Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 3189238, 13 pages
https://doi.org/10.1155/2018/3189238

Research Article

Forecasting of Short-Term Metro Ridership with Support Vector Machine Online Model

Xuemei Wang,[1] **Ning Zhang**,[2] **Yunlong Zhang**,[3] **and Zhuangbin Shi**[2]

[1]*Department of Transportation Engineering, Tongji Zhejiang College, Jiaxing 314000, China*
[2]*Intelligent Transportation System Institute of Ministry of Education, Southeast University, Nanjing 211189, China*
[3]*Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, USA*

Correspondence should be addressed to Ning Zhang; ningzhang1972@gmail.com

Received 18 January 2018; Revised 6 April 2018; Accepted 26 April 2018; Published 27 June 2018

Academic Editor: David F. Llorca

Forecasting for short-term ridership is the foundation of metro operation and management. A prediction model is necessary to seize the weekly periodicity and nonlinearity characteristics of short-term ridership in real-time. First, this research captures the inherent periodicity of ridership via seasonal autoregressive integrated moving average model (SARIMA) and proposes a support vector machine overall online model (SVMOOL) which insets the weekly periodic characteristics and trains the updated data day by day. Then, this research captures the nonlinear characteristics of the ridership via successive ridership value inputs and proposes a support vector machine partial online model (SVMPOL) which insets the nonlinear characteristics and trains the updated data of the predicted day by time interval (such as 5-min). Afterwards, to avoid the drawbacks and to take advantages of the strengths of the two individual online models, this research takes the average predicted values of two models as the final predicted values, which are called support vector machine combined online model (SVMCOL). Finally, this research uses the 5-min ridership at Zhujianglu and Sanshanjie Stations of Nanjing Metro to compare the SVMCOL model with three well-known prediction models including SARIMA, back-propagation neural network (BPNN), and SVM models. The resultant performance comparisons suggest that SARIMA is superior for the stable weekday ridership to other models. Yet the SVMCOL model is the best performer for the unstable weekend ridership and holiday ridership. It shows that for metro operation manager that gear toward timely response to real-world unstable and abnormal situations, the SVMCOL may be a better tool than the three well-known models.

1. Introduction

Short-term ridership forecasting is a vital component of metro operation and management. Accurate predictions can reflect real-time changes in ridership. The prediction results can become important inputs for decision-making in evaluating rail transit service level and system operating status and provide an important basis for station passenger crowd regulation and emergency response. In addition, short-term ridership forecasting is the key to the success of revenue management for railway operators [1].

In the last two decades, traditional metro ridership forecasting is based on travel demand forecasting models including the steps of trip generation, trip distribution, mode choice, and assignment [2, 3]. This type of long-term forecasting has been applied in the planning and construction of metro, but it cannot be adapted to the needs of the operations management.

Though the spatial-temporal characteristics of metro ridership are not completely the same as those for vehicle traffic flow [4], short-term forecasting methods can also be divided into two categories: the theory driven method and the data driven method. Theory driven method is based on traffic flow mechanism to investigate traffic dynamics [5, 6]. The data driven method on the other hand is based on the data of traffic flow series itself to construct models and make predictions. The data driven model is the main method of short-term prediction and can be divided into linear, nonlinear, and hybrid forecasting methods. The linear forecasting method mainly includes time series model [7–9] and Kalman filtering model [10–12]. The nonlinear forecasting method includes nonparametric regression [13, 14], neural network algorithm [15–17],

support vector machine [18–20], and Gaussian maximum likelihood model [21]. The hybrid forecasting method combines at least two methods for prediction to achieve better performance in accuracy and reliability. Hybrid models mainly include wavelet decomposition hybrid model [22, 23], Bayesian decomposition hybrid model [24, 25], empirical mode decomposition hybrid model [26], neural network hybrid model [27, 28], and support vector machine hybrid model [29–35].

Whether it is traffic flow or passenger, time series model has become one of the classic models of short-term flow prediction [36]. Of all the time series models, seasonal autoregressive integrated moving average (SARIMA) model considers the periodicity feature of the time series, so it can capture the inherent periodicity of traffic flow data. Williams et al. [9–11] used the SARIMA model for short-term traffic flow prediction and verified its good performance. But time series model is a linear model, and its prediction performance may worsen significantly if the time series are nonstationary and nonlinear. Nevertheless SARIMA model is widely used to be the benchmark to evaluate the forecasting performance of a novel model.

Neural networks are among the most widely used nonlinear models. A neural network trains neurons based on historical data, maps the complicated nonlinear relation between input and output data, and uses the relationship for predictions for given inputs. Neural network algorithms have the adaptive and learning advantages and are flexible without the need to construct detailed and explicit models like other methods. Vlahogianni et al. [37] optimized neural network structure to forecast urban traffic flow parameters. But the neural network algorithm cannot make expected risk minimization because of the empirical risk minimization principle that may also lead to two major drawbacks: local minima and overfitting [38]. The local minima are associated with the training process of neural network, which is to minimize the difference between the predicted outputs and the observed outputs by optimizing the network weights. Overfitting leads to poor generalization ability and may produce inaccurate predictions with some particular testing data.

Compared with neural network algorithm, support vector machine (SVM) model can strike a compromise between prediction accuracy and generalization ability based on the structural risk minimization principle. With the help of intelligent use of kernel function, SVM can solve the problems of small sample, nonlinearity and the curse of dimensionality, overfitting, and local minima. Zhang and Xie [19] proposed a υ-support vector machine model for short-term traffic volume prediction and showed that it outperformed the multilayer feed-forward neural network (MLFNN) model. Zhang et al. [30] proposed a novel hybrid model that identified the SVM input dimensions via SARIMA model to forecast short-term traffic volume, taking advantage of the individual strengths of the two models. Hong [33] presented a traffic flow forecasting model to forecast interurban traffic flow, which combines the seasonal support vector regression model with chaotic immune algorithm (SSVRCIA), and yielded more accurate forecasting results than the SARIMA, BPNN, and seasonal Holt–Winters models. Wang and Shi [34] constructed a new kernel function using a wavelet function to capture the nonstationary characteristics of the short-term traffic speed data, proposed a short-term traffic speed forecasting hybrid model (Chaos-Wavelet Analysis-Support Vector Machine model, C-WSVM), and achieved the encouraging results. Chen et al. [35] proposed an approach which hybridizes SVR model with adaptive genetic algorithm (AGA) and the seasonal index adjustment, namely, AGA-SSVR, to forecast holiday daily tourist flow.

The research of short-term metro ridership forecasting is a rather new undertaking. Tsai et al. [1] proposed two novel neural network structures based on temporal feature extraction and successfully applied them in railway short-term passenger demand forecasting in Taiwan. Wei and Chen [26] used empirical mode decomposition to extract neural network input variables to forecast the short-term ridership of Taipei Rapid Transit Muzha Line. Sun et al. [29] proposed a novel hybrid model Wavelet-SVM, and the experimental results showed that the approach has appeared to be the promising and robust. These studies indicated that metro ridership has significant characteristics of periodicity and nonlinearity reflecting a variety of factors; however, how these characteristics are embedded into the model without affecting the computational complexity of the model is worth discussing. And, for neural network or support vector model, the previous literature also did not discuss the training time to see if it meets the demand of practical operation. If the training time is too long and leads to serious forecasting delay, the prediction model cannot meet the demand of practical operation even if it has good prediction performance. In addition, most existing research on short-term metro ridership forecasting focused mainly on normal situations; it is not clear how the applicability and the prediction accuracy of the model is when it comes to holidays, inclement weather, large sports events, or emergencies. Sun et al. [29] selected the data including a Valentine's Day (not a major holiday) as training data, not as a predictor. Finally, the short-term prediction interval is long (i.e., 15-min) in these literatures, and, for the actual operation of the metro, it cannot meet the requirement of the operator because the departure intervals are short.

