International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur, Tamilnadu, India.

AN EFFICIENT APPROACH FOR CONSTRUCTING A MODEL FOR DIAGNOSING HEART DISEASE DATASET

V.Hemalatha,
Research Scholar,
M.Phil. Computer Science,
Vellalar College for Women (Autonomous),
Erode,Tamilnadu, India.

C.Usha nandhini,
Assistant Professor,
Department of Computer Applications,
Vellalar College for Women (Autonomous),
Erode, Tamilnadu, India.

Abstract: Classification of coronary heart disease can be valuable for the medical practitioners in the event that is automated with the end goal of quick finding and exact result. Recently, researchers have used different classification and clustering algorithm for diagnosing diseases. The work incorporates the classes of heart disease utilizing K-Nearest Neighbor and Naive Bayes. In this work, we have analyzed the use of K-Nearest Neighbor and Naive Bayes with different normalization techniques. The dataset utilized is the Cleveland heart disease from UCI machine learning repository. Our proposed works analyze the performance of K-Nearest Neighbor and Naive Bayes classification. The results prove that Naive Bayes classification gives better accuracy for diagnosing heart disease.

Keyword: Data Mining, Heart Disease, Classification Techniques

I. INTRODUCTION

The experimental and iterative nature of knowledge discovery is most apparent during data mining and interpretation and evaluation of the knowledge discovery process. Here is a typical scenario for building a supervised or unsupervised learner model:

- Choose training and test data from the pool of available instances.
- Designate a set of input attributes.
- If learning is supervised, choose one or more attributes for output.
- Select values for the learning parameters.
- Invoke the data mining tool to build a generalized model of the data.

Once data mining is complete, the model is evaluated. If an acceptable result is not seen, the just-described steps may be repeated several times. Because of this, the total number of possible learner models created from one set of data is infinite. Fortunately, the nature of the experimental process combined with the fact that data mining techniques are able to create acceptable models with less than perfect data increases our likelihood for success.

- **HEART DISEASE**

The heart is important organ or part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

There are number of factors which increase the risk of Heart disease:

- Family history of heart disease
- Smoking
- Cholesterol
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

- **SYMPTOMS OF HEART ATTACK**

It includes:

- Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone.
- Discomfort radiating to the back, jaw, throat, or arm.
- Fullness, indigestion, or choking feeling (may feel like heartburn).
- Sweating, nausea, vomiting, or dizziness.
- Extreme weakness, anxiety, or shortness of breath.
- Rapid or irregular heartbeats

II.DATA MINING TECHNIQUES: CLASSIFICATION

Classification is one of the Data Mining techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class such that classification error will be least. It is used to extract models that accurately define important data classes within the given dataset. Classification is a two-step process. During first step the model is created by applying classification algorithm on training data set then in second step the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So classification is the process to assign class label from dataset whose class label is unknown.

➢ K-NEARESTNEIGHBOUR (KNN) ALGORITHM

The nearest neighbor (NN) rule identifies the category of unknown data point on the basis of its nearest neighbor whose class is already known. This rule is widely used in pattern recognition text categorization ranking models object recognition and event recognition applications. M. Cover and P. E. Hart purpose knearest neighbour (kNN) in which nearest neighbor is calculated on the basis of value of k that specifies how many nearest neighbors are to be considered to define class of a sample data point. It makes use of the more than one nearest neighbour to determine the class in which the given data point belongs to and hence it is called as K-NN. These data samples are needed to be in the memory at the run time and hence they are referred to as memory-based technique. T. Bailey and A. K. Jain improve kNN which is based on weights

