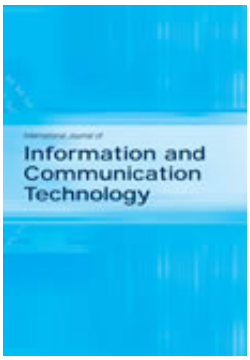


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Prediction method of tourism destination selection behaviour based on nearest neighbour decision tree

Qun Shang

DOI: 10.1504/IJICT.2022.10052406

Article History:
Received: 15 September 2021
Accepted: 17 November 2021
Published online: 05 December 2023

Prediction method of tourism destination selection behaviour based on nearest neighbour decision tree

Qun Shang

College of Culture and Tourism,
Jiangsu Vocational Institute of Commerce,
Nanjing, 211168, China
Email: qunsh@36haojie.com

Abstract: Aiming at the problems of large feature extraction error and poor prediction accuracy in tourism destination selection behaviour prediction method, a tourism destination selection behaviour prediction method based on nearest neighbour decision tree is proposed. With the help of bilinear function, the abstract tourism destination selection behaviour feature data is linearised, the freedom of linear feature data is limited, and the tourism feature is extracted through the scoring matrix. Set the characteristic data matrix, fix the characteristic data in a specific area, and determine the data weight through the cosine similarity algorithm. According to the nearest neighbour algorithm, the maximum attribute value of the selection behaviour data is determined, the tourism destination selection behaviour prediction decision tree is constructed, and the selection error is corrected with the help of the correction function to complete the behaviour prediction. The results show that the accuracy of the proposed method is 97%.

Keywords: nearest neighbour decision tree; tourist destinations; select behaviour prediction; correction function; maximum attribute value.

Reference to this paper should be made as follows: Shang, Q. (2024) 'Prediction method of tourism destination selection behaviour based on nearest neighbour decision tree', *Int. J. Information and Communication Technology*, Vol. 24, No. 1, pp.21–32.

Biographical notes: Qun Shang received her Masters degree in Human Geography from Shanghai Normal University in 2006. She is currently a Lecturer in the College of Culture and Tourism of Jiangsu Vocational Institute of Commerce. Her research interests include tourism enterprise management, hotel and homestay, rural tourism.

1 Introduction

With the rapid development of social economy, people's living standards have been continuously improved. Going out to play has become an important way for people's entertainment (Deng et al., 2021). Tourism has become a new way of leisure and vacation, which plays an important and positive role in relieving the pressure of urban workers and regulating their emotions. Tourism has also become an industry with high growth rate. The tourism industry continues to improve with the new forms of society. Among them, the choice of tourists' destination has become an important direction of

regional development (Cao et al., 2020). Tourism choice behaviour prediction refers to the behaviour of predicting the final decision-making behaviour according to the data of potential tourists, the satisfaction of tourism destinations or desired tourism destinations, and the preference of tourists. Tourists' choice of tourism destination is based on certain preferences, their own ideas and external publicity. In order to improve the recommendation of precision tourism, prediction and Research on tourists' choice of tourism destination has become a hot issue in this field (Claudia-Elena et al., 2020).

The paper proposes a method of forecasting the behaviour of route selection based on the passenger trust network, and predicts the choice of destination by the determination of route. Firstly, the characteristics of the flight tourists are analysed and the potential behaviours of the passengers are predicted effectively. By analysing the large-scale data in PNR data set, the trust network algorithm is introduced, and then the potential destination selection data is processed by system filtering method, and the similar data is deleted, and the remaining data is composed of the set of prediction, and the prediction of travel route selection behaviour is completed. The method can effectively improve the choice of tourist destination by forecasting the travel route. However, the data studied by this method is only for the current passengers who take flight, and consider less about other travel modes, which has some limitations (Feng et al., 2020). A method of travel selection behaviour prediction based on logit model with random coefficients is proposed in literature (Liu and Hao, 2019). This method firstly analyses the psychological state and economic situation of tourists, sets up three potential psychological factors, then uses logit model to deal with the fitting degree of the predicted data. Finally, the prediction of destination selection behaviour is completed by the influence of the random variables of tourists. This method fully considers the conditions of travellers, and provides an effective way to choose their travel capacity and destination. However, there are some deviation in the analysis process, which may lead to the problem of poor accuracy of prediction, and further consideration is still needed.

