Research Article

Real-Time Forecast of Tourists Distribution Based on the Improved k -Means Method

Peiyu Ren, Zhixue Liao, and Peng Ge

Business School, Sichuan University, Chengdu 610064, China

Correspondence should be addressed to Zhixue Liao; law1219@126.com

Received 29 March 2014; Revised 1 May 2014; Accepted 15 May 2014; Published 5 June 2014

Academic Editor: Ker-Wei Yu

Copyright © 2014 Peiyu Ren et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tourist distribution, a vector to reflect the tourist number of every scenic spot in a certain period of time, serves as the foundation for a scenic spots manager to make a schedule scheme. In this paper, a forecast model is offered to forecast tourist distribution. First of all, based on the analysis of changing mechanism of tourist distribution, it is believed that the possibility for a scenic spot to have the same tourist distribution next time is high. To conduct this forecast, we just need to research on the similar tourist distributions of which time and tourist scale are close. Considering that it is time-consuming, an improved K -means cluster method is put forward to classify the historical data into several clusters so that little time will be needed to search for the most similar historical data. In the end, the case study of Jiuzhai Valley is adopted to illustrate the effectiveness of this forecast model.

1. Introduction

Tourism, which benefits transportation, accommodation, catering, entertainment, and retailing sectors, has become a quite prosperous industry in the past few decades [1]. In 1960, the international tourist arrivals were only 69.3 million, while in 2010 the number was 935 million, and that meant an average growth rate of 5.3 per year [2]. The UNWTO forecasted that international tourism demand will double by 2020 to 1.6 billion visitors and it would generate nearly \$2 trillion dollars in economic activities [3]. China's tourism industry develops faster than other countries as the Chinese people's living standards improved significantly along with the rapid development of national economy. As long as a scenic spot can successfully attract tourists and expand its bearing capacity, it will bring huge economic benefits. Therefore, in China, especially the western area which lacks natural resources, tourism has become a driving force of regional economic development. With the rapid development of tourism industry and the rapid increase of tourists, the scenic area has become increasingly congested, so that safety risk is increased and visitors experience is not so good. Congestion can be divided into two categories, true congestion and spurious congestion. The former means the tourist number is far beyond the capacity of scenic

spot, while the latter means tourist number is about to surpass the capacity, and, because of tourist scheduling delay, tourist distribution is of disequilibrium; namely, some spots are crowded while some are vacant. For the first kind of congestion, the capacity should be expanded, while, for the latter, effective scheduling can be used. Nowadays, shunt scheduling has two main problems. The first one is delay. The existing scheduling mode belongs to the schedule that scheduling instruction is offered only after the crowd comes into being, so scheduling instruction is delayed because when the scheduling began to take effect, the congestion degree of the scenic spot has changed, which in the end greatly reduced its effect. The second is that the present scheduling lacks hierarchy so that it fails to effectively control the congestion. A good solution to this problem is the following: managers could accurately forecast the amount of tourists arrivals and forecast under different resolutions. Real-time forecast can help scheduling scheme design in different levels of each node and effectively improve the utilization rate of resources. As a result, the real-time forecast on the amount of tourists arrivals has become the research focus.

The researches on forecasting of tourists quantity began in the 1960s, and, in recent years, scholars have made great efforts to improve the accuracy of the forecast. Exponential model, ARIMA model, and GARCH model were built to

produce accurate forecasting (Kim and Ngo [4], 2001; Chen [1], 2011; Law [5, 6], 2000a, 2004; Qu and Zhang [7], 1996). Some scholars think that nonlinear method can improve the forecasting accuracy because the data characteristics of the tourists quantity is not linear. Law and Au [8] (1999) built the first neural network model to forecast Japanese's demand to travel to Hong Kong, showing that neural network model performed well in forecasting the tourism demand. Since then neural network model had attracted lots of attention to forecasting such as Cho [9] (2003), Law [10] (2000), Taskaya-Temiz and Casey [11] (2005), and Palmer et al. [12] (2006). It was demonstrated that no single forecasting model could produce the best forecast in all situations by empirical results (Cho [13], 2001). Furthermore, there were no certain criteria being used to choose a certain forecasting method which can produce the best performance when a particular tourism demand forecasting was performed. For example Smeral and Wüger [14] (2005) proposed that ARIMA model performed better than exponential model; however Yang et al. [15] (2011) confirmed that ARIMA model performed worse comparing with exponential model.

