ISSN: 2664-9640

ISODATA Clustering Algorithm for Site Selection

Qing Cai#, Jinglin Gong# and Jianfei Huang*

Yangzhou University, Yangzhou, Jiangsu, China

*Corresponding author: jfhuang@yzu.edu.cn

# These authors contributed equally to this work

Abstract

This paper mainly solves the problem of site selection, which provides the weak coverage area of the existing network according to the coverage of the existing network antenna, and has made a coordinate delineation of the area. By establishing a planning model, we select some of the points to build new base stations, so that the total traffic of the sites covering weak coverage points reaches a predetermined value. Then, we uses the ISODATA clustering algorithm to solve the established planning model.

Keywords

K-Means clustering; ISODATA clustering; Cluster analysis; Goal planning.

1. Introduction

With the rapid development of mobile communication technology, people's requirements for the quality of mobile communication networks are increasing day by day. And thus, the network planning optimization is extremely important for operators. As the basis of network planning[1], station location planning not only involves the communication quality of users, but also seriously affects the profits of operators, and also affects people's lifestyle to a great extent. Base station location optimization is an important part of network planning , that is, planning the number, location and type of base stations under comprehensive consideration of construction cost, coverage constraints, service traffic and other constraints. It can also reflect that the network is closely related to people's life. The problem of site selection is to obtain the weak coverage area of the existing network from the coverage of the existing network antennas, and select a certain number of points to build base stations in the weak coverage area to solve the coverage problem in the weak coverage area of the existing network [2]. With the further development of communication networks, this issue has attracted extensive attention from academia and industry. Therefore, solving such problems not only has practical significance, but also can promote the academic and industrial circles.

Based on the given data of site selection, we firstly analyze the data according to the minimum threshold as the Euclidean distance between the existing base station and the weak coverage point to judge the distance, establish ISODATA clustering, find its class center for all uncovered points, and use the class center as a reference click to create a planning model.

Secondly, the coverage capability changes with the angle deviating from the main direction. We simulate the sector curve and use the greedy algorithm to solve it. The coordinates of each base station and the optimal sector are obtained when the traffic volume is greater than 90%.

Finally, we use the ISODATA clustering algorithm and the K-Means clustering algorithm. The root performs cluster analysis on the weak coverage points less than 20, clustering the points less than 20 into one class, and re-clustering the points greater than 20, which can be directly obtained by changing the input parameters in the algorithm. The frequency and percentage of clusters are analyzed, the time complexity is calculated by checking the DBI index and the

ISSN: 2664-9640

number of iterations, and the two are compared. Finally, the total time complexity of ISODATA is considered to be relatively low.

2. Establishment of ISODATA clustering model

The classification of the uncovered points is obtained by introducing the ISODATA clustering model, and on this basis, optimization is carried out with the class center as the goal, so that 90% of the traffic is covered by the planned base station, and the distance between the new site and the existing site cannot be reached. If less than 10, the corresponding site coordinates and type selection can be obtained by using [3]. Now we present a schematic diagram of the distribution of weak coverage points and existing base stations through MATLAB, as shown in Figure 1. In Figure 1, the blue part of the point is the position of the weak coverage point, and the red cross point is the position of the existing base station. The denser the blue, the greater the business demand [4].

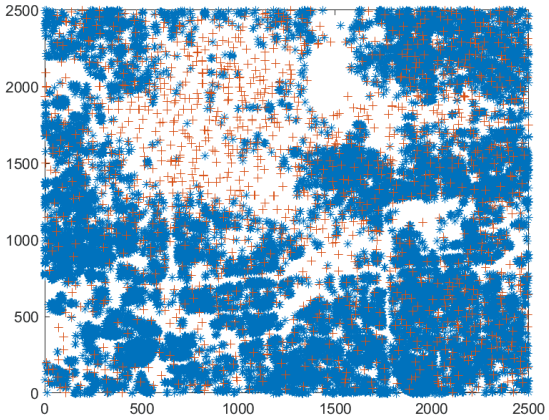

Figure 1: Schematic diagram of the distribution of weak coverage points and existing base stations

By using distance matrix, the distance between each required coverage point and the existing site can be expressed as:

$$D_{ij} = \|x_i - B_j\|, \tag{1}$$

where x_i is the coordinates that the existing site needs to cover, B_j is the existing site coordinates.

To perform preliminary ISODATA clustering on the data[5], we perform preliminary clustering, and give a schematic diagram of the cluster center from this, see Figure 2.

