

A Novel OPTOC-based Clustering Algorithm for Gene Expression Data Analysis

Alan Wee-Chung Liew¹, Hong Yan^{1,2} and Shuanhu Wu³

¹Department of Computer Engineering and Information Engineering
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

²School of Electrical and Information Engineering
University of Sydney, NSW 2006, Australia

³Information School of WuYi University
JiangMen 529020, GuangDong, China
Email: {itwcliew, ityan}@cityu.edu.hk

Abstract

Cluster analysis of gene expression data is useful for identifying biologically relevant groups of genes. However, finding the correct clusters in the data and estimating the correct number of clusters are still two largely unsolved problems. In this paper, we propose a new clustering framework that is able to address both these problems. By using the one-prototype-take-one-cluster (OPTOC) competitive learning paradigm, the proposed algorithm can find natural clusters in the input data, and the clustering solution is not sensitive to initialization. In order to estimate the number of distinct clusters in the data, an over-clustering and merging strategy is proposed. For validation, we applied the new algorithm to both simulated gene expression data and real gene expression data (expression changes during yeast cell cycle). The results clearly indicate the effectiveness of our method.

1. Introduction

Advances in the DNA microarray technology have enabled biologists to monitor thousands of genes simultaneously and measure the whole-genome mRNA abundance in the cellular process under various experimental conditions [1-3]. A large amount of gene expression profile data has become available in several databases. The challenge now is to make sense of such massive data sets and this requires the development of powerful data analysis tools.

A crucial step in the analysis of gene expression data is the detection of gene groupings that manifest similar expression patterns [4-8]. Most current methods for gene expression data analysis rely on the use of clustering algorithms [9-12]. The fundamental biological premise underlying these approaches is that genes that display similar expression patterns are co-regulated and may share a common function.

Recently, a new competitive learning paradigm, called the one-prototype-take-one-cluster (OPTOC), has been proposed [13]. In conventional competitive learning, if the number of clusters is less than the natural clusters in the

data, at least one of the prototypes would win data from more than one cluster. In contrast, OPTOC would win data from only one cluster, while ignoring the data from other clusters. The OPTOC based learning strategy has the following two main advantages: (1) It can find natural clusters, and (2) The final partition of the dataset is not sensitive to initialization.

In this paper, we propose a new clustering framework based on the OPTOC learning paradigm for clustering gene expression data. The new algorithm is able to identify natural clusters in the dataset as well as provides a reliable estimate of the number of distinct clusters in the dataset.

2. The OPTOC Framework

In cluster analysis, we are generally interested in identifying regions of high data concentration in the data space. These regions of high data concentration form natural clusters in the dataset. Ideally, a clustering algorithm should be able to determine the number of natural clusters and their locations in the data space automatically.

However, most conventional clustering algorithms require the prior specification of the correct number of clusters. In conventional clustering, if the number of prototypes is less than that of the natural clusters in the dataset, there must be at least one prototype that wins patterns from more than two natural clusters. Even if the correct number of clusters is given, there is no guarantee that the clusters found do correspond to the natural clusters in the dataset. The implications of not finding natural clusters are that: (i) a natural cluster might be erroneously divided into two or more classes, or worst still, (ii) several natural clusters or part of them are erroneously group into one class. Such behaviors obviously lead to wrong inferences about the data.

In contrast, the one-prototype-take-one-cluster (OPTOC) idea [13] allows one prototype to characterize only one natural cluster in the dataset, regardless of the true number of clusters in the data. The OPTOC clustering framework is

achieved by the introduction of a dynamic neighborhood A_i for each prototype P_i , such that patterns (i.e., data) inside the neighborhood of P_i contribute more to its learning than those outside. Given an input pattern x , and assume that P_i is the winning prototype for x based on the minimum distance criterion, the neighborhood A_i is updated by

$$A_i^* = A_i + (x - A_i) \bullet \Theta(P_i, A_i, x) \bullet \delta_i / n_{A_i} \quad (1)$$

where Θ is a switching function given by

$$\Theta(\mu, v, \omega) = \begin{cases} 1 & \text{if } |\mu v| \geq |\mu \omega| \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

and

$$\delta_i = \left(\frac{|P_i A_i|}{|P_i A_i| + |P_i x|} \right)^2, \quad 0 < \delta_i \leq 1 \quad (3)$$

$$n_{A_i}^* = n_{A_i} + \delta_i \bullet \Theta(P_i, A_i, x) \quad (4)$$

with n_{A_i} initialized to zero. Then, the winning prototype P_i is updated by

$$P_i^* = P_i + (x - P_i) \bullet \delta_i \quad (5)$$

We can see from the above equations that if x is well outside the neighborhood of P_i , i.e., $|P_i x| \gg |P_i A_i|$, it would have very little influence on the learning of P_i . On the other hand, if x is well inside the neighborhood of P_i , i.e., $|P_i x| \ll |P_i A_i|$, both A_i and P_i would shift toward x according to (1) and (5), and P_i would have a large learning rate δ_i .

