An Analysis of Medical Decision Support Systems using Mining Algorithms

B. Suresh
Research Scholar
Department of Computer Science
Adhiparasakthi College of Arts and Science, Kalavai, Thiruvannamalai district Tamilnadu, India

P. Hariharan
Assistant Professor
Department of Computer Science
Adhiparasakthi College of Arts and Science, Kalavai, Thiruvannamalai district Tamilnadu, India

M. Sasikumar
Assistant Professor
Department of Computer Applications
King Nandhivarman College of Arts and Science, Thellar-6040406, Thiruvannamalai District, Tamil Nadu. India

K. Dhakshnamurthy
Head & Assistant Professor
Department of Computer Applications
King Nandhivarman College of Arts and Science, Thellar-6040406, Thiruvannamalai District, Tamil Nadu. India

Abstract

Medical datasets have reached enormous capacities. This data may contain valuable information that awaits extraction. The knowledge may be encapsulated in various patterns and regularities that may be hidden in the data. Such knowledge may prove to be priceless in future medical decision making. The data which is analyzed comes from National Breast Cancer Prevention Program. The aim of this thesis is the evaluation of the analytical data from the Program to see if the domain can be a subject to data mining. The next step is to evaluate several data mining methods with respect to their applicability to the given data. This is to show which of the techniques are particularly usable for the given dataset. Finally, the research aims at extracting some tangible medical knowledge from the set. The research utilizes a data warehouse to store the data. The data is assessed via the ETL process. The performance of the data mining models is measured with the use of the lift charts and confusion (classification) matrices. The medical knowledge is extracted based on the indications of the majority of the models. The data mining models were not unanimous about patterns in the data. Thus the medical knowledge is not certain and requires verification from the medical people. However, most of the models strongly associated patient's age, tissue type, hormonal therapies and disease in family with the malignancy of cancers. The next step of the research is to present the findings to the medical people for verification. In the future the outcomes may constitute a good background for development of a Medical Decision Support.

Keywords: DSS – Decision Support Systems, BP – Back Propagation, SQL –Structured Query Language, Em-Expectation Maximation, OLAM - On-Line Analytical Mining

I. INTRODUCTION

Health care institutions all over the world have been gathering medical data over the years of their operation. The enormous volume of this data is getting to a point where it will exceed available storage capacities. This data may constitute a valuable source of medical information that has potential to be very useful in diagnosing and treatment. Its format, however, may require some additional processing before the data can serve doctors as a precious source of information in their everyday work. This data may comprise thousands of records which may contain valuable patterns and dependencies hidden deep among them. The volume of the dataset and complexity of the medical domain make it very difficult for a human to analyze the data manually to extract hidden information. For this reason the computer science proves to be helpful. Machines are used to mine data for patterns and regularities. Various data mining algorithms have been developed which analyze the data in order to extract underlying knowledge.

This research is focused on breast cancer. The etiologic of the breast cancer, despite broad and thorough research conducted by leading ontological institutions all over the world, is still not well known. The disease can be induced by a variety of factors (carcinogens) such as mutations of genes, age, occurrences of the disease in family, and many others. However, the best results in treatment are achieved when the disease is diagnosed early enough. Otherwise costs are incomparably greater. They encompass costs of treatment (operations, radiology, chemotherapy, etc), complications, sick leaves, pensions, welfare, and so on. Another problem is the social and psychological side-effects of the disease and possibly a death in family. All these reasons confirm the fact that the sooner the disease has been diagnosed it gives not only a greater chance for complete cure but also significantly smaller cost at the same time. The Breast Cancer Prevention Program mentioned above brought a lot of raw data in a form of surveys. Each of them was filled out by both patients and doctors participating in the screening. The former provided general information about the state of their health and information concerning various aspects from their lives (number of births, age at first and last menstruation, etc.). The data contained in these surveys may contain valuable information, encapsulated in various patterns and regularities, which could be used in a diagnosing process in the future. To extract the knowledge the data has to be transformed to an electronic form, then cleansed and finally analyzed.

