International Journal of Statistical Distributions and Applications
2018; 4(1): 22-28
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20180401.13
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Study on Talent Introduction Strategies in Zhejiang University of Finance and Economics Based on Data Mining

Xiao Yang, Caiyun Ying, Yefeng Zhou

School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China

Email address:
xiao_yang1101@163.com (Xiao Yang), 15700156656@163.com (Caiyun Ying), 963368021@qq.com (Yefeng Zhou)

To cite this article:
Xiao Yang, Caiyun Ying, Yefeng Zhou. Study on Talent Introduction Strategies in Zhejiang University of Finance and Economics Based on Data Mining. *International Journal of Statistical Distributions and Applications*. Vol. 4, No. 1, 2018, pp. 22-28.
doi: 10.11648/j.ijsd.20180401.13

Received: March 6, 2018; **Accepted**: March 19, 2018; **Published**: April 27, 2018

Abstract: Current talent introduction strategies are mainly based on staff arrangement, school discipline construction and so on, which depend on experience actually. However, this kind of empirical approach, lacking of scientific basis, usually causes problems in applications such as uneven scientific research level. In this paper, we intend to use data mining to analyze talent information of teachers in Zhejiang University of Finance and Economics, China from 2011 to 2017, and then to predict their capabilities in obtaining National Foundation of China. In a word, this paper aims to provide decision support for universities' talent introduction strategies. After data cleaning and feature engineering, Apriori algorithm is applied to mine the association rules and find key factors that are closely related to teachers' acquisition of National Science Foundation of China. Then we make predictions with four kinds of models, including Logistic Regression Model, Decision Tree Model, Artificial Neural Network Model and Support Vector Machine Model. In the end, in order to get a more accurate model, Logistic Regression Model which has the highest accuracy of prediction is used to do stepwise regression.

Keywords: Talent Introduction Strategies, Apriori Algorithm, Prediction Model, R Language

1. Introduction

The talent introduction strategies are important for the level of human resources and the speed of university development. And teachers' scientific research level is the most important feature in talent introduction strategies. Besides, whether one can obtain National Foundation of China (NFC) could fully reflect his scientific research level. Therefore, it can be decided whether to introduce the talent by predicting whether he or she can get the NFC within some years. This paper achieves above goals by data mining, and it makes full use of talent information data of Zhejiang University of Finance and Economics (ZUFE), China.

As a new interdisciplinary subject, Baker et al. [1] pointed out that data mining generally refers to the process of searching implicit information from a large number of data through algorithms. Data mining is often associated with computer science, which achieves goals by using the statistics, on-line analytic processing, information retrieval, machine learning, expert system, pattern recognition and so on. In recent years, data mining is widely used in lots of fields, especially in constructing talent training scheme. For example, Gong et al. [2] used data mining methods to construct a more effective evaluation system of talent training scheme. Lv et al. [3] explored the cultivation mode of computer talents by data mining methods. However, there is little research about talent introduction strategies based on data mining in universities. This paper applies data mining to analyze talent introduction strategies in ZUFE.

In recent years, only a few of researchers have applied data mining to talent introduction strategies in universities. Some former research works are introduced in this section. Li et al. [4] used data mining and customer classification method to classify university talent into four types by sensitivity, and found out the factors that influence talent development. Zhang et al. [5] reduced and categorized features by explorative data analysis, showed the relationship among teaching, research and social practices based on data mining. Ranjan et al. [6] used data mining to solve problems existed in the university human resource management such as talents division, brain drain and so on. Wei et al. [7] applied a linear assignment method in the hesitant fuzzy talent introduction

environment.

The rest of this paper is as follows. Section 2 illustrates the related works about data mining and the latest research status. Section 3 describes the research idea of the paper and data preprocessing in detail. Section 4 shows the association rules by Apriori algorithm. The predictions with four models and experiment results are also analyzed in section 4. In the last section, the conclusion is presented.

