Hindawi
Journal of Function Spaces
Volume 2023, Article ID 9825379, 1 page
https://doi.org/10.1155/2023/9825379

Retraction

Retracted: Application of Optimized Support Vector Machine Model in Tax Forecasting System

Journal of Function Spaces

Received 3 October 2023; Accepted 3 October 2023; Published 4 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Xin, "Application of Optimized Support Vector Machine Model in Tax Forecasting System," *Journal of Function Spaces*, vol. 2022, Article ID 6212579, 10 pages, 2022.

Hindawi
Journal of Function Spaces
Volume 2022, Article ID 6212579, 10 pages
https://doi.org/10.1155/2022/6212579

Research Article

Application of Optimized Support Vector Machine Model in Tax Forecasting System

Yu Xin

Emerging Economic Formats Research Institute, Shandong Management University, Shandong, Jinan 250357, China

Correspondence should be addressed to Yu Xin; 14438119970177@sdmu.edu.cn

Received 6 June 2022; Revised 27 July 2022; Accepted 5 August 2022; Published 22 August 2022

Academic Editor: Miaochao Chen

Tax forecast has an important impact on financial budget and tax plan. The amount of tax data is greatly increased, the difficulty of tax forecast is improved, and the accuracy is always difficult to keep up with the development demand. Analyze the application of optimized support vector machine model in tax prediction system. Based on the simple analysis of the research situation of tax prediction and the research status of data mining algorithm in tax data classification and prediction, this paper constructs a tax prediction model. Aiming at the problem of too many influencing factors of tax prediction, this paper puts forward the use of principal component analysis to extract the main factors, reduce the dimension of tax data, and reduce the difficulty of analysis. Support vector machine is used to realize tax prediction, aiming at the problem of parameter optimization proposed to optimize the parameters. Finally, the prediction accuracy is evaluated by comparing the error between tax prediction value and real value. The results show that the algorithm used in this paper can optimize the parameters of support vector machine. The tax prediction results show that the predicted value is similar to the real value curve. After grid search optimization, the introduction of principal component analysis reduces the redundancy and improves the prediction accuracy.

1. Introduction

Tax forecast analysis is also called “tax trend analysis.” It refers to an analysis that predicts the trend of future tax revenue and provides decision-making services for leaders according to the mastered historical laws of taxation, using the continuously reflected relevant materials and data, combined with the current economic tax sources and investigation and research materials. This method is applicable to tax planning analysis, economic tax source investigation analysis, and tax accounting analysis. It is an indispensable analytical method for formulating and inspecting tax policies and measures. Especially before the changes of national tax policies, such as the new or suspended collection of a certain type of tax, as well as the provisions on tax increase, tax reduction or exemption, and other major tax measures, the impact of tax revenue should be predicted and analyzed. In the information age, tax analysis and prediction has become an essential content of tax management [1]. Tax data is a kind of time series data, which is not only affected by the economic environment, but also affected by cultural background, political background, and other factors. These data are nonlinear and unstable [2]. Tax forecasting can timely discover the tax change trend and collection; has an important impact on strengthening the organization’s income, resource allocation, and management decision-making; and can help relevant personnel formulate tax policies. Therefore, the accuracy of tax forecasting is very important [3]. At present, there are many researches on tax prediction, from linear analysis and regression analysis to various data mining algorithms, but the accuracy is not high [4].

Based on this background, this paper studies the application of optimized support vector machine in tax forecasting system. In Section 1, we briefly analyze the current tax forecasting research and briefly introduce the chapter arrangement of this research; in Section 2, we mainly introduce the research of support vector machine and prediction algorithm and summarize the shortcomings of the current research. In Section 3, we construct the tax prediction. Aiming at the problem of many tax influencing factors, the principal component analysis method is introduced to analyze the correlation of the main tax factors and extract the main indicators. In Section 4, we simulate and analyze the support vector machine tax prediction model constructed in this paper. After preprocessing, the tax data is brought into the tax prediction model and compares and analyzes the tax data before and after parameter optimization, and the accuracy of tax prediction is significantly improved after the parameter optimization of support vector machine. In the last section, we summarize the whole paper.

The innovation of this paper is that the support vector machine method is used in tax prediction, and the principal component analysis method is introduced, which reduces the influencing factors of tax and improves the prediction accuracy. Aiming at the problems that are used to optimize its parameters, the regularization parameters and radial basis kernel function parameters are applied to tax prediction to improve the accuracy of data analysis.

2. State of the Art

The importance of tax forecasting has been recognized. In terms of relevant research, it has also developed from early qualitative research to current quantitative analysis. Tax forecasting is more scientific and normative [5]. Aprilia and Agustiani adopt K-means clustering analysis algorithm in account data classification, which improves the effect of clustering analysis [6]. In the analysis of tax data, Sun et al. proposed a tax dynamic clustering algorithm, which automatically determines the number of clusters and modifies the cluster center according to the frequency of internal attribute values. TDCA algorithm is a tax dynamic clustering based on weak convergence sequence coefficient judgment function and tax domain knowledge, and the clustering effect is good [7]. In the research of tax risk assessment, Didimo et al. proposed the Maldive method. Through the information diffusion strategy, it is possible to expand the collection of risky taxpayers and display the output to analysts through the network visualization system [8]. In the analysis of financial data, Li constructed financial related indicators and financial report recognition model. The accuracy, precision, recall, and F value show that the performance of the improved BP neural network has been improved [9]. Mwanza and Phiri used intelligent mining algorithm in the research of tax data to realize the outlier algorithm of fraud detection, continuous monitoring based on distance and outlier query based on distance, which improved the accuracy of abnormal data analysis [10]. Miller used KH coder's data mining technology to model individual tax behavior and used unsupervised machine learning text mining and modeling technology to help conduct large-scale analysis of tax behavior methods and find the problems of avoidance and tax evasion in time [11]. Battiston et al. used machine learning algorithm for tax analysis in their research and proposed a loss function to analyze the difference between ideal tax and display [12]. In his research, Hao et al. proposed an algorithm for financial risk prevention and carried out special data preprocessing on convolutional neural network. Combined with the requirements of digital inclusive financial risk method, he constructed a digital inclusive financial risk prevention model [13–15], so as to timely find financial abnormalities and carry out risk early warning.

