Fuzzy forms of the Rand , Adjusted Rand and Jaccard Indices for Fuzzy Partitions of Gene expression and other data

Roelof K Brouwer[1] Senior Member IEEE

Thompson Rivers University

Kamloops Canada

rkbrouwer@ieee.org

***Abstract**: Clustering is one of the most basic processes that are performed in simplifying data and expressing knowledge in a scientific endeavor. Clustering algorithms have been proposed for the analysis of gene expression data with little guidance available to help choose among them however. Since the output of clustering is a partition of the input data, the quality of the partition must be determined. This paper presents fuzzy extensions to some commonly used clustering measures including the rand index (RI), adjusted rand index(ARI) and the jaccard index(JI) that are already defined for crisp clustering. Fuzzy clustering, and therefore fuzzy cluster indices, is beneficial since it provides more realistic cluster memberships for the objects that are clustered rather than 0 or 1 values. If a crisp partition is still desired the fuzzy partition can be turned in to a crisp partition in an obvious manner. The usefulness of the fuzzy clustering in that case is that it processes noise better. These new indices proposed in this paper, called FRI, FARI, and FJI for fuzzy clustering, give the same values as the original indices do in the special case of crisp clustering. Through use in fuzzy clustering of artificial data and real data, including gene expression data, the effectiveness of the indices is demonstrated.*

***Keywords**: Fuzzy clustering; Rand index; Adjusted Rand index; Jaccard index; external cluster quality measures.*

1 Introduction

To understand complicated biological systems, huge quantities of gene expression data has been generated by researchers[1, 2]. Clustering is a useful exploratory technique for analysis of this data [3, 4]. The first stage of knowledge acquisition and reduction of complexity concerning a group of objects is to partition or divide the objects into groups based on their attributes or characteristics. Partitioning objects is one of the most fundamental modes of understanding and learning [5] . This process, called clustering [6], is a form of unsupervised learning whereby similar objects are "put into" the same group or cluster. For example gene clustering is useful for discovering groups of correlated genes potentially co-regulated or associated with a disease such as cancer. The objective of clustering algorithms is to partition the genes into groups exhibiting similar patterns of variation in expression level [3].

Partitions can be crisp or fuzzy although a crisp partition is really a special case of a fuzzy partition. In a fuzzy partition the elements belong to the subsets in a partition to varying degrees whose values lie in [0.1]. The subsets are fuzzy while the total set that is partitioned is crisp. A fuzzy partition induces a crisp partition if the maximum membership value for each object over the various clusters is replaced by a 1 and all other values are replaced by a 0. Even if the objective is to obtain a crisp partition it is still useful to use a fuzzy clustering process as an improved way of handling noise.

[1] Department of Mechanical and Mechatronics Engineering (Visiting Professor) University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa. and Professor Department of Computing Science ,Thompson Rivers University 900 McGill Road Kamloops, BC V2C 5N3Canada E-mail: rkbrouwer@ieee.org

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008,
Supplementary Conf. Proc., pp.104-114, 2008.

To permit comparison of clustering techniques a measure of clustering quality is required. Comparisons of clustering methods[7] are required for example in genomics. The quality of a partition can be measured in various ways. The measures discussed here require knowledge of the correct class of the objects in order to calculate a quantitative quality measure of the clustering result.

The commonly accepted measures of crisp clustering quality, based on comparing a found partition to a given partition, to be discussed here, are defined in terms of 4 parameters *a*,*b*,*c*, and *d* as shown in reference [8] and repeated here.In this case we have: a: number of object pairs whose elements are in the same cluster in partition P1 and also in P2; b: number of object pairs whose elements are in the same cluster in P1 but are in different clusters in P2; c: number of object pairs whose elements are in the same cluster in P2 but are in different clusters in P1. d: number of object pairs whose elements are in different clusters in P1 and are also in different clusters in P2. They are essentially counts of pairs of numbers in the crisp case. In this paper they are redefined so that they can be used for fuzzy partitions and yet yield the same results for crisp partitions.

The rand index(*RI*)[8] is defined as

$$RI = \frac{a+d}{a+b+c+d} \tag{1}$$

It lies in [0, 1] with 0 indicating that the two partitions do not agree on any pair of elements and 1 indicating that the two partitions are exactly the same.

