Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **3**, August **2025**

Regional Data Factor Market Forecasting Based on Non- Parallel Hyperplane Support Vector Machines

Tian-Tian Yang[1,*], Hai-Wei Yang[1,2], Xiao-Ming Lou[1], Yan-Fei Lv[3]

[1]College of Business, Jinhua University of Vocational Technology, Jinhua 321017, P. R. China
vera_yangtiantian@163.com 37265491@qq.com

[2]UCSI University, Kuala Lumpur 56000, Malaysia
1002371054@ucsiuniversity.edu.my

[3]College of Information Engineering, Jinhua University of Vocational Technology, Jinhua 321017, P. R. China
luyanfei78@qq.com

[*]Corresponding author: Tian-Tian Yang

Received July 8, 2024, revised December 1, 2024, accepted March 23, 2025.

Abstract. *Regional data factor market forecasting is of great importance in economic development and policy making, but traditional forecasting methods often show limitations in dealing with high dimensionality, noise and complex data distributions. In this paper, an improved Non-Parallel Hyperplane Support Vector Machine (NPSVM) model is proposed, aiming to enhance the prediction performance of regional data factor markets. Firstly, the computational efficiency and prediction accuracy of the model are improved by the PLS feature extraction method, which effectively reduces the data dimensionality and retains the most relevant information to the prediction target. Second, an anti-noise mechanism based on the absolute loss of L1 paradigm and the improved Hinge Loss function is introduced to enhance the robustness and stability of the model under noisy data. Finally, by optimising the mean and variance of the intervals, the NPSVM model is able to better capture and utilise the overall distributional characteristics of the data, and improve the adaptability to different regional and economic environments. The experimental results show that the improved NPSVM model significantly outperforms the traditional SVM and TSVM models in regional data factor market prediction, with an average error rate of only 0.023, providing higher prediction accuracy. The research in this paper provides an effective solution to the challenge of forecasting complex economic data, which helps to improve the science and reliability of economic decision-making.*
Keywords: Data factor markets; Prediction models; NPSVM; Partial least squares; Anti-noise mechanisms

1. **Introduction.** In modern economy, the forecasting of regional data factor market [1, 2] is of great significance for regional economic development and policy formulation. Regional data factor markets cover key factors such as land, labour, capital and technology, and dynamic changes in these factors directly affect the overall performance of regional economies [3]. Accurately predicting the changing trends of these market factors can provide powerful support for governments and enterprises when making policy and strategic decisions [4]. For example, the supply and demand of land can influence the development of the real estate market, changes in the supply and demand of labour can affect the employment rate and wage level, and the liquidity of capital affects the stability of investment and financial markets [5, 6]. With the intensification of global economic integration, the forecasting of regional data factor markets is not only valuable for regional

economic management, but also key to understanding and responding to global economic changes.

Support Vector Machine (SVM), as a classical machine learning algorithm, has demonstrated strong performance in classification and regression problems. SVM maximises the spacing between different classes of data by constructing an optimal hyperplane to achieve accurate classification and regression analysis. This approach is particularly suitable for dealing with datasets with complex relationships and high-dimensional features. In regional economic forecasting, SVM can effectively capture the nonlinear interactions among factors such as land, labour, capital and technology [7, 8, 9]. In addition, SVM's kernel function technique allows operation in high-dimensional spaces, enabling it to handle nonlinearly distributed data, which is especially critical for dynamically changing and diverse regional economic factor markets. However, SVMs may face some challenges when dealing with high-dimensional and complex data, such as the non-linear and highly noisy nature of the data. In order to solve these problems, Non-Parallel Support Vector Machine (NPSVM) has emerged [10, 11]. NPSVM makes the model more flexible in dealing with complex distributional structures by constructing two non-parallel categorical hyperplanes and defining an optimal hyperplane for each category separately and adaptability [12]. This improvement makes NPSVM have great potential for application in regional data element market forecasting, which can better capture the multidimensional features and nonlinear relationships of data.

The research objective of this paper is to explore and validate the effectiveness and superiority of a regional data element market forecasting model based on an improved non-parallel hyperplane support vector machine. It is also compared with the traditional SVM and TSVM models to demonstrate its potential advantages in practical applications.

1.1. **Related work.** Time series forecasting is a method of predicting future data values by analysing time dependencies in historical data. Time series forecasting methods are widely used to analyse economic indicators, market trends and other key variables in the forecasting of regional data factor markets. Traditional time series forecasting methods mainly include Auto-Regressive (AR) model [13], Moving Average (MA) model [14], and Auto-Regressive Integral Moving Average (ARIMA) model [15]. These models perform well in handling time series data with linear relationships.

In the field of economic forecasting, Abbasimehr and Paki [16] proposed a hybrid time series forecasting method combining statistical modelling and machine learning techniques, which improves the accuracy and stability of the overall forecasts by integrating the forecasting results from multiple models. Although hybrid models perform well in practice, their complex model architecture and parameter tuning process remain a challenge. Although time series forecasting methods have achieved remarkable results in regional data factor market forecasting, the limitations of these traditional methods have gradually emerged in the face of highly noisy, high-dimensional and non-linear data. Therefore, more flexible and adaptable models are needed to cope with the complex data characteristics of the modern economic environment.

