A Comparison of Five Fuzzy Rand Indices

Derek T. Anderson[1], James C. Bezdek[1], James M. Keller[1], and Mihail Popescu[2]

[1] Electrical and Computer Engineering Department, University of Missouri Columbia,
MO, 65211, USA
[2] Informatics Institute, University of Missouri Columbia,
MO, 65211, USA
dtaxtd@mail.missouri.edu, jcbezdek@gmail.com,
{kellerj,popescum}@missouri.edu

Abstract. Five papers have appeared in the last three years that propose different fuzzy generalizations of Rand's classical comparison index for crisp clustering algorithms. We review the five generalizations, compare their complexities, and then give two numerical examples to compare their performance. Our extension (for the pairwise agreements) is ***O***(n), while the other four generalizations are ***O***(n^2).

Keywords: Cluster validity, Rand index, Fuzzy Rand Index.

1 Introduction

Let O={o$_1$,…,o$_n$} denote n objects (fish, cigars, motorcycles, beers, etc.). When each object in O is represented by a (column) vector **x**, the set $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\} \subset \Re^p$ is an *object data representation* of O. When each object in o$_i$ ∈ O has a *physical label*, O is a set of *labeled data*; otherwise, O is unlabeled. Let integer c denote the number of classes, 1 < c < n. Clustering in unlabeled data is the assignment of one of four types of labels to each object in O. The label vectors of the objects are the columns of c-partitions of O, which are sets of (cn) values {u$_{ik}$} that can be conveniently arrayed as (c ×n) matrices, say U = [u$_{ik}$]. The three sets are:

$$M_{pcn} = \left\{ U \in \Re^{cn} : u_{ik} \in [0,1] \forall\, i,k;\ 0 < \sum_{k=1}^{n} u_{ik}\ \forall\, i \right\} = \text{possibilistic c-partitions;} \quad (1a)$$

$$M_{fcn} = \left\{ U \in M_{pcn} : \sum_{i=1}^{c} u_{ik} = 1 \forall\, k \right\} = \text{fuzzy or probabilistic c-partitions;} \quad (1b)$$

$$M_{hcn} = \left\{ U \in M_{fcn} : u_{ik} \in \{0,1\} \forall i,k \right\} = \text{crisp or hard c-partitions.} \quad (1c)$$

It is convenient to have a single name for the set $M_{pcn} - M_{hcn}$, which contains the fuzzy, probabilistic, and possibilistic c-partitions of O. We call $M_{pcn} - M_{hcn}$ the *soft c-partitions* of O. Clustering algorithms map $X \subset \Re^p$ or $R \subset \Re^{nn} \mapsto M_{pcn}$. Let CP = {U$_i$: 1 ≤ i ≤ N} denote N different *candidate partitions* of a fixed object set O that may arise as a result of clustering (X or R) with one algorithm at various values of its

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 446–454, 2010.

parameters; or more generally, with different algorithms, each with its own parameters. Which $U \in CP$ best explains and represents the (unknown) structure in O? This article is about one method for answering this question. Many other methods are nicely discussed in [1-4].

One group of methods for this problem use *comparison indices*, s(U,V). There are various ways to use such indices [5, 6]. The only application we consider in this note is when U is an algorithmically obtained partition, and V is a *reference partition* that purports to represent the "true cluster structure" in O. In this case s(U,V) measures the extent to which U's in CP recover or retrieve the "true" clusters in O, and hence, the sizes of U and V are equal.

We never have an external reference partition in a real clustering situation which, by definition, involves unlabeled data. So, why do this at all? Well, the only way you can evaluate *any* clustering algorithm before using it in a real situation is to see how well it recovers "true but unknown" reference partitions. If nothing else, good recovery rates on data with "known" cluster structure at least provide some psychological reassurance that the clustering algorithm *can* sometimes recover "good clusters".

2 Comparison Indices and the Contingency Table for (U,V)

Let $U \in M_{hrn}$ and $V \in M_{hcn}$ be crisp partitions of O. U and V need not possess the same number of clusters, $r \neq c$. The four classical combinations for pairs of objects from $O \times O$ in clusters of U and V are: (i) paired in U and V; (ii) not paired in U nor in V; (iii) paired in V but not in U; and (iv) paired in U but not in V [6, p. 194]. The comparison of U to V with a similarity measure s begins with the $r \times c$ contingency matrix $N = UV^T$ shown in Table 1 that contains counts of the number of occurrences of each of the four types over the n(n-1)/2 distinct, unordered pairs in $O \times O$. Entry n_{ij} is the number of objects common to classes U_i and V_j.