From the above analysis, it is found that researches on methods to forecast tourist arrivals have been done by lots of researchers, but the research on the real-time forecast is few. Qiu et al. [16] analyzed the space distribution features of tourist flow. Witt and Song [17] pointed out that the tourists distribution is imbalanced in terms of time and space. Yan and Meng [18], Lu et al. [19], and Yan et al. [20] analyzed time and space distribution features of the tourists. Liang and Bao [21] analyzed the seasonal features of theme parks visitors, its influencing factors, and the tourist flow fluctuation law in busy seasons. These studies only focused on the characteristics of tourists' distribution in terms of space and time so that in-depth dynamic forecast researches on nodes in scenic spots are needed.

Based on the analysis of the tourist number of each scenic node in a scenic spot, we initiate a clustering algorithm to search the nearest tourist distribution. Then the algorithm is used to forecast the future tourist distribution of a scenic spot so as to testify the previous forecast. At last, an empirical research is done on Jiuzhai Valley.

Section 2 describes the problem and analyzes tourist distribution forecast in a scenic spot. In Section 3, an improved K -means algorithm is proposed to solve the forecast model. Section 4 takes Jiuzhai Valley as an example to verify the effectiveness of the tourist distribution forecast model for a scenic spot and Section 5 draws a conclusion.

2. Problem Analysis

For a scenic spot with only one entrance, one exit, and many scenic nodes, if we can forecast the tourist number of each node at any time, then the schedule schemes of the scenic spot will become easy and timely. In order to make this real-time forecast, we need to study historical tourist distribution data of the scenic spot. Tourist distribution is defined as

$$N(t) = (N(t)_1, N(t)_2, \dots, N(t)_n). \quad (1)$$

In (1), tourist distribution $N(t)$ at time t is a n -dimension vector which represents n scenic spots' tourists number $N(t)_i$, $i = 1, 2, \dots, m$.

It is clear that the possibility for a scenic spot to have the same tourist distribution next time is high if time and tourist scale are close. For example, time t_1 and t_2 are similar times of different days and $|t_1 - t_2|$ is small. If their respective tourist distributions $N(t_1)$ and $N(t_2)$ are very similar and tourist scales $S(t_1)$ and $S(t_2)$ are also close, then the possibility for their next-time tourist distributions $N(t_1 + 1)$ and $N(t_2 + 1)$ to become similar will be high. So, if, at time t , we want to forecast tourist distribution $N(t + 1)$, we only need to know tourists distribution $N(t')^*$ is close to $N(t)$ and the next-time tourist distribution $N(t' + 1)^*$ of $N(t')^*$ can be taken as the forecast value of $N(t + 1)$.

However, as the purpose is to make real-time forecast of each minute and scenic spots' open duration is generally more than 10 hours which means there are 600 data a day, it is time-consuming to find the similar tourists distribution N_t^* . In order to make the search efficient, we can classify historical data into several clusters by clustering algorithm and then judge which cluster $N(t)$ is belonged to. Thanks to the cluster, we can find the similar tourists distribution $N(t)^*$ within time $(1/k)B$ on average (k is the cluster numbers and B is the time consumed to find $N(t)^*$ without clustering).

3. Modeling

3.1. Principles of Forecast. From the above analysis, to realize the real-time tourist distribution forecast, we should firstly classify the tourist distribution samples into several clusters and abide by the principle that in the same cluster the objects should be similar in respect of tourist distribution, time of that tourist distribution, and the tourist scale at that time. As a result, we should add elements (time of that tourist distribution and the tourist scale at that time) into the tourist distribution $N(t) = (N(t)_1, N(t)_2, \dots, N(t)_n)$ and the new tourist distribution is defined as the time-scale distribution of the tourist distribution $N(t) = (N(t)_1, N(t)_2, \dots, N(t)_n)$.

