Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230807

Constrained K-Means Algorithm and Its Application in Distribution Center Location Problem

Jianjun ZHAN[a], Nan XU[b], Yingqiang XU[c,1], Jufeng WANG[d],
Xu ZHANG[e], and Zhenyu GAO[f]
[a]*National University of Defense Technology, Hunan, China*
[b]*Xianghongqi Road, Haidian District, Beijing, China*
[c]*Department of Industry and Information Technology of Yunnan Province, Yunnan, China*
[d]*East China University of Science and Technology, Shanghai, China*
[e]*Renming University of China, Beijing, China*
[f]*Department of Intelligent Supply Chain, JD Logistics, Beijing, China*

Abstract. To solve the constrained clustering problem, this paper improves the K-means and proposes a constrained K-means algorithm (CK-means). CK-means algorithm takes into account both clustering analysis and constraints, and can effectively deal with clustering problems with constraints, such as distribution center location problem with warehouse capacity constraints, vehicle routing problem with capacity constraints, etc. It has higher practical value and a wider range of applications. There are two core innovations of the CK-means algorithm: firstly, incorporating constraints into the K-means. The second is a search strategy based on sample weights. In addition, this paper also applies the CK-means algorithm to the location problem of distribution stations at the end of JD Logistics' supply chain. The experimental results show that the CK-means can solve the clustering problem with constraints with effect.

Keywords. Cluster Analysis, K-means, Constrained Clustering problem, CK-means

1. Introduction

As one of the commonly used algorithms in Machine Learning (ML) algorithms, clustering algorithm has the characteristics of simplicity, practicality and high efficiency. It has been successfully applied in many fields, such as market segmentation [1-2], image segmentation[3], document analysis [4], location [5] and so on.

Clustering algorithm is a classical unsupervised learning algorithm. In the case of a given sample, the clustering algorithm automatically divides the samples into several categories by measuring feature similarity or distance [6]. Distance measure and similarity measure are the core concepts of clustering analysis. Most clustering algorithms use distance measure [7]. Common distance metrics include Manhattan distance, Euclidean distance, Chebyshev distance and Mahalanobis distance. Common

[1] Corresponding Author: Yingqiang XU, Department of Industry and Information Technology of Yunnan Province; e-mail: xyq408@qq.com

similarity measures include correlation coefficient and angle cosine. There are four common clustering algorithms, distance-based clustering algorithm, density-based clustering algorithm, hierarchical clustering algorithm, spectral clustering algorithm based on graph theory.

The K-means is one of the commonly used clustering algorithms, which was proposed by MacQueen [8] in 1967. The characteristics of K-means are practical and simple; however, it also has some limitations, such as the parameter k needs to be determined in advance, and the initial cluster centers are randomly generated. For the K-means, scholars in various fields have proposed many improvement strategies, mainly focusing on the following directions: Initialization of k-value[9], the selection of the initial clustering center [10], the detection and removal of outliers [11], distance and similarity measurement [12], etc.

The K-means is widely used and often used in location problems. For example, the distribution center location problem, many scholars use K-means to cluster the demand points to form K clusters ; then, a distribution center is established in each cluster. However, this location scheme does not consider whether the capacity of the distribution center can meet the total demand of the demand points in the cluster. Once the total demand of the demand points in the cluster exceeds the distribution capacity of the distribution center, it will lead to problems such as untimely distribution and reduce user satisfaction. Therefore, this paper will improve the K-means and propose a constrained K-means algorithm. While completing the clustering task, some practical constraints are considered to improve the practical application value of K-means. And the CK-means is applied to the distribution center location problem.

2. Constrained K-means Algorithm

2.1. K-means Algorithm

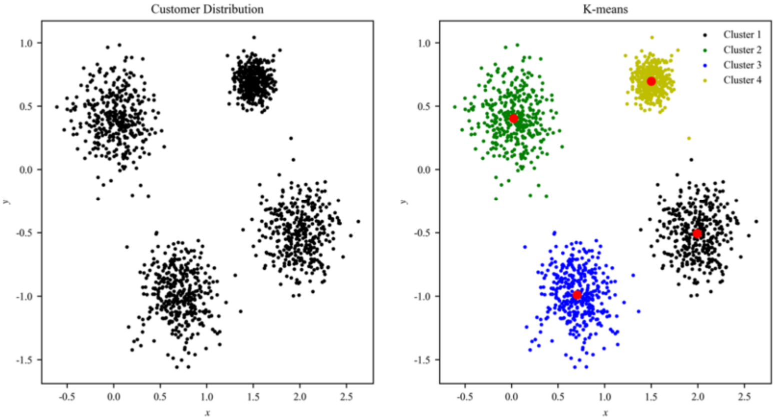

Figure 1. K-means algorithm clustering diagram.

