Bio-Medical Materials and Engineering 24 (2014) 3709–3717
DOI 10.3233/BME-141199
IOS Press

Initial points selection for clustering gene expression data: A spatial contiguity analysis-based approach

Hui Yi[a, b], Cuimei Bo[a], Xiaofeng Song[b] and Yuhao Yuan[a,*]
[a] *College of Automation and Electrical Engineering, Nanjing University of Technology, No. 30, Puzhu South Road, Nanjing 211816, China*
[b] *Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, No. 29, Yudao Sreet, Nanjing 210016, China*

Abstract. Clustering is considered one of the most powerful tools for analyzing gene expression data. Although clustering has been extensively studied, a problem remains significant: iterative techniques like k-means clustering are especially sensitive to initial starting conditions. An unreasonable selection of initial points leads to problems including local minima and massive computation. In this paper, a spatial contiguity analysis-based approach is proposed, aiming to solve this problem. It employs principal component analysis (PCA) to identify data points that are likely extracted from different clusters as initial points. This helps to avoid local minima, and accelerates the computation. The effectiveness of the proposed approach was validated on several benchmark datasets.

Keywords: Gene expression data, k-means, initial points, spatial contiguity analysis

1. Introduction

Clustering is an important tool widely used in genomic studies [1–3]. By clustering the gene expression data that describe how the genetic information converts to a functional gene product, genes with similar biological significance can be grouped together. This helps to identify the functions of unknown genes and extract the potential co-regulation or discriminate pathologies of genes [4–6].

Although well studied with many achievements obtained [7–10], gene expression data clustering remains problematic: as has been pointed out by Jain et al. [8], different selections of initial points lead to very different clustering processes, and result in different classification results. If arbitrary selections of initial points have been made, misleading biological conclusions are liable to be obtained.

A conventional solution to this problem is the random seeds strategy, i.e. to repeat the clustering many times, each with randomly selected data points as initial points. This strategy helps to avoid local minima [9]. However, unless all seeds are investigated, it is not guaranteed to avoid local minima.

*Corresponding author: Yuhao Yuan, College of Automation and Electrical Engineering, Nanjing University of Technology, No. 30, Puzhu South Road, Nanjing 211816, China. Tel.: +8613851415358; Fax: +86-25-58139517; E-mail: yyhmds@sohu.com.

Moreover, it offers little theoretical guide for solving the problem. Methods like competitive learning [10], genetic algorithm [11], particle swarm optimization [12], and density analysis [13,14] have been introduced to optimize the k-means clustering. These methods can effectively help the clustering to jump out of local minima, yet they do not provide an optimized initial points selection strategy. They usually increase the number of iterations in order to achieve global minima. In addition, these methods usually require massive computation.

To offer a facilitated and effective way of selecting the initial points for k-means clustering, a Spatial Contiguity Analysis (SCA)-based approach is proposed in this paper. This approach adopts the idea that the ideal selection of initial points should be the cluster centroids, which, however, could not be found until the iteration is completed. Thus, the SCA-based approach attempts to find samples that are likely to be near the centroids as the initial points. Using this approach helps to avoid local minima, while eliminates the extra computational cost.

2. K-means clustering

Let $X = \{\mathrm{x}_i \,|\, i = 1,\ldots,n\}$ denote a set of n gene expression samples, and k be a positive integer, k-means clustering aims to find a partition that divides sample set X into k disjoint regions. For a given initial set $Q = \{\mathrm{q}_j \,|\, j = 1,\ldots,k\}$ that contains k different data points, sample x_i is considered belonging to the j^*-th cluster according to the following criterion:

$$j^* = \arg\min_j dist(\mathrm{x}_i, \mathrm{q}_j), \quad \mathrm{j} = 1,\ldots,\mathrm{k}\ , \tag{1}$$

where $dist(\mathrm{x}_i, \mathrm{q}_j)$ indicates the Euclidean distance between x_i and q_j. Let D_j be the set of samples that belong to the j-th cluster, the cluster centroids for the first iteration $C_1 = \{\mathrm{c}_j \,|\, j = 1,\ldots,k\}$ could be calculated by:

$$c_j = average(\mathrm{D}_j), \quad \mathrm{j} = 1,\ldots,\mathrm{k}\ . \tag{2}$$

Then, the cluster centroids are iteratively calculated with the k-means algorithm, as shown in Figure 1.

