Research on business area mining using location data based on hierarchical clustering algorithm

Fangliang Huang[a,b*], Huanqing Xu[a], Tongping Shen[a], Li Jin[a]
[a] School of Medicine and Information Engineering, Anhui University of Chinese Medicine, Hefei, China; [b] College of Computing and Information Technologies, National University, Manila, Philippines

ABSTRACT

According to the coverage of cell phone signals in geographic space, combined with time series of cell phone, positioning data can restore the complete realistic activity trajectory of the population, so as to obtain the characteristic information of spatial distribution and activity connection of the population. In this paper, the user location dataset obtained from CMCC is pre-processed, then mined by hierarchical clustering algorithm and combined with data visualization techniques. The experimental results can provide decision support for commercial promotion, so as to carry out targeted business layout.

Keywords: Data mining, hierarchical clustering algorithm, location information

1. INTRODUCTION

With the popularity of personal cell phones and the Internet, cell phones have become a mandatory tool for all people. According to the coverage of cell phone signals in geographic space combined with the time series of cell phone positioning data can restore the complete reality of the population's activity trajectory and thus get the information of the spatial distribution of the population and the characteristics of activity links[1-3].

The business area/circle is an important space for business activities in modern markets, and one of the purposes of business district classification is to study the distribution of potential customers in order to develop appropriate business countermeasures.

The mining target of this project is to adopt data mining technique to subgroup the base station for the historical positioning data of users. We can also conduct feature analysis of different business area subgroups and compare the value of different business district categories so as to select suitable areas for targeted marketing activities.

2. DATASET

The dataset for this project which contains 431 rows and 5 columns is the cell phone subscriber location data obtained by parsing the specific interface provided by CMCC (China Mobile Communications Group Co., Ltd).

Since the goal of mining is to find high-value shopping areas, which requires extracting the pedestrian flow characteristics of the area within the base station, such as average stay time per person and pedestrian flow, based on the user's location data.

High-value business districts are characterized by high pedestrian flow and long average stay time per person in terms of pedestrian flow characteristics.

Office buildings have a fixed base station range where office workers are located during the day and longer hours, and high pedestrian flow.

Residential areas also have the characteristics of fixed base station range, long time and high pedestrian flow. Therefore, the simple dwell time cannot determine the category of business area.

Modern society works, with a week as a small cycle of work, divided into weekdays and weekends. The day is divided

* hfl@ahtcm.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125060P
 · 0277-786X · doi: 10.1117/12.2661792

into working hours and closing hours.

To sum up, four indicators for designing the characteristics of pedestrian flow, namely average stay time per person on working days, average stay time before dawn, average stay time per person on weekends, and average daily pedestrian flow, are also four columns of the dataset.

3. METHODOLOGY

3.1 Data pre-processing

Figure 1 shows the process of this study. Firstly, the original dataset will be pre-processed because there are great differences between the attributes of the dataset. In order to eliminate the impact of order of magnitude data, the deviation standardization method is used to normalize all data to the interval [0,1] before mining.

Figure 1. Research procedure.

Next, we will use three methods to describe the hierarchical diagram of the data set, namely, the nearest neighbour method[4], the variance minimization method[5] and the farthest neighbour method[6]. Figures 2-4 show the results of these three methods.

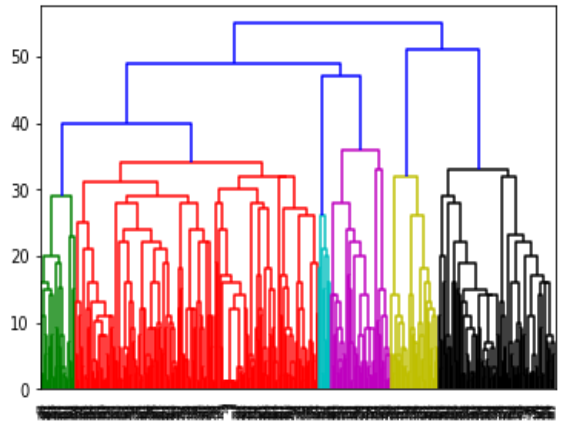

Figure 2. Nearest neighbour.

