Transportation Research Part C 26 (2013) 160–169

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

On feature selection for traffic congestion prediction

Su Yang *

College of Computer Science and Technology, Fudan University, Shanghai 201203, China

ARTICLE INFO

Article history:
Received 2 May 2012
Received in revised form 7 June 2012
Accepted 28 August 2012

Keywords:
Traffic congestion prediction
Pattern classification
Feature selection
Feature ranking

ABSTRACT

Traffic congestion prediction plays an important role in route guidance and traffic management. We formulate it as a binary classification problem. Through extensive experiments with real-world data, we found that a large number of sensors, usually over 100, are relevant to the prediction task at one sensor, which means wide area correlation and high dimensionality of the data. This paper investigates the first time into the feature selection problem for traffic congestion prediction. By applying feature selection, the data dimensionality can be reduced remarkably while the performance remains the same. Besides, a new traffic jam probability scoring method is proposed to solve the high-dimensional computation into many one-dimensional probabilities and its combination.

1. Introduction

Traffic congestion prediction plays an important role in intelligent transportation. For instance, the GPS navigation products equipped with traffic congestion prediction module can make more practical routing decision. Besides, the ability of traffic congestion prediction allows the traffic management department to do better management. There are three prediction problems regarding traffic congestions: Traffic volume prediction, traffic congestion prediction, and travel time prediction. This study is focused on traffic congestion prediction, which is formulated as a binary decision problem on whether the traffic volume will exceed a watching threshold shortly.

So far, traffic congestion prediction has been receiving much attention in the context of civil engineering as well as the information technology. Early researches are focused on single site prediction based on one-dimensional traffic time series such as the ARIMA model (Williams and Hoel, 2003) and the nearest neighbor method (Smith et al., 2002). Recently, the trend has been shifted to prediction based on spatial temporal correlations between traffic flows (Kanoh et al., 2005; Ando et al., 2006; Min and Wynter, 2011; Hu et al., 2009; Ghosh et al., 2009; Romaszko, 2010; He et al., 2010; Ben-Akiva et al., 2012), for instance, the vector ARMA model (Chandra and Al-Deek, 2009) incorporating both spatial and temporal correlations (Min and Wynter, 2011), and the spatial econometrics models focused on congestion propagation over adjacent links (Hu et al., 2009). The core of the existing methods is: They try to predict traffic congestions at one site based on the spatially and temporally correlated information from the sensors distributed on nearby roads, where the number of such sensors contributing to the prediction is referred to as data dimensionality. To the best of our knowledge, for almost all the existing methods, the data dimensionality does not exceed 100. For example, only 15 neighbors are considered in Min and Wynter (2011) and 10 sites in Ghosh et al. (2009). The reasons to limit the consideration in such a narrow field are two folds: (1) The high computational cost induced by higher dimensionality cannot be afforded by such methods. (2) Such methods are cause–effect based, that is, they trace the traffic congestions propagating along nearby roads to foresee the formation of new congestions (Hu et al., 2009). However, our experimental results contradict the widely accepted assumptions, that is, the signatures correlated to the traffic congestions at one site should exist in a very large scale, from more than 100 sensors in general.

* Tel./fax: +86 21 51355520.
E-mail address: suyang@fudan.edu.cn

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.08.005

In another word, the information obtained from more than 100 sensors should be useful in predicting the traffic jams to appear at one site. Through the experiments with the real-world data of more than 4000 loop detectors located around the Twin Cities Metro freeways from 1 January to 22 September 2010, we found that the number of the relevant sensors to reach high-performance traffic congestion prediction at one sensor is over 100 in most cases. We explain such phenomenon as follows: According to the findings turned out from the simulation in Mazloumian et al. (2010), unbalanced vehicle distribution across the network of interest does have a remarkable correlation to traffic congestion from a global point of view at the whole network level. This accounts for why the seemingly irrelevant traffic patterns even far away can act as signals to indicate the possibility of traffic congestion occurrence at the watched site. This is straightforward in that for a specific network, if the traffic loads at considerable links are light, the other links must be undergoing heavy traffic loads due to the unbalanced traffic volume distribution. Accordingly, the global traffic patterns in a wide area can function to indicate possible congestions at a watched site. This gives rise to a new problem, that is, what is the optimal dimensionality of the input data and how to evaluate the significance of every sensor for traffic congestion prediction at a given sensor, which can be formulated as a feature selection/feature ranking problem in terms of pattern recognition. To the best of our knowledge, this is the first study investigating into the feature selection/feature ranking problem for traffic congestion prediction. In the context of pattern recognition, the goal of feature selection/feature ranking is to rank the quality of every attribute and identify the high-quality ones that contribute to improve the classification performance at most. By means of feature ranking/feature selection, irrelevant variables can be rejected and only the highly contributive features are preserved such that the classification performance can in general be improved while the data dimensionality is reduced. In terms of traffic congestion prediction, feature ranking/feature selection functions to identify the most significant features/sensors relevant to traffic jams at a given sensor so as to build a predictor with only relevant sensors data as input, which improves the prediction performance while reduces the data dimensionality.