So let $O = (O^1, O^2, \dots, O^{n+1}, O^{n+2})$ be the time-scale distribution of the tourist distribution $N(t)$ in which O_1 represents time t of the tourist distribution $N(t)$, O_{n+2} represents tourist scale, and $(O^2, \dots, O^{n+1}) = N(t) = (N(t)_1, N(t)_2, \dots, N(t)_n)$ represents tourist number of node n . So time-scale distribution O and tourist distribution $N(t)$ have the following connection:

$$N((O^1)) = (O^2, O^3, \dots, O^{n-1}). \quad (2)$$

$O = (O^1, O^2, \dots, O^{n+1}, O^{n+2})$ can be classified into many clusters based on the cluster algorithm. In the same cluster, tourist distribution N_t , time t of the tourist distribution, and the tourist scale are similar. So, if a certain data $O(i)$ is judged to belong to cluster V_j , then the next-time tourist distribution $N(O^1(i) + 1)$ can be estimated ($O^1(i)$ is the first element of $O(i)$, so it refers to the time) in that it is highly possible that the similar distribution jumps to the similar cluster next time.

FIGURE 1: Changing mechanism of tourist distribution.

FIGURE 2: Forecast process.

For example, in Figure 1, a tourist distribution data $N = (N_1, N_2, \dots, N_n)$ has 3 clusters $V = (A, B, C)$. Any distribution $N(t)_i$ which belongs to cluster A can be changed under three kinds of circumstances. Firstly, next-time distribution $N(t+1)$ may just remain in the same cluster A with the distribution $O(t)$. Secondly, next-time distribution $N(t+1)$ may jump to cluster B. Thirdly, next-time distribution $N(t+1)$ may jump to cluster C.

At time t , the tourist distribution is $N(t) = (N(t)_1, N(t)_2, \dots, N(t)_n)$, and the sample distribution is $N(t)^* = (N(t)_1^*, N(t)_2^*, \dots, N(t)_n^*)$, which is the most similar one to $N(t) = (N(t)_1, N(t)_2, \dots, N(t)_n)$ in the same cluster. In this situation, tourist distribution $N(t+1) = (N(t+1)_1, N(t+1)_2, \dots, N(t+1)_n)$ of next time $t+1$ should be similar to the $N(t+1)^* =$

$(N(t+1)_1^*, N(t+1)_2^*, \dots, N(t+1)_n^*)$. So $N(t+1)^* = (N(t+1)_1^*, N(t+1)_2^*, \dots, N(t+1)_n^*)$ can be taken as the forecast value of $N(t+1) = (N(t+1)_1, N(t+1)_2, \dots, N(t+1)_n)$. The detailed forecast process is shown in Figure 2.

3.2. K-Means Algorithms. Let $O = \{O_1, O_2, O_3, \dots, O_m\}$ be a set of m data objects and each object has $n+2$ natures; then these data can also be represented by a profile data matrix $O_{m \times n+2}$ that has m row vectors and each row vector has $n+2$ dimensions to represent a data object. In this paper, a data object represents a time-scale distribution, so $O_{m \times n+2}$ is the dataset of the time-scale distribution. The i th row vector $O_i = \{O_i^1, O_i^2, O_i^3, \dots, O_i^{n+2}\}$ denotes the i th time-scale distribution from the dataset $O_{m \times n+2}$ and in O_i the first element O_i^1 , and it represents time t , the last element represents tourist scale and other elements represent tourist distribution $N(t)$ at time t . With $O_{m \times n+2}$, the aim of the k -means clustering algorithm is to find a partition $C = C_1, C_2, \dots, C_k$ of k groups. The partition can generate time-scale distributions in the same group that are as similar to each other as possible, while time-scale distributions of different groups are as different as possible. The k -means clustering algorithm's main idea is stated as follows.

- (1) Firstly, randomly select k initial centroids $C = C_1, C_2, \dots, C_k$ and then assign each time-scale distributions O to the closest centroids C_i , $i = 1, 2, \dots, k$. A given time-scale distribution O_i will be assigned to centroid C_t if and only if the following condition holds:

$$\|O_i - C_t\|^2 \geq \|O_i - C_j\|^2, \quad t, j = (1, 2, \dots, k), \quad (3)$$

$$i = (1, 2, \dots, m),$$

where $\|O_i - C_t\|^2$ is the distance between object O_i and centroid C_t .

The distance can be calculated by the following function: $\|O_i - C_t\|^2 = \sum_{\iota=1}^{n+2} (O_i^\iota - C_t^\iota)^2$, where $\iota = (1, 2, \dots, n+2)$ denotes ι th component or feature of the corresponding data object.

- (2) Secondly, generate new centroid by calculating the mean of the objects set assigned to each cluster. The new centroids $C' = C'_1, C'_2, \dots, C'_k$ will be

$$C'_j = \frac{1}{M_j} \sum_{O_i \in C_j} O_i, \quad j = (1, 2, \dots, k), \quad (4)$$

$$i = (1, 2, \dots, n+2),$$

where M_j denotes the number of the objects belonging to cluster C_j , $j = (1, 2, \dots, k)$.

