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 Introduction

1.1 Transportation domain

Public transportation facilities. Common problem to identify regular users.

1.2 Beijing

Resources of public tranpsortation network. Number of subway lines and bus lines. Number of users per day.

1.3 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.

1.4 Scientific context

Usage of machine learning of data mining has been limited

Interdisciplinary study between Artificial Intelligence and Metropolitan Transportation.

2 Related work

Preprocess data by Wang in BJUT lab.

Machine learning for commuters identification. SVM with 94% accuracy.

Ensemble methods

Classifiers in the transportation domain

3 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?

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: [1]
- 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 [2]
- 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.

[1] irrelevant?
[2] as measured by route?

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. [3]

The current training and validation set consists of data from 2015, collected and validated by Tu[1] [4]. 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 Results

5.1 Commuters identification

Accuracy

Confusion matrix

5.2 Variable evaluation

5.2.1 Qualitative

Exploration: Experts opinion

5.2.2 Quantitative

Analysis: Correlation

5.3 Commuters clustering

5.3.1 Expert judgment

6 Conclusion

7 Future work

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.

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

[4] change references