The Prediction Method of Regional Economic Development Potential Along Railway Based on Data Mining

Hui-fang Guo[1(✉)] and Qing-mei Cao[2]

[1] School of Transportation and Municipal Engineering, Inner Mongolia Vocational and Technical College of Architecture, Hohhot 010010, China

[2] Department of Computer Technology and Information Management, Vocational and Technical College of Inner Mongolia Agricultural University, Baotou 014100, China

Abstract. The prediction accuracy and efficiency of regional economic development potential along the railway line are poor when the common methods are used to predict the economic development potential along the railway. In view of this problem, a data mining based prediction method for regional economic development potential along the railway line is designed. Select the influencing factors of regional economic development along the railway, construct the regional economic development index system, collect index data, clean, cluster, sort and standardize the index data, extract the economic development characteristics in the pre-processing index data through data mining, input the economic development characteristics to the neural network, and output the predicted value of economic development potential after training and learning, Divide the economic development potential level. The experimental results show that the design method reduces the deviation of economic development potential prediction, shortens the prediction time, improves the accuracy and efficiency of prediction.

Keywords: Data mining · Economic development potential · Data preprocessing · Economic forecast

1 Introduction

High speed railway is a modern system engineering with high technical difficulty and complexity. It is a public transportation tool with huge carrying capacity. With the construction of high-speed railway, the number of economic zones along the railway increases. Therefore, this paper studies the prediction method of regional economic development potential along the railway, puts forward reasonable decision according to the regional economic trend along the railway, and realizes the optimization of economic decision, It is of great significance.

Foreign research on regional economic development potential prediction is relatively mature. For the study area, the URL address table related to economic development potential is maintained, and the URL address is stored. When the URL address already

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 417, pp. 221–238, 2022.
https://doi.org/10.1007/978-3-030-94554-1_19

exists, it is not added to the address list, otherwise it is added to the list, so as to eliminate the URL address of the Web page and get the characterization information of regional economic development potential, On the basis of completeness and scientificity, the forecast value of regional economic development potential is calculated. The domestic regional economic development potential prediction research has also made great progress. It analyzes the regional engineering cost and land cost, including technical economy, technical design, land acquisition, etc., determines the percentage of the cost in the total project investment, and counts the regional infrastructure construction cost, including all costs related to terrain preparation and platform construction, The amount of these costs varies greatly due to the different terrain characteristics. When encountering the geographical obstacles with special technical difficulties, it is necessary to build the viaduct, bridge and tunnel in the region. In this case, the infrastructure construction cost is doubled. Combined with the engineering cost, land cost and infrastructure construction cost, the regional economic development potential is predicted.

Combined with the above theories, data mining technology is introduced into the prediction of regional economic development potential along the railway. Data mining technology can convert a large number of regional economic development potential data along the railway into useful information and knowledge. Based on this, it can further divide the level of regional economic development potential along the railway and optimize the prediction performance of regional economic development potential along the railway.

2 Design of Economic Development Potential Along Railway Based on Data Mining

2.1 Extract the Influencing Factors of Regional Economic Development Along the Railway

Combined with the cost-effectiveness of the railway, the influencing factors of regional economic development along the railway are extracted. The first type of cost-benefit is infrastructure construction cost. The construction cost of new high-speed railway line is as follows: superstructure cost, which is the specific elements of railway, including guide rail, side rail along the line, signal system, catenary, electrification mechanism, communication and safety facilities. These elements usually account for 5–10% of the total investment respectively. The second type of cost-effectiveness is operation cost. The operation cost of high-speed railway service is as follows: infrastructure operation cost, including labor, energy and other material consumption required for daily operation and maintenance of guideway, wharf, station, energy supply, signal system, traffic management and safety system; Daily operation fixed cost and variable cost include train maintenance cost, electric traction device and catenary maintenance cost, power system maintenance cost, signal system maintenance cost, telecommunication system maintenance cost and other costs, of which track maintenance cost accounts for about 50%; Vehicle operation cost includes dispatching and train operation cost, vehicle and equipment cost, energy cost, sales cost and management cost. Vehicle operation cost mainly includes all labor costs such as vehicle service and driving, while vehicle and facilities mainly refer to the cost of vehicle and equipment depreciation and maintenance.

When analyzing the railway benefits, it is assumed that the service life of high-speed railway is A year, the completion time of high-speed railway infrastructure and superstructure is initial time $t = 0$, and the high-speed railway operators purchase high-speed rolling stock at the initial time. In the A time region, assuming that the ticket price C and annual passenger volume D are constant, and the project investment cost is E, including the infrastructure construction cost and the present value of rolling stock, under the condition of passenger volume D, the fixed cost remains unchanged, only the variable cost is considered, and all costs are calculated as opportunity cost, then the conditional expression of positive net present value is obtained as follows:

$$[F(D) - G(D)]e^{-(C-t)}dt - \int_0^A He^{-Ct}dt \geq E \tag{1}$$

Among them, $F(D)$ is the annual social benefit of the project, $G(D)$ is the annual operation and maintenance cost based on D, and H is the investment cost. It can be seen from formula (1) that during the operation of high-speed railway, the operation and maintenance cost, labor cost and energy consumption of track, station, signal and other real estate will be generated.

