Characterizing measures for the assessment of cluster analysis and community detection

Nejat Arinik, Rosa Figueiredo, and Vincent Labatut

Laboratoire Informatique d'Avignon LIA EA 4128, Avignon, France
{firstname,lastname}@univ-avignon.fr

Keywords: Cluster analysis · Community detection · External evaluation measures.

1 Introduction

The problem of comparing two partitions of the same set occurs in a number of situations, the most widespread being probably the assessment of cluster analysis and community detection results. In these contexts, one has computed the clusters of a dataset, or the community structure of a network. This result takes the form of a partition of the set of data points or set of nodes, respectively. One then wants to compare this estimation with some ground-truth also taking the form of a partition. Alternatively, one has computed several such estimations, for instance using several algorithms, and wants to compare them to each other.

This comparison is traditionally performed through some measures able to quantify the similarity between two such partitions, and often called *external* measures (measures, hereafter), as they allow comparing the output of the partitioning method to some form of ground truth. There are many ways to formalize what one means by "similarity", resulting in the proposition of a very large number of such measures over the years, as well as surveys comparing them [7].

In the literature, authors proposing new external measures follow a relatively standard workflow. First, they list some mathematical properties which they deem desirable in such measures, e.g. not being sensitive to the number of clusters k [6]. They then show that existing measures do not possess these properties. This task is frequently performed through an empirical approach, which consists in applying some predefined transformations to certain partitions, both designed in a way that is related to the property of interest, and to study how the measure reacts to these perturbations by using it to compare those partitions. Finally, the authors solve this issue by proposing a new measure having these properties, or modifying an existing one to this end.

Each application case is likely to bring its own constraints and requirements, so there is no such thing as a "best" measure that would fit *all* situations. One trait considered as positive in one case could very well be perceived as a drawback in another. However, due to the profusion of available measures, selecting the most appropriate one for a given situation is a challenge for the end user. As mentioned before, some survey articles try to compare them, but they focus

only a small number of measures [7] and/or properties [1]. More importantly, the evaluations they perform are specific to these measures and properties [7], preventing the end user to include additional measures or properties in the comparison. In practice, the problem of selecting an appropriate measure to compare partitions is generally overlooked, and researchers tend to follow tradition and use the measures frequently appearing in the literature of their field.

In this work, we propose a new framework to solve this issue. It is based on the empirical approach presented above, and consequently relies on a set of predefined partitions and parametric partition transformations. We study the effect of each parameter on the measure through multiple linear regression, in order to produce results that the end user can interpret. Our framework is not tied to any specific measure, property, or transformation, so it can be applied to any situation. In the full work, we will illustrate its relevance by applying it to a selection of popular measures, and show how it can be put in practice through two concrete use cases. Here, we summarize the framework and give some preliminary results.

2 Method

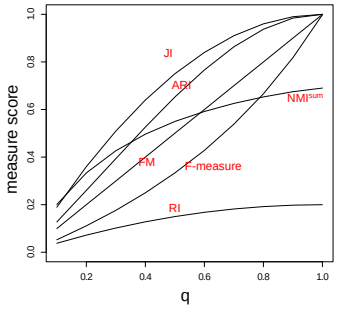
The framework we propose is designed to analyze the behavior of a set of measures, and it is constituted of two parts.

The first part aims at generating the data required to study and compare the considered measures. We proceed in three steps, and unlike the common approach adopted in the literature, we do so in a fully parametric way in order to get a greater control. For the same reason, our approach is deterministic. The first step is to create a so-called *original partition*, controlled by three parameters: number of elements n, number of clusters k and cluster size heterogeneity h. The second step consists in applying a transformation to the original partition, in order to produce a so-called *transformed partition*. This step is controlled by two parameters: type t and intensity q of the transformation. The third step is to compute the selected external measures in order to assess how similar each pair of original and transformed partitions are. We normalize the resulting values so that they all express *dissimilarity* between the partitions, in order to make them comparable. We repeat the process with an adequate number of different parameter values in order to cover the parameter space.

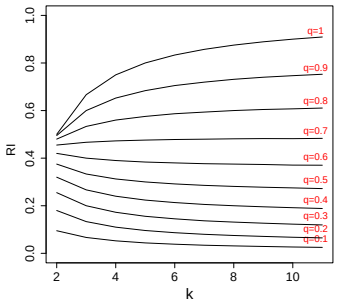
The second part of our framework consists in analyzing all the dissimilarity values obtained at the first part. To do so, we design a multiple linear regression model able to take into account the interactions between the parameters. We perform a relative importance analysis [9] in order to assess how much the framework parameters affect the measures.

3 Preliminary Results

To show the interest of our framework, we discuss here a few preliminary results, directly obtained from the first part of our framework (and preceding the work conducted in the second part).



(a) Scores obtained with all measures, as functions of q, for the *Singleton Clusters* transformation.



(b) The RI scores, as a function of k, for the *1 New Cluster* transformation

Fig. 1: Preview of two types of results obtained with our framework. Figures available at 10.6084/m9.figshare.13109813 under CC-BY license.

First, we consider how measure scores are affected by parameter q (transformation intensity) when applying the so-called *Singleton Cluster* transformation, while other parameters are fixed to arbitrary values. This transformation consists in turning each element concerned by the transformation into a new cluster. Figure 1a shows that all the measure scores increase with q, albeit in different ways. Overall, RI [8] has the smallest slope coefficient, followed by NMI^{sum} [10], and they are therefore the least sensitive to this transformation. We observe that JI [5], ARI [4], FM [3] and the F-measure [2] get similar scores for extreme q values, but are relatively different when q gets closer to 0.5.

In Figure 1b, we focus on the RI and show how it is affected by changes in q (transformation intensity) and k (number of clusters in the transformed partition), for the so-called *1 New Cluster* transformation. This transformation consists in adding to the original partition a single new cluster containing all the elements affected by the transformation. As a function of k, the RI score is always monotonic, however the nature and slope of the trend depends on q: increase trend for $q \geq 0.7$ vs. decrease for $q < 0.7$. This means there is a significant interaction between q and k. This type of joint effect between parameters will be captured by the interaction terms present in our regression model.

The goal of the second part of our framework, whose results are not presented here, is to summarize the information conveyed by figures such as Figure 1, in order to ease the characterization and comparison of the considered measures.

References

1. Albatineh, A.N., Niewiadomska-Bugaj, M., Mihalko, D.: On similarity indices and correction for chance agreement. Journal of Classification **23**(2), 301–313 (sep 2006)
2. Artiles, J., Gonzalo, J., Sekine, S.: The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 64–69. SemEval '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
3. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. Journal of the American Statistical Association **78**(383), 553–569 (sep 1983)
4. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
5. 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**, 547–579 (1901). https://doi.org/10.5169/seals-266450
6. Liu, X., Cheng, H.M., Zhang, Z.Y.: Evaluation of community detection methods. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2019)
7. Nguyen Xuan Vinh, Julien Epps, J.B.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research (2010)
8. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336), 846–850 (dec 1971)
9. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global sensitivity analysis. the primer. John Wiley & Sons (2008)
10. Strehl, A., Ghosh, J.: A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3 (2002) 583-617 (2002)