An experimental validation of some indexes of fuzzy clustering similarity

Stefano Rovetta and Francesco Masulli*

Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, and CNISM
Genova Research Unit, Via Dodecaneso 35, I-16146 Genova, Italy
* Center for Biotechnology, Temple University, Philadelphia, USA
{rovetta,masulli}@disi.unige.it

Abstract. Measuring the similarity between clusterings is a classic problem with several proposed solutions. In this work we focus on measures based on co-association of data pairs and perform some experiments to investigate whether specificities can be highlighted in their behaviour. A unified formalism is used, which allows easy generalization of several indexes to a fuzzy setting. A selection of indexes is presented, and experiments investigate simplified cases and a paradigmatic real-world case, as an illustration of application.

1 Introduction

Fuzzy clustering [1] is a well established procedure and a useful data analysis tool, often providing more flexibility and expressive power than crisp techniques. In general, clustering quality assessment is a problem without a satisfactory solution [2] due to the unsupervised nature of the task. Several cluster validity indexes [3,4] take into account cluster size and composition. However, if we have additionally class labels available, comparisons may be performed according to this external information, although the fact that these reflect the natural grouping of the data is only to be regarded as a working hypothesis. Another form of quality assessment for clusters is stability analysis [5], where comparison is necessary to evaluate cluster variability.

Cluster similarity (or diversity) can also be used to achieve better performance by ensemble clustering [6]. Another use of cluster comparison is biclustering [7] i.e., clustering rows and columns at the same time. Fuzzy biclustering [8] is available. In both crisp and fuzzy cases, many methods only produce one bicluster for each run, so the question arises of whether two biclusters are similar enough to be considered the same.

In all these examples, the problem may be reduced to measuring the similarity between two fuzzy partitions. In this paper, we analyze some methods to compare fuzzy partitions. The contributions of this work consist in:

- A unified formalism useful for the implementation of several clustering comparison indexes, which are all represented by means of co-association matrices;
- The use of this framework to generalize several indexes to a fuzzy setting; some indexes are presented, but the generalization can be extended to many other cases;
- Some experimental insight on the behaviour of these indexes in both simplified cases, where the relationship between the partitions is clear, and a paradigmatic real-world case, as an illustration of application.

2 Measures based on data pairs

Clustering induces a partitions of the data, so measuring the agreement or overlap between two clusterings amounts to measuring the similarity between two partitions.

There are several partition similarities available in the literature. A first distinction can be made between *pairwise* similarity indexes, which apply to pairs of partitions, and *non-pairwise* ones, which can work on an arbitrary number of partitions. The difference is of a practical nature only, since in principle a non-pairwise index can always be obtained by computing a pairwise index for all possible pairs, and then averaging.

A more fundamental distinction refers to the way partitions are compared. The two main approaches include comparing matching subsets, and comparing co-association information. The first approach is not reliable when the partitions are not very similar, and in any case require some criterion for matching subsets. In principle, we can expect these methods to be of linear complexity w.r.t. data cardinality.

In this study, we concentrate on the second approach. Co-association information is obtained by analyzing whether pairs of points in the data set are co-attributed to the same cluster by both partitions. These methods require building a co-association matrix [9], by scanning all possible data pairs, so they run quadratically with cardinality.

Given a data set X, suppose we have two fuzzy or soft partitions A and B of X. Soft partitions means that $\forall x \in X$ there is a membership $\mu(x, a_i)$ for each subset $a_i \in A$ (similarly for B – and this comment applies throughout). We assume normal memberships. For a proper partition we ask that $\sum_i \mu(x, a_i) = 1$. (Note that possibilistic subsets –not proper partitions– are also possible by removing this constraint.)

Each data point is thus represented by a coordinate vector, whose dimension is the number of subsets (clusters) in the partition, and whose components are the membership values, which we assume normalized in $[0, 1]$. Each data pair is described by the degree of similarity between the two objects x and y under the partition A.

