Blind Source Signal Separation Method of Compressed Sensing Sparse Component Analysis

Guo Xukun[1]

[1]*Modern Education Technology Center, Guangzhou Sport University,
Guangzhou Guangdong, 510500, China
guoshaoqun@126.com*

Abstract

Two-step method is the common method for solving the undetermined blind separation of the sparse signals, wherein K-means clustering algorithm is usually firstly adopted to evaluate the aliasing matrix and then the shortest path method is adopted to recover the source signal. The K-means clustering algorithm requests to know the number of the source signals, which is usually not known in practice and should be evaluated. Therefore, the clustering effectiveness evaluation index ---- BWP index is researched in this paper, and meanwhile the particle swarm optimization algorithm is also introduced therein to propose an improved algorithm for determining the number of the source signals, and this algorithm is introduced therein for the undetermined blind separation. The experiment shows that the proposed algorithm can not only ensure the separation accuracy and shorten the separation time, but also save certain memory and have more obvious advantages when there are many observed signals.

Keywords: Particle swarm optimization algorithm; Shortest path method; Undetermined analysis; Blind Signal separation

1. Introduction

In actual signal environment, due to the unknown number of the potential source signals and the limited number of the receiving array elements, number n of the source signals among the received signals is usually more than number m of the array elements, and such blind separation of the composite signals is called as undetermined blind source separation (UBSS) [1]. At present, the undetermined blind source separation is an international difficult problem, but many signals have sparseness or can become sparse through suitable mathematical transformations such as Fourier transformation and wavelet transformation, so the sparse representation [2] of the signals can be adopted to discuss the difficult problems in the blind separation. In this paper, the particle swarm optimization algorithm is adopted to improve Kmeans traversal algorithm, and the improved algorithm is adopted to determine the number of the source signals for the undetermined blind separation. Finally, the undetermined blind separation experiment is carried out for 6 paths of source signals and 2 paths of observed signals in this paper, and the proposed algorithm is compared with other algorithms in the aspects of separation accuracy, separation time and mean memory.

2. K-means Cluster Number Determination Algorithm based on BWP Index

2.1. BWP Index Meaning

The blind source separation problem is adopted to recover the source signals only according to the composite signals observed thereby under the condition of unknown

hybrid parameter. The mathematical model of the blind separation can be represented as follows:

$$X(t) = AS(t) + N(t), t = 1, \cdots, T \tag{1}$$

Where $X(t) = [x_1(t), x_2(t), \text{L}, x_m(t)]^T$ represents the observed signal vector, A represents the aliasing matrix, $S(t) = [s_1(t), s_2(t), \text{L}, s_n(t)]^T$ represents the source signal vector and $N(t) = [n_1(t), n_2(t), \text{L} \ n_m(t)]^T$ represents the noise. The blind separation aims at recovering the waveform of the source signal only through observed signal vector $X(t)$, so the source signal recovered thereby has sequence and amplitude uncertainty. Usually, the noise influence is not considered in the blind separation, so the above model can be also represented as follows:

$$X(t) = AS(\text{t}), \quad t = 1, 2, \text{L}, T \tag{2}$$

Definition 1: $K = \{X, R\}$ is assumed as the cluster space ($X = \{x_1, x_2, \cdots, x_n\}$), n sample objects are clustered as c clusters, and minimum between-cluster distance $b(j, i)$ of the *ith* sample of the *jth* cluster is defined as the minimum value of the mean distance of this sample to other samples in each other cluster, namely:

$$b(j, i) = \min_{1 \square k \square c, k \neq j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} \left\| x_p^{(k)} - x_i^{(j)} \right\|^2 \right) \tag{3}$$

Where k and j represent the cluster labels, $x_i^{(j)}$ represents the *ith* sample of the *jth* cluster, $x_p^{(k)}$ represents the *pth* sample of the *kth* cluster, n_k represents the number of the samples of the *kth* cluster, and $\| \text{g} \|^2$ represents square Euclidean distance.

Definition 2: between-cluster distance $w(j, i)$ of the *ith* sample of the *jth* cluster is defined as the mean distance of this sample to other samples of the *jth* cluster, namely:

$$w(j, i) = \frac{1}{n_j - 1} \sum_{q=1, q \neq i}^{n_j} \left\| x_q^{(j)} - x_i^{(j)} \right\|^2 \tag{4}$$

Where $x_q^{(j)}$ represents the *qth* sample of the *jth* cluster ($q \neq i$), and n_j represents the number of the samples of the *jth* cluster.