To sum up, it can be seen that there are many researches on tax prediction at home and abroad, but different algorithms have their own limitations. Linear regression analysis can only analyze the linear relationship, ignoring the nonlinear relationship of tax influencing factors. The structure of neural network algorithm itself is complex and requires high sample size, so the accuracy of tax prediction is not very high. Support vector machine has its own optimization in the application of prediction, can get the global optimal solution, the data can be analyzed in high-dimensional space, the complexity is not high, and has certain advantages in tax prediction. However, this algorithm needs to determine the regularization parameters and radial basis function parameters, and other methods need to be adopted for parameter optimization in application.

3. Methodology

3.1. Design of Tax Forecasting System. Tax prediction uses various impact indicators of tax revenue to analyze and introduces prediction theory, data mining algorithm, and model to predict. Affected by the economic environment, social environment, and science and technology, the tax changes greatly, and the formed data has a large amount of redundant data, which has great correlation. The traditional economic prosperity index cannot fully reflect the real situation of economic development. In response to these problems, tax big data can give full play to the advantages of complete and dynamic tax coverage, find more accurate and sensitive synchronization indicators and leading indicators in the big data set (for example, take the value-added tax as one of the synchronization indicators and the value-added tax on imported goods as one of the leading indicators), and further adopt the machine learning method to compile the composite index. The economic prosperity index based on tax big data can meet the technical requirements of the above two key steps, and it can more accurately reflect the real situation of economic development to a certain extent. Compared with the traditional economic prosperity index, the economic prosperity index based on tax big data shows its progress in two aspects. First, at the level of computing methods, it has become a trend to apply machine learning methods to study economic problems, such as ridge regression and lasso in machine learning.

Considering the multidimensional and nonlinear characteristics of tax data, it is difficult to find the change characteristics and trends of tax data by a single-factor analysis. Therefore, it is necessary to find the main influencing factors from a large number of data.

In this paper, the main factors affecting the prediction, eliminate the redundancy between various factors, and then establish a model to predict tax. This method makes use of the advantages of nonlinear function prediction and the ability of principal component extraction. The prediction accuracy can be improved. The prediction model is shown in Figure 1.

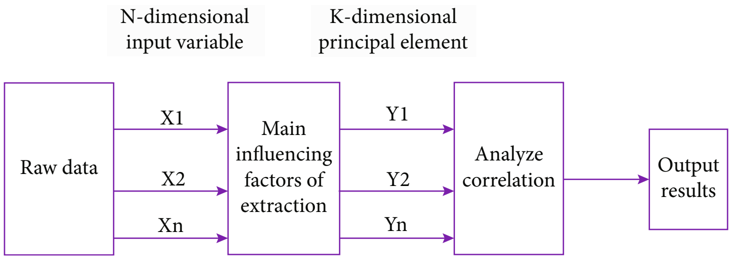
N-dimensional input variable
K-dimensional principal element
Raw data
X1
X2
Xn
Main influencing factors of extraction
Y1
Y2
Yn
Analyze correlation
Output results

FIGURE 1: Design of tax forecasting model.

Reduce the scalar index, and retain the amount of information of the index. This algorithm rotates the n-dimensional spatial coordinates and processes the results without changing the sample data. The principal components obtained are uncorrelated [16]. The introduction of this algorithm can reduce the dimension, eliminate redundant information, and combine it to simplify the architecture and improve the prediction performance. Assuming that the p-dimensional random vector is $X = (X_1, X_2, \cdots, X_p)^T$; these samples form a sample matrix, and the samples are transformed. The formula is

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \tag{1}$$

where i and j are positive integers. After processing, standardized samples are obtained. Calculate the coefficient matrix for standardized samples, and the formula is

$$R = \left(r_{ij}\right)_{p\times p} = \frac{Z^T Z}{n-1}, \tag{2}$$

where i and p are positive integers, and the value range is $(1,p)$. Calculate the characteristic equation $|R - \lambda I_p| = 0$ of the sample correlation matrix to obtain the characteristic root and determine the value of m according to the following formula:

$$\frac{\lim\limits_{m\longrightarrow\infty} \sum_{j=1}^{m} \lambda_j}{\lim\limits_{p\longrightarrow\infty} \sum_{j=1}^{p} \lambda_j} \geq 0.95. \tag{3}$$

After calculation, the information utilization rate reaches more than 95%. For each λ_1, j value, calculate the equation $Rb = \lambda_i b$ to obtain the eigenvector. The variables become principal component factors after orthogonalization. The principal component factors obtained from the comprehensive evaluation are weighted to obtain the final evaluation result value.

3.2. Optimized Support Vector Machine Model. Among the common machine learning algorithms. This algorithm is based on statistics and can be used for classification, as well as regression analysis. The core lies in the minimization of structural risk, which needs to adjust the confidence range and empirical value and improve the generalization ability through the proportion between the two [17–19].

