PROCEEDINGS OF SCIENCE

A Rank-Based K-medoids Clustering Algorithm by a Specific P System

Qian Li[1]
School of Management Science and Engineering, Shandong Normal University, 250014,Jinan, Shandong,China
E-mail: shangrilaying@126.com

Xiyu Liu[2]
School of Management Science and Engineering, Shandong Normal University, 250014,Jinan, Shandong,China
E-mail: sdxyliu@163.com

In this paper, a rank-based K-medoids algorithm by a specific P system is proposed, which exhibits novel aspect of applying membrane computing in clustering. The traditional K-medoids clustering result suffers sensitivity to initial medoids selected randomly. To conquer the defect, we introduce the rank based on similarity between pairs of objects for the initialization. As a biological computing model, P system imitates the structure and function of living cells, and the reactions in it process in a maximal parallel and distributed manner. P system is adequate to solve clustering problem for its high parallelism and lower computational time complexity. A specific P system with a sequence of rule is constructed to realize the rank-based K-medoids algorithm. Through test verification, it can improve the clustering quality.

CENet2015
12-13 September 2015
Shanghai, China

[1]Speaker
[2]Corresponding Author

PoS(CENet2015)009

1. Introduction

Clustering analysis is the process of dividing a set of objects into non-overlapping subsets. Each subset is a cluster with intra-cluster data similar and inter-cluster data dissimilar. Clustering is a rapidly developing area which contributes to research field including machine learning, spatial database technology, biology[1-3] and marketing[4,5], and so on [6].

Membrane computing is a new computing model firstly proposed by Romanian scientist Gheorghe Paun, and that is why the model is called P systems. It simulates the function of living cells, abstracting biochemical reactions and material exchanges to perform calculation process on the cellular level.

The traditional K-medoids clustering result suffers sensitivity to initialization of medoids and local optimum [7]. We adopt the principle of ranking the similarity between pairs of objects to avoid the disadvantages. In this paper, we combine the rank-based K-medoids clustering with a specific P system to reduce the computational time complexity employing the great parallelism of P systems.

2 .The Rank-Based K-medoids Algorithm

As a well-known algorithm of partitioning approach, K-medoids selects k objects in the dataset as centers for each cluster and form k clusters through iterations. It reduces sensitivity to outliers which the K-means algorithm suffers [8]. Assuming a dataset $D=\{a_1,a_2,...a_n\}$ with n objects, the K-medoids algorithm divides D into k<n clusters. The similarity among objects can be defined by applying the Euclidean distance, Manhattan distance and so on[9,10]. In this paper, the Euclidean distance is applied and it is assumed that all the data is in a two-dimensional space.

First of all, a n*n dissimilarity matrix $D_{n,n}$ is constructed as follows:

$$D_{nn}=\begin{pmatrix} w_{11},w_{12},...w_{1n} \\ w_{21},w_{22},...w_{2n} \\ ... \\ w_{n1},w_{n2},...w_{nn} \end{pmatrix} \tag{2.1}$$

Where, ω_{ij} represents the value by rounding the dissimilarity between any the i-th and j-th object.

In the rank-based K-medoids algorithm, we introduce the concept of dissimilarity rank and group. We employ novel method that ranks objects according to their similarities other than the direct use of their similarity values [11]. By this method, the more dissimilar object gets higher rank. That is, rank($\boldsymbol{a}_i$, $\mathbf{a}_j$)= f indicates that $\boldsymbol{a}_i$ is the f-th similar object to $\mathbf{a}_i$ among n objects in the dataset. The similarity rank matrix (denoted as SimRank in this paper) is constructed by sorting the similarity values between any object $\mathbf{a}_j$ and the other objects. The SimRank is expressed as follows:

$$\text{SimRank}=(r_{i,j}),\ \text{rank}(a_i,a_j)=r_{i,j},\ \forall a_i,a_j\in D \tag{2.2}$$

