
The Outliers and Prediction Analysis of University Talents Introduced Based on Data Mining

Junlong Zhang, Dan Zhao, Huijie Wang

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

Email address:

1035168671@qq.com (Junlong Zhang), 526939522@qq.com (Dan Zhao), a450098778@163.com (Huijie Wang)

To cite this article:

Junlong Zhang, Dan Zhao, Huijie Wang. The Outliers and Prediction Analysis of University Talents Introduced Based on Data Mining. *International Journal on Data Science and Technology*. Vol. 4, No. 1, 2018, pp. 6-14. doi: 10.11648/j.ijdst.20180401.12

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

Abstract: To create profits for colleges and universities, introduction of talents is an important indicator of the value evaluation of talent introduction in colleges and universities. It can meet the needs of the large data system demand for abnormal detection and prediction in the process of talent introduction. In this article, after reducing the dimension of data by principal component analysis, using the method based on distance (markov distance), the method based on density (local outlier factor) and the method based on clustering (two-step, k-means), we establish the outlier detection model. We find 15 significant outliers and find that the publication of SSCI papers and the experience in C9 institutions have a significant effect on obtaining National Foundation of China. Finally, we use support vector machine, decision tree (C4.5, C5.0), bayes, and random forest to establish the talent prediction model after eliminating abnormal values. By comparing four methods, we find that support vector machine method and decision tree method's prediction accuracies are higher. After optimization, their accuracies can reach 75.00% and 72.09% respectively.

Keywords: Data Mining, Outlier Excavation, Machine Learning, Talent Identification

1. Introduction

In the era of information explosion in twenty-first century, finding potential knowledge from irregular data and providing decision support is an effective way for many enterprises and departments to enhance their competitiveness. We see data mining as an important knowledge discovery technology and it has accumulated rich achievements in theory, many efficient and intelligent algorithms. What's more, they have been continuously improved and perfected after decades of development. In the field of talent introduction, data mining methods have been used to improve the quality of human resources. However, most of the papers at home and abroad do not carry out a deep research on the problems of the introduction of talents in colleges and universities. In this paper, we use distance-based method [1-2], density-based method [3] and clustering-based method [4] to dig out the outliers. Then, we set up different prediction models for comparison, such as support vector machines [5], random forest [6], decision tree [7] and bayes [8].

Outlier data mining is known as outlier analysis, it is used to discover information in the data collection by analyzing

the data (outlier data). Outlier data is the data which deviates from the majority of objects in the data set and even makes people suspect that they may be generated by a completely different mechanism [9]. With the rapid development of the technology of data mining, the outlier data mining has attracted wide attention of scholars at home and abroad, and it becomes an important branch in the field of data mining. Yu et al. [10] used a new deviation test method based on wavelet exchange to remove the clusters from the original data and then identify the outliers. Banker et al. [11] used super efficiency model to identify and remove the outliers, so that the data is not contaminated by outliers and they can achieve more accurate efficiency estimates. Aggarwal and Yu [12] found a rule. For high dimensional data, the notion of finding meaningful outliers becomes substantially more complex and non-obvious. They find the outliers by studying the behavior of projections from the data set. Meanwhile, data mining has widely used in network intrusion detection and prediction of geological disasters, disease diagnosis, fault detection, false cost, terrorism prevention, credit card fraud, loan fraud and other inspection test [13-14].

However, when it comes to the application of data mining in

the field of talent introduction, scholars want to use available data to make certain demands on accuracy. Scholars at home and abroad have made a great deal of research on the identification of excellent talents. Most of researchers mainly focus on the accuracy rate of talent introduction, due to the lack of deletion and selection of outliers, their prediction accuracy of the model of talent introduction is relatively low. This paper analyzes the introduction of talented personnel in colleges and universities, and analyzes the dissertations, academic qualifications and basic information of the talented people to establish the abnormal value detection model and talent prediction model. This model can help colleges and universities determine whether they can introduce the talents. And the talents can bring certain benefit to the university, so this research is worth undertaking. By using the first-hand data to dig out the outliers successfully. We set up some models to predict after removing those outliers. Finally, our prediction accuracy has been significantly improved. It can be seen from the experimental results that the model has a high precision, and most universities can be used for reference.

2. Data Preprocessing and Modeling

In this paper, the data is pre-processed so that the previous text data can be expressed accurately with the numbers. The

useful information can be dug out from the data

2.1. Overview of Data Preprocessing and Methods

The original data has inconsistencies, noise, higher dimensions and other issues. In this paper, we use data cleaning, data integration, data transformation, data protocol and other methods to preprocess data. For the missing value of vacancy and the different properties, the average of the same kind of samples are used to predict the most likely value and we use the method of removing the property. Data reduction is done by dimension reduction and numerical compression.

Many variables in multivariable sample have relevance, which inconvenience to the analysis. Each index is analyzed separately, the analysis is often isolated and prone to erroneous conclusions. Therefore, in this paper, principal component analysis is adopted to reduce the need to analysis of indicators while minimizing the loss of information contained in the original data to achieve the goal of reducing dimension.

