Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 4, No. 3, 2022

Clustering Analysis of Airport Traffic Similar Days Affected by Epidemic Based on HI-K-means

Derui Kong[1, a], Manzhen Duan[1, b], Ranran Shang[1, c], Yinfeng Li[1, d]

[1]College of Civil and Architectural Engineering, North China University of Science and Technology, Tangshan 063210, Hebei, China
[a]1243188378@qq.com, [b]mz06ss@ncst.edu.cn, [c]1334691069@qq.com, [d]qzll6833@israelar.com

Abstract: During the epidemic period, the daily traffic fluctuation of the airport has a strong correlation with the current control policy. The analysis of traffic similar days can provide a reference for the optimization of airport traffic management. Aiming at the analysis of traffic similar days, a clustering model of traffic similar days based on HI-K-means (Hierarchical k-means clustering algorithm) is proposed. This algorithm combines the advantages of Hierarchical clustering and the K-means clustering algorithm and makes up for the defects of the two algorithms. Taking Tianjin Binhai Airport as an example, cluster analysis is carried out. Finally, it is concluded that the three types of traffic similar days can better match the daily traffic under different policies, indicating that the model has strong availability and high accuracy.

Keywords: Similar day, Hierarchical clustering, K-means, Airport traffic.

1. Introduction

As the connecting area on both sides of the air and field, the airport has complex traffic operations and a large control workload[1]. To improve the forward-looking and predictive ability of airport traffic flow management, the future traffic flow trend can be determined based on the analysis of the airport's historical traffic flow. Since 2020, the novel coronavirus epidemic has had a profound and significant impact on transportation[2-5]. Due to the changeable traffic control policies during the epidemic period and the rapid changes in airport traffic characteristics, a computationally efficient and simple clustering method is needed to analyze airport traffic.

Using similar days obtained by clustering to analyze airport traffic has become a new idea and method[6-7]. Cluster analysis has achieved effective application in many fields because of its simple and effective characteristics[8]. In the aviation field, Grabbe et al. used the maximum expectation algorithm to cluster the airport's weather and expected arrival rate as features to analyze the probability of implementing ground delay procedures under different scenarios[9]. Based on the analysis of traffic characteristics in the terminal area, Shang Ran et al. extracted the daily feature vector, constructed a similar day clustering model based on the SOM-K-means algorithm, and conducted an example analysis to objectively summarize the typical traffic scenarios[10]. Rao et al. proposed a multivariate trajectory deep clustering framework based on a deep neural network and used it for air traffic flow identification and anomaly detection[11]. Chen et al. used machine learning technology to cluster weather factors such as meteorology and wind conditions, and output the results of similar effects of weather on the airport[12]. Based on the historical weather data of Kunming terminal, Shan et al. used the K-mean algorithm for cluster analysis to obtain 7 common instrument flight rule weather types that affect flight operation[13]. Xu et al. used a self-organizing map (SOM) neural network algorithm to cluster the weather at Shanghai Pudong International Airport[14]. Chen et al. studied the prediction of airport peak service rate based on weather classification and used the K-means algorithm to construct an airport peak service capacity analysis model for three weather types.

In summary, the above researches mostly use a single algorithm, and the single algorithm has certain defects. For example, although the K-means algorithm has fast convergence speed and excellent clustering effect, the selection of the k value is not easy to grasp; the SOM algorithm has strong fault tolerance, but the operation efficiency is low. To improve the clustering analysis effect of airport traffic under the influence of the epidemic situation, this paper constructs a traffic-similar day clustering model based on the HI-K-means clustering algorithm based on airport traffic feature extraction. This algorithm combines the advantages of hierarchical clustering and K-means algorithm efficiency and makes up for the defects of the two algorithms.

