Dealing with Missing Values in Data

Jiří Kaiser
Faculty of Civil Engineering, Czech Technical University
Czech Republic
jiri.kaiser@fsv.cvut.cz

DOI: 10.20470/jsi.v5i1.178

Abstract: *Many existing industrial and research data sets contain missing values due to various reasons, such as manual data entry procedures, equipment errors and incorrect measurements. Problems associated with missing values are loss of efficiency, complications in handling and analyzing the data and bias resulting from differences between missing and complete data. The important factor for selection of approach to missing values is missing data mechanism. There are various strategies for dealing with missing values. Some analytical methods have their own approach to handle missing values. Data set reduction is another option. Finally missing values problem can be handled by missing values imputation. This paper presents simple methods for missing values imputation like using most common value, mean or median, closest fit approach and methods based on data mining algorithms like k-nearest neighbor, neural networks and association rules, discusses their usability and presents issues with their applicability on examples.*

Key words: missing data imputation, missing values, missing values imputation, missing data mechanisms, closest fit, k-nearest neighbor, neural networks, association rules, data mining

1. Introduction

Many existing industrial and research data sets contain missing values. Data sets contain missing values due to various reasons, such as manual data entry procedures, equipment errors and incorrect measurements. It is usual to find missing data in most of the information sources used.

Missing values usually appears as “NULL” values in database or as empty cells in spreadsheet table. Some flat-file formats use various symbols for missing values – e.g. arff files uses “?” symbol for missing values. These forms of missing values can be easily detected. However missing values can also appears as outliers or wrong data (i.e. out of boundaries). These data must be removed before intended analysis, and are much harder to find out.

The paper presents an overview missing values problem and strategies for dealing with missing values in data. Main part of the paper is dedicated to missing values imputation methods, discusses their usability and presents issues with their applicability on examples.

2. The Missing Values Problem

Missing value is a value that we intended to obtain during data collection (interview, measurement, observation) but we didn’t because of various reasons. Missing values can appear because respondent did not answer all questions in questionnaire, during manual data entry process, incorrect measurement, faulty experiment, some data are censored or anonymous and many others.

Luengo, J. (2011) introduced three problems associated with missing values as

- loss of efficiency,
- complications in handling and analyzing the data,
- bias resulting from differences between missing and complete data.

Loss of efficiency is caused by time-consuming process of dealing with missing values. This is closely connected to the second problem - complications in handling and analyzing the data. Complications in handling and analyzing data lie in fact that most methods and algorithms are usually unable to deal with missing values and missing values problem must be solved prior to analyses during data preparation phase. The other problem - bias resulting from differences between missing and complete data lies in fact that imputed values are not exactly the same as known values of completed data set and should not be handled the same way. The same problem occurs also if data set is reduced and some cases (rows of data set) are removed or ignored.

3. Characteristics of Missing Data Mechanisms

The missing data mechanism is usually classified as missing completely at random, missing at random or not missing at random (Lakshminarayan, K., Harp S. A. & Samad, T. 1999), (Horton, N. J. & Kleinman, K. P. 2007). Unfortunately identification of missing data mechanism is not always easy.

3.1 Missing Completely at Random (MCAR)

The missing data mechanism is considered as missing completely at random (MCAR), when the probability of a record having a missing value for an attribute does not depend on either the observed data or the missing data. MCAR is sometimes called uniform non-response. An example of a MCAR mechanism would be that a laboratory sample is dropped, so the resulting observation is missing. Data which is missing due to structural reasons cannot be regarded as MCAR (London School of Hygiene and Tropical Medicine, 2013).

3.2 Missing at Random (MAR)

The missing data mechanism is considered as missing at random (MAR), when the probability of a record having a missing value for an attribute could depend on the observed data, but not on the value of the missing data itself. Data which is incomplete only due to structural reasons are MAR.

Horton, N. J. & Kleinman, K. P. (2007) describe MAR mechanism as „states that the missingness depends only on observed quantities, which may include outcomes and predictors (in which case the missingness is sometimes labeled covariate dependent missingness (CDM))".

A special case of MAR is uniform non-response within classes. For example (London School of Hygiene and Tropical Medicine, 2013), suppose we seek to collect data on income and property tax band. Typically, those with higher incomes may be less willing to reveal them. If we have everyone's property tax band and given property tax band non-response to the income question is random, then the income data is missing at random. The reason (or mechanism) for it being missing depends on property band. Given property band missingness does not depend on income itself.

