Missing values: Proposition of a Typology and Characterization with an Association Rule-based Model

Leila Ben Othman, François Rioult, Sadok Ben Yahia, Bruno Crémilleux

Missing Values: Proposition of a Typology and Characterization with an Association Rule-Based Model

Leila Ben Othman[1,2], François Rioult[2], Sadok Ben Yahia[1], and Bruno Crémilleux[2]

[1] Department of Computer Science,
Faculty of Sciences of Tunis. Tunisia
[2] University of Caen Basse-Normandie, France
GREYC - CNRS UMR 6072
{lbenothm,francois.Rioult,Bruno.Cremilleux}@info.unicaen.fr,
sadok.benyahia@fst.rnu.tn

Abstract. Handling missing values when tackling real-world datasets is a great challenge arousing the interest of many scientific communities. Many works propose completion methods or implement new data mining techniques tolerating the presence of missing values. It turns out that these tasks are very hard. In this paper, we propose a new typology characterizing missing values according to relationships within the data. These relationships are automatically discovered by data mining techniques using generic bases of association rules. We define four types of missing values from these relationships. The characterization is made for each missing value. It differs from the well-known statistical methods which apply a same treatment for all missing values coming from a same attribute. We claim that such a local characterization enables us perceptive techniques to deal with missing values according to their origins: the way in which we deal with the missing values should depend on their origins (e.g., attribute meaningless w.r.t. other attributes, missing values depending on other data, missing values by accident). Experiments on a real-world medical dataset highlight the interests of such a characterization.

Keywords: Data mining, missing values, association rules.

1 Introduction

Many data sets are incomplete and handling missing values is a major challenge in data analysis. There are two main approaches to analyze incomplete data: using a data mining method which is adjusted to cope with missing values or completing the data by imputation. The first approach is clearly expensive. Indeed, even if a technique can be updated to handle missing values, the process has to be repeated for each technique. The second is appealing: once a database is completed, it enables us the running of any method. It may explain why many works deal with imputation. Obviously, this approach requires accurate imputations in order to provide proper and unbiased data sets. Basic methods such as the mean, the most common value, a default value, are not satisfactory because they exaggerate correlations [9]. Several techniques are based on the Expectation Maximization (EM). It has been shown that EM provides accurate

probability estimations but it requires a model which has to be chosen according to the data [7]. By using the minimum description length and the idea that the best completion is that allowing for the best compression, a method taking into account how specific values co-occur locally has been proposed in [19]. A rule-based system designed for symbolic data is presented in [5]. More generally, there are several techniques based on the exploitation of local regularities, e.g., association rules [2 4 14 17 20], concise representations of patterns [15] or rough sets [8 11].

Results from the literature show that completion is a very hard task. In these works, completion methods are evaluated by missing values which are artificially introduced in a *reference* dataset. Then, a completion process is applied and the completed dataset is compared to the reference one. However, missing values are introduced according to statistical hypotheses, typically by removing values fully randomly or removing values randomly dependent on a present value. Unfortunately, as recently highlighted in [13], these assumptions may be wrong in real-world datasets and current approaches do not avoid the pitfall of non-random missing values. This issue of different types of missing data mechanisms is addressed in the well-known work of Little and Rubin [10] which distinguishes three types of missing values (cf. section 2.3). This work is interesting because it brings out a better understanding of the origin of missing values and the scope of the completion methods. Not surprisingly, the type *NMAR* [10] (i.e., Not Missing At Random) is generally not tackled by the current completion methods [19]. We also think that these types suffer from limitations in practical uses of the data mining. First, it relies on a global characterization of the missing values. We claim that completion methods can benefit from local characterization. For instance, we will see in Section 4 that the explanation of missing values suggests several completion strategies and it should be illusory to settle for a unique characterization for all the missing values coming from one attribute. Second, the types proposed by Little and Rubin are based on both the true (or complete) data and the available data. But, in real-world data mining tasks, the true data remain unknown.

