BUS ARRIVAL TIME PREDICTION BASED ON PCA-GA-SVM

Z. Peng[*†], *Y. Jiang*[†], *X. Yang*[†], *Z. Zhao*[†], *L. Zhang*[†], *Y. Wang*[†]

Abstract: Considering the correlations of the input indexes and the deficiency of calibrating kernel function parameters when support vector machine (SVM) is applied, a forecasting method based on principal component analysis-genetic algorithm-support vector machine (PCA-GA-SVM) is proposed to improve the precision of bus arrival time prediction. And the No. 232 bus in Shenyang City of China is taken as an example. The traditional SVM and Kalman Filtering model and GA-SVM are also employed to make comparative analysis on the prediction rate, respectively. The result indicates that PCA-GA-SVM obtains more accurate prediction results of bus arrival time prediction.

Key words: *bus arrival time prediction, Principal Component Analysis (PCA), Support Vector Machine (SVM), Kalman filter*

Received: April 23, 2016 **DOI:** 10.14311/NNW.2018.28.005
Revised and accepted: February 8, 2018

1. Introduction

Bus arrival time prediction has already become a key problem of realizing high service level of Intelligent Transport System (ITS) in modern cities. On one hand, precisely bus arrival time prediction in urban public transport system enables passengers to arrange their own itinerary more reasonable, improves the passengers' utility of public transport, and further promotes the development of mass transit; on the other hand, precise bus arrival time prediction makes the real-time information of dispatch of buses achievable, lowers the operating cost of the entire transit system and further improves the service level of ITS. However, the operation of buses is within enormous and complex urban transport system, therefore, bus arrival time is subject to a variety of factors including physical properties of the roads, running time, traffic structure, passengers' willing, weather, land use structure of the surrounding. The emphasis of accurate bus arrival time prediction lies on rational choices of influencing factors, collection of valid observation data and models to predict bus arrival time precisely.

*Zixuan Peng; Collaborative Innovation Center for Transport Studies, Dalian Maritime University, Dalian 116026, China, E-mail: Pengzx_dl@163.com

†Yonglei Jiang – Corresponding author; Zixuan Peng; Xiaoli Yang; Zhigang Zhao; Liu Zhang; Yitian Wang; Transportation Management College, Dalian Maritime University, Dalian 116026, China, E-mail: jiangyl_dl@163.com, Yangxl_dl@tom.com, Zhaozg_dl@tom.com, zhangliu_dl@163.com, wangyitian_dl@163.com

2. Literature review

Bus arrival time forecast is a crucial research topic in the study of urban transit dispatch, and there are abundant researches mainly concentrate on two aspects: models for bus arrival time prediction and suitable input indexes selection.

2.1 Models for bus arrival time prediction

Altinkaya & Zontul [2] divided existing research methods into four categories: Models based on the Historical Data [5, 25, 26, 33], Statistical Models [6, 13], Kalman Filtering Model [15, 23] and Machine Learning Models [8, 9, 20]. A discriminative method, Support Vector Machine (SVM), has been successfully applied to protein sequence classification and shown the superiority to the other methods [7,11,14,16,19]. Relevant researchers compared the forecast effect of these four measures, and found that the Machine Learning Models, represented by SVM, reveals the best performance of sample amounts requirement and prediction effects. Therefore, bus arrival time prediction based on SVM became a hotpot of study [17, 27]. The dynamic selection of both penalty parameter and kernel function parameter is difficult in traditional SVM, therefore "SVM+", as the major hybrid model, emerged for purpose of improve the precision of forecast based on SVM, have been the research trends of bus arrival time prediction [1,4,10,18,20,22,24,28,33,36,37].

2.2 Selection of prediction index

Apart from Models based on the Historical Data, all the other three types of methods need to select and collect input indexes while bus arrival time prediction is conducted. The selection of bus arrival time index in existing researches is showed in Appendix 1. Different studies imply that bus arrival time is affected by the time period, weather, road segment and vehicle operation status. The weight of the identical index varies in different researches. As we know, sufficient input indexes will significantly increase the forecasting accuracy; nevertheless, the correlations between indexes will influence the prediction performance.

Therefore, in order to forecast bus arrival time precisely without the deficiency in value of the penalty parameter and the kernel function parameter as well as the deficiency of the correlation of input indexes, this paper first proposes a PCA-GA-SVM model to predict the bus arrival time. Then, take No. 232 bus in Shenyang of China as an example, and the prediction performance of PCA-GA-SVM is compared to results of traditional SVM model and Kalman Filtering model.

3. Methodologies

3.1 Introduction of PCA-GA-SVM model

3.1.1 Modeling process

The present paper complies with the following modeling process to conduct bus arrival time prediction (Fig. 1). The input index include the time period, weather, road segment, the latest run time of the next link, the running time of the last bus.

First, apply PCA to lower collected data, dimension of characteristic index of SVM's training samples, thereby settle relevant errors of characteristic index of traditional SVM's training samples, and explain redundant problems.

Second, apply GA to determine the penalty coefficient C and RBF kernel function value γ in traditional SVM model, avoiding the influence on prediction accuracy by empirical values.

Finally, apply dynamic slippage to ascertain the training sample set. It is illustrated as follows, when forecasting the time that bus arrives at certain stop at moment $t+1$, dynamically select the time and other characteristic index of preceding n shifts of buses as the training sample to predict the time that bus arrives at certain stop at moment $t+2$. When forecasting moment $t+2$, slippage moves forward a moment in order to retrain sample and continue to predict.

