Journal of Southeast University (English Edition) Vol. 27, No. 2, pp. 174 – 179 June 2011 ISSN 1003—7985

Gaussian mixture models for clustering and classifying traffic flow in real-time for traffic operation and management

Sun Lu[1,2] Zhang Huimin[3] Gao Rong[4] Gu Wenjun[1] Xu Bing[1] Chen Liliang[1]

([1]School of Transportation, Southeast University, Nanjing 210096, China)
([2]Department of Civil Engineering, The Catholic University of America, Washington DC 20064, USA)
([3]Jinzhong Bureau of Highway Administration of Shanxi Province, Jinzhong 030600, China)
([4]Xinzhou Bureau of Highway Administration of Shanxi Province, Xinzhou 034000, China)

Abstract: Based on Gaussian mixture models (GMM), speed, flow and occupancy are used together in the cluster analysis of traffic flow data. Compared with other clustering and sorting techniques, as a structural model, the GMM is suitable for various kinds of traffic flow parameters. Gap statistics and domain knowledge of traffic flow are used to determine a proper number of clusters. The expectation-maximization (E-M) algorithm is used to estimate parameters of the GMM model. The clustered traffic flow patterns are then analyzed statistically and utilized for designing maximum likelihood classifiers for grouping real-time traffic flow data when new observations become available. Clustering analysis and pattern recognition can also be used to cluster and classify dynamic traffic flow patterns for freeway on-ramp and off-ramp weaving sections as well as for other facilities or things involving the concept of level of service, such as airports, parking lots, intersections, interrupted-flow pedestrian facilities, etc.

Key words: traffic flow patterns; Gaussian mixture model; level of service; data mining; cluster analysis; classifier
doi: 10.3969/j.issn.1003 – 7985.2011.02.012

An important contribution of the highway capacity manual (HCM) has been made to the introduction of the concept of level of service (LOS)[1]. The definition and interpretation of LOS are elicited and revised over the years. The latest HCM classifies freeway LOS into six categories, from letters "A" to "F", with "A" representing the best operating condition and "F" the worst[1]. LOS analyses are most useful in the evaluation of facilities that do not exist, and for which no data are available. They are critical in the analysis of alternatives that only exist in concept, where the results of these analyses allow decision makers to compare the relative benefits of potential investments.

Hall et al.[2] pointed out that although different LOS for freeways can be interpreted in terms of driver perception. Very few studies have sought drivers' views about what is important to them. When drivers travel on a segment of roadway, it is the real-time traffic flow state that matters the most to their perception and comfort. The assessment of the real-time traffic flow states is not only valuable to drivers, but also of paramount importance to traffic operations and management[3–5]. Freeway systems in urban areas involve a complex collection of facilities, management and operating strategies. The use of performance measures in many transportation planning and decision-making processes of public agencies has increased significantly. Traffic flow states can be used as a natural performance measure of highway traffic operation.

Public agencies are concerned with preserving the mobility, improving the safety, enhancing the reliability, and meeting the public's expectations for efficient travel. There is a great need to develop a descriptive philosophy-based traffic operation measure that is effective for classifying real-time traffic flow states, and at the same time, to be different from the current scheme of prescriptive philosophy based LOS. Chasey et al.[6] suggested using a comprehensive level of service to measure the performance of civil infrastructure systems. Hua et al.[7–11] adopted artificial neural networks for classifying traffic flow states. While these studies demonstrate the significance of assessing real-time traffic conditions, many of them use HCM as a basis for supervised learning. Recently, Kikuchi and Chakroborty[12] adopted the fuzzy set theory to evaluate traffic flow states.

The motivations of this paper are twofold. First, we propose to use data mining and pattern recognition to conduct real-time classification of traffic flow states. The purpose of real-time classification of traffic flow states is to provide drivers and traffic management an effective performance measure to timely and comprehensively assess roadway traffic operation conditions. Data mining and pattern recognition are data driven learning technologies that can help in analyzing inherent patterns of data[13–14]. The application of these technologies to traffic flows has great potential and involves many challenges[15]. The wide deployment of the intelligent transportation systems (ITS) has made large traffic datasets from field measurements available. This tremendous amount of actual observations provides a basis for this kind of analysis to be representative. Secondly, LOS for many other facilities such as toll plazas, parking lot and airports has not yet been defined explicitly. Several studies have attempted to use various approaches for defining LOS for these facilities with limited success[16–17]. The proposed data mining and pattern recognition based approaches offer different alternatives for achieving this purpose, which have been used successfully in developing multi-regime traffic stream models[18].