It can be seen from the cost-effectiveness of the railway that the stability of the technical safety level of the high-speed railway affects the passenger choice and the change of the traffic volume level. Under certain safety technical conditions, the reasons that affect the regional economic development along the high-speed railway are the length of the high-speed railway, the level of economic development, urban population, visitors, etc. The influence of these factors on the regional economic development along the line is dynamic and relative. In the early stage of the development of high-speed railway, every new railway section can open up the high-speed railway passenger transport market. At this time, the length of high-speed railway becomes the bottleneck of restricting the high-speed railway passenger transport volume. In the state of high-speed railway development has been saturated, the extension of high-speed railway line may not bring significant increase in passenger transport volume. On the other hand, the purpose of passengers taking high-speed railway mainly includes school, commuting, business trip, travel and other private activities. These activities are affected by the level of regional economic development to varying degrees, and have different influence structures in different regions. The size of the target market ultimately depends on the number of people with the ability to pay, who may come from their own countries or may be international tourists, The degree of its influence is directly related to the structural distribution ratio and the regional economic structure. So far, the extraction of influencing factors of regional economic development along the railway is completed.

2.2 Collection of Regional Economic Development Indicators Data Along the Railway

To screen the influencing factors of regional economic development along the railway, build the index system of regional economic development along the railway, and collect the index data. Taking the county as the main research unit, combined with the relevant indicators of county economic development and the availability of data, this paper selects

per capita GDP, per capita gross industrial product, per capita investment in fixed assets, per capita fiscal revenue, per capita savings deposit balance, per capita total retail sales of social consumer goods, per capita net income of rural residents, the proportion of fiscal revenue in GDP, per capita net income of rural residents, per capita net income of rural residents, per capita total retail sales of social consumer goods This paper discusses the economic development in the railway radiation belt from nine factors such as the speed of economic development. The comprehensive development level of county nodes is represented by comprehensive quality index and measured by composite index. From the three levels of economic development, social development and urban construction, 14 indexes are selected to construct the index system of county comprehensive development level. The maximum method is used to standardize the data, and with the help of SPSS software, the county comprehensive development level is preliminarily measured. The index system is shown in Table 1.

Table 1. County economic development potential index system

Target layer	Code layer	Index layer
Development potential of county economy	Economic development	Gross Regional Product
		Per capita GDP
		Proportion of output value of secondary and tertiary industries
		Public revenue
		Investment in fixed assets
		Balance of household savings deposits
	Social development	Number of employees
		Total retail sales of consumer goods
		Urban per capita disposable income
		Business volume of Posts and Telecommunications
		Number of beds in health institutions
	Urban construction	Per capita park green area
		Road area per capita
		Green coverage rate of built up area

The basic data are from the regional statistical yearbook along the line, namely, China Urban Statistical Yearbook in 2020. It is difficult to obtain the relevant logistics data in the county, and the freight capacity is usually measured by the freight volume index. The proportion of freight shifts in County contact is less, and the travel time and passenger flow between counties can reflect the regional accessibility level more. Therefore, passenger transport is used as the contact medium. Considering that railway transportation is suitable for long-distance transportation in County, the railway passenger flow data has the characteristics of high accuracy and easy access, which can better reflect the reality of economic contact between counties. The shortest travel time of county railway passenger transport is selected as the basic data for calculating accessibility and economic connection. The data is from Ctrip. During the query process, the shortest travel time of all direct passenger trains between the two counties is selected. If there is no direct train between two cities, the website will automatically provide the shortest transit route for it, only the distance time will be counted, Transfer time and stop time of transfer are not considered.

Besides the collection source of statistical yearbook, the index data can also be collected by crawler. Taking the regional economic development along the railway as the key word, search the relevant Web pages of development potential indicators in the Internet, and join the crawler queue from the seed UPL, analyze and download the Web page, grab the UPL and get a new URL. In the real network traffic data, statistics the usage heat of various indicators, read the page on the homepage of the Web page, find other link addresses in the Web page, find the next page, set the number of access layers of different pages until all pages of the website are captured [1]. The grab page is preprocessed, the economic development of the railway area is taken as the theme content, the pages that are inconsistent with the theme are filtered out, the table label is used to repair the wrong or irregular tags, the repaired pages are stored in the HTML document, and the HTML file is selected as the root node, that is, an Internet unit, the tag tree is constructed, and the visual information of the Web page is used, Block processing Web pages, starting from the prediction demand of regional economic development along the railway, remove redundant information of Web pages, link useful information together, find text files related to the subject content, mark hypertext, and integrate Web pages [2–4]. Finally, through HTTP protocol, the browser can download the Web page, grab the effective information in the Web page, including the documents such as sound, text and image, obtain the index data of regional economic development field along the railway line, add the relevant contents of regional economic development along the railway line by custom, and collect the video, audio, database, picture, text data and other types of data in the Web page, Eliminate new URL, join new crawl queue, loop above operation. So far, the data collection of regional economic development indicators along the railway has been completed.

