International Journal of Mathematics and Statistics Studies
Vol.5, No.2, pp.1-7, April 2017
Published by European Centre for Research Training and Development UK (www.eajournals.org)

A COMPARATIVE STUDY ON HEART DISEASE DIAGNOSIS USING CLASSIFICATION TECHNIQUES

S.Selvakumar[1], K.Senthamarai Kannan[2]S.Gothai Nachiyar [3]
[1]Department of Statistics, Govt. Arts College, Dharmapuri
[2]Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli.
[3]Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli.

ABSTRACT: *Data mining means mining knowledge from large amount of data. Classification is one of the essential role in the field of healthcare. Diagnosis of health conditions is a very important and challenging task in field of medical science. There are various types of diseases are diagnosis in medical science. Heart disease is the leading cause of death in the world over the past ten years. Heart diseases classification is one of the important problems in medical science because it is directly related to health condition of human body, this type of disease can be solved by proper identify and carefully treatment. In this paper Attribute selection, Naive Bayes, Gini index and Bayesian classification are applied for heart disease data. And also compared for classifying the accuracy.*

KEYWORDS: Attribute Selection Measure, Naïve Bayes, Gini Index, Decision Tree, and Classification.

INTRODUCTION

Data mining is a young and promising field of information and knowledge discovery (Han et. al 2011). Mai Shouman et. al (2012) proposed a model for measuring if applying data mining techniques to heart disease treatment data can provide reliable performance as that achieved in diagnosing heart disease patients. The main objective of this research work is to provide an approach for diagnosing the heart disease of the patients. It was found that more than one in the three adults is found diseased because of heart problems as per the reports of the World Health Organisation (W.H.O.). These disorders of the heart diseases are termed as cardiovascular diseases. The Cardiovasculardiseases affect 17 million people worldwide a year due to heart attacks.

Niti Guru et.al (2007) proposed the prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks. Sellappan Palaniappan et.al. (2008) developed Intelligent Heart Disease Prediction System (IHDPS) using data miningtechniques. Kiyong Noh et.al (2006) has been placed forth a classification method for the extraction of multi parametric features by assessing HRV from ECG, the data pre-processing and the heart disease pattern..

Latha Parthiban et.al (2006) projected a move towards on the basis of coactive neuro-fuzzy inference system (CANFIS) for the prediction of heart disease. Hanen Bouali and Jalel Akaichi (2014) compared different classification techniques. This comparison shows the support vector machine performance with higher accuracy.Acharya et al. (2002) have used heart rate data as the base signal for classification. George Gomes Cabral and Adriano Lorena In´acio de Oliveira (2014) presentenced an analysis of medical data aimed at determining whether or not patients are cardiac.

Kim et al.(2005) evaluated the current treatments for chronic heart failure using a decision tree and compared the results with those of large-scale clinical trails.Das and Sengur (2009)applied classification techniques such as Naive Bayes, Decision Tree, neural network and kernel density and were compared for classifying the accuracy. In this work, decision trees for attribute selection measure, Gini index and Naïve Bayes classification techniques are used for heart disease data.

MATERIALS AND METHODS

We collected a multi-dimensional heart disease dataset. This multidimensional heart disease dataset contains 100 samples with 5variables (chest pain, rest_bpress, rest_electro andexercise_angina).The data is collected from the website: https://archive.ics.uci.edu/ml/**datasets/Heart**+**Disease**. All 100 observations were classified positive and negative using Attribute selection measure, Gini Index and Naïve Bayes classification.

Method

ATTRIBUTE SELECTION MEASURE

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain is chosen as the test attribute for the current node.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for i=1,…,m). Let s_i be the number of samples of S in class C_i . The expected information needed to classify a given sample is given by

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \ldots(1)$$

Where p_i is the probability than an arbitrary sample belongs to class C_i and is estimated by s_i/s. The log function to the base 2 is used since the information is encoded in bits.

