A Verification of Normalization Results using Variable Clustering Methods in cDNA Microarray Data

[†]Gyeongdong Baek, [†]Yountae Kim, [††]Ji-Young Kim, [†]Sungshin Kim, and [††]Choon-Hwan Lee

[†]School of Electrical and Computer Engineering, Pusan National University, Busan, Korea
Pusan National University 30 Changjeon-dong, Keumjeong-ku, Busan 609-735, Korea
E-mail: gdbaek@pusan.ac.kr, dream0561@pusan.ac.kr, sskim@pusan.ac.kr

[††]School of Molecular Biology, Pusan National University, Busan, Korea
Pusan National University 30 Changjeon-dong, Keumjeong-ku, Busan 609-735, Korea
E-mail: dryoung82@pusan.ac.kr, chlee@pusan.ac.kr

Abstract – cDNA microarray analysis has enabled the measurement of thousands of gene expressions at the same time. The gene expression levels are monitored using log ratios between green and red fluorescent intensities. However, imbalances can be caused by different incorporations of dyes, amounts of mRNA, and scanning parameters and these biases result in incorrect conclusions. Normalization makes gene expression data more accurate by removing these systematic variations. Therefore, the study of normalization is important for clustering and also profitable by making groups that show similar expression patterns. We tried to certify the results of normalization by comparing the operation time of *k*-means and fuzzy c-means clustering methods. When it takes less time for *k*-means clustering and more time for fuzzy c-means clustering relatively, we can say the result of the normalization is good. In addition, we analyzed characteristics of standard normalization using two clustering methods. These two methods will be used as the verification methods of any normalization in cDNA microarray.

Keywords: fuzzy clustering, fuzzy c-means algorithm, *k*-means clustering, microarray, normalization

1. Introduction

The latest advancements in genetics has made many things possible [1]. They ascertain the facts how the gene expressed in body and what to do. cDNA microarray maps and sequences all the genes on a small chip. It looks over gene expression patterns. For example, a disease happens because of an interaction between genes and not simply because of one gene.

cDNA microarray is useful to observe the whole expression pattern in this case[2]. The cDNA microarray data needs normalization steps before classification steps are possible [3]. Because there are more noise than any experimentation. The difference the physical properties of two dyes accounts for that noise. This noise is the exact intensity of temperature or fluorescence, sometimes the green dye seems to have high fluorescence intensity. It doesn`t have the same dye intensity for this reason [4]. There are other problems which the efficiency of dye incorporation, data collection, the data scanning process, the difference between pin-groups and slide heterogeneity. All of these problems generate noise. The quality of data is poorer as the noise adds up. The noise differs in degree due to an experimenter`s skill or chemical materials used. The purpose of normalization is to correct errors in the patterns among the kinds of noise. cDNA microarray normalization means the revision of fluorescence intensity and the comparison of gene expression level through experiments or slides. cDNA microarray normalization is divided into three divisions of selection methods. In the first method, let us suppose that a small ratio is expressed in all genes. This normalization applies to almost all of the genes. The second method, genes are normalized based on constant expression. This method applies to a subset of genes instead of all genes. In the third method, genes are either arranged by spiked control or the titration series of control sequence is used. This paper uses the first method in which all genes in the cDNA microarray are used.

This paper verifies the normalization results of the cDNA microarray data. It uses the eigen feature of *k*-means clustering and the fuzzy c-means clustering algorithm [5-8]. The eigen feature makes rules according to experiment data. It influences operation time of the termination tolerance. The notion of cDNA microarray analysis decreased the average operation time of *k*-means clustering and increased the time of fuzzy c-means clustering. The cause of the operation time is an appraised method of cluster position selection. Section 2 explains the difference and features of the two clustering algorithms. Section 3 explains the structure of the experiment data. Section 4 explains the noise reduction method and normalization to be used. Section 5 shows the result based on the proposal theory, and shows the proper objective. The last section presents conclusions and suggests future study.

