Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

© 2018 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-852-5-371

Medical Data Analytics Is Not a Simple Task

František BABIČ[a], Michal VADOVSKÝ[a] and Ján PARALIČ[a]
[a] *Department of cybernetics and artificial intelligence, Faculty of electrical engineering and informatics, Technical university of Kosice, Letná 9, 042 00, Kosice, Slovakia*

Abstract. Data analytics represents a new chance for medical diagnosis and treatment to make it more effective and successful. This expectation is not so easy to achieve as it may look like at a first glance. The medical experts, doctors or general practitioners have their own vocabulary, they use specific terms and type of speaking. On the other side, data analysts have to understand the task and to select the right algorithms. The applicability of the results depends on the effectiveness of the interactions between those two worlds. This paper presents our experiences with various medical data samples in form of SWOT analysis. We identified the most important input attributes for the target diagnosis or extracted decision rules and analysed their interestingness with cooperating doctors, for most promising new cut-off values or an investigation of possible important relations hidden in data sample. In general, this type of knowledge can be used for clinical decision support, but it has to be evaluated on different samples, conditions and ideally in long-term studies. Sometimes, the interaction needed much more time than we expected at the beginning but our experiences are mostly positive.

Keywords. medical data, diagnosis, variables, knowledge

1. Introduction

SWOT analysis has its origins in the 1960s [2]. The SWOT (Strengths, Weaknesses, Opportunities and Threats) represents a study undertaken by an organization to identify its internal strengths and weaknesses, as well as its external opportunities and threats [1]. We decided to use this analytical tool for evaluation of our performed experiments with different medical data samples. We aimed to identify the most important best practices as well as pitfalls based on our real experiences.

2. SWOT analysis

Over the last 3 years we have analyzed several medical data samples like diagnosis of Metabolic Syndrome (MS), Mild Cognitive Impairment (MCI), Parkinson's disease (PD) or hepatitis. We shared our results with the community through various conferences or journal papers [3], [4], [5], [6], [16]; and now we want to summarize our experiences in the form of a SWOT analysis.

2.1. Strengths

Exploratory data analysis helps understanding the data from the overall view. In general, we used this approach in the early stage of the analytical process to understand the data and to create a first bridge between us and medical experts. Typically, we used methods like histograms or boxplots to investigate an attribute values distribution.

Investigation of possible hidden relations between available input attributes. This investigation can improve the quality of analyzed dataset, e.g. we're able to remove some redundancy or multicollinearity. We applied either parametric Pearson r correlation test or non-parametric Spearman correlation test to identify the highest correlation between input attributes, e.g. in case of MCI data sample value 0.96 between hemoglobin and hematocrit. In the case of regression analyses, we used variance inflation factor (VIF) to detect the multicollinearity. The VIF value 1 means no influence of regression coefficients by the collinearity. The value between 1 and 5 means moderate influence and value higher than 5 signalizes high impact of the collinearity [11].

Evaluation of possible relations between the input attributes and the target diagnosis (binary, ordinary). We used statistical tests like Shapiro-Wilk normality test for numeric attributes; non-parametric Mann-Whitney-Wilcoxon test for a combination of numeric attributes without normal distribution and binary target diagnosis; Welch Two Sample t-test for numeric attributes with normal distribution and binary target diagnosis; non-parametric Kruskal-Wallis rank sum test for numeric attributes without normal distribution and ordinary target attribute; Pearson's Chi-squared test or Fisher test for nominal attributes. Typically, we evaluated the null hypotheses on the 0.05 significance level. Also, we used the logistic regression (LR) method; the Odds ratio with the 95% confidence interval and McFadden's R squared Test to evaluate a predictive power of generated LR models.

Investigation of individual cut-off values for particular input attribute. In general, the medical experts use empirically set cut-off values of particular biomarkers with respect e.g. to some disease. It may be useful to calculate individual cut-off values related to the given set of patients. For this purpose, we use the Youden's index [9], [10]. The pair of sensitivity and specificity proportions characterizes the performance of the diagnostic attribute (Figure 1a). Domain expert plays important role in this task, it is always necessary to discuss about potential applicability of computationally derived cut-off values with him/her.

Decision models (often transformed into the form of decision rules) generated based on historical data. The decision tree is a flowchart-like tree structure, where each non-leaf node represents a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent target classes or class distributions [7]. We used this type of prediction model because of its simply understandable representation (Figure 1b) in comparison to other classification models like neural networks, Naïve Bayes or Support Vector Machine. The right choice depends on a specific task and available data sample. For example, IBM Watson team experimented with various machine learning algorithms and finally used a logistic regression as the most robust solution in [8]. Typically, we applied C4.5, C5.0, CART, CHAID or Random Forests. In the case of PD, we used methods like bagging and boosting to generate more power prediction models (increasing accuracy by about 10%).

