Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

Article

Predicting Freeway Travelling Time Using Multiple-Source Data

Kejun Long[1], Wukai Yao[2], Jian Gu[1]*and Wei Wu[1]

[1] Hunan Provincial Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems, Changsha University of Science & Technology, Changsha 410004, China; longkejun@csust.edu.cn; jiaotongweiwu@csust.edu.cn; gujian87@126.com

[2] School of Traffic and Transportation Engineering, Changsha University of Science & Technology, Changsha 410004, China; 185803450@qq.com

* Correspondence : gujian87@126.com; Tel.: +86-731-8525-8575

Abstract: Freeway travelling time is affected by many factors including traffic volume, adverse weather, accident, traffic control and so on. We employ the multiple source data-mining method to analyze freeway travelling time. We collected toll data, weather data, traffic accident disposal logs and other historical data of freeway G5513 in Hunan province, China. Using Support Vector Machine (SVM), we proposed the travelling time model based on these databases. The new SVM model can simulate the nonlinear relationship between travelling time and those factors. In order to improve the precision of the SVM model, we applied Artificial Fish Swarm algorithm to optimize the SVM model parameters, which include the kernel parameter σ, non-sensitive loss function parameter ε, and penalty parameter C. We compared the new optimized SVM model with Back Propagation (BP) neural network and common SVM model, using the historical data collected from freeway G5513. The results show that the accuracy of the optimized SVM model is 17.27% and 16.44% higher than those of the BP neural network model and the common SVM model respectively.

Keywords: Support Vector Machine; Travelling time; Intelligent Transportation System; Artificial Fish Swarm algorithm; Big data.

1. Introduction

Travel time is one of the main indexes that reflect the traffic operation level of a freeway, and it is also the basis for Advanced Traveler Information System (ATIS), Traffic Guidance System(TGS), and Traffic Control System(TCS). The challenges and difficulties of travel time prediction are identified below.

- Diverse influencing factors such as weather, holidays, traffic accidents, out of sample prediction, and mechanisms contributing to congestion. It's difficult to describe and predict the influence mechanism by using traditional conventional mathematical models.
- The complexity and incompleteness of basic data. Although there are many flow detectors and video detection equipment on the freeway, captured data are incompatible, redundant, and include error or loss. To avoid these, techniques which utilize multi-source data to improve the accuracy of travel time prediction is extremely important.

In China, practical application of travel time prediction focuses mainly on the following two aspects.

The first aspect is the prediction of travel time by map navigation providers using their personalized GPS data. Many map service providers employ their personalized data for travel time forecast services and commercial products. For instance, Bai-du, Gao-de, and other Chinese map providers collect real-time GPS data from users while providing map navigation services. Then, a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

correlation algorithm is proposed to obtain the travel time prediction result at road sections, which depends on the market share of the map navigation service. The higher frequency of people using the navigation service, the more complete of GPS data and the higher the prediction accuracy will be. However, according to the Chinese market report, the market share of Bai-du and Gao-de services are presently 29.3% and 32.6%respectively. Therefore, the accuracy of results should be further improved by increasing market share.

The second aspect is the prediction based on the traffic detection data of urban traffic managers and historical data. In recent years, numerous fixed detector devices have been installed in most of urban roads and rural freeways for the prediction of travel time, including inductive loops, video recorder, microwave, and laser detection. However, unavoidable damage to flow detection equipment and transmission error of partial data make traffic detection data incomplete redundant or error. In addition, different detectors have different data formats and data accuracy. With the rough use of mistake data for precise travel-time prediction, the Traveler Information Service system cannot recommend optimal travel routes or warn of potential traffic congestion and users cannot determine the optimal departure time or estimate their expected arrival time based on predicted travel times.

Theoretical research on freeway travel time prediction can be divided into two categories based on single source data and multi-source data.

1.1 Overview ofPrediction Method of Single Source Data

A single data source was earlier method used for predict travel time. Many researches prediction results were obtained upon a single data source.

Gipps, P. G. [1]used the occupancy and arrival time to predict the travel time in a road in 1997. Mehmet Y, Nikolas G [2]set statistical predictive algorithms to predict the future travel time. Shen, L., &Hadi, M. [3]employed data obtained from detector in freeway. Kyung et al. [4,5]used inductive loop detectors to obtain the front position and capture the interactions between trucks and non-trucks. But fixed detector devices is easily affected by external environment and cannot directly access some important parameters, such as travel time, etc.

In addition ,many researches consider using GPS data to predict travel time. Ramezani et al. [6] and Zhang et al. [7,8]considered the diversity of GPS data and investigated the application of Markov chain to travel time estimation and implemented good prediction accuracy. Woodard, Nogin &Paul et al. [9]used the GPS data of the current highest volume GPS data source, and applied the TRIP method to predict the travel time. Based on GPS data sets, Bahuleyan, H., &Vanajakshi, L. D. [10] proposed a prediction method for urban trunk lines which was only suitable for traffic conditions in India. But the GPS data only gets the speed and real-time position information and uncertainty due to the route of the vehicle, it affects the coverage and accuracy of detection data.

Above methods indeed are innovations and improvements in travel time prediction, and results are more accurate. However, many predictions that use a single data source do not consider the impact of other unexpected events or the result was not accurate enough because single data source cannot reflect traffic state of road network exactly. It will result in certain errors between prediction results and true values.