The reliability and the operability of the models play a crucial role in the accuracy and real-time implementation of the prediction, so the choice of the model is very important in a practical application. Since the characteristics of metro ridership are quite different from those in other transportation systems, most of forecasting models provide unsatisfactory prediction effectiveness. After comparing time series model, neural network model, and SVM model, this paper selects SVM model as the base short-term prediction model, considering capturing in real-time the periodicity and nonlinearity characteristics of short-term ridership as mentioned previously. With this base model, this paper proposes a support vector machine overall online (SVMOOL) model, which extracts input features via SARIMA model, trains the updated data by day, and optimizes the parameters by a particle swarm optimization (PSO) algorithm, to capture the periodicity of ridership in real-time. This paper also proposes a support vector machine partial online (SVMPOL) model, which extracts input features based on the temporal continuity of ridership

model, trains the updated data by time intervals (such as 5-min), and also optimizes the parameters by a PSO algorithm to capture the nonlinearity of ridership. Afterwards, the support vector machine combined online (SVMCOL) model is proposed by combining the SVMOOL model and the SVMPOL model.

The main contributions of this paper are as follows.

(1) This paper proposes a novel hybrid model combining the SVMOOL model and the SVMPOL model for short-term ridership forecasting that better captures the periodicity and nonlinearity characteristics by the updated data set. The SVMCOL model takes advantages of the individual strengths of the two models. The actual results of 5-min short-term ridership forecasting show the feasibility and effectiveness of the proposed combined model in real-time implementation.

(2) While the SARIMA model is superior for the stable weekday ridership to other models, experiments results indicate that the SVMOOL model is superior to SARIMA model, BPNN model, or SVM model in terms of MAE and RMSE for the weekend and holiday ridership test. It should be noted that the prediction of ridership under abnormal situations (such as holiday) is evidently more challenging than doing so under normal conditions (such as weekday ridership) and, hence, is much desired by the operator. Therefore, the proposed SVMCOL model is found to be suitable and useful in real-world operations.

(3) The experiments using LibSVM package on desktop computers indicate that the SVMOOL model needs about one hour for three weeks' data (4284 observations) to construct the prediction function and the forecasting time takes less than 1 second for a one-step prediction using SVM. In the process of the implementation experiments, the SVMPOL model needs less than 1 s to construct due to the small data sample and the forecasting time needs less than 1 s for a one-step prediction. Therefore, the training time and the forecasting time can meet the real-time demand for the one-step prediction in implementation as well.

(4) In general, short-term forecasting represents prediction for a specific time interval, such as 5 min, 10 min, and 15 min. For metro ridership, 5-min interval will be more useful for metro operation and management because the departure interval of the metro vehicle is really short. In addition, it is obvious that ridership during workdays is different from that on weekends or holidays. As discussed by Chen et al., some prediction models that work well for workdays data may yield unsatisfactory results for weekends or holidays data. In order to discuss the applicability of the proposed model, three samples were selected. The first sample contains weekdays, weekends, and no holidays, and the second and third samples contain weekdays, weekends, and holidays.

This paper attempts to develop an online hybrid model to improve the forecasting performance of metro ridership. The rest of this paper is organized in the following manner. A brief theoretical background of the SVM model is presented first, followed by detailed description on SVMOOL model, SVMPOL model, and SVMCOL model. After that, a brief description of the data source and the implementation of the models are given. Finally, results analysis and conclusions are presented.

2. Methodology

To introduce the SVMOOL, SVMPOL, and SVMCOL models, SVM model is illustrated here first.

2.1. Support Vector Machine for Regression. A detailed description of SVM algorithm is given in Vapnik [38]. Assume that training input data and the corresponding training output data are (x_i, y_i) $(i = 1, 2, \ldots, m)$, where $x_i \in X \subseteq R^n$ and $y_i \in Y \subseteq R^n$, and m denotes the total number of data. The basic idea of SVM is to map the low-dimensional input space to the high-dimensional feature space using a function $\Phi(x_i)$. The linear regression function can be stated as

$$f(x) = w^T \Phi(x_i) + b, \tag{1}$$

where w and b are coefficients. For SVM, these coefficients can be obtained by solving the following optimization problems:

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{m} (\xi_i^+ + \xi_i^-) \\ \text{subject to} \quad & y_i - w^T \Phi(x_i) - b \le \varepsilon + \xi_i^+ \\ & w^T \Phi(x_i) + b - y_i \le \varepsilon + \xi_i^- \\ & \xi_i^+ \ge 0, \ \xi_i^- \ge 0, \ i = 1, 2, \ldots, n, \end{aligned} \tag{2}$$

where $\varepsilon(\ge 0)$ is the insensitive loss function, ξ_i^+ and ξ_i^- are slack variables, and C is a regularization parameter. The maximal dual function in (2) has the following form:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} y_i (\alpha_i^* - \alpha_i) - \varepsilon \sum_{i=1}^{n} (\alpha_i^* + \alpha_i) \\ & - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \\ \text{subject to} \quad & \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0 \\ & \alpha_i \ge 0, \ \alpha_i^* \le C, \ i = 1, 2, \ldots, n, \end{aligned} \tag{3}$$

where α_i^* and α_i are Lagrange multipliers.

Ultimately, the decision function given by (1) has the explicit form:

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)(\Phi(x_i) \bullet \Phi(x)) + b \\ &= \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(x_i, x) + b, \end{aligned} \tag{4}$$

where $K(x_i, x)$ is the kernel function. There are several types of kernel functions, including polynomial, radial basis, and sigmoid. Generally, a Gaussian radial basis function (see (5)) is widely used because of better prediction performance:

$$K(x_i, x) = \exp\left(-\gamma |x - x_i|^2\right). \tag{5}$$

2.2. Input Features and Parameter Optimization. Identifying input features is crucial step in SVM modeling. Metro ridership has significant characteristics of periodicity and nonlinearity. Abe [39] discovered that excessive features caused not only long training time but also poor generalization ability. Some researchers documented in detail the identification of input features. For example, Zhang et al. [30] identified the SVM input dimensions via SARIMA. Wu et al. [40] extracted input features from successive actual values before the prediction time; that is to say, if the value y_t of future time t is regarded as output, then the real values $y_{t-1}, y_{t-2}, \ldots, y_{t-m}$ of past time $t-1, t-2, \ldots, t-m$ serve as inputs. Cao et al. [41] used principal component analysis, kernel principal component analysis, and independent component analysis for inputs extraction. Huang and Wang [42] and Lin et al. [43] used genetic algorithm (GA) and particle swarm optimization (PSO) algorithm to extract input features, respectively.

