

A new index for clustering validation with overlapped clusters

D.N. Campo^{a,b}, G. Stegmayer^{a,b}, D.H. Milone^a

^a*Research Institute for Signals, Systems and Computational Intelligence, [sinc\(i\)](http://sinc(i).unl.edu.ar/sinc), UNL, CONICET, Santa Fe, Argentina*

^b*Research and Development Center for Information Systems Engineering, CIDISI-UTN-FRSF, CONICET, Santa Fe, Argentina*

Abstract

External validation indexes allow similarities between two clustering solutions to be quantified. With classical external indexes, it is possible to quantify how similar two disjoint clustering solutions are, where each object can only belong to a single cluster. However, in practical applications, it is common for an object to have more than one label, thereby belonging to overlapped clusters; for example, subjects that belong to multiple communities in social networks. In this study, we propose a new index based on an intuitive probabilistic approach that is applicable to overlapped clusters. Given that recently there has been a remarkable increase in the analysis of data with naturally overlapped clusters, this new index allows to comparing clustering algorithms correctly. After presenting the new index, experiments with artificial and real datasets are shown and analyzed. Results over a real social network are also presented and discussed. The results indicate that the new index can correctly measure the similarity between two partitions of the dataset when there are different levels of overlap in the analyzed clusters.

Keywords: overlapped clusters, validation index, external validation, cluster perturbation.

1. Introduction

Clustering algorithms take a dataset as input and, through a non-supervised process, partition the data into a set of clusters or groups. A cluster can be defined as a group of objects that are similar given a relative measure and that are dissimilar to objects grouped in others clusters (Skillicorn, 2007; Xu & Wunsch, 2008). The application of clustering algorithms always returns a solution, even when there is not a clear structure in the data.

Preprint submitted to Expert Systems with Applications

August 6, 2016

Therefore, a reliable mechanism for measuring similarities between partitions is desirable to detect which ones are, for example, more stable when several solutions are considered.

Furthermore, in current practice, most information that is created through social networks, news tags, collaboration networks and other Internet media, is naturally overlapped. As a result, overlapped solutions are expected to be found in the analysis of such type of data. An index for measuring similarities between these partitions would therefore be a valuable tool to study them. Recently, with the spread of social and collaboration networks, the use of clustering with overlapping properties has increased and new algorithms have been proposed (Wang et al., 2014; Alvani et al., 2013; Gopalan & Blei, 2013; Gossen et al., 2014; Chakraborty, 2015; Xie et al., 2013; Amelio & Pizzuti, 2014).

Two types of validation measures can be used for measuring similarities between clustering solutions: internal and external. The first type of metrics measures attributes taken from the data itself and the clusters formed, such as data compactness and separability. The second one makes a comparison between clustering solutions, taking one as a reference and comparing it with other groupings (Halkidi et al., 2001; Handl et al., 2005). Considering external metrics only, three types of measures are available: pair counting measures, set matching measures, and information theory measures. One of the most used and widely known pair counting measure is the Fowlkes-Mallows index (FM), which works with the frequency of pairs of patterns found in two clustering solutions that are being compared (Ben-Hur & Guyon, 2003; Fowlkes & Mallows, 1983). A representative set matching measure is the Maximum Match (Meilă & Heckerman, 2001), which analyses the most similar clusters from both solutions and counts the elements in common in such paired groups. Finally, regarding measures based on information theory, Normalized Mutual Information is extensively used and works by quantifying the information shared between both solutions, through the concept of entropy (Meilă, 2007; Vinh et al., 2010). However, none of these metrics was designed for evaluating similarities between solutions when overlapped clusters are considered.

Lately, given the overwhelming amount of information created through different social and collaboration networks, interest has emerged in clustering analysis to process such

amount of data when there are overlapping clusters (Amelio & Pizzuti, 2014). For example, in (Zhou et al., 2015), the authors proposed an ant-based algorithm to detect communities in networks, where clusters are formed by nodes that may be considered as overlapped. In (Liu et al., 2013), the authors developed a method for characterizing the structure of real-world affiliation networks composed of groups of fully connected, generally overlapped communities. Also, in (Alvari et al., 2013) overlapped clusters in social networks are studied and a framework based on a game theory approach is proposed for detecting communities. Similarly, in (Gopalan & Blei, 2013), a new method based on a Bayesian model is presented. This method enables the detection of large overlapped communities in massive synthetic datasets and in large-scale, real life social, biological and citation networks. In (Kalinka & Tomancak, 2011), the authors present an R package that extends an existing algorithm for clustering, which handles directed and weighted links between nodes in a biological network. These networks would naturally contain nested or overlapped links. In (McGarry, 2013), data acquired from protein networks is clustered and the results are integrated with chemical databases using ontologies. Hence, based on the principle of guilt-by-association (Wolfe et al., 2005; Lacroix et al., 2008; Usadel et al., 2009), the author studies new types of cellular functions. The objects are grouped together in either overlapped or non-overlapped clusters. With an index that enables the evaluation of overlapped clusters, the author would be able to evaluate the resulting groups, according to the study of diseases related to diverse cellular functions. However, although there are plenty of external non-overlapped indexes (Brun et al., 2007; Wu et al., 2009) and a significant amount of research in overlapping clustering, there is a lack of external validation indexes for assessing and comparing overlapped solutions. Finally, taken into account indexes for overlapped clusters, Campo et al. (2014) presented a preliminary study about stability analyses in the context of overlapped clusters and developed an initial index to assess overlapped solutions. However, it failed to show the expected values in some basic cases.