3.3 Not Missing at Random (NMAR)

The missing data mechanism is considered as not missing at random (NMAR), when the probability of a record having a missing value for an attribute could depend on the value of the attribute. Missing data mechanism that is considered as NMAR is non-ignorable. This can be solved by going back to the source of data and obtaining more information about the mechanism or obtain complete data set. Unfortunately it is very rare to know the appropriate model for the missing data mechanism.

Examples NMAR missing data mechanism (Lakshminarayan, K., Harp S. A. & Samad, T. 1999) include a sensor not detecting temperatures below a certain threshold, people not filling in yearly income in surveys if the income exceeds a certain value.

Some sources, e.g. (Horton, N. J. & Kleinman, K. P. 2007), called this mechanism missing not at random (MNAR).

4. Strategies for Dealing with Missing Data

There are three main strategies for dealing with missing data. The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all samples with missing values (Kantardzic, M. 2003). Another solution is to treat missing values as special values. Finally missing values problem can be handled by various missing values imputation methods. Unfortunately missing values imputation methods are suitable only for missing values caused by missing completely at random (MCAR) and some of them for missing at random (MAR) mechanism. If missing values are caused by not missing at random mechanism (NMAR) it must be handled by going back to the source of data and obtaining more information or the appropriate model for the missing data mechanism have to be taken into account.

4.1 Using Missing Values Policy of Used Analytical Method

There is no need to use a special method for dealing missing values if method that is used for data analysis has its own policy for handling missing values. Decision rules extraction methods may

consider attributes with missing values as irrelevant (Luengo, J. 2011). Association rules extraction methods may ignore rows with missing values (conservative approach) or handle missing values as they are supporting the rule (optimistic approach) or are in contrary with the rule (secured approach) as described by Berka, P. (2003).

4.2 Reducing the Data Set

The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all missing values. This can be done by elimination of samples (rows) with missing values (Kantardzic, M. 2003) or elimination of attributes (columns) with missing values (Lakshminarayan, K., Harp S. A. & Samad, T. 1999). Both approaches can be combined. Elimination of all samples is also known as complete case analysis.

Elimination of all samples is possible only when large data sets are available, and missing values occur only in a small percentage of samples and when analysis of the complete examples will not lead to serious bias during the inference. Elimination of attributes with missing values during analysis is not possible solution if we are interested in making inferences about these attributes. Both approaches are wasteful procedures since they usually decrease the information content of the data.

4.3 Treating Missing Attribute Values as Special Values.

This method deals with the unknown attribute values using a totally different approach. Rather than trying to find some known attribute value as its value, we treat missing value itself as a new value for the attributes that contain missing values and treat it in the same way as other values (Grzymala-Busse J. W., Hu M. 2001). Instead of storing value of the attribute we store the information that the value is missing. This approach assumes that we handle these values as they don't influence future analyses.

4.4 Replace Missing Value with Mean

This method replaces each missing value with mean of the attribute (Kantardzic, M. 2003). The mean is calculated based on all known values of the attribute. This method is usable only for numeric attributes and is usually combined with replacing missing values with most common attribute value for symbolic attributes.

4.5 Replace Missing Value with Mean for the Given Class

This method is similar to the previous one. The difference is in that the mean is not calculated from all known values of the attribute, but only attributes values belonging to given class are used. This approach is possible only for classification problems where samples are classified in advance (Kantardzic, M. 2003) or there is a possibility to create the classes.

4.6 Replace Missing Value with Median for the Given Class

Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given attribute is replaced by the median of all known values of that attribute in the class where the instance with the missing value belongs (Acuña, E. & Rodriguez, C. 2004). This method is usable only for numeric attributes and requires existence of classes or possibility to create classes as previous method

4.7 Replace Missing Value with Most Common Attribute Value

This method simply uses most common attribute value for missing value imputation (Grzymala-Busse J. W., Hu M. 2001). The most common value of all values of the attribute is used. This method is usable only for symbolic attributes and is usually combined with replacing missing values with missing values imputation using mean for numeric attributes.

