International Journal of Advanced Information Science and Technology (IJAIST) *ISSN: 2319:2682*
Vol.5, No.11, November 2016 DOI:10.15693/ijaist/2016.v5i11.40-44

A Novel Cluster and Rank Based Method for Prediction of Heart Diseases

Mr.K.Aravinthan
Research Scholar, Department of Computer Science,
J.J. College of Arts and Science Pudukkottai,
Tamilnadu, India

Dr.M.Vanitha
Assistant Professor, Department of Computer Applications, Alagappa University, Karaikudi,
Tamilnadu, India

***Abstract*— In this present work a group of 14 important heart risk attributes is developed from the data set of 76 parameters. The most popular 5 algorithms of J48, Naïve, CART, KNN, and NN are adopted for the prediction. The comparative studies have resulted in finding a new algorithm for higher accuracy. Consequently an algorithm of ArAfPha2016 is introduced. The accuracy of this proposed steps are tested in confusion matrix. The accuracy is found out as 0.9898, which is greater than the values of other methods studied. The above results are discussed with relation to the heart attack risk assessment.**

Keywords - A new novel algorithm for prediction of heart attack, Cluster and rank based prediction of heart diseases, Heart attack risks to assessment by ArAfPha2016.

I. INTRODUCTION

Currently the principle of data processing and data mining for the prediction of heart diseases has been extensively reviewed by many authors these days. Jyoti Soni *et. al.*, [5] has discussed various problems of construction of algorithms used in heart disease diagnosis in their review paper. They have suggested the utilization of real data from health centers and to undertake a comparative study on the optimum accuracy of automatic prediction of heart disease by data mining techniques developed so far. In this same year Subbalakshmi *et. al*., [10] has employed the popular Naïve Bayesian classification technique in heart disease prediction system. She has shown an easy assessment of heart disease risks by retrieval even in complex queries. During 2012 Aditya Sundar *et al*., [1] had different opinion in prediction accuracy. They have compared two systems of classification such as DMX query language and classification matrix method for their reliability in the diagnosis of cardiac diseases out of several complicated attributes. They have also presumed that number of life parameters is needed to be increased for data mining. In line of the above studies in 2013 several research workers have adopted various data mining algorithms (CART, ID3 and DT) for the prediction models (Chitra and Seenivasagam [2]; Dhanashree *et al*., [3]; Vikas Chaurasia and Saurabh Pal [15]).

Recently Sudhakar and Manimekalai [11] followed a latest technique of associative classification to know its efficiency for the diagnosis of heart diseases. They have shown a maximum accuracy nearing 98% in their results. Similarly Rupali [8] has conducted a comparative performance of Naïve Bayesian classification and delink-mercer smoothing technique. Her studies have drawn a conclusion that delink-mercer smoothing is more effective than other methods for the mining of the data from the heart disease attributes. Besides several investigations on the above concept, many research workers like Hlaudi and Mosima [4]; Kodali Lohita *et al*., [6]; Moloud Abdar *et al*., [7] have suggested in their papers that the reliability of the prediction requires an analysis of real data from clinical medical records of hospitals. Further they have proposed that the type of algorithm may be considered on the basis of nature of heart diseases.

Cluster analysis has been employed for pre processing of voluminous clinical data for a particular prediction. Very recently Suganya and Tamije Selvy [12] have approached the above concept in their studies by using Fuzzy-Cart algorithm. She has shown that preprocessing of heart disease data warehouse is mined effectively. Likewise Sharan Monica and Sathees kumar [9] have fully analyzed all the classification by using WEKA tool and found that J48 algorithm can give a better accuracy.

The above thorough scanning of earlier relevant literature allow to attempt much more clear studies on the classification of data, pre-processing and to test with real data of clinical sources. Therefore the present studies and the aim are to bridge certain gaps to ascertain the heart disease forecast by automatic data mining.

II. METHODOLOGY

A. Workflow

Phase I - Pre-processing

- Missing attributes removed

Phase II - Contribution based clustering

- Calculation of contribution score
- Clustering the training data sets based on contribution score

Phase III - Ranking of influencing attributes

- Assigning weights to the clusters
- Rank the attributes and identify the influencing attributes based on threshold

