RANDOM FOREST ALGORITHM OPTIMIZATION OF ENTERPRISE FINANCIAL INFORMATION MANAGEMENT SYSTEM

X.H. LIU †, E.X. WANG‡ and Y.Q. ZHENG†‡

† *School of management, Xiamen University, Xiamen, China.*
‡ *Institute for Financial and Accounting Studies, Xiamen University, Xiamen, China.*
†‡ *Department of architecture and planning, Shenyang Urban Construction University, Shenyang, China.*
kennedyVKvg@yahoo.com

***Abstract*— The optimization of random forest algorithms for enterprise financial information management systems is studied in this paper. A random forest algorithm was proposed to improve the data processing capabilities of the financial system. This paper proposes a random forest model on the premise of referring to the latest results of machine learning. The algorithm was introduced into the real estate business financial management system in this paper. First, the samples are divided into training samples and test samples, and the direct prediction method and the two-step prediction method are applied. Mean SR and MAPE were used to compare the prediction accuracy of different algorithms and it was found that the direct prediction method is better. In the algorithm used in this paper, the random forest effect is the best. Then the linear regression, decision tree, neural network and random forest model fitting effects were compared and the best fitting degree of random forest was found.**

***Keywords*— random forest algorithm; financial information; system.**

I. INTRODUCTION

With the development of information technology and the intensification of market competition, business managers are paying more and more attention to fine management reforms based on "accurate decision-making, precise planning, precise control, and accurate assessment" (Jan *et al*., 2016). Refined management has brought about a drastic expansion of information (Chen *et al*., 2017a). Enterprise managers must effectively use this information to manage and control, so as to increase the competitiveness of enterprises (Lee *et al*., 2017). In order to meet the challenges, locate the differences and the reasons for the differences, the reconstruction of the enterprise information system is the key point (Kaczałek *et al*., 2016). The enterprise information system should integrate information from all aspects of the enterprise to meet the information needs of various levels such as enterprise groups, companies, departments, and posts (Shafique *et al*., 2017).

The enterprise information system referred to in this article is the system of information flow generated when each element of the company's production and operation flows, including the flow of information generated by the flow of people, capital, logistics, technology, etc. of the supply, storage, production, sales, and after-sales service (Chen *et al*., 2017b). This information can be classified as business information and financial information (Halabi *et al*., 2015). However, in the current informatization mode, the business information and financial information of most companies cannot be completely synchronized and matched (Tang *et al*., 2015). This makes it difficult for enterprises to truly realize the integration of accounting informatization and business management informatization and ultimately achieve the goal of high-value information management and decision-making (Boyko *et al*., 2016). The random forest algorithm optimization of enterprise financial information management system is studied in this paper, and a random forest algorithm is proposed to improve the data processing capability of the financial system (Zhu, 2017).

II. STATE OF THE ART

In recent years, the random forest approach has developed rapidly in many areas (Ma *et al.*, 2017). Nico Lai, on the basis of fully studying the random forest proposed by Breiman, proposed a new method based on quantile regression (Zhou *et al*., 2017). This method can display the conditional distribution information of the dependent variable. Sexton found that the best algorithm in Random Forest (RF) is Bootstrap when the data is large (Ming *et al*., 2016). Brence proposed the Booming algorithm through in-depth research and found that this algorithm has unparalleled superiority (Wen, 2016). It should be further promoted and applied. In addition, there are scholars using random forest models to conduct research in the field of biology (Qiu, 2016). In order to find out the variables that have an impact on the density of the bacteria on the beach, the random forest model was introduced by Parkhurst (Choi, *et al*., 2016). The data was subsequently further investigated by Smith. Alonso used the classification function of the random forest model to discriminate and analyze fish populations (Chang *et al*., 2016). In addition, in terms of ecology, Gislason applied Random Forest Model to study the land cover area and found that random forests were significantly faster than other combinatorial algorithms when performing sample training (Gurunathan, 2017). Jan compared and analyzed the eco-hydrological distribution model based on the random forest model and Logistic model. Through

empirical research, it was found that the prediction error based on the random forest model was less than the prediction error based on the Logistic model (Malhotra, 2015).