2. Related Works

Association rules mining is one of methods in data mining. Agrawal et al. [8] first proposed this method. It can mine interesting relationships or casual structures between sets of transaction databases or other databases. As Kamsu et al. [9] pointed out that association rules mining is aimed to find out association rules that satisfy the predefined minimum support and confidence from a given database. Apriori algorithm is one of popular and simple association rules mining methods. There are many researches about the applications of Apriori algorithm, for example, Liu et al. [10] used this algorithm to university management system.

Then we make prediction by four kinds of models, including Logistic Regression Model, Decision Tree Model, Artificial Neural Network Model and Support Vector Machine Model:

Logistic Regression (LR) [11] is a classical method to process Bernoulli's distribution data. It is a machine learning model with simple structure but a high training efficiency. Additionally, LR can be feasible to both categorical and numeric variables without any conversion.

Decision Tree (DT) [12] is one of the basic methods which can be applied to solve various problems by construct a tree-structure model. It has superior quality. For example, it's easy to understand and has good interpretability. In addition, a mature DT can deal with both continuous and discrete variables, avoiding the influence of missing values.

Support Vector Machine (SVM) [13] is a more complex data mining model. In the process of model training and prediction, a few support vectors determine the final result, which help us to grasp the key samples and eliminate a large number of redundant samples.

Artificial Neural Network (ANN) [14] is a kind of mathematical model for simulating biological neural network for information processing. Compared with other networks, ANN are more suitable for modeling and prediction of nonlinear complex systems. ANN has a strong self-learning and self-adaptive ability, through learning the training data, you can train a summary of all the data with a particular neural network significant for prediction.

However, there are few researches for the combination of association rules mining and prediction in domestic and overseas, especially in the field of universities' talent introduction strategies. Therefore, this paper hopes to carry out a variety of data mining using the data of talent information database of ZUFE to extract implicit and useful information, which could offer more scientific advice for the introduction of talents in universities.

3. Research Approach and Data Preprocessing

The research idea of this paper is "raise the problem, analyze the problem and solve the problem". Combining the six processes of data mining, it also means "research introduction and related work - data cleaning and formatting - mining association rules and making predictions", as shown in Figure 1.

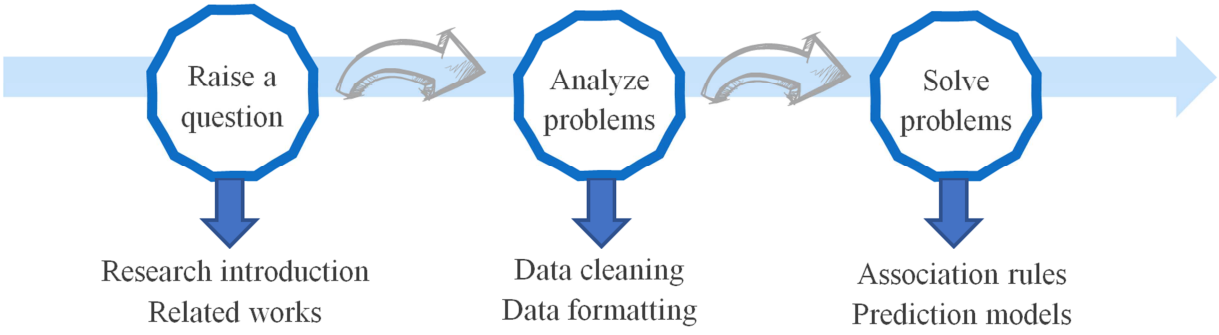

Figure 1. *Basic framework of this paper.*

The data used in this paper is the talent introduction information database from 2011 to 2017 of ZUFE. The sample size is 245.

First of all, data is desensitized to avoid privacy issues. The intended use of the data is limited to academic research. On the basis of desensitization, we clean, identify and correct the data set, including checking data consistency and dealing with invalid value and missing value. In general, there are some errors in the original data, like incomplete data sets, conflicting data records and so on. We clean the dirty data and filter out the data which doesn't accord with the demand.

Due to the association rules mining requires all variables to be Boolean variables, the cleaned data should be further processed. According to the data characteristics, enumeration indexes and quantity indexes are dealt successively. The selected quantity indexes and enumeration indexes are shown roughly in Table 1.

