ISSN: 2321-7782 (Online)
Volume 3, Issue 8, August 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

The Comparative Study for Diagnosing Heart Disease Using KNN and Naïve Bayes

V.KaliyaMeiyar[1]
Research Scholar,
Department of Computer Science,
H.H. The Rajah's College (Autonomous)
India

D.Shanmugasundaram[2]
Assistant Professor,
Department of Computer Science,
H.H. The Rajah's College (Autonomous)
India

Abstract: ***Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. K-Nearest Neighbor (KNN) and Naive Bayes (NB) two of the successful data mining techniques used in classification problems. However, it is less used in the diagnosis of heart disease patients. Recently, researchers are showing that combining different classifiers through voting is outperforming other single classifiers. This research work investigates applying KNN and NB to help healthcare professionals in the diagnosis of heart disease. It also investigates if integrating voting with two classifiers can enhance its accuracy in the diagnosis of heart disease patients. The results show that applying the classifiers could achieve higher accuracy than neural network ensemble in the diagnosis of heart disease patients. When applying voting to decision tree, a different decision rules are extracted from each subset, which helps in extracting new knowledge that could increase the decision tree accuracy. In the case of KNN, the distance between the testing instance and different datasets is measured. However, there is no new knowledge extracted, just the distance is measured. In result shows both classifiers will perform in their area and produce accuracy rate using large dataset.***

Keywords: KNN, Naïve Bayes, Diagnosing Heart Disease, Data Mining

I. INTRODUCTION

This Data mining is rapidly growing in wide range of applications. One of the important data mining fields is medical data mining. There is a wealth of data available in healthcare but there is no effective analysis tool to discover hidden relationships in data. Although millions of people die of heart disease annually, application of data mining techniques in heart disease diagnosis seems to be essential. Discovered Knowledge can help physicians in diagnosis of heart disease.

Data mining is the heart (core) step, which results in the discovery of implicit but potentially valuable knowledge from huge amount of data. Data mining technology provides the user with the methods to find new and implicit patterns from massive data. In the healthcare domain, discovered knowledge can be used by the healthcare administrators and medical physicians to improve the accuracy of diagnosis, to enhance the goodness of surgical operations and to reduce the harmful effects of drug It aims also to propose less expensive therapeutic.

Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. K-Nearest-Neighbors (KNN) is one of the successful data mining techniques used in classification problems. However, it is less used in the diagnosis of heart disease patients. Recently, researchers are showing that combining different classifiers through voting is outperforming other single classifiers. This paper investigates applying KNN to help healthcare professionals in the diagnosis of heart disease. It also investigates if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients. The results show

that applying KNN could achieve higher accuracy than neural network ensemble in the diagnosis of heart disease patients. The results also show that applying voting could not enhance the KNN accuracy in the diagnosis of heart disease.

The knowledge data is classified by using different classification algorithms such as Naive Bayes, K-Nearest Neighbor, Decision Tree and the accuracy of each classification algorithm is noted. From these algorithm, NBperforms, better than other methods for heart disease classification. Medical decision support systems are designed to support clinicians in their diagnosis for heart disease. They typically work through an analysis of medical data and a knowledge base of clinical experts. The quality of medical decisions for heart disease can be increased by using Bayesian algorithm.

II. RELATED WORK

Boshra Bahrami, et al The objective of this paper is to evaluate different classification techniques in heart disease diagnosis. Classifiers like J48 Decision Tree, K Nearest Neighbors(KNN), Naive ayes(NB), and SMO are used to classify dataset. After classification, some performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. The comparison results show that J48 Decision tree is the best classifier for heart disease diagnosis on the existing dataset. In this study four different classification algorithms applied on existing heart disease dataset. The Gain Ratio evaluation technique used as feature selection technique and four features extracted from dataset. Then preprocessed datasets, used to test the four classifiers using 10-folds cross validation. Six different performance measures considered for classifiers. Results of comparison showed that j48 decision tree achieved the highest value in accuracy, sensitivity, specificity, F-measure and precision performance measures. The optimum heart disease predictive model obtained in this study, adopts j48 decision tree as classification algorithm. Hence, it is a suitable candidate for testing in a clinical environment and implementing in decision support systems for helping physicians and healthcare professionals in diagnosis of heart disease[1]