There are two key points to declare about the SimRank. For one thing, SimRank is a matrix that reveals the similarity rank or closeness among objects in the datasets, and also, it shows the extent of dissimilarity through numbers from 1 to n. For another, SimRank is not always a symmetric matrix on account of two points not always being at the same rank of each other.Group is another important concept in our proposed algorithm. The number of members of

a group is determined by a given parameter g. Within a group, it can compute the dissimilarity rank value (denoted as dv) of an object. The dissimilarity rank value of any object a_i in a group G is defined as follows:

$$\mathrm{dv}(a_i)=\sum_{a_j \in G} r_{i,j} \tag{2.3}$$

The steps of the proposed algorithm proceeds as follows:

1 Initialize k medoids

1) Calculate the similarities between any pairs of objects based on their distance,

2) Rank similarity values of any objects and thus construct the SimRank matrix,

3) Select k medoids randomly;

2 Optimize medoids

1) Choose g most similar objects to each medoids so as to form a group, employing the SimRank (g is a given parameter that indicates the number of members of a group),

2) Compute the dissimilarity rank values of every object in a group, and then update the medoids by a specific object with the maximum dv,

3) Go to step 2 2) until the medoids remain unchanged;

3 Allocate the common objects

Distribute the rest common objects to the most similar medoids.

The similarities between any pairs of objects in dataset is computed once, and updating the medoids costs O(k×g) per iteration, where k implies the number of clusters and g is the number of members in a group needed to select the next medoids. While in traditional K-medoids clustering, it distribute every objects in a partition to update the new medoids, and it costs O(n^2) per iteration.

The advantage of the rank based K-medoids algorithm lies in ranking objects according to their similarities other than the direct use of their similarity values. As a consequence, the clustering result gains more accuracy.

3. A Specific P System for the Rank-Based K-medoids Algorithm

3.1 The designed P system

The reader is assumed to be acquainted with the basic prerequisites of P system. It is suggested that the reader refer to papers for further learning[12,13].In this section, a specific P system is designed to realize the rank-based K-medoids algorithm. The P system is a construct incorporating skin membrane and k elementary membranes and especially an output membrane. Object a_i represents the original corresponding point a_i in the data set, and the subscript of a_i means a_i represents the i-th data point [14]. The matrix D_{nn} is applied to compare the distance between the n objects.

The P system for the rank-based K-medoids is designed as follows:

$\pi=\left(O,H,\mu,M_0,M_1...M_k,M_{c_0},R_0,R_1,...R_k,\rho_i,c_0\right)$ Where：

1) $O=\left(\alpha_{11},\gamma_{1,1,0},\theta_1,\xi_1,a_1,a_2,...a_n,s_0\right)$ denotes the collection of objects in the P system;

2) $M_0=\left(\alpha_{11},\gamma_{1,1,0},\theta_1,\xi_1,a_1,a_2,...a_n\right)$ denotes the initial objects in membrane 0;

3) $M_1=M_{2=}...M_k=\left(s_0\right)$ denotes the initial objects in membrane 1, 2…k;

4) $M_{c_0}=(\lambda)$ denotes the output membrane.

Rule in R_0 [15,16]:

$$r1=\left\{\alpha_{ij}a_ia_j\to\alpha_{i(j+1)}a_ia_jU_{ij}^{\omega_{ij}}\middle|1\le i,j\le n\right\}\bigcup\left\{\alpha_{i(n+1)}\to\alpha_{(i+1),1}\middle|1\le i\le n\right\}\bigcup\left\{\alpha_{n(n+1)}\to\lambda\right\}$$

$$r2=\left\{U_{ij}^{\omega_{ij}}\gamma_{i,j,q}\to\gamma_{i,j,q+1}(\varsigma_{ij}^{q+1})_{in_1}(\varsigma_{ij}^{q+1})_{in_2}...(\varsigma_{ij}^{q+1})_{in_k}\varsigma_{ij}^{q+1}\middle|\omega_{ij}=0,1\le i,j\le n,0\le q\le n\right\}$$

