ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 45-50

Double Self-Organizing Maps to Cluster Gene Expression Data

Dali Wang, Habtom Ressom, Mohamad Musavi, Cristian Domnisoru

University of Maine, Department of Electrical & Computer Engineering,
Intelligent Systems Laboratory, Orono, ME 04469, USA

Abstract

Clustering is a very useful and important technique for analyzing gene expression data. Self-organizing map (SOM) is one of the most useful clustering algorithms. SOM requires the number of clusters to be one of the initialization parameters prior to clustering. However, this information is unavailable in most cases, particularly in gene expression data. Thus, the validation results from SOM are commonly employed to choose the appropriate number of clusters. This approach is very inconvenient and time-consuming.

This paper applies a novel model of SOM, called double self-organizing map (DSOM) to cluster gene expression data. DSOM helps to find the appropriate number of clusters by clearly and visually depicting the appropriate number of clusters. We use DSOM to cluster an artificial data set and two kinds of real gene expression data sets. To validate our results, we employed a novel validation technique, which is known as figure of merit (FOM).

1. Introduction

In order to understand complicated biological systems, researchers have generated a large number of gene expression data sets from the microarray chips. The challenge now is to understand such massive gene expression data. Because of the large amounts of genes and the complexity of biological networks, clustering is a useful technique for gene expression analysis. Many clustering algorithms have been proposed to identify genes with similar expression profiles. These algorithms include k-means clustering [1], hierarchical clustering [2], and self-organizing map [3].

Mangiameli [4] suggested the use of artificial data in evaluating the performance of several different clustering algorithms like SOM and hierarchical clustering. As a result of this evaluation, it was concluded that SOM was superior in both accuracy and robustness.

In SOM, the most important parameter is the number of initial nodes. So far, researchers have little information about the number of nodes required for clustering. Therefore, the results are commonly validated heuristically.

As stated in [5], figure of merit (FOM) is a new validation technique for assessing the results of clustering. This method has been used to compare several clustering algorithms on four gene expression data sets and has been proven as a valuable quantitative measure for validating the cluster's quality. However, it is time-consuming and complicated for finding the best number of clusters.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 45-50

Su [6] introduced a novel model of self-organizing neural networks, named double self-organizing map (DSOM). In this paper, we apply this technique to find the number of clusters in a gene expression data set in an easier and more reliable way. DSOM is used to cluster artificial data as well as real gene expression data. We also use FOM to validate the result from DSOM.

2. Algorithm and Implementation

2.1. Clustering and self-organizing map (SOM)

Since there is a tight connection between genes' functions and their expression patterns, genes can be organized according to the similarities of their expression profiles. The clustering methods are used to cluster similar data points together. This technique has been widely applied to gene expression analysis in terms of classifying genes. SOM is one of the most commonly used clustering algorithms, which is similar to k-means. However, SOM has a degree of self-regulation through connected network of nodes and updates all nodes with a particular neighborhood function. Mangiameli [4] compared the SOM with hierarchical clustering method and found that SOM is superior in both robustness and accuracy.

SOM has some fundamental problems including the difficulty of visualizing high-dimensional data and the uncertainty in the number of clusters in the whole data.

In the literature, there is little knowledge about determining the number of clusters in the gene expression data set before clustering. The only way is to heuristically try different number of clusters and choose the best one.

2.2. Double self-organizing map (DSOM)

In DSOM, as described by Su in [6], each node j not only has an n-dimensional synaptic weight vector w_j but also a 2-dimensional position vector p_j. During the self-organizing process, both the weights and the position vectors are updated. All the updating formulae are given in Equations (1)-(6):

$$w_j(k+1) = w_j(k) + \eta_1(k)\Lambda_{j^*}(k)\left[x(k) - w_j(k)\right] \tag{1}$$

$$p_j(k+1) = p_j(k) + \eta_2(k)h_{j^*}(k)\left[p_{j^*}(k) - p_j(k)\right] \tag{2}$$

where $\eta_1(k) = \eta_w \dfrac{1}{k+1}$ (3)

$$\Lambda_{j^*}(k) = \exp\left[-s_w\left(1+\frac{1}{k+1}\right)\left\|p_j(k) - p_{j^*}(k)\right\|^2\right] \tag{4}$$

$$\eta_2(k) = \frac{\eta_p}{1+k}\exp\left[-s_p\left(1+\frac{k}{k_{\max}}\right)\left\|p_j(k) - p_{j^*}(k)\right\|\right] \tag{5}$$

$$h_{j^*}(k) = \exp\left\{-s_x\left(1+\frac{k}{k_{\max}}\right)\left[\left\|w_j(k) - x(k)\right\| - \left\|w_{j^*}(k) - x(k)\right\|\right]^2\right\} \tag{6}$$