2.3 Preprocessing the Data of Regional Economic Development Indicators Along the Railway

According to the massive data of regional economic development indicators along the railway, preprocessing is carried out to make the indicator data accurately express the potential of regional economic development.

Cluster Processing Index Data

The distributed k-means algorithm is used to cluster the related regional economic development index data along the railway. Firstly, the original historical data of county economic development potential indicators are cleaned, and the wrong data, data noise and invalid data are processed. Then, the time series of regional economic development data along the railway are counted in the historical data, and the time series are classified and processed according to the quarterly order, so as to keep the regional economic development index data continuous, and then fill in the missing historical data. Through attribute mapping, the character data of the original data set is transformed into the numerical standardized data. The mapping formula is as follows:

$$\mathrm{m} = \left[\frac{(\mathrm{m}_{\max} - \mathrm{m}_{\min})(\mathrm{n} - \mathrm{n}_{\min})}{(\mathrm{n}_{\max} - \mathrm{n}_{\min})} \right] + \mathrm{m}_{\min} \tag{2}$$

Among them, m is the standardized index data after processing, $\mathrm{m}_{\max}$ and $\mathrm{m}_{\min}$ are the maximum and minimum values of the index data after processing, n is the original historical data of the regional economic development index along the railway, and $\mathrm{n}_{\max}$ and $\mathrm{n}_{\min}$ are the maximum and minimum values of the original data respectively. In the data set, k data objects are randomly selected as the initial cluster center of the regional economic development index data along the railway. By using horse distance, the initial cluster center and the remaining data objects are compared. The calculation formula of horse distance is as follows:

$$H_{ij} = \sum_{k=1}^{n} \frac{\left| x_{ik} - x_{jk} \right|}{x_{ik} + x_{jk}} \tag{3}$$

Among them, n is the original dimension of the indicator data, x_{ik} is the dimension value of the i-th and j-th indicator data of the indicator data object in the k-th data dimension, and H_{ij} is the Mahalanobis distance between the indicator data i and the indicator data j [5]. The closer the Mahalanobis distance H_{ij} is to 1 or -1, the higher the correlation degree and the closer the distance between the two index data are judged. If H_{ij} is closer to 0, the lower the correlation degree and the farther the distance are judged. The remaining index data objects are classified into the nearest initial cluster center, and then the cluster center is re selected, iterating for many times until the criterion function converges, and the k cluster centers do not change. The definition formula of criterion function J is as follows:

$$J = \sum_{r=1}^{k} |Z_r - E_r| \tag{4}$$

Among them, Z_r is the center value of cluster center r, and E_r is the average value of Z_r. After cleaning the clustered index data, delete the irrelevant records of regional economic development along the railway, including image content request, file request and spider crawler program request. When the spider initiates HTTP request, it separates the illogical spider session and records a large number of regional economic development information along the railway through HTTP header. After the index data is partitioned and clustered, the collected data is classified finely to get different local data tuples [6–8]. The refined data items after segmentation are shown in Table 2.

Table 2. Data items of county economic development potential index

Field code	Field type	Meaning
MCI	VarChar2	Gross Regional Product
MC	VarChar	Per capita GDP
SJY	VarChar	Proportion of output value of secondary and tertiary industries
SJYLX	VarChar	Public revenue
SJFL	VarChar	Investment in fixed assets
TXTZ	VarChar2	Balance of household savings deposits
ZXZBX	Date	Number of employees
ZXZBY	Date	Total retail sales of consumer goods
YSZBX	VarChar2	Urban per capita disposable income
YSZBY	Number	Business volume of Posts and Telecommunications
BLC	Number	Number of beds in health institutions
BB	Number	Per capita park green area
SYQX	VarChar2	Road area per capita
SJYSM	VarChar2	Green coverage rate of built up area

Build a distributed SQL database to represent the attribute structure of data items, and provide data support for the prediction of regional economic development potential along the railway through various refined data sets. So far, the clustering of index data is completed.