Table 1. *The selected quantity indexes and enumeration indexes.*

Enumeration indexes	Quantity indexes
Gender	The tenure in universities
Marital status	The number of international papers
Undergraduate school	The number of domestic papers
Doctoral school	The number of scientific research awards
Work experience in company or in research institution	Years of doctoral study
Major category	Age
Professional title	

4. Association Rules Mining

4.1. The Theory of Apriori Algorithm

Given a set of data items and a set of transactions set, the aim of correlation analysis is to analyze the frequency relationship of data item set in the transaction set, then we can excavate the interrelationships hidden in data. The mining method which uses association analysis is called association rules. The most classic association rules mining's algorithm is Apriori algorithm, which is proposed by Agrawal. The basic idea of Apriori algorithm is to traversal search the database. In the first iteration, the data item frequency sets with the number of all elements 1 are calculated. Then repeat the lookup to find all the 2-frequent sets. In the k-th iteration, the data items with all the elements k are calculated. At the end of each traversal, it ends when no data item frequency set has been generated at this time. When transaction collection traverses n+1 times (n is the number of elements in the most elemental set of data items), we could find all the big item sets. The use of the Apriori algorithm is to mine frequent item sets. In other words, Apriori algorithm is used to find out frequent combinations. And then, we finally derive our association rules based on these combinations. Apriori algorithm is widely used in various fields. Through the analysis and mining of the correlation of data, the extracted information has important reference value in the decision-making process.

4.2. The Mining Process of Apriori Algorithm

R is used to conduct association rules of the data. First, when setting minimum support of 0.01 and minimum confidence of 0.5, there are 12520 rules (as shown in Figure 2). Then we control the support and confidence degree at the same time, setting minimum support of 0.03 and minimum confidence of 0.6. There are 161 rules (as shown in Figure 3). As we can see from the graph, with the adjustment of the parameters, the redundant and repetitive rules are greatly reduced. But the distribution of effective rules does not differ greatly. This means that the effective information contained in the 161 rules is no worse than the effective information contained in the 12520 rules.

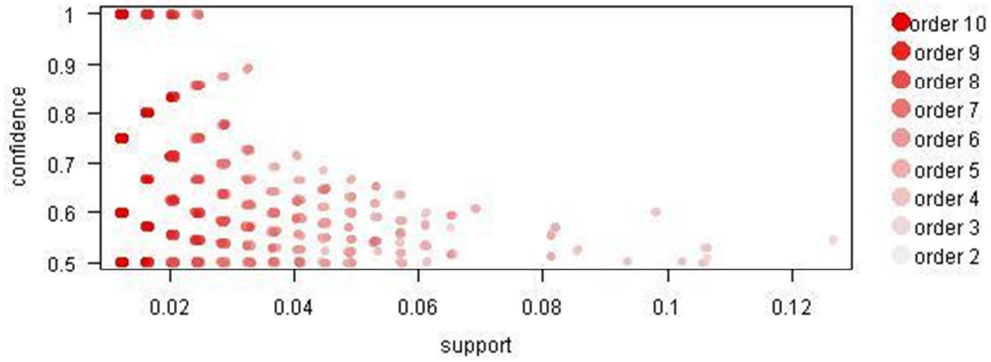

Figure 2. *The scatter plot when the minimum support is 0.01 and the minimum confidence is 0.5.*

