ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

The Diagnosis of Coronary Heart Disease Using Machine Leaning Algorithm

P. Prabaharan[1], H. Poornima[2], S. Pavithra[2], A. Madhumitha[2]

Head of the Department, Department of IT, Vivekananda College of Technology for Women, Elayampalayam, Thiruchengode, Namakkal (Dt), Tamil Nadu, India[1]

B.Tech Student, Department of IT, Vivekananda College of Technology for Women, Elayampalayam, Thiruchengode, Namakkal (Dt), Tamil Nadu, India[2]

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-Neighbour (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. This algorithm uses gain radio for feature selection. It handles both continuous and discrete features. C4.5 algorithm is widely used because of its quick classification and high precision. This paper proposed a C4.5 classifier based on the various entropies instance of Shannon entropy for classification. Experiment results show that the various entropy based approach is effective in achieving a high classification rate.

KEYWORDS: Data Imputation, KNN, Data Preprocessing, Knowledge Base

I.INTRODUCTION

Heart disease is the leading cause of death in the world over the past 10 years. The World Health Organization reported that heart disease is the first leading cause of death in high and low income countries [1]. The European Public Health Alliance reported that heart attacks and other circulatory diseases account for 41% of all deaths [2]. The Economic and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular, cancers, and diabetes diseases [3]. The Australian Bureau of Statistics reported that heart and circulatory system diseases are the first leading cause of death in Australia, causing 33.7% all deaths [4].

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 [5]-[6]. Data mining is an essential step in knowledge discovery. It is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to be detected with traditional

II. RELATED WORK

Quinlan J. R. et al resolved the C4.5 application on continuous attribute. The work proves that the decision tree could have high accuracy in data classification with inclusion of decision tree. In this paper, Continuous attribute is split based on the cut point identified. The cut point is chosen in such a way that the attribute value and its class label

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

varies. If either attribute value or its class label identified to be the same, then the cut point is moved a step ahead. Once the condition is met, the gain ratio for the attribute is calculated at that point. Thus it always gives the binary split or ternary split at a node. Binary split is with the value less than or equal to and greater than. Ternary split is with less than, equal to and greater than of attribute value. Harbinger Chauhan et al propose the implementation of decision tree algorithm C4.5 in weka software. Kalpesh et al compared the performance of ID3 and C4.5 classification algorithms for the student's dataset. C4.5 algorithm proves to be high efficient and accurate compared to ID3 algorithm.

- **C4.5 algorithm**

C4.5 is based on the information gain ratio that is evaluated by entropy. The information gain ratio measure is used to select the test features at each node in the tree. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node. Let D be a set consisting of (D_1… D_j) data instances. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for I = 1… m). Let Djbe the number of samples of D in class Ci. The expected information needed to classify a given sample is given by

$$\text{Splitinfo}_A(D) = -\Sigma \ (|Dj|/|D|)*\log((|Dj|/|D|)) \qquad (1.1)$$

$$\text{Gain ratio}(A) = \text{Gain}(A) \backslash \text{splitinfo}_A(D) \qquad (1.2)$$

Where

$$\text{Gain} = \text{Info}(D) - \text{InfoA}(D) \qquad (1.3)$$

$$\text{Info}(D) = -\Sigma \, Pi \log 2(Pi) \quad \text{and}$$

$$\text{Info}_A(D) = -\Sigma \, (|Dj|/|D|)* \text{Info}(Dj)$$

Where pi = probability of distinct class Ci, D =data Set, A=Sub attribute from attribute, (|Dj|/|D|) =act as weight of j^{th}partition. In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of feature A.