The application of the random forest model is also indispensable in the field of economic management: Bart uses a random forest model to predict customer churn in order to predict customer relationships. Coussement conducted in-depth research and introduced support vector machines and Logistic models. It is found that the prediction result of the random forest model is always optimal, followed by the support vector machine, and the traditional Logistic model performs the worst. Burez improved the random forest model and put forward a weighted random forest (Lundqvist *et al.*, 2018). He also proved that weighted random forest has certain advantages, but the choice of weights is a problem that cannot be solved easily. Buckinx comprehensively compared random forests, artificial neural networks, and multiple linear regression algorithms to predict customer loyalty. He found that random forests have the best prediction effect.

III. METHODOLOGY

A. Random Forest Algorithm

Random Forest is a Machine Learning Algorithm proposed by American scientist Leo Breiman, who combined his Bagging Algorithm (Breiman, 1996) and the stochastic subspace method proposed by Ho, and finally put forward a random forest in 2001. Random forest is a random combination of decision trees, so there are two major categories of applications based on this in practical applications. The first category is the classification forecast, and the second category is the regression forecast. The following is a brief introduction to random forests from these two aspects.

Classification problem: In fact, a random forest is a combination of decision trees. The final result of a single classification regression tree is the result of random forest. In order to improve the classification performance of the decision tree, Breiman combines Bagging and Randomization algorithms.

Assume that the number of samples in the original training set is N. The Bagging algorithm randomly uses the extracted N samples from the original training set to form a new training set. Then calculate the error 1 using out-of-pocket data - unselected data.

The reason why the Bagging method was chosen is mainly because of its two advantages:

(1) According to the bagging algorithm, out-of-pocket data can calculate generalization errors, correlations between classification trees, importance of input feature variables, and performance of each classification tree.

(2) The greatest advantage of the combination of the Bagging algorithm and the Randomization algorithm is that it can improve the accuracy of random forest classification.

Dietterich proved that Bagging and Randomization both have a good tolerance for noise. Therefore, the combination of these two methods will undoubtedly play a significant role in improving the ability of random forests to tolerate noise.

Random forests were trained in the k-round when constructing the classification model, and a classification model sequence $\{h_1(X), h_2(X)... h_k(X)\}$ was obtained. Then a multi-classification model system is generated based on this sequence. Finally, the voting method is used to make the decision. The formula for the decision is as follows:

$$H(\mathrm{x}) = \underset{Y}{\operatorname{argmax}} \sum_{\mathrm{i}=1}^{\mathrm{k}} I\left(h(x) = Y\right) \tag{1}$$

among them, H(x) represents the classification model of the entire combination, h i represents the classification model of a single decision tree, Y represents the target variable, and I(.) is an illustrative function. The above formula shows the process of getting the final classification result of the random forest model. After the classification model sequence {h1 (X), h 2 (X)... h k (X)} is obtained through training, the remaining margin functions can be further deduced accordingly:

$$mg(X,Y) = av_k I\left(h_k(x) = Y\right) - \max_{j \neq k} av_k I\left(h_k(X) = j\right) \tag{2}$$

The headroom function is mainly used to measure the extent that the correct number of classifications exceeds the number of incorrect classifications in the prediction result. Moreover, the accuracy of classification prediction is positively correlated with the value of margin. The generalization error can be written as:

$$PE^* = P_{X,Y} PE^* = P_{X,Y}\left(mg(X,Y) < 0\right) \tag{3}$$

When the decision tree classification model is sufficiently large, $h_k(X) = h(X, \Theta_k)$ obeys the strong law of numbers. This also explains why random forests can well prevent overfitting problems.

Regression problem: Random forests adopt the (Bootstrap) re-sampling technique when performing regression, and random sampling from the original sample constitutes the combined model. The model $\{h(X, \theta_k), k = 1, L, p\}$ selects numerical variables in the prediction to generate a multiple nonlinear regression random forest model. The mean value of these trees $\{h(X, \theta_k)\}$ for k will eventually form a model prediction, and the model must also satisfy a condition that random forests are formed from their respective independent training sets. The generalization error of the numerical prediction vector h(X) selected from the random vector X, Y can be regarded as $E_{X,Y}\left(Y - h(X)\right)$.

B. The Main Evaluation Algorithm

The first is the decision tree algorithm, and the decision tree is a classic prediction model. It mainly reflects the

mapping between object attributes and object values. The entire decision-making process resembles a tree. The tree is composed of multiple nodes, with the top representing the root node and further down the leaf node. Each node represents a classification attribute, and bifurcation represents the inspection result. The main use of this article is CHAID and CRT algorithms.