Received 2010-09-13.
Biography: Sun Lu (1972—), male, doctor, professor, sunl@cua.edu.
Foundation items: The US National Science Foundation (No. CMMI-0408390, CMMI-0644552), the American Chemical Society Petroleum Research Foundation (No. PRF-44468-G9), the Research Fellowship for International Young Scientists (No. 51050110143), the Fok Ying-Tong Education Foundation (No. 114024), the Natural Science Foundation of Jiangsu Province (No. BK2009015), the Postdoctoral Science Foundation of Jiangsu Province (No. 0901005C).
Citation: Sun Lu, Zhang Huimin, Gao Rong, et al. Gaussian mixture models for clustering and classifying traffic flow in real-time for traffic operation and management[J]. Journal of Southeast University (English Edition), 2011, 27(2): 174 – 179. [doi: 10.3969/j.issn.1003 – 7985.2011.02.012]

1 Methodologies

Common unsupervised learning techniques include cluster analysis, self-organizing and association rules, which lead to concept generalization and knowledge discovery. Cluster analysis deals with data labelling without training samples[19-21]. Data with similar properties are grouped into the same cluster. Hastie et al. [14, 22-24] applied the K-means algorithm of cluster analysis to multi-regime traffic stream modeling. Classification belongs to supervised learning techniques, dealing with data labelling with training samples. Supervised methods can be symbolically based, statistically based or neural networks.

The concept of level of service based on descriptive philosophy[3-5] was used as a performance measure of traffic operations. Methods based on objective quantification need to be established in order to classify states of real-time traffic. In this paper, we utilize cluster analysis and classification for such a purpose. Fig. 1 presents a schematic flowchart for establishing classifiers for classifying real-time traffic flow states. The procedure proceeds as follows. First, cluster analysis (left box in Fig. 1) is used to segment massive multivariate traffic data collected from the field into a number of clusters. Each cluster will be labeled uniquely and analyzed statistically. Such information is then used to build a classifier, capable of classifying unseen traffic data into different clusters (right box in Fig. 1).

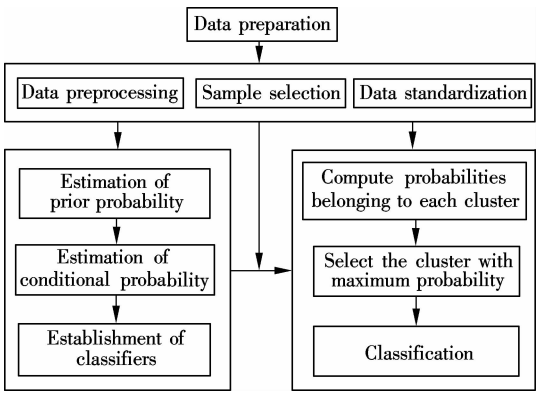

Fig. 1 Procedure for establishing classifiers for classifying real-time traffic flow states

2 Gaussian Mixture Model

This paper uses the Gaussian mixture model for traffic flow clustering analysis. The GMM assumes that traffic flow characteristics are sampled from a mixed Gaussian density. Each component of this mixed density is a multivariate Gaussian distribution. Suppose that $\boldsymbol{x}$ are samples obtained by choosing a state of nature ω_c with probability $P(\omega_c)$ and then selecting $\boldsymbol{x}$ according to the probability law $p(x \mid \omega_c, \theta_c)$. Here $P(\omega_c)$, $c = 1, 2, \ldots, C$ are unknown prior probabilities for each class. Thus, the probability density function for sample $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{c=1}^{C} p(\boldsymbol{x} \mid \omega_c, \boldsymbol{\theta}_c) P(\omega_c) \tag{1}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}^{\mathrm{T}}$. The goal of clustering analysis is to use samples drawn from this mixture density to estimate the unknown parameter vector $\boldsymbol{\theta}$. Once $\boldsymbol{\theta}$ is known, the mixture can be decomposed into its components $P(\boldsymbol{x} \mid \omega_c, \boldsymbol{\theta}_c)$, and a maximum posterior classifier be used on the derived densities[21, 14].