As shown in figure 1, we use the method shown in the figure to perform the outliers mining and prediction tasks:

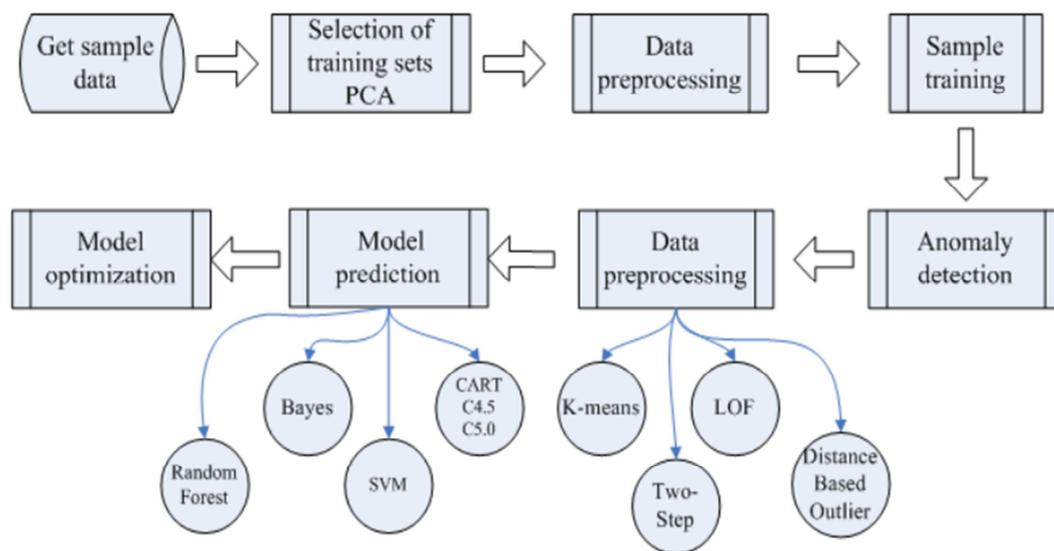


Figure 1. Modeling flowchart.

2.1.1. Outliers Mining Method

Distance based method is mainly based on the distance from a given object, it avoids too much computation, and it can be detected multiple times by changing the setting of the distance to avoid larger errors.

Outlier data mining based on density is built on the basis of density clustering. It determines whether data objects are abnormal by calculating the abnormal factors of data objects.

The basic idea of clustering method is based on outliers the process of clustering outliers, the data set uses the mature model of cluster analysis, divides the data set into multiple clusters and chooses far away from the cluster centroid

samples as outliers.

2.1.2. Prediction Method

1. Support vector machine model is based on the statistics theory and structural risk minimization principle [5]. Based on the limited sample information in the model's complexity, it can get the best promotion ability.

2. Random forest algorithm generates a number of classification trees in a random way, and then it summarizes the results of classification trees [6]. Without significant improvement in the calculation quantity, the prediction accuracy is improved. What's more, the data of missing data and non-equilibrium are relatively stable, which can be used

to predict the function of up to thousands of explanatory variables.

3. Decision tree is a predictive modeling method for classification, clustering and prediction [7]. This method requires building a tree to model the classification process. And it can be divided into CART decision tree, C4.5 decision tree and optimized C5.0 decision tree.

4. Bayes prediction model is a prediction method by using bayes statistics [8]. Bayes statistics is different from general statistical method. It is not only using model information and data information, but also making full use of prior information.

2.2. Methods for Detecting Outliers

2.2.1. Based on the Distance of Outliers (Markov Distance)

The markov distance can be defined as the degree of difference between two random variables that follow the same distribution and whose covariance matrix is Σ , if the covariance matrix is the unit matrix, then the markov distance is simplified as the euclidean distance, if the covariance matrix is diagonal matrix, it can also be called the normalized euclidean distance [1].

