Jurnal Teknologi

Full paper

Comparison of Decision Tree Methods in Classification of Researcher's Cognitive styles in Academic Environment

Zahra Nematzadeh Balagatabi*, Roliana Ibrahim, Hossein Nematzadeh Balagatabi

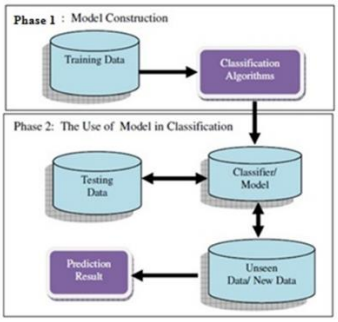
Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: zahra_nematzadeh@yahoo.com

Article history

Received :22 June 2012
Received in revised form :
20 January 2015
Accepted :15 March 2015

Graphical abstract



Abstract

In today's internet world, providing feedbacks to users based on what they need and their knowledge is essential. Classification is one of the data mining methods used to mine large data. There are several classification techniques used to solve classification problems. In this article, classification techniques are used to classify researchers as "Expert" and "Novice" based on cognitive styles factors in academic settings using several Decision Tree techniques. Decision Tree is the suitable technique to choose for classification in order to categorize researchers as "Expert" and "Novice" because it produces high accuracy. Environment Waikato Knowledge Analysis (WEKA) is an open source tool used for classification. Using WEKA, the Random Forest technique was selected as the best method because it provides accuracy of 92.72728. Based on these studies, most researchers have a better knowledge of their own domain and their problems and show more competencies in their information seeking behavior compared to novice researchers. This is because the "experts" have a clear understanding of their research problems and is more efficient in information searching activities. Classification techniques are implemented as a digital library search engine because it can help researchers to have the best response according to their demand.

Keywords: Data mining; classification; cognitive style; decision tree; academic environment

Abstrak

Dalam dunia internet hari ini, memberi maklum balas kepada pengguna berdasarkan apa yang mereka perlukan dan pengetahuan mereka adalah penting. Klasifikasi adalah salah satu kaedah dalam perlombongan data untuk melombong data yang banyak. Terdapat beberapa teknik pengkelasan yang digunakan untuk menyelesaikan masalah klasifikasi, Dalam artikel ini, teknik klasifikasi digunakan untuk mengklasifikasikan penyelidik sebagai "Pakar" dan "Novice" berdasarkan kepada faktor-faktor gaya kognitif dalam persekitaran akademik menggunakan beberapa teknik Pepohon Keputusan. Teknik Pepohon Keputusan yang memberi ketepatan yang tinggi merupakan teknik yang terbaik untuk dipilih sebagai teknik untuk klasifikasi kategori penyelidik sebagai "Pakar" dan "Baru". Persekitaran Waikato Analisis Pengetahuan (WEKA) adalah alatan sumber terbuka yang digunakan untuk pengkelasan. Menggunakan WEKA, teknik Hutan Rawak telah dipilih sebagai kaedah terbaik kerana memberi ketepatan 92,72728. Berdasarkan kajian, kebanyakan pakar penyelidik mempunyai pengetahuan tersendiri yang lebih baik terhadap domain masalah mereka serta mempunyai kecekapan tingkah laku carian maklumat yang lebih tinggi berbanding dengan penyelidik baru. Ini adalah kerana "Pakar" mempunyai pengetahuan yang lebih jelas terhadap masalah penyelidikan mereka dan lebih cekap didalam aktiviti carian maklumat. Teknik pengkelasan dilaksanakan sebagai enjin carian di perpustakaan digital kerana ia dapat membantu penyelidik untuk mempunyai maklum balas yang terbaik mengikut permintaan mereka.

Kata kunci: Perlombongan data; klasifikasi; gaya kognitif ; decision tree; persekitaran akademik

1.0 INTRODUCTION

Information services are prepared for users in internet environments such as Digital Libraries in accordance with their different needs. For this purpose, personalized digital libraries provide a way for different users to express their preferences clearly. The problem confronting users by using this approach is that they may not pay attention to their preferences and thus the research would not be acceptable. To address these problems, this paper investigates an approach that gains user preferences based on cognitive style and recognizes relevant characteristics for information seeking. It also seeks to classify researchers. In this paper, researchers are classified as "Expert" and "Novice" based on cognitive style factors in order to obtain the best possible answers in digital libraries. Data mining is a machine learning approach and includes many tasks such as: 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 [1]. 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 relating to data mining. These methods include: Decision Tree, Fuzzy Logic, Bayesian, Rough Set Theory, Neural Network, Genetic Algorithm and Nearest Neighbor. The criteria for selecting an appropriate technique in some studies are dataset and the accuracy of a model advanced by the techniques [1]. So, based on the dataset, decision tree has been selected as one of the appropriate method in this study.

