INTERDISCIPLINARY APPLICATION OF ALGORITHMS FOR DATA MINING

Jasmin Malkić[1]
Nermin Sarajlić

Original scientific paper

Open Link International GmbH, Germany
University of Tuzla, Faculty of Electrical Engineering

Received: 30.08.2013.
Accepted: 07.09.2013.

UDC: 004.62.021:61
004.8:004.65

ABSTRACT

Interdisciplinary application of data mining is linked with the ability to receive and process the large amounts of data. Although even the first computers could help in executing the tasks that required accuracy and reliability atypical to the human way of information processing, only increasing the speed of computer processors and advances in computer science have introduced the possibility that computers can play a more active role in decision making. Applications of these features are found in medicine, where data mining is used in clinical trials to determine the factors that influence health, and examine the effectiveness of medical treatments. With its ability to detect patterns and similarities within the data, data mining can help determine the statistical significance, pointing to the complex combinations of factors that cause certain effect. Such approach opens the opportunities of deeper analysis than it is the case with reliance solely on statistics.

Key words: *data mining, clinical trials, cluster verification, distance distribution.*

INTRODUCTION

The mutual relation of events in health care and weather conditions is a subject of continuous research effort that put a special emphasis on the diurnal and seasonal variations in the type and number of medical symptoms. The conclusions drawn as a result of such research are generally based on large samples, which primarily increase the statistical significance and the confidence level. However, the main problem of verifying a hypothesis on the basis of its statistical significance is reflected in the fact that such process can overlook a hypothesis that does not seem significant, but in practice is very important. It is therefore necessary to extend the classical statistical approach of hypothesis verification, and explore new ways to use IT methods for processing medical data.

The escalation of health care costs, limitations in human resources and the necessity to build a cost-efficient health care system create a need for predicting the volume of patients in different time periods, which is the feature of data mining. In addition, improved understanding of the processes related to health care makes the medical interventions easier and encourages the positive organizational change. Practical importance of the data mining application in medicine is also reflected through the capability for defining the future improvements and establishing the integrated computing platform for the collection and analysis of medical data, which can be used as the resource for continuous planning of the health care activities.

[1]**Correspondence to:**
Jasmin Malkić, Open Link International GmbH
Friedrichstrasse 200, 10117 Berlin, Germany

CLUSTERIZATION METHODS AND THEIR PROPERTIES

In phenomenological terms, the algorithm for data mining represents mimicry of the process of human reasoning, with an objective to recognize patterns or anomalies and reach conclusions. In order for such algorithm to be functional, it is essential to make the strict formalization of concepts such as similarity or regularity. While notions of equality or identity are axiomatically defined and formalized by a simple binary operation, similarity and regularity are the result of a combination of several criteria. In addition, there is no universal definition for this combination, and for these reasons, data mining algorithms do not produce results in the form of deterministic statements. Final conclusions are instead formulated as reasonable assumptions, where the appropriate probability of an event is a measure of its grounds.

To discover the new information in large amounts of data is the main task of data mining, and the best results in it are achieved by combining human expertise in the problem description and computer capacity for examination of data (Kantardzic, 2003). From the functional point of view, data mining tasks can be divided into prediction and description, where the prediction, determines the variable's future values based on existing data, while description seeks patterns in a given model and displays them in some recognizable form. Most common processes of descriptive data mining are classification and clustering, with the task of identifying a finite set of categories that describe the input data. Unlike classification, where there is a priori knowledge of the number and shape of these categories, and the data is being classified into existing categories that it is most similar to, the task of clustering is first to identify the categories hidden in the input data, and then determine the affiliation of the individual data units. The task of clustering is therefore not predicting values of variables, but to divide the existing data space into relatively homogeneous subgroups or clusters, where the degree of similarity between data within the cluster is maximized, and the similarity with the data outside of cluster is minimized (Larose, 2005).

