Real-time Freeway Traffic State Estimation Based on Cluster Analysis and Multiclass Support Vector Machine

Chao Deng, Fan Wang, Huimin Shi and Guozhen Tan
Department of Computer
Dalian University of Technology
Dalian, China
wangfan@dlut.edu.cn

Abstract—Urban traffic state analysis plays an important role in the solution of traffic congestion problem. To estimate traffic state effectively is a foundational work for improving traffic condition and preventing traffic congestion. In this paper, a novel pattern-based approach is proposed to model the clustering and classification of traffic state. First, fuzzy-set clustering method is utilized to divide the traffic state into a number of patterns. Then multiclass support vector machine (MSVM) is applied to estimate these states with real-time traffic data. The result shows that the proposed approach is promising for the dynamic estimation of road traffic state and can provide forecasted congestion information for the traffic control system and traffic guidance system.

Keywords: cluster analysis; multiclass support vector machine (MSVM); traffic state estimation

I. INTRODUCTION

Nowadays, traffic congestion has become one of the most major and costly problems in the world, especially in many big cities and metropolitan areas. In urban traffic filed, traffic state analysis is the most important approach to solve the traffic problem. Therefore, how to estimate traffic state exactly becomes a very important research field.

Traffic state estimation had been identified as an important task within a traffic control loop already in the 1970s [1]. Traffic state estimation for a freeway network refers to estimating all traffic variables of the network at the current time instant based on available real-time traffic measurements.

Clustering Technology is an essential method in data mining process to reveal natural structures and identify interesting patterns in the underlying data. It seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. And it has been successfully applied to predicting the traffic volume [2] and urban traffic environment quality evaluation [3]. Recently, support vector machine (SVM) is an effective model for classification and regression that has been successfully used to predict traffic factors and incident detection [4]. And later many Multiclass SVM methods are prompted and widely used in multiclass classification such as intrusion detection [5].

In this paper, we present a dynamic traffic estimation model. The model deals with real-time traffic data and converts them into traffic states. In the model the clustering analysis is worked on historical traffic flow data and divides them into clusters with different traffic states. Then the Multiclass Support Vector Machine (MSVM) classification model can be used as a classifier to identify the real-time traffic states.

The rest of the present paper is organized as follows: Section 2 discusses the model of traffic state estimation. In Section 3, the methods for traffic state estimation are introduced. And in Section 4, we discuss the application of this method. Finally, the main conclusions are summarized in Section 5.

II. REAL-TIME TRAFFIC STATE ESTIMATION MODEL

For highway traffic systems, parameters of traffic flow, temporal occupancy, and average vehicle speed are usually used to describe the complex state of the road. These parameters represent the traffic state in perspective of spatial, temporal, microcosmic, and macroscopic properties. Aiming at reaching accurate estimation of the traffic state, the proposed method has three main stages as:

- Data preprocessing: Since in many cases the data will not be 100% correct or even missing to a degree, a preprocessing module is imperative, it performs three tasks: data selecting, data checking and data correction.
- Clustering analysis and real-time classification: Clustering analysis and classification analysis are applied to develop the dynamic traffic state estimation model. Clustering analysis divides the data into clusters with different status. After training of the MSVM with historical data, we can use the model to classify the traffic state with real time traffic data in traffic control system.
- Pattern application: This stage consists of presenting the results to the decision maker, who may solve
potential conflicts with previously believed or extracted knowledge and apply the new discovered patterns.

Fig. 1 shows the whole processes of the model.

III. METHODOLOGICAL APPROACHES

A. K-Means Cluster Analysis Method

K-Means cluster method was first brought out by MacQuen [6]. It takes \( k \) as the parameter and divides the \( n \) objects into \( k \) clusters. Through cluster analysis, the objects are similar to one another in same cluster but dissimilar in different clusters. Similarity between clusters is calculated by the mean value of all the objects in the cluster. This method can handle the situation which the data are concentrated in a cluster but obviously distinct in different clusters. The algorithm is with \( O(nkt) \) time complexity (\( t \) is the iteration number), and has high convergence and efficiency.