Primary and Secondary Relationship of Sort Index Data

This paper uses sprint classification algorithm to sort the primary and secondary relationship of regional economic development index data along the railway. The maximum minimum normalization formula is used to discretize the continuous numerical attribute of the index data, and the linear transformation of the index data is carried out.

$$V = \beta \frac{(L - M)}{(M - N)} \tag{5}$$

Among them, L is the index data value, M and N are the maximum and minimum values of the same attribute index data, β is the mapping interval, and V is the value after

the indicator data is mapped. The neural network center is used to replace the continuous value of index data, and the data attributes are converted to discrete values, and the rules of regularity are displayed on the basis of relative attributes, and the number of values of the same attribute data is reduced [9, 10]. By using sprint classification algorithm, the paper sorts the primary and secondary relations of regional economic development index data along the railway, classifies the regional economic development level along the railway line, divides the areas along the line into multiple sub groups, takes the regions along the line with different development levels as different categories, distinguishes the economic development level of the regions along the line, and the economic development level of the same subdivision area is close to each other. The classification of index data is realized by decision tree. The attribute with the highest priority is selected in the index data. As the root, the pre processed attribute set is provided. The common character is searched from the index data, a series of sorting decisions are made, and the decision tree nodes are divided, and then the attribute data attributes are split, and the attributes are accurately associated with the sub nodes, and the attribute value segmentation data set is obtained [11]. Set the number of data set category as m, and the number of data set categories is the number of leaf node categories. Then the calculation formula of split parameter F is:

$$F = 1 - \sum_{I}^{\mathrm{m}} \mathrm{p}_I^2 \tag{6}$$

Where p_I is the relative frequency of data set category I. A data node is selected in the data set to represent the data of the data set. The logical judgment of the regional economic development level along the line is taken as the internal node of the decision tree, the branch result of the logical judgment is taken as the edge of the decision tree, and the data attribute is associated with the root node of the decision tree to construct the multi tree decision tree. When all the index data belong to the same class, the class label is used to define the leaf node. When the index data do not belong to the same class, the data attribute is measured according to the information entropy, and the data in the original attribute set is deleted. When the candidate set is empty, the leaf node is returned and marked as a common category [12–14]. For different types of index data, the calculation formula of information entropy Q is as follows:

$$Q = -\sum_{I=1}^{\mathrm{m}} \frac{|C_I|}{|\xi|} \log_2\left(\frac{|C_I|}{|\xi|}\right) \tag{7}$$

Where ξ is the given data set of decision tree and C_I is the set of data set belonging to class I objects. The data set ξ is classified according to the attribute characteristics to obtain multiple different objects. The weighted sum of the information entropy Q is carried out through the partition entropy to calculate the information gain attribute of the index data

$$A = -\sum_{I=1}^{\phi} \left(\frac{|\zeta|}{|D|} \times Q\right) \tag{8}$$

Among them, A is the information gain of index data, ϕ is the number of attribute features of data set, D is the amount of information needed for data set classification, and ζ is the amount of information needed for data set partition. In the attribute set, the attribute with the highest information gain A is selected, and the leaf node is marked to get a score of the attribute with the highest information gain, and the elements of the dataset subset can meet the score. When the categories of nodes are the same, the remaining attributes can not be divided again, or the given score has no data, the class label is created, the division of decision tree is terminated, and the classification of regional economic development level along the line is completed. So far, the sorting of the primary and secondary relationship of the index data is completed, and the preprocessing of the regional economic development index data along the railway is completed.

2.4 Mining the Characteristics of Regional Economic Development Along the Railway

Through data mining, in the preprocessed regional economic development index data along the railway, mining the characteristics of regional economic development. This paper defines the semantic attention of index data, highlights the index data of regional economic development along the railway with similar semantics, and applies it to the prediction of regional economic development potential along the railway. The factors influencing semantic distance between data items are selected, which are economic development, social development and urban construction. The three factors are taken as the dynamic characteristics of the index data, so that the index data changes with time, showing different characteristics of regional economic development along the railway. Three kinds of dynamic features are combined to get multidimensional data information. Data exploration is used to reduce the dimension of index data. According to the dimension combination obtained from data exploration, multidimensional dynamic features are transformed into 2D dynamic features [15, 16]. If the characteristic dimension of index data is set as j, the formula of index data combination exploration condition G is as follows:

$$G = j \times (j-1) \times (j-2) \cdots \times 1 \tag{9}$$

In the dynamic features of two-dimensional display, the data items with mutual reference and link relationship are selected to judge the semantic distance between the data items is close, which is more meaningful for the prediction of regional economic development potential along the railway. Abstract features of index data in information space are extracted, and the types of abstract features are divided into three categories: time series, network and hierarchy. The calculation parameters of semantic distance are determined by using the structural relationship among the three categories of index data. Suppose that the dynamic characteristic object of the regional economic development index along the railway is s, and any index data object in the information space is x, then the semantic distance $d(s, x)$ between s and x is:

$$d(s,x) = w_1 f(s,x) + w_2 g(s,x) + w_3 l(s,x) \tag{10}$$

Among them, $f(s, x)$ is x 2-dimensional display of s dynamic feature dimension combination. The 2-dimensional display includes two parts: implicit intention and explicit

intention. The implicit intention is used to determine the impact of index data on regional economic development potential prediction. The explicit intention is used to clarify the regional economic development potential prediction intention. $g(s, x)$ is the structural relationship of abstract feature of index data, which represents the correlation between x and s, $l(s, x)$ is the center distance after semantic representation of x and s, and w_1, w_2 and w_3 are the weights of 2-dimensional presentation, structural relationship and center distance respectively.