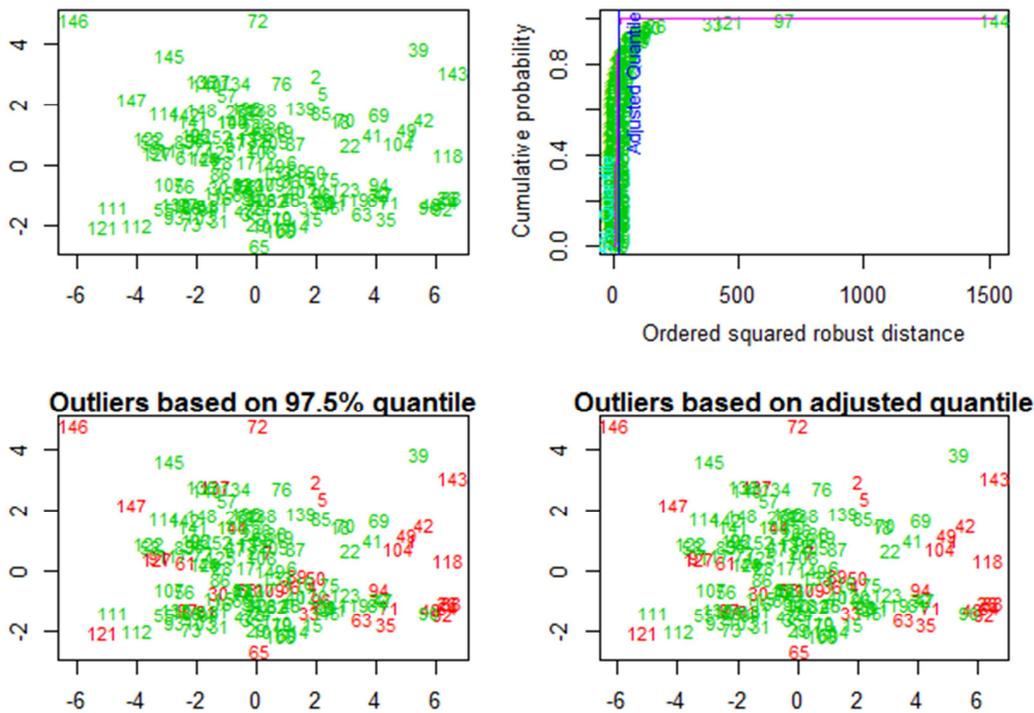


Figure 2. The model of markov distance.

The picture in the upper left corner of figure 2 is the original data, the upper-right corner of the X-axis graph is a sample of stable markov distances, Y axis is the empirical distribution of the distance, red curve is the chi-square distribution, blue vertical can be seen as threshold. When the sample appears on the right side of the threshold, it can be seen as an outlier. Both the lower left and the lower right are shown in different colors, those can be seen as the outliers, but the threshold is slightly different.

It can be observed that the small number of outliers is correctly judged. However, two normal values are misjudged as outliers. So, the parameters still need to be adjusted. We use the ten-fold crossover method to improve.

According to general analysis, the conclusion is: the abnormal value of the top 15 points is 144, 33, 35, 97, 49, 26, 36, 48, 61, 71, 35, 109, 118, 112, 53.

2.2.2. Based on the Distance of Outliers (k-means)

Clustering is the grouping of data in the database, it make the data in each group as similar as possible and make the

data in different groups as different as possible [1-2]. Algorithm is the most widely used clustering method, each of these categories is represented by the average (or weighted average) of all data in the class, which is called the cluster center. The scalability of the algorithm is good, and the time complexity is $O(k)$.

The distance between the data point and the prototype is the objective function of optimization, and the method of using the function to find the extreme value is used to get the adjustment rules of the iterative operation. The k-means algorithm uses euclidean distance as the similarity measure which is the optimal classification of the initial clustering center vector and makes the evaluation index minimum [2].

This paper specifies 3 centroids, calculating the distance from each sample to each centroid, and updating the centroid until the difference between the updated centroid and the centroid is less than the pre-defined tolerance before updating. As shown in figure 3, a scattergram matrix with four principal components is drawn from the 15 outliers which calculated from the mass center. It can seen that these fifteen

points are all far from most points.

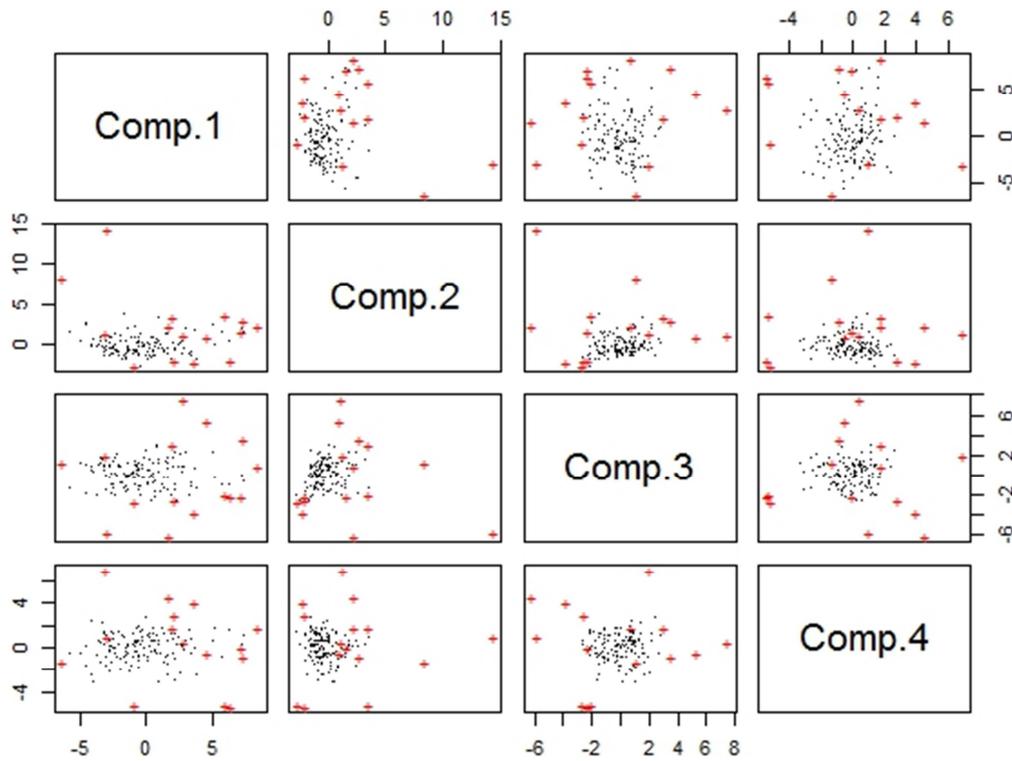


Figure 3. Four principal component scatter plots.