Similarity of strings of memberships can be measured by Hamming distance for crisp bits, which is equivalent to summing the bits of the bitwise-AND between the two words. The fuzzy generalization of this operation is defined once we appropriately define the conjunction connective AND [10]. We adopt the product t-norm [11], which is appealing because it is related to the scalar product operation between vectors, which in turn can be used to define popular distance measures, and also to the concept of a joint probability. This allows some generality in the indexes studied, although the product logic arising from this particular choice does not have some of the properties found in other cases [12] (e.g., Gödel or Łukasiewicz fuzzy logics). In particular, some of the derivations have been obtained by assuming a specific relationship between the AND and OR connectives which is not necessarily satisfied by all possible definitions.

Given two fuzzy memberships/truth values μ and ν, the conjunction logical connective is defined as $\mu \,\mathrm{AND}\, \nu = \mu\nu$. The co-association of a given pair of data points to a given cluster a_i is the conjunction of the respective point memberships to a_i, and the degree of similarity of two points is the average of these values. The *co-association matrix* s^A under partition A is:

$$s^A_{ij} = s^A(x_i, x_j) = \sum_{l=1}^{|A|} \mu(x_i, a_l)\mu(x_j, a_l) \tag{1}$$

Note that in the crisp case this definition collapses to the proposition "partition A puts x and y in the same cluster", but in the fuzzy case it is necessary to take all clusters into considerations because, in general, none of them will be exactly zero or one.

3 Indexes

Once the co-association matrices s^A and s^B are built, we can treat their entries as two paired samples and compare them with appropriate measures.

To simplify notations, we will serialize the matrices s^A and s^B so that they may be indexed as vectors, so that: $s^A_{ij} = \sigma^A_h$. To avoid redundancy, the index h scans only the upper triangular matrix, excluding the diagonal (which is trivial), so $i \in [1, |X| - 1]$, $j \in [i + 1, |X|]$, and $h = |X|(i - 1) - i(i + 1)/2 + j$. Moreover, we define $H = |X|(|X| - 1)/2$ so that $h \in [1, H]$.

Indexes of partition similarity based on the co-association matrix can be computed by several approaches. Some of them are reviewed in [13] and some are experimentally compared in [6]. These may include Ward's linkage criterion [14], Student's t formula [15], information-theoretic criteria as in [6] and [13], and well-known partition overlap measures like Jaccard's [16] or Rand's [17] or Fowlkes and Mallows' [18] indexes, and subsequent work [19].

In general, measures for paired samples can be based on different criteria. Methods like Ward or Student are focused on comparing average and dispersion of the two samples. Another approach is to compute some statistic on the differences between the two data in each pair (that is, $\sigma^A_h - \sigma^B_h \ \forall h$). Yet another criterion is to exploit the $[0, 1]$ range of the values and analyze other types of combination between data, as in the Pearson correlation and the Jaccard index. We will focus on some indexes which are representative of each approach.

Average linkage – This measure of distance is based on comparing the centroids of two sets, i.e., on computing the average of the two sets and measuring their distance.

This criterion is simple-minded, since it is prone to false positives: two sets with the same centroid are considered coincident even if one has a larger variance than the other. However, for normalized data, it may be reasonable.

Correlation – The standard Pearson correlation coefficient, a measure of similarity:

$$\mathrm{corr} = \frac{\frac{1}{H}\sum_h(\sigma^A\sigma^B) - \frac{1}{H}\sum_h(\sigma^A)\frac{1}{H}\sum_h(\sigma^B)}{\sqrt{\frac{1}{H}\sum_h((\sigma^A)^2) - \left(\frac{1}{H}\sum_h(\sigma^A)\right)^2}\ \sqrt{\frac{1}{H}\sum_h((\sigma^B)^2) - \left(\frac{1}{H}\sum_h(\sigma^B)\right)^2}} \quad (2)$$