Suppose that a set $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N unclassified samples of traffic data is drawn independently from a mixture density specified in Eq. (1), in which parameter vector $\boldsymbol{\theta}$ is fixed but unknown, and $\boldsymbol{x}$ is the vector formed by density and speed. By definition, the likelihood of the samples is the joint density.

$$p(D \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(x_n \mid \boldsymbol{\theta}) \tag{2}$$

Let l be the logarithm of the likelihood. The unknown parameter vector $\boldsymbol{\theta}$ and unknown prior probabilities μ_C can be estimated from maximizing the log-likelihood function under constraints

$$\max\ l = \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) \tag{3}$$

$$\text{s. t.} \quad \sum_{c=1}^{C} P(\omega_c) = 1 \qquad k = 1, 2, \ldots, K; P(\omega_c) \geqslant 0$$

Let $\hat{\boldsymbol{\theta}}_c$ be the maximum likelihood estimate for $\boldsymbol{\theta}_c$ and $\hat{P}(\omega_c)$ be the maximum likelihood estimate for $P(\omega_c)$. If the likelihood function is differentiable and if $\hat{P}(\omega_c) \neq 0$ for all c, then the above optimization problem leads to the requirement that $\hat{\boldsymbol{\theta}}_c$ and $\hat{P}(\omega_c)$ must satisfy[18]

$$\hat{P}(\omega_c) = \frac{1}{N} \sum_{n=1}^{N} \hat{P}(\omega_c \mid \boldsymbol{x}_n, \hat{\boldsymbol{\theta}}) \tag{4}$$

$$\sum_{n=1}^{N} \hat{P}(\omega_c \mid \boldsymbol{x}_n, \hat{\theta}) \nabla_{\theta_c} \ln p(\boldsymbol{x}_n \mid \omega_c, \hat{\boldsymbol{\theta}}_c) = 0 \tag{5}$$

where ∇_{θ_c} is the gradient with respect to θ_c and

$$\hat{P}(\omega_c \mid \boldsymbol{x}_n, \hat{\boldsymbol{\theta}}) = \frac{p(\boldsymbol{x}_n \mid \omega_c, \hat{\boldsymbol{\theta}}_c) \hat{P}(\omega_c)}{\sum_{c=1}^{C} p(\boldsymbol{x}_n \mid \omega_c, \hat{\boldsymbol{\theta}}_c) \hat{P}(\omega_c)} \tag{6}$$

Eq. (4) states that the maximum likelihood estimate of the probability of a category is the average over the entire data set of the estimate derived from each sample where each sample is weighted equally.

Assume that the component densities of traffic data are multivariate normal, then $P(\boldsymbol{x} \mid \omega_c, \boldsymbol{\theta}_c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, and $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$, and $P(\omega_c)$ are all unknown. Because only half of the off-diagonal elements of $\boldsymbol{\Sigma}_k$ are independent, the term $\ln p(\boldsymbol{x}_n \mid \omega_c, \hat{\boldsymbol{\theta}}_c)$ in Eq. (6) becomes

$$\ln p(\boldsymbol{x}_n, \boldsymbol{\theta}_c) = \ln\{(2\pi)^{-d/2} \left|\boldsymbol{\Sigma}_c^{-1}\right|^{1/2}\} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_c) \tag{7}$$

where d is the dimension of the dataset. The implementation of the above maximum likelihood estimation is via the expectation-maximization (E-M) algorithm[21, 25-27].

Step 1 Initialization: Start with initial guesses for the

parameters $\hat{\boldsymbol{\theta}}$ and $\hat{P}(\omega_c)$;

Step 2 Expectation: Compute the responsibilities (memberships),

$$\hat{P}(\omega_c|\boldsymbol{x}_n,\hat{\boldsymbol{\theta}}) = \frac{|\boldsymbol{\Sigma}_c^{-1}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_c)\right]\hat{P}(\omega_c)}{\sum_{j=1}^{C}|\boldsymbol{\Sigma}_j^{-1}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{j}}^{-1}(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_j)\right]\hat{P}(\omega_j)}$$

