Heart Disease Prediction System Using Bayes Theorem

Sahana Devanathan[1] , Ambika R[2]
[1]Assistant Professor, [2]Associate Professor, BMS Institute of Technology, Bangalore-560064
sahanadev84@gmail.com, ambika2810@gmail.com

Abstract: *A major challenge facing healthcare organizations (hospitals, medical centers) 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 systems to manage their healthcare or patient data. These 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.

Clinical decisions are often made based on doctors' 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.

The main objective of this project is to develop an Intelligent Heart Disease Prediction System using the data mining modeling technique, namely, Naïve Bayes. It is implemented as web based questionnaire application .Based on the user answers, it can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to services at affordable costs. Quality service implies diagnosing patients correctly and administering reduce treatment costs.

Index terms: Attribute, Bayes Theorem, Cluster analysis**,** Data mining, Decision trees, Diagnosis, Naïve Bayes, Genetic algorithm, Neural networks,

1. Introduction

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. Data mining and statistics both strive towards discovering patterns and structures in data. Statistics deals with heterogeneous numbers only, whereas data mining deals with heterogeneous fields. We identify a few areas of healthcare where these techniques can be applied to healthcare databases for knowledge discovery. Here briefly examine the impact of data mining techniques, including artificial neural networks, on medical diagnosis.

2. Motivation

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality 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 systems 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 research.

3. Problem statement

Many hospital information systems are designed to 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?"," How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like "Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone,

radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease."

Clinical decisions are often made based on doctors' 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.

4. Data mining techniques

The most commonly used techniques are:

4.1 Neural networks

Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an expert in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Neural Networks use a set of processing elements analogous to neurons in the brain. These processing elements are interconnected in a network that can identify patterns in data once it is exposed to the data, i.e, the network learns from experience just as people. The bottom layer represents the input layer, in this case 4 input labels X1 through X4.In the middle is something called as the hidden layer with a variable number of nodes. The output layer in this case Z1 and Z2 representing output values we are trying to the inputs which means that what is learned in a hidden node is based on all inputs taken together.

Neural Networks consists of three layers: input, hidden and output units (variables). Connection between input units and hidden and output units are based on relevance of the assigned value (weight) of that particular input unit. The higher the weight the more important it is. Neural Network algorithms use:

4.2 Decision trees

Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labeled with attribute names, the edges are labeled with possible values for this attribute and the leaves labeled with different classes. Tree shaped structures represent set of decisions. These decisions generate rules for the classification of a data set. Decision trees produce rules that are mutually exclusive and collectively exhaustive with respect to the training data base. Specific decision tree methods include classification and regression trees (CART) and chi square automatic interaction and detection (CHAID)

Decision Tree algorithms include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node. CART uses Gina index to measure the impurity of a partition or set of training tuples. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data.

4.3 Cluster analysis

Cluster analysis is a tool for exploring the structure of data. The core of cluster analysis is clustering which is the process of grouping objects into clusters such that the objects from the same cluster are similar and objects from different clusters are dissimilar. Objects can be described in terms of measurements or by relationship with other objects.

Clustering is sometimes used to mean segmentation. Clustering and segmentation basically partition the database so that each partition are grouped is similar according to some criteria are metric.

4.4 Rule induction

Rule induction is the process of extracting useful if-then rules from data based on statistical significant. Rule induction on a database can be a massive undertaking in which all possible patterns are systematically tooled out of the data and then accuracy and significance calculated, telling users how strong the pattern is and how likely it is to occur again.

4.5 Data visualization

Data visualization makes it possible for the analyst to gain a deeper, more intuitive understanding of the data and as such can work well along side data mining on its own data visualization can be overwhelmed by the

volume of data in a database but in conjunction with data mining can help with exploration

5. Medical database

A total (factors) were obtained from the Cleveland Heart Disease database.Figure (1) lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records).To avoid bias; the records for each set were selected randomly. For the sake of consistency, only categorical attributes were used for all the three models. All the non-categorical medical attributes were transformed to categorical data.

The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value of 909 records with 15 medical attributes "0" for patients with no heart disease. The attribute "Patented" was 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 prototype heart disease prediction system is developed using three data mining classification modeling techniques. The system extracts hidden knowledge from a historical heart disease database.

DMX query language and functions are used to build and access the models. The models are trained and validated against a test dataset. Lift Chart and Classification Matrix methods are used to evaluate the effectiveness of the models. 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 Baye's followed by Neural Network and Decision Trees.

Five mining goals are defined based on business intelligence and data exploration. The goals are evaluated against the trained models. All three models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. Naïve Bayes could answer four out of the five goals; Decision Trees, three; and Neural Network, two. Although not the most effective model, 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. The relationship between attributes produced by Neural Network is more difficult to understand.

According to a landmark National Academies Study , over 98,000 Americans died and more than one million patients were injured due to process and systems failures in health care. The Picker Institute survey [2] pointed out that 75% of patients considered the health care system to be fragmented, convoluted, plagued by duplication of effort, poor communication and conflicting advice. These inefficiencies have resulted in double digit inflation of cost of health care worldwide. However, there is an emerging vision for the future of healthcare that can bring dramatic change in the quality and cost of health care.

