Accounting, Auditing and Finance (2020) 1: 5-8
Clausius Scientific Press, Canada

DOI: 10.23977/accaf.2020.010102
ISSN 2616-3136

Introduction to the K-Means Clustering Algorithm Based on the Elbow Method

Mengyao Cui

Shandong University of Finance and Economics, Jinan, Shandong, China

178742173@qq.com

Keywords: K-means clustering, Elbow method

Abstract: The K-means clustering algorithm is a commonly used algorithm in the financial field, and it is also an unsupervised learning algorithm. It is characterized as an easy and simple algorithm and is widely used in the fields such as machine learning and stock trading. However, the K-means algorithm also has certain shortcomings. The k value is difficult to determine, and the initial center of the cluster is difficult to find. This article introduces the idea of K-means algorithm, using the elbow method to find the most suitable k value.

1. Introduction

2. Clustering Algorithms

Clustering is a data mining technique that divides a data set into multiple different categories by calculating the similarity between data[1]. There is a high degree of similarity within each cluster of data classified by the clustering algorithm, and the similarity between different clusters of data is relatively large. Clustering algorithms usually include hierarchical methods, density-based methods, and partitioning methods. K-means algorithm is a widely used clustering method and is also the method used in this article.

3. K-Means Clustering

3.1 General Principle

k-means clustering is abbreviated as K-means, which is an unsupervised learning model[2]. Unsupervised learning models are used for data sets that have never been labeled or classified. It records the same points in the data set and responds accordingly to these same points in each data point.

The model is a centroid-based algorithm or a distance-based algorithm, and we calculate the distance to assign points to each cluster. First take a K value, and then divide the data into K categories, so that the similarity of the data in the same category is higher, which is convenient for distinguishing.

3.2 Euclidean Distance

Usually, the index to measure the similarity between data is Euclidean distance. Euclidean distance is the distance between two points in two-dimensional and three-dimensional space.

□ Two-dimensional formula:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

□ Three-dimensional formula:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

□ It can also be generalized to n-dimensional space:

$$d = \sqrt{\sum (x_{i1} - x_{i2})^2}$$

where i=1, 2, .., n; x_{i1} represents the i-th dimensional coordinate of the first point; x_{i2} represents the i-th dimensional coordinate of the second point.

The n-dimensional Euclidean space is a set of points, and each point of it can be represented as (x(1), x(2),...x(n)), where x(i)(i=1, 2. ..n) is a real number, called the i-th coordinate of x, the distance between two points x and y=(y(1),y(2)...y(n)) d(x,y) is defined as the above formula.

The smaller the Euclidean distance between data, the higher the similarity between them. At the same time, there are only K centroids at the beginning, and each cluster is associated with a centroid, and data points close to the centroid will be classified as a cluster for classification.

3.3 The Main Goal and Steps of K-Means

The main goal of the K-Means algorithm is to minimize the sum of the distances between the points and their respective cluster centroids, and cluster them in an iterative manner. The steps are as follows [3]:

First, select K objects as the center of the initial cluster according to our research objective. Calculate the Euclidean distance between the data and the cluster center, and divide the data with close distance into one category [4].

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \le j \le k \right\}$$

Second, recalculate the cluster centers of the newly divided clusters, and re-divide the clusters in the previous way according to the new cluster centers.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Third, iterative operation in this way until the cluster center no longer changes, then the algorithm can be stopped.

There are three types of criteria for stopping the algorithm: first, the centroid of the newly formed cluster remains unchanged, second, the points remain in the same cluster, and the third reaches the maximum number of iterations. If the centroid of the newly formed cluster has not changed, we can stop the algorithm. Even after multiple iterations, if all clusters have the same centroid, then the algorithm has not learned any new patterns, and training can be stopped at this time. Another sign of explicitly stopping training is that even after training the algorithm multiple iterations, if these points are still in the same cluster, the training process should be stopped at this time. Finally, if the maximum number of iterations is reached, we can stop training. Suppose we set the number of iterations to 200. Before stopping, the process will repeat 200 iterations.