When using support vector machine for prediction analysis, it is assumed that there is a training sample set, expressed by $\{x_i, y_i\}$, and the mapping condition from input to output is calculated to meet $f(x) = y$. The regression function is established by using support vector machine algorithm, and the formula is

$$y = f(x) = w \times x + b. \tag{4}$$

The nonlinear mapping is expressed as

$$y = f(x) = w^T \times \phi(x_i) + b. \tag{5}$$

The objective function minimization is used to determine the regression function. In the application of support vector machine, in order to avoid the problem of space mapping with different dimensions, it is necessary to introduce kernel function and turn it into a linear analysis problem. Kernel function is to calculate the form of inner product when the data is indeed mapped. For linearly nonseparable data, the data can be mapped to another space to make the data perfect or almost perfect linearly separable. The same is true when we need to map data to a high-dimensional space. Instead of providing an accurate mapping to SVM, we provide the dot product of pairing all points in the mapping space. In the specific application, it is assumed that the data in the low-dimensional space is not linearly analyzable, and the kernel function is used to map these data to Gao Wei to realize the linear analysis of sample data [20]. For support vector machine algorithm, the change of kernel function parameters directly affects the performance, and the influence degree of different kernel functions is also different. We need to choose a reasonable kernel function. At present, there are many kinds of kernel functions. Generally, functions that meet Mercer rules belong to kernel functions. Assuming that a subset of the input space is represented by X and the feature space is represented by H, there is a mapping $\varphi(x)$ from space to feature space, so that all x belong to the input space. All functions that can satisfy $K(x, z) = \varphi(x)\varphi(z)$ are kernel functions. Polynomial kernel functions are

$$k\left(x_i, x_j\right) = \left(x_i x_j + 1\right)^d. \tag{6}$$

The sigmoid kernel function is

$$k\left(x_i, x_j\right) = \tanh\left[b\left(x_i x_j\right) + c\right). \qquad (7)$$

RBF kernel function is expressed as

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\delta^2}\right). \qquad (8)$$

In the support vector machine algorithm, RBF kernel function is selected. This kernel function has the characteristics of wide convergence domain and can handle a variety of samples, which is widely used [21–23]. RBF kernel function only involves one parameter, so its performance is more stable and easy to apply. In the analysis of this paper, assuming $g = 1/\delta^2$, finding the best parameter is to find the best g. the change of G value directly determines the mapping function, which affects the complexity of spatial distribution.

After RBF kernel function calculation, the matrix can be transformed into

$$\begin{bmatrix} A, E \\ E^T, 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \qquad (9)$$

where $A = k + V$ and e represent a matrix with all elements of 1. SVR regression analysis model can be obtained, and the formula is

$$f(x) = \lim_{N \longrightarrow \infty} \sum_{i=1}^{N} a_i k(x, x_i) + b. \qquad (10)$$

It can be seen that after processing, there are only two parameters to be optimized: the regularization. The function of the regularization parameter C is to adjust the range of information and reduce the risk. By adjusting the ratio of the two, the generalization ability of the learning machine can be adjusted. If the regularization parameter C is too large, it can only reduce the empirical risk. If it is too small, it will increase the empirical risk. δ mainly controls the width of Gaussian function and data range. If δ value is too large, the risk will increase. If δ value is too small, the structural risk will increase. In the optimization analysis of these two parameters, the following algorithm is adopted.

This algorithm is used for real number solution. It has strong universality, simple principle, few parameters, and strong collaborative [24]. This algorithm is based on the foraging of birds. Each individual is regarded as a particle. It is assumed that the position of the particle is represented by (C, g), the speed of the particle is represented by v_i, and the best position of the particle is represented by p_i. The particle adjusts its position through its current position, neighbor position, and empirical position and realizes the continuous updating of the position through the equation:

$$v_{ij}^{t+1} = w v_{ij}^{t} + c_1 r_1 \left(p_{ij}^{t} - x_{ij}^{t}\right) + c_2 r_2 \left(p_{ij}^{t} - x_{ij}^{t}\right), \qquad (11)$$

where w represents the inertia factor, which is a positive number, c represents the acceleration constant, and r is the random transformation number, ranging from 0 to 1.

When optimizing parameters by least square method, the optimization performance is closely related to regularization parameters and radial basis function kernel function [25–27]. When using particle swarm optimization algorithm, it can adjust its position and rely on its own experience. Therefore, the direction of iteration is stronger, and the global search ability is improved. Support vector machine algorithm is used to regression analyze the training set, find the best parameters, and then bring the training set into the best parameter model [28–30]. In the optimization using the least square method, the fitness parameter is the content that needs to be determined. Considering that each particle reflects a set of parameters, in the particle swarm optimization design, the fitness function is used to change the fitness. The fitness function selects the mean square error function, and the formula is expressed as

$$\mathrm{MSE} = \lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2, \qquad (12)$$

where n represents the training sample set. Through calculation, it can be obtained that the fitness is a positive integer. The smaller this value is, the higher the fitness is. When the fitness can meet the accuracy requirements, the optimal solution is the optimal parameter.

Although the particle swarm optimization algorithm has obvious advantages, it also has some problems. The search accuracy is insufficient to ensure the global optimal solution. It has strong dependence on specific empirical parameters and lacks independence. Therefore, in the research, the genetic algorithm is used to further optimize the parameters, and the regularization parameters and radial basis kernel function optimized by the least square method are genetically coded, The range of regularization parameters is 0~1. Using genetic algorithm to realize parameter optimization, this problem can be simply understood as chromosome, binary coding the parameters, randomly generate SVR parameter values, select the parameters with high fitness according to the survival of the fittest principle, and get better parameters through cross coding. With the further evolution of genetic algorithm, the parameters with the highest fitness can be obtained. The application of the value range of regularization parameters and kernel function parameters is determined first, and then they are binary coded. Then, the mean square error function is used as the fitness function, and the chromosome is decoded to generate the random initial population. Judge whether the population optimization performance meets the optimization requirements. If so, output the optimal parameters and establish the support vector machine model. If not, continue the optimization algorithm and update the population until the requirements are met.