2. Clustering Feature Extraction

Because the original data volume is large and there are many irrelevant data items, it is necessary to clean the data, eliminate irrelevant data items such as models and parking spaces, and delete missing and abnormal data samples. Based on the cleaned data, the characteristic variables reflecting the characteristics of airport traffic are extracted. The specific processing steps for splitting the data and extracting the daily traffic characteristic indicators used for clustering are as follows:

1) Data cleaning and partitioning: remove missing and abnormal information and retain valid information. After that, the day is divided into 24 time periods, each hour as a time period. Based on the actual take-off time and actual landing time in the original data, statistics are made to obtain the arrival traffic, departure traffic, and total traffic of the hour granularity. On this basis, the daily arrival traffic, daily departure traffic, and daily total traffic are further calculated.

2) Traffic characteristic indicators extraction: Considering that only the daily total arrival and departure traffic is used to cluster each data sample, the characteristic indicators are too few, and it is difficult to obtain the appropriate clustering results. However, if the hourly traffic is included in the characteristic indicators, the sample dimension will be too

high and the calculation efficiency will be too low. Using python to visualize the average daily hourly traffic flow, and draw a daily arrival/departure traffic and total traffic distribution map, additional characteristic indicators can be observed. According to experience, the arrival peak is generally at night, and the departure peak is generally in the morning. According to the traffic distribution map of arrival/departure, the specific arrival/departure peak period can be determined, and then the traffic sum of the evening peak period and the morning peak period of departure can be calculated. Then according to the daily total traffic distribution map, the daily trough traffic and daily peak traffic are observed and calculated. These four additional feature indicators can describe the daily traffic characteristics more concisely without losing accuracy. In the final data matrix, there are seven characteristic indicators for each data sample, which are daily arrival traffic, daily departure traffic, total daily traffic, the sum of arrival traffic in the evening peak period, the sum of departure traffic in the morning peak period, the sum of daily trough traffic and the sum of daily peak traffic, denoted as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (1)$$

The matrix has m data samples and n feature indicators. $[X_1, X_2, ..., X_m]$ represents each data sample.

3. Clustering Model of Traffic Similar Days Based on HI-K-means Algorithm

3.1. Algorithm Overview

Clustering analysis is one of the data mining techniques, which can be divided into a variety of clustering algorithms. Hierarchical clustering and K-means clustering are widely used because of their strong practicability, but they also have their defects. The typical feature of hierarchical clustering is that clustering occurs spontaneously, but once the merging of two clusters is performed, it cannot be corrected[16]. It can be seen that the biggest advantage of hierarchical clustering is that it does not need to specify the number of clusters in advance, and can form a clustering tree directly according to a given distance metric criterion, but this will also lead to a large error in clustering results. K-means clustering has the advantages of high time efficiency, easy description, and good clustering effect[17], which can effectively make up for the shortcomings of hierarchical clustering. However, the K-means algorithm needs to specify the number of clusters *k* in advance before clustering. Therefore, the *k* value is determined by hierarchical clustering, which can make up for the deficiency. Therefore, this paper combines hierarchical clustering and K-means clustering to propose a traffic-similar day clustering model based on the HI-K-means algorithm.

By visualizing the clustering tree obtained by hierarchical clustering, the value range of clustering number k can be obtained. The convergence of the two clustering algorithms is to use the evaluation index to evaluate the K-means clustering results of each cluster number *k* within the value range, and finally select the *k* value corresponding to the optimal clustering results. The original intention of the design of the evaluation index is not designed for a single algorithm[10], so it is inevitable to have subjective factors to evaluate the clustering effect only with one or several indicators. In this paper, 19 evaluation indexes such as Silhouette Coefficient, Davies-Bouldin Index, Kl-divergence, and Dunn in the R language NbClust package are used to evaluate the K-means clustering results. Each index will calculate the value of the optimal clustering number k based on its calculation method. The *k* value with the highest number of occurrences after the calculation of all indicators is input into the K-means clustering algorithm again, and the final optimal clustering result can be obtained.