Parameter optimization is to obtain better forecasting accuracy of the SVM model. The parameters optimized are mainly the penalty coefficient, the insensitive loss coefficient, and the corresponding parameters of kernel function. The LibSVM package [44] uses the grid-searching algorithm combined by cross-validation to determine these parameters but the process takes lengthy computation time. Hong et al. [45], Lin et al. [43], and Hong et al. [45] successfully used GA, PSO, and the ant colony optimization (ACO) algorithm to find the most optimal parameters, respectively. The advantages of PSO lie in easier application, fewer parameters to adjust, and faster convergence to optimum. As a result, PSO is used to optimize the parameters in this study. PSO simulates social behavior, like birds flocking to a promising position, to achieve precise objectives in a multidimensional space [46]. PSO gains the optimal solution through collaboration between individuals.

2.3. Support Vector Machine Online Model

2.3.1. Support Vector Machine Overall Online Model. Support vector machine overall online (SVMOOL) model is based on the theory of SVM, to extract input features, to train the batched updated training data, to use intelligent algorithms, to find the optimal parameters, and to get time-varying prediction function to realize the short-term forecasting.

Due to apparent periodicity feature of the rail transit ridership, SARIMA model is used to extract input features because SARIMA model is able to capture the periodicity of time series. A time series $\{X_t,\ t = 1, 2, \ldots, n\}$ is generated by the SARIMA(p,d,q)(P,D,Q)$_s$ process of Box and Jenkins as described by Williams et al. [8, 10] and Zhang et al. [30] described the process how to extract the features via SARIMA model in detail.

Considering the computation time of the training data and the real-time demand of the one-step prediction, the SVMOOL model is constructed by updating the training data day by day. That is to say, the training data is updated by adding the ridership data of the most recent day, and the time-varying prediction function is then constructed. Stating in simpler words, assume that y_{ij} denotes the ridership value at time j of day i, $i \in \{1, 2, \ldots, i, i+1, i+2, \ldots\}$, $j \in \{1, 2, \ldots, j, j+1, j+2, \ldots, n\}$, where n denotes the number of the data points each day. All of the prediction values of ridership after day i are forecasted by the training data of $i \times n$ ridership values. According to SVMOOL model described above, the prediction function is obtained by using the SARIMA model to extract input features from the training data and using PSO algorithm to optimize parameters, then forecasting value of every time interval j, $j = 1, 2, \ldots, n$, until the real values of day $i+1$ are totally obtained. After that, the training data is updated by adding the actual ridership values of day $i+1$. New prediction function is then constructed to forecast every value of day $i+2$, by retraining data and updating the parameters, and the process repeats. This process of constructing SVMOOL model is shown in Figure 1.

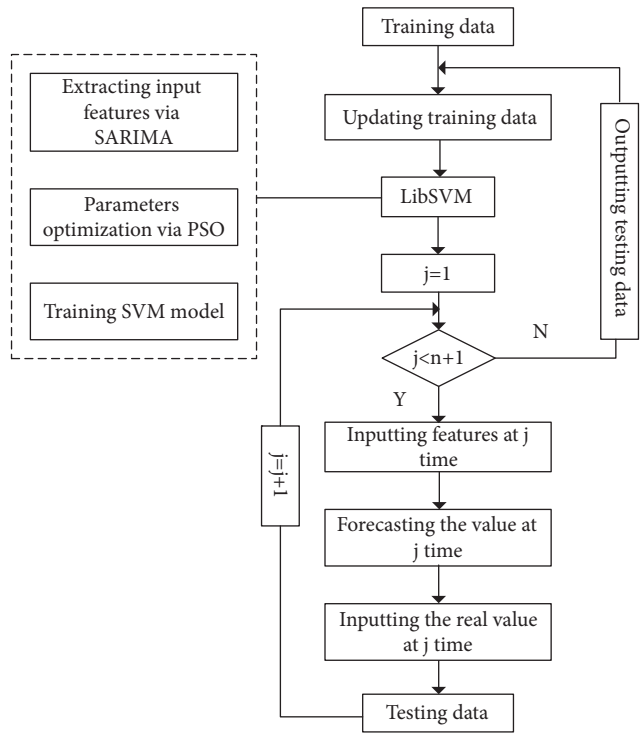

Figure 1: The process of constructing SVMOOL model.

2.3.2. Support Vector Machine Partial Online Model. Support vector machine partial online (SVMPOL) model is also based on the theory of SVM, to extract input features, to train the real-time updated testing data, to use intelligent algorithm, to find the optimal parameters, and to get real-time prediction function to realize the short-term forecasting.

According to the input feature extraction approaches mentioned previously and considering the temporal continuity of the real-time data, SVMPOL model extracts input features from successive actual values before the prediction time to capture nonlinear features of the ridership. In addition, parameters are also optimized by PSO.

The SVMPOL model makes full use of the temporal continuity of ridership data and takes advantage of SVM's capability of addressing small samples. The testing data is updated

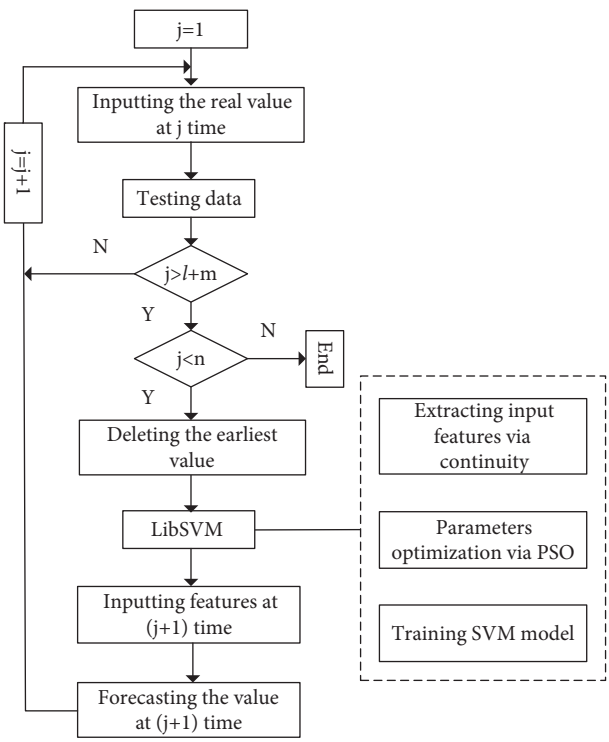

Figure 2: The process of constructing SVMPOL model.

by adding the ridership value of every time interval of the prediction day at same time deleting the earliest value. The real-time forecast function is obtained by training updated data and optimizing parameters in real-time to predict the value in the next time until the end of the prediction day. Stating in simpler words, assume that y_j denotes the ridership value at time j of the prediction day, $j \in \{1, 2, \ldots, j, j+1, j+2, \ldots, n\}$, where n denotes the number of the data points every day. The rest of the ridership after time j needs to be forecasted with the j passenger values. According to SVMPOL model as described above, the prediction function is obtained by extracting successive ridership values prior to the prediction time as the inputs and using PSO to optimize parameters, then the ridership corresponding output in time $j+1$ is achieved. After that, the testing data is updated by adding the actual value of time $j+1$ and deleting the earliest data. New prediction function is then constructed by retraining data and updating the parameters to forecast the ridership values in time $j+2$, and the process continues. This process of constructing SVMPOL model is shown in Figure 2, where l denotes the size of the moving window and m denotes the number of the input features via continuity.