Step 3 Maximization: Compute the weighted means and covariance,

$$\hat{P}(\omega_c) = \frac{1}{N}\sum_{n=1}^{N}\hat{P}(\omega_c|\boldsymbol{x}_n,\hat{\boldsymbol{\theta}})$$

$$\hat{\boldsymbol{\mu}}_c = \frac{\sum_{n=1}^{N}\hat{P}(\omega_c|\boldsymbol{x}_n,\hat{\boldsymbol{\theta}})\boldsymbol{x}_n}{\sum_{n=1}^{N}\hat{P}(\omega_c|\boldsymbol{x}_n,\hat{\boldsymbol{\theta}})}$$

$$\hat{\boldsymbol{\Sigma}}_c = \frac{\sum_{n=1}^{N}\hat{P}(\omega_c|\boldsymbol{x}_n,\hat{\boldsymbol{\theta}})(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_c)(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}}{\sum_{n=1}^{N}\hat{P}(\omega_c|\boldsymbol{x}_n,\hat{\boldsymbol{\theta}})}$$

Step 4 Iterate steps 2 and 3 until convergence.

3 Traffic Datasets and Data Preprocessing

Actual traffic data on July 2, 2001 are collected from TransGuide program (2006), the Advanced Traffic Management System (ATMS)[28] at San Antonio, Texas. TransGuide records speed, volume and occupancy from all roadway lanes at 20-second intervals using loop detectors and video cameras. The traffic data of a whole day is obtained from the second most left lane of basic sections of three roadways: I-10 east bound (sensor ID: L2-0010E-562.581), US-281 north bound (sensor ID: L2-0281N-143.895), and Loop-1604 west bound (sensor ID: L2-1604W-034.326). Each roadway contains 4 320 records during a single day and these sensors are located in basic sections rather than weaving sections of roadway. Before the collected traffic data are used for this study, data quality control and preprocessing are conducted to ensure data integrity and correctness. Some of the records show zero speed and some show null vehicle presence. These data are removed from the datasets, which consist of 24.6% of the entire dataset. Eventually, available records for investigation in this study are 3 258 for each dataset, respectively.

Unlike the existing LOS scheme where only one characteristic (e.g., density) is used to classify traffic flow, in this paper we use flow, speed and occupancy together as a comprehensive characteristic. In other words, every data point is a three-dimensional observation. The reason we choose occupancy rather than density as a feature is because the former can be more easily obtained from spot sensors (loop detector, video camera, etc.).

Since speed, volume and occupancy possess different units, standardization is needed to make the data dimensionless. In this study the following formula is used for standardization.

$$\boldsymbol{x} = \frac{\boldsymbol{x}_{\text{original}} - \boldsymbol{x}_{\text{average}}}{\boldsymbol{x}_{\text{std}}} \text{ with } \boldsymbol{x} = (\text{speed, volume, occupancy}) \tag{8}$$

where $\boldsymbol{x}_{\text{original}}$, $\boldsymbol{x}_{\text{average}}$ and $\boldsymbol{x}_{\text{std}}$ represent original, average and standard deviation of speed, volume and occupancy, respectively. Eq. (8) can also be used to convert a standardized traffic characteristic back into its original counterpart. The Gap statistic is defined as[29]

$$\text{Gap}_n(K) = E_n^*[\log(W_K)] - \log(W_K) \tag{9}$$

where E_n^* denotes expectation under a sample of size n from the reference distribution. A good estimate K^* will be the value maximizing $\text{Gap}_n(K)$[14]. Since the estimate (9) is very general, it can be applicable to any clustering method. Computation of the gap statistic proceeds as follows[29]:

1) Cluster the observed data, varying the total number of clusters from $K=1,2,\ldots,8$, and giving dissimilarity measures W_K, $K=1,2,\ldots,8$.

2) Generate each reference feature uniformly over the range of the observed values for a given feature.