The Chi-Square Automatic Interaction Detection (CAID) algorithm is also known as the Chi-Square Automatic Interaction Detection algorithm. This method specifies the dependent variable as the root node, categorizes each independent variable, and calculates the categorized chi-squared value. If these variables are found to be significant after the calculation, it is possible to select the most significant categorical variable as a child node by comparing the size of the P values of these classifications.

The CHAID method is more appropriate when the predictor is a discrete variable. However, for continuous variables, the CHAID algorithm is not suitable for application. In this case, the CRT algorithm should be considered. The Classification and Regression Tree (CRT) algorithm uses a binary tree to recursively partition the entire prediction space into subsets. Different leaf nodes in the tree also correspond to different areas, and they are further determined based on the branch rules associated with each internal node. By moving from the top to the bottom, the leaf node where each prediction sample is located is determined accordingly. The CRT algorithm is more flexible. When the dependent variable is a categorical variable, the CRT is a classification tree. If the dependent variable is a continuous variable, the CRT is a regression tree.

This is followed by the K-Nearest Neighbor algorithm, which is also known as the nearest neighbor rule classification algorithm. The KNN algorithm can well determine which class the unknown belongs to, and it is widely used in the field of identifying unknown things. The algorithm is mainly based on the Euclidean theorem to determine the characteristics of unknown things and what kind of known things are the closest to the characteristics, so as to distinguish unknown things. The KNN algorithm is suitable for both classification and regression problems.

If the category of a sample in feature space needs to be judged, only one of the most neighboring k samples in the sample needs to be considered, and the sample is also considered to belong to this category. That is, the k nearest neighbors of the samples to be discriminated in the feature space are first found out, and then the average of the attributes of these neighbors is assigned to the sample, so that the attributes of the sample can be obtained.

C. Data Sources and Preprocessing

This article takes a real estate company as an example to test the algorithm. The algorithm of this paper is introduced into the company's financial information management system, which mainly performs statistical processing on sales data, and mainly selects second-hand housing transaction data of the company. For analysis, the data is processed first. First, the discrete variables are processed, and several evaluation indexes are set for the comparative evaluation of the models. The main evaluation indicators are the following three:

One is the average sales ratio (SR). For a certain transaction, the numerator of the sales ratio is an estimate generated by the model, and the denominator is the actual sales price. According to international standards (International Association of Assessing Officers, 2003) 95% confidence interval covers 0.9-1.1 times full moment. The Bootstrap sampling technique is used in this study because the average sales ratio does not obey the normal distribution.

The second is mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i - Y_i}{Y_i} \tag{4}$$

where Yi represents the observed value, and is the predicted value of the i-th observation. Mean absolute percentage error can reflect the accuracy of the model well.

For ease of analysis, sample data is divided into training data and test data. The training data contains 9357 observations, the test data contains 1039 observations, and the ratio of training data to test data is 9:1. In addition, the sample data also has the following characteristics, which also affect the operation of the model in different methods：A large number of missing values, most of the quantitative variables have strong asymmetric distribution, there are few types of data in categorical variables, Categorical variables have multiple levels, and the second-hand housing construction time has heterogeneity and heteroscedasticity.

IV. RESULT ANALYSIS AND DISCUSSION

In this section, random forests will be used for modeling and comparative analysis will be used to detect the prediction of random forest models. The algorithms used mainly include CHAID, CRT, Regression, MLP, RBF, Exhaustive CHAID, KNN (mean), KNN (median), Random Forest, SVM.

The above algorithm is implemented by R. The 95% confidence interval is set for all methods. For the entry method used in the linear regression model, the probability of entry is 0.05 and the probability of deletion is 0.1. For the three methods in the tree (CHAID, CRT, Exhaustive CHAID), the minimum child node is set to 50.

When the nearest elemental analysis method (KNN (mean), KNN (median)) is used, the standardized scale feature values are used and the software automatically selects feature values, among which the selection of the most adjacent element K values is between 1 and 200. In the neural network (MLP, RBF) algorithm, the data used to calculate the prediction error is automatically selected by the software. The minimum relative change

in training errors is 0.0001. The maximum number of units in the hidden layer in the automatic architecture selection is 50, and the smallest number of units is 1.

As for Random Forest, Bootstrap sampling technique is used. The OOB (outofbag) method is used in the error estimation of the classifier. In order to reduce the total error or target category error to be sufficiently low and stable, ntree selects 500 trees and the node size is 3. In support vector machines (SVMs), Sigmoid kernel functions are used and default values are used in other parameters. The accuracy of each algorithm's prediction is compared in Table 1-3 by the indicators SR and MAPE. The sample data is split into training data and test data, in which the training data is 9356 observations and the test data is 1039 observations.