The detailed implementation of the feature ranking/feature selection scheme is as follows: First, we identify the jam times at which the traffic volumes exceed a given threshold and the non-jam times at which the traffic volumes are much less than the threshold such that the positive and negative training data can be obtained, which are the traffic volumes prior to the jam and non-jam times with a certain time interval. Second, we make use of the *p*-test score presented in Golub et al. (1999) to rank the relevance of every sensor in the sense of predicting traffic jams at a given sensor, which is the so-called feature ranking. Third, we establish two Gaussian models from both the positive and negative samples for each sensor. Fourth, for the selected highly relevant sensors/features, we propose to score whether a jam will appear by combining the probabilities of how the input of every relevant sensor fits well into the corresponding Gaussian models. Fifth, we use a wrapper-like scheme (Kohavi and John, 1997) to select the optimal number of features for each sensor that can reach the best performance in the model selection procedure. Then, we use the additional data from 11 December 2010 to 20 September 2011 at the same city to test how the feature selection scheme performs.

Overall, the experiments confirm three points: (1) Signatures correlated to traffic jam prediction at one sensor exist in a wide area involving a large number of sensors in general. (2) Comparable or even better performance can be achieved with reduced dimensionality. (3) The optimal number of features determined by the wrapper-like scheme is not a fixed number but subject to which sensor is the target undergoing prediction, which forms more practical predictors.

The rest of this paper is organized as follows: We present the proposed methodology in Section 2. The experimental results are provided in Section 3. We conclude in Section 4.

2. The methodology

The traffic congestion prediction at a given sensor is formulated as a binary classification problem in the sense of pattern recognition. The goal is to classify the on-line state of a given sensor into two categories, namely jam or non-jam, by referring to the spatial temporal correlated signatures from a large number of sensors. The whole procedure is as follows: First, the training data are collected, where the historical data are partitioned into two sets: The jam set that contains the positive samples prior to the known traffic jams of a certain time lag, and the non-jam set consists of the negative samples prior to the known free travel times with the same time lag. Then, based on the positive and negative training samples, a predictor can be build up, which includes three modules: (1) feature ranking; (2) statistical decision; (3) determination of the optimal number of features. The functions of the three modules are as follows:

(1) For traffic jam congestion prediction at every sensor, the original input is the traffic volumes from all sensors, the dimensionality of which is very high, over thousands in general. However, the contribution of each sensor's data varies much in predicting traffic jams at the sensor of interest. The data from some sensors are more discriminative in distinguishing jam and non-jam while the data from some other sensors are not so discriminative. The goal of feature ranking is to score the discriminative power of each sensor in terms of separating positive training samples from negative ones so as to rank the "quality" of each sensor in predicting traffic jams at the sensor of interest. Once the rank of features/sensors are obtained for a given prediction task, we can make use of only the "high-quality" features in the subsequent decision making module to improve the prediction performance in terms of both precision and time complexity.
(2) Instead of applying classification directly, in this study, we only present the probability regarding whether a jam will occur after a certain time, which results from a statistics based method. In the learning phase, we construct two

Gaussian models for each sensor participating in the prediction at a given sensor, where one model is computed from the positive training samples and the other from the negative training samples. In the decision phase, we first evaluate how well the input traffic volume value from every individual sensor fits into the jam model as well as the non-jam model, and divide the former one by the latter one as the decision made by every individual sensor. Then, we combine the decisions from all sensors into a product to achieve an overall score representing the probability that a jam will occur shortly at the sensor of interest. Note that we do not provide a hard decision on whether a jam will appear but a probability on its possibility. As we aim to reveal how feature selection will affect traffic jam prediction, we rank the prediction scores and examine the top k jam candidates in terms of recall and precision as performance evaluation.

(3) By means of aforementioned feature ranking, we can obtain a list to rank the "quality" of each sensor in the sense of their relevance to traffic jam prediction at the sensor of interest. However, what is the optimal number of features is not solved, say, how to choose parameter K in order to make the top K features lead to the best performance. Actually, it is a model selection problem. The best K can be achieved experimentally by means of the wrapper-like feature selection scheme.

The detailed implementation is described below.

2.1. Preprocessing

We perform data cleaning at first. Suppose that μ_j and σ_j are the mean and standard deviation of the traffic volumes $\{v_1^j, v_2^j, \ldots, v_N^j\}$ from sensor j, where v_t^j represent the traffic volume of sensor j at sampling time t. If $\mu_j \cdot \sigma_j < ThreLow$, then, the data from sensor j should be removed without any further processing. The reason is: If the traffic volume at a sensor maintains always at low level or varies little, it should be independent to the others nodes.

For sensor $j \in J$, we should classify the traffic data into three states, jam, transition, and free travel, according to the volume values. We set two thresholds $ThreHigh = RatioHigh \times \max\{v_t^j\}$ and $RatioNotJam \times ThreHigh$ to achieve the three states, where $\max\{v_t^j\}$ is the maximum volume value among all the available data from all sensors, $RatioHigh < 1$, and $RatioNotJam < 1$. Then, the time set corresponding with congestion state is defined as $T^j = \{t | v_t^j > ThreHigh\}$ and the time set corresponding with free travel state is defined as $\overline{T}^j = \{t | v_t^j < RatioNotJam \times ThreHigh\}$. With the jam time T^j and non-jam time $\overline{T}^j$, we can then form the training sets for predicting traffic jams at sensor j with time lag τ using the data from all sensors, namely, the positive training set $\{v_{t-\tau}^i | t \in T^j\}$ and the negative training set $\{v_{t-\tau}^i | t \in \overline{T}^j\}$.