Since this is a similarity measure between -1 and 1, the correlation distance is

$$C(A, B) = (1 - corr)/2. \quad (3)$$

Jaccard – The Jaccard coefficient [16] is a classic measure of set similarity, and one of the most general. It is the ratio of the intersection of two sets to their union: $J(A, B) = |A \cap B|/|A \cup B|$. In the crisp case, this pairwise index can be practically computed by counting the number $N_{11} = |A \cap B|$ of points put in the same cluster by both partitions,

the number N_{10} of points assigned to the same cluster only by partition A, and the number N_{01} similarly defined, so that $N_{10} + N_{01} + N_{11} = |A \cup B|$ and

$$J(A, B) = \frac{N_{11}}{N_{10} + N_{01} + N_{11}}. \tag{4}$$

In the fuzzy case, the concept of coincidence must be redefined as a degree of coincidence. By taking advantage of the reasonable assumption that a generalization of De Morgan's law holds, for the product t-norm we can define the associated disjunction operator as the *probabilistic sum* t-conorm, so that $\mu \,\mathrm{OR}\, \nu = \mu + \nu - \mu\nu$. Therefore, in terms of σ^A and σ^B, the actual computation in this case is:

$$A \cap B = \sum_h \sigma_h^A \sigma_h^B \qquad \text{and} \qquad A \cup B = \sum_h \left(\sigma_h^A + \sigma_h^B - \sigma_h^A \sigma_h^B\right) \tag{5}$$

The Jaccard distance between A and B is $1 - J(A, B)$.

Student distance – This index exploits the well-known Student's t statistic to obtain a distance measurement which takes into account not only the estimated overlap, but also its significance in terms of variance.

$$S(A, B) = \frac{\left(\sum_h |\sigma_h^A - \sigma_h^B|\right)}{0.5 + \sum_h |\sigma_h^A - \sigma_h^B|^2 / H - (\sum_h |\sigma^A - \sigma^B| / H)^2} \tag{6}$$

With respect to the Student's t formula, this is compensated to avoid a vanishing denominator.

The Rand index – This index was explicitly proposed for comparing clusterings [17], but it is computed similarly to the Jaccard index which was introduced in a more generic setting. It is defined as

$$R(A, B) = \frac{\sum_h (\sigma_h^A + \sigma_h^B) + 1}{H} \tag{7}$$

4 Experiments

We present some experiments aimed at highlighting the behaviors of these indexes. We use three kinds of data: two families of simple datasets, represented through the membership (directly as partitions, regardless of how they were obtained); then we use the Iris data set as an illustration of real-world application.

4.1 Experiments with toy problems

Experiment set 1 – The aim of this experiment set is to compare several synthetic partitions. The family of toy datasets used is composed of 20 data objects in 2 partitions, one of 2 clusters, the other of 3 clusters. The presented methods are insensitive to the number of clusters in partitions, although this sensitivity may be easily introduced. The datasets used are tabulated in Table 1. Table 2 presents the results.

Experiment set 2 – The second toy dataset family is composed of 20 data objects; 2 partitions of 2 clusters each are used. The second partition is fixed as follows:

Table 1. The family of datasets used in the first experiment set.

toy0																		
A	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
B	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
toy1																		
A	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
B	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
toy2																		
A	0.8	0.8	0.8	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	0.2	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
B	0.8	0.8	0.8	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	0.2	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
toy3																		
A	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
B	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.8	0.8	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2
	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8	0.8	0.8	0.8
toy4																		
A	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
B	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0
toy5																		
A	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.0	0.1	1.0	1.0	1.0	1.0	1.0
B	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0

object	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
cluster1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
cluster2	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

In the first experiment of this set we investigate the variation of indexes in a crisp case. The data start equally clustered in both cases (two equal partitions). Then, in 10 further steps, each data object is moved from one cluster to the other in the first partition; the second is left as is. In the second experiment, the variation of indexes in a simple fuzzy case is analyzed. The partitions start identical, with data equidistributed in the two clusters. One data object in the first partition is gradually moved from the first cluster to the second by changing its memberships in steps of 0.1, as follows:

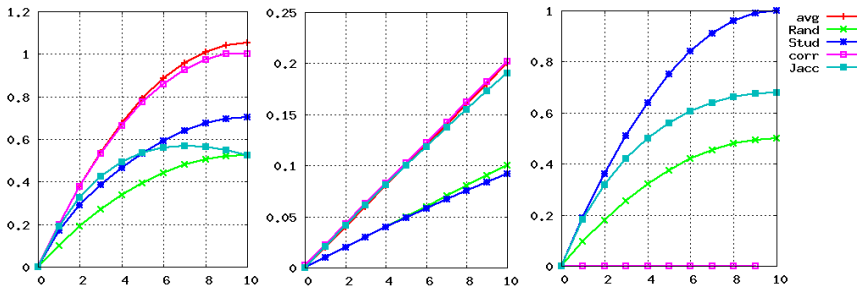
step	0	1	2	3	4	5	6	7	8	9	10
cluster1	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
cluster2	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

The third experiment involves changing the memberships of all data in a partitions from crisp to totally fuzzy, i.e., in the first partition all memberships of the first 10 points in the first cluster are gradually moved from 1 to 0.5 and all memberships of

Table 2. Results on the first toy problem set.

	Average	Rand	Student	Correl.	Jaccard
toy0	0.0000	0.0000	0.0000	0.0000	0.0000
toy1	0.8497	0.4248	0.5707	0.8211	0.6566
toy2	0.0000	0.4352	0.0000	0.0000	0.5976
toy3	0.4455	0.2227	0.3845	0.1610	0.4207
toy4	1.0980	0.5490	0.7344	1.0811	0.8000
toy5	0.9707	0.4854	0.8205	0.9489	0.7595

Fig. 1. Results on the second experiment set.



points 11-20 are moved from 0 to 0.5; memberships in the second cluster are obviously complementary. The second partition is held fixed.

The results of these experiments are illustrated by the graphs in Figure 1. We can observe that in general all the indexes considered feature similar behavior, having value 0 for equal partitions and monotonically increasing behavior (they are not normalized on a single scale, however). The graph also show that, while this similarity holds for most types of variation, there are cases where some index does not agree with others (see Jaccard on the first experiment).

4.2 Real data: Iris

Anderson's Iris data [20] is an almost mandatory testbed, since it is so well-known. Here we use it to show how the properties of the indexes may be exploited. For this dataset the following two features are used: sepal width $\times$ sepal length; petal width $\times$ petal length. These allow a low error with linear separation. In this experiment, two partitions are compared: the ground truth provided by the true classes and the central clustering obtained by taking the averages of each class as centroids. Therefore there are 3 clusters in both the target and class-induced clustering.

Experiments have been made with both crisp and fuzzy partitions. In the crisp case, the memberships are obtained with the simple nearest centroid rule. In the fuzzy case, a membership model similar to the "Maximum Entropy" approach [21] has been used, where membership of data point x_i is computed as $\mu_j(x_i) = \frac{e^{-d_{ij}/\beta}}{\sum_l e^{-d_{Ila}/\beta}}$, where d_{ij} is the (Euclidean) distance between data point x_i and the j-th centroid y_j and β is a scale parameter imposing the degree of fuzziness, here set to 1.

Table 3. Results on the Iris dataset

Crisp case	**Average**	**Rand**	**Student**	**Correl.**	**Jaccard**
Complete	0.5589	0.2795	0.3985	0.5962	0.5552
Setosa vs Versicolor	0.3790	0.1895	0.2899	0.4492	0.5043
Setosa vs Virginica	1.1055	0.5527	0.7397	1.1562	0.8761
Virginica vs Versicolor	0.1523	0.0762	0.1335	0.2175	0.2902
Fuzzy case	**Average**	**Rand**	**Student**	**Correl.**	**Jaccard**
Complete	0.8010	0.4005	0.5639	0.5506	0.6311
Setosa vs Versicolor	0.2916	0.1458	0.2828	0.0763	0.4260
Setosa vs Virginica	0.9465	0.4733	0.8101	0.8406	0.7576
Virginica vs Versicolor	0.1624	0.0812	0.1608	0.0127	0.2939

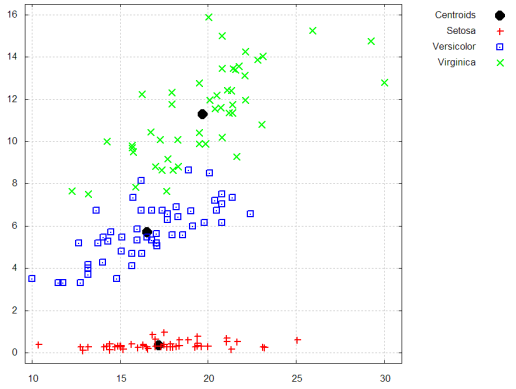


Fig. 2. The Iris data, with class centroids.

