A Noval Clasiification And Prediction Algorithm For Heart Disease Identification

P.DEEPIKA[1], DR. S. SASIKALA[2], S.SARANYA[3], A.KIRUTHIKA[4]
[134] (ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, HINDUSTHAN COLLEGE OF ARTS AND SCIENCE, COIMBATORE.)
[2] (ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, HINDUSTHAN COLLEGE OF ARTS AND SCIENCE, COIMBATORE)
Corresponding Author: P.DEEPIKA

ABSTRACT:*Data mining plays the importantrole in research that is more popular in health organization. Data mining gives an effective participation for uncovering new trends in healthcare organization which is helpful for all the parties associated with this field. Heart disease is the important cause of death in the world over the past 10 years. Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the irregular health condition that directly affects the heart and all its parts. The healthcare industry gathers enormous amount of heart disease data which are not "mined" to discover hidden information for effective decision making. Data mining techniques are useful for analyzing the data from many different dimensions and for identifying relationships. This paper explains the proposed work in classify and predict the disease with high accuracy.*

KEYWORDS: *Data mining, Classification, Decision tree, Sequential Covering Strategy, Predictive Mining*

Date of Submission: 21-01-2018 **Date of acceptance: 05-02-2018**

I. Introduction

Data mining is one of the most vital and motivating area of research. And the aim of using data mining is finding meaningful information from the high volume data sets. Now a day, Data mining provedthat it is more effective in healthcare field because there is a need of efficient analytical methodology for sensing unknown and valuable information in health data. Data mining tools are used to perform data analysis in the data set. Data mining is a more appropriate tool to assists physicians in detecting the diseases by gaining knowledge and information concerning the disease from patient's data. The researchers in the medical field identify and predict the diseases besides proffering effective care for patients with the aid of data mining techniques. One of the major techniques used in heart disease classification is the classification and clustering tasks. The data mining techniques have been utilized by a wide variety of works in the literature to diagnose heart oriented diseases with the following dataset: Heart SPECT, Statlog datasets. Information associated with the disease, prevailing in the form of electronic attributes which taken from electronic dataset, images and more.Several data mining techniques were utilized by several researchers to present prediction and diagnosis approaches for heart diseases. The analysis of different data mining techniques that can be employed in automated heart disease prediction systems. The existing classification algorithms are suffered from the need of large datasets for accurate diagnosis. The decision trees were used for diagnosis, but the problem is the selection of attributes for fast classification.

II. Design And Methodology

The followings are the main contributions of the proposed work.

- The system implements a new decision tree based algorithm with the use of effective forest data structures. The system introduces a new Heart Disease Classification algorithm with sequential covering technique.
- This also creates a new advanced decision tree structure for fast disease classification. The system developed with the intension of high accuracy and less training overhead.
- So the system initially collects and make score for every label, this partially makes an ensemble approach to improve the detection speed.

III. Object Recognition

The proposed system C4.5 based HDC system has the following procedure and introduction about the algorithms and methodologies. This chapter discuss about the algorithms and methodologies. The important process of Heart Disease Classification is the analysis of patterns and grouping those into different subset. Classification is a process which partitions a given dataset into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups. Classification plays an important role in health care field. With the recent advances of medical technology, there has been tremendous growth of the health care equipped data. Identifying co-regulated features to organize them into meaningful groups has been an important research. Therefore, Classification and disease prediction using data analysis has become an essential and valuable tool in advanced data analysis.