The k-means algorithm repeats the process of generating new divisions and new cluster centroids till further iteration causes no change to the cluster centroids. However, k-means clustering does not guarantee unique results, since the converging process is sensitive to the selection of initial points [15]. The following example is designed to illustrate how the selection of initial points influences the clustering performance. Firstly, 17 two-dimensional samples, from three different clusters, are prepared as shown in Figure 2. Then, different sets of initial points are used for clustering the given dataset. Table 1 presents the clustering performances. As seen in the table, when $(x1, x2, x3)$ or $(x8, x9, x10)$ is selected as the initial points set, the k-means clustering falls into local minima. Meanwhile, the set

$(x7, x8, x9)$ results in the global minima, but the computational cost is huge. When using $(x1, x12, x16)$ as the initial points set, the k-means clustering converges correctly within just two iterations. Evidently, the convergence and converging speed are highly dependent on the selection of initial points.

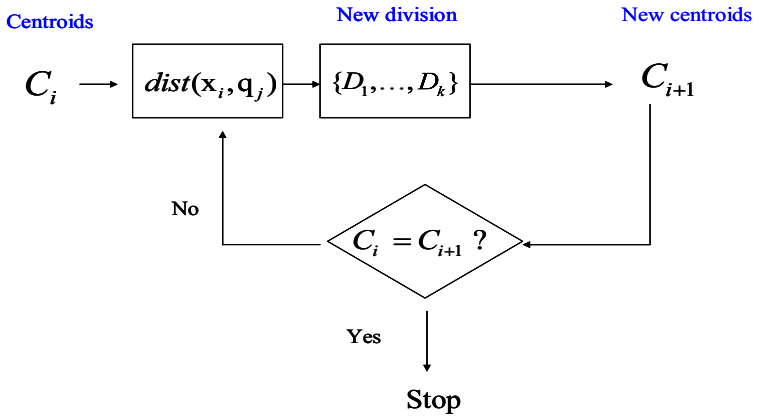


Fig. 1. The process of updating the cluster centroids.

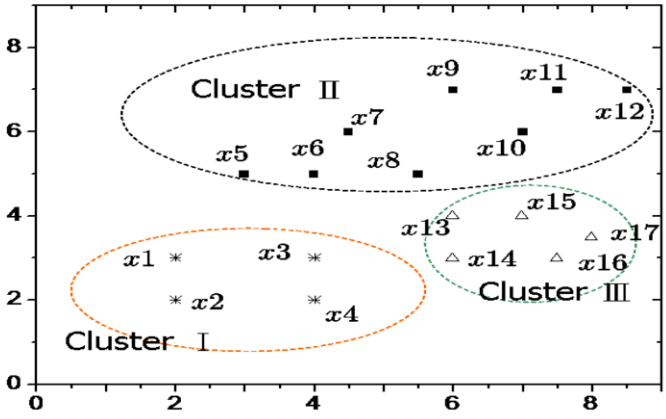


Fig. 2. The dataset for illustration.

Table 1

Performances of k-means clustering with different selections of initial points

Initial Points	(x_1, x_2, x_3)	(x_8, x_9, x_{10})	(x_7, x_8, x_9)	(x_1, x_{12}, x_{16})
Number of Iterations	4	5	4	2
Within-cluster distances	39.8375	39.8375	32.2	32.2

3. SCA-based approach for initial points selection

The performance of iterative clustering that may converge to local minima depends highly on the selection of initial points [16]. There is no denying that the perfect set of initial points should be the cluster centroids. However, the centroids cannot be determined before the k-means clustering is completed. This paper proposes an SCA-based approach, which finds sub-optimal points that are likely to be distributed around the centroids as the initial points for k-means clustering.

