RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2016 Volume 24, Number 39

PREPROCESSING RAW DATA IN CLINICAL MEDICINE FOR A DATA MINING PURPOSE

Andrea PETERKOVÁ, German MICHAĽČONOK

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA,
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA,
INSTITUTE OF APPLIED INFORMATICS, AUTOMATION AND MECHATRONICS,
ULICA JÁNA BOTTU 2781/25, 917 24 TRNAVA, SLOVAK REPUBLIC
e-mail: andrea.peterkova@stuba.sk, german.michalconok@stuba.sk

Abstract

Dealing with data from the field of medicine is nowadays very current and difficult. On a global scale, a large amount of medical data is produced on an everyday basis. For the purpose of our research, we understand medical data as data about patients like results from laboratory analysis, results from screening examinations (CT, ECHO) and clinical parameters. This data is usually in a raw format, difficult to understand, non-standard and not suitable for further processing or analysis. This paper aims to describe the possible method of data preparation and preprocessing of such raw medical data into a form, where further analysis algorithms can be applied.

Key words

Raw medical data, preprocessing techniques, data mining

INTRODUCTION

Data mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. These data sets must be however prepared in a satisfactory form for further processing with data mining methods. Data mining involves methods of artificial intelligence, machine learning and statistical analysis. The overall goal of the data mining process is to extract novel information from a data set and transform this information into an understandable structure for further use (1, 2). Medical data can provide such novel information, because it is usually just stored in the raw format and has not been analyzed in detail yet.

Nowadays, exploring and analyzing medical data is a very topical issue, because it is often stored in a way, in which it cannot be easily analyzed. It is common that this data is also usually of a very low quality. This is why some preprocessing techniques should be used to enhance the quality of the medical data. Some of these techniques includes data cleaning, data normalization, data transformation, feature extraction and selection. The main goal of data pre-processing is to obtain a final training data set, which could be used as an input to the chosen data mining algorithms. The proposed research deals with medical data from the field of cardiology which will be used for cardio-vascular diseases prediction.

In our research, we focus on a cardio-vascular disease called coronary artery disease. It is also known as ischemic heart disease. We are dealing with the data from biochemical and hematology examinations, which were taken during patient hospitalization. Another type of data used is from CT and ECHO screening examinations. The data from these screening examinations are already focused on the target organ, in our case the heart.

DATA PREPROCESSING

Before performing any step of data preprocessing, it is necessary to explore the source data set for possible problems and endeavor to correct the errors. For any real-world data set, doing this task manually is completely out of the question. The process of data cleaning is laborious, time consuming, and itself prone to errors. Useful and powerful tools that automate or greatly assist in the data cleansing process are necessary and may be the only practical and cost effective way to achieve a reasonable quality level in existing data (3).

The medical data about patients is usually stored in non-structured databases, where all the information, such as results from examinations and medical findings are written by a doctor in a text format. For each patient, there usually exists more than one text document with this valuable information. This fact shows the difficulty of extracting all the necessary parameters. To be able to further analyze the medical data with data mining algorithms, it is necessary to extract these parameters from plain text format documents to the structured database.

The first step in the process of parameter extraction was to identify the parameters to be extracted. For this purpose, a medical hypothesis was made by a cardiological expert. This hypothesis describes the aims of the final research and determines all parameters, which will be needed to successfully perform the research. In this step were chosen the mentioned parameters (results from laboratory examinations, results from screening examinations and chosen clinical data).

Before the extraction process itself, it is necessary to prepare the database structure for extracted data. Because of this, each of the chosen parameters had to be assessed and the correct datatype had to be chosen for each parameter.

To automate the process of parameters extraction from the text documents, an application in C# language was developed. This application is able to load necessary documents for each patient and search for the parameters in them. After finding the relevant parameter, it is written to the prepared database.

After the data extraction process, we were able to carry out pre-processing steps. Data pre-processing is used to transform the data into a suitable form for the analysis and further processing using chosen algorithms of data mining. The correct representation and quality of data is a necessity before starting the data analysis. To get the best data quality it is necessary to make these preprocessing steps:

1. Data cleaning – Cleaning of the data is the process of detecting and correcting invalidated or inaccurate records from a database. It involves removing typographical errors or validating and correcting values. In our dataset is was necessary to deal with missing values, where not every patient had 100% of the parameter values. Some of the examination results were not complete. Each patient with missing values had null values for different parameters.
2. Data transformation – The aim of this step is to convert a set of data values from a source data storage system into the data format of a destination data storage system. In this step,

it is important to analyze datatypes of all features in the given data set. With the information about data types it is possible to design the structure of the source database. The medical data from our research had to be transformed from plain text format into the database system.

