International Journal of Mechanical Handling and Automation
Vol. 1: Issue 2
www.journalspub.com

A Comparative Study on Naïve Bayes and Bayes Net Classifiers for the Heart Disease Prediction System

V. Sabarinathan[1], K. Rajesh Khanna[2], V. Sugumaran[2], V. J. Sarath Kumar[3]*
[1]School of Information Technology, SRM University, Chennai, India
[2]School of Mechanical and Building Science, VIT University, Chennai, India
[3]School of Electrical Engineering, SRM University, Chennai, India

Abstract

Cardio vascular diseases are the leading cause of death in the recent years. Large volume of medical data which are available in health care is used to identify the hidden information. From a set of symptoms, one should be able to predict the heart diseases. Here, a predictive model is built using the machine learning approach. This is done in two steps; firstly, feature selection is done through decision tree. Then comparative study on Naïve Bayes and Bayes net classifiers is discussed. Bayes classifiers are built using conditional probability and hence, it requires less number of data points for training. The results show that Bayes net gives maximum of 85% accuracy with three attributes whereas Naïve Bayes yields maximum of 85% with all the 13 attributes. The suitable algorithm is presented in the conclusion.

Keywords: bayes net classifiers, naïve bayes, cardio vascular diseases

**Corresponding Author*
Email ID: v_sugu@yahoo.com

INTRODUCTION

In recent years, cardio vascular diseases are found to be the major causes of death. Numerous research efforts have been made to identify the reasons behind cardio heart diseases. Now, the challenge is to build a predictive model which will tell whether the unknown new patient is having heart diseases or not from the symptoms. Data mining/machine learning algorithms have the capacity to build predictive model from a classified data. Successful models are built on many application areas such as medical data mining; science and engineering; business analysis, etc. Number of classification algorithms are available such as decision tree, *k* means algorithm, bagging algorithm, etc. Many research scholars tested and developed prediction models with classifiers such as decision tree, neural networks, and support vector machine. The proposed model will act as decision support system for the doctors and health care centers to predict the presence or absence of heart disease in a person. This may help in identifying the heart diseases at earlier stage and thus, saving human life.

The research process requires medical data set which includes 13 different attributes such as thalassemia, chest pain type, colored fluoroscopy, etc. In this paper, decision tree is used for feature selection and Naïve Bayes and Bayes net algorithms are used for classification. Decision tree gives structural information hidden in the data in the form of tree. One can easily select the best features that contribute to the classification by merely looking at the tree. The attributes that lie on the top of the tree are the best ones. The attribute that

 © JournalsPub 2016. All Rights Reserved

fall on the bottom of the tree can be consciously ignored due to its meagre contribution towards classifications. Bayes net has advantage that small error in the data does not affect the overall performance of the system. This is because the conditional probabilities do not vary too much due to small number of erroneous data. Also, training these classifiers does not require skilled professionals, since there is no parameter that needs to be tuned in the process of training a classifier. In Bayes classifiers, increase in number of data points gives better conditional class probabilities which in turn give better results. The research paper mainly focuses on comparing the classification accuracy of Bayes net and Naïve Bayes classifiers.

REVIEW LITERATURE

Mai Shouman *et al*., used decision tree for diagnosis of heart diseases. [1] The research process involves data selection through decision tree, data discretization and data portioning. The efficiency in result is improved by equal frequency discretization and gain ratio decision tree. They achieved maximum accuracy of 84.1% by the equal frequency discretization gain ratio decision tree. Chaitrali *et al*., proposed a data mining approach for the prediction of heart disease using neural networks. [2] Initially, they took 573 records and 13 different attributes for prediction and later, for better accuracy they added two more attributes namely obesity and smoking. Chitra *et al.* used hybrid intelligent techniques for the heart disease prediction system. They used neural network to train the dataset. [3] They have achieved 100% accuracy in the result; however, they failed to identify the key attributes which causes cardiovascular diseases and hence, neural network does not provide insight towards heart diseases prediction.

Sen *et al*., used neuro-fuzzy integrated approach for the prediction of coronary heart diseases. [4] They used multi-level approach for the prediction and initially, the critical factors that cause coronary heart disease are placed at first level and the other attributes are placed at second level. Neural network is used to train the dataset and fuzzy logic was used to interpret the factors that cause the heart diseases. Chaurasia *et al*. used data mining tool to detect the heart disease. [5] Classifiers such as j48 decision tree, bagging algorithm and Naïve Bayes were used in the research process. Initially, the dataset contains 303 records with 76 different attributes and later, 11 attributes were selected for further classification. Their result shows 85% accuracy using bagging algorithm. However, there is no justification given for equal number of data point in each class; else, the system will be biased towards the trained datasets that has more instances.

