Prediction of special OD passenger flow based on transfer learning

Yu Wang, Zhengpu Zhang, Xiaodong Wang, Cheng Tao

Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing, China

Luoyu167@163.com

Abstract. The prediction of railway passenger flow has always been an important topic in the field of railway operation. Due to the limitation of the number of allocated tickets, some special OD (Origin to Destination) data are sometimes inaccurate. It will obviously affect the final prediction accuracy if it is directly used as a sample. However, it is very difficult to restore the biased data to the true value. In this paper, we choose case-based transfer learning to remove the unqualified samples from the target domain and replace them with the corresponding samples in the source domain and form the final training sample data set. At the same time, we choose the improved AdaBoost algorithm, the sample weight is adjusted by the error size, and the model is retrained to get the final model. The results show that the prediction accuracy of transfer learning based on improved AdaBoost algorithm is better than that of traditional ensemble learning, ARIMA model and multiple regression model. It also provides a new theoretical view for special OD passenger flow prediction in the future.

1. Introduction

Railway passenger flow prediction can be divided into short-term prediction and long-term prediction according to the prediction time. In recent years, the railway transportation enterprises have reformed the management mechanism with the goal of market-oriented, requiring the railway transportation enterprises to take quick response measures to the market demand, so as to increase the final operating income. In order to respond to the changes of transport market and changes of passenger flow, it is necessary to predict the short-term passenger flow scientifically and accurately. Railway passenger flow is affected by many factors in the short term, such as weather, regional environment, macro-economy, emergencies, etc. the prediction of railway passenger flow is a complex nonlinear problem, which is very difficult.

At present, there are many methods to predict the short-term passenger flow. The most common method is parameter method, including ARIMA and SARIMA. This kind of model mainly describes the short-term periodic and seasonal trend of passenger flow, but the disadvantage is that it is only effective for the stationary time series, while most of the data in reality are non-stationary. There is also a class of nonparametric methods, which have rising up strongly in recent years. It mainly includes Bayes model, support vector machine (SVM), deep learning and so on, especially deep learning has achieved good prediction effect. However, for some special OD (origin to destination), due to the special characteristics of railway transportation, ticket allocation among different OD takes up each other, which leads to sample missing or sample incorrect. Moreover, this kind of missing and distortion is random, they can not be revised by algorithms, and it is difficult to complete the missing data or estimate the real value. How to effectively solve the problem of sample missing and distortion,
the method of transfer learning has been paid attention to. For example, if some special OD data is missing or part of the data is distorted, the passenger flow data of similar OD can be used as supplement and replacement to effectively solve the problem of insufficient and inaccurate data of this special OD.

Transfer learning is a new machine learning method that uses existing knowledge to solve problems in different but related fields. It maps the existing knowledge (source domain) to the target domain by transfer, and solves the learning problem of only a few labeled sample data or even no sample data in the target domain.

Transfer learning can be divided into four types: one is based on relationship which is to apply the source domain learning logic relation network to the target domain. The second is transfer by cases(samples). The idea is to transfer the source domain cases after weighting. The third is transfer by feature. The idea is to transform the source domain and target domain into a feature space to complete the transfer. The fourth is model-based transfer learning. The idea is to apply the source domain model to the target domain. According to the characteristics of this paper, we choose transfer learning by cases(samples). Due to the mutual occupation of ticket allocation between different OD, after selecting the target domain and the source domain, we need to remove the obviously unqualified samples from the target domain, select the appropriate samples from the source domain, and transfer these samples to the target domain. They form the overall samples together with the target domain samples. Each sample is appropriate Finally, we use them to train a new model.

