Traveler Segmentation using Smart Card Data with Deep Learning on Noisy Labels

CHEN Dacheng
Institute of Data Science
National University of Singapore
Singapore
idscd@nus.edu.sg

YANG Ruizhi
Institute of Data Science
National University of Singapore
Singapore
idsyr@nus.edu.sg

SHI Lei
Institute of Data Science
National University of Singapore
Singapore
shilei@nus.edu.sg

THAM Ying Kiat
Institute of Data Science
National University of Singapore
Singapore
idstyk@nus.edu.sg

David LIM Tian Hui
EZ-Link Pte Ltd
Singapore
davidlim@ezlink.com.sg

Jasper KUAN Hon Whye
EZ-Link Pte Ltd
Singapore
jasperkuan@ezlink.com.sg

NG See Kiong
Institute of Data Science
National University of Singapore
Singapore
seekiong@nus.edu.sg

ABSTRACT

Automated Fare Collection (AFC) system is widely used in today’s urban transportation facilities across the globe. While Smart Card data collected by AFC systems contain rich information for understanding the city travelers’ transportation behaviors, the anonymous nature of such information hinders further applications with the data which will require clustering the anonymous smart card holders into meaningful customer groups, such as for transportation policy planning and implementation of customized services or targeted incentives. As manual labelling of the travellers is impractical, heuristic rule-based or frequentist methods can be used but they will lead to noisy labels. In this paper, we propose to tackle the problem using a three-stage procedure involving two-sample testing, heuristic labelling, and deep learning with label denoising to mine the temporal patterns of travelers to classify them into several meaningful customer segments of interest. We experimentally validated the robustness of our label denoising approach for deep learning, and showed that the segment sizes predicted using our method on a nine-month Smart Card dataset are in agreement with the reported population figures.

CCS CONCEPTS

- Applied Computing → Transportation;
- Pattern Recognition → Clustering;
- Database Application → Data Mining;
- Information Systems → Big Data;
- Information Systems → Database Applications;
- Information Systems → Other

KEYWORDS

Transportation, Automatic Fare Collection System, Smart Card Data, Two Sample Test, Deep Learning, Label Noise, Big Data,

ACM Reference Format:


1 INTRODUCTION

Automated Fare Collection (AFC) is widely adopted in public transportation systems in large cities around the globe: EZ-Link in Singapore, Octopus in Hong Kong SAR, Oyster Card in London, Navigo card in Paris, to name a few. In these systems, a stored value contactless smart card (SC) is used by each traveler to tap-in and/or tap-out to record the transactions. The fare for each trip is automatically calculated and deducted upon each use of the transport system by a traveler, typically based on the distance traveled and the means of transportation used. Such systems provide convenience for travelers while at the same time record large amounts of data regarding each trip by each smart card holder. Generally, the information recorded for each trip include the card holder’s ID, the trip’s timings which can include either trip’s starting time or ending time or both, the trip’s geo-information which can include either the starting station or the ending station or both, and the corresponding fare for the trip. For a typical city with several millions of habitants, the total transactions each day can easily go up to tens of millions.

Although the data size is big, the anonymous nature of the data poses a big challenge for further meaningful analysis of the data. Usually, the travel smart cards can be purchased by the commuters without the need to register their personal information.
Clustering of the travelers into meaningful groups is necessary for targeted marketing, service optimization, policy making etc. In the case of Singapore, the smart card can also be used as a mode of cashless payment in non-transportation spendings such as purchases at vending machines, convenience stores, supermarkets, etc. A meaningful smart card segmentation also helps the businesses to better understand their customers.

As manual labeling of the travelers is impossible (and impractical given the large data size), rule-based or frequentist methods can be used for inference but they tend to lead to noisy labels. In this paper, we proposed a three-stage statistical and deep learning combined algorithm to mine the temporal travel patterns of the travelers and cluster them into several meaningful customer groups.

The paper is organized as follows: in Section 2, we will discuss related work on smart card data; followed by the details on the dataset in section 3; the algorithm will be presented in Section 4 together with the performance and related discussions and we will conclude in Section 5.