Table 1. Comparison of prediction accuracy of direct assessment.

Method	Test Data		Training data	
	Mean SR	MAPE (%)	Mean SR	MAPE (%)
CHAID	1.10	10.7	1.07	7.3
CRT	1.17	17.2	1.09	9.7
Regression	1.05	5.4	1.05	4.9
MLP	1.09	9.2	1.07	5.2
RBF	1.26	26.9	1.08	12.0
Exhaustive CHAID	1.18	18.9	1.08	10.9
KNN (mean)	1.05	5.8	1.05	5.3
KNN (median)	1.04	4.3	1.03	3.0
Random Forest	1.00	0.2	1.01	0.6
SVM	1.10	10.3	1.01	1.3

Table 2. Comparison of prediction accuracy of the two-step method.

Method	Test Data		Training data	
	Mean SR	MAPE (%)	Mean SR	MAPE (%)
CHAID	1.10	10.3	1.08	8.5
CRT	1.12	12.7	1.10	10.5
Regression	1.05	12.9	1.05	6.5
MLP	1.09	12.8	1.07	7.5
RBF	1.26	16.3	1.08	7.5
Exhaustive CHAID	1.18	8.3	1.08	9.3
KNN (mean)	1.05	8.4	1.05	6.8
KNN (median)	1.02	5.6	1.03	4.8
Random Forest	1.00	2.4	1.01	3.7
SVM	1.10	13.1	1.02	2.3

Table 1 and Table 2 reflect the comparison of the indicators of the original data. Tables 3 and 4 show the comparison of the indexes after deleting the test data with the number 825-881.

Table 3. Comparison of prediction accuracy of direct assessment after deleting some test data.

Method	Test Data		Training data	
	Mean SR	MAPE (%)	Mean SR	MAPE (%)
CHAID	1.03	3.6	1.07	7.3
CRT	1.05	5.6	1.09	9.7
Regression	1.01	1.3	1.05	4.9
MLP	1.01	1.9	1.07	5.2
RBF	1.09	9.3	1.08	12.0
Exhaustive CHAID	1.05	5.8	1.08	10.9
KNN (mean)	1.01	2.0	1.05	5.3
KNN (median)	1.00	0.5	1.03	3.0
Random Forest	0.99	0.7 1.01 0.6	0.99	0.7 1.01 0.6
SVM	0.99	-0.03	1.01	1.3

Table 4. Comparison of prediction accuracy of two-step method after deleting part of test data.

Method	Test Data		Training data	
	Mean SR	MAPE (%)	Mean SR	MAPE (%)
CHAID	1.03	4.0	1.08	8.5
CRT	1.05	5.7	1.10	10.5
Regression	1.02	5.0	1.05	6.5
MLP	1.02	6.4	1.07	7.5
RBF	1.03	7.1	1.08	7.5
Exhaustive CHAID	1.03	3.7	1.08	9.3
KNN (mean)	1.01	4.9	1.05	6.8
KNN (median)	1.01	4.0	1.03	4.8
Random Forest	0.99	1.8	1.01	3.7
SVM	0.99	2.9	1.02	2.3

From Tables 1 to 4, the following conclusions can be drawn:

1. Mean SR for all training data is in the 0.9-1.1 range. However, it can be seen from Table 1 and Table 2 that the Mean SR obtained by applying the CRT, RBF, and Exhaustive CHAID algorithm exceeds 1.1. After deleting test data numbered 825-881, all test data is within this range.

2. For most algorithms, both the Mean SR and the MAPE (%) have increased. Therefore, using the direct algorithm is better than the two-step method.

3. Although a validity sample was used in the calculation to prevent overfitting, the best-performing neural network algorithm (MRF, RBF) in estimating house prices was not the best algorithm in this study. The main reason is that the observations in this paper have a large number of missing data, so it is difficult to prevent the phenomenon of overfitting.

4. In almost all tables, the random forest algorithm has the best performance. In addition to the table of prediction accuracy comparisons directly evaluated after deleting some of the test data, the MAPE (%) of the

random forest and the support vector machine is negative. It can be seen that random forest algorithms are superior to neural networks in data sets with a large number of missing values. In addition, with the increase in the number of combinatorial trees, the prediction error will always decrease, which also means that the random forest algorithm will not be overfitting.

