RESEARCH ARTICLE

Early-warning model of excess consumption in college students by KPAG method during the COVID-19 pandemic

Zitan Shi[1], Bin Zhao [2]*

Zitan Shi, Bin Zhao. Early-warning model of excess consumption in college students by KPAG method during the COVID-19 pandemic. Clin Pharmacol Toxicol Res.2021;4(3):1-5.

With the arrival of COVID-19, some areas are under closed management, bringing about changes in the way people consume. It also leads to the excessive consumption of some people, especially college students. In order to give early warning to unreasonable consumption behavior, this study designed KPAG algorithm to give early warning to consumption risk. Using Particle Swarm Optimization (PSO) Kernel Principal Component Analysis (KPCA) parameter optimization, optimal polynomial kernel to delete data information, and ant colony genetic algorithm (association) clustering analysis of data dimensionality reduction, according to the consumption behavior of college students are divided into three categories, for the consumption behavior of college students to build an early warning model. Through the classification and verification experiment of real data, the results show that compared with the traditional PCA data fitting method, the accuracy of the model in this paper can reach 90%, which is more reliable than the traditional algorithm, and the accuracy of the model is improved by nearly 20%, which can be used for effective early warning.

Key Words: *Data science in education; Consumption risk; KPAG algorithm; Early-warning model*

INTRODUCTION

With the global outbreak of COVID-19 in 2019, the way people shop has partly changed. College students, as a special group of consumer groups, have the characteristics of low economic burden and high consumption power, which should be paid more attention to. At present, studies have proved that college students are more inclined to consume online than ordinary consumers and online shopping festivals are an important factor affecting consumer behavior [1,2]. Moreover, with the closed management brought by the epidemic, the trend of college students' online consumption and unhealthy consumption has also increased. Bollen Zoé conducted a survey among Belgian college students and found that some college students consumed more alcohol during the closed management period, which required the government to pay attention and give early warning [3]. Therefore, it is very necessary to evaluate the consumption behavior of college students and give early warning to unreasonable behaviors in the period of closed management, so as to prevent a series of campus risks such as excessive campus loans. In terms of research methods, many scholars used to establish logistic regression model to study the main factors affecting consumer consumption [4].

However, there is a high correlation between the related factors affecting college students' consumption behavior, and in most cases, it is impossible to describe and analyze college students' online consumption directly by using the original measurement indicators. Among many current research methods, PCA elimination is a common method to eliminate multicollinearity [5]. In view of the disadvantage that PCA cannot deal with nonlinear problems, some scholars have found that introducing kernel function for processing can obtain higher precision processing effect [6]. Some studies have established an excellent economic management model based on KPCA to solve the problem of information entanglement between data [7,8]. Nadia Souilem designed an appropriate pre-filter for the optimization of the core function of KPCA algorithm [9]. However, all the above studies only processed the data and lacked the continuous evaluation and classification of the corresponding processed objects. In addition, KPCA still has the disadvantages of setting parameters and difficulty in evaluating and classifying indexes. Ag Abo Khalil found that the optimization algorithm could obtain better results in parameter optimization of kernel function [10]. Zhao Min designed a cultural particle swarm optimization algorithm [11].

Rongyi used particle swarm optimization to select kernel function parameters [12]. The above research provides ideas for parameter optimization in this paper. The establishment of academic early warning model based on KPCA and fault detection model also provides an important reference for the establishment of cluster early warning model in this study [13-16]. In this study, combined with the research content, the traditional PCA-linear regression is carried out on the consumption of evaluation model of optimization: three times by introducing kernel function solve the problem of nonlinear data processing, the introduction of optimization algorithm for the optimal kernel parameters optimization solve the defect of the parameters need to be set in combination with clustering algorithm to process the data and implement evaluation and classification of the early warning function.

