Global Journal of Medical research: F
Diseases
Volume 14 Issue 1 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4618 & Print ISSN: 0975-5888

Classification Model for the Heart Disease Diagnosis

By Atul Kumar Pandey, Prabhat Pandey & K.L. Jaiswal
APS University, Rewa (M.P.), India

Abstract- Medical science industry has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data. Advanced data mining techniques can be used to discover hidden pattern in data. Models developed from these techniques will be useful for medical practitioners to take effective decision. In this research work, we have analyzed the performance of the classification rule algorithms namely PART based on K-Means Clustering algorithms. The k-means is the simplest, most commonly and good behavior clustering algorithm used in many applications. Firstly the preprocessed heart disease dataset is grouped using the K-means algorithm with the K =2 values on classes to cluster evaluation testing mode. After that data mining classification rule algorithms namely Projective Adaptive Resonance Theory are analyzed on clustered relevant dataset. In our studies 10-fold cross validation method was used to measure the unbiased estimate of the prediction model. Accuracy of K-Means Clustering, PART and PART based on K-Means Clustering are 81.08%, 79.05% and 84.12% respectively.

Keywords: heart disease, data mining techniques, classification rules, k-means clustering, and part.

GJMR-F Classification : NLMC Code: WG 200, WG 205

Strictly as per the compliance and regulations of:

Classification Model for the Heart Disease Diagnosis

Atul Kumar Pandey [α] Prabhat Pandey [σ] & K.L. Jaiswal [ρ]

Abstract- Medical science industry has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data. Advanced data mining techniques can be used to discover hidden pattern in data. Models developed from these techniques will be useful for medical practitioners to take effective decision. In this research work, we have analyzed the performance of the classification rule algorithms namely PART based on K-Means Clustering algorithms. The k-means is the simplest, most commonly and good behavior clustering algorithm used in many applications. Firstly the preprocessed heart disease dataset is grouped using the K-means algorithm with the K =2 values on classes to cluster evaluation testing mode. After that data mining classification rule algorithms namely Projective Adaptive Resonance Theory are analyzed on clustered relevant dataset. In our studies 10-fold cross validation method was used to measure the unbiased estimate of the prediction model. Accuracy of K-Means Clustering, PART and PART based on K-Means Clustering are 81.08%, 79.05% and 84.12% respectively. Our analysis shows that out of these three classification models Classification based on Clustering predicts cardiovascular disease with improved accuracy.

Keywords: heart disease, data mining techniques, classification rules, k-means clustering, and part.

I. Introduction

Accurate and error-free of diagnosis and treatment given to patients has been a major issue highlighted in medical service nowadays. Quality service in health care field implies diagnosing patients correctly and administering treatments that are effective [11].Hospitals can also minimize the cost of clinical tests by employing appropriate computer-based information and/or decision support systems. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data [10]. These systems generate huge amounts of data which take the form of numbers, text, charts and images.

Data mining is the process of extracting hidden patterns from large data sets. Data mining is a searching process done automatically for hidden patterns present in a large database [2]. Data mining is an iterative process. Its progress is defined by discovery, through either automatic or manual methods. Data mining is reflected in its wide range of methodologies and techniques [8]. These techniques can be applied to a connection of problem sets. Classification deals in generating rules that partition the data into disjoint groups. Classification is a data mining (machine learning) technique used to predict group membership for data instances [4]. The goal of the classification is to assign a class to find previously unseen records as accurately as possible. Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules [9].

There are various kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs [7].

Our goal is to use the publicly available dataset heart disease, and use PART and K-Means data mining algorithms to predict about heart disease, analyses the results and use the rules generated by these algorithms for further predictions. The rest of this paper is organized as following. Section II provides a review of literature. The problem definition is given in Section III. Subsequently, our proposed approach is discussed in Section IV. The experimental results are given in Section V. Finally, Section VI gives the conclusion and future work.