2. The Clustering Characteristic Analysis

Clustering analysis uses general Euclidean distance between an independent entity. The calculated distance indicates similarity and non-similarity. For *k*-means

clustering, it makes using (1) and (2).

$$c_l^i = \frac{\sum_{i=1}^{d} u_{ij}^{(l-1)} m_j}{\sum_{i=1}^{d} u_{ij}^{(l-1)}} \qquad 1 \le i \le c \tag{1}$$

$$u_{ij}^{(l)} = \begin{cases} 1 & if \ d\left(m_j, c_i^{(l)}\right) = \min_{1 \le k \le c} d\left(mj, c_k^{(l)}\right) \\ 0 & otherwise \end{cases} \tag{2}$$

When it satisfies (3), it stops the operation.

$$If \ \left\| U^{(l)} - U^{(l-1)} \right\| \le \delta \ \ then \ \ END \tag{3}$$

This means the change of U , the cluster membership set, is confined to the limits of the termination tolerance. Fuzzy c-means clustering uses Euclidean distance to estimate similarity, too. It uses U , which is a real number $[0,1]$, but not $\{0,1\}$.

$$u_{ij}^{(l)} = \frac{1}{\sum_{k=1}^{c} \left(d_A^2\left(m_j, c^{(i)}\right) / d_A^2\left(m_j, c^{(k)}\right) \right)^{1/(\omega-1)}} \tag{4}$$

The membership matrix changes can be explained by Figure 1.

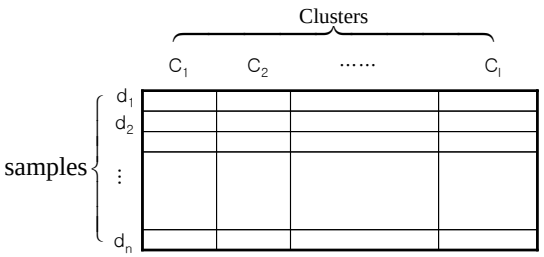

Fig. 1 Membership matrix.

Sample $D = \{d_i \,|\, i = 1, \cdots N\}$ belongs to each cluster with membership values. This means that each sample has an overlap, so it gives the problem careful consideration. For the last section, the normalization corrects fluorescent intensity imbalances and gene expression levels between slides. When it corrects two factors, the cDNA microarray data removes absolute fluorescence intensity and only the variation between samples remains. So it is difficult to divide clusters. We may be thought that it is more convergent than *k*-means clustering, because fuzzy c-means clustering fully considers cluster membership. However, fuzzy c-means clustering time is shortened after normalization. This result can be explained by the cluster membership error effects decreased. If there are more samples, it becomes a remarkable large difference. Unlike fuzzy c-means clustering, *k*-means clustering takes longer and the variance becomes larger after normalization.

3. Experiment Data Structure Analysis

cDNA microarray data is generally represented in matrix form. The matrix expresses the genes in rows, and the cDNA chips or samples in columns. This paper uses the cDNA microarray data from 17,000 genes and 7 chips. The microarray data matrix form is as follows.

Table. 1 DNA microarray matrix

Gene	Chip 1	Chip 2	⋯	Chip 7
1	$x_{1,1}$	$x_{1,2}$	⋯	$x_{1,7}$
2	$x_{2,1}$	$x_{2,2}$	⋯	$x_{2,7}$
⋮	⋮	⋮	⋮	⋮
17,000	$x_{17000,1}$	$x_{17000,2}$	⋯	$x_{17000,7}$
⋮	⋮	⋮	⋮	⋮

For example, $x_{1,1}$ is the expression value of sample 1 and gene 1, and $x_{3,2}$ is the expression value of sample 2 and gene 3. Each expression is a log-ratio value. The microarray software uses this matrix form in the majority of cases.

4. Noise Removal and Normalization

4.1 Noise Removal

When the flag which means reliability level is -50 or -100, it is cut out. In addition, when the sum of gene change rate doesn`t rise above the threshold, it is also cut out. The threshold this paper used is 2. It was able to remove 300 genes.

$$IF \ \left| x_{i,1} + x_{i,2} + \cdots + x_{i,n} \right| < 2$$
$$X(i,:) = [\ ]$$
$$END$$

4.2. 1 Generalization of Normalization

The normalization method is illustrated in figure 2.

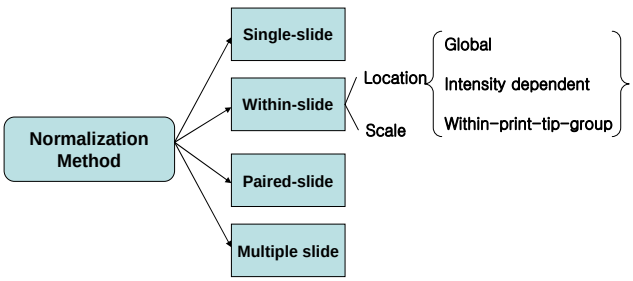

Fig. 2 Normalization Method.

Common normalization methods correct errors using means and variances of the slide by matching the red and green intensities. That is to say, the gene expression level corrects as many errors as α .

$$\log\left(\frac{R}{G}\right)-\alpha \qquad \alpha = \text{Normalization method}$$

α can build up to several normalization. If a special α is made, there will be need to verify.