For to solve the shortcomings of the above methods, this paper proposes a tourism destination selection behaviour prediction method based on nearest neighbour decision tree. It provides a new method. The technical route of this paper is as follows:

1 With the help of bilinear function, the abstract tourism destination selection behaviour feature data is linearised, with the freedom of linear feature data is limited. The tourism destination selection behaviour feature is extracted through scoring matrix;

2 Set up the tourist behaviour characteristic data matrix, fix the behaviour characteristic data in a specific area and divide it into different levels, determine the weight of tourism destination selection behaviour characteristic data through cosine similarity algorithm, remove the characteristic data with high similarity, and complete the pre-processing of tourism destination selection behaviour characteristic data;

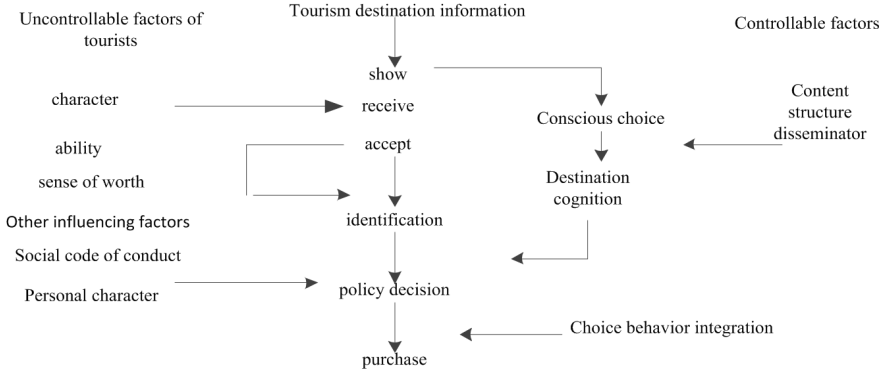
3 This paper analyses the basic operation principle of nearest neighbour decision tree, determines the maximum attribute value of selection behaviour data according to nearest neighbour algorithm, constructs the decision tree, and corrects the prediction error with the help of correction function to complete.

2 Feature data extraction and pre-processing of tourism destination selection behaviour

2.1 Feature data extraction of tourism destination selection behaviour

In order to realise the accurate prediction of tourism destination selection behaviour, it is necessary to obtain the characteristics of tourism destination selection behaviour, which can be used as the basis of tourists' behaviour decision-making, and can improve the prediction effect of tourism destination selection behaviour (Ren et al., 2021). There are many factors that affect tourists' choice in the extraction selection behaviour feature data. Among them, it includes the influence of key factors such as tourism cognition, tourism characteristics and destination characteristics. Tourism cognition mainly focuses on the impact of tourists' destination resources, facilities and services on their decision-making, including some individual active factors, which is a part of their cognition. Destination characteristic data reflects the characteristics and organisation form of tourism objectives (Trull et al., 2019). The environment, scenery and folk customs of the region have become the focus of its research. What ultimately affects tourists' decision-making depends on their own characteristics. The occurrence of tourists' destination selection behaviour is shown in Figure 1:

Figure 1 Occurrence process of tourists' destination selection behaviour



Through the analysis of the occurrence process of tourists' destination selection behaviour, it can be seen that the occurrence of tourism destination selection behaviour has certain characteristics (Priatmoko et al., 2021). Tourists' own characteristic data is an important characteristic data for the occurrence of tourists' choice behaviour. However, it is difficult to extract these abstract data in the feature extraction of tourism destination selection behaviour. Therefore, this paper standardises the abstract tourism destination selection behaviour feature data with the help of bilinear function. Set the bilinear function to:

$$L(a, b) = a^t V b \tag{1}$$

In formula, $L(a, b)$ represents a bilinear function , t represents the selected behaviour characteristic data differential coefficient, a, b represents a generalisation factor for selecting behavioural feature data of tourist destinations (Setiawan et al., 2020). The

bilinear function is nonlinear so that the acquired data is linear data to simplify the extraction of destination selection behaviour feature extraction, namely:

$$G = H\sum U^t CT \tag{2}$$

In formula, H represents constraints for linear data, U^tCT represents an orthogonal matrix.

