Hindawi
Scientific Programming
Volume 2022, Article ID 1691400, 12 pages
https://doi.org/10.1155/2022/1691400

Research Article

Research on Marketing Strategy of Railway Passenger Travel Behaviour Analysis in Competitive Section

Xiaopei Hao,[1] **Jiansheng Zhu**,[2] **Xinhua Shan**,[2] **and Wen Li**[2]

[1]*China Academy of Railway Sciences, Beijing 100081, China*
[2]*Institute of Computing Technologies, China Academy of Railway Sciences, Beijing 100081, China*

Correspondence should be addressed to Jiansheng Zhu; zhujiansheng@rails.cn

Received 2 April 2022; Revised 9 May 2022; Accepted 16 May 2022; Published 1 June 2022

Academic Editor: Muhammad Usman

The deep mining of passengers' travel data can identify competitive segments and gain insights into passengers' characteristics and differentiated demands. This can not only effectively support precise marketing strategy adjustment of railway transport but also improve its competitiveness in the passenger transportation market. In this paper, hidden railway travel behaviour is introduced and integrated with railway travel behaviour to create a complete passenger travel chain, based on existing distance-based competitive segment recognition methods. The loyalty index values of passengers are calculated using this travel chain to identify competitive segments. Furthermore, passenger classification and grouping currently ignore social relationships as well as personal travel characteristics. Therefore, a novel passenger grouping method is proposed; it integrates individuals' travel characteristics and social relations. Individual travel labels are created for travellers based on their travel data. Social relation topologies, such as ticketing relation, the relation of travelling together, and benefit relation via point redemption, can be extracted using these labels. Social relation traits can be retrieved using graph attention networks and multigraph fusion. Finally, travellers are categorised based on their individual travel characteristics. As an example, and a case study, the grouping of Guangzhou–Shanghai passengers in 2020 is taken which shows that the suggested method has the potential to improve both the precision and the feasibility of grouping railway passengers. As a result, new ideas for passenger grouping in railway marketing might be offered.

1. Introduction

Passengers who have a travel demand select appropriate trip plans according to their characteristics in conjunction with various factors related to transportation services, like safety, comfort, convenience, speed, punctuality, and cost-effectiveness. To assist railway passenger transportation administration departments in formulating customised and personalised service strategies based on the travel characteristics of diverse groups, we need to accurately and effectively define the competitive segments of different transportation modes, profoundly investigate the vehicle selection behaviours of passengers, quantitatively analyse the individual factors influencing passengers' travel choices and their social relations in travel, gain insights into the characteristics and differentiated demands of passengers, and finally divide passengers into different groups. This may further promote the passenger service mode innovation, service strategy transformation, and service quality improvement of railway transportation. Theoretical bases can also be provided for railway passenger transport enterprises to reasonably design train service products and implement precision marketing activities.

Regarding the different transportation modes in China's passenger service market, diverse marketing strategies are selected for various segments to meet passengers' demands and attract passengers, thus improving their market competitiveness. In this context, effective recognition of the competitive segments of various transportation means is the basis on which railway passenger service enterprises analyse the advantages of their competitors, discover their weaknesses, and optimise their marketing strategies. According to Dobruszkes et al. [1], supplies are dynamically adjusted by European aviation companies in line with the running time of G-series high-speed trains. The longer the running time, the greater the number of supplies. Supplies are specifically

at the minimum for running times within 2–2.5 h (corresponding travel distance: approximately 500 km). Through a comparative analysis of the superiorities of the passenger transportation means in Taiwan, Cheng [2] stated that civil aviation, G-series high-speed train travel, and highway travel are primarily suitable for distances of 700 km and above, 200–700 km, and 200 km and below, respectively. According to Zhang et al. [3], the travel distance is a crucial aspect that influences passengers' travel decisions. The 600–1,000 km segment is the most competitive between G-series high-speed trains and civil aviation, while the segment with the highest competition is approximately 1,000 km. Due to the vastness of China's territory, imbalanced economic development between various areas, and changes in passenger composition within segments, distance-based division and categorisation of competitive segment may have several drawbacks.

Group segmentation is a foundation of marketing strategy optimisation. In essence, it aims to learn user characteristics, demands, and objectives by analysing historical data to provide users with customised service strategies, maximise benefits, and optimise service quality. For example, in the intuitive target market selection method of Chou et al. [4], the personal features of individuals are established based on demographic variables to identify potential customers. In another approach, the categories and prices of products purchased by customers are analysed to calculate consumer buying behaviour similarity. The simulated annealing algorithm is applied in a behaviour-based customer segmentation model (Yan et al. [5]). According to Holly, self-organising neural networks may also be used for customer segmentation, depending on the particular features of the customers (Rushmeier et al. [6]). Qian [7] created a mixture regression model to investigate how passengers rate safety, comfort, speed, frequency, punctuality, prices, and convenience; he used the expectation maximisation algorithm to evaluate regression coefficients and calculate the distribution probability of passengers. Bayesian statistics is used for this purpose resulting in passenger group segmentation. The recency, frequency, monetary (RFM) model for customer value judgement was introduced and combined with the analytic hierarchy process and fuzzy clustering to segment passengers into five categories and analyse their potential transformation classes; the resulting model was used to identify customer values (Li [8]). The multiclass twin support vector machine (MTWSVM) has been thoroughly explored and experimentally verified to perform well in multiclass classification problems (Zhang et al. [9]). However, the existing travel behaviour research data are mostly collected by means of questionnaire surveys. Questions in these questionnaires usually have certain shortcomings, such as lack of detailed information. Furthermore, although customer segmentation models principally consider the personal features of customers, they neglect the social relations of these individuals. This makes it unlikely for such models to describe customer characteristics comprehensively based on vectors and thereby compromises the performance of the model.

In this study, a passenger railway travel chain that depends on passenger railway travel data is constructed. Hidden railway travel behaviour is introduced to perfect the railway travel chain and then analysed to recognise relevant competitive segments and calculate the railway travel loyalty indices of passengers. Afterwards, we focus on the grouping of railway passengers in competitive segments to analyse their individual travel characteristics and establish social networks during their travels. The loyalty indices of passengers serve as an initial strategy of group segmentation, and the graph attention mechanism is adopted to build a group recognition model. Through passenger group segmentation for competitive segments, passenger transport products are reasonably designed for different competitive segments of railways, and personalised marketing strategies can be made. As a result, passenger experience is improved, and theoretical support is provided for railway resource utilisation efficiency.

2. Travel Chain Analysis

2.1. Travel Chains. Travel, a door-to-door traffic behaviour performed to achieve a certain trip goal, is defined by a set of behaviours that include information such as departure time, departure location, destination, mode of transportation, and journey distance [10]. A travel chain represents the entire passenger travel process. It is made up of connecting links that are placed according to the departure time of a travel behaviour. Generally, passengers select appropriate transportation means to achieve their trip purposes and generate complete travel chains for themselves.