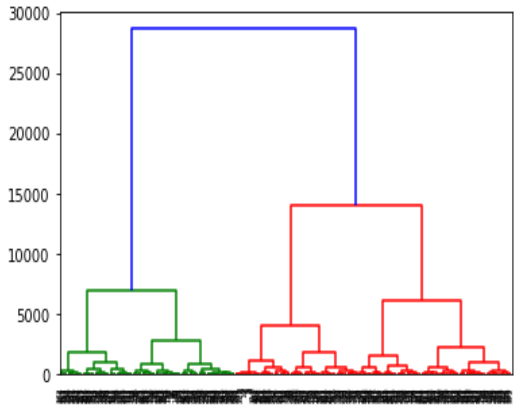

Figure 3. Variance minimization.

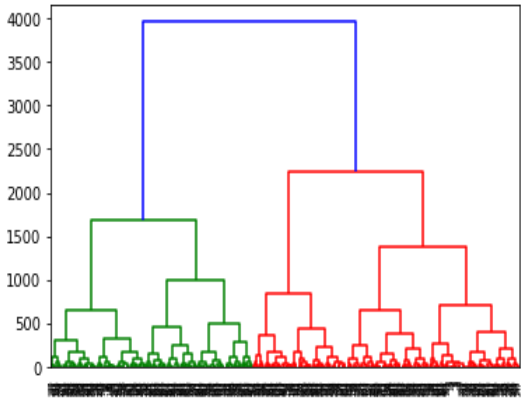

Figure 4. Farthest neighbour.

From the above three figures, we can find that the number of clustering categories is 3, and then the hierarchical clustering algorithm can be used to train the model.

3.2 Hierarchical clustering

Calculating the degree of similarity between groups of data points allows hierarchical clustering algorithms to construct a nested clustering tree in a hierarchical structure[7, 8]. Root nodes of clusters are at the top of a clustering tree, while the original data points of different categories are at the bottom. This study uses the merging approach, one of two ways to construct a clustering tree alongside the splitting approach. The procedure of hierarchical clustering is depicted graphically in Figure 5.

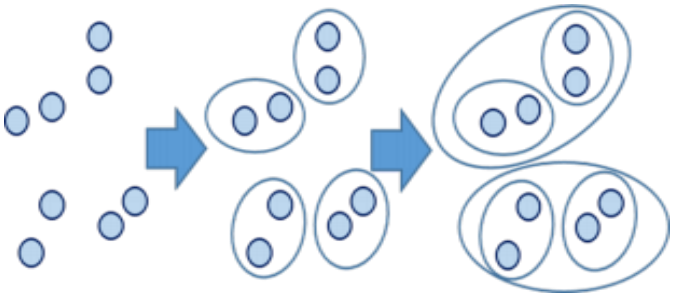

Figure 5. Hierarchical clustering process diagram.

- Bottom-up merging method

By iteratively evaluating the similarity between pairs of data points, the bottom-up merging method combines the two sets of data that are most similar to one another. To put it plainly, the merging algorithm of hierarchical clustering calculates the distance between each pair of data points in each category and all data points and assigns a greater score to pairs with shorter distances. Following this, the clustering tree is constructed by merging the pairs of data points or categories with the smallest distance.

- Euclidean distance matrix

Euclidean distances are used in hierarchical clustering to determine the degree of similarity between data points belonging to distinct categories[9]. Typically, this is accomplished by combining data points with the shortest distance values, a task for which a Euclidean distance matrix is used. The Euclidean distance is represented by equation (1).

$$D = \sqrt{(z_1 - y_1)^2 + (z_2 - y_2)^2} \tag{1}$$

Table 1 shows the example data, and we calculate the Euclidean distance matrix from point A to point G by Euclidean distance, and the results are shown in Table 2. Next, we will create the clustering tree by combining the data points with the highest similarity by merging.

