An Implementation of RCA to Find High Accuracy & Least Error Rate Using Weka Tool over the Various Data Sets

R.Umasundari
Ph.D., Research Scholar, A.V.V.M. Sri Pushpam College (Autonomous),
Poondi, Thanjavur, Tamilnadu, India.
umaamal@gmail.com

Dr.M.Chidambaram
Assistant Professor, Rajah Serfoji Government Arts College (Autonomous),
Thanjavur, Tamilnadu, India.
chidsuba@gmail.com

*Abstract***— Data mining is the approaching inquiries range to solve a number of difficulties and classify datasets is one of the main in the applied of data mining. Data mining denotes to mining knowledge as of big size of records. In this implementation, many datasets are paralleled with RCA (Reliability Classification Algorithm). This classification is used to sort each in set of data in to one of the established in advance set of classes. The data sets are arranged and collected from the Educational Sector, Health care & Agriculture etc., The current study designed to do the great performance result of decision tree Reliability Classification Algorithm using freely available data mining WEKA tool over the different data sets. A reliability classification algorithm for a several data sets are chosen based on excessive classification accuracy and minimum fault rate. This work has been carried out to make a performance evaluation of RCA. The results in this paper demonstrate that the efficiency of RCA is good to perform over various datasets better than other classifiers.**

Keyword - Data mining; Classification; WEKA tool; Reliability Classification Decision tree;

I. Introduction

In recent years Data Mining has turn out to be more popular. It necessary to use some heuristics to be able to analyze the large volume of data that has become available[2]. Data Mining has particularly come to be popular in the areas of fraud analysis, forensic science and healthcare, for it reduces costs in time and money. Classification decision trees are used for the kind of Data Mining difficult which are deals with prediction[1]. Using examples of cases it is possible to construct a design that is able to predict the class of new examples using the attributes of those examples. An experiment has been set up to test the performance of a pruning method which is used in classification tree designing. The pruning methods will be applied on various kinds of data sets.

Educational Data Mining (EDM) is the implementation of information mining systems on educational data. The purpose of EDM[4] is to analyze such data and to find a solution to educational research issues. EDM deals with developing new approaches to search the educational data, and using Data Mining methods to better understand student learning environment.

In Agriculture, decision tree is the classification tools that created as of disorder and random data, the classification exactness of which is high and the result pattern is simple. Used various techniques of depth analysis of data we have included, statistical machine learning, natural trees and alternative data analysis approaches that have studied by agricultural and biological research. This paper covers the latest implementation of decision tree that bring in to play on agriculture. This paper is not meant to be exhaustive[7]. We will give particular attention recent works and then we will briefly mention some other implementation that viewed to us to be the greatest stimulating to report and help to do more on forward which are make a path that how to conclude the concepts of RCA (Reliability Classification Algorithm) that ahead to create true decision making on cultivation.

Data mining is one of the most important domains which help in management of healthcare data. It also helps to discover recently developed vogue healthcare data collected from various hospital sector. The data mining tools and techniques help in analyzing data collected from different hospitals and summarizing it into useful information. There are huge applications of data mining in healthcare sector. In Healthcare, our RCA deals with efficient data mining procedure for predicting the diabetes from medical records of patients. Data mining has proven to be very beneficial in the field of medical analysis as it increases diagnostic accuracy, to reduce costs of patient treatment and to save human resources. There are various data mining techniques[11] such as Association, Classification, Clustering, Neural Network and Regression.

II. Related Work

A. Improved J48 Classification: Prediction of Diabetes

In this paper the authors Gaganjot Kaur & Amit Chhabra et al.[3] discussed about efficient data mining procedure for predicting diabetes patient medical records using Improved J48 classification. It increase accuracy rate of collected data sets. J48 reduce the classification errors which are being produced by in the diabetes data sets. This implementation programmed in WEKA as API in Matlab. The experimental result is improved J48 is effectively predicting the diabetes from medical records. Using J48 it proved that the algorithm can achieve accuracy up 99.87% with comparing the existing classification methods.

B. Using various Classification Techniques on Healthcare datasets to find performance analysis

In this implementation made by Shelly Gupta, Dharminder Kumar and Anand Sharma et al.[1] described on highest classification accuracy and least error rate over the healthcare datasets. In this paper, they are using three various Machine Learning tools namely WEKA, Tanagra & Clementine. Using this tool, the authors carried out the performance analysis of various decision tree algorithms, kNN, SVM, NB, MLP & CART on particular datasets. The outcome of this implementation on the specific datasets depending on the nature of their attributes and size. From the results two prediction methods that are “SVM & kNN” very nearly related with high accuracy rate (96.74% & 97.28%) displays.

