

Estimation of CVD with reference to the People of North East India using Predictive Data Mining

Dulal Ch. Das, Prof. P.H. Talukdar

Abstract: Now-a-days, many diseases are reducing the life time of the human. One of the major diseases is cardiovascular disease (CVD). It has become very common perhaps due to increasing busy lifestyles. The issue of health care assumes prime importance for the society and is a significant indicator of social development. Health is therefore best understood as the indispensable basis for defining a person's sense of well-being. Data mining is the computer based process of analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict future trends, allowing business to make proactive, knowledge-driven decisions. The delivery of health care services thus assumes greater proportion, and in this context the role played by information and communication technology has certainly a greater contribution for its effective delivery mechanism. Data mining tools can answer business questions that traditionally taken much time consuming to resolve. The huge amounts of data generated for prediction of CVD are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. By using data mining techniques it takes less time for the predict on of the disease with more accuracy. In this paper we carry different experiments in which one or more algorithms of data mining used for the prediction of CVD. Result from using neural networks is nearly 100%. So that the prediction by using data mining algorithm given efficient results. Applying data mining techniques to CVD treatment data can provide as reliable performance as that achieved in diagnosing CVD.

Keywords: Cardiovascular disease (CVD), Data mining, Genetic algorithm, EM based cluster, Classification, CVD.

I. INTRODUCTION

The idea of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific aspects. In India healthcare is delivered through both the public sector and private sector. The public healthcare system consists of healthcare facilities run by central and state government which provide services free of cost or at a subsidized rates to low income group in rural and urban areas. The main objective of our paper is to study the different techniques of data mining used in prediction of CVD disease by using different data mining tools. Life is dependent on efficient working of HEART because HEART is essential part of our body. If operation of HEART is not proper, it will involve the other body parts of human such as brain, kidney etc. CVD is a disease that affects on the operation of HEART. There are number of factors which increases risk of HEART disease. . In this work, a detailed survey is carried out on data mining applications in the healthcare sector, types of data used and details of the information extracted. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. There are a large number of data mining applications found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management. To find the useful and hidden knowledge from the database is the purpose behind the application of data mining. Popularly data mining called knowledge discovery from the data. The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data

set, preprocessing, data transformation. Some reasons as follows

Smoking: - smoking is a major cause of CVD attack, stroke and other peripheral arterial disease. Nearly 40% of all people who die of smoking tobacco. A smoker's risk of CVD attack reduces rapidly after only one year of not smoking.

Cholesterol: - Abnormal levels of lipids (fats) in the blood are risk factor of CVD diseases. Cholesterol is a soft, waxy substance found among the lipids in the bloodstream and in all the body's cells. High level of triglyceride (most common type of fat in body) combined with high levels of LDL (low density lipoprotein) cholesterol speed up atherosclerosis increasing the risk of CVDs.

High blood pressure: - High blood pressure also known as HBP or hypertension is a widely misunderstood medical condition. High blood pressure increase the risk of the walls of our blood vessels walls becoming overstretched and injured. Also increase the risk of having CVD attack or stroke and of developing CVD failure, kidney failure and peripheral vascular disease.

Obesity:-the term obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Being obese puts anybody at a higher risk for health problem such as CVD disease, stroke, high blood pressure, diabetes and more.

Lack of physical exercise: -lack of exercise is a risk factor for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD, because it also increases the risk for diabetes and high blood pressure.

II. RELATED WORKS

Many experiments are being carried out for evaluating the performance of Naïve Bayes and Decision Tree algorithm. The results observed so far indicate that Naïve Bayes outperforms and sometimes Decision Tree. In addition to that an optimization process using genetic algorithm is also being planned in order to reduce the number of attributes without sacrificing accuracy and efficiency for diagnosing the CVD.

A. Naïve Bayes

A Naive Bayes classifier predicts that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature This classifier is very simple, efficient and is having a good performance. Sometimes it often outperforms more sophisticated classifiers even when the assumption of independent predictors is far. This advantage is especially pronounced when the number of predictors is very large. One of the most important disadvantages of Naive Bayes is that it has strong feature independence assumptions.

