Validity Index for Fuzzy K-Means Clustering Using the Gap Statistic Method

Chinatsu Arima
arima@brs.kyushu-u.ac.jp

Kazumi Hakamada
hakamada@brs.kyushu-u.ac.jp

Masahiro Okamoto
okahon@brs.kyushu-u.ac.jp

Taizo Hanai
taizo@brs.kyushu-u.ac.jp

Laboratory for Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: gene expression analysis, fuzzy k-means, clustering, gap statistic

1 Introduction
Clustering is an important analysis for the gene expression data. However, there is a problem that the microarray data includes large noise from experimental error and it affects the result of clustering analysis. And it is said that fuzzy logic has the feature that is not influenced by the noise. Therefore, we applied the fuzzy k-means clustering [2] combined fuzzy logic with k-means clustering to the DNA microarray data. We compared the fuzzy k-means clustering with the conventional k-means clustering for the gene expression data, and showed that fuzzy k-means clustering has higher robustness against the noise than k-means clustering. On the other hand, the clustering analysis has a fundamental problem that it is necessary to determine the number of clusters. In this study, because in using k-means clustering, the gap statistic [5] has a good performance to estimate the number of clusters, we modified and applied it to fuzzy k-means clustering. To evaluate the usefulness of the modified gap statistic, we compared it with Xie-Beni index that is the published validity index for fuzzy k-means clustering.

2 Method
2.1 Fuzzy k-Means Clustering
The fuzzy k-means clustering uses following objective function

\[ J(K, m) = \sum_{k=1}^{K} \sum_{i=1}^{N} (u_{ki})^m d^2(x_i, c_k), \]

where \( K \) and \( N \) are the number of clusters and genes in the data sets, respectively \( m \) is a parameter which relate to ‘fuzziness’ of resulting clusters, \( u_{ki} \) is the degree of membership of gene \( x_i \) in cluster \( k \), \( d^2(x_i, c_k) \) is the distance from gene \( x_i \) to centroid \( c_k \), and \( c_k \) shows the representative gene expression profile of a cluster \( k \). The parameters in this equation, the cluster centroid vectors \( c_k \) and the components of the membership vectors \( u_{ki} \) can be optimized by Lagrange’s method. And a fuzziness parameter \( m \) was decided by the method of Dembele et al. [2].

2.2 The Gap Statistic
The gap statistic is a method for estimating the optimal number of clusters. This technique is based on the idea that the change in within-cluster dispersion with the increase of the number of clusters is expected under a reference distribution for random data. First, assume that there are a set of samples \( \{ x_r \} \) then by use of the clustering method, the resultant clusters \( C_1, C_2, \ldots, C_k \) can be obtained. For any cluster \( C_r \), the sum of the pair wise distances \( d^2(x_i, x_r) \), for all points in cluster \( r \) is calculated. And the sum of within-cluster dispersion \( W_k \) is defined the following equation

\[ W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i \in C_r} d^2(x_i, x_r), \]

As to the central concept of the gap statistic, it is to compare the \( \log(W_k) \) with its expectation under a
reference distribution. It is defined

$$\text{Gap}_n(k) = E_n^* \{ \log(W_k) \} - \log(W_k),$$

where $E_n^*$ denotes expectation with a sample of size $n$ of the reference distribution. The optimized number of clusters $k$ is decided by $\text{Gap}_n(k)$. In this study, we modified the gap statistic method by changing $W_k$ to apply the fuzzy k-means clustering.

### 2.3 Data Sets

The three artificial data sets named C4D3, C4D5 and C4D10 were used for analysis. The numeric of name shows the number of clusters and the number of dimensions for the data. For example, for the C4D3 data, there are 4 clusters and the number of dimensions is 3.

The two gene expression data sets, Yeast and ALL-AML were used. Yeast is sporulation data of *Saccharomyces cerevisiae*. Samples were harvested at time $t=0, 0.5, 2, 5, 7, 9$ and 11.5 hours. About 6100 genes of expression profiles are included in this data [7]. Using them, we extracted the genes by following the same method [1] and we finally selected 45 genes, whose functions are biologically characterized by Kupiec et al. [4].

ALL-AML is gene expression data of about 6800 genes in 38 acute leukemia samples [8]. This data set is consisting of 27 acute myeloid leukemia (AML) samples and 11 acute lymphoblastic leukemia (ALL) samples. Using them, we selected 50 genes by the same method of Golub et al. [3] to extract genes that showed significant difference of mRNA levels between ALL and AML samples.

### 3 Results and Discussion

The results of the estimated numbers of clusters by using validity indices, Xie-Beni index (XB) and modified gap statistic (Gap_F) are shown in Table1. The values in parentheses in Table1 are values of correctness ratios [6] showing the average ratios of the gene with the same label in any cluster. For three artificial data and ALL-AML data, the correct number of clusters defined by biological and theoretical knowledge was obtained using the modified gap statistic, but was not obtained using Xie-Beni index. For the Yeast data, the modified gap statistic can not determined the correct number of clusters. However, the correctness ratio of the modified gap statistic was higher than that of Xie-Beni index. As the result, we concluded that the modified gap statistic is a superior validity index of the number of clusters to Xie-Beni index for fuzzy k-means clustering.

<table>
<thead>
<tr>
<th>Data set</th>
<th>C4D3</th>
<th>C4D5</th>
<th>C4D10</th>
<th>Yeast</th>
<th>ALL-AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct number of clusters</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Validity index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XB</td>
<td>2 (0.500)</td>
<td>2 (0.500)</td>
<td>3 (0.500)</td>
<td>3 (0.538)</td>
<td>6 (0.974)</td>
</tr>
<tr>
<td>Gap F</td>
<td>4 (1.000)</td>
<td>4 (0.985)</td>
<td>4 (0.985)</td>
<td>5 (0.667)</td>
<td>2 (0.974)</td>
</tr>
</tbody>
</table>

### References


