A PROPOSAL TO EVALUATE THE SOLUTION OF FUZZY CLUSTERING ALGORITHMS

Carmela Iorio [1], Giuseppe Pandolfo[1] and Antonio D'Ambrosio[1]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: carmela.iorio@unina.it, giuseppe.pandolfo@unina.it, antdambr@unina.it)

ABSTRACT: When the aim is to evaluate the solution of a fuzzy clustering algorithm, the computation of the adjusted version of the Rand index requires converting the soft partitions to hard partitions. Furthermore, in comparing two fuzzy partitions from two different clustering methods, an external validation index should satisfy two desirable properties: *(i)* reflexivity, and *(ii)* a proper interpretation of correction for agreement due to chance. In this paper, we show an extension of the commonly used adjusted Rand index to fuzzy partitions based on normalized degree of concordance.

KEYWORDS: Cluster analysis, Cluster validity, External criteria, Adjusted Concordance Index.

1 Introduction

Cluster analysis is a data mining technique that groups units (or objects) into a finite set of clusters (or groups) based on a distance or a similarity. The purpose of clustering is to partition the objects into distinct groups so that observations within each cluster are similar to each other, while observations in different clusters are different from each other. Many clustering algorithms have been introduced In literature, and many of the methods do not produce a partition, but e.g.hierarchies, or posterior probabilities (e.g. model-based clustering). Furthermore, since groups can be formally seen as subsets of the entire data set, one possible classification of clustering methods can be done according to whether the subsets are crisp (hard) or fuzzy (soft). Hard clustering methods are based on classical set theory and restrict each object in the data set to belong to exactly one cluster. Soft clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. In contrast to hard clustering, each object has a membership value in each cluster: the larger the value of the membership value for a given object with respect to a cluster, the larger the probability of that object being assigned

to that cluster. An extensive overview of cluster analysis can be found in Kaufman & Rousseeuw, 2005, Everitt *et al.*, 2011, Duran & Odell, 2013, Hennig & Meila, 2015. However, clustering is an unsupervised learning problem since the aim is to identify a structure in an unlabeled data set. As a consequence, an important issue in cluster analysis is the evaluation of clustering results. The procedure for evaluating the goodness of the results of a clustering algorithm is known as cluster validation. Generally, there are three approaches to assessing cluster validity involving internal, external, and relative criteria. Internal validation criteria use the information involving the data set used in the clustering process (e.g. Silhouette index). External validation criteria evaluate clustering results by comparing them to an externally known result. Relative validation criteria evaluate the clustering structure by comparing it to other clustering schemes, i.e. by varying different parameter values for the same algorithm. Several external validation criteria have been proposed in the literature to evaluate hard or soft clustering algorithms. Among them the most popular indexes are Rand Index proposed by Rand, 1971 and its corrected versions for fuzzy partitions (see e.g. Campello, 2007, Frigui *et al.*, 2007, Brouwer, 2009, Anderson *et al.*, 2010, Hüllermeier *et al.*, 2012). In this work, attention is put on external validation criteria to evaluate the goodness of fuzzy partitions.

2 The key idea

We think that, in comparing the partitions coming from two, different clustering methods, a good index to be used should satisfy at least two desirable properties: *(i)* reflexivity and *(ii)* a proper interpretation of correction for agreement due to chance. The problem with evaluating the solution of a fuzzy clustering algorithm with the original formulation of the Rand index (RI) is that it requires converting the soft partitions into hard partitions, thus losing information. As Meilă, 2007 and Morey & Agresti, 1984 pointed out, there are other known problems with RI. It approaches its upper limit as the number of groups increases; it is extremely sensitive to the number and size of groups considered in each partition as well as to the overall number of observations considered; the expected value of RI for two random partitions does not take a constant value. To overcome these drawbacks, Hubert & Arabie, 1985 has proposed an adjusted version of RI (ARI) assuming the generalized hypergeometric distribution as the randomness model. Besides the ARI, even the fuzzy generalizations of the RI proposed by Campello, 2007, Frigui *et al.*, 2007, Brouwer, 2009, and Anderson *et al.*, 2010 fail to satisfy reflexivity property and therefore cannot be considered a metric.

