Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 3, No. 2, May 2012), Pages: 23-34
www.jacr.iausari.ac.ir

Comparison of Decision Tree and Naïve Bayes Methods in Classification of Researcher's Cognitive Styles in Academic Environment

Zahra Nematzadeh Balagatabi

Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran

zahra_nematzadeh@yahoo.com

Received: 2012/01/09; Accepted: 2012/03/16

Abstract

In today world of internet, it is important to feedback the users based on what they demand. Moreover, one of the important tasks in data mining is classification. Today, there are several classification techniques in order to solve the classification problems like Genetic Algorithm, Decision Tree, Bayesian and others. In this article, it is attempted to classify researchers to "Expert" and "Novice" based on cognitive style factors in order to have as best as possible answers. The domain of this research is based on academic environment. The critical point of this study is to classify the researchers based on Decision Tree and Naïve Bayes techniques and finally select the best method based on the highest accuracy of each method to help the researchers to have the best feedback based on their demands in the digital libraries.

Keywords: *Data mining, Classification, Cognitive styles, Decision tree, Naïve bayes, Academic environment*

1. Introduction

According to different needs of users in internet environments such as Digital Libraries, information services are prepared for them. For this propose, personalized digital libraries providing a way for different users to express their preferences clearly. Users will be confronting with a problem by using this approach. The problem is that users may not attention to their preferences and cannot have an acceptable research. To address these problems, this paper investigates an approach that gains user preferences based on cognitive style and recognizes relevant characteristics for information seeking and also do classification to classify the researchers. In this paper, researchers are classified to "Expert" and "Novice" based on cognitive style factors in order to have as best as possible answers in digital libraries.

Data mining is a machine learning approach and includes many tasks like concept description; cluster analysis; classification and prediction; trend and evaluation analysis; outlier analysis; statistical analysis and others. The most important tasks in data mining are classification and prediction techniques. The classification methods are known as supervised learning where the classification target and the class level are already recognized. There are several methods for classification specifically in data mining. These methods includes such as Decision Tree, Fuzzy Logic, Bayesian, Rough Set

Theory, Neural Network, Genetic Algorithm and Nearest Neighbor. The criterions for selecting an appropriate technique in some studies are dataset and the accuracy of model advanced by the techniques [1].

2. Classification

Recently, several classification methods have been presented by researchers in machine learning, statistics and pattern recognition. Clustering, association, classification and prediction are the main categories in data mining [2].

Through the years, different techniques have been developed by data mining [3]. These techniques execute the tasks which contains machine learning, database oriented techniques, statistic, pattern recognition, rough set, neural networks and the others. There are many hidden information in data mining and data ware house. This hidden information has an application in intelligent decision making which is alike with human decision.

Also there are two other methods which can provide an intelligent decision making. These two techniques are prediction and classification. They can be used to extract patterns which depicting significant data classes or to predict future data modes [4].

In addition, there are two phases in classification. The first phase is the learning process. In this phase, the training data are analyzed by classification algorithm and rules and patterns are created which are based on learned model or classifier. In the second phase the model is used for classification and test data are used for gaining the accuracy of classification patterns. Then, based on the acceptable accuracy, the rules can be used for the classification of new data or for unseen data (Figure 1) [1].

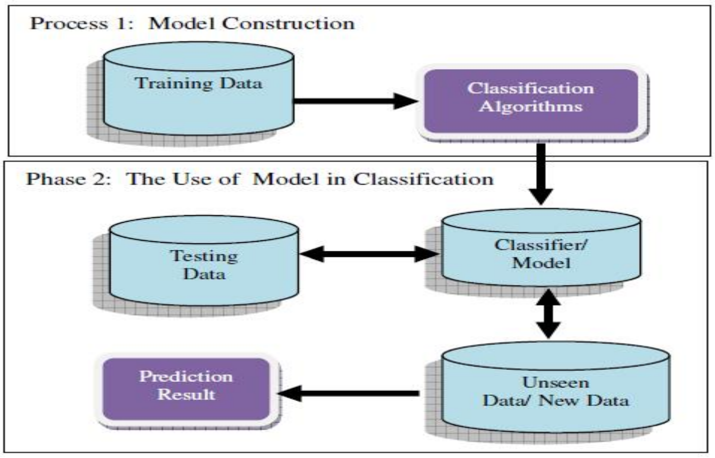
Process 1: Model Construction
Training Data
Classification Algorithms
Phase 2: The Use of Model in Classification
Testing Data
Classifier/ Model
Prediction Result
Unseen Data/ New Data

