Formation of Training and Testing Datasets, for Transportation Mode Identification

Muhammad Awais Shafique
Department of Transportation Engineering & Management, University of Engineering & Technology, Lahore, Pakistan
Email: awais@trip.t.u-tokyo.ac.jp

Eiji Hato
Department of Civil Engineering, The University of Tokyo, Tokyo, Japan
Email: hato@bin.t.u-tokyo.ac.jp

***Abstract*—Recently, a lot of research is done to automatically predict the mode of transportation used by the smartphone carrier by collecting and analyzing the data from sensors like global positioning system (GPS) and accelerometer. The most popular methodology is to train a classification algorithm and then use it to classify the test data. This study provides an insight into how the training and testing datasets should be formed. A comparison is made among the two approaches i.e. randomly splitting the data from all participants into training and testing datasets, or using some participants' data to form training dataset and rest to form testing dataset. For the first method, 50% data from 6 participants was randomly selected to train Random Forest while the rest was used to test it. For the second method, 5 participants' data was used to train the algorithm and a different set of 5 participants' data was used to test it. Results concluded that splitting the data causes over-estimation; therefore different datasets should be utilized for the purpose of mode identification.**

***Index Terms*—accelerometer, classification, GPS, random forest, travel mode**

I. Introduction

Travel data collection methods can be broadly divided into two classes. The first class covers the methods in which the respondents are asked to provide the travel information manually. The second class deals with the inclusion of technology for automatic collection of information. Currently, the methods falling in the first class are being predominantly practiced all over the world. These include paper-based surveys, travel diaries, internet surveys, telephone surveys and interviews. Although the information that can be collected by these conventional methods is very detailed but there are a lot of disadvantages associated with these methods. For example, these surveys are very time-consuming and laborious for the participants as well as for the people collecting the data. The participants are expected to remember every detail of their daily travelling and report it in the survey but mostly this is not the case. They frequently make mistakes while providing the accurate starting and ending time of each trip and occasionally forget to report short trips. Many a times the questions asked in the questionnaire are not fully understood by the respondents and consequently they provide wrong answers or simply give no answer resulting in no-response bias. Moreover, the perception of in-vehicle travel time varies with the mode of transportation. For instance, a person travelling by his personal vehicle will underestimate the travel time compared to if he was making the same trip by public transportation. All this results in decreased accuracy of the collected data.

In order to deal with the shortcomings of the conventional methods, research is being focused on the second class of data collection methods. The most recent approach is to employ smartphones for automatically detecting the mode of transportation used by the smartphone-carrier. Modern smartphones come equipped with a lot of sensors, including Global Positioning System (GPS) and accelerometer. GPS can locate the device anywhere in the world, hence providing the opportunity to constantly track the movement of the device-carrier in real time. On the other hand, accelerometer can record the acceleration of the device in three directions, with respect to the gravitational force. These sensors can collect data, which after suitable processing can be used to infer the mode of transportation used. The current study deals with the development of a methodology wherein suitable features will be extracted from the data and classification will be done.

II. Related Work

A lot of automatic data collection methods have been developed including electronic distance measuring instruments, license plate matching, cellular phone tracking, automatic vehicle identification, automatic vehicle location and video imaging [1]. These methods have certain disadvantages including high cost of the sensors, low accuracy, limitation to specific modes etc. The use of smartphone for automatic data collection has addressed most of the concerns related to other methods. Studies confirm that better data accuracy can be achieved using GPS data collection devices, as compared to the conventional questionnaire methods [2]-[4]. GPS data

Manuscript received August 8, 2014; revised March 30, 2015.

doi: 10.12720/jtle.3.1.77-80

combined with accelerometer data, exhibits better classification accuracy than either of these used alone [5].

For forming the training and testing datasets, different approaches have been used by researchers. In a study, 70% of the collected data was used to train the algorithm while the remaining 30% was used to validate its performance [6]. The classification accuracy highly varied among the participants and ranged from 88% to 97%. In another work, the training dataset comprised of 70% data and the rest was used for testing [7]. Support Vector Machines proved to be a better algorithm with an accuracy of 97.32%. A study investigating the various pre-processing methods used 50% of the data for training and rest 50% for testing [8].

The present study compares the classification accuracies when data from all users is randomly split into two parts for testing and training purpose or when data from some users is used for training while a different set of users' data is used to test the algorithm.