3.2. Model Construction

Based on the above HI-K-means algorithm, clustering is performed according to the daily traffic characteristic indicators, and each cluster in the clustering result represents a similar day. The specific steps of model construction are as follows:

Step 1: Use Equation (2) to normalize the daily traffic characteristic indicators data to balance the weight of each value;

$$x_{ij,scale} = \frac{x_{ij} - x_{j,\min}}{x_{j,\max} - x_{j,\min}} \quad (2)$$

j is the jth traffic characteristic indicator, $1 \le j \le n$; x_{ij} is the ith sample corresponding to the jth indicator, $1 \le i \le m$; $x_{j,\min}$ is the minimum value of the indicator j corresponding to the column data; $x_{j,\max}$ is the maximum value of the indicator j corresponding to the column data; $x_{ij,scale}$ is the normalized traffic characteristic data value; The standardized characteristic indicator matrix is X'.

Step 2: Let each sample in X' be an initial cluster, that is, the number of initial clusters is the same as the number of samples, which is m;

Step 3: The average distance method is used to calculate the distance between each cluster, that is, the sample similarity:

$$d_{avg}(Z_f, Z_h) = \frac{1}{n_f n_h} \sum_{X_e \in C_f} \sum_{X_g \in C_h} | X'_e - X'_g | \quad (3)$$

Z_f and Z_h are the clusters of samples in X'; n_f is the number of samples in cluster Z_f; n_h is the number of samples in cluster Z_h; X'_e is the sample in cluster Z_f; and X'_g is the sample in cluster Z_h.

Step 4: Find the two clusters with the closest average distance and classify them into one category;

Step 5: Repeat steps 3 and 4 until all samples are classified into one category to obtain a clustering tree;

Step 6: Visualize the clustering tree and define the threshold distance to obtain the value range of the number of clusters k;

Step 7: The selected 19 evaluation indexes are sorted and coded, which are marked as {1,2,..., 19} respectively. The

variable l is the coding value of the current evaluation index, starting from l = 1. At the same time, let the initial ' votes ' of each k value be 0;

Step 8: The K-means clustering results corresponding to each k value are evaluated one by one with the evaluation index with the code value l, and the optimal k value under the evaluation index is obtained, so that the 'votes' of the k value are added to 1, and l = l + 1;

Step 9: When l≤19, repeat step 8. When l > 19, select the k value with the largest number of 'votes' as the final value, and mark the final value as k' to obtain the optimal clustering number of K-means algorithm;

Step 10: Take k' sample points as the initial center of k' clusters $C=\{C_1,C_2,...,C_{k'}\}$, $C \in X'$;

Step 11: For each sample point, calculate the Euclidean distance between them and k' centers, and classify it into the cluster with the smallest distance center;

$$dis(X'_p, C_s) = \sqrt{\sum_{t=1}^{n} (x_{pt,scale} - c_{st})^2} \quad (4)$$

X'_p is the pth sample data in X'; C_s is the sth cluster center; $x_{pt,scale}$ is the tth normalized characteristic indicator value in X'_p; c_{st} is the tth normalized characteristic indicator value in C_s;

Step 12: Recalculate each cluster center according to the category of each sample;

Step 13: Repeat steps 10 to 12 until the cluster that the sample points belong to does not change, and the final clustering result is obtained.

The flow chart of model construction is shown in Figure 1.

