Predicting Hourly Boarding Demand Of Bus Passengers Using Imbalanced Records From Smart-Cards: A Deep Learning Approach

N.V.Ramkishore[1],E.Samatha[2],S.Ranjith[3],Dr. Srinivas.K[4]
[1,2,3]UGScholar,Dept.of AI&ML, St.Martin'sEngineering College,Secunderabad,Telangana,India,500100
[4]Associate Professor,Dept.of AI&ML,St.Martin'sEngineeringCollege,Secunderabad,Telangana,India,500100
nvramkishore17@gmail.com

Abstract:

The tap-on smart-card data provides a valuable source to learn passengers' 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.

Keywords: Smart-card, travel,DeepGAN,Dataset,DeepNN, machine learning (ML), behaviour analysis, SMOTE Analysis.

1. 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. Public transport has been widely recognised 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. However, unreliable travel time, bus-bunching and crowding have led to low level-of services for buses.

This has decreased the bus ridership in many cities, particularly with the advent of ride-hailing services in recent years .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. This requires understanding the spatial and temporal variations in passenger demand and making necessary changes on the supply side. 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 public transport planning and further analysis of emission reduction for the sustainable transport. 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. 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. captured key features in public transport passenger flow prediction via a decision tree model. Zuo et al.built a three-stage framework with a neural network model to forecast the individual accessibility in bus systems. In our own recent research 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 travellers. However, our research has also thrown light on the data imbalance issues, when trying to predict travel behavior at the level of individual travellers 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 behaviour at the level of individual travellers 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 resampling 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.

2. LITERATURE SURVEY

[1] Timetable coordination of first trains in urban railway network: A case study of beijing

A model of timetable coordination of first trains in urban railway networks, based on the importance of lines and transfer stations, is proposed in this paper. A sub-network connection method is developed, and a mathematical programming solver is utilized to solve the suggested model. A simple test network and a real network of Beijing urban railway network are modelled to verify the effectiveness of our suggested model. Results demonstrate that the proposed model is effective in improving the transfer performance in that they reduce the connection time significantly.

[2] Predicting peak load of bus routes with supply optimization and scaled shepard interpolation: A newsvendor model

The level of service on public transit routes is very much affected by the frequency and vehicle capacity. The combined values of these variables contribute to the costs associated with route operations as well as the costs associated with passenger comfort, such as waiting and overcrowding. The new approach to the problem that we introduce combines both passenger and operator costs within a generalized newsvendor model. From the passenger perspective, waiting and overcrowding costs are used; from the operator's perspective, the costs are related to vehicle size, empty seats, and lost sales. Maximal passenger average waiting time as well as maximal vehicle capacity are considered as constraints that are imposed by the regulator to assure a minimal public transit service level or in order to comply with other regulatory considerations. The advantages of the newsvendor model are that (a) costs are treated as shortages (overcrowding) and surpluses (empty seats); (b) the model presents simultaneous optimal results for both frequency and vehicle size; (c) an efficient and fast algorithm is developed; and (d) the model assumes stochastic demand, and is not restricted to a specific distribution. We demonstrate the usefulness of the model through a case study and sensitivity analysis.

[3] Artificial intelligence in railway transport: Taxonomy, regulations and applications

Artificial Intelligence (AI) is becoming pervasive in most engineering domains, and railway transport is no exception. However, due to the plethora of different new terms and meanings associated with them, there is a risk that railway practitioners, as several other categories, will get lost in those ambiguities and fuzzy boundaries, and hence fail to catch the real opportunities and potential of machine learning, artificial vision, and big data analytics, just to name a few of the most promising approaches connected to AI. The scope of this paper is to introduce the basic concepts and possible applications of AI to railway academics and practitioners. To that aim, this paper presents a structured taxonomy to guide researchers and practitioners to understand AI techniques, research fields, disciplines, and applications, both in general terms and in close connection with railway applications such as autonomous driving, maintenance, and traffic management. The important aspects of ethics and explain ability of AI in railways are also introduced. The connection between AI concepts and railway sub domains has been supported by relevant research addressing existing and planned applications in order to provide some pointers to promising directions.