III. K-NEAREST-NEIGHBOR (KNN)

K-Nearest-Neighbor (KNN) is one of the most widely used data mining techniques in pattern recognition and classification problems. Recently Paris et al. examined single classifiers and combining different classifiers through voting and showed that voting outperformed other single classifiers. This paper investigates applying KNN in the diagnosis of heart disease on the benchmark dataset to allow comparisons with other data mining techniques used on the same dataset. It also investigates if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients. The rest of the paper is divided as follows: the background section investigates applying data mining techniques in the diagnosis of heart disease, the methodology section explains KNN and integrating voting with it in diagnosing heart disease patients, the heart disease data section explains the data used, the results section presents the KNN and voting results, followed by the summary section.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

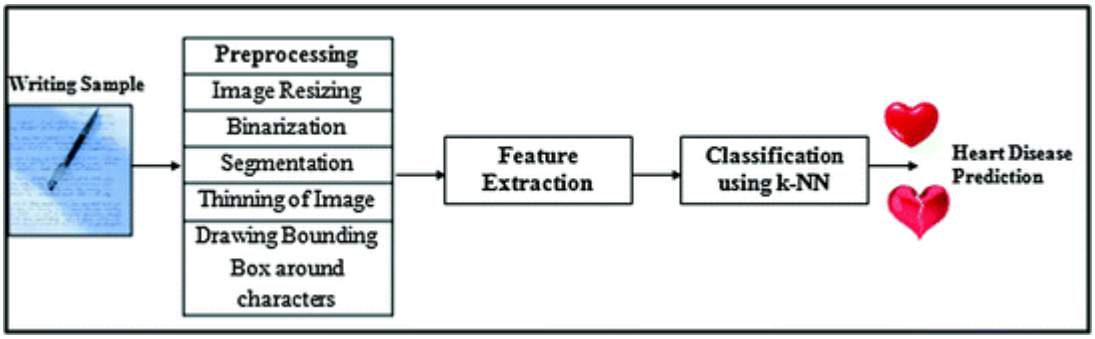

Figure: 1 Heart Disease Prediction Using KNN

KNN classification has two stages

1) Find the *k* number of instances in the dataset that is closest to instance *S*
2) These *k* number of instances then vote to determine the class of instance *S*

4155 Biomed Res- India 2017 Volume 28 Issue 9
Prediction of heart disease using k-nearest neighbor and particle swarm optimization

The Accuracy of KNN depends on distance metric and *K* value. Various ways of measuring the distance between two instances are cosine, Euclidian distance. To evaluate the new unknown sample, KNN computes its *K* nearest neighbors and assign a class by majority voting.

- **Feature subset selection**

The medical data set contains a large number of features which are redundant and irrelevant in nature. The performance of classifier may reduce if the data set contains this type of features. By removing redundant features accuracy of the classifier is improved and shortens the running time. Feature selection methods are broadly classified as

1. Filter
2. Wrapper
3. Hybrid approaches

Feature selection for large data set is a challenging task. Many search techniques used for feature selection suffers from local optima and high computational cost. Hence a cheap global search algorithm is required to develop a feature selection method. In our proposed approach, we used PSO for feature selection problem.

- **Proposed Approach**

Our proposed method aims to enhance the performance of KNN classifier for disease prediction. Algorithm for our proposed method is shown below as Algorithm 1.

Algorithm 1. Heart disease prediction using KNN and PSO.

Step 1: Input: Heart disease data set

Step 2: Output: Classification of data set into patients with heart disease and normal

Step 3: Input the data set

Step 4: Apply pre-processing techniques-Fill in missing values

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

Step 5: select the features based on values obtained after applying PSO as FSS

Step 6: Discard redundant features (features with low values of PSO)

Step 7: Apply (KNN+IQR) on Predominant features

Step 8: Measure the performance of the KNN+PSO model

Algorithm takes the heart disease dataset and classify whether a person is having heart disease or not. The above algorithm is divided into 2 parts. Part 1 (Line 3-6) performs processing and feature subset selection. This part selects only predominant features for further process. In part 2 (Line 7-8), KNN is applied on pre-processed data set and performance is measured. Feature selection measure PSO is used to select the best features to obtain high accuracy.

IV. RESULTS AND DISCUSSION

To predict heart disease the dataset containing 270 instances is collected from UCI repository. Information about heart disease data set is shown in Table 1.