Laksmi *et al*., proposed a data mining approach for heart disease survivability. [6] The research work was based on support vector machine, linear discriminant analysis, bayesian linear regression, partial least square-linear discriminant analysis, apriori algorithm. Totally 2268 instances were taken which include 15 attributes. Various data mining techniques were applied and finally, partial least squares regression gave 86.13% accuracy. Wilson *et al*., developed heart disease prediction system using data mining techniques such as *k*-means and weighted association rule. [7] Their result is based on *k*-means decision tree technique which gives more accurate result than apriority algorithm. Various classifiers discussed above have its own advantages and disadvantages in heart disease prediction system. Bayes classifier uses conditional probabilities and also does not require skilled professionals in training the dataset. In the present study, Naïve Bayes and Bayes net classifiers are

taken up and their classification performances were compared.

FEATURE DESCRIPTION

The dataset is taken from UCI machine learning repository (https://archive.ics.uci.edu/ml). Initially, 13 attributes such as Thalassemia, chest pain type, colored fluoroscopy, exercise induced angina and resting blood pressure were taken for training and testing. Features are defined under 4 different types such as ordered, binary, nominal and real. Age, resting blood pressure, serum cholesterol and maximum heart achieved, old peak and colored fluoroscopy comes under real values. Resting electrocardiographic results, chest pain type and thalassemia comes under nominal. Sex, exercise induced angina and fasting blood sugar comes under binary data type. The slope of the peak exercise segment comes under ordered data type. Sabarinathan *et al*., proposed diagnosis of heart disease using decision tree which gives detailed view on feature selection based on decision tree. [8–10] Feature descriptions about thalassemia, serum cholesterol, chest pain type can be found in the same paper. The remaining features are described below. Exercise induced angina means inadequate flow of blood to the heart muscle from narrowing of coronary arteries. The term angina denotes discomfort or pain in the chest region.

The rate at which heart beats when the body is at rest is termed as resting blood pressure. Resting blood pressure should be low for a normal person and comparatively high for heart disease affected person. The measurement of glucose level in the blood is taken for medication during refraining from eating or drinking any liquids other than water is called fasting blood sugar. Blood sugar test will be taken first for the heart disease affected person.

The electrical activity of heart is measured by Electrocardiogram (ECG) which is used during diagnosis of heart diseases. The electrical pulses can be translated to tracings on the paper. Table 1 gives the list of features that are present in the dataset. Totally 240 instances were present in the dataset. The dataset consists of equal number of healthy and heart disease affected persons. During feature selection, the dataset should contain equal number of instances in both the cases; else the system will be biased towards the case which has more number of instances.

Table 1: *List of Features.*

Order of Features	Name of Features
1	Thalassemia
2	Chest pain type
3	Colored fluoroscopy
4	Exercise induced angina
5	Resting blood pressure
6	Sex
7	Fasting blood sugar
8	Old peak
9	Slope of peak exercise st segment
10	Age
11	Serum cholesterol
12	Resting electrocardiographic results
13	Max. heart rate achieved

CLASSIFIERS

Bayes Net

Bayes net classifier is built using conditional probabilities through directed acyclic graphical (DAG) model. The dependency and independent functions are represented through edges and nodes of directed acyclic graphical model respectively. Bayesian net drastically reduces the space for storing the values. As the number of data points increases, the conditional probability will give better classification accuracy.

Depth first algorithm is possible with Bayes net after the formation of directed acyclic graphical model. Probabilistic queries can also be answered through Bayes net. The joint probability is

computed using the chain rule. [11] Figure 1 shows simple Bayes net DAG.

