Online Traffic Congestion Prediction Based on Random Forest

Xiao Han[1, a], Yijie Shi[2, b]

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China;

[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[a]xhan2015@126.com, [b]yijieshi2000@bupt.edu.cn

Keywords: Intelligent Transportation System, Traffic Congestion, Online Prediction, Random Forest.

Abstract. In recent years, distinction and prediction of urban traffic congestion has become an important part of Intelligent Transportation System (ITS), hence attracting more and more attentions. Road congestion can be predicted by analyzing traffic flow data collected by various data acquisition equipment primarily. However, existing methods not only need to store large amount of historical information, but has not enough suitability for large-scaled and changing traffic flows. Therefore, an online prediction method based on Random Forest (RF) is put forward in this paper and the prediction on congestions is made by real-time data instead of digging the historical data. Simulation and experiment results show that the design presented in this paper improves accuracy of predictions and it has a certain use value.

Introduction

Traffic congestion had restricted the development of urbanization seriously and given rise to economic losses and environmental pollutions. Rapid and accurate identification of traffic condition and predictions on future congestion degree can help traffic administration to enforce reasonable and effective management measures to ease the traffic pressure.

At present, traffic congestion predictions are mainly made by forecasting short-time traffic flow pattern. Traffic flow data can be acquired by all kinds of sensors installed on roads and floating car technology; major parameters consist of speed, density, volumn, occupancy, traffic headway and waiting duration, etc.. In essence, traffic flow data are temporal sequence. Existing in form of data flow, it is featured with large quantities, continuity and real time. Their storage is able to consume a lot of resources.

Currently, existing research methods of short-term traffic can be divided into models based on statistical theory and artificial intelligence.

Statistics-based models include Autoregressive Moving Average Model (ARMA) [1], Time Sequence Model [2] and Kalman Filtering Model [3], etc.. Although simple, they are unable to reflect the uncertainty and nonlinearity of traffic state, because they are based on the linear basis.As a result, their prediction effects are poor.

Models based on artificial intelligence include Neural Network Predictive Model [4,5], Support Vector Machine Regression [6] and Nonparametric Regression Prediction Model [7] which belong to traditional batch learning model. With sophisticated internal structures, they usually rely on a large number of historical data and adapt to concept drift in data flow with quite a latency.

According to shortcomings of models mentioned above, small-scaled real-time data are utilized to perform on-line learning; a sliding window technique is introduced to adapt sudden changes of traffic flow data and the ensemble learning method of RF is also applied to improve both accuracy and stability of classification.

Pre-knowledge

Random Forest Algorithm.

The Random Forest[8] algorithm is developed on the basis of Bagging (Bootstrap Aggregation)

[9] algorithm, an ensemble algorithm, which can improve the prediction accuracy of unstable learning algorithms. With regard to instability of algorithms such as decision tree and neural network, etc., it means that if a small change happens to a training set, prominent influences can be generated on learning results.

Based on Bootstrap sampling of Bagging algorithm, RF selects part of feature attributes to train a CART [10] classifier. In addition to classification accuracy improvement and avoidance of over-fit phenomenon of decision tree, RF doesn't need pruning in the process of training. As a result, not only is the training speed fast, but parallelization can be realized easily and hence modeling efficiency is enhanced. Although RF has been widely applied in fields such as medical science and economics, etc., it is the first time to apply it into traffic congestion prediction.

Decision Tree Algorithm (CART).

The RF algorithm utilizes Classification and Regression Tree (CART) as a weak classifier. CART is a type of decision tree algorithms and featured with many advantages including simple structure, rapid training speed, visual results, being able to overcome data missing, etc.. Moreover, as it has few parameters of itself, there is no need for domain experts to pre-set parameter values before using. CART algorithm is also very suitable to process training set used by texts, because it can deal with both discrete and successive feature attributes. GINI Index is chosen by CART as its objective function when to split a tree node. For a given data set D, which can be classified into K types, and the number of classification k in the data set is |Ck|; then, the GINI Index can be expressed as follows.

$$GINI(D)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k}^{K}p_k^2=1-\sum_{k}^{K}(\frac{|C_k|}{|D|})^2. \quad (1)$$

GINI Index denotes uncertainty of a set D. the larger it is, more uncertain the sample will be.