Figure 1. The process of classification

2.1 Decision Tree

Decision trees are widely used in the classification process. Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. This method is intended to build knowledge structures based on the data set. This method consists of a set of rules that will divide the large group to different smaller and standardized groups based on the targets defined variable. The decision tree usually results in the form of categories and decision tree model is used either to calculate the probability that the existing data set is categorized into the appropriate category. There are various methods in Decision tree, but only 6 of them are used here which includes J48, LMT, Random Forest, REP tree and Decision Stump [5].

In general, Decision Tree performs the classification process without involving many aspects of computation and complexity. Decision Tree is also able to generate rules that are easily understood and even easier to use the database. Decision Tree is the good method for providing guidance to determine the appropriate and most importantly parameters for classification or prediction. In terms of data processing, the Decision Tree does not require the data processor for doing processing for its own data. In fact, if the data is lost, Decision Tree will interpret the data by replacing missing data with new data randomly. In addition, the most important advantage of Decision Tree is to have a very high execution time and still produce a fairly accurate classification results when compared with other classification methods [6].

There is a statistical property which is a good measure for the value of an attribute which is called information gain. It is applicable for selecting the most useful attribute for classifying and it is also useful for measuring how well an existed attribute divides the training examples based on their target classification. This estimation is used to choose between the candidate features at each step during growing the tree.

It is needed to explain a measure which is used in information theory, named entropy for defining information gain accurately. Entropy describes the impurity of a collection of examples. The entropy of set S which includes positive and negative examples of some target concept (a two class problem) is presented below; where p_p is the proportion of positive examples in S and p_n is the proportion of negative examples in S [7].

$$\text{Entropy}(S) = -\, p_p\log_2 p_p - p_n\log_2 p_n \qquad (1)$$

The effectiveness of an attribute in classifying the training data can explain by having entropy which is a measure of the impurity in a set of training samples. This measure is the expected reduction in entropy and is happened by dividing the samples based on this attribute and is called information gain. In information gain, Gain (S, A), A refers to an attribute A and S represents a collection of examples and *Values(A)* is the set of all possible values for attribute A and S_v is the subset of S for which attribute A has value v [7]. The formula is represented as formula 2:

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

2.2 Naïve Bayes

Bayesian classifiers are simple probabilistic classifiers which are based on statistical classifiers. In this method, the probability of a given sample that is a member of a specific class can be predicted. Bayesian classifier which is based on bayes theorem assumes that all the features are independent [8]. According to this assumption, there is no dependency between the attribute value on a given class and the values of the other features. This hypothesis is named class conditional independence and because it makes the computation easier, it is called “naive” [9]. In summary, based on a naïve bayes classifier, there is no relation between the existence or not existence of a specific attribute of a class and existence or not existence of any other attributes. There are various methods in Naïve Bayes, but only 5 of them are used here which includes Naïve

Bayes, Naïve Bayes Multinomial, Naïve Bayes Multinomial Updateable, Naïve Bayes Updateable and Bayes Net [8].

The basic formula of Naïve Bayes classifier is shown in formula 3. Based on this formula, C refers to the training sample set which is divided into K categories C= $\{C_1,C_2,\ldots,C_k\}$; the prior probability of each category represents as p(C_f), where f=1,2…,k. For a given sample represents as d_t= ($W_1,\ldots,W_f,\ldots,W_m$) that feature words are represented as W_f, where f=1,2,…,m belongs to a specific category C_f. To classify the document d_t, the probability of all the documents regarding to a given d_t should be calculated [9]. The posterior probability of category C_f is calculated as follows:

$$p(C_f|d_t)=\frac{p(d_t|C_f)p(C_f)}{p(d_t)} \tag{3}$$

2.3 Data Set

Based on the studies, academic Environment is selected as a domain of this research. The participants are research students in University Technology Malaysia (UTM). They were 34 master research student and 76 PHD students from different faculties. The participants were from Computer Science, Electrical Engineering, Mechanical Engineering, Civil Engineering, Chemical Engineering, Built Environment and Management faculties and number of them were 40, 1, 22, 21, 8, 6 and 12 respectively.