2.3.3. Support Vector Machine Combined Online Model. As described previously, this paper proposes a SVMOOL model to address the periodicity of ridership and a SVMPOL model to address the nonlinearity of ridership. But the SVMOOL model updates the training data day by day and cannot capture the real-time local variations of ridership on the day being predicted. And considering the computation time of the testing data and the real-time demand of the one-step prediction, the testing data contains one-day data at most for constructing the SVMPOL model and the internal mechanism of metro ridership to study is insufficient. To avoid the drawbacks and to take advantages of the strengths of the two individual online models, the average predicted values of two models are the final results, which are called support vector combined online (SVMCOL) model.

3. Data Set and Evaluation Criteria

3.1. Data Set. At present, Automatic Fare Collection (AFC) System has been able to realize real-time data collection of metro passengers in and out station records [47] (though there is a slight delay in data transmission.). By simple statistics, the ridership data can be achieved for the required time interval. That is to say, the short-term ridership data of metro can be collected online, which puts forward higher requirements for short-time prediction. Operators expect faster and more accurate predictions, in order to plan ahead to accommodate the changes in passenger flow.

A ridership dataset of metro is collected to investigate the validity of the proposed SVMOOL, SVMPOL, and SVMCOL model for forecasting short-term ridership. The dataset is collected from the entrance transaction records of Nanjing Metro's Automatic Fare Collection (AFC) Systems. In general, short-term forecasting represents prediction for a specific time interval, such as 5 min, 10 min, and 15 min. For metro ridership, 5-min interval will be more useful for metro operation and management because the departure interval of the metro vehicle is really short. Taking the operation time of Nanjing Metro into consideration, the time period of data collection for each day is from 6:00 AM to 11:00 PM. There are 204 observations collected with a 5-min interval every day. The collected data is divided into two sets of training data plus testing data. In addition, it is obvious that ridership during workdays is different from that on weekends or holidays. As discussed by [48], some prediction models that work well for workdays data may yield unsatisfactory results for weekends or holidays data. In order to discuss the applicability of the proposed model, three samples were selected. The first sample contains no holidays, and the second and third samples contain holidays of Ching-Ming Festival and May Day. The specific sample information is as follows.

3.1.1. Sample 1. The dataset is collected from the entrance transaction records of the Sanshanjie station during the period from November 5 to December 2, 2012, so there are 5712 observations in total for these 28 days. The first training data set is data collected from November 5 to November 25, and the first testing data set contains the remaining seven days' ridership values, or 1428 observations, as shown in Figure 3. The weekend ridership pattern is different from weekday's obviously and the metro ridership shows the weekly periodic characteristics. This is because the weekday ridership is mainly composed of the commuted passenger flow, which is more stable. To the contrary, the weekend ridership mainly consists of the leisure and travel passenger flow, which is relatively fluctuant and has the obvious nonlinear characteristics.

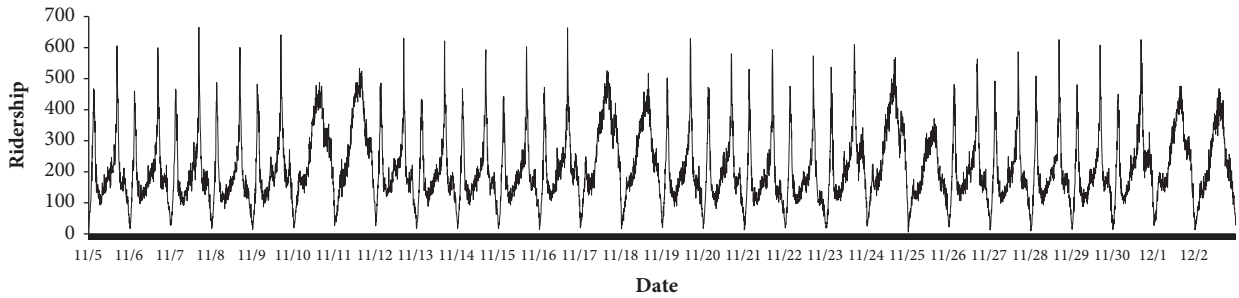

Figure 3: The origin entrance ridership time series at Sanshanjie Station of Nanjing Metro from November 5 to December 2, 2012.

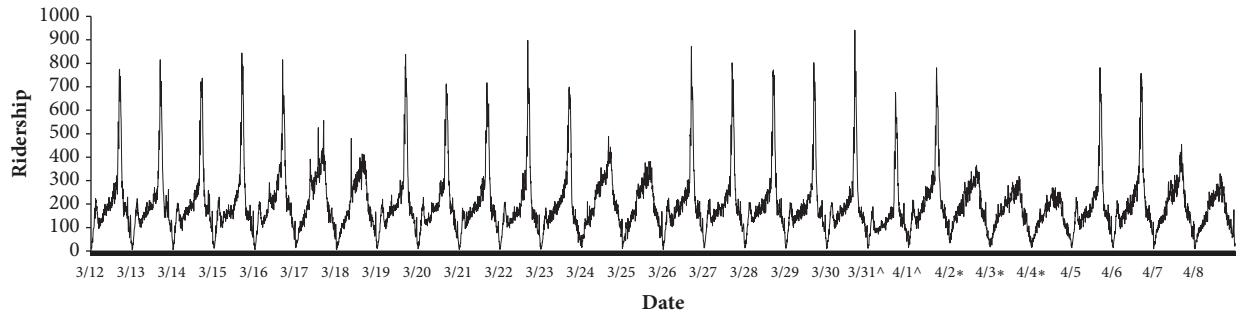

Figure 4: The origin entrance ridership time series at Zhujianglu Station of Nanjing Metro from March 12 to April 8, 2012. *Remark.* ∧ denotes that the weekend changes the weekday; ∗ denotes that of the holiday.

3.1.2. Samples 2 and 3. The dataset is collected from the entrance transaction records of the Zhujinglu station during the period from March 12 to May 6, 2012, so there are 11424 observations in total for these 56 days. The second training data set is data collected from March 12 to April 1, and the third training data set is data collected from April 9 to April 29. Both two training data sets contain three weeks' ridership values, or 4284 observations. Both two testing data sets contain the remaining seven days' ridership values, respectively, or 1428 observations, as shown in Figures 4 and 5. It must be noted that the Ching-Ming Festival is on April 2 (Monday) to April 4 (Wednesday); meanwhile March 30 (Saturday) and April 1 (Sunday) change weekday. The May Day is on April 29 (Sunday) to May 1 (Tuesday); meanwhile April 28 (Saturday) changes weekday. The holiday ridership pattern is similar to weekend's as a whole but is different from the local.

3.2. Data Normalization. Usually, normalizing raw input data can improve the convergence rate and performance of an SVM model. A common practice of data normalization was used to transform the raw data into a range $[-1, 1]$. In this study, each input data point is scaled according to

$$x_i^* = \frac{x_i - 0.5\left(x_{\max} + x_{\min}\right)}{0.5\left(x_{\max} - x_{\min}\right)}, \tag{6}$$

where x_i^* is the normalized value, x_i is any input vector of ridership data, $x_{\max}$ and $x_{\min}$ are, respectively, the maximum value and the minimum value of the training data in the period of training data.

