ISSN 2320 - 2602
Volume 3, No.2, February 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst11322014.pdf

Data Mining Techniques For Heart Disease Prediction

Aswathy Wilson[1], Jismi Simon[2], Liya Thomas[2],Soniya Joseph[2]

[1] Professor, Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy, Thrissur,India
aswathy@jecc.ac.in

[2]Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy, Thrissur,India
,jismysymon22@gmail.com,liyathomas92@gmail.com,soniyajoseph111@gmail.com

ABSTRACT

Data mining is the process of extracting knowledge from hidden information.Today, heart disease is leading cause of death.There exists several data mining techniques to help in diagnosis of heart disease. This paper shows the comparison study of different data mining Algorithms such as K-Means Clustering with Decision Tree, WAC (Weighted Associative Classifier) with Apriori algorithm and Naive Bayes. Different attributes such as age, sex, blood pressure and blood sugar are used in these algorithms to predict the chance of getting a heart disease. The comparison study indicate that K-means clustering with decision tree offers high accuracy.

Keywords: Data mining, WAC, K- Means clustering, Decision tree, Naive Bayes, Apriori.

1.INTRODUCTION

Today heart diseases are most common disease in our society. Around 60% of our population is suffering from heart disease because of their modern daily life[3]. The main reasons for heart disease are food habit , stress , lack of exercise , high blood pressure , diabetes , smoking , alcohol , drug abuse , attack of bacteria , virus and parasites[4]. For reducing the complexity in diagnosing the heart disease, we are introducing a system for predicting the heart disease based on Data Mining[11].

Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining software is one of major analytical tool for analysing data. It allows users to analyse data from different dimensions categorize it, and then summarize the relationships identified. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[1]. Data mining consists of three technical terms such as data, information and knowledge. Data may be facts, numbers, or text that can be processed by a computer. Now, organizations are accumulating large amount of data in different formats and different databases. The patterns, associations, or relationships, provide information.Informationcanbeconvertedinto knowledge about patterns[2].

By utilizing the prediction capabilities of data mining techniques, it can be applied to the medical areas. It will provide services at reasonable costs. Quality service means diagnosing patients correctly and providing treatments that are effective[5]. Health care systems typically generate vast amount of data, which is in the form of numbers, text, charts and images [7].Through this paper we can turn data into useful information that can enable healthcare practitioners to make better clinical decisions.

K-means clustering is one of the most popular and well known clustering techniques. This paper investigates integrating k-means clustering using different initial centroid selection methods with decision tree in the diagnosis of heart disease patients.Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease patients.The research investigates enhancing decision tree performance in the diagnosis of heart disease patients through integrating clustering as a pre-processing step to decision tree classification.

In Weighted Associative Classifier (WAC), different weights are assigned to different attributes according to their predicting capability. Weighted Associative Classifier (WAC) is a new concept that uses Weighted Association Rule for classification.Apriori is an algorithm proposed for mining frequent item sets.Naive Bayes is the basis for many machine-learning and data mining methods. This algorithm is used to create models with predictive capabilities.

2. METHODOLOGY

The following section will discuss the techniques that are used in this paper.

2.1 Weighted Associative Classifier

This is one of the methods that use Weighted Association Rule for classification. Weighted Association Rule Mining uses Weighted Support and Confidence Framework to extract Association rule from data. In Weighted

Associative Classifier (WAC), different weights are assigned to different attributes according to their predicting capability. The major steps of WAC is given as follows.

1) Initially, the heart disease data warehouse is pre-processed in order to make it suitable for the mining process.
2) Each attribute is assigned a weight ranging from 0 to 1 to according to their importance in prediction model . Attributes that have more impact will be assigned a high weight and attributes having less impact are assigned low weight.
3) Once the pre-processing gets over, Weighted Association Rule Mining (WARM) algorithm is applied to generate interesting pattern.
4) Rules generated in this step are known as CAR. It is represented as X Class label where X is set of symptoms for the disease. These rules will be stored in Rulebase.
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 consist of training dataset T= {r1, r2, r3…. ri…} with set of weight associated with each {attribute, attribute value} pair. Each ith record ri is a set of attribute value and a weight wi attached to each attribute of ri record.

Attribute weight:

In medical profile, attributes are stands for the symptoms and the weights are assigned according to their priority. assigning value can be done by the doctors.

Attribute set weight:

Weight of attribute set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute.