Hardik Maniya et al made an effort to predict the most widely spread disease in India named tuberculosis. Using data collected from various TB centers, we made an effort to fetch out hidden patterns and by learning this pattern through the collected data for tuberculosis we can diagnose and predict the disease. In the research work we are comparing naïve bayes classifier and KNN, two the most effective techniques for data classification (especially for medical diagnoses), implemented using C language and using Weka tool respectively and classify the patient affected by tuberculosis into two categories (least probable and most probable). We have used 19 symptoms of tuberculosis and collect 154 cases. We have achieved nearly 78% accuracy with low false negative. This algorithm extracts hidden patterns from available TB database. Naïve Bayes could identify all the significant medical predictors. The prototype can further be improved by incorporating various other attributes and increasing the number of cases for training and testing. The efficiency of results using KNN can be further improved by increasing the number of data sets and for Naïve Bayesian classifier by increasing attributes or by selecting weighted features. [2]

M.Akhil jabber et al Nearest neighbor (KNN) is very simple, most popular, highly efficient and effective algorithm for pattern recognition.KNN is a straight forward classifier, where samples are classified based on the class of their nearest neighbor. Medical data bases are high volume in nature. If the data set contains redundant and irrelevant attributes, classification may produce less accurate result. Heart disease is the leading cause of death in INDIA. In Andhra Pradesh heart disease was the leading cause of mortality accounting for 32%of all deaths, a rate as high as Canada (35%) and USA. Hence there is a need to define a decision support system that helps clinicians decide to take precautionary steps. In this paper we propose a new algorithm which combines KNN with genetic algorithm for effective classification. Genetic algorithms perform global search in complex large and multimodal landscapes and provide optimal solution. Experimental results shows that our algorithm enhance the accuracy in diagnosis of heart disease. In this paper we have presented a novel approach for classifying heart disease. As a way to validate the proposed method, we have tested with emphasis on heart disease on A.P besides other machine learning data sets taken from UCI repository. Experimental results carried out on 7 data sets show that our approach is a competitive method

for classification. This prediction model helps the doctors in efficient heart disease diagnosis process with fewer attributes. Heart disease is the most common contributor of mortality in India and in Andhra Pradesh. Identification of major risk factors and developing decision support system, and effective control measures and health education programs will decline in the heart disease mortality. [3]

Prachi Jambhulkar et al The reduction in supply of blood and oxygen to the heart leads to various heart diseases. Hence, the main objective of our paper is to predict more accurately the percentage of possibility of heart disease with minimum number of attributes. For this we use data mining technique along with wireless sensor network, we can collect different data records from Cleveland database and extract essential attributes required for heart disease prediction using data mining technique This work proposes a wireless sensor network design for real-time monitoring and detection of cardiovascular diseases. This proposed system consists of a wearable wireless sensor system, data mining software system and the warning system. All these together will provide the percentage of possibility of heart disease in a cardiac patient and generate a warning for doctor and patient as well. This system can use for providing enhanced healthcare services to cardiac patient. Thus, the early diagnosis of heart disease detection may reduce the chances of death in cardiac [4]

III. Classification Analysis

All Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. Classification is considered an instance of supervised learning i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

Among several types of knowledge representation present in the literature, classification normally uses prediction rules to express knowledge. Prediction rules are expressed in the form of IF-THEN rules, where the antecedent (IF part) consists of a conjunction of conditions and the rule consequent (THEN part) predicts a certain predictions attribute value for an item that satisfies the antecedent. Using the example above, a rule predicting the first row in the training set.

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. A bank loan officer wants to analyse the data in order to know which customers (loan applicant) are risky or which are safe. A marketing manager at a company needs to analyse a customer with a given profile, who will buy a new computer. In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

The Data Classification process includes two steps. Building the Classifier or Model, Using Classifier for Classification This step is the learning step or the learning phase. In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