The data involved in this study are primarily derived from the real-name system and travel information of railway passengers from Guangzhou to Shanghai. Because of data limitations, no complete travel chains can be formed from the data of passengers who go on tours by multiple modes of transportation. Therefore, hidden railway travel behaviour is introduced, and urban transport is ignored to generate complete travel chains for these passengers. Travel data from 2020 are ranked based on riding time to construct the travel chains of passengers, as shown in Table 1. A travel chain (LC) is formed through an end-to-end connection between the railway travel behaviour and the hidden railway travel behaviour, which are, respectively, defined as TB_j = (train_date, start_time, start_city, to_city) and OB_j = [start_date, stop_time, start_city, to_city]. Integrity (CP) signifies whether the railway travel behaviour of a passenger constitutes a complete travel chain, that is, whether the destination city of the j th trip by train is the departure city selected for the $(j + 1)$th trip by train. TBH is the number of hidden railway travels. It represents the least number of trips that need to be increased when a passenger produces a complete travel chain based on a railway trip. Loyalty to the railway industry, which is denoted by LOY_i, indicates the probability of passenger i to complete intercity displacement by train, and it is expressed as

$$LOY_i = \frac{TBR_i}{TBR_i + TBH_i} * 100, \tag{1}$$

where TBR_i stands for the total number of times passenger i travels by train in a travel chain.

Table 1: Railway passenger travel chain in 2020.

ID	LC	CP	TBH	LOY	ODY	DDY
P_1	(20200119, 1846, Hangzhou, Nanjing, 322 km) -> [20200814, 20200816, Nanjing, Suzhou, 297 km] -> (20200212, 1046, Suzhou, Hangzhou, 692 km) -> (20200718, 0910, Hangzhou, Hefei, 443 km) -> (20200719, 1504, Hefei, Hangzhou, 439 km) -> (20200720, 1739, Hangzhou, Nanjing, 322 km) -> (20200807, 2019, Nanjing, Suzhou, 297 km) -> (20200921, 1304, Suzhou, Hangzhou, 692 km)	No	1	87.5	(Nanjing, Suzhou, 50)	(250 km–350 km, 75)
P_2	(20201004, 1816, Jinhua, Cangnan) -> (20201005, 1730, Cangnan, Jinhua)	Yes	0	100	No	No
P_3	(20200605, 1849, Hangzhou, Jiaxing, 79 km) -> (20200607, 1127, Jiaxing, Hangzhou, 78 km) -> (20201006, 0944, Hangzhou, Yuyao) -> (20201006, 1928, Yuyao, Hangzhou) -> (20201107, 1529, Hangzhou, Jiaxing, 79 km) -> (20201108, 1429, Jiaxing, Hangzhou, 78 km)	Yes	0	100	No	No
P_4	(20200810, 1413, Jiaxing, Hangzhou, 78 km) -> (20200810, 1501, Hangzhou, Jinhua) -> [20200810, 20200814, Jinhua, Hangzhou, 153 km] -> (20200814, 1308, Hangzhou, Jinhua) -> [20200814, 20200816, Jinhua, Quzhou, 110 km] -> (20200816, 2240, Quzhou, Zhuzhou, 666 km) -> (20200822, 2156, Zhuzhou, Hangzhou, 932 km) -> (20200823, 2050, Hangzhou, Jiaxing, 79 km) -> (20200827, 1457, Jiaxing, Hangzhou, 78 km) -> (20200827, 1532, Hangzhou, Quzhou, 263 km) -> [20200827, 20200830, Quzhou, Hangzhou, 260 km] -> (20200830, 2141, Hangzhou, Jiaxing, 79 km) -> (20201115, 1523, Jiaxing, Hangzhou, 78 km) -> (20201115, 1614, Hangzhou, Liling, 887 km) -> (20201116, 1533, Liling, Jinhua, 704 km) -> (20201116, 2050, Jinhua, Hangzhou, 153 km)	No	3	81.25	(Jinhua, Hangzhou, 50) (Jinhua, Quzhou, 0)	(150 km–2000 km, 50) (50 km–100 km)
P_5	(20200820, 1832, Hefei, Yangzhou, 143 km) -> (20200824, 1506, Yangzhou, Shanghai, 324 km) -> (20200824, 1828, Shanghai, Hangzhou, 163 km)	No	0	100	No	No

$\mathrm{ODY}(F,T)_i$, loyalty to a segment, signifies the probability of passenger i, who has a travel demand in segment (F,T) (the segment from departure F to destination T) to select a railway. It is determined as

$$\mathrm{ODY}(F,T)_i = \frac{\mathrm{TBR}(F,T)_i}{\mathrm{TBR}(F,T)_i + \mathrm{TBH}(F,T)_i} * 100, \quad (2)$$

where $\mathrm{TBR}(F,T)_i$ is the total number of times passenger i takes a train in segment (F,T) belonging to his/her travel chain and $\mathrm{TBH}(F,T)_i$ represents the number of times a hidden railway travel behaviour occurs in segment (F,T).

$\mathrm{DDY}(D_f,D_t)_i$, loyalty to travel distance, is the probability of passenger P_i, who has a travel demand to take a train over the trip distances of D_t and D_f. They can be calculated by

$$\mathrm{DDY}\left(D_f,D_t\right)_i = \frac{\mathrm{TBR}(F,T)_i}{\mathrm{TBR}(F,T)_i + \mathrm{TBH}(F,T)_i} * 100, \quad D_f \le D(F,T) \le D_t, \quad (3)$$

where D_f and D_t are the maximum and minimum travel distances, respectively. The travel distances, $TBR(F,T)_i$ and $TBH(F,T)_i$, must be within the range of $[D_f,D_t)$.

According to Table 1, the travel chain (TC_1) of passenger P_1 consists of seven railway travel behaviours. TB_1 represents a travel behaviour involving departure by train from Hangzhou at 18:46, 19 January 2020, and arrival in Nanjing; TB_2 is a travel behaviour involving departure from Suzhou at 10:46, 12 February 2020, and arrival in Hangzhou; TB_3 refers to a departure from Hangzhou at 9:10, 18 July 2020, and arrival in Hefei; TB_4 means that the passenger leaves Hefei at 15:04, 19 July 2020, for Hangzhou; TB_5 stands for departure from Hangzhou at 17:39, 23 July 2020, and arrival in Nanjing; TB_6 means a departure from Nanjing at 20:19, 7 August 2020, and arrival in Suzhou; TB_7 involves leaving Suzhou at 13:04, 21 September 2020, and arriving in Hangzhou. Analysis shows that in the travel chain of passenger P_1, the destination city of TB_1 is not the departure city of TB_2. This reveals that this passenger chooses another mode of transportation to complete his/her travel from Nanjing to Suzhou. In other words, at least one travel behaviour from Nanjing to Suzhou is absent. Therefore, the number of occurrences of hidden railway travel is 1. Considering that 8 is the number of times of travel in a complete travel chain, the number of trips completed by train is 7. From equation (1), the passenger's loyalty to travelling by train is 87.5. Moreover, the hidden railway travel behaviour of this passenger occurs in the segment from Nanjing to Suzhou; hence, either the number of occurrences of railway travel or that of hidden railway travel is 1 in this segment. According to equation (2), loyalty

to the segment of a hidden railway travel behaviour is 50. Equation (3) is utilised to determine the passenger's loyalty to a travel distance of 250–350 km, which is 75.

