INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.113 Volume 15, 2021

A Customer Clustering Algorithm for Power Logistics Distribution Network Structure and Distribution Volume Constraints

Jianying Zhong, Jibin Zhu, Yonghao Guo, Yunxin Chang, Chaofeng Zhu

PINGGAO GROUP CO.,LTD, Pingdingshan 467001, Henan, China

Received: March 17, 2021. Revised: August 8, 2021. Accepted: August 23, 2021. Published: August 25, 2021.

***Abstract*—Customer clustering technology for distribution process is widely used in location selection, distribution route optimization and vehicle scheduling optimization of power logistics distribution center. Aiming at the problem of customer clustering with unknown distribution center location, this paper proposes a clustering algorithm considering distribution network structure and distribution volume constraint, which makes up for the defect that the classical Euclidean distance does not consider the distribution road information. This paper proposes a logistics distribution customer clustering algorithm, which improves CLARANS algorithm to make the clustering results meet the constraints of customer distribution volume. By using the single vehicle load rate, the sufficient conditions for logistics distribution customer clustering to be solvable under the condition of considering the ubiquitous and constraints are given, which effectively solves the problem of logistics distribution customer clustering with sum constraints. The results state clearly that the clustering algorithm can effectively deal with large-scale spatial data sets, and the clustering process is not affected by isolated customers, The clustering results can be effectively applied to the distribution center location, distribution cost optimization, distribution route optimization and distribution area division of vehicle scheduling optimization.**

***Keywords*—Customer Clustering Algorithm, Power Logistics, Distribution Network Structure, Distribution Volume Constraints**

I. INTRODUCTION

Customer clustering problem for power logistics distribution process is a basic problem in the field of logistics distribution research, which can simplify the solution of logistics distribution center location, distribution route optimization, distribution vehicle scheduling optimization and other problems [1-2]. The data scale of logistics distribution customers is generally large, and the difference between them depends on the distance of geographical location [3]. In addition, the distribution demand also requires that the customer clustering process is not affected by the isolated points of scattered distribution, and the sum of the distribution volume of all kinds of customers in the clustering results can not exceed the vehicle load constraint, which belongs to the sum constrained clustering problem, and obtaining the initial solution of the problem belongs to the NP hard problem [4].

Customer clustering problem and constrained distribution problem can be classified as the following categories. The clustering analysis of customers have known one or more distribution center locations. The results of customer clustering are closely related to the location of distribution center, or the customers are directly divided by the distribution center. Sweep algorithm uses the successive approximation method to solve the logistics distribution problem [5]. Taking the distribution center as the origin, the distribution area is divided according to the polar coordinates of each customer. This algorithm can not effectively deal with large data sets. This method can not effectively deal with the uneven distribution of customers, and isolated points tend to be assigned to the last formed class [6]. The center of gravity distance based on Euclidean distance is used to compute the distance between the task and its class center, but it is not clear how to get the initial solution. In reference [7], each customer is divided into distribution centers by ant colony algorithm based on the lowest system cost of distribution center location model. The clustering effect of this method is affected by outliers, and there are many clustering parameters, and the customer distribution volume is not taken as the clustering constraint

The other category is the customer clustering analysis with unknown distribution center location: the customer clustering process is not affected , and the clustering result is independent, which is mostly used for the location of the distribution center. In reference [8], PAM and db-scan are used for two-stage spatial clustering, but PAM algorithm can not effectively deal with large data sets, Db-scan algorithm can not deal with outliers, and it does not consider the constraints of distribution volume. Genetic algorithm has good global convergence ability to solve customer clustering optimization model, but the constraint description ability is not strong [9]. Fuzzy clustering analysis is used to establish multi-attribute distribution customer clustering four, establish index system for customers, and construct fuzzy equivalence relationship, The combination of different level cut sets makes up different clustering results. The number of clusters *k* has nothing to do with the constraints

of distribution volume, so it is difficult to control the load rate of vehicles. In addition, some literatures promote the K-means algorithm to realize the constrained clustering of customers. Because this algorithm itself is sensitive to outliers, so the improved algorithms based on it all have this common fault.