Table 1. The contingency matrix N

Partition V: V_j = row j of V

	Class	V_1 V_2 ••• V_c	Sums
Partition U U_i = row i of U	U_1 U_2 ⋮ U_r	$N = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1c} \\ n_{21} & n_{22} & \cdots & n_{2c} \\ \vdots & \vdots & & \vdots \\ n_{r1} & n_{r2} & \cdots & n_{rc} \end{bmatrix} = UV^T$	$n_{1\bullet}$ $n_{2\bullet}$ ⋮ $n_{r\bullet}$
	Sums	$n_{\bullet 1}$ $n_{\bullet 2}$ $\cdots$ $n_{\bullet c}$	$n_{\bullet\bullet} = n$

The building blocks of many similarity measures for s(U,V) are the four equations (2a)-(2d). These four equations simply count the number of occurrences amongst the n(n-1)/2 pairs of each of the four types of unordered pairs.

$$a = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(n_{ij}-1)\text{; number paired in U and V;} \tag{2a}$$

$$d = \frac{1}{2}\left(n^2 + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2 - (\sum_{i=1}^{r} n_{i\bullet}^2 + \sum_{j=1}^{c} n_{\bullet j}^2)\right)\text{; number paired in neither U nor V;} \tag{2b}$$

$$b = \frac{1}{2}\left(\sum_{j=1}^{c} n_{\bullet j}^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2\right)\text{; number paired in V, not U;} \tag{2c}$$

$$c = \frac{1}{2}\left(\sum_{i=1}^{r} n_{i\bullet}^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2\right)\text{; number paired in U, not V.} \tag{2d}$$

The sums (a+d) and (b+c) are usually interpreted, respectively, as (the total number of) *agreements* and *disagreements* between U and V. Anderson et al. [17] tabulate a [non-exhaustive] list of 14 coefficients that have been proposed for s(U,V) based on functions of a, b, c and d; Sokal and Sneath [8] list many others. In this note, the only index we consider is Rand's index, the classical form of which is

$$s_r(U,V) = (a+d)/(a+b+c+d)\,. \tag{3}$$

3 Generalizing s_r(U,V) When U and/or V Are Soft Partitions

Rand's index first appeared in Sokal and Michener in 1958, where it was called a simple matching coefficient [8]. Rand reintroduced this function in 1971 [5], and the literature has consistently referred to it as "Rand's Index" since then. The resurgence of Rand's index in bioinformatics [9-12] has renewed interest in generalizing it, along with some of the other comparison indices based on the elements in Table 1, to various non-crisp cases. Specifically, we mention the papers (in chronological order) of Campello [13, 2007], Frigui et al. [14, 2007], Brower [15, 2009], Hullermeier and Rifqi [16, 2009], and Anderson et al. [17, 2010]. All of these papers generalize the Rand index to the case of U and/or V being fuzzy partitions of the n objects. Next, we briefly review the method used to generalize (3) in each of these five articles.

3.1 Campello [13]

Campello presents a method for fuzzifying the indices of Rand, Jaccard, Fowlkes-Mallow, Hubert and (one version of) the adjusted Rand. Campello's scheme is based on writing equations (2) in an equivalent form using (cardinalities of) intersections of the crisp subsets of O×O corresponding to each of the four totals, and then replacing the crisp sets with fuzzy ones. Campello's generalization of equation (2a) is:

$$\underbrace{a}_{\text{Campello}} = \sum_{j=1}^{i-1}\sum_{i=2}^{n}\left(\left(\bigvee_{k=1}^{r}\left(u_{ki}\wedge u_{kj}\right)\right)\wedge\left(\bigvee_{k=1}^{c}\left(v_{ki}\wedge v_{kj}\right)\right)\right); \qquad (4)$$