2.2. Analyses of Competitive Segments. The proportion of a hidden railway travel behaviour in a segment can effectively show whether the passenger service products designed for this segment are reasonable, whether the service quality needs to be further enhanced and whether its marketing strategies should be optimised. A large proportion indicates that the existing railway passenger services in the segment fail to meet the travel demands of most passengers. The greater the proportion, the lower the competitiveness of this segment. In this study, the proportions of hidden railway travel behaviours reflect the competitiveness of competitive segments, as expressed by

$$\mathrm{TBHP}(F,T) = \frac{\sum_{i=1}^{n} \mathrm{TBH}_i}{\sum_{i=1}^{n} \left(\mathrm{TBR}_i + \mathrm{TBH}_i\right)}, \tag{4}$$

where $\mathrm{TBHP}(F,T)$ represents the competition intensity from a departure city (F) to a destination city (T); $\sum_{i=1}^{n} TBH_i$ is the number of occurrences of all hidden railway travel behaviours in the segment; and $\sum_{i=1}^{n} TBR_i + TBH_i$ is the sum of the total number of occurrences of railway travel behaviour and that of hidden railway travel behaviour in the segment.

Spark is used to analyse data related to the railway travel behaviour of all passengers in 2020. Being an open source, Apache Spark is a distributed processing system for big data workloads. On this basis, the proportions of hidden railway travel behaviours in segments of different distances are obtained, as shown in Figure 1. An increase in the travel distance is clearly accompanied by a drop followed by an increase in the competition intensity. When the competition intensity of a segment with a travel distance of no more than 50 km exceeds 10%, highways characterised by flexibility, simplicity, and convenience become the main competitors of railways. At travel distances over 1,350 km, the corresponding competition intensity can be raised accordingly. In such segments, flights become the main competitors of railways because of their high speed, safety, and other benefits. The competition is less intense when the travel distance varies from 150 to 1,000 km; therefore, this can be regarded as the dominant segment where railways are superior to other modes of transportation.

3. Construction of Passengers' Travel Behaviour Characteristics

This section presents discussions that are based on railway ticketing data and oriented by the passenger transportation market, and aviation marketing strategies are used as reference. From the perspectives of passengers' personal characteristics and social relations, it aims to fulfil the mining, clustering, regrouping, and deep fusion of digital railway passenger transportation resources. Furthermore, different passenger groups can be segmented, which may help gain insights into the associations between passengers' characteristics and their selection of transportation means.

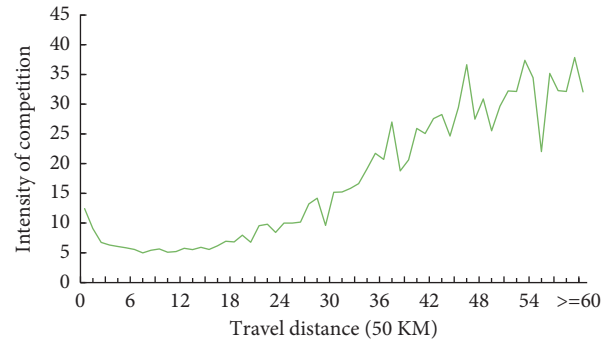

Figure 1: Relationship of travel distance (50 km) and competition intensity.

Hopefully, a data basis can be provided for model improvement, theory refinement, the research process, and the optimisation method.

3.1. Travel Characteristics of Individuals. In the CRNet ticketing system, feature data associated with individual passengers can be divided into two categories: demographic data (about natural attributes) and travel behaviour data. The former relates to information already stored in the system, such as gender, age, and residence. These data are often known as static data since they rarely change and have a relatively constant data structure. A sequence of behaviour records made during a trip, such as ticket booking, travel, ticket check, and inbound/outbound data, fall under the latter group. They are also known as dynamic data because of their high frequency of occurrence. Passengers are shown using multiple data dimensions based on these two kinds of data. As given in Figure 2, the static and dynamic characteristics of passengers are generated with abstract semantic labels that are easy to understand, thus producing a full view of passenger information [11].

3.2. Social Relations. Most passengers in a transportation system do not make decisions on their own, including how their travel requests are generated, how their travel routes are planned, and how their travel times and modes are decided. Passengers are influenced by their social relationships in addition to their preferences and traffic situations.

Since the 12306 Internet ticketing system went online in 2012, massive data capable of embodying social relations have been accumulated by its unique business process. Based on these data, we can extract ticketing relations, relations of travelling together, and relation of benefits by the point redemption mechanism.

3.2.1. Ticketing Relation. This is a relation between a purchaser and a passenger. A single ticketing relation includes the following information: ticket purchaser, passenger, ticket purchasing time, ticket price per kilometre, and number of tickets. Here, $G(i, j)$ represents a behavioural sequence of a ticketing relation in which passenger i buys a ticket for passenger j; $G(i, j)[k] = (b_k, gt_k, atp_j[k])$ denotes a record