METHODS

KPAG-method

Aiming at the problems of the traditional consumption warning model based on PA-Logistics, such as single warning direction, poor data dimension reduction effect, multiple fitting factors and low accuracy, this paper creatively designed the KPAG method. First of all, referred to the dimensionality reduction data processing idea of the kernel method. The sample data set was established, default kernel parameters were set, and the sample set was mapped to a higher dimension by the kernel function. First reference nuclear dimension reduction of the data processing methods, we established X sample data set and design for the default nuclear σ, then using the kernel function K to higher dimensional mapping of sample set, $\Phi(x_j)$ sample data to feature space is obtained, then the principal component is obtained by principal component analysis and the sample on the principal component characteristic vector projection $\overline{\mu_i}$ finally calculated feature space sample data within the class of discrete degree of $\overline{S_w}$ and discrete degree J's difference value between classes. The default kernel parameters are judged according to the characteristic accumulative value of the first three principal components. If the requirements are not met, the objective function J will be established, and the difference between S_w and S_b will be taken as the fitness function $J(\sigma)$, and the particle swarm optimization algorithm will be used to iteratively solve the optimal nuclear parameters σ. KPCA processing is carried out with σ. Finally, by referring to the idea of ant colony clustering algorithm, the early warning objective function $J(w,c)$ is established through characteristic parameters, the uniform two-point crossover operator is introduced to update the information matrix, and the early warning result

[1]*Department of Mechanical Engineering, Hubei University of Technology, Wuhan, P.R.China*
[2]*Department of Mechanical Science, Hubei University of Technology, Wuhan, P.R.China*

Correspondence: Bin Zhao, Department of Mechanical Science, Hubei University of Technology, Wuhan, P. R. China, Tel/Fax: +86 130 2851 7572, E-mail: zhaobin835@nwsuaf.edu.cn

Received: April 13, 2021; Accepted: April 27, 2021; Published: Maty 02, 2021

is obtained through iterative calculation. The algorithm flow chart is shown in Figure 1.

Modeling and derivation

In the derivation, the original consumption data set is represented by X, and x_i 98 × 99 is the K-dimensional consumption impact factor. Remember that the space of $K \times N$ dimension vector is the input space, and the nonlinear mapping Φ is used to map the data sample $X = \{x_1, x_2, ...x_3\}$ to the characteristic space H. Then, the high-dimensional consumption data set $\Phi(\mathrm{x}) = \{\phi_1, \phi_2, ...\phi_3\}$ is obtained. By default, the high-dimensional consumption data set in the feature space has completed centralized processing, satisfying $\Phi(\mathrm{x}_i)\Phi(\mathrm{x}_i)^T = 0$. By PCA analysis of the high-dimensional consumption factor, the covariance matrix C can be calculated, where C satisfies relation $\lambda V = CV$. After diagonalizing C, the consumption eigenvalue λ of C and the corresponding eigenvector $V \in H$ are obtained. Since $V \in$ span $\{\Phi(\mathrm{x}_1), \Phi(\mathrm{x}_2), ...\Phi(\mathrm{x}_i)\}$, it can be known that a group of coefficients $\alpha_1, \alpha_2, ...\alpha_n$ have the following relationship:

$$V = \sum_{k=1}^{n} \alpha_j \Phi(\mathrm{x}_j) \tag{1}$$

For the defined kernel function,

$K : K_{ij} = (\Phi(\mathrm{x}_i) \cdot \Phi(\mathrm{x}_j)) = \mathrm{K}(\mathrm{X}_i, \mathrm{X}_j)$, $\lambda V = CV$ is converted to $n\lambda\alpha = K\alpha$ through the kernel function, and the kernel function $\mathrm{K}(\mathrm{X}_i, \mathrm{X}_j)$ can be calculated in the original space. In order to meet the default centralization hypothesis mentioned above, the kernel matrix was adjusted to $K^{\mu} = K - I_n K - K I_n + I_n K I_n$ in this study, with r representing the number of principal components obtained through dimensionality reduction processing, ω_k representing the contribution of the extracted k-principal component to the data set, α_i^k representing the i-th element of the feature vector, and the k-th principal component of data point X representing $\alpha_i^k K(\mathrm{x}_i, \mathrm{x})$. The first three principal component accumulation values were taken as the representative degree of the original data set. In order to optimize the kernel parameters of the introduced Gaussian kernel function $\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = \exp(-\|x - x_i\|^2 / \sigma^2)$ and polynomial kernel function $\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = [(x \cdot \mathrm{x}_i) + 1]^q$, the problem was transformed into an optimization problem and PSO algorithm was introduced. There are altogether N samples of consumption data $x \in K$ in ω, so the mean vector in the high-dimensional consumption data space is,

$$\mu_i = \frac{1}{N_i} \sum_{i=1}^{n} x \in K \tag{2}$$

$$\delta = \left|\overline{\mu_1} - \overline{\mu_2}\right| = \left|\eta^T (\mu_1 - \mu_2)\right| \tag{3}$$

