EFFICIENT CLASSIFIER FOR PREDICTING HEART DISEASE THROUGH FEATURE RELEVANCE ANALYSIS

S.Venkatesh Babu[1], S.Poonkuzhali[2]

[1]*PG Scholar, Information Technology, Rajalakshmi Engineering College, Chennai (India)*

[2]*Hod of Information Technology, Dept of I.T, Rajalakshmi Engineering College, Chennai, (India)*

ABSTRACT

Feature relevance analysis plays a vital role in clinical data mining as it has a high impact in classifying the disease. In general for detecting a disease number of tests should be required from the patient. But through data mining technique the number of test required for analyzing the disease is greatly reduced. In this paper Heart Disease Dataset is taken from UCI Machine Learning Repository for this research work. Initially preprocessing and feature relevance analysis is done before classification in order to obtain the quality results. Then various classifiers are applied on this dataset. Finally, Cross validation is done on all the classifier for the test data to evolve the best classifier for predicting the hard disease. Here, Random tree classifier given the best classification results.

Keywords: FSF Classifier, Feature Relevance Analysis, Heart Dataset, Rand Tree

I. INTRODUCTİON

Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. It is "the science of extracting useful information from large databases". It is one of the tasks in the process of knowledge discovery from the database. [1]

Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans. It is a process to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. It is a cooperative effort of humans and computers.

Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. There are two primary goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans. The Disease Prediction plays an important role in data mining. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc… This paper analyzes the Heart disease, Diabetes and Breast cancer disease predictions.

II. ARCHITECTURAL DESIGN

The architectural design of the proposed system is given in Fig 1 and each block is explained in the following sections.

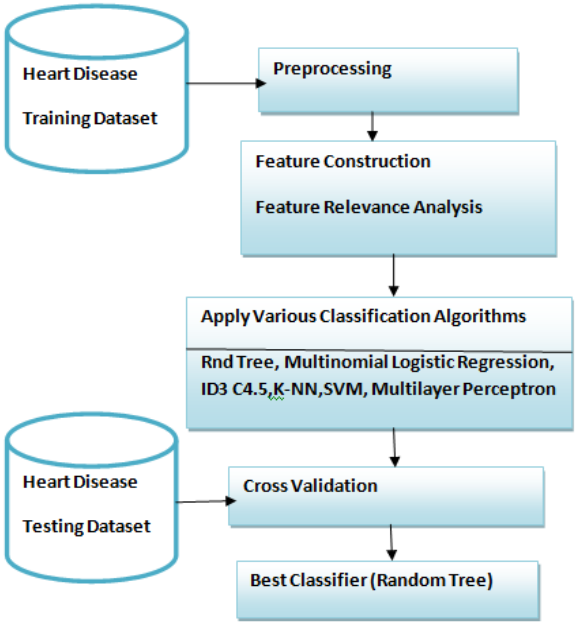

Fig. 1 Artichitecture Design of the Proposed System

2.1 Input Dataset

The Heart Diesease Dataset is taken from UCI Machine Learning repository. It consists of 14 attributes (13-input attributes,1-target attribute) and 303 instances. The dataset field specifies the presence of heart disease in the patient. The description of the attributes of this Heart disease dataset is given in Table 1.

Table. 1 Attribute of Heart Disease Dataset

Attribute No	Attributes
1	age
2	sex
3	chest pain type (4 values)
4	resting blood pressure
5	serum cholestoral in mg/dl
6	fasting blood sugar > 120 mg/dl
7	resting electrocardiographic results (values 0,1,2)
8	maximum heart rate achieved
9	exercise induced angina
10	oldpeak = ST depression induced by exercise relative to rest
11	the slope of the peak exercise ST segment
12	number of major vessels (0-3) colored by flourosopy
13	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

2.2 Data Preprocessing

Nowadays, most of the data in the real world are incomplete containing aggregate and missing values, noisy data containing errors, inconsistent data containing discrepancies in codes and names. As the quality decision depends on quality mining which is based on quality data, pre-processing becomes a very important tasks in any mining related activity. Major tasks in data pre-processing are data cleaning, data integration, data transformation and data reduction. In this dataset data cleaning is done to fill up the missing values with its corresponding mean values.

2.3 Feature Relevant Analysis

Feature relevance analysis is an important area in which pattern recognition, machine learning and statistics are widely done in data mining communities. Here in Feature relevance analysis feature construction and feature selection is done. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. Feature selection techniques provide three main benefits like improved model interpretability, shorter training times, and enhanced generalisation by reducing over fitting with respect to predictive models. Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related.