To sum up, most of the existing customer clustering algorithms for logistics distribution use Euclidean distance as a measure of customer difference [10], which is not in line with the actual situation of distribution. In the real traffic network, due to the existence of various ground objects, the spatial samples are connected by a number of arc segments, and the distance between them is usually quite different from the linear distance between two points. The existing algorithms do not deal with the distribution volume and constraints effectively. If there is no solution, most of them discard the current clustering results and start the clustering process again.

This paper proposes the shortest path distance for the customer clustering problem with unknown location of distribution center. As a measure of difference among customers, the information of distribution traffic network is fully considered, which is in line with the actual situation of power logistics distribution. On this basis, a constraint clustering algorithm based on shortest path is proposed to meet the single load constraints of a single cycle. The single load rate, effective control and constraint clustering are solved. Because it is not necessary to calculate the mean value of all kinds of internal customers as the center point, the clustering process is not affected by isolated customers, and can effectively handle large data sets.

II. The Shortest Path Distance Problem

A. Proposal of the Shortest Main Route Distance

In the clustering algorithm of logistics distribution customers, the distance between customers in a certain sense should be calculated to measure the degree of association between customers. In view of the defect that the traditional European distance can not consider the distribution road information, the concept of the shortest path distance of customers is given.

In geographic information system (GIS), there are several routes from point A to point B, and the distance of each route includes trunk road distance and non trunk road distance. The shortest route is called the shortest route from point A to point B, and the trunk road distance in the shortest route is called the shortest main route distance from point A to point B, which is named SMRD (A, B). Compared with Euclidean distance and Manhattan distance, the shortest path distance can more accurately describe the distance between two customers on the map, so as to provide guarantee for effective customer clustering analysis.

B. The Calculation of the Shortest Path Distance

When the shortest path distance is obtained on the map, the road network should be represented structurally first, and then the shortest path distance should be calculated.

In the road network, roads intersect in an intricate way, forming a number of road nodes and road sections separated by nodes. Therefore, the road network information can be abstracted into node set $G=\{g_i\}$ and road section set $E=\{e_j\}$. For each road section e_j, the numbers of the initial point and the ending point and the weight of the road section (the length of the road section) are recorded. For each node g_i, such as the record of its spatial coordinates, the number of the link and other information. This structure comes from the data organization method in GIS, which reflects the real road entity more truly and reduces the data redundancy effectively.

C. Problem Descriptions and System Model

The descriptions of the vehicle routing problems for the logistics distribution system are specified as follows. Assuming that there are L customer pint, and their requirements are definitive. For satisfying the requirement of each customer, we require that the maximum number of vehicle that can be used is K, and the maximum load capacity for each vehicle is deemed as Q_k. The vehicle routing arrangement will be made to satisfy the total distance of all vehicles having travelled.

Parameter definition: L denotes the total sum of customer points, the quantity q_i at every point of goods is q_i ($i=1,2,..,L$), and K denotes the total sum of vehicles used for transportation. Let $d_{i,j}$ define as the distance between customer point i and j. The quantity of customer served by vehicle k is deemed as ${}^nk$, which follows the following conditions: if ${}^nk=0$, the vehicle k does not serve any customer. The maximum load capacity of vehicle k is Q_k. The set of customers served by the vehicle k is R_k, which satisfies the condition as follows: if ${}^nk=0$, $R_k=\theta$;if not, there is $R_k=\left\{r_k^1,r_k^2,....,r_k^{n_k}\right\}\subseteq\{1,2,...L\}$, where r_k^i means that the order of a certain customer point in the distribution routing of the vehicle k is i. And the mathematical model is given by

$$\min F=\sum_{k=1}^{K}\Big(\sum_{i=1}^{n_k}d_{r_k^{i=1},r_k^i}\Big)\operatorname{sgn}(n_k) \tag{1}$$

Therein (1), $\operatorname{sgn}(n_k)$ is defined as

$$\operatorname{sgn}(n_k)=\begin{cases}1 & n_k\geq 0\\ 0 & n_k=1\end{cases} \tag{2}$$