Genetic algorithm has a large search range; it can compare multiple individuals in parallel. It has simple operation and strong expansibility. However, this algorithm is difficult

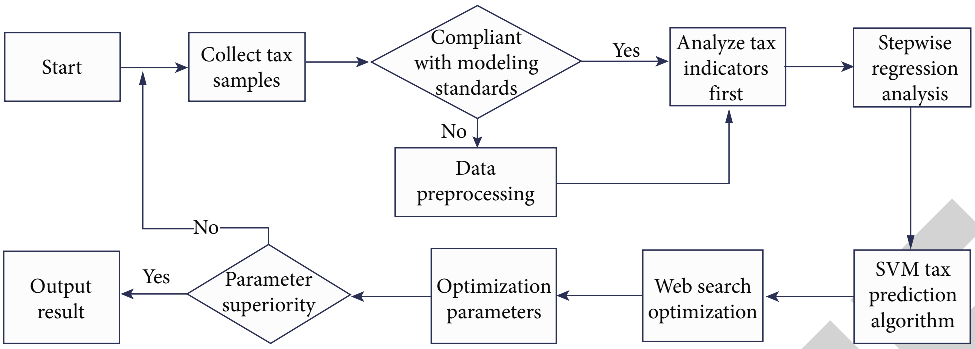

Figure 2: Network search optimization prediction process.

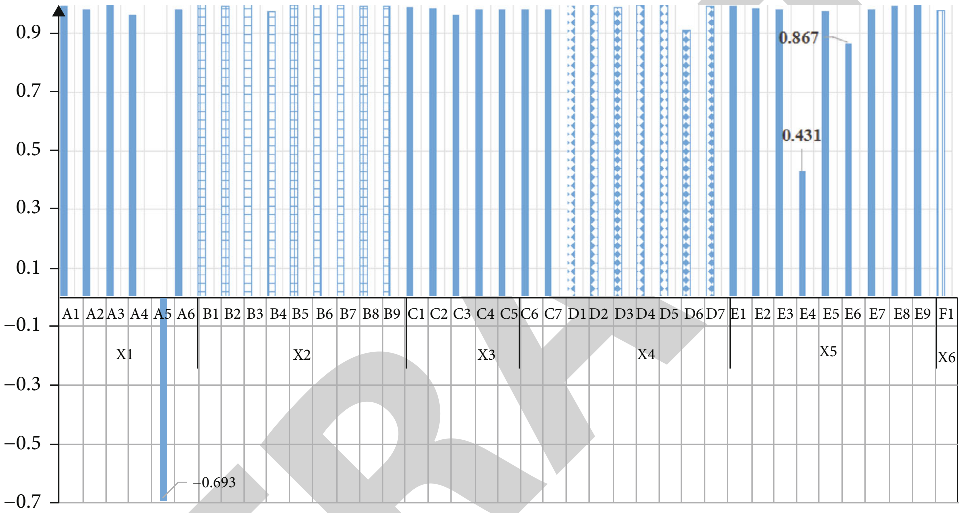

Figure 3: Correlation coefficient analysis between tax and secondary indicators.

to implement in programming, and it needs to continue to calculate after obtaining the optimal solution. The selection of parameters will affect the crossover rate and mutation rate. Moreover, the genetic algorithm cannot effectively feedback the network information, and the search speed is slow. Therefore, the method based on grid search is further used for parameter optimization. Grid search algorithm can search all possible values in the range. Compared with particle swarm optimization algorithm, it is more comprehensive and needs less parameters to be optimized. In the application, the network parameters are generated according to the parameter range and step size of the kernel function and the parameter direction, so as to ensure that all parameter sequences can be analyzed and shrink. Determine the training samples and test samples, test them at the same time, output the errors corresponding to all parameters, and judge whether the error results meet the required accuracy. Reset the parameters and step size, carry out the second search process, and continue this step to find the parameters with the highest accuracy. The specific process is shown in Figure 2. The tax index sequence is represented by X, and the C range and search step are set. This algorithm can avoid blindness. After obtaining the optimal structure, the tax forecast can be determined.

4. Result Analysis and Discussion

4.1. Data Source and Preprocessing. In the tax forecast analysis, the influencing factors need to be determined first. At present, in the domestic statistical yearbook, taxes are divided into six categories. Therefore, these six categories of taxes are taken as the primary research indicators in the research, namely value-added tax (x1), business tax (x2), consumption tax (x3), personal income tax (x4), enterprise income tax (x5), and tariff (x6). The vertical coordinate in Figure 3 is the correlation coefficient of tax. The relevant indicators contained in these taxes are used as secondary indicators, and the secondary indicators are coded. The value-added tax (x1) contains 6 secondary indicators, with the GDP code of A1, the total import value code of A2, the industrial added value code of A3, the retail sales code of A4, the proportion code of industrial added value of A5, and the added value code of A6. The units of these secondary indicators are 100 million yuan. There are 9 secondary indicators of business tax. The code of highway freight volume is B1, the code of highway passenger volume is B2, the code of added value of construction industry is B3, and so on. The code of business income of catering owners above the limit is B9. There are 7 secondary indicators of

