Application of Geometrical Learning for Similarity Index in Clustering DNA Microarray Data

Wenming Cao[1,2], Shoujue Wang [2]

[1]The College of Information Engineering, Zhejiang
University of Technology,
Hangzhou, 310014, China
[2]Lab of Artificial Neural Networks, Institute of
Semiconductors, CAS, Beijing, 100083, China
csann@zjut.edu.cn

Abstract

In this paper, we train Geometrical learning to classify a gene expression pattern as being similar to a pre-specified one ("target"), and therefore create a neural network-based proximity measure. We use the assessed proximity measures in the simple threshold clustering algorithm to verify whether the Geometrical learning-based measure could result in clusters that resembled those formed using functional categories and common regulatory motifs. Finally, we compare the results with other proximity measures for Saccharomyces cerevisiae gene expression data. We show that the clusters obtained using Euclidean distance, correlation coefficients, and mutual information were not significantly different. The clusters formed with the Geometrical learning -based index were more in agreement with those defined by functional categories and common regulatory motifs.

Keyword: Geometrical learning, neural network, DNA Microarray Data

I. Introduction

DNA microarray technology has enabled the study of large-scale gene expression data. A number of analytical methodologies have been introduced to analyze gene expression patterns, and cluster analysis [1] has played a prominent role. Cluster analysis usually requires two steps. The first step is to measure the relations (e.g., distance or similarity) of gene expression by a pre-specified measure, in a pairwise fashion. The second step is to cluster the genes based on the measures derived in the first step. In this paper, we trained Geometrical learning[2,3,4,5,6,7] to classify a gene expression pattern as being similar to a pre-specified one ("target"), and therefore created a neural network-based proximity measure. We used the assessed proximity measures in a simple threshold clustering algorithm to verify whether the Geometrical learning-based measure could result in clusters that resembled those formed using functional categories and common regulatory motifs. We also evaluated clusters derived from Euclidean distance, correlation coefficients, and mutual information.

II. Materials and methods

We used S. cerevisiae gene expression data consisting of 79 measurements of 2467 genes (http://genome-www4.stanford.edu /MicroArray/ SMD/ index. html). The array design is available from that web site. Briefly, the expression data were obtained during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks [8].

A. *Euclidean distance (Normalized Vector)*

The Euclidean distance was calculated based on the normalized dispersion in expression level of each gene across the measurement points (s.d./mean). This normalized expression level is obtained by subtracting the mean across the measurement points from the expression level of each gene, and dividing the result by the standard deviation across the time points:

$$d(x,\overline{x}) = \| x - \overline{x} \| = \frac{| x - \overline{x} |}{1/N\sqrt{\sum_{1}^{N}(x-\overline{x})^2}} \quad (1)$$

where x is the normalized expression level, x is a vector of expression data of a series of N conditions in gene X, and $\overline{x}$ is the mean of X.

B. *Correlation coefficient*

The correlation coefficient was calculated as[9]

$$s(x,y) = \frac{1}{N}\sum_{i=1,N}\left(\frac{x_i - x_{offset}}{\Phi_x}\right)\left(\frac{y_i - y_{offset}}{\Phi_y}\right) \quad (2)$$

Where $\Phi_G = \sqrt{\sum_{i=1,N}\frac{(G_i - G_{offset})^2}{N}}$, G_i is the log-transformed primary data for gene G in condition i, G_{offset} is the mean of observations on G , and X,Y are the vectors of expression data of a series of N conditions for gene X, and Y, respectively.

C. *Mutual information*

The mutual information between gene X and gene Y was calculated as [10]

$$MI(x,y) = H(x) + h(y) - H(x,y) \quad (3)$$

$$H(x) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i)) \quad (4)$$

where $H(x) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i))$. $H(X)$ is the entropy of a gene expression pattern of a series of N conditions in gene X, and $H(X,Y)$ is the joint entropy of genes X and Y defined as[10]

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{N} p(x_i, y_j)\log_2(p(x_i, x_{ji}) \quad (5)$$

where $p(x,y)$ is the joint probability of X and Y.

Note that the number of distinct expression patterns in a neighbor list increases as the threshold becomes large and vice versa. We counted the number of distinct gene expressions in a neighbor gene list at different proximity thresholds. The thresholds were selected so that the number of distinct expression patterns ranged from 50 to 300 by intervals of 10.

