The Handling of Missing Values in Medical Domains with Respect to Pattern Mining Algorithms

Danilo Schmidt[2], Matthias Niemann[1], and Gabriela Lindemann-von Trzebiatowski[3]

[1] Department of Transfusion Medicine, University Hospital Charite
matthias.niemann@charite.de
[2] Department of Nephrology, University Hospital Charite
danilo.schmidt@charite.de
[3] Department of Governing Bodies, Humboldt University of Berlin
gabriela.lindemann@uv.hu-berlin.de

Abstract. Missing values are a wide spread problem in analyzing large data sets. In the medical domain they are unavoidable and complete analyzing methods fail here. In the paper we give an overview of kinds of missingness and common methods to handle missing values in machine learning algorithms. We introduce the Charité Query Language Toolkit which was developed to find out similar patterns in patient data records with respect to post-kidney-transplant patients. The toolkit uses available case analysis methods combined with a preprocessing of missing values as a compromise of simplicity and functionality.

Key words: data mining, medical data, missing values

1 Introduction and Background

Missing values are a wide spread problem for analyzing methods, such as machine learning, pattern recognition or data-mining algorithms, in many domains. For medical data sets missing values are unfortunately unavoidable. In a complete case analysis for these data sets all patient records with missing data would excluded. Performing clinical studies only with complete patient data sets lead to a significantly smaller sample size with reduced statistical expressiveness.
Depending of the choosen method for the statistical analysis missing values can restrict the cohort so much that the whole study is endangered.
In the last decades the amount of electronically collected patient data has grown rapidly and the demand of researchers and physicians for the development of analyzing methods and tools for data-sets with missing values is obvious.
In our work, we will describe the different kinds of missing values and follow here in principle the systematic of Pigott [2] and de Goeij et.al. [1]. We give an impression how to deal with missing values by example of pattern mining algorithms and introduce some useful preprocessing methods for medical data. At last we present a short example for including these methods in our frequent pattern mining toolkit.

2 Kinds of Missingness

Missing values are a common issue when analyzing data in a wide range of research fields. In the medical domain it seems unavoidable, especially in long-term treatments. De Goeij et.al. [1] define a missing value as "hiding the value of an attribute". While analyzing a dataset a missing value occurs when the specific value is not available. This does not necessarily mean that the value does not exist, but it is unknown. The missing value may be one of the attribute values (e.g. a categorical value) or a unique value (e.g. a numerical value). To quantify missingness, a ratio of missing values and all values can formed over all attributes by a simple formula:

$$missingness(B) = \frac{|B_{missing}|}{|B|} \tag{1}$$

There are several reasons for the missingness of values in medical data-sets. Depending on the ratio missingness(B) and the applied analyzing method, missing values may distort the final result and the underlying data missing mechanism may cause a biased statistical analysis. Therefore it is appropriate to spend some considerations into the kind of missingness of the values of a special data-set before choosing an adequate analyzing algorithm. With respect to the reasons of missingness there are distinguished three categories MCAR, MAR and NMAR.

MCAR Ű missing completely at random - is the strongest assumption for missing values of a dataset. The missing of a value neither depends on the observed parameters nor on the unknown value itself. MCAR will not bias the analysis of data, because the missing data has the same distribution as the available data. In medical domain it occurs e.g. if it was forgotten to induce an examination or there were problems in the transmission of laboratory data.

A less strict mechanism is MAR Ű missing at random. The missing of a value is allowed to be dependent on the observed parameters but not on the missing value itself. In long-term treatment of a patient it happens e.g. if the patient was not motivated to come to a medical control round because he had no health problems.

When adjusting a set of variables, MAR can be avoided by selecting highly correlated variables to be observed.

But clearly, this requires a specialist with domain-specific knowledge. To the third category of missingness belong data where values are not missing at random Ű NMAR. Here the missing of a value depends on the value itself. E.g. a creatinine value of a kidney transplanted patient was not measured because of rejection and loss of the transplant.

Without evaluating the dataset, no type of missing value can be ruled out. Even worse, each type of missingness can occur in a single dataset where different mechanisms overlay as our instances above show. Furthermore, it is not trivial to find out what kind of missingness applies.

