Effective Classification of Ground Transportation Modes for Urban Data Mining in Smart Cities

Carson K. Leung(✉), Peter Braun, and Adam G. M. Pazdor

University of Manitoba, Winnipeg, MB, Canada
kleung@cs.umanitoba.ca

Abstract. The increasing amount of digital data in urban research has drawn attention in *urban data mining*. In urban research (e.g., travel studies in urban areas), researchers who conduct paper-based or telephone-based travel surveys often collect biased and inaccurate data about movements of their participants. Although the use of global positioning system (GPS) trackers in travel studies improves the accuracy of exact participant trip tracking, the challenge of labelling trip purpose and transportation mode still persists. The automation of such a task would be beneficial to travel studies and other applications that rely on contextual knowledge (e.g., current travel mode of a person). In this paper, we focus on transportation mode classification. In particular, we develop a system that improves classification accuracy of ground transportation modes (e.g., bus, car, bike, or walk). When compared with related works, our system increases the classification accuracy by uniquely using GPS and accelerometer data together with a window history queue (which uses previously encountered data). Evaluation results show that our system achieves a high classification accuracy.

Keywords: Data mining · Classification · Ground transportation mode
Global positioning system (GPS) data
Geographic information system (GIS) data · Accelerometer data
Window history queue (WHQ)

1 Introduction

In the current era of big data, huge volumes of a wide variety of data of different veracity (e.g., uncertain and imprecise data [1–3]) can be easily collected or generated at a high velocity in many real-life applications. Embedded in these big data are valuable information or knowledge. This calls for *data mining* [4–6], which aims to extract implicit, previously unknown, and potentially useful information from a large amount of data [7]. With the increasing amount of digital data in urban research, the field of *urban data mining* [8–12]—which aims to discover knowledge from data related to urban problems for solving urban issues—has also developed and been given more attention.

In urban research (specifically, travel studies in urban areas), researchers often have been using paper-based and telephone-based *travel surveys* [13]. Those travel surveys

C. Ordonez and L. Bellatreche (Eds.): DaWaK 2018, LNCS 11031, pp. 83–97, 2018.
https://doi.org/10.1007/978-3-319-98539-8_7

can often be biased and contain inaccurate data about movements of their participants. Participants tend to under-report short trips and irregular trips. Additionally, car trips are often reported to be shorter than they are, and public transit trips are reported to be longer than they are [14, 15].

Commute diaries [16, 17] are another approach of collecting data about people's daily commutes, but they also have been shown to be prone to errors. When people are asked to use a diary to keep track of their commutes, they often forget to record their commutes throughout the day. When trips are recorded at the end of the day, diary studies can then inherit the same problems as paper-based and telephone-based travel surveys. Moreover, commute diary studies can also be a mental burden to study participants and cannot be used long term [18]. As people's willingness to record trips accurately throughout the day declines with each day of participation, the accuracy of the commute diaries also drops accordingly [19].

To avoid data collection problems associated with the paper-based travel survey and commute diaries, there is now a shift towards the use of global positioning system (GPS) trackers to collect more objective commute data from participants. Studies show that *GPS-based travel surveys* [20, 21] are more accurate than the aforementioned surveys and diaries.

Although the use of GPS trackers in travel studies improve the accuracy of exact participant trip tracking, the challenge of *labelling trip purpose* and *classifying transportation mode* still persists. Nowadays, with the use of electronic trackers, researchers are often dealing with a large amount of movement trajectories that are collected by participants of a study who use GPS trackers or other sensors (e.g., Bluetooth, Wi-Fi, accelerometers, barometers, etc.). Manual segmentation of trajectories based on transportation mode is a labor-intensive task, and most likely infeasible when performed on big data [22]. The automation of such a task would obviously be beneficial to travel studies and other applications that rely on contextual knowledge (e.g., current travel mode of a person). As an example for a contextual use for transportation, when the person is driving, a device like a smartphone could recognize that the person is driving in a car and give a notification about the current estimated time of arrival—assuming that the phone knows the destination based on previous user interaction or saved frequently visited locations. Another application for transportation mode classification is the automatic trip transportation mode labeling for trip history. This is similar to timeline in Google Maps (which keeps track of a user's location history and attempts to automatically classify trips with the major transportation mode). However, the resulting classifications are observed to be not very accurate and need a lot of corrections by the user. Moreover, it does not track when transportation modes were changed. Hence, a more accurate algorithm or system is needed.

