

Clustering Analysis on the Introduction of Talents in Colleges

Fang Dan, Chen Xinhui, Xi Xin

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

Email address:

805826364@qq.com (Fang Dan), 270818634@qq.com (Chen Xinhui), 1418211498@qq.com (Xi Xin)

To cite this article:

Fang Dan, Chen Xinhui, Xi Xin. Clustering Analysis on the Introduction of Talents in Colleges. *International Journal on Data Science and Technology*. Vol. 4, No. 1, 2018, pp. 15-23. doi: 10.11648/j.ijdst.20180401.13

Received: March 6, 2018; **Accepted:** April 18, 2018; **Published:** April 27, 2018

Abstract: With the development of economy and technology, introducing and training talents have become the key driving force in the world which can enhance the competitive strength of the whole countries. Therefore, the strategies of strengthening the universities and colleges with more talented people and making efforts to implement the construction of “Double top” are put forward in the same time. Methods of clustering analysis have been widely used in the actual researches. In this study, an effective clustering analysis model by comparing the clustering analysis under different dimensionality reduction methods is established. Firstly, preprocess the data about talent introduction which is collected from Zhejiang University of Finance and Economics, and use Principal Component Analysis (PCA), Weighted Principal Component Analysis (Weighted-PCA) and Random Forest (RF) to reduce the dimensions of the data. Next, use K-means clustering algorithm and K-medoids clustering algorithm to cluster the preprocessed data. The classification results indicate that the K-medoids algorithm with Weighted-PCA is superior to other clustering methods in this illustrative case. In addition, the experiment divides talents into high-end talents and mid-end talents. By looking into the analysis of the characteristics of the clustering results, some targeted advices on the talents introduction in colleges can be provided.

Keywords: Clustering Analysis, Dimensionality Reduction, Clustering Algorithm

1. Introduction

Nowadays with the development of modern technology, the concept that science and technology is the primary productive force is playing an extremely important role increasingly. Talent cultivation, which is a key factor for the development of the technology, has gained more and more attentions. Hui et al. [1] pointed out the drawbacks of talent cultivation soft power in the light of awareness, mechanism, methods and facilities through the case study of international trade undergraduate program and discussed measures to improve soft power. Xu [2] through the analysis of the achievements and shortcomings of the existing research literature, discussed the problems of constructing the cultivating mode of undergraduate tourism talents from theoretical basis, practice demand, policy orientation, regional characteristics, developing trend and educational resources. Ge et al. [3] put forward the construction of theoretical teaching system, practice teaching system construction, three aspects of the construction of

comprehensive quality system for the understanding of the application type talents and the main problems in the cultivation of talents in universities. But these studies lack the support of specific empirical models. Therefore, the data mining methods as well as the related research have been extremely necessary. And in this paper, some concrete models are established to explain the problem of talent introduction through the combination of clustering and other methods, and some relevant suggestions are proposed for the results.

The university is the cradle of talent cultivation. And the quantity and the quality of talents introduced by universities as well as the subsequent cultivate may directly affect the efficiency of cultivating excellent talents. Therefore, developing a high level and high quality team of teachers is an important task of all the colleges at hand. As for the current database, a lot of relevant data from various colleges has been collected. Due to the complexity and huge amount

of the data, obtaining some effective information and intrinsic rules from the data itself is difficult. Therefore, using data mining methods to get some useful information is necessary. Clustering analysis is an unsupervised learning method that is commonly used in data mining. This method divides the data into different categories based on the distance between the data. As a result, the interclass variance is minimized and the variance between classes is maximized. In general terms, clustering refers to clustering the same type of data and distinguishing different types of data. Also the clustering method has been widely used in many research fields. In this study, clustering analysis is used to identify the hidden features of the data. Moreover, Random Forest Model (RFM), PCA and Weighted-PCA are used in order to reduce the number of dimensions of the data. Also clustering algorithm of K-means and K-medoids are utilized for clustering analysis. Finally, compare the results of two clustering algorithms under different dimensionality reduction methods and choose an effective clustering model to classify the types of talents and provide relevant suggestions for college personnel training.

Section 2 briefly reviews some researches of other fields about the methods of clustering and introduces some relevant theoretical knowledge. Section 3 describes the data in details, provides the data preprocessing and builds the model based on the data. Section 4 does experiment and depicts the results of classification in detail. Section 5 concludes the research.

2. Related Works

In modern statistical methods, clustering analysis as a classification method divides the objects into several classes according to certain rules. But these classes are not given in advance, they are determined by the characteristics of the data. In the same category, these objects tend to be similar in

a sense and the objects in different categories tend to be dissimilar. And this method is widely used in many fields. In the medical research, aiming to use an unbiased approach based on multivariate classification techniques to phenotype patients with allergic asthma, Sendín-Hernández *et al.* [4] used clustering analysis to cluster three groups, and identified three phenotypes of allergic asthma. Brophy *et al.* [5] used an exploratory clustering analysis to solve the question of whether people on Community Treatment Orders (CTO) can be categorised into statistically reliable, qualitatively distinct groupings. And they identified the possibility of improved targeting of CTO and specific treatment planning. In biology, Eo *et al.* [6] used clustering analysis to divide the ground beetle community into two groups and by deep analysis and comparison included that climate and heterogeneous landscape compositions may influence the distribution of ground beetle communities in agricultural fields. Adamczyk *et al.* [7] aimed to examine the possibility of using cluster analysis to assess physical activity of dairy cows milked in the Voluntary Milking System while taking into account environmental conditions. In diet research, Funtikova *et al.* [8] used clustering analysis to define the validity and reproducibility of the dietary patterns. In addition, clustering analysis is used in other fields. Halčinová *et al.* [9] used clustering analysis as a tool for creation of groups of similar stock items leading to speed up the company's reactions on customers' requirements. Kijewska and Bluszcz [10] used clustering analysis to distinguish the isolation of homogeneous subsets of surveyed objects.