2 RELATED WORKS

The Automated Fare Collection (AFC) system [20] came into existence about 20 years ago, around the turn of this century. An early paper [6] discussed its origin, conception and implementation. Research on the data generated by AFC systems or “smart card data” started to appear several years later, but they were often limited by the lack of data records. Some early works include descriptive statistics for customer behaviors [49][28][15], transfer point estimation, origin and destination inference [50][22][41][18], and estimation of origin-destination matrices [41][52][42], etc.

The problem of card anonymity was also raised in the early days of research on smart card data [3]. While some market segments may be directly based on the pre-defined card types when the cards were sold in certain situations [39], [38], the clustering of the travel card owners often had to be inferred from the smart card data. In [2], a hierarchical Clustering plus k-means algorithm was used on a ten-week data set to infer four customer segments. In fact, k-means was a popular method used by many researchers for clustering. In [40], k-means was applied on preprocessed boarding time feature data to find similar clusters. In [44], the authors preprocessed the data into 20 variables related to various characteristics of the travelers, and used k-means to classify them in to 8 clusters. In [19], instead of manually preprocessing the data, authors converted each of the travelers’ temporal sequences into a vector, representing each traveler as a $x \times y \times s$ vector, with each day having $x$ bins (e.g. 24) across $y$ days and having the status $s$ at each time point. All the sampled travellers were then combined into a matrix upon which PCA (Principal Component Analysis) was applied to project the data into a lower dimensional space. The authors then used the first 8 PCs as input variables to a k-means clustering process to generate clusters.

While the k-means algorithm had been popular with researchers for inferring traveler clustering with smart card data, it is worth noting that k-mean method is measuring similarity with some predefined distance function and such definition needs significant domain expertise. Furthermore, as pointed out in [47], methods that based on measuring distances between the samples may face “curse of dimensionality” and the associated sparsity issues and may yield poor results.

Another popular method named DBSCAN [12] was also used by various researchers. [30] used DBSCAN algorithm to analyze the trip patterns of a traveler and constructed four features based on the regularity profile created by the DBSCAN, before applying the k-means algorithm to cluster the travelers. [26] used DBSCAN to study the spatial and temporal regularity of a passenger and used k-mean to separate the regular from the irregular travelers. In [25], the same group of authors upgraded the DBSCAN to Weighted-Stop DBSCAN (DBSCAN) which further enhanced the computational speed. Although DBSCAN is able to handle more complex cluster shapes than k-means, however, as noted in [47], it is still based on distances between the samples in the original vector space. As such, just like k-means, they are faced with the “curse of dimensionality” and the associated sparsity issues.

A visual method for analysis of temporal user behavior was proposed in [16]. The authors first project the high dimensional data into low dimensional space by a mapping named Semi-Circle Projection (SCP), followed by hierarchical clustering to cluster the data. However, hierarchical clustering-based algorithms typically have high time complexity and not suitable for analyzing very large dataset. In smart card data analysis, a city can easily have multimillion data points in just one day. In this work, we propose a deep neural network approach that can better address the issues such as curse of dimensionality, fixed window size etc that these previous methods had faced.

One interesting approach that has appeared during recent years is the use of the unigram model for classification [43]. The unigram model was initially used to classify unlabeled documents. In [31][32], the authors considered each traveler’s temporal travel record as a “document”, each hour in a week as a “word”, and assumed the proportion of each cluster and the probability of each “word” in a cluster as hidden variables which were estimated similarly with EM algorithm.

While the use of unigram model was an interesting approach compared to the traditional distance-based models, the n-gram models lack flexibility as they rely on exact patterns for the words [5], but the “words” from smart card data (i.e. sequence of travel records per time bin) can vary greatly. Moreover, the unigram model only considers the current “word” and does not take the correlation amongst the words into consideration. This means that when the unigram model is applied to the smart card data, each time bin is assumed to be independent, which is not a valid assumption for transportation. For example, in typical daily commuting, if a person goes to work in the morning, we would expect this person to come back around eight hours later. If a person goes to work late, we could expect him or her to come back late with higher probability. Our deep neural networks approach that we are proposing in this work does not make such simplifying assumptions.