TABLE I. AMOUNT OF DATA USED

Sr. No.	Participant code	Number of data entries per mode						
		Walk	Bicycle	Motor Bike	Car	Bus	Train	Subway
1	Kb01	*264463*	*0*	*0*	*8648*	*66744*	*11969*	*128670*
2	Kb02	*0*	*0*	*0*	*0*	*0*	*0*	*22794*
3	Kb03	*0*	*0*	*0*	*468537*	*81922*	*0*	*0*
4	Kb04	*425487*	*97410*	*0*	*0*	*0*	*8260*	*0*
5	Kb05	*133256*	*34804*	*0*	*0*	*0*	*1409*	*0*
6	Kb06	*469147*	*0*	*440074*	*0*	*0*	*425421*	*0*
7	Kb07	*40284*	*0*	*0*	*0*	*0*	*0*	*162340*
8	Kb08	*1056845*	*0*	*0*	*9143*	*0*	*0*	*0*
9	Kb09	*609414*	*0*	*0*	*11440*	*58593*	*0*	*0*
10	Kb10	*178153*	*0*	*0*	*132788*	*0*	*181703*	*0*
11	Kb11	*130419*	*516920*	*0*	*0*	*0*	*0*	*0*

III. DATA COLLECTION

The data used in this study was taken from a large survey done in Kobe city, Japan. 50 participants used smartphones to collect GPS and accelerometer data during November 2013. The readings were recorded at an average frequency of 16 Hz or 16 readings per second. The collected data covered 7 modes of transportation namely, walk, bicycle, motor bike, car, bus, train and subway.

In this study, 11 respondents' data was selected among the 50 collected. The data comprised of location coordinates with time stamps, collected by GPS and accelerations along the three axes, recorded by accelerometer. Table I shows the amount of data entries for each mode with respect to each participant.

IV. FEATURE EXTRACTION

Nine different features were extracted from the GPS and accelerometer data as follows.

A. Features from GPS Data

Firstly, the entire data was divided into trips by keeping a transition time of 5 minutes. Between two consecutive data entries, if the time difference was more than 5 minutes then it was considered as the start of a new trip. Time difference feature was extracted by calculating the time lapse between consecutive entries within each trip. Time difference was set to zero at the start of each trip.

In addition to the time difference, four other features were extracted using Google Maps. The methodology was similar to the one used in [9]. Google Maps was employed to calculate driving distance, driving time, walking distance and walking time by inputting the coordinates provided by the GPS data. As there is a limitation on the number of queries that can be sent to Google Maps per day, coordinate pairs were spaced at 5 minutes apart. This means that the consecutive data entries were not used. Instead, data entries were taken after every 5 minutes or at the start/end of a trip, and their respective coordinates were used to form the O-D pairs. This step greatly reduced the number of queries to be sent to Google Maps. The calculated distances and times were then divided equally among the entries within each 5 minute window.

B. Features from Accelerometer

Smartphones are usually stored in different places and in different positions, a behavior affected by many factors including the gender and age of the users. For instance, some place their phones in purses or front/back pockets while others hold them while moving or simply place them in a mobile holder. All these different positions introduce variability to the acceleration data collected by the smartphone. To cope with all this, resultant acceleration was calculated and used, instead of using the accelerations along each axis.

Moving average resultant acceleration was calculated by using a window size of 1 minute. Same window size was used to calculate the maximum resultant acceleration and maximum average resultant acceleration. This resulted in four features being extracted from the acceleration data.

C. Normalization of Features

The features finally extracted were

(a) Time difference
(b) Driving distance
(c) Driving time
(d) Walking distance
(e) Walking time
(f) Resultant acceleration
(g) Maximum resultant acceleration
(h) Average resultant acceleration
(i) Maximum average resultant acceleration

The features were normalized from 0 to 1 so that each feature will have comparable effect.

V. Classification

For classification or identification purpose, Random Forest was used. The reason for its selection was its superior performance relative to other popular classification algorithms [10]. It is a supervised learning algorithm, which means that some data with known classes is initially fed to the algorithm. This data known as training/learning data allows the algorithm to recognize the patterns which distinguish one class from the other. Then similar data but without the knowledge of classes, known as test data, is provided to the algorithm. Depending on the lessons learned and the rules formulated from the training phase, the algorithm classifies the test data. Accuracy of the classification is judged by the number of examples correctly classified by the algorithm. Random forest constructs a huge number of decision trees using the training data. Each decision tree is used to individually classify the test data. The final class is decided by applying the maximum vote method on the individually predicted classes.