C. Estimation of missing values using decision tree approach

The authors Gimpy, Dr. Rajan Vohra & Minakshi et al.[5] detailed discussed with missing data or value in a data from the prepared datasets. In this paper they take to estimate of missing data in student records of university using C4.5/J48 classification algorithm and this approach can implementing by data mining tool named WEKA. The input of the data set format is MS.Excell and WEKA is converted in to .csv format. The particular input dataset processed by using Matrix calculation called Confusion Matrix. A confusion matrix contains information about actual and predicted classifications done by a J48 classification system. The J48 algorithm is used and accuracy is calculated for both incomplete data and the imputed data. And as a result accuracy is greater for imputed dataset as compared to incomplete dataset.

D. Decision Tree Approach to Detect Characteristic of Bt Cotton Base on Soil Micro Nutrient

In this paper, the authors Youvrajsinh Chauhan & Jignesh Vania[7] presented to improve crop production and identify crop disease with helps soil systems, used throughout a large amount of crop fields or areas in the environment of agriculture. In this concept a J48 classification to make true decision making on agriculture. To determine and predict true result from the dataset by using data mining machine learning tool WEKA. It gives more accuracy results of predicts soil fertility. The all soil or crop dataset calculating with Bt Cotton gives different ranging values that show yellow crop disease is available or not. This paper proposed algorithm is gives 83.74% true classified results. So, this paper proved to the J48 algorithm is provide highest true result of 91.90% for predicts the crop production and crop disease identification.

E. To Predict Slow Learners in Education Sector Based Data Mining Classification Algorithm

This paper focused on identifying the slow learners among students[9] and displaying it by a predictive data mining model using classification based algorithms such as MLP, NB, SMO, J48 and REPTree, by using open source machine learning tool WEKA. In this research taken to process 152 high school dataset from educational data mining. The dataset inserted in to the attribute evaluator. The data declare in to some variables. Then the author applied classification algorithms to compare, find the output and also applied variables in Ranker Search method technique on WEKA tool. The dataset tested with five classification algorithms and that are provided accuracy results. Finally, the author investigate that MPL (Multi Layer Perception) technique performs best with accuracy 75%. Therefore, performance of MLP is relatively higher than other classification algorithms.

III. Methodology

A. WEKA Tool

Waikato Environment for Knowledge Learning WEKA[12] is a computer software package that was established by the student of the University of Waikato in New Zealand for the resolve of classifying data from large data gather round from agricultural fields. Data preprocessing, classification, grouping, association, regression and feature selection these standard data mining tasks are supported by Weka. It is an open source software which is easily available in web.

In Weka datasets should have arranged to the ARFF file format. The Weka Explorer will use these mechanically if it does not identify an assumed file as an ARFF file format. Classify tab in Weka Explorer is used for the classification purpose to classify data. A huge varying sum of classifiers are used in weka such as bayes, function, tree etc.

Process to put on classification methods on data set and come to be end result in Weka:

→ Process 1: Bring input dataset and convert specify file format.

→ Process 2: Apply the Reliability Classification Algorithm on the collected data set.

→ Process 3: Remark the degree of accuracy given by the RCA and time required for execution.

→ Process 4: Accuracy provided with Reliability classification algorithms for particular dataset.

The experiments are conducted in a system with configuration Intel Core Processor, 2 GB DDR3 Memory and 500 GB HDD. Experiments are conducted 3 times and an average accuracy and time is recorded.

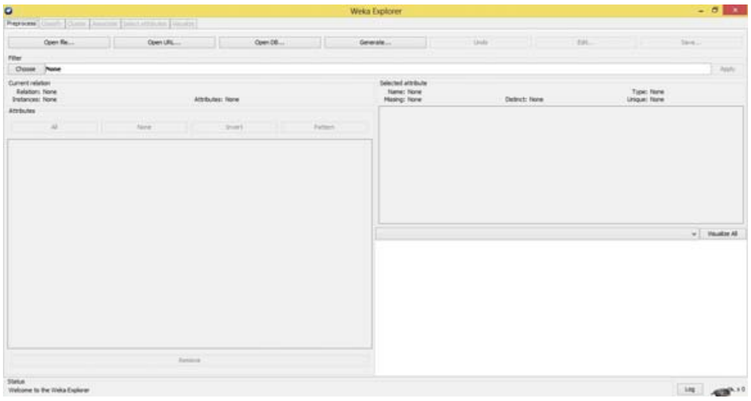

Fig: 1. WEKA Home Screen

Steps to create tree in Weka

- Generate datasets in MS Excel, or any other & save in .CSV format.
- Start the weka Explorer.
- Open .CSV file & save in .ARFF format.
- Get on classify tab & top quality RCA from choose button.
- Choose any suitable test selection.
- Go to Start button & output will be showed.

