International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 11 – No. 2, July 2016 – www.ijais.org

Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods

Rovina Dbritto
Student ME
Computer Engineering
St. Francis Institute of Technology, Mumbai India

Anuradha Srinivasaraghavan
Associate Professor Computer Engineering, St. Francis Institute of Technology, Mumbai India

Vincy Joseph
Assistant Professor Computer Engineering, St. Francis Institute of Technology, Mumbai India

ABSTRACT
A common term heart disease is nothing but a cardiovascular disease or a Coronary heart disease which reduces the efficiency and proper functioning of heart by blocking veins, artery or blood vessels around it. Coronary heart disease causes disability such as damage to the brain resulting in death. Based on Statistics [10] it indicates that range of age group from 25 to 69 have 25% risk of having heart diseases. Some vital causes for cardiovascular disease are, physical inactivity, smoking, consuming more junk food and addiction of alcohol which are major causes for stroke, chest pain, and heart attack. However because of the awareness about factors and symptoms that are responsible for heart problem, it is possible to predict any heart problem based on statistical analysis of medical records. However Data mining, a modern technique has provided an automatic way of analyzing data using standard classification methods. Though many classifiers are available in data mining that can be used to predict the heart problems, this paper emphasizes on finding the appropriate classifier that has the potential to give better accuracy by applying data mining techniques viz. Naïve Bayes , Support Vector machine and Logistic Regression.

Keywords
Coronary, Naive Bayes, Support Vector Machine, Logistic Regression

1. INTRODUCTION
Heart diseases especially coronary heart disease is a very fatal and dangerous disease because if patient ignores its earlier symptoms, which seems to be a warning signs, it gives no time to patient for recovery and eventually may lead to death on spot. This is called as heart attack. It happens because the function of the arteries is to supply oxygen rich blood to the heart but due to the fatty and other substance the plaque is formed which turns normal coronary artery into narrowing of the coronary artery. Coronary heart disease is a disorder in which a waxy element called plaque builds up inside the coronary arteries. Thus the flow of the blood to the heart can either slow down or stop as shown in the Fig 1.

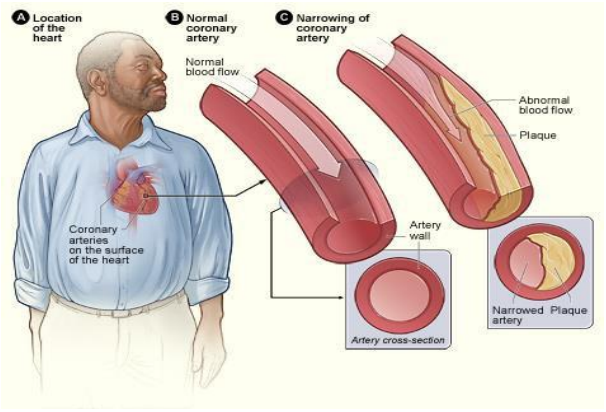

Fig 1: Cause of Heart attack

The main factors which are responsible for coronary disorder are classified as controllable risk factors and uncontrollable risk factors. Some of the controllable risk factors include diabetes, smoking and obesity or excess weight, cholesterol, High blood pressure, lack of exercise and physical activity. Uncontrollable risk factors include family history, age, sex and previous medical disorders. Now a days Electronic Medical Records(EMR) and Electronic Health Records (EHR) has simplify and systematized the analysis part for detecting the patients problems, but prediction of diagnosis with accuracy is still a challenge among present researchers who are contributing in proposing and developing different methodologies in the field of critical Human diseases such as Cancer, heart disease, diabetes etc. Some of the contributions are mentioned in the next section which gives a brief description of activities going on presently.

2. LITERATURE SURVEY
There are different data mining techniques for classification [3].Performance analysis on different classification algorithms such as Decision tree, Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Neural Networks (NN) are carried out. After evaluation it is found that NB gives more accuracy compared to other classifiers. Also the dimensionality of the data can be reduced by attribute selection methods. After reducing the data also it is found that Naïve Bayes gives more accuracy using correlation feature selection method.

