EDUCATION OF DATA MINING AS A NOVEL APPROACH IN CLINICAL AND HEALTH CARE PRACTICE

Jiří Jarkovský*, Klára Komprdová, Ladislav Dušek
Institute of Biostatistics and Analyses, Faculty of Medicine and Faculty of Science, Masaryk University, Brno, Czech Republic
**Corresponding author: jarkovsky@iba.muni.cz*

ARTICLE HISTORY
Received 8 December 2013
Revised 23 December 2013
Accepted 31 December 2013
Available online 31 December 2013

KEYWORDS
data mining
statistics
multivariate analysis
large datasets
databases

ABSTRACT — *Data mining (DM) is a widely adopted methodology for the analysis of large datasets which is on the other hand often overestimated or incorrectly considered as a universal solution. This statement is also valid for clinical research, in which large and heterogeneous datasets are often processed. DM in general uses standard methods available in common statistical software and combines them into a complex workflow methodology covering all the steps of data analysis from data acquisition through pre-processing and data analysis to interpretation of the results. The whole workflow is aimed at one final goal – to find any interesting, non-trivially hidden and potentially useful information. This innovative concept of data mining was adopted in our educational course of the Faculty of Medicine at the Masaryk University accessible from its e-learning portal http://portal.med.muni.cz/clanek-318-zavedeni-technologie-data-miningu-a-analyzy-dat-genovych-expresnich-map-do-vyuky.html.*

INTRODUCTION

The term "data mining" (DM) is currently widespread in all areas related to data analysis. Clinical research belongs to them as well and the application of complex computational methods has become very popular in this area because of increasing amount of available data. The DM concept is nevertheless often overestimated or incorrectly considered as a universal solution for all problems. Although data mining seems to be well defined, the opposite is true. Even its definition is problematic and there are many definitions books and web portals dealing with the data mining. There are two probably the most popular definitions: "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [1] and "The science of extracting useful information from large data sets or databases" [2].

In the article we would like to introduce our educational materials presenting concepts and approaches of data mining for clinicians and other researches in clinical and health care fields.

DM is mostly considered in the relation to large datasets; its usage in the commercial applications is common as well. In fact, the DM is universal methodology applicable to any data analysis and it is not "owned" by any area of science. The DM has been adopted in wide area of applications, such as searching of risk clients, non-legal usage of credit cards, e-mail classification and spam messages detection, text and speech recognition or molecular data analysis. Therefore, the DM is the area of science where its development is multidisciplinary in its nature. Methods applicable in commercial applications can be applied in any other research areas and vice versa.

Data mining is often connected to an idea of genial machine mining previously unknown information from the data and the methodology is often presented as a "black box" with simplified description. The reality is of course more rational. Good knowledge of mathematical background of the DM methods and their limitations is crucial for the correct application of the DM; the most important is expert knowledge and long-term experience. Methods applied in the DM are principally multivariate and have to follow all rules of multivariate data analysis. The benefits of multivariate methods are as follows [3]:

- Visualization of data with multiple variables
- Searching of meaningful views on multivariate data, identification of importance and hierarchy of variables
- Identification of correlations among variables, simplification of their structure
- Analysis of similarities between analysed subjects, their stratification, classification and predictionn

The question is whether the data mining is in any way different from the commonly adopted statistical methods? The answer is both yes and no. DM uses methods available in common statistical packages and "mining" can be sometimes used as a marketing term only. On the other hand, even common statistical methods are used in novel, complex and logically joined context. The real DM is a standardized complex methodology covering all the steps of data analysis from data acquisition through pre-processing and data analysis to interpretation of the results; the example is CRISP-DM, JDM (Java Data Mining) or complex methods of model description such as PMML (Predictive Model Markup Language). The data mining thus brings new quality in data analysis which is more related to innovative combination of methods than to any single method. DM in the hands of experienced data analyst is an important tool of scientific data analysis to be applied on complex heterogeneous multivariate data.

The workflow of data mining can be separated into simple individual steps from data storage and preprocessing to their description and predictive modelling. The individual steps can be performed in various software, such as Statistica, SPSS, SPSS Modeler, S+, Matlab, WEKA or R.

METHODS

Workflow of data mining

As already mentioned, data mining can be considered as an innovative connection of various methods of multivariate data analysis. Methodology of the complex DM approach always incorporates process workflow of analytical steps. Example of such approach is the CRISP DM methodology describing life cycle of DM project and their interconnections [4]; this methodology as one of the most general approaches available was also adopted in our article and educational materials.

According to CRISP-DM methodology the DM project life cycle consists of six phases; their order and direction of crossing between them is not strictly given and the movement in the scheme is based on the results of the previous phase (the arrows in the scheme shows the most common paths). The outer circle symbolizes cyclical nature of data analysis which is repeated until the solution is found. The knowledge gained in one cycle can generate new questions and new cycles utilizing experiences from the previous cycles.