Definition 3: cluster distance $baw(j, i)$ of the *ith* sample of the *jth* cluster is defined as the sum of the minimum between-cluster distance and the between-cluster distance of this sample, namely:

$$baw(j, i) = b(j, i) + w(j, i) \tag{5}$$

Definition 4: cluster separation distance $bsw(j, i)$ of the *ith* sample of the *jth* cluster is defined as the difference between the minimum between-cluster distance and the between-cluster distance of this sample, namely:

$$bsw(j, i) = b(j, i) - w(j, i) \tag{6}$$

Definition 5: between-within proportion (BWP) $BWP(j, i)$ of the *ith* sample of the *jth* cluster is defined as the specific value of the cluster difference distance and the cluster distance of this sample, namely:

$$BWP(j, i) = \frac{bsw(j, i)}{baw(j, i)} \tag{7}$$

2.2. K-means Optimal Cluster Number Determination Method based on BWP Index

Based on BWP index, Zhou Shibing has proposed K-means optimal cluster number determination method ---- Kmeans traversal algorithm, as shown in the following formula:

$$avg_{BWP}(k) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} BWP(j,i) \qquad (8)$$

$$k_{opt} = \underset{2 \le k \le k_{max}}{\arg\max}\{avg_{BWP}(k)\} \qquad (9)$$

Where $avg_{BWP}(k)$ represents the mean BWP index value when the data are clustered into k clusters, k_{opt} represents the optimal cluster number and k_{max} represents the upper search limit of the clustering algorithm.

Figure 1 shows the variation curve of BWP index for K-means clustering of SM2 data set, wherein SM2 data set is composed of the two-dimensional Gaussian distribution data respectively with (0, 0), (5, 5), (10, 10) and (15, 15) as the centers, each cluster includes 225 samples, the covariance matrix of each cluster is $2I_2$, and the structural feature of the data set is that the between-cluster distances of some clusters are very short and there are many overlapped cluster structures. In the experiment, the search scope of the cluster number is $[2, k_{max}]$, and $k_{max} = Int(\sqrt{n})$ is set according to the universal empirical rule $k_{max} \le \sqrt{n}$. According to the figure, when the cluster number is 4, BWP index reaches the maximum value and has high distinction degree, so BWP index can be adopted to well judge the cluster number. Figure 2 shows the clustering effect when the cluster number is 4.
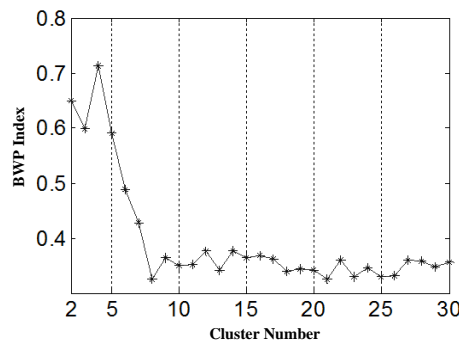

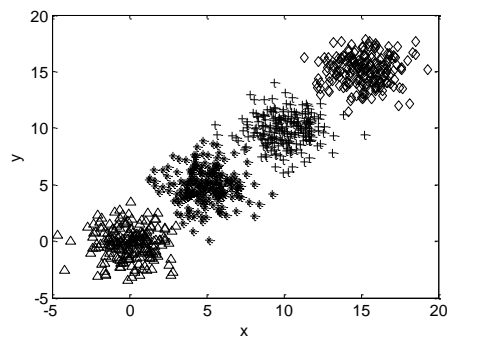**Figure 1. Relational Graph for Cluster Number of SM2 Data Set and BWP Index**

Figure 2. Clustering Result of Data Set SM2 (k=4)

According to Figures 1 and 2, the optimal cluster number can be well determined by this method, but this method needs to traverse all values in the interval of $[2, k_{max}]$, thus having excessive calculation workload when there are many observed signal samples.