Then repeating the above step until centroids $C = C_1, C_2, \dots, C_k$ remains unchanged. The k -means clustering algorithm is shown in Algorithm 1.

Input: the dataset O , and the number of clusters k
Output: Set of k clusters $S = \{S_1, S_2, \dots, S_K\}$
Initialization: $s = \{\}$, Randomly, select k initial centroids
Begin
 Repeat
 Step 1. Vectors are assigned to the closest centroid
 Step 2. The objective function J is computed
 Step 3. New centroids are calculated as the mean of the vectors assigned to each cluster
 Until Convergence (or no change in the objective function)
Return S
End

ALGORITHM 1: K-means clustering: $KM(O, k)$.

FIGURE 3: Truth value and forecast value from 11:01 to 11:15 on May 7, 2013.

FIGURE 4: Truth value and forecast value from 11:16 to 11:30 on May 7, 2013.

3.3. The Improved K-Means Algorithms. In the forecast of tourist distribution, cluster number k has an impact on the accuracy of forecast which is the key to the k -means clustering algorithm. However, the k -means clustering algorithm cannot give cluster number k . To solve this problem, we will propose improved k -means algorithms which can ensure cluster number k to make a better forecast.

3.3.1. Select the Two Optimal Centroids. To make sure that objects in the same group are as similar to each other as possible and objects of different groups differ as much as possible, we should select two centroids whose distance is the farthest when compared with any other two objects. So

they are the two optimal centroids that can be obtained through the method below.

- (1) Randomly select an object O_t from the object data $O = \{O_1, O_2, \dots, O_m\}$, and then calculate the distance d_{ti} between object O_t and any other object O_i :

$$d_{ti} = \sum_{i=1}^{n+2} (O_t^i - O_i^i)^2, \quad i = (1, 2, \dots, n+2), \quad (5)$$

$$i = (1, 2, \dots, n), \quad \text{but } i \neq t.$$

- (2) Select the $\max(d_{ti})$ and the farthest object O_i .
- (3) Calculate distance d_{ij} between the object O_i and any other object O_j and then get the farthest object O_j .

FIGURE 5: Truth value and forecast value from 11:31 to 11:45 on May 7, 2013.

- (4) O_i, O_t are the two optimal centroids if $t = j$ or $d_{ij} < d_{ti}$; else, O_i, O_j are the two optimal centroids,
 (5) C_1, C_2 represent the two optimal centroids.

3.3.2. *Confirmation of the Cluster Number k .* After obtaining the two centroids, we need to determine the cluster number k . The method is stated as follows.

- (1) Select a new centroid.

By calculating the distance d_{ij} between the existing centroids C_j (the current cluster number) and another object O_i we can get each minimum distance $d_i = \min\{d_{i1}, d_{i2}, \dots, d_{ij}, \dots, d_{iq}\}$ of object O_i and object O_k may be taken as a new centroid (named C_{q+1}) if $d_k = \min\{d_1, d_2, \dots, d_i, \dots, d_m\}$.

- (2) Judge whether the new centroid is good or not.

By assigning every other object O to the closest centroid with (1) and (2), the $q+1$ clusters V can be obtained. The new centroid $C' = C'_1, C'_2, \dots, C'_{q+1}$ can be calculated by

$$C'_i = \frac{1}{M_i} \sum_{O_j \in V_i} O'_j, \quad i = (1, 2, \dots, n+2), \quad (6)$$

$$i = (1, 2, \dots, q+1),$$

where C'_i represents i th feature of centroids C'_i in cluster V_i .

