International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 6 Issue 6, September-October 2022 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Analysis of Knowledge Points in the National College Students Mathematics Competition (Non-Professional Group) Based on PCA

He Xujie, Hong Hairu, Li Tingyu, Yang Chen

School of Information, Beijing Wuzi University, Beijing, China

ABSTRACT

The National Mathematical Competition for Undergraduates is a national high-level subject competition for undergraduates. In this paper, 82 knowledge points have been sorted out according to the competition content of the China College Students Mathematics Competition (non-mathematics category) and the real questions of the previous competition preliminaries. The evaluation index system of the importance of knowledge points is obtained by the evaluation matrix, and the principal component analysis (PCA) algorithm is used to establish a model, and statistical analysis is used to obtain the knowledge points with high contribution rate and comprehensive score. The knowledge points to be tested include the calculation of triple integral and double integral, the calculation and application of surface integral, and functions. The knowledge points that will appear with high probability include the geometric meaning of differential, the known binary differential to find the original function, and the properties of power series, Power series to find the limit, effectively helping non-mathematical college students to prepare for exams.

KEYWORDS: *knowledge point prediction; mathematics; principal component analysis; MATLAB*

How to cite this paper: He Xujie | Hong Hairu | Li Tingyu | Yang Chen "Analysis of Knowledge Points in the National College Students Mathematics Competition (Non-Professional Group) Based on PCA" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-6 | Issue-6, October 2022, pp.262-269, URL: www.ijtsrd.com/papers/ijtsrd51813.pdf

IJTSRD51813

INTRODUCTION

(1) Research background: In 2009, the Chinese Mathematical Society and the National University of Defense Science and Technology jointly organized the first Chinese College Students Mathematics Competition (commonly known as the "National College Students Mathematics Competition"). Since then, CMC has been hosted by major universities in China every year. The purpose of the competition is to stimulate the interest of college students in learning mathematics, to discover and select innovative talents in mathematics, to further promote the reform and construction of mathematics courses in colleges and universities, and to improve the teaching level of university mathematics courses. However, the National Undergraduate Mathematics Competition is difficult and has many knowledge points. Many contestants blindly review many knowledge points and fail. Therefore, in order to allow the contestants to have a clear goal when reviewing, we decided to use mathematics and machine learning methods for mathematics competition. analysis of knowledge points,

(2) Problem restatement and analysis: This article analyzes the knowledge points involved in the real questions (non-mathematics) of previous mathematics competitions. First, it summarizes the knowledge points of all the topics in 2009-2021, and then predicts the knowledge points that may be involved in the second year exam according to the summary content.

- Step 1: Summarize all the knowledge points involved in the title according to the real questions of previous competitions and the competition content of the Chinese College Students Mathematics Competition (non-mathematics major), and classify the fragmentary knowledge points in a regular manner.
- Step 2: According to the different proportions of the average score, the classified knowledge point

grades are divided, and an evaluation index system for the importance of knowledge points is established.

- Step 3: Use the Principal Component Analysis (PCA) algorithm to comprehensively evaluate the indicators, establish a model to obtain the contribution rate and score of each knowledge point, and calculate the knowledge points that will be involved in the next year's exam.

Mathematical Analysis

In this paper, a mathematical prediction model based on principal component analysis is established, and multiple indicators are recombined into a few indicators to replace the original indicators, and new information is further extracted [1].

A. Data Information

The data comes from a total of 13 sets of preliminary examination questions for the 1st to 13th Mathematical Contests for Undergraduates (non-mathematics) and the content of the Chinese Mathematical Contests for Undergraduates (non-mathematics majors). Based on the premise of mathematical statistics and regular classification of the knowledge points involved in each year, the frequency and score of each knowledge point in each year are taken as the breakthrough point, and the evaluation index system of the importance of knowledge points is established, and the grades are divided. The evaluation matrix of 82 knowledge points and the 19 knowledge point evaluation matrix based on the competition outline are obtained, which are used for subsequent accurate prediction of knowledge points, and then the comprehensive score and contribution rate of each knowledge point are obtained by principal component analysis. The realization process is handled with MATLAB software.

