A Multi-angle Improved Small Sample Clustering Algorithm

XIAO Bin\(^1\), LIU Chang\(^{1,2,*}\)

\(^1\)Department of management engineering and equipment economics, Naval University of Engineering, Wuhan, Hubei Province, 430033, China
\(^2\)Armed Police Second Mobile Corps, Fuzhou, Fujian Province, 350200, China
\(^*\)Corresponding author’s e-mail: liuchmx@126.com

Abstract. The random selection of initial clustering centers, outliers, and the differences between attributes will affect the clustering effect of k-means. This article first uses the elbow method to determine the number of categories and then uses the square difference radius method to select the cluster seed center to optimize the cluster center's reselection. Finally, the entropy method is used to calculate the difference between attributes. The results show that when the number of categories remains the same and abnormal data is added, the improved clustering algorithm from multiple perspectives is more accurate and stable for small sample data with small dimensions and large differences between categories.

1. Introduction
Cluster analysis is a method of grouping samples according to the sample data itself. It originated from taxonomy\(^1\) and was first proposed by the American scholar Mac Queen\(^2\). It is currently used in many fields. Its general function is to summarize each category's characteristics by mining the in-depth information in the data to make corresponding decisions. For example, companies develop different marketing mixes by studying customer consumption characteristics in the business field, dividing customer categories.

K-means is a distance-based non-hierarchical clustering algorithm\(^3\). The existing drawbacks mainly include the following 4 points: ① The value of k must be determined first. ② The initial center will affect the final result. ③ The classification attribute will lead to the local optimum. ④ Outliers will cause interference. To this end, scholars have proposed a variety of improvement methods. For example, document [4] proposed hierarchical clustering with \(k^*(k>k^*)\) seed initialization. Literature [5] uses genetic operators to improve clustering. Literature adopted the k-mode method that uses the modulus to replace the mean. Literature [6] proposed the CURE method to filter outliers. Literature classifies Hadoop and offers a new model of STING. However, many experts and scholars have optimized the single aspect of the four shortcomings of the k-means algorithm to a certain extent. They have not considered all aspects of the small sample data set.

Given the four deficiencies of the k-means algorithm, using the elbow method to determine the number of categories \(k\) to solve the defects, the number of classes cannot be given in advance; optimizing the initial clustering centers to avoid falling into local minimums; then the clustering centers are recreated. The selection process reduces the sensitivity of the algorithm to outliers; finally, weighing the attributes to correct the Euclidean distance between data objects so that the algorithm can achieve more accurate and significant clustering effects.
2. The algorithm of this paper

2.1. Entropy method to determine attribute weight

Suppose the data set \( A = \{a_1, a_2, \cdots, a_n\} \) and the probability distribution \( p_i = P[X = a_i] \), the self-information amount of a particular data point can be expressed as \( I(a_i) = \log \frac{1}{p_i} \), and the average information amount (information entropy) of the data set can be expressed as:

\[
H(X) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} (i = 1, 2, \cdots, n)
\]  

(1)

Proceed as follows:

**Step 1** Construct attribute value matrix

\[
X = \begin{bmatrix}
    x_{11} & x_{12} & \cdots & x_{1n} \\
    x_{21} & x_{22} & \cdots & x_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{m1} & x_{m2} & \cdots & x_{mn}
\end{bmatrix}
\]

\( n \) is the number of objects, and \( m \) is the number of object dimensions.

**Step 2** The attribute units and values are different, and the dimension is eliminated after standardization. In the process of standardization, it is necessary to judge whether the impact of each index on the classification is positive or negative.

Positive indicator expression:

\[
x' = \begin{cases}
    1 & x \geq x_{\text{max}} \\
    \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} & x_{\text{min}} \leq x \leq x_{\text{max}} \\
    0 & x \leq x_{\text{min}}
\end{cases}
\]

Negative indicator expression:

\[
x' = \begin{cases}
    0 & x \geq x_{\text{max}} \\
    \frac{x_{\text{max}} - x}{x_{\text{max}} - x_{\text{min}}} & x_{\text{min}} \leq x \leq x_{\text{max}} \\
    1 & x \leq x_{\text{min}}
\end{cases}
\]

\( x_{\text{max}} \) is the maximum value of an attribute, and \( x_{\text{min}} \) is the minimum value of an attribute.

**Step 3** The entropy value of the \( j \)-th dimension attribute of the \( i \)-th object:

\[
M_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}
\]  

(2)

In formula (2): \( M_{ij} \) represents the specific gravity, and \( x_{ij} \) represents the attribute value.

**Step 4** Entropy value of the \( j \)-th dimension attribute:

\[
H_j = -\frac{1}{\ln n} \sum_{i=1}^{n} M_{ij} \ln M_{ij}
\]  

(3)

When \( M_{ij} = 0 \), \( M_{ij} \ln M_{ij} = 0 \). If \( x_{ij} \) is equal to a given \( j \), then \( M_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}} = \frac{1}{n} \), and \( H_j \) takes the maximum value.