There are another five models constructed of single and combined data mining technique that cares medical decisions in heart disease analysis and prediction. The five systems offer automatic pattern recognition and endeavors to find relationships among different parameters and symptoms of heart disease. Each system displays set of strengths and

limitations in terms of the type of data it handles, accuracy, and ease of understanding, reliability and generalization ability. Various data mining techniques such as Naïve Bayes, Support vector machine and decision tree are used [4].

Another method suggest to use data mining techniques such as Genetic Algorithm, Support Vector Machine (SVM), association rules, rough set theory and Neural Networks. Out of the above techniques Decision Tree and SVM is most effective for the heart disease. For future work, more Accuracy can be increased by increasing the attributes by using different data mining techniques [5].

This [6] paper discusses and presents the experiment that was executed with Naïve Bayes technique in order to build predictive model as an artificial diagnose for heart disease based on data set which contains set of parameters that were measured for individuals previously. Accuracy of the naïve bays model achieved ratio (100%).

The three data mining techniques are used in this paper such as CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) and also evaluated the performance of the three classifiers. Different classifiers are studied and the research is conducted to find the best classifier for calculating the patients Diagnosis. The best algorithm is CART which gives more accuracy [10].

This paper uses K-means clustering technique which is applied to find out clusters in data which are further used to remove hidden forms related to heart patients. In future work, they have planned to implement an expert system that would predict the probability of patient being Critical or at risk state using logistic regression algorithm and these extracted patterns [11].

3. PROPOSED METHODOLGY

There has been lots of work done on Coronary Heart disease classification and prediction using following data mining techniques.

- Naïve Bayes
- Decision tree
- K nearest neighbor
- Support vector machine

From the literature review it's evident that although naïve Bayes and support vector machines are superior algorithms the accuracy has yet to be verified with larger data sets. So the proposed solution will work in three major modules.

First by increasing the size of the data set the accuracy of prediction would be computed. Suggestion based on the accuracy amongst the two algorithms with increased data sets would be made.

Second since logistic regression can also be used for prediction, a prediction model using logistic regression would be developed

Third a comparison based on the above three modules based on its accuracy would result in a best model for heart disease prediction.

The attributes which are majorly used are taken from the referenced paper [13] where attributes and its usage are mentioned in detail, those are as follows:

- Age
- Gender
- Chest pain
- Resting blood pressure
- Cholesterol
- Fasting blood sugar
- Resting electrocardiographic results
- Maximum heart rate achieved
- Exercise induced angina
- ST depression induced by exercise relative to rest
- Slope of the peak
- Number of major vessels colored by fluoroscopy
- Thalassemia

Steps of implementation are as follows: -

1. Collect Data set that can be used for testing purpose (use of data Set from UCI Repository)
2. Implement Naïve Bayes Algorithm in such a way that it should be able to take data set as an input set of attributes values.
3. Implement Support Vector Machine in a similar way that can also take data set as input values.
4. Build a prediction Model using Logistic Regression Approach
5. Compare the Accuracy based on data and result analysis.

4. DATA SET

Data set from Data mining repository of University of California, Irvine (UCI) has been collected for testing purpose, which consists of collection of data set from Cleveland, Hungary, Switzerland and long beach and Stat log. Some set of Data from UCI Repository are depicted in figure 2.

Age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0

Fig 2: Data set Sample [7]

4.1 Naïve Bayes Algorithm

Naive Bayes algorithm is a good tool in medical diagnosis. For instance given a list of attributes named symptoms, it can speculate probability of a disease. Naïve Bayes consider

conditionally independent attributes. The classifier processes each attribute's probability in a class. The highest posterior Probability class will be considered as result of the classification. Naïve Bayes is simple, efficient and good performance in classification. It also provides good accuracy for general purpose analysis. Due to its good accuracy it can be used in medical diagnosis. The basic approach that can be used is depicted in following figure 3.