Setting \( OD_j(r_j, s_j) \), \( r_j \) is the starting station and \( s_j \) is ending station. \( OD_1(r_1, s_1) \) is shared ticket allocation with \( OD_2(r_2, s_2) \). If the date is \( t \), \( OD_1(r_1, s_1) \) has sold a lot of tickets, which will result in less tickets allocated for \( OD_2(r_2, s_2) \). Even if \( OD_2(r_2, s_2) \) has sold out all tickets, the number of passenger volume on date \( t \) of \( OD_2(r_2, s_2) \) is \( f_2(r_2, s_2) \) which is also under the restrictions. So \( f_2(r_2, s_2) \) is not the real passenger volume. If \( f_2(r_2, s_2) \) is used as samples for prediction, the prediction accuracy will be affected. We can use \( f_1(r_1, s_1) \) as a supplement to \( f_2(r_2, s_2) \) because \( OD_1(r_1, s_1) \) is similar to \( OD_2(r_2, s_2) \). The target domain sample set is composed of \( f_2(r_2, s_2) \), the target domain OD is \( OD_2(r_2, s_2) \), and the source domain sample set is composed of \( f_1(r_1, s_1) \), the source domain OD is \( OD_1(r_1, s_1) \). By selecting the appropriate samples in the source domain and removing the inappropriate samples in the target domain, the final training sample set is composed of samples which are from source domain and the target domain.

2. Algorithm

2.1. Source domain OD selection

As mentioned above, we choose other OD which is similar to the the target domain OD as the source domain OD. We use the distance between two ODs to select the source domain OD. The closer the OD is from the target domain OD, the more suitable this OD is for the source domain OD. If the target domain OD is \( OD_1(r_1, s_1) \), set any OD is \( OD_2(r_2, s_2) \), the distance between \( OD_1(r_1, s_1) \) and \( OD_2(r_2, s_2) \) is shown in equation (1).

\[
d(O_{D_1}(r_1, s_1), O_{D_2}(r_2, s_2)) = \sqrt{\frac{1}{7} \sum_{i=1}^{7} (x_i^1 - x_i^2)^2}
\]  

(1)

In equation (1), \( O_{D_j}(r_j, s_j) \) represents the OD j, \( r_j \) represents the departure station, \( s_j \) represents the arrival station. \( x_i^j \) is the feature i of \( O_{D_j}(r_j, s_j) \). There are 7 features in total, all 7 features have been standardized. The specific meaning is shown in figure 1.
Fig. 1 Feature of OD

The OD with the shortest distance from the target OD is selected as source domain OD by equation (1). This OD is the source domain OD. The appropriate samples are selected from this OD to form the final training sample set.

2.2. Transfer learning based on improved AdaBoost algorithm

At present, the transfer learning based on samples is mainly the classic AdaBoost algorithm. For samples with large error belong to source domain, it does not increase the sample weight, but reduces the weight to avoid negative transfer. If the samples with large error belong to the target domain, we should increase the weight of these sample to improve the prediction accuracy of the model. Based on the above ideas, the improved AdaBoost algorithm is in detail below.

Input: determine the source domain OD and target domain OD, set rules to remove inappropriate samples from the target domain OD, and replace them with source domain OD samples. Finally, the training sample set S is composed of N + M samples. \(\{x_1, y_1\}, \{x_2, y_2\}, \{x_3, y_3\}, ..., \{x_M, y_M\}\) are the samples selected from target domain, and \(\{x_{M+1}, y_{M+1}\}, \{x_{M+2}, y_{M+2}\}, ..., \{x_{M+N}, y_{M+N}\}\) are the samples selected from source domain. All \(N + M\) samples form the training sample set S, and the CART tree is \(RF_1(x)\) and \(l\) represent the-\(l\) iteration, \(L\) is the maximum iteration times.

Output: \(f_{l}(x)\)

\[
f_{l}(x) = \frac{\sum_{i=1}^{L} \alpha_{l} \times RF_{i}(x)}{\sum_{i=1}^{L} \alpha_{l}} \tag{2}
\]

Step1: Initialize sample weight, \(w' = \{w'_1, w'_2, ..., w'_{N+M}\}\), and \(w'_1 = w'_2 = ... = w'_{N+M} = \frac{1}{M+N}\).

Step2: for \(l = 1; l <= L\) do.

Step2.1: According to the weight vector \(w'\) select training samples.

Step2.2: Training model and get model \(RF_{l}(x)\), \(RF_{l}(x)\) is a regression tree.

