

The Area Under the ROC Curve as a Measure of Clustering Quality

Pablo A. Jaskowiak^{1,*}, Ivan G. Costa², Ricardo J. G. B. Campello³

Abstract

The Area Under the the Receiver Operating Characteristics (ROC) Curve, referred to as AUC, is a well-known performance measure in the supervised learning domain. Due to its compelling features, it has been employed in a number of studies to evaluate and compare the performance of different classifiers. In this work, we explore AUC as a performance measure in the unsupervised learning domain, more specifically, in the context of cluster analysis. In particular, we elaborate on the use of AUC as an internal/relative measure of clustering quality, which we refer to as Area Under the Curve for Clustering (AUCC). We show that the AUCC of a given candidate clustering solution has an expected value under a null model of random clustering solutions, regardless of the size of the dataset and, more importantly, regardless of the number or the (im)balance of clusters under evaluation. In addition, we elaborate on the fact that, in the context of internal/relative clustering validation as we consider, AUCC is actually a linear transformation of the Gamma criterion from Baker and Hubert (1975), for which we also formally derive a theoretical expected value for chance clusterings. We also discuss the computational complexity of these criteria and show that, while an ordinary implementation of Gamma can be computationally prohibitive and impractical for most real applications of cluster analysis, its equivalence with AUCC actually unveils a much more efficient algorithmic procedure. Our theoretical findings are supported by experimental results. These results show that, in addition to an effective and robust quantitative evaluation provided by AUCC, visual inspection of the ROC curves themselves can be useful to further assess a candidate clustering solution from a broader, qualitative perspective as well.

Keywords: clustering, clustering validation, internal validation, relative validation, area under the curve, AUC, receiver operating characteristics, ROC, area under the curve for clustering, AUCC, quantitative clustering evaluation, qualitative/visual clustering evaluation

*Corresponding author

Email address: pablo.andretta@ufsc.br (Pablo A. Jaskowiak)

¹Federal University of Santa Catarina, Joinville, Santa Catarina, Brazil.

²Institute for Computational Genomics, RWTH Aachen University Medical Faculty, Aachen, Germany.

³School of Mathematical and Physical Sciences, University of Newcastle, Australia.

1. Introduction

The introduction of Receiver Operating Characteristics (ROC) to the machine learning community is often attributed to the work of Spackman (1989). Since then, ROC analysis has gained popularity in the supervised learning domain, in part as a result of the drawbacks observed with accuracy-based evaluations of classifiers (Bradley, 1997; Provost and Fawcett, 1997; Provost et al., 1998; Huang and Ling, 2005; Fawcett, 2006; Flach, 2010), especially for class imbalanced problems. Currently, ROC analysis stands as a valuable tool to visualize, evaluate and compare the performance of different classifiers (Majnik and Bosnić, 2013; Hernández-Orallo et al., 2013).

Given a classifier and a dataset for which desired classification outcomes (i.e., actual class labels) are available, the first step towards performing a ROC analysis consists in deriving statistics that relate classifier predictions with the corresponding desired outcomes. In the case of a binary classification problem with a positive and a negative class, classifier predictions can be deemed True Positive (TP), False Positive (FP), True Negative (TN), or False Negative (FN) with respect to actual class labels. If the classifier under evaluation produces as output a probability or a score for each object, representing its likelihood or degree of membership to a class, a ROC curve can be derived by plotting the values of False Positive Rate ($FPR = FP/N$) against those of True Positive Rate ($TPR = TP/P$), where P and N stand for the cardinality of the positive and negative classes, respectively. In this case, each point in the curve is associated with a classification threshold, which lies within the interval of the scores produced by the classifier. For each threshold value, objects are deemed as positive or negative according to their classification score relative to the threshold, and the respective TPR and FPR values are calculated and plotted, as illustrated in Fig. 1(a). The diagonal line in this figure ($TPR = FPR$) accounts for the expected performance of a random classifier. A classifier with a curve close to the top-left corner of the graph is usually preferred, whereas a classifier with a curve below the diagonal line performs worse than random. By reversing the classification predictions, its new ROC curve will be mirrored around the diagonal line. For detailed discussions on ROC graphs, see e.g. (Fawcett, 2004, 2006; Flach, 2010).

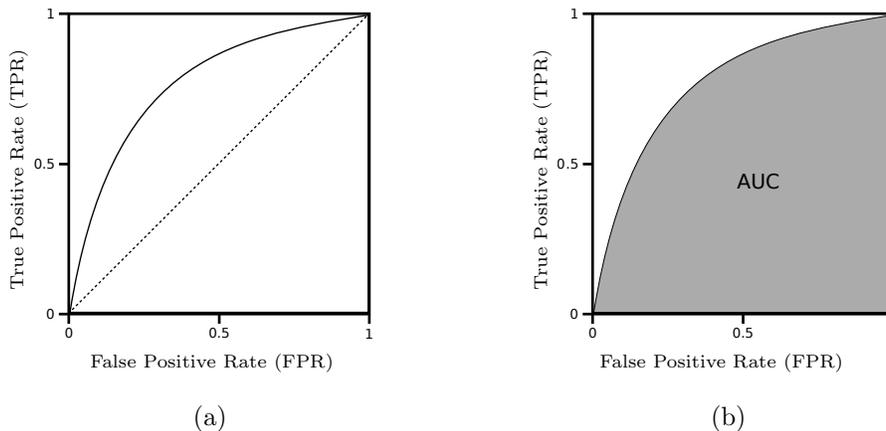


Figure 1: Examples of ROC graphs. A ROC curve (a) can be obtained from a scoring classifier. The dotted line in (a) corresponds to a classifier with performance comparable to that of a completely random classifier. In (b), the Area Under the Curve (AUC) of the corresponding curve is highlighted.

Even though the ROC graph has visual appeal, an aggregated scalar value is typically obtained, in order to make the comparison among different classifiers straightforward. A well-known choice, which is typically regarded as the most important statistics derived from ROC curves (Flach, 2010), is the Area Under the (ROC) Curve (AUC), as illustrated in Fig. 1(b). The AUC of a classifier/model prediction consists of a single value in the $[0, 1]$ interval that, from a statistical standpoint, can be regarded as the probability that it will rank (or score) a randomly selected positive object higher than a randomly selected negative one (Fawcett, 2006; Flach et al., 2011). From this observation it follows that, in general: (i) the larger the AUC value, the better is the performance of the classifier under evaluation; (ii) values of AUC around 0.5 indicate the expected performance of random classifiers¹; and (iii) values below 0.5 indicate a worse than random classifier. Over the years, AUC became one of the standard measures employed to assess classifiers' performance. Nowadays, it is widely accepted that AUC should be favored over accuracy, based on both theoretical and empirical evidence. For instance, Huang and Ling (2005) compared the evaluations obtained with these two measures regarding their consistency and discriminant power. Their results suggest that both measures have a high degree of consistency in their evaluations, i.e., they do not

¹In fact, non-random classifiers can also exhibit such a performance (Flach, 2010).

contradict each other, but AUC has a better discriminant power, i.e., it can discriminate between classification models when accuracy cannot.

In this paper we take a different perspective on AUC, by considering it in the unsupervised learning domain. More specifically, we elaborate on the use of AUC as an internal/relative measure of clustering quality (Jain and Dubes, 1988; Xu et al., 2009; Hennig et al., 2015), which can be employed to evaluate and compare the results obtained from different clustering algorithms or parameterizations of a particular clustering algorithm. Hereafter we shall refer to this measure as Area Under the Curve for Clustering, or simply AUCC. The concept of AUC has been previously considered in the clustering scenario by Jaskowiak et al. (2012, 2013) and Giancarlo et al. (2013). In both cases, however, the AUC was employed within a limited scope, namely, to evaluate the agreement between proximity measures and the *external labels* of a dataset. The goal was to evaluate proximity measures for clustering gene expression microarray data and did not include any theoretical or experimental evaluation of AUC, let alone as an internal/relative criterion.

In contrast, the AUCC measure studied in this paper is defined as an internal, relative clustering evaluation criterion. It operates on the set of all pairs of data objects, with their similarity playing the role of “classification scores” and whether or not they belong to the same cluster (in a candidate clustering solution to be assessed) playing the role of “binary class labels”. In this setting, not only the area under the ROC curve can be computed as a quantitative measure of clustering quality, but visual-based, qualitative interpretation of the ROC curves themselves can also be undertaken.

As one of the main contributions in this paper, we theoretically show that the AUCC of a clustering solution has the same expected value as in classification (0.5) under the assumption of a relevant null model of random clusterings, regardless of the number of clusters or relative cluster sizes in the partitions under evaluation. As we will show, this is particularly desirable in scenarios comparing clustering results with different numbers of clusters. It is worth remarking that, while the study of the expected behavior of evaluation criteria for random clusterings (and the correction of their possible biases, commonly referred to as *adjustment for chance*) has been extensively studied in the context of *external* clustering evaluation, where a ground-truth clustering solution consisting of

external labels exist — see e.g. (Romano et al., 2016), this aspect has been surprisingly overlooked in the context of internal/relative (i.e. fully unsupervised) evaluation.

In addition to the above, we explore and theoretically elaborate on the result — originally and preliminarily described in (Jaskowiak, 2015) — that the quantity defined here as AUCC is actually a linear transformation of the Gamma internal clustering validity criterion introduced more than 40 years ago by Baker and Hubert (1975), for which we also formally derive a theoretical expected value for chance clusterings. Strategies to handle ties in (dis)similarity values are discussed for both AUCC and Gamma in light of their linear relationship as well as their theoretical expected values.

The Gamma criterion was the best performer in a previous evaluation study by Milligan (1981), in which 30 internal clustering validity criteria were assessed on the basis of their (dis)agreement — as measured by correlation — with external evaluations. Notwithstanding, Gamma’s computational time includes a prohibitive $O(n^4)$ term, where n is the number of data objects (dataset size). We show that AUCC has a significantly lower computational complexity, making it computationally tractable in real world problems involving datasets of practical relevance. This allowed us to carry out an empirical evaluation of AUCC in the context of relative clustering validation, relating its performance to that of 28 other commonly employed relative measures from the clustering literature.

The remainder of the paper is organized as follows. In Section 2 we provide a brief overview of performance evaluation in cluster analysis, which is commonly referred to as clustering validity or validation. In Section 3 we introduce AUCC as an internal/relative validity criterion for the unsupervised evaluation of clustering results. We then show that AUCC is equivalent to a linear transformation of the Gamma Index and that, under a null/random model assumption, both AUCC and Gamma have expected values that do not depend on the number (or the sizes) of clusters under evaluation. The section concludes with a discussion of how to handle ties in (dis)similarity values while preserving the theoretical findings in the paper. An empirical study is carried out in Section 4, involving both quantitative evaluations provided by AUCC as well as qualitative evaluations based on visual inspection of ROC curves. Final remarks and conclusions are drawn in Section 5.