Currently, prediction research in data factor markets focuses on how to deal with high-dimensional, noisy and dynamically changing data. In order to improve the accuracy of prediction and the generalisation ability of the model, researchers have explored a lot in the aspects of feature extraction, model robustness and data distribution processing, etc. Sarstedt et al. [17] adopted the feature extraction method based on Partial Least Squares (PLS) in regional economic prediction, and found that PLS can effectively reduce the dimensionality of the data while retaining the feature information which is crucial for the prediction results. It is found that PLS can effectively reduce the dimensionality of the

data while retaining the feature information that is essential for the prediction results. This method performs well on high-dimensional data, but the effectiveness of feature extraction may be limited when dealing with heterogeneous data. He et al. [18] used a Principal Component Analysis (PCA)-based feature extraction method in the prediction of regional economic development, and the results show that PCA can simplify the processing of high-dimensional data, but may not be as good as PLS at retaining important information. These studies highlight the need for a more comprehensive approach to feature extraction in complex data environments. highlight that how to balance the relationship between simplification of feature extraction and information retention in complex data environments is an issue that requires in-depth research. In addition, Beniwal et al. [19] proposed a parameter optimisation method based on Genetic Algorithm (GA) for optimising the parameters of SVM. Experimental results show that GA can effectively search for the optimal parameter combinations and improve the prediction performance of the model. However, the high computational complexity of GA may limit its application on large-scale datasets, and it needs to be combined with other optimisation methods to improve the efficiency.

1.2. **Motivation and contribution.** Existing market forecasting methods for regional data elements often struggle to achieve desirable forecasts in the face of high dimensionality, noise and complex data distributions. Traditional time series models and machine learning methods often rely on linear assumptions or specific data structures, which makes them exhibit limitations when dealing with dynamically changing and multidimensional features of regional data. In addition, many models suffer from significant degradation in prediction accuracy when confronted with highly noisy data, and lack effective anti-noise mechanisms to deal with outliers and disturbances in the data. To address these issues, this paper proposes an improved NPSVM model that aims to enhance the effectiveness and robustness of regional data factor market prediction.

(1) To address the challenges of high-dimensional data in prediction, this paper introduces the PLS method to extract features from the data. PLS extracts the most relevant features to the prediction target by maximising the correlation between the independent and dependent variables, thus reducing the data dimensionality and retaining the key information. This improvement not only simplifies the input dimensions of the model, but also reduces the computational complexity of the

(2) To address the problem of high noise in regional data, this paper designs an anti-noise mechanism based on the absolute loss of L1 paradigm and the improved Hinge Loss function. The absolute loss of L1 paradigm can reduce the sensitivity of the model to large deviation samples, thus enhancing the stability of the model under noisy data. The improved Hinge Loss function further enhances the robustness of the model by controlling the effect of the loss function on sample points far from the classification hyperplane.

(3) In order to improve the model's ability to handle the complexity of data distribution, this paper introduces an optimisation method for interval distribution information in the NPSVM model. By simultaneously maximising the average interval from the sample points to the classification hyperplane and minimising the variance of the interval, the method proposed in this paper is able to better capture and utilise the overall distributional characteristics of the data. This optimisation not only improves the stability of the model's classification decision, but also enhances the model's ability to adapt to different regions and different economic environments, which makes the model perform better when dealing with economic data with significant regional and heterogeneous characteristics.

2. **Principles of market forecasting for regional data elements.**

2.1. **Forecasting characteristics.** In data factor market forecasting, the main challenge we face is how to accurately predict the key factors in regional economic activities. These factors include, but are not limited to, land, labour, capital, technology, etc., which together affect the level of economic development and market dynamics of a region. Effective forecasting of these factors is the basis for economic policy development and market analysis. The following are some of the important features of market forecasting for regional data elements:

(1) Complex multi-dimensional data characteristics

The prediction of regional data factor markets involves a large amount of complex multidimensional data [20]. The data dimensions include not only the time series dimension, but also the geospatial, economic and social dimensions. The existence of such multidimensional data complicates feature extraction and modelling of the data. In this case, how to effectively process and extract features of multidimensional data becomes critical. Mathematically, the multidimensional data can be represented by the matrix X.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{1}$$

where x_{ij} denotes the jth eigenvalue of the ith sample, m is the number of samples, and n is the number of features.

(2) Dynamism and uncertainty

The dynamic and uncertain nature of regional economies and markets is one of the main challenges for forecasting models. Regional markets are affected by a variety of internal and external factors, which include policy changes, technological advances, fluctuations in the global economy, and so on. Forecasting models must be able to adapt and capture such dynamic changes in order to provide accurate forecasts. To represent the dynamic changes in a time series, a time series vector $y(t)$ can be defined to represent the target variable at time t.

$$y(t) = (y_1(t), y_2(t), \ldots, y_p(t)) \tag{2}$$

where $y_i(t)$ denotes the value of the ith element at time t; p is the number of elements.

(3) Regionality and heterogeneity

Data factor markets in different regions are significantly regional and heterogeneous, and there may be large differences in the level of economic development, resource distribution and market demand between regions. Such regionality and heterogeneity require that the forecasting model be flexible and adaptable enough to be able to deal with the specific conditions of different regions. Regional characteristics can be captured by introducing a regional feature vector z into the forecasting model.

$$z = (z_1, z_2, \ldots, z_q) \tag{3}$$

where z_j denotes the jth area feature; q is the number of area features.

(4) Data sparsity and noise

The issues of data sparsity and noise in market forecasting for regional data elements also require special attention. Due to the complexity of the data collection process and the market, there may be missing values, outliers or noisy data in the dataset [21]. These problems can affect the training and prediction accuracy of the model, so anti-noise mechanisms and data preprocessing methods need to be introduced into the model. The noise

in the observed data can be denoted as e, assuming that e is a random noise vector, and its effect can be described by the following model:

$$y = Xw + e \tag{4}$$

where w is the vector of parameters to be estimated; y is the vector of observations.

(5) Timeliness and real-time

Forecasts for regional data element markets typically need to be highly current and real-time. As the market environment changes rapidly, forecasting models need to be able to update and respond quickly to provide up-to-date forecasts. This requires models that not only have high prediction accuracy, but also low computational complexity to be able to make predictions in a short period of time. The need for real-time can be realised by online learning algorithms which continuously update the model parameters in the data stream to adapt to the latest data changes. The process of updating model parameters can be described by the following online updating method.

$$w_{t+1} = w_t - \eta \nabla L\left(w_t, x_t, y_t\right) \tag{5}$$

where w_t is the parameter vector at time t; η is the learning rate; and ∇L is the gradient of the loss function.

