Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks

K.Srinivas B.Kavihta Rani
Associate Professor, Dept. of CSE
Jyothishmathi Institute of Tech & Science
Karimnagar

Dr. A.Govrdhan
Principal and Professor of CSE
JNTUHCE, Nachupally
Jagtial, Karimnagar

Abstract — **The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) are used. This is an extension of naive Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. Discovery of hidden patterns and relationships often goes unexploited. Using medical profiles 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**.

Keywords: Naive Bayes, ODANB, NCC2

I. INTRODUCTION

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows: "Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data" [1]. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient's future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management.

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. Health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care.

The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Likewise, physicians can also confirm their findings with the conformity of other physicians dealing with an identical case from all over the world [2]. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial.

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 [3]. 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.

A. Knowledge discovery in medical databases

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry [4]. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven.

B. Heart Disease

The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease kills one person every 34 seconds in the United States [7]. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death [6]. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart disease (CHD).A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack. Chest pains arise when the blood received by the heart muscles is inadequate [5] and inductive logic programming.

Many healthcare organizations struggle with the utilization of data collected through an organization online transaction processing (OLTP) system that is not integrated for decision making and pattern analysis. For successful healthcare organization it is important to empower the management and staff with data warehousing based on critical thinking and knowledge management tools for strategic decision making. Data warehousing can be supported by decision support tools such as data mart, OLAP and data mining tools. A data mart is a subset of data warehouse. It focuses on selected subjects. Online analytical processing (OLAP) solution provides a multi-dimensional view of the data found in relational databases. With stored data in two-dimensional format OLAP makes it possible to analyze potentially large amount of data with very fast response times and provides the ability for users to go through the data and drill down or roll up through various dimensions as defined by the data structure.

II. DATAMINING TECHNIQUES IN HEALTH CARE

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. We now describe a few Classification data mining techniques with illustrations of their applications to healthcare.

A. Rule set classifiers

Complex decision trees can be difficult to understand, for instance because information about one class is usually distributed throughout the tree. C4.5 introduced an alternative formalism consisting of a list of rules of the form "if A and B and C and ... then class X", where rules for each class are grouped together. A case is classified by finding the first rule whose conditions are satisfied by the case; if no rule is satisfied, the case is assigned to a default class

1) IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) ---->
(Cause—of--- disease)

Example 1: <u>If_then_rule induced in the diagnosis of level of alcohol in blood</u>
IF Sex = MALE AND Unit = 8.9 AND Meal = FULL THEN
<u>Diagnosis=Blood_alcohol_content_HIGH.</u>

B. Decision Tree algorithms

Decision tree 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 [7]. CART uses Gini index to measure the impurity of a partition or set of training tuples [6]. 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. The decision tree shown in Figure 1 is built from the very small training set. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

1) Transformation parameters

In order to set the transformation parameters we must discuss attributes corresponding to heart vessels. The *LAD*, *RCA*, *LCX* and *LM* numbers represent the percentage of vessel narrowing (or blockage) compared to a healthy artery. Attributes *LAD*, *LCX* and *RCA* were partitioned by cutoff points at 50 and 70%. In the cardiology field, a 70% value or higher indicates significant coronary disease and a 50% value indicates borderline disease. A value lower than 50% means the patient is healthy. The most common cutoff value used by the cardiology community to distinguish healthy from sick patients is 50%. The *LM* artery is treated different because it poses higher risk than the other three arteries. Attribute *LM* was partitioned at 30 and 50%. The reason behind these numbers is both the *LAD* and the *LCX* arteries branch from the *LM* artery and then a defect in *LM* is more likely to cause a larger diseased heart region. That is, narrowing (blockage) in the *LM* artery is likely to produce more disease than blockages on the other arteries. That is why its cutoff values are set 20% lower than the other vessels. The nine heart regions (*AL*, *IL*, *IS*, *AS*, *SI*, *SA*, *LI*, *LA*, *AP*) were partitioned into two ranges at a cutoff point of 0.2, meaning a perfusion measurement greater or equal than 0.2 indicated a severe defect. *CHOL* was partitioned with cutoff points 200 (warning) and 250 (high). These values correspond to known medical settings.