According to the different basic ideas, divide clustering methods which have been commonly used into the following four categories: Analytic Hierarchy Process, Division Method, Density-based Method and Grid-based Method. The advantages and disadvantages of the four methods are as follows.

Table 1. Comparison of four clustering methods.

Method	Advantages	Disadvantages
Analytic Hierarchy Process	a. The similarity of distance and rules are easy to define. b. Do not need to predetermine the number of clusters. c. Can find hierarchical relationship.	a. Calculation is more complicated. b. The effect of singular values is very high. c. The algorithm is likely to cluster into a chain.
Division Method	a. Wide range of applications. b. The speed of convergence is fast. c. Can be used for large scale database.	a. Identify clusters with the convex distribution, same size and density. b. Center selection and noise cluster affect the results.
Density-based Method	a. Overcome the shortcomings of distance-based algorithms. b. Find clusters of any shape and it can effectively process noise data in data. c. Insensitive to entry order of the data.	a. Input parameters are very sensitive and it is difficult to determine the parameters. b. Calculation is more complicated.
Grid-based Method	a. The time to process the data has nothing to do with entry order of the data. b. Any type of data can be processed.	a. The time to process the data is related to the number of cells divided by each dimension.

By comparing the merits and demerits of the four clustering methods combining the characteristics of the selected data, finally choose the division method, and use the K-means and K-medoids algorithm to cluster data. The theories of the two algorithms are introduced as follows.

(a) K-means algorithm

One of the most commonly used analysis methods in

clustering analysis is the K-means algorithm. Generally, the K-means algorithm uses Euclidean Distance Formula (EDF) to calculate the distance among the data.

Suppose there is a set of points in m -dimensional Euclidean space [11]:

$$S = \{X_1, X_2, \dots, X_n\}, \text{ and } X_i = (x_{i1}, x_{i2}, \dots, x_{im}) \ (i=1, 2, \dots, n).$$

In this space, select k center points within a certain range $V_i (i=1, 2, \dots, k)$ [11]. Guarantee that the sum of the squares of the distances which are the n points to their own centers is the smallest [11]. Use algebraic formulas to express:

$$E = \sum_{i=1}^k \sum_{X \in C_i} \|X - V_i\|_p \quad (1)$$

In formula (1), E represents the sum of the distances between all points and their own cluster center. X indicates the sample points [11]. V_i represents the average value of each cluster center.

$$\|X - V_i\|_p \quad (2)$$

represents a p -degree measure between X and V_i [11].

(b) K-medoids algorithm

There are some differences between K-means and K-medoids. And the difference is that K-medoids uses one of the most representative observations as a center point. Set k clusters, then use formula (1) to calculate the value of the objective function.

In addition, the technology such as big data, cloud computing and so on, is getting more and more universal. And people are getting more and more convenient ways to collect data. The data which has the characteristics of multi-dimensional and wide range has been used in many empirical researches. However, many attributes are redundant in these panel data and we must find a suitable method to reduce the dimensions of the original data. Simultaneously, due to excessive variables clustering directly is likely to cause greater errors. So it is necessary to reduce the dimensions of the data.

Fei and Liu [12] were motivated to propose a new monitoring method by compensating the principal component analysis with a weight approach. By this method, they increased the weight of the direction of the component with large estimation error and reduced the influence of other directions at the same time. Rodríguez et al. [13] yielded a strong reduction of dimensionality of elevation topography data, to only 19 independent parameters combining Zernike fit and PCA and got the minimum number of orthonormal basis functions. Goudarzi et al. [14] proposed a noble quantitative structure-property relationship technique on the basis of the random forest. In addition, the dimensionality reduction method is used in other studies. RF, PCA and other classifications are very helpful for the research. These literatures have played a guiding role in the following research. Basing on the previous researches, PCA, Weighted-PCA and RFM are selected to reduce the dimensions of the data. And use R programming language to run these three dimensionality reduction methods. And related theories are introduced as follows.

(a) RF

RF is known as decision tree. It can be used to do regression prediction or do classification. It is very useful to

make a feature selection and build an effective classifier, So use this feature to reduce the dimensions of the data in this study. Briefly review the principle of classification of RF. First of all, use Bootstrap Sampling Method (BSM) to take k samples from the original training set (assure each sample size is the same as the original training set). Then create a decision tree model for each sample and get k kinds of classification results. Finally, vote and choose an effective classification result basing on the classification results.

(b) PCA

PCA is a statistical process. And it is a classification system based on linear classification. It transforms the original n -dimensional data set into a new data set called the principal component by orthogonal transformation. In the result, the first principal component has the largest variance, and each subsequent principal component also has a large variance. In the PCA, the top m ($m < n$) of principal components can maintain the vast majority of information of the data. All in all, PCA uses the idea of linear fit to project high dimensional data onto several axes and reduces the dimensions of the data.

