Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
TRA2015-1674

Optimum Training Sample Size for Travel Mode Detection using Smartphone

Muhammad Awais Shafique
PhD Candidate
The University of Tokyo
Japan

Eiji Hato
Professor
The University of Tokyo
Japan

ATINER CONFERENCE PAPER SERIES No: TRA2015-1674

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. This paper has been peer reviewed by at least two academic members of ATINER.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Shafique, M. A. and Hato, E. (2015). **"Optimum Training Sample Size for Travel Mode Detection using Smartphone",** Athens: ATINER'S Conference Paper Series, No: **TRA2015-1674.**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

ISSN: 2241-2891
02/11/2015

Optimum Training Sample Size for Travel Mode Detection using Smartphone

Muhammad Awais Shafique

Eiji Hato

Abstract

Smartphones are increasingly getting popular day by day. They are capturing the major share of the mobile market all over the world. Almost all of the smartphones come equipped with a lot of sensors including a GPS and accelerometer. These sensors have unlocked new possibilities. In this study, mobile technology is used to solve one of the problems present in the transportation sector. A methodology is described where the accelerometer data collected with the help of smartphones can be utilized to automatically predict the mode of transportation used by the phone carrier. A comparison is done among varying percentages of the training datasets to propose the needed optimum training sample size. It is shown that a ratio of 40%:60% between randomly selected training and testing datasets is appropriate when the percentage improvement is restricted to at least 0.1%.

Keywords: Accelerometer, Classification, Random Forest, Smartphone, Travel mode

Introduction

Smartphone penetration is on the rise in all the countries of the world. This vast explosion of smartphones has created a range of new industries and unlocked much greater possibilities. One of the possibilities is to induct this everyday use device into collecting data for transportation studies. Currently information pertaining to travelling is gathered through conventional data collection methods like questionnaire surveys, travel diaries, telephone surveys etc. These methods are laborious and time-consuming. Moreover the accuracy of the data collected is low because the responses depend on the memory of the respondents. They are very likely to forget the exact departure and arrival time for a certain trip; miss reporting short trips and/or having biased responses.

To counter these drawbacks in the existing practice, research is being done to utilize smartphones for an automatic travel data collection. Sensors like a GPS and accelerometer are present in almost all kinds of smartphones. With the help of GPS data, the location of the smartphone carrier can be tracked in real-time. This can provide us with the route information. On the other hand, accelerometer data can be employed to figure out the means of transportation used by the phone carrier. Accelerometer embedded in smartphones records acceleration along three axes with respect to the gravitational force. Each mode for example car or bus, exhibit a range of accelerations but mostly these ranges overlap making it quite difficult to distinguish among different modes. A slow moving car may have the same acceleration as a bicycle or a bicycle going downhill may have the same acceleration as a motor bike. Solving this ambiguity makes the current study interesting and practical.

The application of the developed methodology includes a partial or complete replacement of the conventional travel surveys. Moreover a customer oriented advertisement program can be initiated. For example if a phone carrier is moving by train then after locating him and detecting the mode of transportation (i.e. train), a message can be sent only to him advertising any bargains offered by the train company exclusively for his route or by any other retailers present near the stations along his route.

The rest of the paper is organized as follows: The next section provides an overview of the previous work done in this area. Section “Methodology” gives the methodology adopted in this study. The results and an accompanying discussion are provided in the “Results and Discussion” section whereas last section finishes the study by drawing conclusions and proposing future work.

Related Work

A lot of studies have focused on developing methodologies for detecting physical activities using accelerometer data [1-3]. In [4], IPhones were used to collect the accelerometer data for the purpose of differentiating among various modes of transportation. The collected data was fed to a classification algorithm namely Support Vector Machines (SVM) [5], consequently getting a prediction accuracy of 88% to 97%.

A study [6] did a comparison among two classification algorithms namely Naïve Bayes and SVM with the aim of differentiating among three modes of transportation. The results indicated that SVM performed better than Naïve Bayes with a prediction accuracy of 97.32%. Another study [7] used a wearable device to collect the acceleration data and made a comparison among four classifiers namely SVM, AdaBoost, decision trees and random forest. Random forest outperformed all others with an overall prediction accuracy of 99.8% while differentiating among four modes of transportation i.e. walk, bicycle, car and train. In [8], acceleration data covering six modes of transportation was collected. The average prediction accuracy achieved was 82.4%

This study is different from the previous researches as the acceleration data used is linked to seven modes of transportation including walking, bicycle, motor bike, car, bus, train and subway. Furthermore this study proposes an optimum sample size for training a classification algorithm.