3.3. Performance Indices. The mean absolute error (MAE), the mean absolute percent error (MAPE), and the root mean square error (RMSE) are commonly used criteria to evaluate the forecasting model. Generally, the smaller the MAE, MAPE and RMSE values, the better the prediction performance. The three performance criteria are, respectively, defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - \hat{y}_t\right|, \tag{7}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2}, \tag{9}$$

where y_t is the actual observed value in time t and $\hat{y}_t$ is the forecasting value in time t, and N is the number of the observations every day.

4. Model Implementation

In this section, specific applications of the SVMOOL, the SVMPOL, and the SVMCOL models described previously are addressed.

In the methodology section, several methods of choosing the appropriate input features are introduced. The SVMOOL model's input features are extracted using the SARIMA

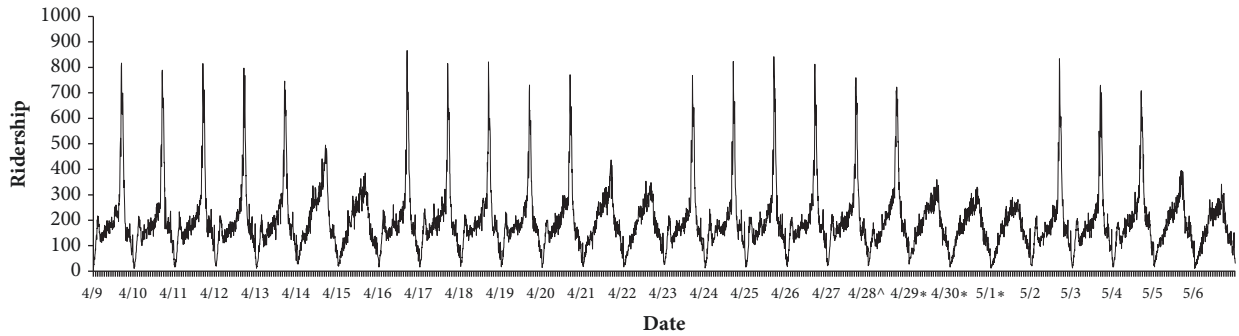

Figure 5: The origin entrance ridership time series at Zhujianglu Station of Nanjing Metro from April 9 to May 6, 2012.

model. The SARIMA model is formulated with statistical software SAS. The model forms generated from the three training data sets are all SARIMA(1,0,1)(0,1,1)$_{1428}$. For example, the specific equation is shown as the following, which constructs by the second training data set at Zhujianglu station:

$$\begin{aligned} y_t = {} & y_{t-1428} + 0.93904_1 \left(y_{t-1} - y_{t-1429}\right) + \varepsilon_t \\ & - 0.52176\varepsilon_{t-1} - 0.49147\varepsilon_{t-1428} + 0.25643\varepsilon_{t-1429}, \end{aligned} \tag{10}$$

where y_t is the real value in time t, ε_t is error between the real value y_t, and the predicted value $\widehat{y}_t$ in time t.

Therefore, for the prediction at time t, the real values for time $t-1429$, $t-1428$, $t-1$ serve as inputs. Afterwards, the ε-SVM model and the Gaussian radial basis function are implemented using the LibSVM software package developed by Wu et al. [40]. The Python codes were developed to integrate the LibSVM package with the PSO algorithm for parameters optimizing. The fivefold cross-validation technique and PSO are applied to obtain the optimal parameters (shown in Table 1) with the training data to construct the final ε-SVM model for future forecasting. The testing data are then used as input to the final ε-SVM model to produce predicted outputs.

The SVMOOL model updates training data set day by day. For example, using the second training data set from March 12 to April 1, predictions of ridership for every time interval on April 2 are made, then actual observed values of April 2 are added to the initial training data set to produce an updated training data set. Then the updated training data set from March 12 to April 2 is used to forecast the ridership of every interval on April 3 and the process repeats.

For the SVMPOL model, the testing data is updated by time interval (i.e., 5-min) for the day being predicted, and the number of input features extracted via continuity, or m value is determined to be 4 through several trails. In each of the 7 testing days, the first 10 data points (from 6:05 am to 6:50) were used as the testing data, with the 11th data point (at 6:55 am) being the target. Then 10-point window "walks", incorporating the 11th data point, which results on a new 10-point window (from 6:10 am to 6:55), having then the 12th data point (at 7:00 am) as the target. The process continues until the last observation (at 23:00 pm) becomes the target.

For the combined model, after the values from SVMOOL and SVMPOL models are calculated, the final prediction value is the average prediction of the previous two models.

5. Results Analysis

After the SVMOOL, SVMPOL, and SVMCOL models are implemented with the data sets, this research selects SARIMA, SVM, and BPNN models (i.e., back-propagation neural network) as the benchmark for one-step prediction are shown in Tables 2, 3, and 4.

5.1. Weekday Ridership Forecasting Results. In addition, the pattern of weekday's ridership is similar, so Table 1 shows the forecasting ridership results of three weekdays. As shown in Table 2, the SARIMA model is the best among them in terms of forecasting accuracy for weekday's ridership. It is demonstrated that the SARIMA model is good at predicting the ridership with periodic and stability characteristics as shown in Figure 6. The SVMOOL model is superior to the two models (BPNN and SVM models) because the updating data set, but the performance of the SVMCOL model, is not very satisfactory and is affected by the SVMPOL model, which is not applicable to the weekday ridership forecasting independently.

5.2. Weekend Ridership Forecasting Results. Table 3 shows that the SVMCOL model is the best among them in terms of forecasting accuracy, which the performance improves 40% compared with SARIMA model and improves 10% compared with the BPNN and SVM models for the value RMSE and MAE. It confirms that the combined model captures the weekly periodic and nonlinear characteristics of time series data for the estimation of short-term ridership (as shown in Figure 7). Though the SVMPOL model is not getting good results, it is much better than the SARIMA model. The SVM and BPNN models are both better than the SARIMA model for the weekend ridership forecasting, which also demonstrate that SVM and BPNN are suitable for nonlinear and fluctuant passenger flow.

5.3. Holiday Ridership Forecasting Results. As shown in Table 4 and Figure 8, it is not difficult to find that the SARIMA model is the worst performance for the holiday ridership and cannot meet the accuracy of short-time prediction. This is because the ridership data in samples 2 and 3 contains the unstable holiday passenger flow of the Ching-Ming Festival and May Day and demonstrates that the

Table 1: The optimal parameter sets of the second training data set at Zhujianglu station.

Training Data Set	C	γ	ε
Data From March 12 to April 1	597.69	0.42	0.57
Data From March 12 to April 2	870.02	0.77	0.76
Data From March 12 to April 3	996.26	0.32	0.57
Data From March 12 to April 4	1001.00	0.28	0.76
Data From March 12 to April 5	577.02	0.46	0.48
Data From March 12 to April 6	797.80	0.47	0.90
Data From March 12 to April 7	1001.00	0.07	1.00

Table 2: Weekday and weekend performance comparison for one-step prediction.