Definition 1. Let $R_{i,j}$ be the intercluster distance, which represents the distance between the i th cluster V_i and the j th cluster V_j . Thus, it can be calculated by

$$R_{ij} = \|\overline{V_i}, \overline{V_j}\| = \sum_{i=1}^{n+2} (C'_i - C'_j)^2, \quad i, j = (1, 2, \dots, q+1), \quad \text{but } i \neq j. \quad (7)$$

Definition 2. Let U_i be the intracluster distance which represents the average distance among the objects $O_i \in V_i$; then it can be calculated by

$$U_i = \frac{1}{M_i} \sum_{O_i \in V_i} \sum_{i=1}^{n+2} (O'_i - C'_i)^2, \quad i = (1, 2, \dots, q+1). \quad (8)$$

Definition 3. Let I be evaluation index which indicates whether the new centroid C_{q+1} is necessary; then I is defined as

$$I_{q+1} = \frac{1}{q+1} \sum_{i=1}^q \max \left(\frac{U_i + U_j}{R_{ij}} \right), \quad j = (1, 2, \dots, q+1). \quad (9)$$

A better cluster algorithm must conform with two conditions. Firstly, objects in the same cluster should be as similar as possible which means intracluster distance U_i must be as small as possible. Secondly, objects in different clusters should be as different as possible which means intercluster

FIGURE 6: Truth value and forecast value from 11:46 to 12:00 on May 7, 2013.

FIGURE 7: Truth value and forecast value from 12:01 to 12:15 on May 7, 2013.

distance R_{ij} must be as big as possible. In (7), we can find that I increases while both U_i and R_{ij} decrease. So I can be regarded as the evaluation index to reflect whether creating

a new centroid is reasonable. In this case when centroid q has been created, we just need to compare I_q with I_{q+1} . If $I_q > I_{q+1}$, then the new centroid is reasonable, and a new cluster by above method. Else, the new centroid is unreasonable and the optimal $k = q$.

The improved k -means algorithm is based on the k -means algorithm, and it has not set the cluster number k , which is determined by the data. So the improved k -means algorithm can perform better than k -means algorithm. The comprehensive description of the improved k -means algorithm is shown in Algorithm 2.

3.4. Forecast Method. In order to forecast tourist distribution, we should judge which cluster the current tourist distribution's time-scale distribution O_i belongs to and then find out the most similar time-scale distribution O_j in that cluster. Therefore, we can take next-time tourist distribution $N(O_j^1 + 1)$ of the similar tourist distribution $N(O_j^1)$ as the forecasting value for the next minute. The detailed forecast process is shown in Figure 2.

4. Empirical Analysis

4.1. Jiuzhaigou Introduction and Errors Evaluation Method. Jiuzhai Valley scenic spot is located in western China and is complimented as a world-class natural scenic spot. The tourism industry there greatly promoted the economic development of western China. The number of tourists every day during statutory holidays, such as Chinese National Day Golden Week, can be as many as 20,000. In the past two years,

FIGURE 8: Truth value and forecast value from 12:16 to 12:30 on May 7, 2013.

as people's living standard improved and the governments vigorously support tourism, tourist number has increased greatly. For example, during the National Day Golden Week in 2012, the tourist number in Jiuzhai Valley per day was more than 40,000. The peak season has posed a great challenge to tourism management so that it is important to forecast tourist distribution.

Jiuzhai Valley Scenic Spot Administration has carried out Digital Jiuzhai Valley Comprehensive Demonstration Project (Digital Jiuzhai Valley for short). Technologies such as RFID and GIS are widely applied in the scenic spot; digitization systems such as Access Control System and RFID card reader have been distributed in each attraction of the scenic spot so that all kinds of data can be collected. In this paper, Jiuzhai Valley is taken as an example to testify the forecast model.

In order to evaluate the effectiveness of the forecast model, this paper takes mean absolute percentage error (MAPE) as evaluation statistics, and the calculating method is as follows:

$$\text{MAPE} = \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \left| \frac{y(t)^i - \hat{y}(t)^i}{y(t)^i} \right|. \quad (10)$$

As MAPE is the mean value of all times forecast errors, we should analyze each times error to find the major characteristics of the errors. We set relative error and its calculating method is as follows:

$$\text{RE}(t) = \sqrt{\frac{\sum_{i=1}^n (y(t)^i - \hat{y}(t)^i)^2}{\sum_{i=1}^n (y(t)^i)^2}}, \quad (11)$$

FIGURE 9: Truth value and forecast value from 12:31 to 12:45 on May 7, 2013.

where $\hat{y}(t)$ represents forecasting result of the tourist distribution and $y(t)$ the truth tourists distribution at time t .

Through (10), it is clear that if the forecast value is closer to the real value, then the result of the equation will be smaller. So the equation can be used to evaluate the effectiveness of the forecast model.