$$\bigcup\left\{\gamma_{i,j,n+1}\to\gamma_{(i+1),1,0}\middle|1\le i,j\le n\right\}\bigcup\left\{\gamma_{n,n+1}\to e\right\}$$

$$r3=\left\{U_{i1}^{\omega_{i1}}U_{i2}^{\omega_{i2}}...U_{in}^{\omega_{in}}\to U_{i1}^{\omega_{i1}-1}U_{i2}^{\omega_{i2}-1}...U_{in}^{\omega_{in}-1}\middle|1\le i\le n\right\}$$

$$r4=\left\{ea_i\theta_t\to eA_{ti}\theta_{t+1}\middle|1\le i\le n,1\le t\le k\right\}\bigcup\left\{e\theta_{k+1}\to\chi_1\right\}$$

$$r5=\left\{\chi_tA_{ti}a_j\varsigma_{ij}^q\xi_p\to\chi_tA_{ti}G_{pj,in_t}\xi_{p+1}\middle|q=0,1\le i,j\le n,1\le t\le k,1\le p\le m\right\}$$

$$r6=\left\{A_{ti}\varsigma_{ij_1}^{q_1}\varsigma_{ij_2}^{q_2}...\varsigma_{ij_n}^{q_n}\to A_{ti}\varsigma_{ij_1}^{q_1-1}\varsigma_{ij_2}^{q_2-1}...\varsigma_{ij_n}^{q_n-1}\middle|1\le i,j_t,q_t\le n\right\}$$

$$r7=\left\{\chi_{k+1}\xi_m\to\xi_1\eta_{1,in_1}\eta_{1,in_2}...\eta_{1,in_k}\right\}$$

$$r8=\left\{\begin{matrix}d^ka_iA_{1j_1}A_{2j_2}...A_{kj_k}U_{ij_1}^{\omega_{ij_1}}U_{ij_2}^{\omega_{ij2}}...U_{ij_k}^{\omega_{ij_k}}\to d^ka_{i,in_p}A_{1j_1,in_1}A_{2j_2,in_2}...A_{kj_k,in_k}\\ \left|\omega_{ij_p}=0,1\le j_1,j_2,...j_k\le n\right.\end{matrix}\right\}$$

$$r9=\left\{U_{ij_1}^{\omega_{ij_1}}U_{ij_2}^{\omega_{ij2}}...U_{ij_k}^{\omega_{ij_k}}\to U_{ij_1}^{\omega_{ij_1}-1}U_{ij_2}^{\omega_{ij2}-1}...U_{ij_k}^{\omega_{ij_k}-1}\middle|1\le j_1,j_2,...j_k\le n\right\}$$

$$r10=\left\{b^id^j\to\phi_{in_1}\phi_{in_2}...\phi_{in_k}\middle|1\le i\le n,0\le j\le k\right\}\bigcup\left\{d^k\to\beta_{in_1}\beta_{in_2}...\beta_{in_k}\right\}$$

$$r10=\left\{\tau\beta\omega\to(\beta\omega)_{in_n}\middle|\omega\subseteq O_{tj}\bigcup\left\{a_p\middle|1\le p\le n\right\}\right\}$$

Rules in membrane t$(1\le t\le k)$:

$$r1'=\left\{(\beta\omega\to\beta\omega\alpha_i)_{a_i}\middle|1\le i\le n,\omega\subseteq O_{tj}\bigcup\left\{a_p\middle|1\le p\le n\right\}\right\}$$

$$r2'=\left\{\beta\omega\to(\tau\beta\omega)_{out}\middle|\omega\subseteq O_{tj}\bigcup\left\{a_p\middle|1\le p\le n\right\}\right\}$$