In this study a novel index is presented based on an intuitive probabilistic approach. The new index works with the probability of finding any pair of objects in each solution and in both solutions simultaneously. The behavior of the proposed index is shown in the

presence of overlapped and disjoint clusters, when two clustering solutions for a same dataset are analyzed. Comparisons with classical external indexes such as FM, Jaccard (JAC) and Adjusted Rand Index (ARI) are performed on artificial and real datasets. Also, a real-life case from YouTube is presented, in which classical indexes fail because they show false differences when clusters become more overlapped.

The remainder of this paper is organized as follows. Section 2 presents a detailed analysis of the new index and an explanation of its factors. Section 3 describes the experiments performed with artificial and real datasets, and with a particular social network, and discusses the results obtained. Finally, conclusions are drawn and future research is suggested in Section 4.

2. A probabilistic approach for designing the new index

In this section, we introduce a new index for evaluating overlapped clustering solutions. First, notation and basic definitions are outlined. Next, a probabilistic approach for analyzing and designing the proposed index is presented, as well as an application example. The new index is also compared with some classical indexes to show its advantages when measuring overlapped solutions.

Given a set $S = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$, which is comprised of N objects, a clustering algorithm partitions them into a collection of subsets $C = \{c_1, \dots, c_k\}$ called clusters. The union of clusters in such partition forms a covering of the original set of objects: $\cup_{i=1}^k c_i = S$. Similarly, another algorithm or an equivalent one with different parameters over the same dataset could generate an alternative partition of k' clusters: $C' = \{c'_1, \dots, c'_{k'}\}$. Since each individual object could be grouped into more than one cluster, it is important to note that the number of elements in the clusters could be greater than or equal to N . For example, two clustering solutions are depicted in Figure 1: C with $k = 1$ and C' with $k' = 2$. In Figure 1.a), the solution is composed of a single cluster, c_1 , that groups all of the objects together. In Fig. 1.b) there are two clusters. Cluster c'_1 groups all objects and c'_2 groups all but one. In this scenario, both clusters of C' share $N - 1$ objects and are said to be *overlapped*. Given that every pair of objects that exists in one solution could be found in

the other one, and vice versa, a similarity value close to 1 should be expected if an external index is applied. However, the values obtained by classical indexes such as FM, ARI and JAC are 0.692, 1.238 and 0.478, respectively.

To overcome the evident misbehavior of classical indexes when overlapped clusters are present, a new index is proposed considering the probability that any pair of objects could be found in a given solution or in both solutions. Consequently, consider the cluster c_i of a given solution. Assuming that all of the objects have the same chance of being grouped into any cluster, this probability can be estimated as

$$Pr((\mathbf{s}_x, \mathbf{s}_y) \in c_i) = \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{|c_i|(|c_i| - 1)}{N(N - 1)}, \quad (1)$$

where $|c_i|$ is the number of elements in cluster c_i . The numerator represents the number of pairs that can be found with $|c_i|$ elements. In order to normalize it, the denominator represents a similar situation where all of the objects are grouped together in a single cluster; hence any possible pair could be found.

Taking into account the previous analysis, consider the solution C , where

$$\tilde{p} = \frac{\sum_{i=1}^k \binom{|c_i|}{2}}{k \binom{N}{2}}, \quad (2)$$

estimates the probability of finding a pair of elements in any cluster c_i for all of the existing clusters k . The numerator accumulates all of the pairs found in each cluster. The denominator represents a normalization factor, which acts as if all of the objects were grouped together. The k factor considers the situation where the overlapping is complete up to all k clusters. An identical reasoning could be applied to obtain a comparable expression for C' ,

$$\tilde{p}' = \frac{\sum_{j=1}^{k'} \binom{|c'_j|}{2}}{k' \binom{N}{2}}. \quad (3)$$

The same analysis described for \tilde{p} and \tilde{p}' can be performed for both solutions together. Therefore,

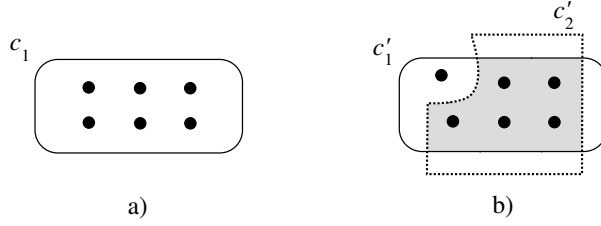


Figure 1: This illustrative example depicts two solutions: a) C with $k = 1$, and b) C' with overlapping and $k' = 2$. The shaded area includes the common elements between the overlapped clusters.

$$Pr((\mathbf{s}_x, \mathbf{s}_y) \in c_i \wedge (\mathbf{s}_x, \mathbf{s}_y) \in c'_j) = \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{|c_i \cap c'_j| (|c_i \cap c'_j| - 1)}{N(N-1)}, \quad (4)$$

can be seen as an approximation to the probability that the pair of data points $(\mathbf{s}_x, \mathbf{s}_y)$ is present in both solutions. In this equation, $|c_i \cap c'_j|$ represents the number of elements in common in clusters c_i and c'_j . The whole expression stands for the event of drawing two objects that are in both clusters c_i and c'_j .