3 THE DATA SET

The data set used in the study is jointly provided by Singapore Land Transport Authority (LTA) and the smart card issuing company EZ-Link. The data set contains EZ-Link’s standard adult travel card data. Data from the concession cards (for example, students, senior
citizens) were excluded from this study, as there is no need to infer their customer segmentation. The data set is from 1st August 2016 to 30th April 2017, covering 9 months of travel data from the whole transport system including Mass Rapid Transit (MRT) and buses in Singapore, by travellers using the standard adult EZ-Link travel cards. The total number of transactions for the nine-month period was 1.2 billion by 5.88 millions unique card IDs. For each transaction, the information recorded are as shown in Table 1.

Table 1: Smart Card Transaction Data Fields

<table>
<thead>
<tr>
<th>Data Field Name</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card ID</td>
<td>Unique ID for each traveller (anonymised)</td>
</tr>
<tr>
<td>Travel Mode</td>
<td>Either Bus or MRT (Rail)</td>
</tr>
<tr>
<td>Ride Start Date</td>
<td>The date of the ride starting point</td>
</tr>
<tr>
<td>Ride End Date</td>
<td>The date of the ride ending point</td>
</tr>
<tr>
<td>Ride Start Time</td>
<td>The time of the ride starting point</td>
</tr>
<tr>
<td>Ride End Time</td>
<td>The time of the ride ending point</td>
</tr>
<tr>
<td>Boarding Station</td>
<td>Ride start location</td>
</tr>
<tr>
<td>Alighting Station</td>
<td>Ride end location</td>
</tr>
</tbody>
</table>

Most previous research were typically conducted on a few weeks' to a month’s worth of data which may not contain enough data records to extract meaningful travel patterns, especially for the non-frequent users. The 9-month duration of the data set in our study is also long enough to observe meaningful travel behaviors. Given that EZ-Link is the primary travel card issuer in Singapore, this data set contained fairly complete information of city state’s public transport usage.

4 METHODOLOGY

From the EZ-Link data set, our goal is to cluster the card holders into five meaningful customer segments of interest in this study: Tourists, Non-Working Locals, Office Workers, Non-Office Workers, and Foreign Domestic Workers1, based on their public transit travel patterns.

Viewing the separate groups’ behaviours, the first distinction is working or non-working. If the person is working, generally speaking, there should be a difference on the behaviours for weekdays and weekend. If the person is not working, either Non-Working Locals or Tourists, for them, there is no much difference for there weekday and weekend behaviours. So what we do is first separate the working and non-working, then fine tuning groups within respective groups.

Hence our travel card customer clustering algorithm comprises three stages. First, given that the working week (Monday to Friday) and the weekend (Saturday and Sunday) often involve rather contrasting routines for the city dwellers, our clustering algorithm starts with a two-sample test to separate the whole data set into two big groups based on the temporal records of the travel activities during weekdays and weekends. In the second stage of the segmentation algorithm, we perform a heuristic labelling step. Using simple heuristic rules, we further subdivide the two groups into roughly the subgroups of interest. As the heuristic rules are simple rules designed to capture stereotypical patterns, the labels generated are inherently noisy. In the third stage of our algorithm, we devise a denoising procedure to select a robust subset from the heuristically-labelled data, and then use it to train a deep neural network for inferring the corresponding segment of each smart card holder. We present the details of of our 3-stage clustering algorithm below.

4.1 Stage One: Two-Sample Testing

As the cities’ economic activities are powered by work, we can expect the activities of most city dwellers to revolve around a weekly cycle with contrasting routines between the working week and the weekend. As such, we will first separate the data into two groups—working versus non-working—based on their weekday-weekend travel patterns. Although our data contained both the spatial (boarding and alighting stations) and temporal (boarding and alighting times) information, we used only the temporal information, which include the boarding and alighting times of each ride by travelers2, which we have found to be sufficient to differentiate the working adults from those who do not work (e.g. tourists).

First, we normalize the temporal data points of a smart card holder as fractions of the day. For example, if a person boarded the MRT at 6 AM in the morning, the temporal data point is converted to be $6/24 = 0.25$ day. Figure 1 shows the histogram of the normalized data of a card holder.

