World Journal of Engineering Research and Technology

WJERT

www.wjert.org

SJIF Impact Factor: 4.326

COMPARATIVE ANALYSIS OF DATA MINING CLASSIFICATION ALGORITHM FOR STUDENTS SOFT SKILLS

[1]A. Komathi, MCA., M.Phil., Ph.D., [2]*R. Kowsalya

[1]Head and Assistant Professor, Department of Computer Science and Information Technology, Nadar Saraswathi College of Arts and Science, Theni.

[2]Research Scholar, Department of Computer Science and Information Technology, Nadar Saraswathi College of Arts and Science, Theni.

Article Received on 18/02/2017 Article Revised on 11/03/2017 Article Accepted on 31/03/2017

***Corresponding Author**
R. Kowsalya
[2]Research Scholar, Department of Computer Science and Information Technology, Nadar Saraswathi College of Arts and Science, Theni.

ABSTRACT

Data mining refers to the finding of relevant and useful information from databases. Data mining is also called Knowledge Discovery in Database (KDD), pattern Analysis, Data Archiology, and Business Intelligences. Data mining techniques are mostly used in educational environments. Educational data mining is an emerging field exploring data in education context by applying different data mining tools and techniques. The ability to predict the students writing skills is very important in educational environments. In this paper focused on comparison of four classification algorithms such as J48, Random Tree, LAD Tree, and ADTree. These algorithms are compared based on Correctly Classified Instances, Incorrectly Classified Instances, Recall, F-Measures, and Precision.

KEYWORDS: Data Mining, Decision Tree, Classification, J48, Random Tree, LAD tree, ADTree.

1. INTRODUCTION

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in the collected data set. EDM exploits statistical, machine learning and data mining algorithm over different types of educational data. Writing has at all times and in all ages been a greater source of knowledge. Today, the ability of writing is highly valued and very important for social and economic advancement. Writing is typically an individual activity, although on occasion a person will write out benefit for other writers or readers. There are seven different attributes are used to find the students writing skills such as college type, qualification, Residence, motivation, language and improvements of writing and hobbies.

The student writing skills data are classified and compared with J48, Random Tree, LAD Tree, and ADTree. Comparison is mainly based on common attributes like Correctly Classified Instances, Incorrectly Classified Instances, Recall, F-Measures and Precision.

2. METHODOLOGY

2.1 Data Collection

The sample data are collected from theni district college students. We have collected 200 sample data from students.Collected data are entered on excel sheet and initial preprocessing is done manually by filling the missing values by standard data and various inconsistence has been removed.

2.2 Preprocessing

Data Preprocessing begins with a collected dataset. Data preprocessing means filling in Missing Values, Smoothing Noisy Data, Identifying or Removing Outliers and Resolve Inconsistencies.

2.3 Classification

Classification is a form of data analysis that extracts models describing important data classes. Many classification methods have been proposed by researchers in machine learning, patterns recognition, and statistics. Classification is a two-step procedure, consisting of a learning step and a classification step. Classification consists of many kinds of algorithms such as, Byes, Functions, lazy, Meta, Rules and Trees.

3. CLASSIFICATION METHODS

3.1. Decision Tree

Decision tree is a flow chart-like tree structure, where each internal node (main node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node (terminal node) holds a class label. A decision tree is a classification scheme which generates a tree and set of rules, representing the model of different classes, from a given data set.

Main advantages of decision tree:

- Decision tree has the ability to generate understandable rules.
- They are able to handle both numerical and categorical attributes.
- They provide a clear indication of which field is most important for prediction or classification.

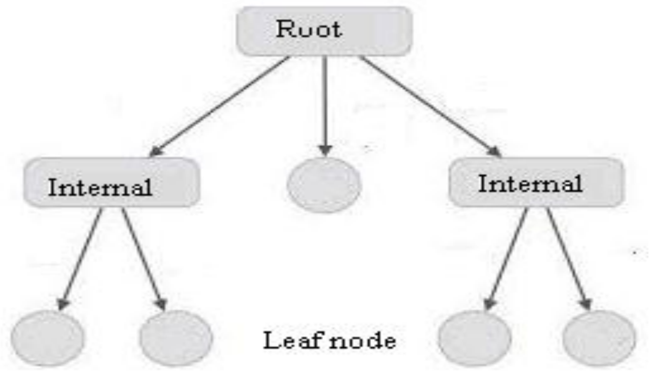


Figure 3.1.1: Basic structure of Decision Tree.