3.2 Frigui et al. [14]

These authors present generalizations of the Rand, Jaccard, Fowlkes-Mallow and Hubert indices. They address only the special case where U and V are both $c\times n$ partitions of O and V is a crisp reference partition. Instead of using the elements from Table 1 to compute equations (2), they first convert U and V into $n\times n$ coincidence matrices, $U^* = U^T U$; $V^* = V^T V$. Frigui et al.'s generalization of equation (2a) is:

$$\underbrace{a}_{\text{Frigui etal.}} = \sum_{j=2}^{n}\sum_{k=1}^{j\text{-}1}\left(\sum_{i=1}^{c}u_{ij}u_{ik}\right)\left(\sum_{i=1}^{c}v_{ij}v_{ik}\right); \qquad (5)$$

If product is used for t-norm and sum is used in place of t-conorm, then Frigui's approach is in effect Campello's [15]. However, Campello's approach is more general. It applies to the cases of fuzzy, probabilistic, and possibilistic U and/or V.

3.3 Brouwer [15]

Brouwer discusses another generalization of the Rand, a (third variant of) the adjusted Rand, and Jaccard's index. His approach is also based on formulating the two potentially very large $n\times n$ (bonding) matrices, $U^* = U^T U$; $V^* = V^T V$. However, instead of dot product for constructing bonding matrix terms, he instead uses cosine correlation, i.e. $u^*_{i,j} = \cos(\mu_i^j)$, where $\cos(\mu_i^j)$ is the angle between vectors $\mathbf{U}_i^T$ and $\mathbf{U}^j$. In the sequel, $\mathbf{A}_i$ and $\mathbf{A}^j$ denote the vectors corresponding to the i-th row and j-th column of any matrix A, and $\langle \mathbf{A}_i, \mathbf{A}^j\rangle$ is the dot product of these two vectors. Brouwer's generalization of equation (2a) is:

$$\underbrace{a}_{\text{Brouwer}} = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\langle U_i^T, U^j\rangle}{\|U_i^T\|\|U^j\|}\frac{\langle V_i^T, V^j\rangle}{\|V_i^T\|\|V^j\|}\right)}{2} - \frac{n}{2} = \frac{\left(\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\cos(\mu_i^j)\cos(v_i^j)\right)\right) - n}{2}. \qquad (6)$$

3.4 Hullermeier and Rifqi [16]

These authors consider only the Rand index. They argue that Campello's fuzzy Rand index is in some sense defective because it is not a metric. They do not formulate their index in terms of equations (2a)-(2d). Instead, their generalization is guided by the fact that Rand's index counts the number of paired agreements (a+d) divided by the total number of possible pairs (a+b+c+d), and this leads them to a direct generalization of the Rand index:

$$s_{\mathrm{FRHR}}(U,V)=1-\left[\sum_{j=i+1}^{n}\sum_{i=1}^{n-1}\left|\left\|V^{i}-V^{j}\right\|-\left\|U^{i}-U^{j}\right\|\right| \Big/ \binom{n}{2}\right]. \tag{7}$$

3.5 Anderson et al. [17]

This paper provides generalizations for 14 comparison indices. It begins by forming the contingency matrix as $N = UV^T$. Anderson et al. note that a modification is needed to accommodate the case when U and/or V are possibilistic. In the possibilistic case, we can have $\sum_{i=1}^{r} n_{i\bullet} > n$, or $\sum_{j=1}^{c} n_{\bullet j} > n$. One or both of the terms $\sum_{i=1}^{r} n_{i\bullet}^{2}$ and $\sum_{j=1}^{c} n_{\bullet j}^{2}$ can make d at (2b) relatively large and negative. Depending on b at (2c) and c at (2d), the (soft) Rand index can result in $s_r(U,V) < 0$ or >1. To remedy this, they scale N with $\phi = n/\sum_{i=1}^{r} n_{i\bullet}$ or $\varphi = n/\sum_{j=1}^{c} n_{\bullet j}$. Since $\sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j}$, $\varphi = \phi$. These authors base their generalization on $N^* = \phi UV^T = [n/\sum_{i=1}^{r} n_{i\bullet}]UV^T$. An advantage of this scaling is that when U and V are crisp, fuzzy or probabilistic partitions, $\phi=1$, thus $N^* = N = UV^T$. This shows that *ANY* index based on only the elements of Table 1 will reduce to the original index when U and V are both crisp partitions of the n objects; and otherwise, they will be valid soft generalizations of those indices. Moreover, in the case of possibilistic partitions, the normalization produces index values in the range [0,1]. Anderson et al.'s generalization of (2a) using $N^* = \phi UV^T$ is:

$$\underbrace{a}_{\mathrm{Anderson,et\,al.}} = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{n\left\langle U_i,(V^T)^j\right\rangle}{\sum_{k=1}^{r}\sum_{p=1}^{c}\left\langle U_k,(V^T)^p\right\rangle}\right)\left(\left(\frac{n\left\langle U_i,(V^T)^j\right\rangle}{\sum_{k=1}^{r}\sum_{p=1}^{c}\left\langle U_k,(V^T)^p\right\rangle}\right)-1\right). \tag{8}$$

