The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 369-380

Recovery Levels of Clustering Algorithms Using Different Similarity Measures for Functional Data

Seong San Chae[1], Chansoo Kim[2], and William D. Warde[3]

Abstract

Clustering algorithms with different similarity measures are commonly used to find an optimal clustering or close to original clustering. The recovery level of using Euclidean distance and distances transformed from correlation coefficients is evaluated and compared using Rand's (1971) C statistic. The C values present how the resultant clustering is close to the original clustering. In simulation study, the recovery level is improved by applying the correlation coefficients between objects. Using the data set from Spellman *et al.* (1998), the recovery levels with different similarity measures are also presented. In general, the recovery level of true clusters was increased by using the correlation coefficients.

Keywords : Agglomerative clustering algorithms; Correlation coefficients; Rand's C statistic

1. Introduction

In some clustering applications, a transformed sample correlation type instead of Euclidean distance is used as a dissimilarity measure if variables rather than objects are clustered. When the correlation coefficients are used as similarity measures, small values are regarded as very dissimilar, while large positive values are regarded as very similar. Negative correlations are replaced by their absolute values in some clustering applications (Johnson and Wichern, 1998).

Eisen *et al.* (1998) found that the correlation coefficient conforms well to the intuitive biological notion to express two genes and captures similarity in shape but places no emphasis on the magnitude of the two series of measurements. Perou *et al.* (1999) used a form of the correlation coefficient that is exactly equal to the Pearson correlation coefficient as a gene similarity metric. Hadjiargyrou *et al.* (2002) established clusters by grouping genes using the Pearson correlation coefficient after excluding inconsistent expression patterns (outliers).

[1] Associate Professor, Department of Information and Statistics, Daejeon Uiversity, Daejeon, 300-716, Korea. E-mail: chae@dju.ac.kr

[2] Assistant Professor, Department of Statistics, Oklahoma State University, Stillwater, OK 74078, USA.

[3] Professor, Department of Statistics, Oklahoma State University, Stillwater, OK 74078, USA.

Wu (2001) suggested the use of the Pearson correlation coefficient and Spearman rank-order correlation coefficient to measure the similarity of each gene's profile in finding a match between genes. The rank-order correlation coefficient may be used rather than precise numerical quantities if a lower specification (or robust against outliers) is wanted (Bickel (2003)). Cherepinsky *et al.* (2003) showed that a shrinkage-based correlation coefficient improves the accuracy of the cluster analysis.

The recovery level of agglomerative clustering algorithms is investigated by using different dissimilarity measures, Euclidean distance and the distance transformed from correlation coefficients. The distance calculated from different types of correlation coefficients is constructed by using the formula $d_{ij} = \sqrt{2(1-\gamma_{ij})}$, where γ_{ij} is the correlation coefficients between i-th and j-th objects. To measure the recovery level of clustering algorithms using different similarity measures, Rand's (1971) C statistic is used. DuBien, Warde and Chae (2004) show that the mean and variance of Rand's C statistic for any K given by Fowlkes and Mallows (1983) are special cases of the mean and variance of C given by DuBien and Warde (1981). It evaluates the results of cluster analysis based on how they partition the data points in the concept of correspondence and is a measure of similarity with $0.0 \le C_k \le 1.0$. When the partition produced by clustering algorithm is identical to the structure within the data treated, then C_k is 1.0. The formulation of the C_k statistic based on the incidence matrix $[n_{ij}]$ is given as follows:

$$C_k = \frac{\binom{N}{2} - \frac{1}{2}\left(\sum_{i=1}^{K} n_{i.}^2 + \sum_{j=1}^{K} n_{.j}^2\right) + \sum_{i \neq j}^{K} n_{ij}^2}{\binom{N}{2}}$$

where n_{ij} is the number of data points in common between the i-th cluster formed from one clustering and the j-th cluster formed from another clustering, $i,j = 1,2,\ldots,K,\ldots,N$. Further, the results of using the different similarities are examined and compared on the cell cycle data from Spellman *et al.* (1998).

