PREDIETING HOURLY BOARDING DEMAND OF BUS PASSENGERS USING IMBALANCED RECORDS FROM SMART-CARDS: A DEEP LEARNING APPROACH

Maturi Rajeshwari [1], Revathy P [2], Dr Dileep P [3]

[1,2] Assistant Professor, Department of Computer Science and Engineering

[3] Professor, Department of Computer Science and Engineering

[1,2] Narsimha Reddy Engineering College,Kompally,Hyderabad,India.

[3] Malla Reddy College of Engineering and Technology,Kompally,Hyderabad,India.

ABSTRACT

The tap-on smart-card data provides a valuable source to learn passenger's boarding behaviour and predict future travel demand. However, when examining the smart-card records (or instances) by the time of day and by boarding stops, the positive instances (i.e. boarding at a specific bus stop at a specific time) are rare compared to negative instances (not boarding at that bus stop at that time). Imbalanced data has been demonstrated to significantly reduce the accuracy of machine-learning models deployed for predicting hourly boarding numbers from a particular location. This paper addresses this data imbalance issue in the smart-card data before applying it to predict bus boarding demand. We propose the deep generative adversarial nets (Deep-GAN) to generate dummy travelling instances to add to a synthetic training dataset with more balanced travelling and non-travelling instances. The synthetic dataset is then used to train a deep neural network (DNN) for predicting the travelling and non-travelling instances from a particular stop in a given time window. The results show that addressing the data imbalance issue can significantly improve the predictive model's performance and better fit ridership's actual profile. Comparing the performance of the Deep-GAN with other traditional resampling methods shows that the proposed method can produce a synthetic training dataset with a higher similarity and diversity and, thus, a stronger prediction power. The paper highlights the significance and provides practical guidance in improving the data quality and model performance on travel behaviour prediction and individual travel behaviour analysis.

I.INTRODUCTION

The rapid progress of urbanization leads to expansion of population in the urban area, increased demand for travel and associated adverse effects in traffic congestion and air pollution [1]–[3]. Public transport has been widely recognized as a green and sustainable

mode of transportation to relieve such transport problems. As a conventional

public transport mode, buses have always played a dominant role in passenger transportation [4], [5]. However, unreliable travel time, bus-bunching and crowding have led to low level-of services for buses [6]–[8]. This has decreased the bus ridership in many cities, particularly with the advent of ride-hailing services in recent years [9]–[11]. To sustain and increase bus patronage, bus operators must find a way to improve its performance and enhance its image and attraction. Advanced operation and management for bus systems can significantly improve the level-of- service and service reliability, which in turn helps increase the bus ridership [12]–[14]. This requires understanding the spatial and temporal variations in passenger demand and making necessary changes on the supply side [15]–[18]. The smart-card system is initially designed for automatic fare collection. As the system also records the boarding information, for example, who gets on buses, where and when, smart-card data has become a ready-made and valuable data source for spatio-temporal demand analysis [19], public transport planning [20]–[23], and further analysis of emission reduction for the sustainable transport [24], [25]. From the smart-card data, we can easily observe the passenger flow at bus stops and on bus lines, and from which to derive the spatial and temporal characteristics of bus trips [26], [27]. However, extracting useful information from big data automatically still poses a significant challenge. In recent years, machine learning techniques have emerged as an efficient and effective approach to analyzing large smart-card datasets. For instance, Liu et al. [28] captured key features in public transport passenger flow prediction via a decision tree model. Zuo et al. [29] built a three-stage framework with a neural network model to forecast the individual accessibility in bus systems.

In our own recent research [30], we demonstrate that smartcard data combined with machine learning techniques can be a powerful approach for predicting the spatial and temporal patterns of bus boarding. The predictions were found to be highly accurate at an aggregated level, averaged over all travelers. However, our research has also thrown light on the data imbalance issues, when trying to predict travel behavior at the level of individual travelers and fine spatial-temporal details. For instance, the boarding of an individual smart-card holder at a specific stop during a particular time window (e.g. an hour) is a rare event: most of the records would denote negative (non-travelling, or not boarding at this bus stop during this time window) instances, and only a few are positive (travelling, boarding at this stop at this time) instances. Such data imbalance