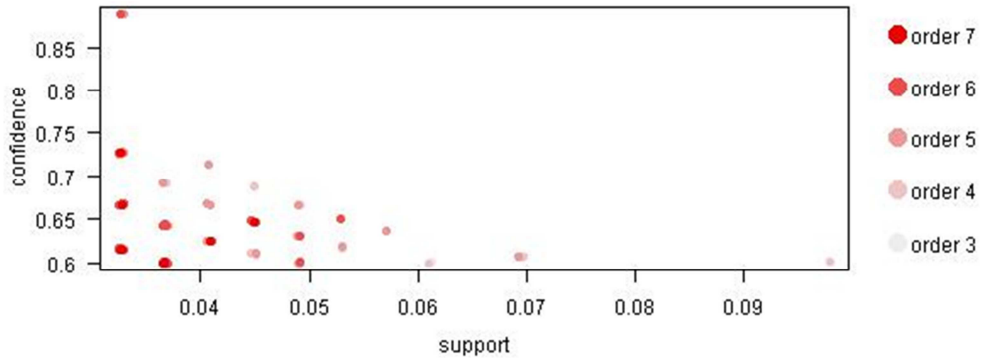

Figure 3. *The scatter plot when the minimum support is 0.03 and the minimum confidence is 0.6.*

In the case of multiple debugging without affecting the valid information, finally the minimum support level is set as 0.05 and the minimum confidence level is set as 0.6. And 14 rules are obtained. The block matrix graph is shown in Figure 4 and the directed graph is shown in Figure 5.

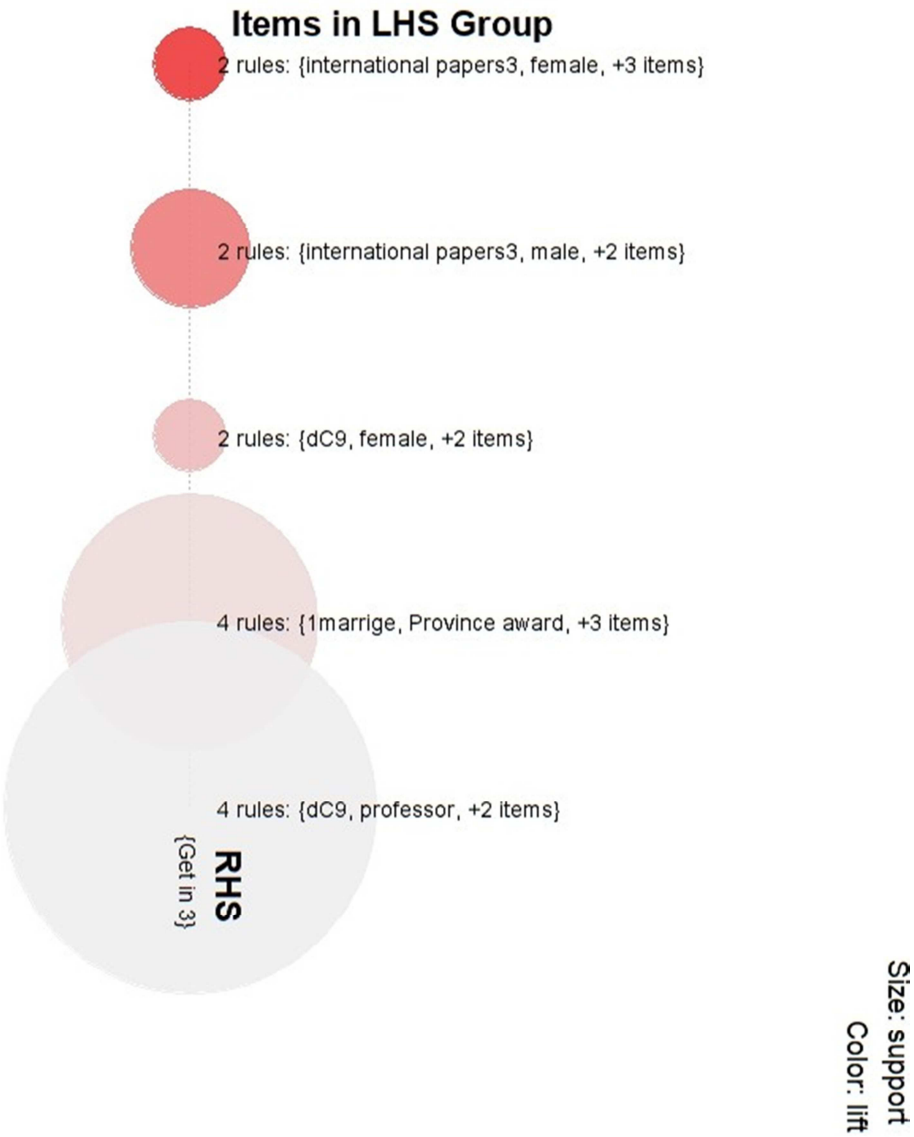

Figure 4. *The block matrix graph when the minimum support is 0.05 and the minimum confidence is 0.6.*