Here, ||•|| denotes the Euclidean distance. η_w and η_p are the initial learning rates. s_w is a scalar parameter which regulates the slope of the function $\Lambda_{j^*}(k)$. $k_{\max}$ is the maximum number of epochs. s_p and s_x are two predetermined scalar parameters which regulate the movement of the position vectors.

Let us assume there are m (actually we don't know the appropriate number of clusters) classes in a data set. Before clustering, we initialize the n nodes and their corresponding position vectors. Here n must be no less than m. Based on these formulae, we can find that the closer the nodes (weights) are, the closer the corresponding position vectors will be. Theoretically speaking, the n position vectors should move into the m groups after clustering.

Because the position vectors are two-dimensional, we can visualize the number of groups of the position vectors by plotting them. Thereby, we are able to determine the number of classes in the gene expression data set.

In this paper, we introduce DSOM to cluster microarray data. We use DSOM to cluster artificial data as well as real gene expression data. Two different kinds of real data were analyzed in our experiments: one in which classification information is known prior to training and the other in which information about classification is unknown. After clustering, we use FOM to validate the result.

2.3. Figure of merit (FOM)

The clusters obtained by different clustering algorithms can be remarkably different. Without validation procedures, results of clustering algorithms may be easily misinterpreted. Figure of merit is an estimate of the predictive power of a clustering algorithm. It is a systematic and quantitative framework to assess the results of clustering algorithms. Yeung [5] introduced FOM to validate clustering performances. They compute FOM for different clustering algorithms so as to solve the problem of choosing an appropriate clustering algorithm for a specific data set.

In the following, we describe FOM as defined by Yeung. A typical gene expression data set contains measurements of expression levels of L genes under B conditions. Assume that a clustering algorithm is applied to the data from condition 1,2,3,..,(e-1), (e+1),…,B and condition e is used to estimate the predictive power of the algorithm. Let there be k clusters, c_1, c_2,…,c_k. Let $R(g,e)$ be the expression level of gene g under condition e in the raw data matrix. Let $Uc_i(e)$ be the average expression level in condition e of genes in clusters c_i. So, the FOM under the condition e is defined as

$$\mathrm{FOM}(e,k)=\sqrt{\frac{1}{L}\sum_{i=1}^{k}\sum_{x\in c_i}\left[R(x,e)-Uc_i(e)\right]^2} \tag{7}$$

And then, the aggregate figure of merit of all conditions is defined:

$$\mathrm{FOM}(k)=\sum_{e=1}^{B} FOM(e,k) \tag{8}$$

After computing FOM for different number of clusters and plotting FOM versus number of clusters, we find that FOM decreases as the number of clusters increases. If a curve enters its saturation region and the corresponding number of cluster at that point is N, we say N nodes are sufficient to cluster the data by using SOM.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 45-50

3. Experiment Data and Results

3.1. Artificial data set

In order to apply DSOM, we create an artificial data set, which has six groups. The centers of the six groups are chosen as shown in Figure 1. Each center is a vector with 17 elements, whose range is between 0 and 1. For every group, 100 random vectors are generated with a variance $v = 0.25$. Hence, each group has 100 sample vectors. The reason for choosing such kind of centers is that the center of the real gene data is similar to one of these centers, or similar to the translation, rotation or combination of these centers.