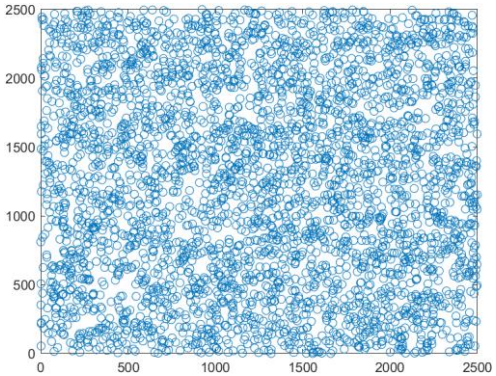

Figure 2: Schematic diagram of cluster centers

Since the threshold distance between base stations is 10, according to the distance formula:

$$d = \sqrt{(x - x_0)^2 + (y - y_0)^2}. \tag{2}$$

Through MATLAB programming, we screen the cluster centers and draw a schematic diagram of the filtered cluster centers, as shown in Figure 3.

ISSN: 2664-9640

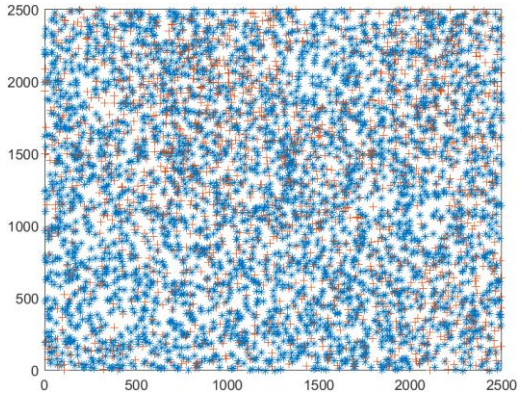

Figure 3: Schematic diagram of the filtered cluster centers

In Figure 3, the blue point is the filtered cluster, and the red point is the original base station location. Then, the 0-1 planning model is established as:

$$w_{min} = 10\sum x_j + \sum y_j. \tag{3}$$

3. Analysis of ISODATA clustering model

As is well known that the ISODATA algorithm is a common algorithm in cluster analysis [6], called dynamic clustering or iterative self-organizing data analysis. The procedure of the ISODATA algorithm can described as follows:

Step 1. Inputting N pattern samples$\{X_i, i = 1,2, \dots, N\}$;

Step 2. Preselecting N_C initial cluster centers $\{Z_1, Z_2, \dots, Z_{N_C}\}$, which may not be equal to the required number of cluster centers, and its initial position can be arbitrarily selected from the sample;

Step 3. Assign N pattern samples to the nearest cluster S_j, if $D_j = min\{\|x - z_i\|, i = 1,2, \dots, N_C\}$, that is, the distance of $\|-z_i\|$ is the smallest, then x ∈ S_j. If the number of samples $S_j < \theta_N$, then cancel the sample subset and subtract 1 from N_C.

Step 4. Correcting each cluster center

$$z_j = \frac{1}{N_j}\sum_{x\in S_j} x\ , j = 1,2 \dots, N_C. \tag{4}$$

Step 5. Calculating the average distance between the pattern samples and each cluster center in each cluster domain S_j

$$\bar{D}_j = \frac{1}{N_j}\sum_{x\in S_j}\|x - z_j\|\ , j = 1,2 \dots, N_C. \tag{5}$$

Step 6. Calculating the total average distance between all pattern samples and their corresponding cluster centers

$$\bar{D} = \frac{1}{N}\sum_{j=1}^{N} N_j\bar{D}_j. \tag{6}$$

Step 7. Compute the standard deviation vector of sample distances in each cluster:

$$\sigma_j = (\sigma_{1j}, \sigma_{2j}, \dots, \sigma_{nj})^T, \tag{7}$$

where the components of the vector are

$$\sigma_{ij} = \sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}(x_{ik} - z_{ij})^2}\ . \tag{8}$$

Step 8. Calculating the distance of all cluster centers

$$D_{ij} = \|z_i - z_j\|, i = 1,2, \dots, N_C - 1, j = i + 1, \dots, N_C. \tag{9}$$

Step 9. Merge the two cluster centers z_{ik} and z_{jk} with a distance of D_{ikjk}, and the new center is

$$z_k^* = \frac{1}{N_{ik}+N_{jk}}[N_{ik}z_{ik} + N_{jk}z_{jk}], k = 1,2, \dots, L. \tag{10}$$

ISSN: 2664-9640

The two cluster center vectors merged in the formula are respectively weighted by the number of samples in the cluster domain, so that z_k^* is the true mean vector.