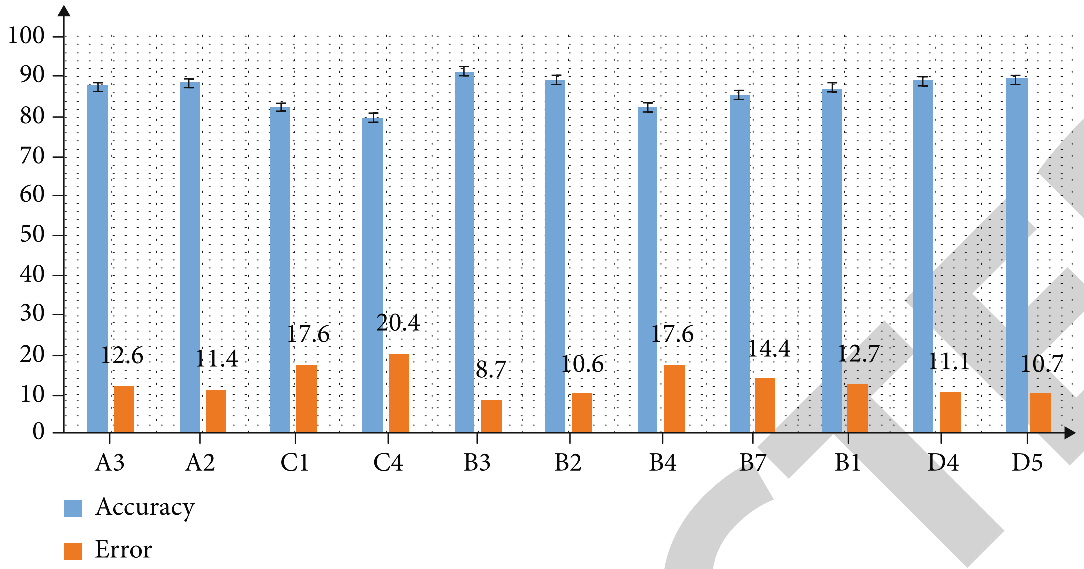

Figure 4: Accuracy and error analysis.

consumption tax, represented by C, 7 secondary indicators of personal income tax, represented by D, 9 secondary indicators of enterprise income tax, represented by E, and the secondary indicator of tariff is the total import and export volume, represented by $F1$.

There are often many indicators and insufficient historical data in economic forecasts. Using neural network to predict such a problem is a typical small sample prediction of large-scale neural network. This situation weakens the generalization ability of neural network. In this case, scholars at home and abroad usually divide these indicators into several subsystems according to their relationship to alleviate the problem of excessive network scale. However, the division of subsystems is very complex, which relies too much on the analysis of the operation mechanism of economic system, and cannot fundamentally solve the problem of multi-index and small sample complex system prediction. In fact, these indicators are often relevant. Therefore, it is necessary to reduce the number of indicators (dimensionality reduction) on the premise of minimizing information loss. Principal component analysis, as a dimension reduction processing technique, starts from reducing the number of input nodes of neural network. It can fundamentally reduce the scale of neural network and improve the generalization ability of neural network in multi index small sample problems.

Considering that there are too many influencing factors of tax prediction, in the correlation analysis, set the limit of correlation coefficient as 0.9 and eliminate the indicators with small correlation. Taking the tax revenue data of China Statistical Yearbook as the data sample, the correlation analysis is carried out by using statistical software. Among the six primary indicators, the correlation with tax is 0.992, 0.991, 0.984, 0.985, 0.994, 0.997, and 0.982, respectively. From the correlation analysis results, it can be seen that the primary indicators have a great impact on tax, and the correlation coefficients exceed 0.95. Continue to analyze the correlation between primary indicators and secondary indicators, as shown in Figure 3. The correlation coefficient between VAT and its secondary indicators is quite different. Except for A5 indicator, the correlation coefficient of other indicators is more than 0.98. The correlation coefficient between business tax and its secondary indicators exceeds 0.99, and all indicators will be analyzed in the next step. The correlation coefficients of consumption tax, individual income tax and their secondary indicators are relatively high, all above 0.96. There is a certain difference in the correlation coefficient between enterprise income tax and its secondary indicators. The correlation coefficient of E4 indicator is only 0.431, excluding this indicator. The correlation coefficient between tariff and secondary indicators is 0.976. It can be seen from the data in the figure that most indicators are highly correlated with their respective tariffs, but there are also negative correlation indicators. Eliminate the negative correlation indicators and the secondary indicators with correlation coefficient less than 0.9.

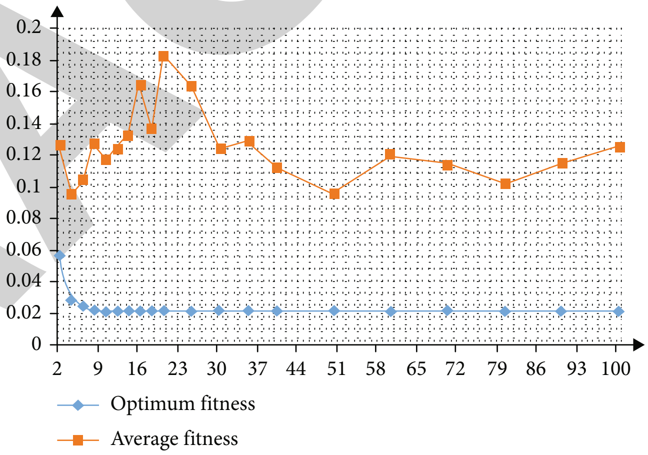

Figure 5: Particle swarm optimization simulation analysis.