The adjusted rand index (*ARI*) [9] corrects the *RI* to give a constant expected value of 0 and may be calculated according to the formula (2).

$$ARI = \frac{2\left(a\times d - b\times c\right)}{c^2 + b^2 + 2\times a\times d + \left(a+d\right)\times\left(c+b\right)} \tag{2}$$

ARI lies in [-1, 1]. To permit direct comparison with *RI*, *ARI* can be converted to $ARI' = \frac{ARI+1}{2}$ without loss of information. In that case it also lies in [0, 1].

Another commonly used method for the comparison of partitions is the Jaccard index(*JI*) [10] defined as:

$$JI = \frac{a}{a+b+c} \tag{3}$$

The Jaccard index lies in [0, 1]. When it is 0 no two elements are together simultaneously in partition $\boldsymbol{P}^{(1)}$ and partition $\boldsymbol{P}^{(2)}$. If the partitions $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ are equivalent then *b*=*c*=0 and the Jacquard index is 1. It makes use of less information than either *RI* or *ARI*. It well known in the clustering community that for comparing two crisp partitions the adjusted rand index possesses the most desirable properties [9, 11, 12].

In this paper a relationship called bonding is defined between two objects that describes the degree to which the two objects are in the same cluster or class. In case of crisp partitions this is always 0 or 1. Two partitions may then be compared on the basis of the bonding similarity in case of both crisp and fuzzy clusters

To allow variable names of more than one letter and thereby permit variable names to be neumonic, all operations are denoted by explicit operators. Multiplication, for example, is defined explicitly using the operator ×. Implicit multiplication will not exist and *na* for example is just a variable name. Names of arrays are in bold font. Rank-1 array variables are in lower case and names of arrays of rank greater than 1 are in capital. $\boldsymbol{A}_{i,.}$ and $\boldsymbol{A}_{.,i}$ and means the i^{th} row and i^{th} column repectively. Division and multiplication between arrays are generally between the components of the arrays.

2 Bonding of Objects in a crisp or fuzzy partition

Partitioning induces a relationship between objects in that two objects are related if they are in the same set of the partition. We call this relationship a bonding. This can be extended to fuzzy partitions as follows.

Definition 1Bonding between objects in a fuzzy partition

Given the membership vectors for the objects and the fuzzy partition a measure of bonding between two elements with fuzzy membership vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ is given by the cosine correlation (cc)

$$b(\boldsymbol{v},\boldsymbol{w}) = \frac{\boldsymbol{v}\bullet\boldsymbol{w}}{|\boldsymbol{v}|\times|\boldsymbol{w}|} \quad \in \quad [0,1] \tag{4}$$

The cosine correlation (abbreviated as *cc* below) between two lists is often used to measure the similarity between two vectors. (Correlation based similarity measures have been widely used in the micro-array literature [4, 13].) A bonding matrix, $\boldsymbol{B}$, for a fuzzy partition or fuzzy membership matrix, $\boldsymbol{M}$, is defined by

Definition 2The bonding matrix *B* for a partition defined by membership matrix *M is*

$$\boldsymbol{B}_{i,j} = cc(\boldsymbol{M}_{i,.}, \boldsymbol{M}_{j,.}) \qquad i, j = 1..ne$$

cc means cosine correlation. $\boldsymbol{M}_{i,.}, \boldsymbol{M}_{j,.}$ correspond to the i^{th} and j^{th} row of $\boldsymbol{M}$ respectively corresponding to membership values for the i^{th} and j^{th} object respectively.

The rows of $\boldsymbol{M}$, $\boldsymbol{M}_{i,.}$, correspond to objects and columns correspond to clusters. Here we see that $\boldsymbol{B}$ is the similarity matrix for the membership vectors in a fuzzy partition. If *a,b,c,* and *d* are defined as in Definition 3 we get the same result for the commonly known indices, *a*, *b*, *c*, and *d*, as they are defined in [8].

Definition 3 Values of *a,b,c,* and *d* for fuzzy partitions and as a special case also for crisp partitions are

$$a = g\left(\boldsymbol{B}^{(1)} \times \boldsymbol{B}^{(2)T}\right) \tag{6}$$

$$b = f\left((1 - \boldsymbol{B}^{(1)}) \times \boldsymbol{B}^{(2)T}\right) \tag{7}$$

$$c = f\left(\boldsymbol{B}^{(1)} \times (1 - \boldsymbol{B}^{(2)})\right) \tag{8}$$

$$d = f\left((1 - \boldsymbol{B}^{(1)}) \times (1 - \boldsymbol{B}^{(2)})\right) \tag{9}$$

$f(\boldsymbol{X})$ represents the function of a square array, $\boldsymbol{X}$, as defined in (10).

$$f(\boldsymbol{X}) = \frac{\sum_{i,j} \boldsymbol{X}_{i,j}}{2} \tag{10}$$

$g(\boldsymbol{X})$ represent the function of a square array of dimension, *n*, as defined in (11).

$$g(\boldsymbol{A}) = \frac{\sum_{i,j} \boldsymbol{X}_{i,j} - n}{2} \tag{11}$$

The product between the matrices is the component wise product.