By using a standalone tracking and logging device, participants of travel surveys would be able to log sensor data reliably and consistently because developers and engineers have full control over the device and the hardware. Moreover, software platforms are the same on every device. Such loggers can be used to log data to local device storage, and then collect the logged data for data retrieval. Some devices could also connect to a smartphone application on a participant's phone via Bluetooth and collect data on regular intervals. The collected data could be further processed, and the users could be prompted with surveys. In such a case, transportation mode

classification could happen on a smartphone. By doing so, computational burden on the logger device is reduced, a cheaper architecture that requires weaker processing units is possible, power consumption is potentially decreased, and thus the battery life is increased.

As a preview, the system proposed in this paper works well for all scenarios described above—*smartphone logging* (*with online and online classification*) and *standalone logging devices*. Moreover, in this paper, our system focuses on the following:

- offline learning (with which the classification model is usually trained on the server);
- offline classification/prediction (with which the classification model classifies the ground transportation modes on the server), and
- online classification/prediction (with which the classification model classifies the ground transportation modes on the mobile device).

In order to increase classification accuracy, our transportation mode classification system not only uses the basis of an online classification approach, but also uses multiple windows of previously encountered data. Online classification focuses on one window at a time during processing and classification. To the best of our knowledge, existing academic works in this area have yet not taken advantage of previously encountered data windows in order to increase the classification accuracy of the currently processed window of data. Hence, a natural question to ask is: Is it possible to compute features based on previously seen data for a single trip and use it to significantly improve real-time window based ground transportation mode classification?

In this paper, we design new classification features based on a *window history queue* (*WHQ*), which focuses on summarizing data from previously encountered data windows. Our goal is to improve accuracy of transportation mode classification when compared with existing systems. Our ***key contribution*** is our classification system for ground transportation modes. To the best of our knowledge, uniquely using new features based on a WHQ, together with accelerometer data and GPS data, in a single system during the classification process improves the classification accuracy.

The remainder of this paper is organized as follows. The next section discusses related works. Section 3 presents our proposed transportation mode classification system. Evaluation results and conclusions are given in Sects. 4 and 5, respectively.

2 Related Works

Most research works [23–25] related to transportation mode classification require at a minimum GPS or accelerometer data for creating a classification model. Much research exists [26], where researchers used some combination of data from different sensors (e.g., GPS, accelerometers) and other modern smartphone sensors (e.g., barometer, magnetometer, etc.). Some researchers [27] also augmented geographic information system (GIS) data to their sensor datasets. However, none of them uniquely combined GPS, accelerometer and GIS data in a single system. In contrast, our system combines GPS, accelerometer and GIS data.

To elaborate, Zheng et al. [23] used supervised decision trees and graph based post-processing after classification to classify transportation modes from *GPS data only*.

In contrast, Hemminki et al. [24] focused on classifying transportation modes ("stationary", "walk", "bus", "train", "metro", "tram") with the use of *only accelerometer data*. To classify transportation modes, three different classifiers were trained with a combination of AdaBoost and Hidden Markov Model for three different classes of modes. Shafique and Hato [25] also used accelerometer data only. They applied multiple machine learning algorithms to perform transportation mode classification, and found that Random Forest [28] gave accurate classification.

Instead of using only GPS data or only accelerometer data, Ellis et al. [26] applied the Random Forest to *both GPS data and accelerometer data* in order to successfully perform transportation mode classification with a relatively high accuracy.

Other than using both GPS data and accelerometer data, Chung and Shalaby [27] developed a system that uses *both GPS and GIS data* instead. Their system classified four transportation modes—"walk", "bike", "bus" and "car"—for GPS-based travel surveys by using a rule-based algorithm and a map-matching algorithm [29] to detect the exact roads people moved on. However, the accuracy of the system is dependent on the corresponding GIS data. Similarly, Stenneth et al. [30] also used both GPS and GIS data when building their real-time transportation mode classification system. To perform the classification, they used the Random Forest as the supervised machine learning algorithm to identify a person's current transportation mode.

3 Our Proposed Classification System

In this section, we describe our ground transportation mode classification system that uses GPS data, accelerometer data, GIS data, and/or window history queue. The system consists of the following five modules:

1. Dataset collection module,
2. Trip segmentation module,
3. Feature extraction module,
4. Model construction module, and
5. Data classification module.

The end result of the system are segmented trips (or trip windows), where each segment is labelled with the ground transportation mode the person used for the period of the segment. See Fig. 1 for the system layout. Figure 2 shows a zoom-in view of the dataset collection module (in which the MongoDB data store is highlighted in yellow) and the analysis component (highlighted in green), which consists of the last four modules (i.e., the trip segmentation, features extraction, model construction, and data classification modules) listed above. This data analysis component interacts with the dataset collection module.

