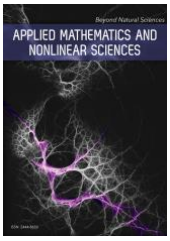
sciendo Applied Mathematics and Nonlinear Sciences, 9(1) (2024) 1-22



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Research and Development of Decision Support System for Tourism Management Based on Big Data Analysis

Yanling Xiao[1,†]

1. Wuyi University, Strait Success College, Wuyishan, Fujian, 354300, China.

Submission Info

Communicated by Z. Sabir
Received March 20, 2024
Accepted June 3, 2024
Available online July 5, 2024

Abstract

With the continuous expansion and change of the tourism market, the massive amount, complexity and dynamic change of tourism data make the establishment of tourism management decision support system has become an important issue in the tourism industry. In this paper, we use the plain Bayesian classification algorithm, the improved Apriori association algorithm, and the gray GM(1, N) prediction model to mine and process the tourism big data and combine the 3S technology and the Agent-based knowledge representation technology to realize the construction of the tourism management decision support system based on big data analysis, and make the optimal decision for the planning of tourist attractions and routes. The attraction classification method's accuracy rate is 72.98%, and its feasibility is high. The integrated error of the adopted GM(1,7) model is only 2.587%, which is smaller than the 3.483% of the GM(1,1) model and the 4.594% of the linear regression model, and the model accuracy is high. The average response time and average TPS of the system are 8.70s and 15.89s, respectively, which generally meet the demand for the system's processing capability. This study provides a reference for the construction of a decision support system for tourism management.

Keywords: Plain Bayesian classification; Apriori algorithm; Gray GM(1,N) prediction model; Tourism management decision support system; Big data analysis.
AMS 2010 codes: 68-02

†Corresponding author.
Email address: m13799146128@163.com

ISSN 2444-8656
https://doi.org/10.2478/amns-2024-1606

sciendo
OPEN ACCESS © 2023 Yanling Xiao, published by Sciendo.

CC BY This work is licensed under the Creative Commons Attribution alone 4.0 License.

1 Introduction

Tourism is one of the industries with the broadest prospects for the application of big data. Still, the lag in the construction of informationization restricts the development of the tourism industry. Tourism consumption is constantly upgrading, and the traditional system can no longer meet the overall development needs of the tourism industry [1-3]. China's tourism informatization application system, especially the overall display of data and service application, is still very low, whether it is tourism management, travel-related enterprises, or tourists who can not see the desired data information [4-5].

The purpose of tourism data resource management is to formulate unified data collection standards. Collect, catalog, and grade data. Achieve classification and archiving of tourism data, authorize external applications [6-7], support major systems, break information silos, form data standard specifications and tourism data assets within the tourism industry, and then carry out various types of excavation and use on this basis to assist in multiple decision-making [8-9]. While introducing big data analysis technology, it should focus on the integration with the original information resource management in order to maximize the use of internal resources to mine the value of information [10]. Through the management of tourism data resources, it can solve the long-standing drawbacks of poor information, untimely information, and inaccurate information in the industry, make tourism data clearer so that the management of enterprises in the industry has a basis, the government has a reference for decision-making, and the service is more intelligent and high-quality, so as to promote the long-term development of the entire tourism industry [11-13]. Tourism data resource management, like other data resource management, goes through the process of data collection, cleaning, analysis, and mining [14].

The purpose of developing a scenic spot management service system based on base station data is to effectively utilize big data technology and apply it to the field of tourism [15] and to promote the breakthrough development of smart tourism by solving the problem of huge amount of data processing [16]. The challenges brought by the era of big data are not only reflected in how to deal with the huge amount of data and obtain valuable information for tourism from it but also in how to strengthen the application of big data technology in the tourism business [17-18], obtain effective and relevant data in a timely manner, and utilize data mining and analysis to build a public opinion analysis system [19], and to build a golden week tourism analysis, a tourism marketing strategy analysis, a tourism topic analysis, a tourism passenger source analysis, advertising efficiency analysis, and so on big data analysis system, and big data reports on government leadership decision-making, government effectiveness, and public safety in the tourism industry [20-21].

In this study, the basic framework of big data analysis is constructed by using the plain Bayesian classification algorithm, the improved Apriori algorithm, and the gray GM(1, N) prediction model, on the basis of which the 3S technology is combined with the Agent-based knowledge representation technology and the problem-solving technology to complete the construction of the decision support system for tourism management based on big data analysis. First, the study classifies and recommends tourist attractions, followed by recommendations for tourist routes based on association rules. The prediction model's performance is evaluated after predicting tourism trends. Finally, the designed tourism management decision support system is tested for performance to provide a reference for the construction of similar systems.

2 Data analysis methods and predictive modeling

2.1 Plain Bayesian classification algorithm

Among many classification algorithms and theories, plain Bayes has been widely used due to its computational efficiency, high accuracy, and solid theoretical foundation. The structure of the plain Bayesian classifier is schematically shown in Fig. 1.

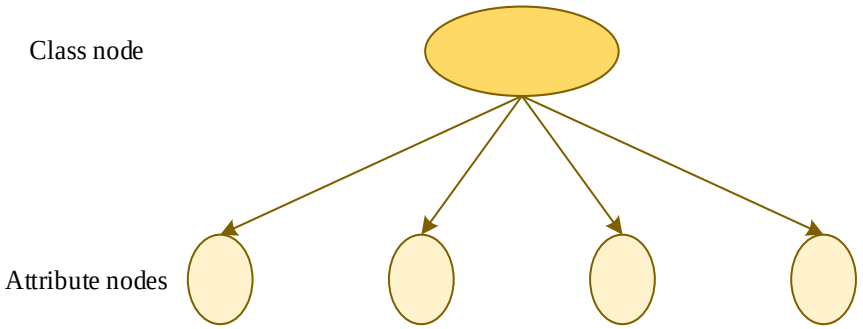


Figure 1. The structure diagram of naive Bayes classifier

The plain Bayesian model assumes that the condition variables are independent of each other with respect to the class or decision variables. In real life, this kind of complete independence is relatively rare, and this assumption that one-time condition attributes are independent of each other affects the classification performance of the plain Bayesian model to some extent. However, plain Bayes has its advantages, especially its simplicity and efficiency, which make it applicable in data mining classification and clustering. The following is its workflow.

1) Represent each data sample by a feature vector. For example, a data sample is represented by a n-dimensional feature vector $X=\{x_1,x_2,\cdots,x_n\}$ and these feature vectors correspond to the n measures of each of the n attribute $A_1,A_2,\cdots A_n$ samples.

2) Assuming that there is m class $C_1,C_2,\cdots C_m$, to classify the sample of data to be processed X without the value of the class label attribute, x is generally estimated to be the class with the highest value of the a posteriori probability, so that it is to assign the unknown samples to the class C_i, which in fact translates into maximizing the value of $P(C_i|X)$. The class C_i that maximizes the value of $P(C_i|X)$ is called the maximum a posteriori assumption and is obtained according to Bayes' theorem:

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)} \tag{1}$$

3) From the formula we see that to maximize $P(C_i|X)$, it is only necessary that $P(X|C_i)P(C_i)$ be maximized. If in the absence of knowledge of the a priori probability, it is generally considered that they are equal probability, that is, $P(C_1)=P(C_2)=\cdots=P(C_n)$

and use this to find the maximization of $P(C_i|X)$, otherwise the maximization of $P(X|C_i)P(C_i)$. According to probability, the a priori probability formula is:

$$P(C_i)=\frac{s_i}{s} \tag{2}$$

s_i denotes the number of training samples and s is the total number of training samples.