issues can significantly reduce the efficiency and accuracy of machine learning models deployed for predicting travel behavior at the level of individual travelers and fine spatial-temporal details. This motivates this current study where we propose an over- sampling method, deep generative adversarial nets (Deep-GAN) model (initially developed in the context of image generation) to address the data imbalance issue in predicting disaggregate boarding demand (i.e. individual passengers boarding behavior during each hour of the day). We show that, with the synthesized and more balanced database, the prediction accuracy improves significantly. The performance of the proposed approach, based on the Deep-GAN method, is further benchmarked against other resampling methods (including Synthetic Minority Oversampling Technique and Random Under-Sampling) and is shown to have superior performance.

The rest of the paper is organized as follows. Section II reviews the key resembling methods and their applications in transport studies. Section III describes the specific data imbalance issue in predicting the hourly boarding demand. Section IV uses a Deep-GAN to provide a synthesized, more balanced training data sample and a deep neural network (DNN) to predict the individual smart- card holders' boarding actions (boarding or not boarding) in any hour of a day. Section V applies the proposed method to a real-world case study, and the results are discussed in Section VI. Finally, Section VII summarizes the main findings and contributions of this paper and suggests future investigations.

II.EXISTING SYSTEM

Smart card data has emerged in recent years and provide a comprehensive, and cheap source of information for planning and managing public transport systems. This paper presents a multi-stage machine learning framework to predict passengers' boarding stops using smart card data.

The framework addresses the challenges arising from the imbalanced nature of the data (e.g. many non- travelling data) and the 'many-class' issues (e.g. many possible boarding stops) by decomposing the prediction of hourly ridership into three stages: whether to travel or not in that one-hour time slot, which bus line to use, and at which stop to board. A simple neural network architecture, fully connected networks (FCN), and two deep learning architectures, recurrent neural networks (RNN) and long short- term memory networks (LSTM) are

implemented. The proposed approach is applied to a real-life bus network.

We show that the data imbalance has a profound impact on the accuracy of prediction at individual level. At aggregated level, FCN is able to accurately predict the rideship at individual stops, it is poor at capturing the temporal distribution of ridership. RNN and LSTM are able to measure the temporal distribution but lack the ability to capture the spatial distribution through bus lines.

Disadvantages

The data generated by SMOTE and ADASYN are susceptible to outliers. They may generate some data in the majority data space due to minority outlier instances (usually noisy data), causing blurred classification borderlines and making the learning difficulties of the classification model.

• The under-sampling methods usually have to pay the price of losing parts of the information of the majority of data because they have to remove a part of the data. Although the Easy Ensemble and Balance Cascade tried to solve the problem of lost information, they increased the number of models tens of times, significantly increasing the computational burden.

• Little study has noticed the loss caused by the data imbalance issue in the public transport system.

III.PROPOSED SYSTEM

• The data imbalance issue in the public transport system has received little attention, and this study is the first to focus on this issue and propose a deep learning approach, Deep-GAN, to solve it.

• This study compared the differences in similarity and diversity between the real and synthetic travelling instanced generated from Deep-GAN and other over-sampling methods. It also compared different resampling methods for the improvement of data quality by evaluating the performance of the next travel behaviour prediction model. This is the first validation and evaluation of the performance of different data resampling methods based on real data in the public transport system.

• This paper innovatively modelled individual boarding behaviour, which is uncommon in other travel demand prediction tasks. Compared to the popular aggregated prediction, this individual-based model is able to provide more details on the passengers' behaviour, and the results will benefit the analysis of the similarities and heterogeneities.

Advantages of proposed system

- The system proposes an over-sampling method, deep generative adversarial nets (Deep-GAN) model (initially developed in the context of image generation) to address the data imbalance issue in predicting disaggregate boarding demand (i.e. individual passengers boarding behavior during each hour of the day).
- The system shows that, with the synthesized and more balanced database, the prediction accuracy improves significantly. The performance of the proposed approach, based on the Deep-GAN method, is further benchmarked against other resampling methods (including Synthetic Minority Oversampling Technique and Random Under-Sampling) and is shown to have superior performance.