We now have a definition of the parameters for the *RI*, *ARI*, and *JI* indices for fuzzy partitions and crisp partitions that gives the same values for the indices in the case of crisp partitions as before. The new indices will be referred to as *FRI*, *FARI*, and *FJI*.

3 Clustering method to which the quality measures are applied

Usefulness of the indices, *FRI*, *FARI*, and *FJI* may be demonstrated through use in fuzzy clustering. The measures are tested by producing fuzzy partitions through fuzzy c-means (FCM) [6, 14] , the traditional fuzzy clustering method that will briefly be described next. The indices are calculated by comparing two partitions, one is a class partition and the other is the clustering partition obtained during clustering.

FCM consists of repeatedly determining prototypes for the clusters to be found and calculating membership values for the objects in the clusters. Every attribute value in a prototype is the weighted mean over all members of the data set to be clustered, with weights equal to a power of the degree to which a pattern belongs to a cluster.

Formally, let the weights or membership values for clusters be designated by $\boldsymbol{M}_{p,k}$; the membership of pattern *p* in cluster *k*. Also let $\boldsymbol{Y}_{p,a}$ represent the parameters for the attribute values themselves within objects. The a^{th} attribute of the prototype for cluster *k* is

$$\boldsymbol{P}_{k,a} = \frac{\sum_{p=0}^{np-1} \boldsymbol{M}_{p,k}^{m} \times \boldsymbol{Y}_{p,a}}{\sum_{p=0}^{np-1} \boldsymbol{M}_{p,k}^{m}} \tag{12}$$

Here, the subscripts have the following meaning

$p = 1..np$	identifies patterns
$a = 1..na$	identifies attributes
$k = 1..nk$	identifies clusters

The cluster membership values are calculated in terms of distance between object representations and the prototypes as in

$$\boldsymbol{M}_{p,k} = \frac{\frac{1}{\boldsymbol{D}_{p,k}^{\frac{2}{m-1}}}}{\sum_{l=1}^{nk} \frac{1}{\boldsymbol{D}_{p,l}^{\frac{2}{m-1}}}} \tag{13}$$

$\boldsymbol{D}_{p,k}$ is the distance between object *p's* representation and prototype *k*. *m* is a measure of fuzziness in the clustering. There are two popular similarity metrics for example in the gene expression analysis domain for clustering [3] : Euclidian distance (for example, [1]) and correlation coefficient (for example [4]). Here the Euclidian distance is used in performing the clustering.

4 Simulations

To determine the usefulness of *FRI*, *FARI*, and *FJI* they are applied in the fuzzy clustering of artificial and real data. The indices are evaluated both during clustering and at the end of clustering. Data for which classes are known are clustered and the partition due to clustering is compared to the class partition using the measures defined in this paper. As clustering proceeds the values of the quality measures are recorded. The values obtained in this way are then examined for sensibility.

Three different partition matrices are considered in each case. There is the class partition matrix or discriminant matrix, the fuzzy membership matrix and the fuzzy membership matrix induced crisp partition matrtix. The crisp membership matrix is obtained from the fuzzy membership matrix by replacing maximums in each row by one's and the other components by zeroes. A confusion matrix [8], C, can be obtained for two crisp partitions as defined by

$$\boldsymbol{C}_{i,j} = \#\, objects\ \ simultaneously\ \ in\ \ class\ \ i\ \ of\ \ partition\ 1\ \ and\ class\ j\ of\ \ partition\ 2 \tag{14}$$

In the comparison tables, *FARI* is replaced by $FARI' = \frac{1 + FARI}{2}$ to map its values into the same range as the others. In the comparison tables cp means class partition, fcp means fuzzy cluster partition, and ccp means crisp cluster partition.