K-means is an unsupervised learning algorithm, that is, clustering analysis of unlabeled data sets. The algorithm measures the similarity between samples by the size of the distance, and divides the sample set into K clusters, as shown in Figure 1, showing an example of the K-means. Its core idea is to minimize the difference between samples in the cluster and maximize the difference between clusters; the core steps are as follows : firstly, K samples are randomly selected as the initial clustering centroid, and each centroid represents a cluster ; secondly, the sample is divided into the nearest cluster ; then, the centroid of the cluster is recalculated ; Finally, determine whether the conditions for terminating the clustering algorithm are met. K-means can be summarized as the partition of the sample set $X = \{x_1, x_2, \dots, x_n\}$. Its learning strategy is to optimize the partition by minimizing the loss function.

Given a sample set $X = \{x_1, x_2, \dots, x_n\}$, its dimension is $m \times n$. The K-means is used to cluster the data set X, and n samples are divided into K clusters ($K \leq n$). The minimum loss function model can be defined as :

$$f = min \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \bar{x}_l||^2 \tag{1}$$

C_i is the $i - th$ cluster ; $\bar{x}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \dots, \bar{x}_{im})$ denotes the centroid of the $i - th$ cluster, i.e., the center point.

2.2. Constrained K-means Algorithm (CK-means)

The traditional K-means measures the similarity between samples by distance, and then realizes the division of sample sets. This clustering algorithm does not consider whether the divided clusters meet other constraints, and has limitations. It can not solve the clustering problems with constraints, such as warehouse location problem with capacity constraints, vehicle routing problem with capacity constraints, etc. Therefore, this paper proposes a K-means algorithm with constraints, referred to as CK-means. The CK-means algorithm takes into account both cluster analysis and constraints, has higher practical value and wider application range.

Given the sample set $X = \{x_1, x_2, \dots, x_n\}$, the dimension is $m \times n$; let $d = \{d_1, d_2, \dots, d_n\}$ denote the demand set of each sample, that is, the demand of sample x_i is d_i; let $t = \{t_1, t_2, \dots, t_K\}$ denote the restriction on the total demand of each cluster. Now we need to cluster the data set X to get K clusters, but to ensure that the total demand of each cluster $C(k)$ does not exceed the limit, that is, $\sum_{x_i \in C(k)} d_i \leq t_k$. The CK-means is used to cluster the data set X, and the n samples are divided into K clusters ($K < n$). The mathematical model can be defined as :

$$f = min \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \bar{x}_l||^2 \tag{2}$$

$$s.t. \sum_{x_i \in C(k)} d_i \leq t_k, \forall k = 1,2, \dots, K \tag{3}$$

C_i is the $i - th$ cluster ; $\bar{x}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \dots, \bar{x}_{im})$ denotes the centroid of the $i - th$ cluster, i.e., the center point. The existing literature[13] has proved that the model of K-means algorithm (i.e., Eq. (1)) is an NP-hard problem; obviously, the mathematical model of CK-means algorithm (i.e., Eqs. (2) and (3)) is also an NP-hard problem. Therefore, this paper designs a heuristic strategy to minimize the loss function of CK-means algorithm. The pseudocode for CK-menas is shown in Table 1.