4) If the number of attributes in the attribute set is relatively large, the time spent on calculating $P(X|C_i)$ may be very large, and in order to minimize the time spent, it is usually assumed that the class conditions are independent, i.e., the individual attribute values are independent of each other. Equation (3) allows for the calculation of $P(X|C_i)$:

$$P(X|C_i)=\prod_{k=1}^{n}P(x_k|C_i) \tag{3}$$

Where both probabilities $P(x_1|C_i),P(x_2|C_i),\cdots,P(x_n|C_i)$ can be valued by the training samples if A_k of them is a discrete property:

$$P(x_n|C_i)=\frac{s_{ik}}{s_i} \tag{4}$$

(1) Where s_{ik} is the number of training samples in class C_i that denotes attribute A_k taking the value of x_i, and s_i denotes the number of training samples in C_i.

(2) If A_k is a continuous-valued attribute, it is generally considered to belong to a Gaussian distribution.

5) Classify x and calculate $P(X|C_i)P(C_i)$ separately for each class C_i. If sample x is classified into class C_i, the condition to be satisfied is:

$$P(X|C_i)P(C_i)>P(X|C_j)P(C_j),1\le j\le m,\ j\ne i \tag{5}$$

In other words, x is classified to the class C_i in which $P(X|C_i)P(C_i)$ has the largest value.

The plain Bayesian classification and its workflow are shown in Fig. 2.

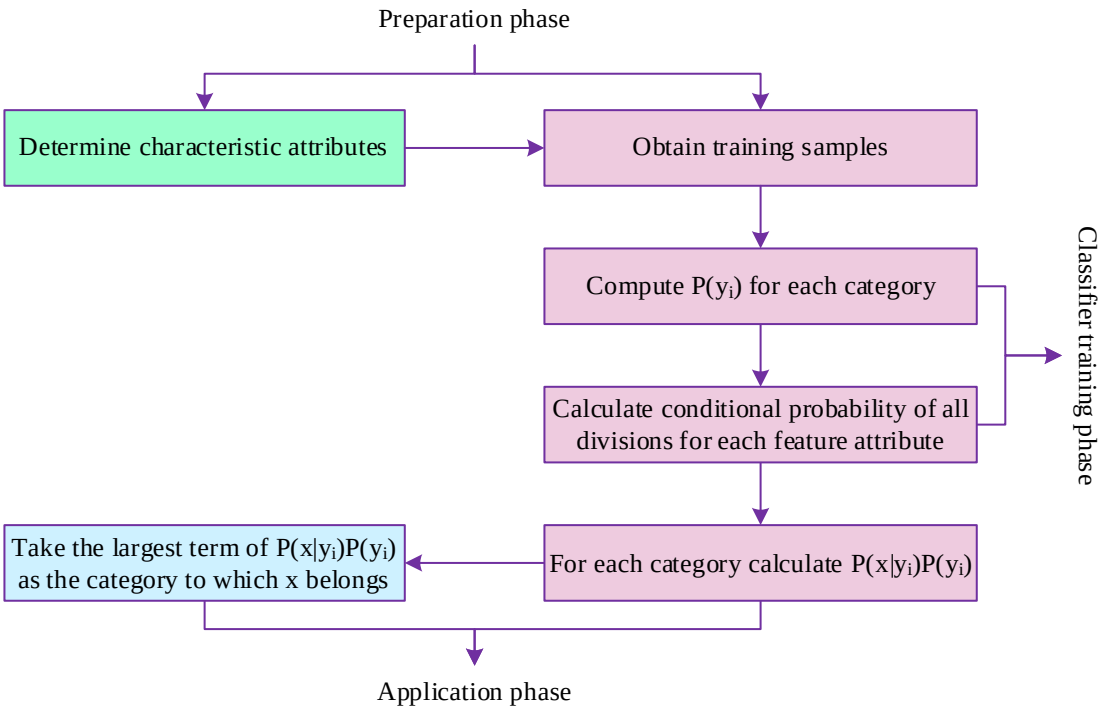


Figure 2. The category of naive Bayes and its work flowchart

2.2 Improved spatial association rule-based mining

Improved Apriori algorithm description:

1) Create array $PAR[n]$ with initial value 0. Scan the database and find the independent occurrence probability $P_1, P_2, \cdots P_n$ and support for each attribute item $A_1, A_2, \cdots A_n$. Each array element $PAR[1], PAR[2] \cdots PAR[n]$ in PAR records the value of independent occurrence probability $P_1, P_2, \cdots P_n$ of each attribute item $A_1, A_2, \cdots A_n$ in the database.

The steps to find the probability of independent occurrence of attribute item A_i are as follows:

(1) If it is not the end of the database, read the record.

(2) If attribute item A_i occurs in that record, add 1 to PAR[i].

(3) Repeat until the end of the database, then $PAR[i] = PAR[i]$/number of records in the database.

(4) Repeat steps (1), (2), and (3) to find the independent occurrence probability of attribute item $A_1, A_2, \cdots A_n$.

2) Set a minimum probability of V_1 to obtain m frequent 1-term sets with probabilities PAR[1], $PAR[j], \cdots, PAR[j], \cdots, PAR[m]$.

3) Let the probability of perfect correlation of A_k and A_m be a, the probability of incomplete correlation of A_k and A_m be b, and $a+b=1, 0<a,b<1$, then P_{km} can be expressed as:

$$P_{km} = a*P_k + b*P_k*P_m \tag{6}$$

From the independent occurrence probability of m frequent 1-item sets, the probability of any two attribute items appearing at the same time is estimated according to Eq. (6), and those whose probability is greater than the pre-set minimum probability of v are the candidate frequent 2-item sets. If the probability is less than the pre-set minimum probability of $v2$, it is directly assigned to 0 to simplify the calculation later. The probability of candidate frequent 2-item sets is stored in array $PUA2[i]$.

V_2 indicates the minimum probability of the candidate frequent 2-term set, and so on, V_{k-2} indicates the minimum probability of the candidate frequent k-term set, and the minimum probability V_{k-1} is set:

$$\begin{aligned} V_{k-1} = {} & a^* \min\left(PAR_{k-1}[1], PAR_{k-1}[2], \cdots, PAR_{k-1}[m]\right) \\ & + b^* \min\left(PAR_{k-1}[1], PAR_{k-1}[2], \cdots, PAR_{k-1}[m]\right) \\ & * \max\left(PAR_{k-1}[1], PAR_{k-1}[2], \cdots, PAR_{k-1}[m]\right) \end{aligned} \tag{7}$$

4) Iterate the above process to find the probability of simultaneous occurrence of k attribute items $A_1, A_2, \cdots A_k$ from the attribute items of $PUA_{k-1}[i] \neq 0$, and so on until the set of n items.

5) From the candidate frequent data item sets obtained in the previous step, scan the database to find the support of each candidate frequent data item set.

   (1) Create array $DMA[m]$, and the initial value of each element of array $DMA[m]$ is 0, and m is the number of candidate data item sets.

   (2) If it is not the end of the database, read the database record.

   (3) If there is a candidate frequent data item set $\left(A_i, A_j, \cdots, A_k\right)$, and in the current data record read, $A_i \neq 0$, $A_j \neq 0, \cdots$, $A_k \neq 0$, then record the support of the candidate data item set $\left(A_i, A_j, \cdots, A_k\right)$. $DMA[k] = DMA[k] + 1$ Repeat until the end of the database, and find the actual support of all candidate frequent data item sets.

Repeat (2) and (3) until the end of the database to find the actual support of all candidate frequent data item sets.