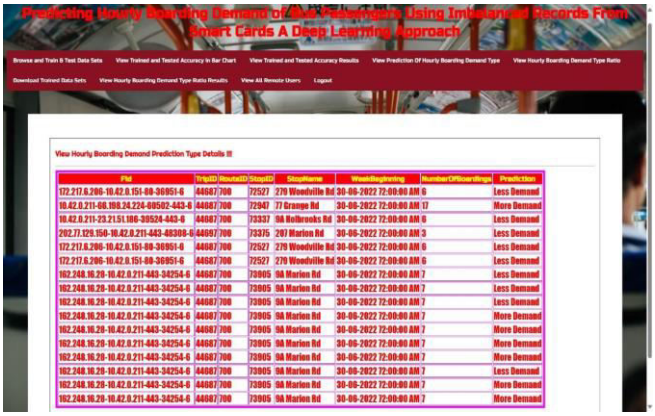
IV.LITERATURE REVIEW

The smart-card system is initially designed for automatic fare collection. As the system also records the boarding information, for example, who gets on buses, where and when, smart-card data has become a ready-made and valuable data source for spatio-temporal demand analysis [19], public transport planning [20]–[23], and further analysis of emission reduction for the sustainable transport [24], [25]. From the smart-card data, we can easily observe the passenger flow at bus stops and on bus lines, and from which to derive the spatial and temporal characteristics of bus trips [26], [27]. However, extracting useful information from big data automatically still poses a significant challenge. In recent years, machine learning techniques have emerged as an efficient and effective approach to analysing large smart-card datasets. For instance, Liu et al. [28] captured key features in public transport passenger flow prediction via a decision tree model. Zuo et al. [29] built a three-stage framework with a neural network model to forecast the individual accessibility in bus systems. In our own recent research [30], we demonstrate that smart card data combined with machine learning techniques can be a powerful approach for predicting the spatial and temporal patterns of bus boarding. The predictions were found to be highly accurate at an aggregated level, averaged over all travellers. However, our research has also thrown light on the data imbalance issues, when trying to predict travel behaviour at the level of individual travellers and fine spatial-temporal details.

V.MODULES

Service Provider

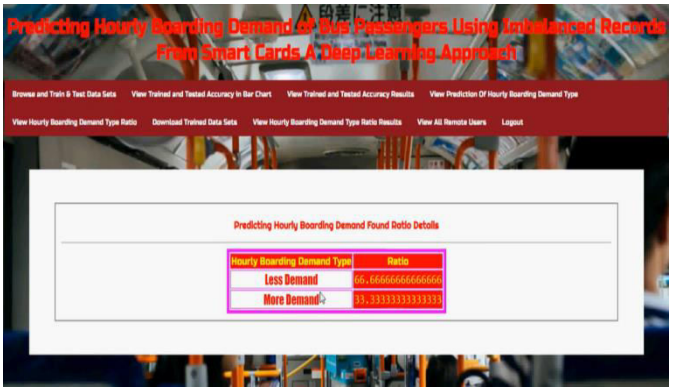
In this module, the Service Provider has to login by using valid user name and password.



After login successful he can do some operations such as Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart,

View Trained and Tested Accuracy Results,

View Prediction Status, View Ratio,



Download Trained Data Sets,

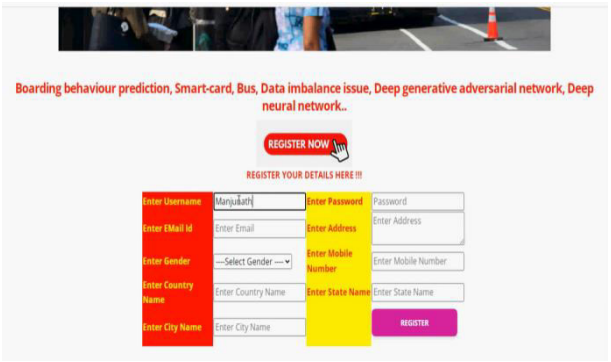
View Status Ratio Results, View All Remote Users.

