International Journal of Information Technology and Knowledge Management
July-December 2012, Volume 5, No. 2, pp. 239-243

COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS

Sunita B. Aher[1] and Lobo L.M.R.J.[2]

ABSTRACT: Course Recommender System in E-Learning is a system which recommend the course to the student based on the choice of various student collected from huge amount of data of courses offered through Moodle package of the college. Here in this paper we compare the five classification algorithm to choose the best classification algorithm for Course Recommendation system. These five classification algorithms are ADTree, Simple Cart, J48, ZeroR & Naive Bays Classification Algorithm. We compare these six algorithms using open source data mining tool Weka & present the result. We found that ADTree classification algorithm works better for this Course Recommender System than other five classification algorithms.

Keywords: ADTree, Simple Cart, J48, ZeroR, Naive Bays Classification Algorithm, Weka

1. INTRODUCTION

Course Recommender System in E-Learning is a system which recommend the course to the based on the choice of various student collected from huge amount of data of courses offered through Moodle package of the college. E.g. If student is interested in course like Database System then he would like to learn the Advanced Database System. Here we use Moodle for data collection & Weka to check the results. A framework for Course Recommender System is explained in [8].

2. LITERATURE REVIEW

In research [1], they conducted experimental comparison of LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest on seven Microarray cancer data sets. The experimental results show that all ensemble methods outperform C4.5. The experimental results also show that all five methods benefit from data preprocessing, including gene selection and discretization, in classification accuracy. In addition to comparing the average accuracies of ten-fold cross validation tests on seven data sets, they used two statistical tests to validate findings.

In the paper [2], they presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. They have investigated three data mining techniques: The Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. The achieved prediction performances are comparable to existing techniques. They found out that C4.5 algorithm has a much better performance than the other two techniques.

In paper [3], they proposed the use of decision tree C4.5 algorithm, bagging with decision tree C4.5 algorithm and bagging with Naïve Bayes algorithm to identify the heart disease of a patient and compare the effectiveness, correction rate among them.

In the paper [4], they conducted experiment in the WEKA environment by using four algorithms namely ID3, J48, Simple CART and Alternating Decision Tree on the spam email dataset and later the four algorithms were compared in terms of classification accuracy. According to their simulation results, the J48 classifier outperforms the ID3, CART and ADTree in terms of classification accuracy.

In paper [5], they presented a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. They also examined the effect that calibrating the models via Platt Scaling and Isotonic Regression has on their performance.

3. CLASSIFICATION ALGORITHMS

Classification is a data mining task that maps the data into predefined groups and classes. It is also called as supervised learning. It consists of two steps. First step is the model construction which consists of set of predetermined classes. Each tuple /sample is assumed to belong to a predefined class. The set of tuple used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae. Second step is model usage which is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the

[1,2] Walchand Institute of Technology, Solapur, Solapur University, India, [1]*E-mail: sunita_aher@yahoo.com*

percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur [7]. Here we consider the brief introduction of each classification algorithm.

3.1 ADTree Classification Algorithm

An alternating decision tree (ADTree) is a machine learning method for classification which generalizes decision trees. An alternating decision tree consists of two nodes. Decision nodes specify a predicate condition. Prediction nodes contain a single number. ADTree 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 [10].

3.2 Simple Cart Classification Algorithm

Simple Cart (Classification and regression tree) is a classification technique that generates the binary decision tree. Since output is binary tree, it generates only two children. Entropy is used to choose the best splitting attribute. Simple Cart handles the missing data by ignoring that record. This algorithm is best for the training data [6].

3.3 J48 Classification Algorithm

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes, the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset. [9].

3.4 ZeroR Classification Algorithm

ZeroR classifier predicts the majority of class in training data. It predicts the mean for numeric value & mode for nominal class.

3.5 Naive Bays Classification Algorithm

Naïve Bays classification is based on Bays rule conditional probability. It makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other [9].

4. EXPERIMENTAL RESULT

Here we are considering the sample data extracted from Moodle database of a college after collection of data for course enrollment by student as shown in Table 1. In this table, we consider 45 student & 15 courses. Fifteen courses are C-programming (C), Visual Basic (VB), Active Server Pages (ASP), Computer Network (CN), Network Engineering (NE), Microprocessor (MP), Computer Organization (CO), Database Engineering (DBE), Advanced Database System (ADS), Operating System (OS), Distributed System (DS), Finite Automata System (FSA), Data Structure (DS-I), Software Engineering (SE), and Software Testing & Quality assurance (STQA). In this table yes represent that the student is interested in that particular course and no represent that student do not like that course.

Table 1
Sample Data from Moodle Database [8]

Courses→ *Roll No. ↓*	*C*	*VB*	*ASP*	*CN*	*NE*	*MP*	*CO*	*DBE*	*ADS*	*OS*	*DS*	*FSA*	*DS-I*	*SE*	*STQA*
1	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	yes	no	no
2	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no
3	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	yes	yes
4	no	no	no	yes	yes	no	yes	no	no	no	no	no	no	no	no
5	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	no	no
6	yes	yes	yes	no	no	no	no	no	no	yes	no	no	yes	no	no
7	no	no	no	yes	yes	yes	yes	no	no	no	no	no	no	yes	no
8	no	no	no	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
9	no	no	no	yes	yes	yes	yes	no	no	no	no	yes	no	no	no
10	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no
11	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
12	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no