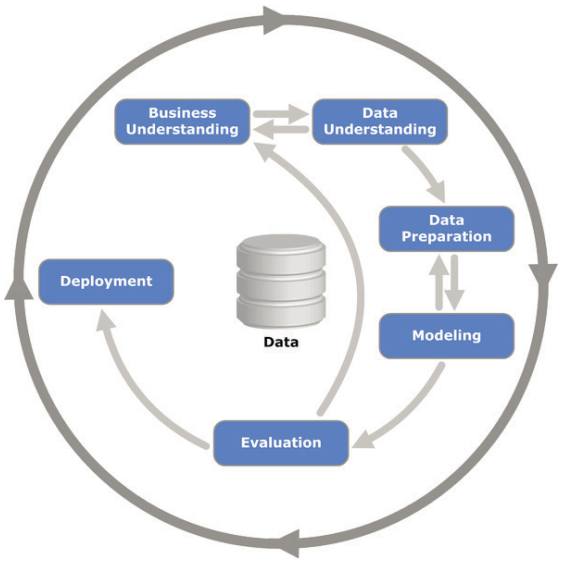

Figure 1: DM workflow according to CRISP-DM methodology (taken from CRISP-DM)

Understanding

This initial phase focuses on understanding the analysis objectives and requirements, and then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. For example, in clinical data analysis this is the preliminary phase of literature review of given clinical problem (terminology, cut-offs, known correlations of variables etc.). Although it looks rather simple, this information is strategically important during the multivariate analysis. Limited knowledge on importance and meaning of variables can resulted into biased or uninterpretable results and during multivariate analysis these problems should not be necessarily revealed. Part of the preliminary phase should be also the power analysis and assessment of the necessary sample size.

Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. Wide set of univariate and multivariate analyses can be adopted for this exploratory analysis (Figure 2).

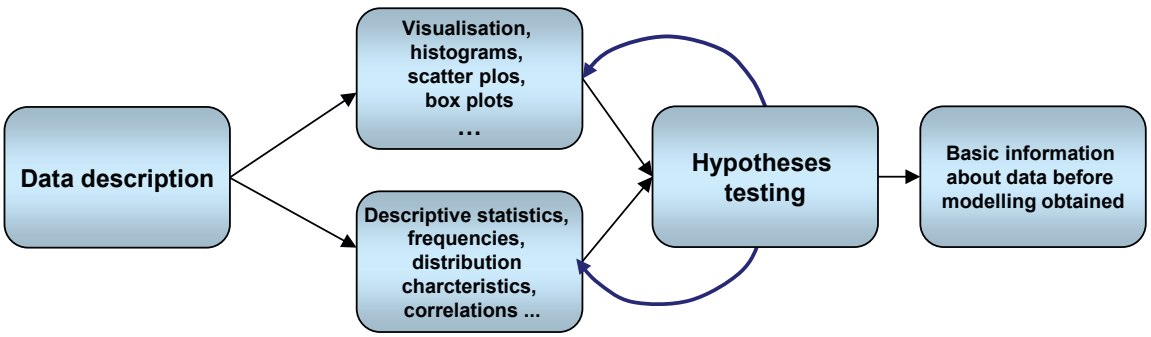

Figure 2: Example of methods applicable for data understanding

Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Tasks include table, record, and attribute selection as well as transformation and cleaning of data. These data processing tasks do not have to be necessarily done in one step; it is more likely to have several data processing steps according to progress of exploratory analysis – for example in the first step we find non-normal distribution of data and data transformation follows; than we return to phase of normality assessment to prove it on transformed data.

Modelling

In this phase, various modeling techniques are selected and applied (typically, there are several techniques for the same data mining problem type). Some techniques have specific requirements on the form of data and, therefore, stepping back to the data preparation phase is often needed (Figure 3). Because the DM is commonly used on large datasets, methods of so called machine learning are often adopted [4]. The best known are neural networks, classification and regression trees, association rules, regression techniques (GLM, GAM, logistic regression as special approach of GLM), time series analysis of for clinical data very common survival analysis (Cox proportional hazards model for example).

The methods can be divided into supervised and unsupervised learning. During the supervised learning the presence of endpoint itself determines the learning process; we have a set of cases with known result [5]. The model is trained on these known cases which serve as a reference dataset for the evaluation of new case. The example can be a predictive model for probability of occurrence of given event in patients according to their initial characteristics. In this situation the model is developed on the base of reference dataset of patients with and without given event that differs in the values of potential predictors. During unsupervised learning we are looking for structure within the dataset based on similarities of cases. We are searching for typical patterns in the data (for example clusters of patients with similar characteristics).

Supervised learning methods are further divided into classification and regression according to the dependent type of variable. Regression is adopted for the continuous variables (blood pressure etc.); classification analysis is used for categorical variables.