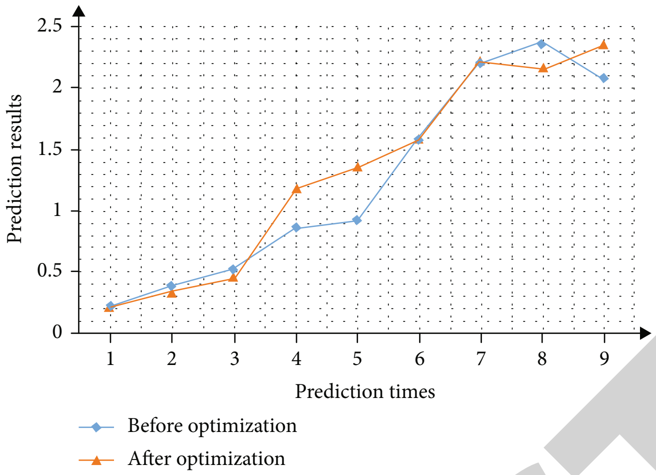

FIGURE 6: Prediction results of particle swarm optimization algorithm.

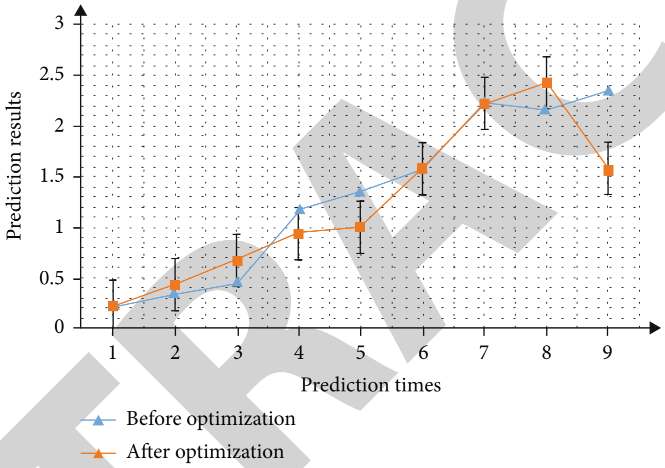

FIGURE 7: Prediction results of genetic algorithm.

Through correlation analysis, excluding the indicators with small correlation coefficient, there are still many remaining tax indicators, so stepwise regression analysis is also needed. Set a dependent variable, analyze the effect of other independent variables on it, and then sort and select the factors with great influence. The original data is collected and processed for simulation analysis. The accuracy and error results are shown in Figure 4. The prediction accuracy has been improved after stepwise regression analysis.

The data dimensions of different indicators are different, so the tax cannot be predicted directly. Therefore, the data need to be standardized. The formula is

$$\bar{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{13}$$

where x_i represents the data of the indicator column, $x_{\max}$ represents the maximum value of the data of the indicator column, and $x_{\min}$ represents the minimum value. After this formula, the range of data is controlled within 0~1.

When the parameters are optimized by the least square method of particle swarm optimization, the range of parameters C and δ^2 is 0~100, with 100 iterations and 20 population evolution. Through experimental simulation, the optimal value of C is 1.032, and the optimal value of δ^2 is 0.01. In order to evaluate the effectiveness of tax forecast, the average absolute percentage error and mean square percentage error are used for analysis. The formulas are as follows:

$$\begin{aligned} \text{MAPE} &= \lim_{n\longrightarrow\infty} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \bar{y}_i)}{y_i}\right|, \\ \text{MSPE} &= \lim_{n\longrightarrow\infty} \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left[\frac{(y_i - \bar{y}_i)}{y_i}\right]^2}, \end{aligned} \tag{14}$$

where MAPE represents the average absolute percentage error and MSPE represents the mean square percentage error.

4.2. Simulation Analysis of Support Vector Machine Optimization. In the simulation analysis, the regularization parameter range is set to 0.1~100, kernel parameter range is set to 0~10, the cross validation is 5 times, the maximum population iteration is 100 times, and the population size is 20. And the results are shown in Figure 5. And the optimal

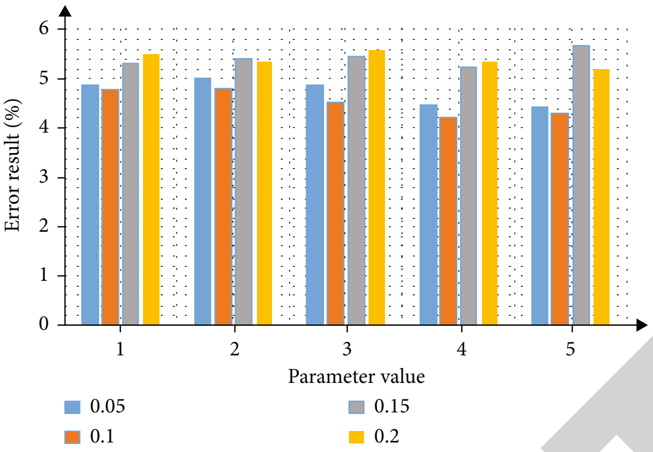

Figure 8: Error results under different parameters.