Table Cont'd

Table 1 Cont'd

13	no	no	no	no	no	no	no	yes	yes	yes	yes	no	yes	yes	yes
14	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	no	no	no	no
15	yes	yes	yes	no	no	no	no	no	no	no	no	no	yes	no	no
16	no	no	no	yes	yes	no	no	yes	yes	yes	yes	no	yes	no	no
17	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	yes	yes
18	yes	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no
19	no	no	no	yes	yes	yes	yes	yes	yes	no	no	no	no	no	no
20	yes	no	no	no	no	no	no	no	no	yes	yes	no	yes	yes	yes
21	yes	no	yes	no	no	yes	yes	no	no	yes	yes	yes	no	no	no
22	no	no	no	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
23	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	no	yes	no	no
24	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
25	no	yes	yes	no	no	yes	yes	yes	yes	yes	yes	no	no	no	no
26	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
27	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no
28	no	no	no	yes	yes	no	no	no	no	yes	yes	no	yes	no	no
29	no	no	no	no	no	yes	yes	yes	yes	no	no	no	no	no	no
30	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	yes	yes
31	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no
32	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
33	no	no	no	yes	yes	no	no	no	no	yes	yes	no	yes	no	no
34	yes	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no
35	no	no	no	no	no	no	no	no	no	yes	yes	no	no	no	no
36	no	no	no	yes	yes	no	no	no	no	no	no	no	yes	no	no
37	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	no	no
38	no	no	no	no	no	no	no	no	no	yes	yes	yes	yes	yes	yes
39	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
40	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	yes
41	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
42	no	no	no	yes	yes	no	no	no	no	no	no	no	no	no	no
43	no	no	no	no	no	no	no	no	no	yes	yes	no	yes	no	no
44	no	no	no	no	no	no	no	no	no	no	no	no	no	no	yes
45	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

In table 2, we are considering only those courses from table 1 for which the classification algorithm classifies this course as "yes". For remaining courses, the classification algorithm gives more percentage of "no" compare to the percentage of "yes".

From table 2, we can observe that ADTree has highest percentage of correctly classified instance & lowest percentage of incorrectly classified instances. ZeroR classification algorithm has lowest percentage of correctly classified instances & highest percentage of incorrectly classified instances. Naive Bays has the 92.77 & 7.23 percentage for correctly & incorrectly classified instances. Simple Cart, J48, & Random Forest classification algorithm has 91.66%, 93.33% & 87.22% correctly classified instances respectively & 8.44%, 6.77%, and 12.88% incorrectly classified instances. Ascending order of classification algorithm considering the classification accuracy into account is ADTree, J48, Naive Bays, Simple Cart, Random Forest, and ZeroR. So we consider the ADTree as classification algorithm for Course Recommender System as classification accuracy for ADTree is highest.

Table 2
Result Using Different Classification Algorithm

Classification algorithm→ *Courses↓*		*Naive Bays*	*Simple Cart*	*ZeroR*	*ADTree*	*J48*
C-programming	Correctly classified instance	42	41	23	43	42
	Incorrectly classified instance	3	4	22	2	3
Operating System	Correctly classified instance	44	44	24	44	44
	Incorrectly classified instance	1	1	21	1	1
Distributed System	Correctly classified instance	43	44	23	44	44
	Incorrectly classified instance	2	1	22	1	1
Data Structure-I	Correctly classified instance	38	36	23	40	38
	Incorrectly classified instance	7	9	22	5	7
Overall result	Correctly classified instance	167 (92.77%)	165 (91.66%)	93 (51.66%)	172 (95.55%)	168 (93.33%)
	Incorrectly classified instance	13 (8.44%)	15 (7.23%)	87 (48.44%)	9 (4.45%)	12 (6.77%)

5. CONCLUSION AND FUTURE WORK

Here in this paper we compare the five classification algorithm to choose the best classification algorithm for recommending the course to student based on various student choices. These five classification algorithms, we consider for comparison, are ADTree, Simple Cart, J48, ZeroR, Naive Bays & Random Forest Classification Algorithm. We use the open source data mining tool Weka to check the result. We found that ADTree classification algorithm works better for this Course Recommender System as incorrectly classified instance for this algorithms are less than other five classification algorithms. Future works include the combination of other data mining algorithms to recommend the course to the student from the data obtained from the Moodle course of the college.

REFERENCES

[1] Hu H., Li J., Plank A., Wang H. and Daggard G., "A Comparative Study of Classification Methods for Microarray Data Analysis", *In Proc. Fifth Australasian Data Mining Conference*, Sydney, Australia (2006).

[2] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques".

[3] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", dasc, pp. 183-187, 2009 Eighth *IEEE International Conference on Dependable, Autonomic and Secure Computing*, 2009.

[4] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", *in International Journal on Computer Science and Engineering (IJCSE).*

[5] Rich Caruana Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms".

[6] "Data Mining Introductory and Advanced Topics", by Margaret H. Dunham.

[7] Sunita B. Aher and Lobo L.M.R.J. "Data Mining in Educational System using WEKA". *IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT)* (3), 20-25, 2011. Published by *Foundation of Computer Science*, New York, USA (ISBN: 978-93-80864-71-13).

[8] Sunita B. Aher and Lobo L.M.R.J. Article: "A Framework for Recommendation of Courses in E-learning System". *International Journal of Computer Applications,* **35(4)**, 21-28, December 2011. Published by *Foundation of Computer Science*, New York, USA ISSN 0975 - 8887 Digital Library URI: http://www.ijcaonline.org/archives/volume35/number4/4389-6091.

[9] http://www.d.umn.edu/~padhy005/Chapter5.html accessed on 30-01-2012

[10] http://en.wikipedia.org/wiki/Alternating_decision_tree accessed on date 02-02-2012