(c) Weighted-PCA

While using the PCA to reduce the dimensions of the data, several principal components are selected which can explain more data. But the contribution and importance of each factor in this model are often ignored. This situation may weaken the importance of the first principal component and then cause inaccurate clustering results. In order to verify this view, the clustering results under PCA and Weighted-PCA are compared.

Using Weighted-PCA to reduce the dimensions of the data has been applied to many researches. Meo et al. [15] proposed a new weighted principal component distance clustering analysis method by defining the weighted principal component distance. So according to the idea of Meo et al. [15], in this study, weighted eigenvalues are used as the classification statistics and objective weight is given to each principal component.

In theory, F_1, F_2, \dots, F_s ($s \leq p$) that is extracted from vector X is the column vector of principal component factors. $X = (x_1, x_2, \dots, x_p)$ is p -dimensional index vector. I_1, I_2, \dots, I_n is the sample row vector after extracting the principal component factor [15]. F_{ij} denotes the j th principal component factor of the i th sample ($i=1, 2, \dots, n; j=1, 2, \dots, s$) [15]. So in s -dimensional space, the data matrix formed by n samples is

$$F = (F_1, F_2, \dots, F_s) = \begin{pmatrix} I_1 \\ I_2 \\ \dots \\ I_n \end{pmatrix} = \begin{pmatrix} F_{11} & F_{12} & \dots & F_{1s} \\ F_{21} & F_{22} & \dots & F_{2s} \\ \dots & \dots & \dots & \dots \\ F_{n1} & F_{n2} & \dots & F_{ns} \end{pmatrix} \quad (3)$$

Assuming that we extract principal component factors F_1, F_2, \dots, F_s and the corresponding eigenvalues are $\lambda_1, \lambda_2, \dots, \lambda_s$.

$$\beta_k \equiv \lambda_k / \sum_{k=1}^s \lambda_k \quad (4)$$

is the corresponding feature weight of F_k , then $\beta_1 \geq \beta_2 \geq \dots \geq \beta_s$,

$$\sum_{k=1}^s \beta_k \equiv 1 \quad (5)$$

3. Data Preprocessing and Modeling

In this section, the source of the data set is introduced and the characteristics of the data are described in detail. At the same time, the specific process of data preprocessing is introduced. Finally, apply the relevant theoretical knowledge that is involved in the second section to establish the clustering analysis models which are under different dimensionality reduction methods.

3.1. Data Description

The data used in this study is from Zhejiang University of Finance and Economics. This data includes the basic information of teachers who are employed from 2011 to 2017, such as the number of papers, the number of awards, whether get the funds or not and so on.

The original data includes 245 instances and 101 attributes. And 46 attributes are discrete data and 55 attributes are continuous data. Because of the different types of the data, the data should be normalized. In addition, some of the attributes in this data set are same or similar. And some observations in the sample are missing values. These issues could affect the final result of the clustering analysis. So preprocess the original data before building the clustering analysis model.

3.2. Data Preprocessing

(a) Integrate or delete the same and similar data.

In the process, use a hierarchical weighting method to integrate the attributes of papers and awards. New attributes will be created after the integration. At the same time, the school level, expertise level and other attributes are quantified.

(b) Delete the missing value in the sample.

Because there are a large number of missing values in the sample, the missing value in the sample is deleted directly.

(c) According to actual needs, the attribute of the National Foundation of China (NFC) that is obtained in five years should be deleted and the data between 2011 and 2015 is also filtered out.

After the data preprocessing, selected 17 attributes to do clustering analysis such as the gender, age, undergraduate school level, the years of reading master and doctor, the total score of theses, the total score of the awards and so on.

3.3. Modeling Method

The purpose of data mining is to find some hidden rules or links in the huge data set. At present, generally two methods are used: regression and classification to do data mining. In this study, the clustering method is used to analyze the talent data and deeply mine effective information. Then make use of the

research result to provide effective guides. In order to improve the effect of clustering analysis and make the result of clustering analysis closer to the reality, we not only reduce the dimensions of the original data, but also use K-means and K-medoids algorithms to analyze. Thus, six kinds of clustering models are established, and specific introductions are as follows:

(a) The K-means algorithm under RF.

RF can calculate the importance of parameters. Considering the NFC within three years as a response variable and the rest of the attributes as dependent variables, R programming language is used to calculate the importance index of each attribute. Then select the first four important attributes and the response variable as the characteristic attributes of clustering analysis. Finally, use K-means algorithm to cluster.

(b) The K-medoids algorithm under RF.

Use the same dimension reduction method as (a) and use K-medoids algorithm to cluster.

(c) The K-means algorithm under PCA.

PCA is a perfect method to reduce the dimensions of the data. It can project high dimensional data into lower dimensions and use several principal components to explain the majority of data. Use the R programming language directly to do PCA. According to the gravel map, select the first three principal components as the characteristic attributes of clustering analysis. Finally, use K-means algorithm to cluster.

(d) The K-medoids algorithm under PCA.

Use the same dimension reduction method as (c) and use K-medoids algorithm to cluster.

(e) The K-means algorithm under Weighted-PCA.

The R programming language is used to calculate the eigenvalues of the principal components. Then the principal component factors are objectively weighted according to the eigenvalues. Consider the weighted principal component factor as the characteristic attributes of clustering analysis. Finally, use K-means algorithm to cluster.

(f) The K-medoids algorithm under Weighted-PCA.

Use the same dimension reduction method as (e) and use K-medoids algorithm to cluster.

