Trip purpose identification using pairwise constraints based semi-supervised clustering

N. Sari Aslam[a], T. Cheng [b], J. Cheshire [c], Y. Zhang [d]

[a b d] Department of Civil, Environmental and Geomatic Engineering, [c] Department of Geography, University College London (UCL), Gower St, London, UK

14 April 2019

Summary

Clustering of smart card data captured by automated fare collection (AFC) systems has traditionally been viewed as an unsupervised method. However, some additional information about human behaviour is available in addition to the smart card data points that can facilitate better partitioning of the data. In this paper, such prior knowledge is translated into pairwise constraints and used with the COP-KMEANS clustering algorithm to identify user activities. The effectiveness of the method was evaluated using performance evaluation measures by comparison of the results with the ground truth. The results demonstrate that pairwise constraints significantly enhance the accuracy of the clusters.

KEYWORDS: Trip purposes, smart card data, COP-KMEANS, semi-supervised clustering

1. Introduction

The availability of digital footprints of user data sourced from the system such as Smart card, GPS devices and mobile phone (Kong *et al.*, 2009) have seen a massive increase in recent decades. Utilising these resources have the potential to help the problems of traffic congestions and urban planning. With this context, the data collected via AFC systems are a valuable resource in transportation networks that can be used to garner a better understanding of human mobility and provide sustainable transportation (Sari Aslam, 2015; Sari Aslam, Cheshire and Cheng, 2015).

Using smart card data is significantly efficient and available for a much larger population compared to traditional survey data and helps to identify primary locations using the characteristics of smart card data (Sari Aslam and Cheng, 2018; Sari Aslam, Cheng and Cheshire, 2019). On the other hand, there are challenges to identify social demographics (Zhang, Cheng and Sari Aslam, 2019) and the purpose of the trip only by looking at the smart card data alone (Devillaine, Munizaga and Trépanier, 2012; Faroqi, Mesbah and Kim, 2018).

This study, therefore, aims to carry out the preliminary work to build a framework of behavioural analysis for the identification of the purpose of the trip using semi-supervised clustering. Pairwise constraints (must-link and cannot-link) based on semi-supervised clustering are applied to understand the meaning of the segments which are related to individuals' activities. The results were evaluated using performance evaluation methods (FMI) by comparison with the ground truth. The results demonstrate that clustering algorithms provide better accuracy in the classification of activities when prior knowledge is added to the algorithm by means of pairwise constraints.

[a]n.aslam.11@ucl.ac.uk,
[b]tao.cheng@ucl.ac.uk,
[c]james.cheshire@ucl.ac.uk,

[d]yang.zhang.16@ucl.ac.uk

2. Methodology

Figure 1 illustrates the framework of the methodology. Label and unlabelled smart card data as an input used for i) data processing to select the right features and ii) to create constraints to apply COP-KMEANS (Wagstaff, Rogers and Schroedl, 2001). After the model validation section, the results presented to infer trip purposes as an outcome.