The semantic distance is regarded as an important parameter of the semantic attention degree of the indicator data. Through $d(s, x)$, the distance between x and s at the semantic level is reflected. In the prediction of regional economic development potential, the prior importance of different data items of the indicator data is determined, and the threshold of semantic attention degree is set to limit the set of indicator data items. The semantic attention H of index data item x to s is:

$$H = \frac{k(s, x)C}{d(s, x)} \tag{11}$$

Among them, $k(s, x)$ is the prior importance of x about s, and C is the threshold of semantic attention. The greater the semantic attention, the closer the semantic of x and s, otherwise, the less the judgment semantics is, the more the index data items with close semantics are aggregated to assist the prediction of regional economic development potential along the railway.

According to the semantically similar regional economic development index data along the railway, the association rules of regional economic development potential along the railway are mined, and the economic development mode of each region along the railway is determined according to the relationship between different attribute characteristics of the index data. In the data set, the attribute information of regional economic development index data along the railway is extracted and divided into continuous attribute set, original invariant attribute set and nominal attribute set [17]. In order to control the data semantic processing process of related attributes, the attribute sets which are similar to the economic development index data of the regions along the railway are generated by using collaborative filtering technology, By classifying the implied semantics, the corresponding concepts of attribute semantics are obtained, and the rule set for mining the characteristics of regional economic development is generated. By using HowNet knowledge base, we define the words in semantic dictionary, take the def item in HowNet as the word concept, and replace the words with similar meaning, so as to make the words have semantic relevance. At the same time, considering the semantic similarity interval between words, we expand the concept of HowNet, delete the same words in def item, and through the minimum semantic similarity and maximum semantic similarity, The semantic similarity of different index data is calculated. The calculation formula is as follows:

$$H = \left(\frac{|K_a \cap K_b|}{\min(K_a, K_b)} + \frac{|K_a \cap K_b|}{\max(K_a, K_b)} \right) \Big/ 2 \tag{12}$$

Among them, K_a, K_b is the concept of a, b, $|K_a \cap K_b|$ is the number of two concepts with the same definition, $\min(K_a, K_b)$, $\max(K_a, K_b)$ is the number of semantic

words with fewer and more semantic words, and H is the semantic similarity of index data. The value of concept similarity is between [0,1]. The smaller the similarity is, the lower the possibility of concept semantic correlation between the feature attributes mined and the prediction of regional economic development potential along the railway line is. The greater the similarity is, the closer the concept semantic is judged. Set the threshold of semantic similarity, select the feature attributes whose H is greater than the threshold, extract the index data and identify their association. The semantic similarity matrix is used to represent the semantic similarity of all index data, and the index features of deep semantic connection are mined. Combined with the semantic elements of regional economic development potential prediction along the railway, the common parts of the semantic elements of index feature attributes are analyzed, the key points of semantic connection are obtained, and the semantic information describing the characteristics of regional economic development is described, According to the semantic bias of economic development characteristics to the prediction of economic development potential, semantic processing index data mining results, the regional economic development characteristics are defined. So far, the mining of regional economic development characteristics along the railway has been completed.

2.5 Preast Economic Development Potential Along the Railway

The characteristics of regional economic development along the railway are input into BP neural network to predict the potential of regional economic development along the railway. BP neural network uses the radial basis function of Multivariable Interpolation, selects three-layer forward neural network as the typical structure of neural network, transforms the characteristic attributes of regional economic development extracted from the input layer in the middle layer, so as to make the category of characteristic attributes of regional economic development closer to the center of the network. Suppose the output value of the i-th neuron is x_i, the sample point of the j-th network center is G_j, and the middle layer neuron of the j-th network center is T_j, then the modified new network center B is:

$$B = \frac{\sum x_i}{G_j} \forall x_i \in T_j \tag{13}$$

The characteristic attributes of regional economic development along the railway line are divided into new network center B, and the network center set is used as the value range to replace the characteristic values of regional economic development along the railway line, so as to eliminate the influence of different dimensions of each dimension data on the prediction of regional economic development potential along the railway line, so as to find the change law of regional economic development potential along the railway line. In the prediction of regional economic development potential along the railway, with the increase of prediction length, the error of prediction value will be larger and larger. Therefore, BP neural network adopts the learning training of fitting error difference to ensure the prediction accuracy of nonlinear factors in the prediction of economic development potential. The learning algorithm of BP neural network is based on the error back propagation algorithm of neural network, and consists of four

processes, In the first stage, the input mode is the forward propagation from the input layer to the output layer through the middle layer. In the second stage, the expected output and the actual output of the network are the error back propagation from the error signal to the input layer through the middle layer, and the connection weight of the neural network is corrected layer by layer. In the third stage, the error back propagation and the mode forward propagation are alternately repeated. In the fourth stage, the neural network tends to converge, The learning convergence process of the global error of the network tends to the minimum [18–20].