B. Data Description

- Since the 84 knowledge points are too small, this paper classifies the 84 knowledge points into 19 knowledge points for mathematical analysis according to the content of the competition. "A point of knowledge. In this way, a fairly high accuracy can be achieved in the prediction of knowledge points in the later stage, and the error can be reduced.
- Calculate the average score (total score/frequency of occurrence) of each knowledge point every year, in which the total score changes according to the location of each knowledge, and the score of the whole question replaces a single knowledge point 's score.
- Count the average score of each knowledge point for partition rating. Based on the frequency and score of knowledge points, it is divided into four levels: 1, 3, 5, and 7. The evaluation index system of the importance of knowledge points is established, and the importance level of each knowledge point is evaluated. The index level table is as follows,

(TABLE 2-1)

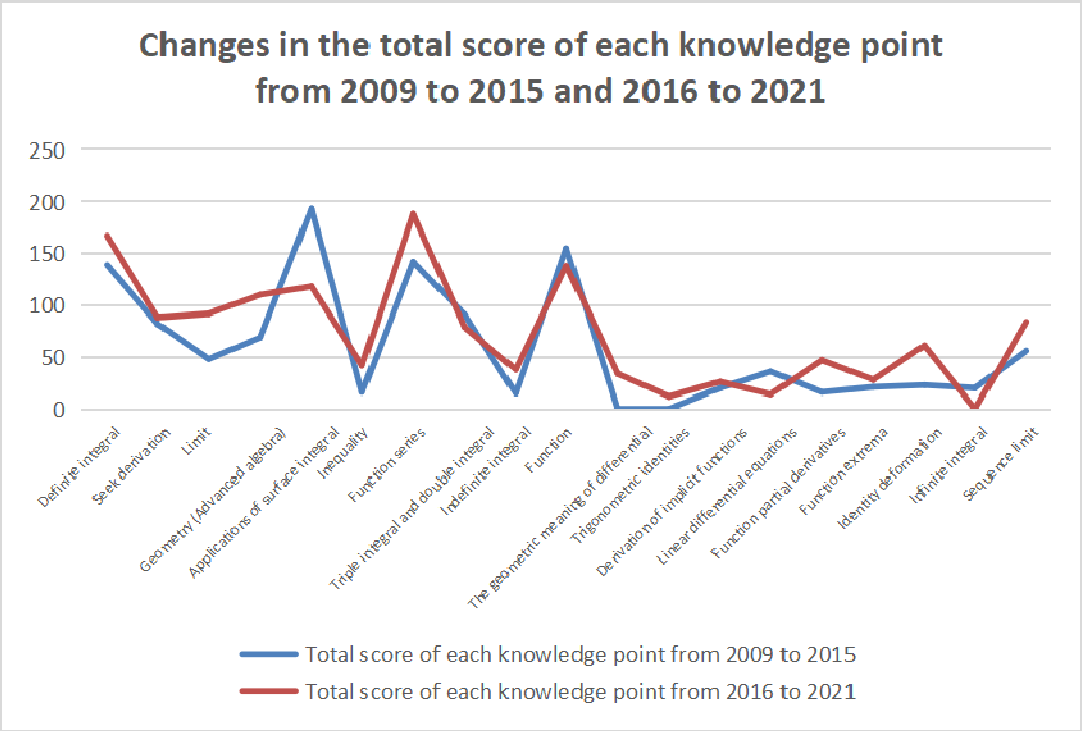
Index level	Importance
7	Very important
5	Important
3	More important
1	Generally important

If the index level is "7", the knowledge point has a high score and a high probability of occurrence, which is more important, such as the calculation of double integral; if the index level is "5", the knowledge point has a high score or a high probability of occurrence. large, such as the geometric meaning of differential; if the index level is "3", the knowledge point has a high probability of occurrence, but the score is not too high, such as the sum of the series; if the index level is "1", the knowledge point appears The lower value is generally a simpler knowledge point, such as equivalent infinitesimal substitution.

The 1357 level distribution map can intuitively see that there are many knowledge points belonging to level 1 and level 7 in the past 13 years.

C. Data Analysis

The more important knowledge points cannot be judged intuitively from the level distribution and the total score of each knowledge point. Therefore, comparing the knowledge points involved in the competition before 2015 with the last 6 years, Figure 2-1 is obtained. From Figure 2-1, it can be seen that the scores of knowledge points such as definite integral, limit, geometry (advanced algebra), function series, and identity deformation in the past six years have increased compared with that before 2015, while the application of surface integral is relatively high. Before 2015, it was lower than the total score.