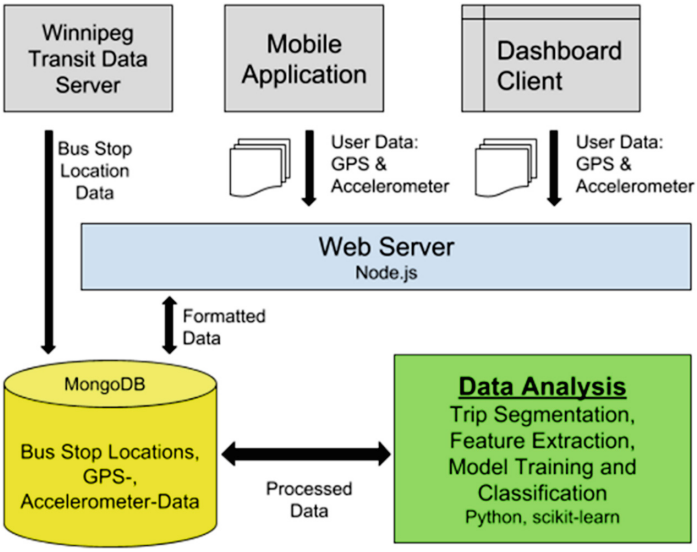


Fig. 1. Layout of our transportation mode classification system.

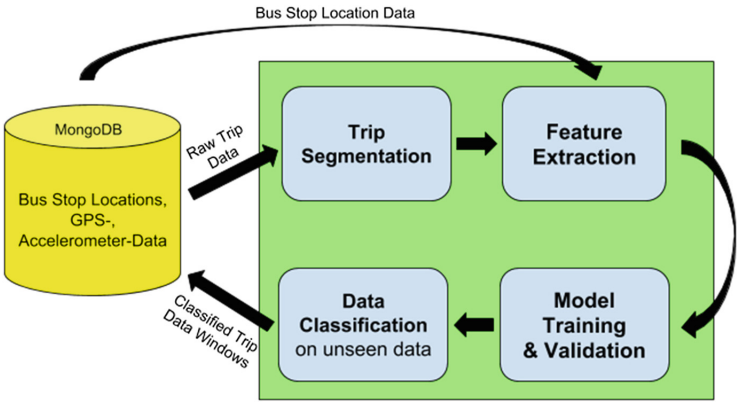


Fig. 2. Layout of the transportation data analysis component of our transportation mode classification system.

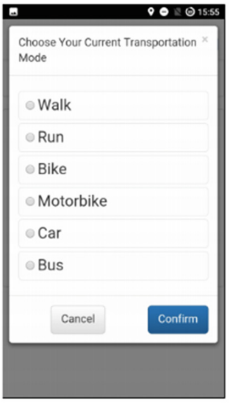
3.1 Dataset Collection Module

First, the ***dataset collection module*** collects trip traces (GPS locations) and trip accelerometer data, and stores them in a database (e.g., MongoDB). In addition, the module also collects the bus stop locations in a city—via its transit application programming interface (API)—when "bus" is one of the ground transportation modes for classification.

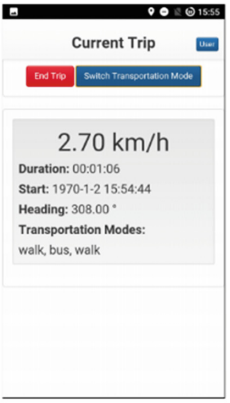
Figure 3 shows the main screens of an iOS application for *data collection*. Figure 3 (a) shows the "current trip" screen of the application, with which users can see their

current trip information (e.g., speed, alerts, map). To start a new trip or trip leg, the users simply end a trip. Then, a new trip will automatically start. After a new trip or trip leg starts, the users will be presented with a pop-up list of transportation modes, as shown in Fig. 3(b). Figure 3(c) shows the screen that lists users' saved trip log, where they can review trip information.

(a) The start screen and the user's list of trips.



(b) Users can select a transportation mode at the beginning of a trip and during a trip when transitioning between different modes.

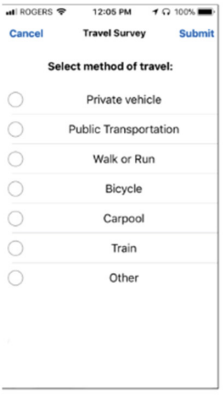


(c) Users can see some information relating to their current trip.

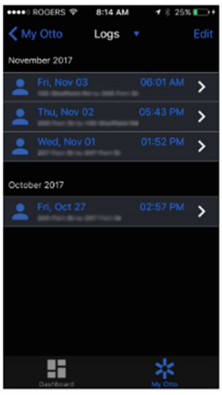
Fig. 3. Application for data collection.