2.0 CLASSIFICATION

Recently, several classification methods have been presented by researchers in the areas of 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 tasks containing machine learning, database oriented techniques, statistics, pattern recognition, rough set, neural networks and others. There is quite a lot of hidden information contained in data mining and the data ware house. This hidden information has an application in intelligent decision making which is comparable with the process of human decision making. There are also two other methods which can provide intelligent decision making. These two techniques are namely prediction and classification. They can be used to extract patterns which depict significant data classes or to predict future data modes [4] . In addition, there are two phases involved in classification. The first phase is the learning process. In this phase, the training data is analyzed by classification algorithm, and rules and patterns are created based on a learned model or classifier. In the second phase, the model is used for classification and test data is used for gaining accuracy of classification patterns. Subsequently, based on the acceptable accuracy, the rules can be used for the classification of new data or for unseen data (Figure 1) [1].

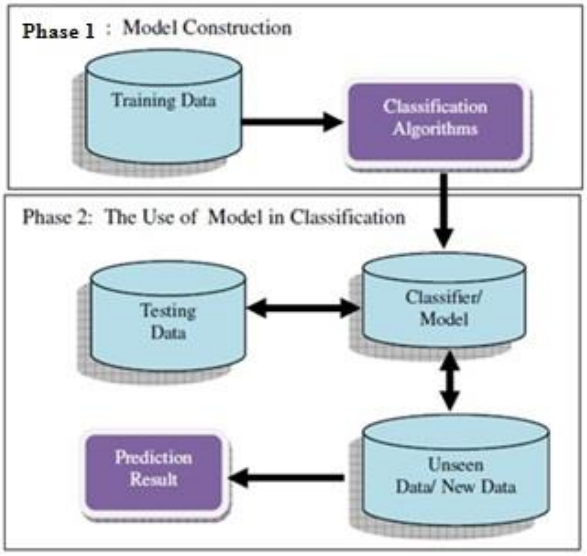


Figure 1 The process of classification

3.0 RESEARCH BACKGROUND

In this section research background in classification techniques and cognitive style are presented:

3.1 Research Background In Classification Techniques

Based on L.jayasimman *et al.,* a questionnaire is prepared to determine the user's cognitive behavior in a web learning environment and was disseminated among 100 students in affiliated colleges of Bharathidasan University. The questionnaire was applied to recognize the areas to improve in layout of the web learning system which is used as the class label for the decision tree algorithm. The algorithms which are used in this research are based on CART, Random Forest, Random Tree and the Naïve Bayes Classifier. The cognitive attributes are applied as the training input for the algorithms. According to the experiments, among these four algorithms, Random Forest was the best classifier and predicted accurately with an accuracy of 75%. [5]

Also another research is done based on predicting the student's performance which is the major concern to the higher education managements. The dataset was collected among 200 postgarduate students of computer scinence course. In this research two algorithms, decision tree (ID3) and Naïve bayes, were conducted and their performance was evaluated. Form the result, decision tree algorithm (ID3) was more accurate than Naïve Bayes and also gave 98% prediction of 50 instances but the error rate was very high in Naïve Bayes and the accuracy was relatively lower than ID3 algorithm. [6]

Based on Charles A. Worrell *et al*. research, an experimental study is done in order to predict ranking of the 12 month risk of defaults in banks. The prediction is done by comparing the classification techniques. The scoring capabilities of different predictive models are compared in this research. Based on the comparison the inductive machine learning can be used for prediction of default risk. According to the achieved results, symbolic rule or decision tree based models conduct higher performance than traditional modeling techniques based on statistical algorithms. [7]

According to Peiman Mamani Barnaghi *et al.* research, four classification techniques are used to recognize the relation of liver disorder and drinking alcohol drink by classification of blood test data. MLP and RBF in neural network, naïve bayes and bayes net in Bayesian, J48 and LMT in decision tree and rough set are used in this work. Observed results demonstrates that neural networks classifier methods have better result than the others. So, MLP attains higher results than RFB, J48 performs better than LMT but rough sets did not perform well in comparison with other methods. Based on the assumption of increasing the size of training set in liver disorder, MLP demonstrates that can provide better results with larger training set. So, experiment results shows that neural network attains best result in this research. [8]