4.2. Data. We have collected data of Jiuzhai Valley's tourist distribution and its corresponding time-scale distribution from May 1 to May 7 in 2012 and 2013, respectively. The data of May 7, 2013, has been used to test the accuracy of the forecast model and other data are taken as the training data to get the clusters and their centroids by the improved k -means cluster algorithm.

4.3. Clustering. With the help of the improved k -means cluster algorithm, we have obtained 9 clusters and their corresponding centroids C are shown in Table 1.

4.4. Forecast. Based on the forecast method proposed in this paper, we have forecast the tourist distribution of the 120 minutes (from 11:01 to 13:00) on May 7, 2013 (see Figures 3, 4, 5, 6, 7, 8, 9, and 10), and then obtained the MAPE = 5.32% by (10). In order to analyze the errors at different times, $\text{RE}(t)$ is calculated by (11) (see Figure 11).

From Figure 11, we can find that 4 forecast errors are less than 3%, and they contribute to 2.9% of the total; 27 are less than 4% but more than 3%, and they contribute to 22.5% of the total; 20 are less than 5% but more than 4%, and they contribute to 16.7% of the total; 26 are less than 6% but more than 5%, and they contribute to 21.7% of the total; 16 are less

TABLE 1: The centroids C .

Cluster	Centroid									
C_1	349	23	13	16	436	618	682	62	450	2300
C_2	241	155	437	544	12	310	138	351	74	2021
C_3	281	225	65	75	27	386	358	74	109	1319
C_4	400	18	11	16	371	179	543	205	226	1569
C_5	311	163	31	37	138	619	648	44	215	1895
C_6	245	93	201	262	9	308	122	214	78	1287
C_7	345	33	21	26	257	420	455	64	230	1506
C_8	410	10	1	16	238	232	364	137	439	1437
C_9	395	15	9	16	377	322	804	176	298	2017

FIGURE 10: Truth value and forecast value from 12:46 to 13:00 on May 7, 2013.

than 7% but more than 6%, and they contribute to 13.3% of the total; 16 are less than 8% but more than 7%, and they contribute to 13.3% of the total; 7 are less than 9% but more than 10%, and they contribute to 5.8% of the total; and 3 are (10.05%, 11.54%, and 10.22% at 11:35, 11:38, and 11:50, resp.) more than 10%, and they contribute to 2.5% of the total. So we can come to the conclusion that the forecast model is effective.

5. Conclusion

Based on the improved k -means cluster algorithm, a forecast model is put forward to forecast the tourist distribution which

FIGURE 11: Forecast relative errors from 11:01 to 13:00 on May 7, 2013.

serves as the foundation for the scheduling. Firstly, with the improved k -means cluster algorithm, time-scale distribution data O is classified into several clusters and then we come to judge which cluster it belongs to. Because $N(t)$ consists of elements $[O_2, O_3, \dots, O_{n-1}]$ in O_i of which the first element $O_i^1 = t$, tourist distribution $N(t)$ can also be classified into the same cluster with O_i , of which the first element $O_i^1 = t$ and $N(t)$ belongs to the same cluster with its time-scale distribution O_i , of which the first element $O_i^1 = t$. Secondly, research is made on the most similar historical data $N'(t')$ in the same cluster with tourist distribution $N(t)$ which is being used to test the forecast model, and tourist distribution $N'(t' + 1)$ is taken as the forecast value of the tourist distribution $N(t + 1)$. Finally, the empirical study in Jiuzhai Valley illustrates that the forecast model is effective.

Input: The dataset O
 Output: Set of k clusters $S = \{S_1, S_2, \dots, S_K\}$ (k is not a certain value)
 Initialization: Select two objects with maximum distance, produce the initial centroids $S = \{S_1, S_2\}$
Begin
 Repeat
 Step 1. Find the longest distance between objects with centroids, and take the it as new centroids
 Step 2. Assign the other objects to the closest centroid
 Step 3. New centroids are calculated as the mean of the objects assigned to each cluster
 Step 4. Calculate the inter-cluster distance, intra-cluster distance and the evaluation index I_{new}
 Until $I_{\text{new}} > I_{\text{last}}$
 Return S
End

ALGORITHM 2: Improved K -means clustering: IKM(O).