4.1 Values of indices obtained during clustering of artificial data

In the artificial data case each data set is generated from four bi-variate normal distributions. Each of the four bi-varite distributions generates a class which is used in the comparisons. The data sets vary in the distinctiveness of the clusters, as can be seen from the plots, from having very distinct clusters to having no apparent partition at all.

4.1.1 Data sets

4.1.1.1 Data set 1

The parameters of the bi-variate normal distributions for data set 1 are listed in Table 1 below. The class of a point (feature vector, pattern, object etc.)is defined by the bi-variate distribution that generated the point.

Table 1 Parameters of bi-variate normal distributions for the first data set

class	X mean	Y mean	X variance	Co variance	Y variance
1	1000	4000	400	10	400

2	4000	1000	400	10	400
3	4000	4000	400	10	400
4	1000	1000	400	10	400

A plot of data set 1 is shown in Figure 1.

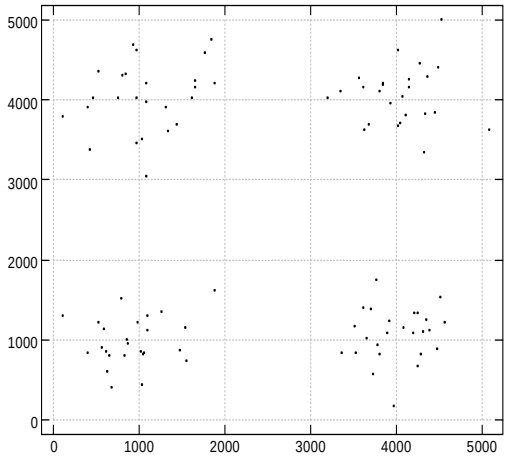

Figure 1 Plot of data set 1

4.1.1.2 Data set 2

Data set 2, as plotted in Figure 2, is produced by the same bivariate normal distributions as data set 1 except that the x and y variances are now 700 to make the classes less distinct geometrically. A plot of the data set appears in Figure 2.

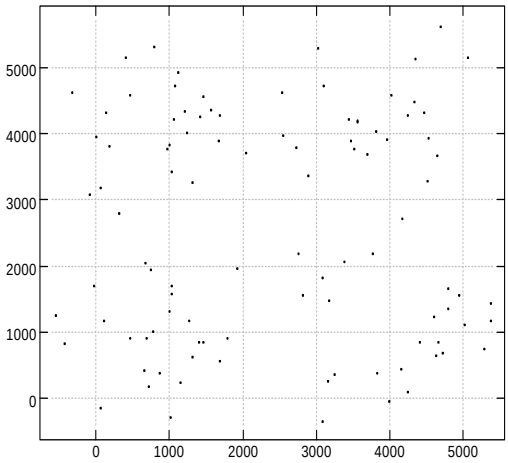

Figure 2 Plot of data set 2

4.1.1.3 Data set 3

Data set 3 is produced by the same bivariate normal distributions as data set 1 except that the x and y variances are now set to 10000. A plot of the data set appears in Figure 3.

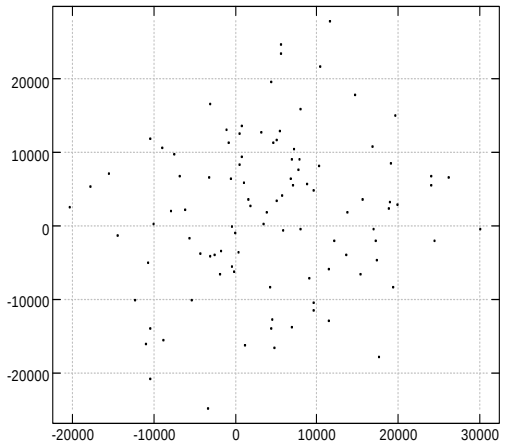

Figure 3 Plot of data set 3

4.1.2 Results obtained during fuzzy clustering

In this case the correct fuzzy membership matrix to include in the comparison is not known unless the actual clusters are an infinite distance apart so that in effect each element belongs to only one cluster. In the case of the first data set this is almost true and we can use the class partition for comparison as before. The class of a member is defined by the bi-variate normal distribution that generated the member as before. The curves in the following plots from top to bottom at the y axis correspond to *FRI*, *FJI*, and *FARI*.

