Traffic Congestion Classification Using Cellular Phone Activity Data

Qing Li, Shuang Wu
Department of Computer Science,
Department of Civil and Environment Engineering

Abstract

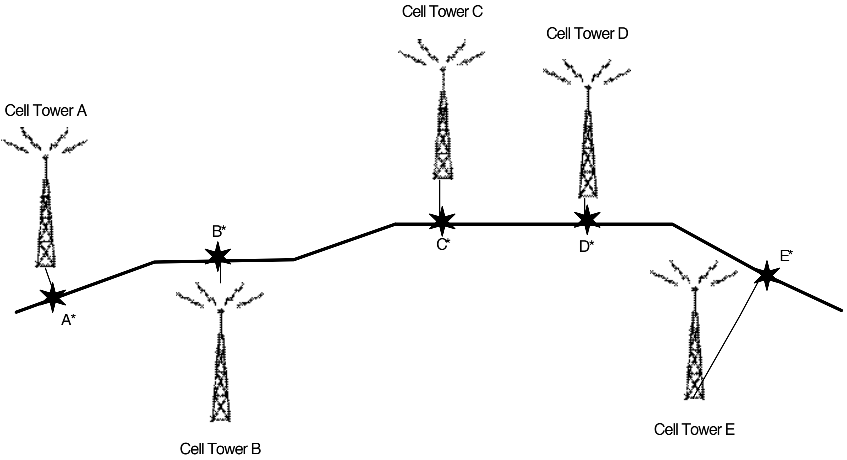
Full cellphone activity (FCA) data refers to the full records of real-time cellphone communication signals generated by cellular towers while maintaining cellular services both on- and off-call. Such signal data may be related to phone calls, texting, web browsing, video and audio streaming, Location-Based Service (LBS) and other cellphone activities. This paper compares different machine learning methods using the full CA data for traffic state detection. Ten features extracted from raw cellular data were used as input.

In this paper, firstly, precisions using support vector machine with different kernel functions and neural network with different hidden units and output functions are compared by 3-fold cross validation. Next, bagging (Bootstrap Aggregation) algorithm is introduced with the SVM and Neural Network chosen by first step, and confusion matrix among three different congestion levels is used to evaluate algorithms. In the last step, we also investigate how predictive accuracy varies as a function of training-set size. Experiments show that ensemble methods outperform single classifiers, and Bagging-SVM with RBF kernel outperforms all other classifiers for this classification task.

Keywords: cellular probe; cellular activity data; freeway congestion detection; SVM; Neural Network; Bagging

1. Introduction

Cellular probe technologies use cellphones as traffic probes to collect and estimate dynamic traffic information. Over the past fifteen years, the technologies have evolved and improved to become reliable technologies complementing the coverage and resolution of existing traffics detection methods due to the high ownership, activity, and widespread of cellphone users [2]. Depending on whether or not the active positioning is triggered by cellular system or a cellphone user, the existing cellular probe methods can be classified into the handset-based and the network-based [3]. The handset-based approach is an active way to collect location information, which is able to provide coordinates of the cell phone, such as triangulation and GPS [4].



Existing network-based cellular probe methods are mostly based on cellphone handoff (HO) data. HO data is generated when the cell phone crossing the coverage boundary between two cellular towers during an active phone call [2]. Two consecutive HO records from a cellphone within a car can provide a travel time sample for the road segment between the two HO locations. Based on a pre-calibrated locations of the HO signal on a roadway link, the travel distance can also be determined. However, the sample size using HO data is insufficient for accurate and timely traffic state estimation.

In this study, we explores the full cellphone activity records and generates ten features from raw cellphone data. To determine which classification algorithm is better for detecting traffic congestion, firstly, precisions using support vector machine with different kernel functions and neural network with different hidden units and output functions are compared by 3-fold cross validation. Next, we introduce bagging (Bootstrap Aggregation) algorithm with the SVM and Neural Network chosen by first step, then compare which algorithm provides the best predictive accuracy on the task and determine if ensembles can lead to better accuracy. In the last step, we also investigate how predictive accuracy varies as a function of training-set size.