And, the constraint condition is as follows

$$\sum_{\mathrm{i}=1}^{n_k}q_{r_k^i}\leq Q_k\ ;\ nk\neq 0 \tag{3}$$

$$\sum_{\mathrm{i}=1}^{n_k}\left(d_{r_k^{i-1},r_k^i}+d_{r_k^i,0\leq D_k}\right);\ nk\neq 0 \tag{4}$$

$$R_{k_1}\bigcap R_{k_2}=\phi\ ;k_1\neq k_2 \tag{5}$$

$$\bigcup_{k=1}^{K}R_{k=\{1,2,...L\}}\ ;\ 0\leq nk\leq L,\sum_{\mathrm{k}=1}^{K}nk=l \tag{6}$$

III. Customer constrained clustering algorithm

A. Algorithm Principle

Traditional clustering methods include various clustering methods[11-13]. Among them, the *k*-center-based partition clustering method divides data objects into *k* classes. Users can specify the value of *k* according to practical application. This kind of algorithm can not identify outliers, but the clustering process is not affected by outliers.

The customer constrained clustering algorithm based on the shortest main road distance proposed in this paper is improved on the basis of CLARANS algorithm, so that it can process the geographic information of data objects, and the clustering results meet the constraints of distribution volume. The algorithm includes the following three parts.

(1) Determine the value of *k* and randomly select *k* initial center points.

To determine the value of *k*, it is necessary to fully predict the total distribution volume of customers in a certain period *T*, and then estimate the number of distribution vehicles in that period according to the single vehicle distribution capacity, which is the value of *k*.

Suppose that the total distribution volume of distribution customers in cycle T is TD, the single vehicle distribution capacity is SC, and the single vehicle load rate is small D, then the customer clustering number k is defined as

$$k = \frac{TD}{SC \cdot load} \tag{7}$$

Among them, the selection of cycle *T* must make each customer be delivered at least once in the cycle, and the value range of load rate of single vehicle is (0,1], which can be used to adjust the load rate of distribution vehicles. In ideal condition, load =100% is required, but it is difficult to achieve in practical application. In addition, by setting the load rate of single vehicle, the unsolved problem of customer clustering with constraints can be solved.

(2) Customer segmentation based on k-center.

The basic idea of the classical *k*-center clustering method is to find *k* representative objects as the representative centers of *k* classes, and assign the non representative objects to the classes represented by the nearest center point. According to a certain quality inspection standard, a representative object is selected to exchange with a non representative object, so as to maximize the clustering quality, Then the non representative objects are redistributed and the above process is iterated until the clustering quality cannot be improved.

The quality of clustering is measured by the total difference degree, which is the sum of the difference degrees between all objects and the center points of their classes. Because the cost of classical k-center clustering method is too large, it can not be effectively applied to large data sets. Therefore, clarans algorithm introduces global sampling technology to improve clustering efficiency and solve the application problems of large data sets

In CLARANS algorithm, a node is an object set composed of *k* objects, representing the selected *k* representative centers [14-15]. If only one object in two nodes is different, the two nodes become neighbors. During iteration, a neighbor sample is randomly selected from all the neighbors of the current node, from which the neighbor node which can improve the clustering quality is selected to replace the current node. Repeat the iteration until there is no node in the neighbor sample of the current node that can improve the clustering quality. This global sampling technique performs random sampling operation in each iteration, so that the search scope is not limited to local, and the algorithm is efficient, and can be applied to large data sets. The iterative process can be restarted many times, and the number of iterations and the neighbor sample size can ensure the effectiveness of the clustering results.

B. Algorithm Steps

The customer clustering method proposed in this paper inherits the iterative idea of clarans algorithm, adopts global random sampling technology, and applies the algorithm to large spatial data sets. The evaluation standard of clustering results is the total distance based on the shortest main road distance.

The steps of CLARANS algorithm are as follows.

(1)Select any node as the current node.

(2)Randomly select a neighbor s of the current node.

(3)Reallocate the non representative objects. If the total difference degree of *S* is smaller, the current node is replaced by *S* until the number of searching current nodes reaches the maximum number of neighbor samples specified by the user.

(4)Repeat the above steps until the number of restart clustering reaches the user specified iteration number threshold, and output the clustering result corresponding to the minimum total distance.