1.2Overview ofPrediction Method of Multiple Source Data

Nowadays, the development of traffic big data environment has progressed rapidly. With the support of a large amount of data, it is possible to clearly visualize traffic flow changes under the joint action of different factors, that is, the traffic state presented, which is more favorable. The construction of the predictive model improves the adaptability and accuracy of the model,[11] and if the same state occurs, it can be predicted based on historical results. The more populated the database is, the higher the quality and the higher the likelihood of finding commonalities and predicting accurate results will be. This concept can be applied by searching for common traffic states for prediction.

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

Owing to progress in the dynamic traffic information acquisition system, various traffic data can be collected more easily. And data fusion is finished in the dynamic traffic information acquisition system, which is jointly determined by the advantages of multi-source data and the characteristics of traffic conditions.

And using multi-source data for prediction can overcome the limitations of the single data source. In other words single data source has low quality and is not comprehensive. The traffic state is described from different angles and directions to improve the accuracy of prediction and reduce disturbance from unexpected factors.

At present, many studies have been conducted on travel time prediction, especially studies based on the historical data travel time of multi-source data.

The common predicting methods and their characters are summarized in the table below.

Table 1. Comparison of common prediction methods

Prediction method	**Author**	**Data source**
Kalman filter	Lin J W C V.2006[12]	Travel time data
	Zhou, J. 2014[13]	Floating vehicle and fixed detector data
	Tang-Hsien Chang.2016[14]	Electronic Toll Collection (ETC) and traditional Vehicle Detector data
Bayesian estimation	Fei, X.2011[15]	The real loop detector data of an I-66 segment in Northern Virginia
	Zhan, X. 2016[16]	A large-scale taxi trip dataset from New York City
Statistical decision theory	Wosyka,J.2012[17]	Two detectors datain Prague and also in the Czech Republic.
Neural network	Innamaa, S.2005[18]	travel time data
	Lin J W C V.2005[19]	Travel data and Some missing or corrupt travel data

Compared to the single source data, the multi-source data method can extract deep information within data, significantly reduce the cost of data acquisition, and make up for lack of information and packet loss of single source data. At present, big data application technology are widely used in the traffic field, many studies have been conducted in the field of freeway travel time prediction based on big data analytics. However, there are several deficiencies including:

- The method pays attention to machine learning algorithms and lacks the mastery of the characteristics of the traffic flow, resulting in the uncoordinated and unsuitable correspondence between the data and the traffic flow.
- With the continuous updating of big data, it provides conditions for traffic travel time prediction, but some advantages and characteristics of these data are not noticed, resulting in many useful data not being used and mined
- Some model parameter calibration is too subjective, which largely depend on researchers' experience;
- Some model mostly aimed at a specific example, and cannot be easily adapted to other situations.

Therefore, in this study, historical data of a freeway toll station were collected, and were categorized using the support vector machine (SVM) algorithm. Although the predecessors have done some work: Wu, Chun-Hsin[20]used the method of support vector regression to predict the time; Vanajakshi[21] obtained the support vector for short-term prediction of travel time by algorithm. Machine technology; Mendes-Moreira [22]obtained a regression method for comparing long-term travel time prediction through intelligent data analysis. But their analysis is based on machine learning algorithms and does not better understand or improve the transportation system. So this paper uses SVM model based on historical data to predict the common traffic state and the method used for model construction was simplified. The practicability of the prediction algorithm was enhanced to overcome assumptions and uncertainties in the existing traffic flow theory.

2. Data Collection and Preprocessing

2.1 Data Description

Data for this study was collected in the FreewayG5513 from Changsha to Yiyang, Hunan Province, starting from the Changsha toll station and ending at Yiyang toll station. This freeway G5513 is a standard freeway with a two-way four lane of 100 km/h and a roadbed width of 26 m. The total length was approximately 63 km, the daily average flow reached 58,000 vehicles, and the peak flow during long vacation was up to 96,600 vehicles. Because of heavy traffic, the freeway has been rated as one of the six most congested sections in the Hunan Province. There are nine toll stations along the road, from east to west, that is Changsha West, Youren, Guanshan, Jinzhou, Ningxiang, Xiangjiang West, Quanjiao, Chaoyang, and Yiyang North, illustrated as Figure1.

Figure 1. Layout of the freeway and toll station

The main data set collected in this study includes:

- Toll data of the whole toll stations along G5513 in February 2018 (vehicles entering and leaving toll station), with a total of 561,081 data items, including the name of the toll station, the time of vehicle entering and leaving the toll station, vehicle type and weight.
- Weather information surveying station located near the freeway, which was collected from the Chinese Weather Network in February 2018, with a total of 672 data items, including 24 hour daily weather, temperature, relative humidity, precipitation, and wind direction.
- Freeway blockage record statistics, which was obtained from the freeways management department, a total of 260 freeway blockage information reports were collected in February, March, April, and May, including blockage location, reasons for the blockage, blockage start time, and blockage end time.
- Freeway traffic control measures report, which was obtained from the Traffic Police Department, with a total of 7 data items collected on April 5 Qingming traditional national Festival, May 1 International Labor Day, and other holiday control information.

2.2 Data Preprocessing

Many abnormal data items were found in the data, which need to be preprocessed before use.

- Data sharing the same entry and exit toll

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

On the freeway, some drivers turn around in the service area or other sections to avoid the charges and even exchange the toll tags, which is likely to make the entry and exit of the vehicles at toll stations consistent.