In the three-layer BP network structure, the number of nodes in the input layer is 2, the number of nodes in the hidden layer is 6, the number of nodes in the output layer is the number of output vectors, and the number of output vectors in the target value of the neural network is 1, which is the prediction result of the regional economic development potential along the railway. The characteristic attributes of regional economic development along the railway are used as training data and test data, BP neural network is trained with the training data, and the predicted value of regional economic development potential along the railway is output [21]. The predicted value of economic development potential is divided into 1–5 levels, as shown in Table 3.

Table 3. Level of regional economic development potential along the railway

Estimate	Grade	Meaning
0–20	1	Great potential for regional economic development along the railway
20–40	2	The potential of regional economic development along the railway is great
40–60	3	The potential of regional economic development along the railway line is general
60–80	4	The regional economic development potential along the railway is poor
80–100	5	The potential of regional economic development along the railway is poor

According to Table 3, the potential level of regional economic development along the railway is determined [22–24]. So far, the prediction of regional economic development potential along the railway has been completed, and the design of prediction method of regional economic development potential along the railway based on data mining has been realized.

3 Experiment and Analysis

The design method is compared with the macro-economic development prediction method based on optimized multi-dimensional grey model in reference [1] and the marine economic development prediction method based on grey prediction model in reference [5], and the accuracy and efficiency of prediction are compared.

3.1 Experimental Preparation

Taking the counties along the Lanzhou Xinjiang Railway economic belt as the experimental object, the absolute economic difference of the county is expanding year by year from 2010 to 2020, and the relative economic difference of the county is fluctuating, which can be roughly divided into two stages: 2010–2013 is the stage of narrowing the economic difference of the county, the coefficient of variation and Searle index fall below the average level, and the Searle index reaches the lowest value in 2012, In 2013, the coefficient of variation reached the lowest value, which was 0.85352013-2019, which was the stage of continuous expansion of county economic differences. The coefficient of variation and Searle index in 2019 were 1.18 and 1.32 times of 2000, respectively.

From the frequency density distribution curve of per capita GDP of each county, the county economic differences along the Lanzhou Xinjiang Railway are positively skewed, reflecting that the regional economic development level depends on a few higher level counties, and the deviation coefficient fluctuates around 2.0. Using geoda software to calculate the per capita GDP and global autocorrelation coefficient from 2010 to 2020, the statistics of per capita GDP and global autocorrelation coefficient are all positive, showing a trend of gradually increasing first, then gradually decreasing, and finally rising again. In terms of time, the per capita GDP and the global autocorrelation coefficient gradually increased from 2010 to 2013, showing the characteristics of spatial agglomeration. The index gradually decreased from 2013 to 2015, and reached the lowest value of 0.1829 in 2015. After 2016, the per capita GDP and the global autocorrelation coefficient increased slowly. Although there was a trend of gradual recovery, the per capita GDP and the global autocorrelation coefficient in 2019 were still lower than those in 2010, It shows that the economic agglomeration of the economic belt along the Lanzhou Xinjiang Railway can not return to the initial level after the shock is reduced, and the spatial agglomeration of the county economy shows signs of recession, reflecting the trend of the expansion of the county economic differences.

According to the section, the economic belt along Lanzhou Xinjiang Railway is mainly divided into Gansu section, Qinghai section and Xinjiang section. According to the spatial decomposition characteristics of Searle index, the overall difference of county economy in the economic belt along Lanzhou Xinjiang Railway is composed of the difference between Gansu section, Qinghai section and Xinjiang section and their respective internal differences. From 2010 to 2020, the differences among the three regions are small and show a trend of gradual narrowing. The Searle index gradually decreases from 0.0353 to 0.0028, and the contribution rate of the overall regional differences also decreases from 11.10% to 0.67%. On the whole, the internal difference of Gansu section is expanding year by year, and the contribution rate to the regional economic difference is the largest, reaching 71.20% in 2019. Although the internal difference of Qinghai section remains at a low level, it narrows slightly in the early stage, but expands in the later stage. The internal difference of Xinjiang section is shrinking year by year, The internal difference of Gansu section is the main contributor to the overall difference of economic belt along Lanzhou Xinjiang Railway.

3.2 Experimental Results

Prediction of Deviation Degree Experimental Results

According to the above historical data, the three methods respectively calculate the predicted value of regional economic development potential along the railway. Combined with the actual development of the region, the deviation degrees of the upper limit, lower limit and basic value of the prediction are calculated.

