THE APPLICATION OF DATA MINING TECHNIQUES TO ANALYZE THE DATA FOR PUBLIC HEALTH POLICIES

.

Ana (Resulaj)Ktona – Informatics Department, Faculty of Natural Sciences, University of Tirana

Ergeta Ktona - Faculty of Nursery, University of Medical Sciences.

Anila Gjonaj - Informatics Department, Faculty of Natural Sciences, University of Tirana

Abstract

Health is a very important aspect of human life. Public health activities are needed to ensure the health of the population of a country. According to the law in Albania Public health includes all organized efforts of the society for life prolongation, disease prevention and health promotion of the whole population. Some basic public health services are early detection and management of disease outbreaks, policymaking etc. Analysis of public health data is important in the delivery of basic services. A modern and effective way to analyze the data is Data Mining. Data mining aims at discovering valuable decision-making information from the data held in databases or text files. Structural features (models, rules, restrictions) could be identified with Data Mining and in certain conditions it may create predictive models. Features and models are important for many organizations, services, and in formulating the more effective risk management plans. In this paper we study the need for finding predictive models previously unknown in the data collected on public health, Data Mining techniques that can be used on these data and how to use them. We find appropriate data mining techniques to analyze data on public health by making a theoretical comparison of the use of these techniques for data with the same nature as our data. These techniques will be tested in practice on data collected by public health in Albania to explore their effectiveness in terms of contributing to the creation of public health policies and the detection and management of spread of diseases. Referring to "Health in Albania, National Background Report" one of the health research priorities is to be created a Health Information System. Adding these techniques to this system makes possible transform it in an expert system and as a result to have a great contribution in Policy Making.

Keywords: *Data Mining, expert system, Public Health, Policy Making*

Introduction

Data mining is a science that is being used more and more in all areas of life. This science aims at transforming the data held in databases or text files into a valuable decision-making information identifying structural features (for example, models, rules, restrictions) and create, in certain circumstances, predictive models. These features are important for many organizations, services and interactions with clients, as well as in the formulation of effective plans to manage risks.

Public health is a social and political concept aimed at improving health, prolonging life and improving quality of life for the entire population, through health promotion, disease prevention

and other forms of health care interventions. Some basic public health services are early detection and management of disease vectors, etc. making policy analysis of public health data is important to provide basic services. Analysis of public health data is important in the delivery of basic services. A modern and effective way to analyze data is data mining.

This paper aims to provide a survey of current techniques of knowledge discovery in data that are in use today in public health using data mining techniques.

The objectives of this paper are as following:

- Presentation of current uses of data mining in medicine and public health and highlighting the importance of the use of data mining in medicine and public health.
- Finding data mining techniques that are used in different fields that can be applied to public health.

Methodology

In the function of our first goal we studied the scientific research carried out in connection with data mining and public health. Then we studied the data mining methods that can be used in the data collected on public health and the usefulness of their use on these data. We found appropriate data mining techniques to analyze data on public health by making a theoretical comparison of the use of these techniques for data with the same nature as our data. These appropriate Data Mining techniques will be tested practically on the Data Warehouse of Public Health Institute. Adding these techniques to the Health Information System makes possible transform it in an expert system and as a result to have a great contribution in Public Health Policy Making.

Findings and Results

Currently, the population size, the volume of electronic data collected, along with the speed of spread of the disease makes it almost impossible for analyzing medical data with traditional methods. Data mining helps in discovering knowledge on voluminous data. Therefore its use is beneficial in meeting basic public health services.

The usefulness of using Data Mining in public health.

The health sector has a greater need for data mining nowadays. There are several arguments that can be provided to support the use of mining in the health sector, particularly in public health.

Data overload.

Acquiring knowledge by the computerized medical records is necessary. However large volume of data stored in these databases makes it extremely difficult for the people to take advantage of this information and discover knowledge.
Data Mining which includes systematic analysis of voluminous data to extract models using automated methods would assist in gaining knowledge from data. Recognition of patterns is necessary to convert the overloaded data at a manageable flow of useful information.

Prevention of hospital errors.

If medical institutions in Albania would apply data mining in their existing data, they can discover new knowledge useful to save lives. Otherwise this information would remain hidden in their databases.

Policy-making in public health.

Using data mining on public health data in Albania, models can be detected across health centers that lead to policy recommendations by the Institute of Public Health. Adding these models to Health Information Systems makes possible transform it in an expert system and as a result to have a great contribution in Policy Making. Expert systems are part of Artificial Intelligence.

According to Edward Feigenbaum an expert system is "an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solutions" (Feigenbaum 1982).

Health experts have started to look at how to apply data mining for early detection and management of pandemics. (Kellogg et al, 2006) presented techniques combining spatial modeling, simulation data mining to find interesting features of disease outbreaks. The analysis resulting from the use of data mining in a simulated environment can then be used to inform policy-makers to detect and manage disease outbreak.

Data Mining in Health Care

The process of Data mining in medical research like in statistics starts with a hypothesis and then this hypothesis is tested with Data Mining techniques. This process differs from the standard practice of data mining, which simply starts the process with some of the data presented which are called training data, without an obvious hypothesis.

Also, while traditional data mining is concerned about patterns and trends in data sets, data mining in Health Care primarily is more interested in the minority that does not match the patterns and trends. One of the main differences in the use of Data Mining methods in Health Care from the use of these methods in other fields is the fact that most traditional data mining is mainly concerned about the description, but not in explaining the patterns and trends. In contrast, Health Care needs these explanations, because a small change can alter the balance between life or death.

Data confidentiality and ethical use of patient information is a major obstacle for data mining in health care. Data mining techniques generally in order to have as an output more accurate results, need a large amount of real data. Healthcare records are private information. The use of these private data may help to prevent deadly diseases.