a) Building the Classifier

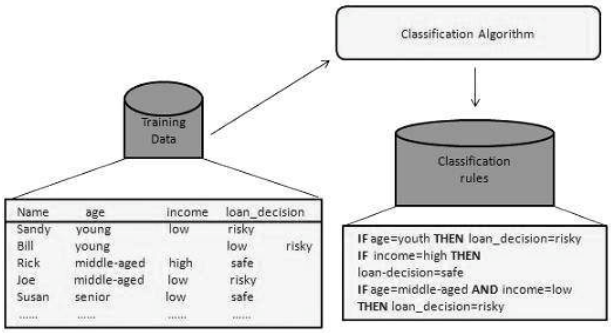


Figure 1: Building the Classifier

b) Using Classifier for Classification

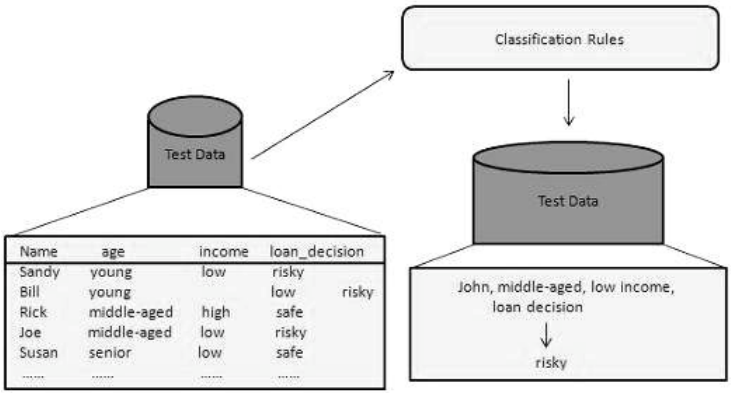


Figure 2: Using Classifier

K Nearest Neighbor (KNN from now on) is one of those algorithms that are very simple to understand but works incredibly well in practice. Also it is surprisingly versatile and its applications range from vision to proteins to computational geometry to graphs and so on. Most people learn the algorithm and do not use it much which is a pity as a clever use of KNN can make things very simple. KNN is a *non parametric lazy learning* algorithm. That is a pretty concise statement. When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made (eg Gaussian mixtures, linearly separable etc). The dichotomy is pretty obvious here – There is a nonexistent or minimal training phase but a costly testing phase. The cost is in terms of both time and memory. More time might be needed as in the worst case, all data points might take point in decision. More memory is needed as we need to store all training data

Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either + or – (for positive or negative classes). But KNN, can work equally well with arbitrary number of classes. We are also given a single number "k". This number decides how many neighbours (where a neighbour is defined based on the instance metric) influence the classification. This is usually a odd number if the number of classes is 2. If k=1 , then the algorithm is simply called the nearest neighbour algorithm.

Here is step by step on how to compute K nearest neighbors KNN algorithm:-

1. Determine parameter K = number of nearest neighbors
2. Calculate the distance between the query instance and all the training samples.
3. Sort the distance and determine nearest neighbors based on the Kth minimum distance.
4. Gather the category of the nearest neighbors.
5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

The Naïve Bayes Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)}$$

IV. METHODOLOGY USED

A. Supervised Learning

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

Supervised learning is the most common technique for training neural networks and decision trees. Both of these techniques are highly dependent on the information given by the pre-determined classifications. In the case of neural networks, the classification is used to determine the error of the network and then adjust the network to minimize it, and in decision trees, the classifications are used to determine what attributes provide the most information that can be used to solve the classification puzzle. We'll look at both of these in more detail, but for now, it should be sufficient to know that both of these examples thrive on having some "supervision" in the form of pre-determined classifications.

Notice something important here: in the classification problem, the goal of the learning algorithm is to minimize the error with respect to the given inputs. These inputs, often called the "training set", are the examples from which the agent tries to learn. But learning the training set well is not necessarily the best thing to do. For instance, if I tried to teach you exclusive-or, but only showed you combinations consisting of one true and one false, but never both false or both true, you might learn the rule that the answer is always true. Similarly, with machine learning algorithms, a common problem is over-fitting the data and essentially memorizing the training set rather than learning a more general classification technique.

As you might imagine, not all training sets have the inputs classified correctly. This can lead to problems if the algorithm used is powerful enough to memorize even the apparently "special cases" that don't fit the more general principles. This, too, can lead to over fitting, and it is a challenge to find algorithms that are both powerful enough to learn complex functions and robust enough to produce generalizable results.

B. Unsupervised Learning

In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning.

Unsupervised learning seems much harder: the goal is to have the computer learn how to do something that we don't tell it how to do! There are actually two approaches to unsupervised learning. The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Note that this type of training will

generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards. This approach nicely generalizes to the real world, where agents might be rewarded for doing certain actions and punished for doing others.

A second type of unsupervised learning is called clustering. In this type of learning, the goal is not to maximize a utility function, but simply to find similarities in the training data. The assumption is often that the clusters discovered will match reasonably well with an intuitive classification. For instance, clustering individuals based on demographics might result in a clustering of the wealthy in one group and the poor in another.