The forecasting of regional data factor markets is characterised by multidimensionality, dynamism, regionality, data sparsity and timeliness. These features make the prediction task complex and challenging. When constructing a prediction model, we need to consider these features comprehensively and adopt appropriate feature extraction, anti-noise processing and efficient computational methods to improve the prediction accuracy and adaptability of the model. In the follow-up, we will design an improved NPSVM model based on these features to solve the prediction problem of regional data factor markets.

2.2. **Classification of forecasts.** In regional data element market forecasting, forecasting tasks can be classified into various types depending on the forecasting objectives. Different types of forecasting tasks require different models and methods to handle them in order to meet specific application requirements.

Time series forecasting is a common type of market forecasting for regional data elements and is particularly suitable for modelling and forecasting time dependencies in economic activities. The goal of this type of forecasting task is to use time trends and cyclical features in historical data to predict future values. For example, forecasting regional GDP growth rates or unemployment rates for the next few quarters. The mathematical description of time series forecasting can be represented by the following recursive relationship:

$$y(t+\tau) = f\left(y(t), y(t-1), \ldots, y(t-p), X(t)\right) \tag{6}$$

where $y(t+\tau)$ denotes the target variable at time $t+\tau$; τ is the time step of the prediction; $X(t)$ is the multidimensional feature vector at time t; f is the prediction function to be learnt; and p is the length of the look-back window for historical data.

To deal with nonlinearity and complex time dependencies, common methods include ARIMA and Long Short-Term Memory Networks (LSTM). ARIMA is suitable for linear time series data.

$$y(t) = c + \sum_{i=1}^{p} \phi_i y(t-i) + \sum_{j=1}^{q} \theta_j \epsilon(t-j) + \epsilon(t) \tag{7}$$

where c is the constant term; ϕ_i and θ_j are the autoregressive coefficients and moving average coefficients, respectively; and $\epsilon(t)$ is the error term.

LSTM is used to capture nonlinear models with long time dependencies and its state update and output methods are:

$$h_t = \text{LSTM}(h_{t-1}, x_t) \tag{8}$$

where h_t is the hidden state at time t; x_t is the input feature; and the LSTM function represents a complex nonlinear mapping.

Cross-sectional data forecasting is another important type of forecasting that applies to analysing and forecasting multiple samples at the same point in time. The goal of this type of forecasting task is to use various types of characteristics at the current point in time to make predictions about specific variables for different samples. For example, predicting the level of house prices or the profitability of firms in different regions at a particular point in time. A mathematical model for cross sectional data forecasting can be expressed in the form of:

$$y_i = g(X_i, z_i, \theta) \tag{9}$$

where y_i denotes the predicted value of the ith sample; X_i is the feature vector of the ith sample, and z_i is the regional feature vector; g is the prediction function to be learnt; and θ is the model parameter.

To deal with heterogeneity and complex feature relationships between samples, commonly used methods include linear regression models (LR) and SVM. LR is used for modelling linear relationships.

$$y_i = X_i\beta + \epsilon_i \tag{10}$$

where β is the regression coefficient; ϵ_i is the error term.

SVMs are suitable for dealing with nonlinear relationships, where features are mapped to a higher dimensional space by means of a nonlinear kernel function.

$$f(X) = \sum_{i=1}^{n} \alpha_i K(X_i, X) + b \tag{11}$$

where α_i is the weight of the support vector; K is the kernel function; and b is the bias term.

Time-series forecasts and cross-sectional data forecasts are the two most common types of tasks in market forecasting for regional data elements. Time series forecasts are appropriate for dynamic data with time dependencies, focusing on the exploitation of historical data trends and cycles. Cross-sectional data forecasts, on the other hand, are suitable for analysing and forecasting specific variables for multiple samples at a specific point in time, focusing on the modelling of characteristic relationships, and are therefore more suitable for regional data factor market forecasting tasks.

3. **Overview of non-parallel hyperplane support vector machines.**

3.1. **Support vector machine.**

SVM is a supervised learning method based on statistical learning theory and is widely used in classification and regression analysis tasks. The core idea of SVM is to maximise the spacing between different classes by constructing one or more hyperplanes to achieve effective classification of data, as shown in Figure 1. SVMs have been performing well in a variety of real-world applications due to their strong theoretical foundations and good generalisation performance.