A. Research Aim and Objectives

The research has three main goals:
- Evaluation of the analytical data gathered from the surveys of the National Breast Cancer Prevention Program
- Evaluation of data mining methods with regard to their applicability to the data from the Program.
- Extraction of medical knowledge from the dataset.

II. DATA MINING METHODS

Data mining, also known as knowledge discovery in databases (KDD) is defined as "the extraction of implicit, previously unknown, and potentially useful information from data". Research fields such as statistics and machine learning contributed greatly to the development of various data mining and knowledge discovery algorithms. The objectives of these algorithms include (among others) pattern recognition, prediction, association and grouping. They can be applied to a variety of learning problems.
They are heterogeneous and contain both discrete and continuous values. There are

A. Several tasks of Learning

- classification – an algorithm is given a set of examples that belong to certain predefined and discrete classes from which it is expected to learn to classify future instances,
- regression (numeric prediction) – an algorithm is expected to predict a numeric class instead of a discrete,
- association – an algorithm seeks for any correlations (dependencies) among features,
- Clustering – an algorithm groups examples that are in some way similar.

Classification and regression are commonly referred to as prediction. They represent a so-called supervised learning. The supervision can be understood by observing the way in which an algorithm learns. It is given a set of instances, each provided with an outcome (class). The method operates as if it was supervised by this training set telling it what class should be assigned to a particular instance.

There are several ways of measuring performance of data mining models. The most common, and the most criticized one is the accuracy. It is computed by applying the model to a dataset and observing the rate of properly classified instances. It gives good results in case of datasets in which the classes are close-to-equally distributed. However, if one class constitutes an overwhelming majority in the dataset the accuracy may always be high if the model favors this class. The actual predictive power of such a model for unseen future instances may be low, though. Alternative measures include lift charts and area under ROC curves (AUC). They have been broadly used in various data mining problems, including medicine.

- Lift chart – a graph which plots a lift factor. The lift is the ratio of concentration of a given class in a sample of data (derived from the entire dataset) to the concentration of the class in the entire dataset. It is calculated according to the equation (3.1) :

ift = class t (sample size)/ population
Where:
Classt is the number of instances belonging to class t,
Sample is the sample size,
Population is the size of the entire population.
In order to be able to plot a lift chart the data mining model has to provide probabilities of predictions (probability that an instance belongs to a certain class). The instances have to be then sorted by these probabilities in the descending order. The creation of the graph is done iteratively, by simply reading a certain number of instances off the list and calculating the lift factor. The graph has on its horizontal axis the size of a subset expressed as a ratio of the size of the entire dataset. The vertical axis shows the number of instances from the given class in the sample. Besides the lift curve the graph also presents a baseline. It depicts the expected lift factor for a sample drawn randomly. The purpose of the lift charts is to determine the best subset of instances which contains the most instances from one class. The Figure 3.1 shows a sample lift chart. The straight diagonal line denotes the baseline and the curve denotes the actual lift plot.