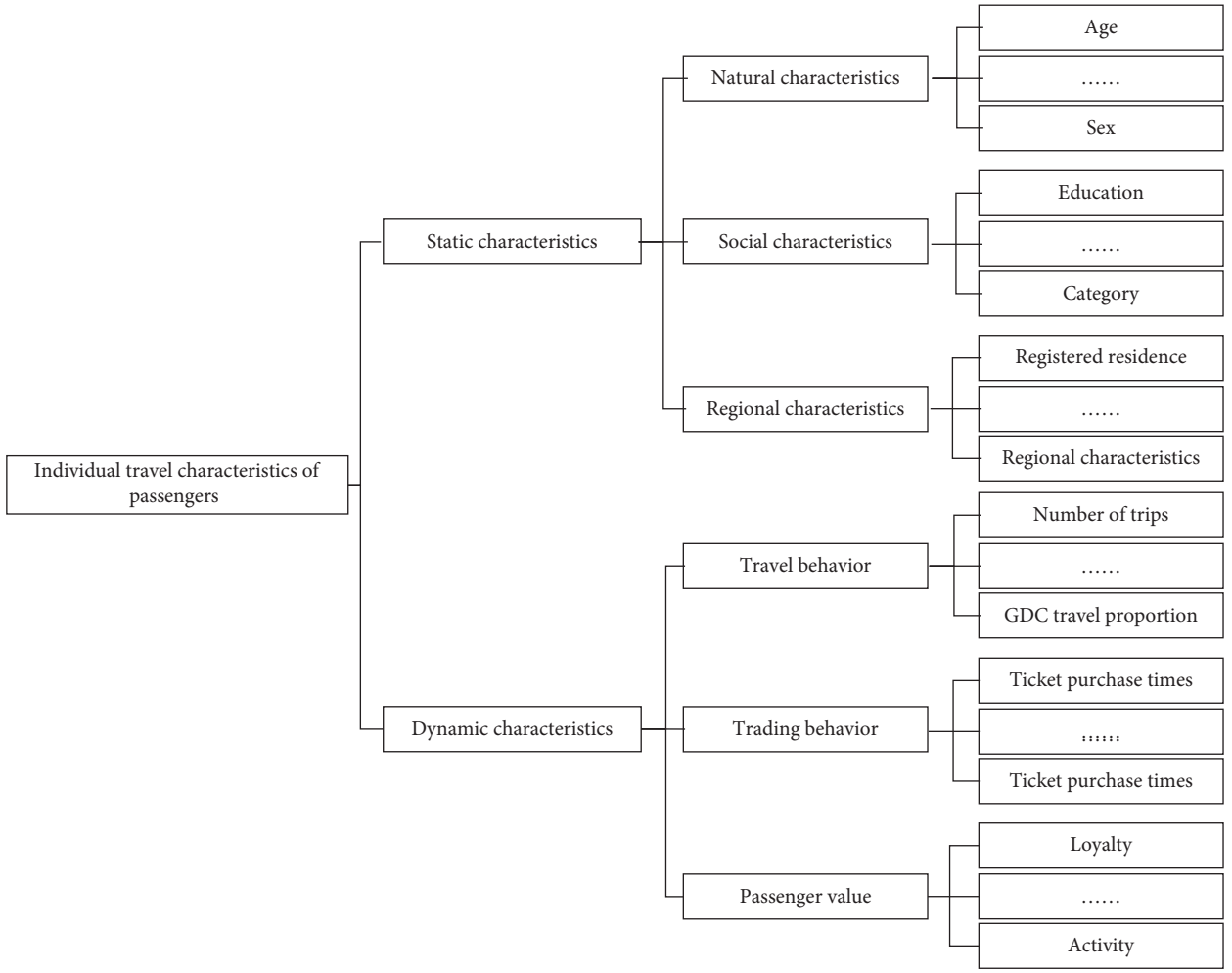

Figure 2: Travel characteristics of individuals.

of the k th ticket purchasing behaviour, where b_k stands for the number of tickets purchased, gt_k the time of buying a ticket, and $atp_j[k]$ the ticket price per kilometre. In accordance with the sequence of a passenger's ticketing relation, the weight of this relation is determined as

$$wr_{i,j}^{g} = \sum_{k} w_g[k] = \frac{1}{\sqrt{b_k \times atp_j[k]} \times \left((ct - gt_k)/(ct - ft)\right)}, \quad (5)$$

where $w_g[k]$ is the weight generated by passenger i when buying a ticket for passenger j the k th time, ct is the current time, and ft is the start date of the sample data. The weight of the ticketing relation is time sensitive and may attenuate as the time window increases.

The 12306 Internet ticketing system has 600 million registered users. According to an analysis of the number of their frequent contact persons (Figure 3), only 34% of these registered users have a single-frequency contact (i.e., the user himself/herself), whereas ticketing relations can be found among over 60% of the passengers when they buy tickets.

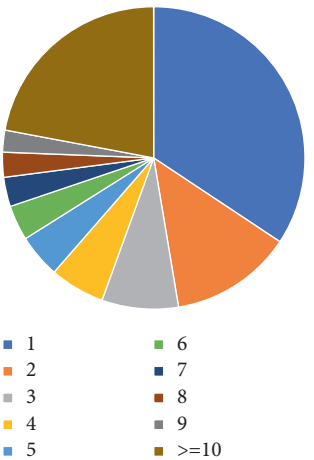

Figure 3: Distribution proportions of the number of frequent contacts.

3.2.2. Relation of Travelling Together. The relation of travelling together exists in passengers under the same ticket booking order, including the specific passengers, riding time, ticket price per kilometre, and number of passengers travelling together. Here, $C(i, j)$ is a behavioural sequence in which passenger i buys a ticket for passenger j; $C(i, j)[k] =$

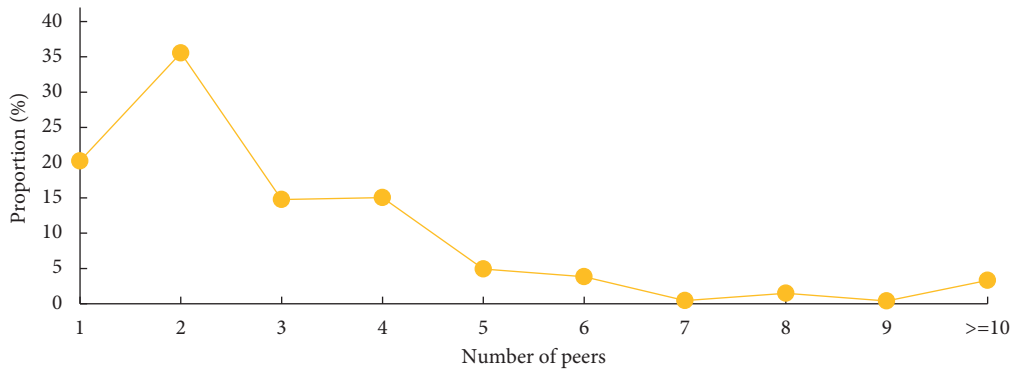

Figure 4: Distribution proportions of the number of passengers travelling together.

$(c_k, t_k, atp_j[k])$ represents a record of the k th ticket purchasing behaviour, where c_k stands for the number of passengers travelling together, t_k for the riding time, and $atp_j[k]$ for the ticket price per kilometre. Depending on the sequence of a relation of travelling together, the weight of this relation can be calculated according to

$$wr^c_{i,j} = \sum_k w_c[k] = \frac{1}{\sqrt{c_k \times atp_j[k]} \times \left((ct - t_k)/(ct - ft)\right)}, \tag{6}$$

where $w_c[k]$ is the weight of the fact that passenger i travels together with passenger j for the k th time.

The number of passengers falling into the same online order numbers in 2020 is statistically analysed, and the results are presented in Figure 4. Only 20% of the passengers travelled alone that year, and a relation of travelling together is found among the remaining passengers.

3.2.3. Benefit Relation by Point Redemption Mechanism. This relation means that the purchaser buys a ticket for another passenger through point redemption. A single benefit relation by the point redemption mechanism consists of the following information: the purchaser, the other passenger, riding time, and ticket price per kilometre. Here, $S(i, j)$ is a sequence representing the act of passenger i buying a ticket for passenger j through point redemption; $S(i, j)[k] = (t_k, atp_j[k])$ is a record of the k th ticket purchasing behaviour based on the point redemption mechanism, where t_k stands for the riding time and $atp_j[k]$ for the ticket price per kilometre. Depending on the sequence of this relation, the corresponding weight is computed according to

$$wr^s_{i,j} = \sum_k w_s[k] = \frac{1}{\sqrt{atp_j[k]} \times \left((ct - t_k)/(ct - ft)\right)}, \tag{7}$$

where $w_s[k]$ is the weight of passenger i buying a ticket for passenger j for the k th time by point redemption.

The benefit relations between purchasers and other passengers in orders made through point redemption in 2020 are analysed, as presented in Figure 5. Nearly 30% of the purchasers paid using their points for other passengers in 2020, thus forming a benefit relation with these passengers by the point redemption mechanism.