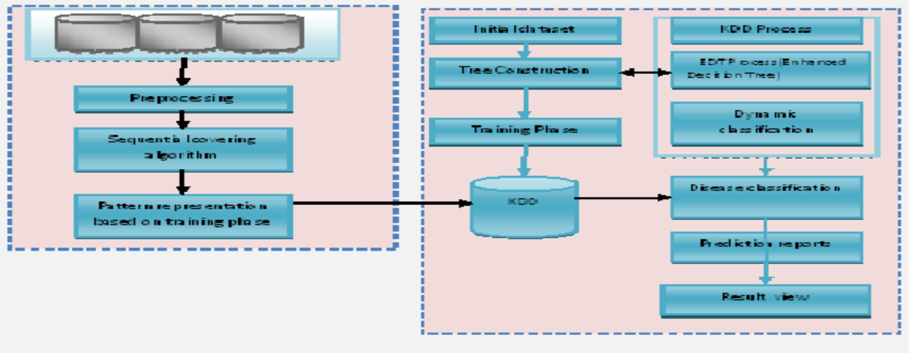

Fig.1:Framework for the Proposed work using Sequential Covering Algorithm

IV. Prediction Scheme

A predictive model analysis in data mining is a process by which a model is generated or selected to predict the best possibility of an outcome. The proposed system performs the prediction model based on the tree structure. The proposed system successfully analyses the heart disease based on the given training dataset. The system also predicts the score for the chance of heart disease based on the boundary calculation. The proposed system implements a semi supervised classifier which does not depend on the training dataset completely. The system performs the statistical properties to evaluate the score of every attribute. The system finally provides the prediction accuracy over the given dataset.

Enhanced Decision Tree (EDT)Model

The reason for selecting heart disease classification, disease prediction is that each dataset is considered as an important in medical domain. Therefore the primary concentration of this research work is on how to classify and predict the disease and score of possibilities to identify the diseases accurately.Based on the clustered data, the modeling function classification is achieved through a decision tree model. It is used to find the relationship between a specific variable, the target variable and other variables among the data. The outcomes are then used to predict the correct class label to previously unseen and unlabeled student and faculty objects. The algorithm applied to the set of data is decision tree therefore the result is presented in terms of a binary tree. There are many reasons of using decision tree in this application since it is relatively fast; it can be converted to simple and easy classification rule; it can be converted to SQL queries for accessing the database and it obtains similar or better accuracy in comparison with other classification techniques.

The major decisions are classified into two in the proposed decision tree model. The first problem involved in any decision making process is to derive whether or not a proposal has to be prepared. The second problem consists of determining if the proposal is considered, the approaches to be considered to satisfy the decision as illustrated in Figure2.

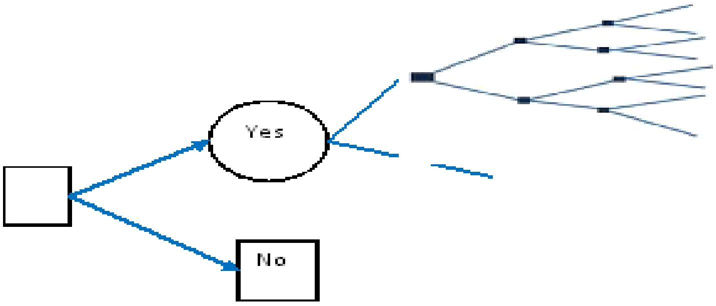

Fig.2:Decision making by analyzing decision tree

An event node in decision tree model comprises of a point where uncertainty is involved and being solved accordingly, where the decision maker notes about the occurrence of a specific event. An event node, is also referred to as a chance node, and also called as a circle. The set of events consists of the event branches which extend from the event nodes and each branch corresponds to one of the possible events that may occur at that particular point. All events must be mutually exclusive in such a way that the event assigned is in subjective probability wherein the sum of probabilities for the events in a set must be equal to one.

Algorithm: EDT

```
1. Create a node N;
2. if samples are all of the same class, C then
3. return N as a leaf node labeled with the class C;
4. if attribute-list is empty then
5. return N as a leaf node labeled with the most common class in samples;
6. select test-attribute, the attribute among attribute-list with the highest information gain;
7. label node N with test-attribute;
8. for each known value ai of test-attribute
9. grow a branch from node N for the condition  test-attribute= ai;
10. letsi be the set of samples for which test-attribute= ai;
11. ifsi is empty then
12. attach a leaf labeled with the most common class in samples;
13. else attach the node returned by
Generate_decision_tree(si,attribute-list_test-attribute)
```

Decision trees are usually unvaried since they use based on a single feature at each internal node. Most decision tree algorithms cannot perform well with problems that require diagonal partitioning. The proposed system overcomes the above problem by applying an optimal tree construction technique.