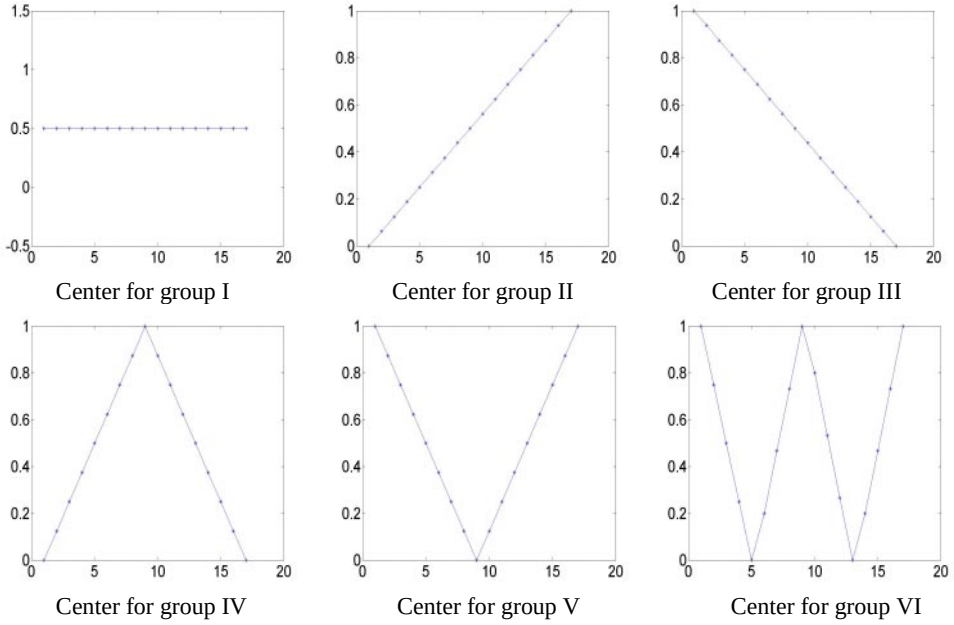


Figure 1: Six centers of the artificial data set. Each center has 17 data points

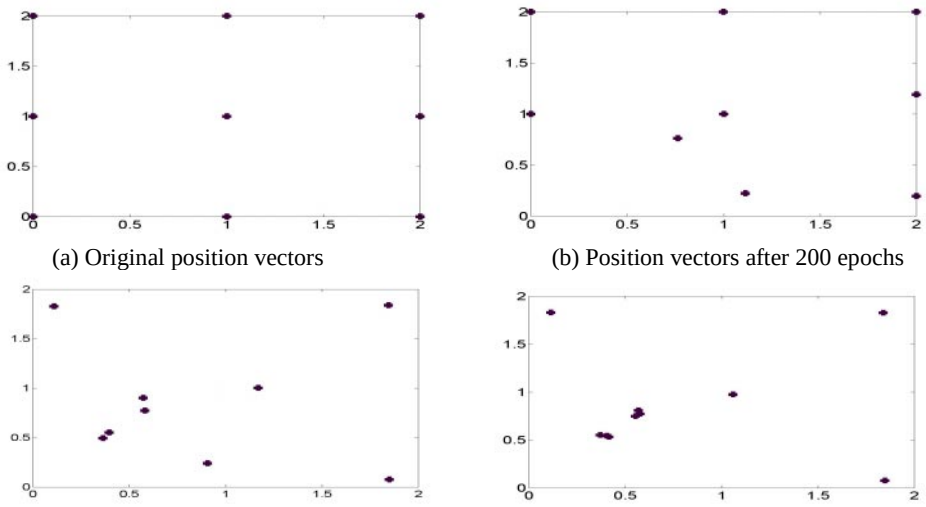
As shown in Figure 2(a), nine original position vectors are placed onto a 3×3 grid. Then the DSOM is run for 1000 epochs. The results are shown in Figures 2(b)-2(d). In Figure 2(d), the nine position vectors fall into the six groups. This shows us that there are six clusters in this artificial data set.

3.2. Real gene expression data set

3.2.1. Yeast data set with known number of clusters

We use some real yeast data, which is classified by the MIPS [7] (the Munich Information Center for Protein Sequences Yeast Genome Database). For details, about the data, check the web-site http://mips.gsf.de/cgi-bin/proj/expression/start.pl.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 45-50



(a) Original position vectors (b) Position vectors after 200 epochs

(c) Position vectors after 500 epochs (d) Position vectors after 1000 epochs

Figure 2**:** Final position vectors after using DSOM for the artificial data set

We cluster this data set using DSOM. As shown in Figure 3, the position vectors of this data set finally fall into four groups. In Figure 3 (left), nine original position vectors are chosen. In Figure 3 (right), four original position vectors are chosen. The results confirm that the number of clusters in this real data set is four.