2. Data Collection

A freeway segment in southeastern China was selected as the test bed. This freeway segment is part of the main corridor connecting two big cities, Shanghai and Nanjing. This 250 km long freeway corridor has four lanes in each direction, and the average daily traffic volume is above 210,000 vehicles, of which 15% are heavy vehicles. The FCA data used in this study was collected from a major cellphone carrier in China.

Fixed-point detector data (loop and microwave detectors) was provided by the local traffic management center (TMC) and used to label the data. We conducted a 12-day test using the data from the one fixed-location detection site. This test period is from September 28th, 2014 to October 9th, 2014, covering the entire China National Day holiday and normal workdays before and after it.

2.1 Feature Description and Data Normalization

Based on the raw cellular data which contains more than one billion records, ten features were generated [5].

TABLE 1 Feature description table

No.	Feature	Unit
1	Activity	
2	Pseudo Speed	Mile per hour
3	MSID Count	
4	Mean of Previous 15 Activities	
5	Standard Deviation of Previous 15 Activities	
6	Mean of Previous 15 Pseudo Speed	Mile per hour
7	Standard Deviation of Previous 15 Pseudo Speed	Mile per hour
8	Mean of Previous 15 MSID Count	

9	Standard Deviation of Previous 15 MSID Count	
10	Average Stay Time	Hour

Due to the range of each feature is not identical, however, all features should be of equal importance, data normalization is applied to all input continuous features, in order to reduce the network estimation error caused by the different magnitudes. After generating all the features listed above, it is normalized by following equation.

$$z_{i,j} = \frac{2x_{i,j} - max(x_{,j}) - min(x_{,j})}{max(x_{,j}) - min(x_{,j})}$$

Where $x_{i,j}$ is the original value of jth feature of ith instance, $z_{i,j}$ is the normalized value of jth feature of ith instance and its range is [-1, 1].

2.2 Congestion Level (Class)

The congestion level is labeled according to the average speed provided by loop detector in the study link. Table 2 shows thresholds for congestion levels.

TABLE 2 Ground truth speed threshold

Traffic Speed (mile/hr)	Congestion Level
>= 40	0
[30, 40]	1
< 30	2

2.3 Training, Testing Set, and Cross Validation

Under the premise of ensuring the stability of the rate for each label in each dataset, stratified sampling method is used to ensure that class proportions are maintained in each selected set. Total dataset contains 1713 samples with 1369 samples in training set and 344 samples in testing set. In the training dataset, there are 1162 samples labeled 0, 22 samples labeled 1, and 185 samples labeled 2. 3-fold cross validation is applied to determine parameters for SVM and Neural Network.

3. Basic Classification Methods and Experiments

3.1 Multi-Label Neural Network

Different traditional binary classification problem, this study has three classes. Thus, for this 3-ary classification problem, 3 output units is adopted. And the predictive label for ith instance is based on the following equation:

$$p_i = \arg\max_j o_{i,j}$$

Where p_i is the predictive label for ith instance, $o_{i,j}$ is jth output for ith instance.

Use two-layer neural network with one hidden layer and one output layer. Number of hidden units (2, 5, and 10) and transfer functions (pure linear, logsig, and tansig) are used as parameters for determination. Batch training method is adopted. The neural network structure used in this study is as follows.

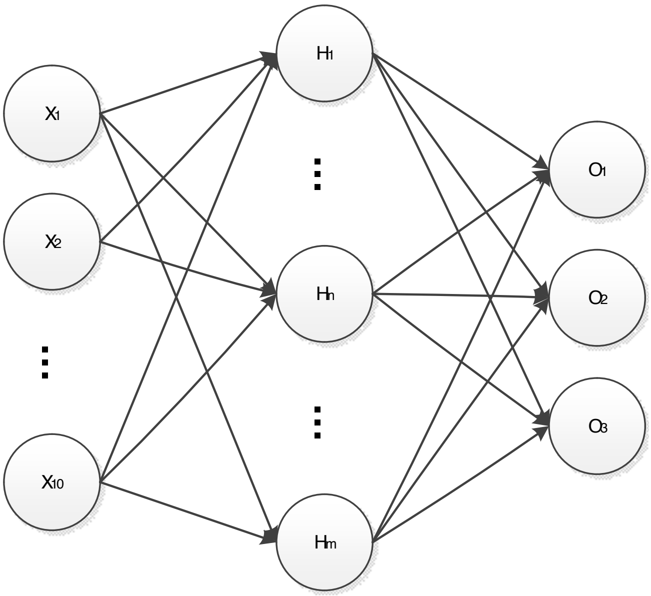


Fig. 1. Neural Network Structure

3.2 Parameter Selections for Neural Network

In this part, 3-fold cross validation is used to determine the accuracy under different parameters of neural network. The result is shown in following table.