2.2. Feature ranking

Suppose that the traffic data are collected from M sensors distributed in a road network. The data from the M sensors do not perform equally in terms of distinguishing jam and non-jam cases at a sensor of interest, so it is natural to evaluate the discriminative power of each sensor. As for sensor i, if the distribution of the positive training samples $\{v_{t-\tau}^i | t \in T^j\}$ overlaps little with that of the negative training samples $\{v_{t-\tau}^i | t \in \overline{T}^j\}$, it means that the data from sensor i is highly reliable in indicating jam or non-jam at sensor j. Less overlap between positive and negative samples corresponds with a higher value in terms of distinguishing the two classes. Hence, we need a mathematical means to measure the overlap quantitatively, which should reflect the "quality" of each sensor in terms of distinguishing the jam and non-jam states at the sensor of interest. This is the so-called feature ranking problem in the context of machine learning. In this study, each sensor corresponds with one feature and the goal is to rank the M features/sensors according to their contribution in identifying jam and non-jam at a given sensor j.

Here, we apply the p-test presented in Golub et al. (1999) to evaluate the power of each sensor in separating jam and non-jam cases at sensor j based on the training data. Without loss of generality, we describe how to compute the p-test score for sensor i as an example. Let μ_{ij} and σ_{ij} denote the mean and standard deviation of the positive training samples $\{v_{t-\tau}^i | t \in T^j\}$, and $\bar{\mu}_{ij}$ and $\bar{\sigma}_{ij}$ those of the negative training samples $\{v_{t-\tau}^i | t \in \overline{T}^j\}$. The score for sensor i is defined as

$$S_{ij} = \frac{|\mu_{ij} - \bar{\mu}_{ij}|}{\sigma_{ij} + \bar{\sigma}_{ij}} \tag{1}$$

A bigger score corresponds with less overlap between the positive and negative samples in terms of statistical distribution, which means that sensor i is able to provide a less ambiguous signal indicating jam and non-jam at sensor j. Accordingly, we rank the scores $\{S_{ij} | i = 1, 2, \ldots, M\}$ in descending order in correspondence with the discriminative power of each sensor in distinguishing jam and non-jam classes at sensor j. Correspondingly, we denote the indices of such sorted list of sensors/features as $[I(1, j), I(2, j), \ldots, I(M, j)]$, where the corresponding scores satisfy $S_{I(1,j),j} \geqslant S_{I(2,j),j} \geqslant \cdots \geqslant S_{I(M,j),j}$. Note that the rank of the sensors $[I(1, j), I(2, j), \ldots, I(M, j)]$ varies with j since the positive and negative training samples are not fixed and subject to which sensor is of interest.

2.3. Prediction based on selected features

For classifying the state of sensor j at time t, the traffic volume values $\{v_{t-\tau}^i | i \in I^K\}$ of the first K sensors with indices $I^K = [I(1, j), I(2, j), \ldots, I(K, j)]$ at time $t-\tau$ are utilized as the input. Then, we compute the probability that a traffic jam will occur

at sensor j at time t based on every individual input volume value at time $t-\tau$. Such a probability is computed via the Gaussian models learnt from the training data, that is,

$$S_{t,\tau}^{ij} = \frac{\Pr\{v_{t-\tau}^i \in N(\mu_{ij}, \sigma_{ij})\}}{\Pr\{v_{t-\tau}^i \in N(\bar{\mu}_{ij}, \bar{\sigma}_{ij})\}} \tag{2}$$

where $i \in I^K$, symbol "Pr" means probability, symbol "N" represents Gaussian distribution, and the definitions of μ_{ij}, σ_{ij}, $\bar{\mu}_{ij}$, and $\bar{\sigma}_{ij}$ refer to Section 2.2. In Eq. (2), $\Pr\{v_{t-\tau}^i \in N(\mu_{ij}, \sigma_{ij})\}$ accounts for how significant $v_{t-\tau}^i$ indicates a possible traffic jam while $1/\Pr\{v_{t-\tau}^i \in N(\bar{\mu}_{ij}, \bar{\sigma}_{ij})\}$ corresponds with in what degree $v_{t-\tau}^i$ is not related to non-jam. Only when both conditions hold significantly, the prediction score in Eq. (2) can be high. In another word, only when $v_{t-\tau}^i$ is highly related to known jams while not obviously related to existing non-jam cases, the probability that $v_{t-\tau}^i$ indicates a jam at time t is high. Note that Eq. (2) presents only the decision based on an individual sensor. We need to combine the decisions of the K sensors to reach an overall decision on the probability that a jam will appear at sensor j at time t, and such overall scoring can be computed via the following equation:

$$S_{t,\tau}^j = \prod_{i=1}^{K} S_{t,\tau}^{ij} \tag{3}$$

For simplicity, we compute the logarithm of Eq. (3) instead of its original form, that is,

$$S_{t,\tau}^j = \sum_{i=1}^{K} \log(S_{t,\tau}^{ij}) \tag{4}$$

Finally, we sort $\{S_{t,\tau}^j\}$ in descending order to indicate the probabilities that traffic jams will appear with time lag τ. We prefer Gaussian models because they are computationally tractable in regard to such massive data to be processed. There are also alternative ways for establishing the statistical models. However, complicated models are often confronted with some difficulties like high computational cost or model selection.