**Step 5** The difference coefficient of the \( j \)-th dimension attribute:
For a given $j$, the smaller the $H_j$ and the larger the $q_j$, the more critical the attribute.

**Step 6** The weight of the $j$-th dimension attribute:

$$w_j = \frac{q_j}{\sum_{j=1}^{n} q_j}$$

In formula (8), $\sum_{j=1}^{n} w_j = 1, j = 1, 2, \ldots, m, 0 \leq w_j \leq 1$.

### 2.2. Determination of the category number $k$

Aiming at the defect that the number of categories of the $k$-means algorithm cannot be given in advance, the general solution is to use the contour map method to determine. Proposed by Peter J. Rousseeuw in 1986, different algorithms can be evaluated based on the same data[9]. However, the more considerable average contour value calculated by the contour map method does not necessarily mean that the average distance between the object and the same object is significant. It may also be the intermediate distance $b$ between the object and other objects and the average distance $a$ between the object and the same object. The latter is much larger. In the clustering of small samples, there may be individual abnormal data, which causes the value of $b$ to be larger than that of $a$. In this case, it may not be reasonable to rely on the contour map method to determine the value of $k$.

Unlike the contour map method, the elbow method determines $k$ based on the sum of squared errors of all objects. The core idea is that when $k$ is less than the actual number of categories, increasing $k$ will significantly reduce the square error; when $k$ reaches the exact number of types, the clustering quality will decline rapidly as $k$ increases, which will lead to the honest mistake of each object and the rate of decrease slows down. The image looks like an elbow, and the value corresponding to the elbow is the actual number of categories[10].

**Input:** $n$ data objects with $m$ features and a data set $D$ with cluster labels and the number of clusters $k$[11].

**Output:** the sum of squared errors corresponding to each $k$ value.

The specific steps of the elbow method are:

**Step 1** The Euclidean distance between the object in the cluster and the cluster centre: Calculate the Euclidean distance between the data object $a$ in the current cluster $C_i$ ($i=1,2,\ldots,k$) and the cluster centre $C_i$. Suppose two vectors $a_i(a_{i1},a_{i2},\cdots,a_{im})$ and $a_j(a_{j1},a_{j2},\cdots,a_{jm})$ represent two objects, and the Euclidean distance is defined as:

$$d(a_i, a_j) = \sqrt{\sum_{d=1}^{m}(a_{id} - a_{jd})^2}$$

**Step 2** Intra-cluster error sum of squares: Euclidean distance from each object in cluster $C_i$ to the cluster centre is squared and summed.

$$SSE_i = \sum_{a \in C_i} (a - C_i)^2$$

**Step 3** Calculate the sum of squares of the total error of the current classification sample according to the result in step 2: Determine $k$, and add the sum of squares of error corresponding to each cluster.

$$SSE = \sum_{i=1}^{k} |a - C_i|^2$$

**Step 4** Calculate the $SSE$ of the data set by incrementing $k$ value ($1,2,\ldots,k$).
Step 5 Determine the optimal $k$ value: store the $k$ different SSE calculated by the loop and their corresponding $k$ values in a two-dimensional array, and draw the two-dimensional coordinate points in a rectangular coordinate system by traversing the array.

![Figure 1. Line diagram of elbow method](image_url)

2.3. Identify high-quality cluster centers

The weighted Euclidean distance is:

$$d_w(a_i, a_j) = \sqrt{\sum_{d=1}^{m} w_d (a_{id} - a_{jd})^2}$$

In formula (9), $w$ is the weight of the $d$-th dimension attribute.

Taking the standard deviation as the standard measurement function, the weighted category objective value function is:

$$\sigma = \sqrt{\frac{\sum_{x \in T_j} d_w(a - c(T_j))}{|T_j| - 1}}$$

In formula (10): $\sigma$ is the weighted standard deviation of a certain type of information entropy; $|T_j|$ is the number of objects contained in $T_j$.

The initial cluster centre of $k$-means is the centre of each category. Based on this, select $k$ samples with different dense domains far from the initial centre. The process is: take the sample with the smallest variance as the initial centre of the first type and make a circle. The sample with the smallest variance outside the process is the initial centre of the next category in turn, and $k$ initial centres are found\cite{12}.

