Journal of Global Information Management
Volume 30 • Issue 7

Stock Price Prediction Based on Data Mining Combination Model

To-Han Chang, Minjiang University, China
https://orcid.org/0000-0001-6737-1226

Nientsu Wang, Minjiang University, China*

Wen-Bin Chuang, National Chi-Nan University, Taiwan

ABSTRACT

Predicting stock indices is a common concern in the financial world. This work uses neural network, support vector machine (SVM), mixed data sampling (MIDAS), and other methods in data mining technology to predict the daily closing price of the next 20 days and the monthly average closing price of the future expected daily closing price on the basis of the market performance of stock prices. Additionally, by the mutual ratio of weighted mean square error, the study achieves the best prediction result. Combining value investment effectively with nonlinear models, a complete stock forecasting model is established, and empirical research is conducted on it. Results indicate that SVM and MIDAS have good results for stock price forecasting. Among them, MIDAS has a better mid-term forecast, which is approximately 10% higher than the forecast accuracy of the SVM model. Meanwhile, SVM is more accurate in the short-term forecast.

KEYWORDS

Data Mining, Neural Network System, Stock Price Prediction, Support Vector Machine

INTRODUCTION

Since the emergence of the stock market, many experts and scholars have been actively engaged in stock market analysis and research. Some effective stock analysis methods have come into being. Many researchers use traditional statistical methods to analyze and predict the stock market. However, some traditional statistical measurement models have strict requirements on data (e.g., the data must have characteristics, such as stagnation, regularity, and low noise). Some limited conditions often exist in the stock market during the analysis process. However, in real life, stock market data have the characteristics of large potential, large amount, and complex data relationships. Therefore, achieving the expected results in the analysis and prediction of the stock market by applying some traditional statistical methods is difficult.

At first glance, there may appear to be no regularity in the stock market. It may seem, on the surface, to lack regularity. When it fluctuates seemingly unpredictably, but that doesn't actually happen. Although accurately predicting future stock prices is impossible, price fluctuations can still be predicted in the short term. Many scholars have raised various stock quality analysis techniques in recent years, with reference to fundamental aspects and technical elements. Specifically, the

DOI: 10.4018/JGIM.296707 *Corresponding Author

introduction of data mining methods provides good technical support for exploring the rules of the stock market and stock price changes. At the same time, it provides an important reference basis for investors' decision making.

The innovation of this paper is based on data mining and market-related techniques. In addition, it applies a wide range of statistical data mining methods to stock market research. It explores the application of data mining in stock market research to recommend stock market analysis and model forecasting; finally, a high-density algorithm is introduced. The model adopts the peak density algorithm for global distribution, which solves the uncertainty problem of the time series model when processing data. The research results show that data mining technology is an important science and application prospect in product analysis and prediction.

RELATED WORK

Jochems et al., (2016) used predictive models to medical care and believed that one of the main obstacles to achieving personalized medicine is to obtain enough patient data to input predictive models. Barbarelli et al., (2016) proposed a 1D numerical code for estimating the performance of a centrifugal pump used as a turbine (PAT). Once these parameters are derived, the loss is calculated, and the characteristic curve of PAT is determined. Bergey et al., (2017) converted and cleaned up diagnostic medical reports for text mining methods. Predictive models are used to characterize clusters obtained through unsupervised classification and potential category modeling of disease events. Chatterjee (2016) proposed a new technique that uses stock market liquidity to predict recessions. The results can be used by professional forecasters and have potential monetary value policy influence.

RESEARCH METHODS OF STOCK PRICE FORECASTING

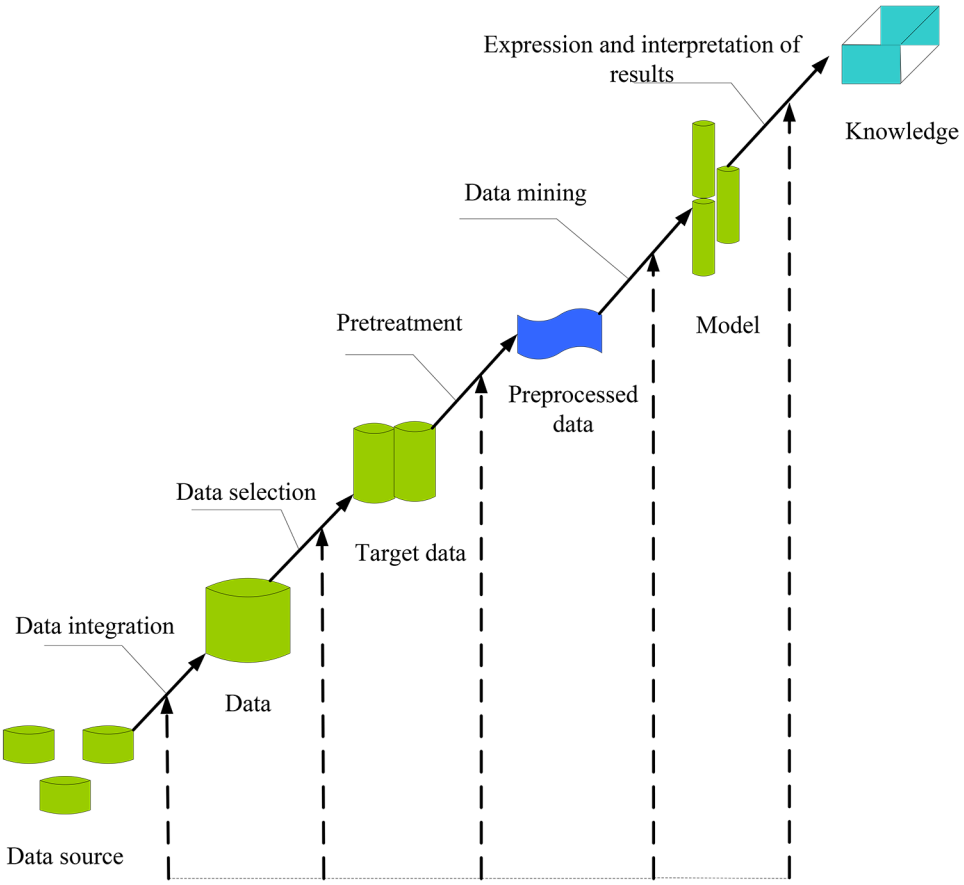
Data Mining

Data mining is based on the needs of users, discovering hidden and important information from complex databases (Forsberg, et al., 2017). Data mining includes machine learning, database technology, statistics, artificial intelligence (AI), pattern recognition, and other technical fields (Djouvas et al. 2017). With the rapid development of Internet technology, people have an increasing number of channels for obtaining information, and their lives have become increasingly convenient (Faria et al., 2017). However, a large amount of information also brings many problems: too much information is difficult to assimilate; the amount of information is difficult to distinguish between true and false; information security is difficult to guarantee (Falkenthal et al., 2017). In this case, data mining technology is introduced. Data mining is a complex process that can help people extract hidden information that was unknown in the past and has great value from a large amount of data (Fan et al., 2018). It's a multidisciplinary field that includes artificial intelligence, database technology, knowledge base systems, neural networks, mathematics, dispenser learning, statistics, knowledge acquisition, information extraction, paralleled computing, with many other aspects.