CART algorithm adopts the technology of Binary Recursive Partitioning to divide a present sample set into two subsample sets. As to the recursive procedure creation for classification tree, the feature attribute with minimum GINI information gain is selected by CART each time as a tree node to classify decision tree. If the Set D is divided into D1 and D2 according to a feature attribute, its GINI information gain can be denoted as,

$$GINI\_gain(D)=\frac{|D_1|}{|D|}GINI(D_1)+\frac{|D_2|}{|D|}GINI(D_2). \quad (2)$$

Online Ensemble Prediction Model

Selection of Sample Attributes.

Attributes of samples used for classification modeling include the properties of traffic flow and the environmental factors.

1) Traffic Flow Parameters

Different traffic flow parameters have different capabilities to reflect changes in the traffic flow. When congestion occurs, speed parameter, followed by occupancy and traffic volume, is most sensitive to reflect such changes[11]. Therefore, the speed parameter can be selected to predict congestions, or, we choose the speed parameter as main parameter while occupancy and traffic volume as auxiliary ones to conduct such predictions so that detection rate of congestions can be improved. In this paper, speed, time occupancy ratio and traffic volume are selected for congestion prediction.

2) Environmental Factors

Environmental factors involved in traffic state are divided into four aspects including weather, time frame, special event and vacation; besides, they are also classified into five levels in line with influence degree[12]. Through our certification, factors of weather and time frame which exert greater influences on traffic flow are reserved in this paper and another factor named as road segment is introduced on this basis. The factor of road segment means that whether special landmarks such as school, hospital and mall, etc. exist around the segment where data are collected

and road segments are divided into 3 levels according to their impacts on traffic flow. To sum up, weather, vacation and road segment are acted a three environmental factor attributes of training sample.

Weather. Weather effects have five levels by means of early warning signal colors of weather forecast. To be specific, if no early warning signal appears, it can be denoted by 0.1; by contrast, 0.3 is represented by blue signal, 0.5 represented by yellow signal, 0.7 by orange signal and 0.9 by red signal.

Time Frame. Rush hours are 7:30-9:00 and 16:30-18:30, represented by 0.9; on and off duty at noon is 11:30-15:00, represented by 0.7; hours from 9:00 to 10:30 are represented by 0.5, from 20:30 to 22:00 are represented by 0.3; while other time intervals are represented by 0.1.

Road Segment. The case that no particular landmark exists around the segment can be denoted by 0.1, while 0.5 for a few such landmarks and 0.9 for many such landmarks.

Online Random Forest Prediction Model Based on Slide Window.

In this paper, traffic flow data are analyzed as data stream to create and update the prediction model dynamically as well as to adapt to continuously changing traffic conditions in real time. A Slide Window Random Forest (SW-RF) model used as the online prediction model of traffic congestion is constructed by combining RF algorithm with slide window technology based on spatio-temporal data of traffic flow.

In order to conform to requirements of data stream, Slide Window is introduced as shown in Figure 1 to maintain a small sample set constituted by real-time data within a slide window of fixed size. Regarding the small sample set as a training set, we select a decision tree of unstable classification algorithm to train weak classifier so as to highlight changes of the data stream. As long as the window slides into another lattice, a new sample will be introduced, while an old one will be abandoned; then, samples within such a slide window can be used to construct prediction model. Different from ordinary online learning algorithms which utilize every arrived sample to update model, this algorithm sets a threshold K; when the number of new samples accumulates to K, model renewal can be triggered and then re-count the new-come sample from 0. As for model updating, n decision trees are generated through the RF algorithm in the first place; then an ultimate classification prediction is produced by means of voting. The details of the whole process are in algorithm 1.

Fig. 1 Slide Window

Algorithm 1 Slide Window Random Forest

Input: Data set D, size of Slide Window W, threshold K, size of decision trees n, size of features m, the instance to be predicted *X*.