$$r3'=\left\{\begin{matrix}A_{tp}G_{1j_1}G_{2j_2}...G_{mj_m}\varsigma_{pj_1}^{q_1}\varsigma_{pj_2}^{q_2}...\varsigma_{pj_m}^{q_n}\to A_{tp}G_{1j_1}G_{2j_2}...G_{mj_m}\varsigma_{pj_1}^{q_1}\varsigma_{pj_2}^{q_2}...\varsigma_{pj_m}^{q_m}\varsigma_p^{Q_p}\\ \left|1\le j_1,j_2,...j_m,p\le n,1\le i\le m\right.\end{matrix}\right\}$$

$$r4'=\left\{\begin{matrix}\eta_iG_{ij_i}G_{1j_1}G_{2j_2}...G_{(i-1)j_{(i-1)}}G_{(i+1)j_{(i+1)}}...G_{mj_m}A_{tp}\varsigma_{j_ij_1}^{q_1}\varsigma_{j_ij_2}^{q_2}...\varsigma_{j_ij_{(i-1)}}^{q_{(i-1)}}\varsigma_{j_ij_{(i+1)}}^{q_{(i+1)}}...\varsigma_{j_ij_m}^{q_m}\varsigma_{j_ip}^{q_p}\to\\ \eta_iG_{ij_i}G_{1j_1}G_{2j_2}...G_{(i-1)j_{(i-1)}}G_{(i+1)j_{(i+1)}}...G_{mj_m}A_{tp}\varsigma_{j_ij_1}^{q_1}\varsigma_{j_ij_2}^{q_2}...\varsigma_{j_ij_{(i-1)}}^{q_{(i-1)}}\varsigma_{j_ij_{(i+1)}}^{q_{(i+1)}}...\varsigma_{j_ij_m}^{q_m}\varsigma_{j_i}^{Q_i}\\ \left|1\le j_1,j_2,...j_m,p\le n,1\le i\le m\right.\end{matrix}\right\}$$

$$r5'=\left\{s_hA_{tp}G_{ij_i}\varsigma_p^{Q_p}\varsigma_{j_i}^{Q_{ji}}\to s_{h+Q_{ji}-Q_p}A_{tp}G_{ij_i}\varsigma_p^{Q_p}\varsigma_{j_i}^{Q_i}\middle|1\le t\le k,1\le j_i,p\le n,1\le i\le m\right\}$$

$$r6'=\left\{s_{h'}A_{tp}G_{ij_i}\to s_0A_{tj_i}G_{ip}\mu\middle|h'>0,1\le t\le k,1\le j_i,p\le n,1\le i\le m\right\}$$

$$\bigcup\left\{s_{h'}\to s_0\nu\middle|h'\le0\right\}\bigcup\left\{\eta_i\to\eta_{i+1}\right\}\bigcup\left\{\eta_{m+1}\to\lambda\right\}$$

$$r7'=\left\{\varsigma_p^{Q_p}\varsigma_{j_i}^{Q_{ji}}\middle|1\le j_i,p\le n\right\}$$

$$r8'=\left\{\mu^i\nu^jG_{1j_1}G_{2j_2}...G_{mj_m}A_{tp}\to(ba_{j_1}a_{j_2}...a_{j_m}A_{tp}\chi_1)_{out}\middle|1\le j_1,j_2,...j_m,p\le n,1\le t\le k\right\}$$

$$\bigcup\left\{\left(\nu^iA_{tp}\to(A_{tp}d)_{out}\right)_{\neg\mu^j}\middle|1\le i,j\le n\right\}$$

$$\rho=\left\{r_i<r_j\middle|1<i<j\le11\right\}\bigcup\left\{r'_i<r_j'\middle|1\le i,j\le8\right\}$$

3.2 The computations in P system

In this section, a comprehensive introduction of the computations and responses in the specific P system designed to realize the ranked K-medoids is presented. We make the clarification in the way the proposed algorithm processing.

