Medical Tracker

Harsh Bhotika, Nilesh Nathani and Dhaval Pathak

GUIDE: PROF. RAHUL JINTHURKAR

Watumull Institute of Electronics Engineering and Computer Science

University of Mumbai

Abstract:

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. As large amount of data is generated in medical organizations (hospitals, medical centers) but this data is not properly used. There is a wealth of hidden information present in the datasets. The healthcare environment is still "information rich" but "knowledge poor". There is a lack of effective analysis tools to discover hidden relationships and trends in data. Advanced data mining techniques can help remedy this situation. For this purpose we can use different data mining techniques. This paper describes about a prototype using data mining techniques, namely Naïve Bayes and multivariable risk algorithm. This system can answer complex "what if" queries which traditional decision support systems cannot. Using medical profile such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It is a web based user friendly system and can be used in hospitals if they have a data ware house for their hospital. Presently we are analyzing the performances of the two data mining techniques by using various performance measures.

Keywords: Disease, prediction, Data mining.

1. INTRODUCTION:

A major challenge facing healthcare organizations (hospitals, medical centres) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information system to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" This is the main motivation for this paper.

1.1. Data Mining

Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Fayyad defines data mining as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database". Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. Each data mining technique serves a different purpose

depending on the modeling objective. The two most common modelling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms. Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. Link classification provides category of an object, not just based on its features, but also on connections in which it takes part, and features of objects connected with certain path. Link analysis in medicine is task of predicting disease type based on people's characteristics or predicting age of people based on disease they are infected with and based on age of people they have been in contact with. Link analysis can be used in order to understand where patients go to receive the healthcare treatment and to identify the components or resources in service that must be addressed.

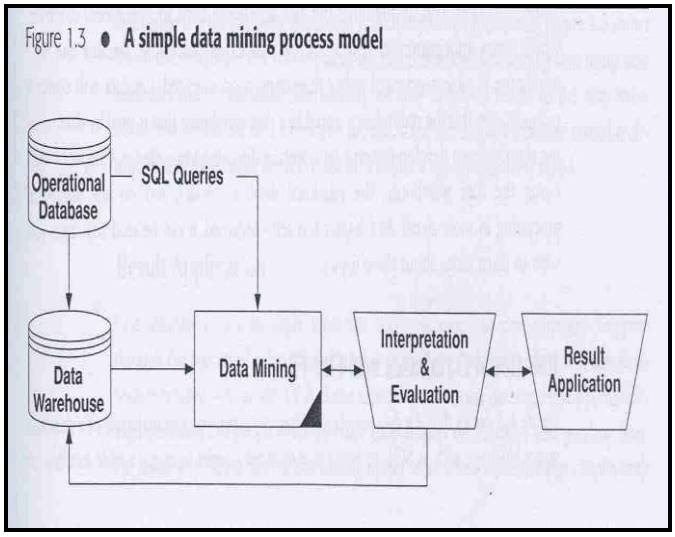

Figure: A simple data mining process model

Knowledge Discovery in Databases is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process.

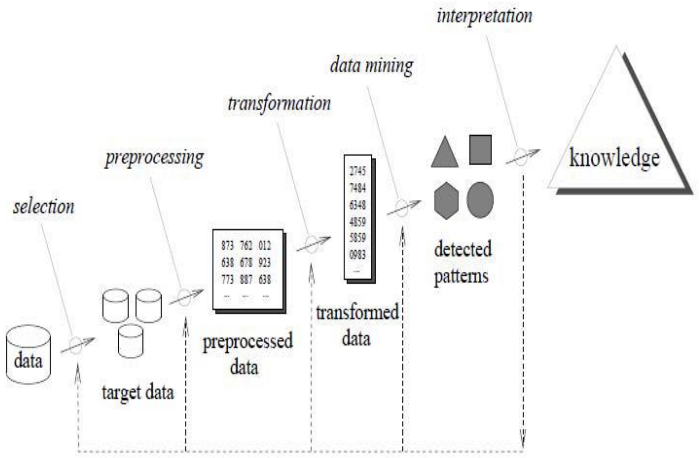

Figure: KDD Process

2. EXISTING SYSTEM:

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", However, they cannot answer complex queries like "Given patient records, predict the probability of patients getting a heart disease."

Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

3. PROPOSED SYTEM:

Clinical databases have accumulated large quantities of information about patients and their medical conditions. The term "Diagnosis" is identified as the predictable attribute with value "1" for patients with disease and value "0" for patients with no disease. "Patient Id" is used as the key, the rest are input

attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

A. Data set Generation:

Questionnaires have advantages over some other types of medical symptoms that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data. However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Here our questionnaire is based on the attribute given in the data set, so the questionnaire contains:

Predictable attribute:

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))

Key attributes:

1. Patient Id – Patient's identification number.

Input attributes:

1. Sex (value 1: Male; value 0: Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0:< 120 mg/dl).
4. Blood Pressure (mm Hg on admission to the hospital).
5. Serum Cholesterol (mg/dl).
6. Thalach – maximum heart rate achieved.
7. Age in Year.
8. Height in cm's.
9. Weight in Kg's.

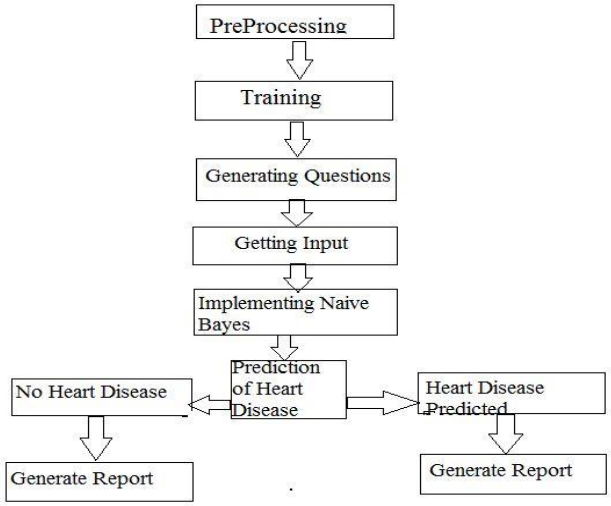

Figure: Block diagram for Heart disease diagnosis

B. Data set Analysis:

The records were split into two datasets such as training dataset and testing dataset. To avoid bias, the records for each set were selected randomly. In artificial intelligence or machine learning, a training set consists of an input vector and an answer vector, and issued together with a supervised learning method to train a knowledge database (e.g. a neural net or a naive bayes Classifier) used by an AI machine. In a dataset a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. After a model has been processed by using the training set, test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute to predict.

4. Algorithms:

Naive Bayes:

1. Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.
2. Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive probability assigns an unknown sample X to the class Ci if and only if:

P (Ci/X)>P (Cj/X) for all 1< = j< = m and j!= i

Thus we maximize P (Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis.

By Bayes theorem,

P (Ci/X)= (P(X/Ci)P(Ci))/P(X) .

3. As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples.

Decision tree algorithm:

Decision trees are one of the most regularly used techniques of data analysis. Decision trees are easy to visualize and understand and resistant to noise in data. Generally, decision trees are used to classify records to a proper class. Besides, they are applicable in both regression and associations tasks. In the medical field decision trees specify the sequence of attributes values and a decision that is based on these attributes. One of the most popularly used decision tree algorithm is Iterative Dichotomized 3 (ID3). Quinlan introduced ID3 algorithm. The algorithm is based on Occam's razor, which means that the smaller trees are preferred. The Occam's razor is formalized using information entropy concept. The construction of a tree is top-down and start with the appropriate attribute for the root node.

5. BENEFITS AND LIMITATIONS:

Health institutions are able to use data mining applications for a variety of areas, such as doctors who use patterns by measuring clinical indicators, decision making based on evidence, identifying high-risk patients and intervene proactively, optimize health care, etc. Integration of data mining in information systems, healthcare institutions reduce subjectivity in decision-making and provide a new useful medical knowledge. Predictive models provide the best Knowledge support and experience to healthcare workers. Data mining is using a technique of predictive modelling to determine which diseases and conditions are the leading trends. This requires a review of medical documentation of a healthcare institution and prescription drugs to determine which problems are the most common amongst patients. Health records are private information, yet the use of these private documents may help in treating deadly diseases.

This system can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a second opinion.