3. Feature extraction - starts from an initial set of measured data and builds derived features intended to be more informative and non-redundant. In some cases it can lead to better human interpretations.

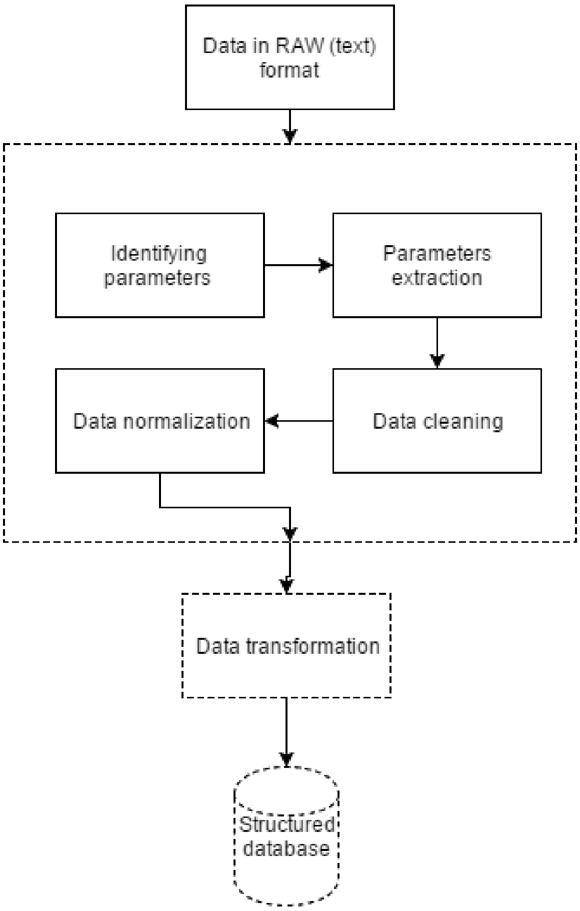

Fig. 1 *Medical data preprocessing diagram*

DATA CLEANING PHASE

There are many issues in data cleaning. One of them is the term “dirty data” known mainly in the business world (4). Recently, Kim (5) proposed a taxonomy for dirty data. However, there is no commonly agreed formal definition of data cleaning. Various definitions depend on the area in which the data cleaning process is applied. The following three phases define data cleaning as a process:

- Definition and determination of the error types.
- Searching and identifying the error instances.
- Correcting the uncovered errors.

The described medical data had also issues, which had to be solved before applying data mining algorithms. One of the most important issues to solve was missing attributes values. As mentioned before, many patients had not complete data from various types of examinations. There can be variety of reasons why data sets are affected by missing attribute values. Some of the attribute values may be not recorded because they are irrelevant at the time of making the record. For example, a doctor was able to diagnose a patient without some medical tests, or the record was not put to evidence, forgotten, damaged or erased. Solving the problem of missing attribute values is very important for further applying data mining methods and also for performing statistical analysis. Various methods exist to deal with missing attribute values. In general, methods to handle missing attribute values belong either to sequential methods or to parallel methods (3). According to the aim of the data analysis, there are many ways, how to correct missing attributes:

- Deleting cases / records with a missing attribute values

This method would ignore patients with missing attributes. All cases with missing values would be deleted from the data set. This approach is however not appropriate for dealing with missing values in medical data. There are more than 40% of patients engaged in the research with missing attribute values. This approach will dramatically decrease the number of records which will cause the inability to perform data mining to a degree of accuracy as is required.

- Add the most common value of a missing attribute

It is one of the simplest methods to handle missing attribute values. These values are replaced by the most common value occurring in the data set. This approach is much more applicable to the given data than deleting the records completely. However, this approach would not provide statistically valuable replacement values because one parameter type will be replaced with one value for every patient with that missing attribute.