When the transformed linear feature data has a certain degree of free change, in order to make the feature data show a stable trend, it is necessary to limit the degree of freedom of the data to obtain:

$$Z_m = H\mu\sum_{i=1}^{m} D^t \tag{3}$$

In formula, D^t represents a diagonal matrix, μ represents the degrees of freedom factor.

According to the linear data of tourism destination selection behaviour characteristics obtained above, the key features in the destination tourism selection behaviour characteristics data are extracted, and selection behaviour data set was set as:

$$S_{data} = (e, r, y) \tag{4}$$

Among them, $e = \{use_1, use_2, \ldots use_m\}$ select the basic behaviour data of m basic tourist destinations, $r = \{r_1, r_2, \ldots r_n\}$ represents the set of items of the n sets. Set the tourist destination selection behaviour characteristic score matrix to y. At this point, the determined tourist destination selection behaviour characteristic is:

$$y = e \times w \times D^t = Q\overline{w^2}^{1}\left(w^T\right) \tag{5}$$

In formula, w represents the orthogonal matrix of the scoring matrix, Q represents an eigensingular value.

In the feature extraction, this abstract tourism destination selection behaviour feature data is linearised with the help of bilinear function, the freedom of linear feature data is limited, and the feature extraction selection behaviour was completed by scoring matrix.

2.2 Data pre-processing of behaviour characteristics

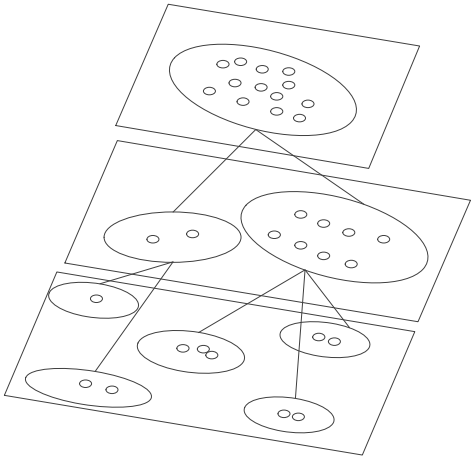
There were some uncertain factors in this above obtained data, which interfere with the prediction of its selection behaviour. Therefore, it is necessary to pre-process the behaviour characteristic data of tourism destination selection (Gomez-Oliva et al., 2019). Set the tourist behaviour characteristic data matrix as:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ & & \cdots & \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \tag{6}$$

In formula, P represents the adjacent matrix, p_{mn} represents the access value of the tourist destination feature interest.

In the pre-processing of the tourist destination selection behaviour data, the characteristic data in its matrix is set at different levels, as shown in Figure 2:

Figure 2 Division of different access levels of selected behaviour characteristic data



According to the selected behaviour feature access level determined above, the feature data with high similarity is effectively removed. The calculation formula of feature data with high similarity is as follows:

$$same(a,b) = k\sum_{a=1}^{n}\frac{\alpha^{a}}{\sum_{a=1}^{n}\alpha^{a}}\times\sum_{b=1}^{n}\frac{T_{ij}\times T_{nj}}{\left|T_{ij}\right|\times\left|T_{nj}\right|} \tag{7}$$

In formula, same (a, b) represents the similarity of the selected behavioural data, n represents the number of hierarchies of the characteristic data matrix. The lower the level of the selected behaviour feature data is, T_{ij} represents the greater the similar weight of the characteristic data in the matrix, T_{nj} represents similar weights for layer I. α^{a} represents the length of a continuous subsequence of similar data, k represents the time vector from which the feature is extracted (Kantsperger et al., 2019).

According to the similarity of the above selection behaviour feature data, the features with high similarity are deleted to obtain the final feature selection behaviour data as follows:

$$S = A(x) + \frac{M\sqrt{2\ln A(x)}}{e^{2}\cdot T_{m}} \tag{8}$$

In formula, $A(x)$ represents similar weights for the selected behavioural characteristic data, T_m represents the proportional factor for selecting the behavioural characteristic data.

In the pre-processing of behaviour feature data of tourism destination selection, set up the behaviour feature data matrix of tourists, fix the behaviour feature data in a specific area, and divide it into different levels. The weight of the behaviour feature data of tourism destination selection is determined by cosine similarity algorithm, and the feature data with high similarity degree is removed. The data pre-processing of behaviour characteristics of tourism destination selection is completed, which provides effective data for the subsequent prediction.

