Heart disease prediction system based on hidden naïve bayes classifier

M.A.Jabbar
Professor,
Vardhaman College of Engineering,
Hyderabad, Telangana, INDIA
email:jabbar.meerja@gmail.com

Shirina samreen
Professor
Anurag Group of Institutions
Hyderabad,Telangana
shirina.samreen@gmail.com

***Abstract*— Coronary heart disease is a major cause of death world wide. The diagnosis of heart disease is a tedious task. There is a need for an intelligent decision support system for disease prediction. Data mining techniques are often used to classify whether a patient is normal or having heart disease. Hidden Naïve Bayes is a data mining model that relaxes the traditional Naïve Bayes conditional independence assumption. Our proposed model claims that the Hidden Naïve Bayes (HNB) can be applied to heart disease classification (prediction). Our experimental results on heart disease data set show that the HNB records 100% in terms of accuracy and out performs naïve bayes.**

Keywords— hidden naïve bayes, data mining, classification, heart disease

I. INTRODUCTION

Heart disease is a condition that affects the heart. As the "coronary arteries narrow, blood flow to the heart can slow down or stop, causing chest pain, heart attack" [1]. Diagnosing heart disease requires highly skilled and experienced physicians [2]. Data mining which is an integration of multiple disciplines is to extract knowledgeable information from huge amounts of data. Data mining is useful to extract health care knowledge for clinical decision making and to generate hypothesis from large medical data [3].

Classification is a pervasive problem that is used for many applications and to find unknown sample. Researchers are focusing on designing efficient classification algorithms for large data sets [1].

Classification system will assist physicians to examine a patient, whether patient is likely to have a heart disease or not. Hidden naïve bayes is proposed by langseth and Nielsen in 2005[4]. Hidden parent is created, in hidden naïve bayes for each feature which combines the influences from all other features [5]. Hidden nave bayes demonstrates remarkable performance than other traditional classification algorithms. In data mining applications especially for medical data mining, we need accurate classification.

In medical data mining, misclassification cost should be taken into consideration. Class probability estimation is required for cost sensitive applications [6].

Our motivation is to apply efficient and accurate classification on HNB for prediction of heart disease. The rest of the paper is organized as follows. Related work on heart disease is presented in section 2. Literature review is discussed in section 3.Our novel classification model is presented in section 4.Section 5 will give implementation details. we Conclude in section 6.

II. RELATED WORK

This section reviews some papers related to CHD are reviewed. "Prediction of heart disease using random forest" was proposed in [2]. Authors developed an approach for classification of heart disease by applying feature selection to random forest classification. Chi square measure is used as feature selection to discard redundant features and achieved an accuracy of 83.70%.

M.A. Jabbar et.al proposed "intelligence data mining technique for diagnosis of heart disease" [7]. Authors attempted to increase accuracy for heart disease. They used discretization preprocessing techniques and genetic algorithm applied on naïve bayes classifier.

Heart disease prediction using GA and associative classification is proposed in [1]. Authors developed a decision support system to identify risk of heart disease. Gini index and Z statistics measures are used to filter number of rules generated by associative classification algorithms**.** Classification of disease using ANN and FSS is proposed in [3].

PCA as a feature selection measure reduces no. of attributes in heart disease data set. Their approach eliminates useless and distortive data, their method thus efficiently classifies heart disease.

Real time patient monitoring system using random forest for disease prediction was proposed by S.Sreejith et.al [8].Their proposed system suggests a frame work for measuring various

risk factors using wearable gadget and the measured parameters are transmitted to android smart phone.

Heart disease prediction using nearest neighbor and GA was proposed by M.A. Jabbar et.al in [9].GA is used as feature selection measure. Highest ranked features are selected for building classifier. Their method achieved good accuracy than other approaches.

Different from earlier work on heart disease classification, we propose a novel approach for efficient classification of heart disease diagnosis using hidden naïve bayes which enhances the accuracy of our model.

III. THEORITICAL BACKGROUND

This section discusses terminology that is used in our proposed method. Data mining has attracted great attention due to its strong functionalities [1]. Classification is the most important task in data mining. Many researchers are focusing on designing most accurate and efficient classifiers for emerging applications like Bio- medical, speech recognition, image classification etc. Classification task takes each instance of a data set and maps it to a distinct class. In medical domain, a binary classifier, classify whether a patient is having heart disease or not.