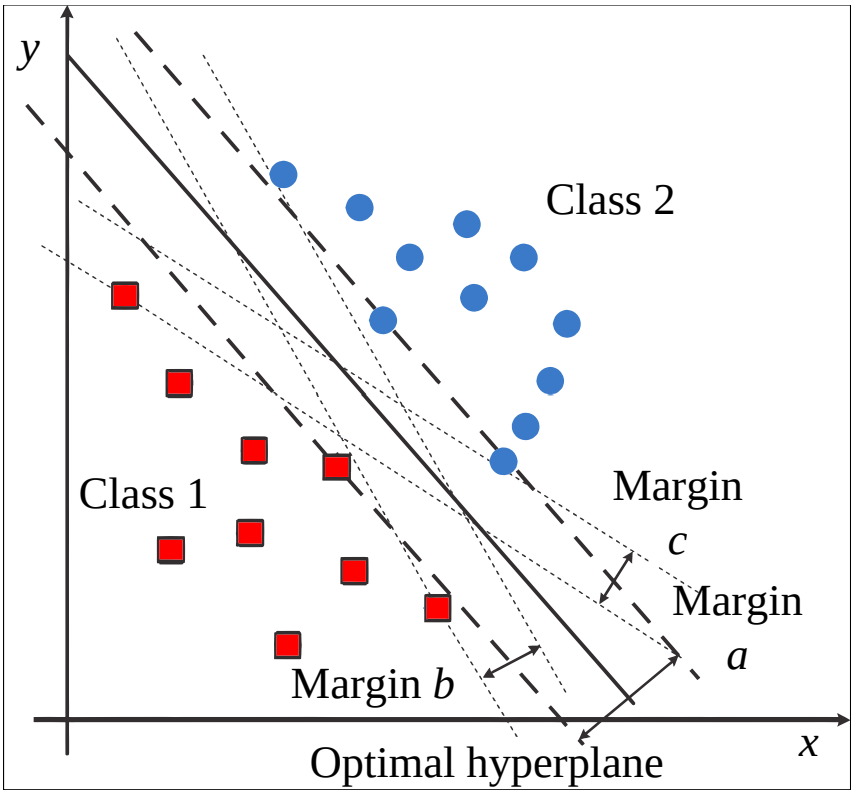

Figure 1. Examples of SVMs

The goal of SVM is to find an optimal hyperplane that divides the sample data into two classes and makes this hyperplane the maximum distance from the nearest sample point of each class. This optimal hyperplane is often referred to as the Maximum Margin Hyperplane [22, 23]. SVM employs different methods and techniques when dealing with linearly and nonlinearly differentiable problems. For linearly differentiable problems, SVM separates the data by hyperplanes of the following form:

$$w \cdot x + b = 0 \tag{12}$$

where w is the normal vector, which represents the direction of the hyperplane; and x is the input eigenvector.

To ensure that the spacing of hyperplanes is maximised, the SVM needs to solve the following optimisation problem.

$$\min_{w,b} \frac{w^2}{2} \tag{13}$$

The following constraints are also satisfied.

$$y_i(w \cdot x_i + b) \geq 1, \qquad i = 1, 2, \ldots, N \tag{14}$$

where y_i is the label of the i-th sample; x_i is the feature vector of the i-th sample; and N is the total number of samples.

The above optimisation problem can be transformed into a dyadic problem by the Lagrange multiplier method, thus simplifying the solution process. The objective function of the dyadic problem is shown as follow:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{15}$$

and satisfies the following constraints.

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \qquad i = 1, 2, \ldots, N \tag{16}$$

where α_i is the Lagrange multiplier, and by optimising these multipliers we can find the optimal hyperplane parameters w and b.

3.2. **TSVM.** TSVM is an improved support vector machine model [24, 25] designed to improve classification efficiency and accuracy by constructing two non-parallel classification hyperplanes. By solving two smaller-scale optimisation problems, TSVM is able to handle classification tasks more efficiently and performs especially well in large-scale datasets and high-dimensional data. Unlike traditional SVM, TSVM performs classification by simultaneously constructing two non-parallel hyperplanes. Specifically, TSVM constructs a hyperplane for each class of data, so that the data points of one class are close to its corresponding hyperplane and at the same time far away from the hyperplane of the other class. This approach allows TSVM to effectively reduce computational complexity while improving the flexibility and robustness of classification.

The two classification hyperplanes of TSVM can be represented as:

$$\min_{w_1, b_1, \xi} \frac{1}{2} w_1^2 + C \sum_{i \in \text{class } 1} \xi_i \tag{17}$$

The following constraints are also satisfied:

$$A_1(w_1 \cdot x_i + b_1) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i \in \text{class } 1 \tag{18}$$

$$A_2(w_1 \cdot x_i + b_1) \leq -1 + \xi_i, \quad \xi_i \geq 0, \quad i \in \text{class } 2 \tag{19}$$

where A_1 and A_2 denote the sample matrices for the first and second classes, respectively; and ξ_i is a slack variable to allow some sample points to be outside the interval.

Similarly, for the second type of data, TSVM determines its hyperplane by the following optimisation problem [26]:

$$\min_{w_2, b_2, \eta} \frac{1}{2} w_2^2 + C \sum_{i \in \text{class } 2} \eta_i \tag{20}$$

The following constraints are also satisfied:

$$A_2(w_2 \cdot x_i + b_2) \geq 1 - \eta_i, \quad \eta_i \geq 0, \quad i \in \text{class } 2 \tag{21}$$

$$A_1(w_2 \cdot x_i + b_2) \leq -1 + \eta_i, \quad \eta_i \geq 0, \quad i \in \text{class } 1 \tag{22}$$

where A_1 and A_2 denote the sample matrices of the first and second categories, respectively; η_i is another set of relaxation variables.

By solving these two optimisation problems, the TSVM can simultaneously determine two classification hyperplanes such that they are each close to their respective categories while being far away from the data points of the other category. This dual optimisation not only improves the classification accuracy but also reduces the computational complexity.

3.3. **NPSVM.** NPSVM is an improvement of traditional SVM and TSVM [27]. NPSVM further enhances the flexibility and adaptability of the classifier by introducing two sets of non-parallel hyperplanes, which is especially superior in dealing with complex and high-dimensional data. NPSVM not only inherits the strong generalisation ability of SVM, but also overcomes the limitations of TSVM in dealing with certain nonlinear data.

NPSVM aims to maximise the interval between different classes by constructing two non-parallel hyperplanes. Unlike SVM, which tries to find a globally optimal hyperplane to maximise the interval between two classes of samples, NPSVM allows the two hyperplanes to be close to their respective classes, and this flexibility allows NPSVM to better adapt to complex distribution structures.

In NPSVM, the two non-parallel hyperplanes can be represented separately as

$$w^+ \cdot x + b^+ = 0 \tag{23}$$

$$w^- \cdot x + b^- = 0 \tag{24}$$

where w^+ and w^- are two different normal vectors defining the directions of the two hyperplanes, and b^+ and b^- are the corresponding bias terms.

The goal of NPSVM is to simultaneously minimise the distances of the two classes of samples with respect to their respective hyperplanes, and for these two hyperplanes to be as non-interfering with each other as possible. Specifically, for the first hyperplane, the NPSVM attempts to move all positive class samples closer to this hyperplane and push negative class samples away from this hyperplane. Correspondingly, for the second hyperplane, the model tries to bring all negative class samples close to this hyperplane and push positive class samples away from this hyperplane.

