Research on Decision Model of Passengers' Travel Route Selection in Urban Rail Transit

Dongyu Kuang, Zhili Liu

MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University

MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University

16120825@bjtu.edu.cn

Abstract. Aiming at the problem of passenger path selection in normal rail transit, using the passenger travel path selection data in urban rail transit operation, analysing the influence factors of passenger path selection during peak hours, and a path selection decision modelling method using machine data for machine learning is proposed. The vector regression machine builds a data-based normal passenger path selection decision model.

1. Introduction

Passenger path selection problem is the core of the passenger flow distribution problem. At present, the discrete selection model based on random utility theory is widely used to implement passenger path selection modeling.

Considering that the data based on mobile positioning is more difficult to obtain and the application is less mature, this paper proposes to use the clearing path data as the data source for data model construction. The clearing route data records historical travel passengers' travel related information, mainly including the following several key data items: passenger number, entering time, entering station, leaving time, leaving the station, road network travel path, and driving time, number of transfers, and so on.

There are three main factors that have studied the impact of passengers on the path selection, including passenger characteristics, OD characteristics, and path characteristics. Among the passenger characteristics affecting the choice of routes on the road, the purpose of travel is the most critical one, considering that most passengers have similarities in the purpose of travel during peak hours. This paper mainly studies the passenger travel route selection decision model from two aspects: OD characteristics and path characteristics.

2. Passenger path selection decision data model

According to the meaning of the stochastic multi-attribute decision problem, the purpose of the passenger path selection decision modeling is to determine the selection probability of each path according to the multiple attribute values of multiple paths in a set for a multipath set under an OD pair. Therefore, the key to constructing a passenger path selection decision model is to determine the mapping relationship between each path attribute value in the path set and each path selection probability.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Therefore, in combination with the above analysis of the influencing factors for passenger travel route selection during peak hours, the input variables in the data model of passenger path selection decisions can be divided into two aspects: OD attributes and path attributes. The output variable is the path selection probability.

2.1. OD property

The OD attribute of an OD pair includes the travel time attribute and transfer reachability attribute of the route, using A to represent the OD attribute of an OD pair, expressed as:

$$A=\left(a\_time_{ij}, a\_tran_{ij}\right) \tag{1}$$

2.2. Path properties

The path attribute of an OD pair includes the set of all paths within the OD and the attributes of each path within the set. For rush hour passengers, through the analysis above, the two most important attributes affecting path selection are selected: the travel time attribute and the transfer number attribute, which represent the attributes of each route. Assuming that there are n paths in an OD pair, use road to represent the path attributes of an OD pair, expressed as:

$$road=\begin{bmatrix} r_{ij}^{1} \\ \cdots \\ r_{ij}^{s} \\ \cdots \\ r_{ij}^{n} \end{bmatrix}=\begin{bmatrix} r\_time_{ij}^{1} & r\_tran_{ij}^{1} \\ \cdots & \cdots \\ r\_time_{ij}^{s} & r\_time_{ij}^{s} \\ \cdots & \cdots \\ r\_time_{ij}^{n} & r\_time_{ij}^{n} \end{bmatrix}, s = 1,2,...,n \tag{2}$$

2.3. A Probability of selection of all paths within the OD

For an OD with n valid paths, using B to represent the probability of selection for all paths of an OD pair, expressed as:

$$B=\left(b_{ij}^{1},\cdots,b_{ij}^{s},\cdots,b_{ij}^{n}\right), s = 1,2,..., n \tag{3}$$

However, in the specific implementation, if the above-mentioned A, road, and B are directly used as input and output of the data model, the complexity of the problem is high and the efficiency of using the classic machine learning algorithm is low. The following is a detailed description:

The structure of the input data unit belongs to a more complex combination type. The top-level structure consists of A and roads. A is a vector form, and road is a matrix form. Some data items in the matrix are successive values, while others are discrete values. For the problem that the input data type is complex and the modeling is difficult, this paper adopts the method of input discrete and dimensionality reduction to solve it.