We propose in this paper a new typology characterizing missing values according to relationships within the data. These relationships are automatically discovered by an association rule-based model. We define four types of missing values leading to a precise characterization of each missing value. Contrary to [10], these types rely on local regularities so that each missing value has its own type independently of the other missing values. Moreover, relationships are only based on the available data. We claim that such a characterization brings out a twofold advantage. First, it provides a better understanding of the underlying reasons of the missing values thus contributing to better control data quality. For instance, in Section 4, we will see that this characterization suggests attributes meaningless w.r.t. other attributes and consequently the data set should be restructured. Second, this characterization enables us perceptive techniques to deal with missing values according to their origins. Experiments in Section 4 show that the type *indirect* of this characterization highlights missing values which are not tackled by the current methods [19]. About completion, previous works have already shown the importance of a preliminary analysis of missing values to propose relevant methods [6].

More generally, we address the following questions that we feel crucial to handle missing values in data mining tasks: what kind of missing values models can be

recognized from the available data? Is it possible to explain the presence of the missing values? How can we characterize these missing values? Could we have different characterizations for the missing values coming from one attribute?

The remainder of the paper is organised as follows. Section 2 presents the terminology used throughout the paper and summarizes missing values models from the literature. Section 3 proposes a new typology of missing values corresponding to our method characterizing missing values. Experiments (Section 4) on a real-world medical dataset highlight the interests of this typology.

2 Preliminaries

This section introduces the technical concepts (real and measured contexts, itemsets and association rules) and the missing values models proposed by Little and Rubin [10].

2.1 Definitions and Notations

Let us consider a database in an "attribute/value" format. Figure 1 provides a toy example. Each object is described by four attributes A_1, A_2, A_3 and A_4. A domain of values is associated to each attribute, *e.g.*, $dom(A_1) = \{a, b\}$, $dom(A_2) = \{c, d\}$, $dom(A_3) = \{e, f, g\}$ and $dom(A_4) = \{h, i\}$. An attribute A_i may have an unknown value, called a *missing value*, noted by "?". We give now the definition of a *real context* (an example of such a context is given by the left part of Figure 1).

Definition 1 (Real context). *A **Real Context** is a triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, where $\mathcal{O}$ is the set of objects or transactions, $\mathcal{I}$ the set of items and $\mathcal{R}$ is a function over $\mathcal{O} \times \mathcal{I}$ which takes its values in $\{present, absent\}$. $\mathcal{R}(o, i) = present$ indicates that the item $i \in \mathcal{I}$ is present in the object $o \in \mathcal{O}$. $\mathcal{R}(o, i) = absent$ means that i is not in o.*

When missing values occur, the real context is converted in a measured context.

Definition 2 (Measured Context)
*A $missing\ value$ modelling operator, noted mv, maps a **real context** $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ into a **measured context** noted by $mv(\mathcal{K}) = (\mathcal{O}, \mathcal{I}, mv(\mathcal{R}))$. The new function $mv(\mathcal{R})$*

	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		×		×			×		×

	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	**?**	**?**		×			×	**?**	**?**

Fig. 1. Boolean context. Left : *real* context. Right : *measured* context.

takes its values in $\{present, absent, missing\}$ *and fulfills the following properties for value* $\in \{present, absent\}$ *:*

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

A real context $\mathcal{K}$ corresponds to the complete dataset (which stays unknown in real-world applications) whereas a measured context $mv(\mathcal{K})$ refers to the available data, i.e., the data that we have to tackle in practice. The modelling operator $mv()$ models a data erasing, *i.e.,* some values were deleted (moved to $missing$). When a value is missing in $mv(\mathcal{K})$, it becomes impossible to guess its real value in $\mathcal{K}$. However, when the value is known in $mv(\mathcal{K})$, it corresponds to the same value in the real context $\mathcal{K}$ (first property of Definition 2). The second property ensures that a value (either present or absent) in $\mathcal{K}$ keeps its value or will be missing in $mv(\mathcal{K})$.