According to their function, algorithms for clustering themselves can be divided into algorithms for understanding the input data and algorithms for further use (Kumar et al., 2005). Application of algorithms for further use relates primarily to the data processing methods in which the degree of spatial complexity of O(m2) or higher, for example the case where the method includes calculating distances between all data points. With larger amounts of data this procedure may be impractical, and there is a need for summarization, in order to reduce the number of data. In such cases, a clustering method can point to the data that are similar to each other and can be ignored, so the distances can be calculated only for the points that represent a particular set of data, which is far more efficient. On the other hand, clustering algorithms for understanding the input data do not provide segregation, but work over the entire data set finding conceptually meaningful data sets that are close to each other on pre-established criteria and share the same characteristics.

Criteria for the classification of data into clusters have a decisive influence on the character of the clusters themselves. In this sense the clusters based on similarities with the prototype could be defined. Given that the central point of a cluster is typically taken as the prototype, clusters obtained by this criterion usually pursue some kind of hyperspherical form. Weakness of this approach is the difficulty of detecting clusters that deviate from such convex forms. This problem can be overcome to some extent if the input data is organized in the form of the graph, and data structure itself is taken as the clustering criterion. Such clusters make interrelated points that have no connection to points outside the cluster. In this case, the main criteria for the formation of clusters are related to the continuity of connection, rather than a form of data. However, even this approach encounters difficulties under certain conditions, because the individual links between distant points can lead that two or more different clusters appear as one. Both of these approaches are also sensitive to the so-called data noise - distant points that do not belong to any cluster. A unique set of criteria that for the arbitrary set of input data give the best result of clustering does not exist, but the criteria that successfully deal with the perceived shortcomings of the current approaches are continuously developed and improved. One of these criteria is the density of the data, where clusters are formed only from the points that are located in the areas with a sufficient number of other points from the same cluster. Clusters based on the density can be of any shape, and the effect of data noise is considerably reduced.

If the data model Y is displayed as a function of the input data set X and the parameters a and c, the process of descriptive data mining is given by equation (1).

$$Y = aX + \sum c_i \qquad (1)$$

If data density is taken as the criteria for cluster formation, then the parameter a is measure of that density, and c represents an additional parameter, such as the data noise threshold. The quality and usability of the data model in terms of shape, size, number, and even the meaning of the clusters depend primarily if the proper choice of parameters has been made. Improvement of methods for determining these parameters opens the possibility of improving the results of data mining too. In clustering algorithms which use the criterion of data density, this explicitly means to improve the methods of data noise isolation.

STATISTICS AND DATA MINING IN CLINICAL MEDICINE

Health care is one of the most information-intensive work disciplines in the modern world. Already at mid-90s, when the computerization of society only took hold, it was estimated that only one clinic can generate volume of about 5 TB of data per year (Huang et al., 1996), and that volume has been rising ever since. Crucial to improving of working conditions and performance in the field of medicine is to use this information in the right way.

Clinical trials are one of the principal sources of clinical data. Conducted for the purpose of a specific medical research, clinical trials are directed primarily at examining the effect of a therapy on the target group of patients. As its conclusion, such tests produce statistically significant findings, where the measure of significance is determined by different statistical methods. P-value method will confirm the finding statistically significant if the probability that it has happened due to the chance is less than or equal to 5%, which is expressed as p = 0,05. P-value is usually associated with the hypothesis that with the considerable probability it may be claimed that between the application of medical treatment, and its omission would not be any difference in the condition of examined patients, or that any observed difference is accidental. If it can be shown that the probability of this hypothesis is less than or equal to 5%, it is considered that the treatment caused a statistically significant change to the target group of patients.

During a clinical trial, there may be a need to prove a number of different hypotheses. To make sure that statistical significance of such hypothesis is correctly identified, they should primarily be formulated as the single independent logical statements. Although statistically correct, this approach can lead to overlooking the significant hypothesis based on a small number of samples, or their importance may be underestimated even if they are discovered.