The performance in terms of traffic congestion prediction can be evaluated as follows: For predication at sensor j, without loss of generality, we denote the sorted scores computed from Eqs. (4) and (2) as $S_{1,\tau}^j \geqslant S_{2,\tau}^j \geqslant, \cdots \geqslant S_{T,\tau}^j$, where the corresponding actual traffic volumes at sensor j with time lag τ are denoted as $[v_t^j | t = 1, 2, \ldots, T]$. Then, the prediction accuracy at sensor j is defined as follows:

$$P_j = \sum_{k=1}^{\#T^j} H(v_k^j > ThreHigh) \Big/ \#T^j \tag{5}$$

where $\#T^j$ means the total number of the true traffic jams happened at sensor j, $H(c) = 1$ if condition c holds, and $H(c) = 0$ otherwise. P_j figures out how many top-ranked results are true compared with the total true jam number. Also, the performance curve regarding precision and recall is defined as

$$p_j(l) = \sum_{k=1}^{l} H(v_k^j > ThreHigh) \Big/ l \tag{6}$$

$$r_j(l) = \sum_{k=1}^{l} H(v_k^j > ThreHigh) \Big/ \#T^j \tag{7}$$

where $p_j(l)$ and $r_j(l)$ are the precision and recall rate in regard to the top-l candidates. The precision $p_j(l)$ figures out how many true jams exist in the top-l returns while the recall rate $r_j(l)$ is the ratio of the correct returns among the top-l returns to all the true jams.

2.4. Optimal number of features

The rank list of the sensors varies with which sensor is under consideration for prediction, that is, the rank list $[I(1, j), I(2, j), \ldots, I(M, j)]$ varies with j due to different training samples. This gives rise to a problem: For different sensor j, how to determine the optimal number K_j that can lead to the best prediction performance with the feature entries $[I(1, j), I(2, j), \ldots, I(K_j, j)]$, which correspond with the top K_j features ranked. This is a model selection problem and can be solved as follows: (1) We partition the training data into two subsets, part A for learning the statistical models defined in Eq. (2) and part B for evaluating the performance. (2) By means of the experiments with the part B data as input, the optimal number of features leading to the highest performance defined in Eq. (5) is recorded for each sensor undergoing prediction. When the optimal number of features for each sensor j is recorded, namely K_j, we can then examine whether such a feature selection scheme outperforms the solution based on all features by using another testing data set as input. The comparison will be presented latter in the experimental section.

3. Experiments

3.1. Overall prediction precision against feature number

We use the real-world data from Traffic Management Center of Minnesota Department of Transportation to conduct the experiment. The data was collected with a 30-s interval from over 4000 loop detectors located around the Twin Cities Metro freeways for 7 days per week from 1 January to 22 September 2010, where the sensors containing missing values, weekends, and the days with incomplete record are not taken into account. Finally, we have a data set of 156 days with the traffic volume data summed per 10-min interval for 4584 sensors. We use the first 126 days for learning and the remaining 30 days for testing.

The prediction is made with 10, 20, 60, and 300 min in advance, respectively. The threshold for distinguishing traffic jams is *ThreHigh* = 320, which is half of the highest traffic volume regarding all observations by setting *RatioHigh* = 0.5. We let *RatioNotJam* = 0.5 to distinguish the non-jam cases. Following the preprocessing with thresholds *ThreLow* = 1000, only 3386 sensors are preserved in the following prediction experiment and the prediction is performed on 888 sensors undergoing frequent traffic jams, at each of which the jam number exceeds a threshold *ThreMany* = 200 in the training data.

We aim to reveal how the prediction precision is subject to the feature number, that is, how the prediction accuracy at every sensor is affected by the sensor number K in Eq. (4). The prediction procedure as well as the performance evaluation is as follows: (1) Without loss of generality, we rank all the sensors according to Eq. (1) with regard to the prediction at sensor j. (2) We select K top-ranked features to perform congestion prediction at sensor j based on Eqs. (4) and (2), where the prediction precision for sensor j is computed via Eq. (5) and denoted as $P_j(K)$. (3) We compute the mean precision of all sensors against the feature number K as follows:

$$P(K) = \frac{1}{M}\sum_{j=1}^{M} P_j(K) \tag{8}$$

The performance curve $P(K)$ obtained with different time lags is illustrated in Fig. 1. It is obvious that the mean prediction precision increases with the increment of feature number for all time lags. When all the features are applied as input, the mean precision reaches the highest one for each time lag listed. The turning points in the performance curves are marked with a symbol "o". Prior to reaching the turning points, the performance curves increase remarkably with the increment of the feature number. Since the turning points, the precisions increase very slowly with the feature number. We list in Table 1 the mean prediction precision as well as the feature number at the turning point and that at the end point for each performance curve. To understand how the feature number affects the mean prediction precision, we provided in Table 2