Now suppose that the same analysis is made for every possible pairing between clusters of C and C' . The probability of finding $(\mathbf{s}_x, \mathbf{s}_y)$ in both solutions can be estimated as

$$\tilde{t} = \frac{\sum_{i=1}^k \sum_{j=1}^{k'} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \frac{\max(n, n')}{N} \min(k, k')}, \quad (5)$$

where n and n' represent the number of objects that can be counted in solutions C and C' , respectively, considering every overlap. For example, in Figure 1.a), $n = 6$, and in Figure 1.b), $n' = 11$. Similarly to \tilde{p} and \tilde{p}' , the numerator of (5) counts all of the effective pairs of objects that can be found in both solutions simultaneously. The denominator acts once again as a normalization term. It basically covers the extreme scenario where all of the objects are clustered together several times. Just as in (2) and (3), $\binom{N}{2}$ counts the number of pairs that can be arranged given all N objects. Since there could be overlaps in both solutions, the given number of pairs should be multiplied by a factor. On the one hand, there could be as many overlaps as k in C and k' in C' . On the other hand, it was found that the matching between clusters of both solutions produces at most $\min(k, k')$ pairs of

clusters in the comparison. Finally, $\max(n, n')/N$ is the average number of objects that can be found considering overlaps.

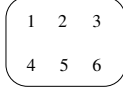
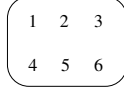
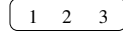
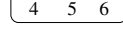
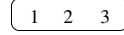
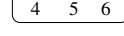


















With these elements in mind, the new index for overlapped clusters (\mathcal{OC}) could be defined as the ratio between the probability of finding two items grouped together in both solutions and the maximum probability of finding them in one of the given solutions. That is,

$$\mathcal{OC} = \frac{\tilde{t}}{\max(\tilde{p}, \tilde{p}')} \quad (6)$$

For the example in Figure 1, the new index produces the following values. When (2) is applied to the solution in Figure 1.a), $\tilde{p} = 1$ is obtained. Then, using (3) in Figure 1.b), $\tilde{p}' = 0.833$, and using (5), $\tilde{t} = 0.909$. Finally, when (6) is employed the new index is $\mathcal{OC} = 0.909/\max(1, 0.833) = 0.909$. The same experiment was performed using 1000 objects and the values obtained for FM, ARI and JAC were 0.707, 1.200 and 0.500, respectively, whereas $\mathcal{OC} = 0.999$. These examples show that the proposed index obtains an intuitively expected similarity between similar solutions with overlapped clusters, given that the probability of finding two objects grouped together in any of them tends effectively to 1.

The FM index tries to reflect the similarity of the two evaluated solutions considering the probability of randomly finding a pair of objects together, for each or both solutions at the same time. The problem is that it does not consider the existence of a pair of objects more than once, when the objects are overlapped in several clusters. With respect to ARI, the behavior with overlapped clusters is inconsistent. It fails to narrow the index score below 1. This behavior is observed because ARI is a corrected-for-chance version of the Rand Index, in which an expected value is subtracted in both the numerator and denominator. In practical applications, when overlapped clusters are present, such adjustment could produce values either below 0 or above 1. As is the case with the FM index, the Jaccard index, is the result of the ratio between a count of objects found in both solutions over the objects found in any of both solutions. As a result, the index does neither consider the overlapping situation nor counts the occurrences of repeated pairs of objects. This is why it is expected to fail in an overlapping scenario. Finally, the proposed index was carefully designed considering

Table 1: Index results for extreme artificial examples

	Solutions		Indexes			
	C	C'	FM	ARI	JAC	\mathcal{OC}
I	c_1 	c_1 	1.000	—	1.000	1.000
II	c_1   c_2	c_1   c_2	1.000	1.000	1.000	1.000
III	c_1   c_2  c_3	c_1   c_2  c_3	0.000	-0.250	0.000	0.000
IV	c_1  c_2  c_3  c_4  c_5  c_6 	c_1  c_2  c_3  c_4  c_5  c_6 	0.000	—	—	0.000

the overlap in the solutions and non-overlapping situations, which is an aspect that has not been considered in the design of the other indexes.

3. Experimental results and discussion

In this section, we present the results of the application of FM, ARI, JAC and the new \mathcal{OC} index. First, a set of artificial examples of extreme¹ situations is given. Next, tests are shown in which the overlap is gradually introduced. These examples show the behavior of the new index compared with standard measures in trivial cases. Then, the application of the index in several real datasets is described. Finally, the application of the \mathcal{OC} index for the analysis of social network data is presented.

3.1. Performance with artificial datasets

The first set of tests was performed over artificial clustering situations, where some extreme cases are analyzed. Also, examples with a gradual degree of overlap are given. The three tables in this subsection present basic tests that can help to better understand the behavior of the indexes under different types of overlap. In Table 1, the first column enumerates the examples given. Columns 2 and 3 depict the solutions that are compared. Finally, columns 4 – 7 show the values for FM, ARI, JAC and \mathcal{OC} , respectively. All of the examples in Table 1 have six data points that were clustered through one to six clusters.

Examples I and II show a pair of identical solutions with different configurations. In Example I, there is only one cluster in each solution, and every pair of objects that can be found in one cluster can also be found in the other one. In Example II, there are two clusters in each solution, and every pair of objects found in one cluster can also be found in a cluster from the other solution. In all of these cases, a value of 1.00 is expected, since the complete equivalence of both solutions is evident. In fact, all of the indexes can detect such situation, except for ARI, which produces no value at all in Example I. This is because the expected and maximum values of the denominator in the definition of ARI (Hubert & Arabie, 1985) are equal and a division by zero is returned. The last two examples in Table 1 present situations where no similarity between both solutions exists. In Example III, none of the possible pairs of objects found in solution C can be found in solution C' . In Example IV, both solutions cannot form any pair of objects at all since each data point is in a different cluster. In this case, a value of 0.00 is expected for each example because no pairs of data could be found in the first and second solutions simultaneously. Once again, the only index that disagrees with this intuition is ARI in Example III. In this case, the numerator of the index is defined as a difference between an observed and an expected value. Therefore, a negative score is computed when the observed value is lower than the expected one. In addition, in Example IV, ARI and JAC present a division by zero since no pairs are formed at all, and the indexes cannot provide a value. The results show clearly that the proposed

¹In the sense of expected 0/1 values for indexes.

index can measure basic situations without overlap.

