Detecting Travel Modes Using Particular Rules Combined with a Naïve Bayesian Classifier

Guangnian Xiao, Corresponding Author
Shanghai Jiao Tong University
535 Fahuazhen Road, Shanghai, China 200052
Tel: +86 21 5230-1396; Fax: +86 5230-1396; E-mail: xiaoguangnian@sjtu.edu.cn

Zhicai Juan
Shanghai Jiao Tong University
535 Fahuazhen Road, Shanghai, China 200052
Tel: +86 21 5230-1396; Fax: +86 5230-1396; E-mail: zcjuan@sjtu.edu.cn

Jingxin Gao
Shanghai Jiao Tong University
535 Fahuazhen Road, Shanghai, China 200052
Tel: +86 21 5230-1396; Fax: +86 5230-1396; E-mail: jausnogao@sjtuits.net

word count: 6,106 words text + 7 tables/figures x 250 words (each) = 7,856 words

Submission Date: 2014/10/26
ABSTRACT
Travel modes are often detected from GPS positioning data with two types of methods, i.e., particular rules and advanced classification methods. The former is intuitive and easily applied, while the latter incorporates relatively preferable generalization and flexibility. However, studies touching the combination of these two methods are limited.
In this study, we employ particular rules to detect subway trips and then utilize a naïve Bayesian classifier with sequential forward selection procedures to detect the remaining modes. Depending on four representative attributes included in the optimal attribute set, more than 98% of subway trips are correctly identified and a total accuracy of 94.85% is achieved. Notably, more than 91% of bus trips and car trips are accurately identified in either training set or test set. The results provide us with a new perspective to select classification methods in detecting travel modes and even other travel characteristics in GPS travel surveys. In addition, a high distinction between car trips and bus trips without GIS sources of bus networks indicates that GIS sources can be replaced to a certain extent by extracting reasonable attributes.
Key Words: travel modes; particular rules; naïve Bayesian classifier; sequential forward selection
INTRODUCTION

Over the past decade, smartphones and dedicated GPS devices have been increasingly utilized to collect location-based data streams in GPS travel surveys, which are wildly regarded as a promising alternative to conventional travel surveys. Low burden in respondents (1), high accuracy in GPS positioning data and rich information included in the data collected merit this type of travel survey method. However, some important trip attributes, such as trip rates, travel modes and trip purposes, cannot be directly obtained from the GPS data. Therefore, it is difficult to explore travel behavior with GPS data alone. According to existing studies, researchers employ particular rules to identify trip ends and utilize GIS sources to infer trip purposes in most cases. In contrast, transportation mode is detected with either particular rules or advanced classification methods. In addition, some studies detecting travel modes take advantage of GIS sources, while others use GPS data alone. A common view about which type of inference method is preferred and whether GIS sources should be utilized in detecting travel modes has not been achieved.

To date, most rule-based imputation algorithms in detecting travel modes apply speed-related measures, such as the average speed, acceleration and percentile speed. Stosopher et al. (11) took three steps to mark a single-mode segment with walk, bicycle, bus or car. First, they identified a segment with an average speed of less than 6 km/hr and a maximum speed of less than 10 km/hr as a walk segment. Second, when a segment matches bus networks, presents periodical stops and incorporates an average speed between 10 and 40 km/hr, it is flagged as a bus trip. Third, for any travel segment to be determined, it is marked with a bicycle mode when the average speed is less than 40 km/hr, otherwise it is identified as a car trip. According to the particular rules as above, they correctly detected about 95% of single-mode segments. It is thus clear that rule-based algorithms are both highly understandable and significantly effective in some cases.

Advanced classification methods, such as fuzzy logic regression and decision tables, are also popular in detecting travel modes. Byon et al. (14) extracted features like the average number of satellites in view, average accuracy of HDOP, speed and acceleration and made a distinction among four types of travel modes with a classification technology of neural network. In order to decrease adverse effects of GPS positioning error on mode inference, Joseph et al. (15) adopted measures of median speed, 95% percentile speed and 95% percentile acceleration and employed a MNL model to automate the four types of transportation modes. As a result, they achieved an accuracy of 87.5%. A recent study by Feng et al. (16) compared particular rules with advanced classification methods and concluded that the latter was potentially more flexible and robust in handling issues in different contexts.