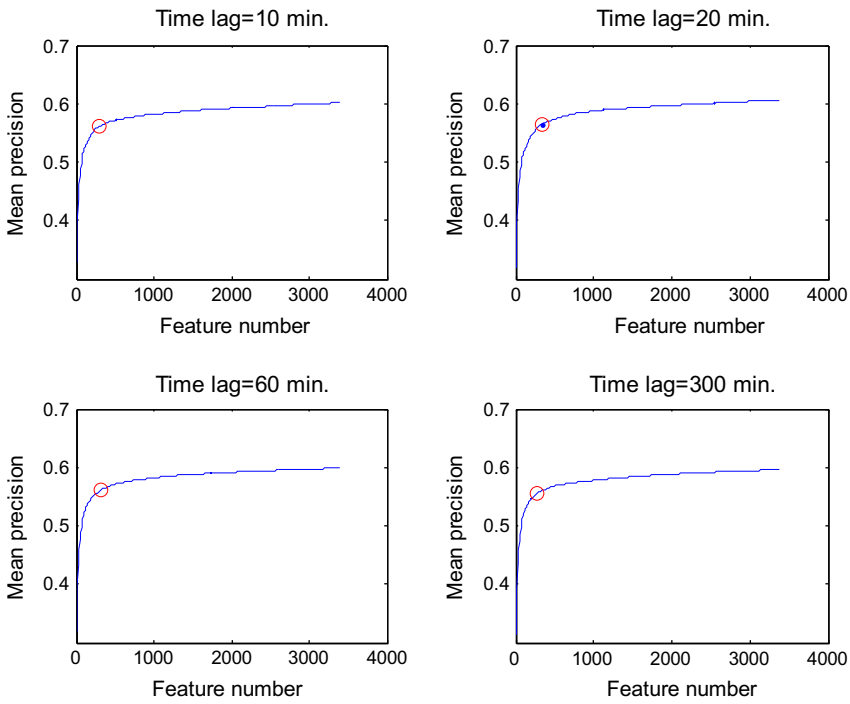

Fig. 1. Mean prediction precision against feature number.

Table 1
Mean prediction precision and feature number at turning point and end point.

Time lag	10 min	20 min	60 min	3000 min
Turning point				
Precision	56%	56.45%	56.17%	55.52%
#Features	286	336	326	276
End point				
Precision	60.13%	60.58%	59.89%	59.62%
#Features	3386			

Table 2
Mean prediction precision against feature number.

#Features	6	26	56	106	156	286	876	1726	3146	3386
10 min	32.67	39.6	48.86	52.26	53.29	56	58.01	59.02	60	**60.13**
20 min	31.75	43.3	48.61	51.72	53.4	55.86	58.54	59.48	60.47	**60.58**
60 min	31.86	43.32	48.82	52.26	53.96	55.75	58.01	58.93	59.74	**59.89**
300 min	31.21	43.58	48.66	52.18	53.84	55.55	57.60	58.61	59.5	**59.62**

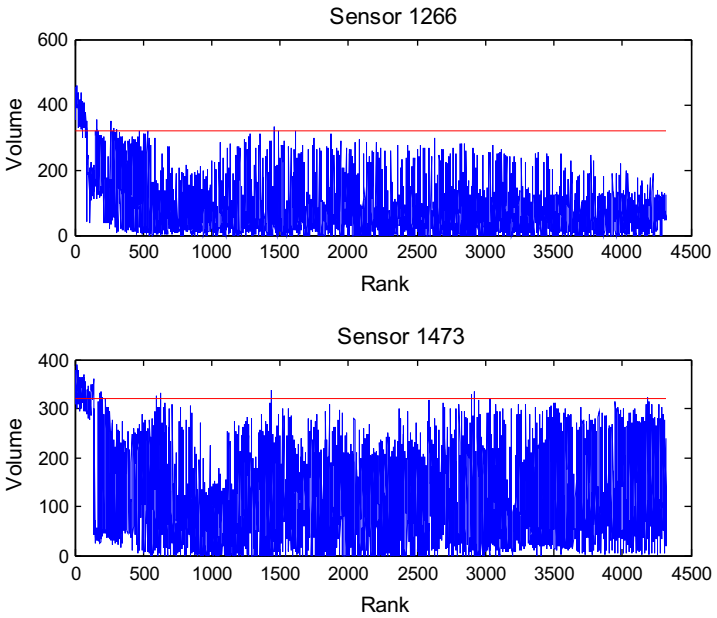

Fig. 2. Actual traffic volumes with time lag of 20 min following the prediction with all features.

the mean prediction precision against the feature number with times lag being 10, 20, 60, and 300 min. The highest precision for each case is bolded, which appears when all the 3386 features are applied in the prediction. From Table 2, we can reach the following observations: (1) the mean prediction performance is quite low if the feature number is less than 100; (2) the highest mean prediction precision is obtained when all the features are applied; (3) when the feature number is greater than 1700 (about half of the total features), the mean prediction precision is close to the highest one, about 1% off.

The above experimental results prove two points: (1) The signatures correlated to traffic jam prediction at a given sensor exist in a quite wide range since relatively high prediction performance can only be achieve with the data from a large number of sensors. (2) The prediction performance improves little when the feature number is high enough, so that the dimensionality can be reduced greatly without sacrificing the precision much.