3) Generate B reference datasets using the uniform prescription described in step 2. Cluster each one giving within cluster dissimilarity measures W_{Kb}^*, $b=1,2,\ldots,B$; $K=1,2,\ldots,8$. Compute the estimated Gap statistic:

$$\text{Gap}_n(K) = \frac{1}{B}\sum_b \log(W_{Kb}^*) - \log(W_K) \tag{10}$$

Let $\bar{l} = \frac{1}{B}\sum_b \log(W_{Kb}^*)$, compute the standard deviation $\text{sd}_K = \left[\frac{1}{B}\sum_b \log(W_{Kb}^*) - \bar{l}\right]^{1/2}$, and define $s_K = sd_K\sqrt{1+\frac{1}{B}}$. Finally, choose the number of clusters via

$$\hat{K} = K_{\min} \text{ such that } \text{Gap}(K) \geqslant \text{Gap}(K+1) - s_{K+1} \tag{11}$$

Fig. 2 shows the Gap statistic against the number of clusters using GMM clustering for three standardized traffic datasets. From Fig. 2, it can be observed that the first local peak in the Gap statistic occurs at $\hat{K}=3$ for all the three roadways, except that the peak for Loop-1604 is not so sharp. So it is plausible that the number of clusters be set to three. Detailed cluster analysis is given in the next section. One may also use a different number of patterns to classify traffic flow. For instance, the HCM sets up six LOSs for basic sections of the freeway (HCM 2001). Generally speaking, a large number of clusters provides more degrees of freedom to describe detailed traffic patterns, at a price of increased model complexity. It is our belief that the right

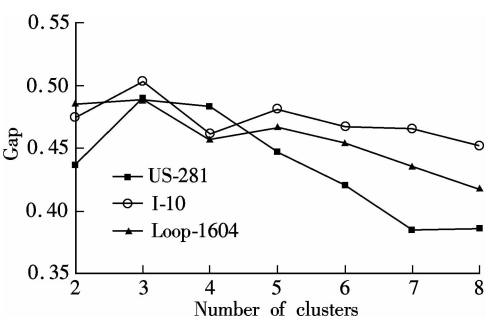

Fig. 2 Gap statistic as a function of number of clusters

number of clusters should be appropriately determined using clustering analysis results, domain knowledge and how the clustered patterns are to be used.

4 Cluster Analysis

In what follows, speed, flow and occupancy corresponding to mean values of clusters are abbreviated as average speed, average flow and average occupancy. Clusters are labeled 1, 2, and 3 in ascending order of the average occupancy of each cluster. Fig. 3 plots clustered flow-speed, flow-occupancy and speed-occupancy for roadways I-10, US-281 and Loop-1604, respectively, in which mean values of each cluster are highlighted using "hollow squares". It can be seen that three clusters are reasonably well distinguished in these figures. Fig. 4 gives the coefficient of variation (CoV) of the average (speed, flow and occupancy) for each cluster. For all three highways, the CoVs of average speed are always less than those of average flow and average occupancy. In addition, the CoVs of average speed for cluster 3 (congestion) are always greater than those of average speed for cluster 1 and cluster 2.