4.2.2 The Normalization Equation to be used

This paper uses the standardization method[3]. The sample is updated to calculate a mean and standard variation. (5) is the mean of each sample, (6) is the standard variation of sample.

$$\overline{x}_i = \frac{x_{i,1} + x_{i,2} + \cdots + x_{i,n}}{n} \tag{5}$$

$$s_i = \sqrt{\frac{\left(x_{i,1}-\overline{x}_i\right)^2 + \left(x_{i,2}-\overline{x}_i\right)^2 + \cdots + \left(x_{i,n}-\overline{x}_i\right)^2}{n-1}} \tag{6}$$

The value of each normalized is updated on the cDNA microarray data matrix.

$$\frac{x_{i,1}-\overline{x}_i}{s_i}, \frac{x_{i,2}-\overline{x}_i}{s_i}, \cdots, \frac{x_{i,n}-\overline{x}_i}{s_i} \tag{7}$$

5. The Experimentation Result

The result of experimentation is drawn in box-plot, changing the sample size and the cluster number. A box-plot can reasonably show the feature affected by the first random centers.

5.1. The Result of *k*-means Clustering

If the normalization of the cDNA microarray goes well, a operation time of *k*-means clustering takes longer. The interquartile range of the box-plot has 50 percent of the observered data. When the value of the median is similar, the large rate of interquartile takers longer, as in figure 3 and figure 4. After the normalization, *k*-means clustering takes longer. That is to say, the rate of change is more emphatic than the absolute fluorescent intensity, because *k*-means clustering has a sensibility for uclidean distance. This feature of normalization also occurs in fuzzy c-means clustering

5.2.1 The Result of Fuzzy c-means Clustering

The features of *k*-means clustering are certainly observed in fuzzy c-means clustering. When the value of the median is similar, the rate of interquartile takes longer, as in figure 5 and figure 6. Figure 5 is ambiguous, but figure 6 shows that the operation does not take long, because fuzzy c-means clustering has a sensibility of cluster membership.

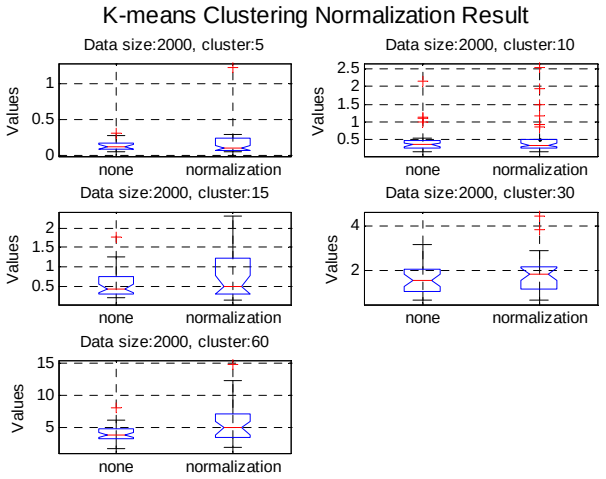

Fig. 3 The result of *k*-means clustering (size:2000).

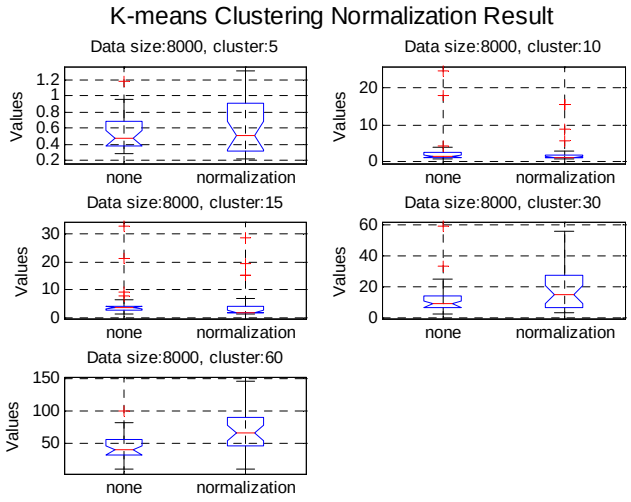

Fig. 4 The result of *k*-means clustering (size:8000).