According to general analysis, the conclusion is: the abnormal value of the top 15 points is 144, 50, 33, 97, 49, 26, 72, 42, 64, 92, 35, 109, 119, 143, 121. These fifteen points are all far from most points, so they are outliers.

2.2.3. Based on the Density of Outliers (LOF)

The definition of density-based outlier is based on the definition of distance. The concept of density is obtained by combining two parameters: the distance between points and the number of points in a given range.

Local anomaly factor: according to the definition of local reachable density, if a data point is far away from the other points, it is obvious that the local reachable density is small [3]. However, the LOF algorithm measures the degree of anomalies in data points rather than the absolute local density [3]. It's the relative density of the neighboring data points. The advantage of doing this is that the distribution of data is not uniform and the density is different. The local anomaly factor is defined by the local relative density. We can see this in figure 4.

`density.default(x = outlier_score)`

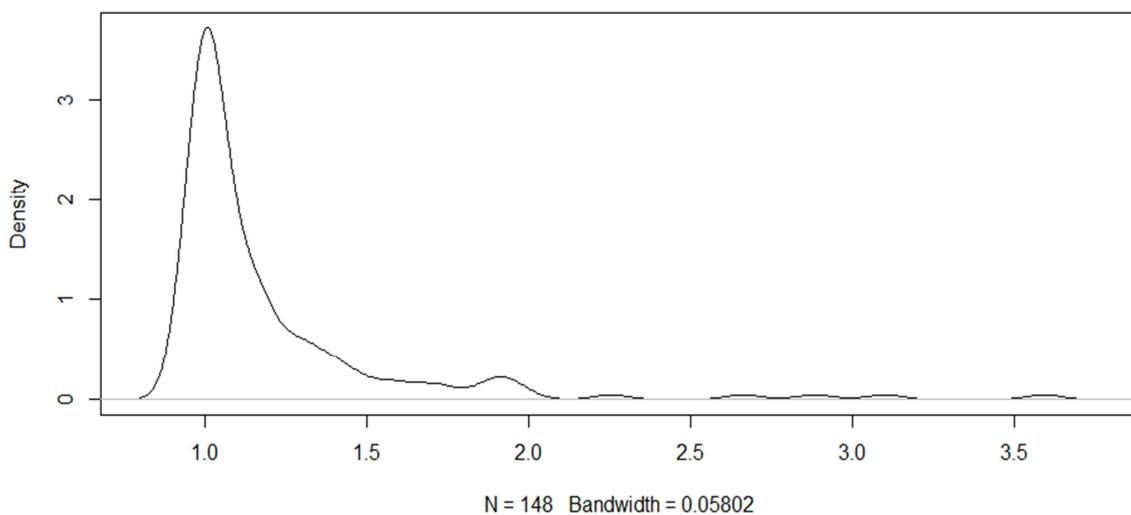


Figure 4. Local outlier factor of density analysis.

According to general analysis, the conclusion is: the abnormal value of the top 15 points is 144, 33, 121, 97, 35, 49, 50, 146, 36, 26, 92, 109, 64, 42, 72.

2.2.4. The Two-Step Clustering Algorithm

The two-step clustering algorithm is divided into two stages:

Pre-clustering stage: The idea of CF tree growing in BIRCH algorithm is adopted. Data points are read one by one in the data set [4]. While CF tree is generated, the data points in dense areas are clustered to form many small sub-clusters in advance.

Cluster stage: With the result of pre-clustering stage, the cluster is the object, and the cluster is merged into the cluster, until the number of clusters is expected [4].

We use the two-step clustering algorithm to dig out the outliers. We use SPSS to find the points, the points that can be expressed are 144, 121, 33, 97, 49, 26, 89, 48, 36 and 50, which are the top ten points of the anomaly index. It is the most remarkable level of abnormality. These 10 points are a single category, so they are the outliers.

2.2.5. Analysis of Outliers Detection Process

After we preprocess the data, four methods are used to find the outliers by clustering analysis of all the data and 148 data separately. We list them in the table 1 and make some analysis.

Table 1. The outliers of the first ten of each method.