Therefore, it is necessary to determine whether a data item is consistent with the toll gates and eliminate invalid data items.

- Abnormal time record data

Owing to the failure of the time system associated with toll station to synchronize or system failure, the time of accessing the toll station can be earlier than the time of exiting the toll station. In addition, there are other factors that can lead to long travel time, such as the breakdown of vehicle on road, accidents, and the situation where driver may have a long rest in service area. All of these situations will result in unusual time record data.

In the process of data preprocessing, abnormal time data record can be eliminated by screening.

- Missing data

There are two main reasons for missing data: on the one hand, it's mainly from equipment problems or road environment includingthe unstable scanning frequency of detector, faulty of transmission equipment, and traffic jam On the other hand, eliminating wrong data items will also lead to partial data missing . Lack of data will cause the road real traffic conditions to change directly or indirectly. Therefore, it is essential to make up for the missing data for historical data. Because of the strong continuity of the traffic flow travel time parameter, the trend in the change of the traffic flow travel time parameter with time is consistent, although its fluctuation will change as the collection period changes. Therefore, the following data fill formula is obtained, as given in equation (1).

$$data(t)=\frac{3}{6}\times data(t-1)+\frac{2}{6}\times data(t-2)+\frac{1}{6}\times data(t-3) \quad (1)$$

Where, *data(t)*represents the current missing data, *data(t-1)*, *data(t-2)* and *data(t-3)* arethe traffic flow travel time data of the past period, two cycles, and three cycles are respectively represented.

3. Support Vector Machine Model

3.1 Problem Descriptionof Freeway Travel Time Prediction

Travel time of a freeway has strong continuity in a certain time range. That is, there are some complex functional relationships between the current travel time and the past travel time. By analyzing the changes in travel time, we can obtain rules and establish a real-time prediction model updated every 5 min. Then, the accuracy and reliability of the predicted results can be improved by using an optimization algorithm to find the optimal solution of the model.

The change in the freeway travel time in different time periods is not a simple linear relationship, and it will neither increase indefinitely, nor decrease indefinitely. But it will only change continuously within a floating interval. Therefore, using a simple least squares regression prediction or similar methods is not sufficient to predict the travel time. The SVM nonlinear regression theory can be employed to solve this problem.

SVM uses nonlinear transformation to map the original variables to a high-dimensional feature space, so that the problem of nonlinear separability in the original sample space is transformed into high-dimensional feature space. The linear separable problem and the application of expansion theorem of the kernel function in the calculation process do not require the explicit expression of nonlinear mapping. In addition, since the linear learning machine is established in the high-dimensional feature space, it can be compared to the linear model. The comparison not only increases the complexity of the calculation, but also solves the problem of "dimensional disaster".

Owing to changes in the traffic environment, sudden traffic accidents, weather, and other special events, the SVM algorithm will eventually be transformed into a quadratic programming problem. In theory, a global optimal solution can be obtained, thus solving the traditional neural network. The network can' avoid the local optimal problem, and should adequately accommodate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

the influencing factors due to these sudden changes to improve the accuracy of the travel time prediction result.

Therefore, this study used the nonlinear support vector machine regression theory[23].

3.2 Model Overview

The SVM is a machine learning method, which is based on the statistical learning theory developed by Vapnik. The theory has been further extended to diversified application algorithms, including the linear SVM classification algorithm, the nonlinear SVM classification algorithm, the linear SVM regression algorithm, and the nonlinear SVM regression algorithm[24]. These SVM algorithms have been widely used in many fields owing to their simple structure and high computational efficiency.

Consider a training sample set of i training samples, $S=\{(x_i,y_i)x_i\in R, y_i\in R\}_{i=1}^{l}$, which is non-linear, where x_i is the input column vector of the i training sample, $x_i=\left[x_i^1,x_i^2\cdots,x_i^d\right]^T$, $y_i\in R$ is the corresponding output of the kernel function $K(x_i,y_i)=\partial(x_i)^T*\partial(x_j)$.

Equation (1) in the following is the linear regression equation established in the high dimensional space, and ε is introduced as a linear insensitive loss function:

$$f(x)=\omega\partial(x)+b \tag{2}$$

$$L(f(x),y,\varepsilon)=\begin{cases}0,|f(x)-y|\le\varepsilon\\ |f(x)-y|-\varepsilon,|f(x)-y|>\varepsilon\end{cases} \tag{3}$$

Where, $\partial(x)$ is the nonlinear mapping function, $f(x)$ is the prediction function, which returns the predicted value, and y is the corresponding real value.

Under the above constraints, we can find the optimal classification hyper-plane, that is, find the solution to the following optimization problem.

$$\begin{cases} min\dfrac{\|\omega\|^2}{2}\\ s.t.\left\|\omega^T\partial(x_i)+b-y_i\right\|\le\varepsilon,i=1,2,\cdots,l\end{cases} \tag{4}$$

This problem can be solved by solving the saddle point of the Lagrange function, and its dual theory can be applied to solve the dual problem.

$$\begin{cases} min\dfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j)+\varepsilon\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)-\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)\\ s.t.\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0,\alpha_i^*,\alpha_i\ge0,i=1,2,\cdots,l\end{cases} \tag{5}$$

To solve the dual problem, a relaxation factor can be set for each data point. After introducing these two relaxation factors, ξ_i,ξ_i^* ($\xi_i,\xi_i^*\ge0,i=1,2,\cdots,l$) ,the function can be optimized as:

$$\begin{cases} min\dfrac{\|\omega\|^2}{2}+C\sum_{i=1}^{l}(\xi_i-\xi_i^*)\\ s.t.\omega^T\partial(x_i)+b-y_i\le\xi_i+\varepsilon,i=1,2,\cdots,l\\ s.t.y_i-\omega^T\partial(x_i)-b\le\xi_i^*+\varepsilon,i=1,2,\cdots,l\end{cases} \tag{6}$$

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

In the above equation, C is the penalty factor; the smaller the value, the smaller the penalty to the error data.