Output: Predicted result *Y.*

1) Initialize Slide Window, begin to move window
2) **If** size of new instance =*K*
3) **For** i=1….n
   Bootstrap sampling from Slide Window, and get training set Di (|Di| = W)
4) Select m features in random
5) Build decision tree Ti with CART algorithm
6) **End for**
7) **For** i = 1….n
8) Predict *X* by Ti, get *Yi*

9) **End for**
10) Vote for the ultimate prediction *Y*
11) **End if**

Experimental Evaluation

We used Weka 3.6 as a modeling tool to conduct experimental evaluation for algorithm. Weka is a data mining tool based on Java language developed by the University of Waikato in New Zealand, fits for small-scaled data mining. During experiment, we adopted classification precision and Mean Square Percent Error (RMSE) which are obtained by 10-fold cross validation method to evaluate classification effects.

Experimental Data.

Those experimental data are taken from measured data and simulated data of simulation software. The former includes loop road microwave detecting data (sampling period is 15s) of a road segment in Beijing City within a week in March 2011 and corresponding historical records of weather forecast; while the latter are also data of one week and obtained by simulating measured road segment based on Vissim software. Concerning simulation, traffic incidents such as road accidents, traffic control, etc. were set by us to generate traffic flow data with sudden changes to make up for the inadequacy of measured data.

After pre-processing, samples for model training totally consist of 9 attributes such as speeds, occupancies and volume acquired by two detectors at the upstream and the downstream of a road segment as well as corresponding weather, time frame and road segment. We lable a sample as three categories according to the Speed: Smooth traffic, speed>=30km/h; general congestion, 30km/h< speed <=10km/h; and severe congestion, speed <10km/h. Sample instances are shown in Table 1.

Table 1 Sample Instances

Time		t1	t2	t3	t4	t5	t6	t7
Parameter								
Speed (km/h)	**upstream**	70	76	53	45	24	19	8
	downstream	65	74	60	50	24	21	9
Occupancy (%)	**upstream**	8	10	15	12	23	24	35
	downstream	10	9	11	16	20	22	32
Volume (Veh/min)	**upstream**	53	55	48	45	32	21	12
	downstream	51	48	49	42	36	25	14
Weather		0.1	0.1	0.1	0.1	0.1	0.1	0.1
Time Frame		0.7	0.5	0.5	0.5	0.5	0.5	0.5
Road Segment		0.1	0.1	0.1	0.1	0.1	0.1	0.1
State		smooth	smooth	smooth	smooth	general congestion	general congestion	severe congestion

Parameter Setting.

The SW-RF algorithm put forward in this paper is required to firstly determine the size of slide window W and threshold K. Considering that the time span is generally 30min at most according to short-term traffic flow prediction, the value of W is set as 120 at the time of computational experiment (calculated by an acquisition interval of 15s) in order to meet the requirements of algorithm to small sample set as well as ensure the real-time performance of data. That is, the slide window maintains sample data of the nearest 30min. In the meanwhile, K is set as 20 and it means that the prediction is carried out every 5 minutes; hence, the proportion taken by new samples in training sample set is guaranteed and the prediction ability of training model on sudden changes is enhanced as well.

With respect to RF algorithm, both n (the number of trees in the Forest) and m (the number of feature attributes selected at the time of constructing trees) need to be set. When the number of trees

(n) is large enough, the precision of integrated model tends to be stable[13]; however, the larger n is, the larger the overhead that model construction needs will be. During experiment, n is chosen to be 20, 40, 60, 80 100, 150 to construct RF and precision of the model is shown in Fig. 2. It can be seen that this precision becomes tend to be stable when n is larger than 80. As a result, 80 is the number of trees in integrated model.

