Learning Customer Usage Patterns From SmartRider Ticketing Logs

Lidia Dokuchaeva

Rachel Cardell-Oliver
School of Computer Science and Software Engineering
University of Western Australia

Sharon Biermann
CEED Client: Planning and Transport Research Centre (PATREC)

Abstract

Understanding public transport customer behaviors is an important problem in urban planning. The goal of this project is to characterize customer types using the ticketing logs of an urban public transport network. Non-negative matrix factorization was used to learn latent customer activities from a database of transport smart card transactions. Clustering customers' activity profiles generated a set of distinct travel patterns. Using a database of trip logs from Perth, Australia, we evaluated the discovered travel patterns by using internal and external cluster quality metrics. Our data driven approach automatically learns understandable customer categories from a large-scale transport smart card database.

1. Introduction

Understanding urban mobility patterns is important for making modern cities liveable and productive. Studying public transport in particular is critical for understanding urban mobility as a whole. This project aims to reveal latent information from smart card data related to customer segmentation for public transport users. In this project, the latest data mining and machine learning techniques were used and evaluated in order to develop and test a system that exploits SmartRider logs to characterise the spatial and temporal habits of individual customers. The paper presents the results obtained for customers with high frequency travel.

1.1 Background and Motivation

Public transport is a vital part of urban infrastructure, and studying the way that people use it is critical for understanding urban mobility as a whole. The paper describes a project focusing on evaluation based on data mining methods in order to develop and test a system that exploits SmartRider logs to characterise the spatial and temporal habits of individual customers and to address the issue of customer segmentation for public transport users.

Clustering is a machine learning technique for identifying similar groups of individuals in a population. For travel logs, it can be used to identify types of customers (as in this project), types of stops or types of journeys. By clustering the ticketing logs, it is possible to find the common ways that public transport is used, and by finding how public transport is being used, it is possible to evaluate the effectives of the public transport system overall - whether by performing further analysis on the data set or by combining this knowledge with other knowledge such as household travel surveys. As such, clustering is an important way of understanding public transport, and thus urban mobility as a whole.

Non-negative matrix factorization (NMF) can be used for clustering transport customers. NMF generates a multi-scale representation of the user activities. NMF has been used to analyse traffic flow (Peng et al. 2012) and profiles of Paris metro customers (Poussevin et al. 2016). Poussevin focused mainly on using the resulting model to cluster stations, not users. Although there have been some studies using smart card data, there is still a lot of potential for mining the data to better understand urban mobility.

Token	**Number**	**%**	**Token**	**Number**	**%**
Standard	23794	31	Health Care	1853	2%
Student 50 cent	14061	18	Pensioner	1507	2%
Senior Off-Peak	9817	13	Freerider	420	1%
Student (Up to Yr 12)	9950	13 %	Convention Pass	5	0%
Student Tertiary	6986	9%	PTA Concession	39	0%
Senior	5358	7%	PTA Free Pass	131	0%
Pensioner Off-Peak	2021	3%	Veteran	84	0%

Table 1 Token Distributions from SmartRider Ticketing Logs Oct 2016

2. Methodology

Currently, Perth's SmartRider data is sent by the Public Transport Authority (PTA) to a third party data warehousing company specialising in ticket log data (netBI). netBI provide a data warehousing service. They clear and archive the log data, and provide statistical reports to PTA. The SmartRider data used for this project has been sourced from netBI under an agreement with PTA and PATREC. The dataset used includes records from October 2016. Each months's data has three parts: a listing with the bus stops, a listing of the train stops, and the listing of SmartRider transactions. This project examines the SmartRider transactions only.

A transaction begins with the user tagging on at a particular location, and ends with the user tagging off. The user ID, the time and location they tagged on, as well as the time and location they tagged off, are all recorded. Additionally, the card type (e.g. student, adult, etc.), the distance travelled and the fare paid are recorded, as well as some other metrics.

