Application of the Sino US Financial Technology in Banks

Rui Li, M.Sc
Rui Li, Lecturer, School of Banking and Finance, University of International Business and Economics, Beijing, 100029, China.
Correspondence LecLi;ybyvcc@163.com

Objectives: With the rise and development of Internet finance, the application of Sino US financial technology in the banking field is becoming more and more widely. **Methods:** In this study, for the data collection of bank customer deposits, data mining and decision tree analysis algorithms were used to classify bank customers.**Results:**The classification accuracy of the traditional algorithm was low, so the optimization algorithm Adaboost and the random forest improvement algorithm were proposed in this paper. The simulation effects of its application in data combination show that the classification effect of the optimization algorithm is obviously better than the traditional classification algorithm. **Conclusion:** The results of this study can help banks gain customers and reduce expenditures.

Keywords:data mining; decision tree algorithm; Sino US financial science and technology
Tob Regul Sci.™ 2021;7(5-2):4366-4374
DOI: doi.org/10.18001/TRS.7.5.2.2

Internet financial technology is now developing very rapidly, and the mainland of China can directly feel the impact of the Internet financial technology [1]. The big data in today's society is black gold. If an enterprise has the connotation of big data, then it is equivalent to the valuable customer information resources and data mining. The explosive growth of big data has brought new opportunities for development of various industries. Sino US financial technology is widely used in the banking field and it can be used to avoid money laundering and to design IC card and consumption platform and so on [2]. On the technical level, traditional data analysis and tools are unable to deal with BT-level financial data, so bank financial big data hidden behind the information is not easy to be tapped. New tools and algorithms need to be designed to deal with such problems[3]. Internet financial technology can build data mining platforms and resources, but it is unable to achieve the data mining effect without scientific algorithms [4]. In data mining, a new algorithm of decision tree was proposed to help banks develop data and excavate new technologies.

The development of the computer industry, the financial industry and the Internet industry has been toward the road of integration, so the relationship between the development of Internet technology and financial theory is getting closer[5]. The computer algorithm can support data mining technology, so bankers can get more high-quality customers and get important information in the existing customers. In fact, financial professionals in the analysis of market and customer big data process have found that there is a big flaw in data and customer information [6]. If the financial products only consider the single market share of the financial products in the process of promotion without analysis of the economic situation of the areas where the financial products are put into production, then it is unable to predict the future development prospects of financial products in the market [7]. Therefore, the introduction of Sino US financial technology and data mining technology and the corresponding algorithms can

dig customer information and data deeply, so as to obtain useful information.

In order to solve the problem of efficiency and accuracy of data mining, researchers have spent much effort on computer algorithms. A typical case is the analysis of the game tree analysis algorithm and the corresponding theoretical basis [8]. The emergence of the first decision tree analysis algorithm takes ID3 algorithm as a typical representative case. The basic data processing principle of maximizing information gain was used to select feature variables, and the decision tree was obtained by computer iterative algorithm [9]. The proposed C4.5 algorithm can maximize the information gain and achieve the set of feature sets, thus solving the problem that the feature set of ID3 algorithm is more centralized. Optimization algorithm is the research goal of many data mining algorithms. Lifting algorithm and update algorithm will lead to the increase of the workload of the algorithm, but the Random Forest algorithm has effectively solved this problem [10]. Random Forest algorithm was used in this paper to analyze and study the decision tree of bank customer data.

METHODS

At this stage, in various industries and data mining related applications, data classification and data regression analysis are the concerns of many researchers. If a bank needs to know whether customers will buy its latest bank financial products, then it needs to predict the quarterly sales of the sales department. The classification algorithm of decision tree can realize the classification rules which are easy to understand, improve the efficiency of computer, simplify the algorithm, and give the newest and most accurate data basis for the decision-maker in the fastest time. Generally, the application steps of decision tree include the generation of decision tree and the pruning process of the tree. Many researchers mainly use ID3 algorithm and C4.5 algorithm, as well as the improved CART optimization algorithm. Firstly, the basic principles of data mining algorithms were introduced based on the feature vector model and the corresponding scientific theories.

In decision tree classification, it is necessary to train data and algorithms. Before data analysis and classification results coming out, researchers are unable to know which variables have the promoting function for categorical data, so the classification effect of variables on data is random. For scientific research, the characteristics of random variables need to be extracted, and the ability characteristics which have a positive effect on data training need to be selected, so as to maximize the classification effect of the algorithm. From the point of view of the model application of the algorithm, the basic principle of the extraction of the classification features of decision tree algorithm is the maximization of information gain.