Phase IV – Prediction

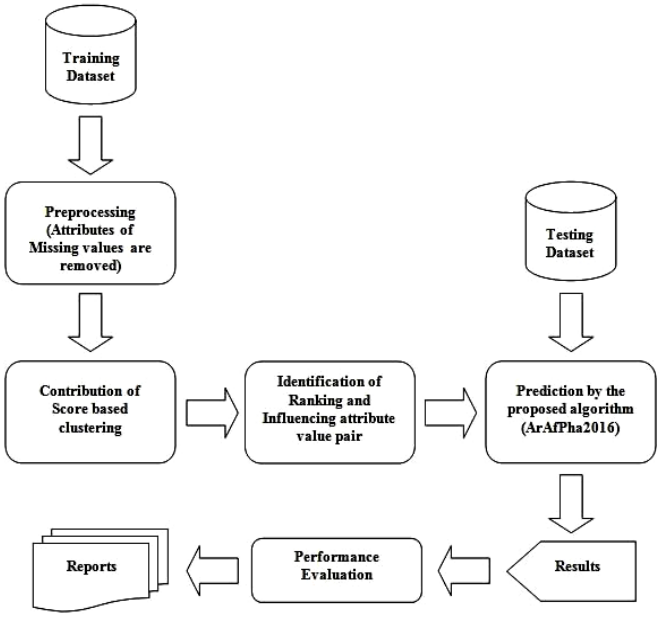


Figure.1 Layout of Methodology Adopted in This Study

B. The source of Hungarian dataset

UCI Machine Learning Repository - Heart Diseases 294 data set [13]

C. The segregation of experimental dataset

Cleveland database have suggested to differentiate presence (values 1,2,3,4) from absence (value 0) we split the dataset into two groups in UCI Machine Learning Repository [14].

Table 1. Classification and Distribution of Hungarian Database.

DATABASE	ABSENCE	PRESENCE				TOTAL
	0	1	2	3	4	
Hungarian	188	37	26	28	15	294

D. Selection of fourteen important of heart attack risk

Based on the earlier studies 14 important parameters were identified. This information was drawn from heart disease directory (UCI Machine Learning Repository) [14] the details of the 14 parameters are explained in the result section.

E. Proposed Algorithm

The proposed algorithm new as ArAfPha2016. Here in denoted as ArAfPha2016.

Step1: Establishment of experimental dataset using Hungarian data in UCI machine learning repository
Step2: Uploading of Pre-processed experimental dataset here 76 attributes are reduced into 14 attributes
Step3: Reduction of 76 attributes into 14 important attributes
Step4: Clustering of 14 important attributes for ranking
Step5: Scoring of positive and negative heart attack risks in sub-clusters.
Step6: Estimation of weightage by the formula given below

$$\text{Weight} = \frac{\text{Count of an attribute having heart disease}}{\text{Total count of entries having heart disease}}$$

Step7: Allotment of threshold value by calculating 75% of the total count of positive cases
Step8: Identification of attributes for prediction
Step9: Determination of prediction value based on minimum rank to the maximum and maximum rank to the minimum contributed attributes
Step10: Stopping the process

III. RESULTS

The present study of automatic computer prediction of heart attack is based on the cluster analysis and construction of new algorithm. It is conducted by using already developed pre fixed Hungarian clinical cardiology data. The data had 76 attributes of cardiology risk factors. For the convenience of analysis the large data set was selected as an important characters of risks containing 14 attributes of small size. There were 294 data set. This was reduced into two groups as one with individuals of having heart attack and other group consists of persons with not having heart attack. The confusion matrix was employed. The result and the calculations are given below.

Table 2. Confusion Matrix.

		PREDICTED CLASS	
		Yes	No
ACTUAL CLASS	Yes	TP	FN
	No	FP	TN

Table 3. Hungarian Database After Confusion Matrix Analysis.

ACTUAL/REALITY	TRUE	FALSE
TRUE	104 (TP)	1 (FN)
FALSE	2 (FP)	187 (TN)

Table 4 is furnished with the above important risk parameters of heart attack reduced by the above mathematical derivations out of 76 attributes.

Table 4. List of Important Risky Parameters of Heart Attack Reduced by Cluster Analysis

SL. NO	ATTRIBUTES	DESCRIPTION

1	Age	Age in years
2	Sex	Sex (1 = male; 0 = female)
3	Cp	Chest pain type Value 1: Typical Angina Value 2: Atypical Angina Value 3: Non-Anginal Pain Value 4: Asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (1=true; 0=false)
7	Restecg	Resting Electrocardiographic Results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina (1=yes; 0=no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment Value 1: upsloping Value 2: flat Value 3: downsloping
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14	Num	Diagnosis of heart disease (angiographic disease status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing

The prediction of heart disease based on this 14 parameters was determined by a new algorithms designed in this present study. The write up of the proposed algorithms is follows.

The adaptation of the new algorithm (ArAfPha2016) suggested in this study gave a predicted accuracy of heart diseases by using Hungarian data set (76 attributes and 294 data set) was yielded 0.9898. To know the efficiency of the new algorithm on the prediction of heart attack the calculated accuracy units were compared with other algorithms (J48, Naive, CART, KNN, NN) studied by earlier workers. Table 5 is provided with the various accuracy values and the comparison of the value of new proposed algorithm.