2.1 Process

First, we pre-process data into an acceptable format, such as removing the errors and converting the logs to appropriate formats for analysis. Next, starting with a database of smart card data, we construct the set of boarding times associated with each user. This representation depends on the events, where an event is a single boarding trip. Using the boarding times, a user model

vector was constructed of probabilities of usage in three frequency bands: high frequency (for events occurring more than twice a week), medium frequency (for events occurring at least once each 10 days), and low frequency (for unusual events). This representation was experimented with to find the best representation for our task. The resulting model allows the construction of a vector of probabilities for each event.

We then collected the vectors for all users with activities in the same frequency band. Using non-negative matrix factorization the significant activities were extracted. Using these activities, a clustering algorithm was applied. Specifically, the CLARA algorithm was used, which is a k-means hierarchical method that clusters around medoids rather than means and is optimized for large datasets (Maechler et al 2017). The non-negative matrix factorization used in this particular case is the non-smooth NMF method (Pascual-Montano et al. 2006).The clusters revealed were evaluated by considering real-world interpretations, considering clustering metrics e.g. distance and coherence, and associations between clusters and other customer information.

3. Results and Discussion

The dataset used for this project was the SmartRider ticketing logs for October 2016. For analysis, a set of roughly 75000 card IDs were chosen from the dataset of 420000 cards. This was done as it is the maximum amount of data the system could handle. These cards were chosen purposefully randomly in order to best approximate the dataset. In order to have a random selection that is similar, the card IDs were chosen to have the same proportion of tokens as the dataset as a whole.

3.1 Frequency Bands

The set of trips was reduced so that only unique user and location couples remained. After this reduction, 355943 trips from a dataset of 1048540 trips remained. The resulting dataset was then divided into the three different frequency matrices (high, medium, and low) described in section 2.1. Note that in future studies, the parameters for the frequency bands may be adjusted for different definitions of 'high', 'medium', and 'low'.

After the frequency bands were separated, the above algorithm was repeated, only per frequency band and aggregating the trips by user (thus dropping the Location data). Table 3 displays summarizes characteristics of the resulting frequency bands.

3.2 Non-negative Matrix Factorization

The activities were extracted by running the nsNMF algorithm on the frequency bands (Gaujoux & Seoighe 2010). There are some properties immediately apparent from this representation.

Frequency:	**High**	**Medium**	**Low**
Number of Trips	12257	12070	30942
Most Frequent 15min	07:30	09:00	09:00

Table 1 Frequency Band Statistics

Some of the more interesting results (represented in Figures 1 and 2) are noted below:

Figure 1 High frequencies activities 1-7, over a 24 hour period and a 7-day period.