Another issue to be considered is whether GIS sources are essential or irreplaceable in detecting travel modes. Joseph et al. (15) compared mode detection accuracy using GPS data alone with that using both GPS and bus network GIS data. They improved detection accuracy from 87.5% to 89.6% by the inclusion of bus
network layers and concluded that bus network data can obviously improve the accuracy of model fits and predictions. Tsui et al. (12) took into consideration availability of GIS data and developed two types of inference methods to detect transportation modes. The first one was developed without a GIS network, while the second one required the intervention of a GIS network. To overcome the disadvantages in detecting transportation modes when GIS data is not available, another study by Zheng et al. (17) extracted several advanced feathers, including the heading change rate, the stop rate and the speed change rate, to improve discrimination performance. These feathers are applied due to the fact that non-motorized segments change heading more frequently than motorized segments and that a bus segment is characterized by periodical stops. With these features, they utilized Decision Trees to match segments with five types of travel modes and achieved a high detection accuracy. Additionally, since no GIS data is involved, an application is developed and employed on the web successfully. Similarly, Schuessler et al. (3) also developed algorithms for detecting travel modes from GPS data streams without any additional information and the algorithms achieve a relatively high accuracy.

As a whole, rule-based algorithms provide an intuitive understanding and a practical method for mode inference; however, the algorithms are struggled with ambiguity among different travel modes. For example, a low-speed bicycle trip might be falsely identified as a walk trip when the average speed is taken as a critical feature to differentiate these two modes. In contrast, advanced classification algorithms can deal with the ambiguity among distinct modes. For example, a naïve Bayesian classifier can exploit value of each feature and priori information sufficiently. More importantly, it is exempt from searching for optimal parameters compared with a Bayesian network classifier. In fact, rule-based algorithms are highly suitable for distinguishing a travel mode that is obviously different from others for a specific feature or several features together. However, little has been studied to explore the possibility of a combination of rule-based algorithms and advanced classification methods in inferring travel modes. We, therefore, aim to combine these two types of methods to deal with issues of travel mode imputation in this study. Specifically, we first identify subway trips using a rule-based algorithm because subway trips match the metro network and most of them are encountered with serious signal loss. Afterwards, we employ a naïve Bayesian classifier to detect the remaining modes. In addition, we extract several targeted features instead of GIS sources to improve the detection accuracy of the classifier. Thus, two objectives of this study are a) to determine whether a combination of particular rules and an advanced classification method can outperform other commonly used methods and b) to determine whether GIS sources are essential or irreplaceable in detecting travel modes.

DATA DESCRIPTION

The data utilized to detect travel modes was collected in a smartphone-based travel survey launched at Shanghai city with three waves from mid-October 2013 to mid-July.
An application is exclusively developed for collecting location-based data by our research group prior to the survey. Android and iOS are selected as platforms to develop the application due to high marketing penetration rates. The application records UTC time, latitude, longitude, altitude, the instantaneous speed, heading, the number of satellites in view and HDOP once every second. Therefore, we present each respondent with an external battery package used to overcome battery drainage because GPS data is recorded in such a high frequency in this study. Also, the package is regarded as a motivation for respondents to participate in this survey. In fact, the application is closed when a stationary state are kept for more than five minutes and restarted when the smartphone move again automatically. This function is designed to decrease battery consumption to a maximum extent, with no adverse effect on normal location acquisition during trips.

Respondents are required to start the positioning application to automate GPS recording before the first trip and close it after the last arrival home every day during the survey period. The reason why the application is required to be closed is to decrease battery consumption and alleviate privacy issues. The positioning data should be uploaded to a database employed on our server. The respondents are recommended to upload the data in the environment of Wi-Fi for the sake of reducing data transmission cost. Afterwards, interviewers ask respondents to check primary travel information detected from the data streams uploaded by respondents and the information is corrected if necessary. In fact, respondents do not bear a heavy burden as expected since most travel information is correctly detected. Such validated travel information is regarded as ‘ground truth’ in the subsequent study. Each respondent is required to participate in the survey for at least five continuous days and 1512 person-day positioning data of 203 respondents are collected. After data screening according to completeness and logicality, 1248 person-day data streams are retained for subsequent analysis.

