Underdetermined Mixed Matrix Estimation Based on DPC-KFCM Algorithm

Cao Wenrui, Xu Yan*, Nan Zhefeng, Wei Yiming

*School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

*511870887@qq.com

Abstract. Aiming at the problem of fuzzy C-means (FCM) in the estimation of underdetermined mixing matrix, that the estimation accuracy is not high and the robustness is poor, a density peak clustering (DPC) based on density peak clustering (DPC) is proposed. Improved Kernel-based Fuzzy C-means (KFCM). The kernel function is introduced into the FCM algorithm to construct the KFCM algorithm based on the Gaussian kernel function, which can effectively overcome the influence of noise points and isolated points on the clustering results and improve the estimation accuracy of the mixed matrix; the traditional DPC algorithm is improved and merged with the KFCM algorithm, and thresholds are set for the local density and high-density distance to achieve the initial clustering center of the KFCM algorithm and the automatic determination of the number of cluster centers improves the robustness of the algorithm. Experimental results show that the algorithm has greatly improved the estimation accuracy and robustness of the underdetermined mixed matrix.

Keywords: fuzzy c-means clustering, density peak clustering, fuzzy c-means clustering, underdetermined mixed matrix estimation.

1. Introduction

Underdetermined blind source separation is a process of restoring the original signal when the number of observation signals is less than the number of source signals and the observation signals are known but the source signal and channel are unknown [1]. This technology is of great significance in many fields such as artificial intelligence, image processing, and speech recognition [2]. In recent years, the effective mainstream method of underdetermined blind source separation is to estimate the mixing matrix first and then reconstruct the source signal according to the result. It can be seen that the accuracy value of the mixing matrix estimation is the key factor that restricts the effect of source signal reconstruction [3].

A large number of experiments show that the clustering algorithm represented by FCM has achieved good results in the underdetermined mixing matrix estimation results. The FCM algorithm is widely used because of its fuzzy, unsupervised, and simple implementation characteristics[4]. However, the initial clustering center of the current FCM algorithm is randomly selected by the program, and its number needs to be determined manually, resulting in a greatly reduced clustering accuracy, making the overall robust performance of the algorithm poor, and the existence of noise...
points and isolated points makes the algorithm difficult. The local search ability is reduced, which is easy to cause unsatisfactory results [5].

Faced with many of the above shortcomings, Chang Xue et al. optimized the bat algorithm and integrated the FCM clustering algorithm to make the initial clustering center more clear [6]. Fu Jiaqi introduced the concept of local density into FCM to solve the problems that the initial cluster centers are easy to fall into the local optimum and the convergence is slow, which has achieved good results [7]. Li Hu integrated the simulated annealing algorithm and genetic algorithm into the FCM algorithm to determine the clustering center and improved the global and spatial search capabilities of the traditional FCM algorithm, but it caused the problem of slowing down the convergence speed [8]. Bi Xiaojun and others integrated the artificial bee colony theory into K-means [9] to enhance the robustness of the overall result. Hou Sizhu et al. Used the density threshold method to determine the clustering center, combined with the distance feature, can complete effective clustering for any scale data set [10]. Liu Cangsheng et al. can achieve the purpose of automatically initializing the cluster center [11] by integrating the idea of density peak into the classic FCM algorithm.

This paper proposes an improved KFCM algorithm based on the DPC algorithm. First, the KFCM algorithm based on the Gaussian kernel function is constructed by introducing the kernel function to overcome the problem that the traditional FCM estimation results are easily affected by noise points and isolated points, and the local search ability is not strong. Use the improved DPC algorithm to further optimize the KFCM, improve the KFCM algorithm's high sensitivity to the initial clustering center and the number of defects that need to be manually given, and prove its effectiveness.

2. Estimation of mixing matrix based on DPC improved KFCM algorithm

2.1. FCM algorithm based on kernel function

In FCM, a kernel function can be introduced to distinguish linearly inseparable points containing noise points and isolated points. The specific approach is as follows:

Construct the objective function:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij}^m \| \varphi(x_j) - \varphi(v_j) \|_1, 1 \leq m \leq \infty$$

Among them, $U$ is the membership matrix, and the set form can be expressed as $U = \{u_{ij} \}_{i=1,j=1}^{n,m}$. The constraints are $\sum_{i=1}^{m} u_{i} = 1, \forall j = 1, 2, ..., n$; $m$ is the fuzzy weighting coefficient; $\varphi(x_j)$ is the eigenvalue corresponding to the point $x_j$ in the high-dimensional space. $\varphi(v_j)$ is the eigenvalue of the cluster center $v_j$. $\| \varphi(x_j) - \varphi(v_j) \|$ is the interval measurement between data points $x_j$ and cluster centers $v_j$, and the specific form is:

$$\| \varphi(x_j) - \varphi(v_j) \| = K(x_j,x_j) + K(v_j,v_j) - 2K(x_j,v_j)$$

Due to $K(x_j,v_j) = e^{-\frac{d^2}{2\sigma^2}}, \sigma > 0$. That is $K(x_j,x_j) = 1, K(v_j,v_j) = 1$ the objective function can be rewritten as:

$$J(x,v) = 2 \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij}^m (1 - K(x_j,v_j)), 1 \leq m \leq \infty$$