Some of the above mentioned methods are sometimes called "black box" but it is not correct to understand this as an unknown principle of reaching the results; exact description of the model is necessary in these methods as well. In the last years the term "white box" is used for DM methods and it will hopefully modify the reputation of these methods; neural networks can serve as a typical example [6].

Evaluation

At this stage in the project, model(s) that appear to be correct and applicable from the data analysis point of view are built. Before proceeding to final decision about the model, it is important to be sure it properly achieves the project objectives. Model(s) should be validated and confronted with the reality to assess their general application for practice. It is also important to find whether there is no problem unaddressed by the analysis. This part of the DM process is based on analytical results, such as metrics of models quality (analysis of variance, BIC, AIC, sensitivity, specificity, etc.) or validation using independent dataset.

Deployment

Development of the predictive model and obtaining the data analysis results is not the final step of the DM project. Even in case of project aimed on description of the data the results have to be adequately presented. Requests on this phase cover wide range of tasks from descriptive report over scientific publication to implementation of the DM process. In most of the projects this phase is connected with results

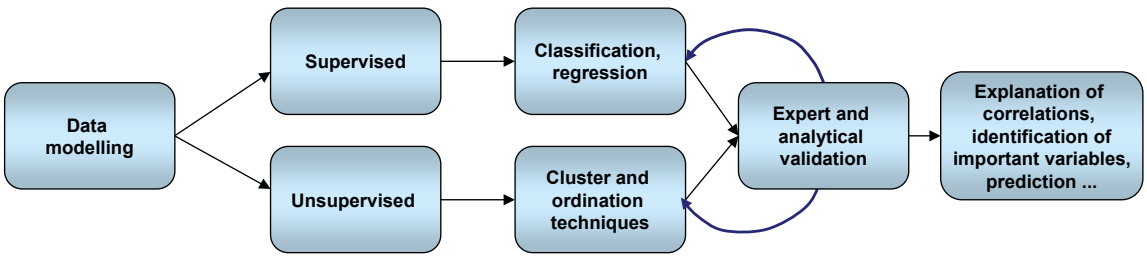

Figure 3: Methods applicable for data modelling

interpretation and requires expert knowledge of the analysed problem.

RESULTS: E-LEARNING COURSE ON DATA MINING

From the above description of data mining it is evident that the DM is a methodological concept based on innovative combination of univariate and especially multivariate statistical methods. Optimised analytical plan of the DM is the main guarantee that useful information will be "mined" from the complex multivariate data. This concept was also adopted in the preparation of educational course of the Faculty of Medicine at the Masaryk University "Introduction of data mining technology and gene expression maps analysis into courses of Faculty of Medicine" [7] accessible from its e-learning portal http://portal.med.muni.cz/clanek-318-zavedeni-technologie-data-miningu-a-analyzy-dat-genovych-expresnich-map-do-vyuky.html.

SUMMARY

Data mining techniques proved to be usefool tool for the analysis of clinical data [8,9]. Our educational materials are aimed on clinicians and other non-statistical users of these techniques to provide them with information about the process of data mining project based on CRISP-DM methodology and overview of the main analytical methods applicable during its steps; we hope these materials will help to spread out correct understanding of the utility of the data mining approach and its advantages and limitations.

RNDr. Jiří JARKOVSKÝ, Ph.D.

ACKNOWLEDGEMENTS

This work was supported by the Czech Science Foundation, project no. 102/09/H083.

REFERENCES

[1] Frawley W, Piatetsky-Shapiro G, Matheus C. Knowledge Discovery in Databases: An Overview. AI Magazine 1992; 13(3): 57–70.
[2] Hand D, Mannila H, Smyth P. Principles of Data Mining. MIT Press: Cambridge, MA 2001. ISBN 0-262-08290-X.
[3] Nisbet R, Elder RJ, Miner GD. Handbook of Statistical Analysis and Data Mining Applications. Academic Press 2009. ISBN-10: 0123747651.
[4] Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehousing 2000; 5(4): 13–22.
[5] Bishop CM. Neural Networks for Pattern Recognition. Oxford University Press: Oxford 1996. ISBN-10: 0198538642.
[6] Hastie T., Tibshirani R. & Friedman J. H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer: New York, 2009. ISBN-10: 0387848576
[7] Jarkovský J, Némethová D, Gelnarová E, Budinská E, Kubošová K, Kokrment L, Dušek L. Zavedení technologie data miningu a analýzy dat genových expresních map do výuky. Multimediální podpora výuky klinických a zdravotnických oborů: Portál Lékařské fakulty Masarykovy univerzity [On-line] Available at WWW: <http://portal.med.muni.cz/clanek-318-zavedeni-technologie-data-miningu-a-analyzy-dat-genovych-expresnich-map-do-vyuky.html>.
[8] Hian Chye H, Tan G. Data mining applications in healthcare. J Healthcare Inform Manage 2011; 19(2): 64–72.
[9] Obenshain M K. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004; 25(8): 690–695.