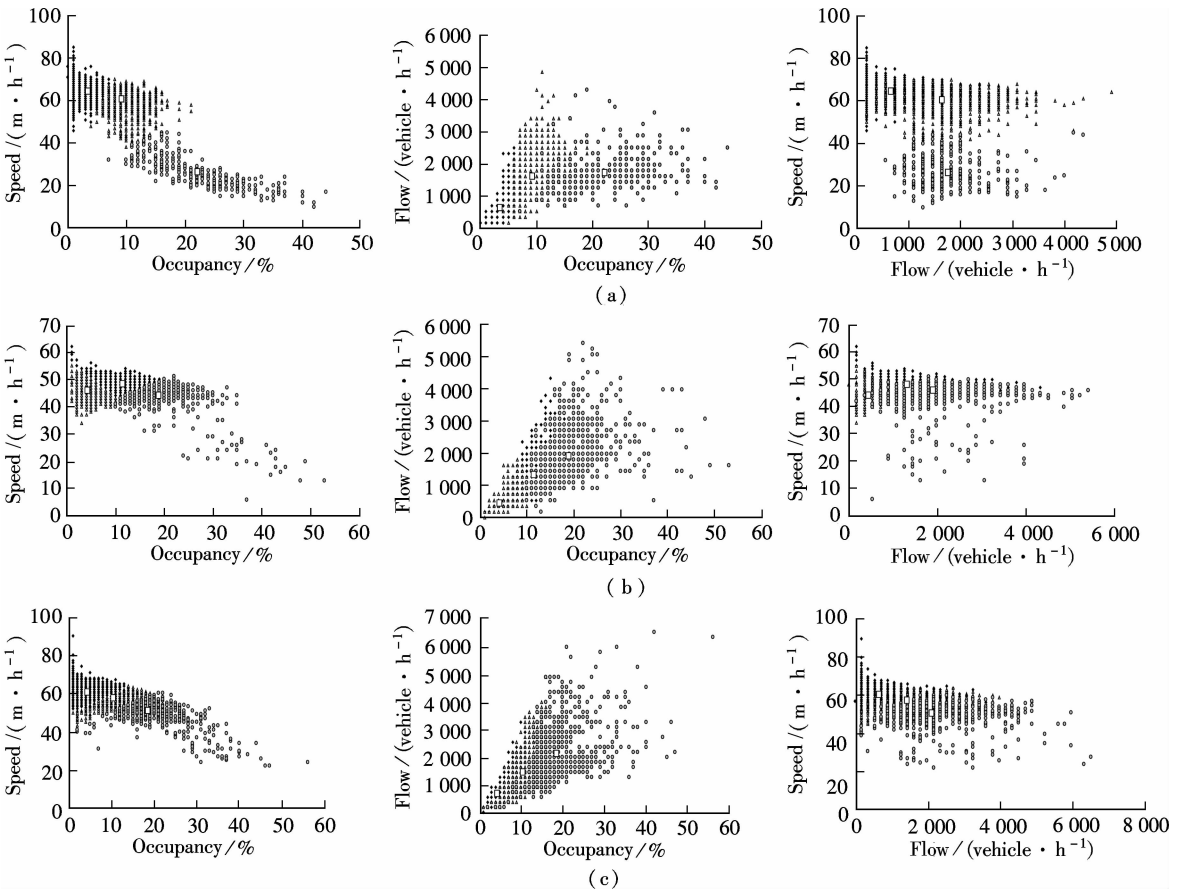

Fig. 3 Three clusters and mean values of clusters. (a) I-10; (b) US-281; (c) Loop-1604

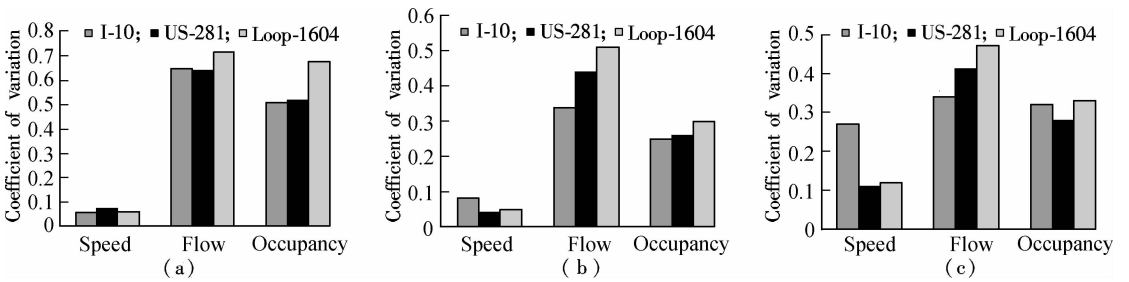

Fig. 4 Coefficients of variation of three cluster centers for three roadways. (a) Cluster 1; (b) Cluster 2; (c) Cluster 3

5 Classifier Design

It is evident that the previous GMM based cluster analysis can be used for identifying features associated with different traffic flow patterns. In addition, the result of the clustered traffic observation after clustering can be also used as a basis for constructing automatic classifiers. The designed classifier can assess real-time highway traffic operations conditions off-line or on-line into different classes (clusters), once new observations of traffic data becomes available. Such classification information can be informative to the traveling public or traffic operations and management teams. Classifi-

cation belongs to supervised learning and a number of methods are available in the literature of pattern recognition and machine learning, such as the k-nearest neighbor method, linear discriminant analysis, artificial neural network, and support vector machine. To be consistent with the clustering method used in the previous section, here we design a classifier that is still based on the GMM.