II. Related Works

A classification rule or classifier is a function that can be evaluated for any possible value specifically given the data it will yield a similar classification. In a binary classification, the elements that are not correctly classified are named false positives and false negatives [12]. Some classification rules are static functions. There are various classification rule algorithms namely OneR, Ridor, Conjuctive Rule etc. There are two types in extracting classification rules namely direct method and

Author α: Assistant Professor of Computer Science, Department of Physics, Govt. PG Science College, Rewa (M.P.) India. e-mail: atul.pandey.it2009@gmail.com.
Author σ: Additional Directorate, Higher Education, Division Rewa (M.P.)-India. e-mail: prabhatpandey51@gmail.com.
Author ρ: Assistant Professor and In charge of BCA, DCA & PGDCA, Department of Physics, Govt. PG Science College, Rewa (M.P.) India. e-mail: drkanhaiyalaljaiswal@gmail.com

indirect method. In direct method the rules are extracted from data [5]. In indirect method the rules are extracted from other classification models. The classification rules are also known as if then rules.

In [1], the author proposed enhanced K-Means clustering algorithm for predicting coronary heart disease. There are two strategies are used for enhancing K-means clustering algorithm. First the author proposed weighted ranking algorithm to overcome the problem of random selection of initial centroids. Second the attributes associated with weights concerned by the physicians are taken into account in both ranking and the K-means algorithm instead of assigning unit weight to all the attributes. The heart dataset was collected from UCI machine learning repository. Moreover 35 conditions are carried out to assign weights to attributes. This paper describes about the rule based classification algorithm namely Part and Simple K-Means clustering algorithm. In this paper we review about the role of those two algorithms in various concepts.

III. Problem Definition

Given a dataset D, a set of classes C, a set of classification rules R over D through the algorithms K-Means, Part and Part based on K-Means, find the best algorithm using some the performance factors.

IV. Proposed System

In the proposed system a clear view of the two algorithms is given. This section discusses a brief description of the two data mining algorithms.

a) K-Means Clustering Algorithm

Clustering the medical data into small with meaningful data can aid in the discovery of forms by supporting the abstraction of several suitable features from each of the collections thereby introducing party into the data and helping the application of orthodox data mining techniques. The k-means is the simplest, most commonly and good behavior clustering algorithm used in many applications [3, 6]. The simplicity is due to the use of squared error as the stopping criteria, which tends to work well with isolated and compact clusters. Its time complexity depends on the number of data points to be clustered and the number of iteration. The K mean algorithm works on the Euclidian Distance Method, is initialized from some random or approximate solution.

K-means groups the data in accord with their individual values into k distinct collections. Data categorized into the identical cluster have a like feature values. K, the positive number representing the number of collections, needs to be delivered in advance. The phases convoluted in a k-means algorithm are given consequently:

Prophecy of heart disease using K – Means clustering techniques

- K points denoting the data to be bunched are positioned into the space. These points signify the primary collection centroids.
- The data are consigned to the group that is nearby to the centroids.
- The points of all the K centroids are again calculated as swiftly as all the data are allotted.
- Steps 2 and 3 are repeated until the centroids stop affecting any further. This results in the isolation of data into groups from which the metric to be diminished can be reflected.

The preprocessed heart illness data is grouped using the K-means algorithm with the K values. Clustering is a type of multivariate statistical examination also known as cluster analysis, unsupervised classify-cation analysis, or numerical taxonomy. K-Means clustering produces a definite number of separate, flat (non-hierarchical) clusters.

b) Classification Rule Based PART Algorithm

Classification is a concept or process of finding a model which finds the class of unknown objects. It basically maps the data items into one of the some predefined classes. Classification model generate a set of rules based on the features of the data in the training dataset. Further these rules can be use for classification of future unknown data items. Classification is the one of the most important data mining technique. Medical diagnosis is an important application of classification for example; diagnosis of new patients based on their symptoms by using the classification rules about diseases from known cases.

PART stands for Projective Adaptive Resonance Theory. The input for PART algorithm is the vigilance and distance parameters [13].

i. *Initialization*

Number m of nodes in F1 layer:=number of dimensions in the input vector. Number m of nodes in F layer: =expected maximum number of clusters that can be formed at each clustering level.