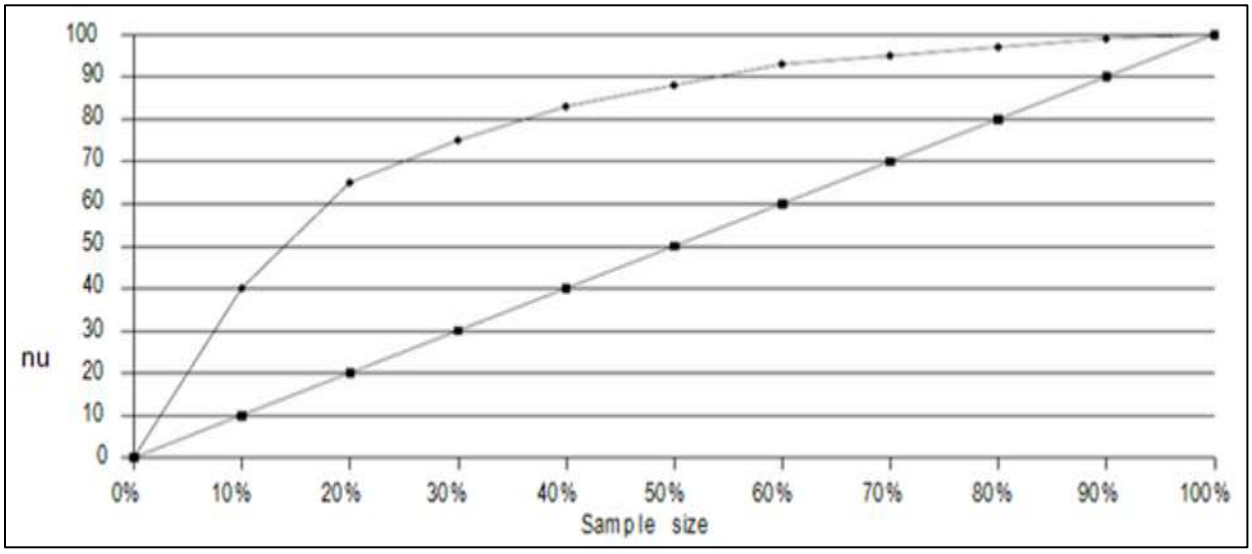

Fig. 3.1: Sample Lift Chart

AUC – a graph closely related to the lift charts. This approach constitutes a good alternative of the accuracy as well. It is also used as a method of comparing the performance of the data mining algorithms. The method utilizes the ROC curves (Receiver Operating Characteristics curves) which originate in the signal detection theory. They plot a number of instances from one class (expressed as a proportion of the total number instances belonging to it on the vertical axis) against a number of instances from the other classes (expressed in the same manner, on the Horizontal axis). In case of multi-class problems each class value has a separate line on the graph. The process of creation of a curve is also iterative, however differs from the one presented for the lift charts. Here, the instances (also sorted by the probabilities), provided with the actual class and the predicted class are analyzed one-by-one in order to draw the line. The bigger the area beneath the plot the better the performance of a model. The Figure 3.2 shows a sample ROC plot for a Boolean problem. The straight line, similarly to the lift charts, denotes the baseline and the curve depicts the ROC plot. In real situations the curve usually is not that smooth, though.