Let s_{ij} be the number of samples of class C_i in a subset S_j. The expected information based on the partitioning into subsets by A is given by

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I\left(s_{1j},\ldots,s_{mj}\right) \qquad \ldots(2)$$

The term $\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}$ acts as the weight of the jth subset and is the number of samples in the subset divided by the total number of samples in S. The encoding information that would be gained by branching on A is

Gain(A) = I(s_1, s_2,…, s_m)- E(A) …(3)

Gini Index

Let us denote by f(i, j) the frequency of occurrence of class j at node i (for m distinct classes of objects). Then, the GINI index is given by

$$\text{GINI(i)} = 1-\sum_{j=1}^{m}f^{2}(i,j)$$

When a parent node is split into p partitions, the quality of split is given by the GINI.

$$GINI_{split}=\sum_{i=1}^{p}\frac{n_i}{n}GINI(i)$$

BAYESIAN CLASSIFICATION

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

Bayesian classification is based on Bayes theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

NAIVE BAYESIAN CLASSIFICATION

Let D be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, X = (X_1, X_2,…,X_n), depicting n measurements made on the tuple from n attributes, respectively, A_1, A_2, … , A_n.

Suppose that there are m classes, C_1, C_2,…, C_m. Given an unknown data sample, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. Thus we maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posterior probabilities.

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)} \qquad \ldots(4)$$

Eqn. (4) derived from the Bays theorem. As P(X) is constant for all classes, only

$P(X|C_i)\,P(C_i)$ need be maximized.

RESULTS

To compute the attribute selection measure for using information gain and Gini index. The highest information gain and Gini index is used for test attribute.

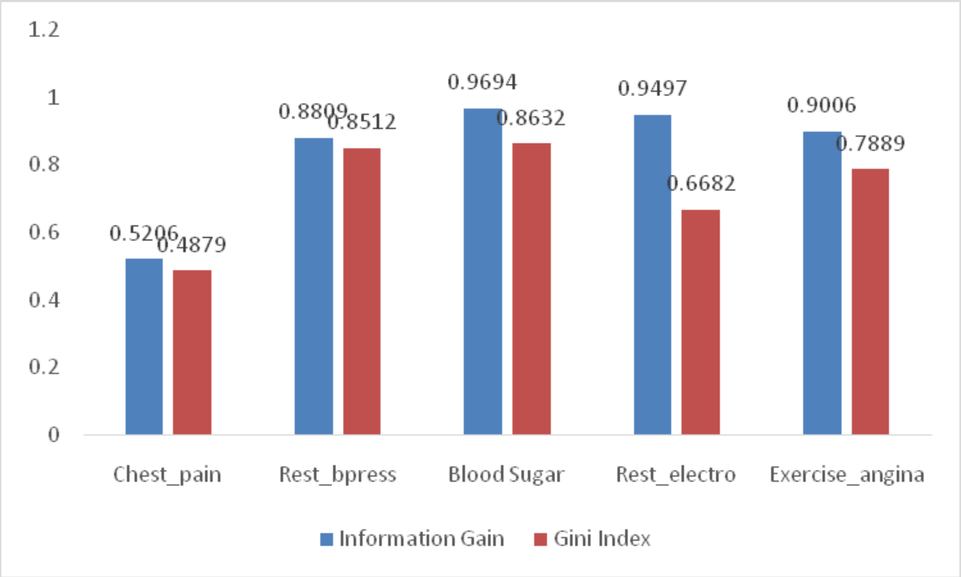
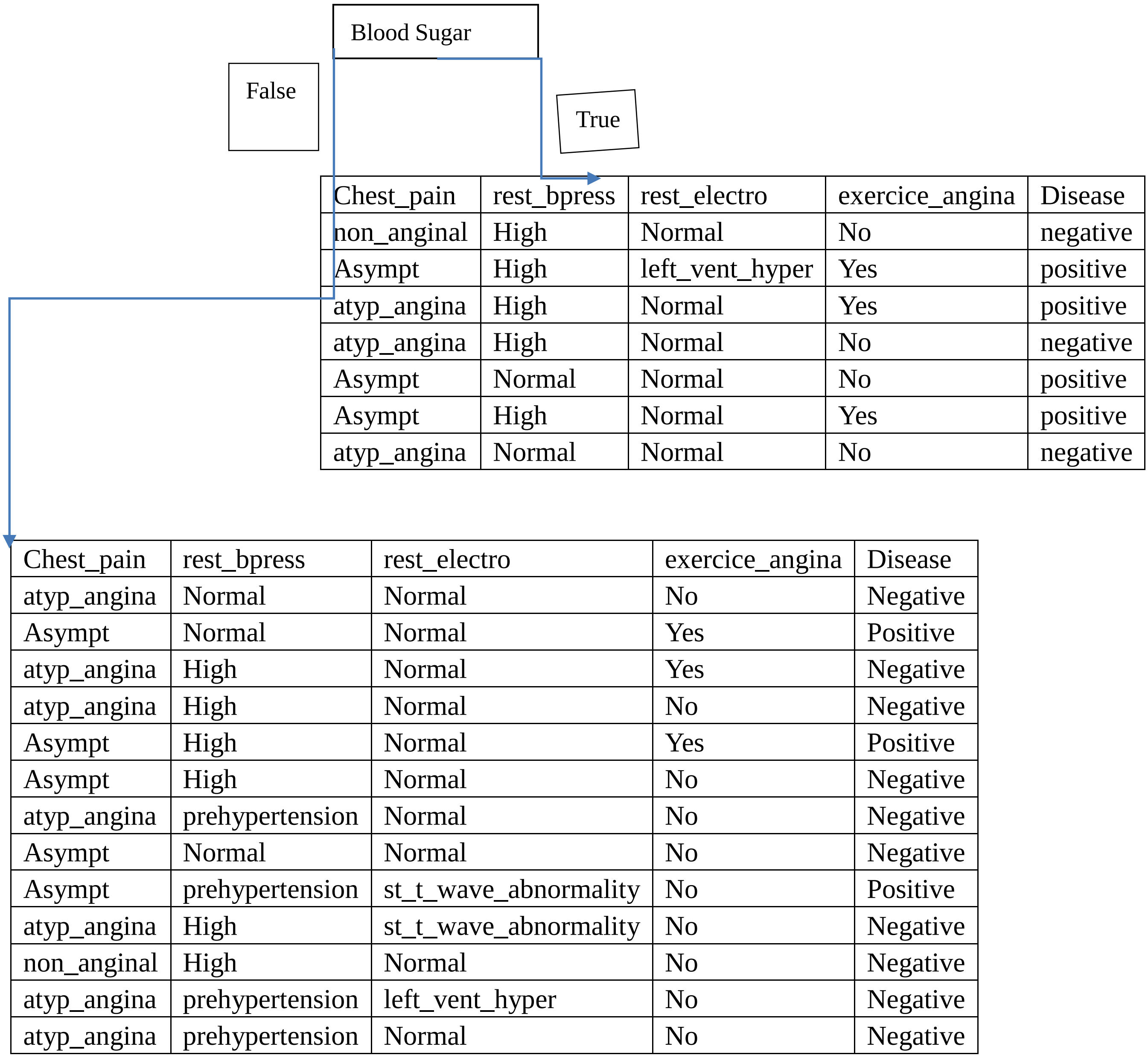