Calculate the difference between the pairwise mean values of the projection direction, calculate the discrete matrix S_b between classes and the discrete matrix within classes:

$$\left\{\begin{array}{l} \overline{S_w} = \sum_{j=1}^{2} \sum_{i \in N} (\eta^T x_i - \eta^T \overline{\mu_j})^T \\ \overline{S_b} = (\overline{\mu_1} - \overline{\mu_2})^2 = (\eta^T \mu_1 - \eta^T \mu_2)^T \end{array}\right\} \tag{4}$$

According to the requirement of consumption warning model that the first three principal components should be greater than 0.85 and the first ten principal components should be greater than 0.95, the corresponding objective function is constructed:

$$\begin{array}{l} \min \sigma > 0 \\ \mathrm{s.t.} \min \mathrm{J}(\sigma) = \dfrac{\overline{S_w}}{S_b}, \max \sum_{i=1}^{3} \alpha_i K(\mathrm{x}_i, \mathrm{x}) \end{array} \tag{5}$$

Taking as fitness function $J(\sigma)$, the value range of kernel parameter $(\sigma_{\max}, \sigma_{\min})$ is determined by the limit of accumulation value of principal component in iteration. Set the number of particles N inertia weight ω and the maximum number of iterations T, randomly generate the initial population and continuously update the particle position until reaching the upper limit of iteration, output the optimal nuclear parameter σ, and complete the dimension reduction of the original consumption data. After obtaining the characteristic samples, x_{if} represents the p characteristic parameter in the i sample and c_{ip} represents the p characteristic parameter in the j category of consumption behavior center. According to the ant colony genetic algorithm, ant colony design clustering target is constructed. The consumption warning model of college students is as follows:

$$\min J(\mathrm{w}, \mathrm{c}) = \sum_{j=1}^{m} \sum_{i=1}^{N_j} \sum_{p=1}^{n} w_{ij} \left\| x_{if} - c_{jp} \right\|^2 \tag{6}$$

$$c_{jp} = \sum_{i=1}^{N_j} w_{ij} x_{ip} / \sum_{i=1}^{N_j} w_{ij} (\mathrm{i} = 1, ..., \mathrm{M}; \mathrm{p} - 1, ..., n) \tag{7}$$

$$w_{if} = \left\{\begin{array}{l} 1, Consumer i is the jth consumer group \\ 0, \mathrm{Otherwise} \end{array}\right\} \tag{8}$$

According to the transition probability formula, the ant solution is updated and the uniform two-point crossover operator is introduced to iterate and update the information matrix. When the clustering target is reached or the maximum number of iterations is reached, the clustering results are output, the ratio of characteristic data and income level is set as the threshold, the extreme samples are tagged into the model, and then different warnings are given to the characteristics of different consumer groups after classification.

RESULTS AND DISCUSSION

From July to November after the COVID-19 outbreak, a questionnaire survey was conducted among undergraduates in a university in Hubei province by random sampling. According to the consumption structure ratio of contemporary college students, the design problems include nutrition, life, clothing, entertainment, and excessive consumption, with a total of 17 consumption influencing factors. Considering the different consumption evaluation of families with different incomes, the household income option is added to the questionnaire as one of the criteria to determine the warning interval in the following part. A total of 89 valid papers were recovered. Four grades were determined according to the Linkert Scale method. According to the extreme values in the questionnaire, 6 unreasonable answers were screened out, and 83 valid data were finally obtained. Part of data is shown in Table 1 below:

Through data transcoding, the indicators are converted into 1,2,3,4 rating 164 according to the consumption amount. Considering the special