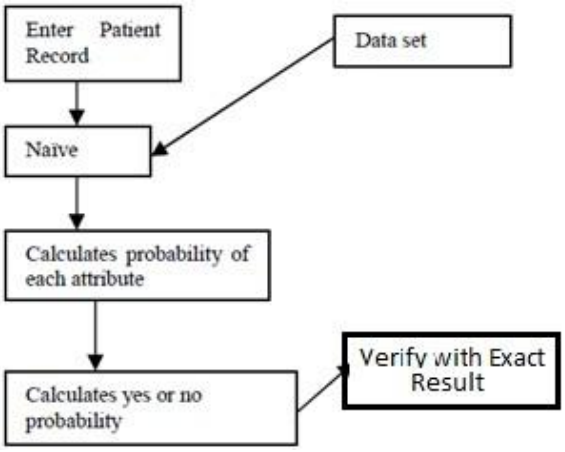

Fig 3: Flow Chart of Naïve Bayes Algorithm on the patient data [7]

4.2 Support Vector Machine (SVM):

A support vector machine is a classification type of method used to scrutinize data and recognize patters in a regression and classification analysis. Support vector machine (SVM) is used when your data is classified as two classes. An SVM recognizes and separates similar data by finding the best hyper plane that separates all data points of one class from those of the other class. Model becomes better when margins are larger between classes. A margin should not have points in its interior part. The support vectors are the data coordinates that are on the boundary of the margin. Mathematical functions are involved in SVM design which is frequently used to model real world problems. Its performance magnify with number of attributes [12].

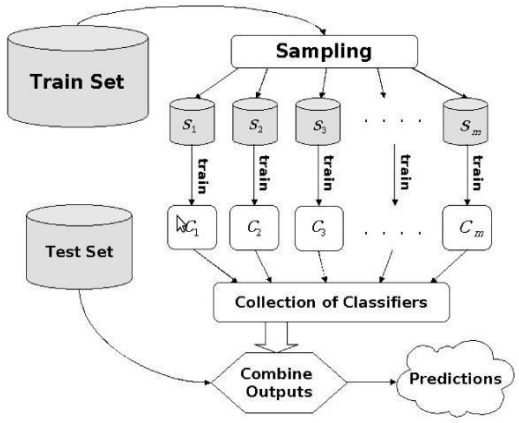

Fig 4: Flow Chart of SVM

4.3 Logistic Regression

It's a type of statistical regression analysis method used for approximation and prediction of result of a definite dependent attributes. Dependent means it can take only some set of values for example binary values such as true or false, good or bad, on or off likewise. Logistic regression is mainly used for prediction besides that it can also be used in calculating the probability of success. Basically Logistic Regression involves fitting an equation of the form to the data:

Y = ß0 + ß1x1 + ß2x2 + … + ßnxn – *eq. 1*

The regression coefficients are usually estimated using maximum likelihood estimation. The maximum likelihood ratio helps to determine the statistical significance of independent variables on the dependent variables. The likelihood-ratio test assesses the contribution of individual predictors (independent variables). Then the probability (p) of each case is calculated using odds ratio,

P/(1-P) = eY – eq. 2

From this p–value is found out. This gives the probability or chance for the individual to have coronary heart disease[12].

5. RESULT ANALYSIS

5.1 Performance Analysis Metrics:

Performance will be evaluated based on following parameters:
Accuracy = (TP + TN)/ (TP + FP + TN + FN)

Where TP =true positives is the number of positive cases correctly classified

TN = true negatives is the number of negative cases correctly classified

FP =false positives is the number of negative cases incorrectly classified as positive

FN= false negatives is the number of positive cases incorrectly classified as negative.

This accuracy vary with the size of data set, so here a comparison table has been created to analyze the impact of size on each methods viz. Naïve , SVM and Logistic while providing accuracy that is depicted in fig. 4.

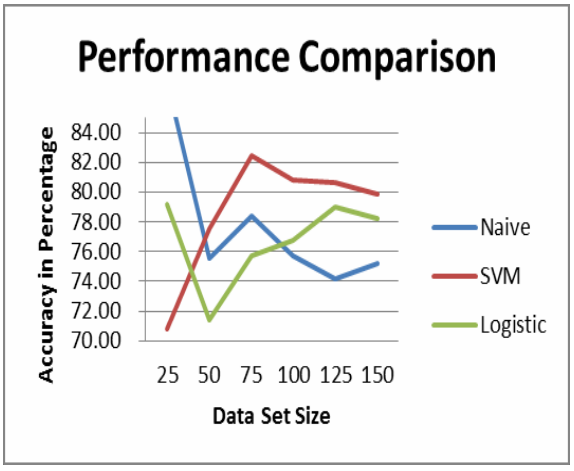

Fig 5: Performance Analysis

5.2 Graphical Analysis

Following graph is showing accuracy wise comparison of Naïve Bayes, SVM and Logistic Regression methods for large data set of more than 1000 entries. From this graph in figure 5, it can be concluded that Support vector Machine is more accurate as compare to other two.