2. Performance Evaluation in Cluster Analysis

Most clustering algorithms from the literature will produce an output regardless of the existence of actual clusters in the data. Even if one assumes that clusters exist, their number and distributions are usually unknown. In order to avoid the use of spurious (i.e., meaningless or poor) clustering results, one can resort to clustering validation techniques (Hennig et al., 2015). According to Jain and Dubes (1988), clustering validation can be defined as the set of tools and procedures that are used in order to evaluate clustering results in a quantitative and objective manner.

Clustering validation techniques can be broadly divided into external, internal, and relative (Jain and Dubes, 1988; Halkidi et al., 2001). External criteria are mostly employed in the evaluation of clustering results against a desired clustering solution known beforehand (i.e., a ground-truth). Although they are very useful for algorithm evaluation and comparison in controlled experiments, external criteria have limited applicability in practical cluster analysis scenarios, where a ground-truth doesn't exist (Färber et al., 2010; Jaskowiak et al., 2016). Internal validity criteria rely their evaluation only on clustering assignments and the data themselves. Internal criteria that are also relative can be used to assess and compare the quality of different partitions in a relative manner. For this reason, relative criteria are frequently employed in practical clustering applications, helping the selection of a final clustering solution for further inspection by a field practitioner.

The literature on relative validity criteria is extensive and a large number of measures have been proposed. These are usually conceived based on the idea that a good clustering solution (partition) should have compact and separate clusters (Halkidi et al., 2001). From different definitions of cluster compactness and separation, different relative validity measures arise. Back in the 1980's, Milligan (1981); Milligan and Cooper (1985) compared the performance of 30 validity criteria, mostly relative ones. Since then, new measures have been introduced, e.g., Rousseeuw (1987); Bezdek and Pal (1998); Halkidi and Vazirgiannis (2008); Moulavi et al. (2014), extensive reviews, assessments and evaluations of those have been performed, e.g., Maulik and Bandyopadhyay (2002); Vendramin et al. (2009, 2010); Arbelaitz et al. (2013), and different implementations of the measures have been made available, e.g., Brock et al. (2008); Charrad et al. (2014); Desgraupes (2016).

3. The Area Under the Curve as an Internal/Relative Measure

Consider a dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with n objects embedded in a space where a measure of similarity between pairs of objects can be defined (e.g., an Euclidean space with d dimensions, i.e., $\mathbf{x}_i = \{x_{i1}, \dots, x_{id}\}$ for $i = 1, \dots, n$). In addition, consider a clustering result in the form of a partition, that is, a labeling of the data into $2 \leq k \leq n - 1$ mutually exclusive clusters. Let $\mathcal{C} = \{C_1, \dots, C_k\}$ denote this partition, with the following properties:

$$C_1 \cup \dots \cup C_k = \mathbf{X},$$

$$C_i \neq \emptyset, \forall i,$$

$$C_i \cap C_j = \emptyset, \forall i, j \text{ with } i \neq j.$$

Notice that we can transform any clustering solution \mathcal{C} as above into a *pairwise* representation \mathcal{C}^p (a binary relation) composed of $n(n - 1)/2$ elements (object pairs), as follows:

$$\mathcal{C}^p(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \text{if } \exists l : \mathbf{x}_i, \mathbf{x}_j \in C_l, \\ 0 & \text{otherwise.} \end{cases}$$

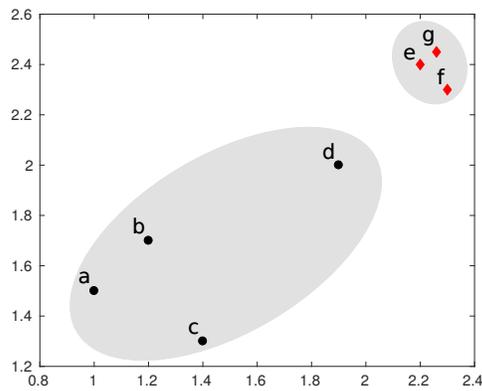
Let \mathbf{D} be a pairwise similarity matrix of the objects in dataset \mathbf{X} , from which a clustering solution \mathcal{C} to be evaluated was derived. The binary relation \mathcal{C}^p of \mathcal{C} can be used, along with the pairwise similarities \mathbf{D} , as input to ROC analysis. The rationale behind this type of evaluation is that object pairs belonging to the same cluster in a good partition \mathcal{C} should have higher similarities (or, conversely, lower dissimilarities) than those belonging to different clusters.

Once a clustering solution (i.e., a partition) is available for a given dataset, its corresponding Area Under the Curve for Clustering (AUCC) can be computed with the following procedure:

1. From the original dataset, compute a similarity matrix of the objects.
2. Obtain two arrays, which indicate, for each pair of objects, their pairwise:

- (a) Similarity: readily available from the similarity matrix;
 - (b) Clustering: 1 if the pair is in the same cluster; 0 otherwise.
3. Provide the two arrays as input to a standard ROC Analysis procedure in order to obtain the corresponding AUC of the clustering solution, which in this particular context we refer to as Area Under the Curve for Clustering (AUCC). The relation to the usual ROC Analysis in classification is straightforward: similarity values correspond to “classification thresholds” whereas pairwise clustering memberships (i.e., binary labels) correspond to the “true classes”.

The toy example in Figure 2 exemplifies the whole process.



(a) Data points and partition with two clusters

	a	b	c	d	e	f	g
a	1.00	0.82	0.72	0.35	0.05	0.03	0.00
b	0.82	1.00	0.72	0.52	0.23	0.20	0.18
c	0.72	0.72	1.00	0.45	0.14	0.15	0.09
d	0.35	0.52	0.45	1.00	0.68	0.68	0.63
e	0.05	0.23	0.14	0.68	1.00	0.91	0.95
f	0.03	0.20	0.15	0.68	0.91	1.00	0.90
g	0.00	0.18	0.09	0.63	0.95	0.90	1.00

(b) Similarity matrix

Pair	Pairwise	
	Clustering	Similarity
ab	1	0.82
ac	1	0.72
ad	1	0.35
ae	0	0.05
af	0	0.03
ag	0	0.00
bc	1	0.72
bd	1	0.52
be	0	0.23
bf	0	0.20
bg	0	0.18
cd	1	0.45
ce	0	0.14
cf	0	0.15
cg	0	0.09
de	0	0.68
df	0	0.68
dg	0	0.63
ef	1	0.91
eg	1	0.95
fg	1	0.90

(c) Arrays of pairwise clustering and similarity for all object pairs

Figure 2: Illustrative example of the Area Under the Curve for Clustering (AUCC) procedure: (a) toy dataset with an arbitrary clustering solution, in which clusters are indicated by a combination of colors and shapes (red diamonds / black circles); (b) similarity matrix between the data objects of the dataset; (c) objects are considered in a pairwise fashion and each pair is associated with the corresponding similarity value and cluster assignment (1 if the pair belongs to the same cluster, 0 otherwise). These pairwise representations can be provided as input to a standard ROC Analysis procedure, resulting in an AUC of 0.9167. This is the AUCC assessment of candidate solution (a).

One ought to note that, in the context of supervised classification, a solution (prediction) is given as real-valued classification scores, while the actual class labels (target result) are represented by binary class labels. In our setup, a clustering solution with any number of clusters is represented as a binary pairwise clustering array, whereas the referential target is represented by real-valued pairwise (dis)similarities intrinsic to the data. Moreover, in the case of clustering, we deal with pairs of objects, as opposed to single objects considered in the traditional classification scenario.

Although the whole validation procedure is described in terms of (dis)similarities, it is important to note that it is not tied to any particular measure. The only requirements are that: (i) the (dis)similarity employed in the validation procedure must be the very same (or equivalent) to the one employed during the clustering phase and; (ii) the measure must satisfy the symmetry, positivity and identity properties. Each measure captures a different aspect of the data and any specific choice will depend on the application scenario in hand (Jain and Dubes, 1988; Jaskowiak et al., 2012, 2014). Yet, regardless of the proximity measure in use, the AUCC validation index captures the same essence, that is, it favors partitions in which objects in the same cluster are more similar than objects from different clusters.

3.1. Equivalence Between AUCC and Baker & Hubert's Gamma

In this section we discuss the equivalence between the AUCC of a clustering result and its evaluation with the Gamma Index, which is a relative validity criterion introduced by Baker and Hubert (1975), based on the Goodman-Kruskal correlation coefficient (Goodman and Kruskal, 1954). We initially show that AUCC and Gamma are equivalent to a linear transformation of one another when there are no ties in proximity values (other than self-proximity values).² We then show that the original Gamma Index can be extended in an intuitive way to account for scenarios in which ties may exist, while preserving both the exact relation with AUCC as well as its expected value under a null hypothesis of random clustering solutions. The theoretical expected values for both Gamma and

²This result was originally and preliminarily described in (Jaskowiak, 2015). An equivalent result, involving the relation between AUC and the 1954 Goodman-Kruskal's rank correlation, was recently rediscovered by Higham and Higham (2019) in an unrelated context, involving measures of resolution in meta-cognitive studies.

AUCC are also derived in this section as part of our contributions.

Before we proceed, let us recall the definition of the Gamma Index, which can be written as:

$$\gamma = \frac{s_+ - s_-}{s_+ + s_-}, \quad (1)$$

or, equivalently, $1 - \frac{2s_-}{s_{total}}$, with $s_{total} = s_+ - s_-$ and:

$$s_+ = \frac{1}{2} \sum_{l=1}^k \sum_{\substack{\mathbf{x}_i, \mathbf{x}_j \in C_l \\ \mathbf{x}_i \neq \mathbf{x}_j}} \frac{1}{2} \sum_{m=1}^k \sum_{\substack{\mathbf{x}_r \in C_m \\ \mathbf{x}_s \notin C_m}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_r - \mathbf{x}_s\|), \quad (2)$$

$$s_- = \frac{1}{2} \sum_{l=1}^k \sum_{\substack{\mathbf{x}_i, \mathbf{x}_j \in C_l \\ \mathbf{x}_i \neq \mathbf{x}_j}} \frac{1}{2} \sum_{m=1}^k \sum_{\substack{\mathbf{x}_r \in C_m \\ \mathbf{x}_s \notin C_m}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| > \|\mathbf{x}_r - \mathbf{x}_s\|), \quad (3)$$

where $\delta(\cdot)$ is equal to 1 if the inequality is satisfied, 0 otherwise. In the equation above s_+ (s_-) is the count of occurrences of *object pairs* from the same cluster that have a smaller (greater) dissimilarity $\|\cdot\|$ than that of *object pairs* that belong to different clusters. Intuitively, s_+ is expected to account for well placed pairs of objects, whereas s_- should account for misplaced pairs of objects.