4.1.2.1 Fuzzy clustering of data set 1

A plot of the 3 parameters *FRI*, *FJI*, and *FARI* as FCM clustering proceeds, is shown in Figure 4 for data set 1. Each parameter reached the value 1 showing that the class partition was equivalent to the fuzzy cluster partition. The classes were sufficiently separated so that the discriminant matrix was almost equal to the correct membership matrix.

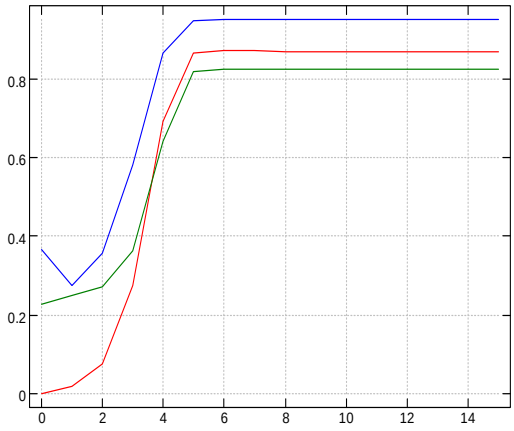

Figure 4 A plot of the 3 parameters *FRI*, *FJI*, and *FARI* as FCM clustering proceeds for data set 1 using fuzzy clustering

Table 2 Values of indices at the end of fuzzy clustering for data set 1

	FRI	*FARI'*	*FJI*
cp vs fcp	0.99983	0.999769	0.9993
cp vs ccp	1	1	1

Table 3 Confusion matrix for fuzzy clustering of data set 1

cluster class	0	1	2	3
0	25	0	0	0

1	0	0	25	0
2	0	25	0	0
3	0	0	0	25

The parameters do not reach 1 because the correct membership matrix is not the same as the discriminant matrix used in the comparison.

4.1.2.2 Fuzzy clustering of data set 2

For data set 3 a plot of the 3 parameters *FRI*, *FJI*, and *FARI* as FCM clustering proceeds, is shown in Figure 5.

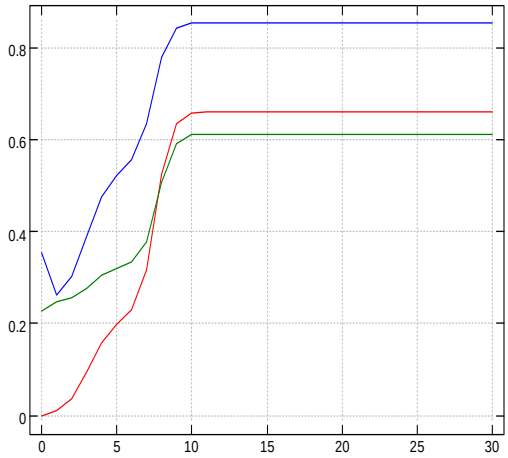

Figure 5 plot of the 3 parameters *FRI*, *FJI*, and *FARI* as fuzzy c-means clustering proceeds for data set 2 using fuzzy clustering

Table 4 Values of indices at the end of fuzzy clustering for data set 2

	FRI	*FARI'*	*FJI*
cp vs fcp	0.975	0.967	0.905
cp vs ccp	1	1	1

Table 5 Confusion matrix for fuzzy clustering of data set 2

cluster class	0	1	2	3
0	1	0	25	0
1	24	0	0	0
2	0	1	0	25
3	0	24	0	0

4.1.2.3 Fuzzy clustering of data set 3

For data set 3 a plot of the 3 parameters *FRI*, *FJI*, and *FARI* as FCM clustering proceeds, is shown in Figure 12.

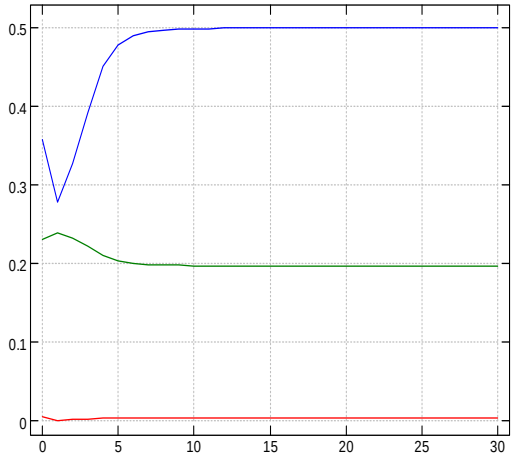

Figure 12 A plot of the 3 parameters *FRI*, *FJI*, and *FARI* as fuzzy c-means clustering proceeds for data set 3 using fuzzy clustering

The curves demonstrate that the correct class membership matrix is not the class discriminant matrix because classes are far from being geometrically distinct in this case.