Fig. 1: Attribute Selection Measure

Thus from the Bar diagram attribute selection measure for information gain and Gini Index highest vales when compared to other variables blood sugar variable as the test attribute.



Chest_pain	rest_bpress	rest_electro	exercice_angina	Disease
non_anginal	High	Normal	No	negative
Asympt	High	left_vent_hyper	Yes	positive
atyp_angina	High	Normal	Yes	positive
atyp_angina	High	Normal	No	negative
Asympt	Normal	Normal	No	positive
Asympt	High	Normal	Yes	positive
atyp_angina	Normal	Normal	No	negative

Chest_pain	rest_bpress	rest_electro	exercice_angina	Disease
atyp_angina	Normal	Normal	No	Negative
Asympt	Normal	Normal	Yes	Positive
atyp_angina	High	Normal	Yes	Negative
atyp_angina	High	Normal	No	Negative
Asympt	High	Normal	Yes	Positive
Asympt	High	Normal	No	Negative
atyp_angina	prehypertension	Normal	No	Negative
Asympt	Normal	Normal	No	Negative
Asympt	prehypertension	st_t_wave_abnormality	No	Positive
atyp_angina	High	st_t_wave_abnormality	No	Negative
non_anginal	High	Normal	No	Negative
atyp_angina	prehypertension	left_vent_hyper	No	Negative
atyp_angina	prehypertension	Normal	No	Negative

Fig. 2: Decision Trees for Attribute Selection Measure

Naïve Bayes Classification

X=(chest pain = asympt, blood sugar = false, excersice_angina = yes, rest_bpress = high, rest_electro = normal)
P(X│positive)= 0.08311
P(X│negative)= 0.0209
The naïve Bayesian classifier predicts “ Diagnosis = positive” for sample X.
Y=(chest pain = non_anginal, blood sugar = True, excersice_angina = no, rest_bpress =

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

normal, rest_electro = left_vent_hyper)
P(Y│positive)= 0.1976
P(Y│negative)= 0.1718
The naïve Bayesian classifier predicts " Diagnosis = positive" for sample Y.