Model	RMSE (ridership/5-min)	MAE (ridership/5-min)	MAPE (%)
November 27, Tuesday, 2012 at Sanshanjie Station			
SARIMA	*23.52*	*18.38*	*11.00*
BPNN	34.84	26.82	15.99
SVM	34.85	26.82	15.93
SVMOOL	34.75	26.73	15.89
SVMPOL	45.29	32.82	20.48
SVMCOL	36.50	26.31	16.52
April 5, Thursday, 2012 at Zhujianglu Station			
SARIMA	*25.38*	*19.05*	*11.79*
BPNN	33.84	23.53	13.89
SVM	33.78	23.58	13.93
SVMOOL	33.68	23.41	13.81
SVMPOL	57.75	32.86	17.90
SVMCOL	41.31	25.76	14.58
May 3, Thursday, 2012 at Zhujianglu Station			
SARIMA	*25.72*	*18.46*	*11.70*
BPNN	39.05	26.99	15.09
SVM	39.18	27.10	15.13
SVMOOL	38.88	26.80	14.62
SVMPOL	58.45	32.02	15.92
SVMCOL	40.06	27.41	15.26

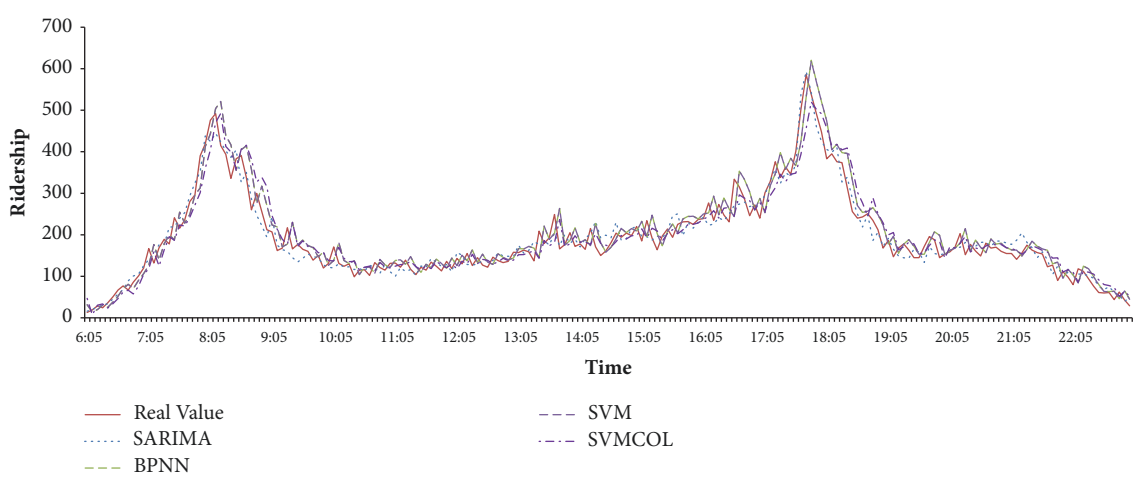

Figure 6: The comparison between the real value and the predicted values by 4 different models using origin ridership time series at Sanshanjie Station on November 27, 2012.

Table 3: weekend performance comparison for one-step prediction.

Model	RMSE (ridership/5-min)	MAE (ridership/5-min)	MAPE (%)
September 1, Saturday, 2012 at Sanshanjie Station			
SARIMA	44.17	33.50	17.07
BPNN	30.99	24.35	12.60
SVM	30.81	24.20	12.53
SVMOOL	29.68	23.32	*12.03*
SVMPOL	32.33	26.20	15.20
SVMCOL	*27.11*	*22.52*	12.23
April 7, Saturday, 2012 at Zhujianglu Station			
SARIMA	46.98	34.65	17.91
BPNN	30.67	23.41	16.84
SVM	30.66	23.43	16.85
SVMOOL	30.25	23.04	*16.51*
SVMPOL	32.04	25.60	21.75
SVMCOL	*28.51*	*21.61*	17.56
May 5, Saturday, 2012 at Zhujianglu Station			
SARIMA	52.09	33.29	18.69
BPNN	30.02	22.78	15.48
SVM	31.48	23.29	15.63
SVMOOL	29.71	22.53	*15.32*
SVMPOL	31.88	23.83	18.87
SVMCOL	*27.89*	*21.25*	15.64

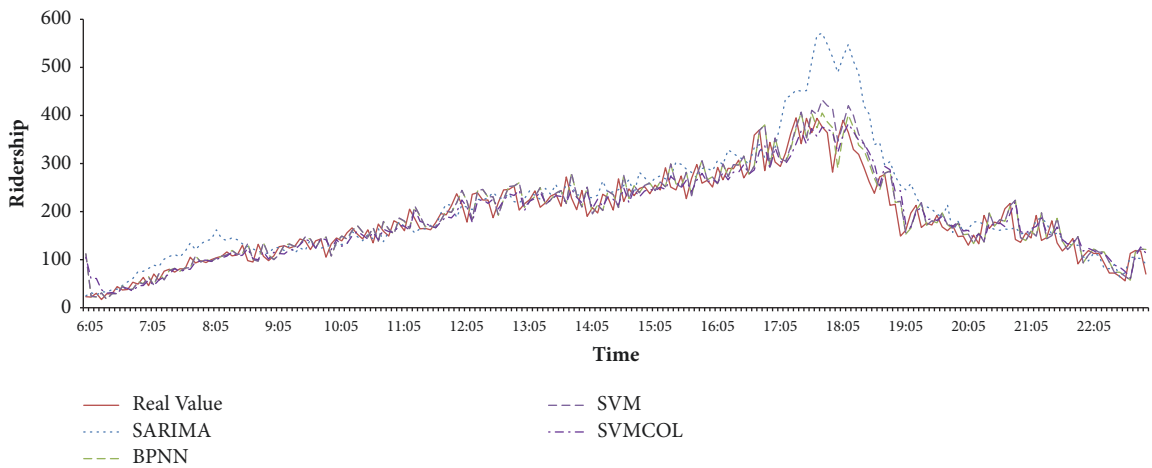

Figure 7: The comparison between the real value and the predicted values by 4 different models using origin ridership time series at Zhujianglu Station on May 5, 2012.

SARIMA model does not apply to nonstationary ridership. The SVMOOL and SVMCOL two models both outperform the three models (such as SARIMA, BPNN, and SVM models) and the SVMCOL model is the best among them in terms of forecasting accuracy, because the two models are constructed on the updated data set and more responsive to the change of passenger flow. The results of the SVMPOL model outperform the SARIMA and SVM models in the case of a small sample (with only 10 samples) as shown in Table 4. It is demonstrated that the SVMPOL model can capture the change of passenger flow in real-time and has special advantages for small sample prediction. In a word, compared with the offline models, the online models achieve better prediction performance. Of course, the prediction performance of BPNN is slightly better than the SVMPOL model, which is due to the small sample information. The results demonstrate the effectiveness of the proposed model. It is noted that the predicted performance of the SVMOOL

Table 4: Ching-Ming Festival and May Day performance comparison for one-step prediction.

