Real-time identification of probe vehicle trajectories in the mixed traffic corridor

Yu Mei ¹, Keshuang Tang *, Keping Li ¹

Department of Comprehensive Transportation Information and Control Engineering, Tongji University, No. 4800, Cao’an Road, Shanghai 201804, China

ABSTRACT

This paper proposes three enhanced semi-supervised clustering algorithms, namely the Constrained-K-Means (CKM), the Seeded-K-Means (SKM), and the Semi-Supervised Fuzzy c-Means (SFCM), to identify probe vehicle trajectories in the mixed traffic corridor. The proposed algorithms are able to take advantage of the strengths of topological relation judgment and the semi-supervised learning technique by optimizing the selection of pre-labeling samples and initial clustering centers of the original semi-supervised learning technique based on horizontal Global Positioning System data. The proposed algorithms were validated and evaluated based on the probe vehicle data collected at two mixed corridors on Shanghai's urban expressways. Results indicate that the enhanced SFCM algorithm could achieve the best performance in terms of clustering purity and Normalized Mutual Information, followed by the CKM algorithm and the SKM algorithm. It may reach a nearly 100% clustering purity for the uncongested conditions and a clustering purity greater than 80% for the congested conditions. Meanwhile, it could improve clustering purity averagely by 21% and 14% for the congested conditions and 6.5% and 6% for the uncongested conditions, as compared with the traditional K-Means algorithm and the basic SFCM. The proposed algorithms can be applied for both on-line and off-line purposes, without the need of historical data. Clustering accuracies under different traffic conditions and possible improvements with the use of historical data are also discussed.

1. Introduction

Probe vehicle systems have been widely implemented in many developed countries as a promising technology. It has also gained increasing attentions in China in recent years. So far, more than 10 mega cities in China including Beijing and Shanghai have already launched the application of probe vehicle systems since 2002, most of which are based on taxi GPS (Global Positioning System) data. Those systems usually upload taxi’s GPS data with a time interval of 5–60 s. Probe vehicle systems have many applications, e.g., dynamic traffic state estimation, OD (Origin Destination) estimation, route travel time estimation. In Shanghai, the numbers of GPS-equipped buses and taxies have reached 15,390 and 48,714 respectively at the end of 2011 according to Shanghai comprehensive transportation annual report (Shanghai City Comprehensive Transportation Planning Institute, 2012). More than 40,000 GPS-equipped taxies provide GPS data to Shanghai Traffic Information Center with a time interval of 10–20 s.

* Corresponding author. Tel.: +86 21 69583656; fax: +86 21 33626318.
E-mail addresses: 90mei@tongji.edu.cn (Y. Mei), tang@tongji.edu.cn (K. Tang), keping_li@vip.163.com (K. Li).

http://dx.doi.org/10.1016/j.trc.2015.06.008
0968-090X/© 2015 Elsevier Ltd. All rights reserved.
A high proportion of urban roads are applied expressways that often exist in parallel with and close to ground arterials in many mega cities in China. It creates the so-called the mixed traffic corridor. Again, taking Shanghai as an example, the total mileage of elevated expressways has reached 308 km at the end of 2011 and most of which are parallel with and close to the arterials (Shanghai City Comprehensive Transportation Planning Institute, 2012). Usually, accurate altitude data together with a three-dimensional (3D) digital map can effectively identify vehicles running on the elevated expressways and the ground arterials. However, the accuracy of altitude data and the availability of 3D digital map often restrict such an application. It is firstly due to that it is difficult to accurately obtain altitude data in an environment full of skyscrapers and elevated expressways because a GPS device needs to receive signals from at least four satellites in order for the estimation of altitude of probe vehicles (Velaga et al., 2009). Secondly, many commercial probe vehicle systems in China only provide the post-processed horizontal GPS positions, i.e., easting and northing fixes (He et al., 2013). Hence, how to identify probe vehicle trajectories in the mixed corridor using horizontal GPS data has become a technical problem in the application of probe vehicle systems in China’s mega cities.

In most of the existing map matching algorithms, common solutions to identify probe vehicle trajectories in the mixed corridor are either applying judgment rules based on topological relations of the mixed corridor, or utilizing travel speed threshold values based on an assumption that vehicles travel faster on the elevated expressways. The former is more often used in the case of offline application, in which a complete trajectory of every probe vehicle during a long time interval is already available. However, it can usually identify a small proportion of sample vehicles due to the limited information available from the topological relations and has rarely been used in the case of online application due to partial trajectories. The latter can be used for both online and offline application purposes as its calculation process is fairly simple and fast. However, proper travel speed threshold values are often difficult to determine particularly for the congested traffic conditions, which need a lot of historical data. In addition, its accuracy would decrease to an unacceptable level for the congested conditions due to small travel speed differences between the elevated expressways and the ground arterials, although it can obtain good performance for the uncongested conditions.