K-means cluster method is one of the simplest unsupervised classification techniques. It is also one of the most popular unsupervised learning algorithms due to its simplicity.

Let \( X = \{X_1, X_2, \ldots, X_n\} \) be a set of \( n \) objects in which each object \( X_i \) is represented as \( \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\} \), where \( m \) is the number of numerical attributes. To cluster \( X \) into \( k \) clusters by the agglomerative fuzzy-K-Means algorithm [7] is to minimize the following objective function:

\[
P(U, Z) = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{i,j} D_{i,j} + \sum_{i=1}^{k} u_{i,j} \log u_{i,j},
\]

s.t. \( \sum_{j=1}^{n} u_{i,j} = 1, \quad u_{i,j} \in (0,1), \quad 1 \leq i \leq n. \)

Where \( U = [u_{i,j}] \) is an \( n \times k \) partition matrix, \( u_{i,j} \) represents the association degree of membership of the \( i \)th object \( x_i \) to the \( j \)th cluster \( z_j \). \( Z = \{z_1, z_2, \ldots, z_k\}^T \) is a dissimilarity measure between the \( j \)th cluster center and the \( i \)th object. Here, the square of the Euclidean norm is used as the dissimilarity measure.

B. Multiclass Support Vector Machine

Support vector machine is originally designed for binary classification. We briefly introduce the formulation for binary classification SVM.

The original idea of SVM is to use a linear separating hyperplane to separate the training data into two classes.

Given training vectors \( x_i \in \mathbb{R}^p, i=1,2,\ldots,l \), in two classes, and a vector \( y \in \mathbb{R}^l \) such that \( y_i \in \{1,-1\} \), \( C - SVC \) [8] solves the following primal problem:

\[
\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i,
\]

s.t. \( y_i[w^T \phi(x_i) + b] + \xi_i \geq 1, \quad \xi_i \geq 0, \quad i=1,\ldots,l. \)

Its dual is

\[
\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha,
\]

s.t. \( y^T \alpha = 0, \quad 0 \leq \alpha_i \leq C, \quad i=1,2,\ldots,l. \)

Where \( e \) is the vector of all ones, \( C > 0 \) is the upper bound, \( Q \) is a \( l \times l \) positive semidefinite matrix \( (Q_{ij} = y_i y_j K(x_i, x_j) \) and the kernel is \( K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \). Here training vectors \( x_i \) are mapped into a higher (maybe infinite) dimensional space by the function \( \phi \).

The decision function is

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b \right)
\]
The main difference between binary classification and multiple class problems is that $y_i \in \{1,2,...,k\}$ instead. There are 3 major methods to extend the binary classification SVM to multiclass SVM.

- **One-against-all**: It constructs $k$ SVM models where $k$ is the number of classes. The $i$th SVM is trained with all of the examples in the $i$th class with positive labels, and all other examples with negative labels. There are $k$ binary classification models and the final result is the class with largest output.

- **One-against-one**: Another major method is called the one-against-one method. This method constructs $k(k-1)/2$ classifiers where each one is trained on data from two classes. Then we predict it in the class with the largest vote. The voting approach which called the “MaxWins” is adapted. The test data $x$ votes for each classifiers and after all $x$ belongs to the class with most vote.

- **Directed acyclic graph SVM (DAGSVM)**: Its training phase is the same as the one-against-one method by solving $k(k-1)/2$ binary SVM classifiers. However, in the testing phase, it uses a rooted binary directed acyclic graph which has internal nodes and leaves. Each node is a binary SVM of $i$th and $j$th classes. Given a test sample $x$, starting at the root node, the binary decision function is evaluated. Then it moves to either left or right depending on the output value. Therefore, we go through a path before reaching a leaf node which indicates the predicted class.

The one-against-one method is chosen in our research for there are only a few traffic states and the one-against-one method has higher accuracy without loss of speed.