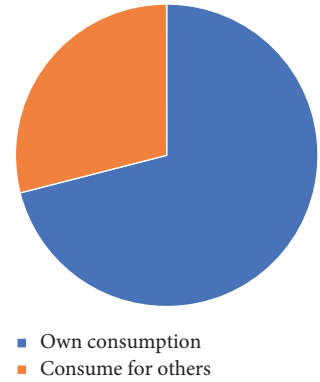

Figure 5: Distribution proportions of the benefit relation by the point redemption mechanism.

3.2.4. Passenger Classification. The railway trips of a passenger are related to their loyalties to railway travel, hidden travel segments, and travel distance. In this paper, passengers of a certain segment are grouped, and the importance of their loyalties is ranked as follows: $O\ DY > LOY > D\ D\ Y$. According to equation (8), LY, passengers' loyalty to a segment, can be calculated in combination with weights and diverse loyalty indices.

$$LY = 0.5 * O\ DY + 0.3 * L\ DY + 0.2 * D\ D\ Y. \tag{8}$$

Based on their indices, passengers are divided into the following groups: low loyalty (0–10), moderate loyalty (11–50), high loyalty (51–80), and very high loyalty (81–100).

4. Railway Passenger Grouping Model

A passenger grouping model integrating social relations is presented based on the travel characteristics and social relations of individual passengers. This model is made up of a personal travel characteristics fusion layer, a social relation fusion layer, an activation layer, and a group categorisation layer, as illustrated in Figure 6. Passengers' personal qualities are initially chosen as input. The feature fusion layer is then used to achieve personal feature vector fusion, and dimensionality reduction is used to lower the complexity of the corresponding algorithm. Following that, a social network topology is built based on passenger social relationships. The

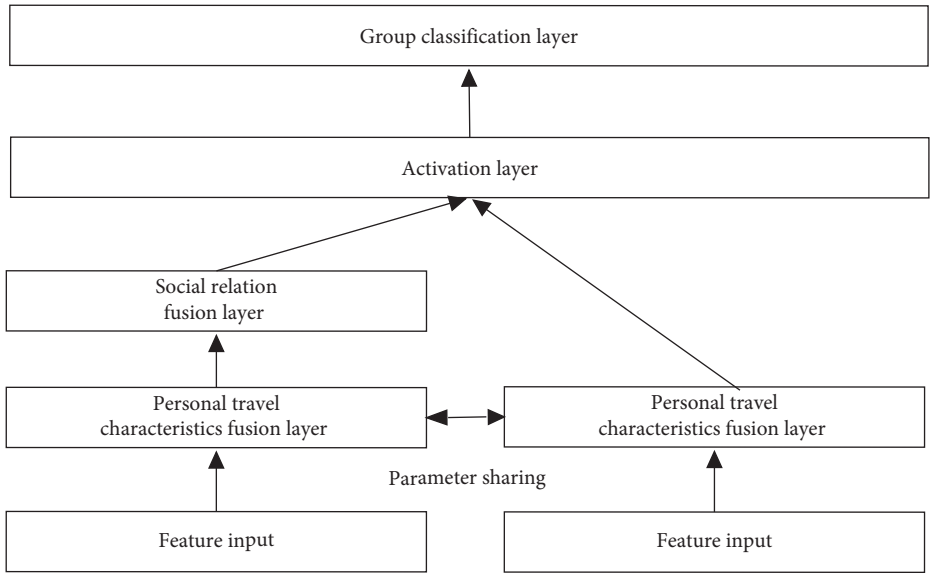

Figure 6: Structure of the graph attention network (GAT) with social relation fusion.

social relation fusion layer receives this topology, as well as the fused personal travel characteristics, as input. This approach is expected to realise feature information interaction between a goal node and a neighbouring node. In addition, the activation layer is designed to acquire the target values of passenger grouping from the passengers' personal characteristics and passengers' characteristics fused with the neighbouring node.

4.1. Personal Feature Fusion Layer. The number of personal travel characteristics already exceeds 2,000 in the user portrait system of railway passengers, which covers redundant and noisy information. This may not only interfere with subsequent data analysis but also affect the algorithm complexity, increase the computation overhead, and eventually influence the accuracy and efficiency of classification. Therefore, an autoencoder is introduced based on feature dimension reduction as the personal feature fusion layer. By virtue of this encoder, data in the high-dimensional feature space of passengers can be mapped to a low-dimensional space to reconstruct the passengers' personal features [12] and acquire the essential structural features of their characteristics. To decrease model complexity and improve training efficiency, personal features are processed through the personal feature fusion layer during personal feature processing and social relation fusion, in addition to parameter sharing.

$$\overrightarrow{f}_i' = \text{Autoencoder}\left(\overrightarrow{f}_i\right), \overrightarrow{f}_i' \in \mathbb{R}^{P'}, \overrightarrow{f}_i \in \mathbb{R}^{P}, \quad (9)$$

where $\overrightarrow{f}_i$ refers to the original feature vector of passenger i, P the number of original features, $\overrightarrow{f}_i'$ the feature vector of passenger i after feature fusion, and P' the number of fused features.

4.2. Social Relation Fusion Layer. The structure of the social relation fusion layer is presented in Figure 7. It consists of three social relation networks and a multigraph feature fusion process.

4.2.1. Social Relation Network. A social network may clearly embody the intended ticketing relation, relation of travelling together, and benefit relation via the point redemption method. Furthermore, the social network of railway passengers is represented by three undirected weighted graphs, namely, $G^g = (P, E^g, F', Wr^g)$, $G^c = (P, E^c, F', Wr^c)$, and $G^s = (P, E^s, F', Wr^s)$, where G^g, G^c, and G^s are the graphs of the ticketing relation, relation of travelling together, and benefit relation by the point redemption mechanism, respectively; P is the set of all railway passengers; E^g, E^c, and E^s are the sets of the ticketing relation, relation of travelling together, and benefit relation by the point redemption mechanism, respectively; F' is the set of the personal travel characteristics of all passengers after feature fusion; Wr^g, the weight of the ticketing relation, comprises $wr^g_{i,j}$; and finally, Wr^c and Wr^s are the weights of the travelling together relation and the benefit relation, respectively (the former is formed by $wr^c_{i,j}$, whereas the latter is composed of $wr^s_{i,j}$).

The 12306 Internet ticketing system has over 600 million registered users. The number of passengers is nearly 900 million. Moreover, there are some abnormal accounts. For these reasons, the relations of travelling together and ticketing are rather complicated for some passengers. In addition, a large difference lies in the number of neighbouring nodes around each node. As the passenger nodes possess a great number of neighbouring nodes, samples are taken from these neighbouring nodes to improve model training efficiency. We assume that the number of neighbouring nodes is N, and the corresponding sampling prescription is as follows.

When $N \leq 20$, all nodes are treated as social relation network nodes. As for $N > 20$, the nodes need to be classified based on the number of times of ticketing and the number of

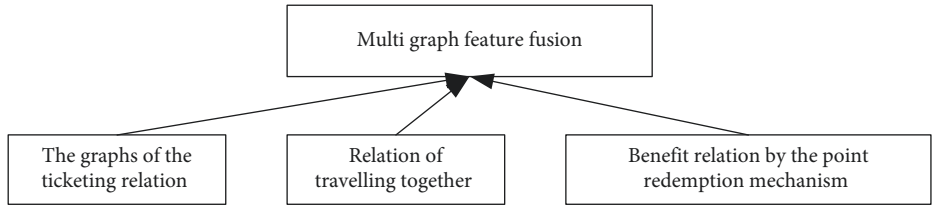

Figure 7: Structure of the GAT.

times of railway travel. For each category, the nodes are ordered based on the number of times and divided into three intervals with proportions of 40%, 40%, and 20% (N_1, N_2, and N_3, respectively). The neighbouring nodes in each interval are sampled in a ratio of $20/N$, and the number of neighbouring node samples can be expressed in $N = N_1 + N_2 + N_3$.

