A new clustering and sorting algorithm for radar emitter signals

MingWei Li\textsuperscript{a}, MingHao He, Jun Han, YuWen Tang

Air Force Early Warning Academy, WuHan, China
\textsuperscript{*}986556395@qq.com

Abstract—The traditional k-means clustering algorithm needs to set parameters manually in radar signal sorting application, which is sensitive to isolated points and prone to "batch" phenomenon, and the selection of initial clustering center has a direct impact on clustering effect. In order to solve the above problems, a radar signal sorting algorithm based on data field and improved k-means clustering is proposed. Firstly, the potential values of all data samples are calculated according to the theory of data field. After the isolated points are removed, the maximum local potential values are found. The nearest sample data is selected as the initial clustering center, and the maximum number of local potential values is selected as the clustering number. Finally, the improved k-means clustering algorithm is used to complete the radar signal sorting. The algorithm can automatically obtain the initial clustering center and the number of clusters, and the clustering results have larger inter cluster distance and smaller intra cluster distance, so the clustering results are better. Simulation results verify the feasibility of the algorithm.

1. INTRODUCTION

Sorting and recognition of radar emitter signal is one of the key technologies in radar countermeasure reconnaissance system, and also an important part of radar countermeasure information processing [1]. That sorting is to separate the signals of each radar from the dense overlapping signal pulse flow and select the useful signals [1]. The recognition of radar emitter is based on the correct signal sorting, so the sorting accuracy is very important for the recognition.

In modern electronic warfare, a large number of new system radar equipment are arrayed, resulting in the current electromagnetic environment, the signal is complex and changeable, and the signal flow density has reached millions to tens of millions of pulses per second. The traditional separation method based on PRI single parameter can not meet the needs. Clustering sorting algorithm has achieved good results in signal sorting and has a broad application prospect. As a classical unsupervised real-time clustering algorithm, K-means algorithm has the advantages of no prior information, simple and feasible, and fast convergence speed. However, it is sensitive to noise and needs to set the number of clusters manually. Therefore, it can not be directly applied to radar signal sorting.

Academician Li Deyi proposed the data field theory based on the concept of physical field [3]. Using the concept of potential in data field, clustering number and clustering center can be formed automatically, which can make up for the deficiency of K-means algorithm. In reference [4-8], data field theory has been applied to radar signal sorting successively. Literature [4] applies data field to multi-mode radar signal sorting, literature [5] uses gray correlation analysis to complete radar signal sorting.
sorting based on data field, literature [6] and combines data field and K-means algorithm to sort frequency agile radar. In reference [7], the data field is combined with k-means algorithm and FCM algorithm to sort non cooperative radar signals. In reference [8], an improved data field clustering algorithm is proposed. The above literature has a more in-depth study on the application of data field in radar signal sorting, but only literature [7] and literature [8] consider the noise signal pulse, which can not be ignored in the actual sorting. Literature [7] and literature [8] propose to remove isolated points according to the potential value of pulse data: literature [7] sets the threshold of pulse number until the number of removed pulses reaches the threshold. The time complexity of this method mainly depends on the number threshold set in advance, and the time complexity of the algorithm is high. In reference [8], the potential threshold is set, and the data whose potential value is less than the threshold is eliminated simultaneously. Although the method saves time, the potential threshold is related to the size of the data amount, and it is not universal for the data with different data amount.

To solve this problem, this paper normalizes the data potential value, sets the normalized potential value threshold to eliminate isolated points; then calculates the potential value of remaining data samples according to the data field theory, obtains the initial clustering center and the number of clusters; finally, uses the improved k-means clustering algorithm to complete the radar signal sorting. The method proposed in this paper can eliminate most of the noise signal pulses, and the normalized potential threshold does not change with the amount of data. The initial clustering center and the number of clusters are automatically obtained by using the data field, without any prior information, which has a high accuracy for radar signal sorting. The improved k-means clustering algorithm has larger distance between clusters and smaller distance within clusters, so the clustering result is better. Simulation results verify the feasibility of the algorithm.

2. BASIC PRINCIPLES OF DATA FIELD THEORY

In the real world, objects interact with each other. The field theory well reflects this point, for example, there is gravitation between all objects in the gravitational field theory, and there is electric field force between the electric charges in the electric field theory. The theory of data field holds that in the whole space of data field, each data is not isolated, but is influenced by the objective data field. For a specific data point, other data points will be affected by the field force of the data point, and the data point will also be affected by the field force of other data points. All data interact and influence each other to form a data field.