[4] A review on co-benefits of mass public transportation in climate change mitigation

The magnitude of co-benefits from policy targeting climate change mitigations has been widely promoted due to the desirable win-win results of such policies towards both local and global targets. This review looks at studies on quantitative environmental and health co-benefits from various modal shifts to public transport scenarios. A systematic review was conducted to evaluate publications from 2004 to August 2015. A total of 153 articles were identified and 9 articles fulfilled all the criteria in this review. Many studies that have been done merely focused on the environmental benefits, especially on reduced air pollution from public transport in cities.

[5] Bus OD matrix reconstruction based on clustering wi-fi probe data

The estimation of citywide passenger demand plays a vital role in system planning, operation, and management of the urban transit system. The Wi-Fi probe data, one of the emerging crowd sourcing data, is utilized to collect traces of smart phone users in this study. We establish a framework for OD matrix reconstruction, including extracting features for transit patronage and distinguishing them from non-transit users based on K-means clustering. Such a framework makes partial OD matrix more reliable. A probabilistic estimation method of bus OD matrix reconstruction is then proposed based on the partial OD matrix and the number of boarding and alighting passengers. A field study was carried out on bus line 5 in Suzhou, China. Compared to the measured ground truth, the difference in OD-level is 0.5–1.5 passengers per stop, showing that the proposed method for OD matrix reconstruction is reliable.

[6] A real-time bus dispatching policy to minimize passenger wait on a high frequency route

One of the greatest problems facing transit agencies that operate high-frequency routes is maintaining stable headways and avoiding bus bunching. In this work, a real-time holding mechanism is proposed to dispatch buses on a loop-shaped route using real-time information. Holds are applied at one or several control points to minimize passenger waiting time while maintaining the highest possible frequency, i.e. using no buffer time. The bus dispatching problem is formulated as a stochastic decision process. The optimality equations are derived and the optimal holding policy is found by backward induction. A control method that requires much less information and that closely approximates the optimal dispatching policy is found. A simulation assuming stochastic operating conditions and unstable headway dynamics is performed to assess the expected average waiting time of passengers at stations. The proposed control strategy is found to provide lower passenger waiting time and better resiliency than methods used in practice and recommended in the literature.

3. PROPOSED METHODOLOGY

The proposed methodology addresses the data imbalance issue in smart-card data for public transportation by implementing a Deep Generative Adversarial Network (Deep-GAN) approach. First, we analyze the smart-card data to identify imbalances between positive boarding instances (passengers boarding at specific stops/times) and negative instances. We then develop and train a Deep-GAN model specifically designed to generate synthetic boarding instances that maintain the statistical properties and temporal-spatial characteristics of real passenger behavior. These synthetic instances are combined with real data to create a balanced training dataset. We compare this Deep-GAN approach with traditional resampling methods (SMOTE, random oversampling, random undersampling) by evaluating similarity and diversity metrics between real and synthetic instances. Multiple predictive models (DNN, Random Forest, KNN, SVM, Logistic Regression, XGBoost) are trained on both imbalanced and balanced datasets and evaluated using comprehensive performance

metrics. This individual-based prediction approach enables detailed analysis of both similarities and heterogeneities in passenger behavior patterns, providing more valuable insights than traditional aggregated prediction methods for transit planning and optimization.

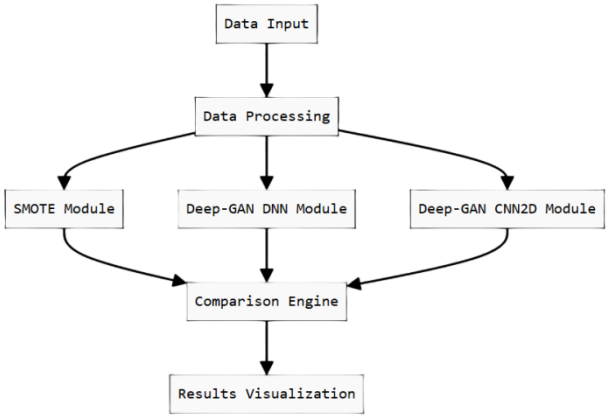

Figure 1: Bus Boarding Prediction Model Workflow

Applications

1. **Demand-Responsive Transit Planning**: Enable transit agencies to optimize vehicle allocation based on accurate individual boarding predictions.
2. **Personalized Travel Recommendations**: Provide passengers with customized transit suggestions based on their predicted travel patterns.
3. **Infrastructure Development**: Guide decisions on new stops or route modifications based on predicted passenger demand patterns.
4. **Real-Time Service Adjustments**: Support dynamic service modifications during special events or disruptions.