Several pieces of information are needed to implement the classifier, including the prior probability and conditional probability estimation. The estimation of the prior probabilities can be obtained from the clustering analysis using

$$P(\omega_j) = \frac{1}{n}\sum_{k=1}^{n} z_{jk} \tag{12}$$

where $P(\omega_j)$ is the prior probability and $z_{jk} = 1$ if the natural state for the k-th sample is ω_j and $z_{jk} = 0$ otherwise. The conditional probability density function corresponding to each category can be obtained from

$$p(\boldsymbol{x}\mid\omega_j) = \frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{13}$$

where $\boldsymbol{x}$ is a d-component column vector (here $d = 3$); $\boldsymbol{\mu}$ is the d-component mean vector; $\boldsymbol{\Sigma}$ is the d-by-d covariance matrix; $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ are the determinant and inverse of the covariance matrix, respectively; $(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}$ is the transpose of $\boldsymbol{x}-\boldsymbol{\mu}$ and ω_j the j-th category. Since the clustered data are now available, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively computed using the maximum likelihood estimation

$$\mu = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k \tag{14}$$

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{x}_k-\boldsymbol{\mu})(\boldsymbol{x}_k-\boldsymbol{\mu})^{\mathrm{T}} \tag{15}$$

The clustered traffic data become a training data set for developing a classified maximum likelihood. The posterior probability can be determined using Bayes' formula

$$P(\omega_j\mid\boldsymbol{x}) = \frac{p(\boldsymbol{x}\mid\omega_j)P(\omega_j)}{p(x)} \tag{16}$$

where $P(\omega_j\mid\boldsymbol{x})$ is the posterior probability and $p(\boldsymbol{x})$ is given by $p(\boldsymbol{x}) = \sum_{j=1}^{c} p(\boldsymbol{x}\mid\omega_j)P(\omega_j)$. The new observation is classified into the cluster that gives the greatest probability.

6 Conclusion

We use three traffic flow characteristics (speed, flow and occupancy) for clustering and classification of traffic flow patterns. The proper number of clusters, which is three in this paper, is determined based on the Gap statistic and the domain knowledge of the traffic flow.

Cluster analysis provides an ideal tool for identifying similarity and partitioning traffic data into different numbers of clusters. The proposed GMM model for real-time traffic flow clustering and classification has a number of advantages. It allows multiple variables to simultaneously enter the clustering and classification process, promising to have a comprehensive evaluation of patterns of traffic flow. No artificial thresholds need to be specified in advance to divide different patterns. Rather, each pattern is recognized and described after clustering analysis. The proposed GMM is data driven and adaptive, and provides a great flexibility to accommodate actual environments of highway traffic operations. The advocated approach gives rise to a mechanism for determining a proper number of clusters by iteratively running clustering analysis for an increased number of clusters and comparing the Gap statistic. Finally, traffic data sets used in this paper are collected from basic sections of freeways.

References

[1] Transportation Research Board. *Highway capacity manual* [M]. Washington DC, USA: National Research Council, 2001.

[2] Hall F L, Wakefield S, Al-Kaisy A. Freeway quality of service: what really matters to drivers and passengers? [J]. *Transportation Research Board*, 2001(1776): 17 − 23.

[3] Hall F L, Hurdle V F, Banks J H. A synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways [J]. *Transportation Research Record*, 1992(1365): 12 − 18.

[4] Helbing D, Hennecke A, Treiber M. Phase diagram of traffic states in the presence of inhomogeneities [J]. *Physics Review Letters*, 1999, **82**(21): 4360 − 4363.

[5] Kerner B S, Rehborn H. Experimental properties of complexity in traffic flow [J]. *Physics Review E*, 1996, **53**(5): 4275 − 4278.

[6] Chasey A D, de la Garza J M, Drew D R. Comprehensive level of service: needed approach for civil infrastructure systems[J]. *Journal of Infrastructure Systems*, 1997, **3**(4): 143 − 153.

[7] Hua J, Faghri A. Dynamic traffic pattern classification using artificial neural networks [J]. *Transportation Research Record*, 1993(1399): 14 − 19.

[8] Mead W C, Fisher H N, Jones R D, et al. Application of adaptive and neural network computational techniques to traffic volume and classification monitoring [J]. *Transportation Research Record*, 1994(1466): 116 − 123.

[9] Lingras P. Classifying highways: hierarchical grouping versus Kohonen neural networks [J]. *Journal of Transportation Engineering*, 1995, **121**(4): 364 − 368.

[10] Saito M, Fan J. Multilayer artificial neural networks for level- of-service analysis of signalized intersections [J]. *Transportation Research Record*, 1999(1678): 216 − 224.

