Mehran University Research Journal of Engineering and Technology
Vol. 40, No. 3, 496 - 501, July 2021
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: https://doi.org/10.22581/muet1982.2103.04

Incorporating MNL Model into Random Forest for Travel Mode Detection

Muhammad Awais Shafique[1], Eiji Hato[2]

RECEIVED ON 26.08.2019, ACCEPTED ON 16.12.2020

ABSTRACT

Mode choice models have been used widely to forecast the relative probabilities of using available travel modes. These depend on mode-related and traveler-related characteristics. On the other hand, smartphones are increasingly being used to collect sensors' data relating to trips made after selection of a suitable mode. Such sensors' data may be correlated with decision-making process of travelers regarding travel mode selection. Discrete Choice Modelling is used to simulate this decision-making process by computing utilities of various travel alternatives, and then calculating their respective probabilities of being selected. In this paper, multinomial logit (MNL) mode choice model is utilized to enhance the prediction capacity of supervised learning algorithm i.e. Weighted Random Forest. To make the procedure less energy-intensive, GPS data was used only to locate the origin and destination of any trip, to be incorporated in mode choice model. Afterwards only accelerometer data was utilized in feature selection for the learning algorithm. One tenth of the classified data was used to train the algorithm whereas rest was used to test it. Results suggested that with incorporation of MNL, the overall prediction accuracy of learning algorithm was increased from 93.75% to 99.08%.

Keywords: Classification, Multinomial Logit, Random Forest, Supervised Learning, Travel Mode

1. INTRODUCTION

Knowledge of traffic movement patterns and decision-making processes has been of big interest to engineers and planners for decades. This information is mostly collected by surveys, whether in person or online, a procedure which introduces errors and uncertainty to the collected data, while limiting the number of participants and restricting the observed time span. Over the past years, vast dissemination of smartphones carrying sensors like GPS and accelerometer, has provided a great possibility to improve the quality and extent of traffic data collected. Research has been conducted to predict the travel mode and trip purpose from sensors' data, occasionally augmented with socio-economic, demographic and land use information. Hence, travel patterns and mode choices of a vast number of persons can be captured more realistically, over extended periods.

Another approach of predicting travel mode choice is by incorporating mode choice models. Being part of the conventional transportation-forecasting model, this method is mainly used to estimate the modal shares of all possible modes between the origin and destination. This paper proposes a novel approach of merging mode choice model and machine learning algorithm to improve the detection of travel mode.

2. LITERATURE REVIEW

Smart phones have become a popular tool to collect data and understand human behavior. Due to this

[1] Department of Civil Engineering, University of Central Punjab, Lahore, Pakistan.
Email: awais.shafique@ucp.edu.pk (Corresponding Author)
[2] Transportation Research and Infrastructure Planning Laboratory, Department of Civil Engineering, The University of Tokyo, Tokyo, Japan. Email: hato@civil.t.u-tokyo.ac.jp

reason, they are an unavoidable part of our daily lives. The data can be collected in chosen formats with help of variety of mobiles apps and software packages. The collected data can be used to infer hidden sequences in our seemingly random activities. This provides better insight into our shopping behavior or travel patterns. Travel mode prediction is one of the areas in which smartphones are providing new grounds to explore. Data collected by sensors integrated in the smartphones (Global Positioning System (GPS), accelerometer, gyroscope *etc.*) present an opportunity to overcome the many shortcomings in the conventional data collection methods like paper-based questionnaires *etc.* [1, 2].

Originally, only GPS data was utilized for travel mode prediction [3-7] but then researchers started collecting and analyzing data from other sources like accelerometer, gyroscope, barometer, orientation sensor etc. [8-12]. Accuracy was further improved by incorporating the Geographical Information System (GIS) data along with sensors' data [13]. Most of the studies incorporating machine learning follow the same methodology. Sensors' data is first cleaned from any noise present and then various suitable features are extracted like average acceleration, distance covered etc. Part of the processed data is used to train one or several machine learning algorithms and based on the heuristics learned, the algorithm is tested against the remaining data known as test data.

Despite the immense advantages of using smartphones as efficient and nonintrusive data collection tool, several challenges still exist. First, most of the studies employ GPS data, which is at times unstable due to signal loss in urban settings or warm/cold starts [14]. Further, GPS data collection is a very energy-consuming operation, which may affect the daily usage of smartphones by draining the battery at a relatively quick pace. This may discourage the smartphone owner from continuing with data collection. Although, a solution can be obtained by improving the hardware, it is much easier to decrease the GPS data collection frequency. In this paper, GPS data is only recorded at the start and end of any trip, so that the origin and destination may be marked. Travel mode identification is achieved using accelerometer data alone. The accuracy of the detection methodology is improved by incorporating multinomial logistic mode choice model.