To overcome those drawbacks of the existing methods, this paper proposed three enhanced semi-supervised clustering algorithms, namely the Constrained-K-Means (CKM), the Seeded-K-Means (SKM), and the Semi-Supervised Fuzzy c-Means (SFCM). The proposed algorithms are able to take advantage of the strengths of topological relation judgment and the semi-supervised learning technique by optimizing the selection of pre-labeling samples and initial clustering centers of the original semi-supervised learning technique based on horizontal GPS data. They were validated and evaluated based on the probe vehicle data collected at two mixed corridors on Shanghai’s urban expressways. Clustering accuracies under different traffic conditions and possible improvements with the use of historical data were also discussed in the paper.

2. Literature review

To date, probe vehicle systems based on the GPS data from private cars, buses or taxis have already applied widely in the Intelligent Transportation Systems (ITS). Traffic state estimation and prediction was one of the basic applications (De Fabritiis et al., 2008; Kong et al., 2013; Lund and Pack, 2010; Pu et al., 2009). Another application of probe vehicle systems was updating digital road maps, which had the advantages of low cost, short updating circle and high accuracy (Li et al., 2012; Vartziotis et al., 2012). Some studies have also applied probe vehicle data to estimate dynamic OD for traffic demand forecasting purpose (Baek et al., 2010; Friedrich et al., 2010). Recently, the extended floating car systems (xFCD) that were based on the floating cars equipped with extra visual cameras and on-board Electronic Control Units have appeared to be a more efficient way for traffic state estimation, the level-of-service estimation, and real-time road safety assessment, e.g., Kyamakya et al. (2011), Messelodi et al. (2009), Pell et al. (2012) and Diaz et al. (2012).

Map-matching (MM) process is vitally important in the application of probe vehicle systems as its performance has a significant effect on obtaining real-time traffic information. MM algorithms usually integrate GPS data with a spatial road map in order to identify the road segment on which a user (or a vehicle) is traveling and the location on that segment. Recent research on MM algorithms has been based on either a conventional topological analysis or a probabilistic approach. Among them, topological MM algorithms are relatively simple, easy and quick, enabling them to be implemented in real-time. Therefore, a topological MM algorithm is used in many navigation devices manufactured by industry. Although the performance of some of these algorithms is good in relatively sparse road networks, they are not always reliable for complex roundabouts, merging or diverging sections of motorways, and complex urban road networks (Velaga et al., 2009). Hence, Quddus et al. (2006) proposed a map matching algorithm based on fuzzy logic theory and tested it on different road networks of varying complexity. It was found that the fuzzy logic-based map matching algorithm could provide a significant improvement over existing map matching algorithms both in terms of identifying correct links and estimating the vehicle position on the links. Velaga et al. (2009) developed an enhanced weight-based topological MM algorithm in which the weights are determined from real-world field data using an optimization technique to improve the performance of matching. Smaili et al. (2014) proposed an approach based on hybrid dynamic Bayesian networks enabling to implement in a unified framework two of the most successful families of probabilistic model commonly used for localization: linear Kalman filters and Hidden Markov Models. The combination of these two models enabled to manage and manipulate multi-hypotheses and multi-modality of observations characterizing MM problems and improve integrity approach.
Meanwhile, some researchers have attempted to tackle a challenging issue for traditional MM algorithms that is mapping large-scale low-frequency FCD onto the road network. Chen et al. (2014) developed a multi-criteria dynamic programming map-matching (MDP-MM) algorithm for online matching FCD. In the proposed MDP-MM algorithm, the MDP technique was used to minimize the number of candidate routes maintained at each GPS point, while guaranteeing to determine the best matching route. They stated that the MDP-MM algorithm was competitive with existing algorithms in both accuracy and computational performance. Other research efforts can be found in He et al. (2013). However, most of the above methods usually fail to directly identify vehicle trajectories in the mixed corridor as two alternative routes of the mixed corridor, i.e., the ground arterial and the elevated expressway, are basically overlapped each other in the digital road map.

The existing methods to identify vehicle trajectories in the mixed corridor could be divided into two categories. The first category is to utilize the auxiliary information gained from the GPS such as the altitude data together with a 3D digital road network map to directly recognize probe vehicle trajectories. For example, Taylor et al. (2001) developed Road Reduction Filter (RRF) algorithm and Virtual Differential GPS (VDGPS) with height information aiding to improve the accuracy of map matching. Moreover, machine vision algorithms made a new direction to traffic monitoring by utilizing GPS data together with a new source of information, i.e., a geographical 3D model of the environment. Map matching was conducted by matching the virtual image provided by the 3D GIS (Geographic Information System) and the real image acquired by the onboard cameras (Quddus et al., 2007). However, such approaches have been rarely applied in the real world so far, as most of the commercial on-board navigation systems are unable to collect accurate altitude information and thus only provide the post-processed horizontal GPS positions (Cappelle et al., 2010).