Table 1. Example data.

Point	Position
A	16.9
B	38.5
C	39.5
D	80.8
E	82
F	34.6
G	116.1

Table 2. The Euclidean distance matrix.

	A	B	C	D	E	F	G
A	0	21.6	22.6	63.9	65.1	17.7	99.2
B	21.6	0	1	42.3	43.5	3.9	77.6
C	22.6	1	0	41.3	42.5	4.9	76.6
D	63.9	42.3	41.3	0	1.2	46.2	35.3
E	65.1	43.5	42.5	1.2	0	47.4	34.1
F	17.7	3.9	4.9	46.2	47.7	0	81.5
G	99.2	77.6	76.6	35.3	34.1	81.5	0

- Distance between data points and combined

After combining data point B with data point C, the distance matrix between each category of data points is recalculated. The distance between data points is calculated in the same way as the previous method. What needs to be explained here is the calculation method between the combined data points (B, C) and other data points. When we calculate the distance from (B, C) to A, we need to calculate the mean value of the distance from B to A and C to A respectively. Equation (2) is shown below.

$$D = \frac{\sqrt{(B-A)^2}+\sqrt{(C-A)^2}}{2} \tag{2}$$

After calculation the distance from data point D to data point E is the smallest among all the distance values, 1.20, as shown in Table 3. This means that among all the current data points (including the combined data points), D and E have the highest similarity. Therefore we combine data point D and data point E and again calculate the distances between other data points.

Table 3. The new Euclidean distance matrix.

	A	(B, C)	D	E	F	G
A	0	22.1	63.9	65.1	17.7	99.2
(B, C)	22.1	0	41.8	43	43.5	77.1
D	63.9	41.8	0	1.2	1.2	35.3
E	65.1	43	1.2	0	0	34.1
F	17.7	4.4	46.2	47.7	47.7	81.5
G	99.2	77.1	35.3	34.1	34.1	0

The later work is to keep repeating the calculation of the distance between data points and data points, data points and combined data points. In this process, we find that the distance value between data point A and data point F is the smallest, so we combine A and F to generate the combined data point (A, F).

Up to this point all the data points have been combined based on the distance values except for the data point G. The bottom level of the clustering tree has been completed. In the following, we will continue to calculate the distances between the combined data points and merge the combined data points with the highest similarity.

- Distance between two combined data points

We employ Average Linkage, one of three approaches for determining the separation between merged data points (the others being Single Linkage and Complete Linkage)[10]. Using the distance between each of the two combined data points and all other data points, we can determine the average linkage. The distance between the two sets of data is calculated as the mean of all the individual distances. This approach requires more processing resources, but yields more plausible findings. Here, the average distance between the two sets of data points (A, F) and (B, C) is determined independently of Formula 3, which determines the distance from the combined data point (A, F) to (B, C).

$$D = \frac{\sqrt{(A-B)^2}+\sqrt{(A-C)^2}+\sqrt{(F-B)^2}+\sqrt{(F-C)^2}}{4} \tag{3}$$

By calculating and comparing the distance between different combinations of data points, the distance from (A, F) to (B, C) is the smallest among all combinations of data points, which is 13.25. The distance from (A, F) to (B, C) is the smallest among all combined data points, which is 13.25. It means that (A, F) to (B, C) has the highest similarity. Therefore, the combination of (A, F) to (B, C) is (A, F, B, C).

Using the same method, the distance between the combined data points (D, E) and G is calculated to be the smallest among all combined data points at 34.70. Therefore, (D, E) and G are combined as (D, E, G).

Finally, by calculating and merging, we obtained two combined data points (A, F, B, C) and (D, E, G). These are also the top two data points of the clustering tree.

