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.

Contents

1 Introduction

1.1 Transportation domain

Public transportation facilities composed by subway, buses, tram, train and others. As a network, they provide service for the majority of citizens. [1]

Environmental impact, air pollution, noise pollution. [2].

Energy problems for fuel consumption. Even economic implications. [3]

1.1.1 Who are the commuters?

Regular users of public transit.

Commuting to work or school is the basis for a routine. It directly influences personal life and impacts quality of life [4]. If the experience is bad, daily travel can bring sorrow to users. Bad experiences may include excessively long commuting time, crowded spaces, inconvenient transfers, elevated prices, low-quality of facilities, and others.

Identifying commuters can help in the long-term planning of public transportation. Policies for improving the overall experience and aid urban areas on a large scale.

Transportation follows swarm behavior. Based on individual travels and routines, on a larger scale travelers exhibit peculiar characteristics. Both levels of understanding are crucial.

1.2 The city of Beijing

Beijing special case for urbanization. Number of people is massive [5]

Pollution is Beijing [6]

Resources of public transportation network. Number of subway lines and bus lines. Number of users per day. [7]

1.3 Motivation

Interdisciplinary study between Artificial Intelligence and Metropolitan Transportation. Introduce data mining techniques to a data rich domain.

Relevance of project on both areas.

1.3.1 Societal context

Commuters use the public transport network regularly to go to work, school or other follow other routines. They need reliable means of transportation.

Government, transport management and operators can gain spatial and temporal insight. This insight can lead to tangible results, policies and counter measures increasing efficiency of network, adjustable travel fares used as incentives to relieve peak hours, urban planning for residential and industrial land use, and others [8]

[1] reference
[2] references
[3] reference
[4] reference
[5] reference
[6] reference
[7] reference
[8] reference

1.3.2 Scientific context

Usage of machine learning of data mining has been limited. Current broadly use method is surveys to reach travelers on individual level, aggregated measurements for gathering their collective behavior. The analysis is usually done with statistical methodology.

Surveys are costly and based on self-report, which by itself has bias problems. Other problems are small population and non-representative samples.

Aggregated methods miss the interactions between individuals that cause the collective behavior.

Technology has reached the data collection point, but has yet to reach the analysis part. Transit cards are capable of recording boarding and alighting stations with their related locations and time stamps.

Many prediction algorithms available. Constant refinement, state of the art must be applied to real life and large impact situations. Domain experts must focus on analyzing insights and using them, not on techniques for curating and making sense out of raw data.

2 Literature review

2.1 Data mining on transit card data

Preprocess data by Wang in BJUT lab. [7]

Data mining to identify transit use cycles in Canadian smart card data [5]

Density Based Scanning Algorithm with Noise to classify travelers according to their travel patterns. [4]

Passenger segmentation by K-means clustering [1]

Machine learning for commuters identification. SVM with 94% accuracy. [6]

11 distinct clusters of users with similar activity and demographic attributes [2]

Latest work using machine learning by [3]

2.2 Classifying and clustering spatio-temporal data

Ensemble methods

Classifiers in the transportation domain

3 Research objective

Objective is to identify and characterize commuters in the city of Beijing by using IC card data.

3.1 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?

3.1.1 Definition of terms

A commuter is someone whose IC card data is repeatable in time and space over a working week (5 days, Monday to Friday).

3.2 Research approach

3.2.1 Ensemble models

Robust and modular. May grow in complexity as needed

3.2.2 Decision trees and random forests

3.2.3 Neural networks

4 Methodology

4.1 Data description

Every record for an IC card contains the following data fields:

- Data date: date that the trip was made
- Card code: card identification number
- Data link: [9]
- Path link: Mode of transportation. B for bus, R for subway. Transfers shown by a dash. Example: B-B is Bus to Bus.
- Travel time: time spent in vehicles, measured in milliseconds
- Travel distance: measured in meters [10]
- Transfer number: number of transfers in the trip
- Transfer time average: time spent in transfer, divided by number of transfers
- Start time: time stamp of when the trip started
- End time: time stamp of when the trip ended
- On traffic:
- Off traffic:
- On middle area:
- Off middle area:
- On big area:
- Off big area:
- On ring road:
- Off ring road:
- On area:
- Off area:
- ID: number| time stamp of beginning of trip | card code
- Transfer detail: Station name, line number, mode of transportation

Every day, more than 50,000 records are collected. This project aims to include data from at least one week.

4.1.1 Training data

Since we perform supervised learning, we need training data for which we know if a record corresponds to a commuter or non-commuter. Such data is expensive and limited since it has only been obtained by asking the users directly if they are commuters or not. Other annotated data is not available, and labeling new records falls beyond the scope of this project. [11]

[9] irrelevant?

[10] as measured by route?

[11] if data is not sufficient (although previous work shows it is) I might need to consider annotating some data myself

The current training and validation set consists of data from 2015, collected and validated by Tu[1] [12]. The data is composed by:

- 6439 records of 481 commuters
- 1628 records of 497 non-commuters

For a total of 978 IC card IDs.

4.1.2 Testing data

Testing data

4.2 Data cleaning

Eliminate records that do not make sense or are faulty, for example having empty fields.

4.3 Redundant variables

Hypothesis: middle area, big area and area overlap. Middle has more precision but maybe not needed.

5 Commuters identification

5.1 Model

5.2 Experiments

5.3 Results

Accuracy

Confusion matrix

6 Variable evaluation

6.1 Qualitative

Exploration: Experts opinion

6.2 Quantitative

Analysis: Correlation

7 Commuters clustering

7.1 Expert judgment

8 Conclusion

9 Future work

References

[1] Ashish Bhaskar, Edward Chung, et al. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3):1537–1548, 2015.

[12] change references

[2] Gabriel Goulet Langlois, Haris N Koutsopoulos, and Jinhua Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16, 2016.

[3] Xiaolei Ma, Congcong Liu, Huimin Wen, Yunpeng Wang, and Yao-Jan Wu. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58:135–145, 2017.

[4] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

[5] Catherine Morency, Martin Trepanier, and Bruno Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[6] Qiang Tu, Jian-cheng Weng, Rong-Liang Yuan, and Peng-fei Lin. Impact analysis of public transport fare adjustment. *Traffic Engineering & Control*, 57(2), 2016.

[7] Yueyue Wang. Research on methods of extracting commuting trip characteristic based on public transportation multi-sourced data. *Beijing University of Technology*, 2014.