The number n of input data roads and output data B is determined by the number of valid paths in the OD pair. Different data units may have different values of n, that is, the input and output scales are not fixed, which increases the difficulty of algorithm design. For this reason, this paper adopts the model segmentation method to decompose the data and perform machine learning respectively.

3. Input data processing

A row in the road indicates the travel time and the transfer time of a path in the OD pair. Among them, the travel time is a continuous variable, and the number of transfers is a discrete variable. There are many possibilities for the distribution of the combination of the two. At this time, if the travel time is discretized according to a certain standard, converts continuous values to discrete points. Then, the combined distribution of the number of transfers and the number of transfers will be reduced, so that one-dimensional distribution can be used to represent the two-dimensional attribute combination distribution, and the input data can be reduced in dimension.

3.1. Discrete processing of travel time

Considering that passengers are not sensitive to time nuances in the actual decision-making process, and passengers often make simple comparisons of the time attributes of multiple paths in an OD pair when making decisions, it is a relative value embodiment. Therefore, a relative qualitative method is adopted to process the original travel time as three kinds of relative discrete values which are "shortest", "middle" and "longest".

3.2. Transfer times simplify processing

This step is similar to the one in the previous section. The purpose is to convert the number of transfer times from the original absolute value to the relative value with respect to the path set of the OD pair. Taking into account that passengers are more sensitive to the number of transfers, the relative values such as "least", "one time", "more than 1 time",..., "more than t times" are used to represent the original absolute value. Through statistics on the distribution of the number of path transfers in the data source, it is found that the difference between the number of passengers in the same OD pair alternate route is generally not more than two times, so the value of t is 2. The value of the original transfer number is handled as three discrete values of "least", "one time" and "more than 1 time" transfer times.

3.3. The two-dimensional attribute combination of the path is converted into a one-dimensional path feature

After the first two steps, the input is discretized, and the path set can be converted from road to road*.

There are 3 possible discrete values for the relative travel times and relative transfer times for any path. Then the combination of these two attributes will have 9 forms, each combination is defined as a path feature, road-new. The corresponding relationship is shown in Table 1.For example, if a route has the shortest relative travel time in its set road* and the least relative transfer times, the path feature of the route corresponds to 1.

Table 1. The correspondence of path feature attributes

Corresponding path features	Shortest	Middle	Longest
Least	1	4	7
One time	2	5	8
More than one time	3	6	9

By converting the two-dimensional attribute combination of the path into a one-dimensional path feature. Then, after two-dimensional to one-dimensional conversion of all the paths in road*, road* can be reduced from the original n×2 matrix to a n×1 matrix. The vector form can be expressed as follows:

$$\mathrm{road}^{*} = \left(\mathrm{roadnew}_{ij}^{1}, \cdots, \mathrm{roadnew}_{ij}^{s}, \cdots, \mathrm{roadnew}_{ij}^{n}\right), s = 1,2, \ldots, n \quad (4)$$

In the formula, $\mathrm{roadnew}_{ij}^{s}$ represents the path characteristic of the s path in the OD, and n represents the number of effective paths of the OD. In addition, road* may contain the same characteristics of the path, that is, these paths take the same value for both the travel time and the number of transfer times. For this case, you can use the build utility function and the MNL model to achieve the same feature path.

In this paper, the paths of the same features are merged into one item when processing, and the sum of the selection probabilities is taken as the selection probability of the path.

3.4. Model segmentation and data decomposition

When different OD pairs have different combinations of path numbers and types, machine learning algorithms face variable learning problems with input and output scales. Considering that the law of

path selection is not only related to the characteristics of a path itself, but also related to the characteristics of other paths in the set. In the set of paths where path features constitute different paths, the path selection rules are different. If the machine learning and model construction are performed separately after data decomposition according to the path set, it is equivalent to dividing the model into multiple sub-models, and the corresponding model may be separately called to perform the operation output according to the path combination characteristics.