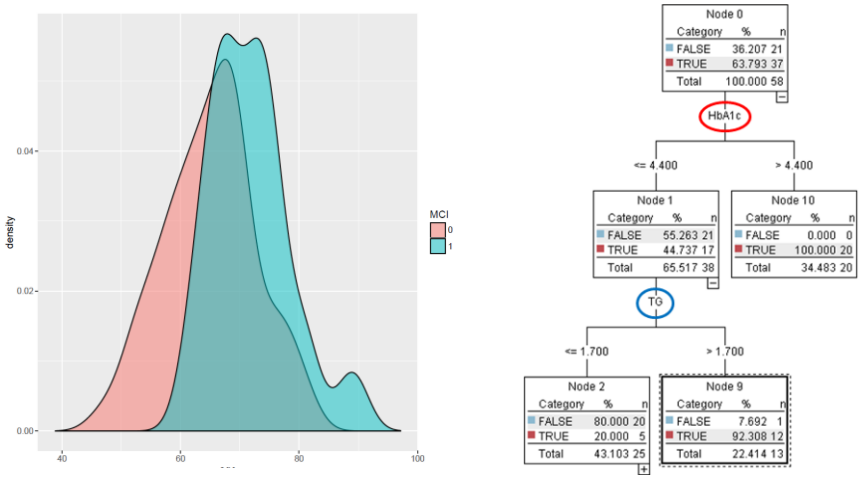

Figure 1a. Distribution function for the attribute Age (x-axis) and the target diagnosis MCI (y-axis) with identified cut-off value (intersection, 73). **1b**. Example of the decision tree for MS diagnosis - women seem to be more prone to diabetes and used metabolic variables associated with diabetes (triglycerides).

2.2. Weaknesses

Clinical decision support system (CDSS) requires a complex knowledge verified in a wider context. Before the CDSS can be deployed in clinical practice [12], important steps need to be performed such as validation of the patient data, modeling of medical knowledge, maintenance of medical knowledge base, reasoning and validation. This approach requires a partner or partners from medical or industry area (data from the intelligent devices) to ensure an effective knowledge transfer. Therefore, we're active in the H2020 proposals preparation.

Lots of results, not each of them is useful. It is not an easy task to select the most useful and helpful knowledge. The data analytics is an iterative and interactive process, for which the most common methodology is CRISP-DM [13]. During modelling phase, we produced typically many different results like models, rules or cut-off values that needed to be evaluated by medical experts. We tried to simplify this evaluation based on appropriately selected graphical form or metrics like confusion matrix, ROC curve (receiver operating characteristic curve), AUC (Area under curve), accuracy. It is important to optimize the models to meet the specified objectives in the best quality.

The results are affected by the size of the input sample and often too small datasets from the statistical point of view are available. From our point of view, this was the main problem we dealt with. For example, we had a good real sample for MCI diagnostics but it contained less than 100 records. We realized a literature review and found out that e.g. maximum likelihood estimation including logistic regression with less than 100 cases is “risky”, that 500 cases is generally “adequate,” and there should be at least 10 cases per predictor [14]. The good inspiration is work of the Holzinger group [15] with its interactive machine learning. Also, we applied some methods for variables reduction like Principal Component Analysis, LASSO or forward and backward stepwise regression (models evaluated by Bayesian information criterion and Mallows’s C_p).

2.3. Opportunities

The volume of collected medical data is increasing. The Stanford Medicine 2017 Health Trends Report expects at least 48% increase per year, in 2020 to 2 314 Exabyte (one Exabyte = one billion Gigabytes). The popularity of wearable intelligent devices is growing very fast; global sales in 2017 will exceed 274 million devices in previous year (Gartner).

Cost containment, effective health care and improved quality of life for patients with various types of diagnoses. Provision of the effective healthcare motivates many initiatives from micro to the EU level. Some experts refer to data analytics as a core to the most successful cost containment strategies based on information-driven solutions (White paper, Discovery Health Partners). The predictive analytics may increase the accuracy of the diagnoses, support the preventive medicine and public health, or improve the costs management.

2.4. Threats

Specific domain knowledge, different vocabularies of data analysts and medical experts. The analytical process starts with agreement on the task, i.e. the domain expert sets up hypotheses and goals from the medical point of view and a data analyst determines the evaluation metrics and transforms the tasks to the data mining goals. Sometimes, we need more iterations to properly analyze medical doctor's expectations and to identify the correct analytical task. The analysts are usually not familiar with the meaning of specific blood tests or genomic characteristics. We can support our understanding with available information sources, but these cannot be grasped fast and easily. Usually face to face meetings are more efficient, because they are interactive and useful for both sides.