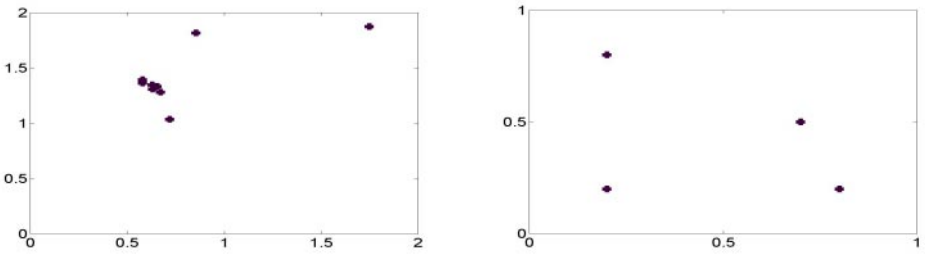


Figure 3. Final position vectors after using DSOM for gene data set, which has 4 groups

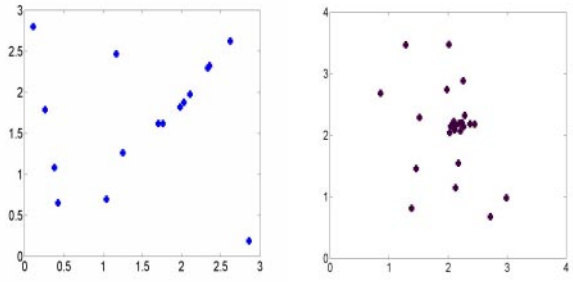


Figure 4: Final position vectors after using DSOM for gene data set with unknown clusters

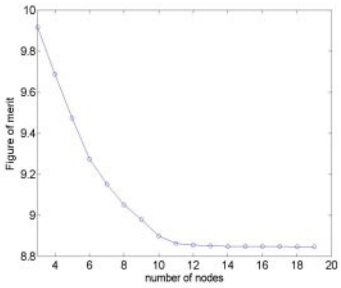


Figure 5: Validation result using FOM

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 45-50

3.2.2 Yeast data set with unknown number of clusters

In this section, we use a yeast data set from [8]. 4×4 and 5×5 original position vectors are initially chosen. Figure 4 shows that, in both cases, all the position vectors fall into nearly 13 groups. As shown in Figure 5, when the number reaches 12-13, the FOM goes into its saturation section. That means, 12-13 is the appropriate number of clusters in the whole data set.

4. Conclusion

In this paper, we use double self-organizing map (DSOM) to cluster the artificial data and gene expression data. Based on the results, it is clear that DSOM can not only cluster the gene expression data very well, but also can efficiently and effectively reveal the appropriate number of classes in those data sets based on the final location of the position vectors. We also use the figure of merit (FOM) to validate our results. It is observed that the number of classes thus obtained using DSOM is quite comparable to the validation values from FOM.

References

1. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, G. M. Church: Systematics determination of genetic network architecture. Nature Genetics, 22, 281-285 (1999).
2. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein: Cluster analysis and display of genome-wide expression patterns. PNAS USA, 95, 14863-14868, (1998).
3. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, E. Kitareewan, E. Dmitrovsky, E. Lander, T. Golub: Interpreting patterns of gene expression with self-organizing maps, method and application to hematopoietic differentiation. PNAS USA, 96, 2907-2912 (1999).
4. P. Mangiameli, S. K. Chen, D. West: A comparison of SOM of neural network and hierarchical methods. European Journal of Operational Research, 93, 402-417 (1996).
5. K. Y. Yueng, D. R. Haynor, W. L. Ruzzo: Validating clustering for gene expression data. Bioinformatics, 17, 309-318 (2001).
6. M. Su, H. Chang: A new model of self-organizing Neural Networks and its application in Data Projection. IEEE transactions on Neural Networks, 12, 153-158 (2001).
7. H. W. Mewes, K. Albermann, K. Heumann, S. Liebl, F. Pfeiffer: MIPS, a database for protein sequences, homology data and yeast genome information. Nucleic Acid Research, 25, 28-30 (1997).
8. M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, M. Ares, D. Haussler: Knowledge-based Analysis of Microarray Gene Expression Data Using Support Vector Machines. PNAS USA, 97, 262-267 (1996).