The theoretical basis of the system is the cornerstone of data mining methods (Kinnebrew et al., 2016). It includes data simplification and simplification of data representation to meet the research needs of people. Data mining uses decision trees, association rules, and other algorithms to compress datasets. Rule discovery, that is, the use of relevant knowledge of probability theory to transform the goal of data mining into a unified search rule, discover the probability distribution of random variables from a large number of datasets, and then conduct research (Pisani et al., 2016). From the perspective of microeconomics, data mining technology is seen as a continuous process of optimization problems (Raff et al., 2017). Optical data mining focuses on data visualization, performs interactive data mining processes, and well presents data mining results, which have become an important part of

data mining (Sharma, et al., 2016). These theoretical frameworks are not isolated; they are intersecting one another. The data mining process is illustrated in Figure 1.

Figure 1. Data mining process



With the deepening of data mining, the mining of one-layer and 1D association rules can no longer meet people's needs (Zhang et al. 2017). Considering that no data are collected, finding strong association rules at the level of data detail is difficult. In general, the support of low-level data is low, whereas that of high-level data is generally high. For example, when analyzing the relationship among customers who purchase a certain type of electronic products, each type has different brands, models, and colors. An obvious hierarchical structure exists among data elements (Zhang et al. 2019). Thus, when extracting the connections among specific products, using single-level association rules to obtain more useful information is impossible. Moreover, useful information must be extracted among multiple levels (Tharwat et al., 2017).

Mining Technology

Data mining is the basic process of discovering a large number of datasets, using machine learning, statistics, and intelligent methods to generate data models at the intersection of database systems (Sun et al., 2017). The initial analysis steps are complemented by database, data administration, data preprocessing, model schemes, rate measurement, complexity, structure discovery, further treatment,

imaging, and online updates (Cavallaro et al., 2015). In this article, the main data mining tools used for stock price modeling and variable selection are: neural network, MIDAS model, and support vector machine (SVM) (Wang et al., 2017, pp. 197-208).

Neural Network Technology

Manufactured artificial neural networks (ANN) have been developed as a computational model founded on the structures and capabilities of biological neural networks (Wang et al., 2017, pp. 650-660). The model influences the structure of ANN, as the neural network in a sense modifies or learns automatically according to influences and exports. It is usually used as a nonlinear statistical modeling tool to identify and shape the complex relationship between input and output during the self-learning process. The basic processing unit of a neural network is a neuron. This cornerstone of human knowledge contains some universal possibilities. Biological neurons receive input from other sources, combine them in a certain way, perform a general nonlinear function in the result, and then extract the final result (Wang et al., 2018). The weights of artificial neurons and connections are usually adjusted as learning progresses. Neurons can have a limit, and only send signals when the total signal exceeds that limit. The algorithm flow of neural network technology is illustrated in Figure 2.

A neural network can gradually improve its performance by "learning" examples, and it is an adaptive system (Gui et al., 2017). The neural network simulates the response of the human nervous system to each stimulus and trains and models communication on the basis of given input and output signals, just as it forms memory rules. Essentially, ANN is similar to the human nervous system, composed of a large number of neurons, which we call a simple information processing unit (Liu et al., 2017). Network connections can systematically adapt to input and output, making it an ideal choice for supervised learning. Figure 3 illustrates the case if it is abstracted as a mathematical model (Maulik et al., 2017).

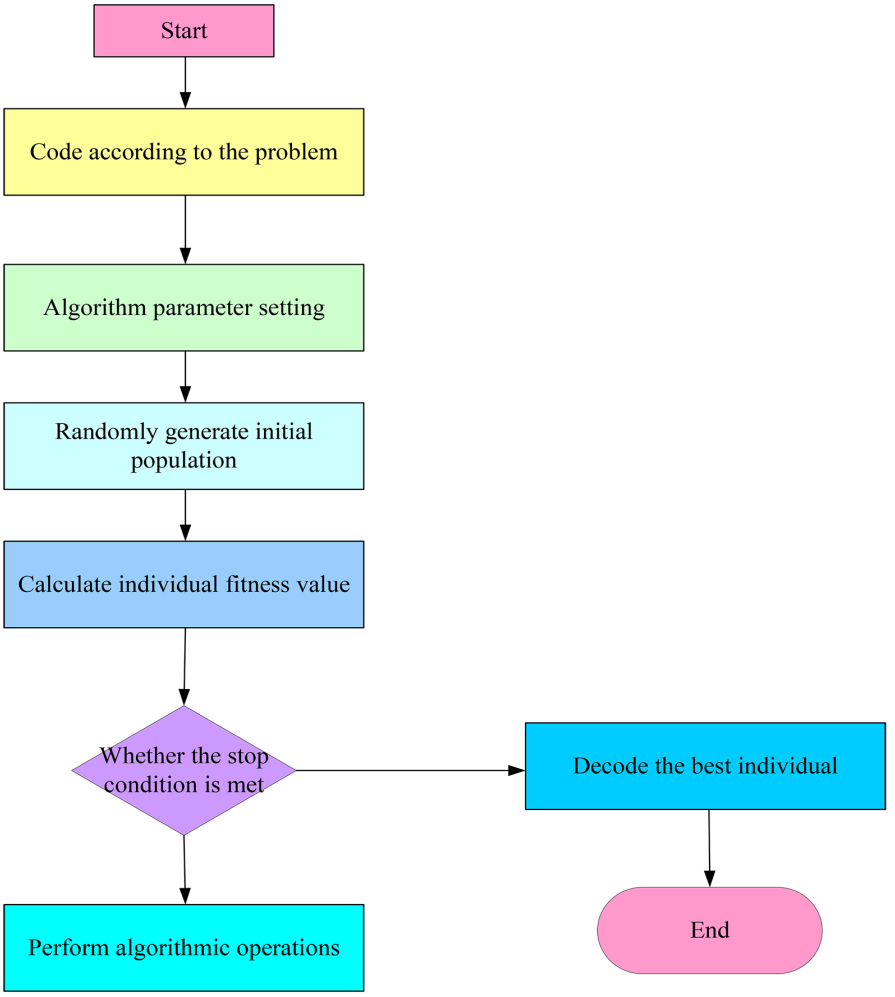
A single neuron can be composed of input signals (including variables $a_1, a_2, \ldots, a_n$) and output signal t, including intermediate transfer functions (Guo et al., 2014). According to the importance of each input variable, weighting it (indicated as w_i) can be regarded as a stimulus of different sizes of neurons from the outside world. Among them, t (output neuron), that is, the calculation method of the target value is (Wang et al., 2017, pp. 650-660):

$$t\left(x\right) = f\left(\sum_{i=1}^{n} w_i a_i + b\right) \tag{1}$$

The backpropagation neural network (BPNN) model is a three-power or multi-power neural network model. It uses a slope algorithm to propagate the error between the calculated output and the actual output and correct any connection weight (Lee et al., 2017). While both training patterns that are introduced into the network, neuron activation values will be propagated out from the input layer through each intermediate layer to the output layer. There will be each neuron in the output layer that will receive the network input the responses and try to shorten the distance between them. Conjugate weights are corrected by the hidden layers in the intermediate layers from the output plane to each layer (Widyantara, et, al., 2017). As this error distribution correction continues, the correct response rate of the network input function will continue. The principle of the learning process is shown in Figure 4.