The examples in Table 2 represent several scenarios with gradual overlap. The table has the same columns as in the previous case and all of the examples given in it have six data points clustered. Solutions labeled with C are always identical (the reference solution) and all of the objects are always grouped into a single cluster c_1 . Solutions named C' have two clusters that are incrementally overlapped between them, ranging from no overlap in Example I to a full double overlap in Example VII. For instance, in Example II the element 3 appears in both clusters of C' , but a repeated pair is not generated. In this case, three new pairs arise from the interaction between object 3 and each object from cluster c'_2 , and the proposed index can identify such situation. By contrast, in Examples III to VII the increasing overlapping effect allows repeated pairs to be produced. Such pairs are formed by objects that belong to both clusters. For example, the pair formed by objects 2 and 3 can be found in both clusters c'_1 and c'_2 . Thus, given the slowly increasing overlap in examples I to IV, all of the indexes but ARI show a corresponding incremental behavior.

Hence, since there is more overlap, more pairs of objects from solution C can be found in C' , with some extra repeated pairs due to the overlap itself. It is expected that, while one solution increases its overlapped clusters, the index tends to raise its value as there are more matchings between solutions. In Example IV and successive ones, all of the possible pairs of objects among six data points can be actually found in both solutions, in addition to some repeated ones in C' . Classical indexes (FM and JAC) increase up to Example V, where they begin to decrease. However, it is expected that the indexes continue to rise given the increasing overlap. The ARI once again shows disagreeing values in these examples. As explained earlier for Table 1, ARI might produce negative values in certain cases. By contrast, as the overlap progresses and duplicated pairs of points are found more frequently, the \mathcal{OC} index follows this behavior with a monotonic increasing value through all of the last examples, achieving the top 1.00 score when a perfect overlap of both clusters exists.

Finally, Table 3 presents more extreme situations. The structure of this table is the same as the previous ones. In the first example, solution C is composed of a single cluster with all of the objects contained in it. By contrast, solution C' has two overlapped clusters

Table 2: Index results for some gradually overlapped artificial examples


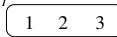
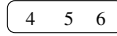
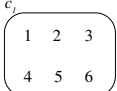
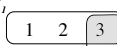
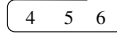
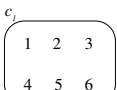
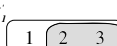
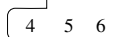
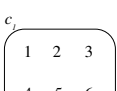

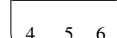
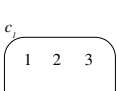


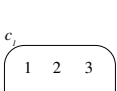

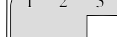
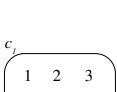


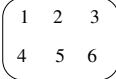



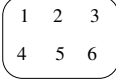
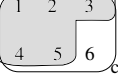
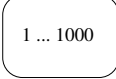
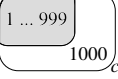
	Solutions		Indexes			
	C	C'	FM	ARI	JAC	\mathcal{OC}
I	c_1 	c_1  c_2 	0.632	0.000	0.400	0.400
II	c_1 	c_1  c_2 	0.665	-2.186	0.442	0.514
III	c_1 	c_1  c_2 	0.695	2.347	0.483	0.650
IV	c_1 	c_1  c_2 	0.721	1.444	0.520	0.800
V	c_1 	c_1  c_2 	0.700	1.324	0.489	0.840
VI	c_1 	c_1  c_2 	0.692	1.238	0.478	0.909
VII	c_1 	c_1  c_2 	0.692	1.179	0.478	1.000

Table 3: Index results for some extremely overlapped artificial examples

	Solutions		Indexes			
	C	C'	FM	ARI	JAC	\mathcal{OC}
I	c_i 	$c'_i = c'_2$ 	0.692	1.179	0.478	1.000
II	$c_1 = \dots = c_{999}$ 	$c'_1 = \dots = c'_{1000}$ 	0.001	1.000	0.001	1.000
III	c_i 	c'_1 	0.692	1.238	0.478	0.909
IV	c_i 	c'_i 	0.707	1.200	0.500	0.999

with all of the objects grouped together. The proposed index shows a complete equivalence between both solutions since any pair of objects can be found in them. None of the other indexes present this similarity. In the second example, something similar occurs but with more overlapped clusters. Solution C has 999 complete overlaps and solution C' has 1000. Once again, \mathcal{OC} and this time ARI present a complete equivalence, while others decrease to almost zero. These cases demonstrate that the proposed index does not change as the overlap increases with a high number. The index maintains a value of 1.00 under any number of complete overlaps, which is the expected behavior since the pairs of objects are maintained through the overlaps. Other indexes fail to show this and their values decrease with higher overlaps.

Example III is exactly the same as Example VI from Table 2. In solution C , a complete overlap is observed among all of the objects, while solution C' contains one cluster with a complete overlap and one cluster that groups all of the objects but one. The proposed index shows a value relatively close to 1.00. This is because almost all of the pairs formed in solution C can be found twice in solution C' , but a few others can be found only once. This is consistent with the fact that not all of the pairs have the same proportion of appearance in the second solution, and the \mathcal{OC} index can reflect this irregular situation.