According to Shweta Kharya's research various data mining approaches have been applied for breast cancer diagnosis in order to enhance the breast cancer diagnosis and prognosis. These data mining techniques include Neural Network, Association Rule Mining, Naïve Bayes, C4.5 decision tree algorithm, Bayesian Networks. Among these various techniques and soft computing approaches, Decision tree is considered as the best classifier with 93.62% accuracy. [9]

Arihito endo *et al.* presented optimal models to predict the survival rate of breast cancer patients. This research was done on the 37, 256 follow-up patients that were diagnosed as breast cancer and registered in the SEER program from 1992 to 1997. The algorithms which were used in this study include Logistic Regression model, Artificial Neural Network (ANN), Naive Bayes, Bayes Net, Decision Trees with naive Bayes, Decision Trees (ID3) and Decision Trees (J48)) besides the most widely used statistical method (Logistic Regression model)in order to produce the prediction models. Based on the results, Logistic Regression model showed the highest accuracy with 85.8±0.2%. Artificial Neural Network display the highest specify and Decision Trees (J48) model with the highest sensitivity tended to be more sensitive to survival prediction and Bayesian model tended to be more sensitive to death .[10]

Mohd Fauzi bin Othman *et al*. investigates the performance of different classification or clustering methods for a set of large data. 6291data are collected from breast cancer data which will be applied to test and evaluate the various classification methods. The methods which are used in this work are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. According to the results, bayes network classifier with an accuracy of 89.71% is the best algorithm. Finally, among other classification methods, bayes network can be the best method in medical or in general, bioinformatics field. [11]

3.2 Research Background In Cognitive Styles

Cognitive styles are used in many fields which are relating to library and information studies (LIS). The following table demonstrates some of these works (Table 1). [12]

Cognitive style also has an effect on information seeking which is the main point of some recent studies. The information-seeking context is classified from databases, hypertext, and virtual information environments to on-line and Web based searching. The following table demonstrates some of related studies (Table 2).

Table 1 Cognitive styles and related studies

Authors	subject
(Davidson 1977)	Linking cognitive styles with document relevance judgments. [13]
(Rholes and Droessler 1984)	Surveys of the incidence of different styles among reference librarians. [14]
(Johnson & White 1981b)	Librarianship students. [15]
(Johnson & White 1982)	The application of cognitive style data to enhance LIS teaching. [16]
(Montgomery 1991)	To studies linking cognitive styles to levels of cooperation between teachers and library media center specialists. [17]
(Huang 1998)	Investigations of cognitive styles and preference for display layouts. [18]
(Crossland *et al*. 2000)	Decision making in geographical information systems. [19]
(Palmquist 2001)	Choice of metaphor for describing the Web. [20]

Table 2 Cognitive style and information seeking

Author	subject
(Ford & Ford 1992)	Conducted an experiment with postgraduate students to discover how they might go about learning from an "ideal" database. [21]
(Ellis *et al*. 1992)	Investigated hypertext navigation by 40 postgraduate students. [22]
(Chou & Lin 1998)	Studied the effects of navigation map types and cognitive styles on performance by 121 university students in searches for information and cognitive map development using a hypertext system. [23]
(Wang *et al*. 2000)	Investigated cognitive and affective aspects of Web searching by 24 Masters students. [24]
(Palmquist and Kim 2000)	Studied the effects of both experience and cognitive style on Web searching. [25]

4.0 DECISION TREE

Decision trees are widely used in the classification process [26]. Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by people and used in knowledge systems such as database. This method is intended to build knowledge structures based on the relevant data set. This method consists of a set of rules that will divide a large group into separate smaller and standardized groups based on the targets defined variable. The decision tree usually results in the form of categories and a decision tree model is used to calculate the probability that the existing data set is categorized into the appropriate category. There are various methods in Decision tree techniques, but only 6 of them are used here. These include J48, LMT, Random Forest, REP tree, Random Tree and Decision Stump [26]. A brief definition of each method is presented below:

Random forest: This algorithm is one of the most accurate learning algorithms which produces highly accurate classifier in many data sets [27]. It was developed by Leo Breiman [28] and Adele Culter. Random forest can run effectively on large data bases. Also, important variables in classification can be estimated

by random forest. And it can be applied in many input variables without any variable elimination. [27]

J48: This algorithm is WEKA's implementation of the C4.5 decision tree learner. In this algorithm a greedy technique is used for inducing decision trees for classification in this algorithm and also uses reduced- error pruning. [29]