2.4 Feature Construction

Feature Construction attempts to discover relations between the existing features and creates new features that reflect those relations. In feature construction good high level features are concentrated for classification.

2.5 Feature Selection Alorithms

2.5.1 Fisher Filtering

Fisher filtering is a supervised feature selection algorithm [13] which processes the selection independently from the learning algorithm. It follows univariate Fisher's ANOVA ranking which ranks the inputs attributes according to their relevance without considering the redundancy aspects of input attributes.

2.5.2 Runs Filtering

Runs filtering [13] are a non parametric test for predictive attribute evaluation. It is an univariate attribute ranking from runs test. It is a supervised feature selection algorithms based upon a filtering approach i.e. processes the selection independently from the learning algorithm. This component ranks the inputs attributes according to their relevance without considering redundancy aspect. A cutting rule enables to select a subset of these attributes.

2.6 ReliefF

ReliefF algorithm [13] detects conditional dependencies between attributes and provides a unified view on the attribute estimation in regression and classification. It is not limited to two class problems, is more robust and can deal with incomplete and noisy data.

2.7 Step Disc

STEPDISC (Stepwise Discriminant Analysis) [13] procedure performs a stepwise discriminant analysis to select a subset of the quantitative variables for use in discriminating among the classes. The set of variables that make up each class is assumed to be multivariate normal with a common covariance matrix. The STEPDISC procedure can use forward selection, backward elimination, or stepwise selection.

2.8 Classification Algorithms

2.8.1 Rnd Tree

A Random forest tree [11] consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. For regression problems, the tree response is estimated for the dependent variable given by the predictors.

2.9 Multinomial Logistic Regression

In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.

2.10 ID3

In ID3 decision tree, each node corresponds to splitting attribute. It uses information gain to determine the splitting attribute. The attribute with the highest information gain is taken as the splitting attribute. Information gain is the difference between the amount of information needed to make a correct prediction before and after splitting. Information gain can also be defined as the different between the entropy of the original segment and the accumulated entropies of the resulting split segments. Entropy is the measure of disorder found in the data. ID3 can handle high-cardinality predictor variables. A highcardinality predictor is a variable which has different possible values thus having different possible ways of performing a split.

2.11 C4.5

C4.5 is an classification algorithm that is used to generate a decision tree. C4.5 is an extension of earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

2.12 K-NN

The k-nearest neighbor algorithm (k-NN) [17] is a method for classifying objects based on closest training examples in an n-dimensional pattern space. When given an unknown tuple the classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuple are the k nearest neighbor of the unknown tuple.

2.13 SVM

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

2.14 Multilayer Perceptron

It is the most popular network architecture [12] in today world. The units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output. The units are

arranged in a layered feed forward topology. The network has a simple input-output model, with the weights and thresholds. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptron are the design specification of the number of hidden layers and the number of units in these layers.

III. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The Heart disease dataset is downloaded from the UCI Machine Learning Repository, the input and target attributes contains only continuous attributes. In order to apply filtering algorithms the entire target attributes have to be transformed either into discrete attributes. Then the filtering algorithms such as ReliefF, Fisher Filtering, Runs Filtering, Stepwise Discriminant Analysis are applied to the feature constructed dataset and the results are given in Table II. Then classification algorithms such as Rnd Tree, Multinomial Logistic Regression,ID3,C4.5,K-NN,SVM, Multilayer Perceptron are applied to each of the above filtering algorithms and the results are given in Table III.

Table. 2 Feature Selection

S.No	Feature selection algorithm	No of attributes Before Filtering	No of attributes After Filtering	Attribute No After Filtering
1	ReliefF	13	6	2,3,7, 9,12,13
2	Fisher Filtering	13	9	1,2,3,8,9,10, 11,12,13
3	Runs Filtering	13	4	3,9,12,13
4	Stepwise Discriminant Analysis	13	8	2,3,7,8,9, 10,12,13

Rnd Tree algorithm produces 100% accurate Results without applying any future relevance analysis Fisher filtering produces 100% accurate results for Rnd–tree algorithms; above 85% accuracy for Fisher filtering and Multinomial Logistic Regression. Fisher filtering produces 85% accuracy for C4.5 algorithm; above 85% for multilayer perceptron filtering and above 85% accuracy for K-NN algorithm and above 85% for SVM. Stepwise discriminant analysis provides above 95% for Rnd –tree algorithm and above 90% accuracy for C4.5 and K-NN classifiers and above 85% for multilayer perceptron classifier. From the results, the Rnd tree classification is considered as a best classifier, as it produced 100% accuracy through all the six classifiers.