Figure 1 (Right) is the measured context associated to the left part of this figure. As this context comes from an attribute/value format, a missing value affects all the possible values of an attribute. For example, the missing values on the items a and b in the object o_8 in $mv(\mathcal{K})$ (Figure 1- Right) hide actually the presence of the item b in $\mathcal{K}$ (Figure 1- Left).

2.2 Association Rules

An *itemset* (or *pattern*) $X \subset \mathcal{I}$ is a set of items. An object $o \in \mathcal{O}$ contains the itemset X and we note $X \subset o$ if $\forall\, i \in X$, $\mathcal{R}(o,i) = present$. The absolute support of X, noted by $Supp(X)$, is defined as follows : $Supp(X) = |\{o \in \mathcal{O} | X \subset o\}|$. An association rule R, based on a pattern Z, is an expression $R : X \rightarrow Y$ where $X \subsetneq Z$ and $Y = Z \backslash X$. The itemsets X and Y are respectively called *premise* and *conclusion* of R. An association rule is quantified by its *support* and *confidence*: the *support* is equal to the one of Z and the confidence is defined as $Conf(R) = \frac{Supp(Z)}{Supp(X)}$. Valid association rules are those whose support and confidence are greater than or equal to minimal thresholds, respectively noted by $minsup$ and $minconf$. If $Conf(R) = 1$, then R is *exact*, otherwise it is *approximative*. For example, $fh \rightarrow c$ is an exact association rule (cf. Figure 1- Left).

Presenting all the steps of the association rule mining is out the scope of this paper. However, we use techniques for non-redundant association rules computation, specially the basis of proper implications [18]. These rules are composed in their premise part by a free-set[1] [3] (*aka* key [12] or minimal generator), and in the conclusion part, by an attribute of its closure[2], which does not appear in the closure of one of the subsets of the premise part. This computational technique considerably restricts the number of redundant exact association rules, while faithfully preserving all the induced knowledge.

2.3 Classical Models of the Missing Values Appearance

Little and Rubbin [10] distinguish the following three types of missing values, which are also called models for missing values appearance.

[1] Since exact association rules cannot be built on part of these patterns.

[2] The closure of a pattern is composed by attributes which are systematically present with it.

– ***MCAR*** *(Missing Completely at Random)*: the probability that an attribute A_i is missing is unrelated to the value of A_i nor to the value of any other attributes. It may affect any record and any object. For example, the missing values in the attributes A_1 and A_2 of Figure 1 (Right) are *MCAR*. These missing values have *a priori* no particular explanation.
– ***MAR*** *(Missing at Random)*: such a missing value depends on the values of other attributes but it does not depend on the true value of any of the missing values. In Figure 1 (Right), the missing values which affect the attribute A_3 are examples of *MAR*. Note that all missing values on A_3 are related to the presence of the itemset ac.
– ***NMAR*** *(Not Missing at Random)*: if the missing value is related to the true value itself, then the missing value is said to be *NMAR*. The missing values affecting the attribute A_4 in Figure 1 (Right) are *NMAR*. If the item i cannot be recorded, then a missing value arises each time that i should be observed.

2.4 Discussion and Position Statement

The use of the Little and Rubbin models is not easy [16]. A first difficulty is that these models require knowledge or assumptions on the real context (i.e., the true values) whereas, in practice, only the measured context is known. With the *MAR* model, the missing values coming from an attribute have to satisfy a property of randomness. The characterization of the *NMAR* model needs assumptions on the real context. For instance, we can assume that a missing weight is more likely linked to an obese people than a healthy one. Such an assumption can be formulated by experts but an expert is not always available. Moreover, there are relationships inside the data which are not taken into account by these models. For instance, in our experiments (cf. Section 4), we note that the size of an invasion of a location is missing when there is no ganglion. In practice, it should be useful to integrate such relationships to deal with missing values (in this example, this relationship suggests that the attribute on the size is meaningless when there is no ganglion and thus the data set should be restructured). Finally, note that these models are based on a global characterization of the missing values: they give the same explanation for all the missing values coming from one attribute. We think that a local characterization is more powerful to deal with missing values, especially to propose completion techniques.