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

Record weight/Tuple Weight*:*

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

$$W(rk) = \sum_{1=|rk|}^{n} (\text{weight (rk)}) / \text{No. of attributes in a record}). \quad (2)$$

Weighted Support:

Weighted support WSP of rule X->Class_label, where X is set of non empty subsets of attribute-value set, is fraction of weight of the record that contain above attribute-value set relative to the weight of all transactions.

$$WSP(X \rightarrow \text{class label}) = \sum_{i=1}^{ri} \text{Weight (rk)} / \sum_{k=1}^{n} \text{weight (rk)}) \quad (3)$$

2.1.1 Apriori Algorithm

Based on Apriori algorithm there are three different frequent pattern mining approaches such as Record filter, Intersection and Proposed Algorithm. Apriori is an algorithm for mining frequent item sets for boolean association rule. Apriori shows an iterative approach known as level wise search, where k item set are used to explore (k+1) item sets.

There are two steps in each iteration. The first step generates a set of candidate item sets. Then in second step we count the occurrence of each candidate set in database and eliminate all disqualified candidates, that means all infrequent item sets. Apriori uses two pruning technique, first technique is based on support count and the second for an item set to be frequent, all its subset should be in last frequent item set. The algorithm is based on the closure property of frequent item sets. If a set of item is frequent then all its proper subsets are also frequent.

In record filter approach, when we count the support of candidate set of length k, we also check its occurrence in transaction whose length may be greater than, less than or equal to the k. Here we count the support of candidate set only in the transaction record whose length is greater than or equal to the length of candidate set because the candidate set of length k, cannot exist in the transaction record of k-1.

In Intersection approach we use the set theory concept of intersection. To calculate the support, we count the common transaction that contains in each elements of candidate set, by using the intersect query of SQL.

In proposed approach, we are presenting an algorithm that uses the concept of both algorithm i.e. Record filter approach and Intersection approach in Apriori algorithm .we use the set theory concept of intersection with the record filter approach.. In proposed algorithm, to calculate the support, we count the common transaction that contains in each elements of candidate set, using the intersect query of SQL. In this approach, we have applied a constraints that we will consider only those transaction that contain at least k items, not less than k in process of support counting for candidate set of k length. This approach requires very less time as compared to all other approaches[5].

2.2K-means Clustering Algorithm

K-Means Clustering Algorithm is most popular and well known clustering technique. The Simplicity and good behaviour of this technique made it popular in many applications. Initial centroid selection is a leading issue in k-means clustering and strongly affects its results. K-

means clustering using different initial centroid selection method such as range, inlier, and outlier, random attribute values, and random row methods in the diagnosis of heart disease patients[14].

Main steps used in K-means Clustering is,

1. Identify the attributes that will be used in clustering.

2. Identify the number of clusters.

3. Apply the Initial Centroid selection using Inliers' method.

4. Assign each of the training instances to the cluster for which it is nearest to the centroid using Euclidean distance.

5. Recalculate the centroid of the k clusters.

6. Repeat 4, 5 until the centroid do not change.

2.2.1Centroid Selection Methods

There are a number of centroid selection methods are available. It includes inlier method, outlier method, Random method, Row method, Random row method. Among these technique Inlier method with two clusters is most popular and best method.

Inlier Method:

In generating the initial K centroids using inlier method the following equation are used:

C_i=Min(X)-i where 0<=i<=k (4)

C_j=min(Y)-j where 0<=j<=k (5)

Where the initial centroid is C (c_i, c_j) and min (Y) are the minimum value of attribute X, and attribute Y respectively. K represents the number of clusters.

2.2.2Decision Tree

Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease patients. Less research have been done on enhancing decision tree performance in disease diagnosis. The gain ratio tree type is used in decision tree .This ratio is based on the entropy approach, which maximizing the information gain. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula:

$E = \sum_1^k pi \log pi$ (6)

Where k is the number of classes of the target attribute.Pi is the number of occurrences of class i divided by the total number of instances.

2.3. Naive Bayes

Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule is used to create models with predictive capabilities. It provides new ways of exploring and understanding data[15].

Algorithm:

1. Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

2. Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive probability assigns an unknown sample X to the class Ci

if and only if:
P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= i
Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,
P(Ci/X)= (P(X/Ci)P(Ci))/P(X) (7)

3. As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples.

3. PERFORMANCE EVALUATION

The analysis of different methods can be evaluated through performance evaluation phase. The purpose was to determine which model gave the highest percentage of accuracy in predictions for diagnosing patients with a heart disease.