2. Similarity Measures and Clustering Algorithms

Let N be the number of data points with p variables. Then an $N \times p$ matrix of measurements, say X, might be represented as X^N indicating that there are N data points in X. In order to examine and compare without standardization, three different transformations (statistical standardization, Mahalanobis standardization and range standardization) were investigated. However, only the results from clustering algorithms by using the range transformation, $z_{il} = \frac{x_{il}}{max(x_l) - min(x_l)}$, where x_{il} is a measurement of the l-th variable on the i-th object and x_l is the l-th variable with N objects in X, are presented.

For either X or Z, a cluster, y_h, is simply a nonempty subset of the object space, and a clustering, $Y = (y_1, y_2, ..., y_h, ..., y_K)$, is any partition of the object space. The number of clusters, K, contained in a clustering shall be referred to as the size of the clustering. Some notations useful for a cluster, a clustering, an hierarchy and a clustering algorithm can be found in Chae and Warde (1991).

In this study, measures of distance, d, imposed on the data points are the Euclidean and the transformed distance from the correlation coefficients between two objects for any series of measurements. Using the fact that the dissimilarities defined by $\sqrt{1-r_{ij}}$ are Euclidean if the similarities r_{ij} satisfy $0 \leq r_{ij} \leq 1$ (Gower, 1966), the formula $d_{ij} = \sqrt{2(1-\gamma_{ij})}$ are used to transform the Pearson correlation coefficient,

$$\gamma_{ij} = \frac{\sum_{l=1}^{p}(x_{il}-\overline{x}_i)(x_{jl}-\overline{x}_j)}{\sqrt{\sum_{l=1}^{p}(x_{il}-\overline{x}_i)^2 \sum_{l=1}^{p}(x_{jl}-\overline{x}_j)^2}},$$

where $\overline{x}_i = \sum_{l=1}^{p} x_{il}$, $i, j = 1, 2, ..., N$, to the Euclidean distance (Johnson and Wichern, 1998). Eisen *et al.* (1998) used $\overline{x}_i = 0.0$; Cherepinsky *et al.* (2003) used a shrinkage-based estimator, $\tau \overline{x}_i$, $0.0 \leq \tau \leq 1.0$, instead of $\overline{x}_i$; Bickel (2003) used the Spearman rank-order correlation coefficient instead of γ_{ij}. The shrinkage-based correlation coefficient suggested by Cherepinsky *et al.* (2003) is excluded since it is not easy to obtain a τ that satisfies the condition $0.0 \leq \tau \leq 1.0$ for real data set. In the primary study, $d_{ij} = \sqrt{2(1-|\gamma_{ij}|)}$ and $d_{ij} = \sqrt{2(1-r_{ij}^2)}$ were also considered, however, the recovery levels of the clustering algorithms were not as good as using the transformation, $d_{ij} = \sqrt{2(1-\gamma_{ij})}$.

Let Y represent the "true" structure of the N data points with K clusters and $Y^{[N,K]}$ be a certain arrangement of Y with K clusters. Let Y' denote a clustering that result from applying an agglomerative clustering algorithm to the N data points with number of clusters K. Then Rand's $C(Y, Y')$ value is a measure of the recovery level of the clustering algorithm to the true structure for K.

For any clustering $Y^{[N,K]}$ in the hierarchy, if the distances d_{ij}, d_{ik}, and d_{jk} between pairs of clusters y_i, y_j and y_k are obtained recursively from clustering $Y^{[N,K+1]}$, $K<N$, then the distance between the new cluster $y_{(ij)} = y_i \cup y_j$ and any other cluster $y_k \in Y^{[N,K]}$ can be computed from the following formula:

$$d_{(ij)k} = \frac{1-\beta+2\pi}{2} d_{jk} + \frac{1-\beta-2\pi}{2} d_{ik} + \beta d_{ij}$$

where $d_{ij} < d_{ik} < d_{jk}$. This formula represents a two parameter (β, π)-family of agglomerative clustering algorithms derived by DuBien and Warde (1979).