Markov distance	k-means	LOF	two-step
144	144	144	144
33	50	33	121
35	33	121	33
97	97	97	97
49	49	35	49
26	26	49	26
36	72	50	89
48	42	146	48
61	64	36	36
71	92	26	50
35	35	92	35
109	109	109	72
118	119	64	112
112	143	42	109
53	121	72	4

Summary, we analyze these fifteen anomalies and find that fifteen outliers do not graduate from the C9 school, so we can initially believe that C9 graduates have a certain degree of scientific research strength and the general emphasis on academic. Second, some national and provincial awards, such as nation awards, provincial awards, national comprehensive awards and provincial comprehensive awards, are evidence of a teacher's scientific research ability. At the same time, among the 15 outliers, the best teacher won the provincial comprehensive awards, so we have reason to believe if the teacher won the nation awards, he will win the NFC. Teaching years, post doctor life years and many other variables have received a certain negative impact on the NFC,

our explanation is that the teacher spent too much time in preparing lessons, classes and correcting homework. So, if teachers reduce the research time, they will reduce the achievements of scientific research. And the publication of the paper is very important, but as long as the teacher issued a minimum level of SSCI journals, he can get a 56.25% chance of the NFC. In the 15 outliers, few teachers have published SSCI papers here, it illustrates that the first top journals published quite sure that one person's level of scientific research and the future development. At the same time, we find an interesting phenomenon. One people has published some 2A and 2B articles, but he didn't get the NFC. By the way, those teachers also appear in our outliers.

2.3. Methods for Prediction

2.3.1. Support Vector Machine

The core theory of support vector machine is that VC dimension theory, the optimal hyper plane concept and nuclear space theory. 1) VC dimension theory minimize the VC dimension of function set in order to control the structure error of the learning machine. 2) The optimal hyper plane concept to minimize the VC dimension of the function. 3) The nuclear space theory maps the input space into the high-dimensional feature space by using the non-linear mapping, which transforms the linearly inseparable problem in the low-dimensional input space into the linearly separable problem in the high-dimensional feature space. By passing the high dimension space, the kernel function makes the operation in low dimension of input space.

Table 2. The prediction ability of radial gaussian kernel in different data.

Data	Number of support vector	Accuracy
primary data	82	63.89%
data without outliers	91	66.67%

From table 2, it can seen that the original data prediction capability is 63.89%, with the accuracy improved to 66.67%.

2.3.2. Bayes

Bayes prediction model is a prediction model based on bayes statistics. Bayes statistics is different from general statistical methods. It not only utilizes model information and data information, but also makes full use of prior information.

The statistical prediction method of Thomas Bayes is a time series prediction method which takes dynamic model as the research object. In statistical inference, the general pattern is: prior information + general distribution information + sample information \rightarrow posterior distribution information.

It can be seen that bayes model not only takes advantage of the earlier data information, but also adds information about decision makers' experience and judgement, and combines objective factors with subjective factors, which makes the occurrence of abnormal situations more flexible. The test of test-data was carried out with the established bayes prediction model, with a prediction rate of 58.14%.

2.3.3. Random Forest

The central idea of random forest is that we need to create a forest in a random way. There are a lot of decision trees in the forest, and there is no correlation between every decision tree in the random forest. After we get the forest, a new input

sample enters the random forest, let each decision tree can make a judgment separately. Then we take a look at the sample should belong to which kind of classification algorithm, and see what kind of selected the most. Finally, the sample for what category should be predicted.

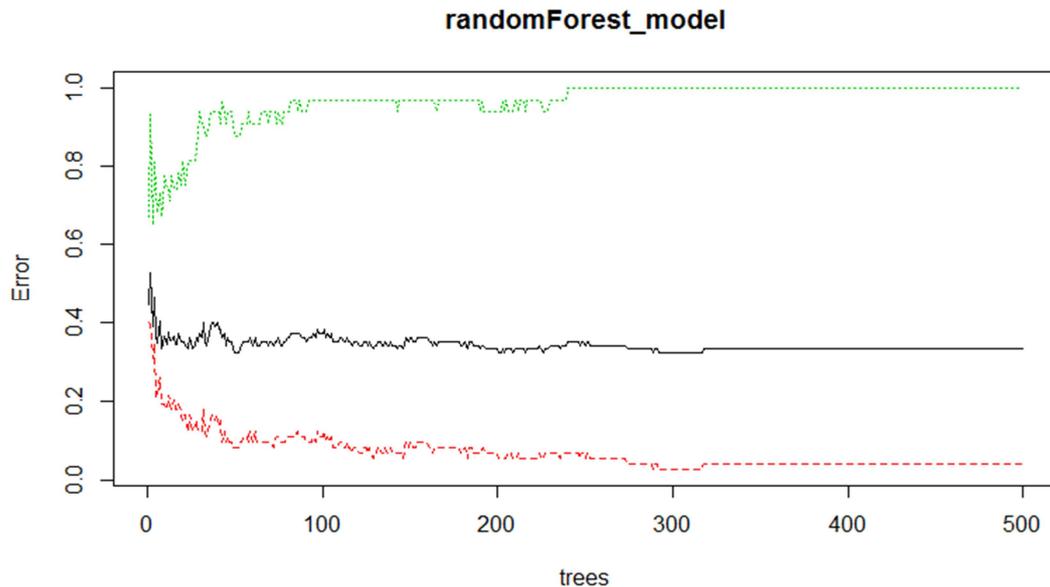


Figure 5. The random forest mode.