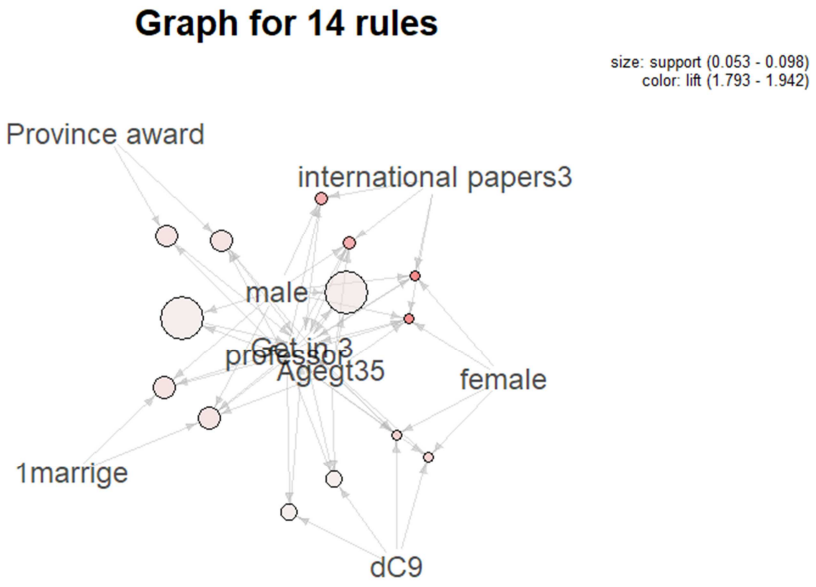

Figure 5. *The directed graph when the size is confidence and the color is lift.*

From Figure 4, we can see that the 14 rules have been clustered into 5 classes. And the depth of the circle color indicates the degree of the lift. The size of the circle indicates the level of the support. So the rules about international papers and gender have the strongest degree of lift and support. Meanwhile, the professional title of professor, gender and age are closely related to the acquisition of NFC in Figure 5. Through comprehensive consideration and screening the repeated rules, 10 rules are listed in Table 2

Table 2. *Selected association rules.*

Rules	Support	Confidence
international papers2*, male, professor => Get in 3*	0.0541	0.8889
National awards, male, professor => Get in 3	0.0541	0.8000
international papers1*, male, professor => Get in 3	0.0608	0.7500
marriage, dC9*, professor => Get in 3	0.0608	0.7500
dC9, male, professor => Get in 3	0.0743	0.7333
dC9, international papers3*, professor => Get in 3	0.0541	0.7273
dC9, teach year, male => Get in 3	0.0541	0.7273
international papers1, professor => Get in 3	0.0676	0.7143
marriage, international papers3, male, professor => Get in 3	0.0676	0.7143
marriage, dC9, male => Get in 3	0.0811	0.7059

*international paper1, international paper2 and international paper3 mean that publishing papers as the first author, the second author and the third author.
*Get in 3 means the teacher could get NFC in 3 years.
*dC9 means the doctoral school is a C9 school.

Based on the selected association rules, the main factors related to acquisition of NFC within three years are publication of international papers, gender, professional title of the researcher, marital status, doctorate school, number of years of teaching in universities before, and the state-level awards received. We can conclude that:

(a) The number of published international papers (as the first author or second author) and the number of scientific research awards at the national level have great effect for getting the NFC.

(b) In most association rules, the variable of professor and the variable of male often appear together, which indicates that the most professors in this school are males.