According to the path selection decision data model proposed above, the original input is A and road, and the output is B. After the input is discretized and dimensioned, the input road is converted to road*and the output B is converted to B*. The input of the data model is A and road*, and the output is B*. After the model is divided and the data is decomposed, road* is used as the basis for data decomposition. The OD in each set of data has the same road*. It may not be used as an input. In this case, for each set of data, the input of the path selection decision model is A, and the output is road*.

In addition, because the travel time in the route attribute is expressed in a relatively qualitative manner as "shortest", "moderate", "longest" and the combination of this and the number of transfers is mapped to a path feature, this approach has This facilitates the division of fuzzy route features, but it also overwhelms the quantitative information of travel times of the route features. Therefore, the travel time difference of the path characteristics has an important influence on the selection probability, and this factor should be added to the input of the data model.

4. Path Selection Model Construction Based on Support Vector Regression

4.1. Support Vector Regression Machine Steps for Path Selection Decision Modeling

In this paper, support vector regression is used as a machine learning algorithm to establish the regression relationship between input and output in the path selection decision model. The modeling process mainly includes the following steps:

Step1: Division of sample data

Each set of learning data is sample data. The random sampling principle is used to divide the samples into training sample sets and test sample sets. The training set is used for the learning and discovery of the path selection law, and a regression model is established, the test set is used to test the established regression model and verify the accuracy and validity of the model.

Step2: Establish a regression model based on training data

Support vector regression was used to complete the learning of training data and a regression model was established. Since the path selection problem belongs to the multi-output learning problem, it is converted into a single output regression problem for each output component when performing machine learning, and then the regression results for multiple single outputs are normalized. This article uses the interface functions provided in the Lib-SVM toolkit developed by Chin-Jen Lin to implement a kernel learning algorithm that supports vector regression. Based on this, the following points need to be done:

Firstly, the selection of support vector regression types and kernel functions.

The constraint optimization conditions of different types of regression machines are slightly different, and the correct selection of kernel functions can improve the accuracy of the model. At present, there is no strict theoretical support for the choice of the two. Generally, based on empirical or experimental methods, this paper selects ε-SVR as the regression model after the experiment and selects the widely used RBF as the kernel function.

Secondly, determining the optimal parameter values for model parameters and nuclear parameters.

After determining the type of regression machine and kernel function, two parameters need to be set by the modeler: one is the error penalty parameter C in the regression model, and the other is the nuclear parameter γ in the RBF kernel function. When the values of the parameters are different, learning using the learning algorithm will produce different performance regression models.

In this paper, we use the grid search algorithm to achieve parameter optimization. This algorithm divides parameter C and parameter γ into grids in the value space according to a certain step, and uses

the kernel learning algorithm to train any value point in the grid. Regression model, then evaluate all the regression models, choose the parameters that make the model evaluation the highest as the optimal parameters. This paper adopts k-fold cross-validation method to complete the model evaluation. This method randomly divides the training data into k equal parts, cyclically selects k-1 of them as the training set, and the remaining one as the test set, which can be finally obtained through the cycle test. The MSE of the evaluation result of the regression model indicates that the smaller the MSE, the higher the model evaluation.

Thirdly, obtaining the optimal training regression model.

After determining the optimal parameter values, the support vector regression kernel learning algorithm is used to learn the training data, and a regression model is established between each attribute combination (independent variable) and the selected probability result (dependent variable).

Step3: Model evaluation based on test set

The independent variable in the test set is used as the input of the established regression model, and the estimated selection probability value of the regression model is obtained, and the estimated value of the model output is compared with the actual value in the test data. In this paper, three performance indexes of absolute error (AE), average absolute error (MAEM) and mean squared error (MSE) are selected as model evaluation criteria.

4.2. Modelling examples

Taking Wuhan rail transit as the background, using support vector regression machine for machine learning, a path selection decision model was established. Because this paper decomposes the data according to the path feature set road*, the path selection decision model is divided into multiple sub-models. The modeling process and method of each sub-model are the same. Therefore, the value of road* is (1,8). A set of samples is used as an example to describe the modeling process of the path selection decision model.