Use the loop function, and the error rate is the lowest when $mytry=1$. The $ntree$ parameter is the number of decision trees in modeling, and the low $ntree$ value can lead to high error rate. The high $ntree$ value can improve the model complexity and reduce the efficiency. From figure 5, we can see that when $ntree=350$, the error in the model is basically stable. In the consideration of insurance, the $ntree$ value is 350, and a random forest model is established for parameters. The test-data is tested with a good random forest model, with a prediction rate of 62.79%.

2.3.4. The Decision Tree

Among the many methods used to solve the classification problem, decision tree is one of the most commonly used method. It is used for classification, clustering and predict the prediction model modeling method. It uses the method of "divide and conquer" and divides problem of the search space into several subsets. This method needs to build a tree to model the classification process. Once the tree is built, it can be applied to the tuple in the data set and get the classification result. In the decision tree method, there are two basic steps: building the tree and applying the tree to the data set, it focuses on how to build the tree's research effectively. Finally, it will come to a conclusion in decision tree leaf nodes, the whole process is based on the new node as the root of the tree to repeat.

1) CART:

Classification and regression tree (CART) algorithm divides the data into two subsets, the samples for each subset have better consistency than before being divided, we do it many times. After the results meet the termination criterion,

we get the final decision tree by building and evaluating.

The ROC curve is a visual tool for displaying the effect of a full range of classification models. ROC is the receiver running curve (Receiver Operation Characteristic). The comparison between the true case rate ($a/(a+b)$) and the false positive rate ($c/(c+d)$) of a given model is shown. The longitudinal axis of the ROC curve represents the true case rate, while the transverse axis represents the false positive rate. The ROC curve is near to the diagonal line, the accuracy of the model is low, and the ROC curve is close to the upper left corner, the accuracy of the model is high. From figure 6, it can seen that the CART model have a high accuracy.

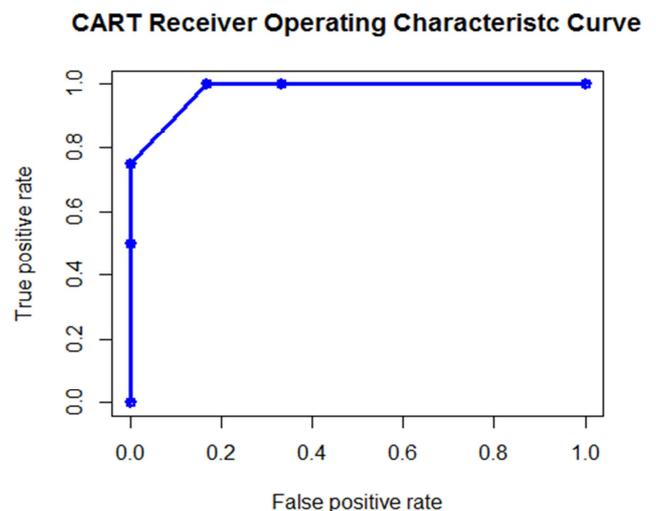


Figure 6. The ROC curve.

In this paper, the CART model has 10 terminal nodes. For 105 sample classification, there are only 10 classification errors, the accuracy rate is 90.48%. And the test-data model has been tested with the established CART model, the prediction rate is 72.09%.

2) C4.5, C5.0:

C4.5 algorithm is used in the classification of machine learning and data mining. It is used to learn to find a mapping relationship from attribute values to categories and it can be

used to classify unknown entities of new categories.

C5.0 algorithm is the revision of C4.5 algorithm. It is known as boosting trees, computing speed is faster in the software and this model has less memory resources.

From figure 7, we can see that samples split by the first Comp13. Then, the samples split by the Comp3 and Comp8, although there are only two branches under Comp3, when $Comp13 < -1.12$ and $Comp3 > 1.88$, the probability of getting the NFC is high.