TABLE 3 Accuracy under different parameters of NN by 3-fold cross validation

Neural Nets	Class 0 Accuracy	Class 1 Accuracy	Class 2 Accuracy	Total Accuracy
2, linear	99.6%	0.0%	80.5%	95.4%
5, linear	99.5%	0.0%	78.9%	95.1%
10, linear	99.7%	0.0%	79.5%	95.3%
2, logsig	99.7%	0.0%	30.8%	88.8%
5, logsig	98.5%	0.0%	69.7%	93.1%
10, logsig	99.3%	0.0%	83.2%	95.5%
2, tansig	**99.0%**	**0.0%**	**87.0%**	**95.8%**
5, tansig	98.2%	0.0%	85.4%	94.9%
10, tansig	98.6%	0.0%	87.6%	95.5%

Based on the results, 2 is set for number of hidden units and tansig is used as transfer function.

3.3 Multi-Label SVM

Compared with other classifiers, SVM could achieve optimal class boundaries by finding the maximum distance between classes. The equation is as follows:

$$\begin{aligned} &\underset{w,b}{minimize}\,\frac{1}{2}\|w\|_2^2 \\ &s.t.\ y^{(i)}(w^T\phi(x^{(i)})+b)\geq 1 \\ &for\ i=1,\ldots,\mathrm{m} \end{aligned}$$

To implement 3-ary classifier, one-vs-the-rest is adopted which consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes.

3.4 Kernel Selections for SVM

In this part, 3-fold cross validation is used to determine the accuracy under different kernel functions of SVM. The result is shown in following table.

TABLE 4 Accuracy under different kernels of SVM by 3-fold cross validation

SVM	Class 0 Accuracy	Class 1 Accuracy	Class 2 Accuracy	Total Accuracy
Linear	93.7%	72.7%	8.1%	81.8%
Quadratic	95.8%	18.2%	72.4%	91.4%
Polynomial-3	97.3%	9.1%	76.8%	93.1%
RBF	**96.6%**	**22.7%**	**80.0%**	**93.2%**

Based on the results, RBF kernel is selected.

4. Bagging Methods and Experiments

4.1 Bagging Algorithm

In this study, bagging [6] which chooses different subsamples of the training set is adopted for comparison. Pseudo code for bagging algorithm is as follows.

TABLE 5 Bagging Algorithm

```
Input: Learner L (SVM or Neural Network), training set D (data size: m ), test instance x
Output: Prediction for x in testing set.
Learning:
    For i from 1 to T (20 in this study) do
        Di ← m instances randomly drawn with replacement from D
        Hi ← model learned using L on Di
Classification:
    Y ← plurality_vote( $h_1(x), h_2(x), \ldots, h_T(x)$)
```

4.2 Experiments

Use whole training set to train the models, we compare confusion matrix and precision under 4 models: NN, Bagging-NN, SVM, Bagging-SVM. The parameters are selected from Part 3. Confusion matrix is as follows.