One of the biggest problems in data mining in medicine is that the raw medical data is voluminous, and heterogeneous . These data can be gathered from various sources such as from conversations with patients, laboratory results, review and interpretation of doctors. All these components can have a major impact on diagnosis, prognosis and treatment of the patient, and should not be ignored. The scope and complexity of medical data is one of the barriers to successful data mining. Missing, incorrect, inconsistent or non-standard data such as pieces of information saved in different formats from different data sources create a major obstacle to successful data mining. It is very difficult for people to process gigabytes of records, although working with images is relatively easy, because doctors are being able to recognize patterns, to accept the basic trends in the data, and formulate rational decisions. Stored information becomes less useful if they are not available in easily apprehensible format. The role of visualization techniques is increasing in this, as the picture are easiest for people to understand, and can provide plenty of information in a snapshot of the results. Within the issue of knowledge integrity assessment, two biggest challenges are: (1) How to develop efficient algorithms for comparing content of two knowledge versions (before and after). This challenge demands development of efficient algorithms and data structures for evaluation of knowledge integrity in the data set; and (2) How to develop algorithms for evaluating the influence of particular data modifications on statistical importance of individual patterns that are collected with the help of common classes of data mining algorithm. Algorithms that measure the influence that modifications of data values have on discovered statistical importance of patterns are being developed, although it would be impossible to develop a universal measure for all data mining algorithms.

Another limitation is that it only uses categorical data. For some diagnosis, the use of continuous data may be necessary. Another limitation is that it only uses two data mining techniques. Additional data mining techniques can be incorporated to provide better diagnosis. The size of the dataset used in this research is still quite small. A large dataset would definitely give better results. It is also necessary to test the system extensively with input from doctors, especially cardiologists, before it can be deployed in hospitals.

6. CONCLUSION:

Health care related data are huge in nature and they arrive from various birthplaces all of them not wholly suitable in structure or quality. Data mining brings a set of tools and techniques that can be applied to the large amount of data in healthcare industry to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making

decisions. The need is for algorithms with very high accuracy as medical diagnosis is a significant task that needs to be carried out precisely and efficiently. A prototype heart disease prediction system is developed using data mining classification modeling techniques such as Decision tree, Naïve bayes, Multivariable risk analysis. The system extracts hidden knowledge from a historical heart disease database. All three models are able to extract patterns in response to the predictable state. The most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Multivariable risk analysis and Decision Trees. Decision Trees results are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees. Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors. This system can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 15 that have been taken. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data.
In future the work can be expanded and enhanced for the automation of various types of disease prediction. It also extended to find other types of diseases with the uses of these attributes. With the future development of information communication technologies, data mining will achieve its full potential in the discovery of knowledge hidden in the medical data.

7. REFERENCES:

[1] Gordon T, Kannel WB. Multiple risk functions for predicting coronary
heart disease: the concept, accuracy, and application. *Am Heart J*. 1982;
103:1031–1039.

[2] Fayyad, U: *"Data Mining and Knowledge Discovery in Databases: Implications for scientific databases"*, Proc. of the 9th Int Conf on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[3] Mehmed, K.: *"Data mining: Concepts, Models, Methods and Algorithms"*, New Jersey: John Wiley, 2003.

[4] Kaur, H., Wasan, S. K.: *"Empirical Study on Applications of Data Mining Techniques in Healthcare"*, Journal of Computer Science 2(2), 194-200, 2006.

[5] Hosseinkhah, F., Ashktorab, H., Veen, R., & Ojaboni, M. O. (2009). Challenges in Data Mining on Medical Databases.

[6] Koh, H. C., & Tan, G. (2005). Data Mining Applications in Healthcare. Journal of Healthcare Information Management -Vol. 19, No. 2 , 64-72.

[7] Eapen, A. G. (2004). Application of Data mining in Medical Applications. Ontario, Canada, 2004: University of Waterloo.

[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 No. 6 June 2011.

[9] G. Parthiban, A. Rajesh, S.K.Srivatsa "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method".

[10] Ruben D. Canlas Jr.,"DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES", August 2009.

[11] Harleen Kaur , Siri Krishan Wasan and Vasudha Bhatnagar,
THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS, Data Science Journal, Volume 5, 19 October 2006 pp119-126.

[12] Awang, R. & Palaniappan, S., "Intelligent heart disease predication system using data mining technique". IJCSNS International Journal of Computer Science and Network Security. Vol. 8, No. 8, 2008.