At the very beginning, rule r1 is performed accordingly to the priority relationship. It calculates the distance as dissimilarity between any two points and produce object U_{ij} . The multiplicity of object U_{ij} ,that is object ω_{ij} , represents the distance distinctly. Moreover, the subscript of object α_{ij} is utilized to control the cyclic process until the distance between any two points is acquired. Then it executes rule r2, r3to construct the SimRankR(r_{ij}) where $\gamma_{i,j,q}$ means the similarity rank of a_i towards a_j is q. Meanwhile, object ς_{ij}^{q+1} is generated and transferred to membranes labeled from 1 to k.This course ends with yielding an object e to stimulate the rule r4 of selecting the initial k medoids randomly. When it accomplishes the last loop, the object e and θ_{k+1} are converted to χ_1 to impel the rule r5.It performs rule r5 and r6 to select the group of the most preferred objects .Once it completes the loop of selecting the group of the most preferred objects for all medoids, it starts to deliver an object η_1 to each membrane labeled from 1 to k to initiate the computations in the corresponding membrane.

Now, the responses in membranes labeled from 1 to k are activated. Rule $r3'$ is performed to calculate the total preference factors of the medoid A_{tp} by adding the preference factors it holds toward other members in the same group. When responses in membranes labeled from 1 to k are accomplished, the membrane 0 with triggering rules repeats the process of clustering until there is no b in it, which indicates that the medoids in all k membranes remain unchanged and the clustering adjustment process terminates [16]. And an object β is generated to stimulate rules in membranes labeled from 1 to k to output the clustering result.

Finally, the result in the form of string enters the output membrane c_0 . One computation process is accomplished.

3.3 Experiment and analysis

In order to give a better interpretation of our P system model for the ranked K-medoids clustering, we take an example to simulate the procedure of the P system. There are 7 points:

$a_1(1,2), a_2(2,2), a_3(3,0), a_4(3,3), a_5(6,1), a_6(7,1), a_7(8,3)$

The P system is supposed to distribute the points into 2clusters with the given parameter m of the value 3. Diagram 3 depicts the original state of the 7 points.

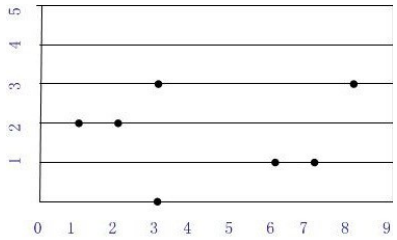

Figure 3: The initial state of the points

First of all, the dissimilarity matrix $D_{7,7}'$ is presented. We calculate the square of the distance between any two points as dissimilarity, thus $D_{7,7}$ is the same to $D_{7,7}'$ .The SimRank

matrix R is constructed by sorting the dissimilarity values between any object a_i and the other objects.

$$D_{7,7} = \begin{pmatrix} 0 & 1 & 8 & 5 & 26 & 37 & 50 \\ 1 & 0 & 5 & 2 & 17 & 26 & 37 \\ 8 & 5 & 0 & 9 & 10 & 17 & 34 \\ 5 & 2 & 9 & 0 & 13 & 20 & 25 \\ 26 & 17 & 10 & 13 & 0 & 1 & 8 \\ 37 & 26 & 17 & 20 & 1 & 0 & 5 \\ 50 & 37 & 34 & 25 & 8 & 5 & 0 \end{pmatrix} \qquad R_{7,7} = \begin{pmatrix} 1 & 2 & 4 & 3 & 5 & 6 & 7 \\ 2 & 1 & 4 & 3 & 5 & 6 & 7 \\ 3 & 2 & 1 & 4 & 5 & 6 & 7 \\ 3 & 2 & 4 & 1 & 5 & 6 & 7 \\ 7 & 6 & 4 & 5 & 1 & 2 & 3 \\ 7 & 6 & 4 & 5 & 2 & 1 & 3 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix} \tag{3.1}$$

Then, the process of the clustering with the designed P system begins.

Eventually, the clustering result was attained that the 7 points were classified into 2 clusters. Consequently, the ranked K-medoids algorithm certified effective. And the clustering effect sketch is shown in Fig.4.