6) The frequent itemset is derived from the support degree of the candidate frequent data itemset, and if $DMA[k]$ is greater than or equal to the minimum support degree of the predetermined frequent itemset, then this frequent itemset is output. Steps 5) and 6) are used to verify the

correctness of each candidate frequent data item set derived by probabilistic estimation method, i.e., whether it meets the requirement of minimum support set by the user.

7) The frequent itemsets generated from step (6) are outputted as association rules.

2.3 Gray GM(1,N)-based tourism volume prediction model

2.3.1 GM(1,N) model fundamentals

The gray $GM(1,N)$ model is used to portray the connection and change between n associated variables, where "1" in $GM(1,N)$ represents a first-order equation and "N" represents multiple variables, and the change trend of the target variable is examined on the basis of the driving factor to portray the functional relationship between independent variables and dependent variables, so as to realize the prediction of the examined target, $GM(1,N)$ the basic principle of the model is as follows:

Consider that there are $X_1, X_2, \cdots, X_n$ these *n* variables, i.e.:

$$X_i^{(0)} = \left(X_i^{(0)}(1), X_i^{(0)}(2), \cdots, X_i^{(0)}(n) \right), i = 1, 2, \cdots, N \tag{8}$$

Doing an accumulation of $X_i^{(0)}$ to generate 1-AGO gives the generating series:

$$\begin{aligned} X_i^{(1)} &= \left(X_i^{(0)}(1), \sum_{m=1}^{2} X_i^{(0)}(m), \cdots, \sum_{m=1}^{n} X_i^{(0)}(m) \right) \\ &= \left(X_i^{(1)}(1), X_i^{(1)}(1) + X_i^{(0)}(2), \cdots, X_i^{(1)}(n-1) + X_i^{(0)}(n) \right), \\ & i = 1, 2, \cdots, N \end{aligned} \tag{9}$$

We view the moments m of series $X_i^{(1)}$ as continuous time variables t and series $X_i^{(1)}$ instead as a function of time t. If series $X_2^{(1)}, X_3^{(1)}, \cdots, X_N^{(1)}$ affects the rate of change of $X_1^{(1)}$, then $X_i^{(1)}$ satisfies the following first-order linear differential equation model:

$$\frac{dX_1^{(1)}}{dt} + aX_1^{(1)} = b_1 X_2^{(1)} + b_2 X_3^{(1)} + \cdots + b_{N-1} X_N^{(1)} \tag{10}$$

Where Eq. (10) is called the $GM(1,N)$-model.

According to the definition of derivative, there are:

$$\frac{dX_1^{(1)}}{dt} = \lim_{\Delta t \to 0} \frac{X_1^{(1)}(t + \Delta t) - X_1^{(1)}(t)}{\Delta t} \tag{11}$$

If expressed in discrete form, the differential term can be written as:

$$\frac{\Delta X_1^{(1)}}{\Delta t}=\frac{X_1^{(1)}(k+1)-X_1^{(1)}(t)}{k+1-k}=X_1^{(1)}(k+1)-X_1^{(1)}(k)=X_1^{(0)}(k+1) \tag{12}$$

Taking $X^{(1)}$ in Eq. (11) as the average of moments k and $k+1$, Eq. (12) reduces to:

$$\begin{aligned}&X^{(0)}(k+1)+a\left[\frac{1}{2}\left(X_1^{(1)}(k+1)+X_1^{(1)}(k)\right)\right]\\&=b_1X_2^{(1)}(k+1)+\cdots+b_{n-1}X_n^{(1)}(k+1)\end{aligned} \tag{13}$$

In the $GM(1,N)$ model, a is referred to as the development coefficient, b_j is referred to as the driving coefficient, and $b_ix_i^{(1)}(k)$ is referred to as the driving term. Denoting the parameter coefficients as $\hat{\alpha}=(a,b_1,b_2,\cdots b_{N-1})^T$, a system of linear equations can be obtained:

$$Y=B\hat{\alpha} \tag{14}$$

$$B=\begin{pmatrix}-\frac{1}{2}\left(X_1^{(1)}(1)+X_1^{(1)}(2)\right) & X_2^{(1)}(2) & \cdots & X_N^{(1)}(2)\\-\frac{1}{2}\left(X_1^{(1)}(2)+X_1^{(1)}(3)\right) & X_2^{(1)}(3) & \cdots & X_N^{(1)}(3)\\\vdots & \vdots & & \vdots\\-\frac{1}{2}\left(X_1^{(1)}(n-1)+X_1^{(1)}(n)\right) & X_2^{(1)}(n) & \cdots & X_N^{(1)}(n)\end{pmatrix} And\ Y=\begin{Bmatrix}x_1^{(0)}(2)\\x_1^{(0)}(3)\\\cdots\\\cdots\\\cdots\\x_1^{(0)}(n)\end{Bmatrix} \tag{15}$$

The above equation is solved by the least squares method and the approximate solution $\hat{\alpha}=\left(B^TB\right)^{-1}B^TY_N$ is obtained, and the corresponding function of time for the model can be obtained by solving $\hat{\alpha}$:

$$\hat{X}_1^{(1)}(k+1)=\left(X_1^{(1)}(0)-\frac{1}{a}\sum_{i=2}^{n}b_iX_i^{(1)}(k+1)\right)e^{-ak}+\frac{1}{a}\sum_{i=2}^{n}b_iX_i^{(1)}(k+1) \tag{16}$$

The above equation is the specific formula for model prediction, and this equation is then cumulatively reduced to obtain the gray prediction model of the original model:

$$\hat{X}_1^{(0)}(k+1)=\partial^{(1)}\hat{X}_1^{(1)}(k+1)=\hat{X}_1^{(1)}(k+1)-\hat{X}_1^{(1)}(k) \tag{17}$$

2.3.2 GM(1,N) model error test methods

After the establishment of the model must be statistically tested to determine the accuracy of the model fit level. We commonly test methods are residual test, post-test difference test, correlation test, and the index value of the difference between the value of the test method. The residual test of point-by-point statistics is the simplest and most frequently utilized among them. The model is evaluated in this paper using the residual test.

Let the original sequence be $X^{(0)} = \left\{ x^{(0)}(k) \right\}$, the simulated sequence of computed values be $\hat{X}^{(0)} = \left\{ \hat{x}^{(0)}(k) \right\}$, and the residual difference between the original and simulated sequences be $q(k)$, i.e:

$$q(k) = x^{(0)}(k) - \hat{x}^{(0)}(k) \tag{18}$$

Based on Eq. (18), the relative error of the prediction model can be calculated to be $\varepsilon(k)$, i.e:

$$\varepsilon(k) = \frac{q(k)}{x^{(0)}(k)} \times 100\% \tag{19}$$

Based on Eq. (19), the average relative percentage error of the prediction model can be calculated to be $\bar{\varepsilon}(k)$, i.e:

$$\bar{\varepsilon}(k) = \frac{1}{n} \sum_{k=1}^{n} \left| \varepsilon(k) \right| \tag{20}$$

Based on Eq. (20), the simulation accuracy of the prediction model can be calculated as:

$$p = 1 - \bar{\varepsilon}(k) \times 100\% \tag{21}$$

3 Design of a decision support system for tourism management based on big data analysis

3.1 Decision-making process based on big data analysis

The analysis and pattern discovery of various types of tourism data helps to make tourism management decisions. Still, according to the development status of various intelligent decision support technologies, none of them can complete all the qualitative and quantitative analyses individually. They can't deal with the decision support of complex events, so it is necessary to synthesize a variety of intelligent decision-support technologies to realize the decision-making process based on big data analysis. In this paper, the process is divided into an iterative process of four stages: problem formulation, data acquisition and model selection, big data analysis, evaluation, and decision making. Figure 3 illustrates the decision-making process that utilizes big data analysis.