Step2.3: Use \(RF_{l}(x)\) and training sample set S to calculate error \(\varepsilon_{l}\)

\[
\varepsilon_{l} = \sum_{i=1}^{N+M} w'_i \times error'_i \tag{3}
\]

\[
error'_i = \frac{|y_i - RF_{l}(x_i)|}{\max(|y_i - RF_{l}(x_i)|)} \tag{4}
\]

Step2.4: if \(\varepsilon_{l} > \beta\), return step2.2.

Step2.5: calculate \(\alpha_{l}\)
\[ \alpha_i = \frac{\varepsilon_i}{1 - \varepsilon_i} \quad (5) \]

Step 2.6: Update sample weight, get \( w_i^{(t+1)} \) and get \( Z_i \), then return to step 2, go on iteration.

\[ w_i^{(t+1)} = \frac{w_i^t \times \alpha_i^{error^i}}{Z_i} \quad 1 < i < M \quad (6) \]
\[ w_i^{(t+1)} = \frac{w_i^t \times \alpha_i^{error^i}}{Z_i} \quad M + 1 < i < M + N \quad (7) \]
\[ Z_i = \sum_{i=1}^{M+N} w_i^t \times \alpha_i^{error^i} \quad (8) \]

Step 3: end

Finally, we get ensemble learning machine \( f_\epsilon(x) \).

3. Case Application

3.1. Data selection

This paper is to predict the daily railway passenger volume from Xinxiang East railway station to Zhengzhou East railway station, so the target domain OD is Xinxiang East railway station to Zhengzhou East railway station. Then selected sample date is from January 1, 2019 to June 1, 2019, and the target domain samples is the daily passenger traffic volume from Xinxiang East railway station to Zhengzhou East railway station, with a total of 151 samples. According to the above, it is necessary to select source domain OD. According to daily operation experience, the tickets allocated from Xinxiang East railway station to Zhengzhou East railway station are shared with those allocated by several similar cities. The specific situation is shown in figure 2.

Fig. 2 Source domain OD selection

Input \( OD_1, ..., OD_5 \) to equation (1). The OD of Hebi East Station to Zhengzhou East station is selected as the source domain OD. The source domain samples are also selected from January 1, 2019 to June 1, 2019. The daily passenger volume from Hebi East station to Zhengzhou East station is 151 samples in total.

Every day, target domain OD from Xinxiang East station to Zhengzhou East station share the same tickets allocation with the source domain OD from Hebi East station to Zhengzhou East station. In some cases, due to the fact that the allocated tickets are occupied by Hebi East station to Zhengzhou East station, the number of tickets allocated from Xinxiang East station to Zhengzhou East station is less than the real number of tickets. At this time, the daily passenger volume from Xinxiang East station to Zhengzhou East station is underestimated, and some samples in target domain is biased. If we use these samples to train the model, the accuracy of the prediction results will be reduced, so this part of samples should be removed from the target domain sample set. But after removing some samples, the samples in target domain may be too few, resulting in the phenomenon of "under fitting". Transfer learning solves this problem very well. For the samples removed from the target domain,
samples in the same date are selected from the source domain to replace it. Whether the target domain sample should be removed is determined by equation (9).

\[ f_i(r, s) <= \varepsilon \times \tilde{f}(r, s) \]  \hspace{1cm} (9)

\[ \tilde{f}(r, s) = \frac{1}{M} \sum_{r=1}^{M} f_i(r, s) \]  \hspace{1cm} (10)

\( \varepsilon \) is the threshold, it is determined by the following. \( \tilde{f}(r, s) \) is the mean value of the samples which has the same month and day of week in the target domain. Though equation (9), we have reason to believe that the sample is distorted.We should remove it and replace with the sample in source domain which has the same date.

### 3.2. Feature selection

Since the railway passenger volume has time series characteristics, this paper choose CART as the basic regression model, which can not totally find out the time series characteristics of the sample. Therefore, it is necessary to add the characteristics of time series to the sample features, and introduce the passenger volume of the previous day as the feature. At the same time, the passenger flow can be divided into peak period and trough period. Different month and day of week also have different influence on daily passenger volume. The final selected feature are shown in figure 3.