(c) Whether applicant's doctorate school is C9, has strong association with getting the NFC greatly. However, there are only 4 persons getting a doctorate in foreign school in the data, so the result doesn't show a strong association between them. In the actual situation, this factor shouldn't be ignored.

(d) The interesting result is that an applicant's marital status associates with the ability of getting NFC, which desired a deep learning.

These association rules are in accordance with common cognitive, yet they are qualitative. In order to quantity these results, different predict methods are used to quantify the association between variables appeared in the result and the ability of getting NFC, and make predictions at the same time in the next section.

5. Modeling

In this section, we use four prediction approaches (LR, DT, ANN and SVM) to predict the applicant's ability to obtaining the NFC. Besides, the 70% of data is set as training data while another 30% is set as testing data. First, training data is used to construct a model which can make a prediction. And then, testing data is used to calculate accuracy and verify the effectiveness of the model. At last, we compare these prediction approaches and analyze the strength of those methods.

5.1. Prediction Results of Four Models

In order to predict the applicant's ability to get the NFC more precisely, a comparison between LR, DT, SVM and ANN is made and the program is coded by R language.

In this prediction, following variables which are obtained from the result of association rules mining are chosen to construct a model. Explanatory variables are the number of published international paper as first author, the number of published international paper as second author, the number of published international paper as corresponding author and conference paper, the number of national award, gender, professor, PHD school, teaching years in university, marital status. Dependent variable is whether one can get NFC within three years. Since some employees work less than 3 years, biased result would be produced by those data. Therefore, we deleted those data on entry in 2015 and beyond. The final sample size is 148. Among these 148 samples, 70% of the data are selected randomly as the training data and 30% of the data as the testing data.

Table 3. *The accuracy of the four prediction models.*

Model	ANN	DT	SVM	LR
Accuracy	62.5%	66.5%	72.5%	75%

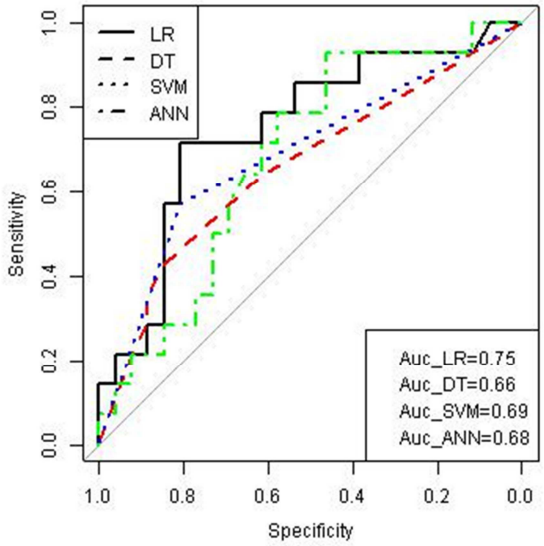

Figure 6. *ROC curves and Auc values of 4 models.*

The prediction accuracy of the four models is shown in Table 3 and the corresponding ROC curves and AUC values are shown in Figure 6. Based on the data used in this paper, the LR has the highest accuracy of prediction. In the next part, LR model is chosen to make stepwise regression, which can help us to understand the data better.

5.2. Stepwise Regression

Stepwise regression can do further variable selection. Therefore, this method is used to construct a more precise model. First, we make regression for each of the explanatory variables respectively, which has the highest accuracy would be chosen as the first explanatory variable in the model. Then repeat previous step for the remaining variables until the accuracy won't be improved. The stepwise regression process is shown in Table 4.