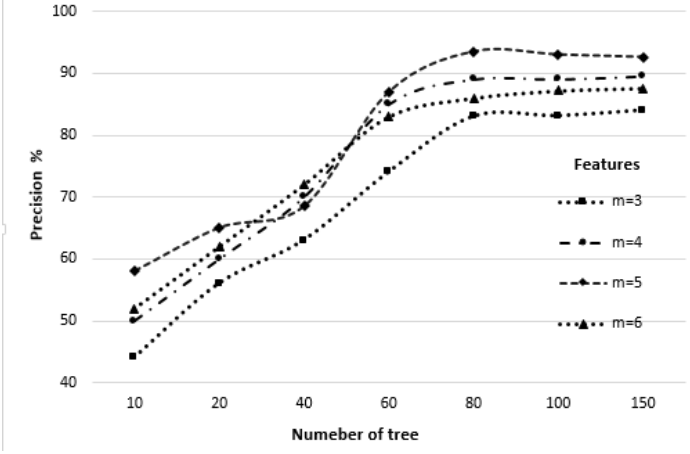

Fig. 2 Precision With Different Number of Trees

Parameter m is the number of candidate features used for node splitting when RF is employed to construct a single decision tree. If the total number of features is M, then feature attributes selected randomly can be $1/2sqrt(M)$, $sqrt(M)$ and $2sqrt(M)$[10]. Therefore, on the premise that n is 80 fixed, m is selected to be 2, 3, 4, 5, 6, 7 to observe changes of classification error. Based on results presented in Fig. 3, we can see that when m is equal to a value ranging from 2 to 5, error rate can be lowered gradually; by contrast, if m exceeds 5, it goes up. Hence, m is chosen to be 5 as the number of feature attributes in this paper.

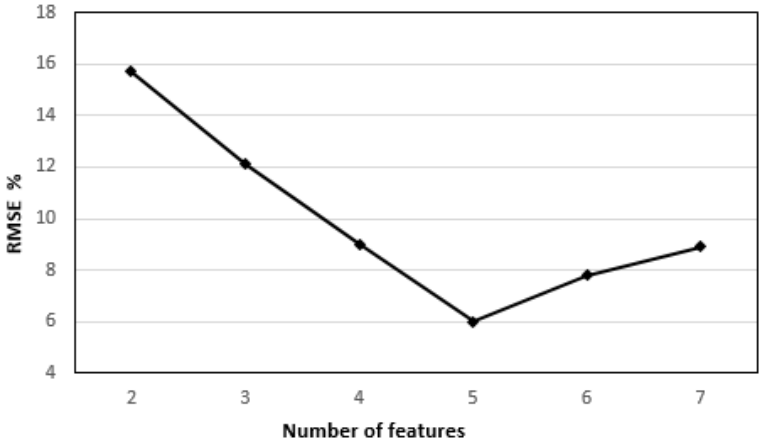

Fig. 3 RMSE With Different Number of Features

Algorithm Comparisons.

A comparative analysis is performed in this paper through typical classification algorithms including Nnaïve Bayes(NB), CART, Back Propagation Neural Network(BP-NN) and SW-RF. For BP-NN, the number of hidden layer neurons and learning algebra need to be set. Through experiment, when these parameters are 8 and 10 respectively, BP-NN shows the best classification effects. Therefore, the number of such neurons is 8 and the learning algebra is selected to be 10 in this paper as its comparison algorithm.

In the first place, measured traffic flow data are chosen to test four models by treating samples of every day as the training set and compare their precisions; afterwards, simulated data are adopted to compare those four algorithms in the same way. Results are shown in Fig. 4 and Fig. 5.

According to these figures, we can see that SW-RF put forward in this paper has the highest precision the average of which is 90% and above; BP-NN takes the second place followed by

CART and NB whose fluctuations are comparatively big. Different from CART decision tree, the precision of NB depends on prior probability of sample so that the precision of CART is higher than that of NB if sample space is lesser. With regard to a single CART decision tree, over-fitting phenomenon can be generated easily in case that no pruning is operated; therefore, classification precision is reduced.

By comparing two result figures, it is found that when the number of sudden changes in data stream increases, precisions of those four algorithms all decline to different extents. Among then, the decline of CART and NB is most obvious, while SW-RF presented in this paper has a minimum reduction and a relatively smooth and steady curve, indicating that the SW-RF algorithm has a good self-adaptation ability to sudden changes of data stream.