(FIGURE 2-1: Changes in the Total Score of Each Knowledge Point from 2009 to 2015 and 2016 to 2021)

After analyzing the total score of knowledge points in the past six years and the first seven years, in order to more intuitively understand the distribution of test points in the preliminaries of the 11th mathematics competition (non-mathematics), the analysis was carried out according to the proportion of the total score. It can be seen from the figure that the function series (13%) has accounted for the largest share in the past 13 years, followed by functions, the application of surface integrals, and definite integrals, all accounting for 12%. Among them, trigonometric identities account for the smallest proportion.

D. Mathematical analysis results

According to data statistics, in recent years, mathematics competition topics have paid more attention to the use of knowledge points, similar to the use of Lagrange's mean value theorem, and the construction of a reasonable form of identity transformation. Candidates are required to deepen their understanding of knowledge points, and to gradually increase the proportion of inspections on definitions and properties. With the continuous development of test questions, the confusion between different test sites has increased, the integration of test sites in the questions is more intricate, and the problem-solving ideas are gradually novel and ingenious. Quickly locate the problem-solving ideas, rather than blindly trying all the ideas, which is a waste of time.

Analysis of Algorithms

A. Model preparation

According to the principles of the evaluation index system for the importance of knowledge points, we conducted an index evaluation on 82 knowledge points from 2009 to 2021, and obtained an evaluation matrix with 1357 levels. Since the 84 knowledge points are too small, the 19 knowledge points extracted from the syllabus of the National Undergraduate Mathematics Competition in this paper also establish an evaluation index matrix.

B. Data preprocessing

Z-Score standardization is a common method of data standardization. It can convert data of different magnitudes into a unified measure and measure how many standard deviations the original data differs from the overall mean of the data [2], so this group adopts the Z-Score method. To standardize the data, the standardized matrix is obtained by the following formula.

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad \text{(formula3-1)}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(formula3-2)}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{(formula3-3)}$$

C. Model establishment

Principal component analysis is a multivariate statistical analysis method that efficiently handles multidimensional data. It transforms a set of possibly correlated variables into a set of uncorrelated variables through orthogonal transformation. The basic idea is to generate a series of uncorrelated new variables through the analysis and calculation of the internal structural relationship of the original data correlation matrix. According to the need to select several new variables less than the original variables, that is, to generate principal components [3]. Therefore, the principal component analysis method is essentially a dimensionality reduction method, and this set of data contains 82 knowledge points, which happens to be high-dimensional data. In order to obtain important knowledge points, dimensionality reduction processing must be performed. It can be seen that the principal component analysis method suitable for this set of data.

- Find the Correlation Coefficient Matrix

Let this group of data be an n-dimensional variable of $(X_1, X_2, X_3, \cdots, X_n)$, and if any correlation coefficient $\rho_{ij}(i, j = 1, 2, \cdots n)$ exists, it is a correlation coefficient matrix, and the formula ρ_{ij} is as follows

$$\rho_{ij} = \frac{cov(x_i, x_j)}{\sqrt{Dx_i}\sqrt{Dx_j}}, cov(x_i, x_j) = E((x_i - E(x_i)) \cdot (x_j - E(x_j))) \quad \text{(formula3-4)}$$

- Find Eigenvalues and Eigenvectors of Correlation Coefficient Matrix

Let this group of data be an n-dimensional variable of $(X_1, X_2, X_3, \cdots, X_n)$, if there are numbers λ and non-zero vectors x such that $Ax = \lambda x, x \neq 0$, x is called an eigenvalue of A, and is the eigenvector of the corresponding eigenvalue of A.

- get the principal components

The principal component ρ_{ij} is obtained by the linear combination of the unit eigenvector of the covariance matrix and the original vector, namely

$$Y = PX = \begin{bmatrix} P_1 \\ \cdots \\ P_n \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ \cdots \\ X_n \end{bmatrix} = \begin{bmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nn} \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ \cdots \\ X_n \end{bmatrix} = \begin{bmatrix} e_{11}X_1 & \cdots & e_{1n}X_n \\ \vdots & \ddots & \vdots \\ e_{n1}X_1 & \cdots & e_{nn}X_n \end{bmatrix}$$

Among them, P_1 is formed by transposing the eigenvector e_1 of the largest eigenvalue λ_1 of the covariance matrix ρ_{ij}, then the first principal component is: $Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1n}X_n$. Similarly, the expression of the kth principal component is $Y_k = e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kn}X_n$.