Finally, Example IV takes the previous example to the limit, where solution C has a thousand objects grouped all together in one cluster. Solution C' has cluster c'_1 , which groups all of the objects, and c'_2 , which groups all but one. In the last two examples (III and IV), solution C has one cluster with all of the elements grouped together. By contrast, solution C' contains two clusters: in the first one, all of the elements are grouped together, but in the second one all of the elements but one are grouped. The difference between Examples III and IV is that the number of elements considered in the latter tends to be high. As demonstrated by these two experiments, the \mathcal{OC} index accurately reflects the fact that all pairs of data can be found in both solutions, obtaining a value close to 1.00, as expected. Other indexes fail when an almost complete overlap is presented. Our proposed index tends to obtain a maximum score when the number of elements is relatively high.

3.2. Benchmarking with real datasets

Four well-known databases² were used for performing the experiments on real datasets: Iris, Wine, Yeast and Glass (Lichman, 2013). The Iris dataset has four attributes and 150 patterns distributed in three classes of 50 patterns each (Fisher, 1936). Only one of the three classes is linearly separable from the others, which have many patterns that are very close in the attribute space. The Wine dataset represents the measure and analysis of 13 chemical attributes of an Italian wine taken from different vineyards. This dataset of 178 patterns is distributed in three groups: A, B and C, with 59, 71 and 48 patterns each, respectively. The Yeast dataset is based on a study of yeast and it is intended to determine the location of its proteins in the cell. It has 1484 patterns distributed in 10 groups with 463, 429, 244, 163, 51, 44, 37, 30, 20 and 5 elements each, and 8 attributes have been measured. Finally, the Glass dataset has 9 attributes and 214 patterns distributed in 7 groups. These datasets are freely available for general purpose use, and they are widely used in the academic community. They were selected for their small size and adequacy for the detailed analysis of the proposed measure.

²<http://archive.ics.uci.edu/ml/datasets/>

Table 4: Results for FM, ARI, JAC and \mathcal{OC} indexes using Iris, Wine, Yeast and Glass databases. The reference solutions C have 4 or 25 clusters and zero overlap ($V_n = 0$). Solutions C' have 25 and 100 clusters, taking $V_n = 0$ and $V_n = 1$

	clusters in C and C'	FM		ARI		JAC		\mathcal{OC}	
		$V_n = 0$	$V_n = 1$	$V_n = 0$	$V_n = 1$	$V_n = 0$	$V_n = 1$	$V_n = 0$	$V_n = 1$
Iris	$k = 4$ vs $k' = 25$	0.33	0.30	0.16	4.26	0.14	0.10	0.14	0.38
	$k = 4$ vs $k' = 100$	0.17	0.16	0.03	-0.66	0.03	0.03	0.03	0.11
	$k = 25$ vs $k' = 100$	0.33	0.23	0.23	-0.34	0.14	0.09	0.14	0.37
Wine	$k = 4$ vs $k' = 25$	0.40	0.32	0.23	9.33	0.17	0.12	0.17	0.48
	$k = 4$ vs $k' = 100$	0.19	0.18	0.06	-0.52	0.04	0.04	0.04	0.14
	$k = 25$ vs $k' = 100$	0.34	0.23	0.26	-0.24	0.15	0.10	0.17	0.39
Yeast	$k = 4$ vs $k' = 25$	0.32	0.23	0.15	6.61	0.13	0.08	0.13	0.35
	$k = 4$ vs $k' = 100$	0.16	0.14	0.04	-0.58	0.03	0.03	0.03	0.11
	$k = 25$ vs $k' = 100$	0.29	0.18	0.21	-0.30	0.13	0.07	0.14	0.33
Glass	$k = 4$ vs $k' = 25$	0.33	0.27	0.10	3.77	0.11	0.08	0.11	0.30
	$k = 4$ vs $k' = 100$	0.15	0.14	0.02	-0.75	0.02	0.02	0.02	0.09
	$k = 25$ vs $k' = 100$	0.38	0.24	0.28	-0.36	0.17	0.09	0.18	0.43

A self-organizing map (SOM) (Kohonen, 1998) was used for clustering the data. Given that several neurons in a region of the map may be considered as a single group, incrementally overlapped clusters can be easily analyzed with different levels of neighborhood between neurons. To process the datasets, each map was trained with different numbers of neurons (clusters). All of the experiments were performed with a rectangular topology, grid shape, principal component analysis initialization and training iterations set to 100 epochs. To consider overlap between clusters, the topological closeness between neurons in the map has been taken into account with a Von Neumann neighborhood. When it is equal to zero ($V_n = 0$), each neuron represents a single cluster. When $V_n = 1$ is considered, each neuron and its four adjacent neighboring neurons (north, south, east and west) are considered as part of the same cluster. Thus, when $V_n = 1$ is used, each neuron and its neighbors may overlap between them, and some patterns are associated with more than one cluster. This is how overlapped clusters are formed in a SOM.

Table 3.2 presents the results obtained over the real datasets. The table is divided as follows: column 1 contains the name of each dataset, column 2 shows the number of clusters considered for C and C' , and columns 3–6 present the values obtained for each experiment and for each index. The last four columns are divided into values with and without overlap ($V_n = 0$ and $V_n = 1$) in the solution C' . Six different experiments were performed for each dataset: $k = 4$ vs. $k' = 25$, $k = 4$ vs. $k' = 100$ and $k = 25$ vs. $k' = 100$. This is with $V_n = 0$ and $V_n = 1$ for solution C' . In all of the cases, no overlap was considered for the reference solution C . For the Iris dataset, a decrease is observed in the FM index, not only when $k = 4$ and k' change from 25 to 100, but also when overlap ($V_n = 1$) is considered. The ARI produces a value over 1.00 when overlap is considered in the experiment $k = 4$ vs. $k' = 25$. The opposite behavior is observed when other sizes of clusters are considered and overlap is taken into account, showing values below 0. The JAC index exhibits a similar behavior to the FM index: it decreases when overlap is considered and when more clusters are taken into consideration in solution C' . Finally, the \mathcal{OC} index decreases when a higher number of clusters is considered in C' , but increases when overlapped clusters are analyzed. This is an expected behavior since it is consistent with the fact that, when overlapped clusters are