Specific flow chart is as follows:

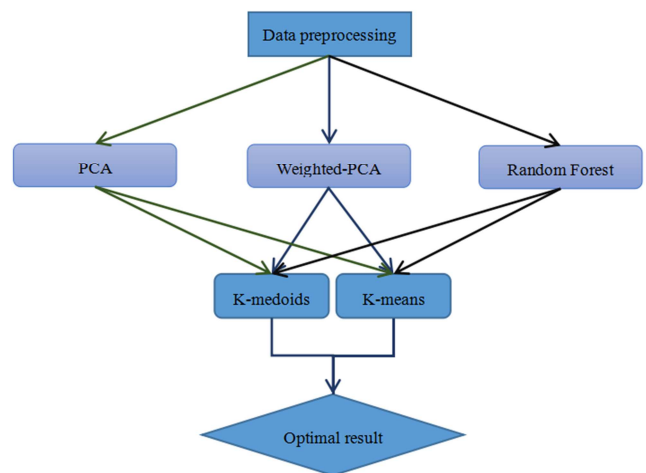


Figure 1. The process of modeling.

4. Experiment and Results

4.1. Dimension Reduction Experiment

In this experiment, in order to find the most reasonable method for clustering analysis, use the R programming language to perform clustering analysis with K-means and K-medoids algorithms after using RF, PCA and Weighted-PCA to reduce the dimensions of the data. The specific experimental operations and the results are as follows:

First of all, data preprocessing and remove missing values of the sample. Then, RF, PCA and Weighted-PCA are used to reduce the dimensions of the preprocessed data set. The results of dimension reduction are as follows:

Table 2. The results of RF.

Variable	Description	Importance Coefficient
age	The age of teachers	8.2428557
TScore	Total score of theses	7.8051357
AScore	Total score of awards	5.2795925
YRead	The years of reading master and doctor	3.8813314
YTeach	The years of teaching in colleges	3.4652020
TRank	The rank of professional technical	3.0067650

Table 3. The results of importance of components.

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6
Standard Deviation	33.2289	19.0036	7.39909	5.31613	2.81661	1.95484
Proportion Of Variance	0.70316	0.22998	0.03486	0.01799	0.00505	0.00243
Cumulative Proportion	0.70316	0.93315	0.96801	0.98601	0.99106	0.99349

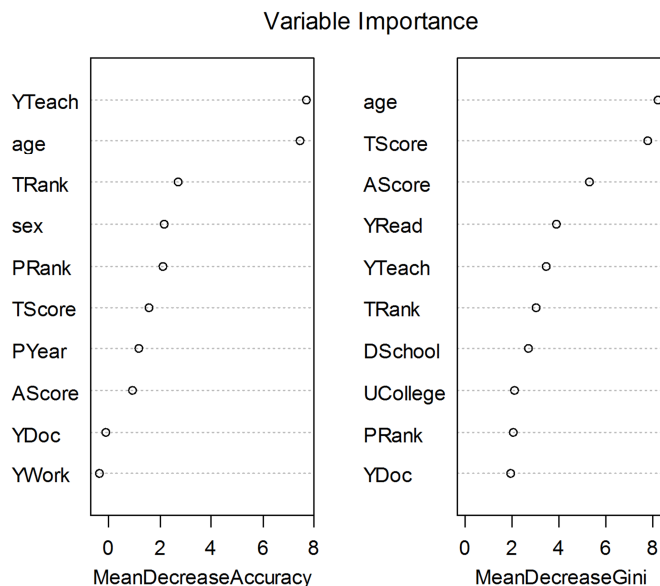


Figure 2. The output of the result of the importance of variables.

Variable	Description	Importance Coefficient
DSchool	Doctor schools	2.6753859
UCollege	Undergraduate colleges	2.1117811
PRank	The rank of the position	2.0295912
YDoc	The years of doctor reading abroad	1.9292135
YWork	The years of working in company	1.7073953
sex	The gender of teachers	1.5520465
PYear	Postdoctoral work years	0.7932718
DLab	Doctoral training team	0.5037148
YIns	The years of working in research institution before entering school	0.4999391
SJoint	The foreign school of joint training	0.3976019
NFC	Get the NFC in three years	

Table 2 and Figure 2 rank 16 attributes which are relative to the response variable. From the experimental results of RF, the first four variables that are related to the response variable are the total score of theses, the age, the total score of rewards and the years of reading master and doctor. Therefore, the sample data that covers these four variables serves as the data source of clustering analysis.

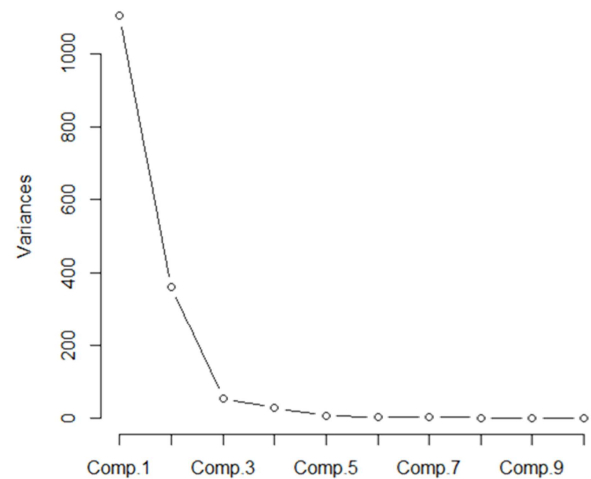


Figure 3. The output of the contribution of the eigenvalue.