LMT: A combination of induction trees and logistic regression produces LMT method. It is a combination of learners which depends on simple regression models, if little or noisy data are available, on the other hand, a more complicated tree structure will be added if there is enough data to warrant such a structure. In comparison with other algorithms it is slower considerably. [30]

Reduced Error Pruning (REP): REP Tree as fast decision tree learner which just sorts values for numeric attributes once. It creats a decision/regression tree by applying entropy as impurity measure and prunes it using reduced-error pruning. [31]

Decision stump: it is a machine learning model which includes one-level decision tree. The prediction in decision stump is based on the value of an individual input feature. It is a decision tree which connects the root (internal nodes) to the leaves (terminal nodes). [32]

Random tree: there is no pruning in random tree method. This method builds a tree which uses K random features in each node. [33]

In general, Decision Tree performs the classification process without involving many aspects of computation and complexity. Decision Tree techniques are also able to generate rules that are easily understood and make it even easier to use the database. Decision Tree is a good method for providing guidance to determine the appropriate and most important parameters for classification or prediction. In terms of data processing, the Decision Tree does not require the data processor for processing its own data. In fact, if the data is lost, Decision Tree will interpret the data by replacing missing data randomly with new data. In addition, the most important advantage of Decision Tree is to have a very high execution time and still produce fairly accurate classification results when compared with other classification methods [34].

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

We need to explain a measure named entropy which is used in information theory for defining an information gain accurately. Entropy describes the impurity of a collection of examples. The entropy of set S, including positive and negative examples of a 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 [35].

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

The effectiveness of an attribute in classifying the training data can be explained 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 occurs 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. 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 [35]. The formula is represented as Equation 2:

$$\text{Gain}(S, A) \equiv \text{Entropy}(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

The process of using information gain in this study is presented in Figure 2. According to the Figure 2, four variables are considered as input dataset. Kuhlthau's stages are a variable which is Information Seeking Behaviour's attribute too, therefore; it is considered as an attribute in evaluation. By recognizing the relation between an input and the targeted outcome, the algorithm will identify the most useful single attribute which obviously separates the outcomes. By calculating information gain, the attribute with the best score will be selected to divide the cases into subsets. This process will be done recursively until the tree cannot be split any more. (Figure 2)

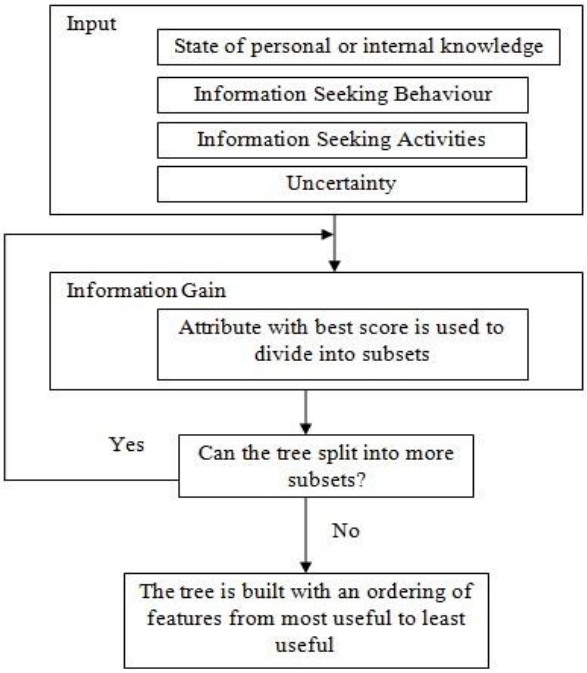


Figure 2 The process of information gain

5.0 CASE STUDY

Based on the studies, an academic Environment was selected as a domain of this research. The participants are research students from UTM. They comprised 34 master research students and 76 PHD students from different faculties. The participants were from the faculties of Computer Science, Electrical Engineering, Mechanical Engineering, Civil Engineering, Chemical Engineering, Built Environment and Management and comprised 40, 1, 22, 21, 8, 6 and 12 participants respectively.

5.1 Dataset

In this step the questionnaire is prepared built on a cognitive style which is based on Ford *et al.* work [36]. The cognitive style instrument was selected in order to provide an 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. However only 120 questionnaires were returned and 10 questionnaires were considered as incomplete data. So, this study is carried out based on 110 questionnaires. The analysis of this study is based on a prediction of the student's status, whether "Expert" or "Novice". Based on this research, "Novice" defines to the beginners with less internal and personal knowledge in the domain of their problem and their information seeking behavior is less clear and pessimistic in comparison with expert researchers. Moreover, the importance of their information seeking activities is recognized vague and also they are not certain in their research problem.

