Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics)

Ka Yee Yeung, Walter L. Ruzzo

May 3, 2001

1 The Adjusted Rand index

In order to compare clustering results against external criteria, a measure of agreement is needed. Since we assume that each gene is assigned to only one class in the external criterion and to only one cluster, measures of agreement between two partitions can be used.

Given a set of n objects $S = \{O_1, \ldots, O_n\}$, suppose $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ represent two different partitions of the objects in S such that $\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$ and $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq C$. Suppose that U is our external criterion and V is a clustering result. Let a be the number of pairs of objects that are placed in the same class in U and in the same cluster in V, b be the number of pairs of objects in the same class in U but not in the same cluster in V, c be the number of pairs of objects in the same cluster in V but not in the same class in U, and d be the number of pairs of objects in different classes and different clusters in both partitions. The quantities a and d can be interpreted as agreements, and b and c as disagreements. The Rand index [Rand, 1971] is simply $\frac{a+d}{a+b+c+d}$. The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1.

A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value (say zero). The adjusted Rand index proposed by [Hubert and Arabie, 1985] assumes the generalized hypergeometric distribution as the model of randomness, *i.e.*, the U and V partitions are picked at random such that the number of objects in the classes and clusters are fixed. Let n_{ij} be the number of objects that are in both class u_i and cluster v_j. Let $n_{i.}$ and $n_{.j}$ be the number of objects in class u_i and cluster v_j respectively. The notations are illustrated in Table 1.

Class \ *Cluster*	v_1	v_2	$\ldots$	v_C	*Sums*
u_1	n_{11}	n_{12}	$\ldots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.C}$	$n_{..} = n$

Table 1: Notation for the contingency table for comparing two partitions.

The general form of an index with a constant expected value is $\frac{index - expected\ index}{maximum\ index - expected\ index}$, which is bounded above by 1, and takes the value 0 when the index equals its expected value.

Under the generalized hypergeometric model, it can be shown [Hubert and Arabie, 1985] that:

$$E\left[\sum_{i,j}\binom{n_{ij}}{2}\right] = \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2} \tag{1}$$

The expression $a + d$ can be simplified to a linear transformation of $\sum_{i,j}\binom{n_{ij}}{2}$. With simple algebra, the adjusted Rand index [Hubert and Arabie, 1985] can be simplified to:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}} \tag{2}$$

Let us illustrate the adjustment for expected value with an example. Example 1 is a contingency table in the same form as in Table 1.

Class \ *Cluster*	v_1	v_2	v_3	*Sums*
u_1	1	1	0	2
u_2	1	2	1	4
u_3	0	0	4	4
Sums	2	3	5	$n = 10$

Example 1

a is defined as the number of pairs of objects in the same class in U and same cluster in V, hence a can be written as $\sum_{i,j}\binom{n_{ij}}{2}$. In Example 1, $a = \binom{2}{2} + \binom{4}{2} = 7$. b is defined as the number of pairs of objects in the same class in U but not in the same cluster in V. In terms of the notation in Table 1, b can be written as $\sum_i \binom{n_{i.}}{2} - \sum_{i,j}\binom{n_{ij}}{2}$ In Example 1, $b = \binom{2}{2} + \binom{4}{2} + \binom{4}{2} - 7 = 6$. Similarly, c is defined as the number of pairs of objects in the same cluster in V but not in the same class in U, so c can be written as $\sum_j \binom{n_{.j}}{2} - \sum_{i,j}\binom{n_{ij}}{2} = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} - 7 = 7$. d is defined as the number of pairs of objects that are not in the same class in U and not in the same cluster in V. Since $a+b+c+d = \binom{n}{2}$, $d = \binom{10}{2} - 7 - 6 - 7 = 25$. The Rand index for comparing the two partitions in Example 1 is $\frac{7+25}{45} = 0.711$, while the adjusted Rand index is $\frac{7-14*13/45}{(14+13)/2-14*13/45} = 0.313$ (see Equation 2 for the definition of the adjusted Rand index). The Rand index is much higher than the adjusted Rand index, which is typical. Since the Rand index lies between 0 and 1, the expected value of the Rand index (although not a constant value) must be greater than or equal to 0. On the other hand, the expected value of the adjusted Rand index has value zero and the maximum value of the adjusted Rand index is also 1. Hence, there is a wider range of values that the adjusted Rand index can take on, thus increasing the sensitivity of the index.