In Table 3, the contribution rate of each factor as well as the degree of interpretation of the original data can be seen. In Figure 3, there is a turning point at the third factor in the gravel map. So select the first three principal components as the data set of clustering analysis. At the same time, the eigenvalues are used to weight the three factors. And it will be convenient for us to compare two methods.

4.2. Clustering Experiments

By reducing the attributes of the data set, choose five

variables and three variables from the RF and the PCA for clustering analysis. The experimental results are as follows:

Table 4. Contour coefficient of clustering result.

Dimension Reduction Method	Clustering Method	Contour Coefficient
Random Forest	K-means	0.2896609
	K-medoids	0.5172726
Principal Component Analysis	K-means	0.3893834
	K-medoids	0.5200169
Weighted Principal Component Analysis	K-means	0.4837657
	K-medoids	0.5996797

As it is shown in Table 4, the contour coefficients of the two algorithms under RF, PCA and weighted PCA are 0.2896609, 0.5172726, 0.3893834, 0.5200169, 0.4837657 and 0.5996797 respectively. Obviously, the K-medoids algorithm with Weighted-PCA has the largest contour

coefficient and the effect of clustering is the best. And the clustering result obtained through the RFM is the worst. Therefore, the K-medoids algorithm with Weighted-PCA is suitable for this case in this experiment.

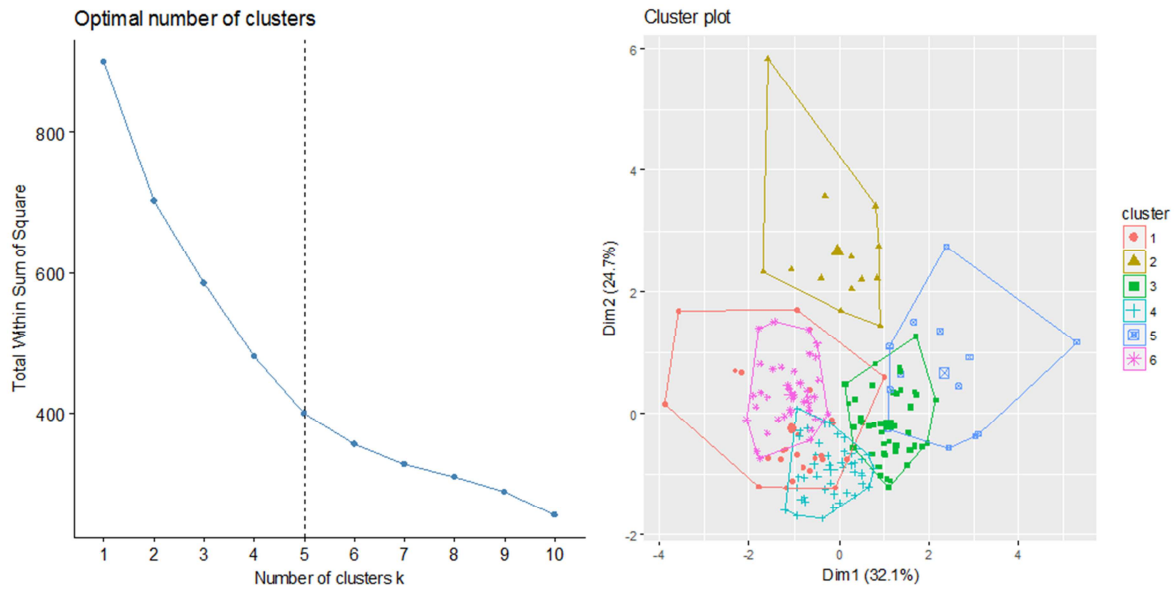


Figure 4. The K-means algorithm under RF.

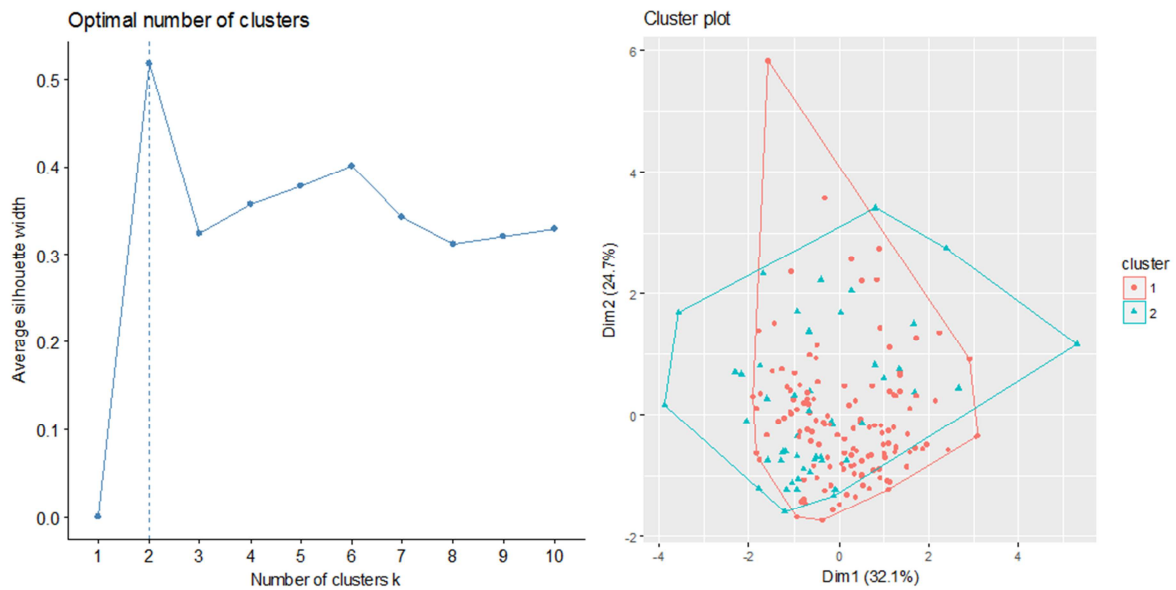


Figure 5. The K-medoids algorithm under RF.

