraph minimal transversal computation. *Fundamenta Informaticae.*, 80(4):415–433.

[Jami *et al.*, 2005] Jami, S., Jen, T., Laurent, D., Loizou, G. et Sy, O. (2005). Extraction de règles d'association pour la prédiction de valeurs manquantes. *ARIMA journal, Numéro spécial CARI'04*, pages 103–124.

[Jen *et al.*, 2009] Jen, T.-Y., Laurent, D. et Sambe, G. (2009). Utilisation de règles d'association pour la prédiction de valeurs manquantes. *In Actes des 5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2009), Montpellier*, volume B-5 de *RNTI*, pages 79–90, Toulouse. Cépaduès.

[Kavvadias et Stavropoulos, 2005] Kavvadias, D. J. et Stavropoulos, E. C. (2005). An efficient algorithm for the transversal hypergraph generation. *Journal of Graph Algorithms and Applications*, 9(2):239–264.

[Kryszkiewicz, 1998] Kryszkiewicz, M. (1998). Representative association rules. *In Proceedings of the fourth Pacific-Asia Conference on Knowldge Discovery and Data mining (PAKDD), Melbourne, Australia.*, pages 198–209.

[Kryszkiewicz, 1999] Kryszkiewicz, M. (1999). Association rules in incomplete databases. *In In Proceedings of The third Pacific-Asia Conference on Knowldge Discovery and Data mining (PAKDD), Beijing, China, 1999. Lecture Notes in Computer Science, Vol. 1574. Springer*, pages 84–93.

[Kryszkiewicz, 2000] Kryszkiewicz, M. (2000). Probabilistic approach to association rules in incomplete databases. *In Proc. of Web-Age Information Management Conference (WAIM), Shanghai, China, 2000. Lecture Notes in Computer Science, Vol. 1846. Springer-Verlag (2000)*, pages 133–138.

[Lefébure et Venturi, 1999] Lefébure, R. et Venturi, G. (1999). *Le Data Mining*. Eyrolles.

[Levene et Loizou, 1993] Levene, M. et Loizou, G. (1993). Axiomatisation of functional dependencies in incomplete relations. *Theoretical Computer Science*, 206.

[Li et Cercone, 2006] Li, J. et Cercone, N. (2006). Assigning missing attribute values based on rough sets theory. *In Proceedings of the IEEE International Conference on Granular Computing (IEEE Grc 2006), Atlanta, USA, May 10-12, 2006*, pages 607–610.

[Li *et al.*, 2007] Li, J., Cercone, N. et Cohen, R. (2007). Addressing missing attributes during data mining using frequent itemsets and rough set based predictions. *In Proceedings of the IEEE International Conference on Granular Computing (IEEE GrC 2007), Silicon Valley, USA, November 2-4, 2007.*

[Li *et al.*, 2001] Li, W., Han, J. et Pei, J. (2001). cmar : Accurate and efficient classification based on multiple class-association rules. *In Proceedings of the IEEE International Conference on Data Mining (ICDM), Vancouver, Canada*, pages 369–376.

[Little et Rubin, 2002] Little, R. et Rubin, D. (2002). *Statistical Analysis with Missing Data, Second Edition*. John Wiley, New York.

[Liu *et al.*, 1997] Liu, W. Z., White, A. P., Thompson, S. G. et Bramer, M. A. (1997). Techniques for dealing with missing values in classification. *In Proceedings of the Second*

International Symposium on Intelligent Data Analysis, Lecture Notes in Computer Science, 1280:527–541.

[Lobo et Numao, 1999] Lobo, O. O. et Numao, M. (1999). Ordered estimation of missing values. *In Proceedings of the Third Pacific Asia Conference on Methodologies for Knowledge Discovery and Data Mining. Beijing, China, April 26-28, 1999.*, pages 499–503. Springer-Verlag.

[Luxenburger, 1991] Luxenburger, M. (1991). Implication partielles dans un contexte. *Mathématiques et Sciences Humaines*, 29(113):35–55.

[MacQueen, 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *In* Cam, L. M. L. et Neyman, J., éditeurs : *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

[Magnani, 2004] Magnani, M. (2004). Techniques for dealing with missing data in knowledge discovery tasks. disponible à l'adresse http ://magnanim.web.cs.unibo.it/data/pdf/missingdata.pdf.

[Nayak *et al.*, 2001] Nayak, J. R., Cook, D. J. et Cook, D. J. (2001). Approximate association rule mining. *In Proceedings of the Florida Artificial Intelligence Research Symposium, Key West, Florida, USA.*, pages 259–263.