Information gain is called the segmentation variable scale of data features in decision tree classification algorithm. First, the set of data D was given, and then the expression of entropy of data set is:

$$Ent(D) = -\sum_{y \in Y} P(y \mid D) l0gP(y \mid D) \quad (1)$$

If the subset of training data set D in training process is very small, then the calculation entropy of the data set will gradually become smaller and smaller. The process of decreasing entropy is called information gain, and the expression of information gain is:

$$G(D : D_1, ..., D_k) = Ent(D) - \sum_{i=1}^{k} \frac{|D_k|}{|D|} Ent(D_k) \quad (2)$$

The feature set of the data set D is A, and it was divided into the class Ck of category number K, and $|C_k|$ represents the number of samples of class Ck.

The empirical entropy of the data set D was calculated as follows:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \quad (3)$$

The computing expression of subsequent empirical conditional entropy of characteristic set D is:

$$H(D \mid A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) \quad (4)$$

The information gain can be calculated by the difference between the formula 3 and formula 4:

$$g(D,A)=H(D)-H(D\mid A) \quad (5)$$

There is a defect in the process of solving information gain, because it will be concerned by the model in the process of feature solving, which may cause the value of some variables to increase. When the principle of information gain maximization is in choosing process, the possibility of variable selection will increase. If the number of samples is the same as the characteristic variable, then the two will be divided into the same node, and the variable will be the most optimal eigenvalue. At this point, variables will lose their functions in the actual decision tree classification process. If we need to deal with two-dimensional classification problems, there will be a unique ID value in the instance. The classification and regrouping of data for the ID of the feature vector will result in a very large decision tree. Accurate classification of data features on ID has no effect on the unknown principle of the model. The criterion of information gain has such an analysis defect that it needs to be solved by the gain ratio of characteristic information. Calculation expression is:

$$P(D:D_1,...,D_k)=G(D:D_1,...,D_k)\left(-\sum_{i=1}^{k}\frac{|D_k|}{|D|}\log\frac{|D_k|}{|D|}\right)^{-1} \quad (6)$$

The decision tree does not need pruning in the early operation, and the branch structure of the decision tree can grow sufficiently and satisfy the condition of the end of the algorithm. However, with the increase of computation data, if the branch of the decision tree is not sheared, then it may lose some important data basic information and produce excessive branch structures, and the data will appear rough problems. Therefore, in order to obtain the best decision tree model, the general algorithm needs the pruning process of decision tree in theory.

In theory, the expression of pruning complexity measure of decision tree is:

$$R_a(T)=R(T)+\alpha|T| \quad (7)$$

In the formula, $R(T)$ represents the cost of training data for decision tree classification; $|T|$ represents the number of branch leaf nodes of a decision tree; α represents the influence factor of applying penalty function to each node.

The meaning of the optimal decision tree is that the process of data mining can only be pruned to the last decision tree according to the minimum classification rules according to the test data. The error of decision tree classification calculation is affected by the individual differences of sampling, so the uncertainty calculation of the model is mainly about the optimal classification of decision tree pruning process. It is a suitable method to analyze the training set by using error curve analysis. The classica l 1-SE calculation criterion was adopted to add a tolerance band to the optimal model, and it became the representation model of the complexity of the model. The classification ability of decision tree was described, and the generalization ability of the model was improved by improving the classification effect.

With the rapid development of Internet banking in China, many banks and financial institutions are launching more categories of financial products constantly, but there is no reasonable financial management technology to achieve the mass financial products management. At the same time, the market prices can't be accurately controlled. Faced with many categories of financial products, issues as that customers prefer what kind of products and how can financial structure withdraw from the market and accept high financial products require management department to make reasonable planning. Next, the algorithm of decision tree analysis will be adopted, and the application of financial technology in the bank's term deposit will be excavated and calculated. According to the training and calculation of data mining, whether customers will choose to buy the bank's time deposits and financial products was analyzed and predicted. Finally, according to the classification speed of the decision tree, the fitting degree of the decision tree algorithm was analyzed.