It is important to note that the formulation of the Gamma Index as presented above is computationally very expensive, turning out to be prohibitive in most practical applications of cluster analysis. Specifically, it has complexity $O(n^4/k)$, where n is the number of data objects and k is the number of clusters in the candidate solution under evaluation (Vendramin et al., 2010)³. Theorem 1 describes the relation between the outcomes of the evaluation of such a candidate clustering solution by both the Gamma Index as well as the AUCC, whereby the Gamma Index can be computed with a significantly lower computational cost (Jaskowiak, 2015):

Theorem 1. *Assume that there are no ties in proximity values (except for self-proximity values),*

³Assuming that (a) all dissimilarities $\|\cdot\|$ are given in advance (otherwise an additional dissimilarity cost would be required — $O(n^2d)$ in case of Euclidean distance, where d is the dimension of the data space), and (b) cluster sizes are balanced (all proportional to n/k , possibly differing by a constant factor) (Vendramin et al., 2010).

i.e., there aren't two pairs of distinct data objects whose (dis)similarity values are exactly the same. Then, the Area Under the ROC Curve for Clustering (AUCC) obtained from the evaluation of a clustering result is equal to $(1+\gamma)/2$, where γ is the value from the evaluation of the same clustering result with the Gamma criterion from Baker and Hubert (1975), given by Equation (1).

Proof. First, let us consider a binary supervised classification problem and a scoring classifier, that is, a classifier that outputs a real-valued score for any data object given as input. Although classification scores may not be interpreted as strict probabilities, the higher their value the higher is the expectancy that the corresponding object should belong to the reference/target class. Scores can thus be associated with a threshold in order to deem objects as negative or positive depending on whether they are below the threshold or not, respectively. Scores can also be used to derive a ranking of the objects: starting from the highest score value, one can rank the objects from 1 (associated with the highest possible rank/score) to a maximum integer associated with the lowest possible rank/score (equal to the number of objects, if there are no ties). Now let us consider the random selection of one positive and one negative object (with respect to their actual class labels). In this case, the AUC obtained with the evaluation of such a classifier has the interesting statistical interpretation of being equivalent to the probability that it will rank the randomly selected positive object *higher* than the randomly selected negative one (Fawcett, 2006).

In the context of clustering validation, each “object” of the evaluation corresponds, in fact, to a pair of data objects from the clustering result. The positive and negative classes indicate whether (1) or not (0) a pair of objects belongs to the same cluster, respectively. Finally, scores readily translate into similarity values between pairs of objects.

Recall from the definition of Gamma that terms s_+ and s_- are equal to the *number* of occurrences of positive (1) pairs from the cluster solution having a higher (s_+) or a lower (s_-) similarity value than negative (0) pairs, respectively. Provided that there are no ties in similarity values, candidate occurrences will be counted either to s_+ or to s_- (there is no alternative outcome), so the total number of possible counts is $s_{total} = s_+ + s_-$, which depends exclusively on the dataset size (n) as well as on the number of clusters (k) and their (im)balance (*i.e.*, relative sizes) in the partition

under evaluation. Notice that we can define the empirical probabilities with relative frequencies of s_+ and s_- by dividing these values by s_{total} . Let us denote such empirical probabilities as $P(s_+)$ and $P(s_-)$. Given that such values are all obtained by dividing s_+ and s_- by a constant value (s_{total}), we can rewrite Gamma as:

$$\gamma = \frac{P(s_+) - P(s_-)}{P(s_+) + P(s_-)}.$$

Since $P(s_+) + P(s_-) = 1$, we have:

$$\begin{aligned} \gamma &= P(s_+) - P(s_-), \\ &= P(s_+) - (1 - P(s_+)), \\ &= P(s_+) - 1 + P(s_+), \\ &= 2P(s_+) - 1. \end{aligned}$$

Notice that $P(s_+)$ is the empirical probability of ranking a positive example (i.e., a pair of data objects belonging to the same cluster) higher than a negative one, which is exactly the same estimate as the Area Under the ROC Curve value (Fawcett, 2006). Hence, $AUCC = P(s_+) = (\gamma + 1)/2$. \square

As a side note, given that $AUC = (Gini + 1)/2$ (Hand and Till, 2001; Fawcett, 2006), then the value obtained with the application of Gamma is the very same one obtained with the application of the Gini Coefficient (Gini, 1912; Ceriani and Verme, 2012). To the best of our knowledge, the relation between Baker-Hubert's Gamma and Gini had not been established elsewhere before. Similarly, notice that there is a known relation between AUC and the classic Wilcoxon-Mann-Whitney *U*-statistic: the U-statistic can be defined as a count of the number of times that observations from one sample (say, class positive) are ranked higher than observations from the other sample (say, class negative) according to their scores; by averaging this quantity over all possible pairwise comparisons, the result can be shown to be exactly equivalent to the AUC (Hanley and McNeil, 1982; Mason

and Graham, 2002) and, by transitivity using Theorem 1, also equivalent to Gamma.

Computational Complexity: As previously mentioned, the original formulation of Gamma (Baker and Hubert, 1975) is computationally prohibitive for most real-world applications, due to its $O(n^4/k)$ complexity (Vendramin et al., 2010). For instance, this has prevented its evaluation in datasets as small as 500 objects in the experimental study performed by Vendramin et al. (2010). As pointed out by Fawcett (2006), computing the AUC for a binary classification problem with n objects has $O(n \log n)$ complexity. Note that in the case of clustering evaluation we are dealing with pairs of objects, thus we have an $O(n^2 \log n)$ time complexity⁴ for the Area Under the Curve for Clustering (AUCC), a considerable reduction when compared to the original Gamma.

3.2. Expected Value Property

In this section we show that, given a dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of finite size n and a dissimilarity measure $\|\mathbf{x}_i - \mathbf{x}_j\|$ associated with each pair of data objects \mathbf{x}_i and \mathbf{x}_j , the expected value of the Gamma Index under a null distribution of random clustering solutions is zero and, accordingly, the expected value of AUCC is 0.5. This results holds true for any given value of k , i.e., it is valid irrespective of the number of clusters assumed in the null model. It also holds true independently of the dataset size n and, as we will show, irrespective of relative cluster sizes, i.e., cluster (im)balance. This is a desirable property for two reasons: (a) it allows for a better interpretation of AUCC values, that is, how far/close a given candidate clustering solution is from random; and, more importantly, (b) it ensures that the use of AUCC as an internal/relative validity criterion is not biased by the number or (im)balance of clusters in the partitions being compared.

Before we formalize these results, it is paramount to stress that the null model is defined at the individual data object level, as a random assignment of objects to clusters; it is *not* directly defined in terms of the binary relation \mathcal{C}^p on which AUCC relies because, by randomly assigning binary labels to pairs of objects, we would actually account for in our calculations encodings that do not correspond to any feasible partition of the data. For instance, consider a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, case

⁴Apart from the cost to obtain the dissimilarity matrix, \mathbf{D} , which is also required by Gamma.

in which there are three possible pairs, $(\mathbf{x}_1, \mathbf{x}_2)$, $(\mathbf{x}_1, \mathbf{x}_3)$, $(\mathbf{x}_2, \mathbf{x}_3)$. A hypothetical random encoding $[1 \ 0 \ 1]$ says that \mathbf{x}_1 and \mathbf{x}_2 are in the same cluster, and so do \mathbf{x}_2 and \mathbf{x}_3 , whereas \mathbf{x}_1 and \mathbf{x}_3 are in different clusters, which is impossible. This prevents us from directly evoking the well-known property that the expected AUC for chance is 0.5, because this result assumes that each and every element has, independently of other elements, the same fixed probabilities of being assigned to each class. The example above shows that the elements of the binary relations \mathcal{C}^p do not satisfy this assumption as they are not independent. Note that independence is also a critical assumption behind the use of the *U-statistic*, which is equivalent to the AUC (Mason and Graham, 2002). Our next result circumvents this hurdle by working at the object (rather than pairwise object) level:

Theorem 2. *Assuming a null model in which every clustering solution with k clusters (as a valid partition of n objects) is equally likely, the expected value of the Gamma Index is zero ($\gamma = 0$).*

Corollary 1. *Assuming a null model in which every clustering solution with k clusters (as a valid partition of n objects) is equally likely, the expected value of AUCC is 0.5.*

Proof. For the sake of simplicity and without loss of generality, we initially assume here that there are no ties in the dissimilarities between pairs of objects (the more general case involving ties will be discussed in Section 3.3). Since there are no ties, the quantity $s_{total} = s_+ + s_-$ depends exclusively on the dataset size as well as on the number and the (im)balance of clusters in the partition under evaluation. For a given dataset of size n and a fixed number of clusters k of interest, s_{total} depends solely on the relative cluster sizes. Once again, for the sake of simplicity and without loss of generality, we will *initially* assume that the relative cluster sizes are also fixed in the null model. This may not be unreasonable in practice since one may want to compare a given candidate clustering solution against a null model of random solutions of exactly the same nature. In spite of that, we will subsequently show that the expected value actually doesn't change if we generalize/extend the null model such that the expectation is computed across random partitions with all possible cluster size proportions. In addition, since the result holds irrespective of the number of clusters, then it can be trivially shown to hold for an even more general null model in which the expected value for chance is computed over all possible partitions of the data, regardless

of k and/or (im)balance.