3.3 Hierarchical clustering dendrogram

The hierarchical clustering tree shows the results of each of the steps before it in the form of a tree. The first seven data points, from A to G, are at the bottom. Figure 6 shows how the second layer of the clustering tree (A, F), (B, C), (D, E), and G are put together based on how similar the seven data points are. This process is repeated until the full hierarchical clustering tree is made.

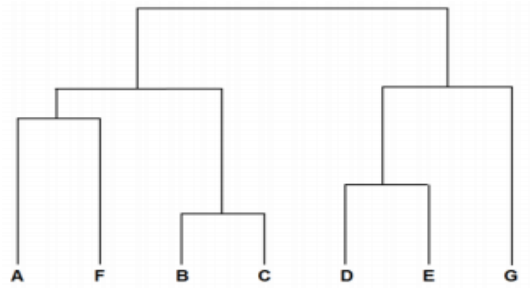

Figure 6. Hierarchical clustering tree.

4. RESULT AND DISCUSSION

Once we understand the principle of the hierarchical clustering method, looking back at Figures 2-4, we can find that the number of clustering categories is 3.

Next we started to use the hierarchical clustering algorithm to mine the dataset by importing sklearn's hierarchical clustering function, setting up the parameters and training the model. The training process outputs in detail the categories corresponding to each sample and draws a line graph for each feature, as shown in Figures 7-9. The horizontal coordinates in the following three figures indicate the average stay time per person in working days (ASPWD), average stay time before dawn (ASBD), average stay time per person on weekends (ASPWS), and average daily pedestrian flow, respectively.

From the three figures, we can obtain the below mining results:

Business area category 1, with low average stay time per person in working days and average stay time before dawn, medium average stay time per person on weekends, and extremely high average daily pedestrian flow, which is in line with the characteristics of business districts.

Business area category 2, the average stay time per person in working days is medium, average stay time before dawn and weekends is long, and the average daily pedestrian flow is low, which is in line with the characteristics of the residential area.

Business area category 3, which has a long average stay time per person in working days, a low average stay time before dawn and weekends stay, and a medium average daily pedestrian flow, which is very much in line with the office business area.

Business area category 2 has a low pedestrian flow, business area category 3 has an average pedestrian flow, and the work area of white-collar workers generally has a high pedestrian flow during work hours and lunchtime, both of which are not conducive to the development of operators' promotional activities. Business area category 1 has a high pedestrian flow and is conducive to operator promotions in such commercial areas.

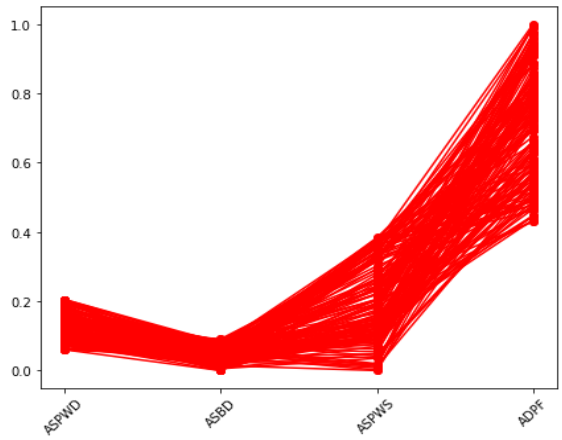

Figure 7. Business area category 1.

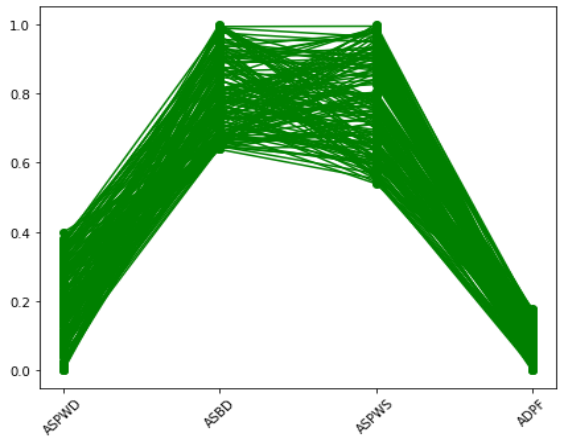

Figure 8. Business area category 2.