3. Improved Optimal Cluster Number Determination Algorithm

3.1. Particle Swarm Optimization Algorithm Principle

Particle swarm optimization (PSO) algorithm is a swarm intelligence based optimization algorithm proposed according to the bird flock behavior rule simulation. The swarm intelligence optimization search generated by the cooperation and competition among the particle individuals in the swarm, and the solution of each swarm has the features of learning the individual optimal solution and the global optimal solution, and each particle is represented by the position vector and the velocity vector. During the iterative optimizing process, each particle determines its flying direction according to its preset flying direction, the traversed optimal direction and the optimal direction of the whole bird flock, wherein the velocity and the position are updated as follows:

$$v_i^{t+1} = wv_i^t + c_1 r_1(pbest_i^t - x_i^t) + c_2 r_2(gbest_i^t - x_i^t) \tag{10}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{11}$$

Where inertia weight factor $w = w_{\max} - \frac{w_{\max} - w_{\min}}{m} \times k$, as the time-dependent linear reduction function, can make PSO have large search area to rapidly locate the rough position of the optimal cluster, and along with the reduction of w, the particle velocity is reduced for fine local search [6]; $w_{\max}$ and $w_{\min}$ are respectively the initial weight factor and the final weight factor, m is the maximum iterations, and k is the present iterations; v_i^t is the particle velocity vector and x_i^t is the present position of the particle; r_1 and r_2 as the random numbers in the interval of $[0, 1]$ are used for keeping the diversity of the swarm; c_1 and c_2 are the learning factors for self-summarization and for learning from the optimal individual; $pbest_i^t$ is the optimal position found by the particle itself, and $gbest_i^t$ is the optimal position found by the whole swarm.

3.2. Optimal Cluster Number Determination Algorithm Improved by Particle Swarm Optimization Algorithm

According to the analysis in Section 1, Kmeans traversal algorithm has such disadvantages of heavy calculation workload and long operation time. However, PSO algorithm does not need to traverse all possible solutions, thus having the ability for reducing the calculation workload and rapidly obtaining the optimal solution. In the following paragraph, we will adopt PSO algorithm to improve Kmeans traversal algorithm.

Firstly, we define the particle swarm as follows:

1) Each particle is one-dimensional and the cluster number corresponding to the position thereof is k ;

2) Mean BWP index for clustering the data set into k clusters is corresponding to the fitness of the particle;

3) The motion scope of the particle is $[2, k_{\max}]$. The position obtained according to Formula (11) is not an integer but K-means cluster number k is an integer, so we need to approximate the new position updated for each time according to Formula (12).

$$x_i^{t+1} = round(x_i^t + v_i^{t+1}) \tag{12}$$

4) The size of the particle swarm depends on the data volume of the observed signals, and is usually set as 10.

Secondly, for the same cluster number, mean BWP indexes obtained through K-means algorithm for several times are approximate to each other [5], different particles in the particle swarm optimization algorithm may be updated to the same position for several times, thus indicating that the clustering operation should be implemented for the same

cluster number for several times and this undoubtedly spends a lot of time. Therefore, we make the following improvements: the position updated for each particle and the fitness of the position are saved as a global variable, and before calculating the fitness of each particle, it is necessary to inquire whether the present position data of the particle have been included in the global variable: if the data are included, then it is necessary to select the corresponding value of the variable; or else, it is necessary to continue the calculation.

4. UBSS based on PSO-Kmeans Algorithm

4.1. Initialization of Observed Signal

When the source signal is a sparse signal, the composite signal has linear cluster characteristic, if the noise influence is ignored, Formula (1) can be expanded as follows:

$$\begin{pmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{pmatrix} = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} s_1(t) + \cdots + \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} s_n(t) \qquad (13)$$

When only a source signal (for example, only $s_i(t)$) is effective at a certain moment, the above formula can be converted as follows:

$$\begin{pmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{pmatrix} = \begin{pmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{pmatrix} s_1(t) \Rightarrow \frac{x_1(t)}{a_{1i}} = \cdots = \frac{x_m(t)}{a_{mi}} = s_i(t) \qquad (14)$$