Since the purpose of this study is to survey the clustering methods available to cluster objects on the basis of their expression patterns, six clustering algorithms are chosen from the (β, π)-family of agglomerative clustering algorithms. In the (β, π)-family, $(.0, -.5)$ is known as single linkage; $(.0, .0)$ as average linkage; $(.0, .5)$ as complete linkage; $(-.25, .0)$ and $(-.5, .0)$ as representations of the flexible strategy; $(-.5, .75)$ is the recommendation by DuBien and Warde (1987).

3. Simulation Study

3.1 Design of Simulation

Few studies provide the method for clustering functional data that is measured over a series of p time points in the experiments. In this study, the simplest auto- regressive model, which is a stationary time series, is chosen to describe a further application for clustering gene expression data. The other types of simulation model may be considered depending on the characteristics of data treated. The Rand's C values that represent the recovery levels of clustering algorithms for true structure with different similarity measures were calculated using the following steps:

1. Simulated data $X_{N \times p}$, $N=150$, $p=24$ was generated using the model,

$$X_{it} = \sigma_{i(k)t} Z_{it} + W_{it},$$
$$W_{it} = \phi W_{i,t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \; Cov(W_{it}, \epsilon_t) = 0,$$
$$Var(W_{it}) = \frac{\sigma_\epsilon^2}{1-\phi^2}, \; -1 < \phi < 1,$$

where $Z_{it} \sim N(0,1)$, $i = 1,2,\ldots,N$, $t = 1,2,\ldots,p$. In fact, W_{it}, that is the mean at time point t if the i-th object belongs to the k-th cluster, $k = 1,2,...,5$, in Spellman's (1998) data set, is used for the autoregressive model $(AR(1))$ where ϕ satisfies the stationary condition, then $\widehat{W}_{it}$ is used instead of W_{it} to generate X_{it};

2. $\sigma^2_{i(k)t}$, the constant to control the variance of X_{it} that belongs to the k-th cluster, is set as $0.5, 0.75$;
3. n_k, the split or the size of the k-th cluster generated from each population;
4. The distance converted from each correlation coefficient between each pair of objects in X was computed and stored in lower triangular matrix order by rows as the vector D_1 of length $\binom{N}{2}$;
5. Each of the six clustering algorithms was applied to D_1 to produce Y';
6. For each of the clusterings Y' generated from above steps, $C(Y, Y')$ was calculated for the six clustering algorithms.

For each setting of the $(\sigma^2_{i(k)t}, n_1;n_2;n_3;n_4;n_5)$, the above sequence of steps was replicated 100 times and the sample mean of the C statistic, $\overline{C}$, was computed for the six clustering algorithms. The $\overline{C}$ value quantifies the recovery of true structure by applying the six clustering algorithms for each setting of $(\sigma^2_{i(k)t}, n_1;n_2;n_3;n_4;n_5)$.

3.2 Results from Simulation

The results from the simulation study are not independent of the fixed structural parameters which were specified previously. However, the results are discussed in terms of changes in parameters $(\sigma^2_{i(k)t}, n_1;n_2;n_3;n_4;n_5)$, the transformation and the clustering algorithms with different similarity measures. Since the results from the single linkage are different from other clustering algorithms, it is excluded from further discussion.

Tables 1-2 give the recovery results of the six clustering algorithms applied to the original and the range-transformed data using the Euclidean distance (ED) and the correlation coefficients. Although several possible settings for $(n_1;n_2;n_3;n_4;n_5)$ were investigated, only the results from $n_k = 30$, for $k = 1,2,...,5$ and $(n_1;n_2;n_3;n_4;n_5) = (24;37;20;55;24)$ are presented.