$P(X_i|parents\ (X_i))\ for\ each\ node\ X_i$

$P(X_1, \ldots\ldots, X_n) = \pi_{i=1}^{n}\ P(X_i|\ X_1, \ldots\ldots, X_{i-1})$ Eq. (1)

$= \pi_{i=1}^{n} P(X_i|parents\ (X_i))$ Eq. (2)

$For\ parents\ (X_i) \subseteq \{X_1, \ldots\ldots, X_{i-1}\}$

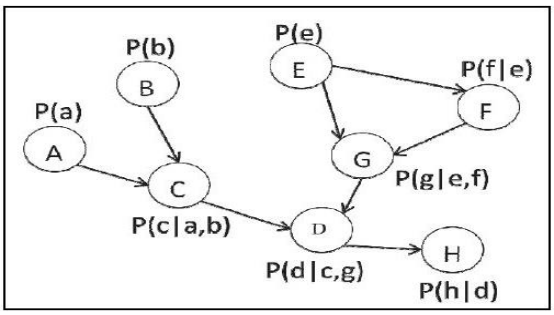

Fig. 1: *Bayes Net DAG.*

Naïve Bayes

Naïve Bayes is a simple classifier which assumes that the value of particular attribute is independent of other attribute present in the class. Each attribute is assumed to contribute independently for the probability that it belongs. In a supervised learning setting, Naïve Bayes can be efficiently trained.

The advantage of Naïve Bayes is that it requires only small amount of dataset for training. The detailed information for the conditional probability equations that can be found in Ref. [11].

$$P(A_1, \ldots. A_n|\ B) = \pi_{i=1}^{n}\ P(A_i|B) \quad \text{Eq. (3)}$$

$$P(B = b_k|\ A_1 \ldots. A_n) = \frac{P(B=b_y)\ P(A_1 \ldots\ldots A_n|\ B= b_k)}{\sum_j P(B=b_j) P(A_1 \ldots\ldots A_n|\ B=b_j)} \quad \text{Eq. (4)}$$

$$P(B = b_k|\ A_1 \ldots. A_n) = \frac{P(B=b_y)\ \pi_i P(A_1|\ B= b_k)}{\sum_j P(B=b_j) \pi_i P(A_1|\ B=b_k)} \quad \text{Eq. (5)}$$

Table 2: *Classification Accuracies of Both Classifiers versus Number of Features.*

No. of Features	Classification Accuracy (%)	
	Bayes Net	Naïve Bayes
1	76.66	76.66
2	75.41	76.66
3	**85.00**	79.58
4	85.00	78.30
5	83.3	82.5
6	84.16	82.5
7	84.16	82.08
8	83.75	83.33
9	82.91	83.75
10	82.50	82.91
11	82.50	83.75
12	82.50	84.16
13	83.30	**85.00**

RESULTS AND DISCUSSIONS

The paper deals with building predictive model using data available. In the present study, Naïve Bayes and Bayes net classifiers are considered. They are trained using data available in UCI repository. From the database, [12] initially 13 attributes listed in Table 1 were taken for classification in both classifiers; however, for the improvement in the accuracy, the attributes can be optimized. All 13 features may not contribute equally for classification. Some features may contribute more compared to others. To find the effect of number of features on classification accuracy, an experiment is carried out. Referring to Sugumaran *et al*., [10] the order of features based on their contribution is listed in Table 1. Initially, top most feature (Thalassemia) alone was considered and a Naïve Bayes and Bayes net model was built. The corresponding classification accuracy was noted down as depicted in Table 2. Then, only two

features (thalassemia and chest pain type) were considered for building Naïve Bayes and Bayes net model and the corresponding classification accuracy was noted and tabled. This procedure was repeated till, the number of features reached 13. The classification accuracy can be found in Table 2.

Bayes Net

Tables 3 and 4 give detailed accuracy of Bayes net classifier and confusion matrix respectively. TP (True Positive) rate should be near to 1 and FP (False Positive) rate should be near to 0. In the present case, TP rate happens to be 0.85 and FP rate is 0.15 which can be accepted practically. Confusion matrix identifies the correct and incorrect classifications. Here, heart disease affected person is identified as HD and normal person is identified as NORMAL.

Table 3: *Classification Accuracy for Bayes Net.*

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.817	0.117	0.875	0.817	0.845	0.866	HD
	0.883	0.183	0.828	0.883	0.855	0.866	NORMAL
Weighted Average	0.85	0.15	0.852	0.85	0.85	0.866	

Table 4: *Confusion Matrix for Bayes Net.*

Classified as		
A (HD)	98	14
B (NORMAL)	22	106

The first row in the confusion matrix denotes heart disease affected person, HD. The second row denotes the normal person, NORMAL. First row has 22 incorrect classified data points that is 22 heart disease affected person is classified as normal person due to which person may not undergo any medication which, in turn may causes loss of life.

Second row has 14 incorrectly classified data points. This may results in creating mental agony to the person.