2.3 Prediction of tourism destination selection behaviour based on nearest neighbour decision tree

The nearest neighbour decision tree is a method to construct KD tree by k-nearest neighbour algorithm. It classifies the research data in different scale space, queries the data around the research sample data through the constructed KD tree, and then forecasts the surrounding data. The method can be used in many fields. In this paper, the algorithm was used to predict this behaviour. Decision tree is a multi-branch tree structure. The branch on each branch is determined by the attribute of the branch (Sugimoto et al., 2019). The data of the left node of the decision tree structure represents the set of tourists' behaviour selection data, and the right represents the attribute value of the current tourism behaviour selection data. When using the decision data to predict, the training sample data should be divided according to the attributes, and then divided according to certain levels until the leaf node disappeared. Leaves in decision tree represent a kind of label of tourism destination selection behaviour data. When forecasting, it needs to forecast from root node down, and finally, the final prediction research is completed according to the determined data attribute.

According to the above analysis, in the prediction of tourism destination selection behaviour in this paper, the behaviour data divided hierarchically on the decision tree and the surrounding data, that is, the properties of its nearest neighbour data, are classified, and it is determined that the predicted selection behaviour data follows a certain attribute division to obtain the maximum attribute value after classification (Hua and Wondirad, 2020). Suppose the selection behaviour data collection on the decision tree root node is N, behaviour data with M properties is present in this data collection. The attribute γ_i for each of the selection behaviour data has the maximum attribute value of n_i of the selection behaviour, that is:

$$\max S = O\frac{\gamma_i}{n_i}\sum_{i=1}^{m} g \tag{9}$$

In formula, $\frac{\gamma_i}{n_i}$ selects the subset values of the behavioural data partition, select the subset values of the behavioural data partition, g represents multiple selection behaviour elements that exist, O represents the classified near-neighbour node selection behaviour data.

According to the maximum attribute value of the tourism destination selection behaviour data obtained above, due to the certain error of the determined large attribute value, it is necessary to correct it to provide more effective data for subsequent prediction. It is corrected according to the nearest neighbour algorithm to obtain:

$$(t) = \sum_{u=1}^{u} \frac{\delta(t)}{\vartheta^u} r(a, b) \tag{10}$$

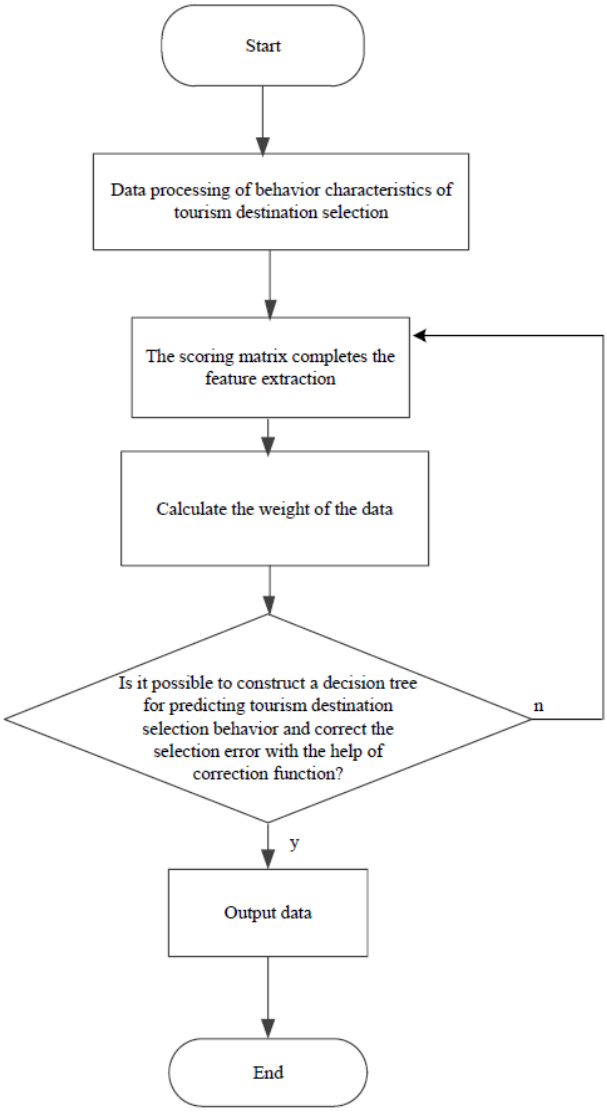
In formula, $\tau(t)$ represents the corrected selected behaviour data attribute values, $\delta(t)$ represents the weight threshold for the selected behaviour data, ϑ^u represents the data set factor.