For given n , k , and cluster size proportions, s_{total} is a constant and, hence, the expected value of γ can be derived from Equation (1) as:

$$E_{\mathbb{C}_k}\{\gamma\} = \frac{E_{\mathbb{C}_k}\{s_+\} - E_{\mathbb{C}_k}\{s_-\}}{s_{total}}, \quad (4)$$

where the expectation is taken evenly across the set \mathbb{C}_k of all possible valid partitions $\mathcal{C}_k = \{C_1, \dots, C_k\}$ consisting of k clusters of fixed relative sizes $|C_1|/n, |C_2|/n, \dots, |C_k|/n$ ($|\cdot|$ stands for set cardinality) or any permutation of these. In order to compute $E_{\mathbb{C}_k}\{s_+\}$ (and, subsequently, $E_{\mathbb{C}_k}\{s_-\}$ in an analogous fashion) it is worth noticing that term s_+ can be written in a completely equivalent form as:

$$\begin{aligned} s_+ = & \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq j \neq i}} \left[\delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_i - \mathbf{x}_s\|) \cdot \mu_{\mathcal{C}_k}(i, j, s) \right] + \\ & + \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_r \in \mathbf{X} \\ r \neq j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq r \neq j \neq i}} \left[\delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_r - \mathbf{x}_s\|) \cdot \phi_{\mathcal{C}_k}(i, j, r, s) \right], \end{aligned} \quad (5)$$

where $\mu_{\mathcal{C}_k}(i, j, s)$ is an indicator function that takes as argument the indexes $i \neq j \neq s$ of three different objects of the dataset (a triple with no duplicates) and returns 1 if and only if the first two objects belong to the same cluster ($\mathbf{x}_i, \mathbf{x}_j \in C_l$) whereas the third object belongs to a different cluster ($\mathbf{x}_s \in C_m, m \neq l$) in partition \mathcal{C}_k ($\{C_l, C_m\} \subset \mathcal{C}_k$); otherwise, $\mu_{\mathcal{C}_k}(i, j, s)$ is equal to zero. Similarly, $\phi_{\mathcal{C}_k}(i, j, r, s)$ is an indicator function that takes as argument the indexes $i \neq j \neq r \neq s$ of four different objects of the dataset (a quadruple with no duplicates) and returns 1 if and only if the first two objects belong to the same cluster ($\mathbf{x}_i, \mathbf{x}_j \in C_l$) whereas the other two objects belong to separate clusters ($\mathbf{x}_r \in C_m, \mathbf{x}_s \notin C_m$);⁵ otherwise, $\phi_{\mathcal{C}_k}(i, j, r, s)$ is equal to zero.

The main advantage of the above representation is that, unlike the previous equivalent definition of

⁵Note that C_m is not necessarily different from C_l , they may or may not be the same cluster in partition \mathcal{C}_k .

s_+ in Equation (2), the summation indexes in Equation (5) do *not* depend on the clustering solution \mathcal{C}_k . Obviously, the summations are now covering an augmented set of terms, namely, terms involving the comparison of pairwise distances from all triples or quadruples of distinct objects in the dataset. The additional/augmented terms (and only those) are, however, cancelled out by a null value of the respective indicator function, namely, $\mu_{\mathcal{C}_k}(\cdot)$ for triples and $\phi_{\mathcal{C}_k}(\cdot)$ for quadruples.

It is worth noticing that functions $\mu_{\mathcal{C}_k}(\cdot)$ and $\phi_{\mathcal{C}_k}(\cdot)$ depend only on the partition \mathcal{C}_k under evaluation, they do not depend on the pairwise dissimilarities between data objects. Conversely, function $\delta(\cdot)$ depends *only* on the pairwise dissimilarities, it does *not* depend on any partition of the data. From this observation, we can write the expectation $E_{\mathcal{C}_k}\{s_+\}$ from Equation (5) as:

$$\begin{aligned}
E_{\mathcal{C}_k}\{s_+\} &= \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq j \neq i}} \left[\delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_i - \mathbf{x}_s\|) \cdot E_{\mathcal{C}_k}\{\mu_{\mathcal{C}_k}(i, j, s)\} \right] + \\
&+ \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_r \in \mathbf{X} \\ r \neq j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq r \neq j \neq i}} \left[\delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_r - \mathbf{x}_s\|) \cdot E_{\mathcal{C}_k}\{\phi_{\mathcal{C}_k}(i, j, r, s)\} \right].
\end{aligned} \tag{6}$$

Notice that terms $E_{\mathcal{C}_k}\{\mu_{\mathcal{C}_k}(i, j, s)\}$ and $E_{\mathcal{C}_k}\{\phi_{\mathcal{C}_k}(i, j, r, s)\}$ can be readily interpreted as the fraction of all partitions $\mathcal{C}_k \in \mathbb{C}_k$ (i.e., the fraction of the population of valid partitions comprised by the null model, \mathbb{C}_k) such that the corresponding indicator functions return a non-zero (unit) value. Since the indicator functions $\mu_{\mathcal{C}_k}(i, j, s)$ and $\phi_{\mathcal{C}_k}(i, j, r, s)$ do *not* depend on any intrinsic property of data objects, they depend instead only on the *cluster labels* imposed to those specific objects indexed by the functions' arguments, the *expected values* $E_{\mathcal{C}_k}\{\mu_{\mathcal{C}_k}(i, j, s)\}$ and $E_{\mathcal{C}_k}\{\phi_{\mathcal{C}_k}(i, j, r, s)\}$ will be the same irrespective of the indexes i, j, r, s . In other words, if we fix any three (four) distinct objects and average $\mu_{\mathcal{C}_k}(i, j, s)$ ($\phi_{\mathcal{C}_k}(i, j, r, s)$) over all partitions in \mathbb{C}_k , across which only the cluster labels of objects are permuted, then the result will be the same (a constant). We will call these constants

$E_{\mathbb{C}_k}\{\mu_{\mathbb{C}_k}\}$ and $E_{\mathbb{C}_k}\{\phi_{\mathbb{C}_k}\}$ for short, whereby we can rewrite Equation (6) as:

$$\begin{aligned}
E_{\mathbb{C}_k}\{s_+\} &= E_{\mathbb{C}_k}\{\mu_{\mathbb{C}_k}\} \cdot \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq j \neq i}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_i - \mathbf{x}_s\|) + \\
&E_{\mathbb{C}_k}\{\phi_{\mathbb{C}_k}\} \cdot \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_r \in \mathbf{X} \\ r \neq j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq r \neq j \neq i}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_r - \mathbf{x}_s\|).
\end{aligned} \tag{7}$$

Following an analogous reasoning, we can also write $E_{\mathbb{C}_k}\{s_-\}$ as:

$$\begin{aligned}
E_{\mathbb{C}_k}\{s_-\} &= E_{\mathbb{C}_k}\{\mu_{\mathbb{C}_k}\} \cdot \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq j \neq i}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| > \|\mathbf{x}_i - \mathbf{x}_s\|) + \\
&E_{\mathbb{C}_k}\{\phi_{\mathbb{C}_k}\} \cdot \sum_{\mathbf{x}_i \in \mathbf{X}} \sum_{\substack{\mathbf{x}_j \in \mathbf{X} \\ j \neq i}} \sum_{\substack{\mathbf{x}_r \in \mathbf{X} \\ r \neq j \neq i}} \sum_{\substack{\mathbf{x}_s \in \mathbf{X} \\ s \neq r \neq j \neq i}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| > \|\mathbf{x}_r - \mathbf{x}_s\|).
\end{aligned} \tag{8}$$

Now, notice that, for every triple (i, j, s) (i.e., $\forall i \neq j \neq s$) such that $\delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_i - \mathbf{x}_s\|) = 1$ in Equation (7), there is a triple (i, s, j) for which $\delta(\|\mathbf{x}_i - \mathbf{x}_s\| > \|\mathbf{x}_i - \mathbf{x}_j\|) = 1$ in Equation (8), and vice versa. Similarly, for every quadruple (i, j, r, s) (i.e., $\forall i \neq j \neq r \neq s$) such that $\delta(\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x}_r - \mathbf{x}_s\|) = 1$ in Equation (7), there is a quadruple (r, s, i, j) for which $\delta(\|\mathbf{x}_r - \mathbf{x}_s\| > \|\mathbf{x}_i - \mathbf{x}_j\|) = 1$ in Equation (8), and vice versa.

Therefore, $E_{\mathbb{C}_k}\{s_+\} = E_{\mathbb{C}_k}\{s_-\}$ and, from Equation (4), it follows that $E_{\mathbb{C}_k}\{\gamma\} = 0$, i.e., the expected value of the Gamma Index under the assumed null model of random clustering solutions is zero. Finally, from this result and using Theorem 1, $E_{\mathbb{C}_k}\{AUCC\} = 0.5$ follows straightforwardly.

Extended Null Model: The above results prove both the theorem and the corollary. However, the proof assumes that the relative sizes in the k clusters contained in any random solution of the null model are fixed. If this condition is not satisfied, $s_{total} = s_+ + s_-$ will no longer be a constant (it will vary across different partitions in the null model), case in which Equation (4) does not hold true, at least not simultaneously across the entire population of random partitions \mathbb{C}_k .

We can extend the above results to cases in which a more general null model is considered, where \mathbb{C}_k contains partitions with any and all possible cluster size proportions, rather than a prefixed

one. This can be achieved by noticing that \mathbb{C}_k is a finite set, and all elements in this set (random clustering solutions) are equally likely, case in which the mathematical expectation of a function (γ) of the elements in this set can be written as the ordinary arithmetic mean of the function evaluations for each element in the set: $E_{\mathbb{C}_k}\{\gamma\} = \frac{1}{|\mathbb{C}_k|} \sum_{\mathcal{C}_k \in \mathbb{C}_k} \gamma(\mathcal{C}_k)$. If we arbitrarily group the set \mathbb{C}_k into any chosen collection of disjoint subsets (of random clustering solutions) $\mathbb{S}_i \subset \mathbb{C}_k$, such that $\bigcup_i \mathbb{S}_i = \mathbb{C}_k$, it is trivial to show that the expectation across the entire set can be written as an average of the expectations within each subset ($E_{\mathbb{S}_i}(\gamma)$) weighted by their cardinalities, i.e.:

$$\begin{aligned} E_{\mathbb{C}_k}\{\gamma\} &= \frac{\sum_{\mathbb{S}_i \in \mathbb{C}_k} |\mathbb{S}_i| \cdot E_{\mathbb{S}_i}(\gamma)}{\sum_{\mathbb{S}_i \in \mathbb{C}_k} |\mathbb{S}_i|} \\ &= \frac{\sum_{\mathbb{S}_i \in \mathbb{C}_k} |\mathbb{S}_i| \cdot E_{\mathbb{S}_i}(\gamma)}{|\mathbb{C}_k|}. \end{aligned} \tag{9}$$

Since this result is valid for any arbitrary subdivision of \mathbb{C}_k as described above, for mathematical convenience we chose a subdivision such that every subgroup \mathbb{S}_i contains all and only the clustering solutions in \mathbb{C}_k that share the same cluster size proportions. In other words, each \mathbb{S}_i is associated with a unique (im)balance of clusters, thence s_{total} is constant for all partitions within \mathbb{S}_i and all the results above in this proof are also valid for \mathbb{S}_i . Specifically, $E_{\mathbb{S}_i}\{\gamma\} = 0, \forall i$. Therefore, it follows from Equation (9) that $E_{\mathbb{C}_k}\{\gamma\} = 0$ and, by evoking Theorem 1 we have $E_{\mathbb{C}_k}\{AUCC\} = 0.5$. \square

3.3. Ties in (Dis)similarity Values

In principle, the relation between the Gamma Index and AUCC established in Theorem 1 assumes that there are no ties in the real-valued thresholds used to compute the area under the ROC curve. To better understand what happens when ties are present, let's consider the toy example in Table 1. In a classification assessment scenario, the six instances would be data objects with binary class labels and a classification score associated with the positive class, whereas in a clustering assessment scenario they would correspond to pairs of objects with binary clustering assignment relations ("labels") and a pairwise similarity value ("score").