4 Examples

The following two examples compare the five fuzzy Rand indices in Sections 3.1-3.5. Candidate fuzzy partitions are generated by the fuzzy c-means (FCM, [1]) algorithm using the *fcm* function from the MATLAB Fuzzy Logic Toolbox with c = 2, 3,…,10, m = 2, maximum number of iterations MAXIT = 100, objective function error EPS=1e-5 and random partition initialization. Scatterplots of the two data sets, X1 and X2, are shown in Figure 1.

In data set X1, each cluster of 500 points is a sample from a mixture of c=6 equiprobable Gaussian distributions in two dimensions. Means of the six component densities are: $[20\ 20]^T$, $[1\ 40]^T$, $[20\ 70]^T$, $[40\ 6]^T$, $[40\ 30]^T$, and $[60\ 50]^T$, and the common covariance matrix was $\Sigma=1.7I_6$. These clusters are fairly compact and well-separated. We expect the best partition in CP to occur at c = 6. The reference partition V is the crisp 6x3000 partition with six "diagonal" blocks of 500 1's in each row.

Data set X2 is three well separated parallel line clusters of size 200 each. Samples are generated according to $\vec{c}+\alpha(\vec{d}-\vec{c})$, where $\alpha\in[0,1]$ is a uniformly distributed random number and $(\vec{c},\vec{d})$ are line segment endpoints. Clusters in X2 include: ($[1\ 1]^T$, $[1\ 10]^T$), ($[6\ 2]^T$, $[6\ 6]^T$), and ($[10\ 0]^T$, $[10\ 12]^T$). We expect the best partition in CP to occur at c = 3. The reference partition V is the crisp 3x600 partition with three "diagonal" blocks of 200 1's in each row.

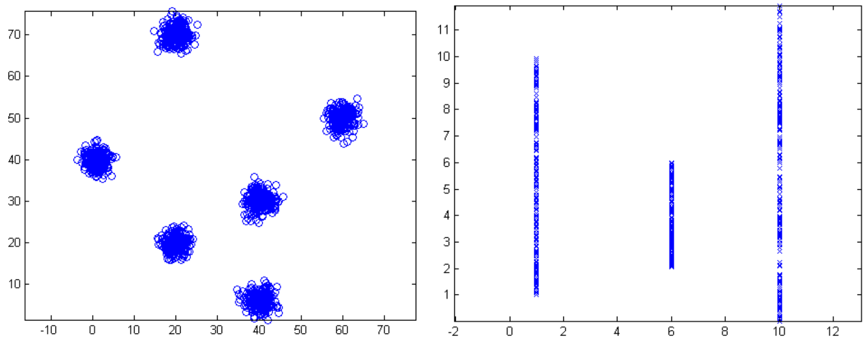

Fig. 1. Gaussian (X1) and parallel line (X2) data sets used for the two examples

Figure 2 shows graphs of the five Fuzzy Rand indices for terminal FCM partitions on X1 as c varies from 2 to 10, so there are 9 candidate partitions in CP. This graph shows two things: first, the five indices are indeed different; and second, they have similar values on this well behaved data set. All five indices have clear maximums at c=6 which points to the most preferable partition in CP. Figure 2 might tempt you to