The combination of data needs to be represented by the box line graph, and it can clearly reflect the distribution of different types of

data sets. In this paper, the parallel distribution of box charts was utilized to compare and analyze data. Which data may affect customers' choice of bank deposit financing products was studied. In this paper, the plot line abscissa Y represents whether to choose the bank's term deposit, and it can well reflect whether there is difference of dependent variable in each variable. Figure 1 is the distribution diagram of the work category. 16 different dependent variables were added to the Bank database, so the number of variables is overmuch from the computational point of view. Therefore, in the process of analysis, we only need to analyze the structure difference of customers choosing time deposits.

Figure 1
The Distribution of Work Types

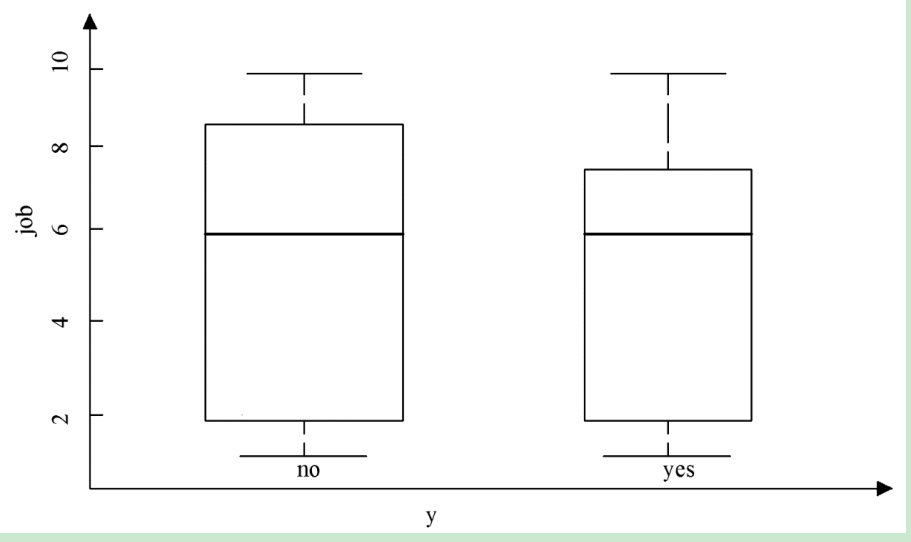


Figure 2 is a state distribution map of continuous time distribution. It can be seen from the figure 2 that in the process of purchasing regular deposits by customers, the status of time deposits is significantly higher than the stage before buying financial products. From the results, it can be considered that in the marketing and financial products, if there are more efficient and close contacts with customers, then investors and customers will be more inclined to buy regular deposit financing products.

Figure 2
State Distribution Diagram of Continuous Time Distribution

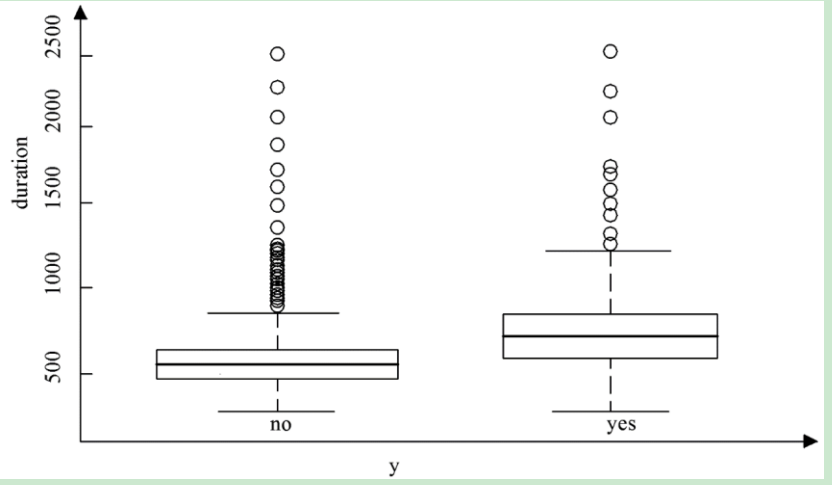
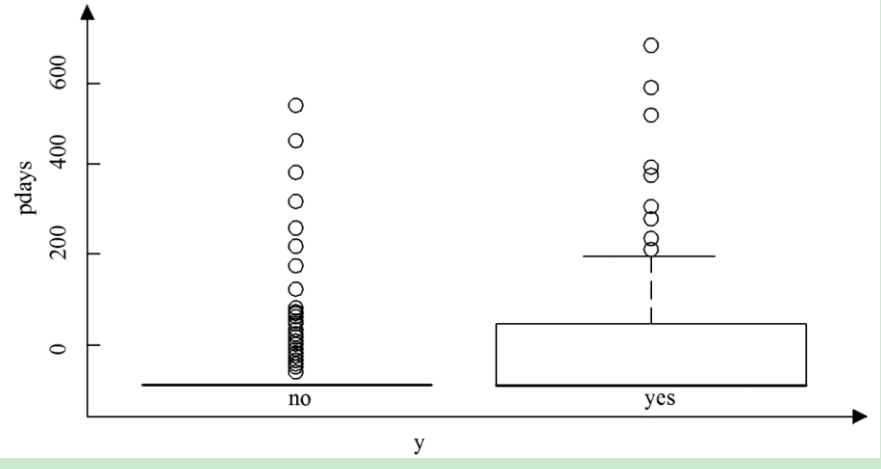