Next, the Lagrangian multiplier method can be used to solve the optimization algorithm, and the nonlinear regression function can be further used to solve the double optimization problem.

$$f(x)=\omega\partial(x)+b=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\partial(x_i)^T*\partial(x_j)+b=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,y_i)+b \tag{7}$$

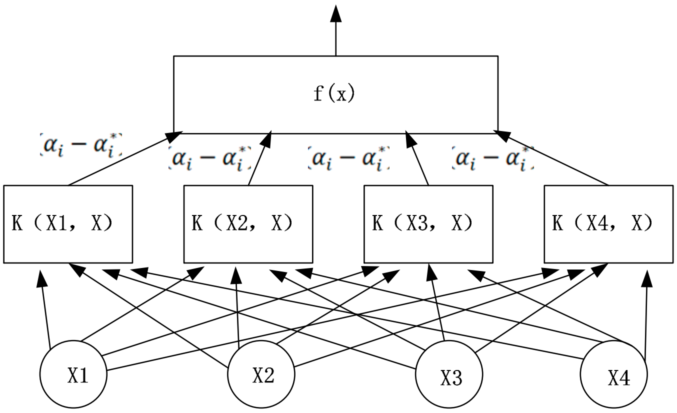

Figure 2.Structure of SVM model

For the freeways, there is essentially no significant error in the travel time between the low peak period and even the flat peak period. However, the impact of different traffic conditions on the travel time is inevitable during peak periods. Therefore, the two cases should be discussed separately. Furthermore, whether or not the working day has a different influence on the travel time of vehicles traveling on the highway. The commuting time, travel purpose, and travel mode will be also different. Therefore, these two points should be viewed separately. In addition, road weather conditions and traffic control will have certain influence on the prediction results and should be considered.

Based on the SVM model, input workdays, non-working days, the morning and evening peak periods, and off-peak hours as original values, finally can get six time periods: the morning peak hours on workdays, the evening peak hours on workdays, off-peak hours on workdays and so on. Weather and traffic control factors for the four scenarios can also be analyzed. However, the difference in highway traffic conditions between the working day and the non-working day, the morning and evening peak periods and flat peak period is not considered due to the limitation of the length of the article.

This study used the travel time prediction of evening peak hours in the classified working days and non-working day peak hours as an example by comparing the traffic data of multiple working days and non-working days.

3.3 Model Construction

The freeway travel time prediction model is based on the SVM algorithm and is constructed based on the relationship between the current travel time of the road segment and the past travel time of the road segment, the current weather, and the possibility of traffic control.

In this study, data from two toll stations with different distances from east to west of G5513 were selected for analysis. Moreover, as the first-class passenger car (7 passenger car) accounts for the vast majority of the data, the travel time of the first-class passenger car was taken as the prediction object, and the analysis time interval is 5 minutes.

The characteristic of the toll station is presented in table 2.

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

Table 2. Analysis objects of freeway travel time prediction

Toll station	Start point	End point	Distance (kilometer)
1	Changsha West	Guanshan	10.6
2	Changsha West	Ningxiang	23.2

The structure of the SVM is similar to the neural network. The output is a linear combination of intermediate nodes, and each intermediate node corresponds to a support vector. To determine the optimal classification function, this study takes the four travel times of the time period before the prediction time as the input, namely $t_{k-1}, t_{k-2}, t_{k-3}, t_{k-4}$.

$$t_k = g\left(t_{k-1}, t_{k-2}, t_{k-3}, t_{k-4}\right) \tag{8}$$

k is the current time period, and t_k represents the average travel time of all vehicles in the current predicted time period.

In the prediction process, variables such as weather, traffic accident that affects the travel time, holiday or non-holiday, and day of the week, are evaluated as follows:

Table 3.Variables type and its meanings

Variable category	Variable name	Variable type	Variable value	Variable meaning
Meteorological	Weather status	Discrete	Clear; Cloudy; Fog; Overcast; Light rain; mod rain; hvy rain	Clear; Cloudy; Fog; Overcast; Light rain; mod rain; hvy rain
Time	Holiday	Discrete	0	N
			1	Y
	Weekly	Discrete	1	Monday
			...	...
			7	Sunday
Accident	Affecting travel time traffic accidents	Discrete	0	N
			1	Y

Many traffic accidents occur on the freeway every day. To simplify the parameters, we divide traffic accidents into two categories: traffic accidents that affect the travel time and those that do not. In this study, traffic accidents that affect the travel time was regarded as invariant.