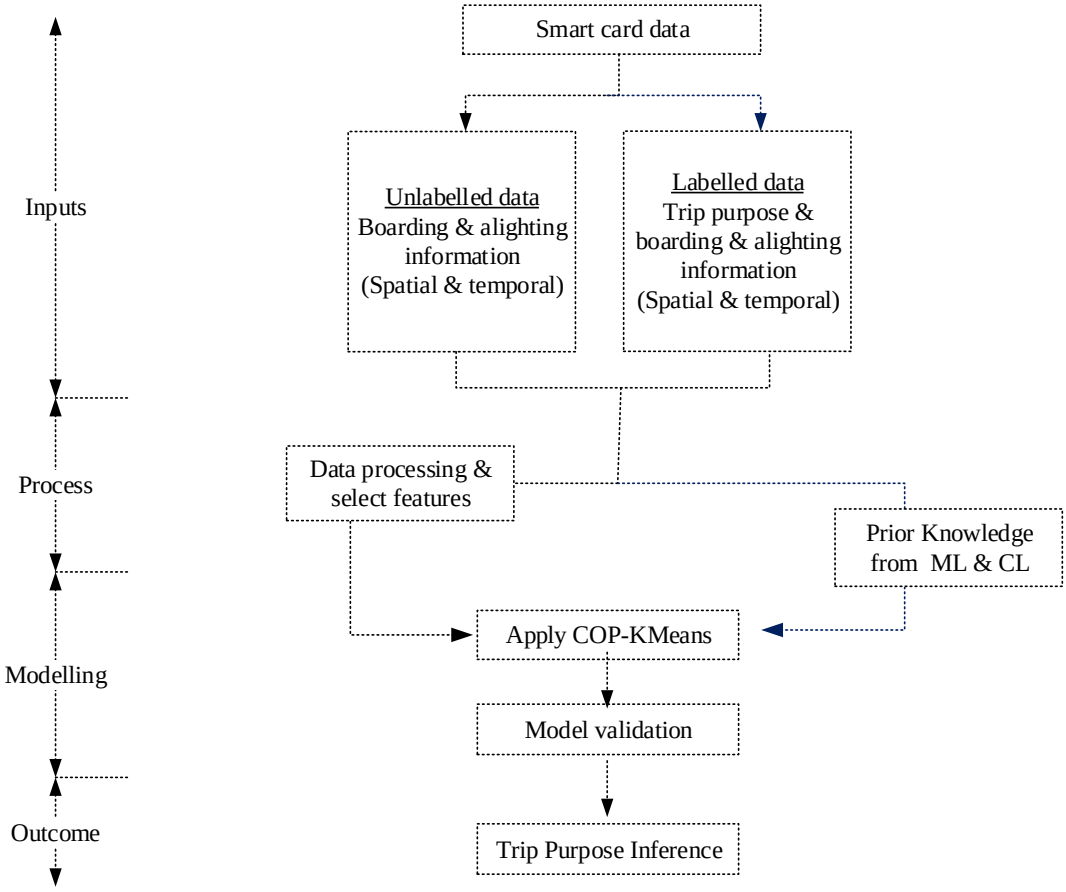

Figure 1: Flowchart of the framework (ML and CL refers to must-link, and cannot-link constrains respectively)

1.1. Data Description

There are two sources of data available for this study. The first one is unlabelled data that contain the transit record of completed journeys by 10000 (randomly selected) individuals from October to November 2013. The second one is labelled data from 40 volunteers' users for two months in 2018. That means approximately 4000 labelled data points available for the purpose of creating pairwise constraints and validation. The key benefit of using semi-supervised learning is that it allows us to leverage the vast amount of unlabelled data with a limited amount of labelled data (Peikari *et al.*, 2018).

The labelled classification includes home and work-related activities as well as other activities such as after-before and midday activities. Both dataset also includes attributes such as entry date/time, entry station, exit date/time, exit station and transport modes such as London Underground, train, London Overground, tram and bus.

2.2. Data Processing

Data pre-processing is fundamentally a series of exclusion steps in order to clean the data. Since an activity defined the time spent (duration) at a specific station between two consecutive journeys, single journeys are excluded from the dataset. Additionally, due to single tap-in, bus journeys were also excluded from the analysis as they do not contain the complete spatial and temporal information of the journey.

2.2.1. Feature Selection

Activity extraction step was followed by the identification of additional *special features* (feature extraction) such as 'home location' and 'work location'. For the majority of the users, the key locations identified using a heuristic approach defined by Sari Aslam, Cheng and Cheshire (2019). That information combined with 'Activity From' and 'Activity To' knowledge as direction. Additionally, *temporal features* extracted such as 'weekend flag',' the day of the activity', 'start hour' and 'end hour' of the activity (Sari Aslam, Cheng and Cheshire, 2018). In the end, features were scaled to normalise the range of independent input variables and the valuable features selected using automated feature selection using sklearn.lib.feature selection function.

2.2.2. Pairwise constraints

One of the most common technique is to create pairwise constrain from prior knowledge by identifying the data points that should or should not be grouped in the same cluster. Usually, pairwise constraints are inferred from the labelled data or the background information known of the dataset (Wagstaff, Rogers and Schroedl, 2001).