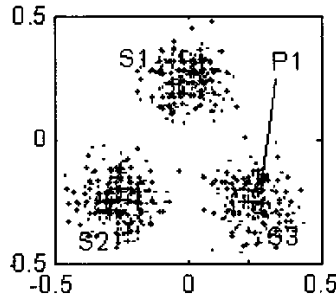


Fig.1: One prototype takes one cluster (OPTOC) and ignores the other two clusters.

During learning, the neighborhood $|P_i A_i|$ will decrease monotonically. When $|P_i A_i|$ is less than a tiny quantity ε , P_i would eventually settle at the center of a natural cluster and the learning stops. Thus, each prototype will locate only one natural cluster and ignore other clusters. Fig.1 shows an example of learning based on the OPTOC paradigm. In this figure, P1 finally settles at the center of S3 and ignores the other two clusters S1 and S2.

3. Self-Splitting and Merging Competitive Clustering

When the number of clusters in the input space is more than one, additional prototype needs to be generated to search for the remaining clusters. Let C_i denotes the center of all the patterns that P_i wins. The distortion $|P_i C_i|$ measures the discrepancy between the prototype P_i found by the OPTOC learning process and the actual cluster structure in the dataset. For example, in Fig.1, C_i would be located at the center of the three clusters S1, S2 and S3 (since there is only one prototype, it wins all input patterns), while P_i eventually settled at the center of S3. After the prototypes have all settled down, a large $|P_i C_i|$ indicates the present of other natural clusters in the data. A new prototype would be generated from the prototype with the largest distortion when this distortion exceeds a certain threshold. Ideally, if a suitable threshold can be given, the cluster splitting process would terminate when all natural clusters in the dataset are found. Unfortunately, due to the high dimension and the complex structure exhibit by the gene expression data, the determination of a suitable threshold to find all natural clusters is very difficult in practice.

In order not to miss any natural cluster in the data, we over-cluster the dataset. After each OPTOC learning, the cluster with the largest variance is split, until the desired number of clusters is reached. When cluster splitting occurs, the new prototype is initialized far away from its mother prototype to avoid unnecessary competition between the two. The location of the possible split, R_i , is also learned dynamically. Initially, R_i is set to be equal to the prototype to which it is associated with. Then, each time a new pattern x is presented, the R_i of the winning prototype P_i is updated as follows,

$$R_i^* = R_i + (x - R_i) \bullet \Theta(P_i, x, R_i) \bullet \rho_i / n_{R_i} \quad (6)$$

where

$$\rho_i = \left(\frac{|P_i x|}{|P_i R_i| + |P_i x|} \right)^2 \quad (7)$$

Note that R_i always try to move away from P_i . After a successful split, (A_i, R_i) of every prototype P_i are reset and the OPTOC learning loop is started again.

With over-clustering, it is possible that a natural cluster in the dataset is split into two or more clusters. Thus, some clusters would be visually similar and should be merged together. The aim of the merging is to produce final clustering result in which all clusters have distinct patterns.

Let us assume that the clusters in a dataset have Gaussian distributions, and that the probability density function (pdf) of a distinct cluster is unimodal. If two clusters are well separated, their joint pdf would be bimodal. When two clusters are close to each other to the extent that their joint pdf form a unimodal structure, then it would be reasonable to merge these two clusters into one. Let C_i be the centers of cluster i and σ_i be its standard deviation. If

$$\|C_i - C_j\| \leq \frac{1}{2}(\sigma_i + \sigma_j) \quad (8)$$

the two clusters should be merged into one. When two clusters are merged into one, the mean and standard deviation of the merged cluster is re-calculated. The merging process is repeated, until no more clusters can be merged together.

4. Experiments

In this section, we verify the performance of the proposed SSMCL algorithm using both simulated and real expression data. We first use simulated gene expression profiles, where the correct solution was known a priori, to validate the effectiveness of our algorithm in finding natural clusters and the correct number of clusters. Then we validate the algorithm by clustering the yeast cell cycle data set provided by Cho et al [14] and examine the biological relevance of the clustering results.