Table 3 presents the results. In both the crisp and the fuzzy case, the first row refers to the whole dataset (three classes), while the remaining rows represent all possible pairing of the three classes. We notice that representation with three centroids is not very good, even if these are chosen as the mean of the respective classes. This can be easily inferred from Fig. 2, showing centroids and data. The results indicate that, although with different numerical values, all the indexes correctly reflect the well-known fact that the separation between the *Versicolor* and *Virginica* varieties is hard.

5 Conclusion

We introduced a possible fuzzy framework for applying traditional and novel partition similarity measures to fuzzy clustering. Some indexes have already been generalized to the fuzzy setting in the literature [22], but we provide a more systematic procedure while also introducing some new indexes. Several other formulations are of course possible in addition to the three proposed here; we only provide these as examples.

The experimental result show that, while all indexes retain the same general behaviour, there are indeed some occasional differences that may deserve to be studied

and exploited. Further research will focus on characterizing these fuzzy indexes and on their application in tasks involving clustering comparison, e.g., ensemble clustering.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA (1981)
2. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey, USA (1988)
3. Bezdek, J.C.: Cluster validity with fuzzy sets. Cybernetics and Systems **3**(3) (1973) 58–73
4. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy Sets Syst. **158**(19) (2007) 2095–2117
5. Filippone, M., Masulli, F., Rovetta, S., Zini, L.: Comparing fuzzy approaches to biclustering. In Tagliaferri, R., Masulli, F., eds.: CIBB 2008 proceedings. (2008) in press.
6. Kuncheva, L.I., Vetrov, D.P.: Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11) (2006) 1798–1808
7. Cheng, Y., Church, G.M.: Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol **8** (2000) 93–103
8. Filippone, M., Masulli, F., Rovetta, S., Mitra, S., Banka, H.: Possibilistic approach to biclustering: An application to oligonucleotide microarray data analysis. In Priami, C., ed.: Computational Methods in Systems Biology, International Conference, CMSB 2006, Trento, Italy, October 18-19, 2006, Proceedings. Volume 4210 of Lecture Notes in Computer Science., Springer (October 2006) 312–322
9. Fred, A.L.N., Jain, A.K.: Data clustering using evidence accumulation. Pattern Recognition, International Conference on **4** (2002)
10. Zadeh, L.A.: Fuzzy sets. Information and Control **8**(3) (June 1965) 338–353
11. Menger, K.: Statistical metrics. Proceedings of the National Academy of Sciences of the United States of America **28**(12) (December 1942) 535–537
12. Klement, E.: A survey on different triangular norm-based fuzzy logics. Fuzzy Sets and Systems **101**(2) (January 1999) 241–251
13. Meilă, M.: Comparing clusterings–an information based distance. Journal of Multivariate Analysis **98**(5) (May 2007) 873–895
14. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58** (1963) 236–244
15. Student: The probable error of a mean. Biometrika **6**(1) (March 1908) 1–25
16. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Société Vaudoise des Sciences Naturelles **37** (1901) 547–579
17. Rand, W.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66** (1971) 846–850
18. Fowlkes, E. B. and Mallows, C.L.: A method to compare two hierarchical clusterings. Journal of the American Statistical Association **78** (1983) 553–569
19. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification (1985) 193–218
20. Anderson, E.: The irises of the gaspe peninsula. Bulletin of the American Iris Society **59** (1935) 25
21. Rose, K., Gurewitz, E., Fox, G.: A deterministic annealing approach to clustering. Pattern Recognition Letters **11** (1990) 589–594
22. Campello, R.J.G.B.: A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters **28**(7) (May 2007) 833–841