2.1 Field Strength Function

With reference to Newton's law of universal gravitation and Coulomb's law formula, the field strength function is used to describe and calculate the effect of single data on the whole space. The field strength at data points \( x \) to \( y \) is defined as follows:

\[
f_{x}(y) = \rho e^{-\frac{d(x, y)^2}{2\sigma^2}} \tag{1}
\]

In formula (1): \( \rho \) is the data amount of reaction data points, generally regarded as 1; \( d(x, y) \) is the Euclidean distance between \( x \) and \( y \); \( \sigma \) is the radiation factor, indicating the influence of data points, generally constant.

2.2 Potential Function

The potential function is defined as the algebraic sum of the field strength of all data in the data field at \( y \):

\[
F_{y} = \sum_{i=1}^{n} f_{y}(x_{i}) = \rho \sum_{i=1}^{n} e^{-\frac{d(x_{i}, y)^2}{2\sigma^2}} \tag{2}
\]

In formula (2), \( n \) is the number of data points. It can be seen from equation (2) that the potential value is inversely proportional to the distance between data points, that is to say, for a data point, the
more dense the position is, the greater the potential value is; the more sparse the position is, the smaller
the potential value is.

2.3 Radiation Factor
In order to analyze the influence of radiation factor on the potential value, we first consider the simplest
case: suppose there is only one data point in the data space. As shown in the figure 1, under different
values, the potential value (equal to the field strength value) changes with the distance.

![Figure 1: The potential value varies with distance](image)

It can be seen from the figure 1 that the smaller the value of $\sigma$ is, the faster the potential value
decays with the increase of distance. Extended to $n$ data points, when value of $\sigma$ is small enough, data
points only have influence on itself, which shows that each data is of its own type; on the contrary,
when value of $\sigma$ is large enough, the potential values of each data point are almost equal, and all data
can be regarded as a data set. Therefore, it is very important to choose a suitable value of $\sigma$ for data
field construction.

In reference [9], the value of $\sigma$ is determined according to the actual engineering experience: when
the number of data is less than 1000, $\sigma$ can take 0.1; when the data set is increased by an order of
magnitude, the value of $\sigma$ is halved. Although the value of $\sigma$ in this method is obtained from
engineering experience, it can not get the optimal value. In reference [10], according to the concept of
entropy in information theory, potential entropy is introduced to describe the influence of value on data
field. In this way, the value of $\alpha$ can be transformed into the problem of finding the minimum value of
potential entropy, avoiding artificial selection. The specific calculation of potential entropy can refer to
[10], as shown in the figure 2 is the schematic diagram of potential entropy changing with radiation
factor.

![Figure 2: The potential entropy varies with $\sigma$](image)
2.4 Isolated Points Removal

Analysis formula (2): for a data point, the more dense the location is, the greater the potential value is; the more sparse the location is, the smaller the potential value is. Therefore, the point with small potential value should be isolated point. According to literature [8], the potential value of isolated point should be larger than and very close to 1. Through simulation, it is found that for most of the noise signal pulses, the potential value is very close to 1; because of the randomness, a few of the pulses are inevitably in the dense place, and these pulses are regarded as normal pulse processing; the remaining pulse potential value is between the two, and also close to 1. In order to improve the sorting accuracy, this part of the noise pulse should also be regarded as isolated point removal. But for different data sets, the degree of approaching 1 is different. The threshold of potential value needs to be set according to different data amount, which is not universal.

Therefore, in this paper, the data potential value is normalized. The normalized potential value is between [0,1]. For data of different data amounts, the potential value of isolated points is greater than and close to 0, and the approach degree is similar. In this way, the threshold of normalized potential value can be set to a fixed value according to the noise environment, which is universal for different data amount.

2.5 Determine The Initial Cluster Center

The line connecting points with equal potential values in the data field is called equipotential line, and the center of the equipotential line surrounding different areas is called potential center. The potential center is described as:

\[ F_{\text{max}} \geq F_y(i, j) \]  

(3)

In the formula (3): \( F_{\text{max}} \) is the potential value of the potential center; \( F_y(i, j) \) is the potential value around the potential center; \((i, j)\) is the position of the data point.

The number of potential centers determines the number of clusters that the data can be divided into, but the potential center does not necessarily coincide with the data points, so the nearest data point is taken as the initial cluster center. The mathematical formula is as follows:

\[ D_{\text{min}} = d[F_{\text{max}}, \text{data}(i, j)] \]  

(4)

\[ F_{\text{center}} = \text{data}(i, j) \]  

(5)

In the formula (4) (5); \( d(\cdot, \cdot) \) is the Euclidean distance between two points; \( \text{data}(i, j) \) is any data in the data set; \( F_{\text{center}} \) is the initial clustering center.