The following figure shows a comparison of travel time affected by an accident and normal travel time on the afternoon of February 17.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

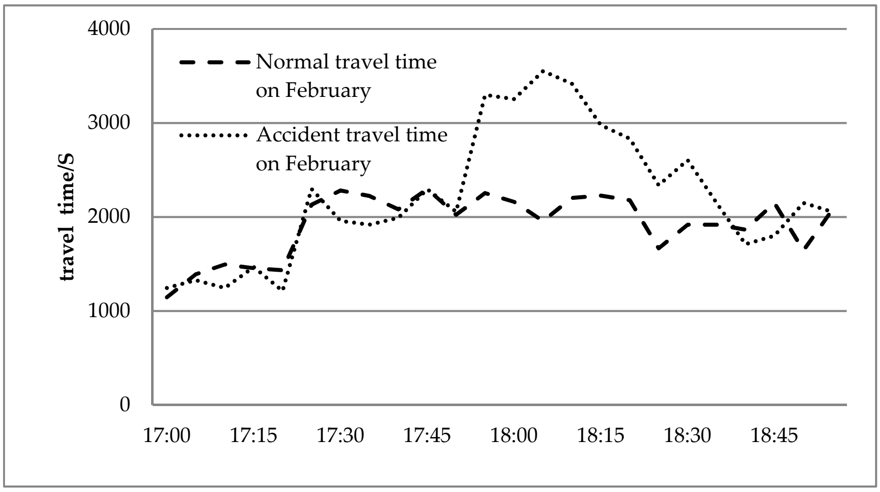

Figure 3. Comparison of accidents and normal travel time

Since SVM is a machine learning model, sample training is required before prediction. Multiple groups of any four consecutive travel times, daily weather, traffic accidents that affect the travel time, holidays, and weekday data were used as training samples to obtain a trained model. In the trained model, t_{k-1}, t_{k-2}, t_{k-3}, t_{k-4}, weather, accident, holiday, and week are used to predict the travel time of the next time period. When a certain number of training samples is achieved, real-time data input can be adopted to predict future results. Moreover, the model can be constantly modified based on the relation between the predicted data and the predicted data to prediction accuracy.

3.4 Parameter Calibration and Optimization

Parameter selection is very important to find the optimal hyper-plane in the SVM model used in this study. Existing studies mainly adopted the traditional grid search method, direct determination method, one-dimensional search method, and inverse ratio method to determine the insensitive loss function parameter ε and penalty parameter C. However, there are many shortcomings associated with these methods, and the resulting errors will significantly influence the accuracy of the prediction results.

Moreover, in the SVM model, kernel function selection is also an important factor that influences the performance of the SVM. The radial basis kernel function $K\left(x_i, y_i\right)=e^{\frac{\|x-x_i\|}{\sigma^2}}$ (RBF) is an adaptive kernel function for low-dimensional space data and high-dimensional space, which have good convergence domains, and this function can be described as an ideal kernel function. Therefore, the RBF was selected as the classification prediction kernel function of the SVM, in which a kernel parameter σ needs to be optimized.

Therefore, three parameters need to be optimized, namely the core parameter σ, the non-sensitive loss function parameter ε, and the punishment parameter C. The kernel parameter σ is the distribution or range of the training sample data. The non-sensitive loss function parameter ε affects the number of support vectors. The larger the value of ε, the lower the regression precision, and the fewer the support vectors. The penalty parameter C is used to control the degree of punishment of samples beyond the allowable error range. The higher the value, the heavier the punishment of samples.

We used the artificial fish swarm algorithm to optimize the parameters of the regression model. The artificial fish swarm algorithm has unique advantages in parameter optimization and overcomes the blindness of traditional algorithms in parameter optimization and the defects of the linear model and neural network in parameter selection. It can be said that the parallel performance of the artificial fish swarm algorithm can ensure that the model parameters converge faster to the global optimization extreme[24,25].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

The first step in the optimization process of the artificial fish swarm algorithm is to feed in the training value and the training target through the SVM model to calculate the fitness of the individual. The most adaptable individual is regarded as the optimal value of the current fish group and the corresponding parameters σ, ε, and C of the current optimal value are saved. In the subsequent iteration, σ, ε, and C corresponding to the maximum fitness value are taken as the final optimization results.

4. Case Study

4.1 Data Selection

The data used in this study was collected in February 2018 on G5513 (from Changsha West to Guanshan/Ningxiang Station)in Changsha, Hunan Province, China. The travel time was detected for all days and the detection interval is 5 minutes. Furthermore, 288 sequences are included in one day.

The daily evening peak (17:00-19:00) data of G5513 (Changsha West to Guanshan /Ningxiang Station) was selected as an example after comparing data of multiple working days and non-working days, which contains 204 to 228 items. Other variable data items also need to be filtered according to the above data. The dimension feature values are based on the time series and the data requirements according to the prediction model.

In this study, the regression SVM model is used to establish the model parameters, and the artificial fish swarm algorithm is used to establish the model parameter optimization algorithm. The optimization results are presented in Table 4. The optimization process for the optimal value of the penalty parameter C is shown in Figure 4.