As shown in Table 1, the difference in recovery represented by $\overline{C}(Y, Y')$ is mainly due to the settings of $\sigma^2_{i(k)t}$ designed into the original data. The recovery levels of the six agglomerative clustering algorithms increase as $\sigma^2_{i(k)t}$ decreases. For the case of equal size, the recovery levels with the three correlation coefficients are higher than with the Euclidean

<Table 1> The $\overline{C}(Y, Y')$ for the original data

Data	$\sigma^2_{i(k)t}$	(β, π)	ED	Eisen	Pearson	Spearman
Equal	0.50	(0.0, -0.5)	.2272	.2414	.2372	.2289
		(0.0, 0.0)	.8548	.9166	.9059	.8918
		(0.0, 0.5)	.8894	.9201	.9219	.8973
		(-0.25, 0.0)	.9143	.9483	.9498	.9357
		(-0.5, 0.0)	.9151	.9468	.9461	.9302
		(-0.5, 0.75)	.8935	.9224	.9255	.8988
	0.75	(0.0, -0.5)	.2271	.2287	.2292	.2288
		(0.0, 0.0)	.7729	.8414	.8480	.8329
		(0.0, 0.5)	.8247	.8386	.8433	.8223
		(-0.25, 0.0)	.8553	.8836	.8856	.8632
		(-0.5, 0.0)	.8613	.8828	.8797	.8593
		(-0.5, 0.75)	.8312	.8348	.8385	.8185
Unequal	0.50	(0.0, -0.5)	.2276	.2742	.2749	.2723
		(0.0, 0.0)	.8857	.9079	.8956	.8703
		(0.0, 0.5)	.9046	.9077	.9000	.8673
		(-0.25, 0.0)	.9249	.9273	.9239	.8995
		(-0.5, 0.0)	.9178	.9213	.9166	.8919
		(-0.5, 0.75)	.8783	.8785	.8737	.8631
	0.75	(0.0, -0.5)	.2732	.2741	.2745	.2701
		(0.0, 0.0)	.8082	.8397	.8381	.8161
		(0.0, 0.5)	.8363	.8358	.8457	.8196
		(-0.25, 0.0)	.8618	.8670	.8687	.8463
		(-0.5, 0.0)	.8582	.8588	.8619	.8454
		(-0.5, 0.75)	.8231	.8246	.8274	.8083

distance. Among three correlation coefficients, the recovery levels using the Spearman correlation coefficient are lower than those using the other correlation coefficients. In particular, the recovery levels using the Spearman correlation coefficient is the lowest for the case of unequal size.

In Table 2, the effect of the range transformation is examined and compared to using no transformation. The recovery levels decrease or increase depending on the clustering algorithms if the range transformation is used. The use of the Spearman rank correlation coefficient is more robust against outliers when the average linkage algorithm is applied to the data with unequal size of clusters. That is why Bickel (2003) suggested using average linkage with the Spearman rank correlation coefficient. However, Kojadinovic (2004) pointed out that it enables only the detection of monotonic functional dependencies.

As shown in Tables 1 and 2, the calculated $\overline{C}(Y, Y')$ values show essential differences

<Table 2> The $\overline{C}(Y, Y')$ for the range-transformed data

Data	$\sigma^2_{i(k)t}$	(β, π)	ED	Eisen	Pearson	Spearman
Equal	0.50	(0.0, -0.5)	.2271	.2312	.2276	.2277
		(0.0, 0.0)	.8342	.8703	.8647	.8538
		(0.0, 0.5)	.8784	.8832	.8800	.8772
		(-0.25, 0.0)	.9087	.9095	.9099	.9047
		(-0.5, 0.0)	.9100	.9076	.9082	.9045
		(-0.5, 0.75)	.8848	.8859	.8822	.8804
	0.75	(0.0, -0.5)	.2270	.2283	.2285	.2290
		(0.0, 0.0)	.7673	.8117	.8184	.8097
		(0.0, 0.5)	.8224	.8276	.8329	.8265
		(-0.25, 0.0)	.8519	.8531	.8560	.8495
		(-0.5, 0.0)	.8495	.8530	.8582	.8487
		(-0.5, 0.75)	.8215	.8311	.8236	.8245
Unequal	0.50	(0.0, -0.5)	.2743	.2776	.2778	.2689
		(0.0, 0.0)	.8776	.8855	.8852	.8801
		(0.0, 0.5)	.8924	.8906	.8912	.8694
		(-0.25, 0.0)	.9113	.9175	.9149	.8998
		(-0.5, 0.0)	.9022	.9017	.9104	.8976
		(-0.5, 0.75)	.8763	.8659	.8747	.8665
	0.75	(0.0, -0.5)	.2733	.2741	.2741	.2701
		(0.0, 0.0)	.7969	.8326	.8314	.8161
		(0.0, 0.5)	.8226	.8176	.8328	.8196
		(-0.25, 0.0)	.8560	.8553	.8558	.8463
		(-0.5, 0.0)	.8479	.8475	.8490	.8454
		(-0.5, 0.75)	.8171	.8146	.8180	.8083