The application also keeps track of users' movement. Figure 4(a) shows the start screen of the application for *movement tracking*, with which users can manage their recorded trips, start new trip recordings and upload existing trips. Figure 4(b) shows

(a) The start screen and information relating to the user's current trip.



(b) Users can select a transportation mode at the beginning of a trip. Trip legs are represented by individual trips.



(c) Users can see a list of their stored trips.

Fig. 4. Application for movement tracking.

the transportation mode selection pop-up that would appear when the users start a new trip. They could also open the popup anytime from the current trip screen when they are switching transportation modes. Figure 4(c) shows the screen for current trip information after the users started a new trip. Here, the GPS sampling rate was set to 1 Hz and the accelerometer sampling was at 22 Hz.

Recall that some related works use GIS data together with GPS location. Knowing the difficulty in obtaining a complete set of GIS information in some real-life situations, the only GIS information required by our data collection module is the bus stop location, which can be easily accessible. For example, in evaluation, we obtained the bus stop locations from Winnipeg Transit Open Data Web Service API.

3.2 Trip Segmentation Module

After collecting raw trip data by the data collection module, our ***trip segmentation module*** segments every trip (which is simply a collection of data points collected during a person's entire commute from origin to destination—say, from home to work) based on the transportation mode used in each segment. For example, for a trip from home to work can be divided into the following three segments, as shown in Fig. 5:

1. Walk from home to the departure bus stop,
2. Bus from departure bus stop to destination bus stop, and
3. Walk from destination bus stop to office.

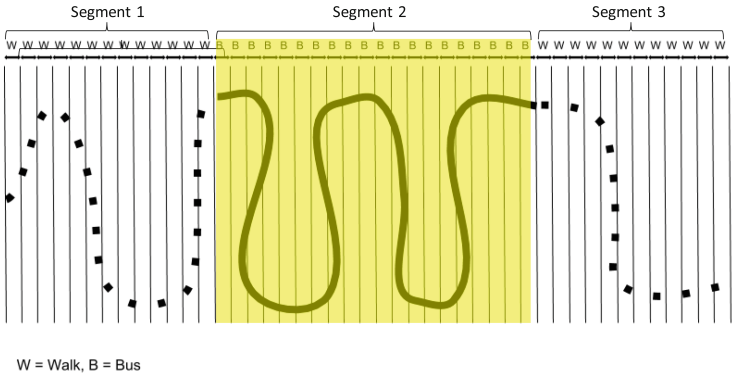


Fig. 5. Segmented trip.

Our trip segmentation module is designed in such a way that it automatically achieves the trip segmentation by simply classifying the transportation mode of each window of data. To elaborate, data of a trip are divided into many small windows of equal time interval. Moreover, when a transportation mode change occurs, data are assigned to different windows so that no two transportation modes are mixed within the same window. Segmenting the data into small windows led to a benefit that classification can be performed in real-time. For instance, as soon as sufficient amount of data

has been collected to fill a new window, the window can be classified with a transportation mode. Once every window is classified with a transportation mode, the user simply concatenates the windows/trip segments (each of which is labelled with a transportation mode) and presents each label on a map with a color-scheme for different transportation modes. Once the trip is rendered on a map, the user can easily identify different legs of the trip by simply looking at the different colors of the trip.

3.3 Feature Extraction Module

With (i) the bus stop location data (i.e., GIS information) collected by the dataset collection module and (ii) the GPS and accelerometer data associated with the trip segmented returned by the trip segmentation module, our ***feature extraction module*** extracts appropriate features for transportation mode classification. Specifically, the module extract the following three key types of features:

- **GIS-based features**, which capture the following GIS information related to bus stop locations:
  a. Stopped at a bus stop, which is a Boolean feature that indicates whether or not a person stopped at any of the nearby bus stops within the window;
  b. The number of bus stops, which captures the number of unique bus stops within the window;
  c. The number of stops at bus stops, which captures the number of stops near all the bus stops within the window; and
  d. Distance to closest bus stop (in meters), which captures the distance to the bus stop that is closest to the person within the entire window.
- **GPS-based features,** which capture the following geo-location and time information provided by GPS sensors:
  a. Maximum speed (in km/h);
  b. Average speed (in km/h);
  c. Average altitude (in meters);
  d. Average location accuracy (in meters);
  e. Travel distance (in meters), which computes the geodesic distance (i.e., shortest possible line between two points p_1 and p_2 on a sphere) in terms of *Haversine distance*. Such a distance d between p_1 and p_2 can be calculated by:

$$d = 2r\sin^{-1}\left(\sqrt{\sin^2\left(\frac{lat_2 - lat_1}{2}\right) + \cos(lat_1)\cos(lat_2)\sin^2\left(\frac{long_2 - long_1}{2}\right)}\right)$$

  where (i) r is the radius of the sphere, (ii) $long_1$ and lat_1 are respectively the longitude and latitude of p_1, and (iii) $long_2$ and lat_2 are respectively the longitude and latitude of p_2; and
  f. GPS signal loss, which is a Boolean feature that indicates whether there is GPS signal or no signal.