Although the algorithm won't have names to assign to these clusters, it can produce them and then use those clusters to assign new examples into one or the other of the clusters. This is a data-driven approach that can work well when there is sufficient data; for instance, social information filtering algorithms, such as those that Amazon.com use to recommend books, are based on the principle of finding similar groups of people and then assigning new users to groups. In some cases, such as with social information filtering, the information about other members of a cluster (such as what books they read) can be sufficient for the algorithm to produce meaningful results. In other cases, it may be the case that the clusters are merely a useful tool for a human analyst. Unfortunately, even unsupervised learning suffers from the problem of over fitting the training data. There's no silver bullet to avoiding the problem because any algorithm that can learn from its inputs needs to be quite powerful.

C. KNN

KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time when dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance. If the first instance is (a1, a2, a3 … an) and the second instance is (b1, b2, b3… bn), the distance between them is calculated by the following formula:

$$\sqrt{\left(a_1 - b_1\right)^2 + \left(a_2 - b_2\right)^2 + \cdots (a_n - b_n)^2}$$

A major problem when dealing with the Euclidean distance formula is that the large values frequency swamps the smaller ones. For example, in heart disease records the cholesterol measure ranges between 100 and 190 while the age measure ranges between 40 and 80. So the influence of the cholesterol measure will be higher than the age. To overcome this problem the continuous attributes are normalized so that they have the same influence on the distance measure between instances.

KNN usually deals with continuous attributes however it can also deal with discrete attributes. When dealing with discrete attributes if the attribute values for the two instances a2, b2 are different so the difference between them is equal to one otherwise it is equal to zero.

D. Naïve Bayes Algorithm

One of the Bayesian methods is Naïve Bayes classifiers which uses the probabilistic formula:

$$\mathrm{P(A/B)} = \frac{\mathrm{P(B/A)} \times \mathrm{P(A)}}{\mathrm{P(B)}}$$

Where A and B are two events (e.g. the probability that the train will arrive on time given that the weather is rainy). Such Naïve Bayes classifiers use the probability theory to find the most likely classification of an unseen (unclassified) instance. The algorithm performs positively with categorical data but poorly if we have numerical data in the training set.

E. Naive Bayes Classifier

NB's main strength is its efficiency; it combines efficiency with good accuracy it is often used as a baseline in text classification research.

1. Classifier Module:-

Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the variable values necessary for classification. Classification problem is given to the classifier as combination of different values of chosen variables with related value of class variable; the classifier then returns a posterior probability distribution over the class variable.

Probability that a training pattern with attribute array A belongs to class CK, also known as the posterior probability is given Bayes theorem of probability,

$$(C_k/A) = P(C_k) * P(A/C_k)/P(A)$$

Where,

A is an array of M >= 1 attributes A1 ,A2 ,….,AM for the patterns of a training set.

P (CK) is the probability that a training pattern belongs to class A, also called prior probability.

P (CK|A) is also called the posterior probability because it is probability of training pattern with attribute array B belongs to class Ck

P (A|CK) is the conditional probability of B given A. It is also called the likelihood shows probability of class A has attribute array B.

P (A) is the probability that a training pattern has attribute array B, regardless of the class to which the pattern belongs.

2. Learning Module:-

A set of 50 cases was taken and the program was trained with these data sets such that the probabilities of all the classes with all the conditions were calculated. Result was stored in database and when the test data was given we got the probabilities for the various classes for the given symptom values on the basis of which we inferred that the patient fell into the class with the highest probability. This is what is called the Naïve Bayes classification. This is a very powerful technique that is instrumental in helping us predict the category a patient falls into.

V. Experimental Results

Accuracy Calculation Accuracy refers to the percentage of correct predictions made by the model compared with actual classifications in the test data. Accuracy = Total no. of Correctly Predicted Record / Total no. of training Record

In this experiment the medical data related to Heart Diseases is considered. This dataset was obtained from Cleveland database. This is publicly available dataset in the Internet. Cleveland dataset concerns classification of person into normal and abnormal person regarding heart diseases. Data Representation: Number of instances: 414. Number of attributes : 13 and a class attribute

Class:

Class0: Normal Person.

Class1: first stroke Class2: second stroke

Class3: end of life Attribute

Attribute Name	Type	Description
Age	Continuous	Age in Years
Gender	Discrete	1=male 0=female
Cp	Discrete	Chest Pain Type: 1=typical angina 2=atypical angina 3=non-angina pain 4=asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting Blood Sugar > 120 mg/dl: 1=true 0=false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes 'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

Table 1: Cleveland dataset

a) Description of Cleveland Dataset

This dataset contains information concerning heart disease diagnosis. The data was collected from the Cleveland clinic foundation, and it is available at UCI Repository. Six instances containing missing values have been deleted from the original dataset.