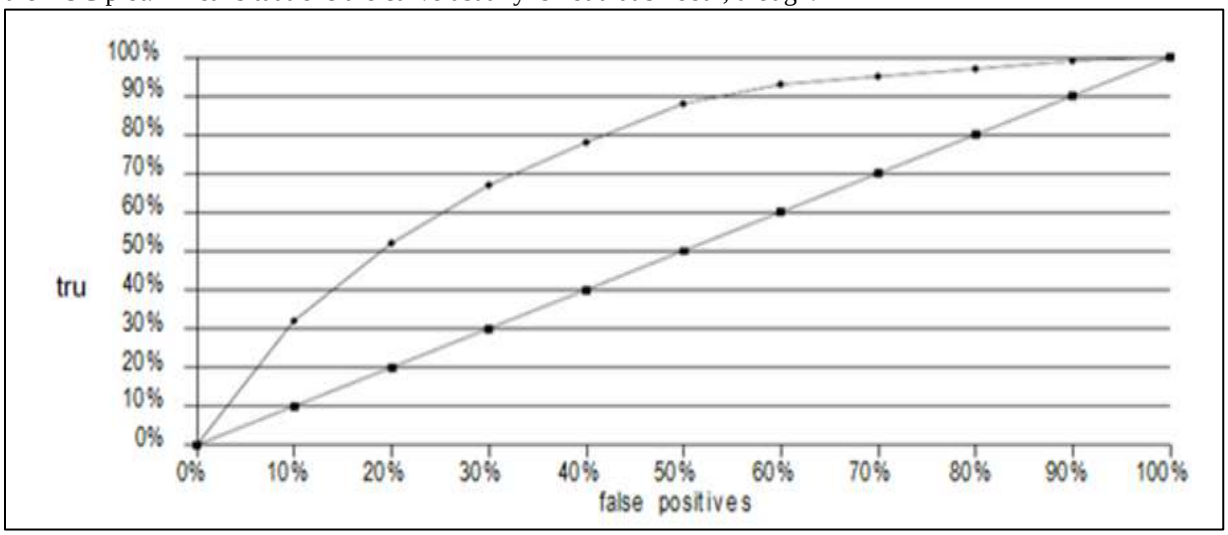

Fig. 3.2: Sample ROC Plot

The research utilizes 6 data mining methods:

- Decision trees,
- Neural networks,
- Logistic regression,
- Clustering,
- Naive Bayes,
- Association rules.

They have been all implemented in the business intelligence part of the Microsoft SQL Server 2005 and this was the main reason for choosing them.

III. DATA ANALYSES

The training dataset, obtained from the surveys of the National Breast Cancer Prevention Program, was used to build several data mining models. They included: decision trees, clusters, neural networks, logistic regressions, association rules and Naive Bayes. However, prior to this the dataset has been pre-analyzed. This was to see how the attributes are represented in terms of their values to determine the initial input set of attributes. This was followed by the analyses.

Initial feature selection

The analytical dataset comprises several attributes. However, some of them do not carry any relevant information from the perspective of the analyses. For instance, the Screening Number in no way can influence the occurrence of the breast cancer. In case of the City one can argue that some cities are more polluted than others and living there may influence the disease. However, in case of this research the patients come from a small area in Poland (one province) and all of the cases are close-to-equally affected in this regard. Thus this attribute will also be skipped. The Table 7.1 shows the attributes both that have been chosen for the analyses and those that have been rejected.

Table - 7.1
Initial feature selection

Attribute	Accepted	Reason for rejection
Family Disease dimension		
Family Line	Yes	
Relative	No	This attribute is dependent on the *Family Line*.
Time dimension		
Month	No	Time span within which the Program was conducted is too short and in no way influences the occurrence of the disease.
Patient dimension		
Age	Yes	
Births	Yes	
City	Yes	
First Menstruation	Yes	
Hormonal Therapies Earlier	Yes	
Hormonal Therapies Now	Yes	
Last Menstruation	Yes	
Mammographies Count	No	Numbers of mammographies taken by a patient does not affect the disease.
Screening Number	No	It is only a statistical attribute introduced for Administrative purposes and carries no medical information.
Self Examination	No	It is uninformative because only one case (out of 1311) bears a different value from all the rest. It may add a lot of noise to the data.
Tissue Type dimension		
TissueTypeName	Yes	
Symptom dimension		
Symptom Description	Yes	
BIRADS dimension		
BIRADS Description	Yes	
BIRADS Value	Yes	

Attribute	Accepted	Reason for rejection
Diagnosis dimension		
Diagnosis Description	Yes	
Screening Result dimension		
Change Count	Yes	
Change Side	Yes	
Change Size	Yes	
Change Location dimension		
Change Location	Yes	

Further feature selection is performed by individual algorithms. This is done automatically and the user does not have any possibility to influence this process.

A. Evaluation of the Algorithms

The last step of the analyses is the comparative evaluation of the generated models. For the evaluation only the "best" models from their groups were taken, i.e. those that yielded the best results in terms of the lift and the classifications. The following models have been chosen:

- The decision tree built with the use of the entropy score method,
- The clusters generated with the use of the EM algorithm,
- The neural network with the 30% split of the dataset
- The logistic regression with the 30% split of the dataset
- The Naive Bayes,
- The association rules.

