Intelligent Statistical Method of Accounting Information Teaching Data Based on SVM

Chen Chen[1(✉)] and Yan Chao[2]

[1] Academic Affairs Office, Fuyang Normal University, Fuyang 236000, China
doublechen85@163.com
[2] School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China

Abstract. Aiming at the problem of poor data fitting results caused by large sample loss in the classification of intelligent statistical method of accounting information teaching data, an intelligent statistical method of accounting information teaching data based on SVM is proposed. Because the distribution of accounting information teaching data is seriously unbalanced, an unbalanced processing mechanism is established to improve the ability of data recognition. Design multiple binary SVM models, synthesize the accounting information teaching data label results predicted by multiple binary SVM based on the fusion strategy, build an intelligent statistical model, and finally output the statistical results of the classification status of the data label. Chi square goodness of fit test and K-S test are carried out for the intelligent statistical method. The results show that the data goodness of fit of the intelligent statistical method of accounting information teaching data based on SVM is higher than that based on latent variable model and weighted distance, so it has better data quality.

Keywords: SVM · Accounting information · Teaching data · Intelligent statistics · Statistical methods · Data statistics · Classifier

1 Introduction

Accounting is a business administration discipline with strong application. This major adheres to establishing morality and cultivating people, and cultivates high-quality workers and skilled talents who can engage in accounting and related posts such as enterprise cashier, salary accounting, warehouse management, transaction accounting, tax declaration, cost accounting, statistics and so on in the front line of operation and management. Through intelligent statistics of accounting information teaching data, we can analyze the trend of students' big data and provide technical support for Discipline Construction and management decision-making of Accounting Specialty in the school [1]. At present, the academic community has carried out extensive research on data intelligent statistical methods. Although there are great differences in starting time nodes and research progress, there is a consensus on the development mode from policy to region

W. Fu and G. Sun (Eds.): eLEOT 2022, LNICST 453, pp. 589–600, 2022.
https://doi.org/10.1007/978-3-031-21161-4_45

and then to institutions. Sun Fenglin et al. Proposed a contour analysis method suitable for multivariate ordered data [2]. Since the ordered data cannot meet the requirements of contour analysis for data normality, the latent variable model is used to assign the ordered variables, and the bootstrap method is used to reconstruct the samples, so that the reconstructed new data can meet the normality and the overall mean value is consistent with the original sample. Therefore, the contour analysis method can be applied to the comparison of the mean vector of ordered data. Lu Haitao et al. Proposed a weighted distance data statistical method using the original characteristics of data [3]. This method obtains the statistical information before the feature binarization of the data set, and uses the hash feature of the query vector to replace the binary coding to calculate the weight value. The weight value is calculated by using data set statistical information, query vector and database binary coding to avoid a large amount of loss of original data information caused by binarization, and better retain the differences between query data. The above intelligent statistical methods can use different applications to analyze and process data with different complexity and depth, but there is still a problem of large sample loss in classification. The application value of education big data should be reflected in the deep integration with the mainstream business of education and the continuous promotion of intelligent reform of the education system, specifically in improving scientific education management, promoting the reform process of education and teaching mode, guiding the practice of personalized learning and teaching activities, promoting the improvement of Education and teaching evaluation system, driving the transformation of scientific research paradigm Leading the humanized education service system suitable for people. SVM theory provides a method to avoid the complexity of high-dimensional space by directly using the inner product function of this space [4]. In the online inseparable mode, the solution method can directly solve the decision-making problem in relatively high-dimensional space. Therefore, an intelligent statistical method of accounting information teaching data is designed based on SVM to continuously and completely record all data related to the whole accounting information teaching activities, the innovation is to put forward the theory of SVM, which is a method used for data analysis, pattern recognition, classification and regression analysis. It has obvious advantages in solving the problems of small samples, nonlinear and multi-dimensional pattern recognition. SVM maps the samples to be classified to a higher dimensional vector space, and constructs a maximum interval hyperplane in this space to maximize the interval between sample points belonging to different classes, so as to ensure the accuracy of classification and provide support for accounting information teaching evaluation and decision-making.