![Fig.3 Feature of sample](image)

### 3.3. Result analysis

Random choose 10 samples as test data, and the rest samples were input model as training samples. We choose RMSE and MAPE as the evaluation criteria for the final prediction results. RMSE and MAPE are as below.Set \( L = 10 \).

\[ RMSE = \sqrt{\frac{1}{10} \sum_{r=1}^{10} (f_i(r, s) - \tilde{f}(r, s))^2} \]  \hspace{1cm} (11)

\[ MAPE = \frac{1}{10} \sum_{r=1}^{10} \left| \frac{f_i(r, s) - \tilde{f}(r, s)}{f_i(r, s)} \right| \times 100\% \]  \hspace{1cm} (12)

To try to different parameter values for \( \varepsilon \), the results are shown in figure 4.Finally set \( \varepsilon = 0.8 \).

![Fig.4 Selection value of \( \varepsilon \)](image)

Firstly, compare the prediction accuracy of transfer learning and traditional ensemble learning and the final results are shown in figure 5. It can be seen that transfer learning, which selecting appropriate samples from the source domain and combining them with the target domain samples to form the training sample set, the result of the pedicion is 11.9% as standard MAPE. Only using the target domain sample set into AdaBoost ensemble learning algorithm, the MAPE of the model
prediction is 13.1%. It shows that some of the samples in the target domain are affected by the tickets allocation, this lead samples biased and inaccurate. The whole training sample set can be modified by replacing with some samples in the source domain, it can effectively improve the accuracy of the model predicion.

|              | Transfer Learning | Adaboost (only target domain) |
|--------------|-------------------|--------------------------------|
| RMSE         | 108               | 141                            |
| MAPE         | 7.1%              | 13.1%                          |

Fig.5 Comparison results about transfer learning and Adaboost

4. Conclusion
The prediction of special OD passenger flow in the field of railway transportation is an old and new proposition. The difficulty is that many sample data are biased because these samples are obtained under the constraint conditions. The traditional idea is to restore biased data to real unbiased data by estimation, but this method is too abstract and the result is difficult to verify. Using transfer learning by cases(samples), some samples in the source domain are transfer to the target domain, which indirectly solve the problem of how to restore biased samples. With the development of railway OD passenger flow prediction study, it is expected that more and more transfer learning methods will be applied to this field.

Acknowledgments
This paper was financially supported by the National Key Research and Development Plan (2018YFB1201400).

References
[1] Keyu Wen, Guotang Zhao, Bisheng He, Jian Ma. An Improved Transfer Learning Based Time Series Prediction Method for the High-speed Rail Short-term Volume[J]. System Engineering. 2020, 5(38): 73-83.
[2] Yufei Wang, Yila Su, Yaping Zhao, Xiaoqian Sun. Mongolian-Chinese Neural Machine Translation Model Based on Parameter Transfer[J]. Computer Applications and Software. 2020, 9(37): 81-87.
[3] Yu Wang, Zhifei Wang, Hongye Wang, Junfeng Zhang, Ruilong Feng. Prediction of Passenger Flow on CNN-LSTM Hybrid Model[C]. ISCID 2020.
[4] Jin Qin, Xuanke Wu, Yan Xu, Yu Wang. Study on Collaborative Optimization of Dynamic Pricing and Ticket Allocation for High-speed Trains[J]. Journal of the China Railway Society, 2020, 3(42): 32-42.
[5] Sheng Hong, Wenxin Li, Hao Liu. Review of Safe Operation State Assessment Based on Transfer Learning Methods[J]. ADVANCES IN AERONAUTICAL SCIENCE AND ENGINEERING A, 2020, (8): 454-460.
[6] Williams B M, Durvasula P, Brow D E. Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing model[J]. Journal of the Transportation Research Board, 2014, 1644(14): 132-141.
[7] WANG Weiwei. Study on Forecast of Railway Passenger Flow Volume under Influence of High-Speed Railways[J]. Railway Transport and Economy, 2016, 38(4): 7-12.
[8] Hao Ren, Bosong Liu, Jinyang Sun. Advances and Perspectives on Knowledge Transfer Based Cross-Domain Recommendation[J]. Journal of Frontiers of Computer Science and Technology, 2020, 9.