4.2.2. GAT Layer. In the GAT, the inherent normalised functions are replaced with an attention mechanism to assign a weight to each passenger node. During the updating of the hidden layer, the nodes and neighbouring nodes are aggregated according to the magnitude of weights [13].

In the present study, three types of social relations are included. For the relation of ticketing, for example, a feature vector set of target passengers and their neighbouring nodes is used as the input of the GAT layer, which can be written as

$$f' = \left\{\overrightarrow{f_o}', \overrightarrow{f_1}', \ldots, \overrightarrow{f_N}'\right\} \overrightarrow{f_i}' \in \mathbb{R}^{P'}, \tag{10}$$

where f' is the feature vector set of nodes (passengers' personal characteristics), $\overrightarrow{f_o}'$ the feature vector set of target nodes, $\overrightarrow{f_i}'$ the feature vector set of the i th neighbouring node of the target node, N the number of neighbouring nodes associated with the target node, and P' the number of passengers' features after fusion.

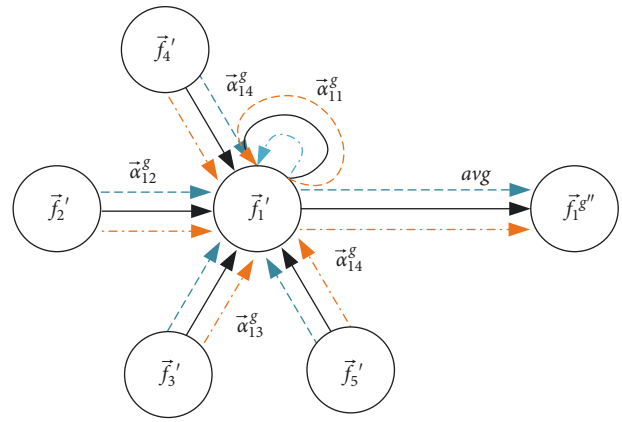

Figure 8: Multigraph attention mechanism.

A graph attention coefficient is constructed to output target node features that contain neighbouring node features. The corresponding computational formula is

$$e^{g} = a^{g}\left(W^{g}\overrightarrow{f}_i', W^{g}\overrightarrow{f}_j'\right), \tag{11}$$

$$\alpha_{ij}^{g} = \text{soft max}\left(e_{ij}^{g}\right) = \frac{\exp\left(\text{LeakyReLU}\left(a^{g\,T}\left[W^{g}\overrightarrow{f}_i', W^{g}\overrightarrow{f}_j'\right]\right)\right)}{\sum_{k\in N+i}\exp\left(\text{LeakyRe}LU\left(a^{g\,T}\left[W^{g}\overrightarrow{f}_i', W^{g}\overrightarrow{f}_k'\right]\right)\right)}. \tag{12}$$

In equation (11), W^{g} is a shared parameter in the network of ticketing relations, and it is used for feature enhancement. $a^{g}(\cdot)$ represents the importance of the target and neighbouring nodes in this network. In equation (12), α_{ij}^{g} is an attention coefficient of nodes i to j, and LeakyReLU to an activation function.

After the normalised attention coefficient is obtained, linear combinations of the corresponding features are calculated and then selected as the final output features of each node. In this paper, multigraph attention is introduced. The multigraph attention mechanism can be utilised to determine the attention coefficients of surrounding nodes, thus stabilising the learning process of the model. An update process for the hidden state is depicted in Figure 8.

Regarding the computational results subjected to K independent attention mechanisms, K-means is adopted and takes the place of a connection. Its computational formula is

$$f_i^{\,g''} = \sigma\left(\frac{1}{K^{g}}\sum_{k=1}^{K^{g}}\sum_{j\in L_i}\alpha_{ij}^{g\,k}W^{g\,k}\overrightarrow{f}_j'\right), \tag{13}$$

where $f_i^{\,g''}$ is a feature vector after a fusion between a target passenger and information of neighbouring nodes in a ticketing relation network formed by this passenger, K^{g} stands for the serial number of an independent attention mechanism, $\sigma(\cdot)$ stands for the activation function, and $\alpha_{ij}^{g\,k}$ stands for the attention coefficient of passenger i relative to passenger j in the network of ticketing relations.

With the use of the abovementioned calculation processes, the ticketing relation network feature fusion, feature fusion $f_i{}^{c''}$ of the relation of travelling together, and feature fusion $f_i{}^{s''}$ of the benefit relation by the point redemption mechanism are obtained.

4.2.3. Multigraph Feature Fusion. A fully connected layer is established for the multigraph fusion of $f_i{}^{g''}$, $f_i{}^{c''}$, and $f_i{}^{s''}$ (vectors of features incorporating social relations), which can be expressed as

$$f_{\mathrm{PSC}\,i} = W_{ps}\left(f_i{}^{g''}, f_i{}^{c''}, f_i{}^{s''}\right), \tag{14}$$

where $f_{PSC\,i}$ is the target passenger feature undergoing fusion with multiple social relations and W_{ps} is a training parameter that denotes the importance of the three relation-generating features.

4.3. Activation Layer. Node feature vectors that incorporate the ticketing, travelling together, and benefit relations are obtained through training by the GAT layer. Personal feature vectors are also acquired through training by the personal feature fusion layer. Afterwards, the node and personal feature vectors are aggregated to generate the final feature vector, which is then transferred to the activation layer. In this way, different groups can be obtained as

$$\widehat{f} = \tanh\left[f_{\mathrm{PIC}} \cdot f_{\mathrm{psc}}\right], \tag{15}$$

$$q(c) = \mathrm{soft\,max}\left(W_q \widehat{f} + b_q\right). \tag{16}$$

In equation (15), f_{PIC}, f_{PIC} stands for the feature vectors outputted from the personal feature fusion layer, f_{psc} for the feature vectors outputted from the social relation fusion layer, and $\widehat{f}$ for the final feature vector of the target passenger. In equation (16), c is the class label of passengers and $q(\cdot)$ is the predicted passenger group.

4.4. Model Training. The methodology divides passengers into four categories based on the passenger loyalty indices. Vectors of passengers' personal travel characteristics are developed based on passenger portraits of railway transportation through supervised training. A network of travellers' social interactions is formed using information from common contacts and online orders (among other things), and then used as the model input through rule-based pruning. Additionally, relevant cross-entropy loss functions are minimised via L_2 normalisation to fulfil model training. The corresponding computational formula is

$$L_{\mathrm{loss}} = -\sum_{c \in C} \widehat{q}_c \cdot \ln q_c + \lambda \sum_{\theta \in \Theta} \theta^2, \tag{17}$$

where $\widehat{q}_c$ and q_c represent the actual class labels of passenger groups and their model-predicted class labels, respectively, and λ stands for the normalised parameter L_2 and Θ for the set of model parameters.