[11] Yang H, Qiao F. Neural network approach to classification of traffic flow states [J]. *Journal of Transportation Engineering*, 1998, **124**(6): 521 − 525.

[12] Kikuchi S, Chakroborty P. Ways to treat uncertainty in level of service determination[C]//*The* 83*rd Annual Meeting of Transportation Research Board*. Washington DC, USA, 2003.

[13] Mitchell T. *Machine learning* [M]. McGraw Hill, 1997.

[14] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*: *data mining*, *inference*, *and prediction* [M]. Springer-Verlag, 2001.

[15] Shekar S, Lu C T, Chawla S, et al. Data mining and visualization of twin-cities traffic data, TR 01-015 [R]. Twin Cities, MN, USA: Department of CSE, University of Minnesota, 2000.

[16] Oh C, Ritchie S. Real-time inductive-signature-based level of service for signalized intersections [J]. *Transportation Research Record*, 2002(1802): 97 − 104.

[17] Klodzinski J, Al-Deek H M. New methodology for defining level of service at toll plazas[J]. *Journal of Transportation Engineering*, *ASCE*, 2002, **128**(2): 173 - 181.
[18] Sun L, Yang J, Mahmassani H, et al. Data mining based adaptive regression for developing equilibrium static traffic speed-density relationships [J]. *Canadian Journal of Civil Engineering*, 2010, **37**(3): 389 - 400.
[19] Hartigan J A. *Clustering algorithms* [M]. New York: Wiley, 1975.
[20] Hartigan J A, Wong M A. A K-means clustering algorithm [J]. *Applied Statistics*, 1979, **28**: 100 - 108.
[21] Duda R O, Hart P E, Stork D G. *Pattern classification* [M]. 2nd ed. New York: John Wiley & Sons, Inc., 2001.
[22] Gordon A D. *Classification* [M]. 2nd ed. London: Chapman & Hall/CRC, 1999.
[23] Rencher A C. *Methods of multivariate analysis* [M]. John Wiley & Sons, 2002.
[24] Sun L, Zhou J. Development of multiregime speed-density relationships by cluster analysis [J]. *Transportation Research Record*, 2005(1934): 64 - 71.
[25] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm (with discussion) [J]. *Journal of the Royal Statistical Society*, *Series B*, 1977, **39**(1): 1 - 38.
[26] Witten I H, Frank E. *Data mining*: *practical machine learning tools and techniques* [M]. 2nd ed. Morgan Kaufmann, 2005.
[27] StatSoft Inc. Electronic textbook: cluster analysis [EB/OL]. (2006) [2010-06-20]. www. statsoft. com/textbook/.
[28] TransGuide Program. The advanced traffic management system (ATMS) at San Antonio [EB/OL]. (2006) [2010-06-20]. http://www. transguide. dot. state. tx. us/.
[29] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic[J]. *Journal of the Royal Statistical Society*, *Series B*, 2001, **63**(2): 411 - 423.

用于交通运营管理的实时交通流状态分类高斯混合模型

孙 璐[1,2] 张惠民[3] 高 荣[4] 顾文钧[1] 徐 冰[1] 陈鲤梁[1]

([1] 东南大学交通学院,南京 210096)
([2]Department of Civil Engineering, Catholic University of America, Washington DC 20064, USA)
([3] 山西省公路局晋中分局,晋中 030600)
([4] 山西省公路局忻州分局,忻州 034000)

摘要:采用高斯混合模型 GMM,同时以交通流量、平均速度和密度 3 种交通流宏观特征为指标,对交通流状态进行聚类和分类. 和其他聚类分类方法比较,高斯混合模型是结构化的模型,适合于各种情形交通流参数. 高斯混合模型中子类的个数通过 Gap 统计量结合交通流的领域知识加以确定,而模型的其他参数则由 E-M 算法进行估计. 所建立的 GMM 模型可以作为实时交通流状态的分类器对新的观察值开展有效的分类识别和预报. 同时,聚类分析和模式识别也可以用来对其他含有服务水平概念的设施进行聚类和分类分析,比如机场、停车场、交叉口等.

关键词:交通流型; 高斯混合模型; 服务水平; 数据挖掘; 聚类分析; 分类
中图分类号:U491