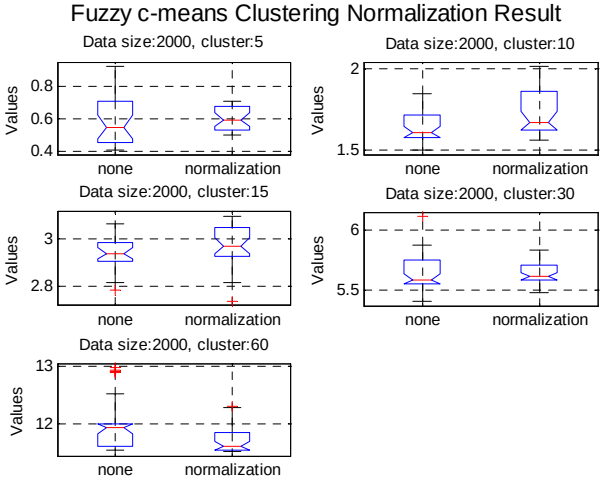

Fig. 5 The result of fuzzy c-means clustering (size:2000)

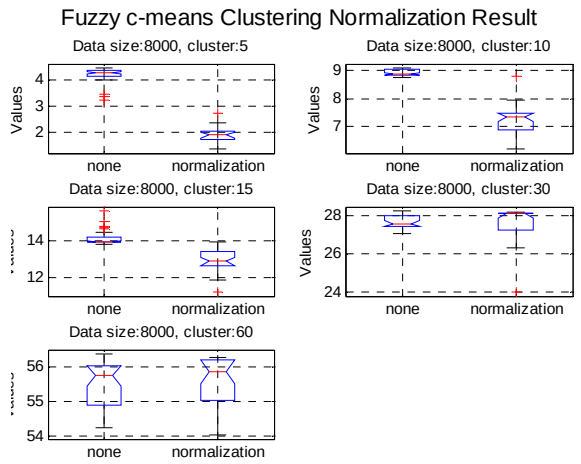

Fig. 6 The result of fuzzy c-means clustering (size:8000).

5.2.2 A High Spot

A high spot is found in figure 6. Figure 7 zooms cluster:30 and cluster:60 in figure 6. It is difficult to show the change of time before and after. Although the normalization emphasizes sample expression, it is difficult to classify to the number of clusters, so it can be assumed that the 8000 samples have 15~30 clusters. When the interquartile rate of fuzzy c-mean clustering keeps their distance, the cluster of gene expression is determined.

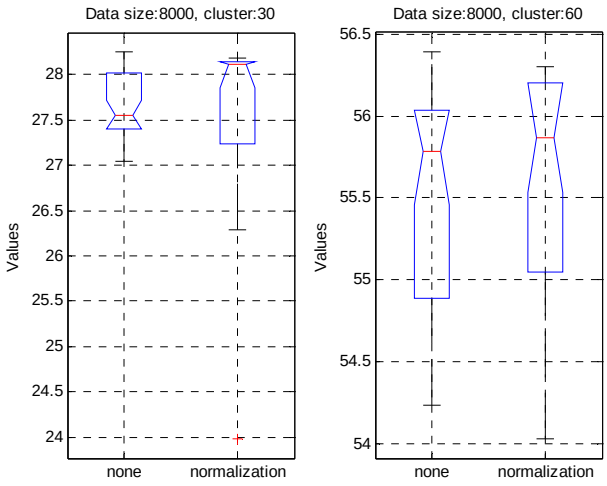

Fig. 7 Zoom in Fig. 6 (cluster:30 and 60).

6. Conclusion

This paper proposes an appraised method of normalization in the cDNA microarray. The proposed model uses the difference of cluster membership update method. However, in this model, it is difficult to determine optimized data-size and cluster number in the microarray. Therefore, in subsequent rearch, the cluster number of gene expression by the interquartile rate is expected to be determined.

ACKNOWLEGEMENT

This work was supported by the Second-stage of the Brain Korea 21 project in 2006.

References

[1] Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999 Jan;21(1 Suppl):33-7

[2] D. Wishart, "Efficient hierarchical cluster analysis for data mining and knowledge discovery." Computing Science and Statistics. pp. 257-263, 1998.

[3] Willian Shannon, Robert Culverhouse, Jill Duncan." Analyzing microarray data using cluster analysis",2003.

[4] J. Derisi, V. Iyer and P. Brosn, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science, vol. 278, pp. 680-686, 1997.

[5] Kjersti Aas, "Microarray Data Mining: A Survey", 2001.

[6] E. Hartuv, A. Schumitt, J. Lange, S. Meier-Ewert,H. Legrach and R. Shamir , “An Algorithm for Clustering cDNAs for Gene Expression Analysis,” Proceedings of the Third International Conference on Computational Molecular Biology (RECOMB 99),pp.188- 197, 1999.

[7] J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Plenum Press, 1981.

[8] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, Fuzzy cluster analysis, Wiley, 2000.