Step1: Sample data division

There are a total of 641 OD sample data with a value of (1, 8) for road* in the data source. They are randomly divided into a training set and a test set. There are 512 ODs in the training set and 129 ODs in the test set.

Step2: Regression model establishment

This paper chooses ε-SVR as the regression model, RBF function as the kernel function, and uses the training function svmtrain provided in the Lib-SVM toolbox to implement the kernel learning algorithm supporting vector regression. Considering that the output is multiple items, this article will perform regression on each output component to establish a regression model.

According to the training set data, a cross-validation method based on grid search was implemented by using Mat lab software. The error penalty parameters in the regression model and the nuclear parameters in the kernel function were optimized. After iteration, when the error penalty parameter is 1.6153 and the kernel parameter in the kernel function is 0.81115, the cross-validation MSE of the resulting regression model is the smallest, thus determining the optimal parameters of the regression model. Further, based on the optimal values of the regression model parameters, 512 ODs in the training set were learned using the training function svmtrain, and a regression model was formed.

The MAE of the regression data of the training data is 0.0610 and the MSE is 0.0053, which means that the deviation of the estimated value from the actual value is smaller on the training set, and the regression model has good expressiveness.

Step3: Evaluation of the regression model on the test set

Tests were performed using 120 ODs in the test set. The regression results for the test set had a MAE of 0.0610 and an MSE of 0.0074. In the regression results on the test set, the values of MAE and MSE are far less than 1, indicating that the established regression model has good generalization ability and estimation ability.

This completes the data learning for the road (1,8) sample group and establishes a path selection decision model. Using this model, the path selection probability can be calculated for any OD with

road* being (1,8). For example, "Han Zheng Road to Fu Xing Road" OD pair, the path feature set road* of the OD is (1,8), and the travel time difference R*=(1,0) of the path feature; the travel time of the OD is 29min, transfer can be The attainability is 1, then the OD attribute A=(29,1); by inputting A and R* into this model, the model will output the selection probability of two paths with different characteristics under the OD, which are respectively b1=0.84355; b2=0.12236.

5. Summary

The data-based model construction method for the urban rail transit passengers travel route needs to use the passenger's travel route results as a data source. Firstly, the passenger travel selection data in the operation of urban rail transit is analyzed, the source of learning data is determined, and passengers traveling in peak hours are analyzed to analyze the influence factors of such passengers' route selection, and the OD characteristics and path characteristics are determined as the path selection influence factors for such passengers; then, for the needs of machine learning, the influence factors are taken as model independent variables, and the path selection probability is taken as the model dependent variable to construct the passenger path selection decision data model. Taking into account the actual learning algorithm to achieve difficulties, the model of this paper adopts the solution of input discrete and dimensionality reduction, model segmentation and data decomposition, and finally forms the final input and output of machine learning. Finally, a support vector regression machine is used as a machine learning algorithm to construct a data-based normal path selection model and taking the rail transit in Wuhan as a modeling example to calculate.

References

[1] Ministry of Transport. Notice on Implementation of Safety Risk Assessment System for Highway Bridges and Tunnels at the Initial Design Stage [Z]. Beijing: Ministry of Transport, 2010.
[2] Li Yuan. Research on Urban Rail Transit Network for Passenger Flow Distribution[D]. Southwest Jiao tong University, 2017.
[3] Chen Tao. Study on dynamic distribution of urban rail transit passengers under uncertain conditions [D]. North University of China, 2017.
[4] LI Wei, XU Ruihua. Simulation model of passengers travel behaviour in subway network under emergencies[J]. Journal of East China Jiao tong University, 2015, 32(02):46-53.
[5] Hu Yongzheng. Research on analysis method of passengers travel behaviour of rail transit based on mobile signalling[D]. Southeast University, 2017.