In this step the questionnaire is prepared based on cognitive style which is based on [10]. The cognitive style instrument was selected in order to provide explanation of observed behavior of students when using Web search engines. Since the tool was self-assessment, students were asked to respond to the questions in a cognitive style in a real life situation.

The questionnaire was disseminated to 130 UTM research students but only 120 questionnaires were returned and 10 questionnaires were considered as incomplete data. So, this study is done based on 110 questionnaires. The analysis of this study is based on prediction of the student’s status whether “Expert” or” Novice”.

3. Researcher’s Cognitive Styles Variables and Attributes

Data is collected based on researcher’s Cognitive styles. The information was designed in questionnaire according to the cognitive style and information seeking variables. The questionnaire consists of 5 variables; where each variable is represented by several attributes. The variables are state of internal and personal knowledge, information seeking behaviour, Kuhlthau’s stages, information seeking activities and uncertainty. Table 1 shows the types of variables and attributes for datasets in general [11][12].

Table 1. Variables and attributes of cognitive styles

Variable	Attribute
State of personal or internal knowledge	• Broad conceptual knowledge of the domain • Specific knowledge or expertise of the problem • Familiarity with the language or terminology used in the problem or domain
Information Seeking Behaviour	• Clarity and focus of thought • Kuhlthau's stages: ▪ Initiation ▪ Selection ▪ Exploration ▪ Collection
Information Seeking Activities	• Ellis's information-seeking activities: ▪ Chaining ▪ Browsing ▪ Differentiating ▪ Maintaining ▪ Systematically working through ▪ verifying
Uncertainty	• Recognizing a real problem to investigate; • Defining the problem appropriately; • Resolving the problem; • Finding an effective way of presenting the results; • Finding relevant information

4. Evaluating the Classification Methods

In this phase, the testing was done in order to do classification. Testing process was developed to select the appropriate classification method. Accuracy is the first factor for evaluating. The selection was based on the accuracy of each method. The classification method with the highest accuracy will be selected. Also the error value for each method is obtained which includes the Square Root Error of Mean (RMSE), Mean Absolute Error for (MAE). In the following formulas, (x) represents the predicted value, (y) represents the actual value, (n) represents the total number:

$$MAE = \frac{1}{n}\sum_{i=0}^{n} x - y \qquad (4) \qquad\qquad RMSE = \sqrt{\sum \frac{(x-y)^2}{n}} \qquad (5)$$

5. Results and Discussions on Decision Tree

Testing methods for selecting a method of classification for decision tree involves 6 decision tree classification methods which are J48, LMT, Random Forest, Random Tree, REP Tree and Decision Stump. First, for preparing train and test data, 10-fold cross validation is done on the data set. In this way, each train data includes 99 data and each test data includes 11 data. Then, based on the training model, testing performed to obtain the accuracy and errors of each method. Each test on the classification method will be recorded based on the value of accuracy, MAE and RMSE. Once the accuracy and the error value for all the tested methods are recorded, the comparison on each of the methods implemented. The results of each method are presented in table 2 to table 7. All the experiments are done in WEKA environment.

Table 2. Results of J48

J48		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	10/11	90.9091
3	11/11	100
4	11/11	100
5	11/11	100
6	10/11	90.9091
7	11/11	100
8	7/11	63.6364
9	11/11	100
10	10/11	90.9091
Average Accuracy (%)	92.72728	

Table 3. Results of LMT

LMT		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	10/11	90.9091
3	10/11	90.9091
4	10/11	90.9091
5	11/11	100
6	10/11	90.9091
7	10/11	90.9091
8	7/11	63.6364
9	11/11	100
10	11/11	100
Average Accuracy (%)	91.81819	

Table 4. Results of random forest

Random Forest		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	10/11	90.9091
3	11/11	100
4	10/11	90.9091
5	11/11	100
6	10/11	90.9091
7	11/11	100
8	8/11	72.7273
9	9/11	81.8182
10	11/11	100
Average Accuracy (%)	92.72728	

Table 5. Results of random tree

Random Tree		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	10/11	90.9091
3	10/11	90.9091
4	10/11	90.9091
5	11/11	100
6	10/11	90.9091
7	10/11	90.9091
8	8/11	72.7273
9	10/11	90.9091
10	10/11	90.9091
Average Accuracy (%)	90.9091	