Algorithm: HDC_EDT

```
Input: training dataset D
 Test samples Ts.
Output: class

Steps:
1.      Read the training dataset D.
2.      For each attribute A and instances I do
3.      Construct the tree using the A and I
4.      Calculate splitting criteria score for each attribute A in dataset D
5.      Calculate the mean and variance for each attribute A and class C.
▪                   Mean(A(I))
▪                   Variance(A(I))
6.      Calculate the score for every attribute and reconstruct the tree structure.
7.      Based on the score perform sequential covering algorithm
8.      Update the training process
9.      Read the test samples and Match the data with the training sample with higher value
10.     Read the score and find the class
11.     Predict the value by applying mutation process
12.     Return the percentage and class as result.
```

This study proposes a new sequential covering strategy for EDT classification algorithms to mitigate the problem of rule interaction, where the order of the rules is implicitly encoded as pheromone values and the search is guided by the quality of a candidate list of rules.This study presents a discussion of the strategy used by EDT_SC classification algorithms to build a list of classification rules and proposes a new strategy that mitigates its potential disadvantages from the existing decision tree algorithms. In particular the system improves the search performed by the EDT algorithm using the quality of a candidate list of rules as a input which represented by pheromone values.

Heart Disease Data SET:

The data used in this study is the Cleveland Clinic Foundation. Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 13 attributes. Consequently, to allow comparison with the literature, we restricted testing to these same attributes.

Patient database is datasets collected from Cleveland Heart Disease Dataset (CHDD) available on the UCI Repository [11]. The 13 attributes considered are age: age, sex, chest pain type, trestbps (resting blood pressure), chol (serum cholesterol in mg/dl), FBS (fasting blood sugar > 120 mg/dl), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise induced angina), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), and CA (number of major vessels (0-3) colored by fluoroscopy). There are a total of 303 patient records in the database.

V. Data Preprocessing

This phase includes extraction of data from the statlog Heart Disease Dataset in a uniform format. The step involves transforming the data, which involves removal of missing fields, normalization of data, and removal of outliers. Out of the 272 available records, 2 tuples have missing attributes. These have been excluded from the data set. For proposed HDC_EDT, data points were automatically centered at their mean and scaled to have unit standard deviation. No changes need be made to the data sets for decision trees or logistic regression.

Training the Models

The dataset has been trained using EDT methods. For decision trees, a node splitting criterion is required so this performs the score calculation function. The best split is one that splits the data into distinct groups. Here Purity is a measure used to evaluate a potential split. A split that divides an attribute into two distinct classes is the most pure. There are many different splitting criterions that can be used in literature; the proposed system uses a statistical deviance.

EDT:

The proposed advance tree with sequential covering algorithm is used the classification technique to predict a response to data. In EDT classification is used when the features are grouped. The proposed EDT Decision tree is one of the main methods in the proposed system. A decision tree is made of a root node, branches, and leaf nodes with the use of forest feature Decision trees must be created using a purity index which will split the nodes as discussed in the training section with higher priority and score. For the statlog dataset, each of the 270 tuples is evaluated down the decision tree and arrives at a positive or negative evaluation for heart disease classification and prediction.

EDT with Predictive Mining

To overcome some hurdles in predictive modeling over heart disease dataset, decision trees are employed here. The decision tree helps to arrange the data or extract the outcome of the predictive model. Decision Trees for predictive mining at the same time contribute to the results for classification as it accommodates large number of rules. In order to reduce the large rule matching problem the system proposed a rule priority calculation.The primary step in constructing a decision tree for institutional data consisting of statlog dataset is to gather set of data values that decision tree is to examine. These sets of data values are known as the training dataset as they are used by the decision tree to discover how the value of a target variable is related to the values of predictor variables.