Suppose *n* customers r_i are divided into *k* classes C_j, and the set of center points of each class is denoted as MQ, then the total distance based on the shortest path distance is from each customer *n* to the center point of its class, denoted as SUM, that is, the total distance based on the shortest path distance.

$$SUM(MQ) = \sum_{j=1}^{k} \sum_{i=1}^{n} SMRD(r_i, mq_j) \tag{8}$$

C. Constraint Condition

When allocating customers, we should ensure to meet the customer distribution constraints. The total distribution amount of each customer class can not exceed the load capacity of a car, which requires that the distribution amount should be taken as a constraint condition in the process of customer clustering to limit the scale of each customer class. In the process of clustering, each customer class should make the total distance minimum on the basis of meeting the constraints of distribution amount. *k* representative points are selected to represent the centers of *k* customer classes. Now we want to divide the non center point customers r_g into one of the classes. The total distribution amount of all customers is not more than the distribution constraint.

The object partition considering distribution constraints is as follows. Firstly, the shortest path distance from r_g to *k* class centers is calculated, and then the *k* centers are sorted according

to the distance from small to large. Finally, it is judged whether the total distribution amount of r_g is greater than the distribution constraint LC after r_g is divided into class C with the smallest distance.

Sum constrained clustering problem may have no solution, that is, some customers can not be divided into any one class under the condition of meeting the distribution constraints. By setting the load rate of single vehicle, the occurrence of no solution can be effectively controlled. Clustering unsolved easily occurs at the end of the clustering process. If the distribution volume of the unassigned customers is large, clustering unsolved easily occurs.

The load rate of single vehicle is used to calculate the value of k. Using single vehicle load rate can avoid the strategy of clustering without solution. When calculating the value of k, it is considered that the vehicle is not full, and the value of single vehicle load rate is set. When customers are divided into different types according to the constraints, LC is considered as the distribution constraints.

In essence, the strategy is to increase the value of k to avoid the cluster without solution. Setting the single vehicle load rate is in line with the actual management, and it also provides a more objective basis for the increase of the value of k.

D. Encoding and Decoding of Vehicle Routing Optimization

The key to solving the vehicle routing problems for logistics distribution system is to find the location of the particle corresponding to the optimal solution. The proposed algorithm constructs $2m$ dimension space corresponding to m customer points, and requires both the vehicles having finished the distribution task and the executive order of these vehicle in the path correspond with each customer point. The optimization of vehicle routing problem for each customer point in the $2m$ dimension space is defined as the vehicles having finished the distribution task for each customer and their executive orders. To particle i, the $2m$ vector can be considered as two m dimension vectors Z_{ix} and Z_{iy}, where Z_{ix} denotes the vehicle number, and Z_{iy} denotes the path order of vehicle travelling among the customers.

The decoding of the proposed algorithm for vehicle routing optimization is as follows. According to the truncating operation for the vector Z_{ix} of any particle i, we can obtain the vehicle j served for the customer i by decoding the vehicle number.

Firstly, look for customer i served by vehicle j, and order them according to the vector Z_{iy} related with customer i, where the ordering principle follows from small to larger numbers so that it can determine the path order of the vehicle by decoding the vehicle path order.

For instance, the number of vehicles is 3 and the number of customers is 7. Table 1 describes the location vectors of particle i before decoding, and the location vectors of particle i after decoding is illustrated in Table 2.

Table 1. Location vectors before decoding

Customer	1	2	3	4	5	6	7
Z_{ix}	1.7	2.3	2.6	3.4	3.9	1.2	3.6
Z_{iy}	0.9	2.8	3.7	1.4	2.6	4.4	1.8

Table 2. Location vectors after decoding

Customer	1	2	3	4	5	6	7
$Int(Z_{ix})$	1	2	2	3	3	1	3
Z_{iy}	0.8	2.6	3.5	1.9	2.9	4.7	2.0

IV. EXPERIMENT

Aiming at the customer constrained clustering algorithm based on shortest path distance proposed in this paper, numerical experiments are carried out by using randomly generated data and actual data of power logistics distribution customers. The purpose of the experiment is to intuitively compare and analyze the clustering effect based on shortest path distance and Euclidean distance, Finally, the clustering effect of considering and not considering the distribution constraints is analyzed.