3 A New Missing Values Typology

In this section, we propose a new typology of missing values based on a local characterization of the missing values computed only from the available data.

– **Direct missing value**: a missing value is said to be *direct*, whenever it has relations with other measured values.
– **Indirect missing value**: a missing value is said to be *indirect*, whenever it has relations with other missing values.
– **Hybrid missing value**: a missing value is said to be *hybrid*, whenever it has relations with both measured and missing values.

– **Random missing value**: a missing value is said to be *random*, whenever it does not have any relation with other measured values or missing ones.

The next section formally defines this typology. It uses association rules for characterizing the missing values.

3.1 Association Rule Based Model for Missing Values Characterization

The definition of association rules characterizing missing values beforehand requires to quantify the degree of the presence/absence of an itemset in a measured context $mv(\mathcal{K})$:

Definition 3 (Present itemset). *An itemset $X \subset \mathcal{I}$ is said to be Present, in $o \in \mathcal{O}$ if and only if $\forall\, x \in X, mv(\mathcal{R})(o,x) = present$, and is noted by $Present(X,o)$.*

Definition 4 (Missing itemset). *An itemset $X \subset \mathcal{I}$ is said to be missing, in $o \in \mathcal{O}$ if and only if $\forall\, x \in X$, $mv(\mathcal{R})(o,x) = missing$, and is noted by $Missing(X,o)$.*

Definition 5 (Partially present itemset). *An itemset $X \subset \mathcal{I}$ is said to be Partially present, in $o \in \mathcal{O}$ if and only if $\forall x \in X, mv(\mathcal{R})(o,x) \neq absent$ and $\exists\, x_1 \in X, mv(\mathcal{R})(o,x_1) = present$ and $\exists\, x_2 \in X, mv(\mathcal{R})(o,x_2) = missing$, and is noted by $PartPresent(X,o)$.*

Example 1. In the measured context depicted by Figure 1 (Right), we have $Present(adf, o_4)$, $Missing(ah, o_8)$ and $PartPresent(bdg, o_8)$.

The regularities allowing the characterization of missing values can be straightforwardly detected by association rules. In practice, these rules are discovered by using a minimal support value, $minsup$ to only care about regularities that appear frequently. We propose here a formalization of this new missing value typology as follows:

Definition 6 (*direct* missing value). *A missing value i is said to be* direct *in $\mathcal{T} \subset \mathcal{O}$ ($|\mathcal{T}| \geq minsup$) if and only if $\exists X \subset \mathcal{I} \setminus \{i\} \,|\, \forall\, o \in \mathcal{T}, Present(X,o) \Rightarrow Missing(i,o)$.*

Definition 7 (*indirect* missing value). *A missing value i is said to be* indirect *in $\mathcal{T} \subset \mathcal{O}$ ($|\mathcal{T}| \geq minsup$) if and only if $\exists X \subset \mathcal{I} \setminus \{i\} \,|\, \forall\, o \in \mathcal{T}, Missing(X,o) \Rightarrow Missing(i,o)$.*

Definition 8 (*hybrid* missing value). *A missing value i is said to be* hybrid *in $\mathcal{T} \subset \mathcal{O}$ ($|\mathcal{T}| \geq minsup$) if and only if $\exists X \subset \mathcal{I} \setminus \{i\} \,|\, \forall\, o \in \mathcal{T}, PartPresent(X,o) \Rightarrow Missing(i,o)$.*

Definition 9 (*random* missing value). *A missing value i is said to be* random *in $\mathcal{T} \subset \mathcal{O}$ ($|\mathcal{T}| \geq minsup$) if and only if $\forall X \subset \mathcal{I} \setminus \{i\}\,, \exists\, o \in \mathcal{T} \,|\, Missing(i,o) \wedge \neg Present(X,o)$.*

Example 2. Let us follow the running example in Figure 1. The rules used for the characterization associated to the context $mv(\mathcal{K})$ (Figure 1 - Right) are given by the left

