Inferring Demographics from Spatial-Temporal Activities Using Smart Card Data

Yang Zhang[*1], Nilufer Sari Aslam[†1] and Tao Cheng[‡1]

[1]SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental & Geomatic Engineering, University College London

January 17, 2020

Summary

Demographic information (e.g., age and income) has been shown great significance in location-based services, behaviour study and other aspects, but it is usually hard to collect. This paper aims to infer passengers' demographic characteristics from spatio-temporal activities extracted from incomplete smart card data (SCD). We introduce a tensor-based method to represent the travellers' spatial-temporal activities as a 3D space-time cube. Latent activity features can be extracted from SCD via tensor decomposition. Finally, a classification model is presented for demographic prediction. Models are tested and validated using London's Oyster Card data and London Travel Demand Survey Data.

KEYWORDS: Smart card data, tensor decomposition, demographic prediction, classification

1. Introduction

Demographic information, such as age, gender, employment status, and socio-economic status are important in various applications in the real world, such as targeted marketing, personalised advertisement, context-aware recommendation system and user-oriented surveillance (Wang et al., 2016; Zhang and Cheng, 2017). However, due to privacy concerns and data ownership, such demographic information is not easy to obtain from the public and usually incomplete. Fortunately, with the development of information technology, researchers show that user's demographics can be inferred through many online digital records of human behaviours, such as web browsing and social network. The limitation is that the discriminative power of mobility in the physical world has received much less attention. As modern public transport (PT) system plays an increasingly significant role in people's daily life, the equipped automatic ticketing system, called Automated Fare Collection (AFC) systems, can collect massive behaviour data. SCD can reveal the public transportation-related physical activity at an individual level and it is significant to understand people's spatial-temporal travel patterns and explore the population demographics (Zhang and Cheng, 2018).

Estimate the user's demographics from traffic SCD is never a simple task because of several challenges. First, raw SCD are always incomplete. For example, the data recorded in London's Oyster Card did not have the boarding and alighting stations of bus journeys because the price of bus tickets does not depend on the travel length. Second, raw SCD cannot be directly used for demographic prediction. The effective representation of the raw SCD for demographic inference should be investigated (Zhang et al., 2019; Zhang and Cheng, 2020).

To solve abovementioned issues, this paper proposes a framework to infer demographics of passengers from incomplete SCD. In this study, SCD is combined with household survey data to explore the

[*] yang.zhang.16@ucl.ac.uk
[†] n.aslam.11@ucl.ac.uk
[‡] tao.cheng@ucl.ac.uk

discriminative power of SCD for demographic prediction, which can be formulated as a classification problem. We first introduce a flowchart to generate the trip chains of individuals from incomplete SCD. Then, passengers' spatial-temporal activities are identified from SCD and a 3D tensor is introduced to represent the individual activities extracted from the SCD. Based on the representation, individual travel features are extracted from the 3D space-time cube via tensor decomposition technique and a classification model is proposed to classify passengers into different demographic groups. Finally, a case study using London's Oyster Card data and London Travel Demand Survey is presented to test and validate the effectiveness of our model.

2. Data description

2.1. London's Oyster card data

The SCD used in this study is a sample of London's Oyster Card data collected from public transit in London, UK, during the full year of 2013. There are two types of SCD, one from the tube system and the other from the bus system. Each transaction is recorded automatically when a passenger taps in/out at a tube station or boards at a bus stop. In summary, the entire data contain about 2.18 million journeys made by 9708 passengers, consisting of 33.7% tube journeys and 66.3% bus journeys. Each transaction record contains the following fields: (1) unique ID, (2) boarding time, (3) alighting time (tube journey only), (4) boarding station, (5) alighting station (tube station only), (6) journey mode (bus or tube).

2.2. LTDS data

London Travel Demand Survey (LTDS), carried out by Transport for London (TfL), is a continuous household survey of the London area, covering all London boroughs and the City of London. Every year, around 8000 randomly selected households undertake the LTDS. All household members aged 5 and over need to complete the questionnaire. The unique Oyster card ID voluntarily provided by interviewed individuals in households for linking LTDS to Oyster card transaction records, which provide the ground truth to validate our method in this study. The social-demographic data used in this study are shown in Figure 1.