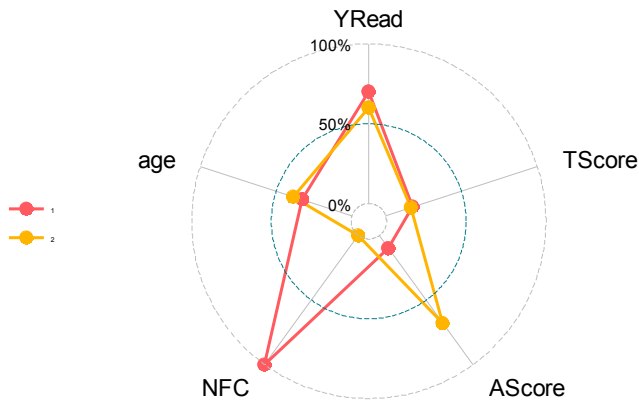


Figure 6. Radar chart of K-medoids algorithm under RF.

Figure 4, Figure 5 and Figure 6 respectively show the results of K-medoids and K-means algorithms under RF, and the general clustering effect can be seen. The RF radar chart demonstrates that the data is divided into two categories. One is with more the NFC, one is more awards.

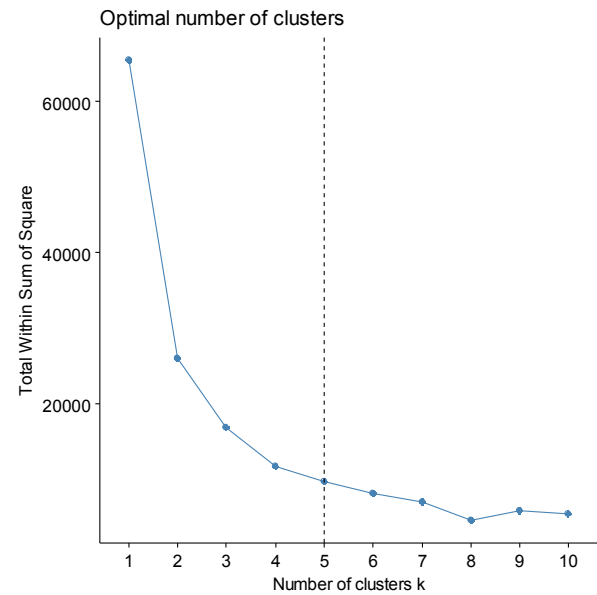


Figure 7. The K-means algorithm under Weighted-PCA.

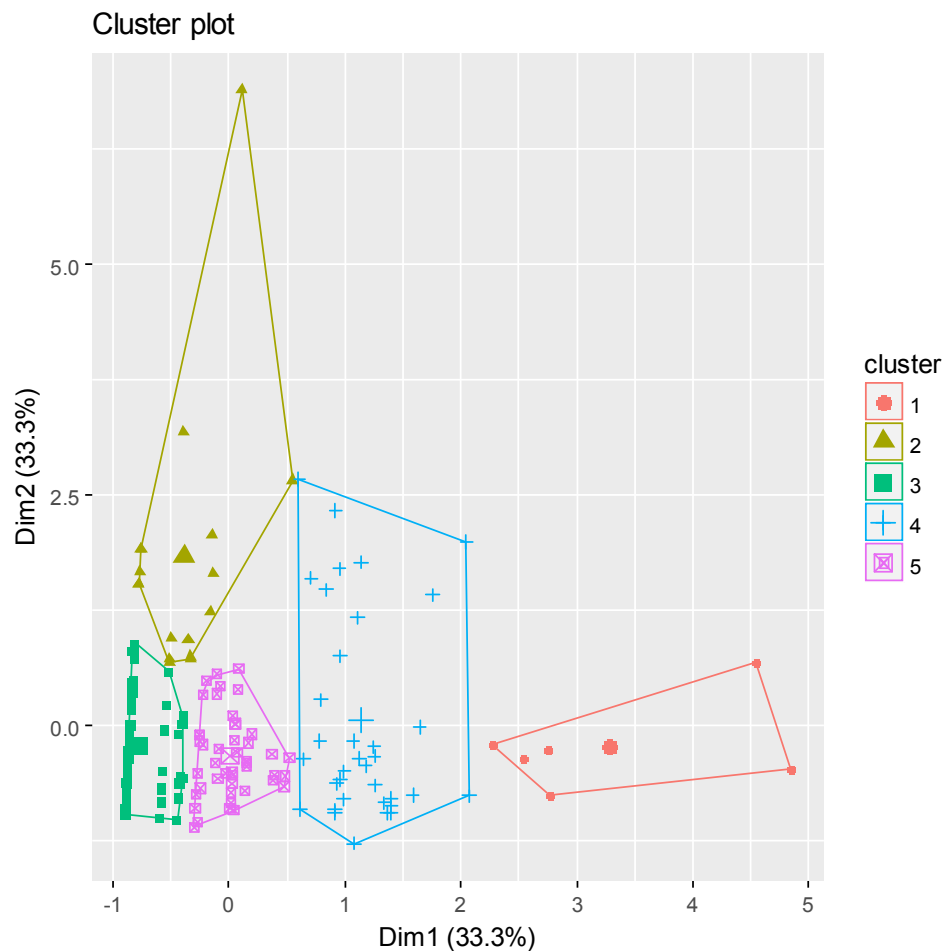


Figure 8. The K-means algorithm under Weighted-PCA.