For each card holder, we obtained two sets of points from his or her travel records over the nine-month duration: one for weekdays another for weekend. Given that there are multiple travel options in the city (for example, taxicabs or bicycles), for some of the city dwellers, the travel records registered on their smart cards can be quite “sparse”. If we had only a few weeks to a month’s data, we may not have sufficient records for each traveller. In this study, with a nine-month duration, we were able to collect sufficient temporal points for weekdays and weekend for the travellers in general.

---

1Foreigners employed as domestic helpers in Singapore, usually as live-in maids.

2Although our algorithm used both the boarding and alighting times, it can also work on just the boarding (or alighting) times as well, as some smart card datasets may contain only the boarding (or alighting) times.
Figure 2 shows the normalized weekday-weekend distribution of typical working and non-working locals. As can be seen, the non-workers (first row) have no easily discernible weekday-weekend difference, while the workers (second row) travel patterns display a clear weekday-weekend difference.

We perform a two-sample test on the two samples for weekday and weekend for each traveller to test for their difference. The test we used is Kolmogorov-Smirnov test (KS test)[1][46][33]. It is a nonparametric test on the equality of one-dimensional probability distributions. The KS-test statistics is:

$$T_{m,n} = \sup_x |F_m^{(1)}(x) - F_n^{(2)}(x)|$$ (1)

where $F_m^{(1)}(x)$ and $F_n^{(2)}(x)$ are empirical cumulative distribution function of sample one and two. The null hypothesis is rejected at level $\alpha$ if $T_{m,n} > -\frac{1}{\sqrt{2n+m}} \ln \frac{\text{min}(n,m)}{\text{max}(n,m)}$. The test is sensitive to the difference in both the position and the shape of the empirical cumulative distribution functions of the two samples. Since it is a non-parametric statistical test, it is able to check whether two data samples come from the same distribution, but it does not need to make any assumption of the underlying distribution.3

Using the two-sample test, we group those travellers having the same temporal distribution between weekdays and weekends in one group ("Non-Working Group"), and those that have different temporal distribution between weekdays and weekends ("Working Group").

### 4.2 Stage Two: Heuristic Labelling

Next, we subdivide each of the two major groups that we have obtained in Stage One using heuristic rules. For the "Non-Working Group", the cardholder can be either a tourist or a non-working local resident. We can easily differentiate these two subgroups based on their duration of stay in the city, as well as the profiles of the locations that they visit—tourists typically stay in Singapore for a short period of time and their travel destinations are usually "touristy", whereas the non-working locals’ travel activities span beyond a couple of weeks. We devised simple heuristic rules based on the stereotypical observations to distinguish between the two subgroups:

1. **Tourists.** Cards with active duration less than 15 days over two continuous months, and more than half the destinations are tourist sites.

2. **Non-Working Locals.** Those in the Non-Working Group that are not tourists.

For those who are working in the city, the office workers have regular nine-to-five-like working hours and exhibit typically regular weekday travel behaviors. In comparison, the foreign domestic workers mostly stay in their employers’ homes during weekdays, and exhibit regular weekend travel behaviors for their weekend breaks outside. The non-office workers (for example, waiters at restaurants) do not follow normal office hours but have their own work schedules during the week. According to these observations, we devised the following simple heuristic rules for the three subgroups in the Working Group as follows:

- **Office Workers.** Cards with mean weekday first tap-in times between 06:30 AM and 09:45 AM, and the standard deviation smaller than 0.5 hour. Under our temporal normalization, the corresponding mean starting hour shall be in the interval [0.271, 0.406] and standard deviation smaller than 0.0208.
- **Foreign Domestic Workers.** Cards with total weekend trips more than weekday trips, and active weekend days more than active weekday days.
- **Non-Office Workers.** Those that do not satisfy any of the above.

Note that for those cards that have less than ten rides, we classify them as "unknown" due to the sparseness of data, unless they have already been detected as tourists above.

### 4.3 Stage Three: Deep Learning on Data Set with Label Noise

After Stage Two, we have roughly subdivide the cardholders into five segments: **Tourists, Non-Working Locals, Office Workers, Foreign Domestic Workers, and Non-Office Workers.** Due to the simplicity of the heuristic rules, we can expect that there is a certain degree of label noise in the data. In this stage, we perform the final clustering step as a classification task using deep learning on the noisy data.