Table 2: Loyalty-based distribution of Guangzhou–Shanghai passengers in 2020.

Passenger types	Class labels	Loyalty ranges	Proportions (%)
Low loyalty	1	0–10	44
Moderate loyalty	2	10–50	22
High loyalty	3	50–80	13
Very high loyalty	4	80–100	21

5. Case Study and Experiments

This section descries the overall result of current study.

5.1. Data Description. The dataset for the case study in this paper is the real-name information of railway passengers and their travel data, both of which underwent masking in 2020. As seen in Section 2.2, competition may become increasingly fierce once the travel distance exceeds 1,500 km. Moreover, passengers in the segment from Guangzhou to Shanghai (travel distance: 1,800 km) are adopted as the research object. An analysis of the travel chain of passengers in 2020 shows a total of 401,300 passengers (railway travel and hidden railway travel behaviours) from Guangzhou to Shanghai. Their loyalty indices are calculated for passenger segmentation. Here, numerals 1, 2, 3, and 4 are the model output of different groups, as shown in Table 2.

For reducing the model complexity, 14 travel features are selected from passengers and listed in Table 3. Features with a large span are normalised and then combined with a social relation network constructed for the 401,300 passengers to serve as the model input.

5.2. Experimental Design. Two experiments are designed for this study: an accuracy test and a compatibility test. Five common classification models (Table 4) are introduced in the accuracy tests for training comparison and accuracy evaluation of the Guangzhou-Shanghai passenger grouping. The compatibility test focuses on grouping prediction for the passengers from January to October 2021 based on the model training of the 2020 passenger data and an analysis of the time-varying performance of the passenger grouping model.

Here, k-fold cross-validation [16] is used to eliminate statistical errors incurred by the use of different training subsets. The dataset is randomly divided into k groups; for model construction, one group is used successively as the test dataset, and the remaining $k - 1$ groups are regarded as training sets. Based on the data size of the experimental samples, the training datasets are randomly classified into five groups.

5.3. Evaluation Indices. Accuracy, precision, recall, and harmonic mean F1 are primarily selected as comprehensive evaluation indices of the passenger grouping model to assess the accuracy of the passenger grouping results.

Table 3: Passenger feature selection.

Serial nos.	Features	Categories	Descriptions
1	Students or not	Enumeration	0: no, 1: yes
2	Business persons or not	Enumeration	0: no, 1: yes
3	Sex	Enumeration	0: male, 1: female
4	Age	Numerical values	Normalisation
5	Number of times of local train ticket booking	Numerical values	Normalisation
6	Number of times of D-series high-speed train ticket booking	Numerical values	Normalisation
7	Number of times of taking local trains	Numerical values	Normalisation
8	Number of times of taking D-series high-speed trains	Numerical values	Normalisation
9	Proportion of passengers taking D-series high-speed trains	Numerical values	Actual values
10	Proportion of passengers purchasing D-series high-speed trains	Numerical values	Actual values
11	Proportion of passengers taking premium seats	Numerical values	Normalisation
12	Proportion of passengers forming a relation of travelling together	Numerical values	Actual values

Table 4: Grouping algorithm description.

Algorithms	Description
Random forest	An algorithm integrating multiple decision trees through ideas of ensemble learning to achieve classification
XGBoost [14]	A gradient boosting decision tree that combines multiple weak classifiers accumulatively into a strong classifier to minimise the objective loss function
MTWSVMs	Primarily fulfilling multiclass problems
LightGBM [15]	A distributed classification algorithm implementing GBDT enabling highly efficient training over large-scale data with low memory cost and high accuracy
DNN	Deep learning

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Re call} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{F1} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \tag{18}$$

In comparison with deep learning models, machine learning models, such as random forest, XGBoost, MTWSVMs, and LightGBM, have lower complexity, fewer training parameters, and shorter training times. However, the classification accuracy of machine learning models is poor. Training time is, therefore, not considered a model evaluation index in this experiment.

5.4. Experimental Results and Analyses. Through threshold adjustment, optimal results of various classification methods are obtained based on the training datasets. The values of the model evaluation indices are determined as well. For each passenger grouping model, the corresponding indices are averaged via fivefold cross-validation. The results are listed in Table 5.

According to Table 5, random forest has the worst accuracy and precision; XGBoost, MTWSVMs, and LightGBM outperform it to a certain extent in terms of these indices. The performance of the proposed model is superior to that of the other models. In some cases, the proposed model even performs the best, followed by DNN. The overall tendency of the proposed model for recall is the same as that for accuracy and precision. According to the F1 values, there are 85% commonalities in the passenger characteristics reflected in the proposed passenger grouping model. Hence, the passenger grouping of the proposed model is highly accurate.

Table 5: Comparison of model-based experiments.

Models	Evaluation indices			
	Accuracy	Precision	Recall	F1
Random forest	59.83	60.13	83.21	69.81
XGBoost	62.63	62.93	84.92	72.28
MTWSVMs	64.36	65.23	85.82	74.12
LightGBM	63.62	65.68	83.91	73.68
DNN	71.42	72.82	88.64	79.95
Proposed algorithm	82.54	84.73	92.36	88.38

The 2021 testing data are separated into 10 parts by month, and passenger groups are predicted using these parts. Figure 9 shows the prediction results, with the x-axis representing the months and the y-axis representing the F1 values. The F1 values of all models gradually decrease with time, and passenger grouping effects turn worse. Previous data training models are no longer sufficient to meet the future segmentation needs of passengers' attributes in this case. The longer the interval from the training time, the worse the passenger grouping results. In particular, the F1 values of random forest and LightGBM are already below 50% in October 2021, showing the worst adaptation. Concerning all tests, the proposed model produces F1 values no less than 70, proving that it is well applicable to future data.

In summary, travel characteristic selection and knowledge extraction are important factors influencing the results of passenger grouping. The features of the random forest model are comparatively static and simple; it ignores feature

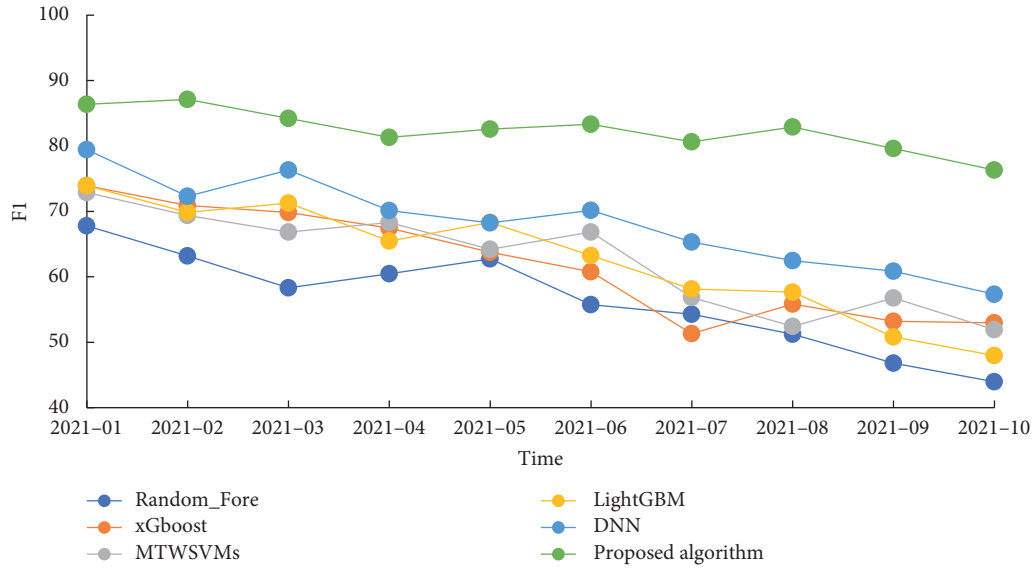

Figure 9: F1 variation prediction of future data based on six approaches.