[Pascual *et al.*, 2010] Pascual, D., Pla, F. et Sánchez, J. S. (2010). Cluster validation using information stability measures. *Pattern Recognition Letters*, 31:454–461.

[Pasquier, 2000] Pasquier, N. (2000). Datamining : Algorithmes d'extraction et de réduction des règles d'association dans les bases de données. Thèse de doctorat, École Doctorale Sciences pour l'Ingénieur de Clermont Ferrand, Université Clermont Ferrand II, France.

[Pasquier *et al.*, 1999a] Pasquier, N., Bastide, Y., Taouil, R. et Lakhal, L. (1999a). Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems*, 24(1):25–46.

[Pasquier *et al.*, 1999b] Pasquier, N., Bastide, Y., Touil, R. et Lakhal, L. (1999b). Discovering frequent closed itemsets. *In* Beeri, C. et Buneman, P., éditeurs : *Proceedings of 7th International Conference on Database Theory (ICDT'99), LNCS, volume 1540, Springer-Verlag, Jerusalem, Israel*, pages 398–416.

[Pearson, 2006] Pearson, R. K. (2006). The problem of disguised missing data. *SIGKDD Explorations*, 8(1):83–92.

[Phan-Luong, 2001] Phan-Luong, V. (2001). The representative basis for association rules. *In Proceedings of the IEEE International Conference on Data Mining (ICDM'01), 29 November - 2 December 2001, San Jose, California, USA*, pages 639–640. IEEE Computer Society.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, anlaysis and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–248.

[Quinlan, 1993] Quinlan, J. (1993). C4.5 : Programs for machine learning. *Morgan Kaufmann Publishers.*

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning. *Lecture Notes in Computer Science, 1997*, pages 81–106.

[Ragel, 1999] Ragel, A. (1999). *Exploration des bases incomplètes : Application à l'aide au prétraitement des valeurs manquantes.* Thèse de doctorat, Université de Caen, Basse-Normandie.

[Ragel et Crémilleux, 1998] Ragel, A. et Crémilleux, B. (1998). Treatment of missing values for association rules. *In Proceedings of the International Pacific-Asia Conference on Knowlege Discovery and Data Mining (PAKDD'98), LNCS Vol. 1394*, pages 258–270, Melbourne, Australia.

[Ragel et Crémilleux, 1999] Ragel, A. et Crémilleux, B. (1999). MVC - a preprocessing method to deal with missing values. *Knowledge-Based System Journal*, 12(5-6):285–291.

[Raghavan et Ip, 1982] Raghavan, V. V. et Ip, M. Y. L. (1982). Techniques for measuring the stability of clustering : A comparative study. *In Proceedings of SIGIR*, pages 209–237.

[Rand, 1971] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

[Rioult, 2005] Rioult, F. (2005). *Extraction de connaissances dans les bases de données comportant des valeurs manquantes ou un grand nombre d'attributs.* Thèse de doctorat, Université de Caen, Basse Normandie.

[Rioult et Crémilleux, 2003] Rioult, F. et Crémilleux, B. (2003). Condensed representations in presence of missing values. *In Proceedings of the 5th International symposium on Intelligent Data Analysis, LNCS Vol. 2810*, pages 578–588, Berlin, Germany. Springer-Verlag.

[Rioult et Crémilleux, 2006] Rioult, F. et Crémilleux, B. (2006). Extraction de propriétés correctes dans les bases de données incomplètes. *In Actes de la Conférence Francophone sur l'Apprentissage Automtique (CAp'06)*, pages 347–362, Trégastel, France.

[Ripley, 1996] Ripley, B. (1996). *Pattern recognition and neural networks.* Cambridge University Press.

[Rubin, 1978] Rubin, D. (1978). Multiple imputations in sample surverys - a phenomenological bayesian approach to nonresponse. *In Proceedings of the survery Research Methods section of the American Statistical Association*, pages 20–34.

[Shafer et Graham, 2002] Shafer, J. L. et Graham, J. W. (2002). Missing data : Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell Sytem Technical Journal*, 27:379–423.

[Shen *et al.*, 2007] Shen, J. J., Chang, C. C. et Li, Y. C. (2007). Combined association rules for dealing with missing values. *Journal of Information Science*, 33(4):468–480.

