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, and transport management and operators to better target their policies for improving 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 is available from the city of Beijing, China.

1 Context

Large cities, like Beijing, China, require efficient and accessible public transit. The usage 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

Commuters must be identified in order for the government and transit operators and management to understand them better.

Furthermore, a more detailed study of commuters behavior can reveal hidden patterns or malfunctions in the public transport system. Policies, and resource management can be informed by mining commuters data.

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?

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

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 nature. Starting from two simple classifiers, assembled via bagging, the model can grow larger or more complex as needed and it may be continued even 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 on 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 and 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 recorded by 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.

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 Transportation under the supervision of Professor Jiancheng Weng.

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.

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.