The Lagrangian method is used to realize the continuous iterative update of the membership matrix and clustering centers, as shown in the following formula:
2.2. Improved DPC algorithm

Improve the local density measurement method: When using different local density calculation methods for an unknown data set, different clustering results will be produced. To ensure more accurate clustering results, this paper uses the idea of K nearest neighbors to replace the idea of using cutoff distance in the original formula. This method provides a unified way for the local density measurement of data sets of any size and also makes the algorithm easier to distinguish cluster centers.

\[
u_j = \frac{1}{\sum_{i=1}^{n} (1 - K(x_j, v_i))^{\alpha-1}} = \sum_{j=1}^{n} u_j^m K(x_j, v_j) x_j
\]

The formula is shown in the following formula:

Among them, \( n \) is the number of data points; \( d^k_i \) is the distance measurement between the point \( i \) and its \( k \) nearest neighbor; \( u^k \) is the average of all \( d^k_i \).

Improve the method of selecting cluster centers: Because the cluster centers in traditional algorithms need to be selected independently by humans, there is a problem of strong human subjectivity. This paper sets thresholds for local density and high-density distance respectively if the sum of a certain point is greater than the threshold. When the point is initially identified as the cluster center, the initial cluster center and number can be automatically determined. The expression is shown in the following formula:

\[
\delta_i > \alpha = u^k + \frac{1}{n-1} \sum_{j=1}^{n} (d^k_i - u^k)^2 \quad \rho_i > \beta = \mu(\rho)
\]

Among them, \( \alpha \) is the threshold of high-density distance \( \delta \); \( \beta \) represents the threshold of local density \( \rho \), and \( \mu(\rho) \) represents the average of all data points.

2.3. Improved KFCM algorithm based on DPC algorithm

In order to overcome the problem that the FCM algorithm is sensitive to noise points and isolated points, and the local search ability is not strong, the algorithm introduces the kernel function into the original FCM algorithm to construct the KFCM algorithm based on the Gaussian kernel function.

In view of the high sensitivity of the KFCM algorithm to the initial cluster centers and the need to manually give the number of clusters, the algorithm incorporates the DPC algorithm into the KFCM algorithm. By using the idea of K-nearest neighbors to improve the local density measurement method, avoid the need for different calculation methods for data sets of different sizes, and set thresholds for the local density and high-density distance respectively to achieve the initial clustering center of the unknown-scale data set and the number is automatically determined.

The algorithm steps are as follows:
Step1: Sample point pre-processing;
Step2: Obtain each initial cluster center and number by improving the DPC method;
Step3: Apply the obtained initial value to KFCM to determine the cluster center and obtain the estimation result of the mixing matrix.

3. Experimental simulation and analysis
3.1. Performance

(1) Normalized mean square error (NMSE) \( \text{NMSE} = 10 \log_{10} \left( \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (w_{ij} - \hat{w}_{ij})^2}{\sum_{i=1}^{M} \sum_{j=1}^{N} w_{ij}^2} \right) \)

In the formula, \( M \) and \( N \) represent the number of rows and columns of the mixed matrix; \( w_{ij} \) and \( \hat{w}_{ij} \) are the row \( i \) and column \( j \) elements of the mixed matrix \( W \) and the estimated matrix \( \hat{W} \) respectively. The smaller the value, the higher the accuracy.

(2) Deviation angle: 

\[
\text{ang}(w_i, \hat{w}_i) = \frac{180}{\pi} \arccos \left( \frac{\langle w_i, \hat{w}_i \rangle}{\|w_i\| \|\hat{w}_i\|} \right),
\]

Among them, \( w_i \) and \( \hat{w}_i \) are the rows \( i \) of \( W \) and \( \hat{W} \) respectively. The smaller the value, the higher the accuracy.

3.2. Experimental simulation and analysis

Randomly select four channels of speech and compose them into two channels of observation signals, where the mixing matrix is:

\[
W = \begin{bmatrix}
-0.3648 & 0.5876 & 0.4322 & 0.3222 \\
0.9656 & -0.6532 & 0.8343 & -0.6438
\end{bmatrix}
\]

The following are the matrix estimation results obtained by each algorithm through experiments. The matrix estimation result obtained by the GASAFCM algorithm is:

\[
\hat{W}_{\text{GASA-FCM}} = \begin{bmatrix}
-0.3628 & 0.5811 & 0.4310 & 0.3208 \\
0.9634 & -0.6471 & 0.8315 & -0.6403
\end{bmatrix}
\]

The matrix estimation result obtained by ABC-K-means is:

\[
\hat{W}_{\text{ABC-K-means}} = \begin{bmatrix}
-0.3542 & 0.5782 & 0.4253 & 0.3153 \\
0.9643 & -0.6471 & 0.8265 & -0.6302
\end{bmatrix}
\]