	Rule	Support
R_1	$a \wedge c \rightarrow MV(A_3)$	2
R_2	$MV(A_1) \rightarrow MV(A_4)$	3
R_3	$a \wedge h \rightarrow MV(A_3)$	2
R_4	$c \wedge MV(A_4) \rightarrow MV(A_1)$	2
R_5	$c \wedge h \rightarrow MV(A_3)$	2
R_6	$d \rightarrow MV(A_4)$	2
R_7	$g \rightarrow MV(A_4)$	2

	A_1	A_2	A_3	A_4
o_1	-		{direct}	-
o_2	{hybrid}	-	-	{indirect}
o_3	-	-	{direct}	-
o_4	-	-	-	{direct}
o_5	{hybrid}	-	-	{indirect}
o_6	-	{random}	-	-
o_7	-	{random}	-	{direct}
o_8	{random}	-	-	{direct, indirect}

Fig. 2. Left: Rules concluding on missing values with *minsup*=2 from the measured context $mv(\mathcal{K})$ (cf. Figure 1). Right: Typology of the missing values associated to $mv(\mathcal{K})$.

part of the Figure 2 (Left) with $minsup = 2$. The notation $MV(A_i)$ indicates a missing value on the attribute A_i (*i.e.*, on all items of the A_i domain). The column *Support* indicates the value of the absolute support of the rule. The characterization of the missing values is given by the Right part of Figure 2. For example, the rule R_4 shows that when c is present and a missing value occurs on the A_4 attribute, then, a missing value is observed on the A_1 attribute. This rule characterizes *hybrid* missing values on the A_1 attribute over the objects o_2 and o_5 (Figure 2 - Right).

3.2 Impact of the Basis of Proper Implications for the Missing Values Characterization

As said in Section 2.2, we use the basis of proper implications for building the rules characterizing the missing values. We now show the interest of this rule basis.

The basis of proper implications can be seen as a nicety of the well-known Bastide's basis [1] which provides a cover of the exact association rules. With the Bastide's basis, every free pattern is the premise of a rule whose the conclusion is the *closure* of its premise (see Section 2.2). The basis of proper implications is a finer cover: a rule is kept only if its conclusion cannot be inferred from the closures of any subset of its premise. The rules of the basis of proper implications satisfy the following suitable properties to characterize missing values:

1. a rule has a minimum premise for a given conclusion. The redundancy, which may lead to conflicts, is limited.
2. the number of rules of the basis of proper implications is very small compared to the size of the Bastide's basis. For example, Table 1 compares the size of these basis under our experimental conditions on the HODGKIN dataset (see Section 4). The number of rules is drastically reduced with the basis of proper implications.

From our example (Figure 1), we illustrate the interest of the basis of proper implication for characterizing missing values. Figure 3 presents the rules concluding on the missing values of the A_4 attribute (noted $MV(A_4)$) which are generated by the Bastide's basis and the basis of proper implications. We note that the rule R'_4 is not generated by the basis of the proper implications since that $MV(A_4)$ has already appeared in the closure of one subset of the premise of R'_4 *i.e.*, $MV(A_4)$ is already in the

Table 1. Number of rules with the Bastide's basis and the basis of proper implications on the HODGKIN dataset

	Rules	Rules concluding on a missing value
Bastide's basis	2 923 070	2 681 045
Proper implications basis	49	15

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

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

Fig. 3. Rules concluding on $MV(A_4)$ Left: Bastide's basis. Right: basis of proper implications

conclusion of the rule R''_1. Consequently, R''_1 is considered as more interesting than R'_4 when characterizing the missing values on the A_4 attribute, since it has a non-redundant and minimum premise. As expected, the non-redundancy limits conflicts between types: objects o_2 and o_5 are characterized only by the type *indirect* with the basis of proper implications whereas the Bastide's basis proposes two types (*indirect* and *hybrid*). For the other objects, the characterization is the same with the two bases. The formalization of this intuition appears in our perspective, *i.e.,* we are particularly focusing on defining properties of this characterization.