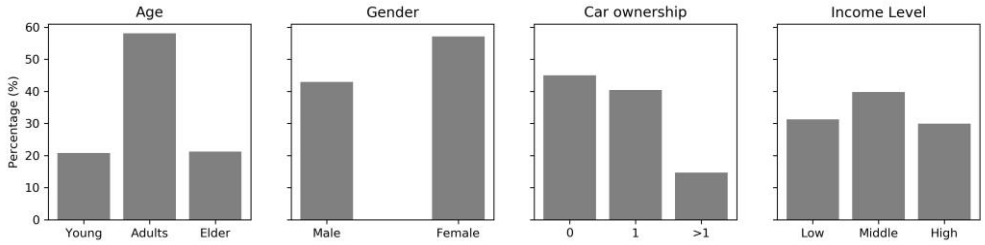

Figure 1 Demographic attributes and corresponding categories

3. Methodology

3.1. Smart Card Data Representation

A central step in this research is the representation of each individual's spatial-temporal activities from SCD for demographic prediction. First, to infer the missing information from the incomplete SCD, we refer to (Zhao et al., 2007). Second, we propose to use a tensor-based method to represent the SCD as a 3D space-time cube and preserve the spatial-temporal activities of individuals. Tensor means the high dimension generalization of vector and matrix. A tensor representation allows us to summarize multivariate categorical data into a multi-dimension array (Sun and Axhausen, 2016). To construct the tensor representation of SCD, we first identify the activity areas of individuals. We refer the algorithm proposed by Goulet Langlois et al. (2016) to define the travelers' activity areas. Then, trip count of individuals can be calculated within spatial areas and a chosen time scale. A pictorial example of a space-time cube of all users is shown in Figure 2 (left). After the three-way tenser constructed, tensor decompositions or factorization is carried out to extract meaningful, latent structure in the data. One extremely tensor decomposition is the Tucker Decomposition. As shown in Figure 2, Tucker decomposes a three-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$ into three factor matrices $\boldsymbol{S} \in \mathbb{R}^{I \times \mathrm{d}S}$, $\boldsymbol{U} \in \mathbb{R}^{J \times \mathrm{d}U}$, and

$\boldsymbol{T} \in \mathbb{R}^{K \times \mathrm{d}T}$, which are also multiplied by a core tensor $\boldsymbol{C} \in \mathbb{R}^{\mathrm{d}S \times \mathrm{d}U \times \mathrm{d}T}$:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{C} \times_1 \boldsymbol{S} \times_2 \boldsymbol{U} \times_3 \boldsymbol{T}$$

Then, each row of the factor matrix $\boldsymbol{U}$ is the factorized feature vector of a user, as the input of the demographic prediction model.