3. IMPROVED K-MEANS ALGORITHM

3.1 K-means Algorithm

K-means algorithm is a widely used unsupervised clustering algorithm. Assuming that the unlabeled data set is \( X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\} \), the task of the algorithm is to cluster the data set into \( k \) clusters: \( C = C_1, C_2, \ldots, C_k \). The objective function of the algorithm is:

\[ E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \]  

(6)

In the formula(6), \( A \) is the midpoint of cluster \( C \):

\[ \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \]  

(7)

The specific steps of the algorithm are as follows:

1. Randomly select sample points in the sample data set as the center of each cluster;
2. Calculate the Euclidean distance from all sample data points to the center of each cluster, and then divide the sample data points into the nearest cluster;
3. According to the sample data points in the cluster, recalculate the cluster center with formula (7);
4. Repeat step 2 and step 3 until the objective formula (6) converges.

3.2 Improved K-means Algorithm
The ideal clustering result is that the smaller the distance between clusters is, the better the distance between clusters is. The traditional K-means algorithm only considers the similarity within the class. In this paper, the objective function is improved. The improved objective function will fully consider the differences between classes. The improved objective function is:

\[
E' = \sum_{m \neq n, m = 1}^{k} \sum_{n=1}^{k} \frac{E}{d(\mu_m, \mu_n)}
\]

In the formula (8): \(E\) is the objective function of traditional K-means algorithm; \(d(\mu_m, \mu_n)\) is the Euclidean distance from the center of cluster \(m\) to the center of cluster \(n\). The improved objective function is determined by the distance between clusters and the distance within clusters, and the clustering effect is better.

4. A NEW CLUSTERING AND SORTING ALGORITHM FOR RADAR_EMITTER SIGNALS
At present, pulse description word (PDW) and pulse modulation parameters are mainly used to sort unknown radar emitter. In this paper, PDW is mainly used to sort the radar without considering the intra pulse modulation parameters. In PDW, pulse repetition frequency (PRI) can be obtained according to the time of arrival (TOA). PRI works in many ways and changes rapidly, which is not generally used as the basis for sorting. Pulse amplitude (PA) changes with the scanning of antenna, which is not generally used as the basis for sorting. Therefore, three parameters of the DOA (carrier frequency) and pulse width (RF) are used to classify the data sets. The specific steps of the algorithm are as follows:

1. Data preprocessing: assume that data set is \(P = [PDW_1, PDW_2, \ldots, PDW_n]^{T}\), \(n\) is the number of data set pulses. The data set parameters are normalized, and the normalized pulse description word is \(PDW' = [RF', PW', DOA']\). According to formula (2), the data field is built with normalized data, and the potential value of any point \(P\) in the data field is \(F_p\). With the method proposed in this paper, the potential value of point \(P\) is \(F_p'\) after the potential value is normalized. The fixed threshold is set to \(T\), and the data satisfying condition \(F_p' \leq T\) is regarded as outlier elimination;
2. According to formula (2), calculate the potential value of each data point after eliminating the isolated points, and then use formula (3) (4) (5) to find out the nearest data point of the maximum value point in the data field as the initial clustering center, and the number of the maximum value as the initial clustering number;
3. Calculate the Euclidean distance from all data points to the initial clustering center, and then divide the sample data points into the nearest clusters;
4. According to the data points in the cluster, recalculate the cluster center with equation (7);
5. Repeat 3 and 4 until equation (8) converges.

5. SIMULATION ANALYSIS
In order to verify the effectiveness of this method, Matlab is used for simulation test. Assume that there are 4 parameters partially overlapped frequency agile radar, and the parameters are shown in Table 1.
### TABLE 1 RADAR PARAMETERS

| Radar label | RF / MHZ | PW / µs | DOA / ° | RF / µs |
|-------------|----------|---------|---------|---------|
| 1           | 3165-3410 (agile) | 8.36-12.68 | 55-58 | 175-180 (jitter) |
| 2           | 2950-3175 (agile) | 10.75-16.24 | 57-60 | 140-185 (jitter) |
| 3           | 3450-3710 (agile) | 16.63-20.42 | 59-62 | 189-241 (jitter) |
| 4           | 2945-3195 (agile) | 19.30-23.66 | 61-64 | 152-181 (jitter) |

5.1 Data Preprocessing

In order to compare the effect of normalized potential threshold and potential threshold in reference [8] on noise elimination. Set 10 data sets with different pulse numbers. The number of radar pulses in each data set is shown in Table 2. The number of noise pulses is 10% of the radar pulses. The parameter value is the uniform random number within the range of each radar parameter value.