3 Preprocessing Missing Values for Application in Pattern Mining Algorithms

There are several ways to handle missing values in pattern mining and data-analysis algorithms. Especially in the medical domain there are several studies on different approaches of dealing with missing values. Van der Heijden et. al. [3] and Marlin [4] developed methods for handling missing values for several machine learning techniques. The most common methods are:

1. Complete Case Analysis (only for remaining complete rows),
2. Available Case Analysis (complete rows for current pattern),
3. Single Unconditional Mean Imputation (impute column's mean),
4. Single Conditional Mean Imputation (impute mean based on conditional columns),
5. Multiple Imputation (regression model generates complete sets),
6. Maximum Likelihood (estimate underlying distributions),
7. Pattern-Mixture Model (user defined patterns of missingness).

All methods are suitable for MCAR data-sets. Furthermore, Single Conditional Mean Imputation, Multiple Imputation, Maximum Likelihood and Pattern-Mixture Model are additional usable for MAR data-sets, but only the last one is practicable for NMAR data-sets. In the following we will introduce in short the handling of missing values for the mentioned methods.

3.1 Complete Case Analysis

The easiest way to handle missing values is to delete all cases (rows) that contain missing values. The remaining data will be complete and all methods requiring complete data sets can be applied without further issues. When applying Complete Case analysis (CC) to table 1, the patients 2, 3 and 4 would be discarded. In subsequent operations, only patient 1 would be considered. If the missing values were not missing completely at random, Complete Case analysis will be biased. Unfortunately MCAR applies rarely (see table 2).
Furthermore, the deletion of many cases is not applicable if there are missing values in almost every case. The loss of information would increase while the significance decreases [3].

Table 1. Data set with missing values

Patient	Test A	Test B	Test C
1	1.0	positive	positive
2	3.0	negative	?
3	?	negative	negative
4	?	?	negative

3.2 Available Case Analysis

Available Case analysis (AC), which is sometimes referred to as pair wise deletion, is less strictly than Complete Case analysis. When analyzing a subset of the observed variables, all complete cases for that subset are viewed. That means only missing values are ignored [3]. Considering tests B and C of table 1, patients 2 and 4 are discarded because for the given set of tests only patients 1 and 3 are complete. This method is rather easily applicable but has some drawbacks as mentioned by [3]. Because of the varying number of observations, errors in estimated covariance matrices might occur. Furthermore, only if missing values are MCAR, the estimates are consistent. It has been shown that Available Case analysis is superior to complete case analysis for weakly correlated variables. For strong correlations, AC is inferior to CC.

Table 2. As A2 is only done if A1 is positive, the missing of A2 is MAR

Patient	A1	A2
1	negative	?
2	negative	?
3	positive	negative
4	negative	?
5	positive	positive
6	positive	positive

3.3 Single Unconditional Mean Imputation

A contrary approach to CC and AC is imputation. Generally spoken, missing values will be filled up (imputed) with calculated values. After that procedure the data can be handled like a complete set. There are different methods of imputing values, which will be introduced briefly in the following.
The Single Unconditional Mean Imputation (sometimes referred to as single value imputation) replaces all missing of an observed variable by the mean of the available values of that variable [3]. In table 1, the mean of test A is 2.0. Hence the missing values of patients 3 and 4 would be replaced with 2.0.
There are several drawbacks of that method: The variance of the imputed variables decreases while the precision is overrated. The results of an unconditional mean imputation will always be biased. This method is rather easily applicable as there is no further information about the dataset required. Adapting this approach to categorical data, the most frequent value of a column is imputed. In order to avoid precision overrating, the unconditional imputation may be extended to analyze columns in order to find the underlying random distribution. The imputation is then based on the columns distribution.

3.4 Single Conditional Mean Imputation

An improvement to unconditional imputation is the conditional imputation method. By linear regression on the conditional (observed) variables with complete data, the missing values are imputed. When considering test B as the condition for test A in table 1, the mean of column A is calculated for all patients where test B is negative (mean = 3.0) and once more for all patients where test B is positive (mean = 1.0). Hence the imputed value for test A of patient 3 would be 3.0. The selection of conditional variables is not trivial. When selecting too many columns, the imputed value may be over fitted.

3.5 Multiple Imputation

In contrast to the previously mentioned single imputation methods, Multiple Imputation (MI) does not calculate a single mean of an observed variable in order to impute a missing value, but creates a set of possible complete data sets. Each imputed parameter is selected by the columns underlying random distribution that was determined by regression. On each complete set the analysis is done. Finally, the results will be brought together. Practical tests show that MI often performs better than CC and AC, especially in the field of nephrology. This holds for MCAR and MAR data as long as the model specification is suitable. An overrated precision is avoided by imputing data several times, while biasing is avoided by applying regression.