D. *Simple threshold clustering: neighbor gene lists*

We applied a simple threshold clustering method to the data after the calculation of the proximity/distance between all pairs of two gene expressions by each of the four measures. For each gene expression pattern, we generated a list of neighbors. In this context, "neighbor" means the genes that have similar expression patterns and it does not mean either physical location of the genes in the genome or similarity of DNA sequences. An element of the neighbor gene list is a set that contains genes within a proximity threshold (radius) from a reference gene (center) (Fig. 2). In other words, the genes in a set show similar expression patterns to the reference gene expression pattern and the degree of similarity is within the proximity threshold from the reference gene expression pattern.

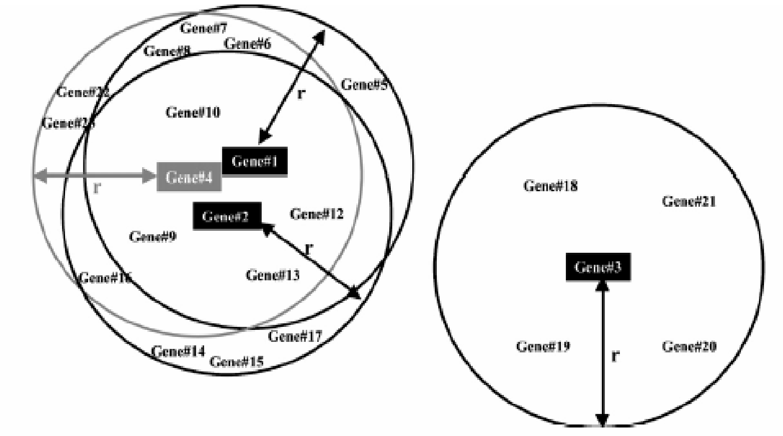

Fig. 1. Simple threshold [r] clustering. A sample of a neighbor gene list in a two-dimensional scheme. The real dimensionality is 79 (the number of measurement points). A neighbor gene list consists of n sets. A set contains gene expression patterns within a threshold distance [radius] from a reference gene expression pattern [center].

E. *Geometrical learning*

Authors are responsible for the accuracy and completeness of the content including the reference lists. Number citations consecutively in square brackets[1]. The sentence punctuation follows the brackets [2]. Numbered references should appear at the end of the article and should consist of the surnames and initials of authors, title of article, name of journal, year, volume, first and last page numbers.

$$f_{GL}(X)=\operatorname{sgn}\left[2^{\frac{d^2(X.\overline{X_1X_2})}{r^2}}-0.5\right] \quad (6)$$

which contains a radius parameter r and the distance between X and the line segment $\overline{X_1X_2}$ as follows:

$$d^2(X,\overline{X_1X_2})=\begin{cases}\|X-X_1\|^2,q(X,X_1,X_2)<0\\ \|X-X_1\|^2,q(X,X_1,X_2)>\|X_1-X_2\|\\ \|X-X_1\|^2-q^2(X,X_1,X_2),otherwise\end{cases} \quad (7)$$

Where $q(X,X_1,X_2)=<(X-X_1),\frac{(X_1-X_2)}{\|X_1-X_2\|}>$ Given a ordered set of expression pattern $P=\{x_i\}_{i=j}^{n}$.

We select a parameter D,the distance between the two contiguous selected expression pattern in S. From P we choose a set $S\{s_i\mid d(s_{i+1},s_i)\approx D,1\le i<m\}$ of n_j expression pattern support points as the sausage parameters $\{X_{j1},X_{j2}\}_{j=i}^{n}$ defined by (7)

Algorithm:

Let S denote the filtered set that contains the expression pattern which determine the network and X denote the original set that contains all the expression pattern in the order.

Begin

1.Put the first expression pattern into the result set S and let it be the fiducial expression pattern s_b, and the distance between the others and it will be compared. Set $S=\{s_b\}$. $s_{\max}=s_b$ and $d_{\max}=0$

2. If no expression pattern in the original set X ,stop filtering. Otherwise , check the next expression pattern in X , then compute its distance to s_b ,i.e., $d=\|s-s_b\|$.

3. If $d>d_{\max}$,goto step 6. Otherwise continue to step 4.

4. If $d < \varepsilon$,set $s_{\max} = s$, $d_{\max} = d$, goto step 2. Otherwise continue to step 5.

5. Put s into the result set: $S = S \cup \{s\}$,and let $s_b = s$, $s_{\max} = s$, and $d_{\max} = d$. Then go to step 2.