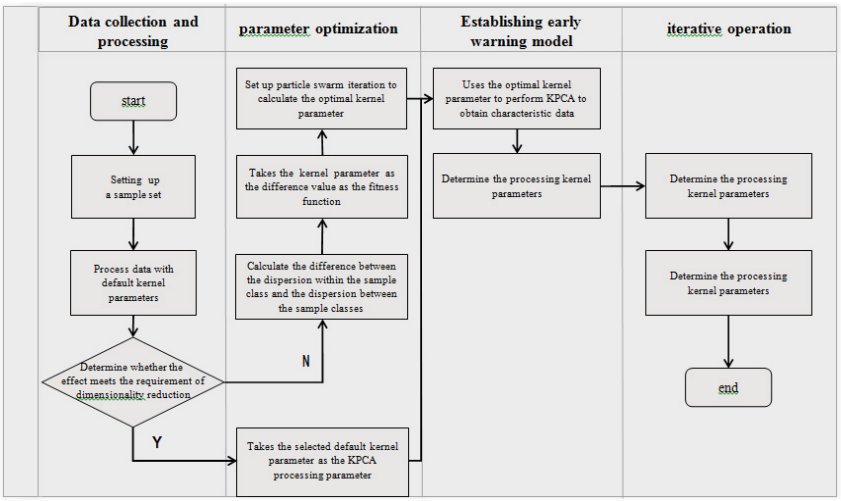

Figure 1) *Algorithm flow chart*

situation of consumption amount 0, the option of consumption amount 0 is separately defined as level 0. In order to represent the data indicators of college students' consumption obtained in the questionnaire, Pearson correlation analysis was carried out on the data and covariance was calculated. Relevant data of various factors were obtained through Matlab programming, as shown in Table 2:

According to Table 2, it can be found that there are information entanglements among various impact factors, and the impact factors need to be dimension-reduced to reflect the consumption behavior of college students. Among all the indicators, the highest correlation with the total consumption of college students is online shopping of clothing, online skin care products and the amount of spending on Singles' Day. It is reasonable to think that today's online shopping culture has become an important part of contemporary university consumption and influence factors. Table 3 is the PSO optimization parameter table. The kernel function is optimized by programming with Matlab. Set the default multinomial kernel parameter of 10 and Gaussian kernel parameter of 287 as the control group for comparison test.

After dimension-reduction processing, it can be found that KPCA is significantly better than PCA algorithm in data processing. After optimizing the Gaussian kernel function and polynomial kernel parameters respectively, it is found that when the kernel parameter is 8, the best dimension reduction effect can be obtained. The comparison of dimensionality reduction effect before and after PSO optimization is shown in Figure 2.

It can be seen from Figure 2 that when the kernel parameter σ is 8, the first three principal components of multiple linear kernel can achieve a cumulative contribution rate of 89.6%, which reaches the expectation. At the same time, the first three principal components obtained can get good aggregation effect in space, so the kernel function is considered to be the most suitable for this model.

Ant colony genetic algorithm was adopted for recognition and classification, and the maximum iteration number was set as 3000. 83 samples were trained by 100 ants under 3 classification modes, and the average training time was 152 seconds. The maximum value of the ratio between consumption amount and income was set as the threshold value. Three samples of obvious consumption amount were too high, four were normal, and three samples of online shopping consumption amount was too high were labeled. After being substituted into the model, the recognition accuracy is 80%, which meets the expected requirements. PCA-regression fitting was used to calculate the consumption fitting curve. K_1, K_2 and K_3 were used to represent the first three principal components respectively to fit the monthly consumption amount to get the consumption function:

$$y = 0.21 - 0.0969k_1 - 0.4074\sqrt[3]{k_3} \quad (9)$$

It can be seen from Figure 3 that the obtained consumption model can divide college students into three categories according to their consumption

TABLE 1
Consumption questionnaire survey raw data table

Expend amount	Diet/nutrition	Snacks/nutrition	Online shopping/ clothing	Live streaming gifts/ entertainment
1500-2000	Less than 300	More than 500	500-1000	Non use
1000-1500	300-600	200-500	Less than 100	0-100
1000-1500	300-600	1-200	100-500	100-500
1500-2000	600-1000	1-200	Less than 100	0-100
More than 2000	600-1000	200-500	100-500	Non use

Source: Questionnaires collected

TABLE 2
Data factor correlation analysis data table

Name of impact factor	Significant correlation factor	Relevant coefficient
Online shopping for clothes/clothing	Snacks and takeaway/nutrition	0
Online shopping for skin care/clothing	Online shopping for clothes/clothing	0.001
Excess consumption on Nov 11/ consumption amount	Dine together for consumption/ entertainment	0.001