3.6 Maximum Likelihood

Maximum Likelihood (ML) methods estimate the parameters of the underlying distributions of the observed variables.
To get the most probable parameters, an EM algorithm can be used. If the algorithm converges, the coefficients with the highest likelihood can be used in linear regression models [3]. In contrast to imputation methods, there are no estimated values filled into the gaps of the data set. Instead, ML methods can help to provide significant estimates for regression models. ML is unbiased for data that is MCAR or MAR and outperforms CC, AC and single imputations methods. But a proper statistical model is fundamental.

3.7 Pattern-Mixture Model

For pattern-mixture models the missing data mechanism may remain unknown. Instead, a mixture of different patterns describes the missingness in the data set, whereas each pattern describes a subset of the missing values. These patterns support the statistical model and can therefore improve the analysis. Hence pattern mixture models can produce good estimates for data that is MCAR, MAR and NMAR. Unfortunately, creating patterns requires a lot of domain specific knowledge about the data.

4 The Charité Query Language Toolkit

The Charité Query Language Toolkit was developed to find out similar patterns in patient data records with respect to post-kidney-transplant patients. Physicians should be

enabled to find out similar courses of diseases and treatments to infer from it for actual cases.
For the development of a toolkit, it might be disappointing that there is no simple general purpose method that handles all missing values in each imaginable query, especially if the data missing mechanism is unknown. Calders et.al. [5] summarizes the common methods and proposes to use different approaches in order to estimate the robustness. Applying multiple methods handling missing values might be confusing for future users of the software, so a compromise has to be found. The toolkit focuses on an easy usage. The user is not expected to provide additional information about the data. For that reason the user cannot be asked for selecting variables to impute the data. Model based approaches introduce better estimations at the cost of higher complexity and therefore have to be avoided too. Since there are different data sources, lots of missing values can be expected. In complex queries, complete case analysis can lead to a drop of all transactions. Even in simple queries the missingness may be very high. That is why complete case analysis has to be avoided either.
The toolkit uses available case analysis methods combined with a preprocessing of missing values as a compromise of simplicity and functionality. It does not focus on creating statistically faultless results. Biased correlations caused by violations of the MCAR-property can be expected and are accepted for that purpose.
When searching for new correlations, the user may not be interested in strong and therefore possibly known rules but in weaker or overlooked associations.
In the design phase of the toolkit, two essential settings have to be done. Firstly, the definition of time slices in order to discretize the time axis is required (see figure 1). In the second step, the definition of norms in order to provide a discretization of the parameters is necessary.

Norm values of parameters can be set in the norms tab (see figure 2). A group of norms for the same parameter was named norm family. This classification is necessary in order to recognize missing values. The norm type depends on the detected parameter as some tests generate qualitative (e.g. HCV, CMV, etc.) and some result in quantitative values (e.g. heart rate, creatinine, etc.). Qualitative norms are simply a mapping of a string value to the norm's name.
This allows to assign several values to the same category (e.g. weak and positive are both mapped to not negative). Quantitative norms are ranges for numeric parameters that are mapped to the name of a norm (e.g. a creatinine value of 1.2 to 6.0 is mapped to bad).
Norms can be created automatically as well. The database contains references for several parameters that may be loaded. Furthermore, a function calculating quartiles and creating norms by these is available.

5 Discussion and Conclusion

In our paper we give a survey of kinds of missing values and common methods to handle them in pattern mining algorithms. We introduce the Charité Query Language Toolkit which was configured to work on post-kidney-transplant patient data. Since the data does not differ from other medical domains, the toolkit may be used in other