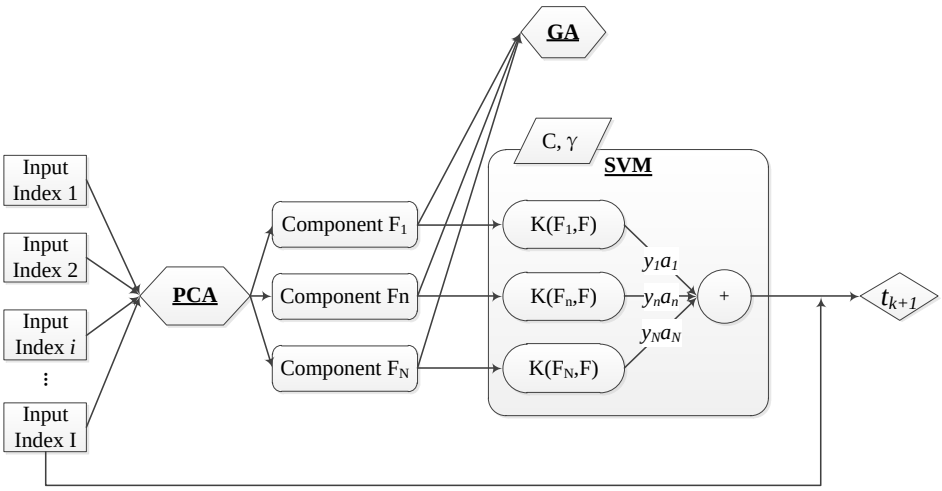

Fig. 1 *Flowchart of proposed PCA-GA-SVM model.*

3.1.2 Regression of SVM

Given a set of samples $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$, among which the relation is unknown. SVM is able to utilize a nonlinear mapping Φ, mapping data x to a high dimensional feature space H and conduct linear approximation in this space. Mapping function is found to nicely approximate the given data sample. According to statistical theory, this function can be expressed as follows

$$f(x) = \omega \bullet \varphi(x) + b. \tag{1}$$

Define this regression estimation problem as a problem of a loss function to minimize branching, the final regression function is minimizing and regularizing risk functional under certain constraints

$$\frac{1}{2}\|\omega\|^2 + C\frac{1}{l}\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i)). \tag{2}$$

Therein, the first item is regularization item, contributing to improve the ability of function generalization; the second item is empirical risk functional, which can be determined by different loss function; constant $C > 0$ controls the degree of penalty in samples which exceed error ε. The distribution is demonstrated as follows when insensitive loss function is applied

$$L_\varepsilon(y_i, f(x_i)) = \max(|y_i - f(x_i)| - \varepsilon, 0). \tag{3}$$

With regard to $L_\varepsilon(y_i, f(x_i))$, if absolute value of deviation between estimated output $f(x_i)$ and desired output y_i is less than ε, it equals to 0; otherwise it equals the absolute value of deviation minus ε. By introducing non-negative slack variables ξ_i, ξ_i^*, the minimizing formula (2) can be converted into

$$\min \frac{1}{2}\|\omega\|^2 + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*), \tag{4}$$

$$\text{s.t. } y_i - \omega \bullet \varphi(x_i) - b \le \varepsilon + \xi_i, \tag{5}$$

$$\omega \bullet \varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*, i = 1, \ldots, l, \tag{6}$$

$$\xi_i^* \ge 0. \tag{7}$$

The minimum of formula (4) is a convex quadratic optimization problem. The Lagrange function is introduced as follows

$$\omega - \sum_{i=1}^{l}(a_i - a_i^*)x_i = 0. \tag{8}$$

Then

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)\varphi(x_i) \bullet \varphi(x) + b. \tag{9}$$

Plug kernel function $K(x_i, x_j)$ into formula (9) we can get the following formula

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x_j) + b. \tag{10}$$

Kernel function $K(x_i, x_j)$ is the core of SVM, different kernel functions can form different SVM. The present paper mainly uses Radical Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2)$, for its sound performance, where γ is the parameter.

3.1.3 Principal component analysis

Train the sample set $(x_i, y_j), (i = 1, \ldots, n, x \in R^d)$ which is the base of SVM model, therein $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ as an indicator vector, where every index may have certain correlations with each other. It may occur that the sample information overlapped too much for there are high correlations between sample indexes when choosing training samples. Therefore, PCA is serviceable to summarize the primary aspect of the numerous data and reduce the dimension of sample index data through these aggregative indicators representing characters in certain field independently and respectively, thereby improve the validity of training samples.

For every training sample, the variable index which it affects is $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, where x_{ij} representing the j-th index value of the i-th sample. Assume that the corresponding index is a random variable $x_{\cdot j}$ to the j-th index, the corresponding sample average is $\overline{x_{\cdot j}}$, sample standard deviation is $S_{\cdot j}$.

First, standardize, $e_{\cdot j} = (x_{\cdot j} - \overline{x_{\cdot j}})/S_{\cdot j}$;

Second, calculate $e = (e_1, e_2, \ldots, e_n)^{\mathrm{T}}$ covariance matrix $\sum$;

Third, use $\sum u = \lambda u$ to calculate and obtain eigenvalue of matrix $\sum \lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and eigenvector matrix $u = [u_1, u_2, \ldots, u_n]^{\mathrm{T}}$, among which $u_j = [u_{j1}, u_{j2}, \ldots, u_{jn}]$;

Fourthly, acquire following principal component by preceding calculations:

$$\begin{cases} F_{\cdot 1} = u_{11} e_{\cdot 1} + u_{12} e_{\cdot 2} + \cdots + u_{1d} e_{\cdot d}, \\ F_{\cdot 2} = u_{21} e_{\cdot 1} + u_{22} e_{\cdot 2} + \cdots + u_{2d} e_{\cdot d}. \\ \cdots\cdots \\ F_{\cdot d} = u_{d1} e_{\cdot 1} + u_{d2} e_{\cdot 2} + \cdots + u_{dd} e_{\cdot d}. \end{cases}$$

Therein, $u_{j1}^2 + u_{j2}^2 + \cdots + u_{jd}^2 = 1$, where $u_j = (u_{j1}, u_{j2}, \ldots, u_{jd})$, then $u_1 u_j^{\mathrm{T}} = 0$. Assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, then name $F_{\cdot d}$ is the d-th principal component.