2.4. Improve the reselection method of cluster centers

K-means takes the mean of the points in the cluster as the new centroid. If there are abnormal points in the cluster, the standard will deviate seriously. To reduce the sensitivity to outliers, one object in each class is taken as a representative. Other items are divided into different categories according to the similarity with the representative point and iteratively repeated so that each representative point becomes the actual centre point of the corresponding class. Compared with the method of taking the average value of the objects in the cluster as the clustering centre, the advantage is that it is not sensitive to the noise point data, and the disadvantage is that it takes a long time and is suitable for clustering of small and medium samples. According to the similarity with the representative points, other objects are divided into different categories and iterated repeatedly to make each representative point the corresponding type's actual centre point, thereby completing the clustering.

Assuming that there are $n_i$ objects $x_1, x_2, \ldots, x_{n_i}$ in a cluster $T_j$ after initial clustering, the sum of the distances between the $i$-th object $x_i$ and other objects in the cluster are:
The new cluster center is defined as the object with the smallest distance from other items, thus completing the cluster center's determination.

2.5. Algorithm description

The steps of the improved algorithm are described as follows:

Input: Data set \( A \), \( A' \) with abnormal data added.

Output: \( k \) optimal clusters.

Step 1 Determine the number of categories \( k \) based on experience, and \( N \) is the number of data. Take \( k \) one by one in \( 2, \sqrt{N} \), and use the elbow method to determine the optimal \( k \) value.

Step 2 Data standardization, entropy weight method calculates the weight of object attributes.

Step 3 Determine \( k \) initial cluster centres. Find the sample \( x_1 \) with the smallest variance as the first initial centre \( C_1 \); take the average distance \( S \) of the data set samples as the radius, collect the sample \( x_2 \) with the smallest variance outside the circle, as the second initial centre \( C_2 \), and find \( k \) initial centres in turn.

Step 4 Scan all data and classify it into the most similar category based on the similarity with various sub-centres.

Step 5 Choose an actual object in each category as a representative, and divide other objects into different categories based on the similarity with the representative point, and repeatedly iterate to make each representative point the actual centre point of the corresponding category, thus completing the aggregation class.

Step 6 Until the number of iterations of the algorithm is completed, the loop terminates and repeats this process.

Step 7 Calculate the standard deviation of each category, and if there are non-numeric data, perform clustering again.

\[
s_i = \sum_{j=1}^{m} \sum_{k=1}^{n} (a_{ij} - a_{ik})^2
\]
Determine classification index

Start

Determine classification index

Obtain the original data of m attribute indicators and n samples

Data standardization

Calculate the information entropy of each indicator

Information entropy determines the index weight

Elbow method to determine the optimal k value

Determination of initial cluster centers by minimum variance radius method

Calculate the entropy weighted Euclidean distance from each data point to the initial cluster center

Cluster and calculate various types of means

Iteratively re-select cluster centers and re-cluster

Determine whether to terminate the operation according to the number of iterations

Determine whether to terminate the operation according to whether there is none-numeric data in the standard deviation

End and output the final classification result

Figure 2. Improved algorithm flow chart

3. Experiment and result analysis

To verify the algorithm's effect, the experimental platform is Intel(R)Core(TM)i7-9750H CPU@2.60GHZ, RAM 16G, Windows 7 operating system, and select Matlab R2019a as the programming tool. Test with a small sample data set of the UCI database.

3.1. The attribute weights of Iris data and the number of categories of the elbow method are shown in Figures 3 and 4.

Figure 3. Iris attribute entropy weight

Figure 4. The elbow method to determine the number of categories
3.2. The attribute weights of Wine data and the number of categories of the elbow method are shown in Figures 5 and 6.

![Figure 5: Wine attribute entropy weight](image1)

![Figure 6: The elbow method to determine the number of categories](image2)

We can see from Figure 3 and Figure 5 that each attribute has different functions. If we ignore the attribute function, the results will be different.

3.3. Clustering accuracy, which refers to the quotient of the same object as the predefined category and the total number, is one indicator for evaluating the quality of clustering[13]. Use the accuracy to assess the clustering effect of the Iris and Wine datasets. See Table 1 and Table 2.

### Table 1. Comparison of the results of the Iris data set

| category | instance | k-means | improved algorithm |
|----------|----------|---------|---------------------|
| 1        | 50       | 50      | 50                  |
| 2        | 50       | 40      | 50                  |
| 3        | 50       | 42      | 50                  |
| accuracy/% |         | 88      | 91.33               |

Note: k-means takes 0.013359s, and the improved algorithm takes 0.051343s

### Table 2. Comparison of the results of the Wine data set

| category | instance | k-means | improved algorithm |
|----------|----------|---------|---------------------|
| 1        | 59       | 59      | 59                  |
| 2        | 71       | 62      | 71                  |
| 3        | 48       | 48      | 48                  |
| accuracy/% |         | 94.94   | 85.39               |

Note: k-means takes 0.010043s, and the improved algorithm takes 0.063154s

After running the algorithm several times to calculate the average, the accuracy of Iris has increased from 88.00% to 91.33%. The range of entropy weight is 0.2405, and the content of entropy weight of Wine is 0.0860. It can be seen from Table 1 and Table 2 that the accuracy improvement effect of Iris is better. The increase in accuracy depends on the data set and attributes. The greater the difference in characteristics, the more significant the improvement.
3.4. After adding abnormal data in each of the Iris and Wine data sets, multiply the last row of each data set by two as the anomalous data, keeping the number of categories unchanged.