The result of the matrix estimation obtained by the DBSCAN algorithm is:

\[
\hat{W}_{\text{DBSCAN}} = \begin{bmatrix}
-0.3582 & 0.5843 & 0.4248 & 0.3172 \\
0.9584 & -0.6503 & 0.8243 & -0.6345
\end{bmatrix}
\]

The matrix estimation result obtained by the FDP-FCM algorithm is:

\[
\hat{W}_{\text{FDP-FCM}} = \begin{bmatrix}
-0.3611 & 0.5827 & 0.4276 & 0.3191 \\
0.9628 & -0.6487 & 0.8298 & -0.6379
\end{bmatrix}
\]

The matrix estimation result obtained by the proposed algorithm is:

\[
\hat{W} = \begin{bmatrix}
-0.3636 & 0.5868 & 0.4298 & 0.3216 \\
0.9642 & -0.6511 & 0.8307 & -0.6418
\end{bmatrix}
\]

The accuracy evaluation of the estimation results of the above algorithms is summarized as follows. estimate \( \hat{W} \)

| Comparison algorithm | Angle between column vectors / (°) | NMSE/dB |
|----------------------|-----------------------------------|---------|
|                      | \( \text{ang}(w_1, \hat{w}_1) \) | \( \text{ang}(w_2, \hat{w}_2) \) | \( \text{ang}(w_3, \hat{w}_3) \) | \( \text{ang}(w_4, \hat{w}_4) \) |
| GASA-FCM             | 0.5739                            | 1.6278  | 1.0271  | 0.8676  | -35.4956 |
| ABC-K-means          | 3.7361                            | 1.8412  | 2.1653  | 2.9316  | -24.5276 |
| DBSCAN               | 2.4573                            | 0.9435  | 2.4292  | 2.3675  | -27.1365 |
| FDP-FCM              | 1.2637                            | 1.3654  | 1.7236  | 1.6297  | -32.4682 |
| Algorithm             | 0.3863                            | 0.7952  | 1.5638  | 0.1528  | -43.4256 |

**Tab.1** Performance evaluation Tab of normalization error and deviation angle of each algorithm
4. Conclusion
In the clustering process of the FCM algorithm, the initial clustering center is randomly given by the program and its number needs to be set subjectively; the existence of noise points and isolated points weakens the local search ability and the robustness. The algorithm uses the idea of a kernel function to turn nonlinear clustering into linear clustering can effectively suppress the influence of isolated points and noise points on the clustering results; using a unified local density measurement method can achieve accurate clustering of data sets of unknown size, and set thresholds for local density and high-density distance respectively, to realize the automatic determination of the initial cluster center and its number. Experiments show that compared with the comparison algorithm in the article, the proposed method has a greater improvement in the accuracy and robustness of matrix estimation, which proves its effectiveness.

References
[1] Liy B, Nie W and Ye F 2016 A mixing matrix estimation algorithm for underdetermined blind source separation J. Circuits, Systems & Signal Processing 35 pp 3367-3379
[2] Xiang W, Zhitao H and Yiyu Z 2014 Underdetermined DOA estimation and blind separation of non-disjoint sources in time-frequency domain based on sparse representation method J. Journal of Systems Engineering and Electronics 25(01) pp 17-25
[3] Li Y, Wang Y and Dong Q 2020 A novel mixing matrix estimation algorithm in instantaneous underdetermined blind source separation J. Signal, Image and Video Processing 11 pp 1001-1008
[4] Yu XC, He H and Hu D 2014 Cover classification of remote sensing imagery based on interval-valued data fuzzy c-means algorithm J. Science China (Earth Sciences) 57(06) pp 1306-1313
[5] Xinyong Y, Ying G and Kunfeng Z 2017 A network sorting algorithm based on blind source separation of Multi-FH signal J. Journal of Signal Processing 33(08) pp 1082-1089
[6] Xue C and Hongyan S 2020 FCM clustering algorithm based on improved bat algorithm optimization J. Computer and Modernization 05 pp 29-33+38
[7] Jiaqi F 2018 Research on intrusion detection technology based on improved fuzzy C-means clustering algorithm D. Lanzhou: Lanzhou University pp 1-39
[8] Hu L 2018 Research on Blind Source Separation Algorithm of Speech Signal Based on Sparse Representation D. Lanzhou: Lanzhou Jiaotong University pp 16-32
[9] Xiaojun B and Rujiang G 2012 Underdetermined blind mixing matrix estimation algorithm based on mixing clustering and mesh density J. Systems Engineering and Electronics 34(3) pp 614-618
[10] Sizu H, Siyu H and Lizhao H 2012 Improved DBSCAN algorithm suiTab for multi-density J. Sensors and Microsystems 34(3) pp 614-618
[11] Cangsheng L and Qinglin X 2018 Fuzzy C-means clustering algorithm based on density peak value optimization J. Computer Engineering and Application 54(14) pp 153-157