The relationship between stock price and its influencing factors is not simply linear. The advantage of neural network lies in its strong nonlinear fitting ability. ANNs are used as random function tools. These types of tools help evaluate the most cost-effective solution when defining computational functions or distributions (Kelley et al., 2016).

Figure 2. Neural network algorithm flow



MIDAS Model

At present, the most commonly used model for data processing is the mixed data sampling model(MIDAS). It is derived from the evolution of the distributed lag model and is essentially a regression model that can directly process the mixed data. The MIDAS model initially uses high-frequency financial data to predict low-frequency financial data. The construction idea of the MIDAS model comes from the distributed lag model whose representation is as follows:

$$Y_t = \beta_0 + B\left(L\right)X_t + \lambda_t \tag{2}$$

Among them, B(L) is a polynomial operator with finite or infinite lag, which is obtained by parameter estimation. The model assumes that the data frequency is consistent.

The most basic unary MIDAS regression model form is as follows(Jeon et al., 2018):

Figure 3. Single neuron model

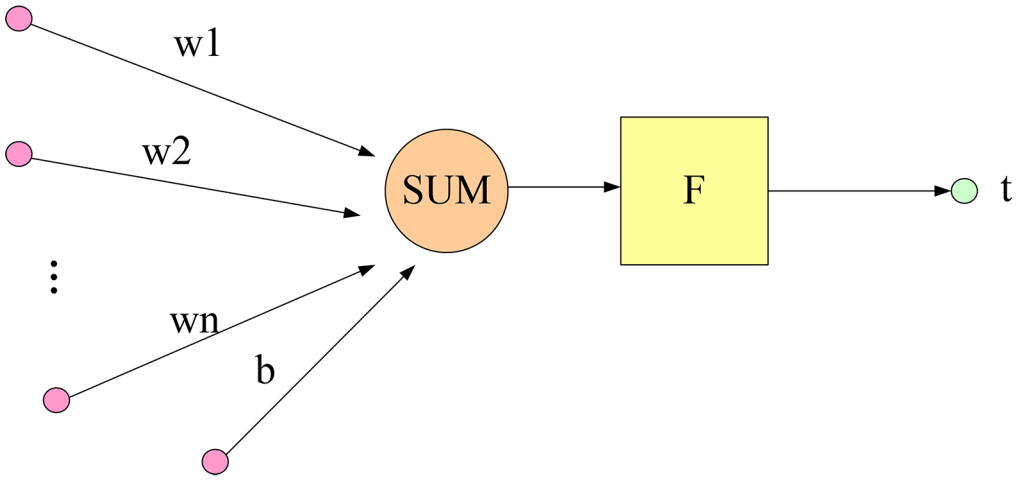
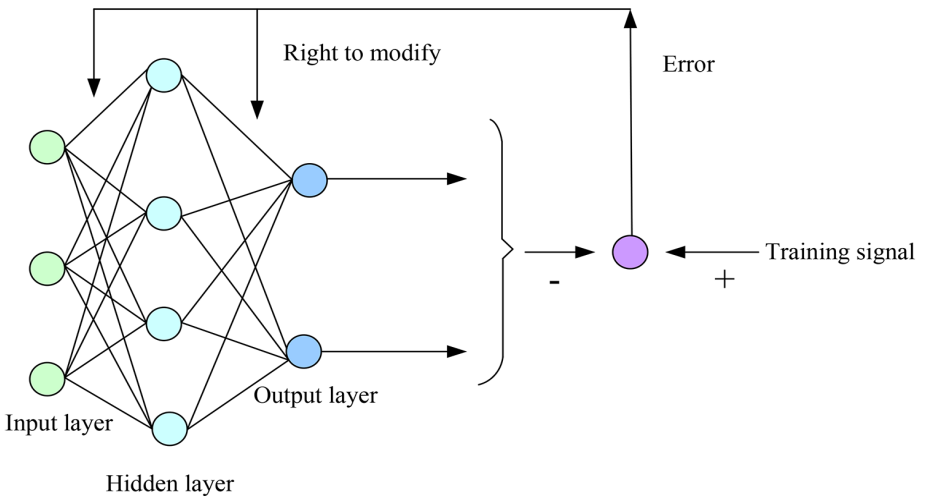


Figure 4. BPNN learning principle



$$y_t = \beta_0 + \beta_1 B\left(L^{K/n};\lambda\right)x_{t-h}^{(n)} + \lambda_t \tag{3}$$

Among them, the left side of equation y_t is a certain low-frequency data, which x^m refers to high-frequency data; frequency y_t is m times the frequency.

We use h to indicate from which period the data are used for model prediction. The three most commonly used forms in weight function are Almon polynomial function, exponential Almon polynomial function, and β polynomial function. The basic expression form of Almon polynomial function is:

$$b\left(k;\lambda\right)=\frac{\lambda_0+\lambda_1 k+\lambda_2 k^2+...+\lambda_n k^n}{\sum_k^K\left(\lambda_0+\lambda_1 k+\lambda_2 k^2+...+\lambda_n k^n\right)} \tag{4}$$

The basic expression form of the exponential Almon polynomial function is:

$$b\left(k;\lambda\right)=\frac{\exp(\lambda_0+\lambda_1 k+\lambda_2 k^2+...+\lambda_n k^n)}{\sum_k^K\exp\left(\lambda_0+\lambda_1 k+\lambda_2 k^2+...+\lambda_n k^n\right)} \tag{5}$$

The basic expression of β polynomial function is:

$$b\left(k,\lambda_1,\lambda_2\right)=\frac{f\left(k\,/\,K,\lambda_1;\lambda_2\right)}{\sum_k^K f\left(k\,/\,K,\lambda_1;\lambda_2\right)} \tag{6}$$

The above three forms of the weight function all implicitly assume that the sum of the weights is 1 and can ensure that the value of the weight function is positive. For simplicity, a weight function containing only two parameters is usually used in the MIDAS model. Nevertheless, the flexibility of calculation can still be guaranteed. The specific form is as follows:

$$b\left(k;\lambda\right)=\frac{\exp(\lambda_1 k+\lambda_2 k^2)}{\sum_k^K\exp\left(\lambda_1 k+\lambda_2 k^2\right)} \tag{7}$$

SVM

Supports vector machines (SVM) avoid the dimensionality in stock prediction disaster. In the support vector machine, the classification function is determined by a small number of support vectors. As a consequence, the "dimensionality curse" can be avoided in a sense.