In [Milligan and Cooper, 1986], many different indices were evaluated for measuring agreement between two partitions in clustering analysis with different numbers of clusters, and they recommended the adjusted Rand index as the index of choice. We adopt the adjusted Rand index as our measure of agreement between the external criteria and clustering results.

1.1 Illustrations of the adjusted Rand index

Two examples from the paper will be used to illustrate the use of the adjusted Rand index. The results on the ovary data set using k-means and Euclidean distance in Figure 5(d) show that the adjusted Rand indices are high for the first 2 and 3 PC's and then drop drastically to below that without PCA. Example 2 and Example 3 show the contingency tables for the clustering results from k-means and Euclidean distance on the ovary data using the first 3 and 4 PC's respectively. In the examples, the rows are the classes and the columns are the clusters. It is clear that the clusters from Example 2 (with the first 3 PC's) separates the four classes in the ovary data. On the other hand, the clusters from Example 3 (with the first 4 PC's) classes 1 and 3 are combined in the same cluster. Using Equation 2, the adjusted Rand index from Example 2 (with the first 3 PC's) is 0.663, while the adjusted Rand index from Example 3 (with the first 4 PC's) is 0.519. The higher adjusted Rand index from Example 2 confirms our visual inspection that the clustering result using the first 3 PC's is of higher quality than that using the first 4 PC's.

Class $\setminus$ *Cluster*	v_1	v_2	v_3	v_4	*Sums*
u_1	55	1	1	1	58
u_2	10	76	1	1	88
u_3	3	2	26	1	32
u_4	6	2	4	45	57
Sums	74	81	32	48	$n = 235$

Example 2: clustering result from k-means and Euclidean distance on the ovary data using the first 3 PC's

Class $\setminus$ *Cluster*	v_1	v_2	v_3	v_4	*Sums*
u_1	1	1	2	54	58
u_2	1	73	4	10	88
u_3	2	1	3	26	32
u_4	45	0	2	10	57
Sums	49	75	11	100	$n = 235$

Example 3: clustering result from k-means and Euclidean distance on the ovary data using the first 4 PC's

When the four classes of the ovary data are viewed in the space of the first 3 PC's (Figure 1), the four classes are reasonably well-separated in the Euclidean space. However, when the ovary data is visualized in the space of the first, second and fourth PC's, the four classes are not as well-separated (see Figure 2). This visualization together with the clustering result from Example 3 suggest that the fourth PC contains probably mostly noise.

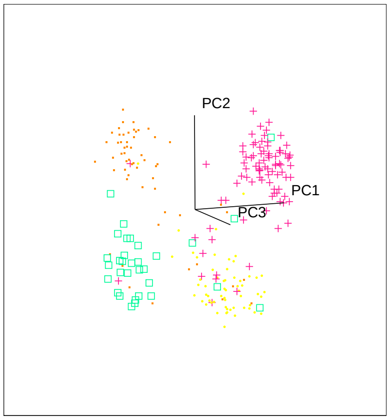

Figure 1: Visualization of the four classes from the ovary data in the space of the first three PC's.

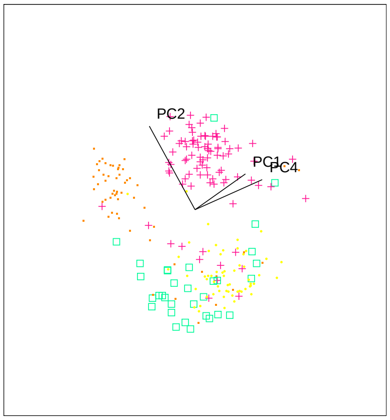

Figure 2: Visualization of the four classes using the first, second and fourth PC's and k-means with Euclidean distance from the ovary data.

2 Clustering algorithms and similarity metrics

We implemented three clustering algorithms: the *Cluster Affinity Search Technique* (CAST) [Ben-Dor and Yakhini, 1999], the hierarchical *average-link* algorithm, and the *k-means* algorithm (with average-link initialization) [Jain and Dubes, 1988].