Fifthly, cumulative $= \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{d} \lambda_j$, where $k \leq d$. When the accumulated variance contribution rate of k principal components is reach 90%, they will be taken as new indexes.

Sixthly, new training samples can be expressed as (F_i, y_i) after lowering the dimension of principal components, where $F_i = (F_{i1}, F_{i2}, \ldots, F_{ik})$.

3.1.4 Genetic algorithm

RBF kernel function adopted by SVM model involves an unknown parameter γ; And the penalty coefficient C is also unknown. Therefore, it is of necessity to preset these two parameters. The present paper applies genetic algorithm to calibrate parameter C and γ to prevent errors from human trial.

This research uses CV_d (D-delete Crossed Validation). It randomly chooses d samples from n training sample sets as predicted sample sets, and select $n - d$ samples as training samples to form learning machine simultaneously, further utilize training sample sets to conduct training and formation of learning machine, eventually forecast the sample sets and observe the accuracy of prediction. For arbitrary training sample set, it is required that E is small enough in the following formula

$$E = \frac{1}{\binom{n}{d}} \sum_{i=1}^{v} \sum_{j=1}^{d} (\hat{y}_{ij} - y_{ij})^2. \tag{11}$$

Therein, y_{ij} represents the actual value of j-th trained sample in i-th combination of predicted sample set, while $\hat{y}_{ij}$ represents the predicted value of j-th trained sample in i-th combination of predicted sample set; d represents sample amounts of each extracted predicted sample set; $\binom{n}{d}$ represents the quantity of all the combination. There are some literatures which have successfully applied heuristic algorithms for parameters for SVM. These results indicate a useful means for our SVM [29, 30]. Furthermore, heuristic algorithms are also tested by lots of researchers as an effective method to solve this kind of complex problems [21, 31, 32, 34]. Thus, we attempt to use genetic algorithm for parameter optimization of SVM. Therefore, the prediction accuracy can reach maximum when the aforementioned fitness function takes the minimum value, through searching for the parameters C and γ with genetic algorithm.

GA optimizes the parameters of SVM as the following steps.

Algorithm 1 GA algorithm for parameter optimization.

Set the initial parameters of GA, such as population size and number of iterations, set num = 1.
Determine the encoding interval of C, ε, γ. Real number coding is chosen to generate the chromosomes.
Mean square error (MSE) is chosen as the fitness function.
repeat
  Roulette selection is used cooperating with elite strategy.
  Crossover and mutation operators are used to create a child population.
  Set num = num +1.
  if fitness agrees **then**
    Output the best individual and optimal solution.
  else
    Run the operators of selection, crossover and mutation.
  end if
until the stopping criterion is met.

3.2 Application of PCA-GA-SVM on forecasting bus arrival time

In aim of proceeding the reasonable prediction of bus arrival time, the present paper sets the parameters as follows.

First, the two variables, weather and time of day, are taken into account as a situation combination in this essay.

Second, the other 6 variables are set as follows: segments of road ($x^1_{k\to k+1}$, from k-th time spot to $k+1$-th time spot); number of lanes in segments ($x^2_{k\to k+1}, k+1 \to k$ in segment); number of intersegments in segment ($x^3_{k\to k+1}, k+1 \to k$ in segment); running time in pre-segment ($x^4_{k-1\to k}$, running time of current bus in $k-1 \to k$

segment); dwell time at last stop ($x^5_{k-1\to k}$, dwell time of current bus at k point in $k-1 \to k$ segment); running time in pre-segment ($x^6_{k\to k+1}$, running time of last schedule). x^2, x^3 represent the static characteristics of a specific operation, while x^1, x^4, x^5, x^6 mark the dynamic operation information of the vehicle, changing under different physical properties of road segments and traffic conditions.

Finally, after obtaining new input variables $((F^1, F^2, \ldots, F^n), n \leq 6)$ based on PCA, use the new n principal components to forecast the bus running time $y_{k\to k+1}$ in $k+1 \to k$ segment. Then, use $x^5_{k-1\to k}$ to output the arrival time t_{k+1} at $k+1$ time spot.

4. Empirical study

4.1 Data collection and analysis

Shenyang is the central city of Northeast in China. In order to testify the prediction effectiveness of the proposed model, this essay takes No. 232 bus in Shenyang, Liaoning province, China, as an object of research (Fig. 2). Its round is from Santaizi to Wanda Plaza. And No. 232 representative the most typical bus in Shenyang. Santaizi is suburbs and Wanda Plaza is a hub in the vicinity. What's more, it is