Table 6. Results of REP tree

REP Tree		
Testing	Number of correctly classified instances	Accuracy (%)
1	11/11	100
2	9/11	81.8182
3	11/11	100
4	9/11	81.8182
5	10/11	90.9091
6	10/11	90.9091
7	11/11	100
8	7/11	63.6364
9	11/11	100
10	10/11	90.9091
Average Accuracy (%)	90.00001	

Table 7. Results of decision stump

Decision Stump		
Testing	Number of correctly classified instances	Accuracy (%)
1	8/11	72.7273
2	9/11	81.8182
3	7/11	63.6364
4	9/11	81.8182
5	7/11	63.6364
6	10/11	90.9091
7	8/11	72.7273
8	7/11	63.6364
9	8/11	72.7273
10	9/11	81.8182
Average Accuracy (%)	74.54548	

Figure 2 shows the average accuracy of each method. Based on figure 2, it is clear that J48 and Random Forest have the same average accuracy with 92.72728, therefore; the value of MAE and RMSE should be measured in order to find the best method.

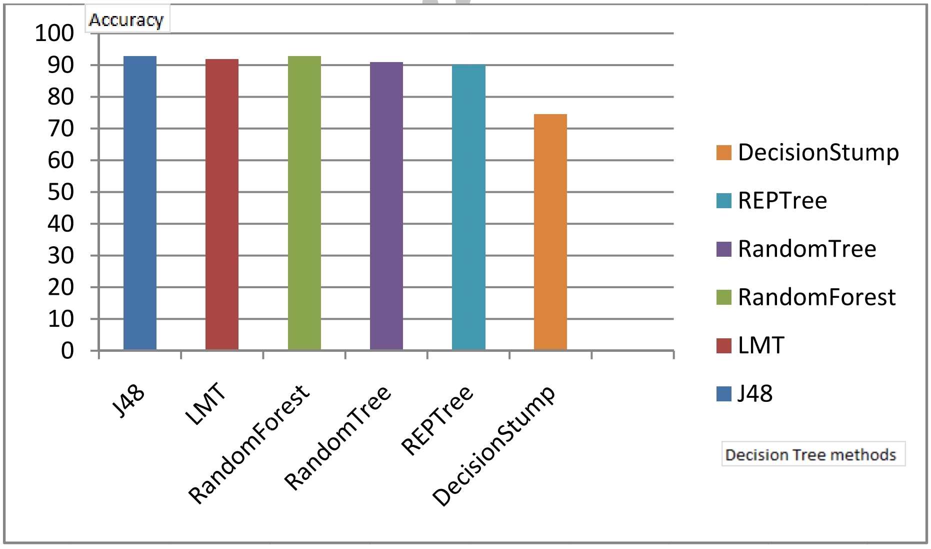

Figure 2. Average accuracy of 6 methods of decision tree

To choose the best method, if the values of the accuracy are same, then total value of the MAE will be measured. The classification method that produces the smallest MAE value will be selected. The next step is to determine the method of classification with the smallest MAE value among the best methods, if the values of MAE were same, the method with the largest RMSE value should be selected. Based on figure 2, it is clear that J48 and Random Forest have the same average accuracy with 92.72728. For choosing the best method among these two methods, it is necessary to find out the average value of MAE and RMSE error (table 8).

Table 8. Value of MAE, RMSE for J48 and random forest

	J48		Random Forest	
Number	MAE	RMSE	MAE	RMSE
1	0.0642	0.1023	0.0439	0.1113
2	0.1166	0.2641	0.0909	0.246
3	0.0812	0.1217	0.0839	0.1733
4	0.1212	0.3178	0.1308	0.3178
5	0.1004	0.1217	0.0646	0.1273
6	0.0994	0.2612	0.1212	0.2701
7	0.1012	0.2249	0.0561	0.1475
8	0.3377	0.5347	0.3545	0.5568
9	0.061	0.1017	0.1686	0.2677
10	0.1221	0.2547	0.0582	0.1365
Average value	0.1205	0.23048	0.11727	0.23543

From table 2 and table 4 and also figure 2, although the accuracy value of J48 and Random Forest are the same, but in terms of average value of MAE in table 8, Random Forest method produces the smallest error (figure 3). So, in this case, it can be concluded that in the decision tree classification method, Random Forest, produces the highest accuracy with the smallest average value of MAE which is the best method.