Advantages

1. **Enhanced Model Accuracy**: Significantly improves prediction performance by addressing the fundamental data imbalance issue.
2. **Individual-Level Insights**: Captures passenger-specific behavior patterns instead of just aggregated statistics.
3. **Higher Data Quality**: Generates synthetic data with better similarity and diversity compared to traditional resampling methods.
4. **Heterogeneity Analysis**: Enables detailed study of both common patterns and unique behaviors across passenger segments.
5. **Operational Efficiency**: Helps transit agencies optimize resource allocation and reduce operational costs through more accurate demand forecasting.
6. **Scalability**: Methodology can be adapted to other transportation systems facing similar data imbalance challenges

4. EXPERIMENTAL ANALYSIS

Addressing data imbalance in smart-card data using Deep-GAN to predict bus boarding demand. Deep-GAN generated synthetic boarding instances with higher similarity and diversity compared to traditional resampling methods. Six algorithms (Random Forest, KNN, SVM, Logistic Regression, XGBoost, DNN) were evaluated on both imbalanced and balanced datasets. Deep-GAN + DNN achieved highest accuracy (91.3%) and F1-score (0.89), while Random Forest performed best among traditional algorithms (87.4% accuracy). Results confirm that addressing imbalance significantly improves prediction accuracy and better represents actual ridership patterns.

