Probability Rule base Clustering Approach for Heart Disease Risk Prediction

K. Chandra Sekhar
Assistant Professor
Department of Computer Science and Engineering
Anil Neerukonda Institute of Technology and Science

P. Naga Srinivasu
Assistant Professor
Department of Computer Science and Engineering
Anil Neerukonda Institute of Technology and Science

ABSTRACT
Data mining is a mechanism to locate divergent patterns that analyze the data and condense it into useful information. The idea of data mining are predictions and descriptions. The current research intends to predict the heart disease risk of patients. Probability rule base Clustering approach for Heart disease Risk Prediction(PbC_HRP) model is proposed in the heart disease risk prediction. In this model there are two approaches, PRBC (Classification approach) and OCPD (Clustering approach). Probability Rule Base Classification (PRBC) constructs knowledge base using medical guidelines and probability values and generates classification rules. Optimized Cluster Pair wise Distance base clustering (OCPD) uses the classification rules from PRBC, calculates fitness values and produces clusters which will represents the risk levels of heart disease. The clusters will give the features to the patients from the respective risk levels of clusters. It helps to warn the patient before disease became sever.

Keywords
Data mining, Risk prediction, Probability Rule Base Classification, Optimized Cluster Pair wise Distance base clustering.

1. INTRODUCTION
Data mining techniques have data processing capability and problem solving nature. Data mining is a knowledge discovery technique to analyze data and confine it into useful information. Data mining performs a significant retrieval of useful information in potential data. In data mining there exists different processes for inspecting data from different perspectives. By using those techniques that can gather knowledge from data. At present researches are going on in discovering knowledge from health care data in health industries. Health care systems focusing on disease prediction, which will be helpful in diagnosis of diseases. Data mining is one of technique to discover the hidden information from the data.

In the health care industry one of major challenging issue is quality of service. This will results providing correct treatment to patients. Poor diagnosis can lead to disastrous consequence which will damage the health system of patients. In the human body heart is an important organ and efficient working of heart results healthy life. If the heart is not working properly it will affect the other organs of body like brain, kidney etc.

At present heart problems are noting down high lethality rates and the supervision of heart diseases is highly difficult [6]. Predicting the heart disease before it goes to critical stage will helpful in our health management and also reduces the treatment costs [7].

In the prognosis of heart illness risk data mining techniques giving better outcomes. Classification and clustering techniques in data mining are more functional in risk prediction of heart illness. Numerous models were proposed to predict the risk of heart illness through data mining. Some of them are c 4.5, Naive bayes, Support vector machine, etc [6, 8, 12].

2. RELATED SURVEY
Huge survey has been done in data mining techniques on medical data. We focused on predictive models in the prediction of heart disease. Predictive techniques will tell what will happen in early days. The predictive techniques are categorized into two classes: Classification and Prediction [1]. Regression, Classification and time series analysis are the crucial job in organizing information in data mining. Numerous types of classification techniques are obtainable in data mining that involves Support Vector Machines (SVM), Neural Networks, decision trees, Bayesian network [1].

In medical and health care research different types of predictive techniques have been exploited. Decision trees, Regression analysis, Genetic algorithms are some of techniques have been used to predict the patient's disease [2, 3, 4, 5]. These predictive models will give the risk of disease in a patient. It will be helpful in monitoring the patient by giving suggestions in their lifestyle, diet, and daily activities [6]. Different predictive techniques have been evolved to predict the disease at premature stage by observing patients in a timely basis [7, 8, 9, 21]. It will be helpful in detecting the severity of disease before patient gets attacked severely by that disease.

In mining the data we encounter different clustering algorithms, every algorithm was developed by distinct properties. [13,14,15,16,17] represents the different algorithms of clustering in data mining. Every clustering algorithm is designed for the same purpose, that is to form clusters by dividing data into groups. But the algorithm that will be apt for us is depends on the type of data in our data set and the expected result.

3. PROPOSED MODEL
3.1 Problem Statement:
In researches related to medical data, prediction of risk of a particular disease is currently in high demand. In that heart disease is noting down high fatality rate [6]. In data mining huge research has been done in the prediction of heart disease. From the data related to heart disease of patients if a research can identify the risk of heart disease then it will be helpful to warn the patients of before its getting dangerous. The goal of this research is to design an efficient heart disease prediction model to identify the risk of heart disease.