A. Naive bayes(NB)

Classification in data mining is based on identifying extracted patterns, used to classify into healthy individual and heart disease patients.Naïve bayes classifier is a simplest form of Bayesian network classifier based on applying bayes theorem, with strong independence of attributes assumption. A Bayesian classifier maps the features A={a1,a2,---a_n} into C classes {c1,c2--c_n} on a data set D,which consists of instances{E1,E2,En},which can be defined as

$$\mathbf{c(E)=arg\ max\ P(c)P(a1,a2\text{----}a_n|C)} \qquad \mathbf{(1)}$$

With the assumption of independence of attributes, naïve bayes classifiers is defined as

$$P(E|C)=P(a1,a2,\text{----an}|C) = \prod_{1}^{n} P(ai \mid c) \qquad (2)$$

Naïve bayes(NB) classification is the most popular model due to its simplicity, efficiency and good performance on data sets. For data sets where complex attribute dependencies are present, NB does not perform well. NB classier will not produce accurate results for large data sets [10]. In medical domain features and their health conditions are correlated. To overcome the drawbacks of NB, Hidden naïve bayes classifier is proposed.

B. Hidden Naive bayes (HNB)

Hidden naïve bayes is more accurate classification compared to naïve bayes, with respect to attribute dependencies .HNB is equivalent to a Bayesian classifier which avoid the intractable complexity and take the influence from all features into account. In hidden naïve bayes, parent is created for each feature, which integrates the influences from other features. The structure of HNB is shown in figure 1.

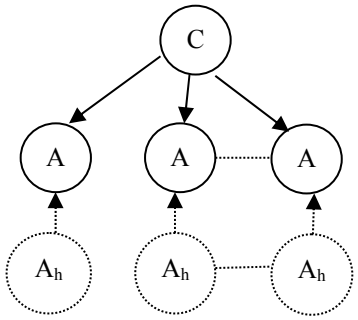

Fig 1: HNB Classifier

Hidden parent in HNB is to combine the weighted influence from all other features. In figure 1, C is class node .This is the parent node for all the features. Each feature Ai has a hidden parent A_{hp1}, A_{hp2}, ----A_{hpn} .

Hidden Naïve Bayes is a structure-extension-based algorithm and needs more training time .Hidden parent in Hidden Naïve Bayes can be seen as aggregating the influences from all other features that are assigned higher weights with higher influences. Due to the increasing need to apply data mining techniques to medical data mining, we applied Hidden naïve bayes for heart disease data set with dependent attributes.

IV. RESEARCH METHOD

In this section, technical aspects of our proposed approach are discussed. Our proposed algorithm is described below.

Algorithm: Heart disease prediction using hidden naïve bayes

Input: Heart disease data set

Output: Classification whether a person is healthy individual or having heart Disease

Step 1: Heart data set is loaded

Step 2: Apply preprocessing filter discretization and inter quartile range **(IQR)**

Step 3: Partition the data sets into training and test set

Step 4: Heart disease data set is trained by HNB

Step 5: The test data set is given to HNB for testing

Step 6: Measure the accuracy of the HNB

Algorithm for hidden naïve bayes(HNB) is as follows

Input: A set of data base

Output: Hidden naïve bayes classifier

Step 1: For each value of c of class C

Step 2: Calculate probabilities P(C) from Database D

Step 3: For attributes A_i and A_j

Step 4: Compute P ($a_i|a_j$, c)from D

Step 5: Compute conditional mutual information

MI=IP(Ai; Aj| C)and weights W_{ij} from D

4.1 Research frame work:

For our experimental analysis we downloaded the heart stalog data set in ARFF from the UCI repository [11].We adopted the following pre-processing techniques to run the experiment.

1. Replace missing values:

We used replace missing values filter to replace all missing feature values. This filter replaces missing values with the mean and mode from the training data.

2. Discretization:

Numeric attributes were discretized by discretization filter using unsupervised 10bin Discretization.

3. Inter-Quartile range filter (IQR):
Inter quartile range (IQR) is a measure of variability. It divides data set into quartiles.
Q1: In a rank ordered data set middle value in first half , Q2: Median value in the data set.
Q3: is the "middle" value in the *second* half of the data set.

IQR= Q3 - Q1.