correlations and produces the worst grouping results under the same conditions. The XGBoost model performs well in terms of prediction precision when applied to low-/medium-dimensional data, but it fails to adapt to large-scale feature inputs. Given a large sample size and large numbers of features and classes, the number of subclassifiers of MTWSVMs exponentially rises, thus excessively increasing the complexity of the corresponding classification system. In this scenario, large quantities of passengers cannot be grouped. The LightGBM model is highly susceptible to noisy information. Finally, DNN can perform feature fusion for passengers' travel characteristics, thereby reducing feature processing complexity and improving model accuracy. Regarding the proposed passenger grouping algorithm, the fusion of social relations and personal features is completed, and the model can correctly extract common features of various passenger groups. From the perspectives of effects, accuracy, and adaptation, the proposed algorithm outperforms the abovementioned existing models.

6. Conclusions

To create a comprehensive travel chain for passengers, hidden railway travel behaviour is introduced and integrated with railway travel behaviour. Passengers' indices of loyalty to railway travel, hidden railway travel segments, and travel distance are determined independently based on passengers' specific information, such as the number of instances of hidden railway travel behaviour, number of railway travels, travel distances, and travel segments. Furthermore, the ratios of segments featuring hidden railway travel behaviour are determined in order to disclose the degree of competition in various segments. The competition is at its peak when the journey distance exceeds 1,350 kilometres, according to the findings. The competition intensity is comparatively low for travel distances of 150 to 1,000 km, and the railway clearly outperforms. As a result, passengers' personal travel characteristics are determined from the dimensions of time and space established on their travel behaviour and real-name data. Furthermore, the point redemption system creates a social relation network of passengers based on their ticketing, travelling together, and benefit relations. Finally, the loyalty of passengers is determined by determining their devotion to railway travel, hidden railway travel segments, and travel distance. The passengers are then divided into four categories based on their level of loyalty: 0–10, 10–50, 50–80, and 80–100.

An autoencoder is employed in addition to the suggested passenger grouping model to minimise the dimensionality of passenger attributes and reduce algorithm complexity. To vectorise the social relations, a graph attention mechanism and a multigraph fusion mechanism are also used. To complete passenger grouping, a fusion of social relation vectors and feature dimensionality reduction outcomes is obtained. The experimental dataset is made up of data from Guangzhou–Shanghai travellers in 2020, which is then trained, tested, and matched to existing classification models including random forest, XGBoost, MTWSVMs, LightGBM, and DNN. According to the findings, the developed passenger grouping model, which adds social ties, beats the other models in terms of accuracy and adaption [17].

Data Availability

The data supporting the results of this study can be obtained from the railway department as reasonably required.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was supported by the Major Scientific and Technological Research Project of China National Railway Group Limited: N2021X034.

References

[1] F. Dobruszkes, C. Dehon, and M. Givoni, "Does European high-speed rail affect the current level of air services? An EU-wide analysis," *Transportation Research Part A: Policy and Practice*, vol. 69, no. nov, pp. 461–475, 2014.

[2] Y.-H. Cheng, "High-speed rail in Taiwan: new experience and issues for future development," *Transport Policy*, vol. 17, no. 2, pp. 51–63, 2010.

[3] X. Zhang, W. X. Luan, and Q. Cai, "Research on the competition between High?Speed rail and air transport," *Journal of Dalian University of Technolgy*, vol. 32, no. 1, p. 4246, 2011.

[4] P. B. Chou, E. Grossman, and D. Gunopulos, "Identifying prospective customer," in *Proceedings of the Six ACM SIGKDD International Conference on Knowledge Discovery and Date Mining*, pp. 447–456, ACM Press, Boston,MAsUSA, 2000.

[5] X.-bin An, L. I. Yi-jun, and Q. Ye, "Research on customer segmentation based on purchase behaviors," *Computer Integrated Manufacturing Systems*, vol. 11, no. 12, pp. 1769–1774, 2005.

[6] H. Rushmeier, R. Lawrence, and A. George, "Case study: visuaiizing customer segmentation produced by self organizing maps," in *Proceedings of the Visualization '97 (Cat. No. 97CB36155)*, pp. 463–466, Phoenix, AZ, USA, October 1997.

[7] B Qian, .. *Research on Revenue Management for Dedicated Passenger Line Based on Passenger Choice-Behavior*, Southwest Jiaotong University, Sichuan, China, 2014.

[8] L. Li, *Research on the Revenue Optimization of High-Speed Railway Basedon Passenger Behavior Analysis*, CHINA ACADEMY OF RAILWAY SCIENCES, Beijing, China, 2017.

[9] R. Zhang, M. A. Yu, and B.-ru Zhao, "Passenger choice behavior of high-speed rail and airline between beijing and Shanghai," *Journal of Transportation Systems Engineering and Information Technology*, vol. 000, no. 001, pp. 223–228, 2016.

[10] L. I. An-juna, D. Wang, and P. E. N. G. Qi-yuan, "Regional intercity railway planning method based on individual travel path," *Journal of Transportation Systems Engineering and Information Technology*, vol. 21, no. 2, pp. 30–36, 2021.

[11] C. Yu and H. Lv, "Research on passenger transport marketing strategy based on big data railway passenger portrait," *Journal of the China Railway Society*, vol. 42, no. 8, pp. 23–28, 2020.

[12] F. Yuan, L. Zhang, J. Shi, X. Xia, and G. Li, "Theories and applications of auto-encoder neural networks: a literature survey," *Chinese Journal of Computers*, vol. 42, no. 1, pp. 203–230, 2019.

[13] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graphattention networks," 2017, https://arxiv.org/abs/1710.10903.

[14] H. Li and Y. Zhu, "Xgboost algorithm optimization based on gradient distribution harmonized strategy," *Journal of Computer Applications*, vol. 40, no. 6, pp. 1633–1637, 2020.

[15] D. I. N. G. Shi-Fei, J. Zhang, X.-K. Zhang, and Y. X. An, "Survey on multi class twin support vector machines," *Journal of Software*, vol. 029, no. 001, pp. 89–108, 2018.

[16] G. Ke, Q. Meng, and T. Finley, *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 1–9, Long Beach, CA, USA, December 2017.