TABLE 3
PSO optimization algorithm parameter table

Project	Particle number *N*	Inertia weight ω	Acceleration sensor *c*	Max iterations *T*	Speed limit *V*
Parameter values	20	1.2	0.4	100	1

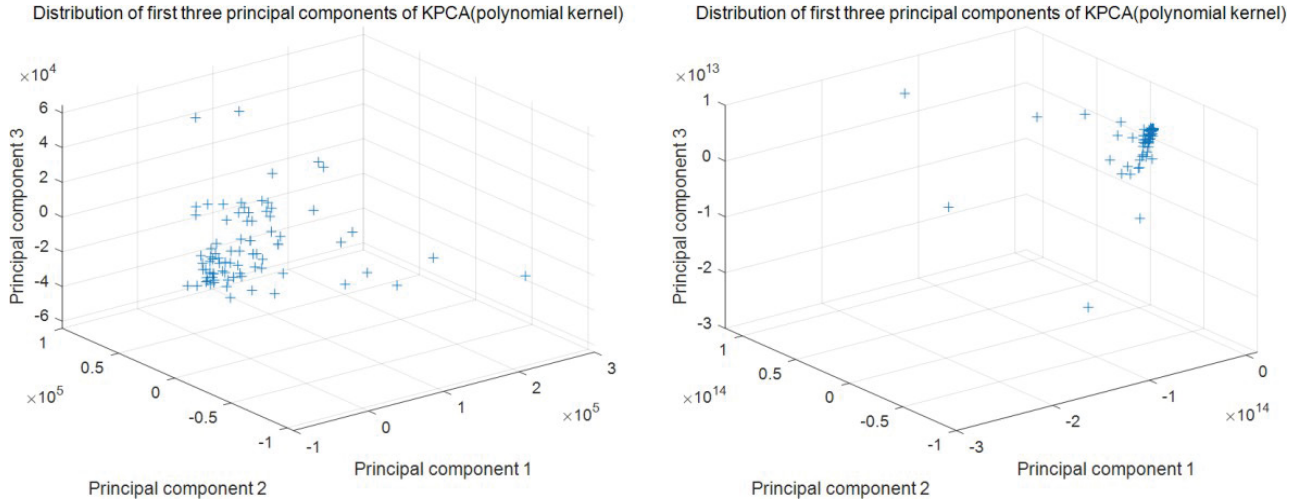

Figure 2) *Comparison of effects before and after PSO optimization*

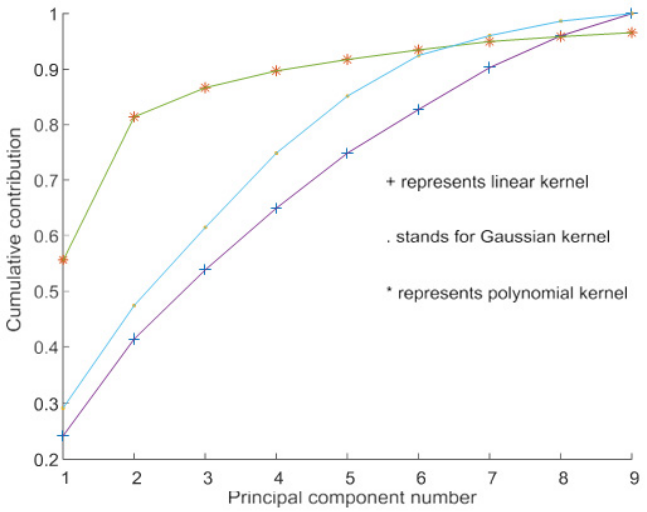

Figure 3) *Composition contribution accumulation chart*

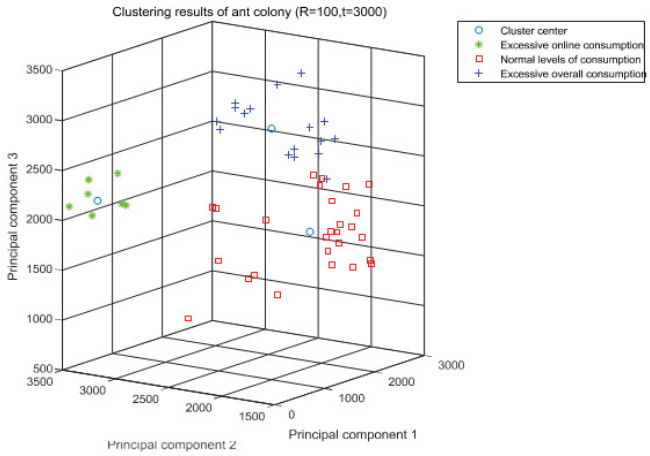

Figure 4) *Classification results of KPAG method*

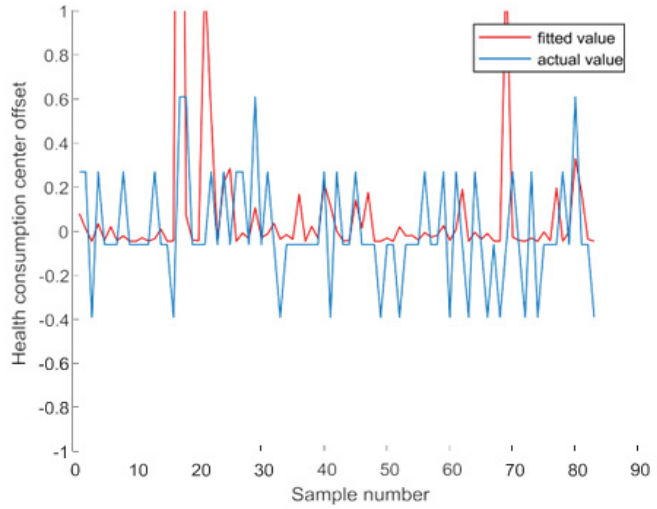

Figure 5) *Traditional PCA-regression consumption curve*

behaviors and give early warning for different consumption behaviors. In the fitting image in Figure 4, the traditional algorithm can partially fit the original monthly total consumption, and the fitting accuracy of the traditional PCA fitting regression algorithm is 58%. Comparison effect of the algorithm is shown in Table 4 (Figure 5).

CONCLUSION

Through the collection of relevant data and the corresponding mathematical optimization analysis, the following conclusions are drawn:

TABLE 4
Comparison of algorithm effects

Algorithm	Principal component contribution	Types of warning	Warning accuracy
Based on KPAG method	91%	3	80%
PCA- ant colony clustering	64%	3	64%
KPCA- linear fitting	80%	2	71%
PCA- linear fitting	64%	2	58%

(1) Online shopping of clothing, online shopping of consumer goods and the consumption amount of the Double 11 are the important influencing factors of college students' consumption, which indicates that the current online consumption has had a significant impact on college students' consumption behavior, which should be paid more attention to.