[Smyth et Goodman, 1992] Smyth, P. et Goodman, R. (August 1992). An information theoretic approach to rule induction from databases. IEEE Transactions *On* K*nowledge And Data Engineering*, 4(4):301–316.

[Song et Shepperd, 2007] Song, Q. et Shepperd, M. J. (2007). A new imputation method for small software project data sets. *Journal of Systems and Software*, 80(1):51–62.

[Stumme *et al.*, 2002] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N. et Lakhal, L. (2002). Computing iceberg concept lattices with titanic. *Data Knowledge Engineering*, 42(2): 189–222.

[Taouil et Bastide, 2001] Taouil, R. et Bastide, Y. (2001). Computing proper implications. *In Proceedings of the 9th International Conference on Conceptual Structures (ICCS'2001)*, pages 49–61, Stanford, CA.

[Vreeken et Siebes, 2008] Vreeken, J. et Siebes, A. (2008). Filling in the blanks - krimp minimisation for missing data. *In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08), Pisa, Italy*, pages 1067–1072.

[Witten et Frank, 2005] Witten, I. H. et Frank, E. (2005). *Data Mining : Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[Wu *et al.*, 2004] Wu, C., Wun, C. et Chou, H. (2004). Using association rules for completing missing data. *In Proceedings of the 4th International Conference on Hybrid Intelligent Systems, (HIS'04), IEEE Computer Society Press*, pages 236–241, Kitakyushu, Japan.

[Wu *et al.*, 2008] Wu, J., Song, Q. et Shen, J. (2008). Missing nominal data imputation using association rule based on weighted voting method. *In Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008*, pages 1157–1162.

[Zaki, 2000a] Zaki, M. (2000a). Generating non-redundant association rules. *In ACM SIGKDD International Conference on Knowledge discovery and data mining, Boston, USA*, pages 34–43.

[Zaki, 2000b] Zaki, M. J. (2000b). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12:372–390.

[Zaki et Hsiao, 2002] Zaki, M. J. et Hsiao, C. J. (2002). Charm : An efficient algorithm for closed itemset mining. *In Proceedings of the 2nd SIAM International Conference on Data Mining, Arlington, Virginia, USA*, pages 34–43.

Résumé L'extraction de connaissances à partir de données incomplètes constitue un axe de recherche en plein essor. Dans cette thèse, nous y contribuons par la proposition d'une méthode de complétion des valeurs manquantes. Nous commençons par aborder cette problématique par la définition de modèles d'apparition des valeurs manquantes. Nous en proposons une nouvelle typologie en fonction des données connues et nous les caractérisons de façon non redondante grâce à la base d'implications propres. Un algorithme de calcul de cette base de règles, formalisé à partir de la théorie des hypergraphes, est également proposé dans cette thèse. Ensuite, nous exploitons les informations fournies lors de l'étape de caractérisation afin de proposer une méthode de complétion contextualisée, qui complète les valeurs manquantes selon le type aléatoire/non-aléatoire et selon le contexte. La complétion des valeurs manquantes non aléatoires est effectuée par des valeurs spéciales, renfermant intrinsèquement les origines des valeurs manquantes et déterminées grâce à des schémas de caractérisation. Finalement, nous nous intéressons aux techniques d'évaluation des méthodes de complétion et nous proposons une nouvelle technique fondée sur la stabilité d'un clustering entre les données de référence et les données complétées.
Mots clés : Exploration de données, Bases de données, Observations manquantes, Hypergraphes.

Title A Missing values completion method based on their appearing models.

Abstract Knowledge Discovery from incomplete databases is a thriving research area. In this thesis, we particularly contribute with the proposal of a missing values completion method. We start approaching this issue by defining the appearing models of the missing values. We propose a new typology according to the given data and we characterize these missing values in a non-redundant manner defined by the basis of proper implications. An algorithm computing this basis of rules, formalized from the hypergraph theory, is also developed in this thesis. We then explore the information provided during the characterization stage in order to propose a new contextual completion method which completes the missing values according to their type and to their context. The non-random missing values are completed with special values intrinsically containing the explanation defined by the characterization schemes. Finally, we investigate the evaluation techniques of the missing values completion methods and we propose a new technique based on the stability of a clustering, when applied on reference data and completed ones.
Keywords : Data mining, Databases, Missing observations, Hypergraphs.

Discipline : Informatique

Laboratoires :

- Unité de Recherche en Programmation, Algorithmique et Heuristique, Faculté des Sciences de Tunis, Tunisie.
- Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (UMR 6072), Université de Caen Basse-Normandie, France.