4.1 Simulated Data

We randomly generated 20 seed patterns of gene expressions with 15 time points each. Then each pattern was transformed into a cluster by generating many profiles from the pattern. Each cluster contains 30 to 165 profiles, with the total number of profiles in the dataset equal to 1785. For each cluster, the data along each time point k were set to have a standard deviation of 0.15. The simulated data is shown in Fig.2.

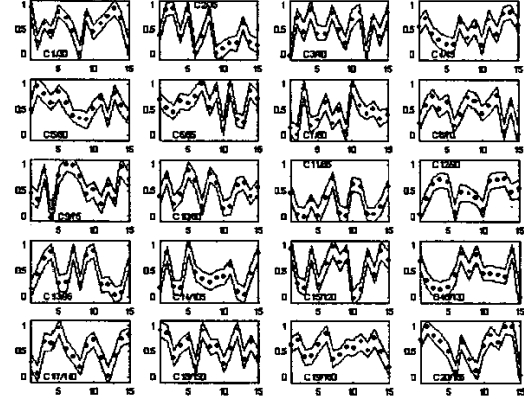


Fig.2: 1785 randomly generated temporal patterns of gene expression grouped in 20 clusters. Each cluster is represented by the average profile pattern in the cluster (dot line). Solid lines indicate the one standard deviation levels of each expression about the mean. C_m/n denotes cluster $\#m$ containing n individual profiles.

We want to verify that the OPTOC clustering framework can find all the natural clusters in the simulated dataset, independent of initialization. The splitting is stopped when 20 clusters have been generated. Fig.3 shows the clustering results. We found that the proposed OPTOC based algorithm was successful in finding all the natural clusters.

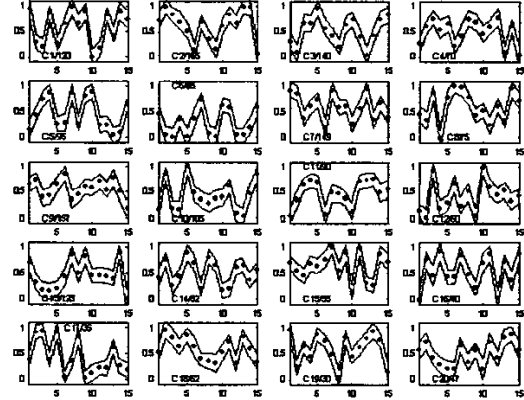


Fig.3: Clustering result for the 1785 randomly generated temporal patterns of gene expression.

In the next experiment, we find out whether our over-clustering and merging strategy can merge similar clusters and stop at the exact number of clusters automatically, when the exact number of clusters in the data is not known. We set the number of clusters to 28. After 28 clusters are obtained, cluster merging is performed. The cluster merging process stopped automatically when exactly 20 clusters were found and the results are shown in Fig.4. Correct clustering of the data is also obtained.

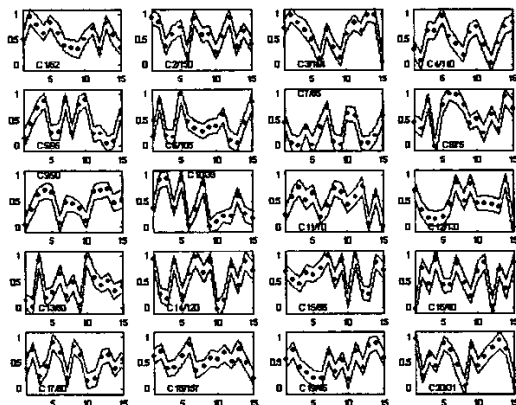


Fig.4: The final clustering result after over-clustering and merging.

4.2 Biological Validation: Yeast Cell Cycle Data

The yeast cell cycle data set has established itself as a standard for the assessment of newly developed clustering algorithm. This data set contains 6601 genes, with 17 time points for each gene taken at 10-min intervals covering nearly two yeast cell cycles (160 min) [5]. The raw expression profiles are downloaded from <http://genomics.stanford.edu>. Firstly, we eliminate those genes whose expression levels were relatively low and did not show significant changes during the entire time course by a variation filter with criteria: (a) the value of expression profile at all 17 time points is equal to or greater than 100 (raw data units); (b) the ratio of the maximum and the minimum of each time-course expression profiles is at least equal to or greater than 2.5. 1368 gene expression profiles passed the variation filter and were normalized to be between 0 and 1.