For a given gene expression dataset $X \in \Re^{n\times m}$, a matrix Y is firstly generated by

$$Y \equiv \frac{1}{\sqrt{n-1}} X^T \quad . \tag{3}$$

Then, the eigenvectors for $Y^T Y$ and principle components for X are calculated as follows:

$$\left|\lambda I - Y^T Y\right| = 0 \quad , \tag{4}$$

$$\begin{pmatrix} \lambda_i - 1 & -r_{12} & \cdots & -r_{1p} \\ -\lambda_{21} & \lambda_{i-1} & \cdots & -r_{2p} \\ & & \cdots & \\ -r_{n1} & -r_{n2} & \cdots & \lambda_{i-1} \end{pmatrix} \begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \vdots \\ \alpha_{ni} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{5}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and $F_i = \alpha_{1i} Y_1 + \alpha_{2i} Y_2 + \cdots + \alpha_{ni} Y_n$ is the i-th principle component of the given dataset X. The required number l of principle components in SCA-based approach can be decided by

$$\frac{k}{2} \leq l \leq \frac{k+1}{2},\ l \in N \ , \tag{6}$$

where k is the given number of clusters is the required number of principle components in our approach.

Principle component reveals the intrinsic distribution of samples. Thus, we use PCA to find the 'corner points', i.e. data points near the tails (or the edges) of the dataset. Given l principle components, for each principle component, the data points with the greatest and smallest values are considered 'corner points'. For the i-th principle component, corner points $\{s_{2i-1}, s_{2i}\}$ are selected.

Remark: When k is an odd number, the corner points should be $\{s_1, s_2, \ldots, s_{2l-1}\}$; otherwise, the corner points should be $\{s_1, s_2, \ldots, s_{2l}\}$.

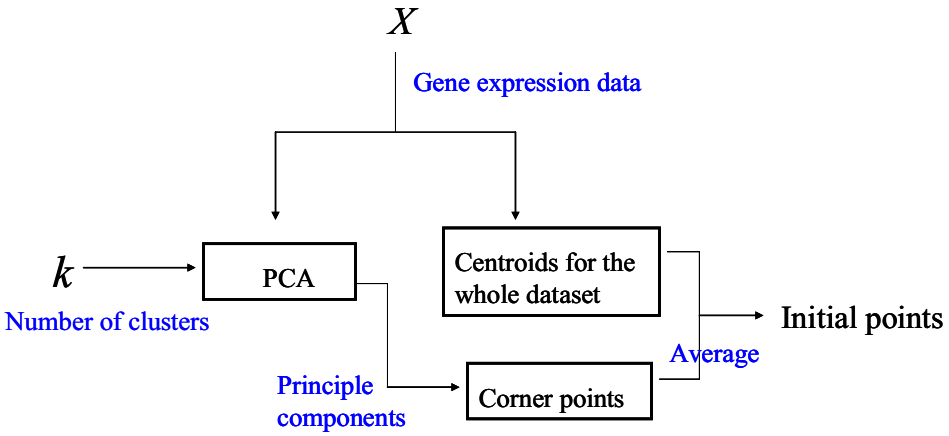


Fig. 3. Flowchart of selecting the initial points for k-means clustering.