2 Intelligent Statistical Method of Accounting Information Teaching Data Based on SVM

2.1 Establishing the Unbalanced Processing Mechanism of Accounting Information Teaching Data

N nearest neighbor samples of a few accounting information teaching data samples are found by using Euclidean distance, Then M samples are randomly selected from N

original nearest neighbor samples as parent samples. If the original number of a few samples is less than M, it can be selected repeatedly. The method to realize this process needs not only learning data, but also prior knowledge. Therefore, inductive learning is inseparable from the prior knowledge of the approximate function of the selected learning method. Then, interpolate between the original sample and its M adjacent samples according to the following formula to generate M new offspring samples. The interpolation formula can be expressed as:

$$A'_M = A_N + \alpha(A_N - A_M) \tag{1}$$

In formula (1), A_M and A'_M represent the eigenvectors of samples before and after interpolation respectively; A_N represents the eigenvector of the nearest neighbor sample; N and M are selected quantities; α is a random number between 0 and 1. Any function set can be used as the function set of the learning method to perform the learning process. The newly generated samples and the set of original samples constitute new accounting information teaching data.

2.2 Design Multiple Binary Classification SVM Models

SVM is a very classic and proven classification model. It bears the brunt in the field of machine learning and has unique advantages [5]. In the secondary classification task, if the sample is negative, it is recorded as - 1, and if the sample is positive, it is recorded as + 1 [6]. The goal of all classification models is to obtain a partition rule that can reasonably distinguish positive and negative samples on the training set, and then apply this rule to the test sample set with unknown real label in order to obtain high classification accuracy. The problem of data classification can usually be summarized as two steps: creating a classification model. The model is obtained by observing and summarizing the limited tuple data. In the rules of the model, the basic characteristics of data class objects are described. Based on simple assumptions, SVM believes that the rule for dividing positive and negative samples should be a hyperplane, and the corresponding equation is:

$$\beta u + \varphi = 0 \tag{2}$$

In formula (2), u represents the sample; β and φ represent hyperplane parameters. In order to reduce the risk of over fitting and improve the generalization performance of the model, SVM finds the unique partition hyperplane by maximizing the interval. Effectively apply the feature function to SVM and import effective data, so that these data can be effectively mapped to the corresponding space and calculated later [7]. The hard interval is the sum of the distances from two heterogeneous support vectors to the partition hyperplane, and the so-called support vector is the sample feature vector with hyperplane inequality constraints. The calculation formula of hard interval is as follows:

$$d = \frac{2}{\|\beta\|} \tag{3}$$

In formula (3), d represents hard interval. According to the promotion of dimension, the classification surface will be more three-dimensional, so a three-dimensional decision

surface can be generated, and the samples can be classified in three-dimensional space to form a multi-dimensional decision plane [8]. In order to solve the above problems, the basic models of SVM with "kernel function" and "soft interval" are introduced. The introduction of kernel function can effectively nonlinear map the eigenvector to a high-dimensional space, and divide the training sample set by hyperplane. The nonlinear mapping function can be expressed as:

$$\gamma(u_1)^T \gamma(u_2) = \theta(u_1, u_2) \tag{4}$$

In formula (4), u_1, u_2 is the data individual in the sample; γ represents the mapped eigenvector; T represents transpose matrix; θ represents and functions. The binary SVM model constructed this time adopts sigmoid kernel function, and the calculation formula is as follows:

$$\theta(u_1, u_2) = \tan\left(\eta_1 u_1^T u_2 + \eta_2\right) \tag{5}$$

In formula (5), η_1 and η_2 are parameters of sigmoid kernel function. The introduction of soft interval is to increase the fault-tolerant mechanism of SVM on some samples. The relaxation is regularized by adding the sum of the relaxation of all samples to the minimization target term to control the overall relaxation. After adjustment, the binary SVM model constructed this time can be expressed as:

$$f = \min \frac{\|\beta\|^2}{2} + \varsigma \sum_u \tau \tag{6}$$

In formula (6), f represents the objective function of the two classification SVM model; τ represents the relaxation of the sample; ς represents the regularization coefficient of the overall relaxation. Divide the category. Use the model created in the first step to divide the unclassified data into one or several categories according to the characteristics of the subclass model. The two classification SVM model is used to find the maximum interval hyperplane of nonlinear accounting information teaching data, so as to provide data classification basis for intelligent statistics.

2.3 Establishment of Multi Classification Data Labels Based on SVM

To solve the problem of excessive sample loss in large-scale data classification, this paper adopts the "progressive" migration method [9], and the process of establishing multi classification data labels based on SVM is shown in Fig. 1.