Table 1. Heart disease data set.

Data set	Instances	Features
Heart disease	270	14

Table 2. KNN specifications.

Sl. no	KNN Specifications
1	KNN=2
2	Cross validation=2
3	NN Search=linear
4	Mean square =false

Features selected by PSO (Dominant features) are listed in Table 4.

Table 3. PSO specifications.

Sl. no	Specification
1	Population size: 100
2	Number of generations: 50
3	Report frequency: 50
4	Random seed=1

Out of 14 features, PSO search selects 8 features (including class). Remaining 6 features will not be considered for classification of heart disease. These 7 features are predominant features which will enhance the accuracy of the classifier. Table 5 shows the accuracy obtained by our model for heart disease data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

The accuracy obtained for various values of *K*. We tested four methods to record the accuracy of the classifier. The Interquartile Range (IQR) is a measure of variability. It divides data set into quartiles.

Table 4. Features selected by PSO.

Sl.no	Feature name
1	Chest
2	Resting_electrocardiographic_results
3	Maximum_heart_rate_achieved
4	Exercise_induced_angina
5	Old peak
6	Number_of_major_vessels
7	Thal

Table 5. Accuracy obtain ed by our model.

Method	K value				
	K=1	K=2	K=3	K=4	K=5
Before FSS	75.18	77.03	78	78	78.14
After normalization	77.7	78.8	81.1	81.4	81.4
After discretization	79.2	79.25	81.1	81.1	80.3
After PSO+KNN	78.8	81.1	81.4	81.4	81.4
KNN+PSO+IQR	100	100	100	100	100

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.

$IQR=Q3-Q1 \rightarrow (1)$

Accuracy recorded by our model before feature subset selection is 75.18 for *k=1* and 78.14 for *k=5*. Discretization filter in WEKA has improved the accuracy from 75.18 to 79.2 for *k=1* and 78.14 to 80.3 for *k=5*. IQR filter along improved the accuracy to 100%. Figure 2 shows accuracy recorded by our model for various values of K. The results obtained by KNN model reveal that our proposed model will improve the accuracy in good level. Experiments for our proposed approach were conducted on four different data sets. Heart disease-1 and Heart disease-2 are real data sets collected from various hospitals in India.

Results of proposed approach for various data sets are shown in Table 6. Values for values performance parameters are recorded. True positive rate and sensitivity are recorded as 100%.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

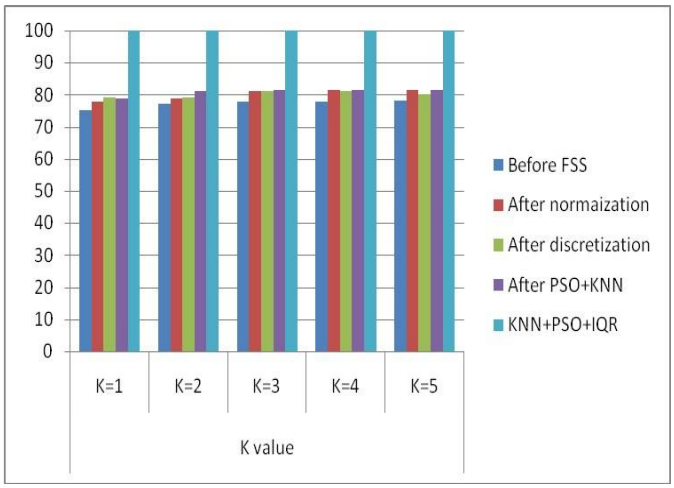

Figure 2. Accuracy recorded by proposed model for various values of k.

Table 6. Accuracy obtained for various data sets.

Sl. no	Data set	Instances	Attributes	Accuracy
1	Heart disease-1	40	10	97.5
2	Heart disease-2	75	12	100
3	Labour data	57	17	100
4	Soyabean	683	36	100

Table 7. Values for various parameters. (Heart stalog data set).