Run the K-means algorithm under Weighted-PCA in R programming language and the output of the results is shown in Figure 7 and Figure 8. According to Figure 7, the best number of center points of clustering is five. So when clustering, the three principal components are clustered into five categories. Use R programming language to cluster five categories, and see the result in Figure 8.

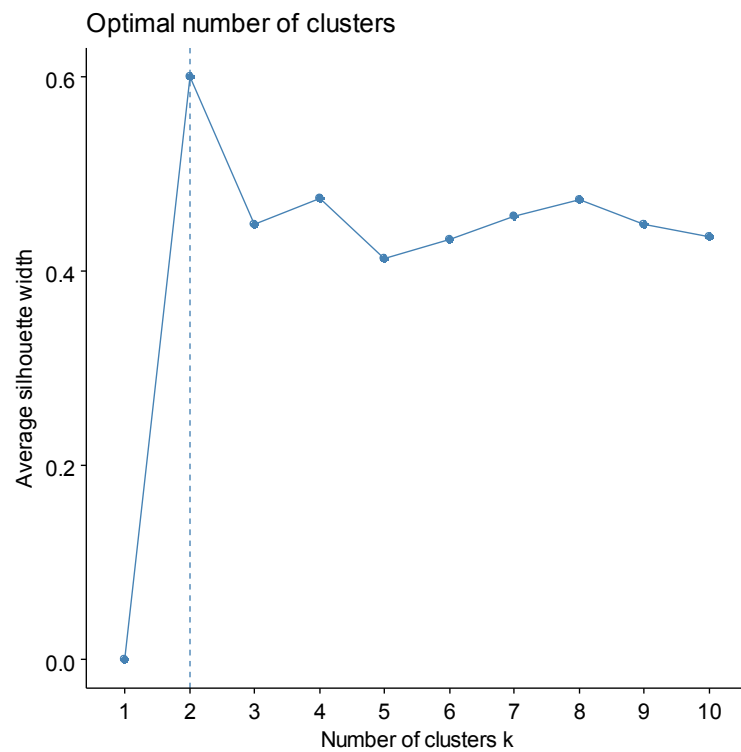


Figure 9. The K-medoids algorithm under Weighted-PCA.

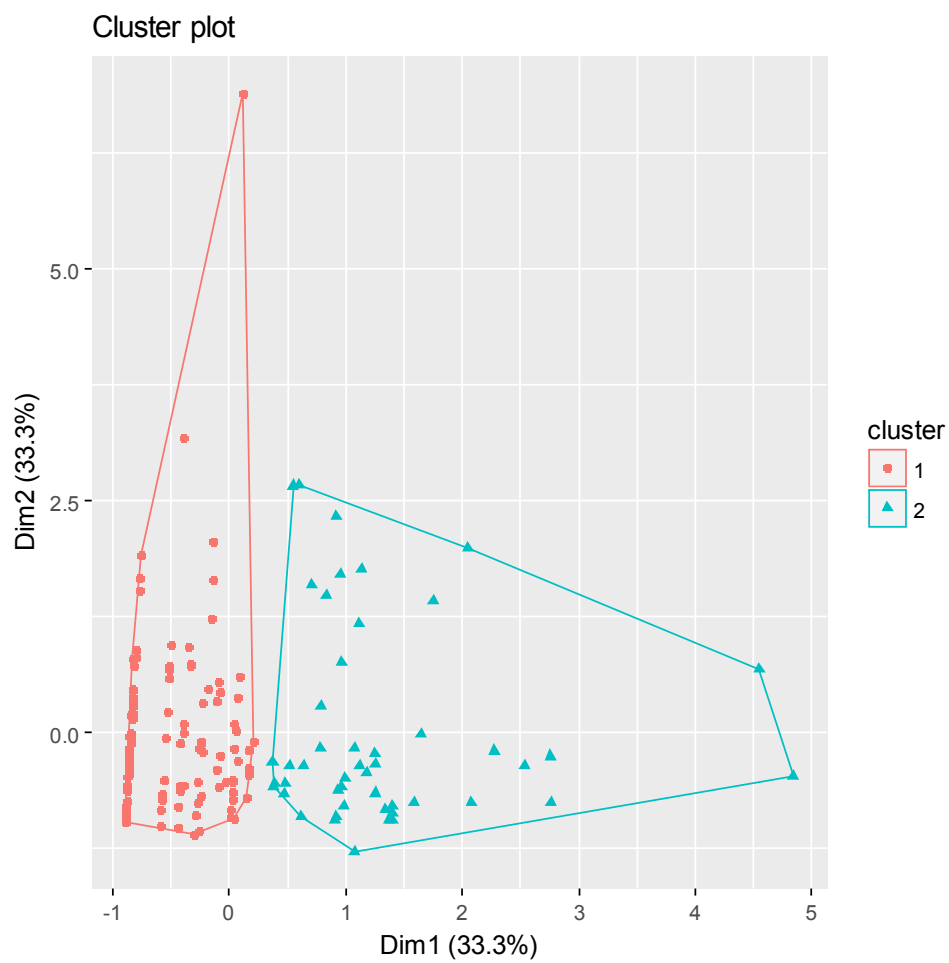


Figure 10. The K-medoids algorithm under Weighted-PCA.