Fig.5 shows the resulting 22 clusters after over-clustering (number of clusters set to 30) and merging. The result shows no apparent visual similarity between clusters. We also checked the resulting 22 clusters using biological knowledge. We used gene expression data from the study of Cho et al [14], where 416 genes have been interpreted biologically and 110 genes passed our filter. Those gene expression profiles include five fundamental patterns that correspond to five cell cycles phases: early G1, late G1, S, G2, and M phase. In Fig.6, we show the five clusters that contain most of the genes belonging to these five different patterns. It is obvious that these five clusters correspond to the five cell cycle phases

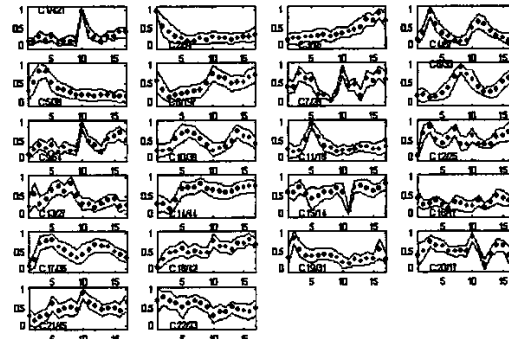


Fig.5: The final clustering results for the yeast cell cycle data. 22 distinct clusters are obtained.

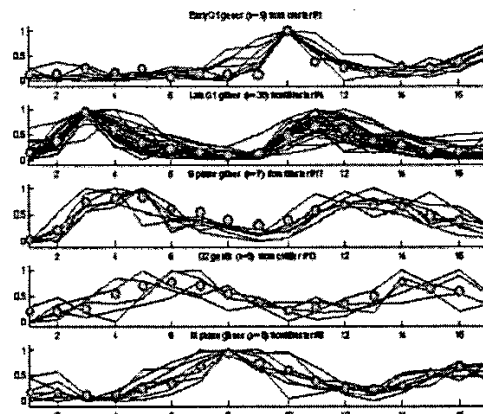


Fig.6: Five fundamental patterns taken from Fig.5 that correspond to the five cycle phases. On each subplot, filled circles represent the average pattern for all profiles in the cluster. The genes presented are only those that belong to this cluster and are biologically characterized and assigned to a specific cell cycle phase [14].

5. Conclusion

Cluster analysis is an important tool in gene expression data analysis. In this paper, we have described a new clustering algorithm that is able to identify the natural clusters, and to estimate the correct number of clusters, in a dataset in a systematic way. The ability to finding natural clusters in a dataset is based on the OPTOC paradigm, which allows one prototype to characterize only one natural cluster in the dataset, regardless of the number of clusters in the data. In order to correctly estimate the number of natural clusters in a dataset, we proposed an over-clustering and merging strategy. The over-clustering step minimizes the chance of missing any natural clusters in the data, while the merging step ensures that the final clusters are all visually distinct from each other. The effectiveness of the algorithm is verified by clustering simulated gene expressions data and real gene expressions

profile data for which the biological relevance of the results is known.

Acknowledgment

This work is supported by a CityU SRG grant (project 7001183) and an interdisciplinary research grant (project 9010003).

References

- [1] D.J. Lockhart and E.A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature*, Vol. 405, pp. 827-836, June 2000.
- [2] D. Shalon, S.J. Smith and P.O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Res.*, Vol. 6, pp. 639-645, 1996.
- [3] R.A. Young, "Biomedical discovery with DNA arrays," *Cell*, Vol. 102, pp. 9-15, Jan. 2000.
- [4] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 14863-14868, Dec. 1998.
- [5] P.T. Spellman, G. Sherlock, M.O. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol. Biol. Cell*, Vol. 9, pp. 3273-3297, 1998.
- [6] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 2907-2912, Mar. 1999.
- [7] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays" *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 6745-6750, 1999.
- [8] G. Zweiger, "Knowledge discovery in gene-expression microarray data: mining the information output of the genome," *Trends Biotechnol.*, Vol. 17, pp. 429-436, 1999.
- [9] J. Hartigan, *Clustering algorithms*. New York: Wiley, 1975.
- [10] A. Jain, and R. Dubes, *Algorithms for data clustering*. Englewood Cliffs, N.J.: Prentice Hall, 1988.
- [11] T. Kohonen, *Self-Organizing Maps*. Berlin, New York: Springer, 2001.
- [12] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, "Optimization by simulated annealing," *Science*, Vol. 220, pp. 671-680, 1983.
- [13] Y.J. Zhang, and Z.Q. Liu, "Self-Splitting competitive learning: A new on-line clustering paradigm," *IEEE Trans. on neural networks*, Vol. 13, pp. 369-380, 2002.
- [14] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart and R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, Vol. 2, pp. 65-73, 1998.