B. Conclusions from the Evaluation

The very important aspect of the data mining is the quality of the data. The distribution of the values of the attributes is mostly unequal. This in particular concerns the values of the class (the Diagnosis Description). First and foremost, it takes 5 different values. In most of the research described the classes took only two values (Boolean) which indicated whether an instance was benign or malignant. In the case of this research such classification was not possible due to the fact that the surveys did not deliver such information. None of the patients' cases have been ultimately assigned to either benign or malignant group. Thus the only value that could be adopted as a class was the final mammography assessment. The evaluation of the data mining models was also limited by the environment used. The Microsoft did not implement evaluation techniques commonly used in medicine– the area under the ROC curves (AUC). However, the main orientation of the SQL Server 2005 is business and in the implementation the system allows for incorporating costs into the lift charts.

IV. CONCLUSIONS AND FUTURE WORK

The main goal of the research was to first and foremost analyze the data from the surveys of the National Breast Cancer Prevention Program and to judge whether it is suitable to be analyzed with the use of the data mining methods. The second step was to evaluate several data mining algorithm in terms of their applicability to this data. Finally, an attempt was made to deliver some tangible medical knowledge extracted by the methods. The quality of analytical data is very important to building good data mining models. This research was only to point to some directions that potentially may bring some new knowledge, which further developed, may turn out to be helpful. The patterns discovered by the data mining methods employed in the research are not supported by any medical knowledge, though. They have been learned only by finding correlations among attributes. Furthermore, the amount of training data was very small comparing to all of the possible combinations of attributes' values. This also may have distorted the findings. It has not been an intention of the research to prove that data mining algorithms can be blindly used in real life. The next step is the evaluation of the results by the medical staff in health unit. In a more distant perspective the analyses along with the Data Retriever software tool will constitute a basis for development of a medical decision support system.

REFERENCES

[1] Abbass H. A., An Evolutionary Artificial Neural Networks Approach for Breast Cancer Diagnosis, CiteSeer, 2002
[2] Appendix No. 4 of the disposition No. 86/2005 of the President of the Polish National Health Fund, Public Information Bulletin, 2005.

[3] Berry, M. J. A., Linoff G. S., Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Second Edition, Wiley Publishing Inc., USA, 2004.
[4] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., Classification and Regression Trees. The Wadsworth Statistics/Probability Series, 1984, 358.
[5] Chen T., Hsu T., A GA based approach for mining breast cancer pattern, Expert Systems with Applications, 30, Elsevier, 2006, 674-681.
[6] Cheng T., Wei C., Tseng V. S., Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches, Proceedings of the 19th IEEE Symposium on the Computer-Based Medical Systems, IEEE, 2006.
[7] Chou S., Lee T., Shao Y. E., Chen I., Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines, Expert Systems with Applications, 27, Elsevier, 2004, 133-142.
[8] Cios K.J., Moore W.G., Uniqueness of Medical Data Mining, Artificial Intelligence inMedicine, 26 (1-2), Elsevier, 2002, 1-24.
[9] Cooper G. F., Herskovits E., A Bayesian Method for the Induction of Probabilistic Networks from Data, Machine Learning, 9, Kluwer, 1992, 309-347
[10] Creswell J. W., Research Design – Qualitative, Quantitative and Mixed Method Approached, Second edition, Sage Publications, Thousands Oaks CA, 2002.
[11] Dawson C. W., The Essence of Computing Projects – a Student's Guide, Prentice Hall, Harlow UK, 2000.
[12] Delen D., Walker G., Kadam A., Predicting Breast Cancer Survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, 34, Elsevier, 2005, 113- 127.
[13] Fogel D. B., Wasson E. C., Boughton E. M., Evolving neural networks for detecting breast cancer, Cancer letters, 96(1), Elsevier, 1995, 49–53.
[14] Han J., Kamber M., Data Mining: Concepts and Techniques, Second edition, The Morgan Kaufmann Publishers, 2006.
[15] Hanley J. A., McNeil B. J., The Meaning and Use of the Area under a Receiver