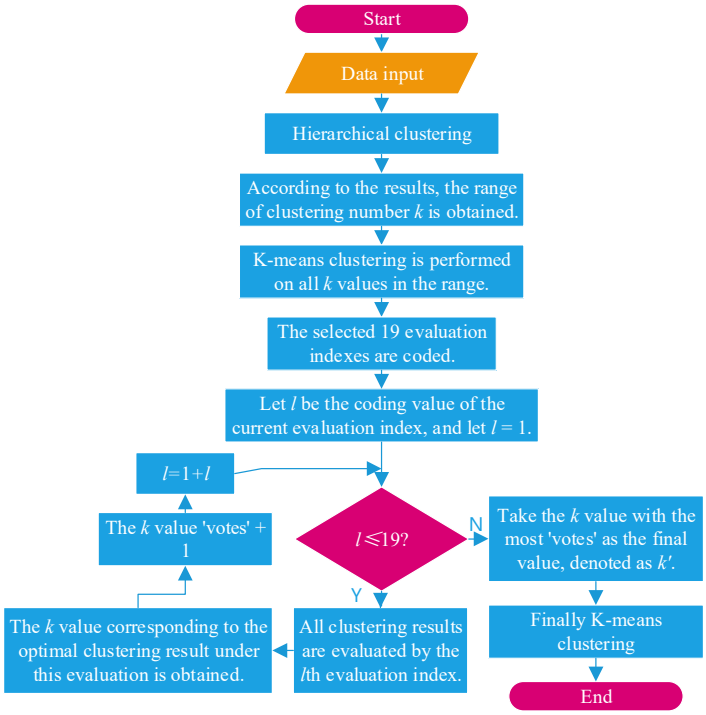

Figure 1. The flow chart of traffic similar days clustering model construction

4. Traffic Similar Day Clustering Example Analysis

4.1. Sample selection and data processing

In this paper, Tianjin Binhai Airport is taken as the research object, and the flight plan data of Tianjin Binhai Airport from January 1,2020 to December 30,2020 are collected, including a total of 114,314 arrival and departure flight information. Among them, the data from January 1,2020 to January 29,2020 are non-epidemic control traffic data, and the data from January 30,2020 to December 31,2020 are epidemic control traffic data, so as to verify whether the model can distinguish airport traffic under different policies.

1) Data cleaning and division: After data cleaning, 113614 valid information are retained, a total of 365 days of traffic data.

2) Traffic characteristic indicator extraction: Firstly, the daily arrival traffic, daily departure traffic and daily total traffic are calculated as the basic characteristic indicators. Then use python to draw the average daily hourly admission and departure traffic to select additional feature indicators, as shown in Figure 2-4. The darker the color of the coordinate point, the more the same traffic value at the same time. At the same time, the average traffic per hour is represented by a broken line on the figure, so as to judge the daily arrival peak period, departure peak period, daily total traffic peak period and daily total traffic trough period, and then add up the hourly traffic of these periods every day to obtain four characteristic indicators. It can be seen from the diagram that the peak of daily arrival is from 20 to 21 every night, the peak of daily departure is from 7 to 8 in the morning, the trough of daily total traffic is from 4 to 5 in the morning, and the peak is from 11 to 12 at noon. The characteristic indicators obtained by adding the traffic corresponding to these periods are the sum of the arrival traffic in the evening peak period, the sum of the departure traffic in the morning peak period, the sum of the daily trough traffic, and the sum of the daily peak traffic.