Table 1: Illustrative example of a classification or clustering problem involving ties.

Instance	Label	Score
1	1	0.75
2	0	0.50
3	1	0.50
4	1	0.50
5	0	0.25
6	0	0.20

Notice that instances 2, 3 and 4 share exactly the same score of 0.5, i.e., they are tied. As the real-valued threshold used to compute the AUC moves from 0.25 to 0.75 (or vice-versa), it is not clear how exactly the ROC curve should move from point $(FPR, TPR) = (\frac{1}{3}, 1)$ to point $(FPR, TPR) = (0, \frac{1}{3})$ in the ROC graph, respectively. This is because the ROC curve depends on the relative rank/order of the instances according to their scores, however, the relative order among instances sharing the same score cannot be uniquely determined. Each rank permutation of those instances would incur a different area under the curve. Notice that the uncertainty around the final, total area comes exclusively from the subarea of the unit square comprising the rectangle with diagonal/opposite vertices $(\frac{1}{3}, 1)$ and $(0, \frac{1}{3})$. The area of this rectangle is $\frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$.

The most accepted and widely adopted approach to compute the ROC curve in case of ties is so-called “walk along the diagonal”, which in our pedagogic example in Table 1 corresponds to connecting the points $(\frac{1}{3}, 1)$ and $(0, \frac{1}{3})$ with a straight line (Fawcett, 2006). In terms of the area under the curve, this corresponds to assigning to the final area precisely half of the subarea subject to uncertainty due to ties (i.e., $\frac{1}{2} \times \frac{2}{9} = \frac{1}{9}$); in the above example, the final area will thus amount to a total of 0.8888. From the probabilistic interpretation of ROC curves previously discussed in Sections 1 and 3.1, this approach corresponds to assigning half of the fraction of probability involved in ties (and whose assignment is unclear) to the computed area under the curve as an estimate of the probability that positive (1) instances will be ranked higher than negative (0) instances. Accordingly, the other half will be assigned to this probability’s complement, i.e., the estimated probability that negative instances will be ranked higher.

In the clustering assessment scenario, the “walk along the diagonal” approach described above

corresponds to assigning half of the fraction of probability involved in ties to the computed area under the curve as an estimate of the probability that pairs of objects belonging to the same cluster (1) will be ranked higher than pairs belonging to different clusters (0). Let's call the fraction of probability (i.e. the subarea) involved in ties as P_t , such that $P_t = \frac{2}{9}$ in our example. When computing the Gamma Index, this is the probability associated with the outcomes that are *not* accounted for by terms s_+ and s_- , namely, those outcomes involving similarity ties. In the presence of ties, the quantity $s_{total} = s_+ + s_-$ is *no longer* equal to a constant that represents the total number of possible counts and depends exclusively on n , k and relative cluster sizes. Rather, such a constant (renamed s'_{total} hereafter) is now equal to $s'_{total} = s_+ + s_- + s_0$, where s_0 is given by:

$$s_0 = \frac{1}{2} \sum_{l=1}^k \sum_{\substack{\mathbf{x}_i, \mathbf{x}_j \in C_l \\ \mathbf{x}_i \neq \mathbf{x}_j}} \frac{1}{2} \sum_{m=1}^k \sum_{\substack{\mathbf{x}_r \in C_m \\ \mathbf{x}_s \notin C_m}} \delta(\|\mathbf{x}_i - \mathbf{x}_j\| = \|\mathbf{x}_r - \mathbf{x}_s\|),$$

such that $P_t = s_0/s'_{total}$. When there are ties, $s_0 \neq 0$, $P_t \neq 0$ and, because $P(s_+) + P(s_-) + P_t = 1$, it follows that $P(s_+) + P(s_-) \neq 1$. In this case, the relation between the Gamma Index and AUCC established in Theorem 1 is no longer valid. As a simple proof by counter-example, in the dataset of Table 1 it is trivial to show that $s_+ = 7$ and $s_- = 0$, so if Equation (1) were to be used, the result would be $\gamma = \frac{7-0}{7+0} = 1$; evoking Theorem 1 one would in that case get $AUCC = (1 + \gamma)/2 = 1$, which is in contradiction with the value obtained by computing the area under the ROC curve walking along the diagonal to resolve ties ($AUCC = 0.8888$). The contradiction is caused by the presence of ties, which violates Theorem 1's assumption.

Under the "walk along the diagonal" assumption for dealing with ties in the AUCC computation, the relation in Theorem 1 can be reestablished by also distributing half of the probability P_t to $P(s_+)$ and the other half to $P(s_-)$ when computing Gamma. In other words, we must assign an additional $s_0/2$ amount to term s_+ and the same $s_0/2$ additional amount to term s_- when

computing γ in Equation (1), which is thereby generalized as:

$$\gamma = \frac{\left(s_+ + \frac{s_0}{2}\right) - \left(s_- + \frac{s_0}{2}\right)}{\left(s_+ + \frac{s_0}{2}\right) + \left(s_- + \frac{s_0}{2}\right)} = \frac{s_+ - s_-}{s_+ + s_- + s_0}, \quad (10)$$

and clearly reduces back to Equation (1) in the absence of ties. By using Equation (10) instead of Equation (1) in the dataset of Table 1, one has $s_0 = 2$, $\gamma = \frac{7-0}{7+0+2} = \frac{7}{9}$ and, in this case, $\text{AUCC} = (1 + \gamma)/2 = 0.8888$ follows from Theorem 1 as expected.

Equation (10) allows Theorem 1 to be stated more broadly without any particular assumption involving ties. It is worth noticing that, by distributing the probability/area associated with ties evenly between s_+ and s_- as in Equation (10) — or equivalently, “walking along the diagonal” when computing AUCC — we do not change the fact that $E_{\mathbb{C}_k}\{s_+\} = E_{\mathbb{C}_k}\{s_-\}$ and, as a consequence, the expected value properties in Theorem 2 and Corollary 1 remain valid for γ and AUCC, respectively.

Finally, it is also worth noticing that at least a couple of (not-so-common) alternative approaches to resolve ties in ROC analysis exist (Fawcett, 2006) that could also be adopted to compute AUCC. In particular, the *optimistic* approach, which fully assigns the total amount of probability/subarea associated with ties to the final area under the curve — would be equivalent to allocating the corresponding additional amount s_0 entirely to s_+ (none to s_-) when computing the Gamma Index. In contrast, the *pessimistic* approach — where none of the probability/subarea associated with ties is assigned to the area under the curve — would be equivalent to allocating the additional amount s_0 entirely to s_- when computing Gamma. While these alternative approaches for AUCC computation and the corresponding aforementioned modifications to the Gamma Index would keep the relation in Theorem 1 valid in spite of the presence/absence of ties, these approaches would clearly bias the expected value of either s_+ or s_- such that $E_{\mathbb{C}_k}\{s_+\} \neq E_{\mathbb{C}_k}\{s_-\}$. Accordingly, Theorem 2 and Corollary 1 would no longer be valid in the presence of ties.

4. Experimental Evaluation

4.1. Agreement with External Evaluation

In order to experimentally assess the use of AUCC in the relative clustering validation scenario, we have employed the same evaluation methodology proposed by Vendramin et al. (2009, 2010). In short, it assumes that the best relative validity criteria should have the highest correlation with an external validity index. External indices are based on comparisons between the clustering results and a ground-truth, which are available for simulated and benchmark datasets. Evaluations of relative criteria based on their correlations with an external index have also been carried out e.g. by (Jaskowiak et al., 2016; Nguyen et al., 2020). The procedure can be summarized as follows:

1. Given a dataset, generate partitions/solutions with different numbers of clusters (k), usually with $2 \leq k \leq \sqrt{n}$ (configuration we adopt here), employing one or more clustering algorithms;
2. Determine the quality of the partitions w.r.t. to the relative validity criteria under scrutiny;
3. Determine the quality of each partition according to one (or more) external validity criterion;
4. Measure the correlation between the unsupervised and supervised evaluations provided by each of the relative and the external validity criteria, respectively.

To generate a diverse collection of clustering partitions (Step 1 of the evaluation methodology), we employ the well-known k-means clustering algorithm (MacQueen, 1967) and four variants of Hierarchical Clustering Algorithms (HCAs) (Jain and Dubes, 1988), namely, Single-Linkage, Average-Linkage, Complete-Linkage, and Ward’s. For each dataset we generate partitions in the range $k \in \{2, \dots, k_{max}\}$, with $k_{max} = \lceil \sqrt{n} \rceil$. In the case of k-means, for each k , 100 initializations are undertaken and the partition with best MSE (Mean Squared Error) is then selected for further evaluation. External agreement of partitions with respect to the true labels are obtained with the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985; Amigó et al., 2009). Correlations between relative and external evaluations are given by the Pearson correlation coefficient (Pearson, 1895).

To place the results of AUCC into perspective we consider a collection of 28 relative validity criteria commonly employed in the literature as baselines (Nguyen et al., 2020; Zhou et al., 2021). These are: Calinski-Harabasz (VRC) (Calinski and Harabasz, 1974), Davies–Bouldin (DB) (Davies and Bouldin, 1979), Dunn’s Index (Dunn, 1974), 17 variants of Dunn’s Index (Bezdek and Pal, 1998), PBM (Pakhira et al., 2004), C-Index (Hubert and Levin, 1976), Point-Biserial (Milligan, 1981), C/Sqrt(k) (Ratkowsky and Lance, 1978; Hill, 1980), Silhouette Width Criterion (SWC) (Rousseeuw, 1987), Simplified Silhouette Width Criterion (SSWC) (Hruschka et al., 2006), Alternative Silhouette Width Criterion (ASWC) (Hruschka et al., 2004), and Alternative Simplified Silhouette Width Criterion (ASSWC) (Vendramin et al., 2009).

We evaluate AUCC alongside the aforementioned baseline validity criteria in 10 real datasets with varied characteristics in terms of their numbers of objects, dimensions and clusters in the reference ground-truth partition. These are: (i) the Yeast Galactose (Yeast) and Cell Cycle from Yeung et al. (Yeung et al., 2001) as well as eight datasets from UCI, as summarized in Table 2.

Table 2: Real datasets employed in experiments.