- **Accelerometer-based features,** which capture the following measurement on acceleration of different transportation modes (e.g., automobile):

  a. Maximum magnitude;
  b. Minimum magnitude;
  c. Average magnitude;
  d. 25th percentile magnitude, which captures the average of all magnitude in the 25th percentile;
  e. 75th percentile magnitude, which captures the average of all magnitude in the 75th percentile;
  f. Magnitude standard deviation;
  g. Lag-1 autocorrelation;
  h. Correlation between *x*- and *y*-axes;
  i. Correlation between *x*- and *z*-axes;
  j. Correlation between *y*- and *z*-axes;
  k. Average *roll*, which captures the average "bank angle" about rotations along the *x*-axis;
  l. Average *pitch*, which captures the average "elevation" about rotations along the *y*-axis; and
  m. Average *yaw*, which captures the average "bearing" about rotations along the *z*-axis.

Recall from Sect. 2 that some existing classifiers use only GPS based data (say, the aforementioned 6 GPS-based features), some use only accelerometer-based data (say, the aforementioned 13 accelerometer based features), some use both GPS and accelerometer based data (say, the aforementioned 6 + 13 = 19 GPS- and accelerometer- based features), and some use both GPS and GIS based data (say, the aforementioned 6 + 4 = 10 GPS- and GIS-based features). However, to the best of our knowledge, *no classifier—except our proposed system—uses all three types of 4 + 6 + 13 = 23 features from these GIS-, GPS- and accelerometer-based data*.

To a further extent, with an aim to enhance the classification accuracy, we propose a novel concept of **window history queue (WHQ)**. The idea behind a WHQ is that previous data windows could help to classify the current data window. For instance, if the current data are similar to the previous one (e.g., similar average speed), then the current data are more likely to have the same classification as the previous data. Conversely, if the current data are quite different from the previous one, then there is a chance that transportation modes have changed. Our classifier is able to determine the subtle data differences during the training phase. For example, the random forest model can learn that a previous high average speed followed by a very low (current) average speed would mean that the current data represent the transportation mode of "walk".

To support this concept of window history queue (WHQ), our feature extraction module extracts 6 + 4 = 10 additional GPS- and accelerometer-based features for WHQ. These 10 additional features are listed as follows:

- **Additional GPS-based features for WHQ** include:

  a. Previous maximum speed (in km/h), which is the maximum speed for all previous windows in the WHQ;

b. Previous average speed (in km/h), which is the average speed for all previous windows in the WHQ;
c. Previous maximum average speed (in km/h), which is the maximum average speed for all previous windows in the WHQ;
d. Delta maximum speed (in km/h), which is the difference between the maximum speed of the current window and that of the previous windows;
e. Delta average speed (in km/h), which is the difference between the average speed of the current window and that of the previous windows; and
f. Delta maximum average speed (in km/h), which is the difference between the maximum average speed of the current window and that of the previous windows.

- **Additional accelerometer-based features for WHQ** include:

a. Previous average magnitude which captures the average of all average magnitudes in the WHQ;
b. Previous 25th percentile magnitude, which captures the average of all magnitude in the 25th percentile of the WHQ;
c. Previous 75th percentile magnitude, which captures the average of all magnitude in the 75th percentile of the WHQ; and
d. Previous magnitude standard deviation, which captures the average of all magnitude standard deviations of the WHQ.

3.4 Model Construction Module

Once features are extracted from GIS information, GPS sensors and accelerometer data, our ***model construction module*** builds, trains and validates a classification model. Specifically, it builds the random forest model by using the extracted features which are calculated based on the collected raw data. The extracted features are stored on a per-trip basis. For each trip, there is a set of feature windows. The trips for each transportation mode are shuffled and then split into two sets: (i) the training set and (ii) the testing/validation set. As a preview, to evaluate our classification system, we used 70% of the data for stratified 10-fold cross-validation with a 50-50 partition split between the training and the testing data for each partition in order to determine the accuracy of the random forest based classifier.