In Fig. 2, we illustrate two examples of the actual traffic volumes corresponding with the rank list obtained via the predictor described previously in Section 2, where the prediction is based on all features with time lag of 20 min. The corresponding precision and recall curves of the two sensors are illustrated in Fig. 3. It can be seen that high volumes are mostly top-ranked. For the case when time lag is 20 min and the prediction is based on all features, the prediction precision ranges from 0% to 94.74% for the 888 sensors undergoing frequent traffic jams. The distribution of the prediction precision over the 888 sensors is illustrated in Fig. 4 in the form of histogram. We can see that for most sensors, the prediction precision falls in [0.5, 0.8].

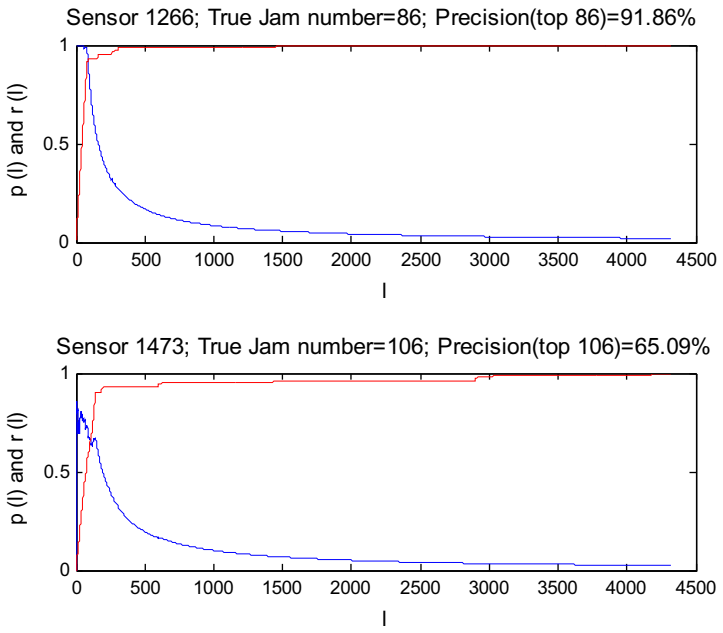

Fig. 3. Precision and recall rate of the two sensors' performance shown in Fig. 2.

3.2. Prediction precision against feature number for each individual sensor

The above performance evaluation is based on the average over the 888 sensors of interest. Regarding every individual feature, however, the optimal number of features at which the highest prediction precision is achieved could be different. Fig. 5 shows the curves of precision against feature number for two sensors, where the points corresponding with the highest precisions are marked with "o". The optimal number of features to reach the highest prediction precision for the two sensors is 156 and 26, respectively, and the highest prediction precision for the two cases is 93.02% and 68.87%, correspondingly. The distribution of the 888 sensors of interest over the maximum precision under different time lag as well as the optimal number of features is illustrated in Figs. 6 and 7, respectively. We can see that the optimal number of features distributes in a wide range from less 10 to the number of all features, 3386. So is the maximum precision, which distributes from 0% to 94.74% for the case with time lag of 20 min. The comparison between the prediction precision using all features and that based on the optimal number of features is illustrated in Table 3. We can see that the prediction precision with the optimal number of features is a bit higher than the performance obtained with all features. This means that it is possible to achieve comparable or higher performance with fewer features. The details will be presented latter. Table 4 provides the details

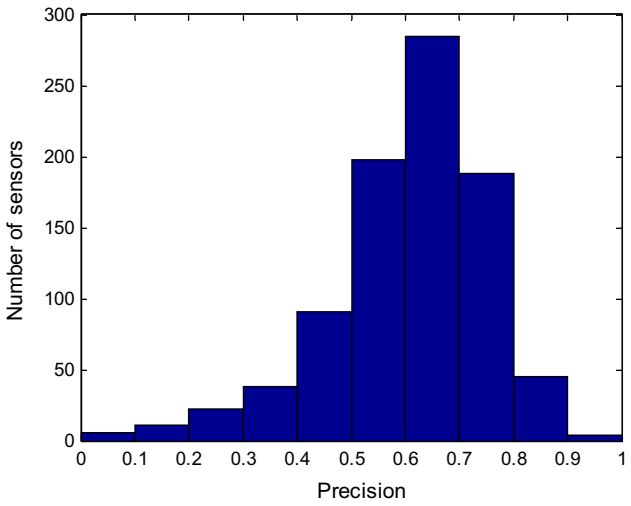

Fig. 4. Distribution of precision with prediction based on all features (time lag = 20 min).

