RESEARCH PAPER | Engineering | Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

A Noval Classification and Prediction Algorithm for Cardio Vascular Disease Identification

KEYWORDS

Seranmadevi V Annamalai	P Vinothini	S Vinitha
Student, Department of Information Technology, Kathir College of Engineering, CBE	Student, Department of Information Technology, Kathir College of Engineering, CBE	Student, Department of Information Technology, Kathir College of Engineering, CBE

ABSTRACT *The accurate diagnosis of diseases with high dominance disease sets, such as breast cancer, diabetes and heart diseases is one of the most important biomedical problems, which cause severe problem. And this is very tedious to classify and predict in the earlier stage. This paper aims at developing a classification and prediction models for heart disease survivability and this provides the subset finding based on the classified results. This paper presents a new method for the classification and prediction of diseases based on the improvement of enhanced decision tree technique with sequential covering algorithm. The dynamic determination of the optimum attribute selection for classification is addressed. The system implements a new enhanced decision tree algorithm with the use of effective forest data structures.*

I.INTRODUCTION:

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The information that can be used to increase revenue, cuts cost, or both. In Health concern, data mining plays a significant task for predicting disease. Data mining is a discipline to realize knowledge from data bases. The data base contains a set of instances.

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. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. But the previous classifiers are failed to produce the maximum level of accuracy and fast detection of disease. 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.

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.

II. HEART DISEASE

The initial diagnosis of a heart attack is made by a combination of general symptoms and characteristic electrocardiogram (ECG) changes. An ECG is a recording of the electrical activity of the heart which have S&T curves. 29.2% of total global deaths are due to Cardio Vascular Disease according to WHO reports in the year 2003. By the end of 2020 year, CVD is expected to be the leading cause for deaths in developing countries due to life style change, work culture and food habits. Hence, we have to examine-cardiac diseases and periodically .

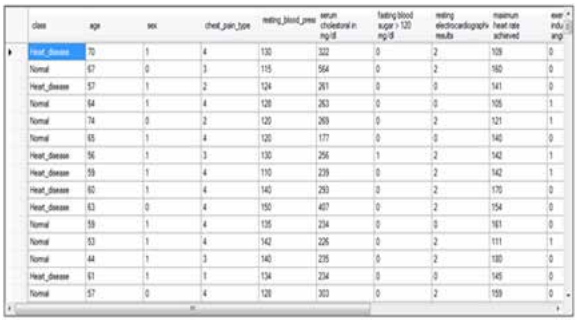
III.DATASETS

The Data set is taken from Data mining repository of University of California, Irvine (UCI). Data set from Hungary data set,Cleveland Data set, Stat logData set, long beach andSwitzerland data set are collected. Cleveland, Hungary, Switzerland and long beach data set contains 75 attributes in all. But only 13 attributes are used. Among all those Cleveland data set ,Statlog dataset,Hungary dataset are the most commonly used data set. Because other dataset has missing values.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pa 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes' criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= possible heart disease

Fig 1:Attributes Used

Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X



class	age	sex	chest_pain_type	resting_blood_press	serum cholesterol in mg/dl	fasting blood sugar > 120 mg/dl	resting electrocardiographic results	maximum heart rate achieved	exer ind ang
Heart_disease	70	1	4	130	322	0	2	109	0
Normal	67	0	3	115	564	0	2	160	0
Heart_disease	57	1	2	124	261	0	0	141	0
Normal	64	1	4	128	263	0	0	105	1
Normal	74	0	2	120	269	0	2	121	1
Normal	65	1	4	120	177	0	0	140	0
Heart_disease	56	1	3	130	256	1	2	142	1
Heart_disease	59	1	4	110	239	0	2	142	1
Heart_disease	60	1	4	140	293	0	2	170	0
Heart_disease	63	0	4	150	407	0	2	154	0
Normal	59	1	4	135	234	0	0	161	0
Normal	53	1	4	142	226	0	2	111	1
Normal	44	1	3	140	235	0	2	180	0
Heart_disease	61	1	1	134	234	0	0	145	0
Normal	57	0	4	128	303	0	2	159	0