4.8 Concept Most Common Attribute Value

This method is similar to the previous one. The concept most common attribute value method is a restriction of the previous method to the concept (Grzymala-Busse J. W., Hu M. 2001). This method

uses most common value of the attribute but uses cases belonging to the given class or concept instead of using global most common value. This method is usable for symbolic attributes and requires existence of classes or possibility to create classes.

4.9 Closest Fit

The closest fit algorithm (Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse, W. J. & Zheng X. 2005) for missing attribute values is based on replacing a missing attribute value with an existing value of the same attribute from another case that resembles as much as possible the case with missing attribute values. The main idea is based on searching through the set of all cases, considered as vectors of attribute values, for a case that is the most similar to the given case with missing attribute values. E.g. if 5 attributes of given case is known and 6th is being searched, the 6th attribute value is taken from the case which has other 5 attributes values most similar to the given case. The proximity measure between two cases x and y is the Manhattan distance between x and y, i.e.,

$$\text{distance}(x, y) = \sum_{i=1}^{n} \text{distance}(x_i, y_i)\,, \tag{1}$$

Where

$$\text{distance}(x_i, y_i) = \begin{cases} 0 \text{ if } x_i = y_i, \\ 1 \text{ if x and y are symbolic and } x_i \neq y_i \text{ or } x_i = ? \text{ or } y_i = ? \\ \dfrac{|x_i - y_i|}{r} \text{ if } x_i \text{ and } y_i \text{ are numbers and } x_i \neq y_i \end{cases}$$

where r is the difference between the maximum and minimum of the known values.

Problem with using only one closest case may lied in replacing missing value by an outlier. The problem is illustrated by following example (Tab. 1):

Tab. 1: Closest fit example

Property table	
floor area [m^2]	rental price [CZK]
45	10 000
52	13 000
52	11 000
54	**?**
55	18 000
62	13 500
62	12 000

Missing value imputation using only one closest case leads to replace missing value with value 18 000 (from the case with floor area 55 m^2), but there are properties with even larger floor area (62 m^2) with lower price.

This problem may be partially solved by using more than one closest case.

Another problem is in assumption of same distance between all values of symbolic attributes. Some attributes can be measured on ordinal scale. E.g. the “C” grade is closer to “B” than to “A”, but it is not possible to specify exactly the distance between these values. Described method uses Manhattan distance but there are several other ways of measuring distance including Euclidean distance and many others.

4.10 Missing Values Imputation Using k-Nearest Neighbor

The closest fit method looks for the most similar case with known attribute value. K nearest neighbor usually looks for more than one (“k”) similar cases with known values of the attribute, whose value is being searched. Nearest neighbor methods are usually used to classify objects. To classify a new object, with input vector y, we simply examine the “k” closest data set points to y and assign the object to the class that has the majority of points among these “k” (Hand, D., Mannila, H. & Smyth, P. 2001).

The k-nearest neighbor algorithm does not create explicit models (Batista, G.E.A.P.A. & Monard, M.C. 2003). There are several ways for using known values of k-nearest neighbor. One possibility for numeric attributes is to impute a weighted mean of the k nearest-neighbors attribute values (The MathWorks, Inc. 2013). The weights are inversely proportional to the distances from the neighboring columns.

The main question is selection of “k” parameter. Another question is handling symbolic attributes. There are several metrics that can be used for measuring the distance (The MathWorks, Inc. 2013). It is possible to use Euclidean distance, Manhattan distance and many others. HUANG, Ch. & LEE, H. (2004) introduced k-nearest neighbor missing values imputation method that uses grey relational analysis to determine the relationships among a referential observation and compared observations by calculating the grey relational coefficient and the grey relational grade.

Some implementations of the k-nearest neighbor method for missing values imputation (Improved outcomes software 2004) are able to compute the k nearest neighbors only on complete datasets so missing values have to be filled in with an initial approximation.

4.11 Missing Values Imputation Using Neural Networks

Neural networks constitute a class of predictive modeling system that works by iterative parameter adjustment (Chen, Z. 2001). The network structure, also called topology or architecture, includes the neural framework (number of neurons, number of layers, neuron model type, etc.) and the interconnection structure. Single-layer network has only an input and output layer. In a multilayer network, one or more hidden layers are inserted between the input and the output layer.