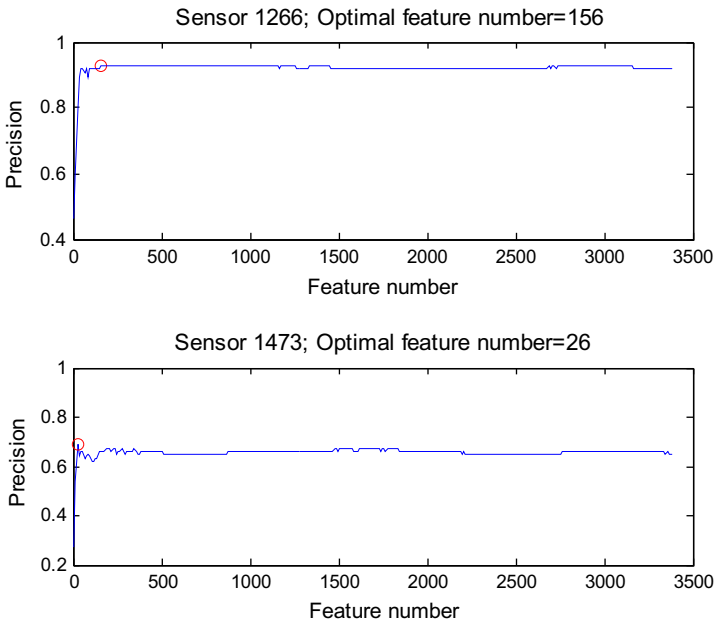

Fig. 5. Prediction precision against feature number for two sensors.

regarding how the 888 sensors of interest distribute over the optimal number of features. It can be seen that for most sensors, the optimal number of features is greater than 100, which indicates that the signatures to warn traffic jams at a given sensor exist widely in the road network.

3.3. Prediction with selected features

We record the optimal number of features for each sensor obtained in section 3.2. Then, we perform another test with a different data set to compare the prediction performance achieved with the optimal number of features and that based on all

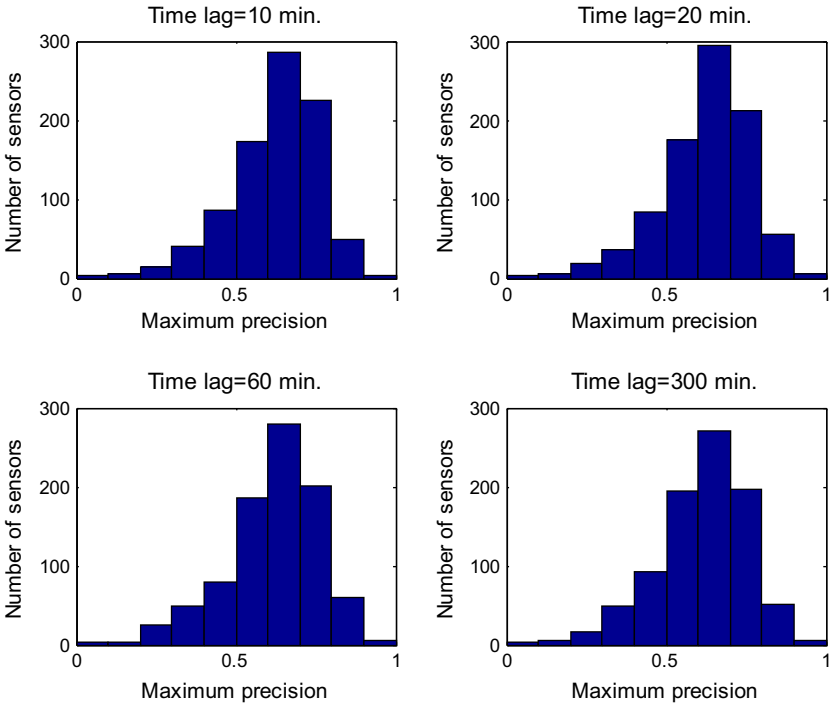

Fig. 6. Distribution of the 888 sensors of interest over the maximum precision.

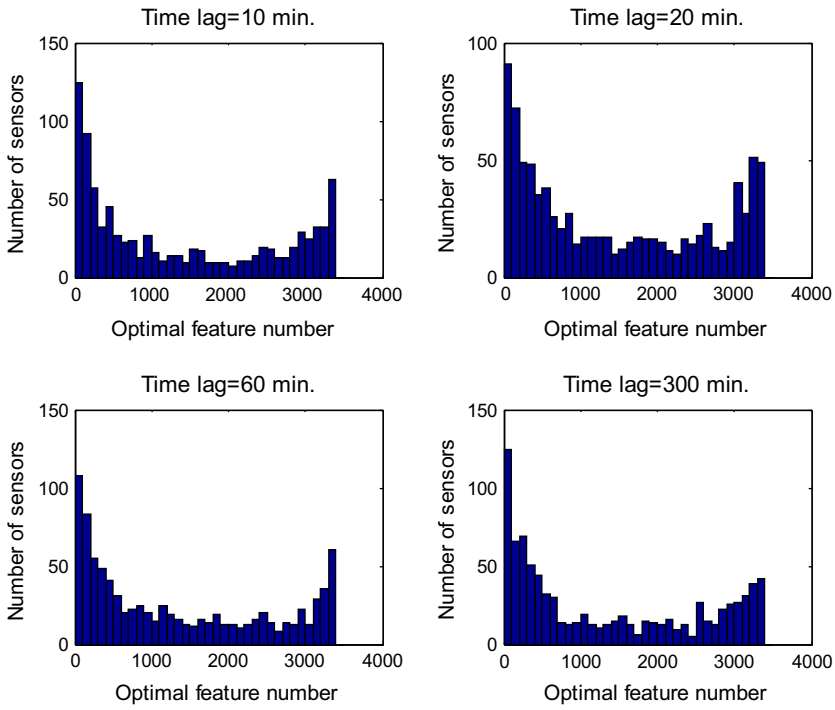

Fig. 7. Distribution of the 888 sensors of interest over the optimal number of features.