3.5 Data Classification Module

After constructing the classification model, our ***data classification module*** classifies unseen data and stores the classified trips back to the MongoDB. Specifically, segments of a trip are classified according to the ground transportation mode (e.g., “walk”, “bike”, “bus”, “car”) used by the user.

4 Evaluation

To evaluate our proposed ground transportation mode classification system, we conducted experiments on a computer running Ubuntu 16.04 LTS as the main operating system. The CPU was an AMD Phenom II X6 1100T with 6 cores clocked at 3.3 GHz to 3.7 GHz. There are 16 GB of RAM and a solid state drive in the computer.

We collected the GIS information (i.e., bus stop locations) from Winnipeg Transit Data Server by querying trip data via the Winnipeg Transit API. Users anonymously and securely uploaded their saved GPS- and accelerometer-based data via the mobile applications or dashboard. These trip information from users were then stored in a MongoDB, which supports basic geo-spatial query capabilities. The trip information was collected throughout a year, which contains trips with different weather and road conditions from summer to winter times. It captures the ground transportation mode (e.g., "walk", "car", "bus") used by the user at the time of commute.

Recall from Sect. 3.2 that the trip segmentation module of our proposed ground transportation mode classification system divides each trip into many small windows of equal time interval. *Our first set of experiments is to determine an appropriate window size.* We varied the window size. The experimental results shown in Fig. 6 reveal that a window size of 4 s gave the most accurate classification.

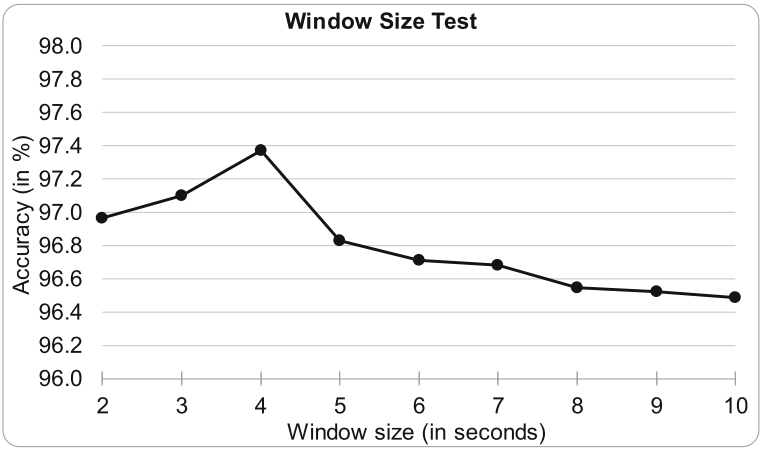


Fig. 6. Experimental result: Segmented window size.

Recall from Sect. 3.3 that the feature extraction module of our proposed ground transportation mode classification system uses a novel concept of window history queue (WHQ), which captures historical data. With WHQ, comparisons can then be made between the GPS information or accelerometer data in the current window and those in the previous windows within the WHQ. So, *our second set of experiments is to determine an appropriate queue length*. We varied the queue length. The experimental results shown in Fig. 7 reveals that a WHQ length of 15 windows (each window of size 4-second interval) gave the most accurate classification.

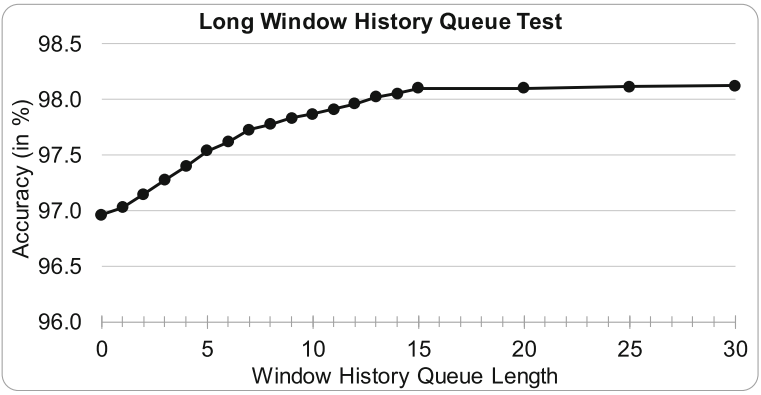


Fig. 7. Experimental result: Window history queue (WHQ) length.

To measure the effectiveness of our proposed ground transportation mode classification system, we use the standard measures of precision and recall. *Precision* measures the positive predictive/classified value, i.e., the fraction of true positives among all positives (i.e., true and false positives):

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positive}}$$

Recall measures the true positive rate or sensitivity, i.e., the fraction of true positives among true positives and false negatives:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

Accuracy measures the fraction of true positives and true negatives among all predications/classifications (i.e., among all positives and negatives):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives. So, *our third set of experiments is to compute the precision, recall, and accuracy of the classification results on unseen data*. The experimental results show that our system accurately classified different segments of unseen data with highly accurate (e.g., above 95%) ground transportation modes.