(2) Compared with the traditional PCA processing idea, 7 principal components need to be extracted if the standard is met. However, the method adopted in this paper can reduce.

(3) The principal components to 2, and the cumulative contribution rate is 89.6%, which well maintains the data characteristics.

(4) In this paper, the processing module of KPCA optimized by PSO can retain the features of the original data more completely and remove miscellaneous information. Compared with the case of no optimization, it shows a higher linear classification effect.

(5) Compared with the traditional algorithm, the consumption risk warning model designed in this paper based on KPAG method is richer in direction and can provide corresponding warnings for different consumption behaviors. The accuracy of the algorithm is higher than that of the traditional algorithm, and it has better feasibility and practical value.

STUDY LIMITATIONS AND FUTURE DIRECTION

Limited access to data

As this method involves questionnaire survey of college students, due to the limitation of actual regions, the questionnaire survey is concentrated in some universities when data collection is carried out. Lack of data collection from students in a wider region and in an international scope makes it impossible to conduct further robustness verification experiments. As a result, the model may have regional adaptation problems in the data clustering link, which has a certain impact on the robustness of the model. Further data collection is needed for modeling and validation experiments. In the following research, relevant data will be further collected through university cooperation, network questionnaire collection and other methods, and the robustness optimization experiment of the model will be conducted.