In the above, the predictions of the ten algorithms have been compared by Mean SR and MAPE (%) indicators. Next, the model fitting conditions of the four major algorithms (Tree, Random Forest, Regression, Neural Networks) are compared by drawing. Since the previous section has confirmed that direct prediction is better than the two-step method, direct prediction is used next. The following figure shows the fitting of the four-in-one algorithm.

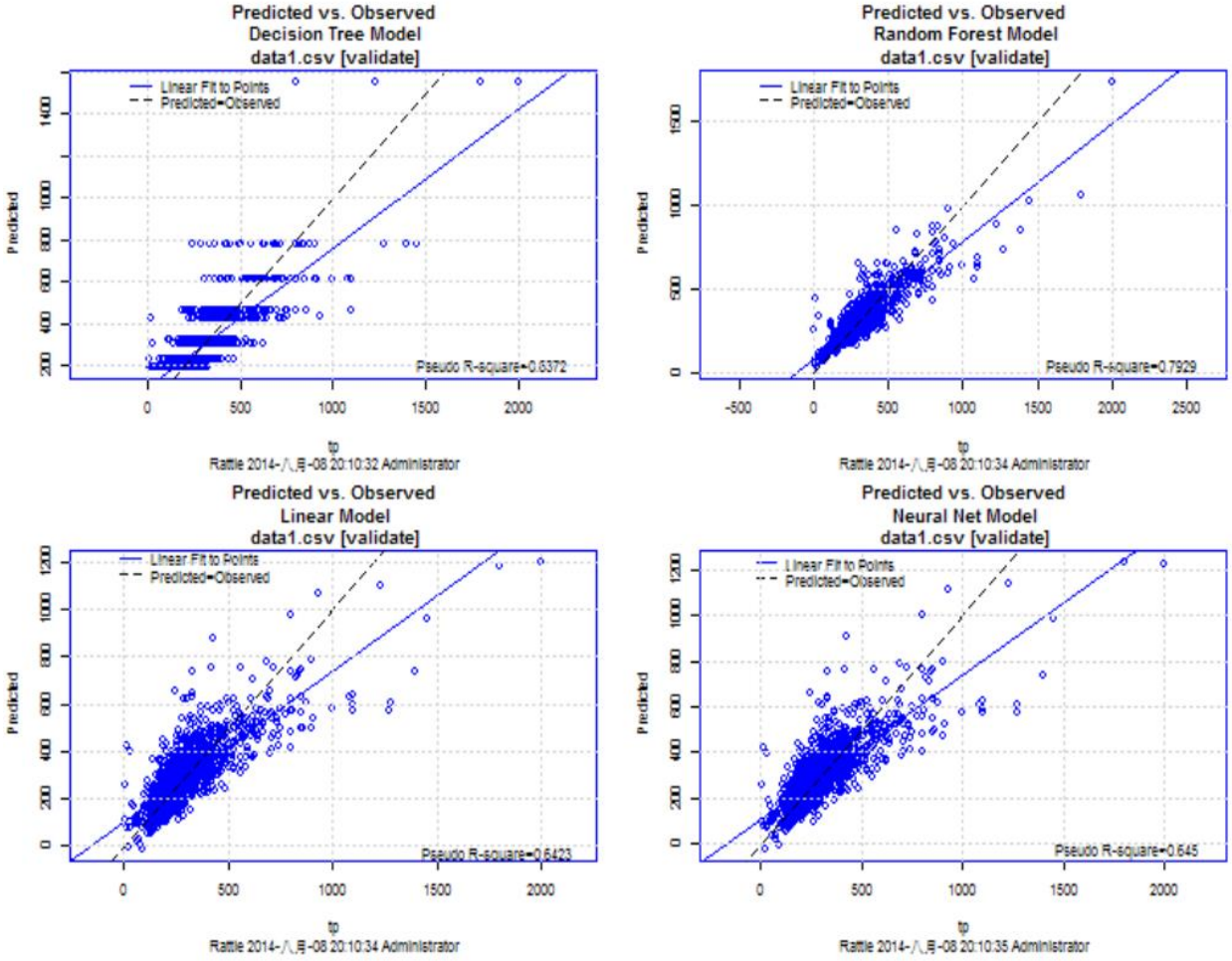

Figure 1. Model Fitting of Four Algorithms.

