Employee Recruitment Method Based on Random Forest

Zhao-Yong An [1,a,*], Hai-Fu Shen [1,b] and Xue-Ting Wang [2,c]

[1] State Grid Shandong Electric Power Company, NO.150, JING ER ROAD, JINAN city ShanDong Province, China

[a] zhaoyong@sd.sgcc.com.cn, [b] shenhaifu@sd.sgcc.com.cn, [c] 15370326@qq.com

*corresponding author

Keywords: Employee Recruitment, Data Mining, Human Resource, Random Forest, Machine Learning

Abstract: Human resource management especially employee recruitment play a crucial role in development and success of enterprises. An employee recruitment method based on random forest has been proposed in this paper. After a survey of related work, data set has been visualized with the quantile grouping method, and random forest has been used to investigate the efficiency of employee recruitment. In the experiments, the employee recruitment of State Grid Shandong Electric Power Company in 2011-2015 was taken as a data source. The results shows that the employee recruitment method based on random forest performs better than the partial least squares discriminant method.

1. Introduction

With the further development of economic globalization and the intensification of global competition, effective human resource management has become the key to development and success of enterprises[1]. Thereinto, employee recruitment is a very important part. As a lot of enterprises are short of talent, there is an especial strategic significance to find appropriate staff through the recruitment of employees. It will help enterprises to achieve the core competitiveness. First of all, through the employee recruitment, enterprises can hire the staff meets the requirements, which ensures the enterprise survival and development; and then, through the employee recruitment, enterprises can be more popular and establish a good corporate image; at last, through the employee recruitment, enterprises can avoid excessive turnover rate, which enhances the enterprise cohesion. However, due to the particularity of employee recruitment, which people are various, there is a common problem that the employment standard is difficult to certify, which impacts the efficiency of employee recruitment. In order to solve this problem, this paper will design an employee recruitment method based on random forest. The employee recruitment of State Grid Shandong Electric Power Company in 2011-2015 will be taken as a data source, in order to investigate and analyse the employee recruitment method based on random forest.

2. Related Work

Random forest is a classifier constructed in a random way and contains multiple decision trees[2]. Its classification depends on the mode of classifications output by each decision tree. It has a great advantage over other algorithms in many current data sets. It can handle very high dimensional data without feature selection. The training speed is fast and the realization is relatively simple.

A random forest based diagnosis approach for rail fault inspection in railways has been proposed[3].After railway image acquisition, the feature extraction such as PCA[4], KPCA[5], SVD[6], HM[7] has been put into operation, and then random forest has been used to classify the healthy data and faulty data. With the train set according to the classification result of history data, the computer

can automatically make decision. It proves the efficiency of random forest algorithm. However, in this approach, the number of decision trees is only 10, and never changes, which means the more experiments with various decision trees should be conducted. Other-wise the value of this approach should be difficult to estimate.

A Random Forest regression model is used to predict popularity of articles from the Online News Popularity data set[8]. The performance of the Random Forest model is investigated and compared with other models. Impact of standardization, regularization, correlation, high bias/high variance and feature selection on the learning models are also studied. Results indicate that, the Random Forest approach predicts popular/unpopular articles with an accuracy of 88.8%.

A rainfall-induced landslide susceptibility assessment using random forest weights (RFWs) has achieved 79.71% accuracy, higher than the entropy weight (EW) of 63.77%[9]. Two experiments were conducted by respectively removing the most dominant and the weakest indexes to examine the rationality and feasibility of RFW; both precision validation and contrastive analysis indicated the assessment results of RFW to be reasonable and satisfactory.

3. Data Set & Data Visualization

In this paper the employee recruitment of State Grid Shandong Electric Power Company in 2011-2015 will be taken as a data source. It involves a total of about 300000 samples, each of which includes 56 indicators. These indicators are composed of classification variables and continuous ones. Classification variables include administrative division code, school code, recruitment mode, educational background, nationality, professional code and so on. Continuous variables include the number of candidates with different degree, the number of candidates in different schools, the number of candidates in different areas, the total number of candidates, expected salary, initial score, interview score, language score the number of rewards, the number of projects, the year-end assessment. The year-end assessment was taken as dependent variable, which actually reflected the recruitment effect. After simple analysis on the original data, the following problems were detected: Some indicators of employees' data were lost, which means some variables should be deleted or updated; the year-end assessment of some employees were incompleted because of job transfer or resignation; the data stability is insufficient and needs to be improved. After data cleaning, employees' year-end assessment has been adjusted to the average value. According to formula 1, the data has been divided into ten even parts, which helps to achieve the balance of data. Let variable X be the yardstick of classification, then the formula will be:

$$Y = -\sum_{i=1}^{10} i * I(quantile(X,(i-1)/10) < X < quantile(X,i/10) \tag{1}$$

4. Proposed Method

In this paper random forest has been adopted to predict employee performance. The process is as follows:

Step1: The entropy of the classification system is derived from the formula (2);

$$H(C) = -\sum_{i=1}^{n} P(C_i) \times \log_2 P(C_i) \tag{2}$$

The variable C is the category field. The variable P (C) is the probability of each category. The variable n is the total number of categories.