Table 5. Comparative Predictive Accuracy Values of Different Algorithms with New Algorithm

Techniques	**Accuracy/ TP**	**Misclassification Rate / FP**	**Precision**	**Recall**
Proposed (ArAfPha2016)	0.9898	0.0102	0.9811	0.9905
J48	0.9478	0.0522	0.0948	0.9510
Naïve	0.7199	0.2801	0.0758	0.7200
CART	0.8599	0.1401	0.0863	0.8600
KNN	0.8837	0.1163	0.8809	0.8809
NN	0.8023	0.1977	0.8378	0.7380

From the table it is evident that proposed new algorithm gave an accuracy value of 0.9898. On the other hand the other cases of algorithms such as J48, Naive, CART, KNN, NN had the corresponding accuracy units as 0.9478, 0.7199, 0.8599, 0.8837, and 0.8023 respectively. In the light of above foregoing results it would reasonable to suggest that the following of new algorithm for heart attack prediction is greater than other algorithms (Table.5, Figure.2 and Figure.3).

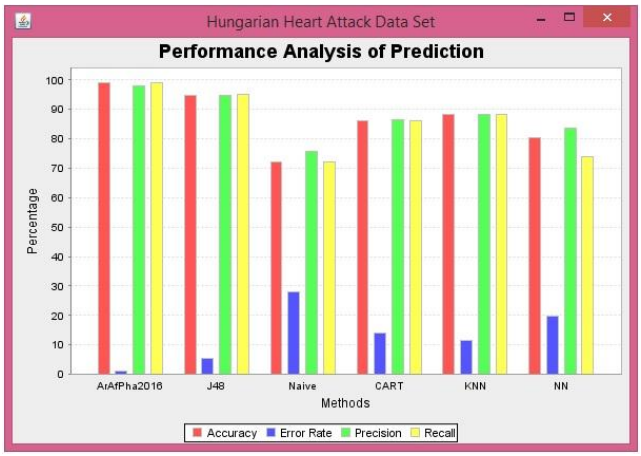
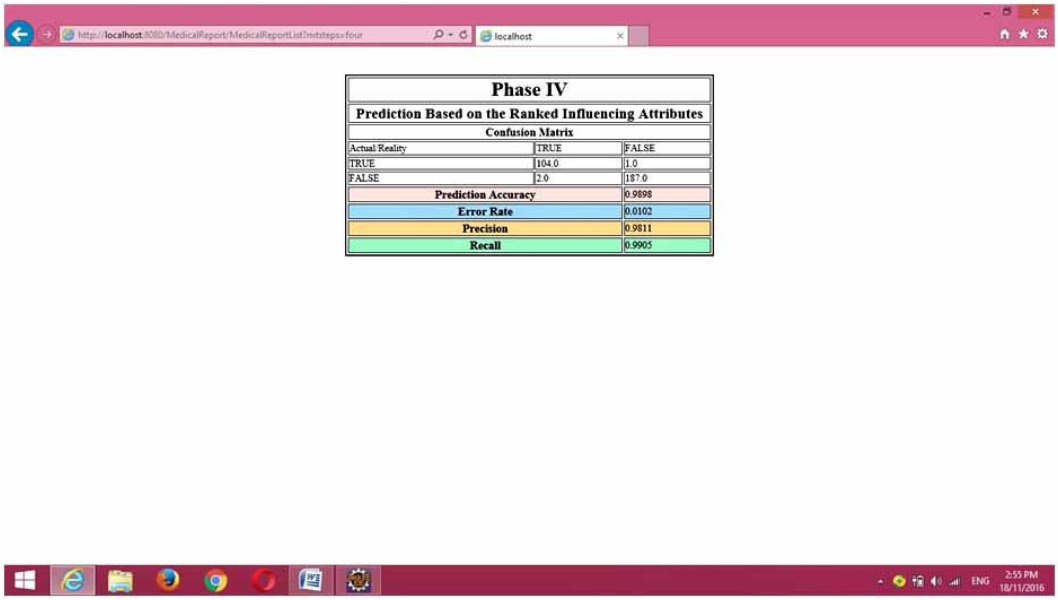


Figure 2. Image of screenshot showing histogram for the performance analysis of prediction.



Phase IV		
Prediction Based on the Ranked Influencing Attributes		
Confusion Matrix		
Actual Reality	TRUE	FALSE
TRUE	104.0	1.0
FALSE	2.0	187.0
Prediction Accuracy		0.9898
Error Rate		0.0102
Precision		0.9811
Recall		0.9905

Figure.3 Image of screenshot showing Prediction Accuracy, Error Rate, Precision and Recall Values.