Gupta, A. & Lam, M. (1998) introduced following procedure for reconstruction of missing values using multilayered networks and backpropagation algorithm:

Step 1 Collect all training cases without any missing value and call them the complete set.

Step 2 Collect all training and test cases with at least one missing value and call them the incomplete set.

Step 3 For each pattern of missing values, construct a multi-layered network with the number of input nodes in the input layer equal to the number of non-missing attributes, and the number of output nodes in the output layer equal to the number of missing attributes. Each input node is used to accept one non-missing attribute, and each output node to represent one missing attribute.

Step 4 Use the complete set and the backpropagation algorithm to train each network constructed in step 3. Since the complete set does not have missing values, different patterns of input-output pairs can be obtained from the complete set to satisfy the input-output requirements for different networks from step 3. As the output of a network is between 0 and 1, data have to be converted to values between 0 and 1 for this reconstruction procedure.

Step 5 Use the trained networks from step 4 to calculate the missing values in the incomplete set.

The main question is strategy of neural networks construction. There are two main approaches. The first one is to start with a simple network and add extra nodes to the network until such addition does not improve the network performance. The second approach is to start with a large network and delete nodes from the network as long as the deletion does not deteriorate the network performance (commonly referred to as the node-pruning method). Fahlman S. E. & Lebiere, Ch. (1991) proposed a network construction method called “cascade-correlation” under the first approach. A node-pruning method, “relevance assessment”, was proposed by Mozer, M. C. & Smolensky, P. (1989).

Another question is handling symbolic attributes, because step 4 of the procedure requires data conversion to values between 0 and 1.

4.12 Missing Values Imputation Using Association Rules

An association rule is a simple probabilistic statement about the co-occurrence of certain events. For binary variables association rule takes the following form (Hand, D. J., Manilla, H. & Smyth, P. 2001):

$$IF\ A = 1\ AND\ B = 1\ THEN\ C = 1\ , \tag{2}$$

where A, B, and C are variables and

$$p = p\left(C = 1 \mid \mathrm{A} = 1, \mathrm{B} = 1\right). \tag{3}$$

i.e., the conditional probability that C = 1 given that A = 1 and B = 1. The conditional probability p is referred to as the "confidence" of the rule, and p (A = 1, B = 1, C = 1) is referred to as the "support". The support can be used as a constraint of minimum count of cases supporting association rule. The "If" part of the rule is often called antecedent and the "then" part is often called consequent (Larose D. T. 2005).

There are several ways of generating association rules, e.g.: apriori algorithm (Berka, P. 2003), (Hand, D. J., Manilla, H. & Smyth, P. 2001), (Kantardzic M. 2003), GUHA method (Berka, P. 2003) and other algorithms derived from apriori algorithm (Burita L. et al., 2012) or (Hipp, J., Güntzer, U. & G. Nakhaeizadeh, G. 2000).

The input for missing values imputation is incomplete data set. Algorithms for association rules generation are usually unable to handle missing values. Some association rules extraction methods may ignore rows with missing values (conservative approach) or handle missing values as they are supporting the rule (optimistic approach) or are in contrary with the rule (secured approach) as described by Berka, P. (2003).

Another possibility is getting complete data set for association rules generation. First way to get complete data set for association rules generation is reducing the data set.by elimination of cases and/or attributes as described in one of previous chapters. Another way to get complete data set for association rules generation is to handle missing values as special values. Using special values to obtain complete data set for association rules can be recommended because association rules with special values can be easily ignored. Association rules can't be often generated directly from numeric attributes. Missing values imputation may be directly used only for symbolic attributes.

Algorithms for association rules generation usually generates rules based on setting of minimum support, minimum confidence and maximum number of rules parameters. Most analyses, that use association rules methods, use only few association rules. For missing values imputation usually many more rules are required because not all association rules are suitable for missing values imputation.

Useable association rule for missing values imputation has the consequent containing value of attribute whose value is being searched and the antecedent correspond to values of other given case attributes. If complete data set for association rules generation was obtained by replacing missing values by special values association rules with these special values in consequent must be omitted. It is possible to use only association rules with consequent length equal to 1 and impute missing values one by one in cases with more than one missing values. Another possibility is using association rules with consequent length equal to the count of missing values in given case.