First, define an optimal hyperplane to maximize the boundary between the two categories; second, extend the above definition for the nonlinear separable problem and specify corresponding penalties for incorrect classification. Finally, the data are mapped to a high-dimensional space; classification is easy when a linear decision surface is used: reconstruct the problem. We find *w* and *b* by solving the following objective function using quadratic programming:

$$\min\frac{1}{2}\left\|w\right\|;s.t.y(w*x_i+b)\geq 1 \tag{8}$$

The algorithm is as follows (Kim et al., 2019):

$$L\left(w,b,\vartheta,\gamma,r\right)=\frac{1}{2}\left\|w\right\|^2+C\sum\nolimits_{i=1}\vartheta_i-\sum\nolimits_{i=1}\lambda_i\left(y_i\left(w^t x_i+b\right)-1\right) \tag{9}$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum\nolimits_{i=1} \lambda_i y_i x_i \tag{10}$$

For partial w, b, ϑ derivatives, we respectively obtain:

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum\nolimits_{i=1} \lambda_i y_i = 0 \tag{11}$$

$$\frac{\partial L}{\partial \vartheta} = 0 \Rightarrow C - \lambda_i - r_i = 0 \tag{12}$$

We finally transform into optimization problem solving

$$\max \sum\nolimits_{i=1} \lambda_i - \frac{1}{2} \sum\nolimits_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \left\langle x_i x_j \right\rangle \tag{13}$$

We introduce an insensitive loss function, and the expression is as follows:

$$L_\phi \left(y, f\left(x\right)\right) = (\left|y - f\left(x\right)\right| - \phi, \left|y - f\left(x\right)\right| - \phi > 0 \tag{14}$$

The soft boundary method is adopted, and the relaxation factor is introduced. Here, the upper and lower relaxation factors are used. Therefore, the optimization problem is obtained:

$$\min \frac{1}{2} \left\| w \right\|^2 + C \sum\nolimits_{i=1}^{l} \left(\lambda_i + \gamma_i\right) \tag{15}$$

$$s.t. \begin{cases} yi - \left(w^T x_i + b\right) < \tau + \lambda_i \\ \left(w^T x_i + b\right) - y_i < \tau + \gamma_i \\ \lambda_i, \gamma_i \geq 0 \end{cases}. \tag{16}$$

Finally, we obtain

$$f\left(x\right) = \sum\nolimits_{i=1}^{l} \left(a_i - a_i^*\right)\left(x_i, x\right) + b \tag{17}$$

$$b = y_i - \left(w, x\right) - \varepsilon, a_i \in \left(0, C\right) \tag{18}$$

$$b = y_i - \left(w, x_i\right) + \varepsilon, a_i^* \in \left(0, C\right) \tag{19}$$

Data are usually mapped to a new space. Then, the inner product of the new vector is subtracted to convert the nonlinearly separable data sample into a linearly separable space, avoiding several mathematical operations. In short, it performs complex data conversion and then discovers the process of separating data according to a specified label or output. Note that the space with the largest feature dimension increases the generalization error of the support machine. However, when multiple samples exist, the algorithm can perform well.

Currently, the most studied kernel functions include:

Polynomial kernel

$$K\left(x, y\right) = \left[\left(x \cdot y\right) + l\right]^q \tag{20}$$

A q-order polynomial classifier can be obtained.

Radial Basis Function Kernel

$$K\left(x, x'\right) = \exp(-\frac{\left|x - x'\right|^2}{2\varepsilon^2}) \tag{21}$$

Each basis function center corresponds to an SVM, and their output weights are automatically determined by the algorithm.

We use u_{ij} to indicate the degree of membership X_i of the cluster centers v_i and calculate each cluster center v_j:

$$v_j = \frac{\sum_i^n = 1 u_{ij}^m x_i}{\sum_i^n = 1 u_{ij}^m} \tag{22}$$

We also calculate the cost function. If the cost function is less than a certain threshold or the change of the cost function during two iterations is less than a certain threshold β, then the algorithm stops, and the cost function is

$$J\left(U, v_1, ..., v\right) = \sum_{i=1}^c J_i = \sum_{i=1}^n u_{ij}^m d^2\left(x_i, v_j\right) \tag{23}$$

We update membership matrix *U*, and then return to the step

$$u_{it} = \sum_{j=1}^c \left(d_{it} \,/\, d_{it}\right)^{-2/\left(m-1\right)} \tag{24}$$

According to the state of the hidden layer unit, the formula for obtaining the visible layer unit in reverse is

$$P\left(v_i = 1 \middle| h, \theta\right) = \sigma\left(a_i + \sum_j W_{ij} h_j\right) \tag{25}$$

After the fuzzification process, the sample data can be fuzzified to obtain a series of fuzzy values. If the influence of non-adjacent data on the fuzzy set is ignored when establishing the fuzzy relationship, and only the establishment of the fuzzy relationship corresponding to two data is considered, then a large amount of information will be lost, and the prediction result will be inaccurate. Therefore, this study establishes the fuzzy corresponding to multiple data relationships.

Application of Data Mining in Stock Forecasting

The prediction model function is mainly to forecast the middle step of the training set sequence according to the specified input, which is preset before modeling, and finally output the result. The specific steps for establishing the prediction model are as follows:

Step 1. Select the components of the input vector
Step 2. Select a continuous time series and use the specified m value to build a training sample set (x, y), i = 1, 2,..., n
Step 3. Choose the type of loss function to determine the type of SVM
Step 4. Select the kernel function and penalty parameters for the conversion of high-dimensional feature space and low-dimensional input space.
Step 5. Solve the planning problem according to the type of SVM and use the obtained Lagrangian sparseness to establish a prediction model. The model is built.

In this article, the relevant data of the Shanghai Composite Index are selected as the modeling data sample. The test period is 20 days, that is, the closing index of the first 20 days is used to predict the subsequent closing index. At the beginning of the model training, this article uses the default parameters of the SVM model. Then, for model training, the output error results of the default and self-selected parameters are compared, as shown in Table 1.