This new vision is characterized by:

- engaged patients with access to a large volume of health-related information online who
- actively contribute to the health decisions made,
- providers who serve patients as coach-consultant,
- personalized medicine guided by genomics and
- Agile, evidence-based care with automated, patient-specific alerts.

This vision is not only a matter of correct policy or business decisions. The introduction of new technologies, such as ubiquitous, wireless telecommunication, web portals as secure bidirectional conduits for communication and documentation of care delivery and advanced clinical decision support systems with automated event monitors, are all required enablers. These technologies need to be tried and deployed requiring innovation and serious investment.

According to Masys, there are several forces at work that push the healthcare industry towards radical change including:

- the dramatic increase in the amount of information required for making health decisions,
- the rapidly growing use of Internet worldwide,
- genome research that opens up opportunity to provide personalized healthcare,
- medical errors caused by failures in information management.

Predictable attribute

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

Key attribute

1. PatientID – 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. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Fig 1: list of attributes

6. Analyzing the Data Set

A **data set** (or **dataset**) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values will normally all be of the same kind. However, there may also be "missing values", which need to be indicated in some way.

A total of 500 records with 15 medical attributes (factors) were obtained from the Heart Disease database lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). To avoid bias, the records for each set were selected randomly.

The attribute “Diagnosis” was identified as the predictable attribute with value “1” for patients with heart disease and value “0” for patients with no heart disease. The attribute “PatientID” was 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.

7. Naïve Bayes Implementation in Data Mining

Baye’s Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

Bayes' Theorem:

Prob (B given A) = Prob(A and B) / Prob (A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone.

8. Designing the Questionnaire

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.

9. Heart Disease in Web

In our Heart disease development the modeling and the standardized notations allow to express complex ideas in a precise way, facilitating the communication among the project participants that generally have different technical and cultural knowledge.

MVC architecture has had wide acceptance for corporation software development. It plans to divide the system in three different layers that are in charge of interface control logic and data access, this facilitates the maintenance and evolution of systems according to the independence of the present classes in each layer. With the purpose of illustrating a successful application built under MVC, in this work we introduce different phases of analysis, design and implementation of a database and web application.

Here in our project we get a data set from .dat file as our file reader program will get the data from them for the input of Naïve Bayes based mining process.

REFERENCES

1. Blake, C.L., Mertz, C.J.: **"UCI Machine Learning Databases"**,2004.

2. Chapman, P., Clinton, J., Kerber, R. Khabeza, T.,Reinartz, T., Shearer, C., Wirth, R.: **"CRISP-DM 1.0: Step by step data mining guide"**, SPSS, 1-78, 2000.

3. Charly, K.: **"Data Mining for the Enterprise"**, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7,295-304, 1998.

4. Chua Sook Ling @ Landa Chua : "**Model-based Healthcare Decision Support using OLAP with Data Mining",** IEEE 6,543-890,2007

5. Fayyad, U: **"Data Mining and Knowledge**" November 22, 2003

6. Grossi , E., **"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.

7. Grossi , E., "**How artificial intelligence tools can be used to assess individual patient risk in cardiovascular disease: problems with the current methods",** Medical Department, Bracco SpA Milan, Italy and 2Centro Diagnostico Italiano, Milan, Italy,2008.

8. Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.

9. Han, J., Kamber, M.: **"Data Mining Concepts and Techniques"**, Morgan Kaufmann Publishers, 2006.

10. Harleen Kaur, Siri Krishan Wasan**,"Empirical Study on Applications of Data Mining Techniques in Healthcare",** Department of Mathematics, Jamia Millia Islamia, New Delhi-110 025, India.

11. Ho, T. J.: **"Data Mining and Data Warehousing",** Prentice Hall, 2005.

12. Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.

13. Mohd, H., Mohamed, S. H. S.: **"Acceptance Model of Electronic Medical Record"**, Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.

14. Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/en-us/virtuallabs/aa740409.aspx, 2007.

15. Obenshain, M.K: **"Application of Data Mining Techniques to Healthcare Data"**, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

16.Sellappan Palaniappan, Rafiah Awang,"**Intelligent Heart Disease Prediction System Using Data Mining Techniques",** Department of Information Technology, Malaysia University of Science and Technology, IEEE Publications,2008.

17. Tang, Z. H., MacLennan, J**.: "Data Mining with SQL Server 2005",** Indianapolis: Wiley, 2005.

18. Thuraisingham, B**.: "A Primer for Understanding and Applying Data Mining"**, IT Professional, 28-31, 2000.

19. Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: **"Tapping the Power of Text Mining"**, Communication of the ACM. 49(9), 77-82, 2006.

20. Wu, R., Peters, W., Morgan, M.W.: **"The Next Generation Clinical Decision Support: Linking Evidence to Best Practice"**, Journal Healthcare Information Management. 16(4), 50-55, 2002.

.