CONCLUSION

- This work focused the implementation of attribute selection measure for decision trees, Gini index and the naive Bayes classification for the Heart disease data.
- From the analysis, it is evident that the classification formed is formation of classifications are different.
- In decision trees and Gini index for attribute selection measure, when compared with other variables blood sugar has the highest information gain.
- In naive Bayes classification the representative probabilities are predicted for an unknown sample.
- In general, Bayesian classifiers have the minimum error rate in comparison to all other classifiers.
- Naive Bayes classifier is better than the decision trees and Gini index.

REFERENCES

1. Acharya, R.U., Lim, C.M. and Joseph, P. (2002). *Heart rate variability analysis using correlation dimension and detrended fluctuation analysis*. *ITBM-RBM*, Vol. 23, No. 6, pp.333–339.
2. Asha, R., Reena, G.S.: Diagnosis of Heart Disease Using Data mining Algorithm. Global
3. B. Sch¨olkopf, J. C. Platt, J. S. Taylor, A. J. Smola and R. C. Williamson(2001). Estimating the support of a high-dimensional distribution. Neural Computation. Vol. 13(7), pp. 1443–1471.
4. Carlos Ordonez(2004). *Improving Heart Disease Prediction Using Constrained Association Rules*. Seminar Presentation at University of Tokyo.
5. Das, R., I. Turkoglu, and A. Sengur(2009). *Effective Diagnosis of Heart Disease through Neural Networks Ensembles*. Export Systems with Applications. Elsevier. vol. 36, pp. 7675-7680.
6. Florin Gorunescu(2011). *Data Mining*. Springer,12th edition.
7. George Gomes Cabral and Adriano Lorena In´acio de Oliveira(2014). *IEEE International Conference on Systems*. Man, and Cybernetics.
8. Han, J., Kamber, M., Jian P. (2011). *Data Mining Concepts and Techniques. San Francisco.* CA: Morgan Kaufmann Publishers.
9. Hanen Bouali and Jalel Akaichi(2014). *Comparative Study of Different Classification Techniques* **.** IEEE**.** 978-1-4799-7415-3/14431.

10. Isaac Niwas, S., Shantha Selvakumari, R. and Sadasivam, V. (2005). *Artificial neural network based automatic cardiac abnormalities classification*. Proceedings of the 6th Internationalconference on Computational Intelligence and Multimedia Applications, IEEE.

11. Kim, J., et al., (2005). *A Novel Data Mining Approach to the Identification of Effective Drugs or Combinations for Targeted Endpoints- Application to Chronic Heart Failure as a New Form of Evidence based Medicine* . Cardiovascular Drugs and Therapy, Springer, vol. 18, pp. 483-489.

12. Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu (2006). *Associative Classification Approach for Diagnosing Cardiovascular Disease*. Springer,Vol:345, pp721- 727.

13. Latha Parthiban and R.Subramanian(2008). *Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm*. International Journal of Biological. Biomedical and Medical Sciences. Vol. 3, No. 3.

14. Mai Shouman , Tim Turner and Rob Stocker (2012). *Using Data Mining Techniques in Heart Disease Diagnosis and Treatment*. Japan-Egypt Conference on Electronics. Communications and Computers. Pp.173-177.

15. Niti Guru, Anil Dahiya, Navin Rajpal (2007). *Decision Support System for Heart Disease Diagnosis Using Neural Network*. Delhi Business Review. Vol. 8, No. 1.

16. Sellappan Palaniappan, Rafiah Awang (2008). *Intelligent Heart Disease Prediction System Using Data Mining Techniques*.International Journal of Computer Science and Network Security, Vol.8 No.8.