4. Elbow Method

To make the centroid of this algorithm no longer change, it must pay attention to the choice of K value, but there is a shortcoming, which is related to the choice of the initial K points. To solve this problem, the performance of the algorithm is calculated for different numbers of centroids. When evaluating, as long as convergence occurs, the distance between the centroid of each cluster and the data point can be calculated. Then add up all the calculated distances as a performance indicator. As the number of cluster centroids increases, the size of the objective function will decrease. In order to select the best K, the elbow method is usually used.

The elbow method is suitable for relatively small k values. The elbow method calculates the squared difference of different k values. As the k value increases, the average distortion degree becomes smaller. The number of samples contained in each category decreases, and the samples are closer to the center of gravity. As the k value increases, the position where the improvement effect of the distortion degree decreases the most is the k value corresponding to the elbow.

Here we introduce a variable, WCSS (Within-Cluster Sum-of-Squares), which measures the variance within each cluster. The better the clustering, the lower the overall WCSS.

For example, at first, the value of K is defined as 1, and the value of WCSS is higher at this time. Then the value of k is equal to 2, and the calculated WCSS value is lower than when k is equal to 1. When k is equal to 3, the value of WCSS also decreases. But when we choose k to be equal to 4, the drop in WCSS value is small and not obvious. The WCSS will equate to zero because every single point has its own cluster. It has its own centroid and that centroid is going to be exactly where that point is so the distance between the point and centroid is going to be zero.

□ Elbow Method Formula for WCSS:

$$WCSS = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

We use the following figure as an example. When k is equal to 3, it is an inflection point of this curve, which can be regarded as the “elbow” of this curve. It determines the k value, that is, the optimal number of clusters in this example is three.[5]

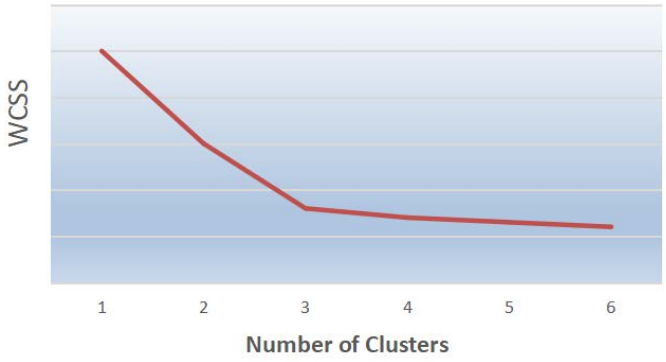

Fig.1 An Example of Wcss

5. Conclusion

Although this algorithm is well developed now, it still faces challenges in the future and has a lot of room for improvement. At the same time, we can apply this algorithm to many fields and accomplish many practical tasks. For example, we can apply the K-mean algorithm to stock trading. When investing on the U.S. stocks, we can use data science techniques to narrow the search scope and group similar stocks using K-means clustering algorithm. We know that the purpose of the

clustering algorithms is to classify data sets and aggregate similar data together.

Since the K-means algorithm was proposed, many scholars have continued explaining and improving to enrich the content of the algorithm. The K-means algorithm has been widely used and will be very useful in the future. For the process of the k-means algorithm and how to select the optimal K value, this article has already elaborated.

References

[1] Dubey A,Choubey A.A systematic review on k-means clustering techniques[J].International Journal of Scientific Research Engineering & Technology,2017,6(6):624-627.

[2] Saroj,Kavita.Review:study on simple k mean and modified K mean clustering technique[J].International Journal of Computer Science Engineering and Technology,2016,6(7):279-281.

[3] Anil K.Jain, Data clustering: 50years beyond k-means, 2009, pp.651-666.

[4] Jain A K,Murty M N,Flynn P J.Data clustering:a review[J].ACM Computing Surveys,1999,3(3):264-323.

[5] Alexander L,Jiang S,Murga M,et al.Origin-destination trips by purpose and time of day inferred from mobile phone data[J].Transportation Research:Part C Emerging Technologies,2015,58:240-250.