5.2 Researcher's Cognitive Styles' Variables And Attributes

Data is collected based on the respective researcher's Cognitive styles. The information was designed in a questionnaire according to cognitive style and information-seeking variables. The questionnaire consists of 5 variables; where each variable is represented by several attributes. The brief discussion of each variables and attributes are presented below [12] and also Table 3 shows the types of variables and attributes for data sets in general. [37][38]

- State of personal or internal knowledge which is divided in classes of: level of conceptual knowledge of the domain; specific knowledge or expertise of the problem; familiarity with the language or terminology used in the problem or domain.[12]
- Clarity and focus of thought. Participants must answer to the question that "How would you describe your thinking about the problem at this stage?" and should determine their position between the two ranges of "general or vague" and "clear or focused".[12]
- Kuhlthau's stages. Participants must specify which of the following stages they were currently at: initiation (having recognized that they needed information information), selection (having identified the general area in which information is needed), exploration (identifying potentially useful information sources), collection (collecting specific information, having focused the problem), formulation (having formed a clearer focus on the problem on the basis of information found), or presentation (in the process of finishing the collection of information). [12]
- Ellis's information-seeking activities. Participants must answer and specify their position in each of these are: chaining (following the chains of citations or other forms of referential connection between documents); browsing (semi directed searching in an area of potential interest); differentiating (distinguishing between different sources of information on the basis of the nature and quality of the material examined); maintaining (keeping awareness of developments in relation to the topic through the monitoring of particular sources); systematically working through (systematically examining a particular source to locate material of interest); and verifying (checking the accuracy of information. [12]
- Uncertainty which is in terms of: a real problem for investigating had been recognized by researcher; the problem is defined by the researcher appropriately; the problem could be resolved; an effective way of presenting the results could be found; relevant information was available and could be found. [38]

Table 3 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
Kuhlthau's stages	• Initiation • Selection • Exploration • Collection
Information Seeking Activities	• Ellis's information-seeking activities: ▪ Chaining ▪ Browsing ▪ Differentiating ▪ Maintaining ▪ Systematically working through ▪ verification
Uncertainty	• Recognizing a real problem to investigate; • Defining the problem appropriately; • Resolving the problem; • Finding an effective way to present the results; • Finding relevant information

5.3 Research Framework

The entire process of this study is shown in the research framework (Figure 3). From the framework can be concluded that the first thing to identify is the data set and preparing the questionnaire subsequently. In this step the questionnaire is prepared built on a cognitive style which is based on Ford et.al [12]. The cognitive style instrument was selected in order to provide an 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.

Then, the questionnaire was disseminated to 130 UTM research students. However only 120 questionnaires were returned and 10 questionnaires were considered as incomplete data. So, this study is carried out based on 110 questionnaires. The analysis of this study is based on a prediction of the student's status, whether "Expert" or" Novice".

In this study, the questionnaire includes two parts which are demographic, and main questions. In the demographic part, the respondents were required to state their major, degree, semester, age and gender. Before the main questions of this questionnaire, there are four other questions which were proposed in this research in order to predict the researchers as expert or novice. In

this part, the respondents must state their background knowledge and experimental skills in their current research problem. Also, they have to mention their publications and the period of time that they spent for their research problem. Based on this measurement, they have been predicted as Expert or Novice.

In the main part of the questionnaire, the questions are divided into four parts. The first part states the personal or internal knowledge of researchers. In the second part, the information seeking behavior is stated and it presents the Kuhlthau's stage. In the third part, the information seeking activities are presented. And finally, the last part presents uncertainty.

There are three steps in data preparation and preprocessing. These steps are assigning the value, normalization and training and testing data.

The last step in preprocessing is to separate the data to train and test data in order to validate the model. This has been done by the 10-fold cross validation technique.

Finally, for validation the accuracy of each method is obtained and among six various methods in decision tree the best method is selected.