As can be seen from Fig. 1, the accounting information teaching data set is trained by SVM model. In this study, one-to-one, one-to-many, error correction output coding and other decomposition strategies are adopted to construct multiple binary SVM single classifiers. For the three decomposition strategies, three fusion strategies are adopted: the one with the highest number of votes, the one with the largest function value and the one with the smallest Hamming distance. The output results of each two classification SVM single classifier are fused to obtain the classification results of accounting information teaching data labels. The label spaces in the source domain and the target domain are

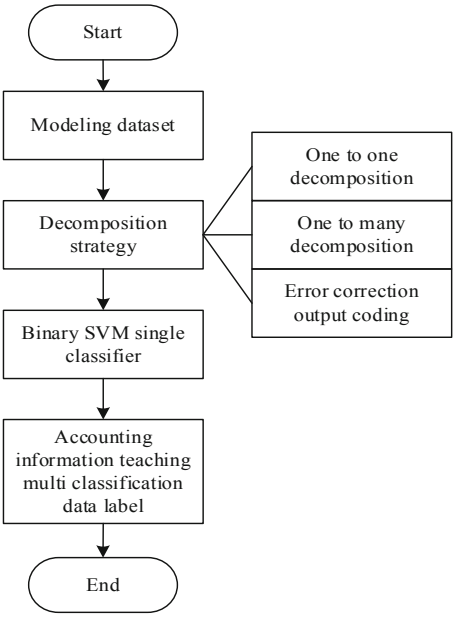

Fig. 1. Multi classification data label modeling process based on SVM

inconsistent. There are no fine-grained category labels in the source domain and the target domain, so feature mapping cannot be constructed unsupervised. Therefore, this paper uses a small number of fine-grained tags in the target domain, which are called "active tags". Due to the limitation of the length of the article, the error correction output coding strategy is mainly described below. The error correction output coding strategy uses the idea of error correction code in model supervision training to decompose the multi category state classification problem into several two category problems [10]. In the coding process, if each category can only be specified as positive or negative examples, it is called binary ECOC code; If each category can be specified as positive example, negative example and inactive example, it is called ternary ECOC code. Ideally, the samples generated by the encoder can "fool" the discriminator, so that the samples in the source domain and the target domain have feature consistency. It is assumed that the output of the discriminator to the source domain sample is 1 and the output to the target domain sample is 0. This paper combines ternary ECOC code with SVM to establish an ECOC-SVM model for multi category data. The framework is shown in Fig. 2, the model is divided into two processes: encoding and decoding.

As can be seen from Fig. 2, in the coding process, the accounting information teaching data is mapped to a set of binary category labels, and the mapping scheme is recorded in the coding matrix. A binary training set is formed by N coding divisions to train N binary SVM classifiers. Each binary SVM classifier evaluates the data to be processed, and synthesizes the prediction results of each SVM classifier to obtain the code sequence. This coding method integrates the process of model learning and data annotation. This method selects the most valuable samples for experts to label according to the set sample

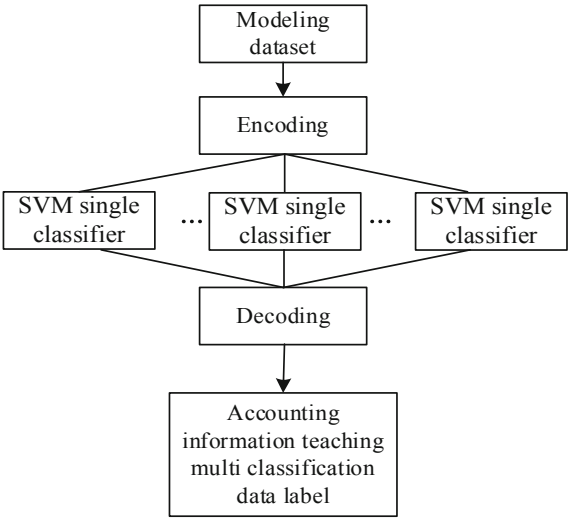

Fig. 2. ECOC-SVM model framework

selection method in the process of model learning. These "valuable" samples are often samples with large losses. The goal of "active learning" method is to use the minimum amount of labels to realize model training. Then calculate the Hamming distance between the code sequence and each line in the code matrix. The calculation formula is as follows:

$$h = \frac{1}{2}\left[1 - \text{sgn}(pq)\right] \tag{7}$$

In formula (7), h represents Hamming distance; sgn represents symbolic function; p represents code sequence; q represents the row vector of the code matrix. Thus, the classification results of accounting information teaching data labels are obtained.