The application of information technology in the health care is focused primarily on the structuring, searching and organizing data, so it can be used to support the problem solving and decision-making processes. Introducing information technology to the health sector opened up many opportunities for interdisciplinary research as well (Kan, 2003). The contribution of information technology in the field of clinical medicine is reflected, among other things, to optimize data processing and facilitating the pattern recognition using different methods of data mining.To recognize similar data, their groups and make a decision on placing the data into clusters is the cluster analysis task. Algorithms for clustering of multi-dimensional data space based on data density (such as DBSCAN), imply division of data on signal (data belonging to one of the clusters) and noise (data that will remain outside the cluster). As already pointed out, separating signal from noise eliminates the influence of outlier points to the formation of clusters and improves the quality of the analysis results. Assuming a sample large enough, even the small samples can be found within the signal and discovering patterns and regularities in the data is not dependent on their size which cannot be controlled, but of the total amount of data, which is controlled input variable. In this way, reliability of the statistical findings also does not depend on the size of observed patterns, so with a sample large enough it is significantly more difficult to overlook the small regularities in the data and hypotheses based on them. So, the application of data mining in taxonomic tasks such as classification and clustering ensures reliability of the final conclusions, particularly in the fields of research that deal with of large quantities of data such as clinical medicine.

CLUSTER ANALYSIS RESULTS VERIFICATION

The process of verifying the quality of a data model formed by clustering is hindered by the fact that there are no universal criteria of similarity of points that form a cluster, or in other words because the precise definition of a data clusters does not exist. In such circumstances ascend the need for developing the reliable methods for verification and validation of cluster analyses, or verifying the correctness of the algorithm and quantitative evaluation of the obtained clusters usability.

Additional confirmation of the results of clinical experiments is achieved by applying statistical tests on their results. A test proves similarity of an examined variable with predetermined test statistics of certain properties. If the test statistic is normally distributed, as in χ or Z test, then a positive result can be interpreted as a proof of randomness of the experimental results. Theoretical basis for this procedure can be formulated using the definition of a cluster based on the distribution randomness of its points, or the normality of distribution of their mutual distances from a certain point. Specifically, distances of the points in a multidimensional space without cluster structure follow the normal distribution, which reflects the randomness of their positions and the absence of patterns, regularities or anomalies that could represent clusters. Any deviation from this rule means a departure from the random distribution of points and indicates some kind of cluster structure within the data. In this way, the cluster can be viewed as a sub-space that has no further cluster structure. A clustering process would then have the objective of identifying specific sections of data space that are not likely to be further subdivided into clusters by using the same algorithm. These sections would then represent the sub-spaces where distribution of point distances follows the normal distribution, which is a confirmation of each discovered cluster.

CONCLUSION

In spite of their limitations and shortcomings, statistical methods of processing medical data met the needs of medical researchers before the computer science and industry succeeded to introduce the databases, able to efficiently store and search large amounts of data over long time intervals. With the increasing volume of available data and the development of efficient algorithms for data mining, a need has developed to use these advances to help statistical analysis and further confirm observed medical findings. Results of clinical experiments and research in contemporary medical practice by now largely depend on the support of IT methods for processing incoming data and presentation of findings. Use of statistical tests and determining the relationship between treatment and the findings through the p-value remain the basis of clinical trial practice, but only the detection of anomalies or regularities in the data, even if they are based on relatively small populations, opens the opportunity for more complex types of diagnostics and new discoveries in medicine. In addition to classification and clustering of data, and supporting the decision making process that data mining provides, application of information technology in medicine reduces the possibility of human error, especially when working with large amounts of data and repeating calculations. In this way, the working standards and efficiency are also lifted to a higher level.

REFERENCES

Huang, H., Tsai, W.T., Bhattacharya, S., Chen, X.P., Wang, Y., &Sun, J. (1996). *Business rule extraction from legacy code,* Proceedings of 20th International Conference on Computer Software and Applications, IEEE COMPSAC'96, pp.162-167. doi:10.1109/CMPSAC.1996.544158

Kan, S.H. (2003). *Metrics and Models in Software Quality Engineering*, 2nd ed., Addison-Wesley, Boston, MA, USA. ISBN: 978-0-201-72915-3

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms,* John Wiley & Sons, New York, NY, USA. ISBN: 978-0-470-89045-5

Kumar, V., Steinbach M., &Tan P.N. (2005). *Introduction To Data Mining*, Addison-Wesley, Boston, MA, USA. ISBN: 978-0-321-32136-7

Larose, T.D. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, New York, NY, USA. ISBN: 978-0-471-66657-8