Step2: For each feature, the conditional entropy is calculated by formula (3);

$$H(C \mid X = t) = P(t)\sum_{i=1}^{n} P(C_i \mid t)\log_2 P(C_i \mid t) + P(\overline{t})\sum_{i=1}^{n} P(C_i \mid \overline{t})\log_2 P(C_i \mid \overline{t}) \tag{3}$$

The variable t is a certain feature. The variable represents the probability that the category is

when a feature t exists. The variable represents the probability that the category is when a feature t does not exist. The variable is the probability that t exists. The variable is the probability that t does not exist.

Step3: Compute the information gain which t brings to the system, as shown in formula (4).

$$IG(t) = H(C) - H(C \mid t) \tag{4}$$

The variable H (c) is the system original entropy. The variable H(C|t) is the system entropy after the feature t is fixed.

The random forest algorithm is implemented on the training set after the feature selection, and the employ-ee recruitment model is determined as follows:

Step4: For a given training set of employee data S with feature dimensions F, random forest parameters have been initialized as follows: the number of classification trees is G; the maximum depth of each tree is d; the number of features used by each node is f; and the termination conditions are the minimum number of samples on the nodes i.e. s, or the minimum information gain on the nodes, i.e. m.

Step5: The training set S (I) randomly selected from S with return, with the same size as S, will be taken as the root node, where the training starts. If the current node satisfies the termination condition, it becomes the leaf node. Its predictive output will be c(j),i.e. the major class in current data set.

Step6: Define p as the proportion of c(j) in current sample set and keep training the other nodes. If the cur-rent node fails to reach the termination condition, randomly no back select f-dimension features from F-dimension feature set. According to them find the best 1-dimension feature k and the threshold th.

Step7: If the kth dimension feature of one sample in the current node is less than the threshold th, then it will be the left child; if not, then will be the right child.

Step8: Keep training the other node. If all the nodes have been trained or labeled as leaf nodes, determine all the classified trees whether have been trained or not. If they are, then end; otherwise, the training set will be selected again.

5. Results

The number of variables and trees will affect the strength and correlation, which finally declines the testing ac-curacy.