Decision tree techniques contain many algorithms such as BFTree, Decision Stump, FT, and j48graft, LAD Tree, LMT, NBTree, REPTree, CART, ID3, C4.5, Simple Cart, Random Forest, Random Tree, AD Tree and User Classifier.This paper is fully based on j48 and Random Tree, LAD Tree, and ADTree Algorithms.

3.2 Precision

It is the determination of exactness. It is the ratio of the predicted positive cases that were correct to the total number of predicted positive cases.

$$\text{Per}=\frac{|\{\text{relevant document}\}\cap\{\text{retrieval document}\}|}{|\{\text{retrival document}\}|}$$

3.3 Recall

Recall is the determination of completeness. It is the proportion of positive cases that were correctly recognized to the total number of positive cases. It is also known as sensitivity or true positive rate (TPR).

$$\text{Recall}=\frac{|\{\text{relevant document}\}\cap\{\text{retrival document}\}|}{|\{\text{relevant document}\}|}$$

3.4 F-Measure

A measure that combines precision and recall is the harmonic mean of precision and recall.

$$F=2*\frac{Precision * Recall}{Precision + Recall}$$

3.5 J48

Decision tree J48 is the implementation of algorithm ID3 developed by weka project team.J48 is an open source java implementation program.J48 is extension of ID3 algorithm. The additional functions of J48 are accounting for missing values, decision tree pruning, continuous attribute value ranges, derivation of rules, etc.

3.6 Random tree

A random tree is a collection of tree predictors that is called forest. It can deal with both classification and regression problems.

The classification works as follows: the random trees Classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of votes. In case of a regression, the classifier response is the average of the responses over all the trees in the forest. The entire tree is trained with the same parameters but on different training sets.

3.7 ADTree

An (ADTree) Alternating Decision Tree is a machine learning method for classification. It generalizes decision trees. An alternating decision tree consists of decision nodes and prediction nodes. Decision nodes specify a predicate condition and prediction nodes contain a single number.ADTrees always have prediction nodes as both root and leaves. An instance is classified by an ADTree by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed.

3.8 LAD Tree

An (LAD) Logical Analysis of Data tree builds a classifier for binary target variable based on learning a logical expression that can distinguish between positive and negative samples in a data set. The basic assumption of LAD model is that a binary point covered by some positive patterns, but not covered by any negative pattern is positive, and similarly, a binary point covered by some negative patterns, but not covered by positive pattern is negative. The construction of LAD model for a given data set typically involves the generation of large set patterns and the selection of a subset of them that satisfies the above assumption such that each pattern in the model satisfies certain requriments in items of prevalence and homogeneity.

4. EXPERIMENTAL RESULT

Writing skills attributes are age, year, college type, qualification, Residence, motivation, language and improvements of writing hobbies. Attribute selection method used to select some attributes.

Thus the sample figures are used for students writing skills attributes. Such attributes are Qualification, Location Wise and Residence Wise.

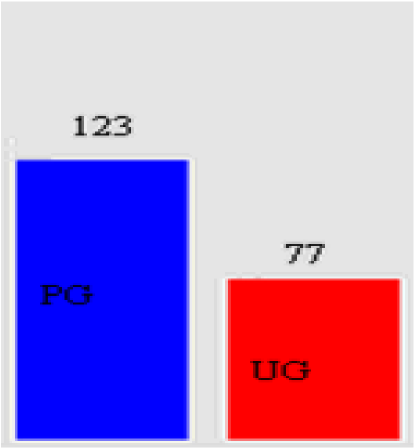


Fig 4.1: (a) Qualification Wise classification.

In the qualification chart compare two attributes such as UG and PG. Finally PG students are improving good writing skills than UG students.

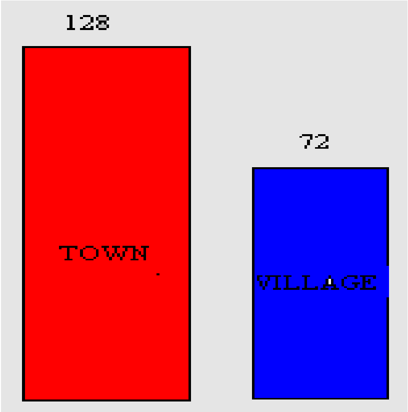


Fig 4.2: (b) Location Wise classification.