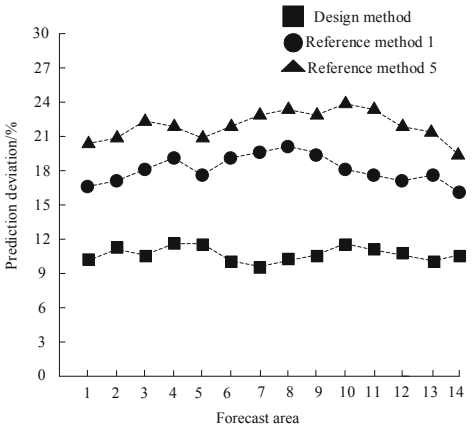

Fig. 1. Results ast upper limit deviation

It can be seen from Fig. 1 that the prediction deviation of the upper limit value of the design method is significantly smaller than that of the two groups of commonly used methods, with an average deviation of 10.4%. The average deviations of reference [1] and reference [5] are 18.7% and 22.1% respectively, and the prediction deviation of the design method is reduced by 8.3% and 11.7% respectively. The lower limit values of the regional economic development potential predicted by the three methods are counted, and the prediction deviation degree is calculated.

It can be seen from Fig. 2 that the prediction deviation of the lower limit value of the design method is also smaller than that of the other two commonly used methods, the average deviation is 10.9%, the average deviation of reference [1] and reference [5] are 18.2% and 21.1% respectively, and the prediction deviation of the design method is reduced by 7.3% and 10.2% respectively. Finally, the basic value of regional economic development potential along the railway is predicted, and the comparison results of prediction deviation are shown in Fig. 3.

As can be seen from Fig. 3, compared with the upper and lower limit values, the prediction deviation of the basic value has decreased, but the deviation of the design method is still less than that of reference [1] and reference [5], with an average deviation of 3.1%, and the average deviation of reference [1] and reference [5] is 9.2% and 14.1% respectively, and the prediction deviation of the design method has decreased by 6.1% and 11.0% respectively. Based on the comprehensive analysis of the prediction upper

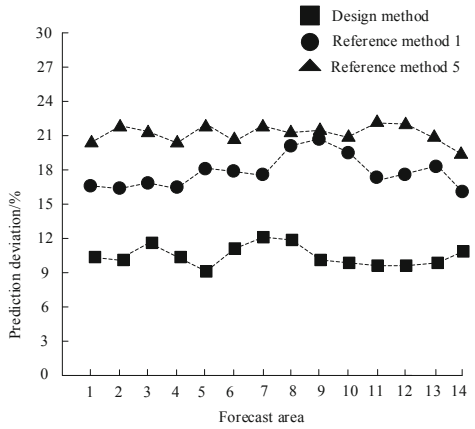

Fig. 2. Comparison result of the lower limit of prediction and deviation

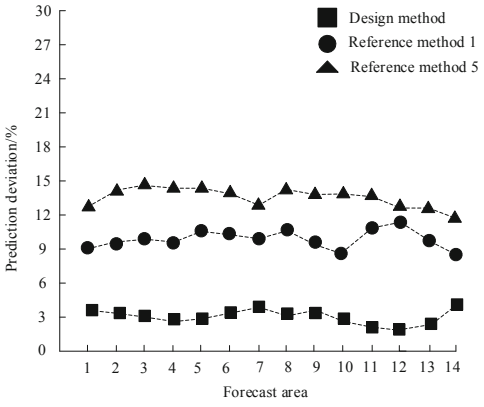

Fig. 3. Predicting comparison result of deviation of base value

limit, lower limit and overall deviation, the introduction of data mining technology into the prediction of regional economic development potential along the railway effectively reduces the impact of influencing factors on the prediction accuracy and the prediction deviation.

Predicted Time-Consuming Experimental Results

The total prediction time of regional economic development potential along the railway in 10 groups of experiments is counted. Each group of experiments is conducted 20 times respectively. The average prediction time of each group of experiments is counted to obtain the total prediction time of regional economic development potential along the railway under different methods. The experimental comparison results are shown in Table 4.

Table 4. Results of development potential (s)

Number of experiments	Design method: Forecast time	Reference [1]: forecast time	Reference [5]: forecast time
1	34.65	100.23	121.73
2	35.23	102.32	120.84
3	35.89	99.34	122.44
4	35.18	101.32	121.37
5	34.04	100.32	120.87
6	35.82	101.76	122.93
7	35.12	99.63	121.83
8	35.46	100.73	120.84
9	35.23	101.83	121.83
10	34.72	100.92	120.83

It can be seen from Table 4 that the average prediction time of the design method is 31.62 s, and the prediction time of reference [1] and reference [5] are 100.84 and 121.55 s respectively. Compared with reference [1] and reference [5], the prediction time of the design method is reduced by 69.22 s and 89.93 s respectively. To sum up, compared with the common methods, this design method reduces the deviation of the predicted value of the regional economic development potential along the railway, and the predicted result is closer to the actual development area of the regional economy. At the same time, it shortens the prediction time, and improves the prediction efficiency while ensuring the accuracy.