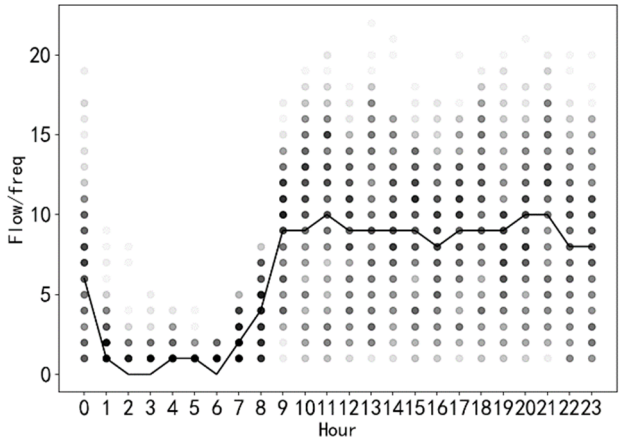

Figure 2. Daily arrival flow distribution

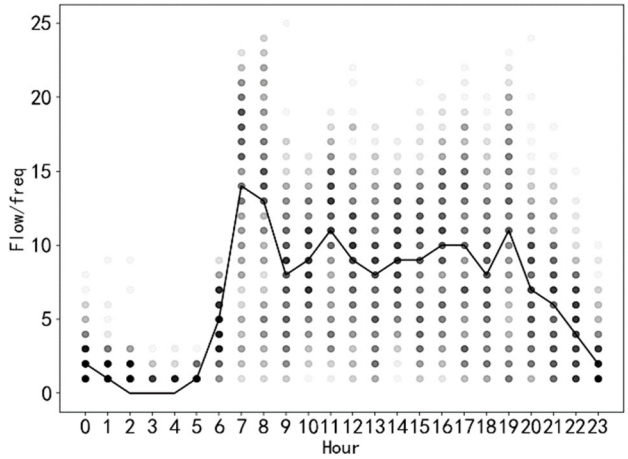

Figure 3. Daily departure flow distribution

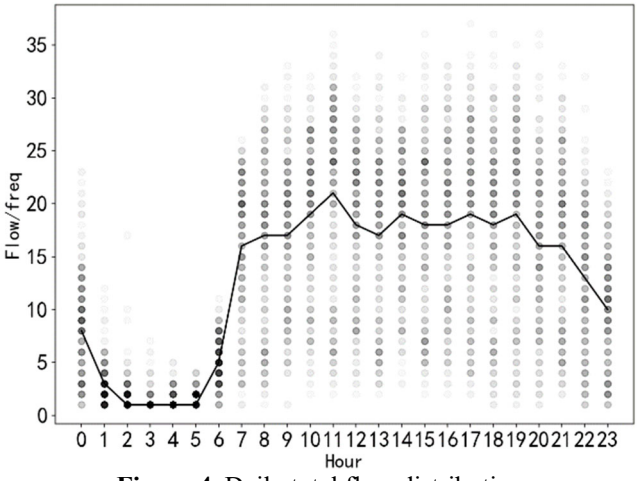

Figure 4. Daily total flow distribution

Finally, a data set with six characteristic indicators is obtained, and some data are shown in Table 1.

Table 1. Daily flow characteristic index (part)

Date	Daily arrival traffic/sortie	Daily departure traffic/sortie	Total daily traffic/sortie	Evening peak arrival traffic/sortie	Morning peak departure traffic/sortie	Daily trough traffic/ sortie	Daily peak traffic/ sortie
2020/1/1	225	231	456	24	38	5	52
2020/1/2	224	225	449	25	38	1	53
2020/1/3	218	231	449	25	40	5	53
2020/1/4	225	219	444	29	40	5	51
2020/1/5	230	242	472	25	44	3	53
2020/1/6	212	225	437	29	30	1	57

4.2. Traffic similar day clustering based on HI-K-means algorithm

1) Hierarchical clustering: Python is used to perform hierarchical initial clustering on 365 daily traffic characteristic indicators data normalized by Tianjin Binhai Airport, and the clustering tree is visualized. The results are shown in Figure 5.