Table 4. *The process of stepwise regression.*

Step	Variables	Accuracy
step 1	international papers1	72.5%
step 2	international papers1, international papers2	72.5%
step 3	international papers1, international papers2, international papers3	72.5%
step 4	international papers1, international papers2, international papers3, professor	77.5%
step 5	international papers1, international papers2, international papers3, professor, national award	77.5%
step 6	international papers1, international papers2, international papers3, professor, national award, dC9	77.5%
step 7	international papers1, international papers2, international papers3, professor, national award, dC9, gender	77.5%
step 8	international papers1, international papers2, international papers3, professor, national award, dC9, gender, teach year	77.5%
step 9	international papers1, international papers2, international papers3, professor, national award, dC9, gender, teach year, marriage	75%

From Table 4, with the increase of the explanatory variables, the accuracy of the model continues to increase, which also indicates that the association rules are reasonable. But at the last step, the marital status has reduced the accuracy. In the end, the accuracy is raised to 77.5%. So there only remain the first nine variables, and the stepwise regression model of LR is that:

$$\text{Get in } 3 = -1.94 + 0.94\text{dC9} + 0.39\,\text{international papers1} + 0.29\,\text{international papers2} + 0.12\,\text{international papers3} + 0.24\,\text{nationalward} + 0.28\,\text{gender} + 0.04\,\text{teach year} \quad (1)$$

We can see from the model equation: applicant's doctorate school has the strongest association with whether getting status fund. And the number of international papers is also very important.

6. Conclusion

This paper takes the data of talent introduction from ZUFE as an example. Through mining association rules of talent information data, the hidden rules related to the availability of NFC in many attribute information are found so as to achieve the goal of selecting important explanatory variables. Then the paper makes predictions by four kinds of models, including LR, DT, ANN and SVM. The most suitable model for the data in this paper is LR model with the accuracy of 75%. Finally, we make a step regression based on the method of LR and get rid of an explanatory variable named marital status. In the end, the accuracy of LR model is 77.5%. Through above steps, the further analysis of the various influencing factors is completed, and the goal of providing a more scientific approach to the introduction of talent programs in universities is realized. However, due to the relatively small amount of data, the missing values and so on, our methods and analysis also have numerous room to improve.

References

[1] R. Baker, E. Duval, J. Stamper, D. Wiley, and S. B. Shum, "Educational data mining meets learning analytics," Technol. Knowl. Learn., 2014, vol. 19, pp. 205-220.

[2] X. Gong, and S. X. Lin, "Construction of evaluation system of sports talent training scheme based on data mining," J. Residuals Sci. Technol., 2016, vol. 13, pp. 343-349.

[3] K. Lv, and Q. Wang, "Data mining in traditional sports talent training program decision making," Advan. Soci. Sci. Educ. Human. Resea., 2016, vol. 71, pp. 767-770.

[4] F. Li, S. Ge, and J. Yin, "Research on talent introduction strategies hazard and training strategy of university based on data mining," Comp. Risk Mgmt., 2011, vol. 5, pp. 219-224.

[5] D. P. Zhang, and D. Jin, "The data mining of the human resources data warehouse in university based on association rule," J. Comp., 2011, vol. 6, pp. 139-146.

[6] J. Ranjan, D. P. Goyal, and S. I. Ahson, "Data mining techniques for better decisions in human resource management systems," Int. J. of Business Information Systems, 2008, vol. 3, pp. 464-481.

[7] G. Wei, F. E. Alsaadi, T. Hayat, "A linear assignment method for multiple criteria decision analysis with hesitant fuzzy sets based on fuzzy measure," Int. J. Fuzzy Syst., 2017, vol. 19, pp. 607-614.

[8] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," Sigmod Rec., 1993, vol. 22, pp. 207-216.

[9] F. B. Kamsu, F. Rigal, and F. Mauget, "Mining association rules for the quality improvement of the production process," Expert Syst. Appl., 2013, vol. 40, pp. 1034-1045.

[10] Y. Liu, C. Wang, and N. N. Wang, "Application of apriori association rule mining algorithm in university management system," J. Residuals Sci. Technol., 2016, vol. 13, pp. 601-604.

[11] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," Ann. Stat., 2000, vol. 28, pp. 337-407.

[12] J. R. Quinlan, "Induction of decision trees," Mach. Learn., 1986, vol. 1, pp. 81-106.

[13] Z. Q. Qi, T. Ying, and Y. Shi, "Robust twin support vector machine for pattern classification," Pattern Recogn., 2013, vol. 46, pp. 305-316.

[14] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nat. Med., 2001, vol. 7, pp. 673-679.