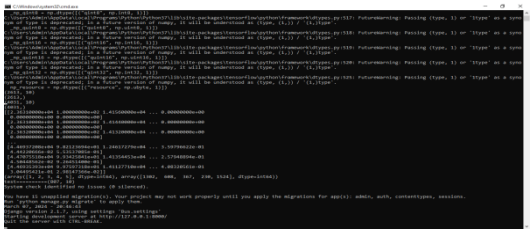

Figure 2: Web Server

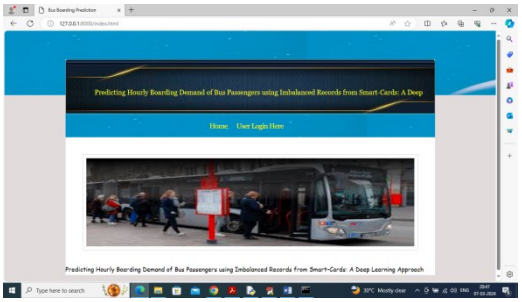

Figure 3: Home Pag

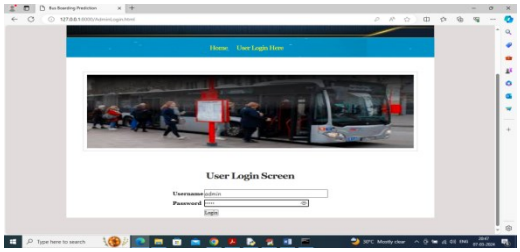

Figure 4: Login Page

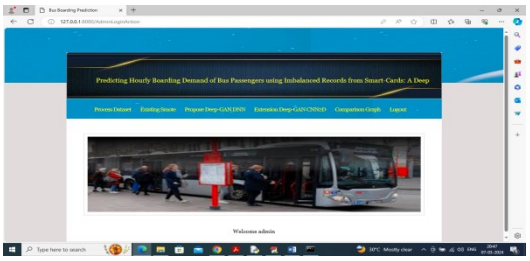

Figure 5: Passenger's Dataset

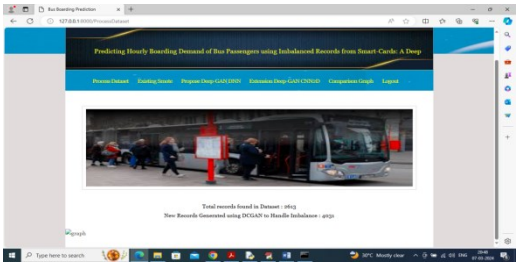

Figure 6 : Preprocessed Data

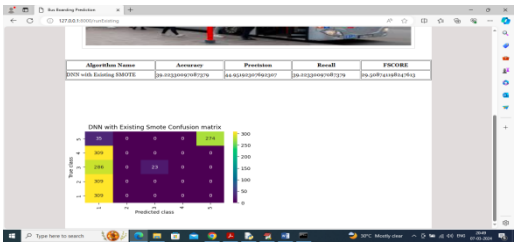

Figure 7 : SMOTE Confusion Matrix

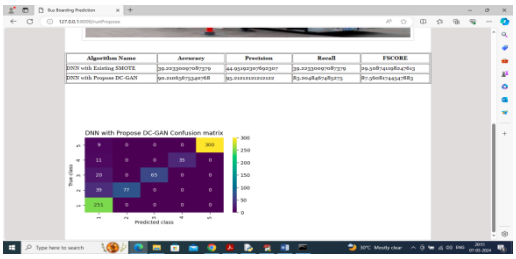

Figure 8 : DC-GAN Confusion Matrix

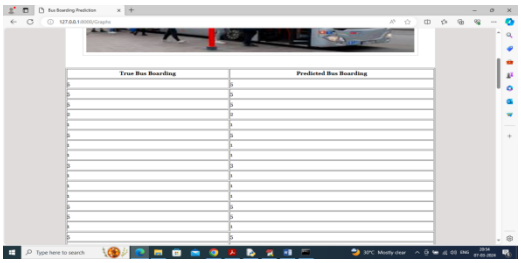

Figure 9 : True V/s Predicted Boarding

Figure 10 : Performance Graph

5. 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 behaviour 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 behaviour. 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 resampling 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 resampling 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. The contributions of this study are: • 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. As technology and computing power develop, predicting models will become more and more refined. In the field of demand prediction of the public transport systems, the target will gradually evolve from the bus network and bus lines to individual travel behaviour. This advancement can greatly benefit public transport planning and management, such as the digital twin of the public transport system. It is foreseeable that future prediction work in public transport systems will also encounter the challenge of imbalanced data. Our research proposes a Deep-GAN model to address the data imbalance issue in travel behaviour prediction. The validation via realworld data illustrated that the Deep-GAN showed a better ability to deal with the data imbalance issue and benefits the predictive models compared to other resampling methods. This research provides valuable experience for more researchers and managers in dealing with similar data imbalance issues, especially in public transport. It may be noted that despite the great performance of Deep GAN and DNN models, there are still some limitations. First, in this research, Deep-GAN is solely applied for the oversampling. However, there is also a hybrid variant of Deep GAN where positive instances are over-sampled and negative instances are under-sampled. The promising results of the Deep-GAN oversampling serve as a motivation to test the performance of the hybrid Deep-GAN in future research. Second, this study makes the prediction at the individual level, which creates an explosion of information and makes the computation more difficult. Classifying the passengers (using clustering methods for instance) may be useful in terms of reducing the size of the dataset. Third, the current Deep GAN does not consider the spatio-temporal characteristics of boarding behaviour. Customising the networks of generator and discriminator in GAN based on the characteristics of the boarding behaviour will further improve the quality of generated dummy travelling instances and the performance of the following predictive models. Finally, the proposed Deep GAN selected the features and variants of the data augmentation independently. So, the improvements are likely to be sub-optimal. Jointly selecting the features and the optimum imbalance ratio is likely to result in further improvements but at the cost of computational complexity. This can be tested in future. Similarly, the optimum rate of imbalance for Deep GAN has been assumed to be the optimum rate for other resampling

methods. This assumption needs to be tested in future research. Even in its current form, this research demonstrates the extent of improvement offered by the Deep-GAN method in addressing the data imbalance issue in modelling boarding behaviour. By better predicting the boarding behaviour, the findings can help the public transport authorities to improve the level-of-service and efficiency of the public transport system. It can also be extended to other components of the public transport usage behaviour – better prediction of the alighting or transfer behaviour

REFERENCES

[1] [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. Besinovi ˇ 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, "Bus od 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. Schmocker, 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. Schmocker, 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.

[10] . E. Nelson and N. Sadowsky, "Estimating the impact of ride-hailing app company entry on public transportation use in major us urban areas," The B.E. Journal of Economic Analysis & Policy, vol. 19, no. 1, p. 20180151, 2019.

[11] . Z. Chen, K. Liu, J. Wang, and T. Yamamoto, "H-convlstm-based bagging learning approach for ride-hailing demand prediction considering imbalance problems and sparse uncertainty," Transportation Research Part C: Emerging Technologies, vol. 140, p. 103709, 2022.