#	Dataset	# Objects	# Dimensions	# Clusters
1	Balance Scale	625	4	3
2	Cell Cycle	237	17	4
3	Control Chart (KDD)	600	60	6
4	E. Coli	336	7	8
5	Iris	150	4	3
6	Karhunen	2000	64	10
7	Sonar	208	60	2
8	Vehicle	846	18	4
9	Wisconsin Breast Cancer	683	9	2
10	Yeast	205	20	4

Evaluation results are depicted in Table 3. Validity criteria are ranked by their average correlation values. Although we evaluated a total of 18 Dunn formulations, for the sake of simplicity, we show the results for the best performer. AUCC ranked second best overall, below Point-Biserial (PB) only. Aggregated results across multiple datasets should be taken with a grain of salt though. Different criteria address the multi-faceted problem of clustering evaluation from different angles and may emphasize more or less certain particular aspects. It is well-known that no single criterion should be expected to outperform all the others in all problems. Instead, different criteria are expected to perform better/worse than others in different problems or scenarios. In fact, notice in

Table 3 that, while a subset of criteria have been outperformed by others within the collection of ten datasets involved in our experiments, five different criteria, namely PB, AUCC, C-Index, C/Sqrt(k) and VRC, exhibited the best/top performance in at least one dataset. For this reason, it is widely accepted that an analyst should not rely on a single criterion for unsupervised clustering evaluation (Bezdek and Pal, 1998; Jaskowiak et al., 2016); naive attempts to elect a single criterion as the best one overall are inevitably fruitless, unless they focus on specific classes of problems/scenarios.

Table 3: Evaluation results of AUCC and baseline relative criteria on 10 real datasets. Each cell displays the Pearson correlation between the relative evaluation and the external evaluation obtained with ARI. Top performance for each dataset (in columns “1” to “10”) is highlighted in bold.

Dataset #	1	2	3	4	5	6	7	8	9	10	Best	Avg.	Worst
PB	0.79	0.91	0.61	0.97	0.69	0.89	0.31	0.40	0.98	0.57	0.98	0.71	0.31
AUCC	0.48	0.60	0.75	0.76	0.13	0.84	0.70	0.78	0.91	0.77	0.91	0.67	0.13
C-Index	0.53	0.75	0.75	0.83	-0.07	0.88	0.64	0.78	0.81	0.57	0.88	0.65	-0.07
C/Sqrt(k)	0.88	0.82	0.10	0.81	0.59	0.76	0.32	0.71	0.73	0.58	0.88	0.63	0.10
SWC	0.76	0.84	0.06	0.65	0.34	0.80	0.38	0.82	0.88	0.73	0.88	0.62	0.06
ASWC	0.70	0.50	0.19	0.58	0.37	0.65	0.17	0.78	0.84	0.70	0.84	0.55	0.17
VRC	0.82	0.72	-0.02	0.62	0.19	0.48	0.13	0.85	0.58	0.72	0.85	0.51	-0.02
SSWC	0.76	-0.22	0.00	0.68	0.53	0.81	0.37	0.57	0.82	0.68	0.82	0.50	-0.22
Dunn 31	0.73	0.60	-0.18	0.65	0.15	0.59	0.36	0.68	0.79	0.46	0.79	0.48	-0.18
ASSWC	0.05	-0.27	0.02	0.64	0.60	0.71	0.12	0.37	0.82	0.69	0.82	0.37	-0.27
PBM	0.49	-0.09	-0.10	0.27	0.56	-0.29	-0.43	0.67	0.43	0.50	0.67	0.20	-0.43
DB	0.57	0.58	-0.54	0.26	-0.67	0.44	0.50	-0.03	0.53	0.04	0.58	0.17	-0.67
Best	0.88	0.91	0.75	0.97	0.69	0.89	0.70	0.85	0.98	0.77	0.98	0.71	0.31
Average	0.63	0.48	0.14	0.64	0.28	0.63	0.30	0.62	0.76	0.58	0.83	0.51	-0.09
Worst	0.05	-0.27	-0.54	0.26	-0.67	-0.29	-0.43	-0.03	0.43	0.04	0.58	0.17	-0.67

Rather than seeking a single, general purpose favorite criterion, a more realistic approach to practical clustering evaluation is to focus on strengths of different criteria to keep a collection of reliable candidates in one’s cluster analysis toolbox. From this standpoint, we argue that AUCC is a candidate to be included in this collection. In terms of reliability as assessed from the lens of robustness, it is noticeable from Table 3 that, in the majority of those cases in which AUCC does not provide the best evaluation, it still produces results close to the best criterion or, at least, far from the worst case. An important aspect that also relates to reliability (and, possibly to a significant strength or weakness) of criteria is their behavior when assessing random solutions without actual cluster structure. This aspect is discussed next.

4.2. Expected Value

AUCC (like its linearly related equivalent, Gamma Index) has the advantage that it allows for a better interpretation in terms of its expected value for chance clusterings, as shown in Corollary 1

following from Theorem 2. We are not aware of other criteria with a theoretical characterization of its value for chance. In order to experimentally assess how the different measures behave in this regard, we ran controlled experiments with 108 synthetic datasets from Vendramin et al. (2009, 2010), consisting of mixtures of multivariate Gaussians with varied characteristics. The 108 datasets have 500 objects each and are obtained from three design factors comprising number of dimensions (2, 3, 4, 22, 23, or 24), number of clusters for the reference partition (2, 4, 6, 12, 14, or 16) and cluster size distribution. Regarding distribution, there are three different settings: (i) balanced clusters; (ii) one cluster with 10% of the objects and the remaining objects evenly distributed among other clusters and; (iii) one cluster with 20% (if $k^* \in \{12, 14, 16\}$) or 60% (if $k^* \in \{2, 4, 6\}$) of the objects, and the remaining objects, again, evenly distributed among the other clusters. The term k^* accounts for the actual number of clusters in the *reference partition* of the dataset.

For each one of the 108 datasets we generated *random partitions* as *candidate clustering solutions* considering the number of clusters (k) in the range of 2 to $\lceil \sqrt{n} \rceil$, where n is the number of objects, therefore $\lceil \sqrt{500} \rceil = 23$. The random partitions were generated considering three balances for the random clusters: (i) balanced clusters; (ii) one cluster with 10% of the objects and the remaining objects evenly distributed among other clusters and; (iii) one cluster with 60% of the objects, and the remaining objects, again, evenly distributed among the other clusters. For each dataset, balance of the random partition, and number of clusters we generated a total of 100 random clustering solutions, which were then assessed by each of the relative validity criteria considered in this study. For the sake of compactness, among the reported top performing criteria from the previous evaluation involving real data, we only show results of a single version of the Silhouette criterion (namely, the original SWC) since other variants exhibited similar behaviour.

Figure 3 summarizes the results. Each row of the plot depicts the results of one criterion, whereas each column accounts for a different balance of the candidate random partitions assessed. In each plot an orange line represents the average of the 100 random partitions for a given dataset. Since we ran experiments on 108 datasets, there are 108 lines per plot, plus a red line that accounts for the mean across all experiments. It is worth mentioning that for each criterion (row) the y-axis is

at the very same range/scale of the criterion's value across the multiple columns.

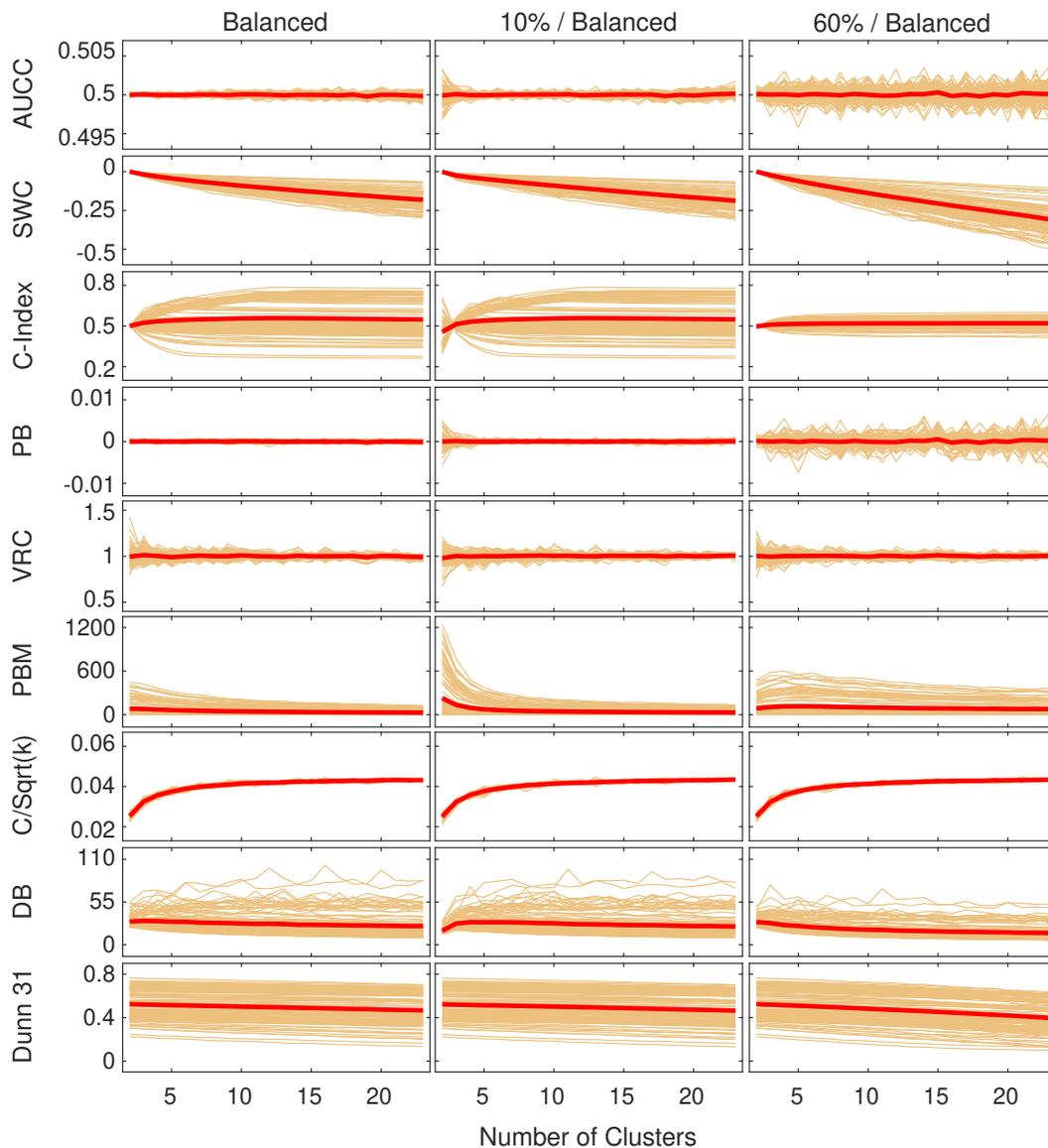


Figure 3: Evaluation of random partitions of 108 synthetic datasets with varied characteristics. Each orange line corresponds to the average of 100 randomly generated partitions when evaluated in a given dataset (total of 108 lines/datasets per plot). The number of clusters *in the randomly generated partitions* is depicted in the x-axis. Results are stratified on the basis of the cluster size distribution of *the generated partitions* (plot columns). The red line accounts for the overall mean.