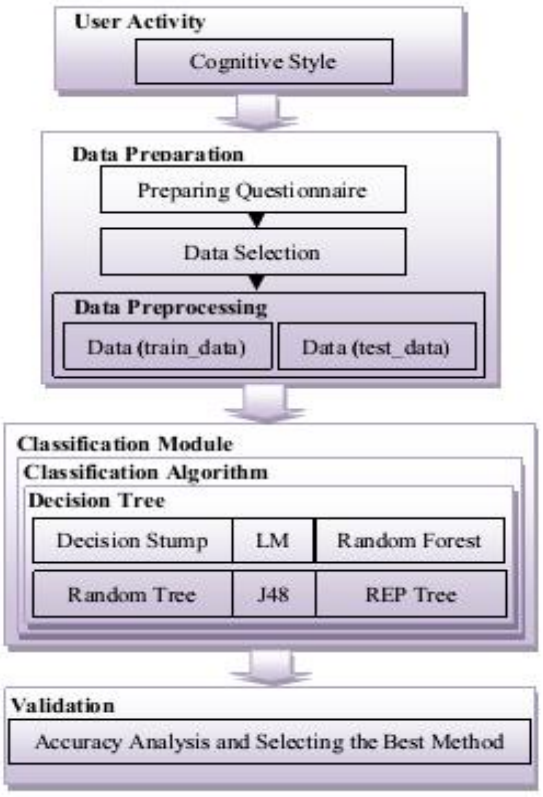


Figure 3 Research frame work

5.4 Evaluating The Decision Tree Classification Methods

In this phase, testing was done in order to perform classification. A testing process was developed to select the appropriate classification method. Accuracy was the first factor in evaluation. The selection was based on the accuracy of each method. The classification method with the highest accuracy was selected. In addition, the error value for each method was obtained, including the Square Root Error of Mean (RMSE) and Mean Absolute Error for (MAE). The MAE is a linear score which means that all the individual differences are weighted equally in the average. It measures accuracy for continuous variables. The RMSE is a quadratic scoring rule which measures the average magnitude of the error. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample.

In the following formulae, (x) represents the predicted value, (y) represents the actual value, and (n) represents the total number:

$$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n} x - y \qquad (3)$$

$$\text{RMSE} = \sqrt{\sum \frac{(x-y)^2}{n}} \qquad (4)$$

6.0 RESULTS AND DISCUSSIONS

Testing methods for selection of a method of classification for decision tree involved 6 decision tree classification methods. These were namely: J48, LMT, Random Forest, Random Tree, REP Tree and Decision Stump. First, in order to prepare train and test data, 10-fold cross validation was performed on the data set. In this way, each train data included 99 data, and each test data included 11 data. Then, based on the training model, testing was performed to obtain the accuracy and errors of each method. Each test on the classification method was recorded based on the value of accuracy, MAE and RMSE. Once the accuracy and error value for all the tested methods were recorded, the comparison on each of the methods was implemented. The results of each method are presented in Tables 4 to 9. All the experiments were performed in the WEKA environment. WEKA is known as a collection of machine learning algorithms which can implement several processing tasks such as classification [8].

Table 4 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 5 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 6 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 7 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 8 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 9 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 4 shows the average accuracy of each method. 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.

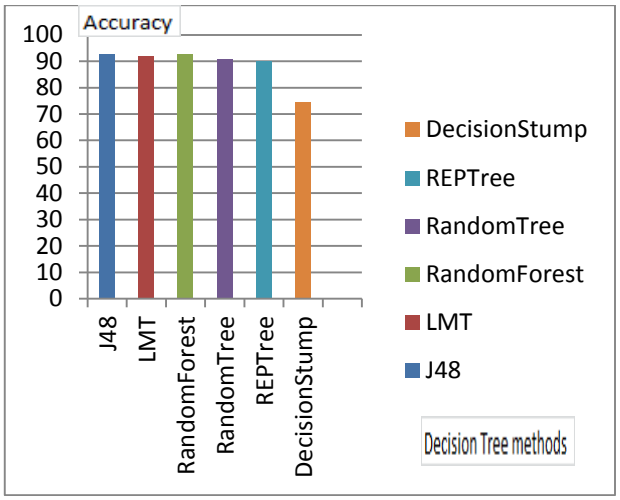


Figure 4 Average accuracy of 6 methods of decision tree

To choose the best method, if the values of accuracy are same, then the 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. The value of MAE and RMSE for J48 and Random Forest is measured in Table 10.

Table 10 Value of MAE, RMSE for J48 and random forest

Number	J48		Random Forest	
	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

Table 5, Table 6 and also Figure 4 depict that, although the accuracy value of J48 and Random Forest are same, in terms of the average value of MAE shown in Table 10, Random Forest method produces the smallest error (Figure 5). So, in this case, it can be concluded that, in the decision tree classification method, the most accurate method, Random Forest, which can be applied in many input variables without any variable elimination, produces the highest accuracy with the smallest average value of MAE, which is ultimately the best method.