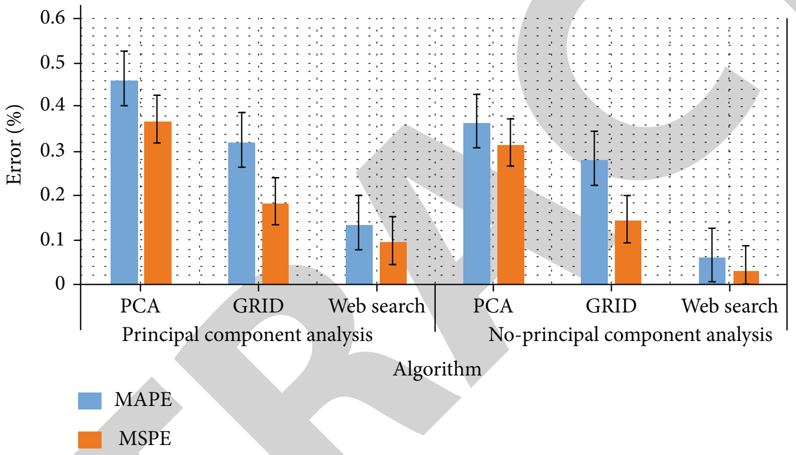

Figure 9: Analysis of application results of principal component analysis.

fitness basically remains unchanged after 20 iterations, indicating that the optimal value of regularization parameter is 1.184 and the optimal parameter of radial basis kernel function is 0.1.

The statistical results are shown in Figure 6. It can be seen that the predicted value is highly consistent with the actual value, indicating a high prediction effect.

Then genetic algorithm is used to optimize. The range of regularization parameters is 0~100, the range of radial basis kernel function parameters remains unchanged, the interactive verification is 5 times, and the maximum iteration is 299 times. Similarly, MATLAB software is used for simulation analysis. It is found that after 4 times, the average fitness function remains unchanged, and the best fitness function also tends to be stable. The best regularization parameter value is 4.457, and the best radial basis kernel function parameter is 0.0175. The obtained parameters are applied to the support vector machine prediction model, and the prediction results are shown in Figure 7.

In grid search optimization, the number of samples will affect the simulation results. Combined with the actual situation, the parameter defaults to $1/n$. Select the values around this value for simulation analysis. The test results are shown in Figure 8. As can be seen from the figure, when the parameter value is 0.1, the accuracy is relatively high.

Under the optimal parameter model, the application effect of principal component analysis is compared and analyzed. The results are shown in Figure 9. It can be seen from the data in the figure that after parameter optimization, the model algorithm error decreases, and the model prediction results of the three methods are improved. After the prediction results of the principal component analysis model, the average absolute percentage error also decreased significantly, indicating that the tax prediction of the data after the redundancy reduces the data dimension and improves the prediction accuracy.

5. Conclusion

Tax revenue is closely related to people's life. Accurate prediction of tax revenue is of great significance to regulate local policies. Considering the tax changes and influencing factors, it is difficult to accurately judge it by using linear analysis alone. Based on this, this paper constructs a support vector machine model in tax prediction and analysis. This machine learning algorithm has great advantages in dealing with nonlinear problems. Considering that the regularization function and radial basis kernel function of support vector machine have a great impact on the tax prediction results, other algorithms are used to optimize these two parameters, and the application effect of principal component analysis method and the accuracy of tax prediction after parameter optimization are analyzed. The experimental results show that the principal component analysis method can reduce the redundancy and reduce the data dimension. The model optimized by the algorithm is applied to tax prediction. The predicted value has a high degree of fit with the actual value, the prediction effect is good, and the accuracy is significantly improved.

However, the independent variable and dependent variable in the actual tax problem are linear. Taxation is affected by many aspects, and its system is complex and uncertain. The relationship between independent variables and dependent variables is mostly nonlinear, and the prediction accuracy still has room for optimization. The following contents need to be further modified in future research. Only one kernel function is selected in this paper. There are many kinds of kernel functions. Without analyzing other types of kernel functions, parameter settings will also directly affect the analysis results. When searching and optimizing parameters in the grid, you can also improve the parameter search process, take tax samples as the search scope, and further narrow the parameter search scope.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Shandong Province Key R&D Program (Soft Science Project): Research on the vertical price stabilization and income increase mechanism of the whole industrial chain of geographical indication agricultural products based on rural revitalization (No. 2021RKY06110).

References

[1] L. M. Mauler, “The effect of analysts’ disaggregated forecasts on investors and managers: evidence using pre-tax forecasts,” *Accounting Review*, vol. 94, no. 3, pp. 279–302, 2019.

[2] A. Habib, M. M. Hasan, and A. Al-Hadi, “Financial statement comparability and corporate cash holdings,” *Journal of Contemporary Accounting & Economics*, vol. 13, no. 3, pp. 304–321, 2017.

[3] O. H. Kwang-Wuk and K. I. Eun-Sun, “Effect of tax-related information on pre-tax income forecast and value relevance,” *Journal of Asian Finance Economics and Business*, vol. 7, no. 1, pp. 81–90, 2020.

[4] J. L. Ye, “The effects of analysts’ tax expense forecast accuracy on corporate tax avoidance: an international analysis,” *Journal of Contemporary Accounting and Economics*, vol. 17, no. 2, article 100243, 2021.

[5] A. H. Miller, L. Alkindi, and A. Alblooshi, “Using a database approach, with big data and unsupervised machine learning to model tax behavior in the expatriate community,” *Solid State Technology*, vol. 63, no. 2, pp. 379–389, 2020.

[6] H. D. Aprilia and D. Agustiani, “Application of data mining using the K-means algorithm in rural and urban land and building tax (PBB-P2) receivables data in Bantul Regency,” *Journal of Physics: Conference Series*, vol. 1823, no. 1, pp. 12–19, 2021.

[7] F. Sun, Z. Wang, and Z. Li, “A tax dynamic clustering method based on weak convergence sequence coefficient,” *Boletin Tecnico/Technical Bulletin*, vol. 55, no. 5, pp. 216–223, 2017.