Methodology

Data Collection

Smartphones were used by 8 respondents in Kobe city, Japan, to collect data during November 2013. The acceleration data was recorded along three axes i.e. x, y and z, at a frequency of 14 readings per second. The modes used by the respondents were walking, bicycle, motor bike, car, bus, train and subway.

Feature Extraction

Mobiles are kept differently by people while travelling. Some prefer to keep it in their purse, others in their front or back pocket. While travelling, the smartphone is repeatedly used to call, message or to check Facebook. All these practices necessitate an approach to cover all possible positions of the smartphone. Therefore instead of using the accelerations in three directions, their resultant was calculated. Then keeping a window size of 450 points, the moving average resultant acceleration was calculated.

Getting the maximum value was another feature to be extracted. Keeping the same window size, the maximum values were calculated for resultant accelerations as well as for the average resultant accelerations.

Finally following four features were extracted for the classification purpose.

- Resultant acceleration
- Maximum resultant acceleration
- Average resultant acceleration
- Maximum average resultant acceleration

Classification

Our previous study [7] proposed that random forest is a better option among classification algorithms, when it comes to classifying travel modes.

Therefore for this study random forest was employed. After the selection of the classifier, the next problem faced is the ratio of training the sample size to test sample size. For this purpose, the training sample was randomly selected from the whole dataset at ratios of 10%, 20%, 30%, 40%, 50%, 60% and 70%, while the rest was used to test the algorithm. Table 1 presents the amount of training and testing data sets used in this study.

Results and Discussion

Tables 2-3 show the confusion matrices for scenario 1 and 7 respectively. In both cases it can be seen that the subway is predicted with the least accuracy, mostly misclassified as walk. As a matter of fact, all modes are mostly misclassified as walk. The reason is the majority representation of the mode walk in the training dataset. Table 4 summarizes the prediction accuracies attained by the 7 scenarios used in the analysis. The results propose that the overall accuracy as well as the accuracy per mode increased with the increase in the amount of the training data. This result is obvious but how much is the improvement?

Table 5 presents the percentage increase in the overall accuracy while moving from one scenario to the next. It is clear that the rate of improvement kept on decreasing with the increase in the training data sample size. If the percentage improvement is restricted to at least 0.1% then scenario 4 gives the appropriate ratio between the training and testing datasets, because afterwards the improvement is less than 0.1%.

Table 1. *Amount of Training and Testing Datasets According to Scenarios*

Scenario	Data	Percentage	Modes						
			Bicycle	Bus	Car	Motor Bike	Subway	Train	Walk
1	Training	10	13221	17069	54485	44007	17845	46500	175406
	Testing	90	118992	153624	490362	396066	160608	418500	1578657
2	Training	20	26443	34139	108969	88015	35691	93000	350813
	Testing	80	105770	136554	435878	352058	142762	372000	1403250
3	Training	30	39664	51208	163454	132022	53536	139500	526219
	Testing	70	92549	119485	381393	308051	124917	325500	1227844
4	Training	40	52885	68277	217939	176029	71381	186000	701625
	Testing	60	79328	102416	326908	264044	107072	279000	1052438
5	Training	50	66107	85347	272424	220037	89227	232500	877032
	Testing	50	66107	85347	272424	220037	89227	232500	877032
6	Training	60	79328	102416	326908	264044	107072	279000	1052438
	Testing	40	52885	68277	217939	176029	71381	186000	701625
7	Training	70	92549	119485	381393	308051	124917	325500	1227844
	Testing	30	39664	51208	163454	132022	53536	139500	526219