The training points are assigned weights according to their distances from sample data point. But still, the computational complexity and memory requirements remain the main concern always. To overcome memory limitation, size of data set is reduced. For this, the repeated patterns, which do not add extra information, are eliminated from training samples .To further improve, the data points which do not affect the result are also eliminated from training dataset . The NN training data set can be structured using various techniques to improve over memory limitation of kNN. The kNN implementation can be done using ball tree, k-d tree, nearest feature line (NFL), tunable metric, principal axis search tree and orthogonal search tree. The tree structured training data is divided into nodes, whereas techniques like NFL and tunable metric divide the training data set according to planes. These algorithms increase the speed of basic kNN algorithm. Suppose that an object is sampled with a set of different attributes, but the group to which the object belongs is unknown. Assuming its group can be determined from its attributes; different algorithms can be used to automate the classification process. With the k-nearest neighbour technique, this is done by evaluating the k number of closest neighbors In pseudo code, k-nearest neighbor classification algorithm can be expressed,

K ← number of nearest neighbors
For each object *X* in the test set **do**
calculate the distance D(*X*,*Y*) between *X* and every object *Y* in the training set
neighborhood ← the *k* neighbors in the training set closest to *X*
X.class ← SelectClass (neighborhood)
End for

The k-nearest neighbors‟ algorithm is the simplest of all machine learning algorithms. It has got a wide variety of applications in various fields such as Pattern recognition, Image databases, Internet marketing, Cluster analysis etc In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes. Here a single number k‟ is given which is used to determine the total number of neighbors that determines the classification. If the value of k=1, then it is simply called as nearest neighbour. K-NN requires an integer k, a training data set and a metric to measure closeness.

➢ NAIVE BAYES ALGORITHM

The Naive Bayes classifier provides a simple approach, with clear semantics, representing and learning probabilistic knowledge. It is termed naïve because is relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it assumes that no hidden or latent attributes influence the prediction process.

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes.

The Bayes theorem is as follows:

Let X={x1, x2, ...,xn} be a set of n attributes.

In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C.

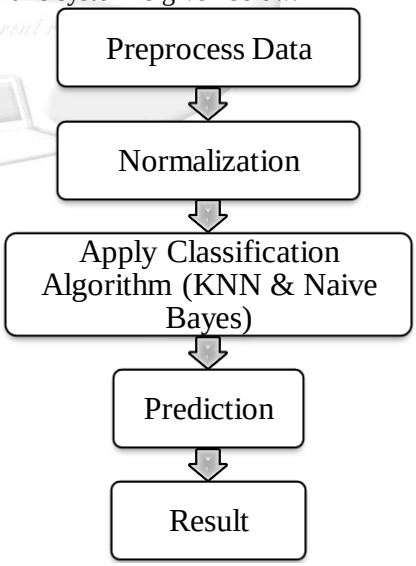
We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X.

According to Bayes theorem the P (H|X) is expressed as P (H|X) = P (X| H) P (H) / P (X)

III. PROPOSED METHOD

Dataset of Cleveland is collected from UCI repository. There are 14 attributes in which the last attribute is the class. Remaining 13 attributes namely Age, Sex, Cp (chest pain), Trestbps (resting pulse), cholestoral and blood sugar in fasting, resting ECG value, thalach – (maximum heart rate achieved), exang (exercise induced angina), oldpeak rate, slope value, ca (number of major vessels (0-3) colored by flourosopy) and thal.

An expert system is created to predict heart disease. The workflow of this system is given below:



Steps of Proposed System

Step 1: For analyzing healthcare data, major steps of data mining approaches like preprocessing data, normalization are applied on train dataset. For preprocessing ignore missing value tuple and for normalization, min-max and z-score techniques are applied.

Step 2: KNN and Naïve Bayes algorithm have executed on preprocessed train dataset and which is the model to classify the test data.
Step 3: Analyze the results of two classification algorithm.

Performance Evaluation:
Performance evaluation is carried out by accuracy calculation as which is the ratio of the number of correctly classified instances to the total number of instances of the test data.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100\%$$

Where,
TP, FP, TN and FN are the number of true positive, false positive, true negative and false negative respectively.