Parameter name	Value
Sensitivity	100 %
TP Rate	100 %
Accuracy	90 %

Table 8. Accuracy comparison with GA and PSO.

Data set name	Approach	Accuracy
Heart disease	KNN+GA	77.7
	KNN+PSO	90

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

Table 9. Accuracy comparis on.

Sl. no	Method	Accuracy (%)
1	Dangare [7]	92.5
2	Krishnail [8]	97
3	Kumar [9]	92
4	Amin [10]	96.2
5	Masetro [11]	99
6	Sonawale [14]	98

Prediction of heart disease using k-nearest neighbor and particle swarm optimization

7	Our approach	90

Biomed Res- India 2017 Volume 28 Issue 9

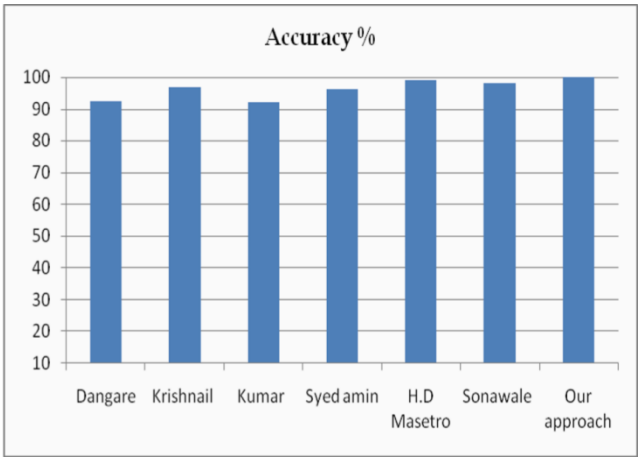

Figure 3. Accuracy comparison.

Before feature subset selection accuracy obtained is 75%. PSO search filters the number of features and selects the features which contribute more to the classification. By applying KNN with PSO accuracy improved to 100%. There is almost 25% increase in the accuracy. We tested OUR model for various values of *k*. Our experiment is limited to *k=5* as there is no much increment in the accuracy. Proposed approach is well suitable for multivariate data set.

V. CONCLUSION

This experiment is performed over eight real datasets using the five methods namely C4.5 decision tree algorithm based on Shannon Entropy, C4.5 decision tree algorithm based entropy, C4.5 algorithm based on Quadratic entropy, C4.5 decision tree algorithm based on entropy and C4.5 algorithm based. As shown in table 5, accuracy of Experimental Method based on three entropies is better than C4.5 algorithm. This paper also shows that comparative analysis between machine learning shown in above table.

To create compact decision trees with successful classification. The size of the decision tree, the performance of the classifier is based on the entropy calculation. So the most precise entropy can be applied to the particular classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

problem. The different entropies based approach can be applied in any classification problem. Such as detecting faults in industrial application, Medical diagnosis, loan approval, pattern recognition, classifying market trends etc. This thesis is a comparative study based on Shannon entropy in parallel and produce more precise classification for data set and a result of this classification is comparable with the other machining learning techniques. This entropy based approach can be applied in real world classification problems.

REFERENCES

[1] Agarwal, S., Pandey, G. N., & Tiwari, M. D. Data Mining in Education: Data Classification and Decision Tree Approach.
[2] Merceron, A., &Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).
[3] Bakar, A. A., Othman, Z. A., &Shuib, N. L. M. (2009, October). Building a new taxonomy for data discretization techniques. In *Data Mining and Optimization, 2009. DMO'09. 2nd Conference on* (pp. 132-140). IEEE.
[4] Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., & Thomson, B. (1995). CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of applied meteorology*, *34*(8), 1848-1862.
[5] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, *13*(1), 21-27.
[6] Dasarathy, B. V. (1980). Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1), 67-71.
[7] deOña, J., López, G., &Abellán, J. (2012). Extracting decision rules from police accident reports through decision trees. *Accident Analysis & Prevention*.