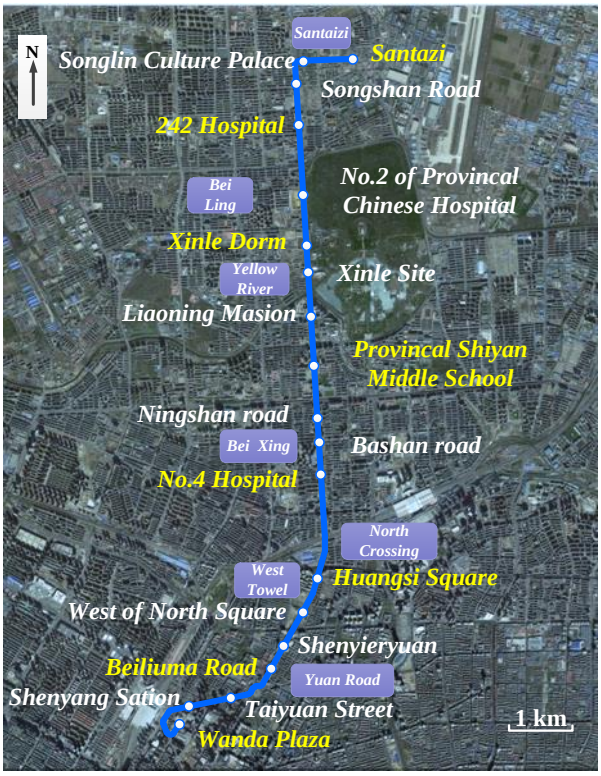

Fig. 2 *No. 232 bus route.*

punctuality to facilitate calculation. The origin of No. 232 bus is Santaizi while the terminal is Wanda Plaza with 11 kilometers long, passing 19 stops (18 segments). The operation time is from 5:20 to 23:00. It runs every 2 minutes in peak hours and every 3 to 5 minutes in normal hours. The present paper selects 8 primary stops (7 segments) as predicted objects among all (Santaizi, 242 Hospital, Xinle Dorm, Provincial Shiyan Middle School, No. 4 Hospital, Huangsi Square, Beiliuma Road and Wanda Plaza).

Considering weather condition (sunny or rainy) and time of day (peak hour (7:00 to 8:00) and normal hour (9:30 to 10:30)) are categorical data, the paper designs 4 groups (sunny-peak hour; sunny-normal hour; rainy-peak hour; rainy-normal hour) to collect data. The present paper surveyed on the vehicle from July 9^{th}, 2015 to July 22^{nd}, 2015 and collected 290 terms of data in 8 major stops (sunny-peak hour: 94 terms; sunny-normal hour: 94 terms; rainy-peak hour: 42 terms; rainy-normal hour: 60 terms). Therein, 240 terms are used as training samples (sunny-peak hour: 80 terms; sunny-normal hour: 80 terms; rainy-peak hour: 35 terms; rainy-normal hour: 35 terms), and the surplus 50 terms are taken as test samples. The statistical information of sample data is depicted in Tab. I and Tab. II.

Index							
Segment length (km)	3.0	2.33	3.3	...	1.9	1.08	1.58
Num. of Lanes	5.5	4.85	6	...	4.7	5.2	4
Num. of intersegments	5.6	7.8	8	...	3	4.47	6.69
Running time of pre-segment (second)	331	265	167	...	31	257	89
Dwell time of pre-segment (second)	46	109	128	...	65	14	79
Running time of last schedule (second)	375	50	265	...	158	172	99

Tab. I *Sample statistics data.*

Index	Max	Min	Avg.	S.E.
Segment length (km)	3.3	1.08	1.62	1.01
Num. of Lanes	6	4	5.44	1.45
Num. of intersegments	8	3	4.72	1.57
Running time of pre-segment (second)	331	31	141.71	83.30
Dwell time of pre-segment (second)	128	14	51.18	34.38
Running time of last schedule (second)	375	50	174.72	97.38

Tab. II *Description of statistical data of sample data.*

4.2 Assessment methods on prediction effectiveness

Prediction effectiveness is mainly measured by mean absolute error (MAE), mean absolute percentage error (MAPE) and root-mean-square error (RMSE), to evaluate prediction effectiveness of arrival time at certain stop.

$$\mathrm{MAE}_i = \frac{\sum \left| t_i^{\mathrm{running}} - \hat{t}_i^{\mathrm{running}} \right|}{N}, \tag{12}$$

$$\mathrm{MAPE}_i = \frac{1}{N} \sum \frac{\left| t_i^{\mathrm{running}} - \hat{t}_i^{\mathrm{running}} \right|}{t_i^{\mathrm{running}}} \times 100\%, \tag{13}$$

$$\mathrm{RMSE}_i = \sqrt{\frac{\sum (t_i^{\mathrm{running}} - \hat{t}_i^{\mathrm{running}})^2}{N-1}}. \tag{14}$$

Therein, t_i^{running} is the observed bus running time at stop i, $\hat{t}_i^{\mathrm{running}}$ is the predicted bus running time at stop i, and N is the quantity of prediction.

4.3 Prediction process of PCA-GA-SVM

4.3.1 Principal component analysis

Aimed at the 6 terms of original index in four groups of training samples, this essay applies PCA to lower the dimension and convert into new training samples.

First, the pair correlations among 6 indicators under four groups are demonstrated in Appendix 2. Obviously, information overlapping exists between indicators by different degrees.

Second, in view of four situation combination, SPSS software is applied to lower the dimensions based on PCA procedure. According to the standard that accumulated variance contribution rate $> 90\,\%$, transform the 6 indicators into 4 principal components by dimension reduction (Appendix 3).

4.3.2 Calibration of C and γ based on genetic algorithm

This essay applies genetic algorithm in order to ascertain the sound value of unknown parameter C in the penalty function and RBF kernel function γ. The present paper primarily adopts real coding, select entities proportionately, cross the single points and the basic bit mutate to proceed setting key steps of genetic algorithm. With regard to fitness of entities, use every C and γ substitute into RBF kernel function from initial groups and then train the previous sample sets by SVM. After the predicted sample values are obtained, substitute them in to formula (6) to get E. This essay sets that d equals 10 in genetic algorithm. For smaller E means bigger fitness value, take E as the standard to filter entities. With accordance to 4 groups of situations, the optimal value C and value γ are acquired through genetic algorithm, illustrated in Tab. III.

4.3.3 Prediction based on SVM

The 3 new indictors are established through PCA. The prediction operates again after training the samples of C and γ, optimized by genetic algorithm.

Meanwhile for the sake of comparison, the present paper uses traditional SVM model to compare, therein the values of C and γ are obtained through trial methods, equaling to 0.2 and 1.54, respectively.