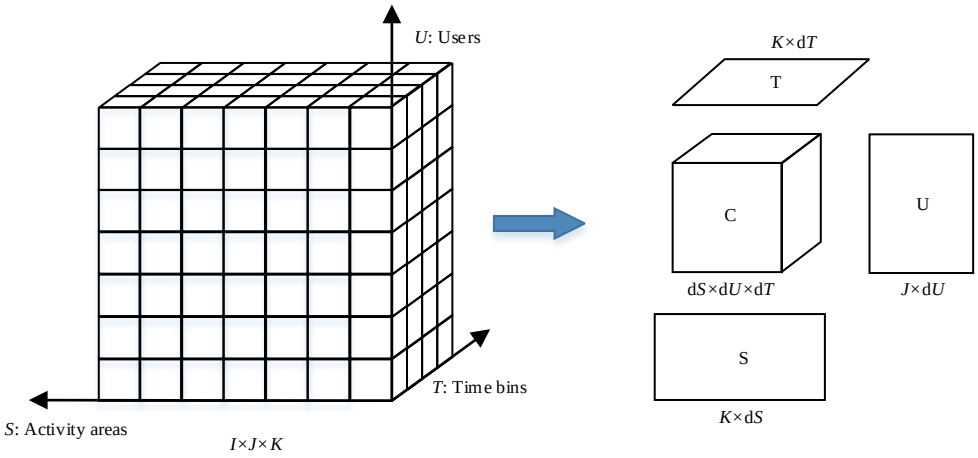

Figure 2 Tucker Decomposition

3.2. Classification Model

In this paper, the basic classifier used as the classification model is support vector machine (SVM), which is a popular machine learning method for classification problem. However, we make no claim about the optimality of the form used, and our aim in this respect is only to explore the capability of the overall approach. The classification method could easily be any other valid standard classifiers, of which several possibilities exist.

4. Case study

The performance of prediction is evaluated by Accuracy (*Acc*), Precision (*Prec*), Recall (*Rec*) and F1 value (*F*1). We conduct a 5-fold cross-validation and calculate the four performance metrics. Results show that the prediction accuracy of 'Age group', 'Gender', 'Income level' and 'Car ownership' can achieve 62.5%, 60.0%, 53.1% and 60.0%, respectively.

5. Conclusion and Discussion

In this paper, we propose a novel framework for demographic prediction using traffic smart card data. The core of task of demographic prediction is a classification or prediction problem which categorises travellers into a demographic group by learning a classifier. Experiment results show the prediction accuracy is quite high. This research can help transport planners to provide better personalised transportation service. In addition, it implies that the travel behaviour of individuals should be protected for privacy concerns. However, there is room for improvement. Currently, we train different classifier for different demographic prediction task, but these tasks are correlated with each other. For example, older people's income is unlikely to be very high. To take advantage of the task correlation, multi-task learning approach can be used to further improve the prediction accuracy.

6. Acknowledgements

The first author's PhD research is jointly funded by China Scholarship Council and the Dean's Prize from the University College London. The data provided by Transport for London (TfL) is highly appreciated.

References

Goulet Langlois, G., et al. 2016. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies,* 64**,** 1-16. http://dx.doi.org/10.1016/j.trc.2015.12.012.

Sun, L. & Axhausen, K. W. 2016. Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation Research Part B: Methodological,* 91**,** 511-524. https://doi.org/10.1016/j.trb.2016.06.011.

Wang, Y., et al. Improving Users' Demographic Prediction via the Videos They Talk about. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP16), 2016.

Zhang, Y. & Cheng, T. Feature Extraction for Long-term Travel Pattern Analysis. Proceedings of the 25th GISRUK conference, 2017 Manchester, UK.

Zhang, Y. & Cheng, T. Inferring Social-Demographics of Travellers based on Smart Card Data. 2nd International Conference on Advanced Research Methods and Analytics, 2018 Valencia, Spain. Editorial Universitat Politècnica de València**,** 55-62. https://doi.org/10.4995/CARMA2018.2018.8310.

Zhang, Y. & Cheng, T. 2020. A Deep Learning Approach to Infer Employment Status of Passengers by Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, 21 (2)**,** 617-629. https://doi.org/10.1109/TITS.2019.2896460.

Zhang, Y., et al. Exploring the Relationship Between Travel Pattern and Social-Demographics Using Smart Card Data and Household Survey. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019 Netherland. Copernicus Publications, XLII-2/W13**,** 1375-1382. https://doi.org/10.5194/isprs-archives-XLII-2-W13-1375-2019.

Zhao, J., et al. 2007. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering,* 22 (5)**,** 376-387.

Biographies

Yang Zhang is a PhD student in the department of Civil, Environmental and Geomatic Engineering at University College London. Her research interest includes spatial-temporal data mining, deep learning and complex network, with applications in crime and transportation.

Tao Cheng is a Professor in GeoInformatics, and Director of SpceTimeLab for Big Data Analytics (http://www.ucl.ac.uk/spacetimelab), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

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.