As shown in Figure 3, for gene expression dataset X and the number of clusters k, the initial points selection process of the proposed algorithm is as follows:

Step 1 The centroid for the whole dataset is calculated by

$$X_{Centriod} = \left[\sum_{i=1}^{n} X_{i1}, \ldots, \sum_{i=1}^{n} X_{in}\right] \Big/ n \quad . \tag{7}$$

Step 2 By PCA, we obtain the first l principle components for X. The points with the greatest and smallest values in each principle component are extracted as the corner points. The corner points set S should be

$$\begin{cases} S = \{s_1, s_2, \ldots, s_{2l-1}\}, & if\ 'k'\ is\ an\ odd \\ S = \{s_1, s_2, \ldots, s_{2l}\}, & else \end{cases} \tag{8}$$

Step 3 The initial points Q are generated by

$$Q = \frac{S + X_{Centriod}}{2} \quad . \tag{9}$$

As shown in Figure 4, when k=3, the red lines indicate the first and second principle components of sample set X. The red circular dots denote the 'corner points' found by PCA. The black circular dot denotes the centroid for the whole dataset. By calculating the average value of the centroid and 'corner points', the initial points are generated, as denoted by the blue circular dots.

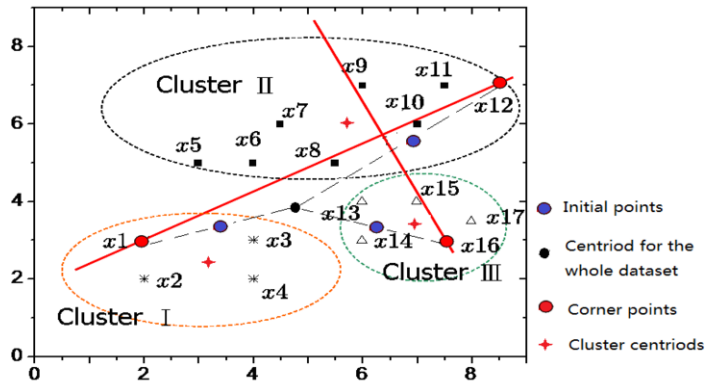
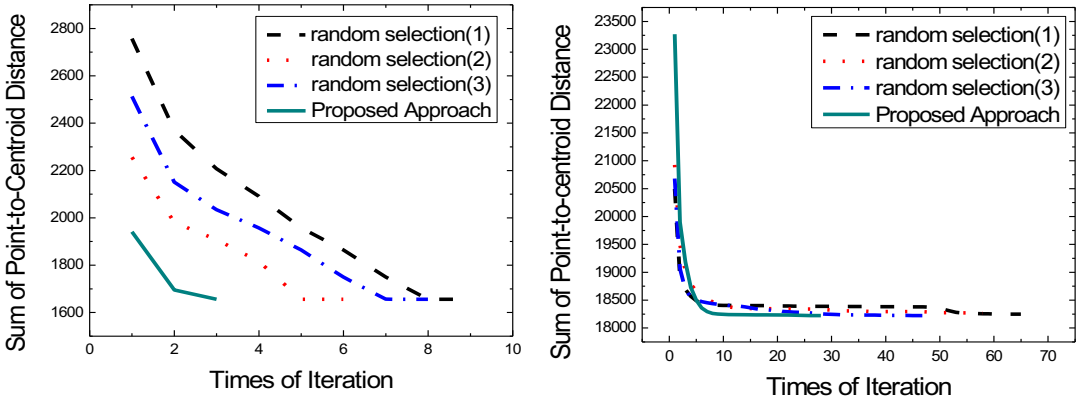
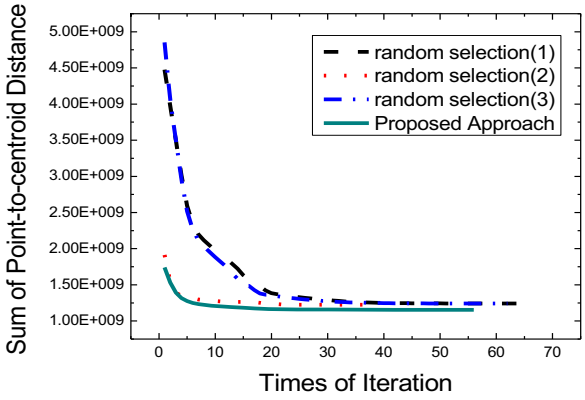


Fig. 4. Illustrative example for SCA-based Initial point selection.