B. Decision Trees

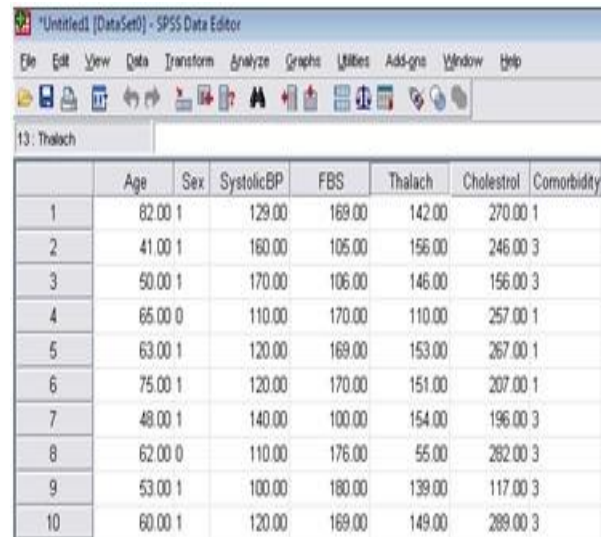
Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. Decision trees are commonly used in operations research, mainly in decision analysis. Some of the advantages are : they can be easily understood and interpreted, robust, perform well with large datasets, able to handle both numerical and categorical data. Decision-tree learners can create over-complex trees that do not generalize well from the training data is one the limitation.

C. Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unverified classification and has no predefined classes.

They are used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Moreover, they are used for data compression, outlier detection, understand human concept formation. Some of the applications are Image processing, spatial data analysis and pattern recognition. Classification via Clustering is not performing well when compared to other two algorithms.

All these algorithms are implemented with the help of SPSS16.0 AND RATTLE3.1 tool for the diagnosis of CVD diseases. Data set of 1000 records with 7 attributes. These Algorithms have been used for analyzing the CVD dataset. The Classification Accuracy should be compared for this algorithm. After the comparison attributes are to be reduced for further purpose.



	Age	Sex	SystolicBP	FBS	Thalach	Cholestrol	Comorbidity
1	62.00	1	129.00	169.00	142.00	270.00	1
2	41.00	1	160.00	105.00	156.00	246.00	3
3	50.00	1	170.00	106.00	146.00	156.00	3
4	65.00	0	110.00	170.00	110.00	257.00	1
5	63.00	1	120.00	169.00	153.00	267.00	1
6	75.00	1	120.00	170.00	151.00	207.00	1
7	48.00	1	140.00	100.00	154.00	196.00	3
8	62.00	0	110.00	176.00	55.00	262.00	3
9	53.00	1	100.00	180.00	139.00	117.00	3
10	60.00	1	120.00	169.00	149.00	289.00	3

Fig:(1) dataset used (comorbidity:-1:T₂DM, 3:HTN, Age:-1:Male, 0:Female

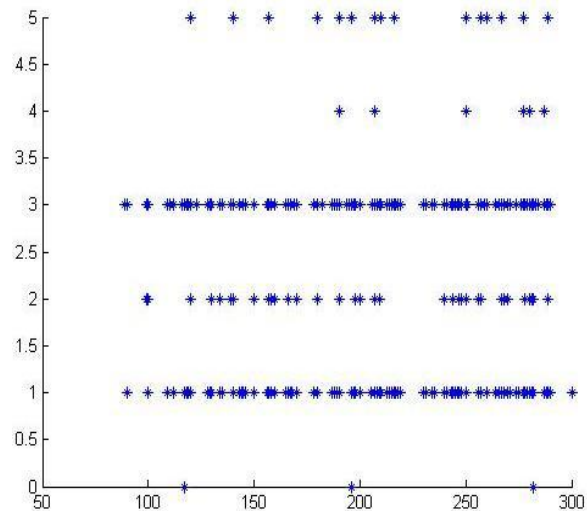


Fig: (2) Clustering of dataset using Scatter Graph

III. PRINCIPLES OF PREDICTIVE DATA MINING

There are many principles which are used for predicting the CVD.

A. Bayes theorem Bayes rule is used in naive bayes algorithm for the manipulation of conditional probabilities.

Bayes' theorem gives the relationship between the probabilities of A and B, P(A) and P(B), and the conditional probabilities of A given B and B given A, P(A|B) and P(B|A).

$$P(A|B) = P(A \cap B) / P(B)$$

B. Entropy

Entropy is one of the principles which is used in decision tree and is to measure the amount of information in an attribute and also the impurity. The general formula is:

$$Entropy(S) = -\sum p(I) \log_2 p(I)$$

IV. PARAMETERS OF PDM

Some of the parameters [4] which are used for Predictive data mining are

A. Sensitivity: It is also known as True Positive Rate. It is used for measuring the percentage of sick people from the dataset.

Sensitivity = $\frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$

B. Specificity: It is also known as True Negative Rate. It is used for measuring the percentage of healthy people who are correctly identified from the dataset.