3.3 Characteristics of the Missing Values Typologies

Table 2 summarizes the differences between the Little and Rubin models and our typology. Obviously, we get back the main features that we have introduced. The Little and Rubbin models need knowledge (or assumptions) on the true data so that their practical use is difficult. That it is why we qualify these models as "theoretical". On the contrary, our typology is only based on the available data. The Little and Rubin models perform a global characterization: they allocate the same type for all missing values coming from an attribute. With our typology, the types are evaluated by using local regularities and

Table 2. Characteristics of the missing values typologies

	Little and Rubin Typology	**Our new typology**
Data	Available + unavailable	Available
Framework	Theoretical	Theoretical + Operational
Characterization	Global	Local
Types	MCAR	Random
	MAR	Direct
	NMAR	-
	-	Indirect
	-	Hybrid

each missing value has its own type independently of the other missing values. The last row compares the types of the two typologies. The *NMAR* model does not appear in our typology since it is based on the unavailable data. *Indirect* and *hybrid* types are not present in the Little and Rubin typology. Besides, even if there is a matching between *MCAR*/*Random* and *MAR*/*Direct*, practical results can be different because our typology stems from local subsets of objects.

4 Experimental Results

This section describes our experiments carried out on a medical dataset about the Hodgkin disease. We have chosen this dataset because it addresses a real-world medical application with many missing values. These missing values are natural (*i.e.,* no simulation was made for artificially introducing them). Furthermore, this database is used by physicians and they can provide advice and feedback on the data and results.

The HODGKIN *dataset.* The Hodgkin disease is a cancer of the lymphatic system. The HODGKIN dataset contains 3904 patients split in three therapeutic trials (H7, H8 and H9) realized over successive temporal periods. Each patient is described by 36 attributes and 29 contain missing values. The percentage of missing values for an attribute varies between 2% and 88%. The attributes include blood and histological characteristics and several information on the locations (cervical, hilum, mediastinum, auxiliaries) and the sizes of the invasions. An invasion is a symptom of a cancer.

Results. The rules were mined with an absolute support equals 700. Only 15 rules concluding on missing values are discovered (recalling that the properties of the rule cover drastically reduce the number of rules as indicated in Table 1). We have also performed experiments with rules allowing few exceptions (i.e., non exact rules with very high confidence) and we found similar results. The 15 rules are reported in Figure 4 (Left). For example, the rule R_4 indicates that all objects containing the item $plaq <= 600$ and a missing value on the attribute ctr (right top cervical ganglion) also contain a missing value on the attribute ctl. It corresponds to an *hybrid* missing value characterization.

The rules R_1 and R_2 conclude on a missing value of the invasions of the left or right top cervical ganglion (*ctl* or *ctr*). These rules contain in their premise only the *trial H7* attribute. Therefore, the type of these missing values is *direct* and the trial H7 explains these missing values. This type highlights a characteristic situation suggesting to investigate the running of the trial H7: actually this trial did not distinguish the top and bottom cervical ganglions and that it is why these values are missing. This case of missing values reveals a classical problem of data merging. Our method enables us to be aware of such issue and therefore it allows to better control the data quality. Note also that some missing values on *ctl* attributes were characterized by other rules as *indirect* (R_3) and as *hybrid* (R_4). It states a multiple missing value characterization.

Rules R_5 until R_{10} (left part of the Figure 4) characterize missing values having the type *direct*. They mean that when a ganglion is not invaded, its invasion size is not measured. It is interesting to check that such a knowledge is automatically discovered. Rules R_{11} and R_{12} characterize *indirect* missing values on the sizes of the ganglions