View and Authorize Users

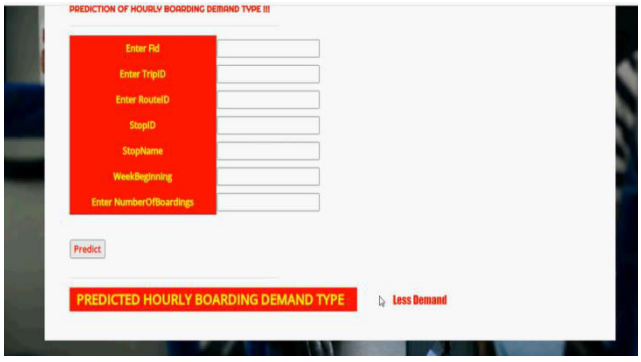
In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations.



Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT TYPE,



VIEW YOUR PROFILE.

VI.CONCLUSION

The motivation of this study was because we have faced the challenge of imbalanced data when we used the real world bus smart-card data to prediction the boarding behavior of passengers at a time window. In this research, we proposed a Deep-GAN to over-sample the travelling instances and to re-balance the rate of travelling and non-travelling instances in the smart-card dataset in order to improve a DNN based prediction model of individual boarding behavior. The performance of Deep-GAN was evaluated by applying the models on real-world smart-card data collected from seven bus lines in the city of Changsha, China. Comparing the different imbalance ratios in the training dataset, we found out that in general, the performance of the model improves with more imbalanced data and the most significant improvement comes at a 1:5 ratio between positive and negative instances. From the perspective of prediction accuracy of the hourly distribution of bus ridership, the high rate of imbalance will cause misleading load profiles and the absolutely balanced data may over predict the ridership during peak hours. Comparison of different resembling methods reveals that both over-sampling and under-sampling benefits the performance of the model. Deep- GAN has the best recall score and its precision scores best among the over-sampling methods. Although the performance of the predictive model trained by the Deep-GAN-data is not significantly beyond other resembling methods, the Deep- GAN also presented a powerful ability to improve the quality of training dataset and the performance of predictive models, especially when the under-sampling is not suitable for the data.

VII.REFERENCES

[1] X. Guo, J. Wu, H. Sun, R. Liu, and Z. Gao, "Timetable coordination of first trains in urban railway network: A case study of beijing," Applied Mathematical Modelling, vol. 40, no. 17, pp. 8048–8066, 2016.

[2] W. Wu, P. Li, R. Liu, W. Jin, B. Yao, Y. Xie, and C. Ma, "Predicting peak load of bus routes with supply optimization and scaled shepard interpolation: A newsvendor model," Transportation Research Part E: Logistics and Transportation Review, vol. 142, p. 102041, 2020.

[3] N. Be˘sinovi´c, L. De Donato, F. Flammini, R. M. Goverde, Z. Lin, R. Liu,S. Marrone, R. Nardone, T. Tang, and V. Vittorini, "Artificial intelligence in railway transport: Taxonomy, regulations and applications," IEEE Transactions on Intelligent Transportation Systems, 2021.

[4] S. C. Kwan and J. H. Hashim, "A review on co-benefits of mass public transportation in climate change mitigation," Sustainable Cities and Society, vol. 22, pp. 11–18, 2016.

[5] Y. Wang, W. Zhang, T. Tang, D. Wang, and Z. Liu, "Busod matrix reconstruction based on clustering wi-fi probe data," Transportmetrica B: Transport Dynamics, pp. 1–16, 2021, doi: 10.1080/21680566.2021.1956388.

[6] S. J. Berrebi, K. E. Watkins, and J. A. Laval, "A real-time bus dispatching policy to minimize passenger wait on a high frequency route," Transportation Research Part B: Methodological, vol. 81, pp. 377–389, 2015.

[7] A. Fonzone, J.-D. Schm¨ocker, and R. Liu, "A model of bus bunching under reliability-based passenger arrival patterns," Transportation Research Part C: Emerging Technologies, vol. 59, pp. 164–182, 2015.

[8] J. D. Schm¨ocker, W. Sun, A. Fonzone, and R. Liu, "Bus bunching along a corridor served by two lines," Transportation Research Part B: Methodological, vol. 93, pp. 300– 317, 2016.

[9] D. Chen, Q. Shao, Z. Liu, W. Yu, and C. L. P. Chen, "Ridesourcing behavior analysis and prediction: A network perspective," IEEE Transactions on Intelligent Transportation

Systems, pp. 1–10, 2020.