Table 1. Default and optional parameters

Output item	Optional parameter	Default parameter
Minimum error	−92.577	−282.432
Maximum error	90.452	175.973
Average error	−3.043	−27.217
Mean absolute error	18.087	−27.217
Standard deviation	29.661	102.07
Linear correlation	0.877	0.771
Number of occurrences	80	80

STOCK PRICE PREDICTION EXPERIMENT AND RESULTS

Stock Price Closing Price and Forecast

The MIDAS and SVM models were used to predict the closing price in SSE (Shanghai Stock Exchange) in order to further proved the effectiveness of the method in this paper with regard to stock price prediction. The prediction results are compared with the true values to obtain the following prediction results, as shown in Table 2.

Table 2. Comparison of actual and predicted results

	Closing price	MIDAS	SVM
2019/1/3	2848.5	2862.47	2917.07
2019/1/4	2996.54	2874.13	2845.37
2019/1/5	2899.86	2852.21	3002.31
2019/1/6	2803.73	2840.03	2897.45
2019/1/7	2865.58	2886.78	2902.49
2019/1/8	2779.58	2886.78	2913.41
2019/1/9	2805.47	2789.47	2776.6
2019/1/10	2827.49	2781.19	2813.23
2019/1/11	2827.41	2781.28	2834.46
...	...	...	...
2019/1/23	2868.32	2869.53	2856.62

The fuzzy time series model based on the density peak is used to predict the closing price of the Shanghai Stock Exchange Index. To make the prediction result further intuitive, the real value is compared with the predicted value. The result is illustrated in Figure 5.

The predicted value of the stock price based on the model is not much different from the true value. The fitting effect is good, and all the true values are almost distributed within the 95% confidence interval, indicating that the predicted result is close to the true values. The standard residual diagram of the prediction results is displayed in Figure 6.

Almost all prediction points fall in the horizontal band-shaped interval from −2 to 2 and present a completely random distribution state without any systematic trend, indicating that the model used is relatively more predictive of the closing price of the Shanghai Stock Exchange Index.

We use earnings per share, ROC, EPF per share, and NAV per share for adopting stock prices as research objects to authenticate the validity of the approach of this research and compare their actual data with predicted data. We first conduct statistics on the relevant variable data of the Shanghai Stock Exchange Index. The results are shown in Table 3.

To verify the effectiveness of the model, we first make a statistical forecast of earnings per share. The results are shown in Figure 7.

From the perspective of the predicted results and actual values, the values between the two prediction methods used in this article are close to the official data. From a horizontal comparison, the difference between the SVM predicted value and the true value is smaller than that in the MIDAS forecast model. Meanwhile, the SVM forecast should be more in line with the actual situation.

For the return on net assets, the actual situation and forecast are shown in Figure 8.

Figure 5. Comparison of predicted values

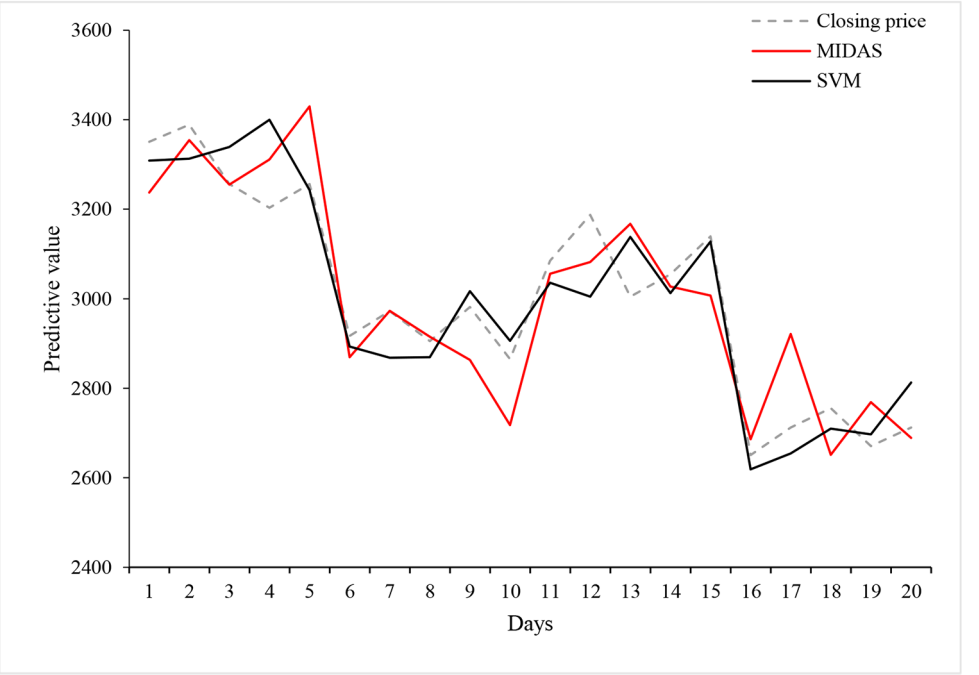


Figure 6. Prediction and standard residual plot

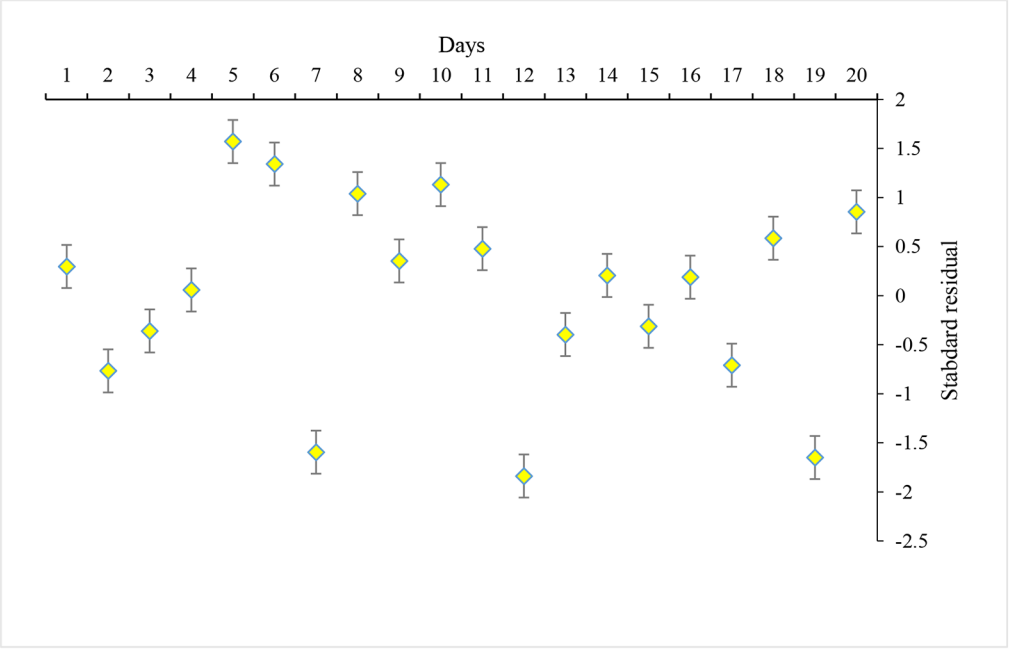


Table 3. Shanghai Composite Index related variables

Date	Opening	Highest	Lowest	Closing price of the day	Volume	Closing price the next day
2019/2/3	1867.45	1963.35	1821.36	1962.45	56632170	1906.21
2019/2/4	2028.05	2035.62	2005.07	2006.21	68131300	1883.55
...	...	...	...	...	...	...
2019/2/23	2739.02	2801.44	2723.51	2801.61	84435545	2783.37