Initialize parameters L, ρo, ρh, σ, α, θ, and e.

1. Set ρ = ρo.
2. Repeat steps 3 – 7 until the stopping condition is satisfied.
3. Set all F2 nodes as being non-committed.
4. For each input vector in dataset S, do steps 4.1-4.6.

Compute hij for all F1 nodes vi and committed F2 nodes vj. If all F2 nodes are non committed, go to step 4.3.

a. Compute Tj for all committed F2 nodes Vj.
b. Select the winning F2 node VJ. If no F2 node can be selected, put the input data into outlier 0 & then continue to do step 4.
c. If the winner is a committed node, compute rJ, otherwise go to step 4.6.
d. If rJ >= ρ, go to step 4.6, otherwise reset the winner VJ and go back to step 4.3.

e. Set the winner VJ as the committed and update the bottom-up and top-down weights for winner node VJ.

5. Repeat step 4 N times until stable clusters are formed (i.e. until the difference of output clusters ay N-th and (N-1)-th time becomes sufficiently small).
6. For each cluster Cj in F2 layer, compute the associated dimension set Dj. Then, set S= Cj and set ρ = ρ + ρh (or ρ = | D |= ρh), go back to step 2.
7. For the outlier O, set S = 0, go back to step 2.

Fig. 1 : Algorithm for PART

V. Experimental Results

The above two algorithms are combined using dataset namely Heart Disease. These dataset are collected from UCI Repository in the website www.ucirepository.com. The heart disease dataset contains 303 instances and 14 selected attributes. Initially dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using ReplaceMissingValues filter from Weka 3.7. This process is known as Data Preprocessing. After pre-processing the data, data mining clustering and classification techniques namely Simple K-Means Clustering and PART were applied.

To measure the stability of the performance of the proposed model the data is divided into training and testing data with 10-fold cross validation. A confusion matrix shows how many instances have been assigned to each class. In our experiment we have two classes or clusters, and therefore we have a 2x2 confusion matrix. The entries of this matrix are used to explain the performance measures. The following charts and figure are based on the combined made of two algorithms namely K-Means and PART for heart disease dataset.

We are evaluating the performance of Simple K-Means algorithm Clustering using the mode of classes to clusters evaluation with the prediction attribute nom. Table 1, Table 2, Table 3 and Table 4 illustrates the confusion matrix of Simple k-means, PART, PART via Simple K-means (Classification via Clustering) and Accuracy of algorithm respectively. Results shows that 169 (56%) records are grouped into cluster 0 and 134 (44%) to cluster 1. Cluster 1 those who have heart disease and cluster 0 has no heart disease.

Table 1 : Confusion Matrix of K-Means

	Actual Class		
Predicted Class		1	0
	Cluster 1 <-- 0	27	138
	Cluster 0 <-- 1	107	31

Table 2 : Confusion Matrix of PART

	Actual Class		
Predicted Class		1	0
	a=1	131	28
	b=0	34	110

Table 3 : PART via Simple K-Means clustering

	Actual Class		
Predicted Class		b	a
	b=cluster 1	125	9
	a=cluster 0	12	157

Table 4 : Comparison of Data Mining Techniques

Classification Techniques	Time(seconds)	Accuracy %
Simple K-Means	0.02	80.858
PART	0.06	79.538
PART via K-Means	0.02	93.0693

Table 5 illustrates the number of rules created by PART algorithm without K-Means, PART based on K-Means. Figure 2 & 3 illustrates the rules generated by Part and Part with cluster relevant data where class value 0 & cluster value 1 has heart disease.