Specificity = $\frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}$

C. Precision and recall It is also known as positive predictive value. It is defined as the average probability of relevant retrieval.

Precision = $\frac{\text{Number of true positives}}{\text{Number of true positives} + \text{False positives}}$

Recall It is defined as the average probability of complete retrieval.

Recall = $\frac{\text{True positives}}{\text{True positives} + \text{False negative}}$

D. Accuracy

A measure of a predictive model that reflects the proportionate number of times that the model is correct when applied to data.

The formula for calculating the Accuracy,

Accuracy = $\frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}$

E. Confusion Matrix It is used for displaying the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is represented in the form of n-by-n, where n is the number of classes. The accuracy of each classification algorithms can be calculated from that.

V. IMPLEMENTATION

The working of the architecture is as follows: the data's of the patients who are having CVD disease has been collected from the hospital. For the diagnosis of CVD disease here two algorithms are being used which are Naïve Bayes and Decision Tree. The prediction of CVD disease is executed with the help of a tool known as SPSS16.0 and Rattle3.1. Here the dataset is being used as the input for the prediction. The dataset consists of attributes and values. This tool will results the accuracy that how many patients are having the CVD disease with in a particular time. In order to improve the efficiency and accuracy an optimizations process is carried out using genetic algorithm. Summary of working of genetic algorithm:

A. ALGORITHM

1. Create a random initial population.
2. To create new population fitness value of current population has to be found.
3. Scales the raw fitness scores to convert them into a

more usable range of values.

4. Selects members, called parents, based on their fitness.
5. Some of the individuals in the current population that have lower fitness are chosen as elite. These elite individuals are passed to the next population.

6. Produces children from the parents and the operation is known as crossover. Children are produced either by making random changes to a single parent called Mutation. The genetic algorithm is being implemented with the help of Matlab. The optimized attributes are fed into SPSS16.0 and Rattle3.1 tool for the prediction purpose. Hence we will get a conclusion that optimization technique is the best method for improving the prediction of CVD. The Implementation has been done for finding the accuracy of decision tree and naïve bayes. The optimization part is the future work which is colored in red box.

B. DATA SET

The data set used in this work is collected from UCI machine learning repository which is a repository of databases, domain theories and data generators. These are the attribute names which is the input given for patients record.

C. PREDICTIBLE ATTRIBUTE

Predictable Attribute

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

Input Attributes

1. Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value3:non-angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
5. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
6. Exang - exercise induced angina (value 1: yes; value 0: no)
7. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
8. CA – number of major vessels colored by floursopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholestrol (mg/dl)
12. Thalach – maximum CVD rate achieved
13. Oldpeak – ST depression induced by
8. CA – number of major vessels colored by floursopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholestrol (mg/dl)

12. Thalach – maximum CVD rate achieved
13. Oldpeak – ST depression induced by exercise

VI. RESULTS

DM Technique	Accuracy
Naïve Bayes	87.06
Decision Tree	74.03

VII. CONCLUSIONS

Many experiments were conducted with the same datasets in SPSS 16.0 tool and Rattle3.1. A data set of 1000 records with 5 attributes is used and the outcome reveals that the Naïve Bayes outperforms and sometime Decision Tree. In Future Genetic algorithm will be used in order to reduce the actual data size to get the optimal subset of attribute sufficient for CVD prediction. Prediction of the CVD disease will be evaluated according to the result produced from it. Improvement is done to increase its consistency and efficiency. Benefit of using genetic algorithm is the prediction of CVD can be done in a short time with the help of reduced dataset. Genetic algorithm will be implemented with the MATLAB

REFERENCES

- [1] Breiman, L. (1996). "Bagging predictors." *Machine Learning*, 24 (2): 123–140.
- [2] Breiman, L. (2001). "Random forests." *Machine Learning*, 45 (1): 5–32.
- [3] Demšar, J. (2006). "Statistical comparisons of classifiers over multiple data sets." *The Journal of Machine Learning Research*, 7: 1–30.
- [4] Domingos, P. (2000). "A unified bias-variance decomposition for zero-one and squared loss." In *Proceedings of the National Conference on Artificial Intelligence*, 564–569.
- [5] Fawcett, T. (2006). "An introduction to ROC analysis." *Pattern Recognition Letters*, 27 (8): 861–874.
- [6] Freund, Y. and Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences*, 55 (1): 119–139.
- [7] Friedman, J. H. (1997). "On bias, variance, 0/1-loss, and the curse-of-dimensionality." *Data Mining and Knowledge Discovery*, 1 (1): 55–77.