Result and Discussion:
The outcomes of clinical decision support system for diabetes disease prediction are presented in this section. It trains and tests this expert system using 90:10 percentage split method to evaluate the classification result. Here randomly chosen 290 instances (about 90%) of original dataset are used for training the system and 90 instances (about 10%) are used for testing purpose. This prediction system has been developed using two well known algorithm i.e. KNN and Naïve Bayes algorithm. The performance of these two algorithms is described below:

1) Performance evaluation of KNN algorithm:
In this table1, the prediction result shows that KNN with Z-score classifier has correctly classified 85 instances and incorrectly classified 5 instances. The accuracy of correctly classified instance is 94.44% and incorrectly classified instance is 5.55%. KNN with Min-Max classifier has correctly classified 83 instances and incorrectly classified 7 instances. The accuracy of correctly classified instance is 92.22% and incorrectly classified instance is 7.77%.

Table1: Classification Summary of KNN Classifier

KNN	Prediction Result		Accuracy	
	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances
Training Instances:290 Testing Instances:90 (Z-score normalization)	85	5	94.44%	5.55%
Training Instances:290 Testing Instances:90 (Min-Max normalization)	83	7	92.22%	7.77%

2) Performance evaluation of Naïve Bayes algorithm:
In this table2, the prediction result shows that Naïve Bayes with Min-Max classifier has correctly classified 88 instances and incorrectly classified 2 instances. The accuracy of correctly classified instance is 97.77% and incorrectly classified instance is 2.22%. Naïve Bayes with Z-score classifier has correctly classified 90 instances and incorrectly classified null instances. The accuracy of correctly classified instance is 100% and incorrectly classified instance null.

Table 2: Classification Summary of Naïve Bayes Classifier

Naïve Bayes	Prediction Result		Accuracy	
	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances
Training Instances:290 Testing Instances:90 (Min-Max normalization)	88	2	97.77%	2.22%
Training Instances:290 Testing Instances:90 (Z-score normalization)	90	-	100%	-

3) Comparison of classification accuracy:
The performance of proposed clinical expert system was analyzed with heart disease dataset using KNN and Naïve Bayes. According to experimental results, Table3 represents the performance comparison of KNN and Naïve Bayes classifier based on percentage split (90:10) technique.

Table 3: Performance Comparison of KNN & Naïve Bayes

Classifier	Number of instances		Accuracy
KNN with Min-Max	Correctly classified	85	94.44%
	Incorrectly classified	5	5.55%
KNN with Z-score	Correctly classified	83	92.22%
	Incorrectly classified	7	7.77%
Naïve Bayes with Z-score	Correctly classified	88	97.77%
	Incorrectly classified	2	2.22%
Naïve Bayes with Min-Max	Correctly classified	90	**100%**
	Incorrectly classified	-	-

IV CONCLUSION
In this paper, we have used Cleveland dataset from the UCI machine learning repository for the diagnosis of heart disease. All the patients' data are trained by using different classifiers such as Naive Bayes and K-Nearest Neighbor. From experiment, it has been found that Naïve Bayes with z-score gives highest accuracy rate of 100%. KNN with z-score classifier gives accuracy of 94.44%. We have compared the results of these techniques and from the experimental outcome it can be concluded that Naive Bayes is more accuracy than KNN. For the future research work, we suggested to develop a system with different feature selection and classification methods which could significantly decreased healthcare cost.

V. REFERENCES

[1] N. Suguna1, and Dr. K. Thanushkodi "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm" 2010.

[2] Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, "A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology",2012.

[3] Aqueel Ahmed, Shaikh Abdul Hannan,"Data Mining Techniques to Find Out Heart Diseases", International Journal of Innovative Technology and Exploring Engineering(IJITEE)ISSN: 2278-3075, Volume-1,Issue-4,September 2012.

[4] Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012.

[5] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[6] M.Akhil jabbar, B.L Deekshatulua, Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" 2013.

[7] Ms. Ishtake S.H, Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research, 2013

[8] Abhishek taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[9] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[10] M.A.Nishara Banu and B.Gomathy, "Disease Forecasting System Using Data Mining Methods",2014.