It can be seen that, as expected, AUCC exhibits values around 0.5 with very small variability regardless of: (i) the number of dimensions in the dataset; (ii) the number of clusters (both in the dataset as well as in the randomly generated partitions); and (iii) the cluster size distribution (once again both in the dataset as well as in the randomly generated partitions). Besides AUCC, two other measures (PB and VRC) did not display noticeable changes in their empirical expected values for random solutions, although we are not aware of any formal proof to support this observation. Notably, other top performing measures such as Silhouette (SWC), C-Index, C/Sqrt(k) and PBM exhibited clear changes/patterns in empirical expected values as a function of the numbers of clusters and/or prominent variability across different datasets (orange lines). For SWC and C/Sqrt(k) the trend of the empirical expected value as a function of the number of clusters (decreasing for the former and increasing for the latter) seems consistent across the different experimental settings and datasets. This is not the case for PBM and C-Index. For C-Index in particular, the empirical expected value can noticeably increase or decrease as a function of the number of clusters, depending on each particular dataset (orange line). It also varies with different size distributions in the randomly generated partitions (different columns of Figure 3).

In a practical scenario, the lack of a known constant expected value for a relative measure under a null model of random clustering solutions can impair evaluation, most noticeably when one wants to compare solutions across different numbers of clusters, because the evaluation result can be biased by the number of clusters irrespective of the quality of the assessed solutions.

4.3. ROC Curves

An important aspect of ROC curves is their visual interpretation, as the curves display the trade-off between sensitivity (TPR) and specificity (1-FPR) for distinct solutions. We analyzed ROC curves for the well-known Ruspini dataset (4 clusters) as well as for a simulated dataset⁶ for candidate solutions with varying numbers of clusters (Figure 4).

⁶This dataset consists of 9 clusters, with 50 objects each, obtained from normal distributions with variance equal to 4.5, centered at (0, 0), (0, 20), (0, 40), (20, 0), (20, 20), (20, 40), (40, 0), (40, 20), and (40, 40).

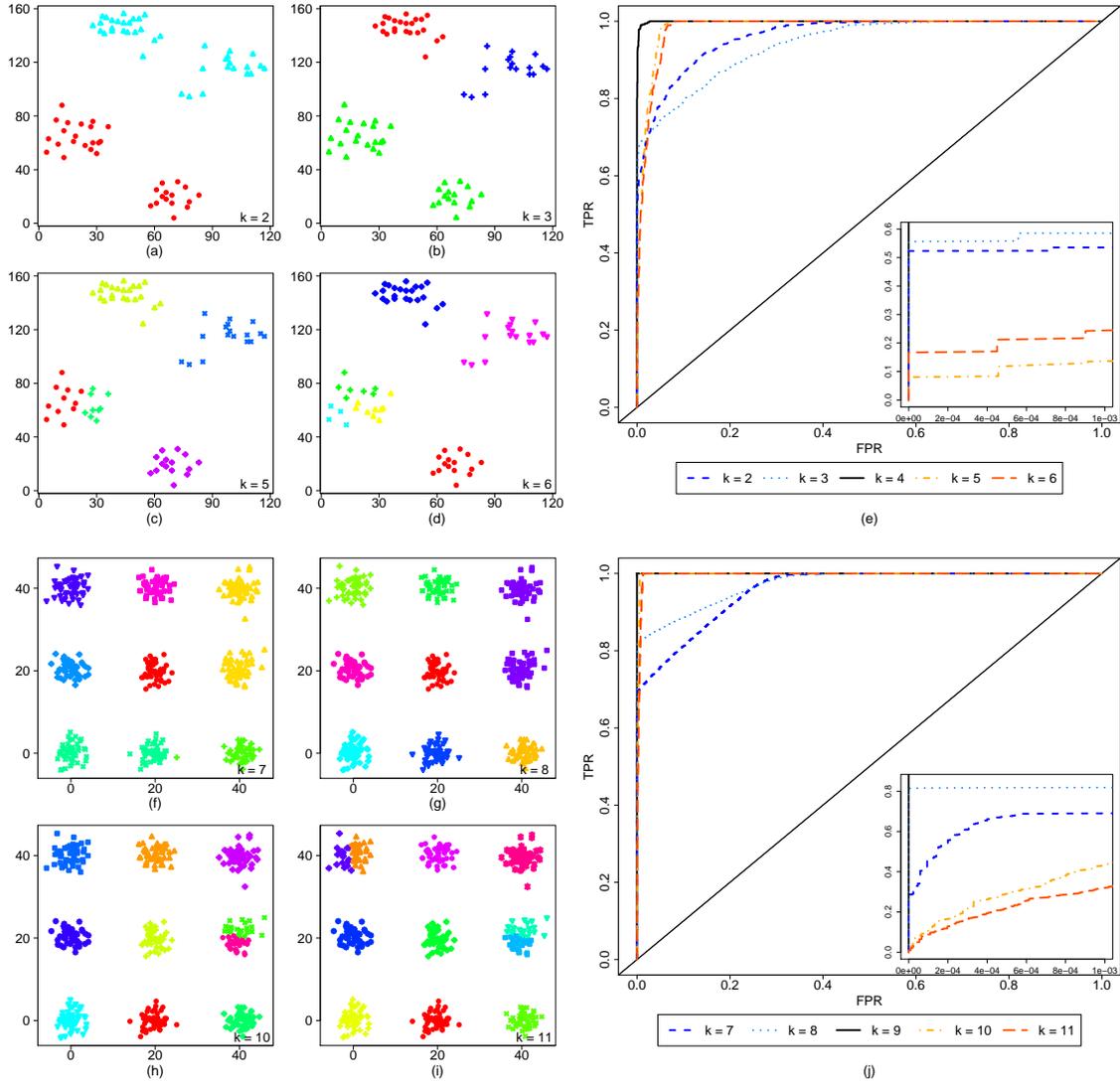


Figure 4: Clustering results and ROC Curves. From (a) to (d) we show clustering solutions with $k = 2, 3, 5$ and 6 for the Ruspini dataset, with (e) their corresponding ROC Curves. From (f) to (i) we show clustering solutions with $k = 7, 8, 10$ and 11 for a simulated dataset, with (j) their corresponding ROC Curves. Partitions were generated with k-means. We zoom in regions at the bottom left corner to highlight the behaviour of FPR and TPR for solutions under- vs over-estimating the number of clusters for small values of the distance threshold. ROC Curves for the (omitted) partitions with the optimal number of clusters are depicted in black in both (e) and (j).

Solutions with the optimal number of clusters (whose partitions are not displayed in Figure 4 for the sake of compactness) resulted in the best AUC values, namely, 0.9994532 for Ruspini and 1.00 for the simulated dataset. Solutions under-estimating the number of clusters produced lower AUC

scores than solutions over-estimating the number of clusters. This is because AUCC is based on pairs of objects and a solution merging two equal sized natural clusters (under-estimation) will produce four times more errors than solutions splitting a natural cluster into two balanced halves (over-estimation). In addition, the main observed trends when comparing these solutions are as follows: starting from the bottom left of the ROC chart, as the distance (resp. similarity) threshold is raised (resp. lowered), under-estimated solutions initially tend to sustain lower FPR values while TPR increases, until a point beyond which there is mixed, alternated increments on TPR and FPR, such that very high, more steady values of TPR (approaching 1) are only reached for higher values of the distance threshold and FPR (see ROC curves in cold colors). This is because under-estimated solutions tend to merge natural clusters, so errors tend to occur from larger distances, i.e., at the lower region of the similarity rank, where there will be mixed negative (0s) and positive (1s) pairs whereas ideally there should only be negative ones. In contrast, over-estimated solutions observe earlier increments on FPR, which however don't prevent high, more steady values of TPR approaching 1 from also being reached earlier, i.e., for lower values of the distance threshold and FPR (see ROC curves in hot colors). This is because over-estimated solutions tend to split natural clusters into more compact sub-clusters, so errors tend to occur at smaller distances, i.e., at the upper region of the similarity rank, where there will be negative pairs (0s, introduced by the splits) mixed with (a reduced number of) positive pairs (1s), whereas ideally there should only be positive ones. This provides insights on how ROC analysis can be used to further investigate under-estimation and over-estimation in cluster solutions.

The ROC Graph and AUCC can also provide valuable insights on the difficulty of the clustering problem itself. In Figure 5 we provide an illustrative example of how different degrees of overlap between clusters can affect ROC Curves. As cluster variances increase, causing clusters to overlap, their corresponding AUCC values decrease considerably. From this perspective, scenarios in which only low to moderate AUCC evaluations are observed for a given ground-truth solution (possibly available for external clustering validation), which cannot be properly recovered despite various different clustering algorithms and parameters having been considered, may be an indicative of the intrinsic difficulties of the problem in hand. Specifically, this may be caused by the fact that the

ground-truth cluster structure in question violates one of the common assumptions about clustering, namely, the assumption that clusters should be compact and separated (Everitt, 1974). This assumption can be interpreted as within-cluster distances expected to be smaller than between-cluster distances. It could also be the case that this is actually mostly true for some suitable underlying distance, which is not the one adopted (to assess the ground-truth or by the algorithms considered) though.