Parameter	Group 1		Group 2		Group 3		Group 4	
	PCA-GA-SVM	GA-SVM	PCA-GA-SVM	GA-SVM	PCA-GA-SVM	GA-SVM	PCA-GA-SVM	GA-SVM
C	0.2532	0.2615	0.2561	0.2387	0.6235	0.6421	0.6451	0.6251
γ	1.5262	1.5036	1.5640	1.6841	1.6284	1.6355	1.6173	1.5972

Tab. III *The Results of Calibration on C and γ in PCA-GA-SVM & GA-SVM.*

4.4 Forecast by variable-parameter state-space model

4.4.1 Kalman filtering model

In order to compare the models, this essay paper applies variable parameter state-space model of Kalman filter to predict bus arrival time

$$y_t = x_t\beta_t + z_t\gamma + u_t, t = 1, 2, \ldots, T, \tag{15}$$

where y_t is the dependent variable, x_t is the explanatory variable vector of $1 \times m$, u_t is the random disturbance term, z_t is the explanatory variable matrix with fixed parameter, γ is the fixed parameter; β_t is unknown parameter vector of $m \times 1$, which is to-be-estimated and time-variant, reflecting the change of relations of dependent variables influenced by explanatory variables. Assume that variable parameter β_t is described by AR (1)

$$\beta_t = \psi\beta_{t-1} + \varepsilon_t \tag{16}$$

It is also feasible to extend into AR (p) model, and suppose that:

$$(u_t, \varepsilon_t)' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\0 & Q\end{pmatrix}\right), t = 1, 2, \ldots, T. \tag{17}$$

u_t and ε_t are independent from each other, and they comply with a normal distribution that the mean, variance and covariance equal to 0, σ^2 and Q respectively.

4.4.2 Variable-parameter state-space model

This paper applies variable-parameter state-space model to conduct bus arrival time forecast simultaneously, with regard Segment length, Num. of Lanes, Num. of intersegments, Weather, Time of day as fixed impact indicators, Running time of pre-segment, Dwell time of pre-segment, Running time of last schedule as the variable parameters impact indicators, to calibrate and forecast the model parameters accordingly. While Num. of Lanes, Weather, Time of day are not significant, the final results of parameter calibration are exhibited in Tab. IV after eliminating them.

	Coefficient	Std. Error	z-Statistic	Prob.
C (Segment length)	44.36805	6.303209	7.038962	0.0000
C (Num. of intersegments)	1.951060	0.362121	5.387867	0.0000
	Final State	Root MSE	z-Statistic	Prob.
SH (Running time of pre-segment)	0.680401	0.044156	15.40891	0.0000
SH (Dwell time of pre-segment)	−0.178132	0.053010	−3.360347	0.0008
SH (Running time of last schedule)	0.588617	0.091026	6.466476	0.0000
Log likelihood	−88.08036	Akaike info criterion		8.189124
Parameters	2	Schwarz criterion		8.288309
Diffuse priors	3	Hannan-Quinn criter.		

Tab. IV *Results of calibration on parameters in variable-parameter state-space model.*

4.5 Comparison of the prediction results

Based on the three evaluation quota values, MAE, MAPE and RMSE, it is apparent to perceive that:

First, under 4 groups of situations, the effectiveness of prediction is PCA-GA-SVM > GA-SVM > SVM > VPSSM. Under 4 groups of situations, the average prediction error rate is about 10 % of PCA-GA-SVM. The one of GA-SVM model is around 12 %. The one of traditional SVM model is around 15 %, while VPSSM model takes the worst prediction effectiveness, reaching nearly 22 %. However, the time of prediction is also PCA-GA-SVM > GA-SVM > SVM > VPSSM. The difference of the computation time is within 5.04 s. GA increases the calculation time of the PCA-GA-SVM. But it does not have much influence in performance of the PCA-GA-SVM.

Second, weather conditions influence the accuracy of four categories of models evidently. The outcome of evaluation indicators unveiled that the prediction precision of sunny days in three models is superior to that of rainy days.

Finally, the effects of the peak hour or the normal hour are of less influence relatively on prediction accuracy. In consider the time of day, the evaluation indicators result show that the forecast precision during normal hour is slightly higher than during peak hour based on these three models. However, the average discrepancy is within 5 % only.

It is discerned that, compared with current prediction methods, the proposed PCA-GA-SVM model is capable of predict bus arrival time more accurately, for this model solves the correlation of information of input index and modifies the precision of traditional SVM kernel function parameter.

5. Conclusions

Considering that the traditional SVM model is flawed on index correlation and unknown parameter calibration of kernel function for bus arrival time prediction, the present paper designs a bus arrival time prediction measure based on PCA, GA and SVM. The proposed PCA-GA-SVM model can reduce dimension of character-

istic indexes and shorten the time of training. Moreover, the PCA-GA-SVM model optimizes the penalty coefficient and kernel function value to improve the classification accuracy. Compared with the traditional SVM, the PCA-GA-SVM is more accurate. Taking No. 232 bus in Shenyang, Liaoning province, China, as an object of research, analyzes and compares the prediction effectiveness between PCA-GA-SVM model, GA-SVM model, traditional SVM model and VPSSM model. By means of comparing three types of prediction effectiveness indicators, it is obvious that though prediction accuracy varies under different weather and time, PCA-GA-SVM model always performs quite well.

Acknowledgement

This research was supported in National Natural Science Foundation of China 71571026, Higher Education Development Fund (for Collaborative Innovation Center) of Liaoning Province, China (Grant No. 20110116401, 20110116101), Liaoning Excellent Talents in University LR2015008, and the Fundamental Research Funds for the Central Universities (YWF-16-BJ-J-40 and DUT16YQ104).

References

[1] ALBA E., GARCIA-NIETO J., JOURDAN L., TALBI E.G. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In: *IEEE Congress on Evolutionary Computation (CEC 07)*, Singapore. New York: IEEE, 2007, pp. 284–290.