(a) The converging process for Rats CNS dataset (k=2)

(b) The converging process for B-cell Lymphoma dataset (k=5)



(c) The converging process for Yeast Saccharomyces Cerevisiae dataset (k=7)

Fig. 5. Clustering performances with different initial point selections.

4. Validations

4.1. Benchmark datasets

Three benchmark datasets, namely, Rats CNS (k=2) [17], B-cell Lymphoma (k=5) [18] and Yeast Saccharomyces Cerevisiae (k=7) [19], were employed for the validation of the proposed approach. These datasets contain different numbers of clusters.

In the experiments, random seeds approach and the proposed approach were implemented to select the initial points for the k-means clustering. The number of iterations and total distances of within-cluster elements were recorded for each clustering process.

As seen in Figure 5(a), with different selections of initial points, the clustering processes all converged to the same value of 'Sum of Point-to-Centroid Distance', implying that all selections yielded the same classification result. However, it only took the proposed approach two iterations to finish the convergence, which means the clustering with initial points selected by the proposed approach converged much faster than those whose initial points were selected by the random seeds approach. Figure 5(b) shows a more complicated case, where the dataset was composed of five clusters ($k = 5$). The proposed approach also yielded the highest converging speed. In Figure 5(c), where the dataset consisted of seven clusters (k = 7), all clustering processes with 'random seeds' selections failed to reach the smallest value of total point-to-centroid distance and fell into local minima, whereas the proposed SCA-based approach converged correctly within 15 iterations.

4.2. Different numbers 'k' of clusters

The SCA-based approach was also validated in conditions where different numbers of clusters were set. In this experiment, Yeast Saccharomyces Cerevisiae dataset was employed, with the setting of different numbers of clusters 'k = 3,4,5,6,7 and 8'. The performances of clustering processes with initial points selected by SCA-based approach and random seeds approach were investigated. As seen in Figure 6, the SCA-based approach always yields the smallest 'Sum of Point-to-centroid Distance', demonstrating its reduced possibility of falling into local minima.

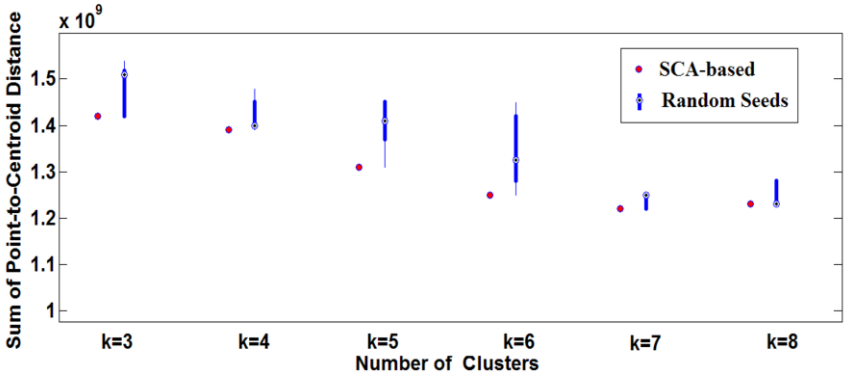


Fig. 6. The clustering performances with different numbers of clusters.

In both experiments, the proposed approach could always select a reasonable set of initial points, and the sets it selected demonstrated superiority to other selections. The effectiveness of the SCA-based initial point selection approach is thus validated.

5. Conclusion

Clustering gene expression data is a useful tool for extracting biological information. The clustering process, however, is greatly influenced by the selection of initial points, and misleading biological conclusions may be made if unnecessary initial points have been selected. A spatial contiguity analysis-based initial point selection approach was proposed in this paper. The proposed approach aims to find samples that are around the centroids as initial points. This strategy may accelerate the convergence and avoid local minima. We validated the proposed SCA-based approach by applying it to benchmark datasets, and satisfactory results were received.