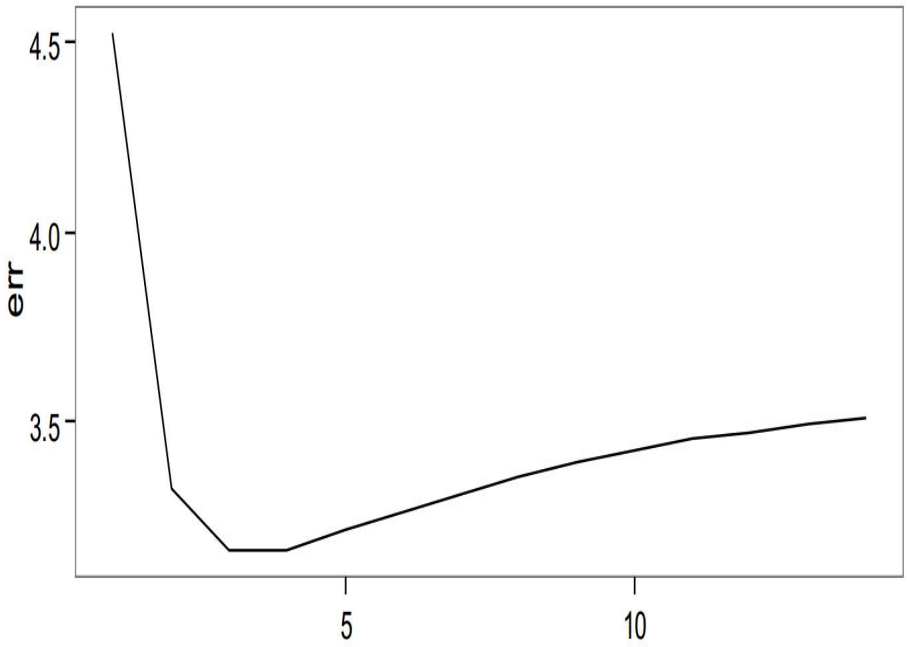

Fig. 1: error V.S. variables' number

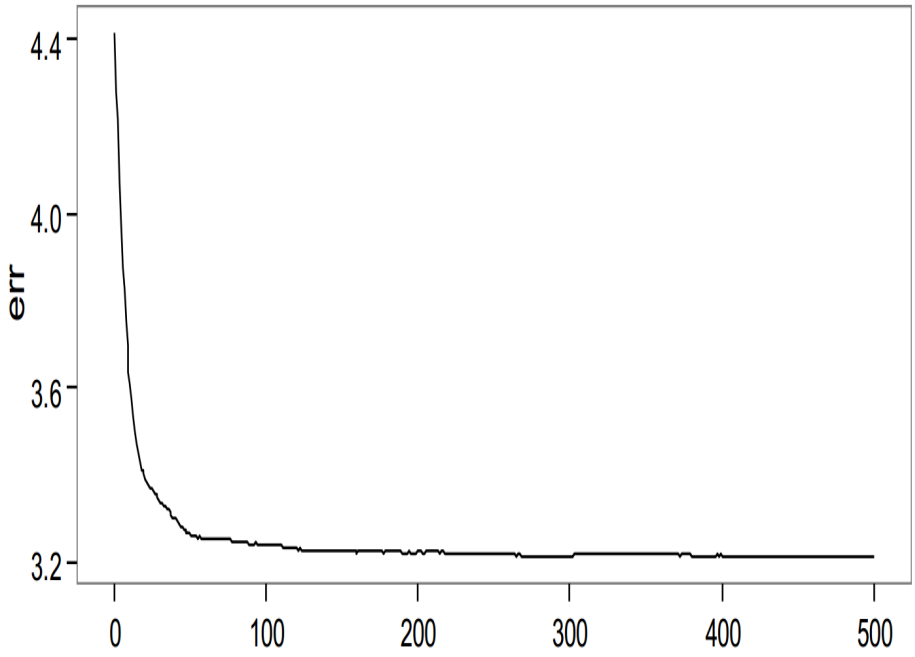

Fig. 2**:** error V.S. decision trees' number

Figure 1 shows the error of random forest with different number of variables. Figure 2 shows the error of random forest with different number of decision trees. The number of selected variables is 3, and the number of decision trees is 200, which makes the out of bag error the least.

Table 1 partial least squares discriminant analysis V.S. random forest

Data set number	Partial least squares discriminant analysis	Random forest
20	0.85	0.95
50	0.88	0.92
100	0.83	0.87
500	0.78	0.85
1000	0.79	0.91
10000	0.77	0.88
100000	0.74	0.89

Compare the random forest method in this paper with the partial least squares discriminant analysis method. The accuracy of both methods under the same amount of data has been investigated. As shown in table 1, the accuracy of partial least squares discriminant analysis is approximately 0.8, and declines with the increase of data; the accuracy of the random forest method in this paper is approximately 0.9, and increases with the increase of data. The investigation shows that the method in this paper performs much better.

References

[1] Becker B, Gerhart B. The Impact of Human Resource Management on Organizational Performance: Progress and Prospects [J]. Academy of Management Journal, 1996, 39(4):779-801.

[2] Breiman L. Random forest [J]. Machine Learning, 2001, 45:5-32.

[3] Santur Y, Karaköse M, Akin E. Random forest based diagnosis approach for rail fault inspection

in railways[C]// Electrical, Electronics and Biomedical Engineering. IEEE, 2017.

[4] H. Abdi, L. J. Williams, “Principal component analysis”. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), pp.433-459, 2010.

[5] J. M. Lee, C. Yoo, I. B. Lee, “Fault detection of batch processes using multiway kernel principal component anal-ysis”. Computers & chemical engineering, 28(9), pp.1837-1847, 2004.

[6] F. Kleibergen, R. Paap, “Generalized reduced rank tests using the singular value decomposition”. Journal of econometrics, 133(1), pp.97-126, 2006.

[7] C. Aytekin, Y. Rezaeitabar, S. Dogru, I. Ulusoy, “Railway Fastener Inspection by Real-Time Machine Vision”. Systems, an, and Cybernetics: Systems, IEEE Transactions on, 2015.

[8] Shreyas R, Akshata D M, Mahanand B S, et al. Predicting popularity of online articles using Random Forest regression[C]// Second International Conference on Cognitive Computing and Information Processing. IEEE, 2017:1-5.

[9] Lai C, Chen X, Wang Z, et al. Rainfall-induced landslide susceptibility assessment using random forest weight at basin scale [J]. Hydrology Research, 2017:nh2017044.