B. RCA-Reliability Classification Algorithm

A Reliability Classification decision tree carry out the classification of a specified data sampling through different stages of decisions to support us reach a final decision. Such a structure of decisions is on behalf of in a tree structure. The tree structure is used in classifying indefinite data records.

All dataset to be studied will be of the definite type and therefore continuous data will not be examined at this stage. The algorithm will however leave room for adaption to include this capability. The algorithm will be tested against for verification purposes.

In Weka, the application of a certain learning algorithm is summarized in a class, and it may be determined by on other classes for some of its functionality. RCA class builds a tree structure data. Each time the Weka executes RCA, it generates an example of this class by assigning retention for building and storing a decision tree classifier. The algorithm, the classifier it builds, and a procedure for outputting the classifier is all part of that instantiation of the RCA class.

Larger datasets are usually split into more than one class. The RCA class does not essentially cover any code for building a tree structure. It consist of positions to instances of other classes that do most of the work. When there are a number of classes as in Weka software they become difficult to comprehend and navigate.

C. Datasets

There are three datasets we have used in our paper. The details of each datasets are shown in Table 1.

Datasets	**Instances**	**Attributes**	**No.of Classes**	**Type**
Educational Dataset	964	26	7	Nominal
Healthcare Dataset	867	19	5	Nominal
Agriculture Dataset	465	8	3	Nominal

The data sets used for the tests come from the UCI Machine Learning repository. We are dealing with classification tasks, thus we have selected datasets of which the class values are nominal. Selection of the datasets further depended on their size, larger data sets generally means higher confidence. We choose different kinds of data sets, because we also wanted to test if the performance of an algorithm depended on the kind of set that is used.

IV.RESULT

For calculating a classifier superiority we can use confusion matrix. Consider the algorithm RC running on various dataset in WEKA, for this dataset we obtain three classes then we have 3x3 confusion matrix. The all data set are applied to the RCA for classify the data and established for constructed the model by using a training model for classify the training data set and see the outcomes of the correctly classified instances. Apply data set to RCA, it gives better result by experiment with high accuracy and low error rate. The confusion matrix helps us to find the various evaluation measures like Accuracy, Recall, Precision etc.

Table: 2. Educational Dataset Result

S.No	Parameters	Classification Algorithm
1	True Positive Rate	0.81
2	False Positive Rate	0.21
3	Precision	0.77
4	Recall	0.81
5	F-Measure	0.79
6	ROC Area	0.85

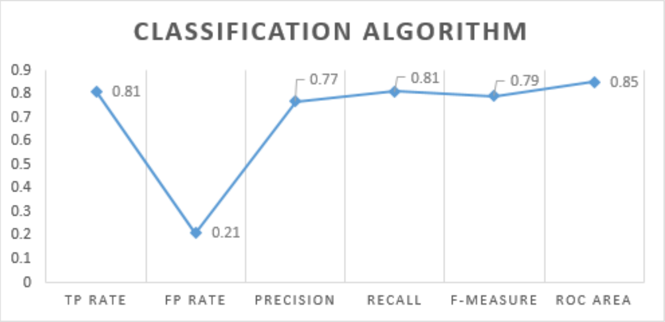

Fig: 2. Graphical Result of Education Dataset

For using Education dataset, accuracy parameters have shown in Table 2 and Fig 2. RC Algorithm is better way to provide accuracy for this dataset.

Table.3. Healthcare Dataset Result

S.No	Parameters	Classification Algorithm
1	True Positive Rate	0.78
2	False Positive Rate	0.35
3	Precision	0.78
4	Recall	0.78
5	F-Measure	0.78
6	ROC Area	0.79