Table 4. Optimization of parameter values

Section	Penalty parameter, C	Nuclear parameter, σ	Insensitive loss function parameter, ε
Changsha West-Guanshan	6.8755	0.0064	0.3461
Changsha West-Ningxiang	8.6485	0.0034	0.6991

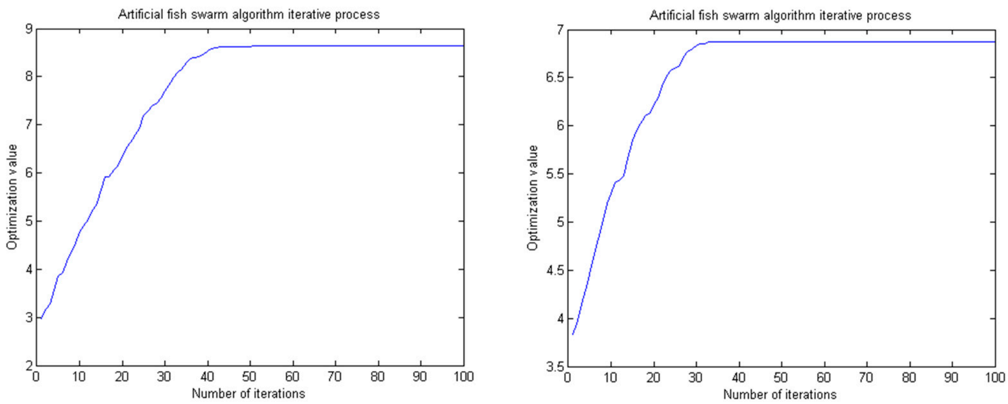

Figure 4. The parameter optimization curve

The parameters of the artificial fish swarm algorithm are set as follows: the maximum number of iterations of the artificial fish is 100; the population size is 5; the maximum number of trials is 5; the crowding factor δ is 0.618; the perceived distance is 0.5; and the moving step is 0.1.

4.2 Results and Comparative Analysis

The data adopted in this study were obtained during the Chinese Spring Festival from February15 to 21, 2018.Therefore, the data was divided into working days and holidays. There were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

336 sets of data from February1 to 14,2018(14 days data) in each group of toll stations; 264 groups (11 days) were randomly selected as training data input, and 72 groups (3 days) were adopted as prediction numbers. There were 168 sets of data from February15to 21,2018(11 days data) in each group of toll stations;96 groups (4 days) were randomly selected as training data input, and 72 groups (3 days) were adopted as prediction numbers.

The detection time is from 17:00 to 19:00 P.M. and the value was taken as test data.

BP neural network, SVM, and optimized SVM were used for the prediction. The root mean square error (RMSE), the mean absolute percent error (MAPE) and the covariance protocol (CP) were selected as the error evaluation criteria in the prediction process[26].

The RMSE is a comprehensive evaluation indicator of the prediction effect, the MAPE is the prediction relative error, while the CP is the error component analysis indicator.

The following figures show the forecasting effect diagram of the holiday evening peak.

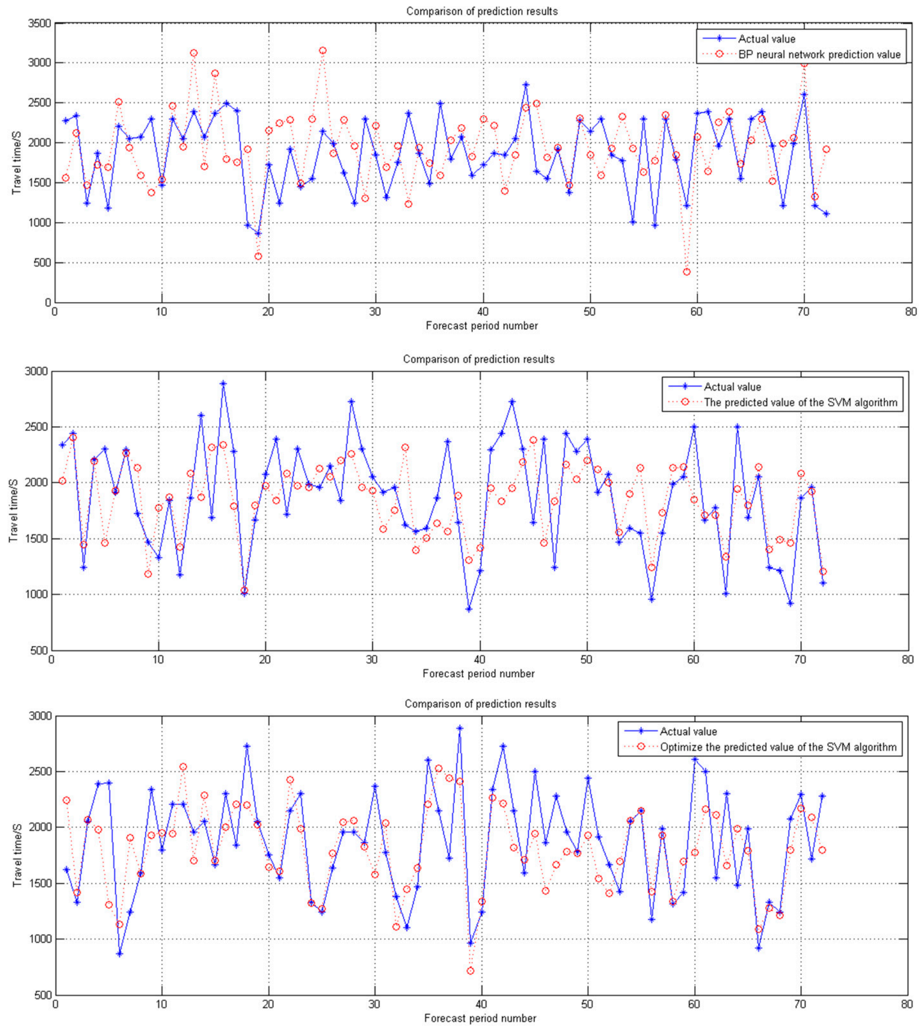

Figure 5. Results of freeway travel time prediction

Table 5. Error evaluation of forecasting freeway travel time

Method		Changsha West- Guanshan Station			Changsha West-Ningxiang Station		
		RMSE	MAPE	CP	RMSE	MAPE	CP
Working day	BP neural network	7.6544	5.0873	0.5489	10.5124	5.5104	0.5809
	SVM	6.3405	4.8225	0.5810	9.0375	4.5422	0.6930