2.4 Establishing an Intelligent Statistical Model of Accounting Information Teaching Data

The in-depth application of big data technology in the field of education can widely mine and analyze education data, analyze the data association and the potential value hidden behind the data, find the deficiencies and defects in education and teaching activities, put forward solutions and optimize education management plans. Big data technology can be used to collect data generated by daily educational activities, store and maintain these data, or establish models to statistically analyze personnel data, educational resource data and daily teaching data of educational institutions, so as to extract useful information and provide data support for scientific decision-making for educational and teaching activities. Based on a certain decomposition strategy, the multi category SVM modeling problem of accounting information teaching data is decomposed into the SVM modeling sub problem of multiple binary classification accounting information teaching data, and then the label results of accounting information teaching data predicted

by multiple binary classification SVM are synthesized based on the fusion strategy, and finally the classification results of data label status are output. It is a feasible idea to mine the characteristics and transferable knowledge of existing data sets and their labels, and apply them to the new and scarce range of labels. Through the intelligent statistics of basic accounting information teaching data, we can get the education data one stage in advance. Through the analysis and sorting of prediction data, we can provide reference for education development and serve for development decision-making. The main problem of data intelligent statistics is to establish a mathematical model. The mathematical model is a mathematical expression describing the dynamic characteristics of the controlled object and an important basis for analyzing and synthesizing the control system. The data intelligence statistical model established in this paper is shown in Fig. 3.

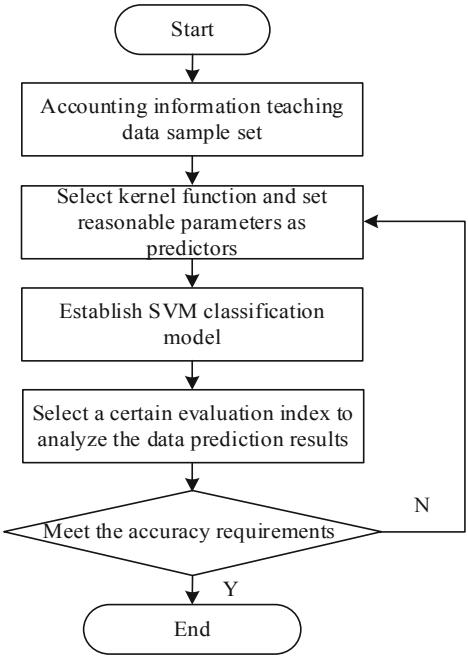

Fig. 3. Data intelligence statistical model

Figure 3 shows that the learning method of the data intelligent statistical model is a processing system. The system is determined through learning, and then corresponding output values are generated for each input vector according to the unknown conditional probability. However, there are few real random outputs in real systems, but the unmeasured inputs are consistent. Therefore, statistically, the impact of these unobserved inputs on the system output can be regarded as random and with a certain probability distribution. The goal of data intelligence statistics is to estimate unknown correlations in a class of approximate functions using available data. The estimation corresponds to the minimum expected risk function, including the general distribution of data. In the process of selecting dimensions, attention should be paid to the relationship between the data. It is

suggested to do Pearson correlation test on the selected data, so as to reasonably select indicators for constructing the dimensions of the training model. When the correlation is guaranteed, the constructed dimensions should also establish as many dimensions as possible to describe the characteristics of abnormal data, Under the description of multiple dimensions, the abnormal characteristics of data will also be more prominent. Traditional statistical methods estimate the potential regularity of things according to the external characteristics of the number of things, so the definition of intelligent statistics is the principles and methods involved in the process of data collection, data analysis and conclusion. The randomness and regularity of things is an important content of statistics. The law of statistical properties when samples tend to infinity is asymptotic theory, which is an important content of intelligent statistics. Any feature or attribute can be represented by a variable in statistics, and the value of the variable represents the thing or individual of a feature studied. So far, the design of intelligent statistical method of accounting information teaching data based on SVM is completed.