### TABLE 2 NUMBER OF RADAR PULSES IN DATA SET1-10

| Radar label | Number of pulses |
|-------------|------------------|
|             | Data set 1 | Data set 2 | Data set 3 | Data set 4 | Data set 5 | Data set 6 | Data set 7 | Data set 8 | Data set 9 | Data set 10 |
| 1           | 27         | 54         | 81         | 108        | 135        | 162        | 189        | 216        | 243        | 270         |
| 2           | 35         | 70         | 105        | 140        | 175        | 210        | 245        | 280        | 315        | 350         |
| 3           | 26         | 52         | 78         | 104        | 130        | 156        | 182        | 208        | 234        | 260         |
| 4           | 28         | 56         | 84         | 112        | 140        | 168        | 196        | 224        | 252        | 280         |

According to the method in this paper, the above data sets are normalized respectively. The value of normalized threshold needs to be set according to the specific noise environment. In this simulation, the threshold equal to 0.15; according to the method in reference [8], the value of potential threshold is a number close to 1, and the value of this simulation is 2. The above two methods were repeated in data set 1-10, the number of experiments was 100, and the average noise rejection rate of the experimental results was 100. Figure 3 shows the situation of eliminating isolated points, and Figure 4 (a) (b) shows the potential values of noise pulses in data set 1 and data set 10 in an experiment.

![Figure 3 Noise pulse rejection ratio](image-url)
It can be seen from Figure 3 that the method in reference [8] gradually reduces the noise rejection rate with the increase of the number of pulses in the data set, so it is necessary to set different potential thresholds for different data sets to achieve good rejection effect; the normalization threshold proposed in this paper has good noise rejection rate for 10 data sets, which has certain universality. According to the analysis of Fig. 4, for data sets 1 and 10, the potential value of most noise signal pulses is very close to 1; because of randomness, a small number of noise pulses are inevitably in a dense place, and these pulses are regarded as normal pulse processing; the potential value of the remaining noise pulses is between the two, also close to 1. In order to improve the sorting accuracy, this part of noise pulses should also be regarded as isolated point removal. In different data sets, the degree of noise pulse approaching 1 is different. In data set 1, the noise pulse potential is between 1-2; in data set 10, it is between 4 to 15. Therefore, if the potential threshold value is 2, the effect of data set 1 is better, and the effect of data set 10 is worse, and the rejection percentage is only 52.2%

5.2 Algorithm Comparison
In order to compare the sorting effect of traditional K-means algorithm and the algorithm in this paper. Two algorithms are used to process data set 10.

Firstly, the outliers of data set 10 (the proportion of noise pulse is 10%) are eliminated by this method, and the data distribution before and after the isolated points are removed is shown in Figure 5.

After data set 10 is preprocessed, according to data field theory, all potential values of data samples are calculated by potential function to describe the distribution of radar signal parameters. The radiation
factor $\sigma$ is 0.068 according to potential entropy theory. As shown in Figure 6, it is the equipotential distribution of the constructed data field.

![Figure 6](image1.png)

**Figure 6** Equipotential distribution

From Figure 6, it can be seen that the data can be divided into four categories. Use equation (3) (4) (5) to solve the data point closest to the maximum value point in the data field as the initial clustering center.

| Radar label | RF     | PW     | DOA    |
|-------------|--------|--------|--------|
| 1           | 0.7614 | 0.1306 | 0.1815 |
| 2           | 0.2912 | 0.2514 | 0.3168 |
| 3           | 0.6089 | 0.6466 | 0.6287 |
| 4           | 0.1724 | 0.8791 | 0.8485 |

**TABLE 3 INITIAL CLUSTERING CENTER**

It should be pointed out that the traditional K-means algorithm can not remove the isolated points and complete the cluster number discrimination correctly. In order to achieve the comparative effect, the data set is preprocessed to remove the isolated points, and then the number of known clusters is set to 4. The initial clustering centers of traditional K-means algorithm are randomly generated, and the experiments are repeated 100 times independently by traditional K-means algorithm. The experimental results of the two algorithms are shown in Table 4. The sorting accuracy in the table is the average sorting accuracy of four radar radiation sources.