The decision tree for predictive mining for healthcare data using benchmark has been improved using the following factors.

- Predictive accuracy is significantly enhanced by having four times those of numerous records.
- The decision tree is intended to switch almost an unrestricted number of records as it is moderately possible to examine datasets with millions of records, though the computation time is reduced in the proposed system.
- Modeling of data is applied. Finally, learning any number of rules is possible using the C#.Net platform that is used in the research for presenting the results.

VI. Results And Comparison

This proposed work was implemented using C#.net. The performance of this proposed work EDT_SCA Scheme was compared with the existing C4.5.

Classification Accuracy Comparison Chart:

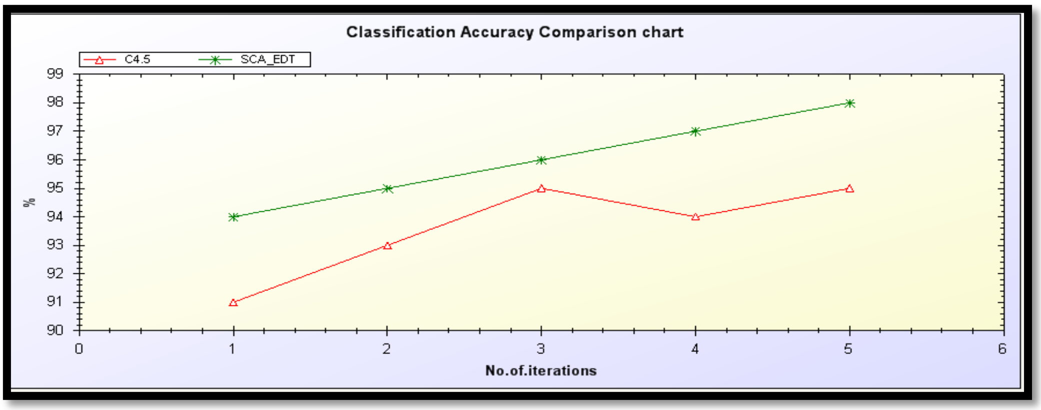

Fig.3: Performance comparison of proposed EDT_SCA with existing C4.5 approaches based on prediction efficiency.

Prediction Accuracy Comparison Chart:

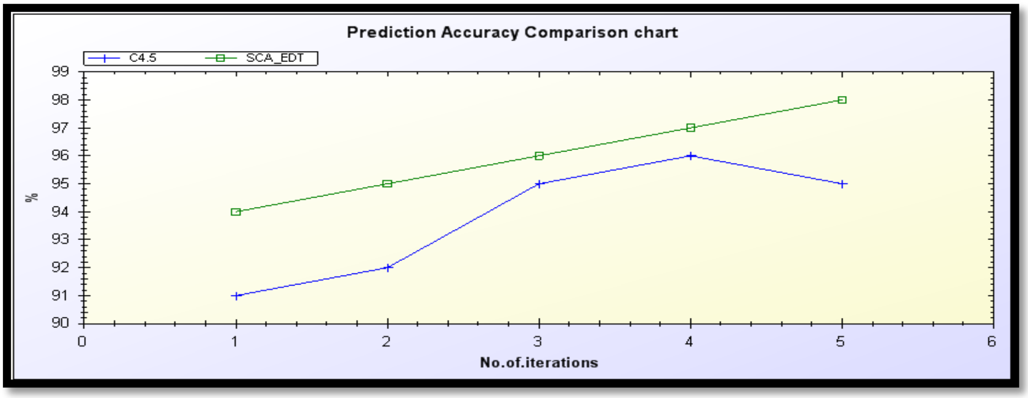

Fig.4:Performance comparison of proposed EDT_SCA with existing C4.5 approaches based on prediction accuracy.