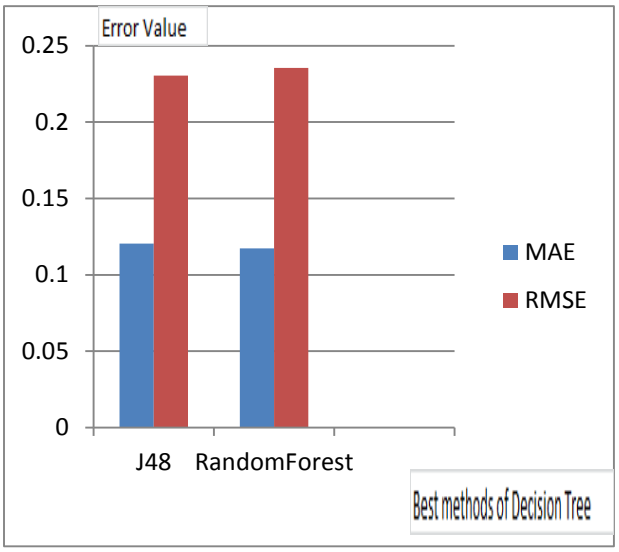


Figure 5 Values of MAE and RMSE for J48 and random forest

7.0 CONCLUSION

Here, we have attempted to classify researchers as "Expert" and "Novice" based on cognitive style factors in order to obtain the best possible answers. For this purpose, a questionnaire was prepared according to cognitive style variables of respective researchers. In addition, the domain of this research is based on an academic environment. An integral point of this study was to classify the researchers based on Decision Tree techniques and finally select the best method of Decision Tree according to the highest accuracy of each method. This would then assist the researchers to obtain the best feedback based on their requirements in digital libraries. In conclusion, the results of 6 methods of decision tree are presented in order to discover the best method of decision tree. In this case, the researchers are classified as expert or novice according to their cognitive styles. Based on the research, most of the expert researchers have better personal or internal knowledge in the domain of their problem and also their information seeking behavior is clearer in comparison with novice researchers. Moreover, the importance of their information seeking activities is recognized clearly and also they are more certain in their research problem As a result of this research; web developers can use the Random Forest technique in order to classify researchers and thereby assist them to obtain the best possible feedback according to their needs in digital libraries.

References

[1] Jantan, H., A. R. Hamdan, and Z. A. Othman. 2009. Classification for Talent Management Using Decision Tree Induction Techniques. *Conference on Data Mining and Optimization*. 15–20.
[2] Ranjan, J. 2008. Data Mining Techniques for Better Decisions in Human Resource Management Systems. *International Journal of Business Information Systems*. 3(5): 464–481.
[3] Chien, C. F., and L. F. Chen. 2008. Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-technology Industry. *Expert Systems and Applications*. 34: 280–290.
[4] Han, J., and M. Kamber. 2006. *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann.
[5] Jayasimman, L., and E. George Dharma Prakash Raj. 2012. Classification Accuracy in Cognitive Load for Users Preference in Web