|               | Algorithm in this paper | Traditional K-means algorithm |
|---------------|-------------------------|-------------------------------|
| sorting accuracy | 97.1%                  | 88.3%                         |
| iterations    | 4                       | 14.2                          |

**TABLE 4 ALGORITHM COMPARISON**
From the experimental results, we can see that under the same data set and the same noise ratio, the sorting accuracy of this algorithm is higher than the traditional K-means algorithm, and the average number of iterations is lower than the traditional K-means algorithm. This is because the initial cluster center given based on the data field will be close to the final cluster center, while the selection of the initial cluster center of the traditional K-means algorithm is random, resulting in low sorting accuracy and high number of iterations.

5.3 Stability Analysis of the Algorithm in This Paper
In practice, the data set size and noise ratio are random. In order to compare the sorting effect of this algorithm and the traditional K-means algorithm under different data sets and different noise ratios. A total of 100 data sets are set, and the total number of pulses in each data set is randomly set at [116-1160]. The pulse data of each radar radiation source in the data set is proportionate according to their PRI size, and the noise pulse proportion is randomly set at 5% - 15%. Two algorithms are used to repeat experiments on each data set, and the number of experiments is 100.

The average distance between the final clustering centers is used to measure the degree of dispersion between clusters, and the sum of squares of errors (SSE) is used to measure the degree of aggregation within clusters. The larger the average distance between the final clustering centers, the higher the degree of separation between clusters, the higher the degree of aggregation within clusters, and the better the clustering effect. The calculation results are shown in Table 5.

| TABLE5 ALGORITHM COMPARISON |
|------------------------------|
| Sorting accuracy | Iterations | Centroid distance | SSE   |
|-------------------|------------|-------------------|-------|
| Algorithm in this paper | 11.2 | 89.7% | 1.0078 | 546.3 |
| Traditional K-means algorithm | 3.4 | 97.4% | 1.1053 | 171.4 |

From the experimental results, it can be seen that, in the case of random data set size and noise ratio, the algorithm in this paper reduces the calculation amount and improves the sorting accuracy compared with the traditional K-means clustering algorithm. In this algorithm, the average distance of the final clustering center is larger than the traditional K-means algorithm, and the sum of error squares SSE is smaller than the traditional K-means algorithm, so the clustering effect is better.

6. CONCLUSION
This paper presents an improved k-means clustering algorithm based on data field. The algorithm first solves the initial clustering center according to the data field theory, and then uses the improved k-means clustering algorithm to complete the radar signal sorting. In this algorithm, most of the isolated points can be removed by setting the normalized threshold of potential value, and the radar signals with partially overlapped parameters can be sorted with high accuracy and good clustering effect. The simulation results show that the algorithm is effective. However, the loss of pulse is not considered in this paper, which needs further study.

REFERENCES
[1] He Minghao, modern radar emitter signal sorting and recognition [M]. Science Press, Beijing, 2016
[2] Yang junhuang, Zhao Chao. Review of K-means clustering algorithm [J]. Computer engineering and application, 2019,55 (23): 7-14 + 63
[3] Meng Haidong, sun Xinjun, song Yuchen. Improved lof algorithm based on data field [J]. Computer engineering and application, 2019,55 (03): 154-158
[4] Guoqiang, Song Wenming, nanpulong, wanjian. Multi mode radar signal sorting algorithm based on data field and cloud model [J]. Journal of Harbin Institute of Technology, 2015, 47 (11): 76-81

[5] Zhang Ran, Xia houpei. A new K-means clustering radar signal sorting algorithm [J]. Modern defense technology, 2015, 43 (06): 136-141

[6] Jia ran, Hu Jin. A radar signal sorting algorithm based on data field clustering [J]. Modern defense technology, 2017, 45 (04): 124-129

[7] Feng Xin. Research on clustering and sorting algorithm of non cooperative radar emitter signal [D]. Xi'an Branch, the Fifth Research Institute of China Aerospace Science and Technology Corporation, 2018

[8] Gao Lipeng, Sha Zuojin. An improved data field clustering algorithm [J]. Applied science and technology, 2019, 46 (06): 30-34

[9] Feng Xin, Hu Xiaoxi, Kuang Yin. Multi mode radar signal sorting algorithm based on data field [J]. Electronic design engineering, 2018, 26 (23): 139-142

[10] Li Xue, Miao daiqian, Feng Qinrong. Rough clustering algorithm based on data field [J]. Computer science, 2009, 36 (02): 203-206 + 244