3. MODE CHOICE

3.1 Discrete Choice

When it comes to mode choice, the individual is provided with a set of available modes, known as the individual's choice set. He/she must choose exactly one mode from the choice set. This setup is known as Discrete Choice, and is different from other common situations where the individuals can select any quantities from any of the available choices. For instance, while buying fruits or vegetables, the buyer is not bound to select only one kind of fruit or vegetable. Examples of discrete choice, other than mode choice, can be the choice of a school for study, a house to buy, or a car to purchase etc. It is obvious that there can only be one selection for the stated examples.

3.2 Understanding Choice

An analyst tries to understand the decision-making process involved in the choice of mode by collecting the observable data regarding all the modes present in the choice set as well as the observable information regarding the traveler. However, can it be claimed with surety that, provided the observable data is unchanged, the traveler's choice of transportation mode will remain the same? Certainly no. Because the traveler takes into account a number of factors, which are not observed by the analyst. These factors include emergencies, extreme weather, good/bad experience with a certain mode, change in routine, preferences of accompanying persons etc. Hence, there is some uncertainty associated with the decision-making process. The answer to this is the use of probability. Therefore, instead of deciding the one mode selected, probabilities for choosing each mode are determined.

3.3 Utility Function

The probability for selecting each mode is determined by calculating its respective utility. Utility is calculated by incorporating all the characteristics of each mode in the choice set. The decision-making

process is assumed to consist of two steps. In the first step, the traveler determines the utility for each mode present in the choice set, and in the second step, the mode with the highest utility is selected. The utility function can be divided into two components, deterministic and random. The deterministic component covers the characteristics of the mode, which are observable to the analyst, while the random component account for the characteristics observable only to the decision-maker. The utility function is formulated as given in equation (1).

$$U_{it} = V_{it} + \varepsilon_{it} \quad (1)$$

where U_{ij} = utility of the alternative i to the decision-maker t, V_{ij} = deterministic or observable portion of the utility, ε_{ij} = randomness or uncertainty, also known as error term.

There are various discrete choice models including Binomial Logit Model and Multinomial Probit Model but in context of the study, only Multinomial Logit Model is discussed.

4. MULTINOMIAL LOGIT MODEL

The mathematical form of a discrete choice model is determined by the assumptions made regarding the error term in the utility function. For Multinomial Logit Model (MNL) the assumptions are as follows,

- The error terms are Gumbel distributed.
- Identically and independently distributed across alternatives.
- Identically and independently distributed across observations.

In statistical and modeling literature, the errors are mostly assumed as normally distributed. In case of choice models, this assumption of normal distribution leads to Multinomial Probit Model (MNP), which is difficult to estimate due to the mathematical complexity involved. So, Gumbel distribution is selected because,

- It has computational advantages when maximization is required.
- It closely approximates normal distribution.
- It produces a closed-form model.

The Multinomial Logit Model gives the choice probabilities of each alternative as a function of the systematic utilities of all the alternatives. From a set of j alternatives, the probability of choosing alternative i (i = 1,2,…,j) is given by equation (2).

$$\Pr(i) = \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}} \quad (2)$$

where $\Pr(i)$ = Probability of decision-maker choosing alternative i, V_i, V_j = systematic utility of alternative i and j respectively.

5. METHODOLOGY

5.1 Data Collection

Smartphone sensors' data was collected by 50 participants in Kobe city, Japan, during the month of November in 2013. They used various android smartphones and a purpose-built mobile application. The respondents were asked to select the mode to be used, and press start once the trip was initiated. Upon reaching the destination, they were expected to press stop, in order to annotate the collected data with that particular trip ID. A recall survey was conducted to verify the accuracy of the collected trips. Six different types of travel modes were observed, including walk, bicycle, car, bus, train and subway. In addition to the trip data, accelerometer data was recorded continuously throughout each recorded trip.

5.2 MNL Model Estimation

60 seconds dwell time was used to segregate the trips [15]. After careful examination of each trip, it was found that some trips were originally misclassified by the participants. For instance, a few trips exceeding 200 minutes were classified as walk, which is completely illogical and cannot be true. Such trips were deleted. The reviewed number of trips and data instances are given in Table 1.

Following three types of attributes, corresponding to the available modes, were extracted using Google Maps.

- Travel time (all modes except bicycle)
- Travel distance (walk and car)
- Transit fare (bus, train and subway)

As mode bicycle was not available for Japan via Google Maps so data for walk was used to calculate the attributes of bicycle. The average walking speed is around 3.6 km/h whereas the average bicycling speed is about 20 km/h. These figures, although conservative, were used to convert the walking time for each trip into bicycling time. The distance was assumed same for both modes. The results for MNL model estimation can be seen in Table 2.