Table 3
Comparison of the mean precision over the 888 sensors of interest achieved by using the optimal number of features to that using all features.

Time lag (min)	10	20	60	300
Mean of max precision (%)	62.21	62.33	61.79	61.46
Mean of precision with all features (%)	60.13	60.58	59.89	59.62

Table 4
Distribution of the 888 sensors of interest over the optimal number of features.

#Features	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)	[50, 100)	[100, 3386)	3386
10 min	11	18	10	18	9	58	753	11
20 min	8	12	8	7	8	48	788	9
60 min	11	4	11	16	5	61	769	11
300 min	13	8	20	13	10	60	761	3

Table 5
Mean prediction precision (%).

Time lag (min)	10	20	60	300
Optimal	52.08	51.92	51.73	51.30
All	52.07	51.80	51.57	50.99

features. The data is also from Traffic Management Center of Minnesota Department of Transportation but the records are from 11 December 2010 to 20 September 2011 excluding weekends. Note that the new data set may contain missing values at the 4584 sensors and we replace the missing values with 0. The comparison regarding the averaged prediction precision over the 888 sensors under different time lags is provided in Table 5. It can be seen that the performance achieved with the optimal number of features is a little bit better than that based on all features while the dimensionality has been reduced greatly (see Fig. 7). This reveals that feature ranking and feature selection should play an important role in traffic congestion prediction, which is so far a miss topic in the literature.

The prediction precision listed in Table 5 is much lower than that obtained in Sections 3.1 and 3.2. The reason is that in Sections 3.1 and 3.2, we use 126 days for training and 30 days for testing but in this section, we use 156 days for training and 283 days with missing values for testing, which degrades the prediction performance.

4. Conclusion

It is the first time in the literature to investigate into the feature selection/feature ranking topic for traffic congestion prediction. As confirmed experimentally, the number of sensors relevant to the prediction task at one sensor is usually over 100. Such a high data dimensionality challenges the existing methods in terms of solving the spatial temporal correlations among massive sensor data. By applying feature ranking and feature selection, only the most relevant features are preserved such that the data dimensionality can be greatly reduce while the performance can be maintained the same or made better. Moreover, we propose a new method in this study to compute the probability of traffic jam appearance with a certain time lag by combining many one-dimensional probabilities into a product.

In the future, we will try to incorporate more feature selection and classification methods into such study.

Acknowledgements

This work is supported by 973 Program (Grant No.2010CB731401), Major Program of NSFC (Grant No.91024011), NSFC (Grant No. 61071133), Ministry of Industry and Information Technology of China (Grant No. 2010ZX01042-002-003-004), and Science and Technology Commission of Shanghai Municipality (Grant No. 09JC1401500).

References

Ando, Y., Masutani, O., Sasaki, H., Iwasaki, H., Fukazawa, Y., Honiden, S., 2006. Pheromone model: application to traffic congestion prediction. In: Brueckner, S.A. et al. (Eds.), Lecture Notes in Artificial Intelligence, vol. 3910. Springer, pp. 182–196.

Ben-Akiva, M.E., Gao, S., Wei, Z., Wen, Y., 2012. A dynamic traffic assignment model for highly congested urban networks. Transportation Research Part C 24, 62–82.

Chandra, S.R., Al-Deek, H., 2009. Predictions of freeway traffic speeds and volumes using vector autoregressive models. Journal of Intelligent Transportation Systems 13 (2), 53–72.

Ghosh, B., Basu, B., O'Mahony, M., 2009. Multivariate short-term traffic flow forecasting using time-series analysis. IEEE Transactions on Intelligent Transportation Systems 10, 246–254.

Golub, T.R. et al, 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

He, J., He, Q., Swirszcz, G., Kamarianakis, Y., Lawrence, R., Shen, W., Wynter, L., 2010. Ensemble-based method for task 2: predicting traffic jam. In: Proc. IEEE International Conference on Data Mining Workshops, pp. 1363–1365.

Hu, J., Kaparias, I., Bell, M.G.H., 2009. Spatial econometrics models for congestion prediction with in-vehicle route guidance. IET Intelligent Transport Systems 3, 159–167.

Kanoh, H., Furukawa, T., Tsukahara, S., Hara, K., Nishi, H., Kurokawa, H., 2005. Short-term traffic prediction using fuzzy C-means and cellular automata in a wide-area road network. In: Proc. 8th International IEEE Conference on Intelligent Transportation Systems, pp. 984–988.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

Mazloumian, A., Geroliminis, N., Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. Philosophical Transactions of the Royal Society A 368, 4627–4647.

Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. Transportation Research Part C 19, 606–616.

Romaszko, L., 2010. IEEE ICDM 2010 contest: traffic prediction – jams. In: Proc. IEEE International Conference on Data Mining Workshops, pp. 1366–1368.

Smith, B.L., Williams, B.M., Oswalsd, R.K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C 10, 303–321.

Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. Journal of Transportation Engineering – ASCE 129, 664–672.