The optimisation problem of NPSVM can be defined by two objective functions. For the first hyperplane, minimise the distance of the positive class samples and push away the negative class samples.

$$\min_{w^+,b^+} \frac{1}{2}(w^+)^2 + C \sum_{i \in \text{class } +} \xi_i \tag{25}$$

The constraints are shown as follow:

$$w^+ \cdot x_i + b^+ \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i \in \text{class } + \tag{26}$$

$$w^+ \cdot x_j + b^+ \leq -1 + \eta_j, \quad \eta_j \geq 0, \quad j \in \text{class } - \tag{27}$$

Similarly, for the second hyperplane, minimise the distance of the negative class samples and push away the positive class samples.

$$\min_{w^-,b^-} \frac{1}{2}(w^-)^2 + C \sum_{j \in \text{class } -} \eta_j \tag{28}$$

The constraints are shown as follow:

$$w^- \cdot x_j + b^- \geq 1 - \eta_j, \quad \eta_j \geq 0, \quad j \in \text{class } - \tag{29}$$

$$w^- \cdot x_i + b^- \leq -1 + \xi_i, \quad \xi_i \geq 0, \quad i \in \text{class } + \tag{30}$$

By solving these two optimisation problems, the NPSVM is able to determine two non-parallel classification hyperplanes simultaneously, leading to efficient and flexible classification.

SVM constructs only one globally optimal hyperplane, which is used to maximise the interval between the two classes of samples. NPSVM on the other hand constructs two non-parallel hyperplanes close to their respective classes. TSVM also classifies by constructing two non-parallel hyperplanes but it focuses on the data in each hyperplane close to its own class and at the same time far away from the data in the other class. NPSVM

goes a step further by maximising the interval between their respective classes and at the same time also takes into account the overall structure of the data distribution.

4. Construction of the INPSVM prediction model.

4.1. **PLS based feature extraction.** The high dimensionality and diversity of data are common challenges in regional data element market forecasting. A large number of features not only increases the computational complexity of the model, but also may lead to overfitting problems of the model. In order to extract important features efficiently and to reduce the dimensionality of the data, PLS is a commonly used technique [28]. PLS not only takes into account the correlation between the independent variables, but also maximises the relationship between the independent variables and the dependent variable, making it a powerful tool for feature extraction.

PLS is a generalised regression technique for scenarios where there is a high correlation between the matrix of independent variables (input data) and the matrix of dependent variables (output data). PLS projects the high-dimensional space of independent variables into a low-dimensional space by finding a set of new linear combinations while maintaining the maximum correlation of these new combinations with the dependent variable.

Given a matrix of independent variables X and a matrix of dependent variables Y, the goal of PLS is to find a new eigenspace by projection of the form:

$$T = XW \tag{31}$$

where T is the projected feature matrix and W is the projection matrix, determined by maximising the covariance between the projected features and the dependent variable.

Firstly, the independent variable matrix X and the dependent variable matrix Y are standardised to ensure that each variable has zero mean and variance. Calculate the covariance matrix C between the standardised X and Y. Perform eigenvalue decomposition of the covariance matrix C to find the principal direction vector (eigenvector). Project the original autocovariance matrix X into the low-dimensional space using the eigenvectors W. On the new low-dimensional eigenspace T, build a regression model to predict the dependent variable matrix Y.

In INPSVM models, PLS is used as a key step in feature extraction, helping to simplify the complexity of the input data and improve the efficiency and generalisation of the model. The features that are most relevant to the target variables are extracted through PLS. These features can effectively represent the main information of the original data while reducing the dimensionality of the data. Using the low-dimensional features extracted by PLS, INPSVM can perform classification modelling in a simplified feature space. The feature extraction process of PLS removes noise and redundant information from the data, allowing INPSVM to have greater generalisation ability and robustness when dealing with high-dimensional and complex data. The INPSVM model can be represented as follow:

$$\min_{w^+,b^+} \frac{1}{2}(w^+)^2 + C \sum_{i\in\text{class }+} \xi_i + \lambda\|T - XW\|^2 \tag{32}$$

$$\min_{w^-,b^-} \frac{1}{2}(w^-)^2 + C \sum_{j\in\text{class }-} \eta_j + \lambda\|T - XW\|^2 \tag{33}$$

where T is the low-dimensional feature matrix extracted by PLS, and λ is the parameter controlling the weights of PLS feature extraction.

4.2. **Anti-noise mechanisms.** In regional data element market forecasting, data are usually disturbed by noise and outliers. These noise and outliers may originate from errors in the data collection process, the impact of unexpected events, or the volatility of the market. In order to improve the stability and robustness of the prediction model, we introduce an anti-noise mechanism in INPSVM, which mainly reduces the influence of noise by introducing the absolute loss of L1 paradigm and the improved Hinge Loss function.

The L1-paradigm absolute loss function performs well when dealing with noise and outliers because the L1-paradigm penalises large deviations (noise) less and is better able to tolerate outliers in the data. Traditional SVMs usually use the L2 paradigm to measure the error, which penalises larger errors more severely and thus may cause the model to be too sensitive to noise. In NPSVM, we introduce the absolute loss of the L1 paradigm into the optimisation problem as an alternative to the traditional loss function. This can be expressed as:

$$L_1(x, y, f) = \sum_{i=1}^{N} |y_i - f(x_i)| \tag{34}$$

where x_i is the feature vector of the ith sample; y_i is the true label of the i-th sample; and $f(x_i)$ is the model's predicted value for the i-th sample.

In the framework of NPSVM, this loss function is used to minimise the loss in two hyperplanes, namely:

$$\min_{w^+,b^+} \frac{1}{2}(w^+)^2 + C \sum_{i\in\text{class }+} [1 - (w^+ \cdot x_i + b^+)] \tag{35}$$

$$\min_{w^-,b^-} \frac{1}{2}(w^-)^2 + C \sum_{j\in\text{class }-} [1 - (w^- \cdot x_j + b^-)] \tag{36}$$

By using the absolute loss of the L1 paradigm [29, 30], the NPSVM model is able to maintain greater stability in the face of noise and outliers, without over-adjusting the weights to accommodate these outliers.