Based Learning. *International Journal of Computer Applications.* 54(16).
[6] Nithyasri, B., K.Nandhini, and E. Chandra. 2010. Classification Techniques in Education Domain. *International Journal on Computer Science and Engineering.* 02(05): 1679–1684.
[7] Worrell, C. A., Sh. M. Brady, and J. W. Bala. 2012. Comparison of Data Classification Methods for Predictive Ranking of Banks Exposed to Risk Of Failure. *IEEE CIFEr* Paper Number: 61. Approved for Public Release: 12-0294.
[8] Mamani Barnaghi, P., V. Alizadeh Sahzabi, and A. Abu Bakar. 2012. A Comparative Study for Various Methods of Classification. *International Conference on Information and Computer Networks.* 27.
[9] Kharya, Sh. 2012. Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. *International Journal of Computer Science, Engineering and Information Technology*. 2(2): 55–66.
[10] Endo, A., T. Shibata, and H. Tanaka. 2008. Comparison of Seven Algorithms to Predict Breast Cancer Survival. *Biomedical Soft Computing and Human Sciences*. 13(2): 11–16.
[11] Othman, M. F. B., and T. Moh Shan Yau. 2006. Comparison of Different Classification Techniques Using WEKA for Breast Cancer. *International Conference on Biomedical Engineering (BIOMED*). 11–14 .
[12] Ford, N., T. D. Wilson, A. Foster, D. Ellis, and A. Spink. 2002. Information Seeking and Mediated Searching. Part 4. Cognitive Styles in Information Seeking. *Journal of the American Society for Information Science*. 53: 728–735.
[13] Davidson, D. 1977. The Effect of Individual Differences of Cognitive Style on Judgments of Document Relevance. *Journal of the American Society for Information Science*. 28(5): 274–84.
[14] Rholes, J. M., and J. B. Droessler. 1984. Online database searchers: Cognitive style. In National Online Meeting Proceedings, New York. 305–311.
[15] Johnson, K. A., and M. D. White. 1982. The Cognitive Style of Reference Librarians.RQ.21. (3): 239–246.
[16] Johnson, K., and M. White. 1981. Individuality in Learning. The Field Dependence/Field Independence of Information Professional Students. Library Research. 3(4): 355–369.
[17] Montgomery, P. 1991. Cognitive Style and the Level of Cooperation Between the Library Media Specialist and Classroom Teacher. *Integration The Vlsi Journal*. 16(3): 185–191.
[18] Huang, C. 1998. The Relationships of Cognitive Styles and Image Matching. *Bulletin of Library and Information Science*. 27: 55–71.
[19] Crossland, M. D., R. T. Herschel, W. C. Perkins, and J. N. Scudder. 2000. The Impact of Task and Cognitive Style on Decision-making Effectiveness Using a Geographic Information System. *Journal of End User Computing*. 2(1): 14–23.
[20] Palmquist, R. A. 2001. Cognitive Style and Users' Metaphors for the Web: An Exploratory Study. *Journal of Academic Librarianship*. 27(1): 24–32.
[21] Ford, N., and R. Ford. 1992. Learning Strategies in an Ideal Computer Assisted Learning Environment. *British Journal of Educational Technology*. 23: 195–211.
[22] Ellis, D., N. Ford, and F. Wood. 1992. Hypertext and learning styles. Final Report of a Project Funded by the Learning Technology Unit. Sheffield: Employment Department.
[23] Chou, C., and H. Lin. 1998. The Effect of Navigation Map Types and Cognitive Styles on Learners' Performance in a Computer Networked Hypertext Learning System. *Journal of Educational Multimedia and Hypermedia*. 7(2/3): 151–176.
[24] Wang, P., W. B. Hawk, and C. Tenopir. 2000. Users' Interaction with World Wide Web Resources: An Exploratory Study Using a Holistic Approach. *Information Processing and Management*. 36: 229–251.
[25] Palmquist, R. A., and K. S. Kim. 2000. Cognitive Style and On Line Database Search Experience as Predictors Of Web Search Performance. *Journal of the American Society for Information Science*. 51(6): 558–566.
[26] Abdelhalim, A., and I. Traore. 2009. A New Method for Learning Decision Trees from Rules. In the Eighth International Conference on Machine Learning and Applications (ICMLA'09).
[27] Caruana, R., N. Karampatziakis, and A. Yessenalina. 2008. An Empirical Evaluation of Supervised Learning In High Dimensions. Proceedings of the 25th International Conference on Machine Learning (ICML).
[28] Breiman, L. 2001. Random Forests. *Machine Learning*. 45(1): 5–32. doi:10.1023/A:1010933404324).
[29] Ross Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. San Francisco, C. A. USA: Morgan Kaufmann.
[30] Landwehr, N., M. Hall, and E. Frank. 2003. Logistic Model Trees. *European Conference on Machine Learning*. Springer-Verlag. 241–252.
[31] Witten, I. H, and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA, USA: Morgan Kaufmann.
[32] Wayne, I., and P. Langley. 1992. Induction of One-Level Decision Trees, in ML92. Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, San Francisco, CA: Morgan Kaufmann : 233–240.
[33] Witten, I. H., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA, USA: Morgan Kaufmann.
[34] Larose, D. 2005. *An Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons.
[35] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
[36] Ford, N., F. Wood, and C. Walsh. 1994. Cognitive Styles and Searching. *On-line & CDROM Review.* 18(2): 79–86.
[37] Spink, A., T. D. Wilson, N. Ford, A. Foster, and D. Ellis. 2002. Information-Seeking and Mediated Searching. Part 1. Theoretical Framework and Research Design. *Journal of the American Society for Information Science*. 53(9): 695–703.
[38] Wilson, T. D., N. Ford, D. Ellis , A. Foster, and A. Spink. 2002. Information Seeking and Mediated Searching: Part 2. Uncertainty and its Correlates. *Journal of the American Society for Information Science and Technology*. 53(9): 704–715.