Naïve Bayes

Tables 5 and 6 give detailed accuracy for Naïve Bayes classifier and confusion matrix respectively. TP rate and FP rate are 0.85 and 0.15 respectively that can be accepted practically. The result shows 85% accuracy and needs all 13 attributes to train the dataset.

Table 5: *Classification Accuracy for Naïve Bayes.*

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.842	0.142	0.856	0.842	0.849	0.896	HD
	0.858	0.158	0.844	0.858	0.851	0.896	Normal
Weighted Average	0.85	0.15	0.85	0.85	0.85	0.896	

Table 6: *Confusion Matrix for Naïve Bayes.*

Classified as		
A (HD)	101	17
B (Normal)	19	103

The interpretation of confusion matrix remains same for Naive Bayes classifier also. The diagonal elements representing correctly classified instances and non-diagonal elements representing incorrectly

classified instances. Out of misclassified instances, 17 instances are normal, but classified as persons having heart disease is a bad decision. This situation is not very critical as it causes only mental agony to the person, whereas 19 instances that are heart disease affected persons were misclassified as normal person.
This situation is very critical as it may causes loss of life.

CONCLUSION

The research paper was mainly focused on comparing the performance of classification in Bayes net and Naïve Bayes for diagnosis of heart disease. As seen in the results and discussion, the classification of Bayes net and Naïve Bayes gives 85% accuracy. Bayes net takes 3 features to train the dataset whereas Naïve Bayes takes 13 features to train the dataset. Bayes net seems to have lesser computational time than Naïve Bayes.

However, the number of critical misclassified instances (heart disease affected person as normal) is comparatively less with Naïve Bayes classifier which helps in saving human life. It should be noted that heart diseases prediction system should have lesser misclassified data points. Saving human life is more important than optimizing the attributes. Thus, a practical system using Naïve Bayes classifier will be a more efficient and useful in diagnosing the heart diseases prediction system.

REFERENCES

1. Shouman M, Turner T., Stocker R. Using Decision Tree for Diagnosing Heart Disease Patients. *Proceedings of the 9th Australasian Data Mining Conference*; 2011; Ballarat, Australia. 2011; 23–9p.
2. Dangare C.S., Apte S.S. A Data Mining Approach for Prediction of Heart Disease using Neural Networks. *International Journal of Computer Engineering & Technology (IJCET)*. 2012; 3(3): 30–40p.
3. Chithra R., Seenivasagam V. Review of Heart Disease Prediction System using Data Mining and Hybrid Intelligent Techniques. *ICTACT Journal on Soft Computing*. 2013; 3(4): 605–9p.
4. Sen A.K., Patel S.B., Shukla D.P. A Data Mining Technique for Prediction of Coronary Heart Disease using Neuro-Fuzzy Integrated Approach Two Level. *International Journal of Engineering and Computer Sciences (IJECS)*. 2013; 2(9): 2663–71p.
5. Chaurasia V., Pal S. Data Mining Approach to Detect Heart Disease. *International Journal of Advanced Computer Sciences and Information Technology*. 2013; 2(4): 56–66p.
6. Lakshmi K.R., Krishna M.V., Kumar S.P. Performance of Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. *International Journal of Scientific Research and Publications*. 2013; 3(6): 1–10p.
7. Wilson A., Wilson G., Likhiya J.K. Heart Disease Prediction using Data Mining Techniques. *International Journal of Computer Sciences Trends and Technology*. 2014; 2(1): 84–8p.
8. Kaur B., Singh W. Review on Heart Disease Prediction System using Data Mining Techniques. *International Journal of Recent and Invoation Trends in Computing and Communication*. 2014; 2(10): 3003–8p.
9. Sudhakar K., Manimekalai M. Study of Heart Disease Prediction using Data Mining. *International Journal of Advanced Research in Computer Science and Software Enginnering*. 2014; 4(1): 1157–60p.

10. Sabarinathan V., Sugumaran V. Diagnosis of Heart Disease using Decisio Tree. *International Journal of Research in Computer Applications & Information Technology*. 2014; 2(6): 74–9p.
11. Vedant, Sugumaran V., Amarnath M., *et al.* Fault Diagonosis of Helical Gear Box using Sound Signal using Naïve Bayes and Bayes Net. *International Journal of Engineering & Technology Research*. 2015; 1(1): 98–105p.
12. Database Collected from Statlog (Heart) Data Set using UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Statlog+Heart.