The approach taken in this paper makes use of the labelled data to create two types of pairwise constraints, which are must-link ((ML) and cannot-link constraints (CL). ML constraints define the relationship between data points (activities) that belong to the same cluster. CL constraints define the relationship between activities that belong to the different cluster. Both ML and CL constraints assume transitivity expressed as a binary relationship between data points (Bair, 2014)

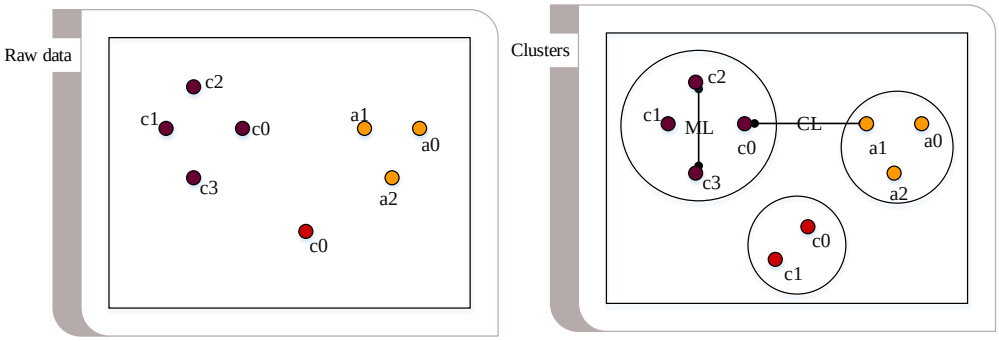

Figure 2: presents the prior knowledge from must-link and cannot-link constraints. ML = [(c0,c1),(c0,c2),(c0,c3).. ,(a0,a1), (a0,a2)..(b0,b1), (b0,b2)…} and CL = [(c0,a0),(c0,a1),(c0,a2).. ,(a0,b0), (a0,b1)..,(b0,c1), (b0,c2)…}

2.3. Clustering selected features using COP-KMEANS

COP-KMEANS is a semi-supervised variant of K-MEANS clustering algorithm. Although K-MEANS was proposed over 50 years ago, it is still one of the most widely used clustering algorithm (Jain, 2010) even with big datasets (Almanza-ortega and Romero, 2018).

The proposed model makes use of the selected set of labelled data points as derived ML and CL constraints, whereas a remaining set is used for the model validation. The difference when compared to the original K-MEANS algorithm, is that the COP-KMEANS algorithm adds a process to check constraints violations. The data points are assigned to the nearest clusters as long as they do not violate CL and ML constraints (Bair 2014).

2.4. Evaluation Methods (Fowlkes Mallows Index)

The evaluation of the model is carried out by comparison of the output with the ground truth or class label for each data point. Determining the density and separation of the clusters, Fowlkes Mallows Index (FMI) was calculated to measure the performance of clustering against the labelled data gathered from the volunteer surveys.

$$FMI = \sqrt{\frac{tp}{tp+fp} + \frac{tp}{tp+fn}} \qquad \text{Equation (1)}$$

FMI, as a clustering Indices, applied to evaluate the results. The number of true positive as tp and the number of false positive as fp and the number of false negatives as fn represented in Equation (1)

2. Results

Figure 3 demonstrates that the accuracy of the model using prior information as constraints improves significantly. As more constraints are added the accuracy tends to get better up to a point when the gains in accuracy plateaus.

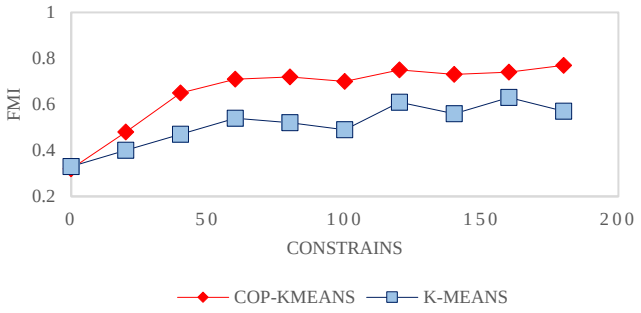

Figure 3: The results of COP-KMEANS versus KMEANS using FMI

Table 1 provides the accuracy of the activities identified in known classifications from the labelled data. 71% of the total activities got mapped to the correct class as known from the ground truth. Within classes' home and work-related activities provide a better percentage of success. In the remaining classes, after work activities were found to identify across multiple clusters.