3.2 Design of heart disease risk prediction model

3.2.1 Data set

We collected data on 318 persons who are suffering from heart disease. We predicted the risk of heart disease of those 318 patients. To generate the prediction model of heart disease, we used eight attribute that include patient id, gender, age, HDL cholesterol, total cholesterol, systolic blood pressure, smoking, diabetes. In addition to these we considered one output variable named CHD Event which will gives the risk of heart disease. The features of these attributes are shown in table 1.

3.2.2 Architecture

Probability rule based Clustering approach in Heart disease Risk Prediction(PbC_HRP) contains Knowledge base and OCPD as main sections.

Components of PbC_HRP Architecture:

In Knowledge base classification has done and In OCPD clustering has done.

a) Knowledge base: It helps in design of classification technique based on rules from the medical expert.
   i) Rule verification: Here the rules are the medical guidelines for the patients data. the rules verification has been done by the medical experts.
   ii) PRBC: Here classification has done based on the probability function which was designed based on the medical guidelines [6].After employment of probability function rule verification should be done.

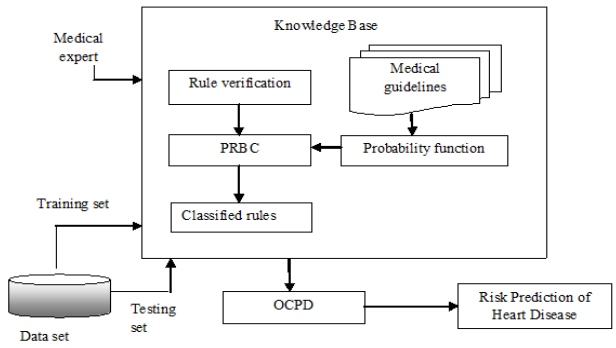

Fig.1 PbC_HRP Architecture

   iii) Classified rules: It contains the result of classification. After classification the resultant data was stored in classified rules.

b) OCPD: The result from rule base was further clustered based on the OCPD clustering. In this OCPD clustering the data records were clustered based on the pair wise distance of data records.

The data set is divided as training set and testing set. Training set contains information related to all attributes regarding symptoms of heart disease.

In the construction of knowledge base based on the rules and probability values one classification algorithm was developed, named PRBC(Probability rule base classification). It gives the classified data based on probabilities which will be helpful in risk prediction of heart disease in a easy way. Before classification the data should satisfy particular constraints known as rules which will be under supervision of medical experts.

The classified data obtained from knowledge base was further clustered on the basis of OCPD(Optimized Cluster Pair wise Distance base algorithm).It results the groups of data which denotes the risk level of heart disease.

3.2.3 Probability Rule base Classification

Before design of PRBC, A study is performed on different classification algorithms in data mining [15].Probability Rule base Classification algorithm is a classification of data based on probabilities. In this process initially the data set was divided as training and testing sets. Training set contains data related to all attributes including risk predicting attribute. Let us consider the total number of records in training data set is n_{tr} . Testing set contains attributes except output attribute CHD Event. Let us consider the total number of records in testing data set is n_{ts} Take first test result attribute value(A1) from the testing dataset and find out posteriori hypothesis i.e. compare first test result value to training dataset of first attribute value. if the first attribute value of training data set is equal to first test result of testing then consider the risk prediction class. This action reiterate until the complete number of training data sets are completed. Count the number of matched attributes in the training data set(nm_{tr})and take those count will calculate probability of first attribute of testing dataset.

$$P(A1 \vee YES) = nm_{tr}/n_{tr}$$

Now it's turn of unmatched attributes as same procedure. Here nu_{tr} is the count of unmatched attributes in the training set.

$$P(A1 \vee YES) = nu_{tr}/n_{tr}$$

It continues till all the attributes completes calculations. Calculate the yes and no probabilities of CHD attribute. That is shown in following equations.

$$P(CHD \vee YES)_{total} = n_y/n_{tr}$$

$$P(CHD \vee NO)_{total} = n_n/n_{tr}$$

Here n_y is number of occurrences of first heart disease risk, n_n is count of non-occurrences of first heart disease risk. Now calculate the total probability of first record of testing data set by following equation.