	Premise	Conclusion	support
R_1	trial H7	MV(ctr)	816
R_2	trial H7	MV(ctl)	816
R_3	MV(axlsiz)$\wedge$ MV(ctr)	MV(ctl)	811
R_4	plaq<=600 $\wedge$ MV(ctr)	MV(ctl)	778
R_5	ctr not invaded	MV(ctrsiz)	2449
R_6	ctl not invaded	MV(ctlsiz)	2407
R_7	cbr not invaded	MV(cbrsiz)	1969
R_8	cbl not invaded	MV(cblsiz)	1690
R_9	axl not invaded	MV(axlsiz)	3295
R_{10}	axl not invaded	MV(axlsiz)	3185
R_{11}	MV(ctr)	MV(ctrsiz)	908
R_{12}	MV(ctl)	MV(ctlsiz)	910
R_{13}	med not invaded $\wedge$ vs <=30	MV(mtr)	920
R_{14}	med not invaded $\wedge$ relapse= no	MV(mtr)	1042
R_{15}	med not invaded $\wedge$ MV(cblsiz)	MV(mtr)	717

attribute	missing values	*direct*	*indirect*	*hybrid*	*random*
ctr	908	90%	0	0	10%
ctl	910	10.7%	10%	79%	0.3%
ctrsiz	3435	71%	3%	24	2%
ctlsiz	3398	71%	3%	24%	2%
cbrsiz	2274	87%	0	0	13%
cblsiz	2027	83%	0	0	17%
axlsiz	3444	96%	0	0	4%
axlsiz	3360	95%	0	0	5%
mtr	1512	32%	0	47%	21%

Fig. 4. Left: Exact association rules discovered from the HODGKIN dataset with $misnusp$=700. Right: Missing values characterization in the HODGKIN dataset.

$ctlsiz$ and $ctrsiz$. Missing values on these size attributes are explained by missing values on other attributes. When we do not know if a ganglion is invaded, then a missing value always occurs on its size. Rules R_{13} until R_{15} characterize missing values on the attribute *ratio mediastinum ganglion width / thorax*. The first two rules characterize *direct* missing values and the third one *hybrid* missing values.

The right part of Figure 4 summarizes the different types of the missing values according to the attributes. An important result is that most of the missing values are not due to randomness but they belong to the *direct*, *indirect* or *hybrid* types. As these types express relationships in the data, it means that our characterization is able to suggest explanations for most of the missing values.

In these experiments, we check that the missing values coming from one attribute may be characterized according to different types. It is the case for the attributes *ctl*, *ctrsiz*, *ctlsiz* and *mtr*. It illustrates the power of the local characterization of our approach, which does not force to consider only a single type for all the missing values coming from one attribute.

The database includes other attributes with a low rate of missing values (between 2% and 9%). As we used an absolute support threshold of 700 corresponding to 18% of the data, we have not discovered rules characterizing these missing values. The characterization depends on the minimal support threshold. Decreasing the minimal support threshold may lead to discover rules characterizing missing values on these attributes but it may provide multiple conflicts of characterization.

5 Conclusions and Perspectives

In this paper, we have proposed a new typology of missing values according to the relationships within the data. These relationships are automatically discovered by an association rule-based model. Contrary to models from the literature, our approach is

only based on the available data and it relies on local regularities so that each missing value has its own type independently of the other missing values. This characterization enables us a better understanding of the underlying reasons of the missing values (e.g., attribute meaningless w.r.t. other attributes, missing values depending on other data, missing values by accident). We claim that it is precious because it suggests explanations about the quality of the data and also more powerful techniques to deal with missing values, especially to propose completion methods. Experiments on a real-world medical dataset highlight the interests of this typology. Among others, they show that many missing values do not stem from randomness. Further work is to show the impact of the variation of the *minsup* over the characterization and on the conflict as well as to investigate the use of this typology for the completion issue. Association rules have been shown to be efficient to complete missing values coming from random processes [2]. Our intuition is that missing values characterized by *direct*, *indirect* or *hybrid* types require the help of background knowledge to be completed since these types express specific behaviors.

Acknowledgments. The authors are grateful to the "Centre Anti-Cancéreux François Baclesse de Caen" and to the Doctor Michel HENRY-AMAR for providing the HODGKIN data. This work is partially supported by the French-Tunisian project *CMCU 05G1412*.