Table 1. Pseudocode for CK-means algorithm

Algorithm 1 CK-means

```
INPUT: Sample set: $X = [xx_1, xx_2, \dots, xx_n]$; Parameters: $K$、$maxiter$、$d = \{dd_1, dd_2, \dots, dd_n\}$、$t = \{tt_1, tt_2, \dots, tt_K\}$
OUTPUT: Cluster results: $C = \{C_1, C_2, \dots, C_K\}$
$pp(0) = \left(pp_1^{(0)}, pp_2^{(0)}, \dots, pp_K^{(0)}\right)$ // Initialize centroids, that is, at the 0th iteration, randomly select K samples as the initial clustering centroids
$numiter = 0$
while numiter<maxiter:
    numiter+=1
    C={}
    data=np.array([[xx1,tt1,dd1],[xx2,tt2,dd2],…,[xxn,ttn,ddn]])// Combine the sample set X, t, and d into an array with a dimension size of (m+2) * n
    data=data[np.argsort(-data[:,-1])]// Sort data in descending order based on d
    X'=data[:,:-2]
    t'=data[:,-2]
    d'=data[:,-1]
    for i=0 to n-1:
        x=X'[i]
        min_=np.inf
        idx=0
        tabu=[]
        while True:
            for k=0 to K-1:
                if k not in tabu:
                    center=p(numiter-1)[k]
                    dist=calculate_dist(x,center)// Calculate the distance between x and center
                    if dist<min_:
                        min_=dist
                        idx=k
                    end if
                end if
            end for
            if sum(d'[C[idx]])+d'[i]<=t'[idx]:
                C[idx].append(i)// Classify sample i into cluster idx
                break
            else:
                tabu.append(k)
            end if
        end while
    end for
    for k=0 to K-1:// Update centroid
        cluster_x=X'[C[k]]
        pp(numiter)[k]=np.mean(cluster_x)
    end for
end while
return C
```

Obviously, the parameter K of CK-means algorithm should satisfy the following conditions : $\sum_{i=1}^{n} d_i \leq \sum_{k=1}^{K} t_k$. And we define the value of $maxiter$ as $n * 10$. There are two core innovations of CK-means algorithm :

- In the clustering analysis, it can cope with the constraints and meet the actual application requirements.
- Before the iteration of the algorithm, all samples are sorted in descending order according to the demand of samples. In the iterative process, the samples are classified in the above order ; this strategy ensures the stability of CK-means algorithm.

3. Experiment and Analysis

CK-means can effectively deal with clustering problems with constraints, such as distribution center location problem with warehouse capacity constraints, vehicle routing problem with capacity constraints, etc. It has great application value. This paper takes the location problem of distribution stations at the end of JD Logistics' supply chain as an example to analyze the application of the CK-means algorithm. Jingdong has n customers in a certain area. It is known that the location set of all customers is $X = \{x_1, x_2, \ldots, x_n\}$, and the dimension of X is $2 \times n$. The customer's demand set is $d = \{d_1, d_2, \ldots, d_n\}$, and the upper limit of the capacity of the distribution station is ub, as shown in Table 2 (Table 2 only shows a portion of the data. If you need a complete dataset, you can contact the author to obtain the complete data). We need to establish several suitable distribution stations in the region based on the above information.

Therefore, we first need to determine the number of distribution stations, which is $K \geq K_0$ ($K_0 = int\left(\frac{sum(d)}{ub}\right) + 1$). Because the construction cost of the distribution station is high, the number of distribution stations should be reduced as much as possible. In this paper, the number of distribution stations is set to $K_0, K_0 + 1, K_0 + 2, \ldots$ until the K that can meet the distribution center location problem with warehouse capacity constraints is found.

Table 2. Customer data set (the data set is desensitized data)

ID	x_{i1}	x_{i2}	d
1	45.7	68.2	8
2	20.8	51.9	7
3	30.9	67.2	9
4	70.6	56.9	2
5	60.2	45.8	1
…	…	…	…
n	67.8	67.2	11

In this case, $ub = 150$. By analyzing the data in Table 2, $K_0 = 6$ can be calculated. Through experiments, $K = 6$ can be calculated. At this time, CK-means can achieve customer clustering analysis when meeting the warehouse capacity constraints. The experimental results are shown in Figure 2, and the (a) represents the distribution of customers in the region; the (b) represents the clustering results without considering the warehouse capacity constraints. The (c) represents the clustering result considering the warehouse capacity limit, and the red point represents the centroid of each cluster. Through comparison, it can be found that there are some differences between the middle subgraph and the right subgraph. This is because the CK-means algorithm considers the constraints when performing cluster analysis, which makes the results different. In the traditional K-menas algorithm, the total demand of customers in the six clusters is (135,140,155,162,138,127). The total demand of some clusters exceeds the capacity limit of the distribution station, which will lead to the failure of the distribution station to complete the distribution task in time and reduce the service level of the enterprise.