Randomly generated data includes road network information and customer data. Four intersecting roads are generated from road network information, and a total of 12 road sections and nodes of 12 road sections are obtained. In the above road network, 40 points are randomly generated as customer locations, and the distribution volume of each customer is assigned with random numbers. Among them, 25 customers are randomly assigned with [2,100] integers; Among the remaining customers, 10 are randomly assigned with [101,300] integers; The other five points are randomly assigned with an integer within the range of [301,500].

The actual data of power logistics distribution customers are described as follows.A certain logistic distribution company has a distribution center with the location coordinates (0,0). There are also 20 customer points which are distributed in the area with 10 square kilometers. The maximum vehicle load capacity is 9 tons, and the locations of customer points and their demands are illustrated in Table 3.

Table 3. Locations of customer points and their demands

No	X /km	Y /km	Demand /q	No	X /km	Y /km	Demand /q
1	0	0	0.0	11	1	3	0.7
2	0	-2	1.6	12	3	4	0.3
3	0	3	1.7	13	-3	0	2.3
4	-2	-2	2.1	14	2	0	1.7
5	-3	-2	0.9	15	1	-3	2.3
6	4	-1	1.4	16	2	-1	0.8
7	-4	0	1.3	17	2	1	0.4
8	-3	-1	2.0	18	1	-4	2.4
9	1	-2	3.1	19	-3	2	3.0
10	1	-1	1.8	20	-1	-1	0.2

All the experiments are operated on the Windows XP system with 4.2GHz Inter Process and 4 GB RAM by using MATLAB. The major parameters are setting as follows. The population is 100, maximum iterative times are 1000, the maximum weight of gradient adaptive inertia is 0.98, the minimum weight of gradient adaptive inertia is 0.08, and initial learning factor value respectively are 2 and 1.4.

According to the calculation, the loading rate of each type in the clustering results considering the distribution capacity constraint is more balanced, and the value range of loading rate is [76.92%, 90.75%] (the loading rates of each type are: Class 1 (923), class 2 (945), class 3 (1089), class 4 (1071), class 5 (1046)). However, in the clustering results without considering the distribution capacity constraint, the loading rate is extremely uneven, and the maximum loading capacity is 1754 (class P), which is about 1.5 times of the vehicle loading capacity constraint (LC = 1200), The minimum carrying capacity is 114 (class q), which accounts for 9.5% of the vehicle carrying capacity, that is, the value range of carrying rate is [9.5%, 150%]. In the actual distribution process, under the condition of a certain number of carrying vehicles, we should try our best to pursue full load and make each vehicle carrying capacity balanced.

V. Conclusion

The customer clustering problem of power logistics distribution is a basic problem in the research field of power logistics distribution. The actual problem requires considering the geographical characteristics of customer distribution, customer distribution volume (i.e. demand) and distribution vehicle load, This paper proposes a clustering algorithm considering the distribution network structure and distribution volume constraints. On this basis, the traditional k-center partition clustering algorithm clarans is improved by using the idea of constrained clustering, so that it can not only consider the geographical information, but also can improve the clustering algorithm, Finally, numerical experiments are carried out on the proposed algorithm.

With the development of economy, the research on power logistics routing will be more in-depth, and the factors affecting the selection of distribution routing will be more and more complex. Looking forward to the future, the paper will also carry out follow-up research from the following aspects.

(1)In this paper, when studying and establishing the power logistics path optimization problem model, there are many assumptions, which are different from the actual situation of the problem. Although the paper considers the change of vehicle running speed in the case of power logistics distribution, there is only a single distribution center and only a single behavior (pick-up or delivery), while in the real power logistics distribution, there are more situations such as multiple distribution centers, simultaneous delivery and pick-up, forward and reverse material flow, and distribution vehicles do not return to the distribution center, Follow up research needs to consider these actual situations.

(2)This paper only considers the power logistics distribution problem under multiple conditions, but in real life, the distribution of goods is a complex system. The follow-up research of this paper will consider the path optimization problem from the perspective of system.

