RESEARCH ARTICLE OPEN ACCESS

Heart Disease Prediction Using the Data mining Techniques

Aswathy Wilson[1], Gloria Wilson[2], Likhiya Joy K[3]
Professor[1], Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy, Thrissur, India

ABSTRACT
Heart disease is a major cause of transience in modern society. Due to time and cost constraints, most of the people rely on health care systems to obtain healthcare services. Healthcare organizations collect and produce large volumes of data on daily basis. Information technology allows automatization of processes for extraction of data that help to get interesting knowledge and regularities. In this paper we are using the Data mining techniques like K Means and Weighted Association rule for the elimination of manual tasks and easier extraction of data directly from electronic records, transferring onto secure electronic system of medical records which will save lives and decrease the cost of the healthcare services. K-means clustering is a usually used data clustering for unsupervised learning tasks. Decision tree is used to prediction process. The WAC has been used to get the significant rule instead of flooded with insignificant relation and the Apriori algorithm is used to find out the frequent item set from the patient database

Keywords- Data Mining, patient records, frequent itemset, decision tree, association rules & clustering**.**

I. INTRODUCTION

Data mining is a process of knowledge discovery in databases. Thus data mining refers to mining or extracting knowledge from large amounts of data. The diagnosis of heart disease depends on clinical data. Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients [1]. Here we are implementing a heart disease prediction system using both weighted association classifier and K means clustering. The healthcare industry collects large amounts of healthcare data and that need to be mined to discover hidden information for effective decision making.

Working on heart disease patients databases can be compared to real-life application. Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

There are many reasons for heart disease. It includes food habit, stress, lack of exercise, high blood pressure, smoking, alcohol, drug abuse, cholesterol, fast blood sugar etc [2].

Because of the fatty food; our blood vessels became weak and this may lead to various heart disease. More pressure to our arteries can make the walls in heart more thick. It can make a block in the flow of blood and lead to heart disease. For reducing the complexity in diagnosing the heart disease we are introducing a method for predicting the heart disease.

Frequent pattern are patterns that appear in a dataset frequently. Frequent item sets play an essential role in data mining tasks that try to find interesting patterns from databases. The basic frequent item sets are collected from the huge database. The problem can be decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database and those are called frequent item sets. The second problem is to generate association rules from those frequent item sets with the constraints of minimal confidence. The first sub-problem can be further divided into two sub-problems: Candidate large item sets generation process and frequent or large item sets generation process.

A. Data Mining

Data mining has attracted a great deal of attention in society and in the information industry as a whole in recent years. It is used for turning large amount of data into useful knowledge and information. Fraud detection, customer retention and prediction, market analysis, exploration and production control science are the various services provided by data mining. Now a day's data mining viewed as a result of the natural evolution of information technology. The aim of data Mining is discovering knowledge out of data and presenting it in a form that is easily compressible to humans. Data mining is a process that is developed to examine large amounts of data routinely collected [3].

A major challenge facing healthcare organizations like hospitals, medical center etc is the provision of quality services at low costs. Quality service implies diagnosing patients perfectly and administering treatments that are helpful. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several

important and critical questions related to health care. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest.

II. RELATED WORKS

A.CRISP-DM Methodology

Classification is one of the most essential and widely used data mining techniques. Approaches for development of classifier performance can be classified according to the data mining process and distinct steps are shown in fig 1. It mainly consists of six major phases. Business understanding phase is used to analyze the business objective and it undergoes the processes like task decomposition i.e.; Break down the objective into sub-tasks, identify constraints like laws, resources etc. Large amount of data is collected, descript and explore during the data understanding phase.

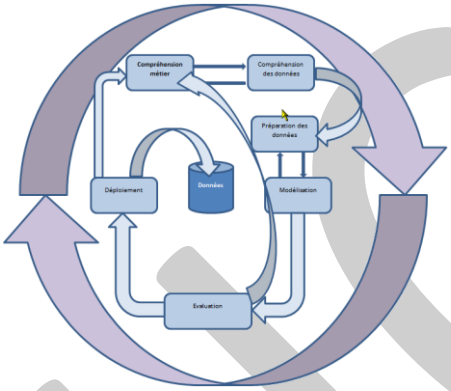

Fig 1: CRISP-DM Methodology

Selected data is integrated by joining multiple data tables and data summarization process is done in data preparation phase. In the modeling phase, appropriate modeling technique is selected and develops a testing system using sampling technique. Evaluate usefulness of results from business point of view in the evaluation phase. At last, in the deployment phase, collected data is finalized [4].