2.1 CAST

The Cluster Affinity Search Technique (CAST) is an algorithm proposed by [Ben-Dor and Yakhini, 1999] to cluster gene expression data. The input to the algorithm includes the pairwise similarities of the genes, and a cutoff parameter t (which is a real number between 0 and 1). The clusters are constructed one at a time. The current cluster under construction is called C_{open}. The *affinity* of a gene g, $a(g)$, is defined to be the sum of similarity values between g and all the genes in C_{open}. A gene g is said to have high affinity if $a(g) \geq t|C_{open}|$. Otherwise, g is said to have low affinity. Note that the affinity of a gene depends on the genes that are already in C_{open}. The algorithm alternates between adding high affinity genes to C_{open}, and removing low affinity genes from C_{open}. C_{open} is *closed* when no more genes can be added to or removed from it. Once a cluster is closed, it is not considered any more by the algorithm. The algorithm iterates until all the genes have been assigned to clusters and the current C_{open} is closed.

When a new cluster C_{open} is started, the initial affinity of all genes are 0 since C_{open} is empty. One additional heuristic that the authors [Ben-Dor and Yakhini, 1999] implemented in their software BIOCLUST is to choose a gene with the maximum number of neighbors to start a new cluster. Another heuristic is that after the CAST algorithm converges, there is an additional iterative step, in which all clusters are considered at the same time, and genes are moved to the cluster with the highest average similarity.

2.2 Hierarchical average-link

Agglomerative hierarchical algorithms build clusters bottom up. Initially, each object is in its own cluster. In each step, the two clusters with the greatest cluster similarity are merged. This process is repeated until the desired number, k, of clusters is produced. In average-link, the cluster similarity criterion is the average pairwise similarity between objects in the two clusters. Refer to [Jain and Dubes, 1988] and [Anderberg, 1973] for detailed discussions on hierarchical algorithms.

2.3 K-means

The number of clusters, k, is an input to the k-means clustering algorithm. Clusters are described by *centroids*, which are cluster centers, in the algorithm. In our implementation of k-means [Jain and Dubes, 1988], the initial centroids consist of the clustering results from average-link. Each object is assigned to the centroid (and hence cluster) with the closest Euclidean distance. New centroids of the k clusters are computed after all objects are assigned. The steps of assigning objects to centroids and computing new centroids are repeated until no objects are moved between clusters.

2.4 Similarity metrics

There are two popular similarity metrics used in the gene expression analysis community: Euclidean distance (for example, [Wen *et al.*, 1998]) and correlation coefficient (for example, [Eisen *et al.*, 1998]). In our experiments, we evaluated the effectiveness of PCA on clustering analysis with both Euclidean distance and correlation coefficient, namely, CAST with correlation coefficient[1], average-link with both correlation and distance, and k-means with both correlation and distance.

If Euclidean distance is used as the similarity metric, the minimum number of components in sets of PC's (m_0) considered is 2. If correlation is used, the minimum number of components (m_0) considered is 3 because there are at most 2 clusters if 2 components are used.

When there are 2 components, the correlation coefficient is either 1 or -1. Suppose there are two genes g_1 and g_2 with two components. Let $x_{i,j}$ (where $i, j = 1, 2$) be the expression level of gene i under component j. The correlation coefficient between g_1 and g_2 can be simplified to:

$$\frac{(x_{1,1} - x_{1,2}) * (x_{2,1} - x_{2,2})}{\sqrt{(x_{1,1} - x_{1,2})^2 * (x_{2,1} - x_{2,2})^2}} \tag{3}$$

Since the denominator in Equation 3 represents the product of the norms of genes g_1 and g_2, the denominator must be positive. From Equation 3, the correlation coefficient between genes g_1 and g_2 is 1 if $(x_{1,1} - x_{1,2}) * (x_{2,1} - x_{2,2}) > 0$, the correlation coefficient is -1 if $(x_{1,1} - x_{1,2}) * (x_{2,1} - x_{2,2}) < 0$. If $x_{1,1} = x_{1,2}$ or $x_{2,1} = x_{2,2}$, the correlation coefficient is undefined. Since there are only two possible values that the correlation coefficient can take when there are two components, there are at most two clusters.

References

[Anderberg, 1973] Anderberg, M. R. (1973) *Cluster analysis for applications*. Academic Press.

[Ben-Dor and Yakhini, 1999] Ben-Dor, A. and Yakhini, Z. (1999) Clustering gene expression patterns. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France.

[Eisen *et al.*, 1998] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA*, **95**, 14863–14868.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 193–218.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

[Milligan and Cooper, 1986] Milligan, G. W. and Cooper, M. C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, **21**, 441–458.

[1] When Euclidean distance is used in CAST, the algorithm usually does not converge in practice.

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

[Wen *et al.*, 1998] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. and Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Science USA*, **95**, 334–339.