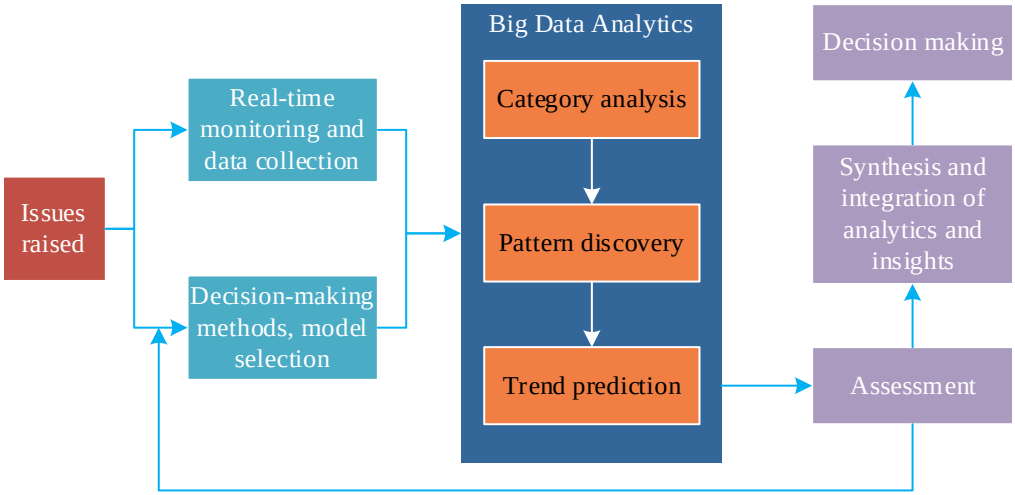


Figure 3. Decision making process based on big data analytics

The key to this process is the ability to compare and select various analytical methods and decision-making models based on the assessment results, as well as to make continuous adjustments to the data collection strategy in order to improve the adaptability and effectiveness of big data analytics.

3.2 System architecture

In order to provide effective decision support for the process and adapt to the selection and use of a variety of decision-making techniques, this paper comprehensively integrates a variety of intelligent decision-making techniques, such as data mining, Agent and group decision-making, to construct a decision support system for tourism management based on big data analysis. The system architecture is shown in Figure 4.

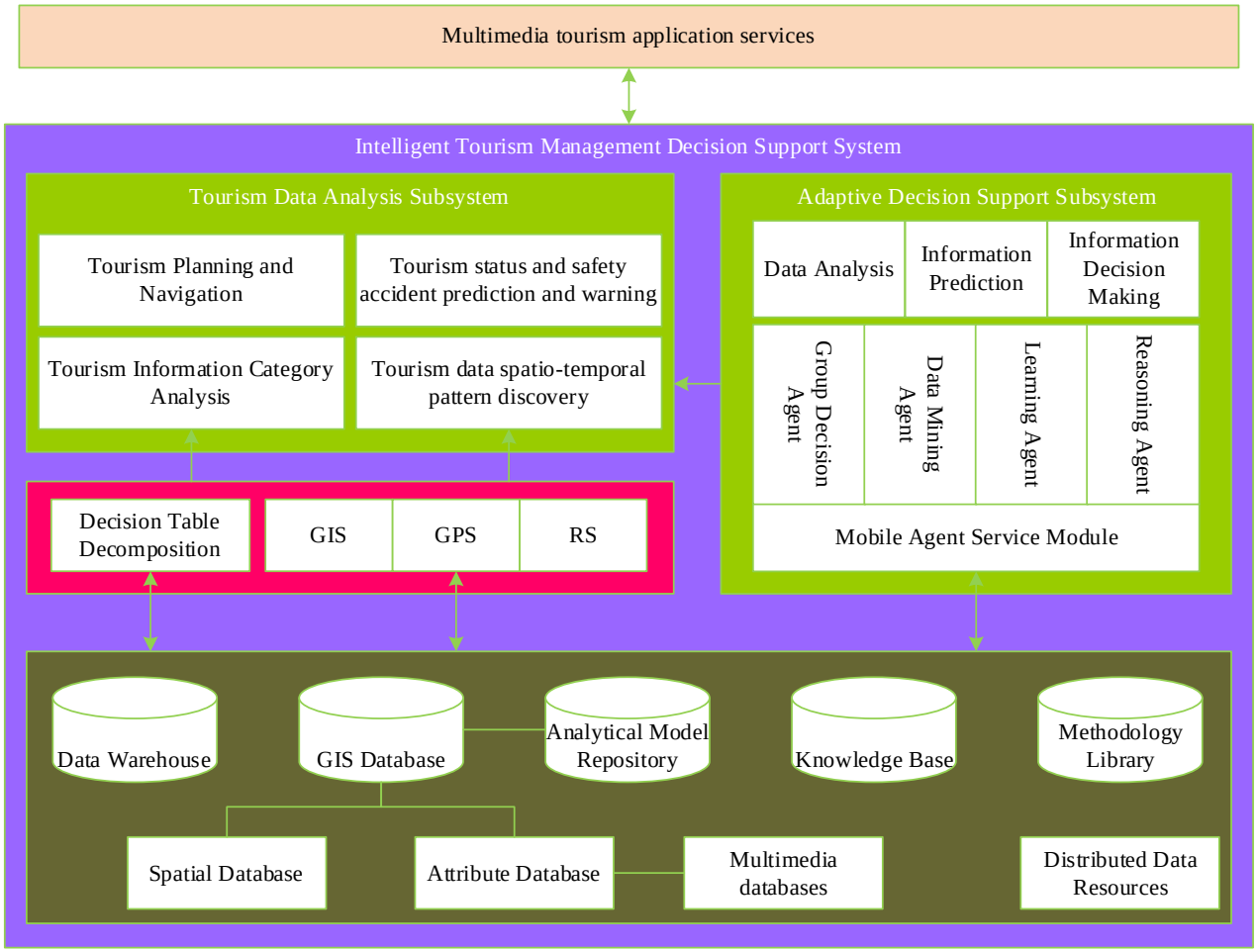


Figure 4. Architecture of tourism management decision support system based on Big Data Analytics

The system's technical support level is enhanced by the use of 3S technology for spatial data collection and processing. Decomposition technology solves the complexity of large decision tables. Using data mining technology and intelligent reasoning technology, the system's database, model library, and knowledge base are utilized for model decision-making, for example, reasoning Web mining, etc., to assist experts in forming opinions and views. Use adaptive decision support technology in the system to achieve quantitative analysis support for complex tourism management problems and dynamic data. Group/organization decision support technology, is used to provide the human-human interaction environment, which encourages collaborative interaction among experts and relies on group experts to solve unstructured intuitive thinking problems.

The data analysis subsystem relies on the tools and technical environments provided by 3S, adaptive decision support technology, etc., and employs various data analysis methods to classify, mine, and discover spatiotemporal patterns in tourism data, thus realizing holiday tourism planning, tourism status, and safety accident prediction and warning, and ultimately providing decision-making support for various types of tourism application systems.