Model	RMSE	MAE	MAPE (%)
April 2, Monday, 2012 at Zhujianglu Station			
SARIMA	103.92	62.34	34.38
BPNN	29.01	22.98	14.39
SVM	30.48	23.98	14.68
SVMOOL	30.48	23.98	14.68
SVMPOL	31.37	25.06	14.89
SVMCOL	*27.87*	*22.01*	*14.09*
April 3, Tuesday, 2012 at Zhujianglu Station			
SARIMA	102.98	54.73	32.40
BPNN	27.61	21.32	15.32
SVM	29.53	22.55	15.78
SVMOOL	27.42	21.21	15.26
SVMPOL	27.01	22.01	16.25
SVMCOL	*24.84*	*19.48*	*14.18*
April 4, Saturday, 2012 at Zhujianglu Station			
SARIMA	121.96	69.05	51.69
BPNN	24.06	18.26	15.42
SVM	28.67	20.36	16.38
SVMOOL	23.88	18.08	*15.35*
SVMPOL	23.74	18.70	18.07
SVMCOL	*21.90*	*16.95*	15.64
April 30, Monday, 2012 at Zhujianglu Station			
SARIMA	106.88	60.02	38.55
BPNN	28.68	22.17	15.42
SVM	32.25	24.51	16.22
SVMOOL	32.25	24.51	16.22
SVMPOL	28.46	22.58	*14.02*
SVMCOL	*26.92*	*21.22*	14.73
May 1, Tuesday, 2012 at Zhujianglu Station			
SARIMA	117.53	65.01	53.22
BPNN	24.98	19.16	16.66
SVM	32.13	22.54	18.16
SVMOOL	24.28	18.80	16.34
SVMPOL	26.08	20.48	19.25
SVMCOL	*21.16*	*16.24*	*15.59*

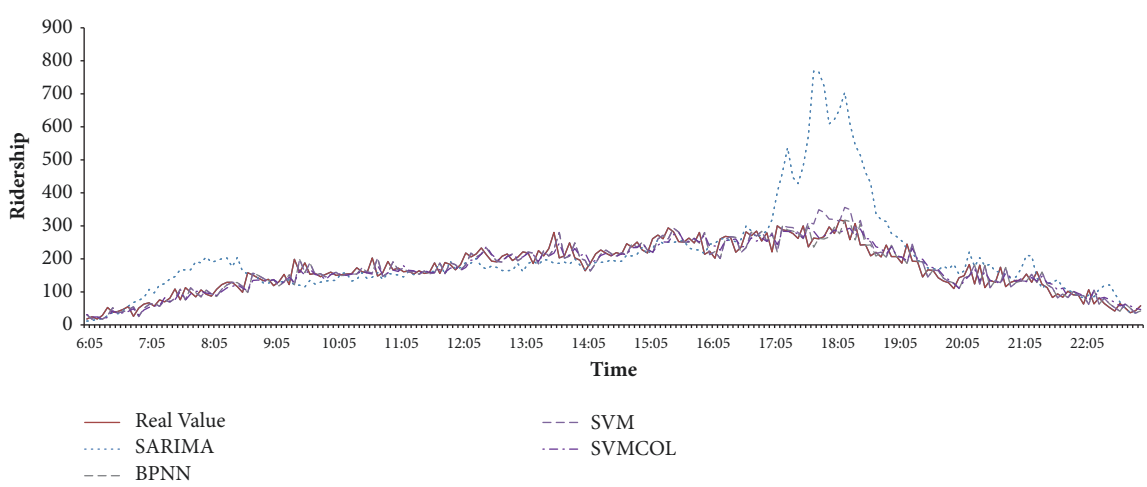

Figure 8: The comparison between the real value and the predicted values by 4 different models using origin ridership time series at Zhujianglu Station on April 3, 2012.

model on April 2 and 30, 2012, at Zhujianglu station is equal to the SVM model because the two models both own the same training sample.

6. Discussion

The training time and the forecasting time are the keys to real-time implementation. The experiments using LibSVM package on desktop computers indicate that the training time needs about one hour for three weeks' data (4284 observations) to construct the prediction function and the forecasting time needs less than 1 second for a one-step prediction using SVM. According to the SVMOOL model updating testing data set by day, the SVMOOL model uses the training data sample size from 21 days' observations to 22 days' or 27 days' observations, but the training time only increases 10 min and the forecasting time needs less than 1 s. Because the SVMOOL model is retrained once a day, the obtained forecasting function can be used for one-step predictions for the day, therefore real-time implementation is possible. The SVMPOL model is retained in real-time in 5-min interval. The obtained forecasting function can be used to one-step prediction for the next 5-min. In the process of the implementation experiments, the SVMPOL model needs less than 1 s to construct due to the small data sample and the forecasting time needs less than 1 s for one-step prediction. Therefore, the training time and the forecasting time can meet the real-time demand for the one-step prediction in the implementation as well.

7. Conclusions

The key to metro operation and management is based on the changes of the ridership to effectively deploy and use the system resources and to timely adjust operation strategy to ensure that metro is safe to complete the transportation service task. The results of short-term ridership forecasting can provide useful information to decision makers of metro system, and the prediction accuracy directly influence the legitimacy and effectiveness of any changes in operations, such as adjustments to headway, train dispatching, and the activation of station passenger crowd regulation plan or emergency response plan.

This paper proposes a novel hybrid model combining the SVMOOL model and the SVMPOL model for short-term ridership forecasting that better captures the periodicity and nonlinearity characteristics by the updated data set. The SVMCOL model takes advantages of the individual strengths of the two models. While the SARIMA model is superior for the stable weekday ridership to other models, experiments results indicate that the SVMOOL model is superior to SARIMA model, BPNN model, or SVM model in terms of MAE and RMSE for the weekend and holiday ridership test. The actual results of 5-min short-term ridership forecasting show the feasibility and effectiveness of the proposed combined model in real-time implementation.

It should be noted that the prediction of ridership under abnormal situations (such as holiday) is evidently more challenging than doing so under normal conditions (such as weekday ridership), and hence, much desired by the operator. Therefore, the proposed SVMCOL model is found to be suitable and useful in real-world operations, particularly in prediction under abnormal conditions. And, further studies need apply the proposed model to other abnormal situations (such as horrible weather, large sports events or emergencies, this study chooses the weekday, weekend, and holiday ridership as the demonstration). In addition, different characteristics (the impact of different meteorological conditions, the number of metro station entrances, etc.) can be considered as the input features in further studies. Jia et al. [49] indicate that, with the consideration of additional rainfall factor, the traffic flow prediction accuracy is improved.

Data Availability

Detailed data are included within the supplementary materials.

Disclosure

An earlier version of this paper has been presented in the Transportation Research Board 92nd Annual Meeting (Washington DC, 2013).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research has been supported by the Fundamental Research Funds for the Central Universities (no. KYLX16_0270). The authors thank the Nanjing Metro for providing the data used in this research. Thanks are due to Chih-Chung Chang and Chih-Jen Lin for permission to use the LibSVM package in this research.

Supplementary Materials

APPENDIX Table 1: the origin 5-min entrance ridership data at Sanshanjie Station of Nanjing Metro from November 5 to December 2, 2012. APPENDIX Table 2: the origin 5-min entrance ridership data at Zhujianglu Station of Nanjing Metro from March 12 to April 8, 2012. APPENDIX Table 3: the origin 5-min entrance ridership data at Zhujianglu Station of Nanjing Metro from April 9 to May 6, 2012. *(Supplementary Materials)*

References

[1] T.-H. Tsai, C.-K. Lee, and C.-H. Wei, "Neural network based temporal feature models for short-term railway passenger demand forecasting," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3728–3736, 2009.