According to the maximum attribute value of the corrected behaviour data of tourist destination selection, a decision tree of tourist destination selection behaviour prediction is constructed (Lekovi et al., 2020). The data set of the tourist destination selection

behaviour is assumed as a decision tree and its predicted data blade of φ_i, decision tree branches The number of tree branches is ω_j, At this time, before performing the tourist destination selection behaviour, the decision tree needs to be built to a certain extent, and the decision tree for obtaining the trimmed tourist destination selection behaviour data prediction is as follows:

$$\aleph_i = \frac{\left[\omega + \frac{1}{m}\right]}{\varphi_i} \tag{11}$$

Figure 3 Prediction process of tourism destination selection behaviour based on nearest neighbour decision tree



According to the decision tree of tourism destination selection behaviour data prediction constructed, the error in the prediction process needs to be corrected by using the correction function, and the final correction result is the result of tourism destination selection behaviour prediction:

$$^{\circ}F_i = I\frac{\sum \aleph_i + \frac{1}{m}}{\sum \aleph_i + \frac{1}{m}(\sigma)} \tag{12}$$

In formula, $^{\circ}F_i$ represents the results of tourist selection of tourist destinations, σ selects behavioural interference items on behalf of the tourist destination, I represents the correction function in the prediction (Huang et al., 2019).The prediction process of tourism destination selection behaviour based on nearest neighbour decision tree is shown in Figure 3:

The basic operation principle of nearest neighbour decision tree is analysed, the maximum attribute value of selection behaviour data is determined according to the nearest neighbour algorithm, the decision tree of tourism destination selection behaviour prediction is constructed, and the prediction error is corrected with the help of correction function to complete the prediction of tourism destination selection behaviour.

3 Experimental analysis

3.1 Experimental scheme design

In order to verify that the prediction method proposed in this paper can effectively predict the behaviour of tourism destination selection, an experimental analysis is carried out. In the experiment, 20 people from three different age groups were selected for the study. Among them, 10 men were 18–35 years old, 10 were women, 10 were men in 36–45 years old and 10 females. The results showed that the number of male and female in the 46–60 years old period was 10, and the behaviour of destination selection was predicted. Among them, 90% of the subjects in the 18–35 years old stage have achieved economic independence, but their income is less; the subjects in 36–45 years old stage were the group with higher income in three stages, and the income of the subjects in 46–60 years old stage was relatively stable. A total of 30 specific destinations are set up to serve as the predicted destination data for the selected behaviour of the subjects. Among them, 15 scenic spots are relatively perfect tourism development areas, and the rest are the tourist attractions to be developed with low visibility. The results were collected by setting up questionnaires, and the recovery rate of the questionnaire was up to 95%, which was used as the data for the analysis of the experimental results.

3.2 Experimental index design

On the basis of the above experimental scheme design, the experiment is carried out by comparing the methods in this paper, the travel route selection behaviour prediction method based on passenger trust network and the travel selection behaviour prediction method based on random coefficient logit model. The set indicators are the feature

extraction error of tourism destination selection behaviour and the prediction accuracy of tourism destination selection behaviour. Among them, the tourism destination selection behaviour feature extraction error refers to the effectiveness of the subject's selection behaviour feature extraction. The lower the error, the better the representative performance and the better the later prediction effect. The prediction accuracy of tourism destination choice behaviour refers to the final decision accuracy of choice. The closer it is to 100%, the better the prediction performance. In order to ensure the accuracy of the experiment, the data related to the predicted behaviour of the questionnaire will be effectively processed by SPSS software to improve the accuracy of the data.