depending on the similarity measures between objects. It implies that the use of the correlation coefficient in applying the clustering algorithm has a significant effect on the recovery of the true clustering. Based on the recovery levels, more similar clustering is retrieved when the correlation coefficient instead of the Euclidean distance is used as a measure between objects.

4. An Example

An application using a set of data that includes yeast (*Saccharomyces cerevisiae*) genes from Spellman *et al.* (1998) is given. The primary data can be obtained at *http://cellcycle-www.stanford.edu.* In their normalization procedure on the primary data, a total of 800 yeast genes are identified as being periodically regulated and meeting an objective minimum criterion for cell cycle regulation. The statistical methods used in analyzing gene expression data might be found in Kim (2004).

For convenience, 630 genes with 24 variables (the results of a series of 24 time points in the experiments) are taken out of the identified 800 genes that have no missing values on the data set with five clusters. According to Spellman *et al.* (1998), these five clusters approximate the commonly used cell groups and provide a natural basis for organizing yeast gene expression in the literature. The sizes of clusters to which it belongs are (102-159-82-231-56) for *S/G2*, *G2/M*, *M/G1*, *G1* and *S* groups of genes.

As shown in Table 3, the five clusters are identified by the agglomerative clustering algorithms with different similarity measures on the original data and the range-transformed data. The recovery level of identified clusters in Spellman *et al.* (1998) is increased by using the correlation coefficients instead of using the Euclidean distance on the original and the range-transformed data. For discussion on the effect of range transformation, the recovery levels of the six clustering algorithms are first discussed for the original data. The recovery levels of using the Eisen correlation and the Pearson correlation coefficients are the same since the mean of each gene by the normalization procedure is close to 0.0. The use of complete linkage, $(.0, .5)$, and one of the flexible strategies, $(-.25, .0)$ with the correlation coefficients might be recommended instead of the average linkage for the cell cycle data of Spellman *et al.* (1998). Above all, the results of using the three correlation coefficients are better than those using the Euclidean distance.

At this point, the results from the clustering algorithms are discussed for the range-transformed data. The recovery levels are increased or decreased, depending on the different algorithms when the results are compared with the results with the range transformation. No single clustering algorithm or transformation has proven free from ambiguity in establishing well specified and carefully validated procedures in cluster analysis. However, the use of average linkage with the Spearman rank correlation coefficient on the range-transformed data gives the best recovery level for the cell cycle data of Spellman *et al.* (1998).

As we have little a priori knowledge of expected gene expression patterns, it is difficult to say one specific clustering algorithm which is best overall. The choices of a clustering algorithm and a measure of similarity depends on the structure and characteristic of the data. Comparing the recovery levels of clustering algorithms presented by Rand's C value, the use of the correlation coefficients recovers the arbitrarily defined clusters by Spellman *et al.* (1998) better than using Euclidean distance.