Walk, bicycle, e-bicycle, car, bus and subway are differentiated in our travel survey. All track data are split into single-mode segments according to reported travel information. A reported travel mode is assigned to each segment. In total, 5898 single-mode segments are extracted from the track data and separately counted according to travel modes. As shown in FIGURE 1, walk takes a maximum percentage with an absolute advantage. This result is due to the fact that walk trips usually play a transitional role when inter-modal transfers are applied. For example, one usually needs to walk to a subway station after takes off a bus when subway and bus are jointly applied for a trip. Subway takes the second largest percentage among all travel modes and undertakes more than 20% of the total trips. This percentage is due to a high travel time reliability and a well-connected large-scale network of the subway mode, which is especially outstanding in Shanghai city. Up to May 2014, 16 metro lines with over 300 stations and about 1000 entrances/exits have been in operation in Shanghai. Subway, therefore, is usually considered to be a preferable choice when it competes for passengers with other modes. It is thus obvious that a high detection accuracy of subway
provides an opportunity to effectively improve the overall detection accuracy of travel modes. There does not exist a significant difference of percentages among other travel modes. It should be pointed out that e-bicycle trips, which are usually regarded as non-motorized trips, take a relatively considerable percentage due to convenience and a low cost. Travel modes other than subway will be detected using a naïve Bayesian method after subway trips are detected with particular rules in this study.

![FIGURE 1 The Number of Segments with Six Reported Travel Modes.](image)

**METHODOLOGY**

In this study, detecting travel modes includes identification of subway trips using particular rules and identification of trips with the remaining transportation modes with a naïve Bayesian classifier. Before employing the classifier, we utilize sequential forward selection to achieve an optimal feature set from six representative features extracted from GPS positioning data streams.

**Detecting Subway Trips**

It is highly appropriate to detect travel modes of trips with subway using particular rules. Unlike a bus network, a subway network includes much less lines and stations. More importantly, the subway network in Shanghai does not overlap with roads on which vehicles drive in most cases. Moreover, the quality of GPS signal is much lower during subway trips than that during trips with other transportation modes. In general, there does not exist any GPS points during subway trips or only exists a small quantity of points. These distinctive characteristics reduce the possibility of confusing subway trips with trips taking other transportation modes. Intuitively, one might think that several representative features, such as distance from nearest metro lines, signal quality and the percentage of GPS points recorded, can be added to the feature set for implementing the classification to improve detection accuracy of subway trips instead of particular rules. However, this way of operation might increase computation cost significantly and even decrease detection accuracy of other modes considering the interaction of all
features on determining travel modes. Therefore, it is more reasonable to employ particular rules to detect subway modes than advanced classification methods. Two scenarios, i.e., without GPS signal and with incomplete GPS signal, need to be considered when travel modes are detected (As shown in FIGURE 2).

FIGURE 2 Subway Trips with Incomplete GPS Signal and without GPS Signal.

Subway trips without GPS signal are most often seen in reality. Subway stations and lines are located at underground or overground locations. Most metro lines incorporate both underground and overground parts. It is probable not to record any points for smartphones located in an underground metro train, while it is also likely not to receive the sufficient number of satellites (four satellites are enough in most cases) to position themselves for smartphones located in an aboveground metro train blocking GPS signal severely. Therefore, most cases are displayed like FIGURE 2 (a), in which GPS points are not recorded at all during a subway trip. A partially enlarged version of the subway trip is shown in FIGURE 2 (c), which shows that the GPS signal does not
recover until the respondent walk out of the station from exit 1. If a segment satisfies
the following conditions: a) the distance between the start point of the trip and a certain
entrance/exit is less than 100 meters and b) the distance between the end point of the
trip and a certain entrance/exit of a different station is less than 200 meters, we identify
it as a subway trip. In contrast, Tsui et al. (12) employed 630 meters as a critical distance
for both the start point and the end point. In fact, we collect coordinates of all
entrances/exits of each station other than the stations as a whole so that we can decrease
critical distance and maintain a high detection accuracy at the same time. A key
motivation to decrease the critical distances is to reduce the possibility of falsely
identifying non-subway trips as subway trips. Additionally, GPS signal can be received
until one walk into a station from a certain entrance, while it might not recover
immediately after one walk out an entrance/exit due to cold/warm start. Thus, the
critical distance for the end point is set to be twice as large as that for the start point.