References

1. Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., Stumme, G.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Palamidessi, C., Moniz Pereira, L., Lloyd, J.W., Dahl, V., Furbach, U., Kerber, M., Lau, K.-K., Sagiv, Y., Stuckey, P.J. (eds.) CL 2000. LNCS (LNAI), vol. 1861, pp. 972–986. Springer, Heidelberg (2000)
2. Ben Othman, L., Ben Yahia, S.: $GBAR_{MVC}$: Generic Basis of Association Rules based approach for Missing Values Completion. The International Journal of Computing and Information Sciences (to appear)
3. Boulicaut, J.-F., Bykowski, A., Rigotti, C.: Approximation of frequency queries by means of free-sets. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS, vol. 1910, pp. 75–85. Springer, Heidelberg (2000)
4. Calders, T., Goethals, B., Mampaey, M.: Mining itemsets in the presence of missing values. In: Proceedings of the ACM Symposium on Applied Computing, Seoul, Korea, pp. 404–408. ACM Press, New York (2007)
5. Dardzinska, A., Ras, Z.W.: CHASE-2: Rule based chase algorithm for information systems of type lambda. In: Tsumoto, S., Yamaguchi, T., Numao, M., Motoda, H. (eds.) AM 2003. LNCS (LNAI), vol. 3430, pp. 258–270. Springer, Heidelberg (2005)
6. Delavallade, T., Dang, T.: Using entropy to impute missing data in a classification task. In: Proceedings of the International Conference of Fuzzy Systems (FUZZ-IEEE 2007), London, UK, July 2007, pp. 23–26 (2007)
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
8. Grzymala-Busse, J.W.: Three approaches to missing attribute values - a rough set perspective. In: Workshop on Foundations of Data Mining, associated with the fourth IEEE International Conference on Data Mining (2004)

9. Grzymała-Busse, J.W., Hu, M.: A comparison of several approaches to missing attribute values in data mining. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS, vol. 2005, pp. 378–385. Springer, Heidelberg (2001)
10. Little, R., Rubin, D.: Statistical Analysis with Missing Data. John Wiley, New York (1987)
11. Nelwamondo, F., Marwala, T.: Rough set theory for the treatment of incompltete data. In: Proceedings of the IEEE International Conference of Fuzzy Systems (FUZZ-IEEE 2007), London, UK, July 2007, pp. 23–26 (2007)
12. Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., Lakhal, L.: Generating a condensed representation for association rules. Journal of Intelligent Information Systems 24, 29–60 (2005)
13. Pearson, R.K.: The problem of disguised missing data. SIGKDD Explorations 8(1), 83–92 (2006)
14. Ragel, A., Crémilleux, B.: Treatment of missing values for association rules. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) PAKDD 1998. LNCS, vol. 1394, pp. 258–270. Springer, Heidelberg (1998)
15. Rioult, F., Crémilleux, B.: Mining Correct Properties in Incomplete Databases. In: Džeroski, S., Struyf, J. (eds.) KDID 2006. LNCS, vol. 4747, pp. 208–222. Springer, Heidelberg (2007)
16. Shafer, J.L., Graham, J.W.: Mising data: Our view of the state of the art. Psychological Methods 7(2), 147–177 (2002)
17. Shen, J.J., Chang, C.C., Li, Y.C.: Combined association rules for dealing with missing values. Journal of Information Science 33(4), 468–480 (2007)
18. Taouil, R., Bastide, Y.: Computing proper implications. In: Proceedings of the 9th International Conference on Conceptual Structures (ICCS 2001), Stanford, CA, pp. 49–61 (2001)
19. Vreeken, J., Siebes, A.: Filling in the blanks - krimp minimisation for missing data. In: Perner, P. (ed.) ICDM 2008. LNCS, vol. 5077, pp. 1067–1072. Springer, Heidelberg (2008)
20. Wu, C., Wun, C., Chou, H.: Using association rules for completing missing data. In: Proceedings of 4th International Conference on Hybrid Intelligent Systems (HIS 2004), Kitakyushu, Japan, December 5-8, 2004, pp. 236–241 (2004)