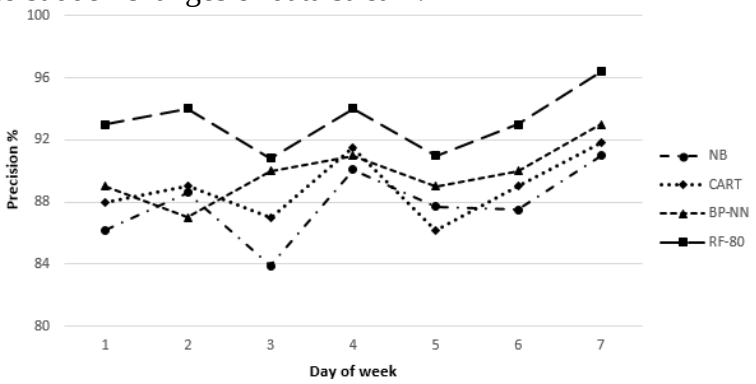

Fig 4. Algorithm Contrast With Measured Data

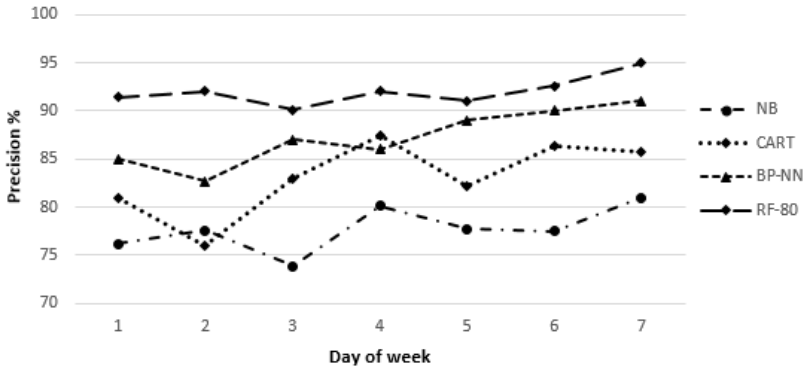

Fig 5. Algorithm Contrast With Stimulated Data

Conclusions

Combined with characteristics of traffic flow data, a traffic congestion prediction algorithm which integrates slide window with RF algorithm is put forward and a real-time prediction system based on Storm framework is also designed in this paper. Experiment shows that the algorithm presented in this paper is clearly more precise and more stable comparing with other single classifier algorithms and it is more suitable for changeable traffic congestion predictions. Next, our researches will focus on dynamically adjusting the size of slide window W and threshold K in accordance with prediction results to further improve the performance of such an algorithm and expand the scale of the experiment.

Acknowledgements

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

References

[1] G Yu, C Zhang, et al. Switching ARIMA model based forecasting for traffic flow. IEEE International Conference on Acoustics, Speech, Signal Processing. Quebec: IEEE, 2004:429-432.

[2] B Ghosh, B Basu, M O'Mahony, et al. Bayesian time-series model for short-term traffic flow forecasting. Journal of Transportation Engineering, 2007, 133(3):180-189.

[3] ZS Yang, Z Zhu, et al. A real-time traffic volume predict ion model based on the kalman filtering theory[J]. China Journal of Highway and Transport, 1999,12(3):63-67.

[4] MA Jun, LI Xiao-Dong, Y Meng, et al.Research of Urban Traffic Flow Forecasting Based on Neural Network. Acta Electronica Sinica, 2009, 37(5):1092-1094.

[5] VJ Hodge, R Krishnan, J Austin, J Polak, T Jackson, et al. Short-term prediction of traffic flow using a binary neural. Neural Compute & Application, 2014, 25(7-8):1639-1655.

[6] G Fu, GQ Han, F Lu, ZX Xu, et al. Short-Term Traffic Flow Forecasting Model Based on Support Vector. Journal of South China University of Technology, 2013, 41(9):71-76.

[7] T Zhang, X Chen, MP Xie, YJ Zhang, et al. K-NN based nonparametric regression method for short-term traffic flow forecasting. Journal of Systems Engineering-Theory & Practice, 2010,30(2):376-384.

[8] L Breiman. Bagging predictors. Machine Learning. 1996, 24(2):123-140.

[9] L Breiman. Random Forests, Machine Learning.2001, 45, 5-32.

[10] L Breiman, JH Friedman, R Olshen, CJ Stone, et al. Classification and regression trees. Biometrics, 1984, 1(1):14–23.

[11] XW Tan. Congestion Prediction Based on Classifiers Combination Technology. Maeter Degree, Fuzhou University, China, 2005.

[12] LI Chun-Ying, ZK Tang, YD Cao,et al. Study on traffic congestion prediction model of multiple classifier combination. Computer Engineering and Design, 2010, 31(23):5088-5091.

[13] XU Wenhua, Z Wei, et al. An Online Short-term Traffic Flow Prediction Model Based on Data Stream Ensemble Learning. Journal of Transport Information and Safety, 2014,31(23):14-19.