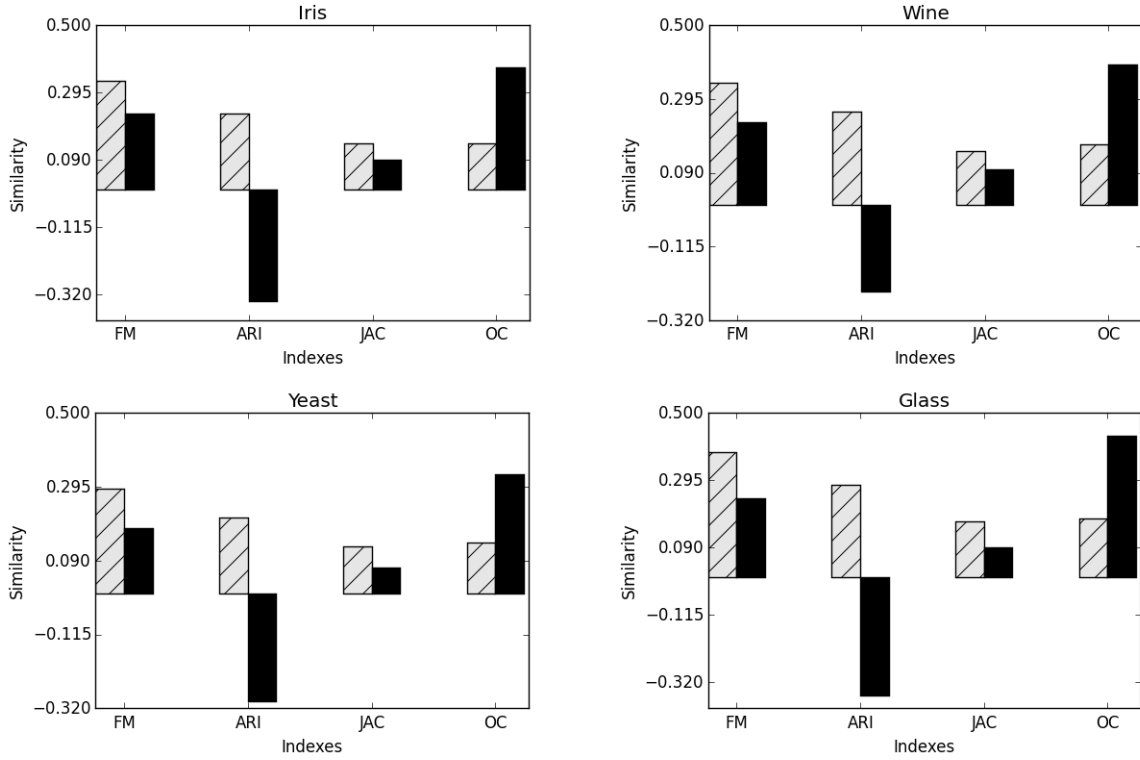


Figure 2: Bar plots of FM, ARI, JAC and \mathcal{OC} indexes for Iris, Wine, Yeast and Glass databases. The reference solutions C have 25 clusters and zero overlap ($V_n = 0$). Solutions C' have 100 clusters with $V_n = 0$ (gray diagonal striped bars) and $V_n = 1$ (black bars).

used, the probability of finding more matching pairs of points between solutions is higher.

The analyses for Wine, Yeast and Glass datasets are very similar to the previous one. In these experiments, the values of FM and JAC indexes decrease when a higher number of clusters is considered. This is also the case when overlap is taken into account. The ARI shows exactly the same behavior as in the previous dataset. With respect to the \mathcal{OC} index, a remarkable increase is observed when overlapped clusters are analyzed. However, it decreases when solution C' has more clusters.

Figure 2 presents the results of the experiment where the reference solution C have $k = 4$ and $V_n = 0$, and the C' solutions has $k = 100$ and either $V_n = 0$ or $V_n = 1$. For all of the datasets, when the overlap increases, classical indexes show a notable decrement, while the proposed index shows an increment. This behavior is consistent with the intuition that,

given the existence of overlapped clusters, it should be more likely to find a pair of objects in common in both solutions.

In these experiments, an increment is observed in all of the indexes but ARI when the number of clusters of C is close to the number of clusters of C' . This is due to the data dispersion in C' : when there are more clusters, the data patterns are spread through more neurons, thereby reducing the value of the index. This is observed when the experiment $k = 25$ vs. $k' = 100$ is analyzed. In the case of FM and JAC, the scores also decrease when overlap is considered, while the proposed index always shows an increment for $V_n = 1$ with respect to $V_n = 0$. This is because the classical indexes do not handle overlapped clusters properly, whereas \mathcal{OC} does. With this in mind, should be noted that when there are overlapped clusters, both FM and JAC indexes do not count the matchings between groups appropriately. This explains why FM and JAC barely decrease, and \mathcal{OC} hardly increases with overlap.

In summary, with artificial or real datasets, the proposed index is effective for assessing clustering solutions in which there can be overlapped clusters. Moreover, \mathcal{OC} shows reliable and confident results with extreme overlapping cases, thus enabling a better understanding and comparison of the outcome of clustering algorithms.