Figure 7. Earnings per share forecast

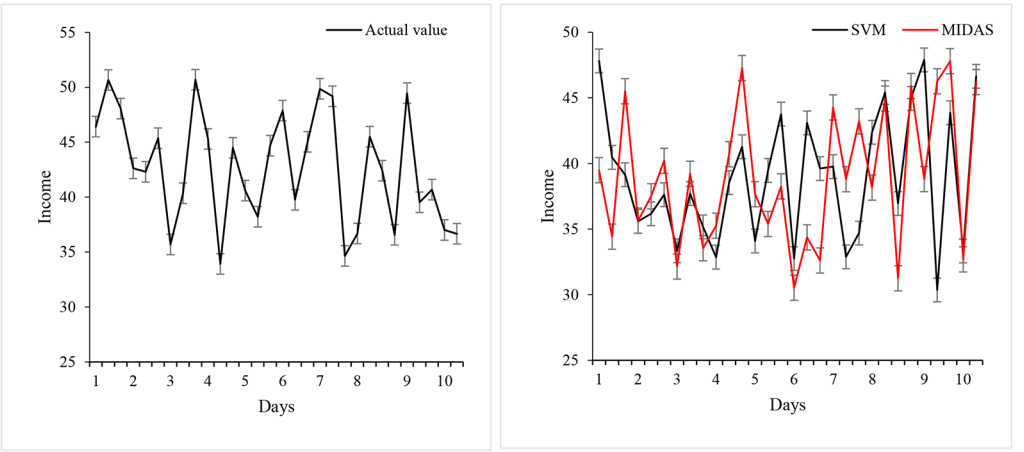


Figure 8. Comparison of return on net assets

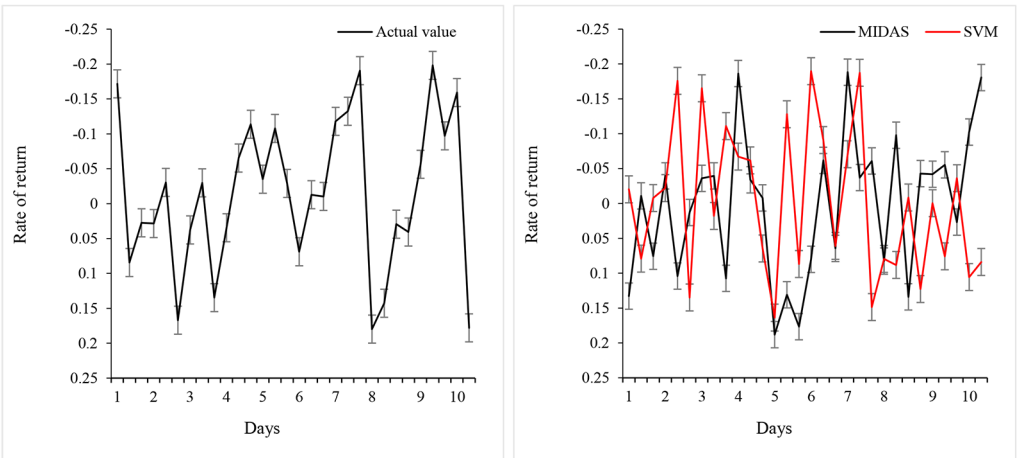
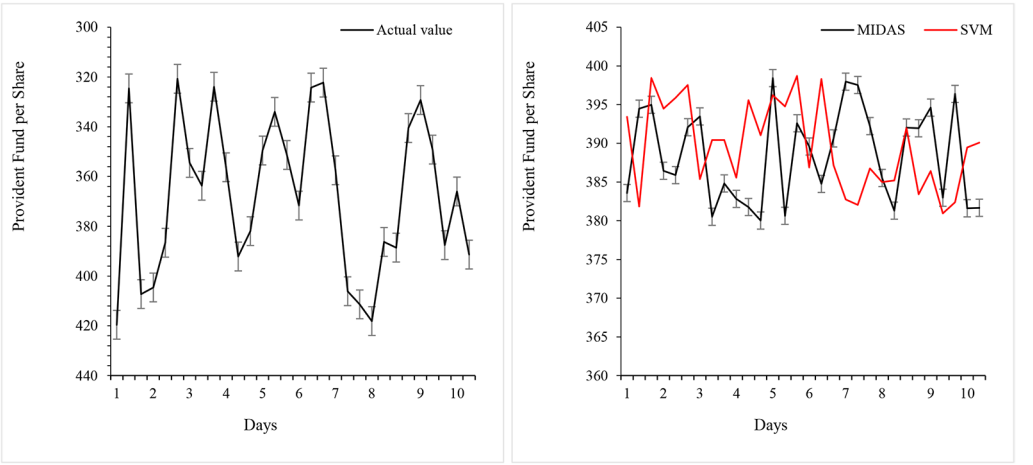


Figure 9. Provident fund per share and forecast



In the forecast of the return on equity, we can see that in the first three days, the forecast accuracy of SVM is better than that of the MIDAS model; however, in the middle five days, the forecast accuracy of SVM is slightly low. As time passes, the prediction accuracy of SVM also improves.

We make statistics on the provident fund per share and the forecast situation in these statistics. The results are presented in Figure 9.

Figure 10. Net assets per share

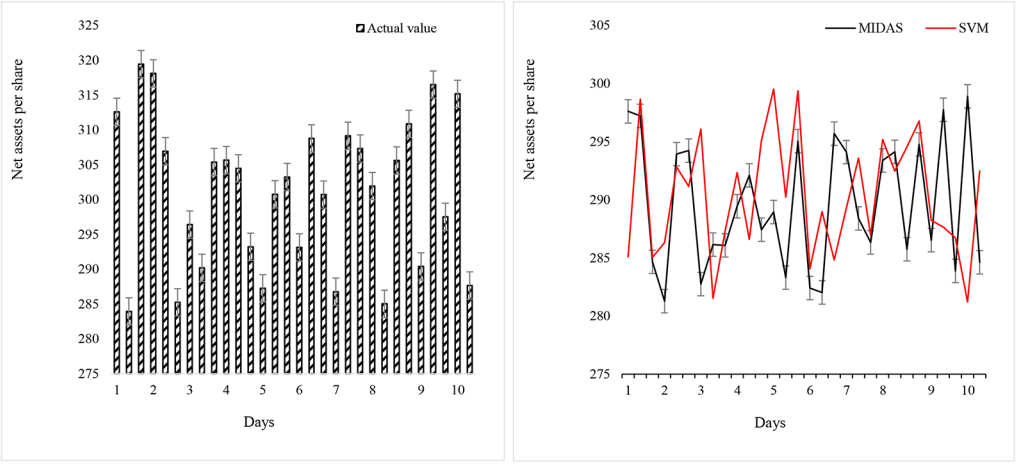


Figure 9 displays that when predicting the provident fund per share, due to insufficient relevant data, the two models used in the 10-day forecast are not that good in the forecasting effect. The forecast result is relatively close to the median, but a certain gap is observed in the specific number of forecast days. This gap may be caused by insufficient relevant data parameters.