The main objective of this study was to compare the effect of using same participants' data and different participants' data for travel mode detection. For this purpose two analysis were done. In the first analysis, first 6 users' data was taken (Kb01 – Kb06). 50% of this data was randomly selected to form the training dataset and the rest was used as the test data. In the second analysis, all 11 users' data was used. It is clear from the table 1 that motor bike was used by only one participant i.e. Kb06. In order to include this mode, the data from Kb06 was divided into two parts. One part was added to the data by first 5 users to form the training dataset, while the other part was added to the data by last 5 users to form the test dataset.

TABLE II. CLASSIFICATION RESULTS WHEN USING SAME USERS' DATA

	Bicycle	**Bus**	**Car**	**Motor Bike**	**Subway**	**Train**	**Walk**	**Accuracy (%)**
Bicycle	***66104***	*2*	*0*	*0*	*0*	*0*	*1*	*100.00*
Bus	*0*	***74321***	*4*	*0*	*2*	*5*	*1*	*99.98*
Car	*0*	*0*	***238472***	*5*	*38*	*8*	*69*	*99.95*
Motor Bike	*0*	*0*	*0*	***220031***	*0*	*0*	*6*	*100.00*
Subway	*0*	*2*	*17*	*0*	***75712***	*0*	*1*	*99.97*
Train	*1*	*32*	*52*	*0*	*0*	***223441***	*3*	*99.96*
Walk	*0*	*22*	*5*	*0*	*3*	*14*	***646132***	*99.99*

TABLE III. CLASSIFICATION RESULTS WHEN USING DIFFERENT USERS' DATA

	Bicycle	**Bus**	**Car**	**Motor Bike**	**Subway**	**Train**	**Walk**	**Accuracy (%)**
Bicycle	***31491***	*28708*	*25682*	*1612*	*20751*	*20226*	*64780*	*16.3*
Bus	*1576*	***2995***	*11517*	*31353*	*1697*	*3888*	*5567*	*5.11*
Car	*1117*	*4833*	***3389***	*9555*	*27*	*14394*	*43997*	*4.38*
Motor Bike	*0*	*1*	*0*	***220235***	*0*	*0*	*0*	*100*
Subway	*1997*	*1430*	*3534*	*0*	***26709***	*6368*	*20990*	*43.77*
Train	*0*	*0*	*38*	*0*	*0*	***212481***	*1*	*99.98*
Walk	*19388*	*242890*	*60955*	*305766*	*47163*	*98073*	***614167***	*44.24*

VI. Results and Discussion

The confusion matrices along with the classification accuracies for both analysis are given in Table II and Table III. We can see that same users' data generate high classification accuracy for all modes. The least accurate is car with a prediction accuracy of 99.95%. On the other hand bicycle and motor bike are predicted with approximately 100% accuracy. But the picture is totally opposite when using different users' data for training and testing the same algorithm. Here the accuracy is very low.

The least accuracy is displayed by car, bus and bicycle respectively. The reason for such a strange behavior might be because when the data collected from the same users is divided into two parts, where one part is used to train the algorithm while the other part is used to test it, then both these datasets share some meta-information.

This sharing causes the test data to be classified more accurately. But when no such sharing is present, the prediction accuracy might reduce, as is evident from this study. Fig. 1 and Fig. 2 provide an example as to how this sharing of meta-information might affect the results. Fig. 1 depicts the scenario where data from two users only, covering two modes i.e. walk and car, is split to train and test the algorithm, whereas Fig. 2 shows the situation where data from user 1 is used to train the algorithm while testing is done by the data from user 2.