Run the K-medoids algorithm under Weighted-PCA in R programming language and the output of the results is shown in Figure 9 and Figure 10. According to Figure 9, the best number of center points of clustering is two. So when clustering, the three principal components are clustered into two categories. The result is shown in Figure 10.

5. Conclusion

In this study, the K-means and K-medoids algorithms under RF, PCA and Weighted-PCA are compared in order to get a best clustering model. Then, some useful conclusions are drew. So divide the talents into high-end talent and mid-level talent at two levels. The high-end talent is the people who should be focused on, introduced and developed. And the mid-level talent requires us to focus on their development and to tap their potential as much as possible and to take appropriate measures to enhance their capabilities. So two types of talents are concluded, and the basic information is as follows:

Table 5. Clustering results.

Variable	age	YDoc	YTeach	TScore	AScore	NFC
Group1	36.23	1.88	1.73	18.63	11.44	0.25
Group2	36.66	3.2	0.42	23.9	71.64	0.34

In Table 5, the high-level talents who have published more papers, get more awards and have the longer years of studying in foreign high school should be focused on. And as for the people publishing little papers, getting little awards and having shorter years of studying in foreign high school, the introduction and post-cultivation in order to maximize their potential are vital. In addition, there is no difference in age between the two groups, so the consideration of age in introducing talents can be referred as appropriate.

Due to the limitation and the lack of the data set in this study, the classification of talents is too simple. Therefore, in the future work, more accurate and abundant methods and data are used to analyze. In addition, some hybrid algorithm can be used to further improve the accuracy of clustering.

References

- [1] H. Li, Y. R. Liu, X. W. Jia, "Soft power study on higher education talent cultivation mode of international trade," IBM, 2014, vol. 9, pp. 111-117.
- [2] D. Xu, "Study on the cultivating mode of undergraduate talents in tourism management: literature review, analysis and discussion," JSSM, 2015, vol. 8, pp. 496.
- [3] Y. Ge, C. C. Che, J. F. Liu, et al., "Discussion and practice of application oriented personnel training system in university," CSS, 2015, vol. 11, pp. 132-138.
- [4] M. P. Sendin-Hernández, C. Ávila-Zarza, C. Sanz, et al., "Cluster analysis identifies 3 phenotypes within allergic asthma," The Journal of Allergy and Clinical Immunology: In Practice, 2017.
- [5] L. M. Brophy, J. E. Reece, and F. McDermott, "A cluster analysis of people on community treatment orders in Victoria, Australia," Int J Law Psychiatry, 2006, vol. 29, pp. 469-481.
- [6] J. Eo, M. H. Kim, H. S. Bang, et al., "Effects of climate and landscape heterogeneity on the distribution of ground beetles (Coleoptera: Carabidae) in agricultural fields," J. ASIA. PAC. ENTOMOL, 2016, vol. 19, pp. 1009-1014.
- [7] K. Adameczyk, D. Cywicka, P. Herbut, et al., "The application of cluster analysis methods in assessment of daily physical activity of dairy cows milked in the voluntary milking system," COMPUT ELECTRON AGR, 2017, vol. 141, pp. 65-72.
- [8] N. Funtikova, A. A. Benítez-Arciniega, M. Fitó, et al., "Modest validity and fair reproducibility of dietary patterns derived by cluster analysis," NUTR RES, 2015, vol. 35, pp. 265-268.
- [9] J. Halčinová, P. Trebuňa, I. Janeková, et al., "The proposal of stock items reconfiguration on the basis of cluster analysis results," Procedia Eng, 2014, vol. 96, pp. 143-147.
- [10] Kijewska and A. Bluszcz, "Research of varying levels of greenhouse gas emissions in European countries using the k-means method," ATMOS POLLUT RES, 2016, vol. 7, pp. 935-944.
- [11] L. H. Chen, Z. S. Xu, H. Wang, et al., "An ordered clustering algorithm based on K-means and the PROMETHEE method," Int. J. Mach. Learn. & Cyber, 2016, 1-10.
- [12] Z. Fei and K. Liu, "Online process monitoring for complex systems with dynamic weighted principal component analysis," Chin. J. Chem. Eng, 2016, vol. 24, pp. 775-786.
- [13] P. Rodríguez, R. Navarro, and J. J. Rozema, "Eigencones: application of principal component analysis to corneal topography," OPTHAL PHYSI OPT, 2014, vol. 34, pp. 667-677.
- [14] N. Goudarzi, D. Shahsavani, F. Emadi-Gandaghi, et al., "Quantitative structure-property relationships of retention indices of some sulfur organic compounds using random forest technique as a variable selection and modeling method," J. Sep. Sci, 2016, vol. 39, pp. 3835-3842.
- [15] M. Meo, V. Zarzoso, O. Meste, et al., "Catheter ablation outcome prediction in persistent atrial fibrillation using weighted principal component analysis," BIOMED SIGNAL PROCES, 2013, vol. 8, pp. 958-968.