Table 5 : No. of Rules generated by Algorithm

Classification Techniques	No. of Rules
PART	26
PART via Simple K-Means Clustering	11

PART DECISION LIST

1) ca > 0.674497 AND cp = asympt AND sex = male: 0 (62.0/2.0)
2) thal = normal AND ca <= 1 AND slope = up AND fbs = f AND exang = no AND age <= 56: 1 (54.0/1.0)
3) oldpeak > 2.4 AND thal = reversable_defect: 0 (16.0/1.0)
4) thal = normal AND ca <= 1 AND slope = up AND fbs = t: 1 (10.0)
5) thal = normal AND age <= 45: 1 (16.0)
6) exang = no AND sex = female AND fbs = f AND thal = normal: 1 (32.0/3.0)
7) slope = flat AND sex = female: 0 (17.0/2.0)
8) sex = female: 1 (6.0/1.0)
9) slope = up AND fbs = t: 1 (5.0)
10) ca <= 1 AND thal = normal AND exang = yes: 1 (7.0/1.0)
11) exang = yes AND chol > 243: 0 (10.0)
12) fbs = t AND ca <= 0: 1 (5.0)
13) oldpeak > 0.7 AND slope = flat: 0 (19.0/3.0)
14) exang = yes: 1 (6.0/1.0)
15) ca <= 1 AND thal = fixed_defect: 1 (4.0)
16) ca <= 1 AND thal = normal AND slope = up AND chol <= 271: 1 (4.0/1.0)
17) ca <= 1 AND thal = normal AND slope = flat: 1 (3.0)
18) restecg = left_vent_hyper AND cp = typ_angina: 0 (3.0/1.0)
19) cp = atyp_angina AND slope = up AND trestbps > 122: 0 (3.0/1.0)
20) cp = atyp_angina: 1 (5.0/1.0)
21) ca <= 1 AND slope = flat AND ca <= 0: 1 (3.0/1.0)
22) slope = up AND thal = normal: 0 (3.0)
23) slope = up AND cp = asympt AND restecg = normal: 1 (3.0/1.0)
24) cp = asympt: 0 (2.0)
25) slope = up: 1 (2.0)
26) : 0 (3.0/1.0)

Fig. 2 : Generated Rules by PART

PART DECISION LIST ON CLUSTERED RELEVANT DATA

1) exang = yes AND num = 0: cluster1 (76.0)
2) thal = normal AND exang = no AND slope = up: cluster0 (92.0)
3) restecg = normal AND cp = non_anginal: cluster0 (23.0)
4) restecg = normal AND cp = atyp_angina: cluster0 (11.0)
5) restecg = normal AND sex = female: cluster0 (7.0)
6) cp = atyp_angina AND age <= 60: cluster0 (9.0)
7) restecg = left_vent_hyper AND cp = asympt: cluster1 (29.0)
8) age <= 53: cluster0 (17.0/3.0)
9) exang = no AND slope = flat AND thal = reversable_defect: cluster1 (15.0)
10) exang = no AND fbs = f AND oldpeak <= 3.6: cluster0 (13.0)
11) : cluster1 (11.0)

Fig. 3 : Rule Generation by PART via K-Means

Figure 4 & 5 illustrates the threshold curve of PART algorithm for class 1 & 0. We can say that the area under ROC= 0.831.