V. EXPERIMENTAL RESULTS

To apply hidden naïve bayes classifier, we used WEKA 6.4 tool. We ran our experiments on heart data downloaded from UCI [11]. Heart stalog data set consists of 14 attributes and 270 instances. HNB evaluation is performed using 10 fold cross validation. Heart data set information is shown in table 1. Statistics of data set is described in table 2.

Table I: Data set information

Sl. no	Data set	Instances	Number of Attributes
1	Heart disease	270	14

Table II: Heart stalog data Statistics

Data set	Instances cross validation (10 fold)	
Heart disease data set	Test	Train
	27	243

We used accuracy, sensitivity, specificity, positive predictive value measures in our experiment to evaluate the performance of HNB. These performance measures are the most important in medical field and are derived from confusion matrix. Confusion matrix is is used to visualize the performance of an algorithm. It is used to measure incorrect and correct prediction made by the classifier. Table 3 shows Confusion matrix.

Table III: confusion matrix

		Predicted	
		Yes	No
Actual	Yes	TP	FP
	No	FN	TN

Table IV: Accuracy of data set

name of the data set	**Method**	**Accuracy**
Heart stalog	HNB+IQR	100

TP=> True positive, TN=> True Negative, FN=> False negative FP=> False positive
According to confusion matrix following classification measures are defined [2]
1) Sensitivity= TP/FN+TP 2) Specificity=TN/FP+TN
3) Accuracy= (TN+TP)/(TN+FN TP+FP)
4) Positive predictive value (PPV) = TP/FP+TP
5) Negative predictive value (NPV)= TN/FN+TN
6) True positive rate (TPR) = TP/FN+TP
7) False positive rate (FPR) =FP/FP+TN

PPV calculates probability that the disease is present, when the s test is positive.
NPV calculates the probability that disease is absent, when the test is negative.

Table 4 records accuracy of HNB classifier on heart stalog data set.
Table 5 presents the result of HNB and NB.
Table 6 shows accuracies obtained by various approaches on heart stalog data set.

Table V: Results for heart disease data set for various

MEASURE	HNB		NB	
	Cross validation	Full training	Cross validation	Full training
Sensitivity	90.56	92.5	79.16	80
Specificity	86	96	85.6	87.85
Accuracy	83.3	94.4	83.70	85.18
PPV	82.0	94.8	83	85
NPV	81	94.1	88	84
TPR	80	92.5	79	80.83
FPR	14	4	14	12.1

Table VI: Accuracies obtained by various approaches

Sl.no	Author	Method	Accuracy
1	Peter john [12]	CFS+NB	85.18
2	Maishouman[13]	NB	84.5
3	Rupali [14]	NB	78
4	Elma [15]	KNN+NB	82.96
5	**Our approach**	**HNB+IQR**	**100**

Sensitivity, specificity and accuracy of HNB and NB is shown in figure 2.

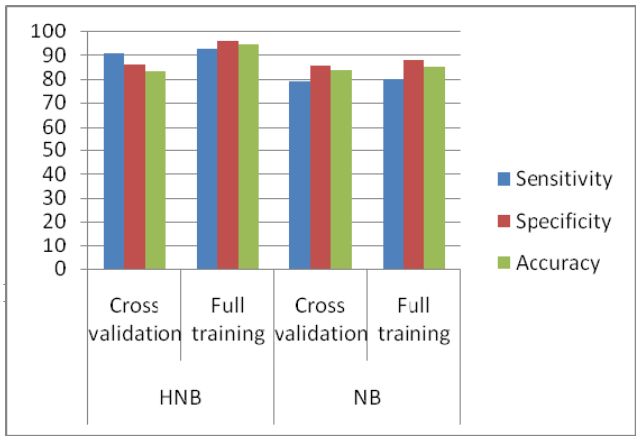

fig.2 sensitivity and accuracy of proposed approach

PPV calculates probability of a disease when it is present, when the test result is +ve, whereas NPV calculates the probability that disease is absent, when the test result is -ve.
PPV and NPV Values measured for HNB and NB are shown in figure 3.