The other category is to use the normal 2D GPS data which is commonly available from the current probe vehicle systems and apply typical clustering algorithms such as K-Means or decision rules based on topological relations of the mixed traffic corridor and travel speed threshold values to divide the vehicles running on the elevated expressways and ground arterials. Miwa et al. (2008) developed a shortest path based on map matching algorithm to identify vehicle trajectories in the mixed traffic corridor, in which the elevated expressways were assigned a higher weight if the sample vehicle traveled faster than a previously set speed threshold. It is an algorithm combined topological relations with constant travel speed threshold values. Zhu et al. (2012) established a model dividing the expressways into straight roads and interchanges, and aimed at precise matching for interchange areas. Schroedl et al. (2004) applied K-Means clustering algorithm for inducing high-precision road network map to directly recognize probe vehicle trajectories. For example, Taylor et al. (2001) developed Road Reduction Filter (RRF) algorithm and Virtual Differential GPS (VDGPS) with height information aiding to improve the accuracy of map matching. Moreover, machine vision algorithms made a new direction to traffic monitoring by utilizing GPS data together with a new source of information, i.e., a geographical 3D model of the environment. Map matching was conducted by matching the virtual image provided by the 3D GIS (Geographic Information System) and the real image acquired by the onboard cameras (Quddus et al., 2007). However, such approaches have been rarely applied in the real world so far, as most of the commercial on-board navigation systems are unable to collect accurate altitude information and thus only provide the post-processed horizontal GPS positions (Cappelle et al., 2010).

As a new direction in machine learning research, the semi-supervised learning technique has been widely applied in picture processing, text classification, etc. Unlike the conventional clustering algorithms such as K-Means, the semi-supervised learning technique is a sort of advanced clustering method that makes use of both labeled and unlabeled data, typically a small amount of labeled data with a large amount of unlabeled data. The semi-supervised clustering algorithms have also been adopted to deal with similar issues in the field of transportation. Melo et al. (2006) applied the semi-supervised K-Means clustering technique to classify highway lanes. Kianfar and Edara (2010) used it to optimize sensor locations on freeways by clustering speed and travel-time data obtained from the GPS-equipped probe vehicles. Wang et al. (2013) incorporated it into the Delphi method to categorize the level-of-service for Beijing’s urban expressways. However, no study was found in literature to apply the semi-supervised learning technique in identifying probe vehicle trajectories in the mixed traffic corridor. As discussed previously, a part of probe vehicle trajectories in the mixed corridor could be directly identified via topological relations and that part of successfully identified probe vehicles could be regarded as labeled data in the semi-supervised learning technique so as to improve clustering performance. Based on this idea, this study innovatively developed three enhanced semi-supervised clustering algorithms based on the original semi-supervised learning technique to identify probe vehicle trajectories in the mixed corridor.

3. Methodology

Although inherent feature of the semi-supervised learning technique matches with the objective of this study, the original semi-supervised learning technique has two limitations which significantly constrain its capability in solving the problem of interest.

Firstly, it is incapable of optimizing the selection of pre-labeled samples and thus only works with a given collection of pre-labeled samples. As discussed earlier, topological relations can be used to identify probe vehicle trajectories in the mixed corridor as most of the existing clustering algorithms do. In addition, mean stopping time of probe vehicles can provide supplementary information contributing to the identification of probe vehicle trajectories. The underlying rationale is that, due to the existence of intersections, vehicles running on the ground arterials usually experience a larger number of stops than those on the elevated expressways during the off-peak time periods, and are more possible to encounter longer mean stopping time during the peak time periods when both the elevated expressways and the ground arterials get congested.
Utilizing mean stopping time together with topological relations is possible to enlarge the collection of pre-labeled samples so as to improve clustering performance.

Secondly, in the original semi-supervised learning technique, the determination of initial clustering centers is merely based on the samples’ distances to the clustering center, i.e., the travel speed differences between each sample and the speed at the clustering center in the problem of discussion. However, travel speeds of vehicles on the elevated expressways usually have smaller variance within a short time interval, even during a building or dissolving stage of traffic congestion, owning to traffic flow characteristics of uninterrupted flow facilities. Therefore, the density of samples representing the concentration level of samples could help in determining the initial clustering centers too.

In light of the above thoughts, this paper proposed three enhanced semi-supervised clustering algorithms based on the original semi-supervised learning technique, namely the Constrained-K-Means (CKM), the Seeded-K-Means (SKM) and the Semi-Supervised Fuzzy c-Means (SFCM) (Basu, 2005; Bezdek, 1981). The CKM and SKM algorithms are based on the clustering algorithm of K-Means, both of which improve the clustering performance by utilizing labeled samples to optimize the initial clustering centers. Their differences lie in that the CKM algorithm keeps the original class of labeled samples in the iteration of clustering while the SKM algorithm has no constraint of that. The SFCM algorithm applies membership function to describe the degree of samples’ belongingness. The clustering center and membership function which are deduced from objective function are updated in each iteration until convergence. The SFCM algorithm utilizes labeled samples to optimize initial clustering centers and keep the membership function of labeled samples as 0 or 1. The objective function is:

\[
J(U, c_1, c_2) = \sum_{h=1}^{2} J_h = \sum_{h=1}^{2} \sum_{i=1}^{n} u_{ih}^2 ||x_i - c_h||^2
\]

where \( x_i \) is the \( i \)th sample; \( c_h \) is the clustering center of group \( h \), i.e., \( h = 1 \) for the group of vehicles on the ground arterials and \( h = 2 \) for the group of vehicles on the elevated expressways; \( u_{ih} \) is the membership function with the following constraint:

\[
u_{ih} \in [0, 1], \sum_{h=1}^{2} u_{ih} = 1 \tag{2}
\]