From Figure 1, in the effective prediction data, the fitting effect of the random forest is obviously better than the decision tree, linear regression and neural network.

V. CONCLUSIONS

This paper proposes a random forest model on the premise of referring to the latest results of machine learning. In this paper, the algorithm is introduced into the real estate enterprise financial management system. First, the samples are divided into training samples and test samples. The direct prediction method and the two-step prediction method are used. Mean SR and MAPE are used to compare the prediction accuracy of different algorithms. It is found that the direct prediction method is better. In the algorithm used in this paper, the random forest works best. Then by comparing the linear regression, decision tree, neural network and random forest model fitting results, it is found that the random forest has the best fitting degree. This paper empirically finds that the random forest model has obvious advantages in large-scale house price prediction. Especially when a certain variable has a large number of missing values or there are a large number of classification variables, the random forest model has good prediction effect and it is worthy of popularization and application.

REFERENCES

Boyko, K. and I. Derun, "Disclosure of non-financial information in corporate social reporting as a strategy for improving management effectiveness", *Journal of International Studies*, **9(3)**, 159-177 (2016).

Chang, S., T. Wang and H.J. Li, "A study of ITG mechanism in post-implementation phase of an ERP system", *PACIS Proceedings, Nº. 7*, 1-16 (2016).

Chen, J., K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng and K. Li, "A parallel random forest algorithm for big data in a spark cloud computing environment", *IEEE Transactions on Parallel and Distributed Systems*, **28(4)**, 919-933 (2017a).

Chen, Y., M. Luo, J. Peng, J. Wang, X. Zhou and S. Li, "Classification of land use in industrial and mining reclamation area based grid-search and random forest classifier", *Transactions of the Chinese Society of Agricultural Engineering*, **33(14)**, 250-257 (2017b).

Choi, Y., X. Ye, L. Zhao and A.C. Luo, "Optimizing enterprise risk management: a literature review and critical analysis of the work of Wu and Olson", *Annals of Operations Research*, **237(1-2)**, 281-300 (2016).

Gurunathan, T., "Risk Assessment of ERP Implementation Using Generalized Stochastic Petri Net", IIMK *Working Papers*, Nº. 239, 1-15 (2017).

Halabi, A.K. and B. Carroll, "Increasing the usefulness of farm financial information and management: A qualitative study from the accountant's perspective", *Qualitative Research in Organizations and Management: An International Journal*, **10(3)**, 227-242 (2015).

Kaczałek, B. and A. Borkowski, "Urban road detection in airbone laser scanning point cloud using random forest algorithm", *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, **41**, 255-259 (2016).

Lee, J., D. Park and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier", *KSII Transactions on Internet and Information Systems*, **11(10)**, 5132-5148 (2017).

Lundqvist, S.A. and A. Vilhelmsson, "Enterprise risk management and default risk: evidence from the banking industry", *Journal of Risk of Insurance*, **85(1)**, 125-157 (2018).

Ma, L. and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests", *BMC Bioinformatics*, **18(1)**, 169 (2017).

Malhotra, Y., "Toward integrated enterprise risk management, model risk management & cyber-finance risk management: bridging networks, systems and controls frameworks", *SSRN Electronic Journal*, (2015).

Ming, D., T. Zhou, M. Wang, and T. Tan, "Land cover classification using random forest with genetic algorithm-based parameter optimization", *Journal of Applied Remote Sensing*, **10(3)**, 035021 (2016).

Qiu, J., "Analysis on the risk and control strategy of ERP project implementation in large group enterprises", *Jiangsu Science & Technology Information*, **9**, 269-270 (2016).

Shafique, M.A. and E. Hato, "Classification of travel data with multiple sensor information using random forest", *Transportation Research Procedia*, **22**, 144-153 (2017).

Tang, G., Z. Luan and B. School, "Non-Financial information disclosure, management control capability and corporate performance", *Journal of Beijing Technology and Business University*, (2015).

Wen, Q. "Research on the internal control of enterprise modern accounting management under the ERP system platform", *Value Eng.*, **4(8)**, 15-17 (2016).

Zhou, T., D. Ming, R. Zhao, et al, "Land cover classification based on algorithm of parameter optimization random forests", *Science of Surveying and Mapping*, (2017).

Zhu, D., "Research on the impact of financial information management on financial internal control", *Heilongj. Sci.* (2017).

Received: December 15th 2017
Accepted: June 30th 2018
Recommended by Guest Editor
Juan Luis García Guirao