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

	Optimized SVM	6.0369	4.6536	0.5952	8.3069	4.4524	0.6404
Holiday	BP neural network	11.6308	9.1023	0.5990	16.0845	9.6148	0.6454
	SVM	11.5152	7.8645	0.6460	13.3215	8.1153	0.6905
	Optimized SVM	9.6218	6.2451	0.7621	12.2548	7.8651	0.7245

From Table 5, it can be observed that the working day was predicted by the RMSE, the MAPE, and the CP data between the two toll stations. It could be found that all three models can be used for predicting travel time.

Although the prediction error of the BP neural network may be larger than those of the SVM and the optimized SVM models, there is no deviation between the error of the SVM and the optimized SVM model.

However, when forecasting holidays with large traffic and long travel time, the RMSE of the optimized SVM model is significantly better than those of the BP neural network and the SVM model.

In the prediction of Changsha West-Guanshan, the accuracy of the optimized SVM model using artificial fish swarm algorithm is 17.27% higher than that of the BP neural network model and 16.44% higher than the conventional SVM model.

In the prediction of Changsha West-Ningxiang, the accuracy of the optimized SVM model using artificial fish swarm algorithm is 23.80% and 8.01% higher than those of the BP neural network model and the conventional SVM model, respectively.

The optimized SVM model described in this paper has higher travel time prediction accuracy in the road segment, and the mapping law of the input and output are better represented by the optimized SVM model.

In terms of the relative prediction errors of the three prediction models, the MAPE of the optimized SVM model is lower than the prediction errors of the BP neural network and the conventional SVM model when using holiday and working day data. This indicates that the optimized SVM model described in this paper has certain advantages in terms of the travel time prediction model of the road segment, and the data requirements are lower.

4.3 Analysis of Influencing Factors of Travel Time

The freeway travel time is determined by various factors, such as weather, traffic accident, holiday, and week day. However, owing to limitations of sample data, only traffic accidents and holidays were considered.

4.3.1 Effect of traffic accidents on travel time

First, the optimized SVM model described in this paper is superior to the BP neural network and SVM model in terms of the CP.

In the prediction of Changsha West-Guanshan, the CP of the optimized SVM model is 21.40% higher than that of the BP neural network model, which is 7.28% higher than that of the conventional SVM model;

In the prediction of Changsha West-Ningxiang, the CP of the optimized SVM model is 10.9% and 6.53% higher than those of the BP neural network model and the conventional SVM model, respectively.

The results presented in this paper indicates that the optimized SVM model has better inclusiveness and stability when unexpected factors such as traffic accidents that affect the travel time are encountered, thereby avoiding the need for repeated trial and error to address network problems.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2018 doi:10.20944/preprints201810.0601.v1

Peer-reviewed version available at Appl. Sci. 2018, 9, 104; doi:10.3390/app9010104

Second, freeway traffic accidents will cause the traffic capacity of certain sections of the road network to decrease, and queues will be formed near the accident site, which increases the travel time of the vehicle.

4.3.2 Effect of holidays on travel time

A comparison of the working day and holiday forecasting error evaluation criteria presented in the previous section indicates that the BP neural network has a larger prediction error than the working day prediction results of the other models probably due to the problem of construction of the network structure. However, the conventional SVM and the optimized SVM models have similar prediction error results. In the above analysis, there are large gaps in the holiday prediction results of the three different models.

It can observed that holidays have significant influence on the travel time. The effect of holidays on the travel time, which was obtained from the analysis of the original toll station data, is that it significantly increases the volume of traffic on the highway network.

5. Conclusions

This study performed an in-depth analysis of freeway travel time prediction to provide high-quality travel experience for users and found that the freeway travel time is affected by travel time, weather, and traffic. The effect of different factors was analyzed, such as accidents and holidays.

Bad weather reduces the overall traffic rate, which increases the travel time. Traffic accidents lead to reduced road traffic capacity, which affects the travel time. Free passage on highways during holidays and the increased demand for travel result in increased vehicle flow, which also affects the travel time.

In this study, basic data was analyzed, and the traffic state prediction method based on SVM data mining technology was proposed to transform the problem into a quadratic programming problem using artificial fish swarm algorithm, which reduces the computational and local optimal problems of traditional neural networks. The parameters of the SVM were optimized using traditional network optimization, and a global optimal solution was obtained.

Results show that the accuracy of the optimized SVM model is 17.27% and 16.44% higher than those of the BP neural network model and the conventional SVM model, respectively. Accurate prediction of the travel time on the freeway was realized, which can provide data support for monitoring, early warning, and decision analysis for the freeway operation status.

In this study, influencing factors such as weather, traffic, accidents and holidays were included in the optimization of the SVM prediction model. However, owing to limitations of the number of samples, the model was not fully trained. Therefore, a certain error occurred in the prediction results.