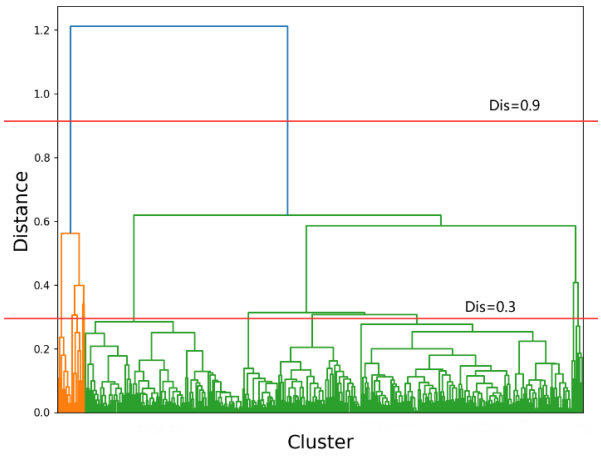

Figure 5. Hierarchical clustering results

Each point in the abscissa represents a sample, and the ordinate represents the distance between each cluster. It can be seen that as the distance increases, the number of clusters gradually decreases, the minimum number of clusters is 2, and the maximum number of clusters is 365 (the total number of samples). The distance between each cluster is too large or too small, which will make the clustering result inaccurate. Therefore, two threshold distances are set to determine the range of cluster number k. Here, the minimum distance is 0.3, the maximum distance is 0.9, and the value range of k is [2,11].

2) Determine the optimal number of clusters: The K-means clustering results of each cluster number are evaluated by 19 evaluation indexes, and the final voting results are shown in Figure 6.

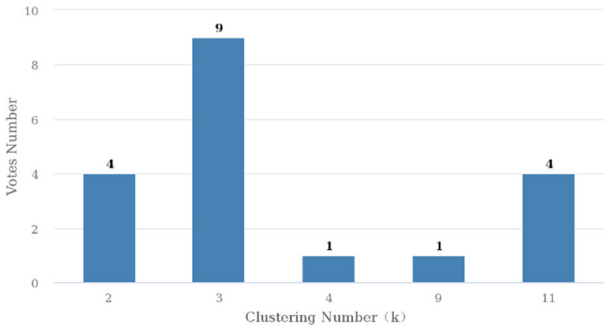

Figure 6. Evaluation index 'votes' results

From the graph, it can be seen that the number of clusters k = 3 gets the most 'votes', which is 9 'votes', that is to say, the evaluation results of 9 of the 19 evaluation indexes are the best when k = 3. Thus, the optimal number of clusters is determined to be 3.

3) Finally K-means clustering: K-means clustering is performed based on the optimal number of clusters obtained above. Finally, the clustering results are shown in Table 2. Among them, daily arrival traffic, daily departure traffic, total daily traffic, the sum of arrival traffic in the evening peak period, the sum of departure traffic in the morning peak period, the sum of daily trough traffic, and the sum of daily peak traffic are shown as the mean value of each category.

Table 2. The clustering results of traffic similar days of Tianjin Binhai Airport flow in 2020

Category	Quantity	Characteristic indicators						
		Daily arrival traffic/sortie	Daily departure traffic/sortie	Total daily traffic/sortie	Evening peak arrival traffic/sortie	Morning peak departure traffic/sortie	Daily trough traffic/ sortie	Daily peak traffic/ sortie
1	92	68	69	137	9	11	2	17
2	146	204	209	414	26	36	2	51
3	127	158	161	319	19	28	2	42

5. Analysis of Results

1) Similar day distribution: From Table 2, it can be seen that the daily total flow of Category 2 is the largest, reaching 414 sorties, and the peak hours of arrival and departure and daily peak flow are also the highest. At the same time, the number of samples in this category is the largest, with a total of 146, accounting for 40 % of the total number of days in 2020. The daily total flow of category 1 is the smallest, only 137 sorties, and the peak flow and daily peak flow are also the least, with only 93 samples in this category. The total daily flow of Category 3 is high, reaching 319 sorties. In addition, the number of samples in this category is 127. Combining the above clustering results with the month can better analyze, and the distribution of each category in each month is shown in Figure 7 below.