We make statistics and forecasts on the net assets per share of the Shanghai Composite Index, and the results are shown in Figure 10.

For forecasting net assets per share, there are better forecasting results of MIDAS and SVM used in this paper. In general, MIDAS has better medium-term to medium-term forecasting results, while the prediction accuracy of the SVM model is 10% higher than that of MIDAS. However, with respect to support vector machines, the prediction results outperformed those of MIDAS in the early and short term. This occurrence is also due to the difference in construction between the two models.

DISCUSSION

Stock price prediction has been a question that many investors have been discussing since the emergence of the stock market. As a result, two schools of fundamental and technical analyses have gradually formed. In the 1980s, data mining technology began to be applied to stock price forecasting, and it has gradually become an important tool for domestic and foreign scholars to explore the laws of the stock market.

The article selects technical and financial indicators. Given that technical indicators reflect the information of the transaction data in terms of trend, momentum, volatility, and transaction volume, we do not use the screening method and mainly use more than 20 indicators common in quantitative investment work. These indicators are data converted from the original transaction data through mathematical formulas. After normalization, these data can be directly used as input vectors for neural networks and SVMs. When selecting financial indicators, we first use the random forest method to filter important variables, and then filter according to the target company's own accounting subjects, collinearity, and missing data.

When selecting financial indicators, we first obtain all the data of the Shanghai Stock Index and use random forest to filter. Random forest does not consider the collinearity among variables when selecting variables and only filters on the basis of the importance of the variables. In the empirical analysis, the actual accounting subjects and actual meanings of the target company are selected, both of which are comprehensive and pertinent.

By drawing on the latest theoretical results regarding classical MIDAS series models, neural networks and support vector machines, we present in this paper. For the problem of objective data forecasting in stock index forecasting, integrated multidisciplinary research in methodology and results, which constructs a data forecasting model, and applies it to the economic forecasting of SSE index. Modeling predictions are also made. First, regardless of other factors, based solely on market trading performance, MIDAS and SVM methods are used to predict the daily closing price of individual stocks in the next month. The prediction results of the two methods are compared with the real value, and the prediction of MIDAS is found. MIDAS is found better than SVM. Second, multiple linear regression is conducted to predict the average daily closing price of the next month after the financial report is disclosed; we find a strong collinearity among financial index data. A ridge regression, which can overcome the multicollinearity problem, is also adopted; we observe that it is the same as the true value. The residual error is relatively large, and the forecast effect is larger than the daily forecast. However, the overall trend is roughly the same. Finally, the weighted combination of the two prediction results using the inverse mean square error ratio method is better than MIDAS. Therefore, we believe that combining market transaction forecasts with financial index forecasts is meaningful.

CONCLUSION

The model constructed in this research can deal with missing data and inconsistencies among independent variables, a characteristic that is theoretically innovative. According to the historical transaction data of individual stocks and the previous financial report data, the daily closing price and the average closing price of the next month are respectively predicted. The former adopts neural network and SVM methods, which are suitable for processing a large amount of historical transaction

data and have strong self-learning ability. Through comparison, we find that neural network has more advantages in short-term prediction than SVM. In terms of forecasting based on financial indicators, due to the uncertainty about the length of time, the financial report will affect the stock price; moreover, the average daily closing price of the next month after disclosing the financial report is used as the response variable. This article also has certain shortcomings. Data mining technology, stock data analysis, and data mining applications are involved in this study. However, due to the existence of data acquisition, basic theoretical knowledge, technology and model applications, few defects are noticed. Moreover, the actual application of correct stock prediction in the field of data mining still needs to be further explored and studied.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors have declared to have no competing interests.

REFERENCE

Barbarelli, S., Amelio, M., & Florio, G. (2016). Predictive model estimating the performances of centrifugal pumps used as turbines. *Energy*, *107*, 103–121. doi:10.1016/j.energy.2016.03.122

Bergey, F., Saccenti, E., Jonkers, D., van den Heuvel, T., Jeuring, S., Pierik, M., & Martins dos Santos, V. (2017). P316 New approaches for IBD management based on text mining of digitalised medical reports and latent class modelling. *Journal of Crohn's and Colitis*, *11*(suppl_1), S237–S238. doi:10.1093/ecco-jcc/jjx002.441

Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A., & Plaza, A. (2015). On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(10), 4634–4646. doi:10.1109/JSTARS.2015.2458855

Chatterjee, U. K. (2016). Do stock market trading activities forecast recessions? *Economic Modelling*, *59*, 370–386. doi:10.1016/j.econmod.2016.08.007

Djouvas, C., Mendez, F., & Tsapatsoulis, N. (2016). Mining online political opinion surveys for suspect entries: An interdisciplinary comparison. *Journal of Innovation in Digital Ecosystems*, *3*(2), 172–182. doi:10.1016/j.jides.2016.11.003

Falkenthal, M., Barzen, J., Breitenbücher, U., Brügmann, S., Joos, D., Leymann, F., & Wurster, M. (2017). Pattern research in the digital humanities: How data mining techniques support the identification of costume patterns. *Computer Science-Research and Development*, *32*(3), 311–321. doi:10.1007/s00450-016-0331-6

Fan, R., Taylor, S. J., & Sandri, M. (2018). Density forecast comparisons for stock prices, obtained from high-frequency returns and daily option prices. *Journal of Futures Markets*, *38*(1), 83–103. doi:10.1002/fut.21859

Faria, M., Prats, E., Padrós, F., Soares, A. M., & Raldúa, D. (2017). Zebrafish is a predictive model for identifying compounds that protect against brain toxicity in severe acute organophosphorus intoxication. *Archives of Toxicology*, *91*(4), 1891–1901. doi:10.1007/s00204-016-1851-3 PMID:27655295

Forsberg, D., Rosipko, B., & Sunshine, J. L. (2015). Analyzing PACS Usage Patterns by Means of Process Mining: Steps Toward a More Detailed Workflow Analysis in Radiology. *Journal of Digital Imaging*, *29*(1), 47–58. doi:10.1007/s10278-015-9824-2 PMID:26353749