Predicting healthy arteries:

IF (*SA* <= 0.37 *AP* <= 0.66 *Age* <= 78)
THEN *not*(*LAD* >= 50) ls=76% cf=0.58
IF (*SA* > 0.37 *Age* <= 53 *AS* > 0.67)
THEN *not*(*LAD* >= 50) ls=0.3% cf=1.00

Predicting diseased arteries:

IF (*SA* <= 0.37 *AP* > 0.66)
THEN *LAD* >= 50 ls=10% cf=0.80
IF (*SA* <= 0.37 *AP* <= 0.66 *Age* > 78)
THEN *LAD* >= 50 ls=4% cf=0.74
IF (*SA* > 0.37 *Age* <= 53 *AS* <= 0.67)
THEN *LAD* >= 50 ls=1% cf=0.85
IF (*SA* > 0.37 *Age* > 53)
THEN *LAD* >= 50 ls=8% cf=0.98

Figure 1. Decision tree rules with numeric dimensions and automatic splits.

C. Neural Network Architecture

The architecture of the neural network used in this study is the multilayered feed-forward network architecture with 20 input nodes, 10 hidden nodes, and 10 output nodes. The number of input nodes is determined by the finalized data; the number of hidden nodes is determined through trial and error and the number of output nodes is represented as a range showing the disease classification. The most widely used neural-network learning method is the BP algorithm [8]. Learning in a neural network involves modifying the weights and biases of the network in order to minimize a cost function. The cost function always includes an error term a measure of how close the network's predictions are to the class labels for the examples in the training set. Additionally, it may include a complexity term that reacts to a prior distribution over the values that the parameters can take. Neural networks have been proposed as useful tools in decision making in a variety of medical applications. Neural networks will never replace human experts but they can help in screening and can be used by experts to double-check their diagnosis. In general, results of disease classification or prediction task are true only with a certain probability.

D. Neuro-Fuzzy

Stochastic back propagation algorithm is used for the construction of fuzzy based neural network. The steps involved in the algorithm are as follows: First, initialize weights of the connections with random values. Second for each unit compute net input value, output value and error rate. Third, to handle uncertainty for each node, certainty measure (c) for each node is calculated. Based on the certainty measure the decision is made. The level of the certainty is computed using the following conditions.

a. If 0.8 \< c ≤ 1, then there exists **very high certainty**

b. If 0.6 \< c ≤ 0.8, then there exists **high certainty**

c. If 0.4 \< c ≤ 0.6, then there exists **average certainty**

d. If 0.1 \< c ≤ 0.4, then there exists **less certainty**

e. If c ≤ 0.1, then there exists **very less certainty**

The network constructed consists of 3 layers namely an input layer, a hidden layer and an output layer. Sample trained neural network consisting of 9 input nodes, 3 hidden nodes and 1 output node is shown in Figure 2. When a thrombus or blood clot occupies more than 75% of surface area of the lumen of an artery then the expected result may be a prediction of cell death or heart disease according to medical guidelines i.e. R is generated with reference to the given set of input data.

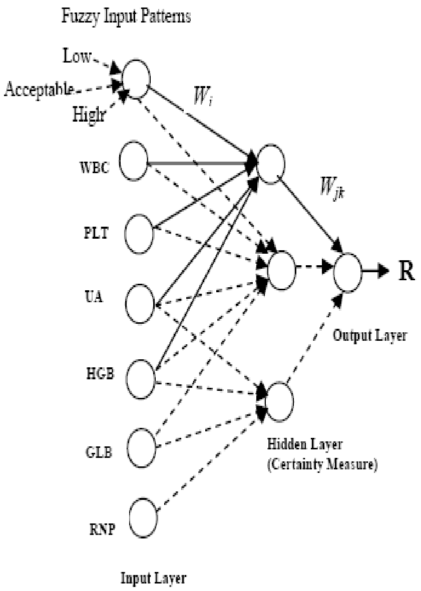

Figure 2. Trained Neural Network for Thrombosis

E. Bayesian Network Structure Discoveries

A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*. This probability is denoted as P*(C | E)* where

$$P\ (C \mid E) = \frac{P(E \mid C)\ .\ P(C)}{P\ (E)} \qquad (1)$$

Bayes' theorem is the method of finding the converse probability of the conditional,

$$P\ (E \mid C) = \frac{P(C \mid E)\ .\ P(E)}{P(C)} \quad = \frac{P(C,E)}{P(C)} \qquad (2)$$