Another issue to be handled is to detect a subway trip with incomplete GPS signal.
As shown in FIGURE 2 (b), GPS signal does not disappear immediately after the subway
start to run. The points recorded during the subway trip are located at two sides of the
subway line and the distance between these points and the subway line differs due to
GPS positioning error (as shown in FIGURE 2 (d)). A segment with incomplete GPS
signal is flagged as a subway trip when it satisfies the following conditions: a) it meets
all the requirements set to identify subway trips without GPS signal and b) all points
must be located in a distance of less than 30 meters from the subway line. Intuitively,
30 meters is considered as a relatively large critical distance as opposed to GPS
positioning error, which ranges from 5 to 10 meters in most cases. However, the error
might become large occasionally; on the other hand, as mentioned earlier, positioning
quality is adversely affected by blockage of subway trains. Such blockage improves the
possibility of recording inaccurate points. Therefore, the rules for detecting subway
trips can be summarized as two steps: (1) if a segment satisfies all the conditions for
detecting subway trips with signal loss, it is flagged as a subway trip; and (2) if a
segment whose travel mode is to be detected satisfies all the conditions for detecting
subway trips with incomplete signal, it is also flagged as a subway trip.

Feature Description

A feature set incorporating a high distinctiveness provides an opportunity to highly
improve the performance of classification. According to existing literatures, the average
speed, the average acceleration, the trip distance and the 95% percentile speed are
usually utilized to infer travel modes. However, it is hard to distinguish bus trips from
car trips when only aforementioned speed-related features are utilized (11; 18). In most
cases, the inclusion of a transit network layer can significantly improve the distinction
degree of bus trips from car trips (15). However, the transit network layer is not
available for most researchers, and we are now struggled with the same difficulty. In
addition, the bus network is updated every month and even every day at times in such
a megacity as Shanghai. It requires a timely update to effectively infer bus trips so that
it is costly to maintain an up-to-date GIS layer of the bus network. All these factors motive us to develop a new feature named as ‘low-speed point rate’, i.e., the rate of points with speed of less than 1 m/s, which is expected to capture the characteristics of periodical stops of buses. In fact, we investigate four types of critical speed, i.e., 0.5, 1.0, 1.5 and 2.0 m/s, for this feature. Ultimately, 1 m/s is applied since it can best differentiate bus trips from car trips. Another obstacle potentially decreasing detection accuracy of travel modes stems from a high level of uncertainty of speed-related features (19). For example, an e-bicycle trip with a relatively high speed is probably falsely identified as a bus or a car trip. However, a motorized vehicle can only drive on roads and does not usually change to a new lane unless necessary. In contrast, a respondent might usually stop for a certain reason or overtake others with greater randomness when he rides on a bicycle. Therefore, we extract a feature named ’heading change rate’, which is represented as the average heading change between any two adjacent points. In short, in addition to four commonly used features, two extra ones are applied to further differentiate bus trips from car trips and non-motorized trips from motorized trips.
Different features incorporate different distinction ability among travel modes. Overall, the probability mass function of the average speed and the 95% percentile speed are similar in the shape in the sense that the former seems to be prolonged about twice to form the latter on the horizontal axis (as shown in FIGURE 3 (a) and (b)). This similarity indicates the stability of these two feathers, which are different from the maximum speed prone to be disturbed by positioning errors. These two features seem to classify the five modes as three classes, i.e., walk as a single class, bicycle and e-bicycle as a class, the other two as a class. The average acceleration is another important feature utilized to detect travel modes. In terms of the average acceleration, e-bicycle, bus and car trips are distributed closely to each other, while they are highly different.
from other two modes (as shown in FIGURE 3 (c)). Such characteristics might potentially play an important role in distinguishing bicycle trips from e-bicycle trips. In terms of trip distance, there exists a high similarity between bus trips and car trips and a relatively low affinity among trips with three other modes (as shown in FIGURE 3 (d)). As expected, the ‘low-speed point rate’ can obviously distinguish walk trips from trips with other travel modes and differentiate bus trips from car trips to a certain extent (as shown in FIGURE 3 (e)). In addition, trips with motorized modes and non-motorized modes seems to present two distinct patterns for ‘heading change rate’. (as shown in FIGURE 3 (f)).