Gui, G., Pan, H., Lin, Z., Li, Y., & Yuan, Z. (2017). Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection. *KSCE Journal of Civil Engineering*, *21*(2), 523–534. doi:10.1007/s12205-017-1518-5

Guo, J., Chen, Z., Ban, Y. L., & Kang, Y. (2014). Precise enumeration of circulating tumor cells using support vector machine algorithm on a microfluidic sensor. *IEEE Transactions on Emerging Topics in Computing*, *5*(4), 518–525. doi:10.1109/TETC.2014.2335539

Jeon, S., Hong, B., & Chang, V. (2018). Pattern graph tracking-based stock price prediction using big data. *Future Generation Computer Systems*, *80*, 171–187. doi:10.1016/j.future.2017.02.010

Jochems, A., Deist, T. M., Van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P., & Dekker, A. (2016). Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, *121*(3), 459–467. doi:10.1016/j.radonc.2016.10.002 PMID:28029405

Kelley, B. P., Klochko, C., Halabi, S., & Siegal, D. (2016). Datafish multiphase data mining technique to match multiple mutually inclusive independent variables in large PACS databases. *Journal of Digital Imaging*, *29*(3), 331–336. doi:10.1007/s10278-015-9817-1 PMID:26572132

Kim, G. H., & Kim, S. H. (2019). Variable selection for artificial neural networks with applications for stock price prediction. *Applied Artificial Intelligence*, *33*(1), 54–67. doi:10.1080/08839514.2018.1525850

Kinnebrew, J. S., Killingsworth, S. S., Clark, D. B., Biswas, G., Sengupta, P., Minstrell, J., Martinez-Garza, M., & Krinks, K. (2016). Contextual markup and mining in digital games for science learning: Connecting player behaviors to learning goals. *IEEE Transactions on Learning Technologies*, *10*(1), 93–103. doi:10.1109/TLT.2016.2521372

Lee, M., & Lee, H. J. (2017). Stock Price Prediction by Utilizing Category Neutral Terms: Text Mining Approach. *Journal of Intelligent Information Systems*, *23*(2), 123–138. doi:10.26493/1854-6935.17.335-351

Liu, C., Wang, W., Wang, M., Lv, F., & Konan, M. (2017). An efficient instance selection algorithm to reconstruct training set for support vector machine. *Knowledge-Based Systems*, *116*, 58–73. doi:10.1016/j.knosys.2016.10.031

Maulik, U., & Chakraborty, D. (2017). Remote Sensing Image Classification: A survey of support-vector-machine-based advanced techniques. *IEEE Geoscience and Remote Sensing Magazine*, *5*(1), 33–52. doi:10.1109/MGRS.2016.2641240

Pisani, F., Facini, C., Pelosi, A., Mazzotta, S., Spagnoli, C., & Pavlidis, E. (2016). Neonatal seizures in preterm newborns: A predictive model for outcome. *European Journal of Paediatric Neurology*, *20*(2), 243–251. doi:10.1016/j.ejpn.2015.12.007 PMID:26777334

Raff, A. B., Weng, Q. Y., Cohen, J. M., Gunasekera, N., Okhovat, J. P., Vedak, P., Joyce, C., Kroshinsky, D., & Mostaghimi, A. (2017). A predictive model for diagnosis of lower extremity cellulitis: A cross-sectional study. *Journal of the American Academy of Dermatology*, *76*(4), 618–625. doi:10.1016/j.jaad.2016.12.044 PMID:28215446

Sharma, A., Hostetter, J., Morrison, J., Wang, K., & Siegel, E. (2016). Focused decision support: A data mining tool to query the prostate, lung, colorectal, and ovarian cancer screening trial dataset and guide screening management for the individual patient. *Journal of Digital Imaging*, *29*(2), 160–164. doi:10.1007/s10278-015-9826-0 PMID:26385814

Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. *Knowledge-Based Systems*, *120*, 4–14. doi:10.1016/j.knosys.2016.12.019

Tharwat, A., Hassanien, A. E., & Elnaghi, B. E. (2017). A BA-based algorithm for parameter optimization of support vector machine. *Pattern Recognition Letters*, *93*, 13–22. doi:10.1016/j.patrec.2016.10.007

Wang, C., Wang, X., Zhang, C., & Xia, Z. (2017). Geometric correction based color image watermarking using fuzzy least squares support vector machine and Bessel K form distribution. *Signal Processing*, *134*, 197–208. doi:10.1016/j.sigpro.2016.12.010

Wang, H., Zhou, Z., & Xu, Y. (2018). An improved ν-twin bounded support vector machine. *Applied Intelligence*, *48*(4), 1041–1053. doi:10.1007/s10489-017-0984-2

Wang, W., Xi, J., Chong, A., & Li, L. (2017). Driving style classification using a semisupervised support vector machine. *IEEE Transactions on Human-Machine Systems*, *47*(5), 650–660. doi:10.1109/THMS.2017.2736948

Widyantara, I. M. O., Asana, I. M. D. P., Wirastuti, N. M. A. E. D., & Adnyana, I. B. P. (2017). An Automated Approach of Shoreline Detection Applied to Digital Videos using Data Mining. *Research Journal of Applied Sciences, Engineering and Technology*, *14*(3), 101–111. doi:10.19026/rjaset.14.4152

Zhang, W., Zhang, S., Zhang, S., Yu, D., & Huang, N. (2017). A multi-factor and high-order stock forecast model based on Type-2 FTS using cuckoo search and self-adaptive harmony search. *Neurocomputing*, *240*, 13–24. doi:10.1016/j.neucom.2017.02.054

Zhang, Y. J., & Wang, J. L. (2019). Do high-frequency stock market data help forecast crude oil prices? Evidence from the MIDAS models. *Energy Economics*, *78*, 192–201. doi:10.1016/j.eneco.2018.11.015

To-Han Chang is a Professor in Department of History at Minjiang University, China. He received his Ph.D. in Economics from National Central University, Taiwan. His research interests are empirical industrial organization and quantitative analysis. He has published his research in Technology Analysis & Strategic Management, Total Quality Management & Business Excellence, and Asia-Pacific Journal of Accounting & Economics.

Nientsu Wang is a Professor of Management at Minjiang University, China. He earned his Ph.D in Management from Peking University, China. He received his master degree of Communication from Nanhua University, Taiwan and a bachelor of history from National Chung Hsing University, Taiwan. His research interests are bibliometrics and digital publication.

Wen-Bin Chuang is an associate professor in management at the Department of International Business Studies, National Chi-Nan University in Taiwan. He received his Ph.D. in Economics from National Central University, Taiwan. His research focuses on Technology Management, International Business, and Financial Management. He has published his research in Technology Analysis & Strategic Management, Total Quality Management & Business Excellence, and Asia-Pacific Journal of Accounting & Economics.