The proposed algorithms are able to optimize the selections of pre-labeling samples and clustering centers of the original semi-supervised clustering technique, by integrating topological relations and mean stopping time in identifying the pre-labeled samples as well as applying a weighted function of density and distance in optimizing initial clustering centers. The inputs of the algorithms are the clustering samples collection and the initial labeled collection of vehicles on the ground arterials. The outputs are two classified collections, i.e., the ground arterial category and the elevated expressway category. Major components of the proposed algorithms, including preparing sample vehicle data, pre-labeling sample vehicle data, optimizing initial clustering centers and solution algorithm, are explained in detailed as follows.

3.1. Preparing sample vehicle data

3.1.1. Mixed corridor

A mixed traffic corridor can be divided into the ramp areas and the between-ramps areas. Ramp area may be an on-ramp or an off-ramp area and the between-ramps area represent the parallel road segments between two adjacent ramps in the direction of driving. Fig. 1 illustrates the mixed traffic corridor and the between-ramps areas in an example digital map. In Fig. 1(a), the deep lines indicate the mixed corridors and the light lines present ground roads, and the dotted ellipses refer to the between-ramps areas. Fig. 1(b) shows a series of recorded GPS plots of a sample probe vehicle traveling from the north to the south in one of the between-ramps areas. Zooming in the top part of Fig. 1(b), Fig. 1(c) exhibits the detailed road configuration, in which the two inner lines are the elevated expressways and the two outer lines are the ground arterials.

3.1.2. Time window

For the sake of adapting dynamic traffic state estimations, vehicle samples need to be identified within a short time interval. This study adopted a 5-min time window based on the considerations of obtainable sample size and traffic state variations. Due to the fixed time window, a vehicle may drive in a between-ramps area within one time window or multiple time windows. Fig. 2 exhibits time-dependent speeds of a sample vehicle in a between-ramps area. The dotted lines are two adjacent time windows from 7:00 am to 7:10 am that divide the vehicle trajectory into two clustering sets based on the boundaries of the between-ramps area. The left clustering set contains those speed plots during the time window of 7:00–7:05, which are used in clustering for this between-ramps area and this time window. Similarly, the speed plots on the right are used for the time window of 7:05–7:10. In this way, speed plots of all the sample vehicles can be grouped for each between-ramps area on the basis of 5-min time window.

3.1.3. Data filtering

Typically, very few speed plots can be found for a specific time window if the sample vehicle entered the between-ramps area at the end of the time window or left at its beginning. The limited number of speed plots as well as short traveling distance may lead to unreliable results in the estimation of travel speed for the whole between-ramps area and thus need to be
excluded from the clustering. In such cases, a larger number of speed plots are likely to appear in the adjacent time window as shown in Fig. 2. Therefore, both the number of speed plots and the total traveling distance were used for data filtering in order to pick up the right time window that includes more speed plots. The filtering rule is explained by Eq. (3).

\[
\text{if } N_i < \text{Median}(\bigcup_{j=1}^{n} N_j) \quad \text{and} \quad D_i < \text{Median}(\bigcup_{j=1}^{n} D_j) \quad \text{then}
\]

where \( N_i \) and \( D_i \) are the number of speed plots and the total traveling distance for sample vehicle \( i \) of a clustering set; \( \text{Median}(\bigcup_{j=1}^{n} N_j) \) and \( \text{Median}(\bigcup_{j=1}^{n} D_j) \) represent the median values of \( N \) and \( D \) of the entire population observed in a between-ramps area that may include multiple clustering sets.

3.2. Pre-labeling sample vehicle data

3.2.1. Topological relations

Topological relations of digital road network could help to identify the vehicle trajectories running on the ground arterial in the mixed corridor. Specifically, a vehicle traveling from a crossing road to the mixed corridor area must be on the ground arterial until it encounters an on-ramp, as shown in Fig. 3(a); vice versa, a vehicle traveling from the mixed corridor to a crossing road must be on the ground arterial after it leaves an off-ramp.
However, a small proportion of samples can usually meet the requirements of the above judgment conditions. In other words, topological relationship judgment may fail in most cases. Fig. 3(b) shows an unsuccessful example. A vehicle traveling in the between-ramps area, whose route before entering the subject corridor area could not obtained within one time window. It is unable to identify the vehicle's trajectory through the above judgment conditions as the target vehicle did not appear on any crossing road of the mixed corridor.