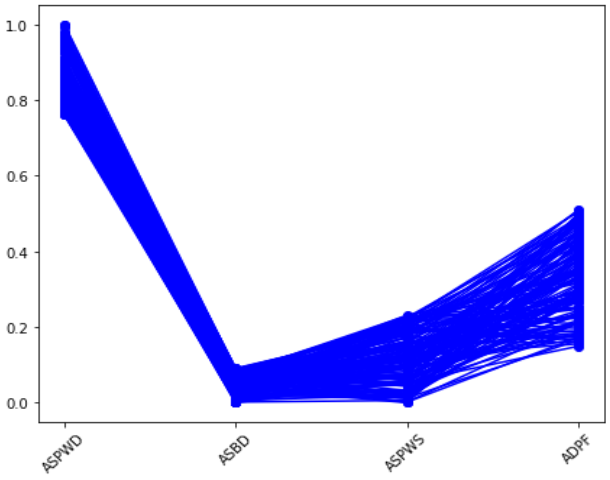

Figure 9. Business area category 3.

5. CONCLUSION

In this paper, the deviation standardization method is used to pre-process the dataset, which solves the influence caused by the large order of magnitude difference of each column in the original dataset. In addition, data mining technology is used to provide decision support for business analytics, thereby achieving the effect of reducing business costs and improving the efficiency of decision-making.

In the next step, we will use more mining algorithms based on this dataset for comparison experiments, and consider both the time complexity and space complexity of each algorithm in order to find an optimal mining method and promote it.

ACKNOWLEDGMENTS

Authors would like to thank Albert A. Vinluan who is a professor at New Era University for enriching discussions and kindly review. This study was supported by the Key Project of Scientific Research in Anhui Higher Education Institutions of China under Grant NO. KJ2021A0587 and NO. SK2020A0244, the Provincial Quality Engineering Project in Anhui Higher Education Institutions of China under Grant No.2020jyxm1029.

REFERENCES

[1] Jia, J. S., Lu, X., Yuan, Y., Xu, G., Jia, J., and Christakis, N. A., "Population flow drives spatio-temporal distribution of COVID-19 in China," Nature, 582(7812), 389-394(2020).
[2] Masso, A., Silm, S., and Ahas, R., "Generational differences in spatial mobility: A study with mobile phone data," Population, Space and Place, 25(2), e2210(2019).
[3] Dzwinel, W., Yuen, D. A., Boryczko, K., Ben-Zion, Y., Yoshioka, S., and Ito, T., "Cluster analysis, data-mining, multi-dimensional visualization of earthquakes over space, time and feature space," Nonlinear Processes in Geophysics, 12, 117-128(2005).
[4] Liu, J., Zhao, L. Y., "Clustering algorithms research," Journal of Software, 19(1), 4861(2008).
[5] Lurka, A., "Spatio-temporal hierarchical cluster analysis of mining-induced seismicity in coal mines using Ward's minimum variance method," Journal of Applied Geophysics, 184, 104249(2021).
[6] He, L., Wu, L. D., and Cai, Y. C., "Survey of clustering algorithms in data mining," Application Research of Computers, 1, 10-13(2007).
[7] Chi, Y., [Hierarchical Cluster Analysis], http://www.rtutor.com/gpu-computing/clustering/hierarchical-cluster-analysis, 2020-02-22/2022-3-04.
[8] [Hierarchical Clustering], http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html, 2018-05-12/2021-3-04.
[9] Tan, P. N., Steinbach, M., and Kumar, V., "Data mining cluster analysis: Basic concepts and algorithms," Introduction to Data Mining, 487, 533(2013).
[10] Križanić, S., "Educational data mining using cluster analysis and decision tree technique: A case study," International Journal of Engineering Business Management, 12, 1847979020908675(2020).