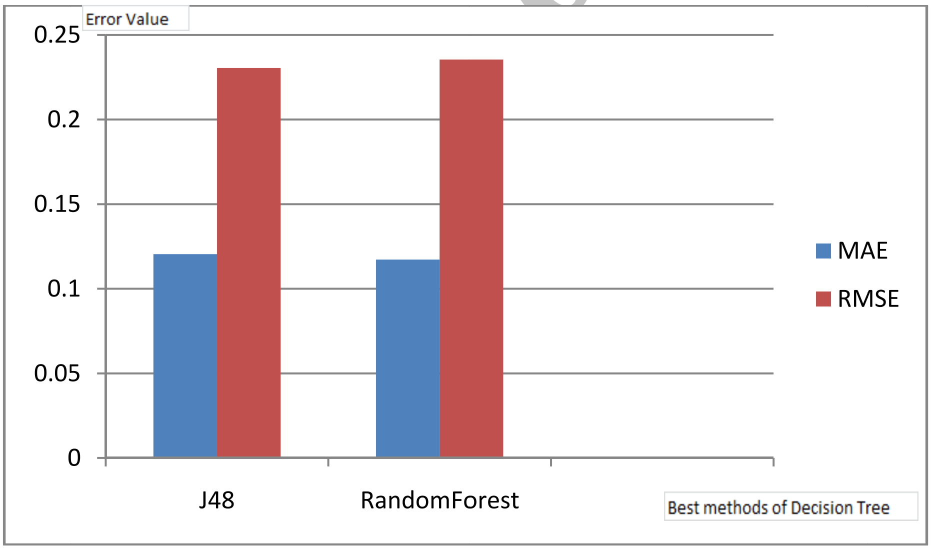

Figure 3.Values of MAE and RMSE for J48 and random forest

6. Results and Discussion on Naïve Bayes

Testing methods for selecting a method of classification for Naïve Bayes involves 5 naïve bayes classification methods which are Naïve Bayes, Naïve Bayes Multinomial, Naïve Bayes Multinomial Updateable, Naïve Bayes Updateable and Bayes Net. Here, the experiment is done on the same train and test data which was used in Decision Tree. Testing performed to obtain the accuracy and errors of each method. Tables 9 to13 show the results of running experiments to compare the accuracy for each classification methods.

Table 9. Results of Naïve Bayes

Naïve Bayes		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	10/11	90.9091
3	11/11	100
4	10/11	90.9091
5	11/11	100
6	10/11	90.9091
7	11/11	100
8	7/11	63.6364
9	11/11	100
10	11/11	100
Average Accuracy (%)	93.63637	

Table 10. Results of Naïve Bayes updateable

Naïve Bayes Updateable		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	10/11	90.9091
3	11/11	100
4	10/11	90.9091
5	11/11	100
6	10/11	90.9091
7	11/11	100
8	7/11	63.6364
9	11/11	100
10	11/11	100
Average Accuracy (%)	93.63637	

Table 11.Results of Naïve Bayes multinomial

Naïve Bayes Multinomial		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	7/11	63.6364
2	5/11	45.4545
3	8/11	72.7273
4	5/11	45.4545
5	7/11	63.6364
6	7/11	63.6364
7	8/11	72.7273
8	5/11	45.4545
9	8/11	72.7273
10	6/11	54.5455
Average Accuracy (%)	60.00001	

Table 12. Results of Naïve Bayes multinomial updateable

Naïve Bayes Multinomial Updateable		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	7/11	63.6364
2	5/11	45.4545
3	8/11	72.7273
4	5/11	45.4545
5	7/11	63.6364
6	7/11	63.6364
7	8/11	72.7273
8	5/11	45.4545
9	8/11	72.7273
10	6/11	54.5455
Average Accuracy (%)	60.00001	

Table 13. Results of Bayes Net

Bayes Net		
Testing	**Number of correctly classified instances**	**Accuracy (%)**
1	11/11	100
2	9/11	81.8182
3	11/11	100
4	9/11	81.8182
5	10/11	90.9091
6	10/11	90.9091
7	10/11	90.9091
8	7/11	63.6364
9	11/11	100
10	11/11	100
Average Accuracy (%)	90.00001	

To choose the best method, if the values of the accuracy are same, then total value of the MAE will be measured. The classification method that produces the smallest MAE value will be selected. The next step is to determine the method of classification with the smallest MAE value among the best methods, if the values of MAE were same, the method with the largest RMSE value should be selected. Among these 5 classification methods, there are 2 methods with the highest and also the same accuracy which is shown in figure 4.