Format: A data frame with 303 observations on the following 14 parameters: P1 - Age P2 - Gender P3 – CP (chest pain) P4 - trestbps : resting blood pressure P5 – cholesterol P6 – fbs: fasting blood sugar>120 ? yes=1,no = 0 P7 – restecg: resting electrocardiographic results 0,1,2 P8 – thalach : maximum heart rate achieved P9 – exang : exercise induced angina (1= yes ; 0= no) P10 – oldpeak = ST depression induced by exercise relative to rest P11 – slope : the slope of the peak exercise ST segment P12 – ca: no. of major vessels (0 to 3) colored flurosopy P13 – thal :3 =normal ,6=fixed defect ,7= reversable defect P14 – diagnosis of heart disease

In this study, MATLAB that is a powerful data mining tool, was used to apply the data mining algorithms. In this section experimental results from implementation of selected classification algorithms, decision tree, Naive Bayes, KNN on heart disease dataset are analyzed and compared. After comparison, results showed that the best classification accuracy is 80% that achieved by Naïve Bayes. Table shows classification efficiency indicator values of models constructed with three algorithms. Comparison of accuracy that is achieved by three classifiers is shown in Figure 1. Although accuracy is the most common measure in classification performance, other important performance measures such as sensitivity, Specificity, F-Measure, precision and ROC indicators considered to evaluate and compare classification efficiency of three selected algorithms. Figure 2, shows the Comparison of sensitivity, specificity, precision, F-measure and area under Roc curve, achieved by four classification algorithms on existing heart disease dataset.

Data Set	KNN	Naive Bayes
K=1	60.7407	66.6667
K=2	61.8519	85.1852
K=3	64.0741	77.7778
K=4	67.4074	81.4815
K=5	68.5185	85.1852
K=6	67.037	85.1852
K=7	65.9259	74.0741
K=8	65.1852	77.7778
K=9	64.4444	88.8889
K=10	63.3333	74.0741
Accuracy Rate	**64.8519**	**79.6296**