[2] H. Bar-Gera and D. Boyce, "Origin-based algorithms for combined travel forecasting models," *Transportation Research Part B: Methodological*, vol. 37, no. 5, pp. 405–422, 2003.

[3] G. Jovicic and C. O. Hansen, "A passenger travel demand model for Copenhagen," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 4, pp. 333–349, 2003.

[4] M.-C. Chen and Y. Wei, "Exploring time variants for short-term passenger flow," *Journal of Transport Geography*, vol. 19, no. 4, pp. 488–498, 2011.

[5] L. Chang-Jen and S.-P. Miaou, "Real-time prediction of traffic flows using dynamic generalized linear models," *Transportation Research Record*, no. 1678, pp. 168–178, 1999.

[6] S. Sundaram, H. N. Koutsopoulos, M. Ben-Akiva, C. Antoniou, and R. Balakrishna, "Simulation-based dynamic traffic assignment for short-term planning applications," *Simulation Modelling Practice and Theory*, vol. 19, no. 1, pp. 450–462, 2011.

[7] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transportation Research Record*, no. 1644, pp. 132–141, 1998.

[8] B. M. Williams, "Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling," *Transportation Research Record*, no. 1776, pp. 194–200, 2001.

[9] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[10] Z. R. Ye, Y. L. Zhang, and D. R. Middleton, "Unscented Kalman filter method for speed estimation using single loop detector data," *Transportation Research Record*, no. 1968, pp. 117–125, 2006.

[11] J. Xia, M. Chen, and Z. Qian, "Predicting freeway travel time under incident conditions," *Transportation Research Record*, no. 2178, pp. 58–66, 2010.

[12] Z. Ye and Y. Zhang, "Speed estimation from single loop data using an unscented particle filter," *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 7, pp. 494–503, 2010.

[13] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, 2002.

[14] S. Clark, "Traffic prediction using multivariate nonparametric regression," *Journal of Transportation Engineering*, vol. 129, no. 2, pp. 161–168, 2003.

[15] T. Tsung-Hsien, L. Chi-Kang, and W. Chien-Hung, "Design of dynamic neural networks to forecast short-term railway passenger demand," *Journal of the Eastern Asia Society for Transportation Studies*, no. 6, pp. 1651–1666, 2005.

[16] Y. Mo and Y. Su, "Neural networks based real-time transit passenger volume prediction," in *Proceedings of the 2009 2nd Conference on Power Electronics and Intelligent Transportation System, PEITS 2009*, pp. 303–306, December 2009.

[17] Y. Huang and H. Pan, "Short-term prediction of railway passenger flow based on RBF neural network," in *Proceedings of the 4th International Joint Conference on Computational Sciences and Optimization, CSO 2011*, pp. 594–597, April 2011.

[18] Y. Zhang and Y. Xie, "Forecasting of short-term freeway volume with v-support vector machines," *Transportation Research Record*, vol. 2024, no. 1, pp. 92–99, 2007.

[19] Y. Zhang and Y. Xie, "Travel mode choice modeling with support vector machines," *Transportation Research Record*, no. 2076, pp. 141–150, 2008.

[20] Q. Chen, W. Li, and J. Zhao, "The use of LS-SVM for short-term passenger flow prediction," *Transport*, vol. 26, no. 1, pp. 5–10, 2011.

[21] W.-H. Lin, "A Gaussian maximum likelihood formulation for short-term forecasting of traffic flow," in *Proceedings of the 2001 IEEE Intelligent Transportation Systems*, pp. 150–155, Oakland, Calif, USA, 2001.

[22] Y. Xie, Y. Zhang, and Z. Ye, "Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 326–334, 2007.

[23] D. Boto-Giralda, F. J. Díaz-Pernas, D. González-Ortega et al., "Wavelet-based denoising for traffic volume time series forecasting with self-organizing neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 7, pp. 530–545, 2010.

[24] C. P. I. van Hinsbergen, J. W. C. van Lint, and H. J. van Zuylen, "Bayesian committee of neural networks to predict travel times with confidence intervals," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 5, pp. 498–509, 2009.

[25] X. Fei, C. C. Lu, and K. Liu, "A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1306–1318, 2011.

[26] Y. Wei and M. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.

[27] H. Xiao, H. Sun, and B. Ran, "Special factor adjustment model using fuzzy-neural network in traffic prediction," *Transportation Research Record*, no. 1879, pp. 17–23, 2004.

[28] Y. Zhang and Z. Ye, "Short-term traffic flow forecasting using fuzzy logic system methods," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 12, no. 3, pp. 102–112, 2008.

[29] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.

[30] N. Zhang, Y. Zhang, and H. Lu, "Seasonal autoregressive integrated moving average and support vector machine models: prediction of short-term traffic flow on freeways," *Transportation Research Record*, vol. 2215, pp. 85–92, 2011.

[31] W.-C. Hong, Y. Dong, F. Zheng, and S. Y. Wei, "Hybrid evolutionary algorithms in a SVR traffic flow forecasting model," *Applied Mathematics and Computation*, vol. 217, no. 15, pp. 6733–6747, 2011.

[32] W.-C. Hong, "Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm," *Neurocomputing*, vol. 74, no. 12-13, pp. 2096–2107, 2011.

[33] W.-C. Hong, "Application of seasonal SVR with chaotic immune algorithm in traffic flow forecasting," *Neural Computing and Applications*, vol. 21, no. 3, pp. 583–593, 2012.

[34] J. Wang and Q. Shi, "Short-term traffic speed forecasting hybrid model based on chaos–wavelet analysis-support vector machine theory," *Transportation Research Part C: Emerging Technologies*, vol. 27, no. 1, pp. 219–232, 2013.

[35] R. Chen, C.-Y. Liang, W.-C. Hong, and D.-X. Gu, "Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm," *Applied Soft Computing*, vol. 26, pp. 435–443, 2015.

[36] S. Ishak and H. Al-Deek, "Performance evaluation of short-term time-series traffic prediction model," *Journal of Transportation Engineering*, vol. 128, no. 6, pp. 490–498, 2002.

[37] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 211–234, 2005.

[38] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[39] S. Abe, "Analysis of support vector machine," in *IEEE Signal Processing Society Workshop*, pp. 88-89, IEEE Press, New York, NY, USA, 2002.

[40] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, pp. 276–281, 2004.

[41] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, and Q. M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1-2, pp. 321–336, 2003.

[42] C. Huang and C. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Systems with Applications*, vol. 31, no. 2, pp. 231–240, 2006.

[43] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008.

[44] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2001.

[45] W.-C. Hong, Y. Dong, F. Zheng, and C.-Y. Lai, "Forecasting urban traffic flow by SVR with continuous ACO," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 35, no. 3, pp. 1282–1291, 2011.

[46] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks (ICNN' 95)*, vol. 4, pp. 1942–1948, Perth, Western Australia, November-December 1995.

[47] W. Zhu, W. Wang, and Z. Huang, "Estimating train choices of rail transit passengers with real timetable and automatic fare collection data," *Journal of Advanced Transportation*, vol. 2017, Article ID 5824051, 12 pages, 2017.

[48] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 103–118, 2012.

[49] Y. Jia, J. Wu, and M. Xu, "Traffic flow prediction with rainfall impact using a deep learning method," *Journal of Advanced Transportation*, vol. 2017, Article ID 6575947, 10 pages, 2017.