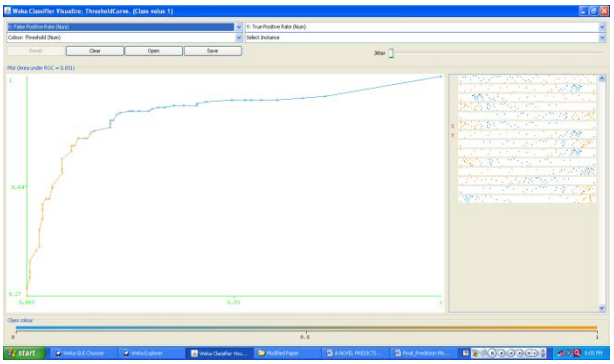

Fig. 4 : Threshold Curve of PART for Class 1

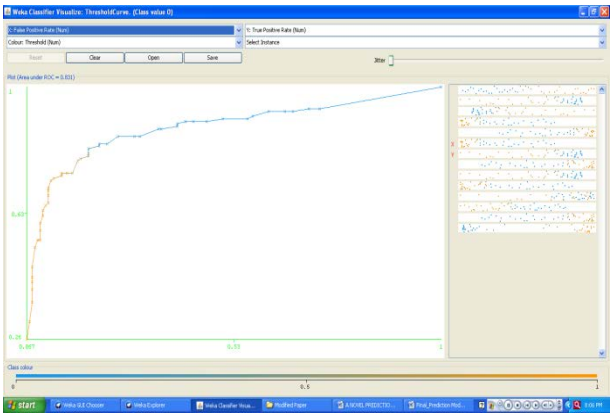

Fig. 5 : Threshold Curve of PART for Class 0

Fig. 5: Threshold Curve of PART for Class 0

Figure 6 & 7 illustrates the threshold curve of PART algorithm for class value cluster1 & cluster0. We can say that the area under ROC= 0.959.