Limited generalization of resulted models in case of small data samples. This situation represents often a group of patients with relevant characteristics depend on their life style, living conditions and family history. We can extract some interesting and important attributes or their cut-off values but we cannot say that they are applicable in broader sense. We propose a concept of an intelligent knowledge base allowing an automatically or semi-automatically update of the imported knowledge in the form of decision rules.

It is complicated to obtain the real medical data. Who owns patient/medical/healthcare data? This question is more and more popular in several aspects. The patients want to get the best healthcare to improve their health status and quality of life; the doctors want to determine a correct diagnosis taking into account all conditions and relations; and the providers prefer the quality services in combination with cost containment. In all aspects, data analytics can help but only with appropriate access to the relevant data in the electronic medical records. And, the new EU General Data Protection Regulation (GDPR) is the most important change in data privacy regulation in recent years, which may even complicate the current situation.

3. Conclusion

Some experts point out that the doctors will become the data analysts in the near future. John Mattison, the chief medical information officer at Kaiser Permanente, predicts

a new data environment integrating all peoples' personal data and this ecosystem will be important source for the healthcare. The supporting tools for data analytics are every day more and more user-friendly and comfortable for the users without or with less analytical background. Meantime, the role of data analyst is still important and success of the performed analytical process depends on the collaboration with participating medical experts. We evaluated our experiments within SWOT analysis. Even though our findings depend on the characteristics of the given samples, we confirmed their applicability in some general context and will reuse them in our future work, e.g. data sample related to the brain attack or quality of sleeping.

Acknowledgment. This work was supported by the Slovak Research and Development Agency under the contract No. APVV-16-0213; the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16 and by the COST Action TD1405.

References

[1] M.M. Helms, J. Nixon, Exploring SWOT analysis – where are we now?: A review of academic research from the last decade, *Journal of Strategy and Management*, **3(3)** (2010), 215-251.
[2] E.P. Learned, C.R. Christensen, K.E. Andrews, W.D. Guth, *Business Policy: Text and Cases*, Irwin, Homewood, IL, 1965.
[3] F. Babič, L. Majnaric, A. Lukáčová, J. Paralič, A. Holzinger, On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive modelling based on Machine Learning, *Lecture Notes in Computer Science, Springer International Publishing* **8649** (2014), 118-132.
[4] A. Lukáčová, F. Babič, Z. Paraličová, J. Paralič, How to Increase the Effectiveness of the Hepatitis Diagnostics by Means of Appropriate Machine Learning Methods, *Lecture Notes in Computer Science* **9267** (2015), 81-94.
[5] F. Babič, M. Vadovský, M. Muchová, J. Paralič, L. Majnarič, Simple Understandable Analysis of Medical Data to Support the Diagnostic Process, *Proceedings of the SAMI Conference* (2017), IEEE, 153-158.
[6] F. Babič, J. Paralič, M. Vadovský, M. Muchová, A. Lukáčová, Z. Vantová, What is a Relation between Data Analytics and Medical Diagnostics?*, International Journal on Biomedicine and Healthcare* **5(1)** (2017), 8-12.
[7] S.K. Murthy. *Automatic* Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery* **2(4)** (1998), 345-389.
[8] R.E. Hoyt, D.H. Sniner, C.J. Thomson, S. Mantravadi, IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics, *JMIR Public Health Surveill* **2(2)**, e157, 2016
[9] E.F. Schisterman, N.J. Perkins, A. Liu, H. Bondell, *Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples*, *Epidemiology* **16** (2005), 73-81.
[10] J. Yin, L. Tian, Joint confidence region estimation for area under ROC curve and Youden index, *Statist. Med*. **33(6)** (2014), 985–1000.
[11] G. James, D. Written, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer-Verlag New York, 2013
[12] M.A. Musen, B. Middleton, R.A. Greenes, Clinical Decision-Support Systems, *Shortliffe E., Cimino J. (eds) Biomedical Informatics* (2014), Springer, London, 643-674.
[13] C. Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of Data Ware-housing* **5(4)** (2000) 13–22.
[14] J.S. Long, *Regression models for categorical and limited dependent variables*, Thousand Oaks, CA: Sage, 1997
[15] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, *Brain Informatics* **3(2)** (2016) Springer, 119-131.
[16] M. Vadovský, J. Paralič, Parkinson´s Disease patients classification based on the speech signals, *Proceedings of the SAMI Conference (*2017), IEEE, 321-325.