Table 6 Values of indices at the end of fuzzy clustering for data set 3

	FRI	*FARI'*	*FJI*
cp vs fcp	0.593	0.491	0.148
cp vs ccp	0.614	0.492	0.138

Table 7 Confusion matrix for fuzzy clustering of data set 3

cluster class	0	1	2	3
0	4	4	5	3
1	7	3	5	6
2	7	10	9	11
3	7	8	6	5

4.2 Values of indices obtained during fuzzy clustering of iris data

In this simulation a data set with 148 random samples of flowers from the iris species *setosa*, *versicolor*, and *virginica* collected by Anderson [15] was used. There are 50, 50 and 48 observations respectively for the species for sepal length, sepal width, petal length, and petal width in cm. This dataset was used by Fischer [16] in his initiation of the linear-discriminant-function technique.
A visual representation of the values of the indices as clustering proceeds is provided in

Figure 14. The curves from top to bottom at the y axis in the following plots correspond to *FRI*, *FJI*, and *FARI* respectively.

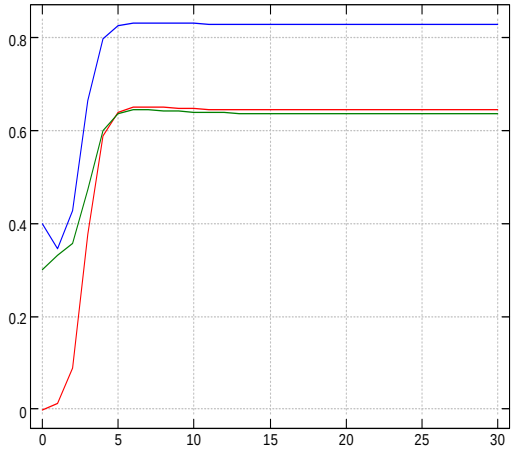

Figure 14 Plot of indices *FRI*, *FARI*, and *FJI* as clustering proceeds for iris data set with 3 clusters

Table 8 Values of indices at the end of fuzzy clustering for iris data set

	FRI	*FARI'*	*FJI*
cp vs fcp	0.876	0.8625	0.694
cp vs ccp	0.879	0.865	0.696

Table 9 Confusion matrix for fuzzy clustering of iris data set

cluster class	0	1	2
0	0	48	14
1	0	2	34
2	50	0	0

The fact that the crisp clustering induced by the fuzzy clustering does not completely match the class partition does not mean that the clusters that were obtained by the clustering process are of low quality and due to a poor clustering method. It may just be due to the fact that the classification is not compatible with measures of distances between the feature vectors. It may be due to feature vectors being poor representations of the objects that they represent.

4.2.1 Fuzzy clustering of a gene expression data set with 2 clusters

The data matrix pertaining to the article by Alon et.al. [17] contains the expression of 2000 genes with highest minimal intensity across 62 tissues. Gene expression in 40 tumour and 22 normal colon tissue samples was analyzed [18] . In this simulation the gene expression values are treated as values of attributes with each gene treated as an attribute. The tissue samples are the objects that are clustered. Each object vector is normalized so that the sum over its components is zero and the magnitude of the vector is one.
Following is the result of applying FCM clustering. As before the curves from top to bottom along the y-axis correspond to FRI, FJI and FARI respectively.

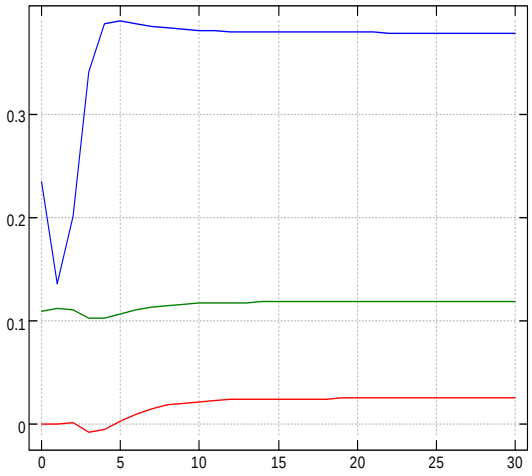

Figure 6 Plot of indices *FRI*, *FARI*, and *FJI* as clustering proceeds for gene expression data set

Table 10 Values of indices at the end of fuzzy clustering for gene expression data set

	FRI	*FARI'*	*FJI*
cp vs fcp	0.305	0.514	0.122
cp vs ccp	0.518	0.5305	0.132

Table 11 Confusion matrix for fuzzy clustering of gene expression data set

cluster class	T	N
N	**6**	18
T	34	**4**

We have labelled the clusters on the basis of the maximum in each row which is possible since each class has a unique cluster maximum . We get the result of : false positives is 6 out of 40 or 15% and , false negatives is 4 out of 22 or 18%. To obtain better results more pre-processing is required in terms of gene selection possibly.

