Commuter classification and behavior clustering: Beijing use case

Selene Baez Santamaria
s.baezsantamaria@student.vu.nl

Abstract

Public transportation, centered on subway and bus networks, is an data-rich domain that can benefit from data mining and machine learning techniques. The classification of commuters versus non-commuter/occasional travelers can help government, transport management and operators to better target their policies in order to improve the transportation network in large cities. Furthermore, characterizing commuters by behavior clustering can bring deeper insight into their needs and routines as a whole. This project proposes the usage of ensemble models for classification and clustering of public transport users. For this purpose, transit card data will be used, available from the city of Beijing, China.

1 Context

Large cities, like Beijing, China, require efficient and accessible public transit. The implementation of transit cards for usage and payment of transit (like IC card in Beijing) has made is possible to track users and gather large amounts of invaluable data on travelers behaviors.

Commuters represent a large part of daily travelers, as they consistently follow the same routes at approximately the same time. If correctly isolated, their routine trips can be targeted for improvement thus benefiting not only said commuters, but public transit users in general.

In 2014, previous work by Tu[1] used machine learning techniques to classify commuters versus non-commuters in the city of Beijing. A Support Vector Machine (SVM) was trained and reported 94.24% accuracy in their results.

2 Problem statement

Different shareholders in the transportation domain -such as government, transit management and operators- may gain a better understanding of commuters if these can be singled out from other less predictable types of travelers. As regular users of the public transportation network, commuters display preferences in mode of transportation (bus or subway in this case), lines (or routes), transfer connections, waiting times and economic value. Therefore, a detailed study of commuters can reveal hidden patterns in their behavior or exhibit shortcomings the public transport system.

For all these reasons, mining commuters data can enlighten the decision making processes for policy making and resource management of public transportation networks. This work aims to apply techniques of Data mining in the transportation domain and aid experts to explore and cope with the substantial amounts of data accessible to them.

3 Research questions

1. How accurately can commuters and non-commuters be identified using an ensemble model? How does this compare to the previous SVM model?

2. What is the minimal set of information needed from IC card data to reach an acceptable accuracy in classification?
3. To what extent is clustering commuters by its behavior informative to transportation specialists?

4 Method

In order to classify commuters versus non-commuters, a ensemble model was chosen. The choice of model is motivated by both its robustness and modularity by nature. Starting from two simple classifiers, assembled via bagging, the model can grow larger or more complex as needed and it may be extended beyond the scope of this Thesis Project. Furthermore, as suggested by Tu[1] results, the data is almost linearly separable thus simple classifiers such as decision trees may suffice.

The model will be trained, validated and tested using Beijing IC card data from 2014. A first instance of the model will use all available variables in the data (starting and ending stations and their related time stamps as well as travel mode, line ID, etc), as used by Tu[1] for a fair model comparison. Next, the set of variables will be revised to disregard redundant information. A second comparison with Tu[1]'s SVM model will be made.

Finally, commuters will be further clustered according to patterns in their behaviors that will emerge from all variables of the IC card data. The clusters will be analyzed and interpreted to find distinctive characteristics that may be judged as useful by transportation specialists.

Overall, the proposed project includes both supervised and unsupervised techniques for mining transportation data. Challenges on preprocessing massive amounts of IC card data will be faced. Furthermore, we anticipate possible drawbacks when clustering commuters behavior as detailed analysis of clusters must be done and interpreted by domain experts.

The project will be coded in Python, using libraries for data and numerical manipulation (*Pandas*, *numpy*), machine learning (*sklearn*), neural networks(*Theano*). The models will be evaluated based on their confusion matrices, which include its accuracy, precision, recall, as well as negative predictive value and specificity. Additional metrics such as cross entropy or hinge loss will also be reported.

5 Plan

This project is part of a collaboration between Vrije Universiteit Amsterdam (VU) and Beijing University of Technology (BJUT). As part of an exchange program, I will spend three months in Beijing, working at The College of Metropolitan Transportation under the supervision of Professor Jiancheng Weng. His PhD student Liang Quan will serve as the domain expert for the project and will provide with guidance in understanding the data and the context of the problem.

The project started on March 1st, 2017 and is programmed to be finished by August 31st, 2017. The main VU supervisor is Professor Zhisheng Huang.

A first presentation of the project (equivalent to a first KIM presentation) was presented on March 20th, 2017 to Professor Weng's team in the College of Metropolitan Transportation in Beijing. The final presentation (equivalent to a second KIM presentation) will be presented in Amsterdam, close to the end of the project.

The main steps to follow in this project are as follows:

Stage 1: Classification of commuters vs non-commuters

(a) Data exploration
(b) Determine appropriate type of classifiers for the task [1]
(c) Data preprocessing
(d) Individual classifiers set up
(e) Ensemble model by bagging
(f) Tuning: First round of experiments

[1] Most likely Random Forest, Bayesian methods or simple Multilayer Perceptron

Stage 2: Variable space reduction

(a) Input variable evaluation: qualitative analysis by domain experts and quantitative analysis by correlation tests

(b) Variable set selection: Second round of experiments

Stage 3: Clustering of commuters

(a) Determine hyper-parameters for clustering: input variables, number of clusters, number and type of classifiers [2]

(b) Data preprocessing

(c) Individual classifier set up

(d) Preliminary results analysis: qualitative analysis by domain experts

(e) Ensemble model by bagging

(f) Tuning: Third round of experiments

References

[1] Tu, Q. Weng, J. C. & Yuan, R. L. Impact Analysis of Public Transport Fare Adjustment on Travel Mode Choice for Travelers in Beijing. *CICTP 2016.*, pp. 850–863.

[2] Most likely Neural Networks