Detecting Travel Modes with a Naïve Bayesian Classifier
Feature selection is an important procedure to decrease the complexity and improve the prediction accuracy of the classification model. Most existing studies detect transportation modes directly based on pre-defined features. However, some features might be unnecessary for detecting travel modes. It is possible achieve higher detection accuracy by using less features. Therefore, we employ sequential forward selection (SFS) algorithms to select the most valuable attributes. In fact, the main objective of SFS is to choose appropriate features to minimize the total classification error. SFS is a bottom-up search method starting from an empty feature set \( S \) and gradually adds features from the remaining features that have not been added to \( S \). Each time only one feature can be added to \( S \) by an evaluation function, which refers to the overall detection accuracy in our study. If the inclusion of any feature in \( S \) does not lead to a more preferable value of the evaluation function, then the search procedure stops. Since the features included in \( S \) cannot be moved out of the set, the method has a high efficiency in feature selection process. Therefore, it is appropriate to employ this method to handle a big dataset, especially for GPS travel survey.

In this study, we apply the naïve classifier with SFS to detect travel modes. The naïve Bayesian classifier is selected to detect travel modes in our study. This method incorporates the following advantages. First, any observation is assigned to a class with a maximum probability instead of absolutely being allocated to a specific class. The probability is calculated based on Bayes’ Theorem and maximum posteriori hypothesis, which adequately utilize existing information and potentially improve detection accuracy. Second, all attributes potentially contributes to the determination of travel modes. It can overcome ambiguity of a specific attribute to utilize the information of all features. For example, it might not differentiate e-bicycle trips from bus trips to depend on the average speed alone, but the intervention of the ‘heading change rate’ contributes to the classification. Third, this method avoids searching for an optimal solution. It requires a high computation cost and even might be struggled with a local optimal solution to implement a search. Though a basic hypothesis that attributes is independent of each other does not completely hold, existing studies comparing classification methods have found that the naïve Bayesian method is comparable of selected neural network classifiers in terms of performance (20).
The naïve Bayesian classifier works as follows:

1) Let $T$ be a training set of observations, each with an actual class. In total $k$ classes, i.e., $C_1, C_2, \ldots$ and $C_k$, exist. Each observation is expressed as an n-dimensional vector, $X = \{x_1, x_2, \ldots, x_n\}$, expressing $n$ measured values of the $n$ attributes, $A_1, A_2, \ldots, A_n$, respectively.

2) Given a sample $X$, the classifier will determine that the sample belongs to the class which has the highest posteriori probability, conditioned on $X$. In other words, the sample is assigned to the class $C_i$ if and only if

$$P(C_j | X) > P(C_i | X), 1 \leq j \leq k, j \neq i$$

(1)

Therefore, the class that maximizes $P(C_i | X)$ is found. The reason for selecting $C_i$ as the objective class is based on the maximum posteriori hypothesis. According to Bayes’ theorem

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}$$

(2)

3) Since $P(X)$ is the same for all classes, we need only to maximize $P(X | C_i)P(C_i)$. It should be noted that $P(C_i)$ represents the priori probability of the class $C_i$ in the training set.

