Deep Learning for Demographic Prediction based on Smart Card Data and Household Survey

Yang Zhang[*1], Tao Cheng[†1] and Nilufer Sari Aslam[‡2]

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

January 30, 2019

Summary

This study devotes to investigating the possibility of inferring demographics of passengers using smart card data (SCD) and household survey. We first represent SCD as a two-dimension image to capture travel patterns. Then, we propose to use a convolutional neural network for automatic feature extraction and demographic prediction, including age group, gender, income level and car ownership. The household survey data is used to train the deep learning model. Finally, a case study using on London's Oyster Card and survey is presented and results show it is a promising opportunity for demographic study based on people's mobility behaviour.

KEYWORDS: Smart card data, travel pattern, deep learning, demographic prediction

1. Introduction

Demographic information is usually treated as private and sensitive data (e.g., age and income), but has been shown great significance in personalized services, behavior study and other aspects. Lately, demographic inference based on an individual's behaviour has become an emerging topic in many research areas. However, most literature focuses on the demographic prediction based on user's activities in the virtual internet world (Hu et al., 2007; Saste et al., 2017), the discriminative power of mobility in the physical world has received much less attention.

Nowadays, public transit (PT) equipped with Automated Fare Collection (AFC) systems generate big smart card data (SCD) and provide an opportunity to reveal large population's mobility pattern (Zhang and Cheng, 2017). Leveraging household survey, it provides an opportunity to learn individual's demographics from SCD. Very recently, several works have begun to explore the possibility of inferring demographics from human mobility patterns. For example, Zhu et al. (2017) extracted features from GPS trajectories to predict individuals' social-demographic information, including age group, gender, and employment status. Similar work can be seen in (Zhang and Cheng, 2018). However, these studies are based on manually extracted features, which cannot capture the full scale of the travel behaviours.

In this paper, we represent the SCD as an image to capture the travel pattern of passengers, and we propose to predict demographics using a convolutional neural network (CNN), a deep learning method, which can automatically extract useful features from images without priori knowledge. We validate our method via London's Oyster Card data and London Travel Demand Survey (LTDS) provided by Transport for London (TfL). The details of data representation, methodologies and case study are given as below.

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

2. Data description

2.1. London's Oyster card data

The Oyster card data used in this paper is a sample of the whole Oyster card transaction records 2012 provided by TfL. Summarily, the entire dataset contains around 2.18 million journeys made by 9708 passengers, made up of 33.7% tube journeys and 66.3% bus journeys. In Oyster card data, each record contains: (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). As the fare of a bus trip does not depend on the travel distance or zone, the alighting time and station of bus trips are never recorded.

2.2. LTDS data

The LTDS is a continuous survey based on the household for collecting individual or household demographics and travel-related information. The unique Oyster card ID voluntarily provided by interviewed individuals in households for linking LTDS to Oyster card transaction records. The LSTD data used in this study are visualised in Figure 1.

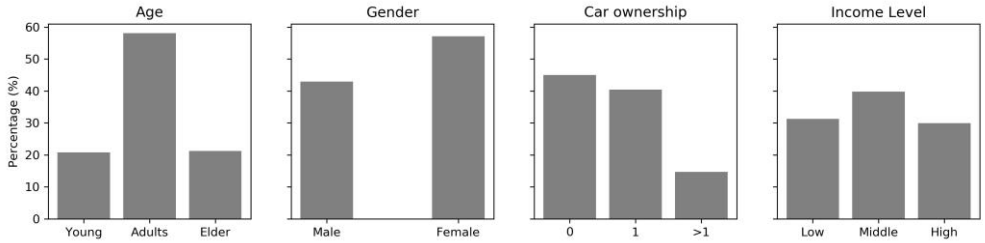

Figure 1 Demographic attributes and corresponding categories

3. Methodology

3.1. Smart Card Data Representation

SCD cannot be directly used for demographic prediction. To represent the spatio-temporal (ST) movement pattern of passengers, we propose to reconstruct the SCD as a two-dimension matrix. First, as the alighting time and location of bus trips are missing in the SCD, we need to first infer this information referring to (Trépanier et al., 2007; Zhao et al., 2007). In the two-dimension matrix, each column along the x-axis shows a day and each row along the y-axis shows the time of day. The value of each grid indicates a user's travel frequency during the corresponding time period of that day. In this way, the one-month SCD of a passenger can be represented as a two-dimensional image (hours×days), capturing his/her travel patterns.