5 Conclusion

The simulations have shown that the fuzzy extensions of the standard cluster quality measures, *RI*, *ARI* and *JI* to *FRI*, *FARI*, and *FJI* respectively, proposed in this paper give desired results. Of these 3, *FARI* is most meaningful in the fuzzy case just as *ARI* is in the strictly crisp partition case.

In this paper, partitions obtained by fuzzy clustering, are compared to class partitions. There are two reasons why the class partition may not match the cluster partition. One is that the clustering process itself is an approximation. Another more serious reason is that the class partition only defines the source of the data. Classes of elements or objects, that are represented by feature vectors, are not assigned on the basis of distance between the feature vectors while assignment of elements to clusters is based on distance. Thus a low value for an index may mean no more than that the original data is not very separable in terms of the distance measure used in clustering.

In general the correct partition is not known since clustering is unsupervised. The indices in this paper however only measure similarity between two partitions whatever their source. Unless the clusters are very distinct in the original data, the correct partitioning in the fuzzy case is not known and it is difficult to determine the parameters for the indices.

Of course there are also measures that do not require comparison to a correct partition. These will be discussed in another paper [19].

Acknowledgement The support of an NSERC grant 227338-04 from the Canadian Government is greatly appreciated..

References

1. Wen, X., et al. *Large-scale temporal gene expression mapping of central nervous system development*. in *National Academy of Science*. 1998. USA.
2. DeRisi, J.I., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**: p. 680-6.
3. Yeung, K.Y., D.R. Haynor, and W.L. Ruzzo, *Validating clustering for gene expression data.* Bioinformatics, 2001. **17**(4): p. 309-318.
4. Eisen, M.B., P.T. Spellman, and P.O. Brown. *Cluster analysis and display of genome-wide expression patterns*. in *National Academy of Science of the U.S.A.* . 1998.
5. Jain, A.K. and R.C. Dubes, *Algorithms for Clustering Data*. 1988, Upper Saddle River, NJ: Prentice Hall. 321.
6. Bezdek, J., *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981: Plenum.
7. Thalamuthu, A., et al., *Evaluation and comparison of gene clustering methods in microarray analysis.* Bioinformatics, 2006. **22**(19): p. 2405-12.
8. Rand, W.M., *Objective Criteria for the Evaluation of Clustering Methods.* Journal of the American Statistical Association, 1971. **66**: p. 846-50.
9. Hubert, L. and P. Arabie, *Comparing partitions.* Journal of Classification, 1985. **2**: p. 193-8.
10. Saporta, G. and G. Youness. *Comparing two partitions: Some Proposals and Experiments."* in *COMPSTAT, 15th Conference on Computational Statistics*. 2002.
11. Hubert, L.J. and R.G. Golledge, *A Heuristic Method for the Comparison of Related Structures.* Journal of Mathematical Psychology 23, pp214-226 81.
12. Milligan, G.W. and M.C. Cooper, *A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis.* Multivariate Behavioral Research, 1986. **21**: p. 441-58.
13. Gentleman, R., et al., eds. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. 2005, Springer: New York.
14. Hoppner, F., et al., *Fuzzy Cluster Analysis*. 1999: John Wiley and Sons. 289.
15. Anderson, E., *The irises of the Gaspé peninsula.* Bulletin of the American Iris Society, 1935. **59**: p. 2-5.
16. Fisher, R.A., *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, 1936. **7**: p. 179-188.
17. Alon, U., et al., *Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc. Natl. Acad. Sci. USA, 1999. **96** (12): p. 6745-50.
18. Alon, U., et al., *Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(12): p. 6745-6750.
19. Brouwer, R.K., *A Clustering Quality Measure based on the Proximity Matrices for the Pattern Vectors and the Membership Vectors.* IJPRAI, 2008. under review.