Figure 3 is the date distribution chart of the last contact with the bank. As can be seen from figure 3, the greater the value of the variable is, the more customers and investors will prefer to buy the bank's fixed deposit financial products. Then we can know that variables have big differences for customers to buy time deposits. At the same time, it is also an important analysis variable value of data mining decision tree model.

Figure 3
The Distribution Diagram of the Nearest Contact Distance Today



CART algorithm combines the overall advantages of ID3 and C4.5 algorithm. CART algorithm can deal with continuous attribute problems, and the classification efficiency of decision tree is very high, so the later pruning process can be more convenient. In R software, the classification tree type of the algorithm was constructed and it was precisely pruned. Then, the CART algorithm was used to build the structure classification model in decision tree.

RESULTS

The results of the training data are shown in table 1. The ID data are 1~10, and the X range is 0~9, with different Y values.

Table1
Training Data Set Used in This Paper

ID	x	y
1	0	1
2	1	1
3	2	1
4	3	-1
5	4	-1
6	5	-1
7	6	1
8	7	1
9	8	1
10	9	-1

The distribution of initialized data was calculated as follows:

$$D_1 = (w_{11}, w_{12}, ..., w_{110}), w_{1i} = 0.1, i = 1, 2, ..., 10 \quad (8)$$

In the weight distribution of the data, for the distribution result D1, the limit value V was set to 2.5. The expression of the classifier is:

$$G_1(x) = \begin{cases} 1, x < 2.5 \\ -1, x > 2.5 \end{cases} \quad (9)$$

The calculation results of classification error in data training process were given by classifier calculation:

$$e_1 = P(G_1(x_i) \neq y_i) = 0.3 \quad (10)$$

By calculation, the result of coefficient $G_1(x)$ is:

$$\alpha_1 = \frac{1}{2}\log\frac{1-e_1}{e_1} = 0.42 \quad (11)$$

Through the combination of three update data and training, the weight distribution calculation results are as follows;

$$D_4 = (0.125, 0.125, 0.125, 0.102, 0.102, 0.102, 0.065, 0.065, 0.125) \quad (12)$$

Well, the following formula was obtained;

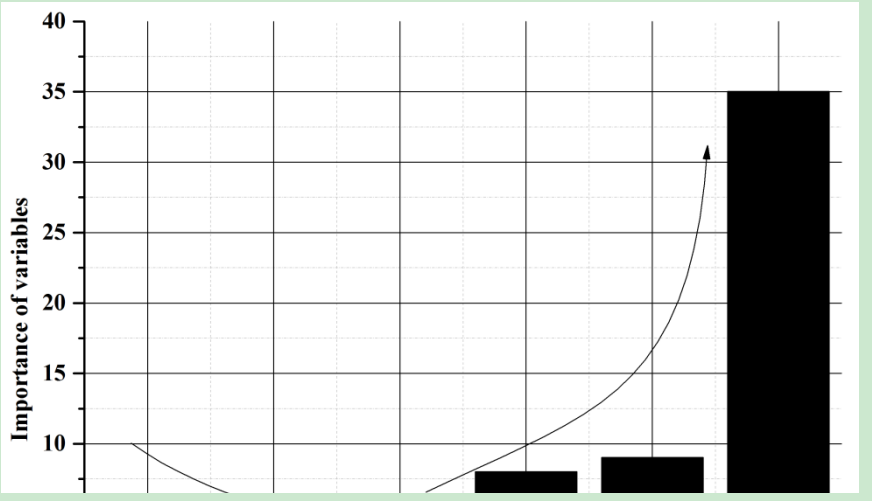
$$f_3(x) = 0.423G_1(x) + 0.649G_2(x) + 0.751G_3(x) \quad (13)$$