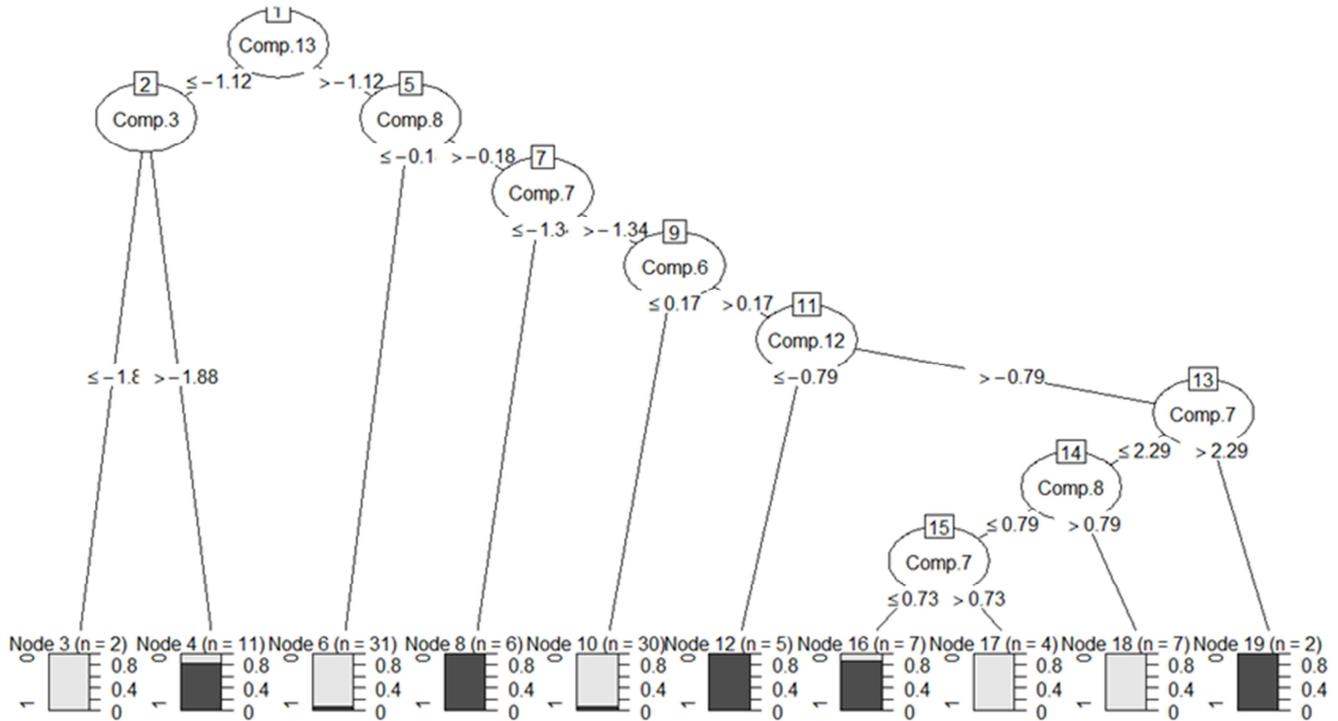


Figure 7. The C4.5, C5.0 tree model.

So, it need to satisfy multiple conditions. What's more, it has a chance to improve the probability of getting the NFC. For C5.0 decision tree. We use 12 principal components to predict and see from the decision tree that the prediction uses 105 data, and it can seen that the error is 30.50% from the confusion matrix. So, C5.0 decision tree's prediction accuracy is 69.50%. The accuracy rate of this algorithm is good and we will optimize it later.

2.3.5. Comparison of the Effects of Various Prediction Models

Table 3. The results of various selected methods.

Method	Accuracy
SVM	66.67%
CART	72.09%
C5.0	69.50%
K-NN	75.00%
ANN	71.05%

For the selected kernel function, degree parameter is the parameter of the kernel function polynomial, and the default value is 3. The parameters of all functions are given in the

gamma parameter and the default value is 1. The coef0 parameter is the parameter of the inner product function and sigmoid of the kernel function, and the default value is 0. In addition, parameter cost is the separation point weight of the soft interval model.

From table 3, we can see the results of various selected methods. KNN has the highest accuracy.

3. Optimization of the Model

The prediction accuracy of each model did not meet our expectations. Therefore, the model is optimized.

3.1. Ten Fold Cross Validation

Ten fold cross validation is used to test the accuracy of the algorithm. It is a common test method. First, the data set is divided into ten parts, taking nine of them as training data and one as test data. Then, each test draws the corresponding accuracy, we use the average of 10 fold cross-validation as estimation of the accuracy of the algorithm. We need to do 10 fold cross-validation many times and seek its mean. Finally, we need to estimate the accuracy of the algorithm.

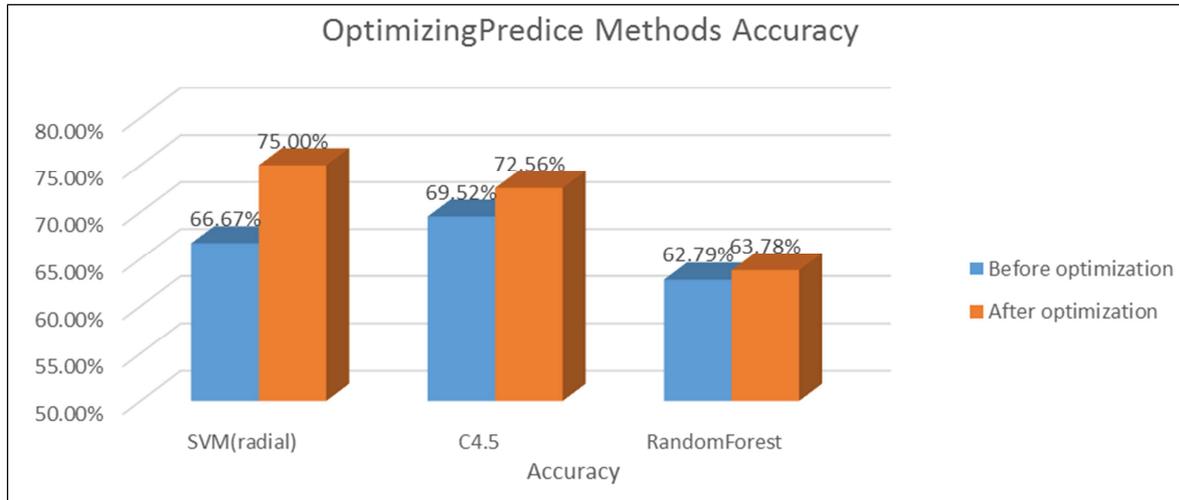


Figure 8. 10 cross-validation accuracy to improve the comparison diagram.