3.3 System key technologies

3.3.1 Integrated application of 3S technology

Due to the spatial and multimedia characteristics of tourism data, the application of Geographic Information System (GIS), Remote Sensing (RS), and Global Positioning System (GPS) technology in the tourism management decision-making system can effectively enhance the system's data acquisition and updating capabilities.RS has strong functions in the acquisition of spatial information of tourism resources, image processing, etc.GPS has a greater role in the spatial positioning of tourism resources, GPS plays a greater role in the spatial positioning of tourism resources field data acquisition and can be used to quickly obtain the parameters of the ground control points of the resource monoliths, and can also be used to measure and get the spatial information data of tourism resources.GIS carries out the management of the acquired spatial data as well as the spatial analysis. These technologies are integrated and applied to the tourism management decision support system, which can automatically and in real-time collect, process, and update tourism data. Supplemented with 3D technology, virtual reality (VR) technology can also realize the visualization of tourism resource data in geospatial form, which can simulate the effect of tourism planning schemes in tourism planning decision-making in order to further improve the planning scheme.

Combining 3S technology with agent-based knowledge representation and problem-solving technology it has unique advantages when applied in decision support systems. In addition to traditional spatial information management and query functions, it can also combine expert knowledge and spatial data models for intelligent analysis and adopt different analysis strategies according to various user conditions.

3.3.2 Decomposition of large decision tables

Decision table data analysis is currently facing difficulties due to the massiveness and complexity of data in practical tourism management decision-making applications. Many decision tables have a complex structure that involves many objects and attributes. If they are mined directly, the computational complexity increases, and the quality of the rules obtained decreases.

To address the complexity of decision tables, the processing of attribute sets is first considered, which aims to reduce the size of attribute sets. By removing irrelevant or unimportant attributes from the decision table, attribute approximation methods aim to reduce the size of the attribute set while keeping the same classification capability. In certain situations, attribute reduction techniques still have some drawbacks. For example, the approximated attribute set may still be large, the results of the approximation algorithms depend on the number of objects in the training set, and some approximation algorithms are less efficient and computationally complex for large decision tables.

Decomposition techniques can also be used to solve the problem of large decision tables, i.e., the original decision table is decomposed into a number of smaller sub-tables from the point of view of the object set or attribute set. After the decomposition of the decision table, each sub-table itself is still a complete decision table. Still, the size of the object set or condition attribute set is smaller, and the analysis and processing are simpler than the original decision table. The acquisition of rules in sub-tables reduces the amount of data processed each time, avoids the difficulties and defects of modeling directly in complex systems, and improves the efficiency and quality of data analysis. Decomposition methods used in decision table analysis include:

1) Functional decomposition method: the decision attributes in the original table are regarded as a function of the conditional attributes, which is decomposed into a composite function according to the functional decomposition theory, and the set of conditional attributes is divided into two subsets in the external and internal functions. Through recursive decomposition, the decision table is decomposed into sub-tables with hierarchical relationships and smaller sizes. Meaning intermediate decision concepts are obtained by establishing a hierarchical structure on the attribute set, thus describing the complex decision logic.

2) Decomposition method based on attribute kernel: According to rough set theory, the kernel refers to the important attributes in the decision table that play a key role in decision-making. Therefore, based on the kernel and non-kernel attributes in the decision table, the original decision table can be decomposed into two sub-tables, and an intermediate decision attribute can be generated.

3) Decomposition method based on attribute clustering: the decision table is decomposed by calculating the dependency between conditional attributes and thus clustering the conditional attributes to form multiple independent subsets of conditional attributes.

Since different decomposition methods have some differences in many features and performances and are applied differently, it is necessary to choose a suitable decomposition method according to the characteristics of the decision table data itself when using it. It can be evaluated and selected from the aspects of rule quality and classification accuracy, information nondestructiveness and decision equivalence, time complexity, etc., and strives to achieve a balance between decomposition efficiency and classification quality under the premise of ensuring decision equivalence.

4 Analysis of tourism management decisions

4.1 Recommended classification of tourist attractions

In this paper, a classifier is obtained by the Bayesian classification algorithm using the user-attraction combination vector as a sample and the user's rating of attractions as a category label. 7856 pieces of data were used to train the classifier, and then 818 pieces of data not involved in the training were used to test the accuracy of the classifier, and the specific results are analyzed as follows.

1) Overall distribution of classification

The distribution of the classification results of the test samples is shown in Figure 5, where the orange color indicates the true distribution of user scores on attractions in the 818 test samples, and the purple color indicates the distribution of attraction scores predicted by Bayesian classification. The predicted results with the highest and lowest scores are more than the true results, with a 2 almost unchanged difference. The predicted results for scores of 2 and 3 are smaller than the true results, with a difference of 12 and 7, respectively, while the weight of the predicted results for scores of 4 rises significantly, and the predicted results are larger than the true results, with a difference of 16. This suggests that there is a kind of overfitting phenomenon in the classifier and that the high proportion of high scores in the original distribution leads to a further increase in the proportion of high scores in the predicted results.

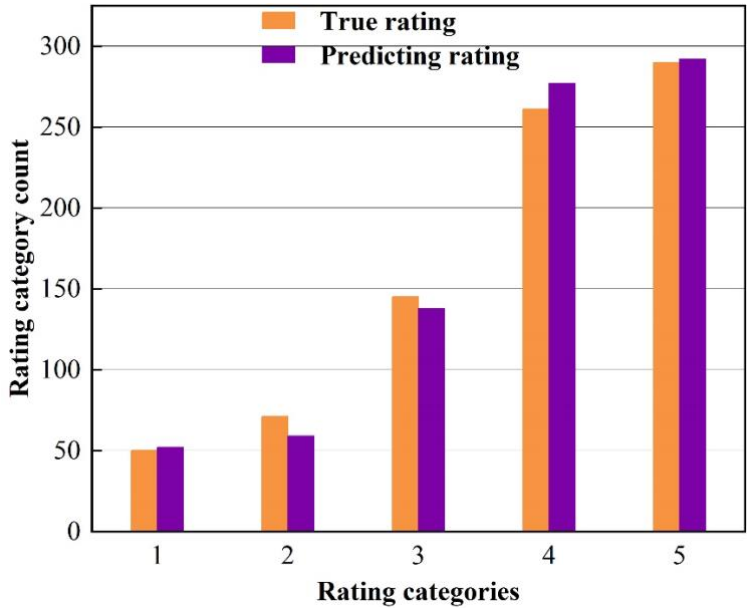


Figure 5. Distribution of test sample classification results

2) Evaluation of classification results

In general, the classification accuracy is in the acceptable range, and the specific classification correctness is analyzed below. The interconversion of all original and predicted categories can be seen in Table 1.

Each row of the table indicates the source of records for a predicted classification, e.g., the first row indicates that there are 22 records with true ratings of 1 and 9 records with true ratings of 2 out of all records with a predicted rating of 1, and so on. Each column indicates how records with specific ratings in the true records were converted to other rating categories after prediction, e.g., the first column indicates that of the records with original ratings of 1, there are 22 records with predicted ratings that remain 1, 5 records with predicted ratings incorrectly predicted to be 2, and so on. The last line in the table is the accuracy of the correct classification of each category, i.e., The ratio of the number of records correctly categorized in each category to the number of original records in the category.

From Table 1, it is easy to see that categories 4 and 5 have the highest classification accuracy rate, reaching more than 81%, which is more satisfactory. On the other hand, the correct categorization of categories 1 and 2 is less effective, with accuracy rates of less than 45%.

Table 1. Transformation of test sample classification results

Original / Forecasting	1	2	3	4	5	Total
1	22	9	11	8	2	52
2	5	28	9	13	4	59
3	8	10	92	15	13	138
4	4	14	15	214	30	277
5	11	11	18	11	241	292
Total	50	71	145	261	290	818
Classification accuracy	44.00%	39.44%	63.45%	81.99%	83.10%	72.98%

In order to explore the reasons for this phenomenon, this paper uses the number of training samples of a category as the independent variable and the classification accuracy as the dependent variable to produce a scatter plot between the two and perform curve fitting. The fitting results are shown in Figure 6. The correctness of category classification and the number of samples used for training in the category show a significant positive correlation. This phenomenon is consistent with the law that the larger the sample size, the more knowledge the classifier can learn. The method proposed in this paper for evaluating attractions based on user and attraction information has an overall accuracy rate of 72.98%, which indicates feasibility.