In the Location chart compare two attributes such as VILLAGE and TOWN. Finally TOWN area students are improving better writing skills than VILLEGE area students.

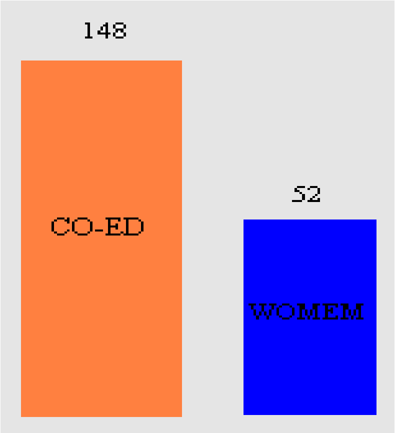


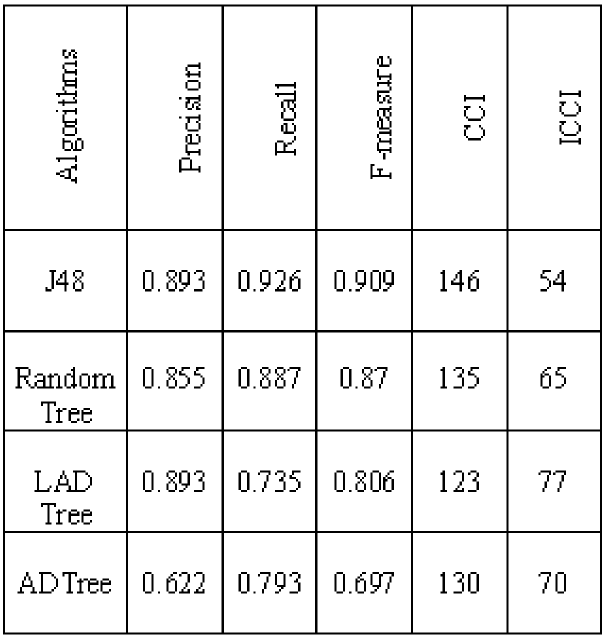
Fig 4.3: (c) College Type Wise classification.

In the College Type chart compare two attributes such as CO-ED and WOMEN. Finally CO-ED college students are improving writing skills than WOMEN college students.

5. RESULT AND DISCUSSION

Compare the classification algorithms (J48, Random Tree, LAD Tree, and ADTree) based on Correctly Classified Instances (CCI), Incorrectly Classified Instances (ICCI), Precision,F-Measure and Recall.

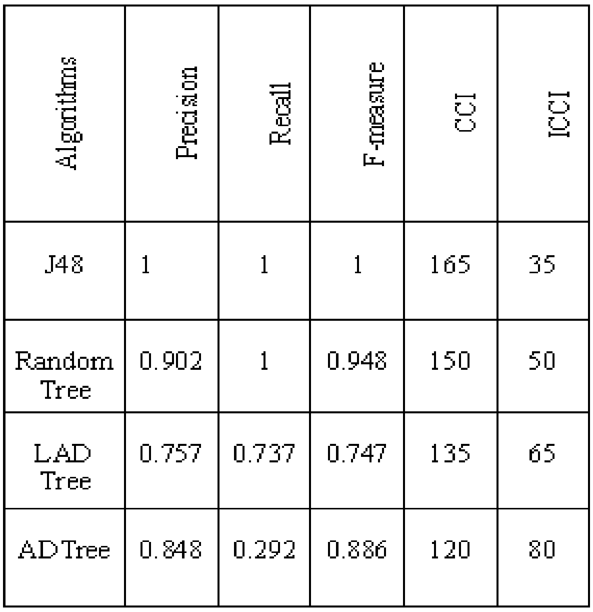
Tab 5.1: Qualification Wise Comparisons in J48, RT, ADTree, and LAD Tree.