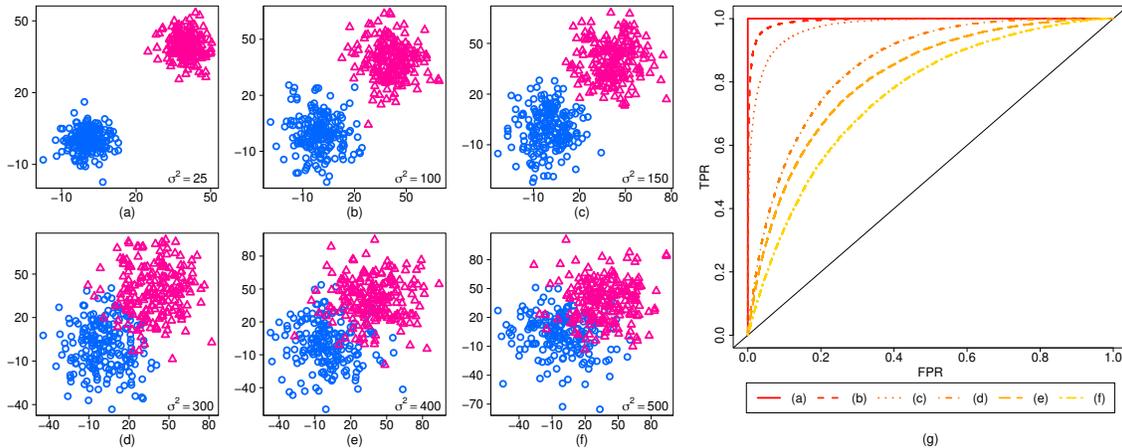


Figure 5: From (a) to (f) we depict six datasets with two clusters (200 objects per cluster) obtained from normal distributions centered at $(0, 0)$ and $(40, 40)$. For each dataset, cluster variances are the same for both clusters and each of their coordinates, fixed at 25, 100, 150, 300, 400, and 500, from (a) to (f). In (g) we depict the ROC Curves for the ground truth-partition of each dataset. Their respective AUC values are 1.00, 0.9923144, 0.9739683, 0.8554201, 0.8129828, and 0.7548744.

5. Conclusions

AUC has been extensively employed in the supervised learning domain as a valuable tool to evaluate and compare different classification models. In this work, we showed that it can also be employed in the unsupervised learning domain, more specifically, in the relative evaluation of clustering results. In this particular setting we introduced the Area Under the Curve for Clustering, AUCC. We theoretically showed that its expected value under a null model of equally likely random clustering solutions is 0.5, irrespective of the number or (im)balance of clusters. To our knowledge, no other relative measure has been theoretically shown to have this property.

We also showed that in the context of internal/relative clustering validation AUCC is a linear transformation of the Gamma Index from Baker and Hubert (1975), for which we have also derived a theoretical expected value under a null model of random clustering partitions. In that context, we showed how ties in (dis)similarity values can be handled consistently across AUCC and Gamma so that their relationship and expected value properties are preserved. We discussed the computational complexity of these criteria and showed that AUCC represents a much more efficient algorithmic way to implement Gamma. We also showed that a visual inspection of the AUCC provides insights on clustering solutions producing under-estimated and over-estimated solutions, as well as on the difficulty of the clustering problem.

In addition to its theoretical, computational and visual appeals, AUCC exhibited very competitive results in our experimental evaluations using well-known classification benchmark datasets. These results need, however, to be taken with a grain of salt. In fact, as argued e.g. by Hennig (2015), cluster analysis can have different aims in different contexts, and is not necessarily or only about finding a unique “true” clustering. Given a diverse collection of candidates, finding a clustering that captures well the structure in the data according to a given notion of similarity may be of interest for reasons other than recovery performance on benchmark datasets with ground-truth, and the AUCC/Gamma Index is one way of doing it. In summary, since there is no free lunch in clustering evaluation, we believe that AUCC can be a useful additional tool in an analyst’s toolbox as a standalone measure or in combination with other measures. Indeed, the combination of relative validity measures has gained attention in the literature recently (Vendramin et al., 2013; Jaskowiak et al., 2016; Kim et al., 2018).

References

- Amigó, E., J. Gonzalo, J. Artiles, and F. Verdejo (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12(5), 613.
- Arbelaitz, O., I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition* 46(1), 243 – 256.

- Baker, F. B. and L. J. Hubert (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association* 70(349), 31–38.
- Bezdek, J. C. and N. R. Pal (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics, Part B* 28(3), 301–315.
- Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7), 1145 – 1159.
- Brock, G., V. Pihur, S. Datta, and S. Datta (2008). cValid: An R package for cluster validation. *Journal of Statistical Software* 25(4), 1–22.
- Calinski, R. and J. Harabasz (1974). A dendrite method for cluster analysis. *Commun Stat* 3, 1–27.
- Ceriani, L. and P. Verme (2012). The origins of the gini index: extracts from *variabilità e mutabilità* (1912) by corrado gini. *The Journal of Economic Inequality* 10(3), 421–443.
- Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *J. of Statistical Software* 61(6), 1–36.
- Davies, D. and D. Bouldin (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1, 224–227.
- Desgraupes, B. (2016). *clusterCrit: Clustering Indices*. R package version 1.2.7.
- Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *J. of Cybernetics* 4, 95–104.
- Everitt, B. (1974). *Cluster analysis*. Heinemann Educational for the Social Science Research Council London.
- Färber, I., S. Günemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek (2010). On using class-labels in evaluation of clusterings. In *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings, Washington, DC*.
- Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Technical report.

- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27(8), 861–874.
- Flach, P., J. Hernández-Orallo, and C. Ferri (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Intl. Conference on Machine Learning — ICML*.
- Flach, P. A. (2010). *Encyclopedia of Machine Learning*, Chapter ROC Analysis, pp. 869–875. Boston, MA: Springer US.
- Giancarlo, R., G. Lo Bosco, L. Pinello, and F. Utro (2013). A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis. *BMC Bioinformatics* 14(Suppl 1), S6.
- Gini, C. (1912). *Variabilità e mutabilità*. Tipogr. di P. Cuppini.
- Goodman, L. and W. Kruskal (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association* 49, 732–764.
- Halkidi, M., Y. Batistakis, and M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems* 17(2-3), 107–145.
- Halkidi, M. and M. Vazirgiannis (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters* 29, 773–786.
- Hand, D. J. and R. J. Till (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 45(2), 171–186.
- Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1), 29–36.
- Hennig, C. (2015). Pattern recognition letters. *What are the true clusters?* 64, 53–62.
- Hennig, C., M. Meila, F. Murtagh, and R. Rocci (2015). *Handbook of cluster analysis*. CRC Press.
- Hernández-Orallo, J., P. Flach, and C. Ferri (2013). ROC curves in cost space. *Machine Learning* 93(1), 71–91.

- Higham, P. A. and D. P. Higham (2019). New improved gamma: Enhancing the accuracy of Goodman–Kruskal’s gamma using ROC curves. *Behavior Research Methods* 51(1), 108–125.
- Hill, R. S. (1980). A stopping rule for partitioning dendrograms. *Botanical Gazette* 141, 321–324.
- Hruschka, E. R., R. J. G. B. Campello, and L. N. Castro (2004). Improving the efficiency of a clustering genetic algorithm. In *Ibero-American Conference on Artificial Intelligence — IBERAMIA*, Volume 3315, pp. 861–870.
- Hruschka, E. R., R. J. G. B. Campello, and L. N. de Castro (2006). Evolving clusters in gene-expression data. *Information Sciences* 176(13), 1898–1927.
- Huang, J. and C. X. Ling (2005, March). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 17(3), 299–310.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2(1), 193–218.
- Hubert, L. J. and J. R. Levin (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* 10, 1072–1080.
- Jain, A. K. and R. C. Dubes (1988). *Algorithms for Clustering Data*. Prentice Hall.
- Jaskowiak, P. A. (2015). *On the evaluation of clustering results: measures, ensembles, and gene expression data analysis*. Ph. D. thesis, University of São Paulo, Brazil (DOI: 10.11606/T.55.2016.tde-23032016-111454).
- Jaskowiak, P. A., R. J. G. B. Campello, and I. G. Costa (2012). Evaluating correlation coefficients for clustering gene expression profiles of cancer. In *7th Brazilian Symposium on Bioinformatics (BSB2012)*, Volume 7409 of *LNCS*, pp. 120–131. Springer / Berlin Heidelberg.
- Jaskowiak, P. A., R. J. G. B. Campello, and I. G. Costa (2014). On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics* 15 Suppl 2(Suppl 2), S2.
- Jaskowiak, P. A., R. J. G. B. Campello, and I. G. Costa Filho (2013). Proximity measures for clustering gene expression microarray data: A validation methodology and a comparative analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10(4), 845–857.

- Jaskowiak, P. A., D. Moulavi, A. C. S. Furtado, R. J. G. B. Campello, A. Zimek, and J. Sander (2016). On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems* 47(2), 329–354.
- Kim, B., H. Lee, and P. Kang (2018). Integrating cluster validity indices based on data envelopment analysis. *Applied Soft Computing* 64, 94–108.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematics, Statistics, and Probabilistics*, Volume 1, pp. 281–297.
- Majnik, M. and Z. Bosnić (2013, May). Roc analysis of classifiers in machine learning: A survey. *Intell. Data Anal.* 17(3), 531–558.
- Mason, S. J. and N. E. Graham (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128(584), 2145–2166.
- Maulik, U. and S. Bandyopadhyay (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(12), 1650–1654.
- Milligan, G. W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* 46(2), 187–199.
- Milligan, G. W. and M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2), 159–179.
- Moulavi, D., P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander (2014). Density-based clustering validation. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, pp. 839–847.
- Nguyen, T., J. Viehman, D. Yeboah, G. R. Olbricht, and T. Obafemi-Ajayi (2020). Statistical comparative analysis and evaluation of validation indices for clustering optimization. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 3081–3090.

- Pakhira, M. K., S. Bandyopadhyay, and U. Maulik (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition* 37, 487–501.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Proceedings of the Royal Society of London* 59, 69–71.
- Provost, F. and T. Fawcett (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 43–48. AAAI Press.
- Provost, F. J., T. Fawcett, and R. Kohavi (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, San Francisco, CA, USA, pp. 445–453. Morgan Kaufmann Publishers Inc.
- Ratkowsky, D. A. and G. N. Lance (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal* 10, 115–117.
- Romano, S., N. X. Vinh, J. Bailey, and K. Verspoor (2016). Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research* 17(1), 4635–4666.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20, 53–65.
- Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, San Francisco, CA, USA, pp. 160–163. Morgan Kaufmann Publishers Inc.
- Vendramin, L., R. J. G. B. Campello, and E. R. Hruschka (2009). On the comparison of relative clustering validation criteria. In *Proceedings of the 9th SIAM International Conference on Data Mining (SDM)*, Sparks, NV, pp. 733–744.
- Vendramin, L., R. J. G. B. Campello, and E. R. Hruschka (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining* 3(4), 209–235.

- Vendramin, L., P. A. Jaskowiak, and R. J. G. B. Campello (2013). On the combination of relative clustering validity criteria. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), Baltimore, MD*, pp. 4:1–12.
- Xu, R., D. Wunsch, and D. Wunsch II (2009). *Clustering*. IEEE Press.
- Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17(10), 977–987.
- Zhou, S., F. Liu, and W. Song (2021, Apr). Estimating the optimal number of clusters via internal validity index. *Neural Processing Letters* 53(2), 1013–1034.