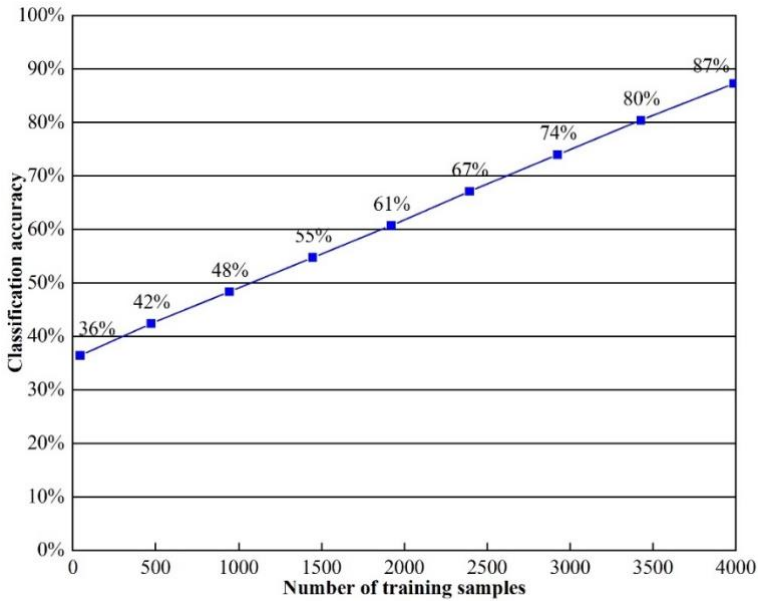


Figure 6. Relationship between sample classification accuracy and number of samples

4.2 Recommendation of travel routes based on association rules

In this paper, we mainly use the improved Apriori association algorithm to explore the associations of tourist attractions in the same region of travel users so that we can generate tourist routes in the same area with the highest degree of user preference. In view of the availability of data, this paper is based on the travel route nodes and user code data of online travel users visiting Nanjing during the whole travel process, and the association mining of travel nodes staying in Nanjing. In order to ensure the authenticity and validity of the data, the screening conditions for the travel route data are established:

First, data with incomplete or garbled user attributes and tourism data in the web crawler are eliminated.

Second, visiting users are non-local residents (excluding users with 1 day of tour).

Third, have visited at least one attraction in the destination.

Through screening to obtain 1080 travel route data in line with the correlation mining and the entry of tourism nodes for some scattered nodes to be merged, such as Fuzimiao and Fuzimiao scenic area collectively referred to as Nanjing Fuzimiao, etc.

In this paper, Nanjing Fuzimiao, Zhongshan Mausoleum, Jiangning Weaving Museum, Jiangnan Weaving Museum, Yuhuatai, Nanjing Massacre Memorial Hall, Yunjin Museum, Xuanwu Lake Park,

Jiangnan Tribute Academy, Ancient Jiming Temple, Zhanyuan Garden, Ming Dynasty City Walls, Old Mendong, Pioneer Bookstore, 1912 Fashion District, Summer Road, Nanjing Yangtze River Bridge, Qixia Mountain, Qixia Temple, Yuejiang Tower, Xinjiekou, and Zijing Mountain are denoted with letters as A1, A2, A3, A4, B1, B2, B3, B4, C1, C2, C3, C4, D1, D2, D3, D4, D5, E1, E2, E3, E4, E5, and the minimum support was set at 0.001. The confidence level was set at 0.01, and the results of the obtained correlation analysis are shown in Table 2.

According to Table 2, in terms of correlation support, code 7 has the largest support of 0.0468. This indicates that Nanjing Fuzimiao and Xuanwu Lake Park are the most frequently chosen combination of travel nodes for visiting online travel users. Codes 2, 4, and 9 also have greater support, at 0.0181, 0.0159, and 0.0181, respectively, which suggests that the proportion of those choosing the combination of tourist nodes such as the Jiangning Weaving Museum, the Rain Flower Terrace, and the Memorial Hall of the Nanjing Massacre with the Nanjing Fuzimiao, the hotspot of Nanjing's tourism, is also higher.

In terms of association confidence, among the 20 coded association rules, A1 appeared 12 times with the highest frequency. This indicates that there is a correlation between most of the travel nodes and Nanjing Fuzimiao, which has become a must-visit attraction for most online travel users visiting Nanjing. Pioneer Bookstore, Nanjing University, 1912 Fashion Block, Yihe Road, Nanjing Yangtze River Bridge, Qixia Mountain, and other tourist nodes, on the other hand, constitute a negative and weak correlation with Nanjing Fuzimiao. The visit of these tourist nodes rather reduces the probability of visiting Nanjing Fuzimiao, which indicates that these tourist nodes are more different from Nanjing Fuzimiao and that these attractions can play a diversionary role for Nanjing Fuzimiao in the peak travel period.

In terms of the degree of lift (Lift) of the association rules, the range of the degree of lift for codes 11 to 20 is [1.8986, 9.2606], all of which are low. This indicates that the tourist nodes such as Pioneer Bookstore, Nanjing University, 1912 Fashion Block, Summer Palace Road, Nanjing Yangtze River Bridge, and Qixia Mountain constitute a negative and weak correlation with Nanjing Fuzimiao, and the visits to these tourist nodes rather reduce the probability of visiting Nanjing Fuzimiao. The tourist nodes are distinct from Nanjing Fuzimiao and can be an alternative for Nanjing Fuzimiao during the peak tourist season. The range of the degree of enhancement of code 1~10 is [29.1015, 93.7942], which is much higher than the degree of enhancement of number 11~20. This indicates that the enhancement degree of the tourism node combination association model of Nanjing Fuzimiao and Sun Yat-sen Mausoleum, Nanjing Fuzimiao and Jiangning Weaving Museum, Yuhuatai and Nanjing Fuzimiao, Nanjing Massacre Memorial Museum and Yunjin Museum, and Jiangnan Tribute Academy and Ancient Jixing Temple is high. There exists a large correlation, where the visit to the former tourism node of any of these combinations will greatly increase the probability of the visit to the latter tourism node. The association's recommendation based on the appearance of the first will be more recognized in tourism itinerary planning or tourism ticket recommendations.

Table 2. Results of association analysis

Encoding	Association rule	Support degree	Confidence degree	Lift
1	A1→A2	0.0080	0.2776	93.7942
2	A1→A3	0.0181	0.3011	82.5349
3	A4→A1	0.0084	0.4052	82.8847
4	B1→A1	0.0159	0.0496	60.0947
5	B2→B3	0.0082	0.1026	46.6881
6	B3→B2	0.0091	0.0919	45.4672
7	A1→B4	0.0468	0.0635	43.7940
8	C1→C2	0.0421	0.3155	34.1514
9	A1→B2	0.0181	0.3008	31.1849
10	A2→C3	0.0082	0.0230	29.1015
11	C4→D1	0.0031	0.0459	9.2606
12	A1→D2	0.0027	0.0275	3.6016
13	D3→A1	0.0019	0.0332	2.1587
14	D2→C2	0.0012	0.0165	4.1798
15	A1→D4	0.0018	0.0158	3.2661
16	A1→D5	0.0021	0.0099	2.0589
17	E1→C2	0.0011	0.0154	2.5436
18	A1→E2	0.0021	0.0237	2.1808
19	E3→E4	0.0014	0.0198	1.8986
20	A1→E5	0.0033	0.0209	2.2880

4.3 Assessment of Tourism Trend Forecasting Model Applications

4.3.1 Gray GM(1,7) simulation of China's outbound tourism arrivals

In this paper, we propose to use the multivariate gray prediction GM(1,7) model for the predictive modeling of China's outbound tourism trips. At the same time, we add two traditional prediction models, the univariate gray prediction GM(1,1) model, and the linear regression model, as the comparative models in order to complete the simulation simulation of China's data on the number of outbound tourism trips. Table 3 displays the specific simulation and prediction results.