As shown in Table 3 and Table 4: Iris accuracy increased from 91.33% to 95.33%, Wine accuracy rose from 91.01% to 91.57%. The execution time is almost the same.

| Table 3. Comparison of the results of the Iris data set |
|------------------------------------------------------|
| category | instance | right | instance | right |
| 1        | 50       | 50    | 50       | 50    |
| 2        | 50       | 50    | 50       | 47    |
| 3        | 50       | 37    | 50       | 46    |
| accuracy/% | 91.33    | 95.33 |

Note: k-means takes 0.010068s, and the improved algorithm takes 0.057593s

| Table 4. Comparison of the results of the Wine data set |
|-------------------------------------------------------|
| category | instance | right | instance | right |
| 1        | 59       | 58    | 59       | 58    |
| 2        | 71       | 57    | 71       | 58    |
| 3        | 48       | 47    | 48       | 47    |
| accuracy/% | 91.01    | 91.57 |

Note: k-means takes 0.013933s, and the improved algorithm takes 0.074604s

Different types of data have a certain degree of randomness. From the test results, the improved method has better accuracy.

4. Conclusion

Clustering is more critical in practical tasks. Since k-means has no objective standard, it is the field where new algorithms appear the most in machine learning. This paper uses the elbow method to solve the problem of relying on experience to determine the number of categories. The seed center is selected in advance by the square difference radius method so that the selected initial center solves the disadvantage of randomly selecting the initial cluster center and can better reflect the classification decision surface. Subsequently, the reselecting cluster centers method was optimized for small sample data, which reduced the outliers' sensitivity. Simultaneously, the contribution of different indicators to the quality of clustering is considered. The entropy weight method highlights the representative role of essential indicators so that the clustering is more targeted, and the clustering is closer to the actual classification. Experiments have shown that for data with low dimensionality and large differences between categories, even if there is abnormal data and the execution time is almost unchanged, the improved algorithm can obtain more accurate and stable results.

Acknowledgments

This research was funded by National Natural Science Foundation of China (grantnumber61802425), National Social Science Foundation Military Science Project of China (grantnumber16GJ003-105).

References

[1] Zhang, L.J., Tan, L.Y., Liu, M.J., et al. (2020) Hands-on Data Analysis and Data Mining with Python. Mechanical Industry Press, Beijing, 125-135. (In Chinese)

[2] MacQueen, J.B. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, 281-297.

[3] Zhou, Y., Zhuo, J.W., & Bian, Y.Q. (2016) Big Data Mining: Systematic Method and Case Analysis. China Machinery Industry Press, Beijing, 187-218. (In Chinese)
[4] Qi, J., Yu, Y., Wang, L., Liu, J., & Wang, Y. (2017). An effective and efficient hierarchical k-means clustering algorithm. International Journal of Distributed Sensor Networks, 13(8(2017-8-01), 155014771772862.

[5] Islam, M. Z., Estivill-Castro, V., Rahman, M. A., & Bossomaier, T. (2018). Combining k-means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. Expert Systems with Applications, S0957417417306103.

[6] Debora, de, Chiusole, Luca, Stefanutti, & Andrea, et al. (2016). A class of k-modes algorithms for extracting knowledge structures from data. Behavior Research Methods.

[7] Ma, L., & Fan, S. (2017). Cure-smote algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinformatics, 18(1), 1-18.

[8] Phu, V. N., Tran, V. T. N., Dat, N. D., Chau, V. T. N., Nguyen, & Tuan, A. (2017). Sting algorithm used english sentiment classification in a parallel environment. International Journal of pattern recognition and artificial intelligence.

[9] Rousseeuw, P. J. (1987). Rousseeuw, p.j.: silhouettes: a graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. Journal of Computational and Applied Mathematics, 20, 53-65.

[10] Yang, Z.F. (2019) Research on Multipath Components Clustering for High Speed Railway Wireless Channels Using Machine Learning. Beijing Jiaotong University, Beijing. (In Chinese)

[11] Wang, J.R, Ma, X.,Duan, G.L. (2019) Improved K-means clustering k-value selection algorithm. Computer Engineering and Applications, 55(8): 27-33. (In Chinese)

[12] Xie, J.Y, Wang, Y.E. (2014) K-means Algorithm Based on Minimum Deviaton Initialized Clustering Centers[J]. Computer Engineering, 40(8), 205-211. (In Chinese)

[13] Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 63(2), 503-527.