[2] ALTINKAYA M., ZONTUL A.M. Urban bus arrival time prediction: a review of computational models. *International Journal of Recent Technology & Engineering*. 2013, 2(4), pp. 164–169.

[3] CHEN M., LIU X., XIA J.A. Dynamic bus-arrival time prediction Model based on APC data. *Computer-Aided Civil and Infrastructure Engineering*. 2004, 19(5), pp. 364–376, doi: 10.1111/j.1467-8667.2004.00363.x.

[4] CHIEN S., DING Y., WEI C. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. *Journal of Transportation Engineering*. 2002, 128(5), pp. 429–438, doi: 10.1061/(ASCE)0733-947X(2002)128:5(429).

[5] CHUNG E.H., SHALABY A. Expected time of arrival model for school bus transit using real-time global positioning system based automatic vehicle location data. *Journal of Intelligent Transportation Systems*. 2007, 11(4), pp. 157–167, doi: 10.1080/15472450701649398.

[6] DELURGIO S.A. Forecasting principles and applications. Bonston: Irwin/Mcgraw-Hill, 1998.

[7] DING C., DUBCHAK I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 2001, 17(4), pp. 349–358, doi: 10.1093/bioinformatics/17.4.349.

[8] HAGAN M.T., DEMUTH H.B., BEALE M. Neural Network Design. Boston: Pws Pub., 1996.

[9] HUANG S.H., RAN B. *An Application of Neural Network on Traffic Speed Prediction under Adverse Weather Condition*. Wisconsin, 2003. PhD thesis, University of Wisconsin-Madison. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?, doi: 10.1.1.453.8901.

[10] HUERTA E.B., DUVAL B., HAO J.K. A hybrid GA/SVM approach for gene selection and classification of microarray data. In: *Workshops on Applications of Evolutionary Computation*, Budapest, Hungary. Berlin: Springer, 2006, pp. 34–44.

[11] JAAKKOLA T., DIEKHANS M., HAUSSLER D. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*. 2000, 7(1-2), pp. 95–114, doi: 10.1089/10665270050081405.

[12] JEONG R., RILETT R. Bus arrival time prediction using artificial neural network model. In: 7th IEEE International Conference on Intelligent Transportation Systems, Washington, DC. New York: IEEE, 2004, pp. 988–993, doi: 10.1109/ITSC.2004.1399041.

[13] KALMAN R.E. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*. 1960, 82(1), pp. 35–45, doi: 10.1115/1.3662552.

[14] LESLIE C., ESKIN E., COHEN A., WESTON J., NOBLE W.S. Mismatch string kernels for discriminative protein classification. *Bioinformatics*. 2004, 20(4), pp. 467–476, doi: 10.1093/bioinformatics/btg431.

[15] LESNIAK A., DANEK T., WOJDYLA M. Application of kalman filter to noise reduction in multichannel data. *Schedae Informaticae*. 2009, 1718(1), pp. 63–73, doi: 10.2478/v10149-010-0004-3.

[16] LIAO, L., NOBLE, W. S. Combining pairwise sequence similarity and support vector machines for detecting remoteprotein evolutionary and structural relationships. *Journal of Computational Biology*. 2003, 10 (6), pp.857-868, doi: 10.1089/106652703322756113.

[17] LIN Y., YANG X., ZOU N., JIA L. Real-Time Bus Arrival Time Prediction: Case Study for Jinan, China. *Journal of Transportation Engineering*. 2013, 139(11), pp. 1133–1140, doi: 10.1061/(ASCE)TE.1943-5436.0000589.

[18] LIU H., ZUYLEN H., LINT H., SALOMONS M. Predicting urban arterial travel time with state-space neural networks and Kalman filters. *Transportation Research Record: Journal of the Transportation Research Board*. 2006, 1968, pp. 99–108, doi: 10.3141/1968-12.

[19] MARKOWETZ F., EDLER L., VINGRON M. Support vector machines for protein fold class prediction. *Biometrical Journal*. 2003, 45(3), pp. 377–389, doi: 10.1002/bimj.200390019.

[20] PARK T., LEE S. A Bayesian Approach for Estimating Link Travel Time on Urban Arterial Road Network. In: International Conference on Computational Science and Its Applications (ICSSA 2004), Assisi, Italy. Berlin: Springer, 2004, pp. 1017–1025.

[21] PENG Z.X., SHAN W.X., GUAN F., YU B. Stable Vessel-Cargo Matching in Dry Bulk Shipping Market with Price Game Mechanism. *Transportation Research Part E: Logistics and Transportation Review*. 2016, 95, pp. 76–94, doi: 10.1016/j.tre.2016.08.007.

[22] REN Y., BAI G. Determination of optimal SVM parameters by using GA/PSO. *Journal of Computers*. 2010, 5(8), pp. 1160–1168, doi: 10.4304/jcp.5.8.1160-1168.

[23] THOMAS T., WEIJERMARS W., VAN BERKUM E. Predictions of urban volumes in single time series. *IEEE Transactions on Intelligent Transportation Systems*. 2010, 11(1), pp. 71–80, doi: 10.1109/TITS.2009.2028149.

[24] VAN LINT J.W.C., HOOGENDOORN S.P., VAN ZUYLEN H.J. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies*. 2005, 13(5-6), pp. 347–369, doi: 10.1016/j.trc.2005.03.001.

[25] VANAJAKSHI L., SUBRAMANIAN S.C., SIVANANDAN R. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. Intelligent Transport Systems. 2009, 3(1), pp. 1–9, doi: 10.1049/iet-its:20080013.