P(R || YES) = P (A_1 || YES) * P (A_2 || YES)*…..*P (A_n || YES) * P (CHD || YES)
P(R || NO) = P (A_1 || NO) * P (A_2 || NO)*…..*P (A_n || NO) * P (CHD || NO)

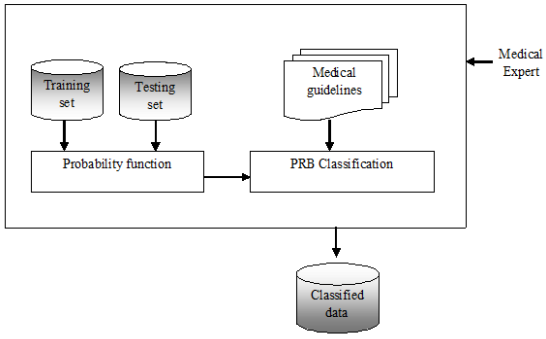

Fig.2 PRBC Architecture

After calculation of probability, classification was done based on rules given from medical expert.

In PRBC on training and testing data sets probability function is applied. After calculation of probabilities consider the highest probability CHD risk and classification done based on guidelines given by medical expert. The result is the classified data.

3.2.3 Optimized Cluster Pair wise Distance base Clustering

In this process the input for OCPD algorithm is classified dataset obtained from PRBC. OCPD clustering approach formed heart disease risk levels as clusters. On the resultant dataset which is obtained five clusters such as Very low, Low, Moderate, High, Very high which will gives the risk of heart disease. There exists different clustering algorithms in data mining [13, 14, 15,16,17,18,19]. In on the basis of study of clustering algorithms a developed OCPD clustering is proposed.

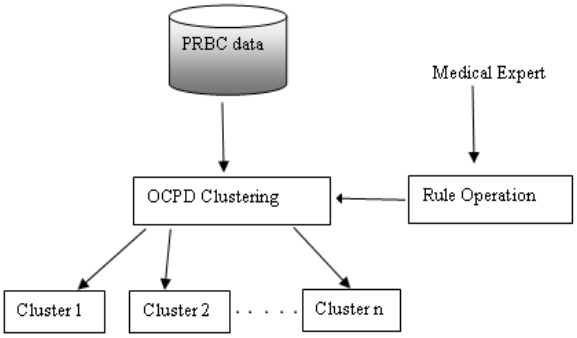

Fig.3 OCPD Architecture

The data obtained from the PRBC classification is supplied to the OCPD clustering. Each time while doing clustering the rule verification should be done. Those rules were designed by medical experts.

In this risk prediction of heart disease we have five risk levels of heart. Those are Very low, Low, Moderate, High, Very high. To get the risk levels as clusters, the count of clusters was taken as five. Centroids of each cluster ware chosen randomly. Now the data record(x_i other than the centroid (c) was chosen randomly and the mean square error (MSE) with each centroid was calculated by the following equation.

$$MSE = 1/n \sum_{i=1}^{n} (x_i - c)$$

This process goes next iteration and compare MSE with previous result and take the better result from both. This procedure repeats till there will not be further change in MSE value and consider that as centroid. Consider those centroids and calculate pair wise distance between cluster, pick the clusters which are having smaller pair wise distance. This procedure results five clusters which are having heart disease risk level.

4. EXPERIMENTAL EVALUATION

To measure the accuracy of our proposed model PbC_HRP, It was employed using R console. The dataset of heart disease patient was also used. We examined the CHD risk percentages grabbed from Heart disease patients dataset with those using PbC_HRP model.