The Rho value shows good model fitting. Can this model assist in improving the detection accuracy? To answer this question, probabilities for choosing each mode were calculated for all the trips.

5.3 Feature Extraction

Features were extracted from accelerometer data by calculating moving averages using 15 seconds data windows. Features included resultant acceleration, average resultant acceleration, maximum resultant acceleration, maximum average resultant acceleration, standard deviation, skewness and kurtosis. To incorporate MNL into Random Forest, the probability values calculated by the MNL model were added as features. The data instances falling in a single trip were assigned the same probability values as calculated for the entire trip. Thus, the list of features now included probability of choosing walk, bicycle, car, bus, train and subway respectively, in addition to the features extracted from accelerometer data.

Table 1: Data Cleaning

Modes	Original No. of Trips	No. of Trips after Cleaning	Amount of Data
Walk	513	447	32566
Bicycle	10	9	1582
Car	31	29	3825
Bus	26	10	784
Train	43	33	5190
Subway	16	9	679
Total	639	537	44626

Table 2: MNL Model Estimate

Coefficients	Estimate	t-value	Pr (>|t|)
Bicycle (intercept)	-2.1594	-5.2584	0.00001 ***
Car (intercept)	-2.1973	-6.9320	0.00001 ***
Bus (intercept)	-2.0516	-2.7035	0.00355 **
Train (intercept)	-1.7399	-3.4604	0.00029 ***
Subway (intercept)	-1.8243	-2.2834	0.01141 *
Duration	-5.1948	-5.0254	0.00001 ***
Fare (Transit)	-0.1968	-0.5108	0.30478
Observations	537		
Initial log-likelihood	-305.925		
Final log-likelihood	-192.016		
Rho-squared	0.34619		
Signif. Codes: 0.001 '***' 0.01 '**' 0.05 '*'			

5.4 Weighted Random Forest and Post-Processing

To counter the problem of imbalanced data, as evident from Table 1, Weighted Random Forest was utilized for the prediction purpose. Here, instead of using the individual predictions made by each tree as basis for the voting system, the probabilities of each mode were multiplied by weights, calculated from the mode distribution in the training dataset, and then voting was performed using the maximum weighted probability mode predicted by each tree. The weights (W), calculated by equation 3, depended on the amount of data available in each class.

$$W_i = 0.5 + \frac{D_{min}}{D_i} \quad (3)$$

where W_i = Weight for class i, D_i = Data size of class i, D_{min} = Minimum data size among all classes.

This way the issue of imbalanced data was addressed, subsequently improving the prediction accuracy of the algorithm. One tenth of the data was used to train Weighted Random Forest, which was then applied to predict the remaining data. Prediction by the algorithm was followed by a 2-step post-processing. In the first step, voting was used to correct the prediction error, similar to the approach developed by [16]. It was found that the upper bound value of five, instead of

four provided better error correction. In the second step, the mode having maximum predictions within a trip was assigned to the entire trip.

6. RESULTS AND DISCUSSION

Firstly, only the MNL model was used to predict the mode selected, from the calculated probabilities of choosing each mode. Therefore, the estimation was performed solely based on data extracted from Google Maps. Secondly, the features extracted from accelerometer data were fed to Weighted Random Forest and mode detection was performed. In the third run, both approaches were combined. MNL Model was applied to determine the relative probabilities of choosing each mode, which were included with the accelerometer data as features for Weighted Random Forest. The results are summarized in **Table 3.** It is evident from the table that MNL is unsuitable for mode detection on its own. The results show that two modes i.e. bicycle and bus are totally left out with zero predictions. Poor results may be caused by the wrong perception of analyst regarding users' selection process, resulting in overlooking important decision-making variables for inclusion in the calculation of relative utilities.

Table 3: Prediction Results (%) using MNL, Machine Learning and Both

Modes	MNL	Machine Learning	MNL + Machine Learning
Walk	95.29	99.76	99.58
Bicycle	0.00	94.19	99.67
Car	17.20	88.31	99.08
Bus	0.00	81.42	93.72
Train	21.19	66.39	97.36
Subway	21.21	58.63	93.36
Overall	73.80	93.75	99.08
Time	-	76.04	75.57

Next, the use of machine learning algorithm demonstrates an overall accuracy of 93.75%. Only train and subway fall below 80% accuracy level. The performance is not satisfactory but when both approaches are combined, the resulting figures are remarkable. Unable to predict travel mode independently, MNL when paired with learning algorithm improved its prediction accuracy from 93.75% to 99.08%, with none of the modes showing a predicted with less than 93% accuracy. As evident from Table 3, detection accuracy is increased for all modes except "Walk". In fact, a slight decrease is observed from 99.76% to 99.58%. This is because number of trips recorded for "Walk" is strikingly disproportionate with rest of the modes. This over-sampling results in increased sensitivity of the mode towards slight changes in algorithm training. Nevertheless, the drop is quite miniscule.