IV. CONCLUSION

The Present study is on the prediction of heart diseases by cluster analysis and using five algorithms such as J48, Naïve, CART, KNN, NN. Since there are large number of clinical parameters are appeared to be consider for the estimation of heart attack. The results have revealed that efficiency of assumption of the risks for heart disease by adopting the above principle is very much appreciable. Besides the comparative examination of the above described algorithms have allowed to develop a new algorithm with more accurate predictable value for the heart diseases of any causative factors. Further it may reduce any artifacts happened in the course of automation risk assessment. Nevertheless the utilization of Hungarian medical data set is a fixed one, applied elsewhere by many research workers and it is a secondary data in nature. Therefore it is desirable to utilize a real sophisticated cardiology specialized hospital data with variable attributes in future to evaluate this present novel prototype automatic solution for the rapid diagnosis of heart disease risks. Such studies may find a suitable computer automatic technique in clinical practices in Indian context.

REFERENCES

[1] Aditya Sundar. N, P.Pushpa Latha, M.Rama Chandra. Performance Analysis of Classification Data Mining Techniques Over Heart Disease Database. *International Journal of Engineering Science & Advanced Technology*, 2012; 2(3): 470-478.

[2] Chitra R and V.Seenivasagam. Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques. *ICTACT Journal On Soft Computing*, 2013; 3(4): 605-609.

[3] Dhanashree S. Medhekar, Mayur P. Bote and D. Deshmukh. Heart Disease Prediction System using Naive Bayes. *International Journal of Enhanced Research in Science Technology and Engineering*, 2013; 2(3): 1-5.

[4] Hlaudi Daniel Masethe and Mosima Anna Masethe. Prediction of Heart Disease using Classification Algorithms. *Proceedings of the World Congress on Engineering and Computer Science (WCECS)*, 2014; II: 22-24.

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal on Computer Applications*, 2011; 17(8): 43-48.

[6] Kodali Lohita, Adusumilli Amitha Sree, Doreti Poojitha, T. Renuga Devi, A. Umamakeswari. Performance Analysis of Various Data Mining Techniques in the Prediction of Heart Disease. *Indian Journal of Science and Technology*, 2015; 8(35): 1-7.

[7] Moloud Abdar, Sharareh R. Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto, Goli Arji. Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*, 2015; 5(6): 1569-1576.

[8] Rupali R. Patil. Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing. *International Journal of*

Advanced Research in Computer and Communication Engineering, 2014; 3(5): 6787-6789.

[9] Sharan Monica L,Sathees Kumar B. Analysis of Cardiovascular Heart Disease Prediction Using Data Mning Techniques. *International Journal of Modern Computer Science (IJMCS)*, 2016; 4(1): 55-58.

[10] Subbalakshmi. G, K. Ramesh, M. Chinna Rao. Decision Support in Heart Disease Prediction System using Naive Bayes. *Indian Journal of Computer Science and Engineering*, 2011; 2(2): 170-176.

[11] Sudhakar. K and M.Manimekalai. Study of Heart Disease Prediction using Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2014; 4(1): 1157-1160.

[12] Suganya. S, P. Tamije Selvy. A Proficient Heart Disease Prediction Method Using Fuzzy-Cart Algorithm. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, 2016; 2(1): 1-6.

[13] UCI Machine Learning Repository (Database), *https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/hungarian.data*.

[14] UCI Machine Learning Repository (Heart Disease Directory), *https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names*.

[15] Vikas Chaurasia and Saurabh Pal. Early Prediction of Heart Diseases Using Data Mining Techniques. *Carib.j.SciTech*, 2013; 1: 208-217.

Authors Profile

Mr.K.Aravinthan is a Ph.D. Research Scholar Department of Computer Science in J. J. College of Arts and Science Pudukkottai, Tamilnadu, India. His research work on Data mining applications to Prediction of Heart Diseases. He obtained his Bachelor's Degree in physics and Master's degree MCA in computer science during the year 2005 and 2008, respectively both from the AVVM Sri Pushpam College Poondi, affiliated to Bharathidasan University, Tamilnadu, India. He has published number of research papers in International journals and international conferences.

Dr.M.Vanitha is a Assistant Professor, Department of Computer Applications, Alagappa University, Karaikudi, Tamilnadu, India. She Obtained her Bachelor's Degree in Mathematics during the year 1995 from SRC, Trichy. She did her Master's degree M.Sc (OR and CA) during the year 1997 at NIT, Trichy, M.Phil in Computer Science from Bharathidasan University, Trichy and Ph.D in Computer Science from Mother Teresa University, Kodaikanal. She published many research papers and acted as session Chairperson and reviewer.