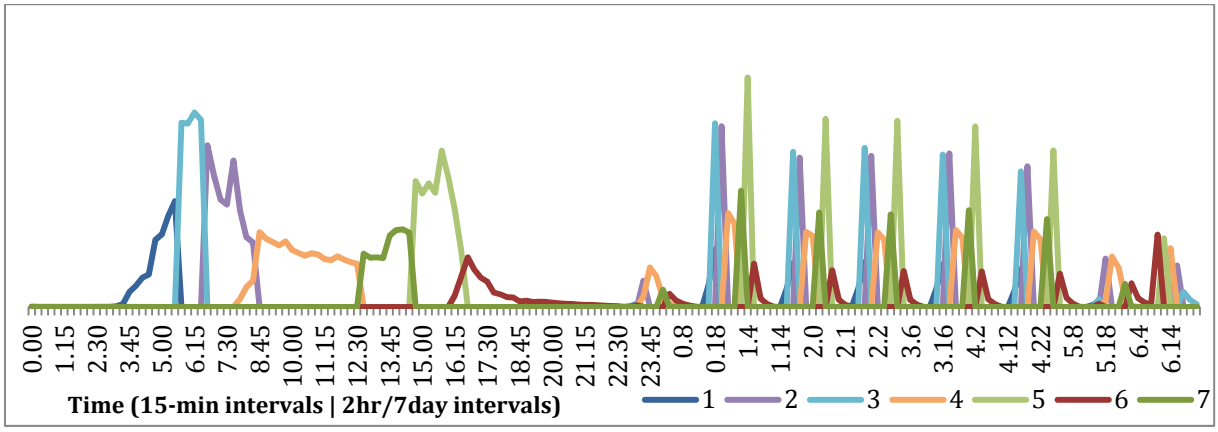
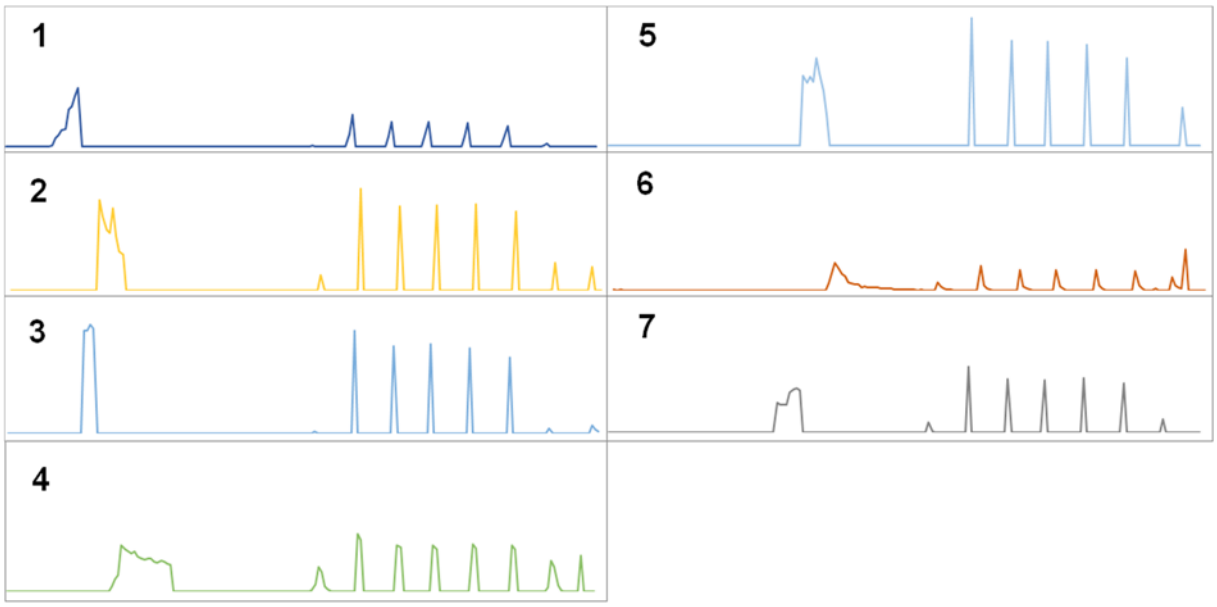


Figure 2 High Frequency Activities – Individual. Activities 2, 3, and 5 can be classified as going to work activities, with activities 1, 6 and 7 as going home activities, with activity 6 having a significant peak on Saturdays. Activity 4 is a lunch activity, and notably has a significant presence on the weekends.



- Activity 3 can be described as the ’going to work’ activity, given the fact that it is almost exclusively around 6:15, and occurs only on weekdays.
- Activity 7 is similar to the ’going to work’ activity, in that it is a monomodal peak that only occurs on weekdays, except that it can be characterized as a ’going home’ activity. It is less concentrated than Activity 3, and does not peak as high. This can be explained due to the lack of time pressure on going home.
- Activity 4 ranges from 8:00 to 14:00, and thus covers most of the morning and early afternoon - it does not peak as high as the other activities, but it exists on every single day, and is the most significant weekend activity. This could potentially be a going to lunch or shopping activity.

Token	A	B	C	D	E	F	**Total**
Senior	11	**41**	28	8	6	7	100
Pensioner	16	**36**	23	8	6	10	100
Standard	9	2	17	**38**	20	14	100
Student Tertiary	16	12	**31**	12	9	20	100
Health Care	14	11	**28**	19	10	18	100
Veteran	0	33	0	**67**	0	0	100
Student (Up to Yr	**43**	5	19	20	3	10	100
Student 50 cent	**51**	3	18	20	2	6	100
PTA Concession	8	0	13	**46**	15	18	100
Freerider	0	0	0	0	0	0	0
PTA Free Pass	**42**	0	8	**33**	8	8	100
Total	12	18	22	24	13	12	100