<Table 3> The sizes of clusters and $\bar{C}(Y, Y')$ for Spellman's data

Data	Similarity	Group	S/G2	G2/M	M/G1	G1	S	*	$\bar{C}$
	Measures	(β, π)/Sizes	102	159	82	231	56		
original	Euclidean	(0.0, 0.0)	2	362	5	258	3	359	.6561
		(0.0, 0.5)	143	129	28	326	4	409	.7178
		(-0.25, 0.0)	37	186	46	287	74	341	.7036
		(-0.5, 0.0)	100	157	82	271	20	369	.7248
		(-0.5, 0.75)	210	146	57	112	105	296	.7133
	Eisen Pearson	(0.0, 0.0)	93	166	28	179	164	290	.7221
		(0.0, 0.5)	107	145	121	171	86	382	.7821
		(-0.25, 0.0)	164	184	105	143	34	384	.7747
		(-0.5, 0.0)	145	211	45	145	84	345	.7565
		(-0.5, 0.75)	72	122	122	208	116	339	.7633
	Spearman	(0.0, 0.0)	7	195	85	252	91	340	.7225
		(0.0, 0.5)	172	76	116	161	105	311	.7417
		(-0.25, 0.0)	116	132	160	163	69	320	.7573
		(-0.5, 0.0)	109	193	82	162	84	347	.7553
		(-0.5, 0.75)	55	211	130	147	87	319	.7450
range transform	Euclidean	(0.0, 0.0)	2	321	2	302	3	329	.6069
		(0.0, 0.5)	3	169	105	293	60	266	.6408
		(-0.25, 0.0)	120	166	53	246	45	383	.7443
		(-0.5, 0.0)	88	181	37	249	75	365	.7367
		(-0.5, 0.75)	139	98	113	172	108	264	.7342
	Eisen	(0.0, 0.0)	47	187	122	181	93	319	.7455
		(0.0, 0.5)	211	111	86	181	41	297	.7228
		(-0.25, 0.0)	141	159	136	102	94	342	.7558
		(-0.5, 0.0)	173	179	92	144	42	330	.7468
		(-0.5, 0.75)	137	143	122	132	96	304	.7625
	Pearson	(0.0, 0.0)	85	106	108	250	81	394	.7749
		(0.0, 0.5)	92	185	59	166	128	374	.7709
		(-0.25, 0.0)	170	102	132	156	70	352	.7605
		(-0.5, 0.0)	174	161	83	99	113	331	.7580
		(-0.5, 0.75)	94	180	122	201	33	392	.7696
	Spearson	(0.0, 0.0)	96	218	90	209	17	436	.8002
		(0.0, 0.5)	172	76	116	161	105	311	.7417
		(-0.25, 0.0)	131	126	119	147	107	352	.7529
		(-0.5, 0.0)	109	193	82	162	84	347	.7553
		(-0.5, 0.75)	55	211	130	147	87	319	.7450

* : maximum numbers of genes which are assigned to the "target" clusters as defined by Spellman *et al.* (1998)

5. Concluding Remarks and Future Study

Several clustering algorithms with different similarity measure are commonly used to find an optimal clustering or close to original clustering. In applying a clustering method, the first step is to choose a mathematical description of similarity in the behavior of two objects.

Different similarity measures between objects are applied to produce groups with similar patterns of expression which form the basis of a classification scheme useful in later studies for predictive purposes. The recovery levels from agglomerative clustering algorithms using Euclidean distance and different correlation coefficients are evaluated and compared to using Rand's C values.

In the biological literature, the correlation coefficient conforms well to the intuitive biological notion of similarity between two genes, since this statistic captures similarity in shape. Few studies provides experimental information whether or not a transformation is necessary or desirable. According to our primary study, the Mahalanobis transformation that takes into account the off diagonal elements of the correlation matrix should not be used if any form of standardization is necessary on sample data with a large number of variables. In applying the agglomerative clustering algorithms, the effect of transformation was examined and compared to results without transformation for the various structural settings of the parameters.

In simulation study, $(-.25, .0)$ and $(-.5, .0)$, which are the representations of the flexible strategies, give better recovery level when the five agglomerative clustering algorithms using different measures are applied to the same data with or without transformation. Differences in recovery levels are found between the original and the range-transformed data.