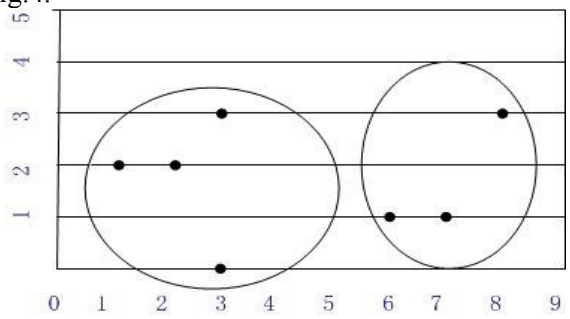

Figure 4: The final clustering result

4. Conclusion

In this paper we propose a specific P system to solve the clustering problem in the framework of the cellular computing with membranes. Tests prove that the P system is adequate to implement the ranked K-medoids algorithm for its high parallelism. However, some questions remain to be discussed furthermore. On the one hand, the feasibility and the efficiency of the model in large database need to be studied. On the other hand, the number of members in a group is determined by a given parameter g which is determined by data experiment. Lastly, it is of great significance to realize other clustering methods by a P system.

References

[1] J. Ponomarenko, T. Merkulova, G. Orlova, O. Fokin, E. Gorshkov, M. Ponomarenko, *Mining DNA sequences to predict sites which mutations cause genetic diseases*, Knowledge-Based Systems. 15 (4) 225–233.

[2] J. Shi, Z. Luo, 2010,*Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples*, Computers in Biology and Medicine 40 (8) 723–732.

[3] D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson, O. Piot, 2011, *Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections,* Laboratory Investigation. 91 (5) 799–811.

[4] P. C. Chang, C. H. Liu, C.Y. Fan, 2009,*Data clustering and fuzzy neural network for sales forecasting: a case study in printed circuit board industry*, Knowledge-Based Systems. 22 (5) 344–355.

[5] E. Hadavandi, H. Shavandi, A. Ghanbari, 2010, *Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting, Knowledge-Based Systems.* 23 (8) 800–808.

[6] J. Han and M. Kambr, 2012,*Data Mining Concepts and Techniques*. USA: Elsevier Inc., ch.8

[7] R. Xu, D. Wunsch, 2005,*Survey of clustering algorithms*, IEEE Transactions on Neural Networks. 16 (3) 645–678.

[8] S. Khan, A Ahmadb, 2004,*Cluster center initialization algorithm for K-means clustering*, Pattern Recognition Letters, 25(1):1293–1302.

[9] Park, H.S., Jun, C.H. 2009,*A simple and fast algorithm for K-medoids clustering*. Expert Systems with Applications. 36(2), 3336–3341.

[10] Q. Zhang, I. Couloigner, 2005, *A new and efficient K-medoid algorithm for spatial clustering*, Computational Science and Its Applications_ICCSA. Springer 207–224.

[11] S. Mohammad ,R. Zadegan, M. Mirzaie, F. Sadoughi, 2013, *Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets*. Knowledge-Based Systems.133–143.

[12] G. Paun, 2010, *A quick introduction to membrane computing*, The Journal of Logic and Algebraic Programing, 79(3):291-294.

[13] G. Paun,G. Rozenberg and A. Salomaa. 2010,*Membrane Computing*. New York：Oxford University press, 282-301.

[14] J.D. Martin-Guerrero, A. Palomares, E. Balaguer-Ballester, E. Soria-Olivas, J. Gomez-Sanchis, A. Soriano-Asensi, 2006,*Studying the feasibility of a recommender in a citizen web portal based on user modeling and clustering algorithms,* Expert Systems with Applications. 30 (2) 299–312.

[15] Y. Zhao, X. Liu, J.Qu,2012,*The K-medoids clustering algorithm by a class of P system*, Information and Computational Science,J.,9(18):215-22.

[16] L. Han, L. Xiang , X. Liu, J. Luan,2014, *The K-medoids Algorithm with Initial Centers Optimized Based on a P System,* Information and Computational Science J., 11(6):132-144.