B. Weighted Association Rule

Employing association rule discovery for classification, improves the predictive correctness of classification system. Weighted Association Rule Mining uses Confidence Framework and Weighted Support to take out Association rule from data repository. Classification rule mining provides a training data set and generates a small set of rules to classify future data. To construct a classifier based on association, we use a particular subset of association rules known as Class Association Rules [5]. In Weighted Associative Classifier (WAC), different weights are assigned to different attributes according to their predicting ability. The major steps of WAC are given as follows.

1) To make the Heart disease data warehouse suitable for the mining process, initially this is pre -processed.
2) According to their importance in prediction model each attribute is assigned a weight ranging from 0 to 1. More impact attribute will be assigned a high weight and less impact are assigned low weight.
3) After pre-processing the database, Weighted Association Rule Mining (WARM) algorithm is applied to generate interesting pattern.
4) Generated rules are known as CAR. It shown as X□□ . Class label where X is set of symptoms for the disease. These rules will be stored in Rule base**.**
5) Whenever a new patients record is provide, the CAR rule from the rule base is used to predict the class label.

Weighted associative classifiers contain training dataset T= {r_1, r_2, r_3…. r_i} with set of weight associated with each {attribute, attribute value} pair. A weight w_i attached to each attribute of r_i record. Where each i^{th} record r_i is a set of attribute value.

1) Attribute Weight

In medical profile, the weights are assigned according to the priority of attributes. Here the attributes are stands for the symptoms. The task of assigning value is done by the doctors. Attribute set weight:

Weight of attribute set X is denoted by W(X) and which is calculated as the average of weigh

$$W(X) = \sum_{i=1}^{n} (\text{weight } (a_i) / \text{Number of attributes in X})$$

Record weight/Tuple Weight:

The record weight or tuple weight can be defined as type of attribute weight. And it is average weight of attributes in the corresponding tuple.

$$W(rk) = \sum_{1=|rk|}^{n} (weight(rk) / \text{no of attributes})$$

2) Weighted Support

Weighted support WSP of rule Class label, where X stands for set of non-empty subsets of attribute-value set, which is X□□ fraction of weight of the record that consist of attribute-value set relative to the weight of all transactions.

$$WSP\ (X \square\square Class_{label}) = \sum_{i=1}^{ri} \text{Weight (rk)} / \sum_{k=1}^{n} \text{weight (rk))}$$

C. Apriori Algorithm

Frequent pattern mining is a heavily researched area in the field of data mining with wide range of applications. Mining frequent patterns from large scale databases has emerged as

an important problem in data mining and knowledge discovery community [6]. In this paper we are using the Proposed Algorithm based on classical Apriori algorithm.

1. *Apriori property*

A frequent itemset can be a defined as a subset of frequent itemset i.e., if {PQ} is a frequent itemset, both {P} and {Q} should be a frequent itemset.
1. Iteratively discover frequent itemsets with cardinality from 1 to k (k-itemset).
2. Use the frequent itemsets to produce association rules. Join Step: Ck is generated by joining Lk-1with itself Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
Initialize: K: = 1, C1 = all the 1- item sets; read the database to count the support of C1 to determine L1. L1 := {frequent 1-item sets};
k: =2; //k represents the pass number//
While (Lk-1 ≠ □) do
begin
Ck: = gen_candidate_itemsets with the given Lk-1
Prune (Ck)
for all candidates in Ck do
count the number of transactions of at least
k length that are common in each item Ck
Lk := All candidates in Ck with minimum
support;
k := k + 1;
end

Key attributes:

1. Patientid – Patient's identification number

Input attributes:

1. Sex (value 1: Male; value 0: Female)
2. Chest Pain Type (value 1: typical type 1 angina, value2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value2: showing probable or definite left ventricular hypertrophy)
5. Slope – the slope of the peak exercise ST segment (value1: unsloping; value 2: flat; value 3: down sloping)
6. Exang – exercise induced angina (value 1: yes; value 0: no)
7. CA – number of major vessels colored by fluoroscopy (value 0 – 3)
8. Trest Blood Pressure (mm Hg on admission to the hospital)
9. Thal (value 3: normal; value 6: fixed defect; value7: reversible defect)
10. Cholesterol (mg/dl)
11. Oldpeak – ST depression induced by exercise relative to rest
12. Thalach – maximum heart rate achieved

Excluding these data records it also includes personal questions about the patient. They are smoking, overweight, alcoholic, daily fast food habit and going through regular exercise.

D. K-Means Clustering

K-means clustering is one of the most accepted and well known clustering techniques because of its simplicity and good behavior in many applications. For identifying the attributes that will be used in the clustering and these attributes are apparent clustering attributes for heart disease patients.
Initial centroid selection is an important matter in K-means clustering and strongly affects its results[7] .The generation of initial centroids are based on actual data points using inlier method, outlier method, range method, random attribute method, and random row method .Here we deals with inlier method.
In generating the initial K centroids using the inlier method the following equations are used:
$Ci = Min (X) – i$ where $0 \leq i \leq k$ (1)
$Cj = Min (Y) – j$ where $0 \leq j \leq k$ (2)
Where C (ci, cj) is the initial centroid and min (X) and min (Y) is the minimum value of attribute X and Y respectively. K represents the number of clusters.

E. Decision Tree

Decision tree is one of the data mining techniques showing considerable success when compared to other data mining techniques. Applying decision tree in diagnosing heart disease patients showed accuracy about 94%.The performance of Decision tree can be enhanced by the use of K-Means clustering [8].
Our paper investigates integrating K-means clustering with decision tree can improve the classifier's performance in diagnosing heart disease patients. Importantly, the research involves a logical investigation of which initial centroid selection method can provide better performance in diagnose heart disease patients.

III. IMPLEMENTATION DETAILS

By using the techniques, WAC and K-means the heart disease prediction is performed on the patient databases. The CRISP-DM methodology is used to improve the classification performance and efficiency of data retrieval. Apriori algorithm is used for finding the frequent itemsets from candidate itemset. And weights are assigned on the given itemsets to perform the prediction. The rule discovery of classification improves the predictive correctness of the classification system.

In K-means clustering, clusters are generated from the datasets through inlier method. Then prediction is performed under basis of decision tree, these two methods are effective when evaluating the results and it can be used in real time prediction process.

IV. PERFORMANCE EVALUATION

The effectiveness of models was tested using the pie chart method. The purpose was to determine which pie chart model gave the highest percentage of correct predictions for diagnosing patients with a heart disease.

The pie chart is perhaps the most widely used statistical chart in the business world and the mass media. Pie charts presented in fig 2 and fig 3 can explain clearly about the performance level of techniques, WAC and K-means by undertaking 500 patient records.

Given pie chart shows the testing result of these 500 datasets. After analyzing the result, accuracy of two methods are plotted. It shows that K-Means had better accuracy when compared to weighted association classifier.