Using results from simulation and application to Spellman's data, the recovery of true clusters as evaluated using Rand's C values was increased by using the correlation coefficients instead of the Euclidean distance between objects. If we have enough knowledge on expected gene expression patterns, we might say that the shape of clusters formed by one specific clustering algorithm is the best for this data set that reach the goal of the research.

For further research, the robust cluster analysis by using robust estimator is considered. Our focus is on the way to find the robust estimators of γ_{ij} not only for $\bar{x}_i$, since the correlation coefficients suggested by Bickel (2003) and Cherpinsky *et al.* (2003) work only for the special cases according to our study.

References

[1] Bickel, D.R. (2003). Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically, *Bioinformatics,* Vol. 19, 18-824.

[2] Chae, S.S. and Warde, W.D. (1991). A method to predict the number of clusters, *Journal of the Korean Statistical Society*, Vol. 20, 162-176.

[3] Cherpinsky, V., Feng, J., Rejali, M. and Mishra, B. (2003). Shrinkage-based smilarity metric for cluster analysis of microarray data, *Proceeding of National Academy Sciences in USA*, Vol. 100, 9668-9673.

[4] DuBien, J.L. and Warde, W.D. (1979). A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms, *The Canadian Journal of Statistics*, 7, 29-38.

[5] DuBien, J.L. and Warde, W.D. (1981). Some distributional results concerning a comparative statistic used in cluster analysis, *ASA Proceedings of the Social Statistics Section*, 309-313.

[6] DuBien, J.L. and Warde, W.D. (1987). A comparison of agglomerative clustering method with respect to noise, *Communications in Statistics, Theory and Method*, Vol. 16, 1433-1460.

[7] DuBien, J.L., Warde, W.D. and Chae, S.S. (2004). Moments of Rand's C statistic in cluster analysis (accepted by *Statistics & Probability Letters*).

[8] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceeding of National Academy Sciences in USA*, Vol. 95, 14868-14868.

[9] Fowlkes, E.B. and Mallows, C.L. (1983). A method for comparing two hierarchical clusterings, *Journal of American Statistical Association*, Vol. 78, 553-569.

[10] Gower, J.C. (1966). Some distance properties of latent root and vector mehtods used in multivariate analysis, *Biometrika*, Vol. 53, 325-338.

[11] Hadjiargyrou, M., Lombardo, F., Zhao, S., Ahrens, W., Joo, J., Ahn, H., White, D.W. and Rubin, C.T. (2002). Transcriptional profiling of bone regeneration: insight into the molecular complexity of wound repair, *Journal of Biological Chemistry*, Vol. 277, 30177-30182.

[12] Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*, 4th Edition, Prentice Hall.

[13] Kim, Choongrak (2004). Statistical methods for gene expression data, *The Korean Communications in Statistics*, Vol. 10, 59-77.

[14] Kojadinovic, I. (2004). Agglomerative hierarchical clustering of continuous variables based on mutual information, *Computational Statistics & Data Analysis*, Vol. 46, 269-294.

[15] Perou, C.M., Jeffrey, S.S., Rijn, M.V., Rees, C.A., Eisen, M.B., Ross, D.T. Pergamenschikov, A., Williams, C.R., Zhu, S.X., Lee, J.C.F., Lashkari, D., Shalon, D., Brown, P.O. and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proceeding of National Academy Sciences in USA*, Vol. 96, 9212-9217.

[16] Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, Vol. 66, 846-850.

[17] Spellman P.T., Sherlock, G., Zhang, M.Q., Iyer V.R., Eisen M.B., Brown, P.O., Botstein,

D. and Futcher B. (1998). Comprehensive identification of cell cycle- regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol. 9, 3273-3297.

[18] Wu, Thomas D. (2001). Analysing gene expression data from DNA microarrays to identify candidate genes, *Journal of Pathology*, Vol. 195, 53-65.

[Received March 2004, Accepted June 2004]