Acknowledgement

The work described in this paper is supported by grants from the Doctoral Fund of Ministry of Education of China (20113218110011), the National Science Foundation of China (61171191, 61203020), the Science Foundation of Jiangsu province (BK20140953), the Science Foundation of Jiangsu High Schools (13KJB510013) and the Open project of key laboratory of Hunan province (2013NGQ004).

References

[1] K.Y. Yeung, D.R. Haynor and W.L. Ruzzo, Validating clustering for gene expression data, Bioinformatics **17** (2001), 309–318.
[2] Julia Handl, Joshua Knowles and Douglas B. Kell, Computation cluster validation in post-genomic data analysis, Bioinformatics **21** (2005), 3201–3212.
[3] M.A. Fadhl, Analysis and visualization of gene expression data using biclustering: a comparative study, African Journal of Biotechnology **11** (2013), 1744–1753.
[4] P. Maji and S. Paul, Rough-fuzzy clustering for grouping functionally similar genes from microarray data, IEEE Trans. on Computational Biology and Bioinformatics **10** (2013), 286–299.
[5] A.K. Das, D. Mandal, M. Adhikary and A. K. Sen, Fuzzy mining approach for gene clustering and gene function prediction, International Journal of Computer Science and Mobile Computing **3** (2014), 309–318.
[6] Hui Yi, Xiaofeng Song, Bin Jiang and Yufang Liu, Reducing the subjectivity of gene expression data clustering based on spatial contiguity analysis, Database Theory and Application, Bio-Science and Bio-Technology **258** (2011), 118–124.
[7] Rui Xu and Donald Wunsch Ⅱ, Survey of clustering algorithms, IEEE Trans. on Neural Networks **16** (2005), 645–678.
[8] A.K. Jain, M.N. Murty and P.J. Flynn, Data clustering: A review, ACM Comput. Surveys. **31** (1999), 264–323.
[9] Patrik D'haeseleer, How does gene expression clustering work? Nature Biotechnology **23** (2005), 1499–1501.
[10] Hong Jia, Yiu-ming Cheung and Jiming Liu, Cooperative and penalized competitive learning with application to kernel-based clustering, Pattern Recognition **47** (2014), 3060–3069.
[11] M. Laszlo and S. Mukherjee, A genetic algorithm that exchanges neighbouring centres for k-means clustering, Pattern Recognition Letters **28** (2007), 2359–2366.
[12] Jie Zhang, Yuping Wang and Junhong Feng, A hybrid clustering algorithms based on PSO with dynamic crossover, Soft Comput. **18** (2014), 961–979.
[13] Xiang Li and Suhong Liu, The optimal initial centers clustering algorithm based on local outlier factor, Microelectronics & Computer **30** (2013), 109–112.
[14] Desheng Fu and Chen Zhou, Improved k-means algorithm and its implementation based on density, Journal of Computer Applications **31** (2011), 432–434.

[15] Edward R. Dougherty and Marcel Brun, A probabilistic theory of clustering, Pattern Recognition **37** (2004), 917–925.
[16] M.N.M. Sap and Ehsan Moheb, Hybrid self organizing map for overlapping clusters, International Journal of Signal Processing, Image Processing and Pattern Recognition (IJSIP) **1** (2011), 11–20.
[17] Hui Yang, Wenqin Cai and Kecheng Zhang, Distribution of GAP-43 mRNA ub the CNS of adult rats, Chinese Science Bulletin **40** (1995), 1296–1299.
[18] A. Alizadeh, M. Eisen, R. Davis et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature **403** (2000), 503–511.
[19] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang et al., Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Molecular Biology of the Cell **9** (1998), 3273–3297.