This conditional relationship allows an investigator to gain probability information about either *C* or *E* with the known outcome of the other. Now consider a complex problem with *n* binary variables, where the relationships among them are not clear for predicting a single class output variable (e.g., node 1 in Figure 3). If all variables were related using a single joint distribution, the equivalent of all nodes being first level parents, the number of possible combinations of variables would be equal to (2*n*-1).. This results in the need for a very large amount of data [9, 10]. If dependence relationships between these variables could be determined resulting in independent variables being removed, fewer nodes would be adjacent to the node of interest. This parent-node removal leads to a significant reduction in the number of variable combinations, thereby reducing the amount of needed data. Furthermore, variables that are directly conditional, not to the node of interest but to the parents of the node of interest (as nodes 4 and 5 are with respect to node 1 in Figure 3), can be related, which allows for a more robust system when dealing with missing data points. This property of requiring less information based on pre-existing understanding of the system's variable dependencies is a major benefit of Bayesian Networks [10]. Some further theoretical underpinnings of the Bayesian approach for classification have been addressed in [11] and [12]. A Bayesian Network (BN) is a relatively new tool that identifies probabilistic correlations in order to make predictions or assessments of class membership.

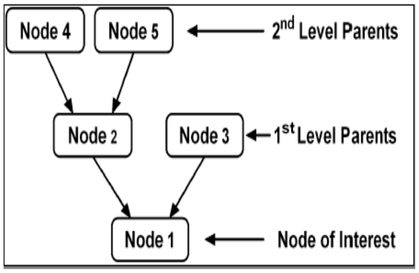

Figure 3. Basic Bayesian Network Structure and Terminology

While the independence assumption may seem as a simplifying one and would therefore lead to less accurate classification, this has not been true in many applications. For instance, several datasets are classified in [13] using the naïve Bayesian classifier, decision tree induction, instance-based learning, and rule induction. These methods are compared showing the naïve classifier as the overall best method. To use a Bayesian Network as a classifier, first, one must assume that data correlation is equivalent to statistical dependence.

1) Bayesian Network Type

The kind of Bayesian Network (BN) retrieved by the algorithm is also called Augmented Naïve BN, characterized mainly by the points bellow

- All attributes have certain influence on the class
- The conditional dependency assumption is relaxed (certain attributes have been added a parent)

2) Pre-Processing Techniques

The following data pre-processing techniques applied to the data before running the ODANB [18] algorithm.

- *ReplaceMissingValues*: This filter will scan all (or selected) nominal and numerical attributes and replace missing values with the modes and mean.

- *Discretization*: This filter is designed to convert numerical attributes into nominal ones; however the unsupervised version does not take class information into account when grouping instances together. There is always a risk that distinctions between the different instances in relation to the class can be wiped out when using such a filter.

F. Some Implementation Details

JNCC2 loads data from ARFF files this is a plain text format, originally developed for WEKA (Witten and Frank, 2005). A large number of ARFF data sets, including the data sets from the UCI repository, are available from http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html.2636. As a pre-processing step, JNCC2 [19] discretizes all the numerical features, using the supervised discretization algorithm of Fayyad and Irani (1993). The discretization intervals are computed on the training set, and then applied unchanged on the test set. NCC2 [19] is implemented exploiting the computationally efficient procedure.