Support and confidence of the rule are two criteria that should be maximized during making decision about using the association rule for missing value imputation. For missing value imputation can be used the rule with maximum confidence along with required support. It is also possible to ignore support of the rule and use only confidence. Setting required min. support of the rule highly depends on given data set.

There is a possibility to combine association rules approach with most common attribute value method. At first the most common attribute value method can be used if it is not possible to find suitable association rule with required support for missing value imputation. At second the most common attribute value method can be used if there is no suitable association rule with required support and confidence not lower then relative frequency of occurrence of most common attribute

value. It is also possible to use only few association rules to improve missing values imputation accuracy using the most common attribute value.

Following tables and figures show how missing values imputation accuracy may be influenced by setting of requirement of minimum support of association rules. Analyses were made using 5 categorical attributes from STULONG study data set. Missing values were imputed using combination of association rules and most common attribute value method. There were 5% of missing values generated. Missing values were generated under MCAR conditions.

Following table and figure show example where minimum of incorrectly imputed values is reached with ignored support parameter (minimum required support was set to 0 counts).

Tab. 2: Incorrectly imputed values using combination of association rules and most common attribute values with different requirement of minimum association rules support (1st example)

Requirement of min. association rules support	Incorrectly imputed values	
[count of cases]	[count]	[%]
<0;52>	538	20,3865
(52;65>	541	20,5002
(65;78>	547	20,7275
(78;233>	548	20,7654
(233; 2323>	546	20,6897
(2323; ∞)	591	22,3948

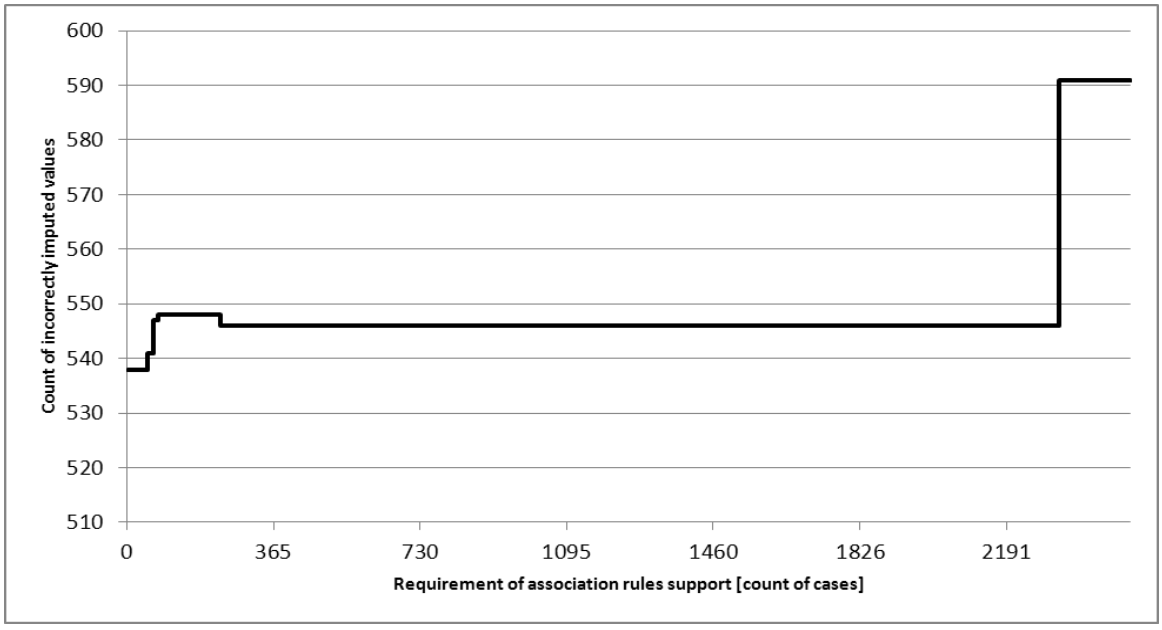

Fig. 1: Count of incorrectly imputed values using combination of association rules and most common attribute values with different requirement of minimum association rules support (1st example)

Next table and figure show example where minimum of incorrectly imputed values is reached with minimum required support set within the interval (60; 86> counts).