Deep Neural Networks (DNN) have been widely used in analyzing data sets that are high dimensional and have complex structure and multiple levels of abstraction. It has been successfully applied in various fields including image processing[27][13][48], human voice processing[21][34], DNA and gene analysis[9][29] and even particle accelerator data[8]. We use the DNN here since the temporal behavioral profile for each card holder is high-dimensional and complex.

#### 4.3.1 Temporal Behavioral Profile

For each card holder, we construct a "temporal behavioral profile" as a high-dimensional vector based on his or her travel records consisting of multiple journeys over the period of observation. We convert each trip of the card holder into a vector $J_n = (j_n^{(1)}, ..., j_n^{(192)})$ in which the first 96 dimensions represent each quarter of an hour in a day during weekdays, and the other 96 dimensions for weekends. For example, a person goes to work on Monday, board the MRT at 7:35 AM and alight at 8:10 AM. The vector for this trip is as follows:

$$J_n = (0, ..., 0, 1, 1, 1, 0, ..., 0)$$

Suppose the total number of trips that a card holder made during the period is $N$. We define the temporal behavioral profile vector for this person as follows:

$$X_k = \frac{1}{\sum_{n=1}^{N} j_n^{(1)} j_n^{(2)}} \sum_{n=1}^{N} J_n$$ (2)

Note that each dimension on this vector can be viewed as the probability of this person being on the public transport within this quarter of hour during weekdays or weekend.

It is likely that a commuter may need to transfer between different buses and MRTs for one single trip. To avoid mistaking the
transfer trips in a single journey as separate trips, previous researches [4][24] had tried to "glue" the trips together by setting a certain threshold (e.g., 30 minutes) as the allowed time interval between two trips as transfers. In our temporal behavioral profile, as we sum over the "time intervals" of the trips to compute the probability of the card holder being on the public transport within a particular time interval of a weekday or weekend, any small and random gaps between two transfer trips are likely to be covered by the time intervals of another similar trip on a different day, especially for routine commuting. As such, the values in the person’s temporal behavioral profile vector, namely the probabilities of the person being on the public transport at that point of time, is not likely to be affected by random temporal gaps between transfers of a single journey with transfers.

4.3.2 Label Denoising. It has been shown that deep learning is robust to label noise provided that the number of clean labels is sufficiently large [45]. In [17], the authors have proposed and proved a sufficient condition for noise tolerance under class conditional noise (Theorem 3 of [17]). That condition can be satisfied only when more than half of labels are labeled correctly.

To ensure that we have a reliable subset of labelled training data for deep learning, we devise a filtering procedure to select a suitable subset of the noisy data for training. The procedure is as follows:

1. divide the data set into $K$ subsets that contain similar amount of records.
2. upon each of the subset, train a neural network model with noisy labels to have $K$ models. For extremely large data set, this step can be parallelized.
3. classify each record with all the $K$ previously trained models and keep those records that are predicted to be the same class by all the $K$ models.
4. use the records kept in the previous step to train a new final model.

In this work, we divide our datasets of Working Group into $K = 4$ subsets to train four neural networks to filter noisy labels with ensemble voting. Each neural network has the following architecture: the input dimension is 192, the output dimension is 3 with three hidden layers having 96, 64 and 32 nodes respectively. The training of the neural network is done through Proximal Adagrad algorithm [51].

4.3.3 Re-Classification by Deep Learning. After applying the noisy label filtering step, we train a final neural network model using the filtered subset to re-classify the cardholders in the Working Group into their three corresponding subgroups, namely Office workers, Foreign Domestic Workers, and Non-Office Workers. The architecture of the final neural network model is the same as the ones used for the label denoising steps—192 dimensions for input, 3 dimensions for output, with three hidden layers each having 96, 64 and 32 nodes.

4.4 Results

4.4.1 Customer Clustering. We applied our DNN model that was trained on the travel data with noisy labels to classify the corresponding customer segment for each of 5.88M EZ-Link card

---

*The probabilities of one class being labeled as each of the other classes are different from class to class.*
holders in our dataset. The clustering results are displayed in Figure 3. The corresponding proportions of the five customer clusters of interest are shown for each of the nine months. We can observe that there is a high degree of agreement across all the months in our nine-month study.