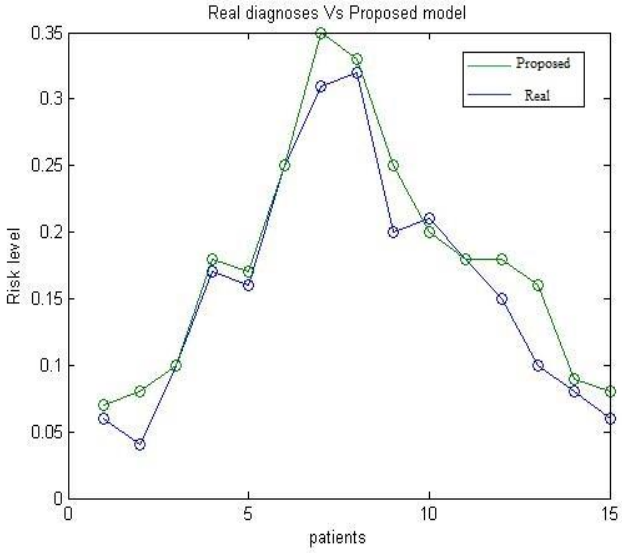

Fig. 4 Real diagnoses Vs Proposed model

Fig. 4 illustrates the compatibility of the results from the proposed model PbC_HRP with those from the Heart disease patients dataset. In the development of PbC_HRP model we implemented PRBC classification to predict the risk of heart disease. After gathering information related to patients suffering from heart illness to which set we applied PRBC classification to predict the heart disease risk. In fig. 4 we have shown the comparative results between patient data and classified data.

Here the graph was drawn between proposed model and actual diagnoses. From the actual diagnoses which are measured the percentage of risk for patients and in the same way for the proposed model also. The graph was drawn for 15 patients risk levels.

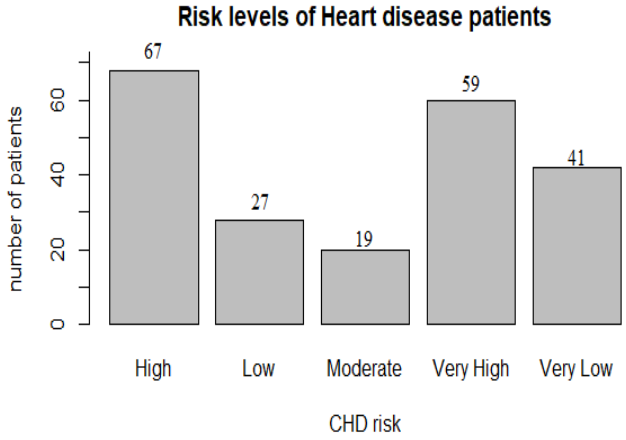

Fig.5 Clusters obtained from OCPD

In final result of PbC_HRP model is clusters of CHD risk of patients. Using OCPD clustering approach that cluster a classified data obtained from PRBC classification. On experimenting obtained five clusters based on the risk levels of heart disease. Fig. 5 illustrates the number of patients in the probability levels of heart illnes. The clustering of risk levels

helpful to identify the symptoms of patients according to the risk level of heart disease.

Table 1 Dataset features

No.	Feature	Units	Range	Type
1	Patient id	Number	100101-100318	Numeric
2	Sex	[1: Male, 2: Female]	29-73	Categorical
3	Age	Year	1, 2	Numeric
4	Total Cholesterol	mg/dL	104-357	Numeric
5	HDL Cholesterol	mg/dL	25-91	Numeric
6	Systolic BP	mmHg	56-154	Numeric
7	Diabetes	[0: No, 1: Yes]	0, 1	Categorical
8	Smoking	[0: No, 1: Yes]	0, 1	Numeric
9	CHD Event	[Very Low risk, Low risk, Moderate risk, High risk]	VL, L, M, H	Categorical

The total testing records were 213. In the 213 recorded and wanted to find out the risk levels of patients based on the training records. In this risk prediction on these 213 records PRBC classification has done, it results classified data. On the classified data obtained from PRBC classification OCPD clustering technique has applied. The OCPD clustering results five clusters in accordence with the risk levels of heart disease. Those are High-67, Low-27, Moderate-19, Very High-59, Very Low-41.