3.2. CNN model for demographic inference

We reformulate the demographic inference task as a supervised classification problem and we propose to use CNN model to infer passenger's demographics based on the spatial-temporal matrix. CNN is a deep learning model, which has achieved great success in the field of image classification (Krizhevsky et al., 2012). A CNN model consists of several convolutional layers, pooling layers and full-connected layers. A typical architecture of CNN is provided in Figure 2. The convolution layer conducts convolution operation by sliding the filter over the input two-dimension matrix. The pooling layer is to subsample the feature map extracted by the convolutional layers to progressively reduce its dimensionality, decreasing the number of parameters and the computation cost. Finally, the full-connected layer generates the final outputs.

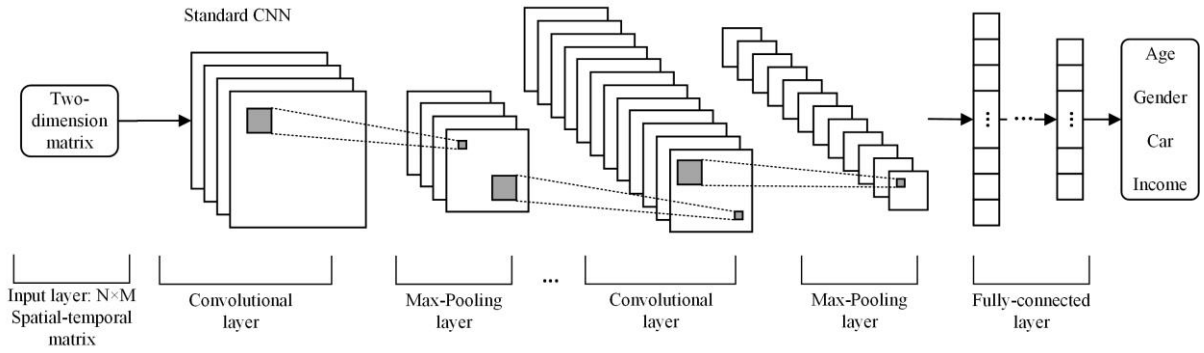

Figure 2 The architecture of CNN

4. Case study

We use London's Oyster card data and the household survey to infer the age, gender, car ownership and income level of passengers at an individual level. The performance of prediction is evaluated by accuracy. We conduct five-fold cross-validation and calculate the prediction accuracy. Results show the best prediction accuracy of 'Age group', 'Gender', 'Car ownership' and 'Income level' can achieve 67.1%, 60.24%, 57.84% and 61.94%, respectively.

5. Conclusion and Discussion

This work explores the possibility of using smart card data to infer demographics of passengers using deep learning method. 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.

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

Hu, J., Zeng, H.-J., Li, H., Niu, C. & Chen, Z. (2007), Demographic Prediction Based on User's Browsing Behavior, *Proceedings of the 16th international conference on World Wide Web*, ACM, pp. 151-160.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems*, pp. 1097-1105.

Saste, A., Bedekar, M. & Kosamkar, P. (2017), Predicting Demographic Attributes from Web Usage: Purpose and Methodologies, *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 381-386.

Trépanier, M., Tranchant, N. & Chapleau, R. (2007), Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System, *Journal of Intelligent Transportation Systems*, **11**(1), 1-14.

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

Zhang, Y. & Cheng, T. (2018), Inferring Social-Demographics of Travellers Based on Smart Card Data, *2nd International Conference on Advanced Reserach Methods and Analytics*, Editorial Universitat Politècnica de València, Valencia, Spain, pp. 55-62.

Zhao, J., Rahbee, A. & Wilson, N. H. (2007), Estimating a Rail Passenger Trip Origin-Destination

Matrix Using Automatic Data Collection Systems, *Computer-Aided Civil and Infrastructure Engineering*, **22**(5), 376-387.
Zhu, L., Gonder, J. & Lin, L. (2017), Prediction of Individual Social-Demographic Role Based on Travel Behavior Variability Using Long-Term Gps Data, *Journal of Advanced Transportation*, **2017**.

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.