For analysing the accuracy of these algorithms we consider 54 patient records. Out of these records,44 records can be predicted correctly by the integration of K-Means Clustering with Decision tree and 42 records can be predicted correctly by WAC with apriori algorithm. Similarly 39 records can be predicted by Naïve Bayes.

When comparing Decision tree with k-means clustering and other data mining techniques applied on the records, it enhance the accuracy in diagnosing heart disease. The accuracy of different algorithms can be shown below Table 3.1.The corresponding bar chart is shown in the figure 3.1.

Table 3.1 : Accuracy of Different techniques

Sl.no	Techniques	Accuracy
1	Naïve Bayes	81%
2	Decision tree with K-Means Clustering	83.24%
3	WAC(Weighted Associative Classifier) with Apriori Algorithm	74%

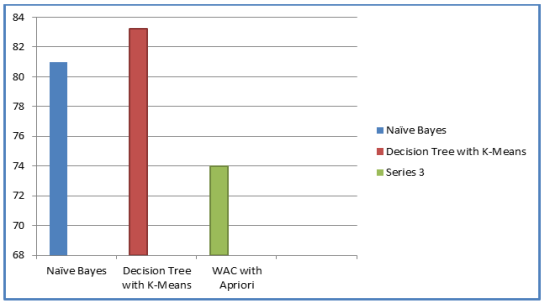

Figure 3.1 Accuracy of Different techniques

4.CONCLUSION

This paper is a comparison study of different data mining algorithms such as K-means clustering with Decision tree, WAC with Apriori Algorithm,Naive Bayes. K-Means algorithm is a clustering method where large data set is partitioned into various clusters.it evaluates continuous values.WAC is used for classifying the data set and it evaluates discrete values. Apriori algorithm is used to find the frequent itemset. Naivebayes rule is used to create models with predictive capabilities.The gain ratio decision tree is used in the decision tree algorithm. Several attributes such as age, sex, blood pressure, blood sugar are used in these algorithms. The comparison study indicate that the K-means clustering with decision tree should have high accuracy than other algorithms. The best accuracy is achieved by inlier method with two clusters. How ever increasing the no of clusters more than three clusters decreases their accuracy in the diagnosis of heart disease.

REFERENCES

[1]. 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.

[2] Obenshain , M.K: *"Application of Data Mining Techniques to Healthcare Data"*, Infection Control and Hospital Epidemiology, 25(8), 690–695, 200.Medical Journal of Australia,2003.

[3] RakeshAgrawalRamakrishnanSrikant*"Fast algorithms for mining association rule in large database"*, IBM AlmadenResearchcenter.SanJose,California,June 2004

[4]. Fayyad, U: *"Data Mining and Knowledge Discovery in Databases: Implications for scientific databases"*, Proc. of the 9th IntConf on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[5] Goswami D.N., ChaturvediAnshu. Raghuvanshi:*"An Algorithm for Frequent Pattern Mining Based On Apriori"*In Computer Science Jiwaji University Gwalior Computer Application Department MITS Gwalior,GICTS College of Professional Education Gwalior,1999.

[6]. Mehmed, K.: *"Data mining: Concepts, Models, Methods and Algorithms"*, New Jersey: John Wiley, 2003.

[7]. Han, J., Kamber, M.: *"Data Mining Concepts and Techniques"*, Morgan Kaufmann Publishers, 2006.

[8]. Charly, K.:*"Data Mining for the Enterprise"*, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[9] Lee, I.-N., S.-C. Liao, and M. Embrechts,:*"Data mining techniques applied to medical information"*, Med. inform, 2000

[10]Thuraisingham, B :*"A Primer for Understanding and Applying Data Mining"*, IT Professional IEEE, 2000.

[11]Porter, T. and B. Green:*"Identifying Diabetic Patients, A Data Mining Approach*", Americas Conference on Information Systems, 2009.

[12]Panzarasa, S., et al.:*" Mining techniques for analyzing stroke care processes*", Proceedings of the 13th World Congress on Medical Informatics, 2010.

[13]Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark R:, *"Data mining techniques for cancer detection using serum proteomic profiling, Artificial Intelligence in Medicine"*, Elsevier, 2004.

[14]Das, R., I. Turkoglu, and A. Sengu:,*"Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications*", Elsevier, 2009. 36 (2009): p. 7675–7680.

[15]Andreeva, P:*"Data Modelling and Specific Rule Generation via Data Mining Techniques*", International Conference on Computer Systems and Technologies - CompSysTech, 2006.