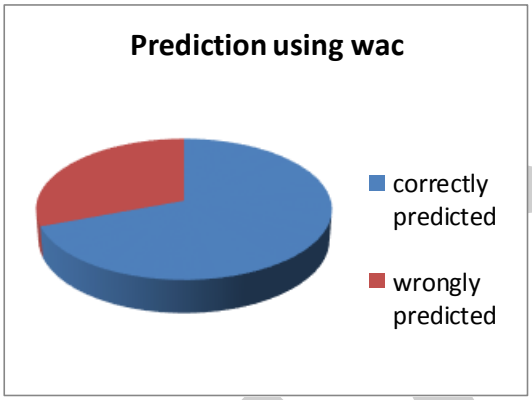

Fig 2: Prediction using WAC

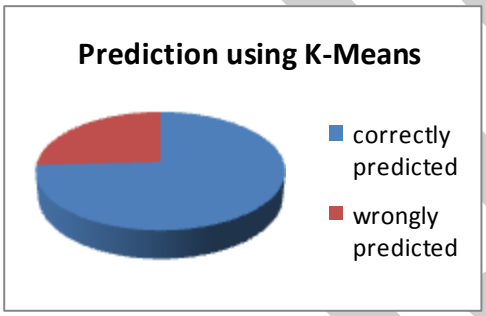

Fig 3: Prediction using K- Means

A. Screenshots

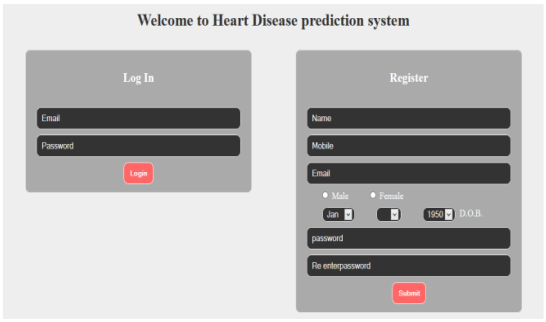

Fig 4: Login page

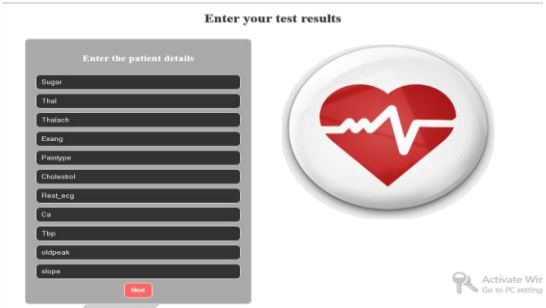

Fig 5: Data input page

Fig 6: Questionnaire page

Fig 7: Output page

V. CONCLUSIONS

The implementation paper provides a comparison study of two data mining techniques; K-means with decision tree and the weighted association classifier with Apriori algorithm over heart disease prediction system.

The weighted association classifier is a idea of weighted association rule for classification. Weighted Association Rule Mining uses Weighted Support and Confidence Framework to extract Association rule from data warehouse. Classification rule mining takes a training data set and generates a small set of rules to organize future data. The k-means clustering is the technique to cluster the attributes from the patient record. The decision tree with K-means clustering can enhance the classifier's performance in diagnosing heart disease. The initial centroid selection technique among the five K-means clustering techniques is used here because it can provide better performance over heart disease prediction. The experiments shows that

K-means with decision tree technique make the system more accurate and efficient when compared to the weighted association rule with Apriori algorithm.

REFERENCES

1. NidhiBhatla, KiranJyoti" *An Analysis of Heart Disease Prediction using Different Data Mining Techniques*," IJERT, Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181
2. K.Srinivas, G.Raghavendra ,Govardhan,"*Analysis of Attribute Association in Heart Disease Using Data Mining Techniques*, " IJERT ISSN: 2248-9622 www.ijera.com Vol. 2, Issue4, July-August 2012, pp.1680-1683
3. Jiawei Han,MichelineKamber," *Data Mining:Concepts and Techniques,"* Second Edition, University of Illinois at Urbana-Champaign
4. Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "*CRISP-DM 1.0: Step by step data mining guide*", SPSS, 1-78, 2000.W.-K. Chen.
5. N. Aditya Sundar1, P. Pushpa Latha2, M. Rama Chandra3,"*Performance analysis of classification data mining techniques over heart disease data base*," IJESAT, Volume-2, Issue-3, 470 – 478
6. Goswami D.N,Chaturvedi Anshu, Raghuvanshi C.S*," An Algorithm for Frequent Pattern Mining Based OnApriori*," (IJCSE) International Journal Vol. 02, No. 04, 2010, 942-947
7. Chris Ding ,Xiaofeng He," *K-means Clustering via Principal Component Analysis,Chris*", Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
8. Mai Shouman, Tim Turner, Rob Stocker,"*Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients",*Northcott Drive, Canberra ACT 2600