3.2.2. Mean stopping time

As explained earlier, Mean Stopping Time (MST) can also be used to pre-label other possible samples after the step of pre-labeling samples based on topological relations. In this study, a dynamic threshold value of MST that is determined as the average MST of the pre-labeled samples by topological relationship judgment for each time window is used to label the vehicles on the ground arterials. If a sample vehicle’s MST is larger than the threshold value, then it is more likely to travel on the ground arterials, as given in Eq.(4).

$$MST_i > \frac{1}{|S|} \sum_{j \in S} MST_j$$

where $MST_i$ is the mean stopping time for sample vehicle $i$; $S$ is the set of pre-labeled samples based on topological relationship judgment; $|S|$ is the total number of samples in the set.

A part of vehicles traveling on ground arterials could thus be pre-labeled based on mean stopping time and topological relations, which are regarded as the labeled samples in the proposed algorithms. All the labeled samples belong to the category of ground arterials. The remaining part of sample vehicles are regarded as unlabeled samples, which are composed of the rest of vehicles driving on the ground arterials and all the vehicles traveling on the elevated expressways. The following explains how to cluster the unlabeled samples based on the successfully labeled samples.

3.3. Optimizing initial clustering center

3.3.1. Initial clustering center of ground arterial category

Initial clustering center is of great importance for clustering efficiency (Steinley and Brusco, 2007). The initial center of the ground arterial category is the average of the labeled samples as given in Eq. (5). Mean Speed (MS) and Standard Deviation of Speed (STDS) were used as clustering variables that can represent travel speed differences between the group of vehicles on the ground arterials and those on the elevated expressways.

$$c_1 = \frac{1}{|S|} \sum_{i \in S} x_i$$

where $c_1$ is the initial clustering center for the group of vehicle on the ground arterials; $S$ is the labeled samples collection of the arterial category; $x_i$ is the $i$th sample, which is a two-dimensional variable [MS, STDS].

3.3.2. Initial clustering center of elevated expressway category

As explained previously, density can act as another important parameter to optimize the initial clustering centers in addition to distance. Thus, a weighted function of density and distance was proposed to determine the initial clustering centers of the elevated expressway category, as shown in Eqs. (6) and (7).
\[ c_2 = x_k, k = \arg \max \{ f \}, \]

where \( c_2 \) is the initial clustering center of the elevated expressway category; \( x_k \) is the sample owing the maximum \( f \), which is the weight function for vehicle \( i \) and calculated based on Eq. (7).

\[ f_i = \alpha \cdot \text{density}_i + (1 - \alpha) \cdot \text{distance}_i \]

where \( \text{density}_i \) is defined as the number of data points located in the circle with a center of \( c_2 \) and a diameter of \( r_v \), which is determined as the median value of the distances between the given vehicle \( i \) and all the other sample vehicles, as illustrated by Eq. (8); \( \text{distance}_i \) is the distance of vehicle \( i \) to \( c_1 \) defined by Eq. (9), which is to ensure the intra-cluster distance between \( c_1 \) and \( c_2 \) to be far as much as possible; \( \alpha \) is the weight coefficient of density, ranging between 0 and 1. \( \alpha = 0 \) represents that the distance dominated the initial clustering center, while \( \alpha = 1 \) refers to that the density is the dominant factor.

\[ r_i = \text{Median} \left( \sum_{j=1}^{n} d(x_i, x_j) \right), \quad i \neq j \]

where \( \text{Median} \left( \sum_{j=1}^{n} d(x_i, x_j) \right) \) refers to the median value of \( d(x_i, x_j) \) of the entire population; \( d(x_i, x_j) \) is the Euclidean distance between sample \( x_i \) and \( x_j \).

\[ \text{disatance}_i = d(x_i, c_1) \]

where \( d(x_i, c_1) \) is the Euclidean distance between sample \( x_i \) and \( c_1 \).

### 3.4. Solution algorithm

The proposed algorithms can be solved by applying the calculation procedures of the original semi-supervised learning techniques presented in Table 1, with the proposed optimization methods for pre-labeling samples and determining initial clustering centers.

### 4. Data collection and reduction

#### 4.1. Site descriptions and data collection

Two mixed corridors on Shanghai urban expressways were selected as study sites. The first site, i.e., A, is located in the South-North Elevated Expressway and the second site, i.e., B, is located in the Inner-Ring Elevated Expressway, as shown in Fig. 4. A is approximately 1.7 km long in total and includes an off-ramp, an on-ramp, and 5 ground intersections. B is about 0.85 km long and includes an on-ramp, an off-ramp and 2 ground intersections. The speed limits are 50 km/h for the ground arterials and 80 km/h for the elevated expressways. Four scenarios representing typical situations defined by the length of between-ramps area, number of intersections and time periods were included in the evaluation works, as summarized in Table 2.