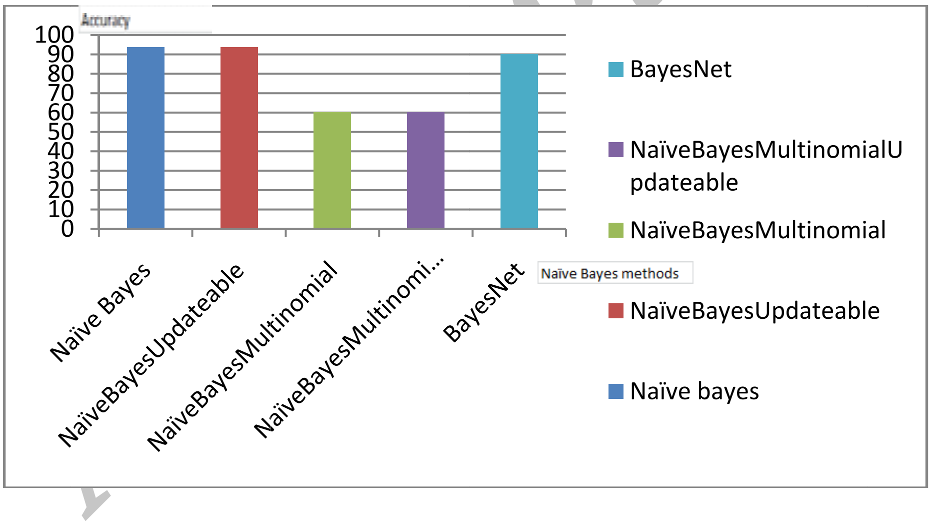

Figure 4. Average accuracy of 5 methods of Naïve Bayes

From figure 4, although the accuracy value of Naïve Bayes and Naïve Bayes updateable are same with 93.63637, in terms of average value of MAE, both methods produce same MAE and RMSE error (table 14). So, based on figure 5, it can be concluded that in the Naïve Bayes classification method, Naïve Bayes and Naïve Bayes Updateable produce the highest accuracy with the smallest MAE which are the best methods.

Table 14. Value of MAE, RMSE for Naïve Bayes and Naïve Bayes updateable

	Naïve Bayes		Naïve Bayes updateable	
Number	MAE	RMSE	MAE	RMSE
1	0.662	0.092	0.6662	0.0992
2	0.1121	0.271	0.1121	0.271
3	0.0704	0.1299	0.0704	0.1299
4	0.1358	0.2622	0.1358	0.2622
5	0.0442	0.0684	0.0442	0.0684
6	0.1359	0.2984	0.1359	0.2984
7	0.0542	0.1179	0.0542	0.1179
8	0.3367	0.5149	0.3367	0.5149
9	0.111	0.1756	0.111	0.1756
10	0.1021	0.175	0.1021	0.175
Average value	0.17686	0.21125	0.17686	0.21125