Algorithms	Precision	Recall	F-measure	CCI	ICCI
J48	0.893	0.926	0.909	146	54
Random Tree	0.855	0.887	0.87	135	65
LAD Tree	0.893	0.735	0.806	123	77
ADTree	0.622	0.793	0.697	130	70

Thus the result, When evaluate the four algorithm J48, Random Tree, LAD Tree, and ADTree. J48 provides a superior result in correctly classified instances, Precision, Recall and F-Measure than other algorithms.

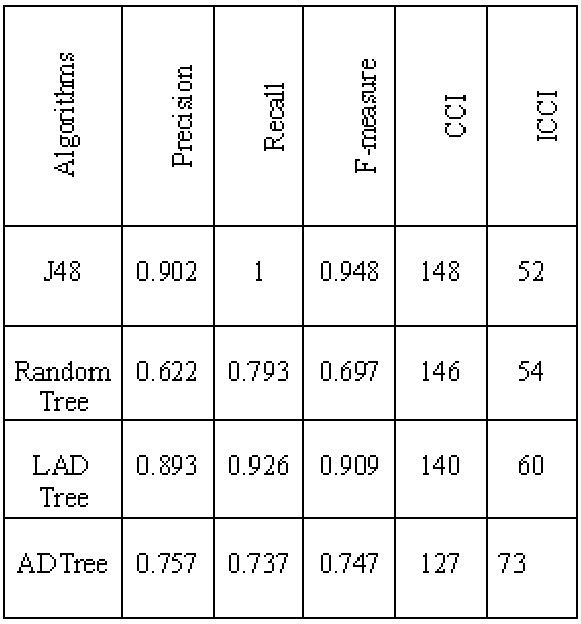
Tab 5.1: Location Wise Comparisons in J48, RT, ADTree, and LAD Tree.



Algorithms	Precision	Recall	F-measure	CCI	ICCI
J48	1	1	1	165	35
Random Tree	0.902	1	0.948	150	50
LAD Tree	0.757	0.737	0.747	135	65
ADTree	0.848	0.292	0.886	120	80

Thus the result, When evaluate the four algorithm J48, Random Tree, LAD Tree, and ADTree. J48 provides a greater result in correctly classified instances, Precision, Recall and F-Measure than other algorithms.

Tab 5.1: College Type Wise Comparisons in J48, RT, ADTree, and LAD Tree.



Algorithms	Precision	Recall	F-measure	CCI	ICCI
J48	0.902	1	0.948	148	52
Random Tree	0.622	0.793	0.697	146	54
LAD Tree	0.893	0.926	0.909	140	60
ADTree	0.757	0.737	0.747	127	73

Thus the result, When evaluate the four algorithm J48, Random Tree, LAD Tree, and ADTree. J48 provides a grander result in correctly classified instances, Precision, Recall and F-Measure than other algorithms.

6. CONCLUSION

From the consequence, the highest Precision, Recall and F-Measure values obtained for the J48. It can be seen J48 achieve increased classification performance and yield results that are exact. These sceneries also cover the missing values problem in the datasets and thus besides accuracy. It also overcomes the overfiting problem generated due to missing values in the datasets. For the classification troubles, if one has to a classifier among the tree based classifier set, we suggest to use J48 with confidence for variety of classification problem.

7. REFFERENCES

1. Jiawei Han, micheline kamber and jian pei,”Data Mining: Concepts and Techniques”, Third Edition, Elsevier, Microsoft Reasearch, 2012.
2. Leo Breiman,”Random Forset”, Machine Learning, 45(1): 5-32, 200.
3. H. Written and E. Frank, Data Mining: Pratical machine learning tools and techniques, Morgan Kaufmann, 2005.
4. Arun K Pujari, Data Mining Techniques Universities Press (India) Private Limited.
5. Margaret H.Dunham, S. Sridhar-Data /mining /introductory and advanced topics-pearson Education.
6. “A survey on predicting student performance” a.dinesh kumar, (ijcsit) international journal of computer science and information technologies, 2014; 5(5): 6147-6149.
7. "Importance of data mining in higher education system" bhise r.b,thorat s.s,supekar a.k.,iosr journal of humanities and social science(iosr-jhss), jan-feb 2013; 2(2).
8. “Mining students’ academic performance” azwa abdul aziz, nur hafieza ismail, fadhilah ahmad, journal of theoretical and applied information technology, 31july 2013; 53(3).