3 Experimental Study

In order to verify the application effect of intelligent statistical method of accounting information teaching data based on SVM, an experiment is designed and analyzed. The output results of the intelligent statistical method of accounting information teaching data based on SVM are compared with the data intelligent statistical method based on latent variable model and weighted distance. Collect and sort out the accounting information teaching data of a university, a total of 20896 data. This experiment mainly tests the location probability distribution of statistical data. The goodness of fit test was carried out under 5000 and 20000 accounting information teaching data respectively. This paper tests the intelligent statistical method of accounting information teaching data by using the results of chi square goodness of fit test and K-S test. Goodness of fit test is an effective method for statistical significance test using chi square statistics. According to the overall distribution of data, it obtains the expected frequency of each category in the classification variables, compares the classification frequency obtained by the actual test, obtains the difference between the actual frequency and the expected frequency, and analyzes the classification variables in detail. After chi square goodness of fit test, it can be judged whether the frequency distribution of the first digit in the statistical data is significantly different from Benford's law. The calculation formula of chi square goodness of fit test can be expressed as:

$$w_1 = U \sum_{x=1}^{9} \left[\frac{(g_x - y_x)^2}{y_x} \right] \tag{8}$$

In formula (8), w_1 represents the chi square goodness of fit test result; U represents the total amount of experimental test data samples; g_x represents the number of times the number x is found in the first (second or third) place in the statistical data; y_x represents the number of times the number x is found based on the first bit (second or third bit) of Benford's law. The chi square test results are shown in Table 1 and Table 2 respectively.

According to the statistics of 5000 data, the chi square goodness of fit of the intelligent statistical method of accounting information teaching data based on SVM is 18.021,

Table 1. Chi square test results of 5000 accounting information teaching data

Number of experiments	Chi square goodness of fit		
	Intelligent statistical method of accounting information teaching data based on SVM	Intelligent statistical method of accounting information teaching data based on latent variable model	Intelligent statistical method of accounting information teaching data based on weighted distance
1	17.264	13.407	11.446
2	18.488	12.815	12.675
3	16.555	13.548	12.288
4	17.246	12.656	11.357
5	18.873	13.922	12.064
6	17.622	12.535	13.195
7	18.331	12.204	13.558
8	19.264	13.521	12.229
9	18.045	13.852	12.036
10	18.522	13.663	12.322

which is 4.809 and 5.704 higher than that based on latent variable model and weighted distance.

According to the statistics of 20000 data, the chi square goodness of fit of the intelligent statistical method of accounting information teaching data based on SVM is 15.277, which is 4.056 and 4.925 higher than that based on latent variable model and weighted distance. K-S test refers to the method of observing whether the sample is a specific theoretical distribution. Compare and analyze the cumulative distribution function of the sample data and the cumulative distribution function of the specific theoretical distribution, and compare the differences to find out the maximum of all the absolute values of the differences. Then, the table is looked up to show whether the maximum value meets the confidence interval. The relative size of the maximum value and the zero bound value reflects the theoretical distribution of the data. The calculation formula of K-S test can be expressed as:

$$w_2 = \max(z_{1x} - z_{2x}) + \max(z_{2x} - z_{1x}) \tag{9}$$

In formula (9), w_2 represents the K-S test result; z_{1x} represents the first cumulative distribution function in the actual statistical data; z_{2x} represents the theoretical distribution, that is, the cumulative distribution function based on the first digit in Benford's law. K-S test results are shown in Table 3 and Table 4 respectively.

According to the statistics of 5000 data, the K-S test result of the intelligent statistical method of accounting information teaching data based on SVM is 1.204, which is 0.274 and 0.276 higher than that based on latent variable model and weighted distance.

Table 2. Chi square test results of 20000 accounting information teaching data

Number of experiments	Chi square goodness of fit		
	Intelligent statistical method of accounting information teaching data based on SVM	Intelligent statistical method of accounting information teaching data based on latent variable model	Intelligent statistical method of accounting information teaching data based on weighted distance
1	15.490	11.426	10.444
2	15.764	11.744	9.185
3	15.851	10.875	9.556
4	14.585	11.557	9.663
5	14.648	11.685	10.390
6	15.316	10.298	10.082
7	15.222	11.566	10.255
8	14.565	10.853	11.529
9	15.233	11.622	11.945
10	16.100	10.585	10.471

Table 3. K-S test results of 5000 accounting information teaching data

Number of experiments	Chi square goodness of fit		
	Intelligent statistical method of accounting information teaching data based on SVM	Intelligent statistical method of accounting information teaching data based on latent variable model	Intelligent statistical method of accounting information teaching data based on weighted distance
1	1.154	0.925	0.919
2	1.168	0.932	0.926
3	1.246	0.945	0.935
4	1.129	0.916	0.938
5	1.155	0.923	0.944
6	1.283	0.940	0.937
7	1.225	0.931	0.938
8	1.202	0.914	0.915
9	1.231	0.928	0.906
10	1.247	0.945	0.923