Fig 2:Sample Data Set

IV. DATA MINING TECHNIQUES USED IN HEART DISEASE PREDICTION

Data mining techniques such as clustering, Classification, Regression and Association Rule mining are used in extracting knowledge from database. Classification is the best algorithm that is suited for Heart Disease Prediction. Among classification we have chosen Decision Tree. Some of the Decision Tree algorithms are stated below

CHID:

CHAID (CHi-squared Automatic Interaction Detector) is a fundamental decision tree learning algorithm. It is easy to interpret, easy to handle and can be used for classification and detection of interrelationship between variables. It is an extension of the AID (Automatic Interaction Detector) and THAID (Theta Automatic Interaction Detector) procedures. It bases on principal of adjusted significance testing.

CART:

Classification and Regression tree (CART) constructs binary trees which is also refer as Hierarchical Optimal Discriminate Analysis (HODA). This is a non-parametric decision tree learning technique which produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. It uses Gini index as impurity measure for selecting attribute. The attribute with the largest reduction in impurity is used for splitting the node's records.

ID3:

ID3 stands for Iterative Dichotomiser 3.The decision tree information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. By this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest.

C5.0/Sec 5:

C5.0 algorithm is an extension of C4.5 algorithm which is the extension of ID3. This is the classification algorithm which applies in big data set. It is better than C4.5 on the speed, memory and the efficiency. This model works by splitting the sample based on the field that provides the maximum information gain. This model can split samples on basis of the biggest information gain field. This sample subset that is get from the former split will be split afterward. This process will continue until the sample subset cannot be split and is usually according to another field.

IV:Proposed Algorithm-EDT
Algorithm: HDC_EDT

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.
6. Mean(A(I))
7. Variance(A(I))
8. Calculate the score for every attribute and reconstruct the tree structure.
9. Based on the score perform sequential covering algorithm
10. Update the training process
11. Read the test samples and Match the data with the training sample with higher value
12. Read the score and find the class
13. Predict the value by applying mutation process
14. Return the percentage and class as

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. This evaluates the impact of the new strategy in terms of both predictive accuracy and size of the classification model (discovered list of rules), and compare the results against the previous algorithms.

V.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. The figure below shows the results and comparison of the proposed system

Classification Accuracy Comparison Chart:

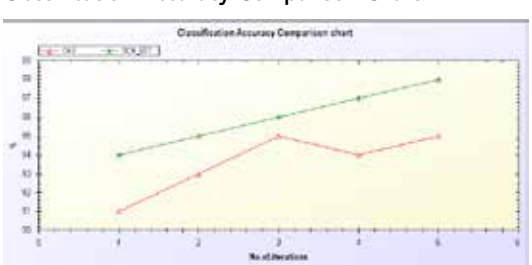


Fig: Performance comparison of proposed EDT_SCA with existing C4.5 approaches based on prediction efficiency.

Prediction Accuracy Comparison Chart:

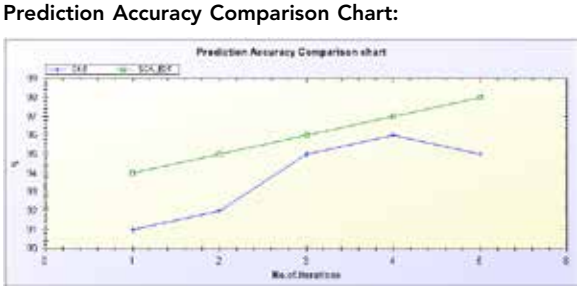


Fig: Performance comparison of proposed EDT_SCA with existing C4.5 approaches based on prediction accuracy.

Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

EVALUATION PROCESS:
Assume the following:

- A dataset contains 120 records on a particular drug
- Analysis was conducted on 120 patient details
- From the 120 test records, 85 are normal 45 were abnormal.

Calculate the precision and recall scores for the search.

Comparison Matrix/ Evaluation Parameters
The comparisons are made on the basis of the least value of Accuracy, Precision, and Recall values. In case of the two clusters based problem, the confusion matrix has four categories: True positives (TP) are modules correctly classified as faulty modules. False positives (FP) refer to fault free modules incorrectly labeled as faulty modules. True negatives (TN) correspond to fault-free modules correctly classified as such. Finally, false negatives (FN) refer to faulty modules incorrectly classified as fault-free modules.

Prediction of the best and worst drug using both the EDT_ SCA algorithms proves that SCA is a better fit for EDT. The existing algorithms are compared for higher accuracy and efficiency using the metric "Error Rate". The evaluation parameters are the correctly classified and incorrectly classified data points. Based on these parameters, the error rate is evaluated using the following formula.

Per Species,

Actual Total – Correctly Classified = Incorrectly Prediction

Overall Error Rate,

Error= (TI) / (TC+TI)

Here,

TI - total number of incorrectly classified species

TC - total number of correctly classified species

	Real Data Value of Project Status	
Predicted class	Success	Failure
Success	TP	FP
Failure	FN	TN

Table: 1 Confusion Matrix of Prediction Outcomes
With help of the confusion matrix values measurement of the precision and recall values are calculated described below:

Precision:
Precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). The equation is:

Precision = TP / (TP + FP)

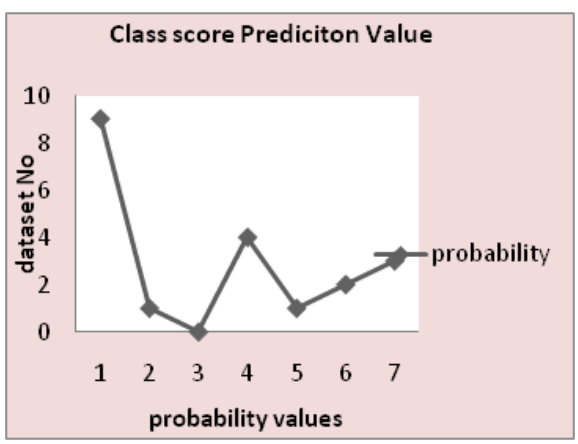
Recall:
Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been.) The

Recall can be calculated as:
Recall = TP / (TP + FN)

Accuracy:
The percentages of the predicted values are match with the expected value for the given data. The best system is that having the high Accuracy, High Precision and High Recall value.

Disease Prediction:
This section evaluates the proposed EDT_SCA with dynamic score based prediction framework in terms of both accuracy and performance. The system applied statlog dataset from UCI;

Table: Probability based comparison



V.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.

REFERENCE [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] G. D. Magoulas and A. Prentza, "Machine learning in medical applications," Mach. Learning Appl. (Lecture Notes Comput. Sci.), Berlin/Heidelberg, Germany: Springer, vol. 2049, pp. 300–307, 2001. | [3] L. Breiman, "Bagging predictors," Mach. Learning, vol. 24, no. 2, pp. 123–140, 1996. | [4] Gordan.V.Kass(1980). An exploratory Technique for inverstigation large quantities of categorical dataApplied Statics, vol 29, No .2, pp. 119-127. | [5] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California. | [6] Quinlan J. R. (1986). Induction of decision trees. Machine Learning, Vol.1-1, pp. 81-106. | [7] Zhu Xiaoliang, Wang Jian YanHongcan and Wu Shangzhuo(2009) Research and application of the improved algorithm C4.5 on decision tree. | [8] Prof. Nilima Patil and Prof. Rekha Lathi(2012), Comparison of C5.0 & CART Classification algorithms using pruning technique. | [9] Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining | [10]. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.