Hinge Loss is a loss function commonly used in SVMs to measure the distance between sample points and the classification hyperplane. The traditional Hinge Loss is defined as follows:

$$L_{\text{hinge}}(x, y, f) = \sum_{i=1}^{N} \max(0, 1 - y_i f(x_i)) \tag{37}$$

The Hinge Loss imposes a larger penalty on sample points near the hyperplane or on the wrong side, which is effective for handling clean data, but may cause the model to be overly sensitive to noise and outliers in the presence of these outliers.

In order to enhance the noise immunity of the model, we design an improved Hinge Loss function, which has less effect on noise and outliers. The improved Hinge Loss can be expressed as follow:

$$L'_{\text{hinge}}(x, y, f) = \sum_{i=1}^{N} \max(0, \gamma(1 - y_i f(x_i))) \tag{38}$$

where γ is a tuning parameter that controls the sensitivity of the loss function to sample points at greater distances. When γ is small, the loss function penalises noise and outliers further away from the hyperplane less, and vice versa.

In the framework of INPSVM, the improved hinge loss is introduced into the optimisation objective:

$$\min_{\mathbf{w}_+,b_+} \frac{1}{2}\|\mathbf{w}_+\|^2 + C_1 \sum_{i\in \text{ class }+} \max\left(0, \gamma\left(1-(\mathbf{w}_+\cdot\mathbf{x}_i+b_+)\right)\right) \tag{39}$$

$$\min_{\mathbf{w}_-,b_-} \frac{1}{2}\|\mathbf{w}_-\|^2 + C_2 \sum_{j\in \text{ class }-} \max\left(0, \gamma\left(1-(\mathbf{w}_-\cdot\mathbf{x}_j+b_-)\right)\right) \tag{40}$$

Combining the absolute loss of the L1 paradigm and the improved Hinge Loss, we can construct an anti-noise optimisation problem for NPSVM. This optimisation problem not only considers the maximisation of the classification interval, but also pays special attention to the effects of noise and outliers to enhance the robustness and generalisation of the model. After comprehensive consideration, the anti-noise optimisation problem of INPSVM can be expressed as follow:

$$\min_{\mathbf{w}_+,b_+} \frac{1}{2}\|\mathbf{w}_+\|^2 + C_1 \sum_{i\in \text{ class }+} \left(|1-(\mathbf{w}_+\cdot\mathbf{x}_i+b_+)| + \max\left(0, \gamma\left(1-(\mathbf{w}_+\cdot\mathbf{x}_i+b_+)\right)\right)\right) \tag{41}$$

$$\min_{\mathbf{w}_-,b_-} \frac{1}{2}\|\mathbf{w}_-\|^2 + C_2 \sum_{j\in \text{ class }-} \left(|1-(\mathbf{w}_-\cdot\mathbf{x}_j+b_-)| + \max\left(0, \gamma\left(1-(\mathbf{w}_-\cdot\mathbf{x}_j+b_-)\right)\right)\right) \tag{42}$$

4.3. **Interval Distribution Information.** In regional data element market forecasting, the distribution structure information of the data has an important impact on the performance of the classification model. Traditional NPSVM models mainly focus on the distance (i.e., interval) between the sample points and the classification hyperplane, while ignoring the overall distribution structure of the data. This neglect may lead to the model's lack of generalisation ability on complex datasets. To enhance the performance of NPSVM, we introduce the concept of interval distribution information to improve the robustness and generalisation ability of the model by optimising the mean and variance of the intervals.

In NPSVM, the interval is the distance from the index data point to the support hyperplane of the class it belongs to. For positive category samples, the interval can be expressed as follow:

$$\gamma_+ = w_+ \cdot x_+ + b_+ \tag{43}$$

For negative class samples, the interval can be expressed as

$$\gamma_- = w_- \cdot x_- + b_- \tag{44}$$

where x_+ and x_- are the eigenvectors of the sample points of the positive and negative classes, respectively.

In order to better describe the distributional information of the data, we introduce the mean and variance of the intervals:

$$\bar{\gamma}_+ = \frac{1}{N_+}\sum_{i\in \text{ class }+} \gamma_{i+}, \quad \bar{\gamma}_- = \frac{1}{N_-}\sum_{j\in \text{ class }-} \gamma_{j-} \tag{45}$$

$$\sigma^2_{\gamma+} = \frac{1}{N_+}\sum_{i\in \text{ class }+} (\gamma_{i+}-\bar{\gamma}_+)^2, \quad \sigma^2_{\gamma-} = \frac{1}{N_-}\sum_{j\in \text{ class }-} (\gamma_{j-}-\bar{\gamma}_-)^2 \tag{46}$$

where N_+ and N^- are the number of positive and negative class samples respectively; γ_{i+} and γ_{j-} are the intervals of the respective sample points; $\overline{\gamma}_+$ and $\overline{\gamma}_-$ are the mean intervals of the positive and negative class samples; σ^2_+ and σ^2_- are the variances of the positive and negative class sample intervals respectively.

In the INPSVM model, we optimise the distributional information of the data by simultaneously maximising the mean of the interval and minimising the variance of the interval. This optimisation method helps to improve the adaptability and robustness of the model to differently distributed data.

5. **Empirical studies.**

5.1. **Experimental indicators and data.** The indicators that reflect the changes in the data factor market are very many and varied. If all of them are selected, it will make the difficulty of the experiment rise sharply, because the increase of its workload will lead to the actual operation is time-consuming and labour-intensive, for example, there are indicators of a certain year's data is difficult to obtain the situation. On the other hand, there is usually some correlation between variables that reflect fluctuations in economic development.

Combined with previous experience and after repeated consideration, this paper selects 14 operational indicators from a large number of data factor market indicators that reflect the macro changes in the region as much as possible. The specific indicators for each category are shown in Table 1.