Since we are interested in comparing fuzzy partitions and ARI is still the most popular measure used for clustering comparison, we show an extension of ARI to fuzzy partitions. The proposed index, named Adjusted Concordance Index (ACI), is based on the fuzzy variant of the ARI proposed by Hüllermeier *et al.*, 2012. These authors based their proposal on the fuzzy equivalence relation and this allows us to rewrite every partition as a similarity matrix based on the normalized city block. Thus, the ACI is given by:

$$ACI = \frac{NDC - \overline{NDC}}{1 - \overline{NDC}},$$

where the normalized degree of concordance (NDC) is a direct generalization of the RI and $\overline{NDC}$, is the mean value of the NDC over all the permutations. Since, $NDC(\mathbf{P},\mathbf{Q}) = 1 - d(\mathbf{P},\mathbf{Q})$, where $\mathbf{P}$ and $\mathbf{Q}$ are two fuzzy partitions, the NDC is the only extension of the RI to the fuzzy partition which fulfills the reflexivity property that always guarantees that its maximum value is equal to one.
For further details and comments on ACI, the interested reader may refer to D'Ambrosio *et al.*, 2021.

3 Conclusion

To evaluate the fuzzy clustering results, the external validation criteria proposed in the literature fail two desiderata: reflexivity, and a proper expectation. To compare fuzzy clustering algorithms, the adjusted Rand index (ARI), is commonly used to measure agreement between partitions. Following similar reasoning to Hubert & Arabie, 1985, we have provided the adjusted version of the normalized degree of concordance (NDC) index defined by Hüllermeier *et al.*, 2012. We named it the adjusted concordance index (ACI). It normalizes the difference between NDC itself and the point estimate of its expected value. Since NDC is the only fuzzy extension of the Rand index that possess the reflexivity property, thus the resulting ACI is itself a reflexive index. In this regard, our proposal works with any raw fuzzy index, provided that the two above-mentioned desiderata are satisfied.

References

ANDERSON, D.T., BEZDEK, J.C., POPESCU, M., & KELLER, J.M. 2010. Comparing fuzzy, probabilistic, and possibilistic partitions. *Fuzzy Systems, IEEE Transactions on*, **18**(5), 906–918.

BROUWER, R.K. 2009. Extending the Rand, adjusted Rand and Jaccard indices to fuzzy partitions. *Journal of Intelligent Information Systems*, **32**(3), 213–235.

CAMPELLO, R. JGB. 2007. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, **28**(7), 833–841.

D'AMBROSIO, A., AMODIO, S., IORIO, C., PANDOLFO, G., & SICILIANO, R. 2021. Adjusted concordance index: an extensionl of the adjusted rand index to fuzzy partitions. *Journal of Classification*, **38**, 112–128.

DURAN, B.S., & ODELL, P.L. 2013. *Cluster analysis: a survey*. 2 edn. Heidelberg, Germany: Springer Science & Business Media.

EVERITT, B.S., LANDAU, S., LEESE, M., & STAHL, D. 2011. *Cluster analysis*. 5 edn. Chichester, UK: Wiley.

FOWLKES, E.B., & MALLOWS, C.L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**(383), 553–569.

FRIGUI, H., HWANG, C., & RHEE, F.C.H. 2007. Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition*, **40**(11), 3053–3068.

HENNIG, C., & MEILA, M. 2015. Cluster analysis: an overview. *Chap. 1, pages 1–20 of:* HENNIG, CHRISTIAN, MEILA, MARINA, MURTAGH, FIONN, & ROCCI, ROBERTO (eds), *Handbook of cluster analysis*. Boca Raton, FL: CRC Press.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

HÜLLERMEIER, E., RIFQI, M., HENZGEN, S., & SENGE, R. 2012. Comparing fuzzy partitions: A generalization of the Rand index and related measures. *Fuzzy Systems, IEEE Transactions on*, **20**(3), 546–556.

KAUFMAN, L., & ROUSSEEUW, P.J. 2005. *Finding groups in data: an introduction to cluster analysis*. 2 edn. Hoboken, NJ: John Wiley & Sons.

MEILĂ, M. 2007. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, **98**(5), 873–895.

MOREY, L.C., & AGRESTI, A. 1984. The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, **44**(1), 33–37.

RAND, W.M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.