4) Since the class conditional independence is considered to hold, i.e., the values of the attributes are conditionally independent of one another, conditioned on the class label of the sample. Therefore, we can calculate the posteriori probability by

$$P(X | C_i) = \prod_{k=1}^{n} P(x_k | C_i)$$

(3)

The probabilities of $P(x_1 | C_i), P(x_2 | C_i), \ldots, P(x_n | C_i)$ can easily calculated from the training set. As mentioned earlier, $x_k$ refers to the value of the attribute $A_k$ for the sample. Since all the attributes in our study are continuous-valued, we assume that the values have a Gaussian distribution with a mean $\mu$ and the standard deviation $\sigma$ defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

(4)

so that

$$P(x_k | C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

(5)

where $\mu_{C_i}$ and $\sigma_{C_i}$ represent the mean and standard deviation of values of attribute $A_k$ for training samples of class $C_i$.

5) To predict the class label of $X$, $P(X | C_i)P(C_i)$ is calculated for each class $C_i$. The naïve Bayesian classifier predicts the class label of $X$ as $C_i$ if and only if it is the class that maximizes $P(X | C_i)P(C_i)$.

RESULTS AND DISCUSSIONS

SFS method is utilized to select an optimal attribute set in our study. Ten-fold cross validation is applied in this study to measure the performance of the methods. The
evaluation function of SFS is the total accuracy, which represents an integrated index
of detection accuracy of both training set and test set. According to this evaluation
function, the average speed is first chosen and more than 50% of travel modes are
correctly detected by the average speed alone (as shown in TABLE 1). This promising
result proves that the average speed is essential and irreplaceable in detecting travel
modes. The second attribute included in the optimal attribute set is the average
acceleration, which improves the total accuracy by more than 30%. The obvious
enhancement by average acceleration stems from its powerful distinction among walk,
bicycle and e-bicycle trips. ‘low-speed point rate’ is subsequently selected since it can
effectively differentiate bus trips from car trips. Consequently, it improves the total
accuracy by nearly 7%, which represents a large improvement in terms of travel mode
detection. ‘Heading change rate’ is the fourth attribute selected to construct the
classification model. Though an improvement of 3% seems not to be great, it is highly
difficult to further improve the results considering that the total accuracy has exceeded
90%. Unfortunately, the total accuracy decreases after trip distance is included in the
classification model. Therefore, the SFS procedures stops with four attributes included
in the optimal attribute set and the total accuracy of 94.89%. Trip distance and the 95%
percentile speed are not utilized to detect travel modes eventually. The role of trip
distance has been replaced by both average speed and average acceleration. Similarly,
the 95% percentile speed is not used because it has nearly the same shape of distribution
as the average speed.

TABLE 1 The Process of Sequential Forward Selection

<table>
<thead>
<tr>
<th>Serial ID</th>
<th>Attributes added</th>
<th>Accuracy of training set</th>
<th>Accuracy of test set</th>
<th>Total accuracy</th>
<th>Stop the procedures?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Average Speed</td>
<td>53.14%</td>
<td>51.65%</td>
<td>52.99%</td>
<td>False</td>
</tr>
<tr>
<td>2</td>
<td>Average acceleration</td>
<td>84.63%</td>
<td>83.24%</td>
<td>84.49%</td>
<td>False</td>
</tr>
<tr>
<td>3</td>
<td>Low-speed point rate</td>
<td>91.47%</td>
<td>90.46%</td>
<td>91.37%</td>
<td>False</td>
</tr>
<tr>
<td>4</td>
<td>Heading change rate</td>
<td>94.97%</td>
<td>94.28%</td>
<td>94.89%</td>
<td>False</td>
</tr>
<tr>
<td>5</td>
<td>Trip distance</td>
<td>93.73%</td>
<td>93.11%</td>
<td>93.67%</td>
<td>True</td>
</tr>
</tbody>
</table>