Algorithm 1 Pseudo code for validation via testing file.

```
validateTestFile()

/*loads training and test file; reads list of non-Mar features;
discretizes features*/

parseArffFile();
parseArffTestingFile();
parseNonMar();
discretizeNumFeatures();
/*learns and validates NBC*/
nbc = new NaiveBayes(trainingSet);
nbc.classifyInstances(testSet);
/*learns and validates NCC2; the list of non-Mar features in
training and testing is required*/
ncc2 = new NaiveCredalClassifier2(trainingSet,
nonMarTraining, nonMarTesting);
ncc2.classifyInstances(testingSet);
/*writes output files*/
writePerfIndicators();
writePredictions();
```

JNCC2 can perform three kinds of experiments: training and testing, cross-validation, and classification of instances of the test set whose class is unknown. The pseudo code of the experiment with training and testing is described by Algorithm 1.

The ODANB has been compared with other existing methods that improves the Naïve Bayes and with the Naïve Bayes itself. The results of the comparison prove that the ODANB outperforms the other methods for the disease prediction not related to heart attack.

The comparison criteria that have been introduced are

• Accuracy of prediction (measures defined from the confusion matrix outputs). The table-I below recaps the benchmarked algorithms accuracy for each dataset consider. In each row in bold the best performing algorithm:

TABLE 1 COMPARISON OF RESULTS

DATASETS	**ODANB**	**NB**
Heart-c	80.46	**84.14**
Heart-h	79.66	**84.05**
Heart-statlog	80.00	**83.70**

We focus on the results which clearly states that TAN(Tree Augmented Naïve Bayes) [19] works efficiently for the comparison of data sets of general and regular things like vehicles, anneal(metallurgy) over ODANB, Naïve Bayes. But for prediction of heart disease Naïve Bayes observes better results.

III. CONCLUSIONS

We studied the problem of constraining and summarizing different algorithms of data mining. We focused on using different algorithms for predicting combinations of several target attributes. In this paper, we have presented an intelligent and effective heart attack prediction methods using data mining. Firstly, we have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weightage, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Five mining goals are defined based on business intelligence and data exploration. The goals are to be evaluated against the trained models. All these models could answer complex queries in predicting heart attack

In our future work, this can further enhanced and expanded. For predicting heart attack significantly 15 attributes are listed. Besides the 15 listed in medical literature we 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. We can also use Text Mining to mine the vast amount of unstructured data available in healthcare databases.

REFERENCES

1. Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases:An Overview. The AAAI/MIT Press, Menlo Park, C.A.
2. Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464.
3. Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999.
4. Glymour, C., D. Madigan, D. Pregidon and P.Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.
5. Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999.
6. "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.
7. "Heart disease" from http://wikipedia.org
8. Rumelhart, D.E., McClelland, J.L., and the PDF Research Group (1986), *Parallel Distributed Processing,* MA: MIT Press, Cambridge. 1994.
9. Heckerman, D., *A Tutorial on Learning With Bayesian Networks*. 1995, Microsoft Research.
10. Neapolitan, R., *Learning Bayesian Networks*. 2004, London: Pearson Printice Hall.
11. Krishnapuram, B., et al., *A Bayesian approach to joint feature selection and classifier design.*Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. **6**(9): p. 1105-1111.
12. Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.
13. Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.
14. Pedro Domingos , Michael Pazzani , On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, 29, 103–130 (1997) c° 1997 Kluwer Academic Publishers. Manufactured in The Netherlands.
15. Richard N. Fogoros, M.D, The 9 Factors that Predict Heart Attack 90% of heart attacks are determined by these modifiable risk factors, About.com Guide.
16. Harleen Kaur and Siri Krishan Wasan, Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, 2006 ISSN 1549-3636 © 2006 Science Publications.
17. Bressan, M. and J. Vitria, *On the selection and classification of independent features.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003. 25(10): p. 1312-1317.
18. Domingos, P. and M. Pazzani, *On the optimality of the simple Bayesian classifier under zeroone loss.* Machine Learning, 1997. 29(2-3): p. 103-30.
19. Juan Bernabé Moreno, One Dependence Augmented Naive Bayes, University of Granada, Department of Computer Science and Artificial Intelligence.
20. Giorgio Corani, Marco Zaffalon, JNCC2: The Java Implementation Of Naive Credal Classifier 2, Journal of Machine Learning Research 9 (2008) 2695-2698