Recall from Sects. 2 and 3 that some existing classifiers use only GPS-based data (say, the 6 GPS-based features), some use only accelerometer-based data (say, the 13 accelerometer-based features), some use both GPS- and accelerometer-based data (say, the 6 + 13 = 19 GPS- and accelerometer-based features), and some use both GPS- and GIS-based data (say, the 6 + 4 = 10 GPS- and GIS-based features). In contrast, *our ground transportation mode classification system uses all GPS-, GIS- and*

accelerometer-based features, which is a key contribution of this paper. Hence, *our fourth set of experiments is to compare the accuracy of the classification results of our classification system with related works*. The experimental results shown in Table 1 reveals that our system led to higher classification accuracy.

Table 1. Classification accuracy of our classifier with compared with related works.

	Data used	Accuracy
Related Works	Only GPS data	88.87%
	GPS + GIS data	88.96%
	Only accelerometer (Acc) data	93.13%
Our proposed classification system	GPS + Acc data	95.58%
	GPS + GIS + Acc data	96.97%

Recall from Sect. 3 that *another key contribution of this paper is our proposal of the concept of window history queue* (*WHQ*). Hence, *our fifth set of experiments is to measure the benefits* (*e.g., increase in classification accuracy*) *of using WHQ*. The experimental results shown in Table 2 reveals that our system (which uses the WHQ) led to higher classification accuracy.

Table 2. Classification accuracy of not using vs. using the WHQ.

Without WHQ	Accuracy	With WHQ	Accuracy
Only GPS data	88.87%	GPS + WHQ	94.70%
GPS + GIS data	88.96%	GPS + GIS + WHQ	94.63%
Accelerometer (Acc) data	93.13%	Acc + WHQ	94.85%
GPS + Acc data	95.58%	GPS + Acc + WHQ	98.08%
GPS + GIS + Acc data	96.97%	GPS + GIS + Acc + WHQ	98.09%

5 Conclusions

This paper focuses on urban data mining. In particular, we proposed a ground transportation mode classification system. The system consists of five modules. Among them, the dataset collection module collects the GIS information about bus stop locations, the geo-location and time information provided by GPS sensors, and accelerometer-based data capturing the acceleration of different transportation modes. These data are collected from different sources: GIS information from the city or transit company, whereas GPS- and accelerometer-based data are collected by users. Then, the trip segmentation module segments data about a trip into many windows of size 4 s, and each of these windows captures a single mode of transportation. Afterwards, the feature extraction module extracts the standard GIS-, GPS- and accelerometer-based data. Most, if not all, of the existing related works do not use all GIS-, GPS- and accelerometer-based data. So, we proposed in this paper to use the combined GIS-, GPS- and accelerometer-based data. Consequently, the model construction module

builds and trains a random forest classification model using these combined data. The data classification module then classifies future unseen data. Experimental results show that the use of combined GIS-, GPS- and accelerometer-based data led to high classification accuracy.

Moreover, with an aim to improve classification accuracy, we also proposed in this paper a novel concept of window history queue (WHQ) so that the model construction module can compare the data in the current window with data in the previous windows within the WHQ. To facilitate our proposed concept of WHQ, the feature extraction module extracts additional GPS- and accelerometer-based data for WHQ. Experimental results show that the use of WHQ led to higher accuracy than their counterparts that do not use the WHQ. These results demonstrate the effectiveness of our system in classifying ground transportation modes for urban data mining in smart cities.

As ongoing and future work, we are examining relationships (e.g., correlations) among the extracted features to see if the number of features can be reduced. At the same time, we are also exploring new features that could further enhance the classification accuracy. Moreover, we are also conducting more exhaustive evaluation (e.g., comparisons with fitness trackers).

Acknowledgements. This project is partially supported by NSERC (Canada) and University of Manitoba. Thanks F. Franczyk—President of Presen Technologies Inc. (PERSENTECH)—for his introduction of OttoFleet Mobile™ (an application for GPS-enabled iPhones®), which inspired the design of the dataset collection module of our classification system.