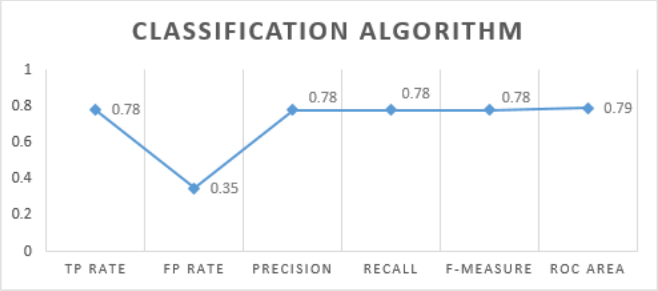

Fig: 3. Graphical Result of Healthcare Dataset

Table: 4. Agriculture Dataset Result

S.No	Parameters	Classification Algorithm
1	True Positive Rate	0.69
2	False Positive Rate	0.14
3	Precision	0.70
4	Recall	0.70
5	F-Measure	0.70
6	ROC Area	0.85

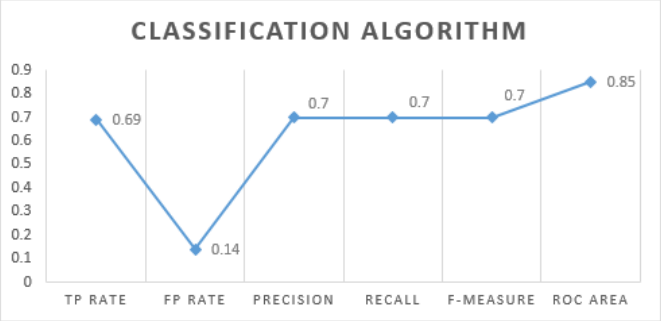

Fig: 4. Graphical Result of Agriculture Dataset

Various decision tree algorithms can be used for prediction and classification for different datasets. This studies showed that Reliability Classification Algorithm gives 83.74 % accuracy; hence it can be used as a base learner. We make better prediction model that help to improve prediction and classification of data.

V. Conclusion

This research has conducted a study on a various dataset which is using data mining toolkit Weka to find high accuracy and low error rate. After analyzing the results, we found that are able to generate tree model in very less time. Weka tools is very efficient in generating decision trees. However, in terms of classifiers applicability, we conclude that the Weka tool is better in terms of the ability to run the classifier and in terms of error rate. Also, Weka is faster than other tree generation as its internal structure is organized in columns in memory. Through this study, we conclude that Weka is better tool for our proposed Reliability Classification Algorithm to predict the data. Also, we found that Reliability Classification Algorithm works well in decision tree induction. In future, we can implement this algorithm with more data and larger set of patient records, Educational records and Agriculture records to produce better results better than other classification algorithms.

References

[1] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang & Laurence T. Yang, “Data Mining for Internet of Things: A Survey”, in Communications Surveys & Tutorials, 2013.
[2] Chakchai So-In, Nutakarn Mongkonchai, Phet Aimtongkham, Kasidit Wijitsopon & Kanokmon Rujirakul, “An evaluation of data mining classification models for network intrusion detection”, Digital Information and Communication Technology and it's Applications, 2014.
[3] Gaganjot Kaur & Amit Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes”, in IJCA 2014.
[4] Gimpy, Dr. Rajan Vohra & Minakshi, “Estimation of Missing Values Using Decision Tree Approach”, in IJCSIT 2014.
[5] M. Mayilvaganan & D. Kalpanadevi, “Comparison of classification techniques for predicting the performance of students’ academic environment”, Communication and Network Technologies 2014.
[6] Parneet Kaura, Manpreet Singhb & Gurpreet Singh Josanc, “Classification and prediction based data mining algorithms to predict slow learners in education sector”, in ICRTC-2015.
[7] Rakesh Rana, Miroslaw Staron, Jörgen Hansson, Martin Nilsson & Wilhelm Meding, “A framework for adoption of machine learning in industry for software defect prediction”, Software Engineering and Applications, 2014.
[8] Rahmad Kurniawan, Mohd Zakree Ahmad Nazri, M. Irsyad, Rado Yendra & Anis Aklima, “On machine learning technique selection for classification”, Electrical Engineering and Informatics 2015.
[9] R. P. T. H. Gunasekara, M. C. Wijegunasekara & N. G. J. Dias, “Comparison of major clustering algorithms using Weka tool”, Advances in ICT for Emerging Regions, 2015.
[10] Shelly Gupta, Dharminder Kumar & Anand Sharma, “Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data”, in IJCSIT 2011.
[11] Youvrajsinh Chauhan & Jignesh Vania, “J48 Classifier Approach to Detect Characteristic of Bt Cotton base on Soil Micro Nutrient”, in IJCTT 2013.
[12] Zhenni Feng & Yanmin Zhu, “A Survey on Trajectory Data Mining: Techniques and Applications”, IEEE Access in 2016.