Table. 4 Error Rate of Heart disease Dataset Classification

Classification Algorithm	Error Rate Before Filtering	Error Rate After Filtering			
		ReliefF	Fisher Filtering	Runs Filtering	StepDisk
Rnd Tree	0.0000	0.1111	0.0037	0.1296	0.01111
Multinomial Logistic Regression	0.1444	0.1481	0.1407	0.1407	0.137
ID3	0.237	0.237	0.237	0.237	0.237
C4.5	0.0111	0.1333	0.1148	0.1333	0.1185
KNN	0.1333	0.1556	0.1148	0.137	0.1481
SVM	0.1481	0.1519	0.1444	0.1556	0.137
Multi Layer	0.1333	0.137	0.1222	0.1481	0.137

3.1 Error Rate

Error rate of a classifier was defined as the percentage of the dataset incorrectly classified by the method. It is theprobability of misclassification of a classifier

$$\textbf{Error rate} = \frac{\textbf{No.of incorrectly classified samples}}{\textbf{Total no of Sample in the class}}$$

3.2 Accuracy

Accuracy of a classifier was defined as the percentage of the dataset correctly classified by the method. The accuracy of all the classifiers used for classifying this TP53 germline dataset are represented graphically in Figure2

$$\textbf{Accuracy} = \frac{\textbf{No of correctly Classified Samples}}{\textbf{Total no of Sample in the class}}$$

3.3 Recall

Recall of the classifier was defined as the percentage of errors correctly predicted out of all the errors that actually occurred. The recall of the best classifier Rnd Tree is given in Figure 3.

$$\textbf{Recall} = \frac{\textbf{True Positive}}{\textbf{True positive + False Negative}}$$

3.4 Precision

Precision of the classifier was defined as the percentage of the actual errors among all the encounters that were classified as errors.

$$\textbf{Precision} = \frac{\textbf{True Positive}}{\textbf{True positive + False Positive}}$$

The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to classifier's expectation.

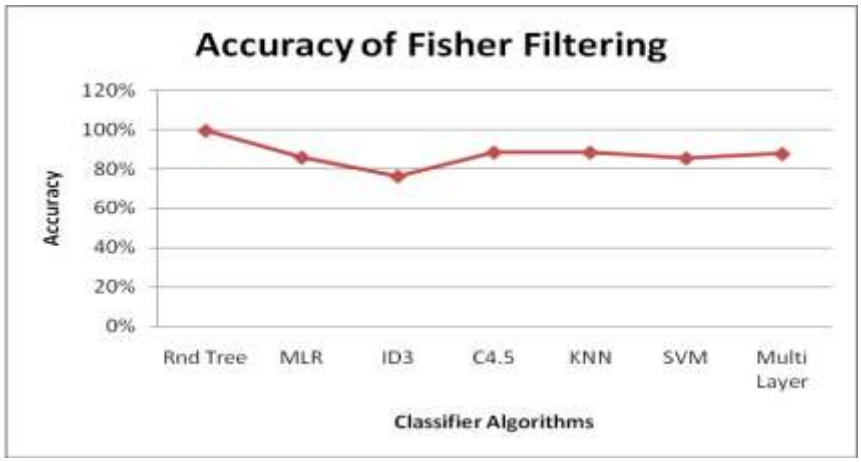

Fig. 2 Accuracy of the Fisher Filtering

IV. CONCLUSION

This work finds the best classifier for predicting the Heart dataset through feature relevance analysis. The evolved classifier has produced 100% accuracy for the Random forest tree classifier using Fisher Filter selection algorithm and it also accurately classified the test dataset.

REFERENCES

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

[2] "Data mining: Introductory and Advanced Topics" Margaret H. Dunham

[3]. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 No. 6 June 2011.

[4]. Carloz Ordonez, "Association Rule Discovery with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[5] M. ANBARASI, E. ANUPRIYA, N.CH.S.N.IYENGAR, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370- 5376.

[6] G. Parthiban, A. Rajesh, S.K.Srivatsa "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method" .

[7] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques,"Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[8] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

[9] Delen Dursun , Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005.

[10] Ruben D. Canlas Jr.,"data mınıng ın healthcare: current applıcatıons and ıssues", August 2009.

[11] R. Geetha Ramani, G. Sivagami, Parkinson Disease Classification using Data Mining Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 32– No.9, October 2011.

[12] Shomona Gracia Jacob, R.Geetha Ramani, Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.

[13] S. Poonkuzhali, R. Geetha Ramani, R. Kishore Kumar,Efficient Classifier for TP53 Mutants using Feature Relevance Analysis, in International Multiconference of Engineers and computer scienctists, Vol 1, 2012.

[14] Tanagra-Data Mining tutorials http://data-mining- tutorials.blogspot.com.