3.3 Analysis of experimental results

3.3.1 Error analysis of behaviour feature extraction of tourism destination selection by different methods

The experiment analyses the error of the prediction method in this paper, the prediction method of tourism route selection behaviour based on passenger trust network and the prediction method of travel selection behaviour based on random coefficient logit model in extracting the characteristics of tourism destination selection behaviour of sample tourism data. The results are shown in Table 1:

Table 1 Extraction error of behavioural features of tourism destination selection by different methods (%)

Extraction/times	*Paper method*	*Prediction method of travel route selection behaviour based on passenger trust network*	*Travel choice behaviour prediction method based on random coefficient logit model*
10	1.2	2.5	3.2
20	1.3	2.8	3.4
30	1.2	2.7	3.4
40	1.4	2.6	3.6
50	1.2	2.8	3.8
60	1.3	3.0	4.0
70	1.4	3.1	4.1
80	1.3	3.2	4.1
90	1.2	3.1	4.2
100	1.1	3.0	4.3

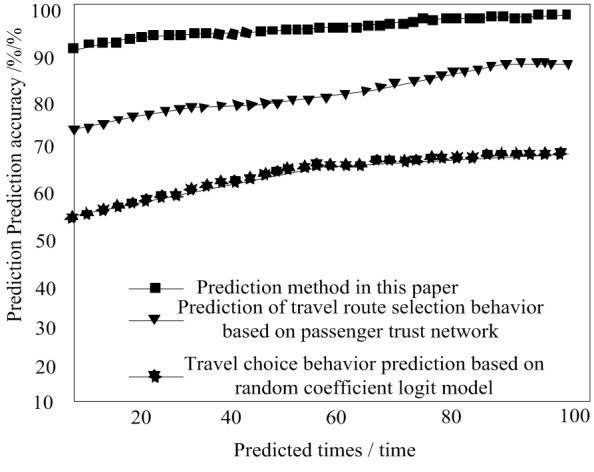
By analysing the experimental results in Table 1, it can be seen that there are some differences in the extraction error of tourism destination selection behaviour features using the sample tourism data. Among them, the minimum extraction error of the tourism destination selection behaviour feature of the sample tourism data by this method is about 1.1%, and the extraction error changes little. The minimum extraction error of the tourism route selection behaviour feature of the sample tourism data by the tourism route selection behaviour prediction method based on the passenger trust network is about 2.5%, and there are some fluctuations in the extraction process; The travel choice behaviour prediction method based on random coefficient logit model has the lowest

extraction error of tourism destination choice behaviour features of sample tourism data, which is about 3.2%, and its fluctuation is greater than the other two methods in the experiment.

3.3.2 Analysis on prediction accuracy of tourism destination selection behaviour by different methods

Based on the above feature extraction of tourism destination selection behaviour, the experiment further analyses the prediction method in this paper, the tourism route selection behaviour prediction method based on passenger trust network and the travel selection behaviour prediction method based on random coefficient logit model. The prediction accuracy of sample tourism destination selection behaviour is analysed. The results are shown in Figure 4.

Figure 4 Comparison of prediction accuracy of different methods for tourism destination selection behaviour



The experimental results in Figure 4 show that with the change of prediction times, the accuracy of the three methods for the prediction of sample tourism destination selection behaviour is always rising. When the iteration times are 40, the prediction accuracy of the method is about 92%, the prediction accuracy of the travel route selection behaviour based on the passenger trust network is about 81%, and the prediction accuracy of the travel choice behaviour based on the random coefficient logit model is about 58%; when the iteration number is 80, the prediction accuracy of the method is about 96%, the prediction accuracy of the sample tourism destination selection behaviour based on the passenger trust network is about 85%, and the prediction accuracy of the travel choice behaviour based on the random coefficient logit model is about 62%; when the iteration times are 100, the prediction accuracy of the method is about 97%, the prediction accuracy of the travel route selection behaviour based on the passenger trust network is about 86%, and the prediction accuracy of the travel choice behaviour based on the random coefficient logit model is about 62%; compared with the other two methods, the accuracy of this method is much higher than that of the other two methods.