### Table 1

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Calculation procedures</th>
</tr>
</thead>
</table>
| **The Constrained-K-Means algorithm (CKM)** | 1. Initialize clustering center: using aforementioned initialization method  
2. Assign cluster: For \( x_i \in S \) assign \( x_i \) to the cluster ground road \( X^{(1)}_b \), for \( x_i \notin S \), assign \( x_i \) to the cluster \( X^{(2)}_b \), for \( h' = \arg \min_{h \in \{1,2\}} \sum_{x_i \in X^{(h)}_b} d(x_i, x_j) \)  
2a. Update center: \( c^{(h+1)}_b = \frac{1}{n^{(h+1)}} \sum_{x_i \in X^{(h)}_b} x_i \)  
2b. Repeat until convergence (\( t = t + 1 \))  |
| **The Seeded-K-Means algorithm (SKM)** | 1. Initialize clustering center: using aforementioned initialization method  
2. Repeat until convergence (\( t = t + 1 \))  
2a. Assign cluster: Assign each data point \( x_i \) to the cluster \( X^{(2)}_b \), for \( h' = \arg \min_{h \in \{1,2\}} \sum_{x_i \in X^{(h)}_b} d(x_i, x_j) \)  
2a. Assign cluster: Assign each data point \( x_i \) to the cluster \( X^{(2)}_b \), for \( h' = \arg \min_{h \in \{1,2\}} \sum_{x_i \in X^{(h)}_b} d(x_i, x_j) \)  
2b. Update center: \( c^{(h+1)}_b = \frac{1}{n^{(h+1)}} \sum_{x_i \in X^{(h)}_b} x_i \)  
2b. Repeat until convergence (\( t = t + 1 \))  |
| **The Semi-Supervised Fuzzy c-Means algorithm (SFCM)** | 1. Initialize clustering center: using aforementioned initialization method  
2. Repeat until convergence (\( t = t + 1 \))  
2a. Update \( u \): For \( x_i \in S \) keep \( u^{(2)}_i = 1 \), \( u^{(1)}_i = 0 \); For \( x_i \notin S \), update \( u^{(1)}_i, u^{(2)}_i = \frac{1}{\sum_{k=1}^{2} \frac{1}{d(x_i, c^{(k+1)}_{b})}} \)  
2b. Update center: \( c^{(h+1)}_b = \frac{1}{n^{(h+1)}} \sum_{x_i \in X^{(h)}_b} u^{(h+1)}_i x_i \)  
3. Assign \( x_i \) to the cluster \( X^{(h+1)}_b \), for \( h' = \arg \max_{h \in \{1,2\}} \{ u \} \)  |
4.2. Data reduction

To reduce data reduction efforts and without the loss of generality of the proposed methodology, a 30 min of GPS data was collected for each scenario, i.e., 6 time windows. GPS data received from taxies on April 1, 2011 was used in the clustering analysis. The taxi’s GPS data updated every 10 s and mainly included the information of taxi ID, longitude coordinate, latitude coordinate, velocity, heading angle and global time. With the uploaded GPS data, a map-matching process was firstly done based on the used GPS data of one day and the digital map of Shanghai shown in Fig. 4. A rough estimation of GPS data errors was then conducted and results showed that the positioning errors are smaller than 50 m, which ensures that there are only two alternative routes in the study mixed corridors. Afterwards, a program based on the C-language was coded to extract the trajectories of probe vehicles running on the study corridors. The program was able to automatically pick up all the GPS data plots of each sample vehicle crossing the study areas within the analysis time periods according to its longitude and latitude coordinates and global time. Those GPS data plots were then linked together to form complete vehicle trajectories. Eventually, the total number of obtained taxi samples was 138, 122, 61 and 37 for the four scenarios respectively, averagely 23, 21, 10 and 6 for each time window. Considering the duration of time window, i.e., 5 min, the sample sizes are generally acceptable as they are normally sufficient to support travel speed estimation. The percentages of the successfully pre-labeled vehicles were 44%, 61%, 18% and 43% for Scenario 1, 2, 3 and 4 respectively.

Note that it is difficult to measure real trajectories of probe vehicles as it is very costly and the uploaded taxi ID is a random code that cannot match with the real number plate of the vehicle due to privacy issue. Hence, those probe vehicles actually running on arterials or expressways were identified by a detailed manual tracking of the vehicles’ complete trajectories in a broad area for more than 30 min, i.e., before and after the analysis time window. A vast majority of sample vehicles included in the analysis were successfully classified based on topological relationships of the subject road network explained in Section 3.2.1.
5. Online evaluation of the proposed algorithms

5.1. Performance indicators

Clustering Purity and Normalized Mutual Information (NMI) were selected to assess the performance of the proposed algorithms. Both of them are commonly used indicators in the evaluation of clustering algorithms (Melo et al., 2006; Wang et al., 2013; Steinley and Brusco, 2007). The first indicator simply represents the hit ratio of the correctly classified samples, as defined in Eq. (10).

\[
Purity = \frac{n_c}{n}
\]

where \(n\) is the total number of samples and \(n_c\) is the total number of correctly classified samples.

High purity is easy to achieve when the number of clusters is large. In particular, purity is 1 if each document gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters. A measure that allows us to make this tradeoff is Normalized Mutual Information or NMI, as defined in Eq. (11).

\[
I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)
\]

where \(X\) is the collection of clustering results and \(Y\) is the collection of correct classification; \(p(x)\) and \(p(y)\) represent the probability of samples belong to the class arterial and expressway in \(X\) and \(Y\) respectively; \(p(x, y)\) is the probability of correct classification.