Table 2: Results

b) Classifier Performance Evaluation

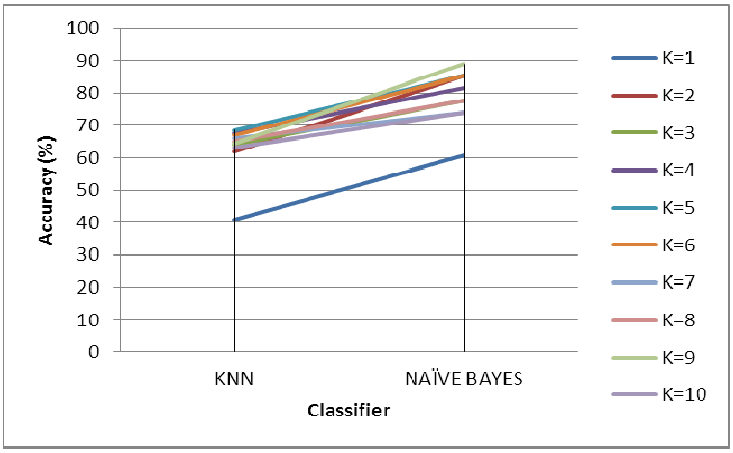


Figure 3: Accuracy Rate for Two Classifiers with Different Iteration

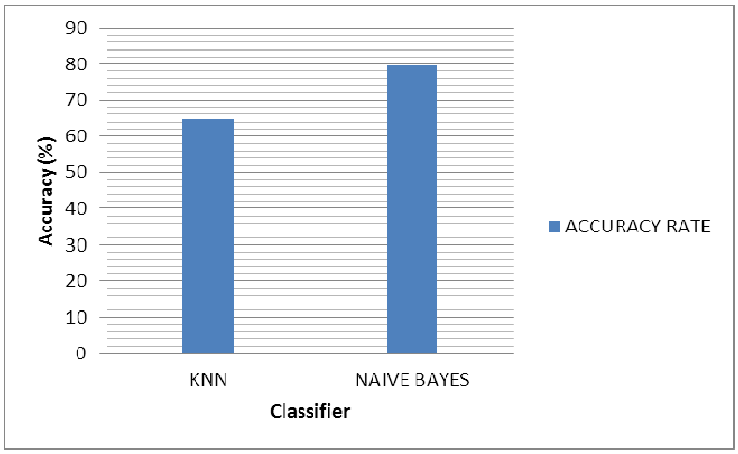
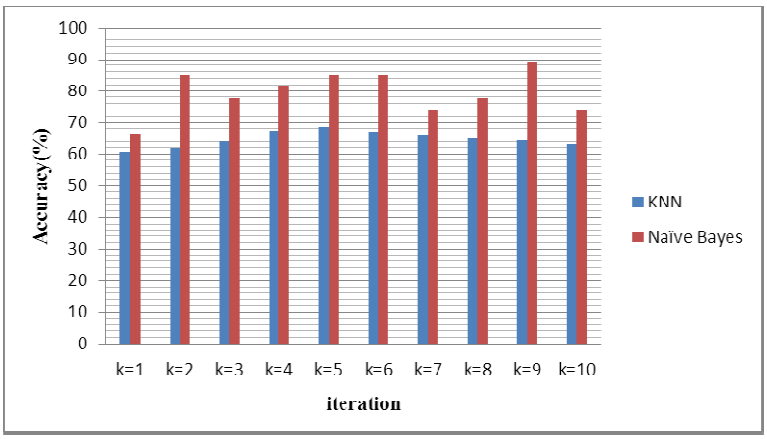


Figure 4: Overall Accuracy Rate

c) Performance Analysis

Heart disease is the leading cause of death all over the world in the past ten years. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data that could be used to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease. In this research work we report research that applied decision tree, KNN and Naïve bayes on a benchmark dataset to investigate its efficiency in the diagnosis of heart disease. We also investigated if integrating voting with KNN with others could enhance its accuracy even further. Our results show that applying Naïve bayes achieved an accuracy of 80% which is higher than any other published findings on that benchmark dataset. The results also show that applying voting could not enhance the three algorithms accuracy in the diagnosis of heart disease. Of course, while Naïve bayes has produced excellent results, the work needs to be verified against other and larger datasets.



VI. CONCLUSION

In this study four different classification algorithms applied on existing heart disease dataset. The Gain Ratio evaluation technique used as feature selection technique and four features extracted from dataset. Then preprocessed datasets, used to test the three classifiers using 10-folds cross validation. Six different performance measures considered for classifiers. Results of comparison showed that naïve bayes achieved the highest value in accuracy, sensitivity, specificity, and precision performance measures. The optimum heart disease predictive model obtained in this study, adopts naïve bayes as classification algorithm. Hence, it is a suitable candidate for testing in a clinical environment and implementing in decision support systems for helping physicians and healthcare professionals in diagnosis of heart disease.

In future we investigate to combine each and every algorithm to other and get the accuracy rate for combining algorithm to measure the accuracy on heart disease dataset. Also we use a different dataset to measure the accuracy of prediction.

References

1. Prediction and Diagnosis of Heart Disease by Data Mining Techniques by Boshra Bahrami, Mirsaeid Hosseini Shirvani
2. Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis by Hardik Maniya, Mosin I. Hasan,Komal, P. Patel
3. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm by M.Akhil jabbar , B.L Deekshatulua, Priti Chandra
4. Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network by Prachi Jambhulkar, Vaidehi Baporikar
5. Backpropogation neural network for prediction of heart disease byNABEEL AL-MILLI
6. J. Han and M. Kamber, "Data Mining Concepts and Techniques," Morgan Kaufmann Publishers, 2006.
7. I.-N. Lee, S.-C. Liao, and M. Embrechts, "Data mining techniques applied to medical information," Med. inform, 2000.
8. M. K. Obenshain, "Application of Data Mining Techniques to Healthcare Data," Infection Control and Hospital Epidemiology, 2004.
9. J. Sandhya et al., "Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques," International Journal of Engineering and Technology, vol.2, no. 4, 2010.
10. B. Thuraisingham, "A Primer for Understanding and Applying Data Mining," IT Professional IEEE, 2000.
11. D. Ashby and A. Smith, The Best Medicine? Plus Magazine - Living Mathematics, 2005.
12. S. C. Liao and I. N. Lee, "Appropriate medical data categorization for data mining classification techniques," MED. INFORM, vol. 27, no. 1, pp. 59–67, 2002.
13. T. Porter and B. Green, "Identifying Diabetic Patients: A Data Mining Approach," Americas Conference on Information Systems, 2009.
14. S. Panzarasa et al, "Data mining techniques for analyzing stroke care processes," in Proc. of the 13th World Congress on Medical Informatics, 2010.