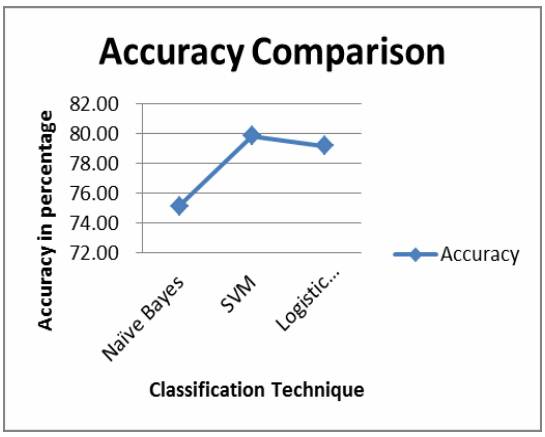

Fig. 6 Accuracy Comparison

6. CONCLUSION AND FUTURE WORK

In this paper, have discussed about the major classification techniques used in data mining for prediction of heart disease is discussed and with the help of accuracy analysis we have shown that SVM is better than other two methods when we have large data set of entries.

The system with SVM will help for early diagnosis of the heart disease for any given patient. This system when deployed can complement traditional heart disease detection system and can help not only the doctors but also the patients. This system can act as a boon for the medical industries for detection of the heart disease with better accuracy.

7. REFERENCES

[1] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis," Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions On Information Technology In Biomedicine, VOL. 14, NO. 3, MAY 2010.

[2] Raghunath Nambiar, Adhiraaj Sethi, Ruchie Bhardwaj, Rajesh Vargheese," A Look at Challenges and Opportunities of Big Data Analytics in Healthcare", 2013 IEEE International Conference on Big Data.

[3] T.John Peter, K. Somasundaram," An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques", IEEE, International conference on Advances in engineering, science and management,pp.514-518, 2012.

[4] Eman AbuKhousa, Piers Campbell," Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems", IEEE, International Conference on Innovations in Information Technology, pp.267-272, 2012.

[5] Aqueel Ahmed, Shaikh Abdul Hannan," Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012.

[6] Lamia AbedNoor Muhammed," Using Data Mining technique to diagnosis heart disease", IEEE, International conference on statistics in science, Buiseness and Engineering, pp.1-3, 2012.

[7] Sivagowry S, Dr. Durairaj. M and Persia. A & Research Scholar, "An Empirical Study on Applying Data Mining Techniques for the Analysis and Prediction Heart Disease", IEEE, International Conference on Information Communication and embedded system, pp.265-270, 2013.

[8] M.Akhil jabbar , Dr.Priti Chandra, Dr.B.L Deekshatulu," Heart Disease Prediction System Using Associative Classification and Genetic Algorithm", ICECIT, 2012.

[9] Ranganatha S., Pooja Raj H.R., Anusha C., Vinay S.K.," Medical Data Mining And Analysis For Heart Disease Dataset Using Classification Techniques",IEEE, National conference on challenges in research and technology in the coming decades,pp.1-5,2013.

[10] Vikas Chaurasia, Saurabh Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech, 2013, Vol.1, 208-217.

[11] Mamuna Fatima, Iqra Basharat, Dr. Shoab Ahmed Khan, Ali Raza Anjum,, "Biomedical (Cardiac) Data Mining: Extraction of significant patterns for predicting heart condition", IEEE conference on Computational Intelligence in bioinformatics and computational biology, pp.1-7, 2014.

[12] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", IJCA, Vol.68- No.16 April 2013.

[13] Carlos O., Edward O , Levien de Braal, and team "Mining Constrained Association Rules to Predict Heart Disease", IEEE, International Conference on Data Mining p.433-440, 2001.