7. CONCLUSION AND FUTURE WORK

This paper demonstrates that by incorporating MNL Model into the framework of Random Forest, the prediction power of the algorithm can be enhanced considerably. It not only improves the overall prediction accuracy but also enhances the classification of relatively less-represented modes. Further, with the infrequent collection of GPS time stamps (only to locate the origin and destination), and reliance on accelerometer data, the methodology is inherently made more energy-efficient. Hence, the smartphone users will not be bothered by the continuous collection of accelerometer data in the background, as it will not be affecting battery consumption significantly. The precise effect of data collection under various scenarios need to be investigated. Further, the MNL Model should be improved by including socio-economic variables specified by zones. This will surely improve the entire methodology. It would be intriguing to include other less common modes of transportation into the analysis, to come up with a methodology that would be easily applicable to different travel cultures.

REFERENCES

1. McGowen P., McNally M., "Evaluating the potential to predict activity types from GPS and GIS data", *Proceedings of the 86th meeting of the Transportation Research Board*, Washington D.C., U.S.A., 21- 25 January 2007.
2. Hato E., "Development of behavioral context addressable loggers in the shell for travel-activity analysis", *Transportation Research Part C: Emerging Technologies,* Vol. 18, No. 1, pp. 55-67, 2010.

3. Stopher P., FitzGerald C., Zhang J., "Search for a global positioning system device to measure person travel", *Transportation Research Part C: Emerging Technologies,* Vol. 16, pp. 350-369, 2008.
4. Bohte W., Maat K., "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands", *Transportation Research Part C: Emerging Technologies,* Vol. 17, No.3, pp. 285-297, 2009.
5. Chen C., Gong H., Lawson C., Bialostozky E., "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study", *Transportation Research Part A: Policy and Practice,* Vol. 44, pp. 8430-840, 2010.
6. Bolbol A., Cheng T., sapakisI. T, Haworth J., "Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification", *Computers, Environment and Urban Systems,* Vol. 36, pp. 526-537, 2012.
7. James J., "Travel Mode Identification With GPS Trajectories Using Wavelet Transform and Deep Learning", *IEEE Transactions on Intelligent Transportation Systems,* Vol. 22, No.2, pp. 1093-1103, 2020.
8. Shafique M. A., Hato E., "Use of acceleration data for transportation mode prediction", *Transportation,* Vol. 42, pp. 163-188, 2015.
9. Frendberg M., "*Determining Transportation Mode through Cellphone Sensor Fusion*," Bachelor of Science Undergraduate Thesis, Massachusetts Institute of Technology, 2011.
10. Su X., Caceres H., Tong H., He Q., "Travel Mode Identification with Smartphones", *Proceedings of the 94th Transportation Research Board Annual Meeting*, Washington D.C., U.S.A., 2015.
11. Sankaran K., Zhu M., Guo X. F., Ananda A. L., Chan M. C., Peh L.-S., "Using mobile phone barometer for low-power transportation context detection", *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, pp. 191-205, Memphis Tennessee, U.S.A., November 2014.
12. Wang C., Luo H., Zhao F., Qin Y., "Combining Residual and LSTM Recurrent Networks for Transportation Mode Detection Using Multimodal Sensors Integrated in Smartphones", *IEEE Transactions on Intelligent Transportation Systems,* pp. 1-13, 24 April 2020.
13. Zheng Y., Chen Y., Li Q., Xie X., Ma W.-Y., "Understanding transportation modes based on GPS data for web applications," *ACM Transactions on the Web,* Vol. 4, No. 1, 2010.
14. Gong H., Chen C., Bialostozky E., Lawson C. T., "A GPS/GIS method for travel mode detection in New York City", *Computers, Environment and Urban Systems,* vol. 36, pp. 131-139, 3// 2012.
15. Shafique M. A., Hato E., "Travel Mode Detection with varying Smartphone Data Collection Frequencies", *Sensors,* Vol. 16, No. 5, p. 716, 2016.
16. Yu M.-C., Yu T., Wang S.-C., Lin C.-J., Chang E. Y., "Big data small footprint: The design of a low-power classifier for detecting transportation modes", *Proceedings of the VLDB Endowment,* Vol. 7, pp. 1429-1440, 2014.