Table 2. *Confusion Matrix for Scenario 1*

	Bicycle	Bus	Car	Motor Bike	Subway	Train	Walk	Accuracy (%)
Bicycle	111511	99	393	46	526	557	5860	93.71
Bus	247	146773	1200	313	324	274	4494	95.54
Car	252	325	469995	4536	304	750	14202	95.85
Motor Bike	25	45	2924	389552	6	2	3512	98.36
Subway	416	125	700	8	150501	469	8389	93.71
Train	851	193	1036	77	554	406643	9146	97.17
Walk	1534	706	5572	3426	1942	2225	1563253	99.02

Table 3. *Confusion Matrix for Scenario 7*

	Bicycle	Bus	Car	Motor Bike	Subway	Train	Walk	Accuracy (%)
Bicycle	39476	0	5	2	14	20	147	99.53
Bus	10	51026	27	18	4	5	118	99.64
Car	17	12	162868	135	11	47	364	99.64
Motor Bike	2	0	101	131822	1	0	96	99.85
Subway	7	4	27	0	53276	9	213	99.51
Train	18	2	21	1	12	139187	259	99.78
Walk	42	39	156	109	42	132	525699	99.90

Table 4. *Prediction Accuracy (%) for Different Scenarios*

Mode	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7
Bicycle	93.71	97.42	98.56	98.92	99.18	99.46	99.53
Bus	95.54	98.31	99.04	99.35	99.48	99.63	99.64
Car	95.85	98.18	99.03	99.29	99.47	99.60	99.64
Motor Bike	98.36	99.24	99.56	99.69	99.77	99.82	99.85
Subway	93.71	97.23	98.51	98.93	99.11	99.44	99.51
Train	97.17	98.71	99.28	99.48	99.63	99.69	99.78
Walk	99.02	99.56	99.75	99.82	99.86	99.90	99.90
Overall	97.62	98.96	99.42	99.59	99.68	99.76	99.80

Table 5. *Rate of Improvement between Scenarios*

Scenario	Improvement (%)
1 ➔ 2	1.368
2 ➔ 3	0.471
3 ➔ 4	0.163
4 ➔ 5	0.099
5 ➔ 6	0.081
6 ➔ 7	0.031

Conclusion and Future Work

Smartphones have become a vital part of our daily lifestyles. Due to their significant and vast spread, they can be successfully utilized to detect the mode of transportation used by the smartphone user. This will greatly assist the transportation personals in their task to capture the travel patterns of the general public so that new facilities could be efficiently designed and tailored to the attitudes of the people.

This study proposes a practical approach for detecting the transportation mode of the phone user by utilizing the accelerometer data collected. The overall accuracy achieved ranges from 97.62% for scenario 1 to 99.80% for scenario 7. Moreover, this study provides a comparison among the varying training data sample sizes to propose the optimum one. Restricting the rate of improvement to 0.1%, scenario 4 proves to be the optimum one with 40% of the data randomly selected to train the algorithm, while the rest 60% to test it.

The optimum sample size proposed in this study might be specific only to random forest. A similar analysis should be done for other algorithms, to verify their optimum levels.

References

[1] Bao, L., & Intille, S. S. (2004). "Activity recognition from user-annotated acceleration data," In Pervasive computing (pp. 1-17). Springer Berlin Heidelberg.

[2] Lester, J., Choudhury, T., & Borriello, G. (2006). "A practical approach to recognizing physical activities," In Pervasive Computing (pp. 1-16). Springer Berlin Heidelberg.

[3] Tapia, E. M., Intille, S. S., Haskell, W., Larson, K., Wright, J., King, A., & Friedman, R. (2007, October). "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor," In Wearable Computers, 2007 11th IEEE International Symposium on (pp. 37-40). IEEE.

[4] Nham, B., Siangliulue, K., & Yeung, S. (2008). "Predicting mode of transport from iphone accelerometer data," Machine Learning Final Projects, Stanford University.

[5] Chang, C. C., & Lin, C. J. (2011). "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

[6] Nick, T., Coersmeier, E., Geldmacher, J., & Goetze, J. (2010, July). "Classifying means of transportation using mobile sensor data," In Neural Networks (IJCNN), The 2010 International Joint Conference on (pp. 1-6). IEEE.

[7] Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. Transportation, 42(1), 163-188.

[8] Hemminki, S., Nurmi, P., & Tarkoma, S. (2013, November). "Accelerometer-based transportation mode detection on smartphones," In Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (p. 13). ACM.