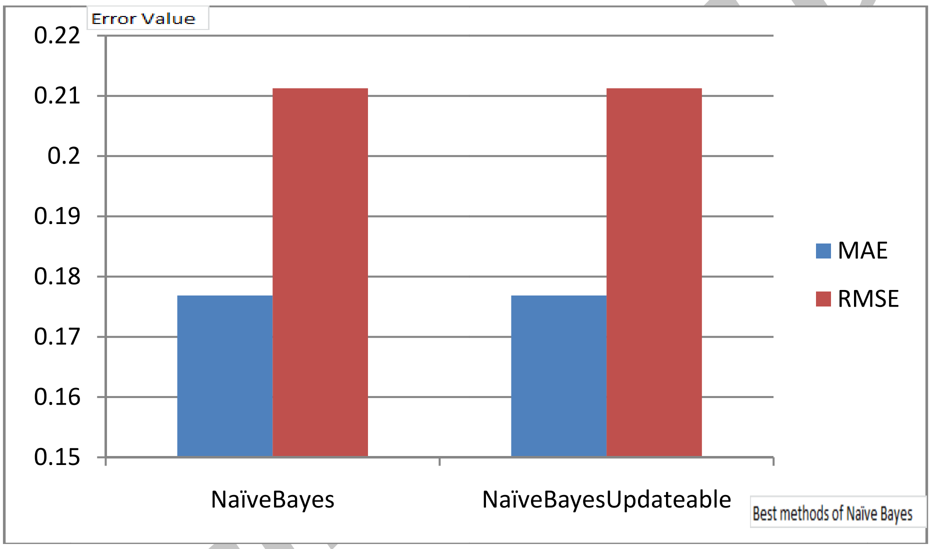

Figure 5. Values of MAE and RMSE for Naïve Bayes and Naïve Bayes updateable

7. Discussion on the Results

In this section the results produced with Decision Tree and Naïve Bayes algorithms were compared and a brief discussion was given on the results. Moreover, as the table 15 shows, among these two methods decision tree had the worst results with the accuracy of 92.7278 and Naïve Bayes with the accuracy of 93.63637 is the best method. However, if we want to compare these two methods based on the average accuracy of algorithms used in them then we can conclude that Decision Tree with the average accuracy of 88.78789% is better than Naïve Bayes with the average accuracy of the 79.454554.

Table 15. Comparison of results of Naïve Bayes and decision tree methods

Method	Naïve Bayes	Decision Tree
Best accuracy	93.63637	92.72728
Average accuracy	79.454554	88.78789

8. Conclusion

Here, it is attempted to classify researchers to “Expert” and “Novice” based on cognitive style factors in order to have as best as possible answers. For this purpose, the questionnaire is prepared based on researcher’s cognitive styles variables, also the domain of this research is based on academic environment. An important point of this study is to classify the researchers based on Decision Tree and Naïve Bayes techniques and finally select the best method of classification based on the highest accuracy to help the researchers to have the best feedback based on their demands in the digital libraries. In conclusion, the results of 6 methods of Decision Tree and 5 methods of Naïve Bayes are presented in order to find out the best method of each technique. In this case, the researchers are classified to expert or novice based on their cognitive styles. Based on the best accuracy, it can be concluded that web developers can use Naïve Bayes or Naïve Bayes updateable technique in comparison with Decision Tree in order to classify the researchers and help them to have a best feedback based on their demands in digital libraries.

9. References

[1] H. Jantan, A.R. Hamdan, and Z.A. Othman, "Classification for Talent Management Using Decision Tree Induction Techniques," *Science And Technology*, 2009.

[2] J. Ranjan, "Data Mining Techniques for better decisions in Human Resource Management Systems," *International Journal of Business Information Systems*, 2008.

[3] C.F. Chien and L.F. Chen, ""Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry,,"" *Expert Systems and Applications*, 2008.

[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2006.

[5] A. Abdelhalim and I. Traore, "A new method for learning decision trees from rules," *In the Eighth International Conference on Machine Learning and Applications (ICMLA'09), Miami, Florida, USA, December*, 2009.

[6] D. Larose, "An Introduction to Data Mining," 2005.

[7] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, "Classification and regression trees," *Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software*, 1984.

[8] L. Yuan, "An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing," *Science*, 2010, pp. 267-269.

[9] Kim.S, Han.K, Rim.H, and Myaeng.S, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Transactions on Knowledge and Data Engineering*, 2006.

[10] N. Ford, T.D. Wilson, A. Foster, D. Ellis, and A. Spink, "Information Seeking and Mediated Searching. Part 4 .Cognitive Styles in Information Seeking," *Journal of the American Society for Information Science*, vol. 53, 2002, pp. 728 -735.

[11] N. Ford, F. Wood, and C. Walsh, "Cognitive styles and searching. On-line & CD ROM Review," 1994.

[12] A. Spink, T.D. Wilson, N. Ford, A. Foster, and D. Ellis, "Information-Seeking and Mediated Searching. Part 1 .Theoretical Framework and Research Design," *Journal of the American Society for Information Science*, 2002.