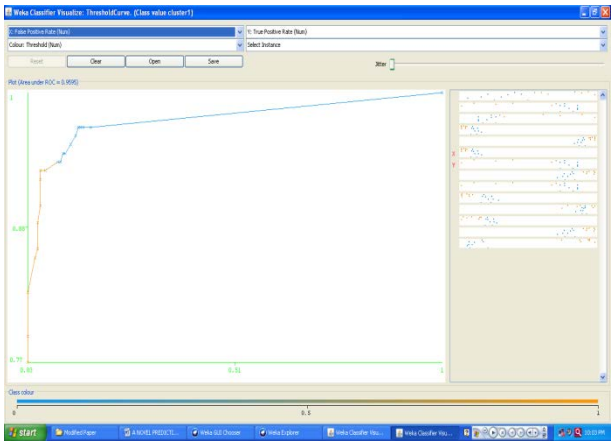

Fig. 6 : Threshold Curve of PART for Class value cluster 0

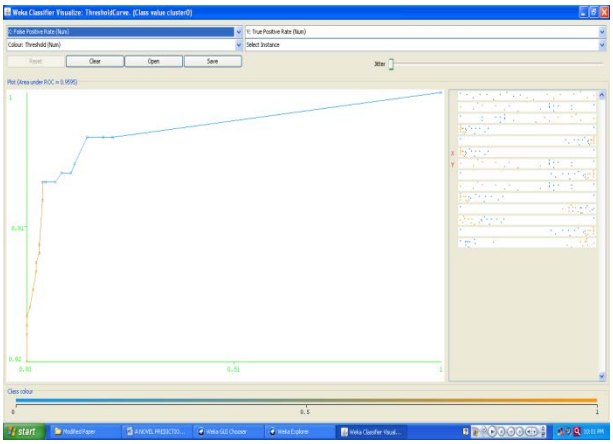

Fig. 7 : Threshold Curve of PART for Class value cluster 0

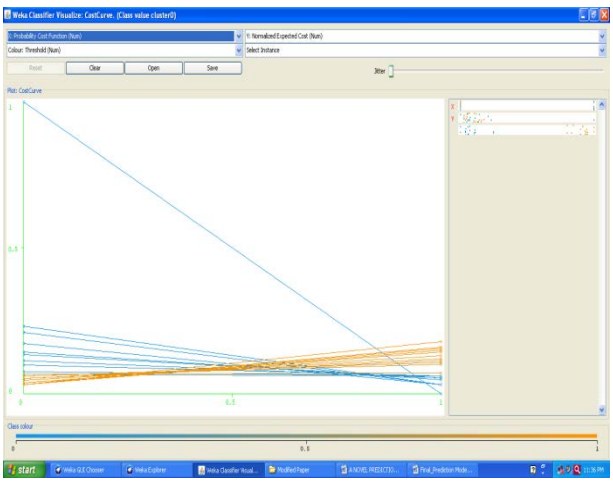

Figure 8 & 9 : Cost Curve for Class value Cluster0

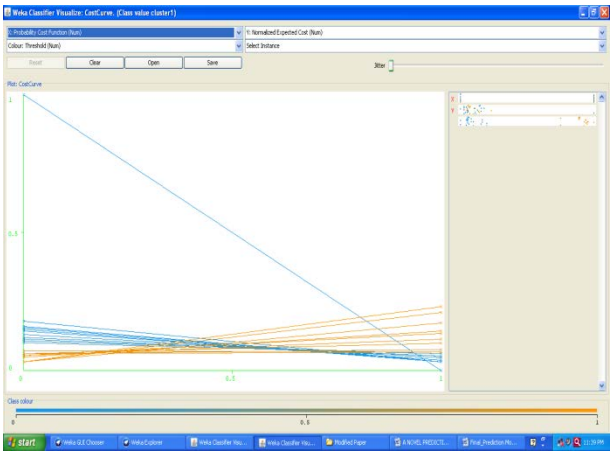

Figure 9 : Cost Curve for Class value Cluster1

VI. Conclusion and Future Work

Around 18 million people 7% of the Indians are affected by heart disease. Heart disease is mostly affected the person under the age of 65. In this paper, we have compared PART and PART based on K-Means Clustering algorithms which are very suitable for generating rules in classification technique. The classification rule generation algorithms generates classification rules which is both sensitive and non sensitive. There are different data mining techniques that can be used for the identification and prevention of cardiovascular disease among patients. Our studies showed that Part based on K-Means Clustering turned out to be best classifier for cardiovascular disease prediction.

In our future work, we have planned to design and develop an efficient heart attack prediction system with Patient Prescription Support using the web mining and data warehouse techniques. New algorithms and techniques are to be developed which overcome the drawbacks of the existing system. In future some privacy preserving technique can be induced for the rule generation in the classification technique. We intend to improve performance of these basic classification techniques by creating Meta model which will be used to predict cardiovascular disease in patients.

References Références Referencias

1. R. Sumathi, E. Kirubakaran "Enhanced Weighted K-Means Clustering Based Risk Level Prediction for Coronary Heart Disease" European Journal of Scientific Research ISSN 1450-216X Vol.71 No.4 (2012), pp. 490-500 © Euro Journals Publishing, Inc. 2012.
2. Jeffrey W. Seifert, "Data Mining An Overview", CRS Report for Congress.
3. Wu, X., et al., Top 10 algorithms in data mining analysis. Knowl. Inf. Syst., 2007.
4. Jens Hühn, Eyke Hüllermeier, "FURIA: An Algorithm for Unordered Fuzzy Rule Induction", Philipps-

Universität Marburg, Department of Mathematics and Computer Science.
5. Jianyu Yang, Rutgers, "Classification by Association Rules: The Importance of Minimal Rule Sets", the State University of New Jersey, New Brunswick, NJ 08903 USA.
6. Bramer, M., Principles of data mining. 2007: Springer.
7. Minoru SASAKI, Kenji KITA,"Rule-based Text Categorization Using Hierarchical Categories", Faculty of Engineering, Tokushima University, Tokushima, Japan, 770-8506.
8. Murlikrishna Vishwanathan, Geoffrey I.Webb, "Classifi-cation Learning Using All Rules", Proceedings of the Tenth European Conference on Machine Learning (ECML '98), Springer, pp. 149-159.
9. Srivatsan Laxman, P S Sastry, "A survey of temporal data mining", Sadhana Vol. 31, Part 2, 2006, India, pp. 173– 198.
10. Herbert Diamond, Michael P. Johnson, Rema Padman, Kai Zheng, "Clinical Reminder System: A Relational Database Application for Evidence-Based Medicine Practice" INFORMS Spring National Conference, Salt Lake City, Utah-April 26, 2004
11. Sellappan Palaniappan , Rafiah Awang "Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques" Proceedings of iiWAS2007.
12. Veronica S. Moertini, Jurusan Ilmu Komputer,"Towards the use of C4.5 algorithm for classifiying banking dataset", Fakultas Matematika dan Ilmu Pengetahuan Alam universitas, Katolik Parahyangan Bandung.
13. Yongqiang Cao, Jianhong Wu, "Projective ART for clustering data sets in high dimensional spaces", Elsevier Science Ltd, Neural Networks 15, 2002, pp. 105-120.