It can comprehensively evaluate the classification method to explore the
confusion matrix with detection accuracy of training set and test set (as shown in TABLE
2 and TABLE 3). On the one hand, subway trips are detected with the highest accuracy
of more than 98% for either training set and test set, which indicates that it is advisable
to employ particular rules to detect subway trips. Since we employ the same pre-defined
rules to detect subway trips, the detection accuracy of subway trips for training set is
analogous to that for test set. Additionally, detection accuracy of all travel modes
exceeds 86%. On the other hand, it can be found from the two tables that travel mode
detection for test set achieves nearly the same accuracy with that for training set. In fact,
more subway trips in test set are correctly identified than that in training set. These
preferable results prove a promising classification ability and an excellent
generalization of the classification methods in our study.
TABLE 2 Detection Accuracy of Training Set

<table>
<thead>
<tr>
<th>Identified</th>
<th>Subway</th>
<th>Walk</th>
<th>Bicycle</th>
<th>E-bicycle</th>
<th>Bus</th>
<th>Car</th>
<th>Total</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subway</td>
<td>1068</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>12</td>
<td>1089</td>
<td>98.07%</td>
</tr>
<tr>
<td>Walk</td>
<td>2</td>
<td>1989</td>
<td>47</td>
<td>19</td>
<td>9</td>
<td>6</td>
<td>2072</td>
<td>96.00%</td>
</tr>
<tr>
<td>Bicycle</td>
<td>2</td>
<td>3</td>
<td>529</td>
<td>16</td>
<td>7</td>
<td>3</td>
<td>560</td>
<td>94.46%</td>
</tr>
<tr>
<td>E-bicycle</td>
<td>6</td>
<td>2</td>
<td>15</td>
<td>347</td>
<td>12</td>
<td>9</td>
<td>391</td>
<td>88.75%</td>
</tr>
<tr>
<td>Bus</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>15</td>
<td>616</td>
<td>13</td>
<td>662</td>
<td>93.05%</td>
</tr>
<tr>
<td>Car</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>14</td>
<td>11</td>
<td>488</td>
<td>530</td>
<td>92.08%</td>
</tr>
</tbody>
</table>

TABLE 3 Detection Accuracy of Test Set

<table>
<thead>
<tr>
<th>Identified</th>
<th>Subway</th>
<th>Walk</th>
<th>Bicycle</th>
<th>E-bicycle</th>
<th>Bus</th>
<th>Car</th>
<th>Total</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subway</td>
<td>122</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>124</td>
<td>98.39%</td>
</tr>
<tr>
<td>Walk</td>
<td>0</td>
<td>219</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>230</td>
<td>95.22%</td>
</tr>
<tr>
<td>Bicycle</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>63</td>
<td>93.65%</td>
</tr>
<tr>
<td>E-bicycle</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>38</td>
<td>2</td>
<td>1</td>
<td>44</td>
<td>86.36%</td>
</tr>
<tr>
<td>Bus</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>68</td>
<td>2</td>
<td>74</td>
<td>91.89%</td>
</tr>
<tr>
<td>Car</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>54</td>
<td>59</td>
<td>91.53%</td>
</tr>
</tbody>
</table>

A further analysis about detection accuracy of other modes besides subway can help to understand advantages of this combination method and the way of improving detection accuracy in future study. Walk trips are detected with the second highest accuracy, which is due to that walk has distinct characteristics with other travel modes in terms of both the average speed and the average acceleration. More than 93% of bicycle trips are correctly detected, which can be attributed to the inclusion of heading change rate distinguishing non-motorized trips from motorized trips effectively. The detection accuracy of bus and car modes exceeds 90%. More importantly, most falsely detected bus trips are incorrectly flagged as e-bicycle trips instead of car trips, and vice versa. The low confusion rate between bus trips and car trips can be attributed to application of ‘low-speed point rate’. The high distinction between bus trips and car trips indicates that some specific attributes can replace a bus network GIS layer to a certain extent. In contrast, e-bicycle trips are identified with the lowest accuracy, which is possibly due to that there exists a relatively high similarity between bicycle and e-bicycle in terms of the average speed and between other modes with e-bicycle in terms of the average acceleration. Therefore, we need to extract attributes having larger distinction between e-bicycle and other travel modes in next stage.