In a m-dimensional space, it is a straight line passing the origin, and the straight line direction depends on the ith column vector $(a_{1i}, L, a_{mi})^T$ of the aliasing matrix. On this basis, n source signals $s_i(t), i \in (1,2,L\ n)$ will determine n straight lines. Therefore, the evaluation of A is converted into the evaluation of the straight line direction in the observation space. Four paths of long whistling signals in literature [7] are left multiplied with a random 2X4 aliasing matrix to obtain two paths of observed signals. The long whistling signals are not obviously sparse in the time domain but obviously sparse in the frequency domain, so Fourier transformation is implemented for the two paths of observed signals. The scatter diagram of the aliasing signal after transformation is as shown in Figure 3.
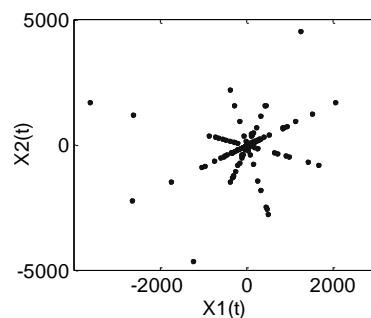

Figure 3. Scatter Diagram of Two Paths of Observed Signals

According to Figure 3, the four paths of source signals can determine four straight lines.

4.2. Estimated Aliasing Matrix

The scatter diagram of n paths of source signals presents n straight lines passing the origin after aliasing, and the straight line directions are corresponding to relevant column vectors of the aliasing matrix. The clustering algorithm is adopted to classify the observed signals according to the straight line directions in the scatter diagram, and each cluster is

corresponding to one straight line, and the direction represented by each cluster center is approximate to one straight line direction in the scatter diagram, namely corresponding to a column vector of the aliasing matrix. In this way, the estimation of the aliasing matrix can be obtained through clustering.

Firstly, the straight line clustering is converted into dense clustering in this paper in order to standardize the observed signals, wherein the standardization treatment includes dimension normalization, direction mirroring[8] and calculation of the arc distance to point (1, 0). Specifically, the dimension normalization of observed signal $X(t)$ is as follows:

$$X'(t) = X(t) / \|X(t)\| \tag{15}$$

Where $\|X(t)\|$ represents Euclidean norm of $X(t)$; the direction mirroring means to map the vector at the lower half of the plane to the upper half of the plane. During standardization, the data points near the origin are overlapped with each other or are the noise points, so these data points can be eliminated in order to effectively reduce the calculation workload and improve the estimation accuracy. After dimension normalization and direction mirroring, the observed signals are all at the upper circumference, as shown in Figure 4.
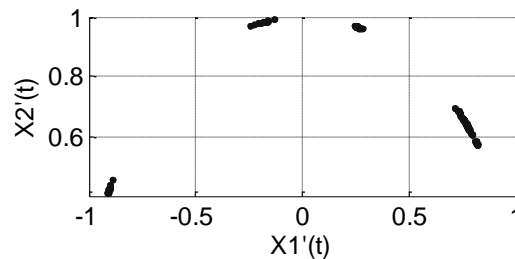

Figure 4. Observed Signals after Dimension Normalization and Direction Mirroring

Then, arc distance $d(t)$ of the observed signal to point (1, 0) after dimension normalization and direction mirroring is calculated as follows:

$$d(t) = \begin{cases} \arctan\left(\dfrac{X'_2(t)}{X'_1(t)}\right), & X'_1(t) > 0 \\ \arctan\left(\dfrac{X'_2(t)}{X'_1(t)}\right) + \pi, & X'_1(t) < 0, \, t = 1, 2, \cdots, T \\ \pi / 2, & X'_1(t) = 0 \end{cases} \tag{16}$$

The points on the same straight line belong to the same cluster, so the arc distances of the observed signals of the same cluster to point (1, 0) are slightly different from each other. Figure 5 shows the distribution of the arc distances of the observed signals to point (1, 0) after dimension normalization and direction mirroring.
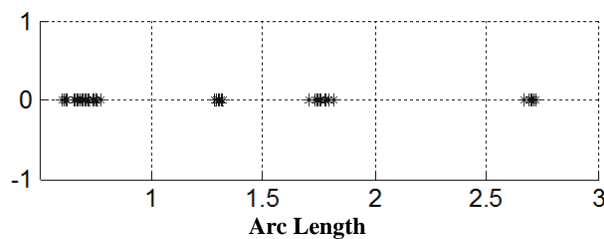