[26] WILLIAMS B.M., HOEL L.A. Modeling and forecasting vehicular traffic flow as a seasonal ARMA process: theoretical basis and empirical results. *Journal of Transportation Engineering*. 2003, 129(6), pp. 664–672, doi: 10.1061/(asce)0733-947x(2003)129:6(664).

[27] WU C.H., HO J.M., LEE D.T. Travel time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*. 2004, 5(4), pp. 276–281, doi: 10.1109/TITS.2004.837813.

[28] YAO B.Z., CHEN C., CAO Q., JIN L., ZHANG M.H., ZHU H.B., YU B. (2016a); Short-term traffic speed prediction for an urban corridor. *Computer-Aided Civil And Infrastructure Engineering*. 2016, 32(2), pp. 154–169, doi: 10.1111/mice.12221.

[29] YAO B.Z., HU P., ZHANG M.H., JIN M.Q. A support vector machine with the tabu search algorithm for freeway incident detection. *International Journal of Applied Mathematics and Computer Science*. 2014, 24(2), pp. 97–404, doi: 10.2478/amcs-2014-0030.

[30] YAO B.Z., YAO J.B., ZHANG M.H., YU L. Improved support vector machine regression in multi-step-ahead prediction for rock displacement surrounding a tunnel. *Scientia Iranica*. 2014, 21(4), pp. 1309–1316.

[31] YAO B.Z., YU B., HU P., GAO J.J., ZHANG M.H. An improved particle swarm optimization for carton heterogeneous vehicle routing problem with a collection depot. *Annals of Operations Research*. 2016, 242(2), pp. 303–320, doi: 10.1007/s10479-015-1792-x.

[32] YAO B.Z., HU P., YU L., ZHANG M.H., GAO J.J. Merged Automobile Maintenance Part Delivery Problem Using an Improved Artificial Bee Colony Algorithm. *Scientia Iranica*. 2015, 22(3), pp. 1258–1270.

[33] YOU J.S., KIM T.J. Development and evaluation of a hybrid travel time forecasting model. *Transportation Research Part C: Emerging Technologies*. 2000, 8(1-6), pp. 231–256, doi: 10.1016/S0968-090X(00)00012-7.

[34] YU B., KONG L., SUN Y., YAO B.Z., GAO Z.Y. A bi-level programming for bus lane network design. *Transportation Research Part C: Emerging Technologies*. 2015, 55, pp. 310–327, doi: 10.1016/j.trc.2015.02.014.

[35] YU B., LAM W. H., TAM M.L. Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*. 2011, 19(6), pp. 1157–1170, doi: 10.1016/j.trc.2011.01.003.

[36] YU B., SONG X.L., GUAN F., YANG Z.M., YAO B.Z. K-nearest neighbor model for multiple-time-step prediction of short-term traffic condition. *Journal of Transportation Engineering*-ASCE. 2016, 142(6), 04016018, doi: 10.1061/(ASCE)TE.1943-5436.0000816.

[37] YU B., WANG Y.T., YAO J.B., WANG J.Y. A comparison of the performance of ann and svm for the prediction of traffic accident duration. *Neural Network World*. 2016, 26(3), pp. 271–287, doi: 10.14311/NNW.2016.26.015.

[38] YU B., YANG Z. Z., CHEN K., YU B. Hybrid model for prediction of bus arrival times at next station. *Journal of Advanced Transportation*. 2010, 44(3), pp. 193–204, doi: 10.1002/atr.136.

[39] YU B., YANG Z. Z., WANG J. Bus travel-time prediction based on bus speed. *In Proceedings of the Institution of Civil Engineers-Transport*. 2010, 163(1), pp. 3–7, doi: 10.1680/tran.2010.163.1.3.

[40] YU B., YANG Z.Z., YAO J.B. Bus arrival time prediction using support vector machines. *Journal of Intelligent Transportation Systems*. 2006, 10(4), pp. 151–158, doi: 10.1080/15472450600981009.

Appendix 1

Authors	[12] Jeong & Rilett	[40] Yu et al.	[38] Yu et al.	[39] Yu et al.	[35] Yu et al.	[3] Chen et al.	[17] Lin et al.
Time of day		√	√	√		√	√
Weather		√	√				
Segment		√	√	√		√	
Bus time interval					√		√
Bus running time	√	√	√	√	√		√
Arrival time				√			√
Departure time						√	√
Stop distance							
Dwell time	√						
Day of week						√	√

Tab. V *Comparison of calculation methods.*

Appendix 2

Scenario 1

	Index 1	Index 2	Index 3	Index 4	Index 5	Index 6
Index 1	1					
Index 2	0.1729	1				
Index 3	0.6024	0.1644	1			
Index 4	−0.3130	−0.0957	−0.2892	1		
Index 5	0.0522	−0.1540	0.0733	0.8467	1	
Index 6	0.4989	0.1577	0.9980	−0.2906	0.0512	1

Scenario 2

	Index 1	Index 2	Index 3	Index 4	Index 5	Index 6
Index 1	1					
Index 2	0.1654	1				
Index 3	0.5504	0.1662	1			
Index 4	−0.3178	−0.0575	−0.2544	1		
Index 5	0.0741	−0.1467	0.1364	0.8549	1	
Index 6	0.5462	0.1730	0.9984	−0.2687	0.0370	1

Scenario 3

	Index 1	Index 2	Index 3	Index 4	Index 5	Index 6
Index 1	1					
Index 2	0.1654	1				
Index 3	0.5473	0.1638	1			
Index 4	−0.3175	−0.0582	−0.2529	1		
Index 5	0.0799	−0.1419	0.1539	0.8432	1	
Index 6	0.5504	0.1662	0.9997	−0.2531	0.1496	1