TABLE 5 Confusion matrix for different machine learning methods

NN	Actual 0	Actual 1	Actual 2
Pred 0	285	5	1
Pred 1	1	1	4
Pred 2	3	5	39

Bagging NN	Actual 0	Actual 1	Actual 2
Pred 0	287	0	4
Pred 1	2	0	4
Pred 2	4	0	43

SVM	Actual 0	Actual 1	Actual 2
Pred 0	285	5	1
Pred 1	1	1	4
Pred 2	3	5	39

Bagging SVM	Actual 0	Actual 1	Actual 2
Pred 0	287	0	4
Pred 1	2	2	2
Pred 2	2	2	43

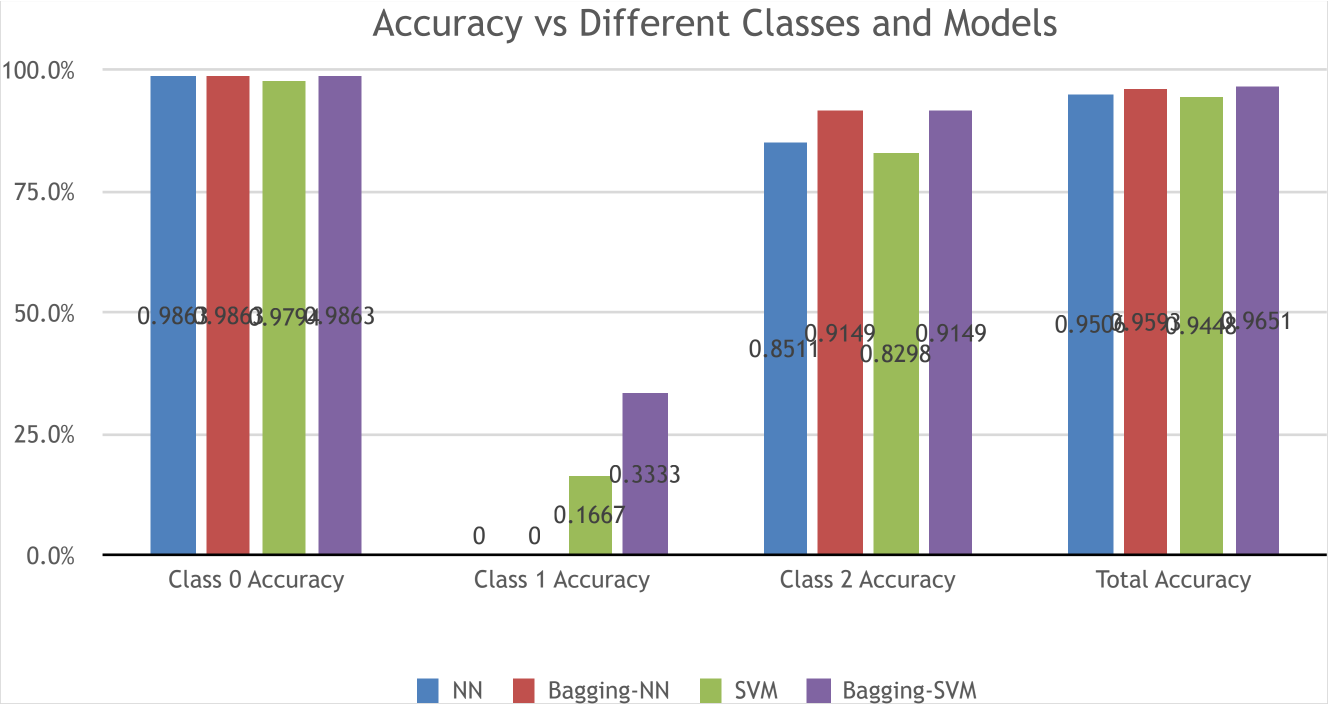
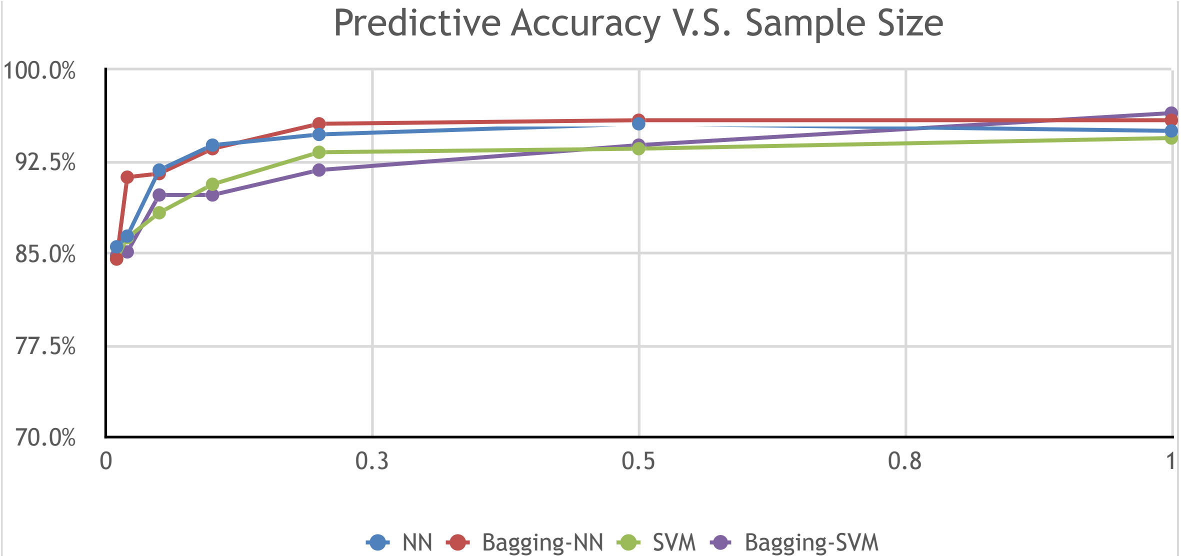