Through the above method, we finally calculate that the number of macro base stations to be constructed is 3021, the number of micro base stations is 431, the total of them is 3452, the total cost is 30641, the proportion of the new base station business volume to the total business is 90.87%, the Euclidean distance zoom in 30 times for visual analysis, and present a schematic diagram of the distribution of macro base stations (Figure 4) and a schematic diagram of the distribution of micro base stations (Figure 5).

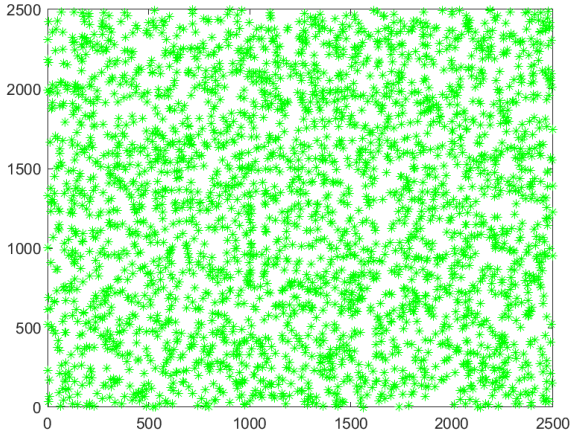

Figure 4: Schematic diagram of macro base station distribution

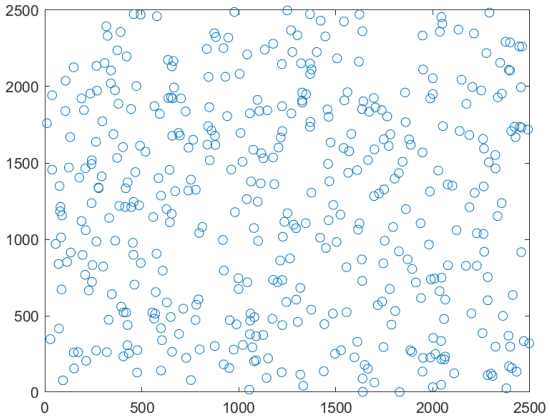

Figure 5: Schematic diagram of micro base station distribution

By using MATLAB programming, we finally obtain the coordinate points of the new macro base station and micro base station [7]. For the ISODATA clustering algorithm, we use the cyclic and progressive algorithm many times to calculate the resulting data. Compared with the stable target value, it can be found that the error is not large, which shows that the model we established and the solution result are more accurate.

4. Analysis of coverage

Assuming the direction of the existing base station, the positive direction of the x-axis is the main direction of the first sector, the azimuth angle is $\theta = 0$ in the positive direction of the x-axis, and the value range of the counterclockwise azimuth is $[0, 2\pi]$.

From this, it can be judged whether the grid points are covered, so as to determine whether all grid points are covered to obtain the uncovered point set, and perform ISODATA clustering on the uncovered point set to obtain the class center. The distance with the base station excludes the point that cannot be used as the base station, the grid point in the center of each class is simulated, and the optimal angle with a radius of 10 and a radius of 30 is obtained.

Now, we can draw the change curve of angle and coverage, see Figure 6.

ISSN: 2664-9640

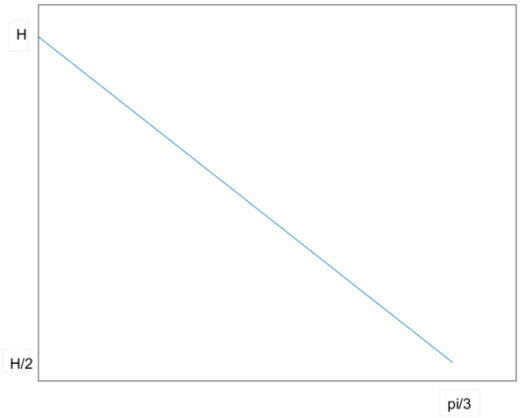

Figure 6: Changes in angle and coverage

Let us take the macro base station 1 as an example to calculate, and solve the angle change between the macro base station and the weak coverage point within the constraints of the threshold and coverage. The visual analysis is shown in Figure 7.