Given that we do not have the ground truths for the smart card data, we verify the classification performance indirectly using the population and workforce statistics that are published by the government in various official reports for 2017.

In the 2017 report by the Ministry of Manpower of Singapore[36], it was reported that the resident labour force was 2.2697 millions and in the Ministry’s report on foreign workforce numbers[35], it was stated that the foreign workforce in 2017 was 1.368 million. In [11], it was indicated that non-resident population was 1.650 millions. Thus, we deduce that the number of foreigners who were not working then was around 0.282 million.

For the official number of non-working locals, we can compute using the resident labour force participation rates published in [10] and the population figures in [11]. We show the information (broken down in various age groups) in Table 2 for reference. The overall number of non-working locals in 2017 was thus $2.359 M - 1.927 M = 0.422 M$. The total non-working locals in Singapore then was $0.282 M + 0.422 M = 0.704 M$. The official ratio of non-working locals over working adults was thus $0.704/(2.2697 + 1.368) = 19.35\%$.

### Table 2: Official Local Population [11] and Labour Participation Rate [10] by Age Groups

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Population</th>
<th>Labour Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-24</td>
<td>259,072</td>
<td>62.2%</td>
</tr>
<tr>
<td>25-29</td>
<td>290,198</td>
<td>90.2%</td>
</tr>
<tr>
<td>30-34</td>
<td>279,340</td>
<td>91.8%</td>
</tr>
<tr>
<td>35-39</td>
<td>300,956</td>
<td>89.8%</td>
</tr>
<tr>
<td>40-44</td>
<td>311,484</td>
<td>87.7%</td>
</tr>
<tr>
<td>45-49</td>
<td>305,457</td>
<td>86.8%</td>
</tr>
<tr>
<td>50-54</td>
<td>312,814</td>
<td>82.8%</td>
</tr>
<tr>
<td>55-59</td>
<td>301,678</td>
<td>63.6%</td>
</tr>
<tr>
<td>Overall</td>
<td>2,358,999</td>
<td>1,937,043</td>
</tr>
</tbody>
</table>

As for the foreign domestic workers, it was published [35] to be 0.247 million. Hence, the official ratio of the number of foreign domestic workers over the total number of working adults in Singapore was $0.247/(2.269 + 1.368) = 6.78\%$.

We compare the ratio of our predicted non-working locals over working adults, and the ratio of foreign domestic worker over working adults, for which the official figures (19.35\% and 6.78\% respectively) are available as ground truths. The corresponding ratios for all the 9 months are displayed in Table 3, showing a high degree of agreement with the official ratios. As can be seen from the table, our predicted proportion of Non-Working Locals over Working adults is around 20\% across nine months, while the official proportion was 19.35\%. Our predicted proportion of foreign domestic workers is around 8\%, while the proportion computed from government data is 6.78\%.

The predicted figures are a bit different (in both cases, higher) from the official figures. This is expected as the predicted figures were computed using the EZ-Link population whereas the official figures were based on the entire city state’s population. On the other hand, given that public transportation is a main mode of transportation and almost all in Singapore owns an EZ-Link card for travel payment, the closeness between the predicted ratios and the actual ratios are also expected.

Finally, in Figure 4, we visualize the temporal weekday and weekend distributions of the four segments: Office Workers, Non-Office Workers, Foreign Domestic Workers, and Non-Working Locals\(^5\). As can be seen from the figure: the office workers’ weekdays have two peeks, with one in the morning, and another in the evening, as expected. The non-office workers have a big peak during lunch time, as lunch is the main regular feature amongst this diverse group. Interestingly, we observed that the foreign domestic workers’ tend to go out in the evenings on weekdays besides their weekend off-day activities. One possibility could be that they are picking up their employers’ children after school. The non-working locals generally do not have a clear pattern, as expected.