Table 4. K-S test results of 20000 accounting information teaching data

Number of experiments	Chi square goodness of fit		
	Intelligent statistical method of accounting information teaching data based on SVM	Intelligent statistical method of accounting information teaching data based on latent variable model	Intelligent statistical method of accounting information teaching data based on weighted distance
1	1.046	0.894	0.886
2	1.078	0.888	0.864
3	1.085	0.895	0.897
4	1.060	0.873	0.888
5	1.119	0.882	0.895
6	1.065	0.895	0.906
7	1.086	0.909	0.882
8	1.075	0.883	0.865
9	1.094	0.895	0.870
10	1.108	0.874	0.882

According to the statistics of 20000 data, the K-S test result of the intelligent statistical method of accounting information teaching data based on SVM is 1.082, which is 0.193 and 0.198 higher than that based on latent variable model and weighted distance. Therefore, the data fitting degree of the statistical method designed in this paper is higher, and the statistical results of teaching data have higher quality.

To sum up, the chi square goodness of fit and K-S test results of the method in this paper are both high and have good performance.

4 Conclusion

Intelligent statistics of accounting information teaching data refers to systematically collecting all aspects of school accounting information teaching information according to certain purposes and standards, adopting scientific attitudes and methods, making qualitative and quantitative value judgments on the status and performance of activities, personnel, management and conditions in educational work, and accurately understanding the actual situation of accounting information teaching activities, Evaluate the school running level and education quality, so as to provide a reliable basis for the school improvement work to carry out education reform and the education management department to improve macro management. The innovation of the research content is to use SVM theory to make the intelligent statistical method of accounting information teaching data have better performance. The follow-up research will refine the problems to be studied. It is necessary to decompose large problems into small problems. The

weight of data indicators can be reasonably selected according to different data types, and SVM and other methods can be used to test the data quality in more detail.

Fund Project. 1. 2021 Key Project of University-level Young Talents: Research on Budget Performance Management of colleges and universities from the mid-term perspective; Project Number: rcxm202109

2. Key project of Higher Education Department of Anhui Province in 2021: Research on carbon information disclosure and enterprise performance in the context of carbon neutrality; Project Number: SK2021A0329

References

1. Chen, D., Zhan, Y., Yang, B.: Analysis of applications of deep learning in educational big data mining. E-education Research **40**(2), 68–76 (2019)
2. Sun, F., Lu, T., Lei, S.: The profile analysis of multi-ordinal data based on underlying variable model. Statistics & Information Forum **34**(5), 3–9 (2019)
3. Lu, H., Tian, A., Wang, Z., et al.: Data statistics-based query adaptive weighted ranking algorithm. Computer Engineering and Design **40**(12), 3538–3544 (2019)
4. Morozova, I.M., Lyusev, V.N., Gladkova, M.N., et al.: Information technologies in teaching humanitarian disciplines. J. Educ. Psychology - Propositos y Representaciones **9**(1), 817–825 (2021)
5. Luo, X., Zhao, L., Liu, J., et al.: Mining outliers in multi-scale time series data based on neural network technology. Computer Simulation **38**(1), 231–235 (2021)
6. Haiyang, M.A., Lejun, L.I., Liu, X., et al.: Teaching design of tourism management based on information-based teaching method: a case study of selection of hotel construction site. Asian Agric. Res. **13**(12), 55–61 (2021)
7. Sun, S.Y., Cui, Y.M.: Information teaching design of "Basic Nursing Technology" based on blended teaching model. Education Teaching Forum **18**, 244–256 (2020)
8. Yang, L., Zhang, D., Luo, J., et al.: Automatic recognition for cotton spider mites damage level based on SVM and AdaBoost. Trans. Chin. Soc. Agric. Mach. **50**(2), 14–20 (2019)
9. Zhang, J., Chen, Z., Ma, J., et al.: Investigating the influencing factors of teachers' information and communications technology-integrated teaching behaviors toward "Learner-Centered" reform using structural equation modeling. Sustainability **13**(22), 1–17 (2021)
10. Gu, S., Wang, S.: Research of carbon financial risk early warning model based on SVM. East China Economic Management **33**(3), 179–184 (2019)