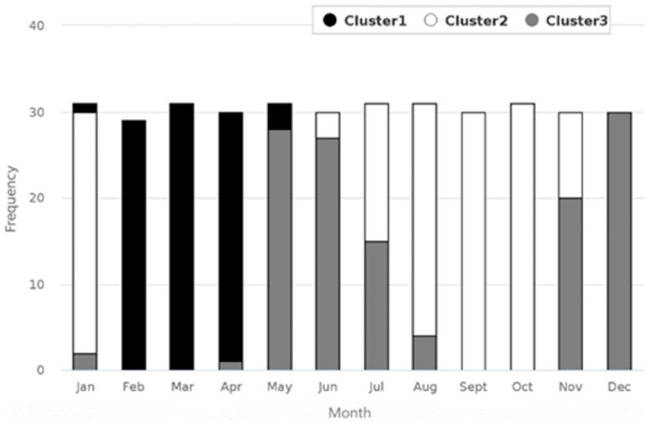

Figure 7. Distribution of traffic similar days in 2020

In January 2020, the coronavirus epidemic began to show signs, but traffic control has not been issued yet, and the flow of Binhai Airport conforms to the law of the past years. It can be seen that January is mostly a similar day of category 2, which is close to the Spring Festival. Students are taking holidays one after another, and working people are returning home, so the airport traffic is large.

Since January 30, 2020, Tianjin began to seal the city, which lasted until the beginning of May 2020. The similar days of category 1 gathered in this period. During this period, the traffic control policy was the most stringent. Even during the Spring Festival in late January and early February, the airport traffic decreased significantly, only 33 % of the daily traffic in January, completely deviating from the usual traffic rules, and a large number of people were trapped in their hometowns.

In early May 2020, the epidemic was under control in most parts of the country. Tianjin began to implement a 'dynamic clean-up' policy, that is, people did not have to go out, traffic control was relaxed, and daily traffic at coastal airports began to increase slowly, returning to 77 % of daily traffic in January. Most of the days before July were similar days in category 3. In July 2020, the 'dynamic clean-up' policy was initially effective, the domestic epidemic was effectively controlled, and traffic control was further liberalized. At the same time, with the arrival of the students' summer vacation, the traffic volume of the Binhai airport also increased significantly, and the departure and arrival volume returned to normal levels. This situation continued until early November. From July to November, most of the dates are similar days in category 2.

In mid-November 2020, in order to avoid the possible spread of the epidemic caused by the tide of returning home, people's travel was once again limited, and the flow of Binhai airports was slightly reduced. From mid-November to the end of the year, it is a category three similar day.

From the above, category 3 can be matched as the similar day of airport traffic under the normal control policy of the epidemic, category 2 can be matched as the similar day of normal traffic under the condition of traffic control liberalization, and category 1 can be matched as the similar day of low traffic under strict traffic control during the epidemic. It can be seen that the clustering model of traffic similar days based on the HI-K-means algorithm can accurately distinguish airport traffic under different traffic control policies during the epidemic.

Airport managers can match future dates with these three similar days and the epidemic control guidance documents issued by their superiors, to design three different traffic management methods to deal with the management of different expected traffic conditions, so as to achieve the fullest utilization of airport human and physical resources and avoid waste of resources.

2) Principal component analysis dimension reduction: In order to more intuitively verify the accuracy of the HI-K-means clustering results, the principal component analysis is used to reduce the 7-dimensional data used in the clustering to 2-dimensional, and a scatter plot is drawn on the plan to show the clustering results, as shown in Figure 8 below.

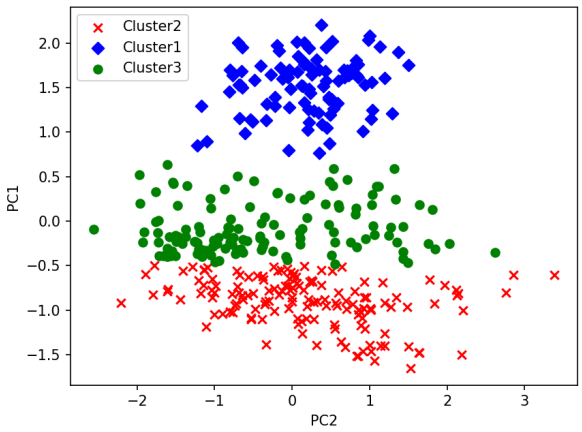

Figure 8. Clustering results PCA dimension reduction display

In the figure, PC1 represents the first principal component and PC2 represents the second principal component. It can be intuitively seen that similar samples are clustered into one category, and the clustering results are ideal.