Table 2 High Frequency - Percentage of Clusters per Token

3.3 Clustering

For each of the previously identified NMF activities, the CLARA clustering algorithm has been run. The high frequency activities generated ten clusters; the medium frequency activities generated seven clusters; the low frequency activities generated nine clusters. All are dominated by one or more of the activities extracted by the NMF. For instance, the dominant activity of cluster D is a going to work activity combined with a going home activity. This implies that, as might be expected, the most common high frequency activity is the morning commute activity. However, it also has hints of other activities as well – a little of a slightly earlier going to work activity, the dominant going home activity, and even less of the other going home activities. In general, each activity corresponds to roughly one cluster that is dominated by this activity.

These clusters exhibit some of the trends of high frequency customers. Cluster E peaks significantly around 7:30, and then has a somewhat less significant, yet still apparent, peak around 17:15. This can be easily categorized as a typical commuter pattern. Cluster F peaks earlier in the morning, but has a similar afternoon peak, if less significant - this can be another commuter class. In fact, other than cluster D, which does not have obvious peaks around daylight hours, all the high frequency clusters seem to be commuter clusters. Cluster D can perhaps be explained when compared to card types, since it is heavily dominated by Senior card types.

4. Conclusion and Future Work

This paper examines some of the customer clusters that arise in Perth SmartRider data. The results presented here include some of the common high frequency activities found across the dataset. Using non-negative matrix factorization, the dataset was decomposed into a series of activities, and the customers were clustered based around those activities. This project discovers that decomposing the ticketing logs into frequency matrices and running NMF on these matrices produces a viable set of activities that can be used to improve the quality of discovered clusters. This approach can be used to run over different sets of the data, or over different frequency

matrices, in order to discover interesting customer classes, some of which have been discussed in this paper.

Overall, this has been yet a preliminary study of the user clusters in Perth's SmartRider data. There is still work that can be done on this dataset and other smart card data of this type. For instance, the effects of other clustering algorithms can be studied in order to compare the various types of clusters found. Additionally, comparison to different types of NMF, such as the Brunet method or the ALS method, can be undertaken.

Other information than just boarding time and location can be used, such as alighting time and location, in order to create more finely grained clusters. Additionally, boarding location can be expanded to include boarding zone, whether it is the STEM zones used by the Department of Transport or some other zones. Other information, such as the total amount of kilometers travelled, can also be considered. In short, full use can be made of the rich dataset available.

5. Acknowledgements

I would like to extend a thank you to my supervisor, Rachel Cardell-Oliver, the staff at CEED and PATREC for supporting me with this project, and PTA/Transperth for SmartRider data access. Additionally, I would like to thank the Department of Transport for their support, in particular Renlong Han and Alan Kleidon.

6. References

Gaujoux, R. & Seoighe, C. (2010). "A flexible R package for nonnegative matrix factorization", *BMC Bioinformatics*, **11**(1), p. 367.

Kieu, L. M., Bhaskar, A. & Chung, E. (2015). "Passenger Segmentation Using Smart Card Data", *IEEE Transactions on Intelligent Transportation Systems,* **16**(3), pp. 1537-1548.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2017). "cluster: Cluster Analysis Basics and Extensions.", R package version 2.0.6.

Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D. & Pascual-Marqui, R. D. (2006), "Nonsmooth nonnegative matrix factorization", *Pami*, **28**(3), pp. 403-415.

Pelletier, M.-P., Trépanier, M. & Morency, C. (2011), "Smart card data use in public transit: A literature review", *Transportation Research Part C: Emerging Technologies,* **19**(4), pp. 557-568.

Peng, C., Jin, X., Wong, K.-C., Shi, M. & Lio, P. (2012), "Collective Human Mobility Pattern from Taxi Trips in Urban Area", *PLoS ONE*, **7**(4), pp. 1-8.

Poussevin, M., Tonnelier, E., Baskiotis, N., Guigue, V. & Gallinari, P. (2016). Mining Ticketing Logs for Usage Characterization with Nonnegative Matrix Factorization", *SenseML*, **9546**, pp. 147-164.