References

1. Braun, P., Cuzzocrea, A., Jiang, F., Leung, C.K.-S., Pazdor, A.G.M.: MapReduce-based complex big data analytics over uncertain and imprecise social networks. In: Bellatreche, L., Chakravarthy, S. (eds.) DaWaK 2017. LNCS, vol. 10440, pp. 130–145. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64283-3_10
2. Chen, Y.C., Wang, E.T., Chen, A.L.P.: Mining user trajectories from smartphone data considering data uncertainty. In: Madria, S., Hara, T. (eds.) DaWaK 2016. LNCS, vol. 9829, pp. 51–67. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43946-4_4
3. Hoi, C.S.H., et al.: Supporting social information discovery from big uncertain social key-value data via graph-like metaphors. In: Xiao, J., Mao, Z.-H., Suzumura, T., Zhang, L.-J. (eds.) ICCC 2018. LNCS, vol. 10971, pp. 102–116. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94307-7_8
4. Egho, E., et al.: MiSeRe-Hadoop: a large-scale robust sequential classification rules mining framework. In: Bellatreche, L., Chakravarthy, S. (eds.) DaWaK 2017. LNCS, vol. 10440, pp. 105–119. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64283-3_8
5. Leung, C.K.: Big data analysis and mining. In: Encyclopedia of Information Science and Technology, 4th edn., pp. 338–348 (2018)
6. Leung, C.K., Jiang, F., Pazdor, A.G.M., Peddle, A.M.: Parallel social network mining for interesting 'following' patterns. Concurr. Comput. Pract. Exp. **28**(15), 3994–4012 (2016)
7. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1–34 (1996)

8. Behnisch, M., Ultsch, A.: Urban data mining using emergent SOM. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 311–318. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78246-9_37
9. Andrienko, G., et al.: Mining urban data (Part A). Inf. Syst. **54**, 113–114 (2015)
10. Andrienko, G., et al.: Mining urban data (Part B). Inf. Syst. **57**, 75–76 (2016)
11. Andrienko, G., et al.: Mining urban data (Part C). Inf. Syst. **64**, 219–220 (2017)
12. Sokmenoglu, A., Cagdas, G., Sariyildiz, S.: Exploring the patterns and relationships of urban attributes by data mining. In: eCAADe 2010, pp. 873–881 (2010)
13. Murakami, E., Wagner, D.P., Neumeister, D.M.: Using global positioning systems and personal digital assistants for personal travel surveys in the United States. In: Transport Surveys: Raising the Standard, art. III-B (2000)
14. Ettema, D., Timmermans, H., van Veghel, L.: Effects of Data Collection Methods in Travel and Activity Research (1996)
15. Stopher, P.R..: Household travel surveys: cutting-edge concepts for the next century. In: Conference on Household Travel Surveys, pp. 11–23 (1995)
16. Maat, K., Timmermans, H.J.P., Molin, E.: A model of spatial structure, activity participation and travel behavior. In: WCTR 2004 (2004)
17. Stopher, P.R.: Use of an activity-based diary to collect household travel data. Transportation **19**(2), 159–176 (1992)
18. Schlich, R., Axhausen, K.W.: Habitual travel behaviour: evidence from a six-week travel diary. Transportation **30**(1), 13–36 (2003)
19. Arentze, T., et al.: New activity diary format: design and limited empirical evidence. TRR **1768**, 79–88 (2001)
20. Forrest, T., Pearson, D.: Comparison of trip determination methods in household travel surveys enhanced by a global positioning system. TRR **1917**, 63–71 (2005)
21. Wolf, J., Guensler, R., Bachman, W.: Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. TRR **1768**, 125–134 (2001)
22. Biljecki, F., Ledoux, H., van Oosterom, P.: Transportation mode-based segmentation and classification of movement trajectories. IJGIS **27**(2), 385–407 (2013)
23. Zheng, Y., Chen, Y., Li, Q., Xie, X., Ma, W.: Understanding transportation modes based on GPS data for web applications. ACM TWeb **4**(1), art. 1 (2010)
24. Hemminki, S., Nurmi, P., Tarkoma, S.: Accelerometer-based transportation mode detection on smartphones. In: SenSys 2013, art. 13 (2013)
25. Shaque, M.A., Hato, E.: Use of acceleration data for transportation mode prediction. Transportation **42**(1), 163–188 (2015)
26. Ellis, K., et al.: Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. Front. Pub. Health **2**, art. 36 (2014)
27. Chung, E., Shalaby, A.: A trip reconstruction tool for GPS-based personal travel surveys. Transp. Plann. Technol. **28**(5), 381–401 (2005)
28. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
29. Greenfeld, J.: Matching GPS observations to locations on a digital map. In: TRB 81st Annual Meeting (2002)
30. Stenneth, L., Wolfson, O., Yu, P.S., Xu, B.: Transportation mode detection using mobile phones and GIS information. In: ACM SIGSPATIAL GIS 2011, pp. 54–63 (2011)