- Calculate contribution rate

The variance of the principal components can determine the variance of the data set, that is, the eigenvalues λ of the covariance matrix ρ_{ij} need to be used, then the variance λ_k of the kth principal component is, and the variance contribution rate Y_k of the principal component is

$$Y_k = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_n} \quad \text{(formula3-5)}$$

The cumulative contribution rate of variance of the first k principal components is

$$D = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_n} \quad \text{(formula3-6)}$$

- Calculate principal component scores

The principal component score is obtained by adding the score of each component and the weight score, that is, the feature vector of the extracted principal component is multiplied by the normalized original data, and then added row by row.

D. Algorithm implementation

- Data normalization

  Standardize the Z-Score data of the original matrix in MATLAB software. Since the row and column elements of the matrix are opposite, the original evaluation table needs to be transposed before normalization.

- Calculate the coefficient matrix, find the eigenvalues and eigenvectors of the coefficient matrix.
- Calculate contribution rate
- Calculate principal component scores

The specific 82 knowledge point principal component analysis codes are shown in Appendix, and the 19 knowledge point principal component analysis codes are shown in Appendix.

E. Model results

- 82 knowledge points scoring matrix analysis results

First, use Excel to implement the transposition operation to standardize the data, and then calculate and obtain thirteen principal components. The eigenvalues, contribution rates and cumulative contribution rates of the thirteen principal components are

(TABLE 3-1)

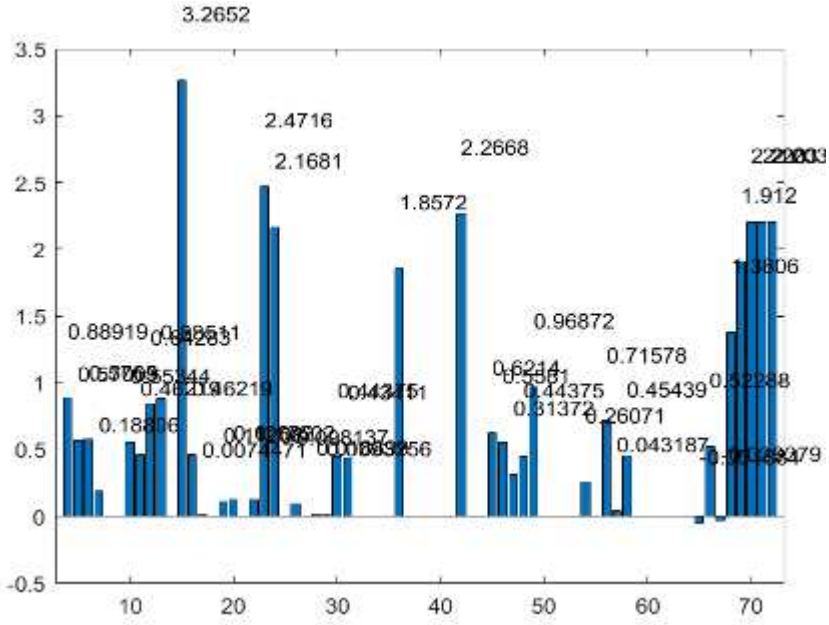
Main ingredient	Eigenvalues	Contribution rate	Cumulative contribution rate
1	1.98	0.15	0.15
2	1.61	0.12	0.28
3	1.48	0.11	0.39
4	1.29	0.10	0.49
5	1.16	0.09	0.58
6	1.05	0.08	0.66
7	0.86	0.07	0.73
8	0.78	0.06	0.79
9	0.71	0.05	0.84
10	0.65	0.05	0.89
11	0.56	0.04	0.93
12	0.46	0.04	0.97
13	0.41	0.03	1.00

When the information retention rate is 0.9, the number of principal components is 11, and the comprehensive score and ranking of each knowledge point are finally obtained. The scores of the top ten knowledge points are as follows:

(TABLE 3-2)

Ranking	Principal component score	Knowledge point
1	3.265152	Monotonicity of functions
2	2.471554	Lagrange's mean value theorem
3	2.266804	Calculation of triple integral
4	2.202964	The geometric meaning of differential
5	2.202964	Properties of Power Series
6	2.202964	Power series to find the limit
7	2.168107	Properties of Surface Integrals
8	1.911999	Parametric Equations of Spatial Curves
9	1.857208	Calculation of Surface Integral
10	1.380553	Knowing the total differential of a binary function to find the original function

Visualize the principal component scores of the top 40 important knowledge points, and the results are as follows



(FIGURE 3-1)

➢ 19 knowledge points scoring matrix analysis results

First, standardize the data of the transposed evaluation matrix, and then obtain 13 principal components. The eigenvalue, contribution rate and cumulative contribution rate of the principal components are

(TABLE 3-3)

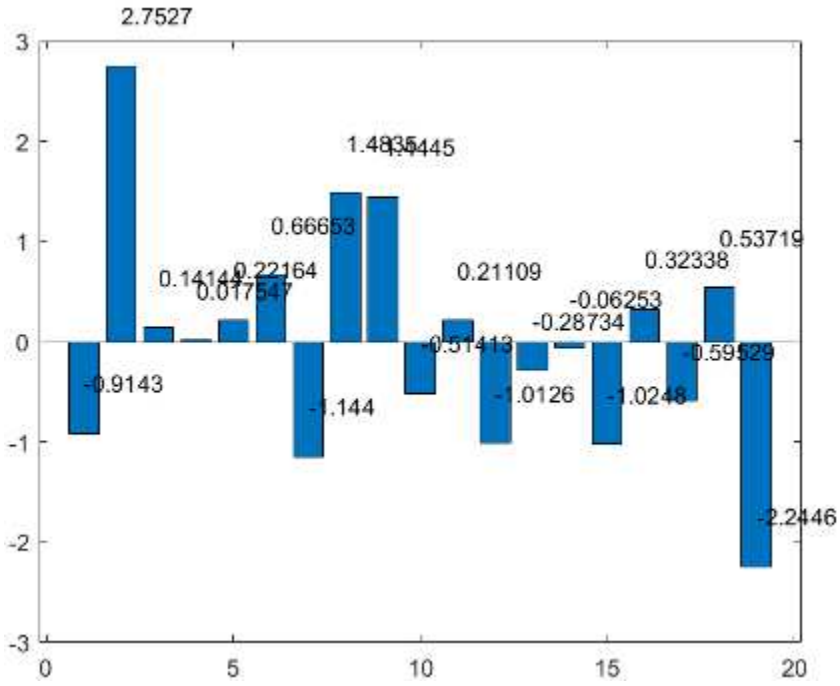
Main ingredient	Eigenvalues	Contribution rate	Cumulative contribution rate
1	4.72	0.36	0.36
2	1.72	0.13	0.50
3	1.67	0.13	0.62
4	1.27	0.10	0.72
5	0.95	0.07	0.79
6	0.82	0.06	0.86
7	0.56	0.04	0.90
8	0.48	0.04	0.94
9	0.38	0.03	0.97
10	0.26	0.02	0.99
11	0.11	0.01	1.00
12	0.04	0.00	1.00
13	0.02	0.00	1.00

When the information retention rate is 0.9, the number of principal components is 7, and the principal component scores and rankings of 19 knowledge points are finally obtained. The scores of the top ten knowledge points are as follows

(TABLE 3-4)

Ranking	Principal component score	Knowledge point
1	2.752697	seek derivation
2	1.483501	Triple Integral and Double Integral
3	1.444518	indefinite integral
4	0.666528	inequality
5	0.537194	infinite integral
6	0.323381	extreme value of the function
7	0.221645	Applications of Surface Integral
8	0.211089	The geometric meaning of differential
9	0.14144	limit
10	0.017547	Analytic Geometry of Space (Advanced Algebra)

Visually analyze the principal component scores of the 19 knowledge points, and the results are as follows



(FIGURE 3-2)