5.2. Results analysis

It is worth mentioning in advance that as the sensitivity coefficient of clustering center optimization \(\alpha\) might be sensitive to the clustering accuracy indicated by clustering purity and NMI and thus needs to be analyzed concurrently. Table 3 compares the clustering accuracies of the three enhanced algorithms under the defined four scenarios, with \(\alpha\) ranging between 0 and 1.

It was found that all the three algorithms could obtain very high, approximately 100%, clustering purity for the off-peak time periods shown in Scenario 2 and Scenario 4 and their performance was very insensitive to \(\alpha\) especially when \(\alpha\) is larger than 0.2. There is almost no discrepancy between the performance of the SFCM and the CKM in terms of clustering purity, while the purity of the SKM is slightly low. In addition, the NMI of SFCM and CKM reached 100% in Scenario 2 and 95% in Scenario 4 when \(\alpha\) is larger than 0.2. However, that of SKM remains at 80% for both of the scenarios.

Furthermore, the performance of the three algorithms was found to slightly decrease and become more sensitive to the value of \(\alpha\) for the peak time periods shown in Scenario 1 and Scenario 3. The clustering purity and NMI basically rise as the value of \(\alpha\) increases. Meanwhile, the SFCM algorithm performed best and its maximum purity could reach approximately 95% and 80% under Scenario 1 and Scenario 3 respectively, if \(\alpha\) was no less than 0.5. Under the same situations, the purity of CKM and SKM are 80% and 70% respectively in Scenario 1 and while both of them are 80% in Scenario 3, close to that of SFCM. In contrast, their NMI showed more remarkable differences when \(\alpha\) is larger than 0.5. More specifically, the NMI of SFCM stayed at 0.7, CKM at 0.5, and SKM at 0.3 in Scenario 1, while they became 0.5, 0.4 and 0.3 respectively in Scenario 3.

Regarding time consuming performance, it was found that all the three algorithms and its map-matching process could be completed within 0.1 s for all scenarios on a laptop computer and the converging speed of CKM was the fastest, followed by SKM and SFCM. Hence, it can be expected that the proposed algorithms could facilitate short-term (e.g., 5 min) traffic state estimations for a large scale road network with a large dataset of probe vehicles as well, aided by the parallel computing technique and more powerful computers.

Overall, the SFCM algorithm performed best in terms of purity and NMI for all the scenarios. With the SFCM algorithm and the optimal value of \(\alpha\), Fig. 5 further exhibits a time series analysis of travel speeds on the elevated expressways and the ground arterials under Scenario 1, in order to look into the resulted travel speed estimation accuracy via a comparison of the successfully clustered samples and the real ones. It can be found that the SFCM could produce fairly good clustering results close to the measured, which is consistent with the conclusions discussed previously.

Comparisons with the traditional K-Means clustering algorithm and the basic SFCM algorithm were also conducted to further look into the superiority of the proposed SFCM algorithm with the optimal value of \(\alpha = 0.6\). In the traditional K-Means algorithm, no pre-labeled sample was used and the clustering process was completed merely based on the two-dimensional variable, i.e., \(x_i = [MD, STDs]\). In the basic SFCM, only the successfully identified vehicles on ground arterials based on topological relations were used as pre-labeled samples and there was no optimization in the step of initializing clustering center. The results are presented in Fig. 6. It showed that the enhanced SFCM algorithm could improve clustering purity averagely by 21% and 14% for the congested conditions represented by Scenario 1 and 3, and 6.5% and 6% for the uncongested conditions represented by Scenario 2 and 4, as compared with the traditional K-Means algorithm and the basic SFCM. It well demonstrates the efficiency of the proposed optimizations in pre-labeling samples and initializing clustering centers in the enhanced SFCM algorithm.
5.3. Discussions

5.3.1. Clustering accuracies under various traffic conditions

For the uncongested conditions represented by Scenario 2 and Scenario 4, the sensitivity parameter $\alpha$ was found to be almost irrelative to the clustering accuracy. It translates that the initial clustering center has little influence on the final clustering results for the uncongested conditions, i.e., both density and distance are effective in identifying probe vehicle trajectories. Most of the sample vehicle trajectories could be identified according to the travel speed differences between the ground arterials and the elevated expressways.
However, the travel speed differences were smaller under the congested conditions represented by Scenario 1 and Scenario 3. It might raise the possibility of clustering failure. In some cases, the two collections of travel speeds could actually come together and then cross over within a same time window, e.g., starting with expressway faster and ending with arterial faster. Such an effect is more pronounced in Scenario 1 due to a larger number of intersections in area A, leading to a higher probability of stopping. A high frequency of stopping could not only increase the number of labeled samples based on mean stopping time but also enlarge the travel speed differences between the ground arterials and the elevated expressways. This is perhaps the reason why the clustering performance is relatively good in Scenario 1, as compared with Scenario 3.