In the future, it is necessary to categorize traffic accidents, clarify the impact of each type of accident on the travel time, categorize increase in the holiday traffic, and clarify the impact of each level of accident on the travel time. Furthermore, the number of training samples, database capacity, and prediction accuracy should be continuously increased. In this study, only data from freeway toll stations was validated, and application to actual large-scale road networks should be further explored in the future.

Author Contributions: This work was conducted by Kejun Long and Wei Wu with the help of graduate student Wukai Yao. It was mainly drafted by Kejun Long and Wukai Yao, and checked and revised by Wei Wu and Jian Gu. Kejun Long and Wukai Yao designed and analyzed the proposed model. Wukai Yao and Jian Gu performed the simulation. Kejun Long is responsible for the English polish and proofreading of the work.

Funding: This research was funded by National Natural Science Foundation of China (NSFC),grant number 51678076, Hunan Provincial Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems ,grant number2017TP1016.

Acknowledgments: The authors express their thanks to all who participated in this research for their cooperation. The authors would like to give great thank to the hard work by the peer reviewers and editor.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gipps,P.G. The estimation of a measure of vehicle delay from detector output. Newcastle-Upon-Tyne University, England1977, 18 p.
2. Yildirimoglu, M; Geroliminis, N. Experienced travel time prediction for congested freeways. Transp. Res. Part B-Method, 2013, 53(4):45-63.
3. Shen,L; Hadi,M. Practical approach for travel time estimation from point traffic detector data. J. Adv Transportation, 2013, 47(5):526-535.
4. Hyun ,K; Tok,A, Ritchie S G. Long distance truck tracking from advanced point detectors using a selective weighted Bayesian model. Transport. Res. Part C-Emerg. Technol, 2016, 82:24-42.
5. Hyun, K. K.; Jeong, K. Assessing crash risk considering vehicle interactions with trucks using point detector data. Accid. Anal. Prevent, 2018, 17p.
6. Ramezani, M.; Geroliminis, N. On the estimation of arterial route travel time distribution with Markov chains. Transp. Res. Part B, 2012, 46(10): 1576-1590.
7. Zhang, J. T.; Zhou, J. An Arterial Travel Time Estimation Model Based on Discrete Time Markov Chains. Syst. Eng, 2014,5:98-104.
8. Zhang, J. T.; Zhou, J. Travel time estimation model based on spatial Markov chains. Syst. Eng,2015,12 :72-77.
9. Woodard, D.; Nogin,G. Predicting travel time reliability using mobile phone GPS data. Transport. Res. Part C-Emerg. Technol, 2017, 75:30-44.
10. Bahuleyan, H.;Vanajakshi, L. D. Arterial path-level travel-time estimation using machine-learning techniques. J. Comput. Civil. Eng, 2016, 31(3), 04016070.
11. Tan, Y. Current Situation of Short-term Flow Forecasting and Discussion on Forecasting Method in Big Data Environment. ITS China:2017:10.
12. Lin J W C V. Incremental and online learning through extended kalman filtering with constraint weights for freeway travel time prediction. ITSC. IEEE, 2006:1041-1046.
13. Zhou, J.; Zhang, C.B. Travel Time Prediction Model for Urban Road Network based on Multi-source Data. Procedia - Social and Behavioral Sciences, 2014,138.
14. Tang-Hsien Chang; Albert Y. Chen. Freeway Travel Time Prediction Based on Seamless Spatio-temporal Data Fusion: Case Study of the Freeway in Taiwan. Transportation Research Procedia,2016,17.
15. Fei X; Lu C C. A Bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. Transp. Res. Part C, 2011, 19(6):1306-1318.
16. Zhan, X.; Ukkusuri, S. V. A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data. Autom.Constr,2016, 72, 237-246.
17. Wosyka,J; Přibyl, P. Real-time travel time estimation on highways using loop detector data and license plate recognition. Elektro. IEEE, 2012:391-394.
18. Innamaa, S. Short-Term Prediction of Travel Time using Neural Networks on an Interurban Highway. Transportation, 2005, 32(6):649-669.
19. Lint, J.W.C.V; Hoogendoorn, S P. Accurate freeway travel time prediction with state-space neural networks under missing data. Transp. Res. Part C, 2005, 13(5):347-369.
20. Wu, Chun-Hsin; Jan-Ming Ho. Travel-time prediction with support vector regression. IEEE.Trans.Intell.Transp.Systems.5.4 ,2004: 276-281.
21. Vanajakshi, L; Rilett, L R. Support Vector Machine Technique for the Short Term Prediction of Travel Time. Intelligent Vehicles Symposium. IEEE, 2007:600-605.
22. Mendes-Moreira, João. Comparing state-of-the-art regression methods for long term travel time prediction. Intell. Data. Anal. 16.3 (2012): 427-449.
23. Wang, X.;Chen, X. H.; Yang, X. M. Short term prediction of expressway travel time based on k nearest neighbor algorithm.Chin.J.Highw.Transport,2015,28(1), 102-111.
24. Li, S.; Yuan, Z. C.; Wang, C. Optimization of support vector machine parameters based on group intelligence algorithm. CAAI TIS,2018,13(01):70-84.

25. Wang, Q.; Liu, Z.; Peng, Z.A PSO-SVM Model for short-term travel time prediction based on Bluetooth Technology. J. Harbin. Inst. Technol, 2015,22(3), 7-14.
26. Yang, Z. S. Study on the Synthetic Link Travel Time Prediction Model of Key Theory of ITS.J.Traff.Transp.Eng,2001,01:65-67.