(3)When the algorithm designed in this paper is used to solve the problem, it adopts the method of distributed calculation, and does not solve the problem from the overall point of view. If the index selection in the algorithm solving process is unreasonable, it will have a great impact on the result of the problem and limit the flexibility of solving the problem. The follow-up research of this paper will use artificial bee colony algorithm combined with other intelligent optimization algorithms to solve the optimization problem of power logistics distribution path as a whole.

References

[1] A. Lim, F. Wang. Multi-depot vehicle routing problem: a one-stage approach, *IEEE Transactions on Automation Science and Engineering*, 2019, Vol. 2, No. 4, pp. 397-402.

[2] Xiang Y, Wang Y, Su Y, et al. Reliability correlated optimal planning of distribution network with distributed generation. *Electric Power Systems Research*, 2020, 186:106391.

[3] Suboh Alkhushayni, Taeyoung Choi, DuAlzaleq, Data Analysis using Representation Theory and Clustering Algorithms, WSEAS Transactions on Computers, Volume 19, 2020, Art. #38, pp. 310-320.

[4] B. Zhao. A multi-agent-based particle swarm optimization approach for optimal reactive power dispatch, *IEEE Trans on Power Systems*, 2020, pp.1070-1078.

[5] J W, Zeemering S, Isaacs A, et al. Female sex, atrial fibrillation, and heart failure, but not ageing, are associated with endomysial fibrosis in atrial myocardium: results from the CATCH ME consortium. EP Europace, 2021(Supplement_3), pp.399-411.

[6] Greco L , Lucadamo A , Agostinelli C . Weighted likelihood latent class linear regression. *Statistical Methods & Applications*, 2021, 30, pp. 274-288.

[7] Y. Gong, J. Zhang, O. Liu and R. Huang, Optimizing the vehicle routing problem with time windows: a discrete particle swarm optimization approach, *IEEE transactions on systems, man, and cybernetics*, 2020, Vol. 42, No. 2, pp. 254-267.

[8] S. L. Ho, S. Y. Yang, G. Z. Ni and K. F. Wong. An improved PSO method with application to multimodal Functions of Inverse Problems, *IEEE Transactions on magnetics*, 2017, Vol. 43, No. 4, pp. 1597-1600.

[9] Khushboo Jain, Anoop Bhola, An Optimal Cluster-Head Selection Algorithm for Wireless Sensor Networks, WSEAS Transactions on Communications, Volume 19, 2020, Art. #1, pp. 1-8.

[10] Y. SONG, C. YE and Z. HUANG. Cuckoo search algorithm for multi-resource leveling optimization, *Journal of Computer Applications*, 2018, Vol. 34, No. 1, pp. 189-193. (in Chinese)

[11] C. QU, Y. FU. An optimal Cuckoo search algorithm based on hybrid mutation operator, *Science technology and engineering*, 2018, Vol. 13, No. 27, pp. 8008-8009. (in Chinese)

[12] N. V. Dieu, P. Schegner and W. Ongsakub, Cuckoo search algorithm for non-convex economic dispatch, *IET Generation, Transmission & Distribution*, 2019, Vol. 7, No. 6, pp. 645-654.

[13] R. V. Kulkarni, G. K. Venayagamoorthy. Particle swarm optimization in wireless-sensor networks: a brief survey, *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, 2021, Vol. 41, No. 2, pp. 262-267.

[14] F. Wang, Q. Wu, Particle swarms optimization for vehicle routing problem with time window, *in Proc. Int. Conf. Risk Rel. Manage.*, 2018, pp. 962-966.

[15] H. Huang, C. Wu and Z. Hao, A pheromone-rate-based analysis on the convergence time of ACO algorithm, *IEEE transactions on systems, Man, and Cybernetics*, 2019, Vol. 39, No. 4, pp. 910-923.

[16] Ferraz B P , Resener M , Pereira L A , et al. MILP model for volt-var optimization considering chronological operation of distribution systems containing DERs. *International Journal of Electrical Power & Energy Systems*, 2021, 129.

[17] Abbasi A, Mohammadi B. A clustering ‐ based anonymization approach for privacy, reserving in the healthcare cloud. *Concurrency and Computation Practice and Experience*, 2021(6).