Tab. 3: Incorrectly imputed values using combination of association rules and most common attribute values with different requirement of minimum association rules support (2nd example)

Requirement of min. association rules support	Incorrectly imputed values	
[count of cases]	[count]	[%]
<0; 2>	568	21,5233
(2; 3>	569	21,5612
(3; 38>	570	21,5991
(38; 60>	568	21,5233
(60; 86>	566	21,4475
(86; 2289>	567	21,4854
(2289; ∞)	647	24,5169

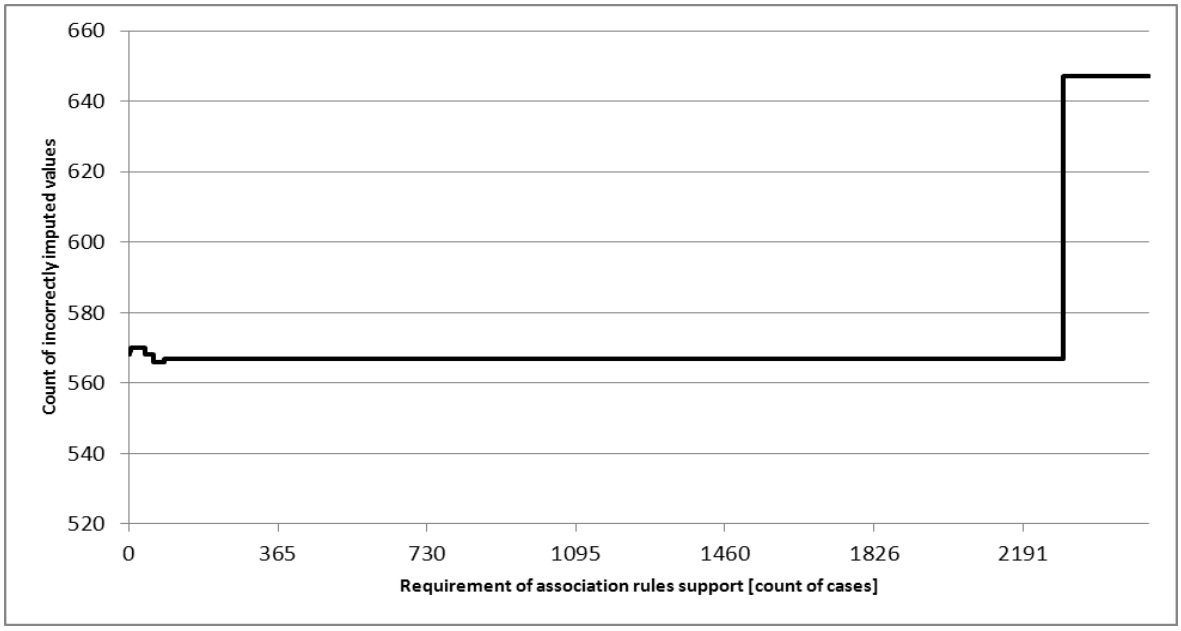

Fig. 2: Count of incorrectly imputed values using combination of association rules and most common attribute values with different requirement of minimum association rules support (2nd example)

5. Conclusion

Selection of missing values imputation method highly depends on given data set, structure of attributes and missing data mechanism. Missing data mechanism is a key factor to decide if missing values can be imputed using some of described methods. Missing data mechanism can be considered as missing completely at random, missing at random or not missing at random. If missing data mechanism is considered as not missing at random, imputation can not be done without knowledge of this mechanism. Unfortunately missing data mechanism is usually unknown.

Some analytical methods have their own mechanism for dealing with missing data so missing values imputation methods should be used only if necessary. It is also possible to use data set reduction by eliminating all missing values. This can be done by eliminating cases (rows) or/and attributes (columns) with missing values but this approach usually decrease the information content of the data.

Most often used missing values imputation methods are simple solutions like imputation using mean or most common value of given attribute. These methods don't consider dependencies among attributes.

Another possibility for missing values imputation is using described methods that are based on data mining methods like k-nearest neighbor, neural networks or association rules. These methods are more complicated and often don't represent exact procedure for imputation of missing values. Results of missing values imputation may vary based on setting of various parameters as shown on examples of missing values imputation using association rules.

Selection of missing values imputation method must be also done with consideration of structure of given dataset attributes. Some methods more suits for numeric attributes and some for symbolic attributes. Methods can be often combined.

Acknowledgements

The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomečková, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvárová, DrSc). The data resource is on the web pages http://euromise.vse.cz/challenge2003. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107.