Comparison of data mining methods

Data mining techniques can be divided into classification, clustering and association techniques.

Classification Techniques

Classification techniques classify instances exclusively in one of the predefined classes. Classification is one of the most common tasks of data mining, which is also prevalent in human life. Humans usually classify or categorize in order to understand and communicate about the world. For any object or instance, classes are designed based on the value of a particular field. Classification includes examining features of new or invisible cases and assigns it to one of the predefined classes. In fact, in the case of data mining instances to be examined come from a database. Classification task involves updating each record by filling in a field with a class code. This task begins with a group of instances often called training group. These instances belong to predefined classes and are used to train and to build a data mining model so that the model can be applied to classify new or invisible objects. Prediction techniques are similar to classification techniques except that the results pertaining to the future.
Any of the techniques that can be used for classification task can be adapted to be used also in prediction task. This is made by using training examples where the value of the variable being predicted is already known, together with historical data about those examples. Historical data are used to build a model that explains the behaviour being monitored at the moment. When the model is applied on the current input, the result is a prediction of future behaviours.

Clustering techniques

Clustering techniques find groups that are very different from each other but the members of the group are very similar to each other. The only difference between classification and clustering is that unlike the case of classification clustering has no predefined classes. There are no predefined classes and instances in the clustering methods grouped together on the basis of similarity between them.

Techniques to find association rules

These techniques discover rules that associate two or more attribute. Association learning in data mining is a scheme in which each combination is required, not just a combination that predict a unique class (Witten et al, 2011). Association is appropriate if the problem is to extract any structure from some given data. The aim of the association is to examine which cases or instances you have most likely to be grouped together. The analysis of market basket is a typical practical application association. Association rules can be developed in order to determine the positions of the items on the shelves of a shop in a supermarket so that certain items that are often purchased together will be found together. Association, unlike classification, can predict every area, not just classes and it can predict more than one attribute value at a time. For this reason we can find a bigger number of association rules comparing to classification rules.

The most appropriate techniques for our data.

Based on the nature of our data and in the fact that it's necessary the explanation of patterns and trends the most appropriate techniques for this kind of data are prediction techniques. One of the most popular algorithms of predictive techniques is decision tree algorithm. The method of

decision trees has wide usage in data exploration, classification and listing. They can also be used to estimate ongoing values even though they are rarely a first choice because decision trees generate garbled estimation, all the records that go to the same leaf are assigned the same estimation value. They are a good choice when the data mining task is classification of records or prediction of discreet results. Decision trees are used when the purpose is to assign every record to one or more categories. Theoretically decision trees can assign lines to a certain number of classes, but tend to make errors when the number of training examples per class is small. This can occur very quickly in a multi-level and/or many braches per node tree. Decision trees are also a natural choice when the goal is to generate explanative, understanding rules. The ability of decision trees to generate rules that can be translated in understandable, natural language or SQL is one of the strongest points of this technique. Even in complex decision trees is generally easy to follow any path from tree to a leaf. So the explanation for each separate classification or prediction is relatively direct. The decision trees require less preparation of data than many other techniques because they are capable enough to handle the ongoing and categorical variables.

Categorical variables, which pose problem for neural networks and statistical techniques are divided by forming class groupings.

Ongoing variables are divided by forming the interval of their values. Since decision trees do not use current values of numerical variables, they are not sensitive to external and distorted distributed. This versatility is due to non-use of information that is valuable to training data, so a neural network and a well-organized regression model often give a better use of the same area than the decision tree. For this reason decision trees are often used to select a good set of variables which will be used as inputs of another modeling technique. Time oriented data require large preparation of data. The series of time data must be expanded so that sequential and favorite models become apparent. Decision trees show a lot about data where they apply so the authors can use them in early phases of data mining projects even when the final projects are created using other techniques.

Conclusions

Analysis of public health data is important in the delivery of basic services in public health which are of a crucial importance for the population. We have presented on this paper data mining techniques (that are modern and effective ways of analyzing data) comparison. We presented also the most appropriate techniques for our data the prediction techniques based on the nature of our data and in the fact that it's necessary the explanation of patterns and trends. The most known algorithm of these techniques decision tree algorithm will be used on data of the Public Health Institute to gain knowledge about the usefulness of the application of Data Mining Techniques for Public Health in Albania. In this paper is presented also an overview of actual practices and challenges on the application of Data Mining on public health. Healthcare organizations can look at these applications to find ideas on how to gain knowledge from databases of their systems.

Referring to "Health in Albania, National Background Report" (Hajdini 2009) one of the health research priorities is to be created a Health Information System. Adding data mining techniques to this system makes possible transform it in an expert system and as a result to have a great contribution in Policy Making. As a result Institute of Public Health can use this system to find trends in the spread of disease or death (i.e., infant mortality) per region and per hospital. Institute of Public Health can reveal hidden patterns of mortality or diseases that can lead to better health policies such as better planning of vaccination, identification of factors of diseases such as malaria, prevention of hospital errors, sporadic outbreaks of influenza in certain areas that are thought of as the root of these diseases.

References

(C. Bailey-Kellogg, N. Ramakrishnan, and M. Marathe. 2006) Spatial data mining to support pandemic preparedness. ACM SIGKDD Explorations, 8:80-82, 2006

(Feigenbaum 1982), Eduard A. Feigenbaum, "Knowledge Engineering in the 1980's", Stanford University , Stanford, 1982

(Hajdini, G. 2009): Health in Albania. National Background report. Tirana: Ministry of. Health, Medicine Faculty at UT, Institute for Public Health, 2009.

(Ian H. Witten, Eibe Frank, Mark A. Hall 2011) Data Mining Practical Machine Learning Tools and Techniques Third Edition, Morgan Kaufmann series in data management systems. Morgan Kaufman.