Conflict of Interests

The authors (Peiyu Ren, Zhixue Liao, and Peng Ge) declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Major International Joint Research Programme of the National Natural Science Foundation of China (Grant no. 71020107027), the National High Technology Research and Development Major Programme of China (863 Programme) (Grant no. 2008AA04A107), the National Natural Science Foundation of China (Grant no. 71001075), the National Natural Science Foundation of China (Grant no. 71371130), the China Postdoctoral Science Foundation (Grant no. 2012M521704), Doctoral Fund of Ministry of Education of China (20110181110034), 985 and 211 Projects of Sichuan University, and Central University Fund of Sichuan University.

References

- [1] K. Y. Chen, "Combining linear and nonlinear model in forecasting tourism demand," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10368–10376, 2011.
- [2] H. Y. Song and R. J. Hyndman, "Tourism forecasting: an introduction," *International Journal of Forecasting*, vol. 27, no. 3, pp. 817–821, 2011.
- [3] S. E. Levy and D. E. Hawkins, "Peace through tourism: commerce based principles and practices," *Journal of Business Ethics*, vol. 89, no. 4, pp. 569–585, 2009.
- [4] J. H. Kim and M. T. Ngo, "Modelling and forecasting monthly airline passenger flows among three major Australian cities," *Tourism Economics*, vol. 7, no. 4, pp. 397–412, 2001.
- [5] R. Law, "Demand for hotel spending by visitors to Hong Kong: a study of various forecasting techniques," *Journal of Hospitality and Leisure Marketing*, vol. 6, pp. 17–29, 2000.
- [6] R. Law, "Initially testing an improved extrapolative hotel room occupancy rate forecasting technique," *Journal of Travel and Tourism Marketing*, vol. 16, pp. 71–77, 2004.
- [7] H. Qu and H. Q. Zhang, "Projecting international tourist arrivals in East Asia and the Pacific to the year 2005," *Journal of Travel Research*, vol. 35, no. 1, pp. 27–34, 1996.
- [8] R. Law and N. Au, "A neural network model to forecast Japanese demand for travel to Hong Kong," *Tourism Management*, vol. 20, no. 1, pp. 89–97, 1999.
- [9] V. Cho, "A comparison of three different approaches to tourist arrival forecasting," *Tourism Management*, vol. 24, no. 3, pp. 323–330, 2003.
- [10] R. Law, "Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting," *Tourism Management*, vol. 21, no. 4, pp. 331–340, 2000.
- [11] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5–6, pp. 781–789, 2005.
- [12] A. Palmer, J. J. Montañó, and A. Sesé, "Designing an artificial neural network for forecasting tourism time series," *Tourism Management*, vol. 27, no. 5, pp. 781–790, 2006.
- [13] V. Cho, "Tourism forecasting and its relationship with leading economic indicators," *Journal of Hospitality and Tourism Research*, vol. 25, pp. 399–420, 2001.
- [14] E. Smeral and M. Wüger, "Does complexity matter? Methods for improving forecasting accuracy in tourism: the case of Austria," *Journal of Travel Research*, vol. 44, no. 1, pp. 100–110, 2005.
- [15] X. Z. Yang, C. L. Gu, and Q. Wang, "Study on the driving force of tourist flows," *Geographical Research*, vol. 30, pp. 23–36, 2011.
- [16] Y. Q. Qiu, P. Ge, and P. Y. Ren, "A study on temporal and spatial navigation based on the load-Balance of tourists in Jiuzhaigou Valley," *Resources Science*, vol. 32, pp. 118–123, 2010.
- [17] S. F. Witt and H. Song, "Forecasting tourism flows," in *Tourism and Hospitality in the 21st Century*, vol. 3, pp. 106–118, 2002.
- [18] F. Yan and J. D. Meng, "The application of logistic growth model in forecast of tourists amount: a case study of Suiyang, Guizhou Province," *Human Geography*, vol. 20, pp. 87–91, 2005.

- [19] S. Lu, L. Lu, and L. Wang, "Temporal characteristics of tourist flows to ancient villages: a case study of two world cultural heritages, Xidi Village and Hongcun Village," *Geographica Sinica*, vol. 24, pp. 250–256, 2004.
- [20] L. Yan, X. G. Xu, and X. P. Zhang, "Analysis to temporal characteristics of tourist flows on Jiuzhaigou world natural heritage," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 45, no. 1, pp. 171–177, 2009.
- [21] Z. X. Liang and J. G. Bao, "A seasonal study on tourist flows in theme parks during golden weeks: a case of theme parks in Shenzhen Overseas Chinese Town," *Tourism Tribune*, vol. 27, pp. 58–65, 2012.