The results of the classifier are as follows:

$$G(x) = \text{si}\,gn[f_3(x)] = sign[0.423G_1(x) + 0.649G_2(x) + 0.751G_3(x)] \quad (14)$$

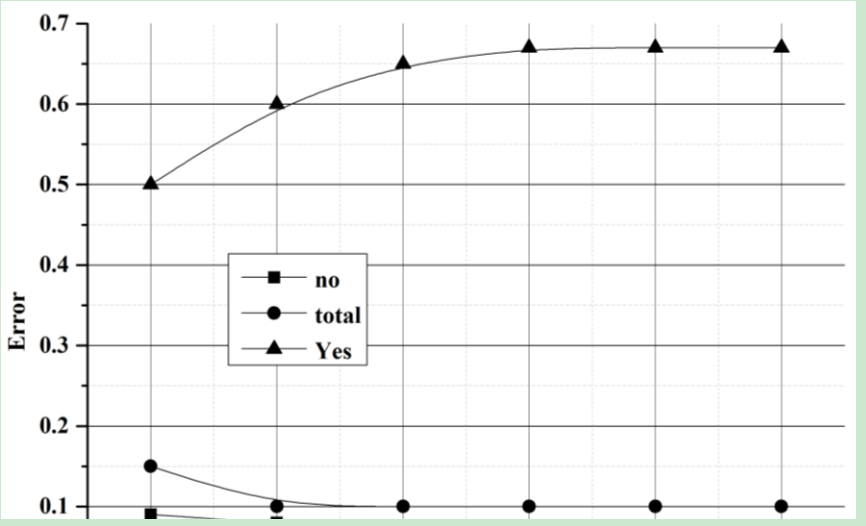
The decision tree algorithm will often encounter non balance data set classification efficiency problems in dealing with practical problems, and sometimes the classification efficiency is not high. At this point, the decision tree algorithm will have an over fitting problem. The *AdaBoost* algorithm can adjust the weights of data in the process of updating data step by step. If the classification effect is not ideal and the branch weight of decision tree becomes smaller, then the classifier with smaller classification error will increase the weight, and the final classification results will play a leading role. In this paper, a non-equilibrium set model was proposed to improve the accuracy of classification and update the accuracy of the algorithm under the condition of data. Figure 4 shows the distribution of the importance of dependent variables in the model construction of the algorithm established in this paper.

Figure 4
Model of the Algorithm Established in This Paper Constructs the Distribution of Dependent Variables



The *AdaBoost* algorithm has some shortcomings in the actual application of the error rate, so in practical application, this paper hopes to add random forest measurement method. By removing the correlation characteristics of the classification tree in the forest, a forest structure was established by means of calculation and training samples. In the process of forest construction, the classification of each decimal fraction is not allowed to split most of the classification trees. The complexity of the *AdaBoost* algorithm is relatively large, and it is more complex than the classical algorithm of the decision tree. The computer operating efficiency is relatively low, so the complexity of the algorithm needs to be further weakened. Forest random algorithm can greatly reduce the amount of calculation, so forest random algorithm was applied in Bank data set, and the prediction and evaluation model of data was obtained. Figure 5 is the quantitative relation graph between model error and decision tree. It can be seen from figure 5 that when the number of decision trees is more than 200, the error analysis results of the model can be stabilized. Therefore, in this paper, the number of decision trees in the analysis model was set to 200.

Figure 5
Quantitative Relation Diagram Between Model Error and Decision Tree



According to the characteristics of three different algorithms of CART, *AdaBoost* and random forest, the error results were predicted, as shown in table two. According to the calculation model of three kinds of decision tree classification, as can be seen from the collection of test data, in terms of the balanced state of data, the performance of the classical algorithm is poor because it is unable to correct the unstable and uneven data sets in time. The accuracy of the proposed algorithm is higher than that of the classical decision tree model, but in the process of updating the data classifier to calculate the weights, it will calculate data repeatedly, so the complexity of the algorithm and the amount of computation of the computer will be greatly improved. Random forest algorithm can repair unstable data and improve the accuracy rate. At the same time, it will not affect the computational efficiency of the computer.