In the CK-means algorithm, the total demand of six clusters is (147,145,150,147,139,129), and the total demand of all clusters does not exceed the limit conditions of the distribution station, which is conducive to the normal operation of each distribution station and thus improves the service level of the enterprise.

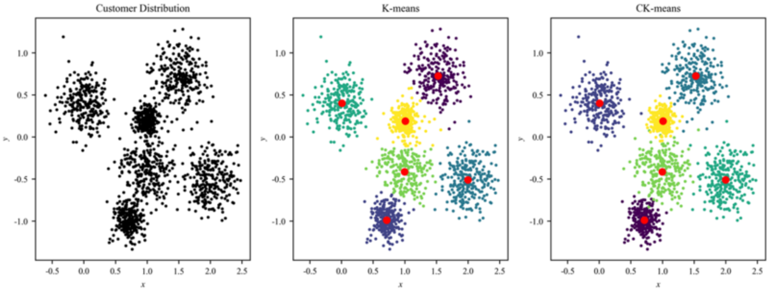

Figure 2. Experimental results of distribution center location problem with warehouse capacity constraints.

4. Conclusion

The K-means has the characteristics of simple principle, easy implementation and high stability, and its application is very extensive. However, the traditional K-means also has some limitations, for example, it cannot solve the constrained clustering problem. To solve the constrained clustering problem, this paper improves the traditional K-means and proposes a constrained K-means (CK-means). When clustering, the CK-means algorithm will take into account the constraints that the clustering results need to meet. CK-means provides an effective solution to solve constrained clustering problems, such as warehouse location problem with warehouse capacity constraints, vehicle routing problem with capacity constraints, etc. Finally, this paper takes the location problem of distribution stations at the end of JD Logistics' supply chain as an example, and applies the CK-means to solve the location problem with capacity constraints. The experimental results show that the CK-means can effectively solve the clustering problem with constraints.

References

[1] Tleis, M. , Callieris, R. , & Roma, R. . (2017). Segmenting the organic food market in lebanon: an application of k-means cluster analysis. British food journal(7), 119.

[2] Hung, P. D. , Ngoc, N. D. , & Hanh, T. D. . (2019). K-means Clustering Using R A Case Study of Market Segmentation. the 2019 5th International Conference.

[3] Dhanachandra, N. , Manglem, K. , & Chanu, Y. J. . (2015). Image segmentation using k -means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54, 764-771.

[4] Li, W. , Feng, Y. , Li, D. , & Yu, Z. . (2016). Micro-blog topic detection method based on btm topic model and k-means clustering algorithm. Automatic Control and Computer Sciences, 50(4), 271-277.

[5] Wu J., Liu X, Li Y.Y., Yang L.P., Yuan W.Y. and Ba Y.L..(2022). A Two-Stage Model with an Improved Clustering Algorithm for a Distribution Center Location Problem under Uncertainty.Mathematics, 10(14): 2519

[6] Aggarwal, C. C. , & Reddy, C. K. . (2013). Data Clustering: Algorithms and Applications. Chapman & Hall/CRC.

[7] Shang, T. , Zhao, Z. , Ren, X. , & Liu, J. . (2021). Differential identifiability clustering algorithms for big data analysis. Chinese Science: Information Science (English Version).

[8] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

[9] Rodriguez, A. , & Laio, A. . (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492.

[10] Tanir, D. , & Nuriyeva, F. . (2017). On selecting the Initial Cluster Centers in the K-means Algorithm. 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT). IEEE.

[11] Fan, Z. , Sun, Y. , & Luo, H. . (2017). Clustering of college students based on improved k-means algorithm. Journal of Computers (Taiwan), 28(6), 195-203.

[12] Xu Y.. (2018). A k-means algorithm based on feature weighting. Computer Science and Application, 08(8), 1164-1171.

[13] Guan Y.J.. (2019). Location of emergency logistics center based on k-means clustering algorithm. Pure Mathematics, 09(7), 809-812.