Figure 5. Standardized Observed Signal

Then, PSO-Kmeans algorithm is adopted for the standardized observed signals to obtain optimal cluster number k and corresponding cluster center set $C = \{c_1, c_2, \cdots, c_k\}$. In a unit circle, the arc length is equal to the radian, so each cluster center is collinear with a certain column of aliasing matrix A, and this column of A can be obtained according to Formula (17).

$$a_i = [\cos(c_i), \sin(c_i)]^T, \ i = 1, 2, \cdots k \tag{17}$$

4.3. Source signal recovery

After the aliasing matrix is estimated, the shortest path method can be adopted to recover the source signals. The blind source separation of the sparse signal can be concluded as the following optimization problem [10]:

$$\min_{A,S} \frac{1}{2\sigma^2} \|AS - X\|^2 + \sum_{i,t} |s_i(t)| \tag{18}$$

Where σ^2 is the noise variance, the first item is the quadratic sum of the reconstruction error, and the second item is the non-sparse item. Under the precondition of not considering the noises, since aliasing matrix A has been estimated, the above formula is equivalent to the following formula:

$$\begin{cases} \min_{s(t)} \sum_i^n |s_i(t)|, \\ As(t) = x(t), t = 1, 2, \cdots, T. \end{cases} \tag{19}$$

According to the above formula, an optimization problem is determined at each moment t so as to estimate the source signal.

4.4. Steps of Undetermined Blind Source Algorithm

According to the above analysis, the undetermined blind source algorithm includes the following steps:

(1) Standardize the observed signals;

(2) Adopt PSO-Kmeans algorithm to obtain the optimal cluster number of the standardized observed signals, namely the number of source signals;

(3) Find the center corresponding to the optimal cluster number and adopt Formula (17) to calculate separation matrix A;

(4) Adopt the shortest path method to recover the source signal.

5. Experimental Analysis

The experimental simulation is carried out in Matlab 7.12 environment, and the six paths of long whistling signals are regarded as the source signals [7] to generate two paths of aliasing signals, namely $m = 2$ and $n = 6$, and the aliasing matrix is $A = \begin{bmatrix} 0.7660 & 0.5000 & 0.2588 & -0.1736 & -0.7071 & -0.9063 \\ 0.6428 & 0.8660 & 0.9659 & 0.9848 & 0.7071 & 0.4226 \end{bmatrix}$, and the method mentioned in 3.2 is adopted to standardize the observed signals so as to obtain the standardized observed signals, as shown in Figure 6.
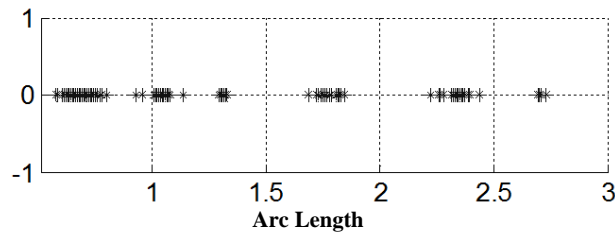
Figure 6. Standardized Observed Signal

The standardized observed signals totally include 6,405 data points.

Table 1. Included Angle (°) of the Corresponding Columns of Aliasing Matrix and Estimated Aliasing Matrix

Algorithm	Included Angle (°) of the Corresponding Columns of Aliasing Matrix and Estimated Aliasing Matrix					
	$\langle a_1, \hat{a}_1 \rangle$	$\langle a_2, \hat{a}_2 \rangle$	$\langle a_3, \hat{a}_3 \rangle$	$\langle a_4, \hat{a}_4 \rangle$	$\langle a_5, \hat{a}_5 \rangle$	$\langle a_6, \hat{a}_6 \rangle$
The Proposed Algorithm	0.3797	0.0748	0.9154	0.1333	0.3422	0.4447
Conventional K-means Algorithm	2.7439	1.5991	9.4873	5.4032	4.2728	1.9053