### IV. Experiment Procedure

#### A. Data Description

The detailed traffic data on southbound US 101 Freeway, in Los Angeles, USA are used in this research (http://ngsim.camsys.com). The US 101 dataset contains loop detector data, raw and processed video, weather data, and other data. These data are collected in every 5 minutes. Loop 717486 is chosen and we adapt the traffic flow, occupancy and average speed data from Jun. 13th to Jun. 17th, 2005 for research. Since there is few traffic in late night, we use the data form 5:00 am to 10:00 pm, and there are totally 1020 observations. The first 4 days’ data (204 observations) are used for train and the last day’s data for test.

#### B. Cluster Analysis

We choose the K-means algorithm to cluster the traffic flow data. We take the parameters with $N = 4$ (cluster number) and $S = 10$ (seed number). TABLE I shows the information about the cluster centers of each cluster.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Cluster 0</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>345</td>
<td>422</td>
<td>114</td>
<td>139</td>
</tr>
<tr>
<td>Flow</td>
<td>511.6812</td>
<td>721.4242</td>
<td>855.5263</td>
<td>718.0576</td>
</tr>
<tr>
<td>Occupancy</td>
<td>0.0591</td>
<td>0.0825</td>
<td>0.1363</td>
<td>0.1982</td>
</tr>
<tr>
<td>Speed</td>
<td>70.6261</td>
<td>71.4398</td>
<td>52.5728</td>
<td>33.5201</td>
</tr>
</tbody>
</table>

In this experiment, the traffic parameters are clustered in 4 clusters. Cluster 0 has lowest occupancy, low flow and high speed, it represents the sparse traffic. Cluster 1 takes high flow and speed yet low occupancy, it represents the free traffic. Cluster 2 and 3 take high flow and occupancy but low speed, they represent the congestion traffic. Cluster 3 takes higher occupancy but lower flow and speed, so it must be a serious congestion. Cluster 2 has higher flow and speed, but lower occupancy, so it is a lower congested situation.

Fig. 2 shows the results of the cluster analysis with consideration to the relationship between occupancy and flow. Obviously, we can see there are 4 clusters with different traffic states, and we can get a direct impression of the relationship between the traffic parameters flow and occupancy.

![Figure 2. The cluster results by using K-means algorithm](image)

#### C. Multiclass Support Vector Machine Classification

The algorithm uses 75% of the data as the learning set to train the model and uses the remaining 25% of the data to test the validity of the model. Using the data developed by the cluster analysis described in the previous section, the class variable for classification model is defined as the cluster number. Multiclass SVM is used to do classification on the normalized samples. LIBSVM toolbox is used to compute the results and radial basis function is chosen as the kernel function with parameters $\gamma = 0.0001$, $C = 3.15$.

To evaluate the classification performance of our algorithm, classification accuracy (CA) is employed as follow:

$$CA(100\%) = \frac{t}{n} \times 100\%$$

Where $t$ is the number of sample cases been correctly classified, and $n$ is the total number of sample cases.
The model is 96.5686% accurate on the test data. And this indicates that Multiclass SVM is an effective way to solve the traffic state estimation problem.

V. CONCLUSIONS

Traffic state is a kind of perceived measure, which is influenced by the number of vehicle on the road and the mental endurance of people. In the paper, the clustering analysis and the classification analysis, are worked on real traffic data, and develop the clustering model in fuzzy theory view and classification model in multiclass support vector machine. The classification model then can be used as a classifier to estimate the real-time traffic states. According to experimentation, we can investigate the capability of each node and link, which is the basic component of road network. More importantly, it can help us research into the true condition of road traffic, which is our main purpose. In particular, the presented approach can estimate traffic state for the given road network, which could aid traffic managers in evaluating traffic demand level and making demand management.

ACKNOWLEDGMENT

This paper was supported in part by Grand 60873256 of National Natural Science Foundation of China and Grand 2005CB321904 of The National Basic Research Program.

REFERENCES