Table 1. Data factor market predictors

Name	**Variant**	**Unit**
Gross primary sectoral product	x_1	Billions
Gross secondary sectoral product	x_2	Billions
Gross tertiary sectoral product	x_3	Billions
Total retail sales of consumer goods	x_4	Billions
Local fiscal expenditures	x_5	Ten thousand yuan
Local revenues	x_6	Ten thousand yuan
Year-end deposit balances of financial institutions	x_7	Billions
Foreign trade imports	x_8	Ten thousand dollars
Value of foreign trade exports	x_9	Ten thousand dollars
Total investment in fixed assets	x_{10}	Billions
Actual utilisation of foreign capital	x_{11}	Ten thousand dollars
Total energy consumption	x_{12}	Million tonnes of standard coal
Number of students enrolled in general higher education	x_{13}	Man
Gross domestic product (GDP)	y	Billions

The experimental data comes from the official website of the National Bureau of Statistics, and considering the availability of economic operation data of a province, the experimental data is finally determined to be the data factor market operation data from 2015 to 2022.

In the pre-processing stage of the data, the raw data are standardised, which is conducive to the smooth progress of the experiment, in order to avoid confusion in the use of the data, the standardised data of the independent variable and the dependent variable are denoted as $x\%_1, x\%_2, \ldots, x\%_{13}, y\%$ respectively. The process of standardisation is as follows.

$$x\% = \frac{x - \overline{x}}{S_x} \tag{47}$$

where $\overline{x}$ is the mean of variable x; S_x is the variance of variable x. For the accuracy of the experimental results, all the data used in the following experiments are standardised data.

In this paper, Python 3.6 is used as the tool, and the main third-party tool libraries used to make regression predictions. The improved NPSVM method has kernel function

parameter $\gamma = 0.01$, penalty parameter $C = 10$, anti-noise parameter $\gamma = 1$, and PLS score $k = 10$. These parameter settings were selected through cross-validation and performance evaluation of the model, and are intended to optimise the model's performance for regional data element market forecasting.

5.2. **Comparison of Prediction Models.** To judge whether a new prediction model is superior or not, it is necessary to compare the prediction effect with similar prediction models, so this paper introduces three prediction models similar to the structure of INPSVM for comparison, the first one is GA-SVM. The second one is TSVM. The second one is the PCA method which is similar to the PLS ideology of this paper, i.e., PCA is used to extract the components of the original independent variable data before using NPSVM regression modelling and prediction, which is also known as PCA-NPSVM model. The second one is the PLS method, which is similar to the idea of PLS in this paper.

The prediction results of the above three comparison models and the INPSVM model are put together for comparative analysis and the results are shown in Figure 2. It can be seen that the INPSVM prediction model has the smallest error rate, and its average error rate is 0.023. The PCA-NPSVM prediction model has a prediction average error rate of 0.031, the TSVM prediction model has a prediction average error rate of 0.041, and the GA-SVM prediction model has a prediction average error rate of 0.057. The error probabilities of these three comparison models are basically all above 3% or more. It can be seen that the prediction effect of INPSVM is better compared to the prediction performance of the other three models.

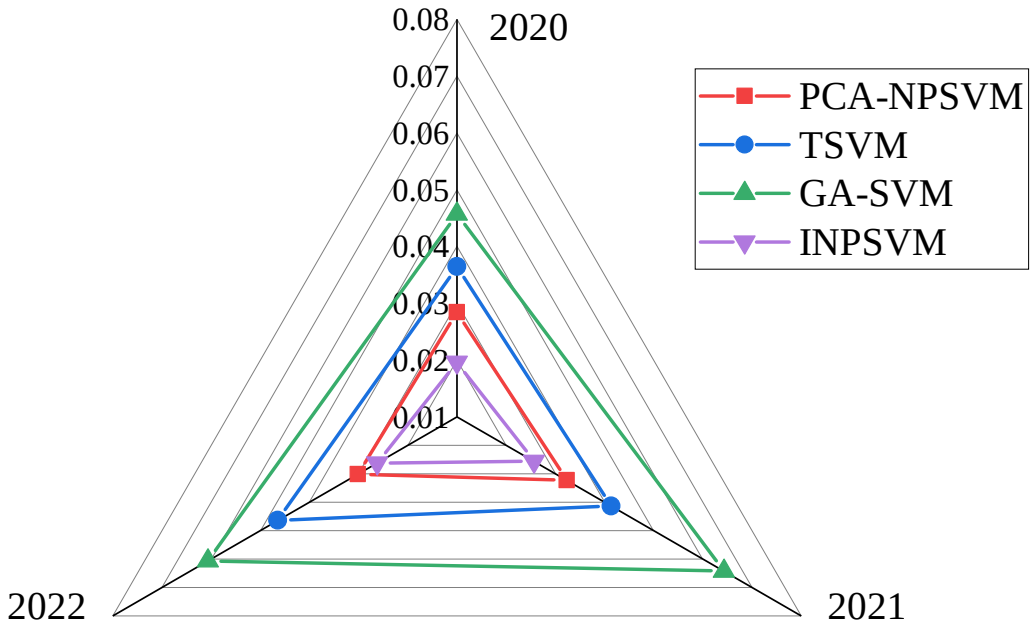

Figure 2. Comparison of error rates of prediction models

6. **Conclusion.** An INPSVM model for regional data factor market forecasting is proposed. First, in terms of feature extraction, INPSVM utilises the PLS method to effectively reduce the dimensionality of the data while retaining the information that best reflects the target variables. This not only reduces the computational complexity of the model, but also enhances the ability to handle high-dimensional data. Second, in order to address the impact of noise on the model, this paper introduces an anti-noise mechanism

based on the absolute loss of the L1 paradigm and an improved Hinge Loss function. In addition, this paper also improves the model's ability to exploit the overall distributional characteristics of the data by optimising the interval distribution information. The optimisation of interval distribution information plays a key role in improving the generalisation ability of the model, which enables the INPSVM model to better adapt to diverse data. The experimental results show that the PLS feature extraction method shows better performance than the traditional PCA feature selection method in the experiments, in addition to the anti-noise mechanism can effectively reduce the sensitivity of the model to outliers. Future research will incorporate more advanced feature extraction to improve the efficiency of model application on large-scale datasets.