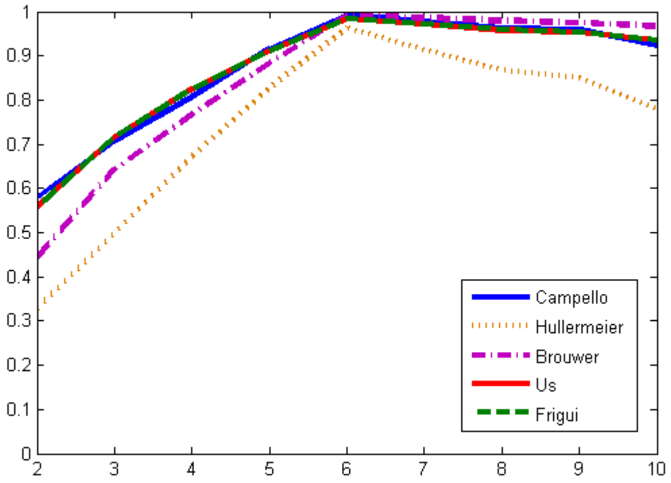

Fig. 2. Comparison of the five fuzzy Rand indices on FCM

conjecture that Hullermeier and Rifki is bounded above by the other four indices, but we have not attempted a proof of this. The other indices all cross each other.

Figure 3 shows graphs of the five Fuzzy Rand indices for terminal FCM partitions on X2 as c varies from 2 to 10. The graph again shows that the five indices are indeed different and they have similar values on this data set. We expect FCM to fail on this example. The crisp V for c = 3 will NOT match well with the FCM c = 3 partition. As expected, no indices have a clear maximum at c = 3. Hullermeier and Rifki have a maximum at c = 4 and the others are at c = 5. Also, Hullermeier and Rifki appear to again be bounded above by the other four indices, which all cross each other.

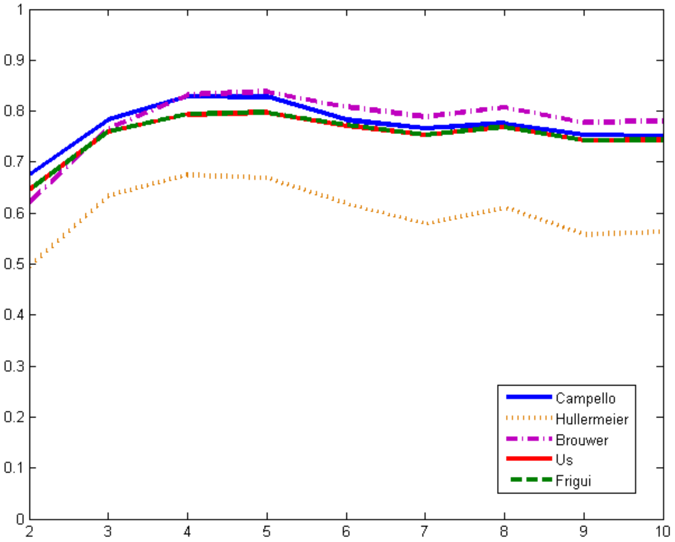

Fig. 3. Comparison of the five fuzzy Rand indices on FCM

5 Computational Complexity

Assuming similar cost for different operations, the cost of evaluation of formula (2a) for all but the index of Hullermeier and Rifqi (who form the fuzzy Rand index directly) are reported in Table 2. The example in the last column of Table 2 shows that in terms of computational costs, Anderson et al.'s method is (at least 3 and at best 5) orders of magnitude less than the other generalizations of Rand's index. Computation of the Rand index at (3) involves calculation of all four equations, (2a)-(2d). Combining the factors as in (3) uses only addition and subtraction, and will cost all methods equally. Hence, we can extend the results of Table 2 from just equation (2a) to equation (3) without loss.

Table 2. Computational complexity for the five fuzzy Rand indices

Method	**Computational Complexity**		**n=1000, r = c = 5**
Anderson et al. [17] for (2a)	$\mathbf{O}(2rcn+3rc)$	$\mathbf{O}(n)$	50,075
Brouwer [15] for (2a) (assuming cosine as a single operation)	$\mathbf{O}(2n^2+n+2)$	$\mathbf{O}(n^2)$	2,001,002
Brouwer [15] for (2a) (using dot product and magnitude form)	$\mathbf{O}(4n^2+ rn^2+ cn^2-n+3rn+3cn+2)$	$\mathbf{O}(n^2)$	14,029,002
Campello [13] for (2a)	$\mathbf{O}(rn^2+cn^2+(n-n^2)/2-rn-cn)$	$\mathbf{O}(n^2)$	9,490,500
Hullermeier and Rifqi [16] (for the fuzzy Rand)	$\mathbf{O}((3rn^2+3cn^2-3rn-3cn)/2+5)$	$\mathbf{O}(n^2)$	14,985,005
Frigui et al. [14] for (2a)	$\mathbf{O}(rn^2+cn^2+(n-n^2)/2-rn-cn)$	$\mathbf{O}(n^2)$	9,490,500