Table 3. Simulation results of Chinese outbound tourism number by each model

Year	Actual value	GM(1,7) model		GM(1,1) model		Linear regression model	
		Simulated value	Relative simulation percentage error	Simulated value	Relative simulation percentage error	Simulated value	Relative simulation percentage error
2015	1.17	1.17	0	1.17	0	1.117	1.449
2016	1.14	1.16	1.509	1.12	1.89	1.042	2.299
2017	1.43	1.329	10.412	1.36	3.21	1.236	3.907
2018	1.37	1.403	1.995	1.49	1.37	1.531	1.913
2019	1.73	1.699	2.297	1.674	4.408	1.615	6.664
2020	0.93	0.889	3.674	0.96	2.06	0.968	3.403
2021	1.05	0.983	2.068	0.75	2.38	0.833	2.964
2022	0.87	0.799	1.265	0.97	4.82	0.834	0.402
2023	1.47	1.492	1.502	1.38	2.452	1.354	3.428
Integrated error		2.587%		3.483%		4.594%	

The fitting curves of the GM(1,7) model, GM(1,1) model, and linear regression model for the Chinese outbound tourism trips are shown in Figure 7. According to the simulation effect in Fig. 7, it can be seen that the fitting effect of the GM(1,1) model and the traditional linear regression model are very similar, and their simulated trend movements have a relatively close height. The gray GM(1,7) prediction model constructed in this paper is consistent with the original data in the trend direction, and its fitting effect is closer than the other two models in the simulation of the original data, except for the existence of a slightly larger deviation in the three years of 2017, 2021 and 2022.

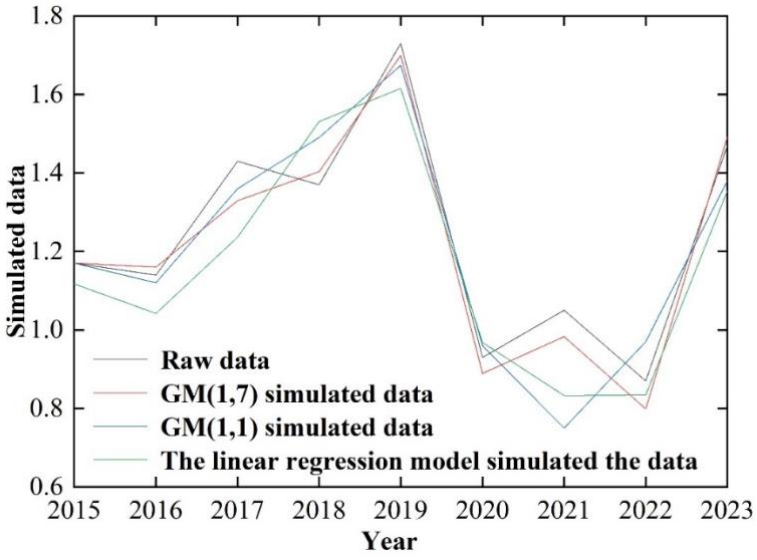


Figure 7. Fitting curves of each model to the number of Chinese outbound tourists

4.3.2 Model Performance Analysis

In this subsection, it is proposed to use the average relative error to evaluate and analyze the performance of different prediction models. The commonly used accuracy test reference standards are shown in Table 4. If the relative error is below 1%, the prediction model's model accuracy class is Class I. The prediction model's model accuracy class is Class II when the relative error is between 1% and 5%. When the relative error is higher than 5%, the model is usually difficult to be considered a qualified prediction model.

According to the experimental data in Table 3 and the error reference standards in Table 4, the average relative integrated error grades of the three prediction models can be determined. The experimental results show that the combined error of the GM(1,7) prediction model for China's outbound tourism trips is smaller compared with the combined errors of the other two prediction models, i.e., the GM(1,7) prediction model constructed in this paper has a higher modeling accuracy and a better prediction effect.

First, the simulation errors of the three models are greater than 1% and less than 5%, which indicates that the comprehensive performance of all three models is better. And on the other hand, it also shows the reliability of the two comparison models chosen.

The GM(1,7) model stands out among the three models due to its combined error of only 2.587%, which is close to me according to the prediction model rank judging scale. The GM(1,1) model, on the other hand, has a slightly worse performance with a prediction error of 3.483%. The prediction error of the linear regression model is 4.594%, and its grade is close to the II level according to the prediction model grade judging table.

Third, except for 2017, the simulation errors of the GM(1,7) model are all smaller, which is due to the model characteristics of GM(1, N). The GM(1, N) model has to adjust the data in the early stage in order to achieve the ideal expected prediction accuracy when modeling, adjusting for its oscillatory nature, and the errors are generally larger. When the prediction years are longer, the data in the previous period can be partially deleted, and the average error of the subsequent four years is only 2.098%, which indicates that the accuracy of the GM(1,7) model is higher.

Table 4. Prediction model accuracy level evaluation table

Model accuracy level	Critical points of relevant indicators			
	Relative error	Grey correlation degree	Posterior difference ratio	Small probability of error
I	0.01	0.90	0.35	0.95
II	0.05	0.80	0.50	0.80
III	0.10	0.70	0.65	0.70
IV	0.20	0.60	0.80	0.60

Predictive models are used to predict the future development trends of variables. The good performance of the model is mainly reflected in its prediction accuracy, and its pre-data is only used for modeling purposes. The GM(1,7) model is a better option for predicting the analysis of China's outbound traveler data than the GM(1,1) model and linear regression model.

5 System performance testing

The test indexes mainly refer to the number of concurrent users, the number of things per second (TPS), throughput, response time, resource utilization, and so on. In the performance test of this paper, the response time and TPS are mainly chosen to stress test the performance of the system at 100 user concurrency.

Mark the initial test time as 0; the average system response time obtained from the test for 12 consecutive hours is shown in Figure 8. From the test results, we can see that when 100 users operate the system, the minimum response time of the system is 0.87s, the maximum response time is 41.53s, and the average response time is 8.70s. During the whole test process, there is a request waiting phenomenon that leads to a response time of 41.53s, but in general, the system's performance is relatively stable. It can be maintained for some time. However, overall, the system performance is relatively stable and can be retained for a long period. As a whole, it is capable of meeting the system response time requirements.