6. If $d_{\max} - d > \varepsilon_2$, go to step 2. Otherwise put $s_{\max}$ into the result set: $S = S \cup \{s_{\max}\}$,and let $s_b = s_{\max}$, $d_{\max} = \|s - s_{\max}\|$ go to step2.

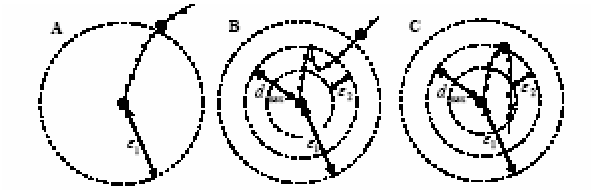

Fig. 2. The three case of the algorithm

The above algorithm constructs the one dimensional topological framework of the HSN. The other key task is to determine r, the radius of the hyper-sphere moving along the framework. As before, we use S, a subset of X, to construct the framework.

III. Results and Discussion

Using *Saccharomyces cerevisiae* gene expression data, we compared the performances of the Geometrical learning -based similarity index and other similarity (or distance) measures in forming clusters that resemble those defined by functional categories or the presence of common regulatory motifs. A simple clustering method based on similarity thresholds was used for comparison. The clusters obtained using Euclidean distance, correlation coefficients, and mutual information were not significantly different. The clusters formed with the neural network-based index were more in agreement with those defined by functional categories or common regulatory motifs. Non-linear similarity measures such as the one proposed may play a role in complex microarray data analysis. Further studies are necessary to demonstrate their applicability beyond this "proof-of-concept" experiment.

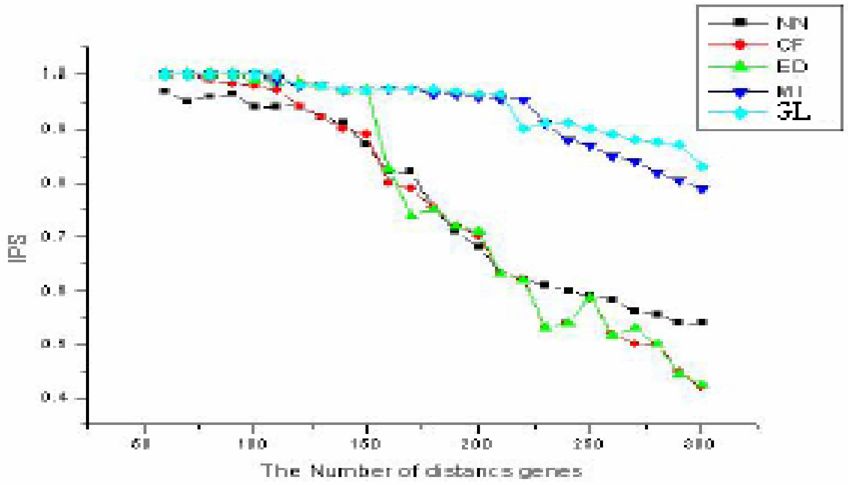

Fig.3. Comparison of performance based on motifs. On the x-axis: proximity thresholds are represented by the number of distinct expression patterns. On the y-axis:the performance in integrated performance score (IPS) is shown. The Geometrical learning-based measure was superior in all IPS scores (p 0:01 by the modified log-rank test). There was no significant statistical difference in the IPSs of the neighbor gene lists produced using neural networks (NN), correlation coefficients (CF), Euclidean distance (ED), mutual information (MI), and Geometrical learning.

IV. Conclusion

In this study, we developed a Geometrical learning-based measure of gene expression proximity and evaluated its performance in a single dataset. The clusters were based on simple proximity threshold cutoffs. We test these measures on data consisting of 79 measurements from 10 different experimental conditions. The cluster performance was evaluated based on the motif DNA sequences and MIPS functional categories. The performance is compared statistically. There is no significant difference among results obtained using Euclidean distance, correlation coefficients, mutual information and Geometrical learning. The performance of the geometrical measure is significantly different. Non-linear proximity geometrical learning methods such as the one derived from high dimension space may play a role in the analysis of gene expression data.

References

[1] M.S. Aldenderfer and R.K. Blashfield : Cluster Analysis, Sage. Newbury Park, CA (1984).