6 Discussion and Conclusions

We compared our generalization of the classical Rand index with four other fuzzy generalizations of it both experimentally, and in terms of computational complexity. Our extension of the Rand index is $\boldsymbol{O}(n)$, while the other four are all $\boldsymbol{O}(n^2)$. More examples using different types of data, algorithms and other indices that involve comparing partitions appear in [17]. The advantage of using $N^* = \phi UV^T$ is that this formulation directly generalizes *all* indices that depend only on equations (2) to every combination of (U, V). There are sixteen possible pair types according as each of U, V are crisp, fuzzy, probabilistic or possibilistic, so we have, for example, 16 Rand indices, 16 Jaccard indices, and so on. Each formula there is recovered when U and V are crisp, i.e., these are true generalizations to every case - by definition.

The use of comparison indices for validation of clustering algorithms has the significant advantage of being independent of the correspondence problem for comparing clustering solutions to known reference partitions. When U is soft, one approach to retrieval assessment is to first harden any soft partition U. Then the hardened version of U, say H(U), defines the function $s_e(H(U),V) = 1 - \sum_{k=1}^{n} \left\| [H(U)]^k - V^k \right\|_1 \Big/ 2n$, which counts the number of label matches. This comparison method is *similar* to assessment by s(U,V), but before using $s_e(H(U),V)$, we *must* register the reference clusters to their algorithmic counterparts. This complication is avoided by s(U,V), because the indices in Table 2 depend only on the values in Table 1; double sums, row sums, or column sums of the entries of $N=UV^T$. Consequently, crisp comparison indices such as Rand's index are *independent of the correspondence problem* which plagues evaluation of retrieval success for soft clustering algorithms by the "harden and count" method represented by $s_c(H(U), V)$. This important advantage for the comparison index method remains true even when $r \neq c$ and the resubstitution error rate cannot even be computed!

Acknowledgments

Derek Anderson is a pre-doctoral biomedical informatics research fellow funded by the National Library of Medicine (T15 LM07089).

References

1. Bezdek, J.C., Keller, J.M., Krishnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer, Norwell (1999)
2. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
3. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, NY (2006)
4. Hoppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis. Wiley and Sons, Chichester (1999)
5. Rand, W.M.: Objective criteria for the evaluation of clustering methods. JASA 66(336), 846–850 (1971)
6. Hubert, L.J., Arabie, P.: Comparing partitions. J. Classification 2, 193–218 (1985)
7. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy - The Principles and Practice of Numerical Classification. W. H. Freeman, San Francisco (1973)
8. Sokal, R.R., Michener, C.D.: A Statistical Method for Evaluating Systematic Relationships. The University of Kansas Scientific Bulletin 38, 1409–1438 (1958)
9. Duan, F., Zhang, H.: Correcting the loss of cell-cycle synchrony in clustering analysis of microarray data using weights. Bioinformatics 20(11), 1766–1771 (2004)
10. Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.: Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics 22(19), 2405–2412 (2006)
11. Wong, D.S.V., Wong, F.K., Wood, G.R.: A multi-stage approach to clustering and imputation of gene expression profiles. Bioinformatics 23(8), 998–1005 (2007)
12. Yu, Z., Wong, H.S., Wang, H.: Graph-based consensus clustering for class discovery from gene expression data. Bioinformatics 23(21), 2888–2896 (2007)
13. Campello, R.J.G.B.: A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Patt. Recog. Letters 28, 833–841 (2007)
14. Frigui, H., Hwang, C., Rhee, F.: Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition (40), 3053–3068 (2007)
15. Brower, R.K.: Extending the Rand, adjusted Rand, and Jaccard indices to fuzzy partitions. J. Intell. Inf. Systems 32, 213–235 (2009)
16. Hullermeier, E., Rifqi, M.: A fuzzy variant of the Rand index for comparing clustering structures. In: Proc. IFSA, Lisbon, Portugal, pp. 1–6 (2009)
17. Anderson, D., Bezdek, J.C., Keller, J.M., Popescu, M.: Comparing soft partitions. IEEE Trans. Fuzzy Systems (2010) (in review)