3.3. Social networking application

This subsection describes the experiments performed over a real dataset taken from a social network. The results of the application of the proposed measure over the YouTube dataset are analyzed and discussed (Yang & Leskovec, 2013). YouTube is a video-sharing social network. Users can create groups or communities to share their videos, and other users can join them. This dataset captures the relation of a group of users of the social network through communities. The dataset is comprised of communities that are defined as groups of two or more users who share similar interests. Each community, which is considered as a cluster of users, is described in the dataset as a list of user IDs. One user may belong to one or more communities. When a user belongs to several communities, those communities are said to be overlapped. The level of overlap of a community depends on how many of its

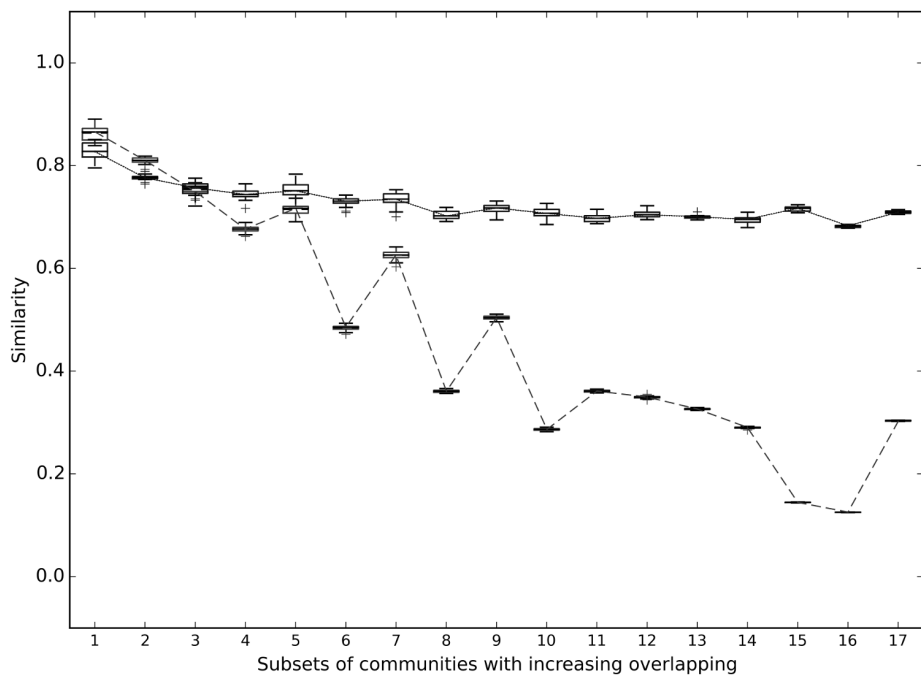


Figure 3: Boxplot of FM and \mathcal{OC} indexes for the social network (YouTube) dataset. The dashed line corresponds to the FM index and the continuous one, to the \mathcal{OC} index.

participants also belong to other communities. The resulting dataset, after preprocessing and removing communities with less than 10 users, contains 37038 users and 2087 communities. Communities with less than 10 users showed an almost non-existent overlapping behavior, which would affect the focus and interpretability of indexes when overlap is tested.

The experiment performed over this dataset involved sorting the groups C_j by their degree of overlap. Solution C'_j was taken from solution C_j with different levels of perturbation. Random modifications were applied and users were added to random communities. The original dataset was then divided into several subsets. Since the communities are arranged by levels of overlap, the first subset of 35 communities has zero overlap. Each of the following subsets considered in this study represents a different level of increasing overlap. Communities are grouped into subsets with a similar level of overlap until no more communities are available. Each subset has at least three times the number of communities as the first one (with zero overlap) in order to ensure that all of the solutions have a minimum number of elements for calculating the indexes. The last subset includes the communities with a higher level of overlap. Therefore, the original dataset was divided into 17 disjoint subsets ranging from zero to the maximum overlap.

Figure 3 shows boxplots for the FM and \mathcal{OC} indexes for a 10% random perturbation of users within communities. Each box corresponds to the median of 20 runs over the corresponding subset of the dataset. The abscissa axis shows the increasing level of overlap of each subset from zero (subset labeled 1) to the maximum overlap (subset labeled 17). Communities with no overlap reach similar values, for both indexes, between 0.8 and 0.9. As the overlap increases, a remarkable decrease is observed in the FM index, reaching values barely upon 0.1 when the overlap is very high. The values obtained by the \mathcal{OC} index are more stable when the overlap increases, thereby demonstrating the capability of \mathcal{OC} to be immune to the overlap. Moreover, with a higher overlap in the subsets, the FM curve falls in a fluctuating manner. By contrast, the \mathcal{OC} curve shows a smooth behavior, maintaining high values. Therefore, we conclude that the proposed index is effective for measuring similarities in real scenarios where there are overlapped clusters. Furthermore, the \mathcal{OC} index exhibits a more stable behavior than classical measures such as FM, irrespective of the presence of

overlap between subsets.

4. Conclusions and future work

In this study, we proposed a new index (\mathcal{OC}) for comparing solutions that may have a certain degree of overlap. The proposed index was designed from an intuitive probabilistic approach and was then compared with classical approaches, such as Fowlkes-Mallows, Adjusted Rand and Jaccard indexes. For simple artificial examples, these indexes showed unexpected behaviors, while a more reliable situation was observed with the \mathcal{OC} index. Experiments performed with benchmark datasets and real data from a social network confirmed these findings. On the one hand, classical indexes tended to show fewer similarities between solutions as the overlap increased. On the other hand, the proposed index was immune to the overlap and performed accurately, showing the level of similarity between clustering solutions. It should be noted that the \mathcal{OC} index also performed well when there was no overlap. Thus, the proposed index can be applied to any type of solution, regardless of the presence of overlapped clusters.