VII. Conclusion And Future Work

The study proposed a new classification and prediction scheme for heart disease data. The system studied the main two problems in the literature, which are prediction accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced decision tree with sequential covering algorithm. The EDT represents with the effective splitting criteria which has been verified by the sequential covering algorithm. The system performs pre pruning and post pruning to eliminate irrelevant results. The system effectively identifies the disease and its sub types, the sub type which is referred as the percentage of class such as normal and disease.The experimental results are evaluated using the C#.net. The experimental result shows that integrated extended decision tree with sequential covering algorithm shows better quality assessment compared to traditional C4.5 techniques. From the experimental results, the execution time calculated for classification object is almost reduced than the existing system.

Future work:

The proposed framework model can be used to analyze the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of research in the field of data mining in other dataset and use of other classification algorithms. As further work, use this model as a functional base to develop an appropriate data mining system for classification performance.

References:

[1]. I. Kononenko, “Machine learning for medical diagnosis: History, state of the art and perspective,” *Artif. Intell.Med.*, vol. 23, no. 1, pp. 89–109, 2001.

[2]. Moustakis, V. and Charissis, G. (1999). "Machine learning and medical decision making". In Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence.

[3]. Vrushali Y Kulkarni, Aashu Singh and Pradeep K Sinha, “An Approach towards Optimizing Random Forest using Dynamic Programming Algorithm”.

[4]. Gordan.V.Kass(1980). An exploratory Technique for inverstigation large quantities of categorical data Applied Statics, vol 29, No .2, pp. 119-127.

[5]. Wei-Yin Loh, “Classification and regression trees”.

[6]. Quinlan J. R. (1986). Induction of decision trees. Machine Learning, Vol.1-1, pp. 81-106.

[7]. Zhu Xiaoliang, Wang JianYanHongcan and Wu Shangzhuo(2009) Research and application of the improved algorithm C4.5 on decision tree.

[8]. Rinkal Patel and RajanikanthAluvalu, “A Reduced Error Pruning Technique for Improving Accuracy of Decision Tree Learning”.

[9]. S. M. Kamruzzaman and Ahmed RyadhHasan, “Pattern Classification using Simplified Neural Networks with Pruning Algorithm”.

[10]. V´ıctor Soto, Gonzalo Mart´ınez-Mu˜noz, Daniel Hern´andez-Lobato, and Alberto Su´arez, “A double pruning algorithm for classification ensembles”.

[11]. C. Aflori, M. Craus, “Grid implementation of the Apriori algorithm Advances in Engineering Software”, Volume 38, Issue 5, May 2007, pp. 295-300.

[12]. Srinivas, K., “Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques”, IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[13]. Shantakumar, B.Patil,Y.S.Kumaraswamy, “Predictive data mining for medical diagnosis of heart disease prediction” IJCSE Vol .17, 2011

[14]. NidhiBhatla and KiranJyoti, “A Novel Approach for Heart Disease Diagnosis using”.

[15]. Data Mining and Fuzzy Logic

[16]. VikasChaurasia and Saurabh Pal, “Early Prediction of Heart Diseases Using Data Mining Techniques”.
Chaitrali S. Dangare and Sulabha S. Apte, “A Data Mining Approach For Prediction Of Heart Disease Using Neural Networks”.

[17]. E. E. Tripoliti, D. I. Fotiadis, and G. Manis, “Modifications of random forests algorithm,” *Data Knowl. Eng.*, to be published.

[18]. JieGu, “ Random Forest Based Imbalanced Data Cleaning and Classification”.

International Journal of Engineering Science Invention (IJESI) is UGC approved Journal with Sl. No. 3822, Journal no. 43302.

P.DEEPIKA “A Noval Clasiification And Prediction Algorithm For Heart Disease Identification” International Journal of Engineering Science Invention (IJESI), vol. 07, no. 01, 2018, pp. 12–17.