The label evaluation function needs to be optimized

In this method, the evaluation label in the test sample is defined mainly through the self-evaluation and the deviation between the actual consumption value and the average consumption value of the students surveyed in the questionnaire survey. However, there are slight differences in excessive consumption behaviors due to different household incomes. As a result, a small number of unreasonable classification results may occur in the actual model classification process. In the following research, the influencing factor of family income will be introduced to optimize the label evaluation function more scientifically and quantitatively. In the following research, the influencing factor of family income will be introduced to optimize the label evaluation function more scientifically and quantitatively.

Fewer kernel function types tested

In the process of model and kernel method, linear kernel, Polynomial kernel and Gaussian kernel are used in this study. There are still Radial Basis Function kernel, Laplacian kernel, sigmoid kernel and other kernel functions that have not been tested. Therefore, the model may not reach the optimal dimension reduction processing effect and the highest warning accuracy. Further testing and optimization can be done. Further testing and optimization is required. In the following research, different kernel functions will be added to the kernel function processing link for analysis and processing, in order to seek better dimension reduction effect and higher classification accuracy.

DISCLOSURE STATEMENT

We have no conflict of interests to disclose and the manuscript has been read and approved by all named authors.

FORMATTING OF FUNDING SOURCES

This work was supported by the Philosophical and Social Sciences Research Project of Hubei Education Department (19Y049) and the Staring Research Foundation for the Ph.D. of Hubei University of Technology (BSQD2019054), Hubei Province, China.

REFERENCES

1. Kuswanto H, Pratama WB, Ahmad IS. Survey data on students' online shopping behaviour: A focus on selected university students in Indonesia. Data in Brief. 2020;29:105073.
2. Shang Q, Jin J, Qiu J. Utilitarian or hedonic: Event-related potential evidence of purchase intention bias during online shopping festivals. Neurosci Letters. 2020;715:134665.
3. Bollen Z, Pabst A, Creupelandt C, et al. Prior drinking motives predict alcohol consumption during the COVID-19 lockdown: A cross-sectional online survey among Belgian college students. Addictive Behaviors. 2021;115:106772.
4. Shulan C, Li Y. Analysis on influencing factors of college students' consumption level in guangxi based on order logistic model. J Guangxi Aca Sci. 2011;(3):186-89.
5. Moeller S, Pisharady PK, Ramanna S, et al. Noise Reduction with Distribution Corrected (NORDIC) PCA in dMRI with complex-valued parameter-free locally low-rank processing. Neuro Image. 2021;226:117539.
6. Alsenan SA, Alturaiki IM, Hafez AM. Auto-KPCA: A two-step hybrid feature extraction technique for quantitative structure-activity relationship modeling. IEEE Access. 2020;21(9):1-2.
7. Yuanyuan G. Research on elimination of multicollinearity based on Kernel Principal Component Regression (KPCR). Sci Eng A. 2014.
8. Wenyan P, Zongjun W. Evaluation of low carbon economic development level based on kernel principal component analysis. Fin Econom. 2016;(004):55-9.
9. Nadia S, Ilyes E, Hassani M. On the use of KPCA pre-filtering for KCCA method. The International Journal of Advanced Manufacturing Technology. 2017;91(9-12):4331-40.
10. Abokhalil A. Maximum power point tracking for a PV system using tuned support vector regression by particle swarm optimization. J Eng Res. 2020;8(4):139-52.
11. Min Z, Huixian Y, Xunyong O, et al. KPCA feature extraction based on cultural particle swarm optimization. In: 2009 International workshop on intelligent systems and applications. Wuhan, China. 2009: pp. 2908-11.
12. Rongyi L, Jin Z, Zhongyu S. Kernel parameter optimization based on particle swarm optimization. J Jiangnan Univ. 2010;(04):444-47.
13. Junling Z. KPCA-based early warning model of college students' academic performance and its application. Edu Social Sci Integ. 2015.
14. Xiao Y, Tian X. Dark background image-denosing based on KPCA method. Adv Comput Sci Res. 2017:1128-31.
15. Hongfang Y, Shuang X, Huaqing W. Fault diagnosis of gearbox based on kpca and improved ant colony genetic algorithm. Measure Cont Tech. 2015;34(6):17-20.
16. Fan D, Chen R, Xue J, et al. Quality classification and evaluation of human-machine composite translations of scientific text based on KPCA. In: 2020 IEEE 3rd International conference on computer and communication engineering technology (CCET). Beijing, China. 2020: pp. 163-66.