### Table 3: Comparison of the predicted ratios of Non-Working Locals (NWL) over working adults and Foreign Domestic Workers (FDW) over working adults with the Official Ratios

<table>
<thead>
<tr>
<th>Months</th>
<th>NWL (%)</th>
<th>FDW (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.08</td>
<td>22.25%</td>
<td>7.78%</td>
</tr>
<tr>
<td>16.09</td>
<td>21.60%</td>
<td>8.19%</td>
</tr>
<tr>
<td>16.10</td>
<td>21.03%</td>
<td>8.42%</td>
</tr>
<tr>
<td>16.11</td>
<td>20.17%</td>
<td>8.07%</td>
</tr>
<tr>
<td>16.12</td>
<td>22.30%</td>
<td>8.44%</td>
</tr>
<tr>
<td>17.01</td>
<td>21.73%</td>
<td>8.48%</td>
</tr>
<tr>
<td>17.02</td>
<td>20.42%</td>
<td>7.76%</td>
</tr>
<tr>
<td>17.03</td>
<td>20.78%</td>
<td>7.79%</td>
</tr>
<tr>
<td>17.04</td>
<td>21.77%</td>
<td>8.00%</td>
</tr>
</tbody>
</table>

| Official Ratios | 19.35\% | 6.78\% |

\(\text{\textsuperscript{5}}\)Residents includes both the citizens and the permanent residents.

\(\text{\textsuperscript{6}}\)Non-residents include foreigners who are working in Singapore holding Employment Pass, S Pass, etc, those who are dependents of the residents such as the foreign spouse or the family member of a resident, as well as those who are studying in Singapore holding student passes.

\(\text{\textsuperscript{7}}\)We do not show Tourists here as they do not have meaningful temporal weekday and weekend distributions.
for training[7]. A good survey of various techniques [14]. In this work, we have chosen the latter approach, by employing an ensemble of trained voters to filter out the more reliable labeled data for training the classification model.

Given that our smart card data do not have the ground truths for evaluation, we evaluate our proposed denoising procedure on the popular MNIST data set. To create the wrong labels, we defined an error mapping to mimic class conditional noise. For example, to wrongly label a sample, we map the label of 1 to 7, 7 to 1 or 9, 6 to 4 etc. We applied the error mapping to wrongly label different proportions of the data. In Figure 5, we show the classification performance of the models trained directly on noisy data of different degrees of noise labels, against the performance of the models trained on filtered data using our proposed label denoising procedure.

As we can see from Figure 5, DNN is indeed fairly robust against label noise. Even with 40% of the data corrupted, the DNN can still maintain a reasonably good performance. However, our denoising procedure can further improve the performance of the DNN. When the noisy label proportion is low, the performance difference between models that are trained on filtered data and raw noisy data is

DNN with input 784, output 10, and 3 hidden layers with 96, 64, and 32 nodes respectively.
small. However, when the noisy label proportion grows, the model that are trained on filtered data has better prediction performance.

5 CONCLUSION

Customer clustering is important for understanding the underlying structures of the customer population for implementing targeted marketing and customized services. However, manually labeling the customers into their respective segmentation is often impractical. Simple rules based on basic assumptions of the respective segments can be used, but they are inaccurate and generate noisy labels.

In this work, we have looked into the segmentation of public transit travellers into five meaningful customer segments: Tourists, Non-Working Locals, Office Workers, Non-office Workers and Foreign Domestic Workers based on their travel patterns over a nine month period from August 2016 to April 2017. As we did not have known labels for the five customer segments of interest, we performed the segmentation as a classification task by deep learning on noisy labels generated by heuristic rules. Our segmentation algorithm comprised three stages. First, we employed a two-sample test to broadly partition the traveller population into two major groups, namely the working adults and the non-working adults. Then, we used simple heuristic rules to further subdivide the two broad groups into Tourists and Non-Working Locals (from the non-working group), and Office Workers, Non-office Workers and Foreign Domestic Workers (from the working group). As the labels generated by the heuristic rules were noisy, in the third stage we performed a label denoising step by using ensemble voting to robustly select a subset of reliable labelled data for training the final deep neural network. We employed the DNN to re-classify the 5.88 millions EZ-Link cardholders into their respective customer segments of interest.

The customer segmentation methodology that we have proposed here can be easily extended to classify other customer segments of interest. As our future work, we would like to explore the integration of additional data source (such as non-transit card spending data) for traveller segmentation, and to apply our segmentation by classification on noisy label methodology to other application domains.

REFERENCES