Analysis and summary of results

In this paper, principal component analysis is carried out on 82 knowledge points that have appeared frequently in previous mathematics competitions. The highest principal component scores are the monotonicity of functions and the Lagrange median theorem. It can be seen that these two knowledge points are in the title. The probability of occurrence is high; secondly, the principal component scores of the four knowledge points, namely the calculation of triple integral, the geometric meaning of differential, the calculation of surface integral, and the total differential of known binary functions to find the original function, are also relatively high, and it can be seen that integral and differential appear. The probability is very high; the two knowledge points of the nature of the power series and the limit of the power series occupy the fifth and sixth in the ranking of the principal component scores. It can be seen that these two knowledge points play an important role and have a high probability of occurrence; the surface integral The properties of , the parametric equations of space curves, and the calculation of surface integrals rank seventh, eighth, and ninth in the ranking of principal component scores. It can be seen that spatial analytic geometry also occupies an important part in the competition, and most of them appear in the same big topic. middle.

Principal component analysis is performed on 19 knowledge points. The knowledge point with the highest principal component score is the derivation. It can be seen that although this knowledge point is relatively simple, the frequency of occurrence is very high, and the relative transformation is difficult. ; Secondly, the geometric meanings of triple integral and double integral, indefinite integral, infinite integral, and differential also rank higher in the ranking of principal component scores; inequalities also occupy a higher ranking in the ranking of principal component scores, so the mastery of inequality-related knowledge It is also very important; the application of surface integral and the principal component score of spatial analytic geometry are also high. This knowledge point does not appear often, but it will appear in a major topic. It can be seen that the score is not low, and it is an important knowledge point.

Combining the results of these two analyses, the calculation and application of triple integrals account for a high proportion. This part of the knowledge points is an important knowledge point. The calculation of triple integrals is based on double integrals, so you should keep the formula in mind and do more In terms of question types, it is flexible to face various alternative questions; secondly, functions account for a large proportion of knowledge points, such as monotonicity of functions, extreme values of functions, etc. Although the score is not large, it is the solution to the following. The question has paved the way and is more important and should not be ignored; spatial analytic geometry is also a key knowledge point, including the properties, calculation and application of surface integrals, parametric equations of spatial curves, etc., which account for a higher score; power series in Competition content also accounts for a large proportion.

To sum up, through mathematical analysis and principal component analysis, the analysis results of the two methods are compared, and it is found that

the knowledge points of the competition are mostly concentrated in the content of medium difficulty. Among them, the knowledge points that must be tested every year are: the calculation of triple integral and double integral, the calculation and application of surface integral, and functions. Including the application and calculation of definite integrals, functions including function limits, function series, etc., and the forms of knowledge points are changeable, so the review of relevant knowledge points can be carried out at the same time during review, so that you can comprehensively review important knowledge points, and It can save review time. Contestants not only need to have a solid foundation, but also have flexible thinking ability. Only by doing more questions and cultivating multiple ways of thinking can they have ideas when seeing new question types. Secondly, the knowledge points that are likely to appear are: the geometric meaning of differentiation, the known binary differentiation to find the original function, the properties of power series, and the limit of power series. These knowledge points account for a large number of points, and the question type changes. There are many questions, and the questions are relatively flexible. First of all, you must lay a solid foundation and have a solid grasp of basic knowledge points such as the nature of power series. At the same time, you can get familiar with various question types by doing more questions. The geometric meaning and the function of finding the original by the total differential of known binary functions account for a large proportion. It can be seen that these two knowledge points appear frequently and have various forms. Contestants can repeat these two contents and prepare for full review. Do relevant questions and improve your familiarity with knowledge points. Finally, the basic knowledge points should not be underestimated, such as function derivation, integral calculation, inequality and other knowledge points, although they are relatively simple, but they are the basis for later problem solving, so the review of basic knowledge is very important, and it is essential to consolidate the basic links.

References

[1] Zhao Haixia, Wu Jian. Analysis of principal component analysis method [J]. Science and Technology Information, 2009, (2): 87.

[2] Wang Zhengpeng, Xie Zhipeng, Qiu Peichao. Comparison of Data Standardization Methods in Semantic Relationship Similarity Calculation [J]. Computer Engineering, 2012.

[3] Zhou Zhihua. "Machine Learning". Beijing: Tsinghua University Press, 2016.