From figure 8, we can see that the accuracy of various methods has been improved after optimization.

3.2. Support Vector Machine Model Improvement

In this paper, the tune function is used to select the optimal model parameter cost penalty coefficient C and γ . We

discuss different kernel functions in this article. Different kernel functions have different results, we show it in the table 4. It is easy to find that polynomial support vector machine model is better than others. So, we have reasons to believe that polynomial support vector machine model can be used to predict the accuracy of talent introduction.

Table 4. Different kernel functions support vector machine optimization.

Method	Cost	Gamma	Number of support vector	Accuracy before optimization	Accuracy after optimization
linear	100	1e-05	74	66.67%	60.72%
radial	10	1e-04	90	66.67%	75%
polynomial	100	1e-05	82	60.71%	75%

From table 4, it can be seen that the optimized precision is higher than before optimization.

4. Conclusion

This article is aimed at the large amount of information. In this article, in order to meet the need of system big data anomaly detection and prediction in the process of talent introduction, after reducing the dimension of data by PCA, the method based on distance (markov distance), the method based on density (local outlier factor) and the method based on clustering (two-step, k-means) are used to establish the outlier detection model. We find 15 significant outliers. The two-step method can make full use of the information of the data set and excavate outliers efficiently. By analyzing the common features of the outliers, we find that the teachers in the outliers are not graduated from C9 schools, and the number of articles published in SSCI papers is generally less than 3. What's more, most of them have not won the national awards for scientific research. So, we think that the more C9 school graduates, the more SSCI papers published. The talents can win the prize at the national level which are generally strong in scientific research ability, and this kind talent should be introduced to our school.

After removing the outliers, we use SVM, decision tree (C4.5, C5.0), bayes and random forest to build a talent

prediction model for the rest of the data. The prediction results of SVM, decision tree (C4.5, C5.0), bayes and random forest are 66.67%, 69.50%, 58.14%, 62.79%, respectively. By comparing four methods, we finally optimize the two prediction methods with high accuracy, and we can get the highest prediction accuracy of the optimized SVM model is 75.00%. According to the results of the experiment, the SVM (radial gauss core) has a certain advantage in predicting whether the teacher will be able to get the NFC in 3 years. Theoretical analysis and experiments show that the algorithm proposed in this paper is effective and feasible.

References

- [1] E. Knorr and V. Tucakov, "Distance-based outliers: algorithms and applications," *Vldb Journal*, 2000, vol. 8, pp. 237-253.
- [2] F. Jiang, J. W. Du, Y. F. Sui, et al, "Outlier detection based on boundary and distance," *Acta Electronica Sinica*, 2010, vol. 38, pp. 700-705.
- [3] M. M. Breuing, H. P. Kriegel and R. T. Ng, "LOF: identifying density-based local outliers," *ACM Sigmord Record*, 2000, vol. 29, pp. 93-104.
- [4] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, 1999, vol. 31, pp. 264-323.

- [5] L. V. Utkin, A. I. Chekh and Y. A. Zhuk, "Binary classification svm-based algorithms with interval-valued training data using triangular and epanechnikov kernels," *Neural Networks*, 2016, vol. 80, pp. 53-66.
- [6] L. Breiman, "Random forest," *Machine Learning*, 2001, vol. 45, pp. 5-32.
- [7] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," *Machine Learning: Sixteenth International Conference*, 1999, vol. 99, pp. 124-133.
- [8] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, 2004, vol. 3, pp. 1-25.
- [9] R. K. Pearson, "Outliers in process modeling and identification," *IEEE Transactions on Control Systems*, 2008, vol. 10, pp. 55-63.
- [10] D. Yu, G. Sheikholeslami and A. Zhang, "Findout: finding outliers in very large datasets," *Knowledge and Information Systems*, 2002, vol. 4, pp. 387-412.
- [11] R. D. Banker and H. Chang, "The super-efficiency procedure for outlier identification, not for ranking efficient units," *European Journal of Operational Research*, 2006, vol. 175, pp. 1311-1320.
- [12] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," *ACM Sigmod Record*, 2001, vol. 30, pp. 37-46.
- [13] M. S. Chen, J. Han and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and Data Engineering*, 1996, vol. 8, pp. 866-883.
- [14] F. Jiang, J. W. Du, Y. F. Sui, et al, "Outlier detection based on boundary and distance," *Acta Electronica Sinica*, 2010, vol. 38, pp. 700-705.