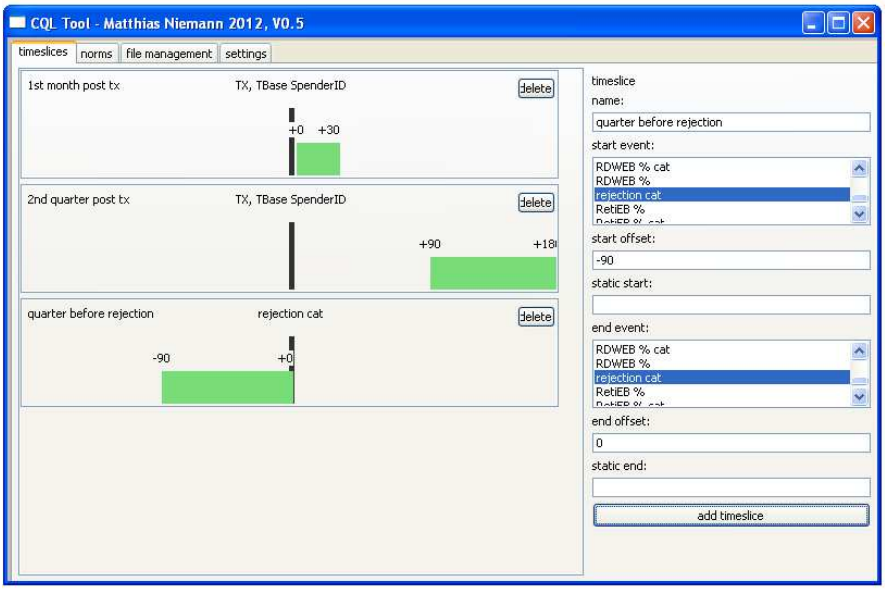

Fig. 1. time slices

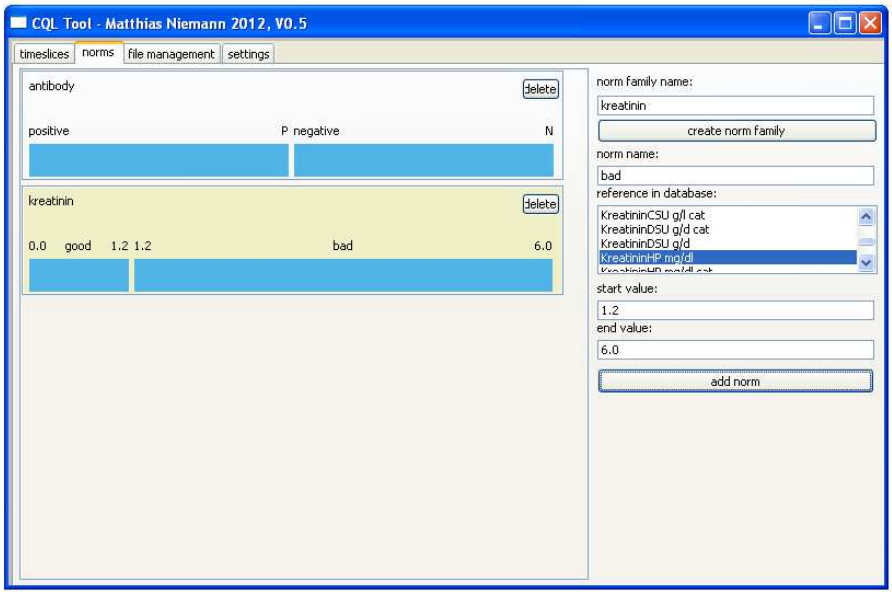

Fig. 2. norms

departments as well. Either a separate database is provided or the data is loaded into the current database. Depending on the domain, individual preprocessing plug-ins might be necessary in order to provide proper data transformation abilities.

References

1. de Goeij MC, van Diepen M, Jager KJ, et al. Multiple imputation: dealing with missing data. Nephrol Dial Transplant. 2013 Oct;28(10):2415-2420.
2. Pigott TD. A Review of Methods for Missing Data. Edicational Research and Evaluation. 2001;7(4).
3. van der Heijden GJMG, Donders ART, Stijnen T, et al. Imputation of missing values is superior to complete case analysis and the missingindicator method in multivariable diagnostic reaearch: A clinical example. Journal of Clinical Epidemiology. 2006;59:1102Ű 1109.
4. Marlin BM. Missing Data Problems in Machine Learning. Canadian theses. Library and Archives Canada Ű Bibliotheque et Archives Canada; 2008. http://books.google.de/books?id=5FlBPwAACAAJ (accessed 25 June 2015).
5. Calders T, Goethals B, Mampaey M. Mining Itemsets in the Presence of Missing Values. In: Proceedings of the 2007 ACM Symposium on Applied Computing. SAC Š07. New York, USA: ACM; 2007:404Ű408. http://doi.acm.org/10.1145/1244002.1244097 (accessed 25 June 2015).
6. Niemann M, Schmidt D, Lindemann von Trzebiatowski G, et al. First Steps towards a Frequent Pattern Mining with Nephrology Data in the Medical Domain. In: CSP, editor. Proceedings of the 21th International Workshop on Concurrency, Specification and Programming. vol. 2;2012:261-268