Table 1 Example of missing values of chosen parameters

Patient ID	**Parameters**			
	GLU	KREAT	UREA	K+
0122	?	83	5.6	5.25
0123	5.45	59	3.4	4.27
0124	5.81	?	?	4.34
0125	5.45	83	7.5	?
0126	6.28	71	3.4	4.69
0127	8.47	83	6.9	4.5
0128	5,45	?	?	4.34

Table 2 Replacement of missing values by adding the most common value based on the values from the table 1

Patient ID	**Parameters**			
	GLU	KREAT	UREA	K+
0122	5.45	83	5.6	5.25
0123	5.45	59	3.4	4.27
0124	5.81	83	3.4	4.34
0125	5.45	83	7.5	4.34
0126	6.28	71	3.4	4.69
0127	8.47	83	6.9	4.5
0128	5,45	83	3.4	4.34

- Global closest fit, this method is based on replacing the missing values by the know value, that resembles as much as possible the case with the missing value. In the process of searching the closest fit, two vectors of attribute values are computed. One of the vector corresponds to the case with the missing values and the other vector the candidate for the closest fit. For each case the distance is computed. The case for which the distance is the smallest one is the closest fitting case that is used to determine the missing values.

Let x and y be two cases. The distance between cases x and y is computed as follows:

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

where

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

Where *r* is the difference between the maximum and minimum of the known values of the numerical attribute with a missing value. If there is a tie for two cases with the same distance, a kind of heuristics is necessary. This method of computation of missing values is most appropriate for the case of the medical data in the described research because the value of the computed replacement values is much higher than in previous possible methods.

The following two tables (Table 3 and Table 4) show the example of calculating missing values using the global closest fit method. The example shows computation of parameter called GLU for the patient with the ID 0122. First the distances between the case of patient 0122 and other cases were calculated. The missing value was then replaced with the value of the case which had the smallest distance from the case of the patient 0122, which was (according to table 3) the case of the patient 0127.

Table 3 Illustration of calculated distances between cases from table 1

d (122,123)	*d (122,124)*	*d (122,125)*	*d (122,126)*	*d (122,127)*	*d (122,128)*
3,54	3,93	2,46	4,57	2,09	3,93

Table 4 Replacing missing value with the global closest fit

Patient ID	**Parameters**			
	GLU	KREAT	UREA	K+
0122	8,47	83	5.6	5.25
0123	5.45	59	3.4	4.27
0124	5.81	?	?	4.34
0125	5.45	83	7.5	?
0126	6.28	71	3.4	4.69
0127	8.47	83	6.9	4.5
0128	5,45	?	?	4.34

CONCLUSION

In our paper, we have focused on the issue of raw clinical data preprocessing. The first step in data preprocessing was the initial raw data analysis and choosing appropriate parameters to be extracted for further use. After the analysis, a new database for parameters was designed. For the extraction of parameters from the plain text raw format an application in C# language was developed. This application automated the process of extracting all necessary parameters about the patients from various text sources to the new designed database. After parameters extraction, the data cleaning phase was performed. In this phase the data were repaired and missing values were computed using the global closest fit method. Cleaned data will be further used for data mining purposes. Data preprocessing is one of the most important issues in the process of data mining, because all raw data sources are different and therefore it is hard to develop a global methodology for data preprocessing. This applies especially to medical data, when the format of collected data is fully dependent on the doctor (for example using abbreviations). If the data is not clean and transformed correctly at the beginning, the incorrect results may occur in the further phases of data analysis. That is why it is necessary to define the types of errors in the data and correct these errors to prevent false information.

Acknowledgement

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

References:

1. BERRY, M. J., LINOFF, G., 1997. *Data mining techniques: for marketing, sales, and customer support.* John Wiley & Sons, Inc.
2. LAROSE, D. T., 2014. *Discovering knowledge in data: an introduction to data mining.* John Wiley & Sons.
3. GRZYMALA-BUSSE, J. W. Handling missing attribute values. Data mining and knowledge discovery handbook. Second edition. Springer New York Dordrecht Heidelberg London. ISBN 978-0-387-09822-7
4. HERNÁNDEZ, M. A., STOLFO, S. J., 1998. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery 2.1*, pp. 9-37.
5. KIM, WON, et al., 2003. A taxonomy of dirty data. *Data mining and knowledge discovery 7.1*, pp. 81-99.