In future research, we will perform experiments using the proposed index in order to analyze the stability of clustering solutions with any degree of overlap.

5. Funding and acknowledgments

This study was supported by the National Scientific and Technical Research Council (CONICET) [PIP 2013-2015 117], Universidad Nacional del Litoral (UNL) [CAI+D 2011 548] and National Agency of Science and Technology Promotion (ANPCyT) [PICT 2014 2627].

6. References

Alvari, H., Hashemi, S., & Hamzeh, A. (2013). Discovering overlapping communities in social networks: A novel game-theoretic approach. *AI Communications*, 26, 161–177.

- Amelio, A., & Pizzuti, C. (2014). Overlapping Community Discovery Methods: A Survey. In Ş. Gündüz-Öğüdücü, & A. Ş. Etaner-Uyar (Eds.), *Social Networks: Analysis and Case Studies* Lecture Notes in Social Networks (pp. 105–125). Springer Vienna.
- Ben-Hur, A., & Guyon, I. (2003). Detecting Stable Clusters Using Principal Component Analysis. In M. Brownstein, & A. Khodursky (Eds.), *Functional Genomics* number 224 in Methods in Molecular Biology (pp. 159–182). Humana Press. doi:doi:10.1385/1-59259-364-X:159.
- Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, *40*, 807–824. doi:doi:10.1016/j.patcog.2006.06.026.
- Campo, D., Stegmayer, G., & Milone, D. (2014). Stability analysis in overlapped clusters. *Iberoamerican Journal of Artificial Intelligence*, *17*, 79–89.
- Chakraborty, T. (2015). Leveraging disjoint communities for detecting overlapping community structure. *J. Stat. Mech.*, *2015*, P05017. doi:doi:10.1088/1742-5468/2015/05/P05017.
- Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*, 179–188. doi:doi:10.1111/j.1469-1809.1936.tb02137.x.
- Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, *78*, 553–569. doi:doi:10.1080/01621459.1983.10478008.
- Gopalan, P. K., & Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, *110*, 14534–14539. doi:doi:10.1073/pnas.1221839110.
- Gossen, T., Kotzyba, M., & Nürnberger, A. (2014). Graph clusterings with overlaps: Adapted quality indices and a generation model. *Neurocomputing*, *123*, 13–22. doi:doi:10.1016/j.neucom.2012.09.046.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *J. Intell. Inf. Syst.*, *17*, 107–145. doi:doi:10.1023/A:1012801612483.
- Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*, 3201–3212. doi:doi:10.1093/bioinformatics/bti517.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. doi:doi:10.1007/BF01908075.
- Kalinka, A. T., & Tomancak, P. (2011). linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics (Oxford, England)*, *27*, 2011–2012. doi:doi:10.1093/bioinformatics/btr311.
- Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*, 1–6. doi:doi:10.1016/S0925-2312(98)00030-7.
- Lacroix, V., Cottret, L., Thébault, P., & Sagot, M. F. (2008). An Introduction to Metabolic Networks and Their Structural Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *5*,

594–617. doi:doi:10.1109/TCBB.2008.79.

Lichman, M. (2013). UCI machine learning repository.

Liu, D., Blenn, N., & Mieghem, P. V. (2013). Characterising and modelling social networks with overlapping communities. *International Journal of Web Based Communities*, *9*, 371–391. doi:doi:10.1504/IJWBC.2013.054909.

McGarry, K. (2013). Discovery of functional protein groups by clustering community links and integration of ontological knowledge. *Expert Systems with Applications*, *40*, 5101–5112. URL: <http://www.sciencedirect.com/science/article/pii/S0957417413001905>. doi:doi:10.1016/j.eswa.2013.03.027.

Meilă, M. (2007). Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, *98*, 873–895. doi:doi:10.1016/j.jmva.2006.11.013.

Meilă, M., & Heckerman, D. (2001). An Experimental Comparison of Model-Based Clustering Methods. *Machine Learning*, *42*, 9–29. doi:doi:10.1023/A:1007648401407.

Skillicorn, D. (2007). *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. CRC Press.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., & Provart, N. J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment*, *32*, 1633–1651. URL: <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3040.2009.02040.x/abstract>. doi:doi:10.1111/j.1365-3040.2009.02040.x.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.*, *11*, 2837–2854.

Wang, Z., Zhang, D., Zhou, X., Yang, D., Yu, Z., & Yu, Z. (2014). Discovering and Profiling Overlapping Communities in Location-Based Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*, 499–509. doi:doi:10.1109/TSMC.2013.2256890.

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, *6*, 227. URL: <http://dx.doi.org/10.1186/1471-2105-6-227>. doi:doi:10.1186/1471-2105-6-227.

Wu, J., Chen, J., Xiong, H., & Xie, M. (2009). External validation measures for K-means clustering: A data distribution perspective. *Expert Systems with Applications*, *36*, 6050–6061. doi:doi:10.1016/j.eswa.2008.06.093.

Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study. *ACM Comput. Surv.*, *45*, 43:1–43:35. doi:doi:10.1145/2501654.2501657.

Xu, R., & Wunsch, D. (2008). *Clustering*. John Wiley & Sons.

Yang, J., & Leskovec, J. (2013). Defining and evaluating network communities based on ground-truth.

Knowl Inf Syst, 42, 181–213. doi:doi:10.1007/s10115-013-0693-z.

Zhou, X., Liu, Y., Zhang, J., Liu, T., & Zhang, D. (2015). An ant colony based algorithm for overlapping community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 427, 289–301. doi:doi:10.1016/j.physa.2015.02.020.