Table 1: Proportions of trip purposes of inference results in two methods

	The percentage of the activity	
	COP-KMEANS	K-MEANS
Activities	(%)	(%)
Home Related Activities	70%	45%
Before Work Activities	61%	36%
Work-Related Activities	50%	25%
Midday Activities	33%	22%
After Work Activities	79%	69%
Total	71%	54%

4. Conclusions and Future Work

The study establishes that the accuracy of the clusters can be improved using pairwise constraints compared to the traditional clustering algorithm. It also enables the identification of the trip purpose using the minimal number of labelled data points.

The main challenge in the application of this approach is the performance of the model with a large dataset. Therefore, the future work is focusing on the optimization of the model to allow parallel processing of large volumes using the map-reduce framework.

5. Acknowledgements

I am grateful to the Economic and Social Research Council for funding my studentship at UCL.

6. Biography

Nilufer Sari Aslam is currently PhD student at Department of Civil, Environmental and Geomatic Engineering at UCL. Nilufer's research interests are big data analysis, spatial-temporal analysis and machine learning.

7. References

Almanza-ortega, N. N. and Romero, D. (2018) 'Balancing effort and benefit of K -means clustering algorithms in Big Data realms', pp. 1–19.

Bair, E. (2014) 'Semi-supervised clustering methods', 5(5), pp. 349–361. doi: 10.1002/wics.1270.

Devillaine, F., Munizaga, M. and Trépanier, M. (2012) 'Detection of Activities of Public Transport Users by Analyzing Smart Card Data', *Transportation Research Record: Journal of the Transportation Research Board*, 2276(3), pp. 48–55. doi: 10.3141/2276-06.

Faroqi, H., Mesbah, M. and Kim, J. (2018) 'Applications of transit smart cards beyond a fare collection tool : A literature review', *Advances in Transportation Studies*, 45(July), pp. 107–122. doi: 10.4399/978255166098.

Jain, A. K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*. Elsevier B.V., 31(8), pp. 651–666. doi: 10.1016/j.patrec.2009.09.011.

Kong, Q.-J. *et al.* (2009) 'An approach to Urban traffic state estimation by fusing multisource information', *IEEE Transactions on Intelligent Transportation Systems*, 10(3), pp. 499–511. doi: 10.1109/TITS.2009.2026308.

Peikari, M. *et al.* (2018) 'A Cluster-then-label Semi- supervised Learning Approach for Pathology Image Classification', *Scientific Reports*. Springer US, (April), pp. 1–13. doi: 10.1038/s41598-018-24876-0.

Sari Aslam, N. (2015) 'Analysis of Demand Dynamics and Intermodal Connectivity in London Bicycle Sharing System', *UCL*, (September 2015), p. 74. doi: 10.13140/RG.2.2.31912.42248.

Sari Aslam, N. and Cheng, T. (2018) 'Smart Card Data and Human Mobility', in Longley, P., Cheshire, J., and Singleton, A. (eds) *Consumer Data Research*. London,UK: UCL Press, pp. 111–119. Available at: https://www.jstor.org/stable/j.ctvqhsn6.11.

Sari Aslam, N., Cheng, T. and Cheshire, J. (2018) 'Behavioural Analysis of Smart Card Data', in *Proceedings of the 26th GIScience Research UK Conference, GIS Research UK (GISRUK).* Leicester, UK.

Sari Aslam, N., Cheng, T. and Cheshire, J. (2019) 'A high-precision heuristic model to detect home and work locations from smart card data', *Geo-spatial Information Science*. Taylor & Francis, 22(1), pp. 1–11. doi: 10.1080/10095020.2018.1545884.

Sari Aslam, N., Cheshire, J. and Cheng, T. (2015) 'Big Data analysis of population flow between TfL Oyster and bicycle hire networks in London', *Proceedings of the 23rd Conference on GIS Research UK*, pp. 69–75. Available at: https://figshare.com/articles/GIS_Research_UK_GISRUK_2015_Proceedings/1491375.

Wagstaff, K., Rogers, S. and Schroedl, S. (2001) ‘Constrained K-means Clustering with Background Knowledge’, pp. 577–584.

Zhang, Y., Cheng, T. and Sari Aslam, N. (2019) ‘Exploring the relationship between travel pattern and social - demographics using smart card data and household survey’, in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Enschede, The Netherlands, pp. 10–14.