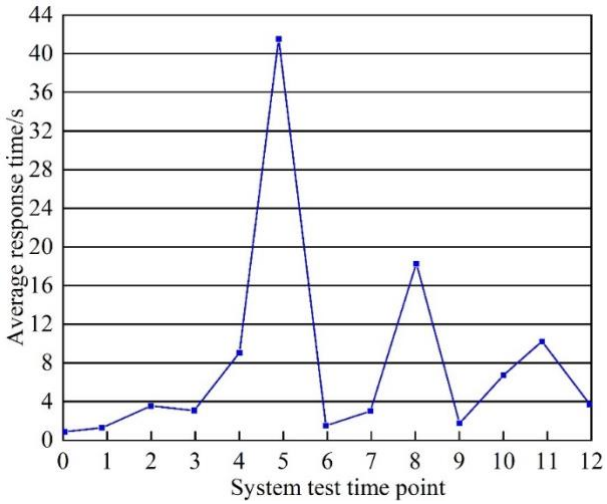


Figure 8. System response time

The number of things per second (TPS) at each time point during the test is shown in Fig. 9. TPS can reflect to some extent the maximum capacity of the system to process the business at the same time, and the higher the value, the better. In Fig. 9, it can be seen that the maximum TPS is 23.25, the minimum is 4.73, and the average value is 15.89. In the 0~4th hour and the 7th hour, the TPS value of the time point where the system response time increases has been reduced, which affects the system's ability to process the transactions, but overall, it is at a good level, which is able to satisfy the demand for the system's processing capability.

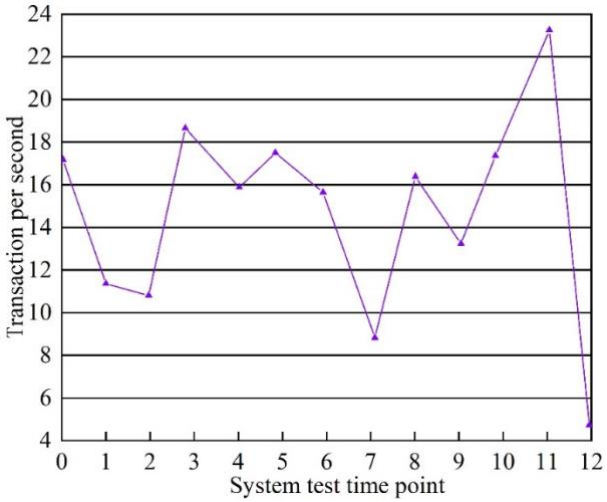


Figure 9. Transaction per second

6 Conclusion

In this paper, a decision support system for tourism management based on big data analytics is designed and tested after processing tourism data for classification, mining, and prediction.

When utilizing the plain Bayesian classification algorithm to classify tourist attractions, the correct rate of category classification and the number of samples used for training in that category show a significant positive correlation, which is in line with the law that the larger the sample size, the more knowledge the classifier can learn. The proposed method for evaluating attractions based on user and attraction information is highly feasible and has an accuracy rate of 72.98% for all categories.

When using the improved Apriori association algorithm for association mining of tourist nodes staying in Nanjing, the recommendation of tourist routes was realized based on the support, confidence, and association rule enhancement, and the feasibility was high.

When predicting the tourism trend, the gray GM(1,7) prediction model constructed in this paper is consistent with the original data in the trend direction. Its fitting effect is closer to the other two models in the simulation of the original data, except for a slight deviation in the three years of 2017, 2021, and 2022.

In the performance test of the tourism management decision support system constructed in this paper, the minimum response time, the maximum response time, and the average response time of the system are 0.87s, 41.53s, and 8.70s, respectively, which can satisfy the requirements of the system's response time as a whole. The TPS is 23.25 at the maximum, 4.73 at the minimum, and the average value is 15.89, which can satisfy the requirements of the system's processing capacity in general. Demand. It provides references for improving tourism decision-making technology and building a support system for tourism management decisions.

Fund projects:

1) Social Science Research Project of Fujian Provincial Department of Education: "Feasibility Study on Building a "New Rural Health Tour" Base in the Green Hinterland on the West Coast of the Taiwan Straits", No.: JAS160560.

2) Social Science Planning Project of Fujian Province: Innovative Research on Constructing a New Type of Health Tourism Chain in the Old Revolutionary Area on the West Coast of the Taiwan Straits Green-Red-Remember, No. FJJKCG16-411.

References

[1] Raúl Tarazona-Montoya, Peris-Ortiz, M., & Devece, C. (2020). The value of cluster association for digital marketing in tourism regional development. Sustainability, 12(23), 9887.

[2] Nuenen, T. V., & Scarles, C. (2021). Advancements in technology and digital media in tourism. Tourist Studies(1).

[3] Akhtar, N., Khan, N., Khan, M. M., Ashraf, S., Hashmi, M. S., & Khan, M. M., et al. (2021). Post-covid 19 tourism: will digital tourism replace mass tourism?. Sustainability, 13.

[4] Henar, Salas-Olmedo, Maria, Moya-Gomez, Borja, & Carlos, et al. (2018). Tourists' digital footprint in cities: comparing big data sources. TOURISM MANAGEMENT, 66(Jun.), 13-25.

[5] Ramos, V., Maurici Ruiz-Pérez, & Alorda, B. (2021). A proposal for assessing digital economy spatial readiness at tourism destinations. Sustainability, 13.

[6] Valls, F., & Roca, J. (2021). Visualizing digital traces for sustainable urban management: mapping tourism activity on the virtual public space. Sustainability, 13.

[7] Floros, C., Cai, W., Mckenna, B., & Ajeeb, D. (2019). Imagine being off-the-grid: millennials' perceptions of digital-free travel. Journal of Sustainable Tourism(6), 1-16.

[8] Kuzior, A., Lyulyov, O., Pimonenko, T., Kwilinski, A., & Krawczyk, D. (2021). Post-industrial tourism as a driver of sustainable development. Sustainability, 13(15), 8145.

[9] Iorio, C., Pandolfo, G., D'Ambrosio, A., & Siciliano, R. (2020). Mining big data in tourism. Quality & Quantity, 54(2).

[10] Filieri, R., D'Amico, E., Destefanis, A., Paolucci, E., & Raguseo, E. (2021). Artificial intelligence (ai) for tourism: an european-based study on successful ai tourism start-ups. International journal of contemporary hospitality management(11), 33.

[11] Luis, E., Boavida-Portugal Inês, Carlos, C. F., & Jorge, R. (2017). Identifying tourist places of interest based on digital imprints: towards a sustainable smart city. Sustainability, 9(12), 2317.

[12] Poux, F., Valembois, Q., Mattes, C., Kobbelt, L., & Billen, R. (2020). Initial user-centered design of a virtual reality heritage system: applications for digital tourism. Remote Sensing, 12(16), 2583.

[13] Qiong, S., Xiankai, H., & Zheng, L. (2022). Tourists' digital footprint: prediction method of tourism consumption decision preference. The Computer Journal(6), 6.

[14] Martins, J., Goncalves, R., Au-Yong-Oliveira, M., Moreira, F., & Branco, F. (2020). Qualitative analysis of virtual reality adoption by tourism operators in low-density regions. IET software(6), 14.

[15] Hu, M., Li, H., Song, H., Law, R., & Li, X. (2022). Tourism demand forecasting using tourist-generated online review data. Tourism management(Jun.), 90.

[16] Mai, Thanh, Smith, & Carl. (2018). Scenario-based planning for tourism development using system dynamic modelling: a case study of cat ba island, vietnam. Tourism Management.

[17] Lu, Y., & Cui, B. (2022). Intelligent tourism marketing and publicity methods for revenue enhancement. Mobile information systems(Pt.24), 2022.

[18] Du, J. (2021). Research on intelligent tourism information system based on data mining algorithm. Mobile Information Systems.

[19] Wei, C., Wang, Q., & Liu, C. (2020). Research on construction of a cloud platform for tourism information intelligent service based on blockchain technology. Wireless Communications and Mobile Computing.

[20] Chen, J., Xue, H., & Tsa, S. B. (2021). An empirical study on intelligent rural tourism service by neural network algorithm models. Complexity.

[21] Yu, Y. (2022). Analysis and study on intelligent tourism route planning scheme based on weighted mining algorithm. Scientific programming(Pt.15), 2022.