6. Conclusion

Taking Tianjin Binhai Airport as an example, this paper processes its flight plan data in 2020 to obtain 7-dimensional traffic characteristic indicators for calculation. Then, a clustering model of traffic similar days based on HI-K-means algorithm is proposed. The model is used to cluster the data of traffic characteristic indicators, and then three types of traffic similar days are obtained. After drawing the distribution of similar days, it is found that the clustering results can better match the daily traffic under different policies, which proves the effectiveness of the model. Finally, the clustering results are visualized by principal component analysis, which proves the accuracy of the model.

Acknowledgment

China Civil Aviation North China Air Traffic Administration Science and Technology Project (201904,202002)

References

[1] XIONG Ting. Research on evaluation indicator system of airspace design and operation in terminal[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2020.

[2] TIRACHINI A, CATS O. COVID-19 and public transportation: current assessment, prospects, and research needs[J]. Journal of Public Transportation, 2020,22(1):1-21.

[3] LOSKE D. The impact of COVID-19 on transport volume and freight capacity dynamics: an empirical analysis in german food retail logistics[J]. Transportation Research Interdisciplinary Perspectives, 2020,6(1):1-6.

[4] WANG P F, CHEN K Y, ZHU S Q, et al. Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak[J]. Resources, Conservation and Recycling, 2020,158(3):1-9.

[5] YEN M Y, SCHWARTZ J, CHEN S Y, et al. Interrupting COVID-19 transmission by implementing enhanced traffic control bundling: implications for global prevention and control efforts[J]. Journal of Microbiology, Immunology and Infection, 2020,53(3):377-380.

[6] KUHB K D. A methodology for identifying similar days in air traffic flow management initiative planning[J]. Transportation Research Part C: Emerging Technologies, 2016,69:1-15.

[7] KUHN K, SHAH A, SKEELS C, et al. Characterizing and classifying historical days based on weather and airtraffic[C]//IEEE. 2015 IEEE/AIAA 34th Digital Avionics Systems Conference(DASC). Prague, Czech Republic: IEEE, 2015: 1C3-1-1C3-12.

[8] FANG Ka-tai. Cluster analysis(I)[J]. Mathematics in Practice and Theory, 1978, 01:66-80.

[9] GRABBE S, SRIDHAR B, MUKHERJEE A. Clustering days and hours with similar airport traffic and weather conditions[C]//AIAA. 14th AIAA Aviation Technology, Integration, and Operations Conference. Reston, VA: AIAA, 2014:751-763.

[10] SHANG Ran-ran. Research on short-term traffic flow prediction technology in terminal area based on data mining[D]. Hebei: North China University of Science and Technology, 2021.

[11] RAO Dan, SHI Hong-wei. Study on air traffic flow recognition and anomaly detection based on deep clustering

[12] CHEN J T, RAFAL K, STEVEN M S, et al. Using weather translation and machine learning to identify similar weather impact days[C]//ARC. AAIA Guidance, Navigation, and Control Conference. Published Online: ARC, 2017: https://doi.org/10.2514/6.2017-1727.

[13] SHAN Le. Research on related technologies of terminal capacity assessment on weather conditions[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2014.

[14] XU Yi-fan, LI Jie, WEI Yi-tao. Clustering analysis on airport weather based on SOM network[J]. Mathematics in Practice and Theory, 2016,46(17):210-217.

[15] CHEN Si. The forecast of airport peak service rate based on weather[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2019.

[16] DUAN Ming-xiu. Application and research of hierarchical clustering algorithm[D]. Changsha: Central South University, 2010.

[17] WU Su-hui, CHENG Ying, ZHENG Yan-ning, et al. Survey on K-means algorithm[J]. Data Analysis and Knowledge Discovery, 2011(5):28-35.