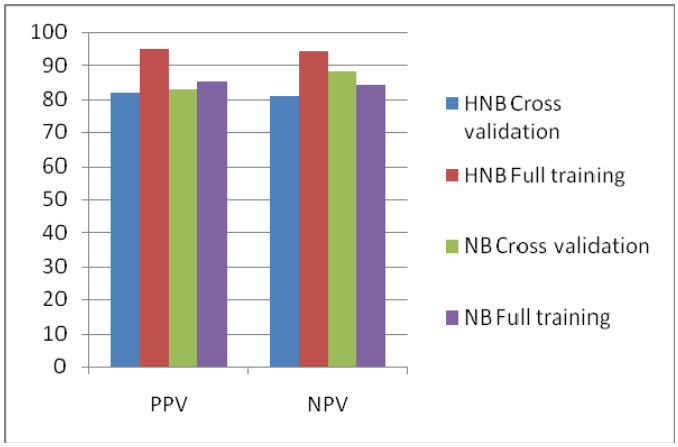

Fig 3.PPV and NPV calculation for HNB and NB

It is observed that performance of proposed model is superior performance than other model discussed in this research paper. Highlights of proposed model are summarized below.
1) Accuracy obtained by our model is 100%.to the best of our knowledge this is the highest accuracy obtained by other models reported in the literature.
2) Accuracy of proposed approach is 14.82% higher than naïve bayes approach, with CFS Feature selection.
3) HNB based model obtained higher accuracy without feature selection than other approach.
4) Proposed model overcomes the shortfalls of NB classifier to classify heart disease
5) Sensitivity of proposed model is much higher than NB.
6) True positive rate of proposed model is considerably higher than NB.

Considering the overall performance of proposed model with other models, this model is best suited as an intelligent DSS for heart disease diagnosis.

VI. CONCLUSION

In this paper, we applied HNB classifier for diagnosis of heart disease. We applied HNB and tested performance for heart stalog data set. Experimental result shows that HNB model exhibits a superior performance compared with other Approaches. Proposed approach applies discretization and IQR filters to improve the efficiency of Hidden naïve bayes. Proposed model recorded highest accuracy (100%) compared with NB classification model. HNB model helps reliable decision support system (DSS) for automatic diagnosis of disease.

As a performance result HNB classifier is a promising model for medical data sets like heart disease with dependent attributes for diagnosis of disease.

References

[1] M.A.Jabbar,"Heart Disease Prediction System using Associative Classification and Genetic Algorithm", ICECIT, pp 183-192, Elsevier, vol 1(2012)

[2] M.A.Jabbar et.al" Prediction of heart disease using Random forest and Feature subset selection ",AISC SPRINGER, vol 424,pp187-196(2015)

[3] M.A.Jabbar, " Classification of heart disease using artificial neural network and feature subset selection", GJCST, Volume 13 Issue 3 , Ver 1.0, 15-25,(2013)

[4] Helge Langseth," Classification using hierarchical naive Bayesian models", Machine learning, pp 135-159,63(2)(2006)

[5] Liangxiao, jiang, harry zhang, zhihua cai, "novel bayes model: Hidden naïve bayes" IEEE Transactions on knowledge and data engineering",vol 21,no 10, (2009)

[6] Domingos," A general method for making classifiers" Proc of fifth Intl. Conf, KDD, pp 155-164 (1999)

[7] M.A. Jabbar," Computational intelligence technique for early diagnosis of heart disease", ICETECH 2015, IEEE, pp1-6(2015)

[8] Sreejith,Rahul jisha," patient monitoring system for disease prediction using random forest", AISC, pp 485-500(2016)

[9] M.A.Jabbar," classification of Heart Disease Using K-Nearest Neighbor and Genetic lgorithm ", Procedia technology, pp 85-94 (2013)

[10]] Yaguang J,Songnan, yafend,"A novel bayes model: Package hidden naïve bayes" ,ITAIC, pp 20-22(2011)

[11] Kohavi," Scaling up the accuracy of NB classifiers: A decision tree hybrid", In Proc.KDD 1996, pp 202-207(1996)

[12] archieve.ics.uci.edu/ml/data sets

[13] peter John ,"Study and development of FSS for disease prediction", IJSRP, Vol 2,Issue 10,(2012)

[14] shouman ," Integrating NB and K-Means for disease diagnosis", CSCP, pp 125-137(2012)

[15] Rupali "Heart disease prediction using NB and JM Smoothing", IJARCCE,pp 6787-6792(2014)

[16]Elma ," Combination of NB and K-NN in the predictive models", CIS, Vol 6, No 3, pp 48-56(2013)