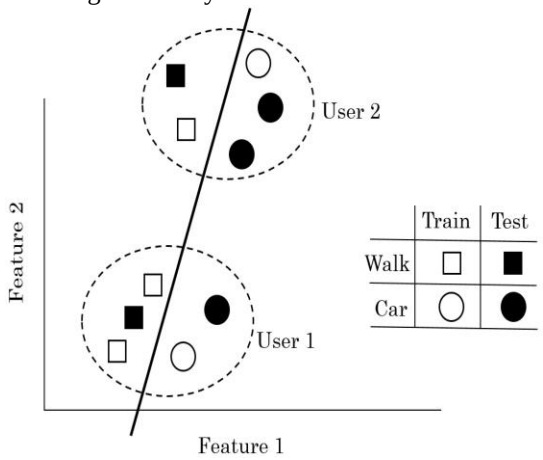

Figure 1. All users' data for training and testing

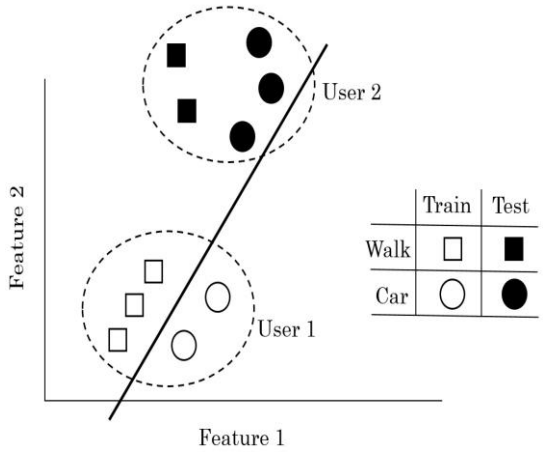

Figure 2. Different users' data for training and testing

VII. Conclusion and Future Work

Different sensors' data collected by smartphones can be successfully utilized to identify the mode of transportation used by the mobile carrier. Automatic mode identification by smartphones will have a huge impact on the way travel behavior data is collected and analyzed. This study focused on difference in accuracy achieved when data from all the participants is divided into two parts for training and subsequent testing purpose and when data from some participants is used to train while the rest is used for testing.

Results proposed that splitting the data provides high accuracy but the accuracy is very poor when different datasets are used. The reason might be because of the sharing of meta-information in case of split data. This indicates that splitting the data might result in over-estimation, therefore different datasets should be used in order to get the actual picture. The accuracy achieved in this study is very low therefore efforts should be made to improve it substantially. Testing non-conventional algorithms and extracting better features might ameliorate the situation.

References

[1] S. M. Turner, "Advanced techniques for travel time data collection," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1551, no. 1, pp. 51-58, 1996.

[2] T. L. Forrest and D. F. Pearson, "Comparison of trip determination methods in household travel surveys enhanced by a Global Positioning System," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1917, no. 1, pp. 63-71, 2005.

[3] N. Ohmori, M. Nakazato, and N. Harata, "GPS mobile phone-based activity diary survey," in *Proc. Eastern Asia Society for Transportation Studies*, vol. 5, 2005.

[4] J. Wolf, M. Oliveira, and M. Thompson, "Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1854, no. 1, pp. 189-198, 2003.

[5] T. Feng and H. JP Timmermans, "Transportation mode recognition using GPS and accelerometer data," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 118-130, 2013.

[6] B. Nham, K. Siangliulue, and S. Yeung, "Predicting mode of transport from Iphone accelerometer data," Machine Learning Final Projects, Stanford University, 2008.

[7] T. Nick, E. Coersmeier, J. Geldmacher, and J. Goetze, "Classifying means of transportation using mobile sensor data," in *Proc. International Joint Conference on Neural Networks*, 2010, pp. 1-6.

[8] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 645-662, 2010.

[9] M. A. Shafique and E. Hato, "Classification of travel data with multiple sensor information using random forest," Submitted to Transportation, Springer.

[10] M. A. Shafique and E. Hato, "Use of accelerometer data for transportation mode prediction," *Transportation*, Springer, 2014.

Muhammad Awais Shafique is currently a Ph.D. student in Transportation Research and Infrastructure Planning Laboratory at the University of Tokyo and an Assistant Professor in the department of Transportation Engineering and Management at the University of Engineering and Technology, Lahore. His research interests include application of machine learning and pattern recognition in transportation.

Eiji Hato is currently a Professor in Behavior in Networks Studies Unit at the University of Tokyo. He works in the area of travel behavior modelling, multi scale parallel simulation based on data assimilation and electric vehicle sharing services. He has received numerous awards including honorable mention in 2002 WCTR young prize from WCTRs and 2011 Kometani-Sasaki Award from ISSR.