4 Conclusions

In order to promote the rapid development of tourism market economy, a prediction method of tourism destination selection behaviour based on nearest neighbour decision tree is proposed. Firstly, the method extracts the characteristics of tourism destination selection behaviour, with calculates the weight according to the obtained characteristic data. On this basis, the maximum attribute of behaviour data is determined with the help of nearest neighbour decision tree, the decision tree of selection behaviour prediction is constructed, and the prediction error is corrected with the help of correction function to complete the prediction of tourism destination selection behaviour. It is found that this method has the following advantages:

1 Using this method to extract the characteristics of tourism destination selection behaviour of sample tourism data, the minimum error is about 1.1%, which has certain reliability.

2 Using this method to predict the tourism destination selection behaviour of sample tourism data, the highest accuracy is about 97%, which shows the effectiveness of this method.

Acknowledgements

This work was supported by Excellent Innovation Team of Philosophy and Social Science in Jiangsu Universities — Rural Tourism Development Research Team.

References

Cao, M.Q., Liang, J. and Li, M.Z. (2020) 'TDIVis: visual analysis of tourism destination images', *Frontiers of Information Technology & Electronic Engineering*, Vol. 21, No. 4, pp.536–557.

Claudia-Elena, Ț., Diana-Maria, V. and Carmen-Eugenia, N. (2020) 'The role of social media in health safety evaluation of a tourism destination throughout the travel planning process', *Sustainability*, Vol. 12, No. 2, pp.14–20.

Deng, B., Xu, J. and Wei, X. (2021) 'Tourism destination preference prediction based on edge computing', *Mobile Information Systems*, No. 1, pp.1–11.

Feng, X., Zhang, C. and Lu, M. (2020) 'Prediction of route selection behaviour based on passenger trust-network', *Modern Electronics Technique*, Vol. 43, No. 4, pp.78–82, p.86.

Gomez-Oliva, A., Alvarado-Uribe, J., Parra-Meroo, M.C. et al. (2019) 'Transforming communication channels to the co-creation and diffusion of intangible heritage in smart tourism destination: creation and testing in Ceutí (Spain)', *Sustainability*, Vol. 11, No. 2, pp.14–20.

Hua, H. and Wondirad, A. (2020) 'Tourism network in urban agglomerated destinations: implications for sustainable tourism destination development through a critical literature review', *Sustainability*, Vol. 13, No. 14, pp.23–28.

Huang, B., Saaty, T.L. and Li, Y. (2019) Collaborative R&D and pricing policy of supply chain under the selection behaviour of heterogeneous customer', *Mathematical Problems in Engineering*, Vol. 45, No. 5, pp.1–9.

Kantsperger, M., Thees, H. and Eckert, C. (2019) 'Local participation in tourism development — roles of non-tourism related residents of the Alpine Destination Bad Reichenhall', *Sustainability*, Vol. 11, No. 31, pp.743–756.

Lekovi, K., Tomi, S. and Mari, D. (2020) 'Cognitive component of the image of a rural tourism destination as a sustainable development potential', *Sustainability*, Vol. 12, No. 22, p.9413.

Liu, J-R. and Hao, X-N. (2019) 'Travel mode choice in city based on random parameters logit model', *Journal of Transportation Systems Engineering and Information Technology*, Vol. 19, No. 5, pp.108–113.

Priatmoko, S., Kabil, M. and László, V. (2021) 'Reviving an unpopular tourism destination through the placemaking approach: case study of Ngawen Temple, Indonesia', *Sustainability*, Vol. 13, No. 1, pp.15–21.

Ren, X., Li, Y. and Zhao, J.J. (2021) 'Tourism growth prediction based on deep learning approach', *Complexity*, Vol. 41, No. 15, pp.1–10.

Setiawan, B., Arief, M., Hamsal, M. et al. (2020) 'Local's perspective of community participation in Lake Toba as a tourism destination', *Solid State Technology*, Vol. 63, No. 4, pp.1021–1038.

Sugimoto, K., Ota, K. and Suzuki, S. (2019) 'Visitor mobility and spatial structure in a local urban tourism destination: GPS tracking and network analysis', *Sustainability*, Vol. 11, No. 3, pp.56–62.

Trull, O., Peiró-Signes, A. and García-Díaz, J.C. (2019) 'Electricity forecasting improvement in a destination using tourism indicators', *Sustainability*, Vol. 11, No. 20, pp.45–52.