5. CONCLUSION

Nowadays heart disease is noting down high lethality rates. Prediction of heart disease is complex situation for medical experts also. This paper exhibit the importance of associating the medical knowledge from medical expert and data mining techniques. It conquer the improbability of prediction of heart disease from medical knowledge. This proposed the Probability Rule base Clustering approach for Heart disease Risk Prediction. The PbC_HRP model is designed based on the medical knowledge and data mining techniques. The medical knowledge was collected from medical experts. For the medical knowledge that could be applied PRBC classification to predict the heart disease risk, which was developed based on the probabilistic values. OCPD clustering gives the clusters of heart disease risk levels. It will helpful to know the symptoms of patients according to the risk level. In further work will examine using huge dataset in order to improve the risk prediction of heart disease.

6. REFERENCES

[1] S. Sivapalan, A. Sadeghian, H. Rahnama, and A. M. Madni. Recommender systems in e-commerce. In World Automation Congress (WAC), pp. 179-184, 2014.

[2] T. W. Joo, and S. B. Kim. Time series forecasting based on wavelet filtering. In Expert Systems with Applications, 2015.

[3] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra Diagnosis of lung cancer prediction system using data mining classification techniques. In International Journal of Computer Science and Information Technologies, vol. 4, no. 1, pp. 39-45, 2013.

[4] M. Kurosaki, N. Hiramatsu, M. Sakamoto, Y. Suzuki, M. Iwasaki, A. Tamori, K. Matsuura, S. Kakinuma, F. Sugauchi, and N. Sakamoto. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C In Journal of hepatology, vol. 56, no. 3, pp. 602-608, 2012.

[5] M. H. Lee, H. I. Yang, J. Liu, R. Batrla-Utermann, C. L. Jen, U. H. Iloeje, S. N. Lu, S. L. You, L. Y. Wang, and C. J. Chen. Prediction models of long-term Cirrhosis and hepatocellular carcinoma risk in chronic hepatitis B patients: Risk scores integrating host and virus profiles. In Hepatology, vol. 58, no. 2, pp. 546-554, 2013.

[6] Jae-Kwon Kim · Jong-Sik Lee · Dong-Kyun Park · Yong-Soo Lim · Young-Ho Lee · Eun-Young Jung. Adaptive mining prediction model for content recommendation to coronary heart disease patients. Springer Science+Business Media New York 2013, 25 September 2013.

[7] M. Sabibullah, V. Shanmugasundaram and R. Priya. Diabetes patient's risk through soft computing model. In International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 2, no. 6, pp. 60-65, 2013.

[8] F. Siraj, and M. A. Abdoulha. Mining enrolment data using predictive and descriptive approaches. In Knowledge-Oriented Applications in Data Mining, pp. 53-72, 2007. [9] S. Tuff—ry. Data mining and statistics for decision making. John Wiley & Sons, 2011.

[10] Clocksin,W.F.: Artificial intelligence and the future. Philos. Trans. R. Soc., Math. Phys. Eng. Sci. 361(3), 1721–1748 (2003)

[11] Subramanian, G.H., Yaverbaum, G.J., Brandt, S.J.: An empirical evaluation of factors influencing expert systems effectiveness. J. Syst. Softw. 38(3), 255–261 (1997)

[12] S. Radhimeenakshi. Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network. : Computing for Sustainable Global Development (INDIACom),3rd International Conference,2016

[13] Raj bala, Sunil Sikka, Juhi Singh.A Comparative Analysis of Clustering Algorithms. International Journal of Computer Applications (0975 – 8887) Volume 100 – No.15, August 2014

[14] Sharmila, R.C Mishra. Performance Evaluation of Clustering Algorithms. International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013

[15] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu,

Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. Top 10 algorithms in data mining. Knowledge and Information Systems, January 2008, Volume 14, Issue 1, pp 1–37

[16] Garima, Hina Gulati, P.K.Singh. Clustering techniques in data mining: A comparison. Computing for Sustainable Global Development (INDIACom), 2015.

[17] Ma Hong, Kang Jing, Liu Li-xiong. Research on clustering algoriths of data streams. Information Management and Engineering (ICIME), 2010.

[18] Hem Jyotsana Parashar, Singh Vijendra, and Nisha Vasudeva. An Efficient Classification Approach. International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012 for Data Mining