4 Conclusion and Prospect

In order to improve the accuracy and efficiency of the prediction of regional economic development potential along the railway, a prediction method of regional economic development potential along the railway based on data mining is proposed, which gives full play to the technical advantages of data mining. The prediction of regional economic development potential along the railway is realized by extracting influencing factors, collecting index data, pre-processing index data and mining development characteristics.

However, there are still some shortcomings in this study. In the future research, we will adopt the custom way to provide the semantic class view of the economic development characteristic attributes, and delete the attributes that will not have a related impact on the potential prediction, so as to further improve the reliability of the prediction results.

References

1. Wu, P., Qiu, S.: Prediction of macroeconomic development based on optimized multidimensional grey model. Stat. Decis. **36**(3), 42–45 (2020)
2. Li, Y., Lu, S., Yuan, X., et al.: Forecasting algorithm of macroeconomic indicators based on correlation analysis. Command Inf. Syst. Technol. **11**(1), 84–88+100 (2020)
3. Lu, B., Ming, Q., Guo, X., et al.: Current and future aspects of coupling situation of tourism technological innovation-regional economy in China. Geogr. Geo-Information Sci. **36**(2), 126–134 (2020)
4. Yang, R., Xu, T.: Development scale of Chongli Ski tourism industry under Winter Olympics: from the perspective of economic forecasting. J. Shenyang Sport Univ. **38**(6), 1–7 (2019)
5. Xu, X., Zhu, R.: The development forecast of Shanghai marine economy based on grey prediction model. Ocean Dev. Manage. **36**(10), 44–46 (2019)
6. Fan, J., Duan, H., Shu, L.: Forecast analysis of coordinated development of highwaytransportation and national economy. J. Chang'an Univ. (Philos. Soc. Sci. Ed.) **21**(3), 49–58 (2019)
7. Zhou, W., Yang, W.: Research on macroeconomic forecasting based on interval-valued financial time series data. On Economic Problems **487**(3), 35–41 (2020)
8. Wang, J.: Monitoring and forecasting economic performance with big data. Data Anal. Knowl. Discov. **4**(1), 12–25 (2020)
9. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big Data Driven e-Learning and e-Education. Mobile Netw. Appl. 1–4 (2021)
10. Chen, H.-J., Hu, X.-B., Deng, X.: A short-term macroeconomic Forecasting model based on GMDH. J. Sichuan Univ. (Nat. Sci. Ed.) **57**(5), 915–919 (2020)
11. Wu, R., Zhou, X.: Research on short-term prediction of local economy based on big data. J. Univ. Sci. Technol. Liaoning **43**(4), 304–308 (2020)
12. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. **3**, 1–23 (2020)
13. Liu, S., Liu, X., Wang, S., Muhammad, K.: Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT assisted complex environment. Neural Comput. Appl. **33**(4), 1055–1065 (2021)
14. Feng, X., Li, N., Wang, G., et al.: Development of a liver cancer risk prediction model for the general population in China: a potential tool for screening. Ann. Oncol. **30**, ix46–ix47 (2019)
15. Pei, C., Liu, Y.: The simulation of the prediction model of the economic development potential of the Coastal Area. J. Coastal Res. **112**(sp1), 211–215 (2020)
16. Hoffmann, A., Ponick, B.: Method for the prediction of the potential distribution in electrical machine windings under pulse voltage stress. IEEE Trans. Energy Convers. **36**(2), 1180–1187 (2020)
17. Bushuk, M., Msadek, R., Winton, M., et al.: Regional Arctic sea-ice prediction: potential versus operational seasonal forecast skill. Clim. Dyn. **52**(5–6), 2721–2743 (2019)
18. Sturniolo, S., Liborio, L.: Computational prediction of muon stopping sites: a novel take on the unperturbed electrostatic potential method. J. Chem. Phys. **153**(4), 044111 (2020)
19. Bai, K., Li, K., Chang, N.-B., Gao, W.: Advancing the prediction accuracy of satellite-based PM2.5 concentration mapping: a perspective of data mining through in situ PM2.5 measurements. Environ. Pollut. **254**, 113047 (2019)
20. Wang, H., Huang, Z., Zhang, D., et al.: Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in Kawasaki disease. IEEE Access **8**, 97064–97071 (2020)

21. Wang, C., Bi, J., Sai, Q., et al.: Analysis and prediction of carsharing demand based on data mining methods. Algorithms **14**(6), 179 (2021)
22. Li, H., Lu, Y., Zheng, C., Yang, M., Li, S.: Groundwater level prediction for the arid oasis of northwest china based on the Artificial Bee Colony Algorithm and a back-propagation neural network with double hidden layers. Water **11**(4), 860 (2019)
23. Trumpis, M., Chiang, C.H., Orsborn, A.L., et al.: Sufficient sampling for kriging prediction of cortical potential in rat, monkey, and human ECoG. J. Neural Eng. **18**(3), 036011 (2021). (18pp)