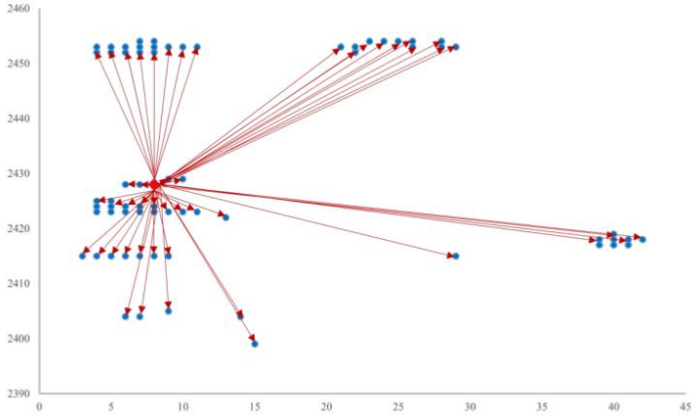

Figure 7: The angle between macro base station 1 and the corresponding weak coverage point

After solving the base station Number 1, we remove all the weak signal points that have been covered by the three sectors of the base station Number 1. Then, examine all the weak signal points around the base station Number 2, and perform clustering again to get three angles [8]. Similarly, some steps are also performed for other base stations, and the final results are summarized in Table 1.

Table 1: Main direction service volume and corresponding service ratio

Main direction 1	Main direction 2	Main direction 3	Does it meet the requirements	Service volume	Proportion
193.2324	-223.255	-42.342	No	402482	65.987%
223.4657	-235.634	-8.224	No	96263	78.943%
223.4326	-123.145	-29.750	No	21546	80.024%
-41.7524	-143.673	-111.632	No	45467	83.621%
235.634	175.901	-82.0873	Yes	64245	94.996%
81.3436	-190.5	-50.3116	Yes	35421	92.641%
223.4657	-237.215	33.8257	Yes	32461	93.563%
128.3237	14.587	-26.9471	Yes	126575	91.325%
2698.4323	-90.4346	-24.1437	Yes	63564	90.234%
131.4366	67.884	-1.7437	Yes	78456	91.864%

ISSN: 2664-9640

We choose ISODATA clustering algorithm to cluster weak coverage points and calculate the time complexity. Let the workload of a single loop be $f_i(n)$: If there are multiple loops distributed in parallel in the algorithm, the number of each loop is

$$f_{1i}(n) = n_1 + n_2. \tag{11}$$

If there are multiple loops, the number of inlays in the body is

$$f_{2i}(n) = n_3 n_4. \tag{12}$$

Thus, we have

$$f_i(n) = \sum f_{1i}(n) + \sum f_{2i}(n). \tag{13}$$

For the space complexity, let the amount of work cells of different lengths be *m(n)*, and the corresponding value can be obtained through statistics.

5. Conclusion

Based on the above analysis and results, the ISODATA clustering algorithm can give priority when searching for solutions, and will aggregate many points into a few points before searching. However, it cannot carry out global search, so the accuracy is not high. When time permits, it can be improved to use particle swarm algorithm to solve the problem. This algorithm can perform global search, generate an initial population and then substitute it into the function. If it is satisfied, the solution is outputted, and a solution suitable for the problem can be searched first. All points can be entered and retrieved, and the optimal value can be obtained by continuous iteration after meeting the conditions, but the program runs slowly due to the large data.

References

[1] L.P. Kang: 5G base station site planning strategy research. Electronic World, (2021) No. 16, p. 9-10.
[2] C.W. Zhu, M. Huang and Z.Q. Fu: Research on the optimization of ship-to-air missile firing scheme based on dynamic programming theory. Ship Electronic Engineering, Vol. 42 (2022) No. 2, p. 25-31.
[3] Y. Liang, S. Chen and Y.Y. Tang: Research on site selection for urban and rural logistics network optimization under FA-kmeans algorithm. Comprehensive Transportation, Vol. 43 (2021) No. 5, p. 115-122.
[4] F.B. Song: *Improvement of genetic clustering algorithm and its application in gene expression data analysis* (Ph.D., Anhui University, China 2019), p. 17-19.
[5] N. Litvinenko, O. Mamyrbayev, A. Shayakhmetova and M. Turdalyuly: Clusterization by the Kmeans method when K is unknown. ITM Web of Conferences, Vol. 24 (2019) p. 01013.
[6] J.Y. Zeng. The principle and implementation of ISODATA algorithm. Science Mosaic, (2009) No. 7, p. 126-127.
[7] X.M. Yang, Y. Luo: Implementation and analysis of ISODATA algorithm. Mining Technology, Vol. 6 (2006) No. 2, p. 66-68.
[8] N. Zeng, J.H. Chen and Q.Q. Fu: Dynamic programming strategy based on greedy algorithm. Computer Knowledge and Technology, Vol. 17 (2021) No. 20, p. 141-143.