Scenario 4

	Index 1	Index 2	Index 3	Index 4	Index 5	Index 6
Index 1	1					
Index 2	0.1654	1				
Index 3	0.5451	0.1615	1			
Index 4	−0.3184	−0.0561	−0.2405	1		
Index 5	0.0823	−0.1557	0.1400	0.8742	1	
Index 6	0.5473	0.1638	0.9992	−0.2501	0.1361	1

Tab. VI *Person correlation coefficient matrix.*

Appendix 3

Index	Scenario 1 Extraction loading Square			Scenario 2 Extraction loading Square			Scenario 3 Extraction loading Square			Scenario 4 Extraction loading Square		
	Eigen-value	Percent-age	Accumulated Percentage	Eigen-value	Percent-age	Accumulated Percentage	Eigen-value	Percent-age	Accumulated Percentage	Eigen-value	Percent-age	Accumulated Percentage
1	2.626	43.770	43.770	2.643	44.048	44.048	2.634	43.903	43.903	2.617	43.624	43.624
2	1.215	20.258	64.028	1.226	20.435	64.482	1.195	19.918	63.822	1.208	20.130	63.754
3	0.939	15.645	79.673	0.904	15.074	79.557	0.897	14.947	78.769	0.948	15.795	79.549
4	0.721	12.021	91.694	0.729	12.152	91.708	0.764	12.734	91.503	0.734	12.232	91.781
5	0.654	10.374	101.753	0.684	10.574	102.879	0.674	10.465	101.879	0.634	10.268	101.298
6	0.593	8.685	109.864	0.623	8.765	110.196	0.603	8.699	110.054	0.549	8.679	108.979

Tab. VII *Extraction results by principal component analysis.*

Appendix 4

Stage one

Perf.	Model	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6	Seg. 7
MAE	PCA-GA-SVM	15.30	15.76	15.24	16.61	18.77	14.26	21.26
	SVM	20.16	16.57	22.70	26.92	22.24	27.88	27.65
	VPSSM	42.30	63.74	41.89	52.19	38.33	64.96	43.43
MAPE	PCA-GA-SVM	7.65 %	11.18 %	6.66 %	4.34 %	11.45 %	7.58 %	10.27 %
	SVM	10.08 %	11.75 %	9.91 %	7.03 %	13.56 %	14.83 %	13.36 %
	VPSSM	21.15 %	45.21 %	18.29 %	13.63 %	23.37 %	34.55 %	20.98 %
RMSE	PCA-GA-SVM	7.42	9.40	6.64	7.12	8.49	8.25	9.14
	SVM	11.76	7.61	9.67	11.49	11.23	15.62	11.70
	VPSSM	25.03	35.79	17.73	23.30	19.96	36.12	23.26

Stage two

Perf.	Model	Seg. 1	Seg.t 2	Seg.t 3	Seg. 4	Seg.t 5	Seg. 6	Seg. 7
MAE	PCA-GA-SVM	16.30	21.50	12.79	12.42	24.15	20.64	17.61
	SVM	21.50	30.05	26.84	20.23	29.59	33.78	18.27
	VPSSM	40.30	57.90	50.45	52.83	35.54	53.41	52.67
MAPE	PCA-GA-SVM	7.84 %	7.62 %	4.49 %	8.39 %	10.11 %	15.52 %	5.91 %
	SVM	10.34 %	10.66 %	9.42 %	13.67 %	12.38 %	15.40 %	6.13 %
	VPSSM	19.37 %	20.53 %	17.70 %	25.70 %	14.87 %	20.16 %	17.68 %
RMSE	PCA-GA-SVM	7.28	11.50	6.02	5.36	13.81	9.38	9.20
	SVM	10.35	14.35	13.27	10.45	13.44	16.96	7.66
	VPSSM	16.33	33.89	27.71	26.03	19.39	29.66	25.30

Stage three

Perf.	Model	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6	Seg. 7
MAE	PCA-GA-SVM	16.47	14.59	28.24	21.31	15.52	23.72	24.45
	SVM	22.56	31.57	30.63	28.91	34.09	34.47	36.90
	VPSSM	47.30	44.40	47.71	50.17	44.44	74.16	38.07
MAPE	PCA-GA-SVM	8.23 %	10.35 %	12.33 %	5.56 %	9.46 %	12.62 %	11.81 %
	SVM	11.28 %	12.39 %	13.38 %	7.55 %	15.78 %	18.33 %	17.83 %
	VPSSM	23.65 %	31.49 %	20.84 %	13.10 %	27.10 %	39.45 %	18.39 %
RMSE	PCA-GA-SVM	9.37	6.84	14.94	11.92	7.92	10.36	14.40
	SVM	9.78	16.53	16.37	13.12	16.18	15.90	16.88
	VPSSM	26.60	24.13	26.30	21.28	22.79	42.16	22.16

Stage four

Perf.	Model	Seg. 1	Seg.t 2	Seg.t 3	Seg. 4	Seg.t 5	Seg. 6	Seg. 7
MAE	PCA-GA-SVM	15.66	24.79	13.61	19.54	15.27	25.82	25.63
	SVM	21.23	35.88	32.35	34.42	19.30	28.12	29.77
	VPSSM	45.21	63.74	41.89	52.19	38.33	64.96	43.43
MAPE	PCA-GA-SVM	7.53 %	10.35 %	12.33 %	5.56 %	9.46 %	12.62 %	11.81 %
	SVM	10.21 %	12.39 %	13.38 %	7.55 %	10.78 %	18.33 %	17.83 %
	VPSSM	21.74 %	22.60 %	14.70 %	25.26 %	16.04 %	28.84 %	14.58 %
RMSE	PCA-GA-SVM	6.44	12.46	7.73	9.76	6.30	12.08	11.52
	SVM	11.96	19.63	17.94	19.62	7.88	14.64	15.44
	VPSSM	25.91	33.38	19.74	26.73	20.70	27.33	22.55

Tab. VIII *Comparison of prediction effectiveness.*