To compare the detection accuracy of the combination method with other commonly used methods, we select four representative methods to deal with the issue discussed in the paper. They are support vector machine, multinomial logit, BP neural network and Bayesian network. As shown in Table 4, the method proposed in this paper achieves the highest detection accuracy. This result demonstrates the reasonability of combining particular rules and a naïve Bayesian classifier to detect travel modes.
Furthermore, the accuracy difference between training set and test set of the combination method is obviously lower than that of other methods. Therefore, it is proved that the method proposed incorporates a favorable generalization ability.

Table 4 Comparison of detection accuracy

<table>
<thead>
<tr>
<th>Methods</th>
<th>Training set</th>
<th>Sample size</th>
<th># of trips correctly detected</th>
<th>Detection accuracy</th>
<th>Test set</th>
<th>Sample size</th>
<th># of trips correctly detected</th>
<th>Detection accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support vector machine</td>
<td>5304</td>
<td>4793</td>
<td>90.37%</td>
<td></td>
<td>594</td>
<td>523</td>
<td>88.05%</td>
<td></td>
</tr>
<tr>
<td>Multinomial logit</td>
<td>5304</td>
<td>4412</td>
<td>83.18%</td>
<td></td>
<td>594</td>
<td>474</td>
<td>79.80%</td>
<td></td>
</tr>
<tr>
<td>BP neural network</td>
<td>5304</td>
<td>4697</td>
<td>88.56%</td>
<td></td>
<td>594</td>
<td>506</td>
<td>85.17%</td>
<td></td>
</tr>
<tr>
<td>Bayesian network</td>
<td>5304</td>
<td>4845</td>
<td>91.35%</td>
<td></td>
<td>594</td>
<td>532</td>
<td>89.56%</td>
<td></td>
</tr>
<tr>
<td>The method here</td>
<td>5304</td>
<td>5037</td>
<td>94.97%</td>
<td></td>
<td>594</td>
<td>560</td>
<td>94.28%</td>
<td></td>
</tr>
</tbody>
</table>

SUMMARY AND CONCLUSIONS

We employed particular rules to detect subway trips and utilized a naïve Bayesian classifier with SFS procedures to identify other travel modes. Before constructing the classifier, we extracted six attributes, including the average speed, the 95% percentile speed, the average acceleration, trip distance, low-speed point rate and the heading change rate. By SFS procedures, four attributes were selected other than the 95% percentile speed and trip distance. Finally, subway trips are detected with the highest detection rate of more than 98%, while more than 86% trips are correctly detected for each mode. Notably, more than 91% of bus trips and car trips are accurately identified with little confusion between them.

The combination of particular rules and advanced classification methods in detecting travel modes has been hardly used in existing research. However, the promising result by this combination provides us with a new perspective in terms of travel mode imputation and even other issues, including trip purpose detection, involved in GPS travel surveys. Besides, depending on extracted attributes besides commonly used features in existing literature, we effectively differentiate bus trips from car trips. This preferable result demonstrates that we can improve detection accuracy of bus trips and car trips by extracting representative attributes even without a bus network layer. Compared with its competitors, the method achieves the highest detection accuracy and shows a favorable generalization ability. However, it is inappropriate to compare the performance of multiple methods in different studies since the difference in detection accuracy depends on the number of reported travel modes, selection of input variables, urban setting and data utilized to validate the methods. In a study by Stopher et. al (11), for example, the detection accuracy of 95% is incomparable to that of this study since e-bicycle and subway modes are added to our list of reported travel modes.

The method proposed is highly suitable for detecting travel modes in a GPS travel survey undertaken in a city where subway trips take a relatively high percentage.
As a large-scale subway network is gradually constructed and put into operation in a great many countries, especially the emerging economies, the methods are expected to be applied more extensively. For the cities without subway lines, it is recommended that travel modes are detected for all trips with a naïve Bayesian classifier with SFS procedures since there does exist a significant ambiguity between trips with travel modes other than subway mods. In a further study, features incorporating high distinctiveness between e-bicycle trips and trips with other travel modes need to be extracted due to with a relatively low accuracy for detecting e-bicycle trips.

ACKNOWLEDGEMENTS

This research has been supported by the National Natural Science Foundation of China (51278301).

REFERENCE