Table2
CART, Adaboost, Random Forest Algorithm in Three Different Prediction Error Results in Data _ Test Platform

Classification model	CART	AdaBoost	Random forest
Total prediction error	11.4%	10.62%	7.88%
Yes prediction error“no”	3.30%	3.69%	1.80%
Prediction error forno“yes”	75%	64.80%	55.50%

The author believes that the randomness of the experiment is relatively large, so the experiment at one time can't fully explain the problem. Therefore, by repeating the test data collection, data mining algorithms in the model calculation was carried out, and the error of the calculated result of the same index was no more than 7.8%. The random forest algorithm can guarantee the stability of data calculation from the error degree of the algorithm. The optimized decision tree algorithm has better classification effect in the *Bank* data set, and the optimized decision tree algorithm can well fit the *Bank* data set.

DISCUSSION

With the continuous development of the Internet in the financial sector, operation and management of bank financial products are based on the financial and technical supports. Sino US financial technology has been widely used in the management of bank financial products. In terms of whether the financial products are liked by customers and how to design financial products with purposes, the method of data mining algorithm was proposed in this paper, and the decision tree correlation between the attributes of financial products and the attributes of customer demand was calculated. The traditional decision tree algorithm may appear over fitting or low classification accuracy in the process of data classification, but the optimized decision tree algorithm can improve the accuracy of the prediction algorithm, and can also maintain the stability of the algorithm. When the classical algorithm is used to deal with imbalanced data, there is an unstable rate of accuracy, so the improved decision tree algorithm can calculate the weakening of the classifier and avoid the over fitting problem. The optimized decision tree analysis model can improve the accuracy of the algorithm, and can be well fitted with the *Bank* data set. On the *Bank* platform, the three algorithms as CART, *AdaBoost* and random forest were compared and analyzed. It can be said that the random forest algorithm can maintain stable computing efficiency and improve the accuracy of calculation.

Human Subjects Approval Statement

This paper did not include human subjects.

Conflict of Interest Disclosure Statement

None declared.

References

1. Acheampong N K. The Effects of Foreign Bank Entry on Financial Performance of Domestic-Owned Banks in Ghana. *Social Science Electronic Publishing,* 2013, 7:93-104.
2. Akka, Akin L, Evren, et al. Application of Decision Tree Algorithm for classification and identification of natural minerals using SEM-EDS. *Computers & Geoscie nces,* 2015, 80(C):38-48.
3. Barbosa R M, Nacano L R, Freitas R, et al. The use of decision trees and naive Bayes algorithms and trace element patterns for controlling the authenticity of free-range-pastured hens' eggs.. *Journal of Food Science,* 2014, 79(9):C1672.
4. Chen L, Wang Y, Zhang F. Research and application of dynamic rule extraction algorithm based on rough set and decision tree. *Computer Knowledge & Technology,* 2015, 16.
5. Jiang F, Sui Y, Cao C. An incremental decision tree algorithm based on rough sets and its application in intrusion detection. *Artificial Intelligence Review,* 2013, 40(4):517-530.
6. Khanbabaei M, Alborzi M. The Use of Genetic Algorithm, Clustering and Feature Selection Techniques in Construction of Decision Tree Models for Credit Scoring. *International Journal of Managing Information Technology,* 2013, 5(4):13-32.
7. Kubikkomar A, Kubera E, Piotrowskaweryszko K, et al. Application of decision tree algorithms for discriminating among woody plant taxa based on the pollen season characteristics. *Archives of Biological Sciences,* 2015(00):89-89.
8. Kumar A R S, Goyal M K, Ojha C S P, et al. Application of ANN, Fuzzy Logic and Decision Tree Algorithms for the Development of Reservoir Operating Rules. *Water Resources Management,* 2013, 27(3):911-925.
9. Soleimaniangharehchopogh F, Mohammadi P, Hakimi P. Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study. *International Journal of Computer Applications,* 2013, 52(52):21-26.
10. Zheng G, Yu W. Financial Conditions Index's Construction and its Application on Financial Monitoring and Economic Forecasting ☆. *Procedia Computer Science,* 2014, 31:32-39.