[8] W. Didimo, L. Grilli, G. Liotta, L. Menconi, F. Montecchiani, and D. Pagliuca, “Combining network visualization and data mining for tax risk assessment,” *IEEE Access*, vol. 8, pp. 16073–16086, 2020.

[9] S. L. Li, “Data mining of corporate financial fraud based on neural network model,” *Computer Optics*, vol. 44, no. 4, pp. 665–670, 2020.

[10] M. Mwanza and J. Phiri, *Fraud Detection on Big Tax Data Using Business Intelligence, Data Mining Tool: a Case of Zambia Revenue Authority, [Ph.D. thesis]*, University of Zambia, Zambia, 2016.

[11] A. H. Miller, “Data modeling and visualization of tax strategies employed by overseas American individuals and firms,” *Individuals and Firms*, vol. 29, pp. 5–28, 2019.

[12] P. Battiston, S. Gamba, and A. Santoro, “Optimizing tax administration policies with machine learning,” *Working Papers*, vol. 7, pp. 436–447, 2020.

[13] Y. Hao, “Digital inclusive finance risk prevention based on machine learning and neural network algorithms,” *Journal of Intelligent and Fuzzy Systems*, vol. 2, pp. 1–11, 2021.

[14] Y. He, K. F. Tsang, Y. C. Kong, and Y. T. Chow, “Indication of electromagnetic field exposure via RBF-SVM using time-series features of zebrafish locomotion,” *Sensors*, vol. 20, no. 17, p. 4818, 2020.

[15] I. Unceta, J. Nin, and O. Pujol, “Risk mitigation in algorithmic accountability: the role of machine learning copies,” *PLoS One*, vol. 15, no. 11, pp. 241–286, 2020.

[16] P. F. Dai, X. Xiong, and W. X. Zhou, “A global economic policy uncertainty index from principal component analysis,” *Finance Research Letters*, vol. 40, no. 4, article 101686, 2020.

[17] Y. Zhu, L. Zhou, C. Xie, G. J. Wang, and T. V. Nguyen, “Forecasting SMEs’ credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach,” *International Journal of Production Economics*, vol. 211, no. 5, pp. 22–33, 2019.

[18] P. Li, Y. Peng, P. Jiang, and Q. Dong, “A support vector machine based semiparametric mixture cure model,” *Computational Statistics*, vol. 35, no. 3, pp. 931–945, 2020.

[19] D. Zhao, J. Ding, and S. Chai, “Systemic financial risk prediction using least squares support vector machines,” *Modern Physics Letters B*, vol. 32, no. 17, article 1850183, 2018.

[20] J. Zhu, P. Sun, Y. Gao, and P. Zheng, “Clock differences prediction algorithm based on EMD-SVM,” *Chinese Journal of Electronics*, vol. 27, no. 1, pp. 128–132, 2018.

[21] K. Adhikary, S. Bhushan, S. Kumar, and K. Dutta, "Evaluating the performance of various SVM kernel functions based on basic features extracted from KDDCUP'99 dataset by random forest method for detecting DDoS attacks," *Wireless Personal Communications*, vol. 123, no. 4, pp. 3127–3145, 2022.

[22] A. E. Kitali, S. Mokhtarimousavi, C. Kadeha, and P. Alluri, "Severity analysis of crashes on express lane facilities using support vector machine model trained by firefly algorithm," *Traffic Injury Prevention*, vol. 22, no. 1, pp. 79–84, 2021.

[23] X. Tao, Q. Li, C. Ren et al., "Affinity and class probability-based fuzzy support vector machine for imbalanced data sets," *Neural Networks*, vol. 122, pp. 289–307, 2020.

[24] S. Dong, Y. Zhang, Z. He, N. Deng, X. Yu, and S. Yao, "Investigation of support vector machine and back propagation artificial neural network for performance prediction of the organic Rankine cycle system," *Energy*, vol. 144, no. 1, pp. 851–864, 2018.

[25] B. Zhu, S. Ye, P. Wang, J. Chevallier, and Y.-M. Wei, "Forecasting carbon price using a multi-objective least squares support vector machine with mixture kernels," *Journal of Forecasting*, vol. 41, no. 1, pp. 100–117, 2022.

[26] J. Kang, L. Liu, S. D. Zhou, D. Y. Wang, and Y. C. Ma, "A novel recursive modal parameter estimator for operational time-varying structural dynamic systems based on least squares support vector machine and time series model," *Computers & Structures*, vol. 229, article 106173, 2020.

[27] Y. Guo, Z. Zhang, and F. Tang, "Feature selection with kernelized multi-class support vector machine," *Pattern Recognition*, vol. 117, no. 5324, article 107988, 2021.

[28] G. Sharma, A. Panwar, I. Nasiruddin, and R. C. Bansal, "Non-linear LS-SVM with RBF-kernel-based approach for AGC of multi-area energy systems," *IET Generation, Transmission & Distribution*, vol. 12, no. 14, pp. 3510–3517, 2018.

[29] A. Nah, B. Ari, and C. Fs, "An optimized support vector machine (SVM) based on particle swarm optimization (PSO) for cryptocurrency forecasting," *Procedia Computer Science*, vol. 163, pp. 427–433, 2019.

[30] C. Zhang, Q. W. Gong, T. Wang, and K. Koyamada, "Visual extraction system for insulators on power transmission lines from UAV photographs using support vector machine and color models," *Journal of Visualization*, vol. 23, no. 6, pp. 1101–1112, 2020.