Acknowledgment. This work is supported by the Zhejiang Province Philosophy and Social Science Planning Project (No. 24NDJC048YB) and the Zhejiang Provincial Department of Education Research project (No. Y202353033).

REFERENCES

[1] W. A. Kamakura, M. Wedel, F. De Rosa, and J. A. Mazzon, "Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction," *International Journal of Research in Marketing*, vol. 20, no. 1, pp. 45-65, 2003.

[2] S. G. Fifield, D. M. Power, and C. D. Sinclair, "Macroeconomic factors and share returns: an analysis using emerging market data," *International Journal of Finance & Economics*, vol. 7, no. 1, pp. 51-62, 2002.

[3] T. Elrod, and M. P. Keane, "A factor-analytic probit model for representing the market structure in panel data," *Journal of Marketing Research*, vol. 32, no. 1, pp. 1-16, 1995.

[4] J. J. Choi, and M. Rajan, "A joint test of market segmentation and exchange risk factor in international capital market," *Journal of International Business Studies*, vol. 28, pp. 29-49, 1997.

[5] L. Brandt, T. Tombe, and X. Zhu, "Factor market distortions across time, space and sectors in China," *Review of Economic Dynamics*, vol. 16, no. 1, pp. 39-58, 2013.

[6] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, 40, 2019.

[7] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, p. 2339, 2023.

[8] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, p. 1977, 2023.

[9] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, no. 10, pp. 2513-2522, 2005.

[10] L. Liu, M. Chu, R. Gong, and L. Zhang, "An improved nonparallel support vector machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 5129-5143, 2020.

[11] Y. Tian, Z. Qi, X. Ju, Y. Shi, and X. Liu, "Nonparallel support vector machines for pattern classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1067-1079, 2013.

[12] G. Liu, L. Wang, D. Liu, L. Fei, and J. Yang, "Hyperspectral image classification based on nonparallel support vector machine," *Remote Sensing*, vol. 14, no. 10, 2447, 2022.

[13] C. S. Wong, and W. K. Li, "On a mixture autoregressive model," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 62, no. 1, pp. 95-115, 2000.

[14] Y. Xin, J. Gao, X. Yang, and J. Yang, "Maximum likelihood estimation for uncertain autoregressive moving average model with application in financial market," *Journal of Computational and Applied Mathematics*, vol. 417, 114604, 2023.

[15] T. Dimri, S. Ahmad, and M. Sharif, "Time series analysis of climate variables using seasonal ARIMA approach," *Journal of Earth System Science*, vol. 129, pp. 1-16, 2020.

[16] H. Abbasimehr, and R. Paki, "Improving time series forecasting using LSTM and attention models," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 1, pp. 673-691, 2022.

[17] M. Sarstedt, C. M. Ringle, J.-H. Cheah, H. Ting, O. I. Moisescu, and L. Radomir, "Structural model robustness checks in PLS-SEM," *Tourism Economics*, vol. 26, no. 4, pp. 531-554, 2020.

[18] M. He, Y. Zhang, D. Wen, and Y. Wang, "Forecasting crude oil prices: A scaled PCA approach," *Energy Economics*, vol. 97, pp. 105189, 2021.
[19] M. Beniwal, A. Singh, and N. Kumar, "Forecasting long-term stock prices of global indices: A forward-validating Genetic Algorithm optimization approach for Support Vector Regression," *Applied Soft Computing*, vol. 145, 110566, 2023.
[20] M. A. M. Al-Afeef, "Factors affecting market capitalization: A practical study ase 1978-2019," *International Journal of Scientific and Technology Research*, vol. 9, no. 3, pp. 7049-7053, 2020.
[21] Y. Suhara, M. Bahrami, B. Bozkaya, and A. S. Pentland, "Validating gravity-based market share models using large-scale transactional data," *Big Data*, vol. 9, no. 3, pp. 188-202, 2021.
[22] A. Rizwan, N. Iqbal, R. Ahmad, and D.-H. Kim, "WR-SVM model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification," *Applied Sciences*, vol. 11, no. 10, 4657, 2021.
[23] K. Qi, and H. Yang, "Elastic net nonparallel hyperplane support vector machine and its geometrical rationality," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7199-7209, 2021.
[24] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, and M. Ganaie, "Comprehensive review on twin support vector machines," *Annals of Operations Research*, pp. 1-46, 2022.
[25] Z. Liang, and L. Zhang, "Uncertainty-aware twin support vector machines," *Pattern Recognition*, vol. 129, 108706, 2022.
[26] Y. Li, H. Sun, and W. Yan, "Domain adaptive twin support vector machine learning using privileged information," *Neurocomputing*, vol. 469, pp. 13-27, 2022.
[27] W.-J. Chen, Y.-H. Shao, C.-N. Li, Y.-Q. Wang, M.-Z. Liu, and Z. Wang, "NPrSVM: Nonparallel sparse projection support vector machine with efficient algorithm," *Applied Soft Computing*, vol. 90, 106142, 2020.
[28] W. Wu, Y. Xu, and X. Pang, "A hybrid acceleration strategy for nonparallel support vector machine," *Information Sciences*, vol. 546, pp. 543-558, 2021.
[29] L. Zhang, L. Sun, W. Li, J. Zhang, W. Cai, C. Cheng, and X. Ning, "A joint bayesian framework based on partial least squares discriminant analysis for finger vein recognition," *IEEE Sensors Journal*, vol. 22, no. 1, pp. 785-794, 2021.
[30] C. Ruiz, C. M. Alaíz, and J. R. Dorronsoro, "Convex formulation for multi-task L1-, L2-, and LS-SVMs," *Neurocomputing*, vol. 456, pp. 599-608, 2021.