Where the upper limit of the cluster number is $k_{\max} = Int(\sqrt{6405}) \approx 80$. Particle number is set as 10, the research scope of each particle is[2,80], the maximum iterations are 50, the initial weight factor is $w_{\max} = 0.9$, the final weight factor is $w_{\min} = 0.4$, the learning factor is $c_1 = 1.5$ and $c_2 = 2$. Then, PSO-Kmeans clustering algorithm is adopted to calculate the optimal cluster number as 6 and the corresponding avg_{BWP} as 0.9584. The clustering result is as shown in Figure 7.
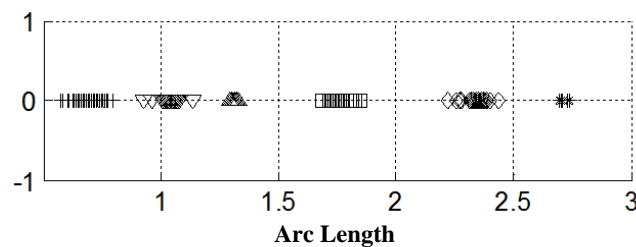

Figure 7. Clustering Result of Standardized Observed Signal

Then, the estimated aliasing matrix is calculated to obtain $\hat{A} = \begin{bmatrix} 0.7703 & 0.5050 & 0.2601 & -0.1893 & -0.7016 & -0.9073 \\ 0.6377 & 0.8630 & 0.9656 & 0.9819 & 0.7126 & 0.4205 \end{bmatrix}$. In order to explain the advantages of the proposed algorithm for estimating the aliasing matrix, the conventional K-means clustering algorithm is adopted to estimate the aliasing matrix as follows under the condition of knowing the cluster number, namely knowing the number of the source signals: $\hat{A}_c = \begin{bmatrix} 0.7956 & 0.5317 & 0.3477 & -0.0782 & -0.6078 & -0.9113 \\ 0.6052 & 0.8643 & 0.7636 & 0.9726 & 0.7059 & 0.4624 \end{bmatrix}$; then, Formula (20) is adopted to respectively calculate the included angles of the corresponding columns of $\hat{A}$, $\hat{A}_c$ and A so as to verify the accuracy for aliasing matrix estimation[9].

$$ang(a, \hat{a}) = \frac{180}{\pi} \arccos \left\{ \frac{\langle a, \hat{a} \rangle}{\|a\| \cdot \|\hat{a}\|} \right\} \tag{20}$$

Where a represents the column vector in original aliasing matrix A, and $\hat{a}$ represents the corresponding column vector of aliasing matrix $\hat{A}$ estimated according to the above algorithm. The included angle of the column vector of the estimated aliasing matrix and the column vector of the original aliasing matrix is as shown in Table 1, and we can know that the proposed algorithm is superior to conventional K-means clustering algorithm.

Additionally, the operation time and the average memory of Kmeans traversal algorithm and PSO-Kmeans algorithm for determining the optimal cluster number, namely the number of the source signals, are respectively shown in Figures 8 and 9. According to Figure 8, along with the increase of the upper limit of the cluster number, the operation time of Kmeans traversal algorithm is linearly increased while that of the proposed algorithm is slowly increased; in case of a large cluster number, namely a large amount of observed signal samples, the proposed algorithm can save a lot of time. According to Figure 9, when the upper limit of the cluster number is relatively small, the average memory consumption of the two algorithms is slightly different from each other, but in case of large cluster number, the average memory consumption of Kmeans traversal algorithm tends to be increased.
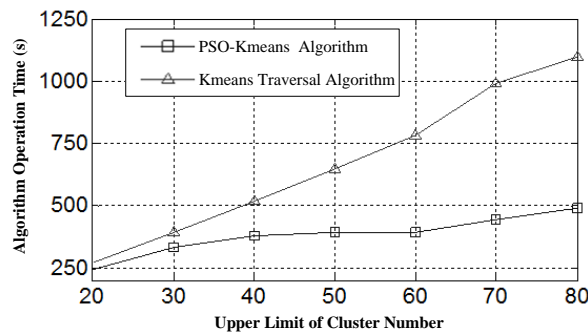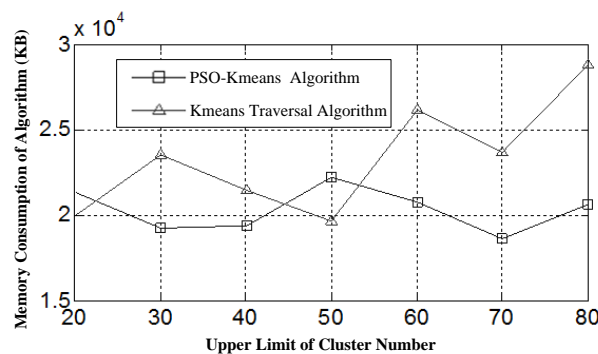