Furthermore, the clustering accuracy under the congested conditions was relatively high when \( \alpha \) became closer to 1.0, i.e., density appeared to be the dominant factor. A possible explanation is that the vehicles traveling on the elevated expressways have more uniform speeds and the distribution of speeds is thus more concentrated. Consequently, the initial clustering center tends to appear in the vicinity of the sample vehicles traveling on the elevated expressways.

5.3.2. The role of historical trajectory data

In the proposed algorithms, vehicle trajectory identification proceeded with a time step of 5 min and its corresponding clustering set, as illustrated in Fig. 2. The probe vehicle trajectories included in the adjacent clustering sets are correlated with each other. For instance, the trajectory on an individual between-ramps area could be divided into two clustering sets if the length of between-ramps area was long as shown in the figure; then, contradictory results, i.e., one part of the trajectory was identified on the elevated expressway and the other part on the ground arterials, would occur if without bringing the recognizable historical trajectory based on the topological relations in the previous time step into the current time step. Therefore, in this study, the probe vehicle trajectory in the previous time step was also utilized together with the trajectory information available from the connected ramp areas.

However, there is still a potential to improve the proposed algorithms by the use of historical trajectory data of a longer time period. The SFCM algorithm could allocate a fuzzy ratio to each route for each time step, which can be considered as a probability that the sample vehicle is running on this route. Therefore, the accumulative historical trajectory data over a long time period is possible to pave a way to improve the current SFCM algorithm, based on a maximum likelihood method.
As depicted by the gray dotted arrows in Fig. 7, the sample vehicle could be identified on the ground arterial in the time step of 7:00–7:05, based on its bigger fuzzy ratio of 0.7. However, the same vehicle could be identified on the elevated expressway in the next time step, i.e., 7:10–7:15, as the accumulative likelihood of two adjacent time steps is 0.24 (i.e., $0.3 \times 0.8$) for the elevated expressway and is larger than that of the ground arterial 0.14 (i.e., $0.7 \times 0.2$). Based on this idea, it is possible to calculate the cumulative probability for each route and then infer real trajectory of the sample vehicle via maximum likelihood estimation.

In summary, the proposed methodology is applicable for both the on-line and the off-line applications of probe vehicle systems. For the on-line applications, e.g., short-term dynamic traffic state estimation, the clustering process can be completed within a short time step, i.e., 5 min, and the clustering results can then be used for travel speed estimation for the defined time step. Moreover, the proposed clustering method doesn’t require the accumulation of historical trajectory data, which most of the current probe vehicle systems rely on. For the off-line applications, e.g., OD estimation or long-term traffic state estimation, the identified trajectories in every time step could be easily stored in a database and then used for more precise traffic state estimation and route choice analysis. In addition, along with the maximum likelihood method explained above, the proposed methodology could be expected to provide more reliable classification of probe vehicle trajectories.

6. Conclusions and future works

This paper proposed three enhanced semi-supervised clustering algorithms, i.e., Constrained-K-Means (CKM), Seeded-K-Means (SKM), and Semi-Supervised Fuzzy c-Means (SFCM), to classify probe vehicle trajectories in the mixed traffic corridor based on horizontal GPS data. The proposed algorithms can integrate topological relation judgment and the semi-supervised learning technique by optimizing the selection of pre-labeling samples and initial clustering centers of the original semi-supervised learning technique based on horizontal GPS data. They were validated and evaluated based on the probe vehicle data collected at two mixed corridors on Shanghai’s urban expressways.

Results indicate that the enhanced SFCM algorithm could achieve the best performance in terms of clustering purity and NMI. It may reach a nearly 100% clustering purity for the uncongested conditions and a clustering purity greater than 80% for the congested conditions. Meanwhile, it could improve clustering purity averagely by 21% and 14% for the congested conditions and 6.5% and 6% for the uncongested conditions, as compared with the traditional K-Means algorithm and the basic SFCM. The CKM algorithm was slightly worse than the SFCM, while it had faster converging speed which is beneficial for the real-time application. The SKM algorithm was the worst in terms of clustering purity and NMI. The sensitivity coefficient $\alpha$ for the clustering center initialization was found to have significant impacts on the performance of the algorithms. Clustering purity and NMI could reach the best when $\alpha$ became greater than 0.5. Scenario analysis results showed that better performance of the proposed algorithms could be achieved for uncongested conditions and long between-ramps area including a large number of intersections. The proposed algorithms can be applied for both on-line and off-line purposes, without the need of historical data.

The proposed algorithms need to be reinforced and improved in future in order to facilitate widespread applications. Firstly, the clustering accuracy could be further improved if the historical data can be utilized to compensate the number of samples when the current identifiable samples are insufficient; Secondly, the proposed algorithms should be tested under various aggregate time intervals so as to find out the best time interval to adopt the proposed algorithms; Thirdly, the algorithms should be modified and extended to cover more complicated road networks such as large multi-level interchanges that exist in some of Chinese cities.

Acknowledgements

This study was supported by National Key Technology Support Program (No. 2014BAG03B02). The authors would like to convey sincere appreciations to Shanghai Yootu Information Technology Co., Ltd. for providing probe car data used in this study.
References