Bibliography

Acuña, E. & Rodriguez, C., 2004: The Treatment of Missing Values and its Effect on Classifier *Accuracy, Classification, Clustering, and Data Mining Applications,* pp. 639-647, Available at: http://link.springer.com/chapter/10.1007%2F978-3-642-17103-1_60 [Accessed 19 September 2013]

Berka, P., 2003: *Dobývání znalostí z databází* 1st ed., Academia

Burita L., Gardavsky P., Vejlupek T. 2012: K-GATE Ontology Driven Knowledge Based System for Decision Support. *Journal of Systeme Integration* Vol 3, No 1, pp. 19-31

Chen, Z., 2001: *Data mining and uncertain reasoning: an integrated approach*. Wiley

Fahlman S. E. & Lebiere, Ch. 1991: The Cascade-Correlation Learning Architecture, Available at: http://www.ifi.uzh.ch/ailab/teaching/neuralnets2013/fahlman.pdf [Accessed 16 September 2013]

Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse, W. J. & Zheng X., 2005: A Closest Fit Approach to Missing Attribute Values in Preterm Birth Data, *Lecture Notes in Artificial Intelligence vol. 3642*, pp. 342–351, Available at: http://sci2s.ugr.es/MVDM/pdf/grzymala_busse_goodwin05.pdf [Accessed 12 September 2013]

Grzymala-Busse J. W., 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,* pp. 378-385, Available at: http://sci2s.ugr.es/keel/pdf/specific/congreso/grzymala_busse_hu01.pdf [Accessed 19 September 2013]

Gupta, A. & Lam, M., 1998: The weight decay backpropagation for generalizations with missing values, *Annals of Operations Research 78*, pp. 165-187. Available at: http://link.springer.com/article/10.1023%2FA%3A1018945915940 [Accessed 16 September 2013]

Hand, D. J., Manilla, H. & Smyth, P. 2001: *Principles of Data Mining*, A Bradford Book

Horton, N. J. & Kleinman, K. P., 2007: Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models, *The American Statistician* 61(1). Available at: http://www.math.smith.edu/~nhorton/muchado.pdf

HUANG, Ch. & LEE, H. 2004: A Grey-Based Nearest Neighbor Approach for Missing Attribute Value Prediction, Applied Intelligence 20, pp. 239–252. Available at: http://80.link.springer.com.dialog.cvut.cz/content/pdf/10.1023%2FB%3AAPIN.0000021416.41043..pdf

Improved outcomes software, 2004: *Nearest Neighbors Missing Value Estimation* [Online] (Updated 21 December 2004) Available at: http://www.improvedoutcomes.com/docs/WebSiteDocs/PreProcessing/Estimation_of_Missing_Values/Nearest_Neighbors_Missing_Value_Estimation.htm [Accessed 12 September 2013]

Kantardzic M. 2003: Data Mining – Concepts, Models, Methods, and Algorithms, *IEEE*, pp. 165-176.

Lakshminarayan, K., Harp S. A. & Samad, T., 1999: Imputation of Missing Data in Industrial Databases, *Applied Intelligence 11,* pp. 259–275. Available at: http://www.ime.unicamp.br/~wanderson/Artigos/imputation_industrial_databases.pdf

London School of Hygiene and Tropical Medicine, 2013: *Missingness mechanisms* [Online] Available at: http://missingdata.lshtm.ac.uk/index.php?view=category&id=40%3Amissingness-mechanisms&option=com_content&Itemid=96 [Accessed 2 September 2013]

Luengo, J., 2011: *Missing Values in Data Mining* [Online] Available at: http://sci2s.ugr.es/MVDM/index.php [Accessed 2 September 2013]

The MathWorks, Inc., 2013: *Impute missing data using nearest-neighbor method* [Online] Available at: http://www.mathworks.com/help/bioinfo/ref/knnimpute.html [Accessed 12 September 2013]

Mozer, M. C. & Smolensky, P., 1989: *Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment.* Available at: http://digitool.library.colostate.edu///exlibris/dtl/d3_1/apache_media/L2V4bGlicmlzL2R0bC9kM18xL2FwYWNoZV9tZWRpYS8xNjcwMjE=.pdf [Accessed 16 September 2013]

JEL Classification: C45, C82