Figure 8. Operation Time Comparison

Figure 9. Average Memory Consumption Comparison

6. Conclusion

BWP index for evaluating the clustering result and Kmeans traversal algorithm are researched in this paper, and the particle swarm optimization algorithm is also introduced in this paper to propose the improved algorithm for estimating the number of the source signals ---- PSO-Kmeans algorithm. Meanwhile, PSO-Kmeans algorithm is introduced therein for the undetermined blind source separation, and the separation process of the proposed algorithm is introduced in this paper in detail. Finally, the experiment is carried out for six paths of source signals and two paths of observed signals. The undetermined blind source separation of the signals is compared with conventional K-means algorithm in the aspect of the separation accuracy, and is also compared with Kmeans traversal

algorithm to obtain the operation time and the memory consumption for obtaining the number of the source signals. The experiment result shows that the proposed algorithm can not only ensure the recovery accuracy and shorten the separation time, but also save certain memory and present more obvious advantages when there are many observed signals.

References

[1]   Wei Gu, Zhihan Lv, Ming Hao. Change detection method for remote sensing images based on an improved Markov random field. Multimedia Tools and Applications. 2016.

[2]   Zhihan Lu, Chantal Esteve, Javier Chirivella and Pablo Gagliardo. A Game Based Assistive Tool for Rehabilitation of Dysphonic Patients. 3rd International Workshop on Virtual and Augmented Assistive Technology (VAAT) at IEEE Virtual Reality 2015 (VR2015), Arles, France, IEEE, 2015.

[3]   Zhihan Lv, Alaa Halawani, Shengzhong Feng, Haibo Li, and Shafiq Ur Rehman. Multimodal Hand and Foot Gesture Interaction for Handheld Devices. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 11, 1s, Article 10 (October **2014**), 19 pages.

[4]   Yancong Lin, Jiachen Yang, Zhihan Lv, Wei Wei, Houbing Song. A Self-Assessment Stereo Capture Model Applicable to the Internet of Things. Sensors. **(2015)**.

[5]   Wei Ou, Zhihan Lv, Zanfu Xie. Spatially Regularized Latent topic Model for Simultaneous object discovery and segmentation. The 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC **2015**).

[6]   Yi Wang, Yu Su, Gagan Agrawal. A Novel Approach for Approximate Aggregations Over Arrays. In Proceedings of the 27th international conference on scientific and statistical database management, ACM, **(2015)**.

[7]   Zhihan Lv, Alaa Halawani, Shengzhong Feng, Shafiq ur Rehman, Haibo Li. Touch-less Interactive Augmented Reality Game on Vision Based Wearable Device. Personal and Ubiquitous Computing. **(2015)**.

[8]   Jiachen Yang, Shudong He, Yancong Lin, Zhihan Lv. Multimedia cloud transmission and storage system based on internet of things. Multimedia Tools and Applications. **(2016)**.

[9]   Dingde Jiang, Xu Ying, Yang Han, Zhihan Lv. Collaborative Multi-hop Routing in Cognitive Wireless Networks. Wireless Personal Communications. **(2015)**.

[10] Weixi Wang, Zhihan Lu, Xiaoming Li, Weiping Xu, Baoyun Zhang, Xiaolei Zhang. Virtual Reality Based GIS Analysis Platform. 22th International Conference on Neural Information Processing (ICONIP **2015**), Istanbul, Turkey.

[11] Zhihan Lv, Chantal Esteve, Javier Chirivella and Pablo Gagliardo. Clinical Feedback and Technology Selection of Game Based Dysphonic Rehabilitation Tool. 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth2015), IEEE, **(2015)**.

[12] Jiachen Yang et al. A Low-Power and Portable Biomedical Device for Respiratory Monitoring with a Stable Power Source. Sensors 15.8 **(2015)**: 19618-19632.

Author

Guo Xukun, He received his M.S. degree in computer software and theory from South China Normal University in Guangzhou, Guangdong, China. He is currently a lecturer in the Modern Education Technology Center at Guangzhou Sport University. His research interest is mainly in the area of computer network, big data applications, computer software and theory. He has published several research papers in scholarly journals in the above research areas.