Fig. 2. Accuracy vs Different Classes and Models

From the confusion matrix and figure above, we can see that the total accuracy among four models don't differ a lot. This is because that most models chosen can distinguish class 0 and class 2, which have most samples. However, the accuracy of class 1 is quite low. One reason is that class 1 has fewer samples compared with other classes. Moreover, since congestion level 1 is a transition level between level 0 and level 2, it is hard to train a classifier to accurately classify this class. Even though, we can still see that Bagging-SVM is better for this classification task among four models. And ensemble methods actually will improve performance.

4.3 Predictive Accuracy and Sample Size

In this part, we use total accuracy as predictive accuracy. The relationship between predictive accuracy and sample size under different models is as follows. For each sample size, we run ten times and mean accuracy is used.



From the figure above, neural networks will get better accuracy when sample size is smaller. However, when exploring the accuracy in different classes, SVM would achieve better accuracy in class 2. Neural network outperforms SVM in class 0, which has more samples.

5. Conclusion and Future Work

This paper compares different machine learning methods using the full CA data for traffic state detection. Firstly, precisions using support vector machine with different kernel functions and neural network with different hidden units and output functions are compared by 3-fold cross validation. Next, bagging (Bootstrap Aggregation) algorithm is introduced with the SVM and Neural Network chosen by first step, and confusion matrix among three different congestion levels is used to evaluate algorithms. In the last step, we also investigate how predictive accuracy varies as a function of training-set size. Experiments show that ensemble methods outperforms single classifier, and Bagging-SVM with RBF kernel outperforms all other classifiers for this classification task.

Under the existing conditions, the size of the training set is not sufficiently large especially for class 1. This limitation affects the classification accuracy in this class. Moreover, since class 1 is a transition level between level 0 and level 2. Further work will investigate and improve performance by changing vote function in ensemble methods.

References

[1] Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell, eds.*Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[2] Ran, B., Jin, P.J., Cebelak, M.K., Cheng, Y., 2014. Cellular Probe Technologies Moving Forward: The Current Trends and Perspectives on 3g, 4g, and Smartphone Applications, *the 21st ITS World Congress*, , Detroit, Michigan.

[3] Yang, F., Cheng, Y., Jin, J., Xia, J., Yang, D., Ran, B., 2012. Wireless Communication Simulation Model for Traffic Monitoring Systems Based on Dynamic Cellular Handoffs. *Transportation Research Record: Journal of the Transportation Research Board* 2291(-1), 26-34.

[4] Herrera, J.C., Work, D.B., Herring, R., Ban, X., Jacobson, Q., Bayen, A.M., 2010. Evaluation of Traffic Data Obtained Via Gps-Enabled Mobile Phones: The Mobile Century Field Experiment. *Transportation Research Part C: Emerging Technologies* 18(4), 568-583.

[5] Ran, B., Cheng, Y., Jin, J.J., Ding, F., Li, Q., A Feature Based Approach to Large-Scale Freeway Congestion Detection Using Full Cellular Activity Data. Manuscript sumbitted for publication.

[6] Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.