[2] Shoujue Wang: A New Development on ANN in China - Biomimetic Pattern Recognition and Multi Weight Vector Neurons. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 2639: 35-43 2003

[3] Wenming Cao, Feng Hao, Shoujue Wang: The application of DBF neural networks for object recognition. Information Science 160(1-4): 153-160 (2004)

[4] Wenming Cao, Shoujue Wang: Study of Adaptive Equalizers Based on Two Weighted Neural Networks. CIT 2004: 612-615

[5] S.Amari, Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements. IEEE Trans. Computers, Vol.C-21, No. 11, pp.1197-1206, November 1972.

[6] Wang Shoujue. A new development on ANN in China - Biomimetic pattern recognition and multi weight vector neurons, LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 2639: 35-43 2003

[7] Wang Shoujue,etc. Multi Camera Human Face Personal Identification System Based on Biomimetic pattern recognition ,Acta Electronica Sinica 2003,31(1): 1-3

[8] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein Proc. Natl. Acad. Sci. USA 95 (1998), pp. 14863–14868.

[9] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church Nat. Genet. 22 (1999), pp. 281–285.

[10] R. Herwig etc. Genome Res. 9 (1999), pp. 1093–1105.

[11] Wang Shoujue,etc. Geometrical learning, descriptive geometry, and biomimetic pattern recognition , NEUROCOMPUTING 67: 9-28 AUG 2005

Acknowledgements: This work was supported by the National Natural Science Foundation of China Grant No. 60474703

Wenming Cao was born in China, 1965. He received the B.Sc.degree in Maths form Jiangsu Normal College, and the M. Sc. degree in Operation Research from institute of system science of Chinese Acad Sci, China in 1984 and 1991 respectively. He received the Ph.D. degree in Control theory in 2003 from Automatic Department, Southeast University,Jiangsu, China. Since July 2003,He has been with the AI, Zhejiang University of Technology, China where he is currently a professor of Artificial Intelligent Dr. Cao interests in Artificial Intelligent, machine learning, and their applications in control systems, fault diagnosis and medical engineering and bioinformatics.

Prof. Wang Shoujue, specialist on information science, was born in June 1925 in Shanghai. He graduated from the Department of Electrical Engineering, Tong Ji University in 1949. After graduation, he served as Research Assistant in the Institute of Radium, Peking Academy in Shanghai. He joined the Institute of Applied Physics, Chinese Academy of Sciences in Beijing in 1956, and served as director of the Semiconductor Devices Laboratory since the founding of the Semiconductor Institute, CAS in 1960. He became deputy director and director of the Semiconductor Institute, CAS in 1977 and 1983 respectively. He was elected the Member of Chinese Academy of Sciences in 1980. He is now a research fellow in the Semiconductor Institute, CAS, in charge of the research laboratory on the Artificial Neural Networks and Machine Thinking in Image. He is also a guest Professor at the Tong Ji University in Shanghai and the Zhejiang University of Technology in Hangzhou in China. Professor Wang developed the first domestic made high frequency switching transistor in 1958, for using in the early high speed transistorized computer, for computing in research works on the nuclear physics. He also developed the earliest silicon planar transistors and solid state circuits in China in 1963 and 1965, the Pattern Generator for LSI mask making in 1971. By using his invention on the Integrated High Speed Fuzzy Logic Circuit DYL, published in 1978, he reduced the converting time of 8 bit D/A converter chips form 80 ns to 4 ns. In the last decade of the twentieth century, he worked and got significant achievements on Artificial Neural Networks, including chips, hardware, mathematical models, algorithms and applications in pattern recognition as well as process optimization. He proposed (2002) a new model of pattern recognition principles, witch is based on “matter cognition” instead of “matter classification” in traditional statistical pattern recognition, named point-set covering recognition or topological recognition. This new model is better closer to the function of human being, therefore it also be called Biomimetic Pattern Recognition (BPR).In 2002 he developed a new tool named “Computational Descriptive Geometry in High Dimensional Space” for analyzing the point-set covering problem. For hardware implementation of Neural Networks to solving the Point-set Covering Problem in high dimensional space, he created the Multi-weighted Neural Networks in 2001.Professor Wang received the first-class award of National New Products Prize in 1964, the first-class as well as the second-class award of Scientific and Technological Achievement Prize from CAS in 1980, 1983, 1992, and 1996, the first-class award of Advanced Science and Technology Prize Beijing in 2001. He also has received twice the National Invention Prize in 1964 and again in 1996. He was given the Award of Progress in Science and Technology in 2001.