A comparison of external clustering evaluation indices in the context of imbalanced data sets

Marcilio C. P. de Souto†, André L. V. Coelho‡, Katti Faceli§, Tiemi C. Sakata‡, Viviane Bonadia†, Ivan G. Costa†
† Centro de Informática, Universidade Federal de Pernambuco, Brazil, Email: {mcps, igcf}@cin.ufpe.br
‡ Programa de Pós-graduação em Informática Aplicada, Universidade de Fortaleza, Brazil, Email: acoelho@unifor.br
§ Departamento de Computação de Sorocaba, Universidade Federal de São Carlos, Brazil, Email: {katti, tiemi}@ufscar.br

Abstract—For highly imbalanced data sets, almost all the instances are labeled as one class, whereas far fewer examples are labeled as the other classes. In this paper, we present an empirical comparison of seven different clustering evaluation indices when used to assess partitions generated from highly imbalanced data sets. Some of the metrics are based on matching of sets (F-measure), information theory (normalized mutual information and adjusted mutual information), and pair of objects counting (Rand and adjusted Rand indices). We also investigate the BCubed metric, which takes into account the concepts of recall, precision, as well as counting pairs. Furthermore, in order to avoid the class size imbalance effect, we propose a modification to the Rand index, referred to as the normalized class size Rand (NCR) index. In terms of results, apart from NCR, our experiments indicate that all the other analyzed indices are not able to deal properly with the problem of class size imbalance.

Keywords—Clustering Algorithms, External Evaluation Indices, Imbalanced Data Sets.

I. INTRODUCTION

A. Motivation

Clustering is a type of unsupervised learning that has as aim to find the underlying structure, composed of clusters, present in the data [1], [2]. Examples (objects or instances) belonging to each cluster should share some relevant property (similarity) regarding the data domain. The definition of effective evaluation measures is important to provide the users with a degree of “goodness” (quality) for the clustering results derived from the algorithms used. These assessments should be objective and have no bias to any algorithm.

Indeed, several clustering validity measures have been proposed to evaluate the quality of clustering results [3], [4], [5], [6], [7]. In this work, we are interested in external validation indices, which are employed to measure the extent to which cluster labels match externally supplied class labels.

Recently, there have been some works trying to characterize, based on some formal constraints, which aspects of the quality of a partition are captured by different external validation indices. Cluster homogeneity and cluster completeness are examples of basic constraints (or “desirable properties”) that clustering accuracy measures should display [4], [5], [8]. For instance, cluster homogeneity states that clusters should not mix objects belonging to different classes, that is, they should be homogeneous.

In this paper, we are concerned with the behavior of such external evaluation measures when one has to compare partitions generated from highly imbalanced data sets (or data with a highly skewed class distribution). For highly imbalanced data sets, almost all the instances are labeled as one class (majority class), whereas far fewer examples are labeled as the other classes. Examples of this are cancer gene expression data sets, which usually have a high number of genes (features) and a low number of patients (examples). Moreover, some types of tumors/tissues (classes) have fewer examples compared to others [9], [10].

In the context of supervised learning, standard machine learning algorithms tend to be biased towards the majority class and ignore the minority class. Many researchers have addressed the problem of how to evaluate supervised learning algorithms in the case of class size imbalance [11]. In contrast, up to now, there are few studies trying to characterize to which extent class imbalance can influence the measures used to compare the quality of the partitions generated with clustering algorithms [3], [12] (see next subsection).

The measures investigated in our work are based, respectively, on matching of sets (F-measure), information theory (normalized mutual information and adjusted mutual information) and pair of objects counting (Rand, adjusted Rand indices and BCubed). Furthermore, we also present a modification to the Rand index, in order to deal with the class size imbalance effect. We call this modification as normalized class size Rand (NCR) index.

This work is organized as follows: first we will describe the external validation indices analyzed, including a detailed description of the proposed NCR index (Section II). Next, in Section III, we will describe the formal validation constraints that will be used to evaluate how these indices behave under distinct desirable characteristics. This also includes a newly proposed constraint for evaluating partitions with class imbalance. In Section IV, we describe the data sets and
cluster solution generation strategies used in the experiments and experimental results.

B. Related work

The authors in [12] presented a study of the effect of skewed data distributions on k-means clustering, showing that clustering validation measures such as entropy and purity may not capture this aspect, providing the user with misleading information on the clustering performance. They also showed that the F-measure can provide a somewhat consistent indication about the loss of “true” clusters caused by k-means clustering.

In [3] the authors analyzed several internal validity measures in the context of imbalanced datasets. They showed that the measures investigated do not work well for clusters with different densities and/or sizes. According to the results, such measures usually have a tendency of ignoring clusters with low densities. The authors also proposed a new internal validity measure, which can deal with this situation.

In this paper, like in [3], we present an empirical investigation of several different clustering evaluation measures when used to assess partitions generated from highly imbalanced data sets. However, differently from [3], whose work was in the context of internal evaluation indices, we address external indices. In this sense, our work is closer to the one in [12]. But, in contrast to the latter, which approached a very specific context, our study covers seven different indices and is not dependent on a specific clustering algorithm.

II. EVALUATION INDICES FOR CLUSTERING ALGORITHMS

Before defining the group of external evaluation indices addressed in this work, we will provide some background. A common basis for most of these measures is that they can be computed from a contingency table. Such a table can be built as follows. Given a set of n objects X = \{x_1, x_2, \ldots, x_n\}, consider: (1) \( U = \{U_1, U_2, \ldots, U_R\} \) to be the partition of X given as output by a clustering algorithm, and (2) \( V = \{V_1, V_2, \ldots, V_C\} \) to be the partition formed by a priori information independent of partition \( U \) (a.k.a. the gold standard, ground truth or actual class structure).

Then, we can build a contingency table to represent the overlap between pairs of groups in the two partitions, where \( n_{ij} = |U_i \cap V_j| \) is the number of common objects in cluster \( U_i \) and class \( V_j \); \( n_i \) is the number of objects in \( U_i \); and \( n_j \) is the number of objects in \( V_j \). Also, let \( p(i) \) be \( n_i / n \), \( p(j) \) be \( n_j / n \), and \( p(i,j) \) be \( n_{ij} / n \).

With the exception of the normalized class size Rand index, which will be presented later in this section, Table I presents the definition of all indices used in our analysis: Rand index (R), adjusted Rand index (AR), normalized mutual information (NMI), adjusted mutual information (AMI), F-measure (FM), and Bcubed (B3) index. For most indices, the value 1 indicates high quality of the partition against the ground truth and values close or lower than 0 indicates no agreement. As an exception, we have the two indices adjusted for random agreement (AMI and AR), which give a value of 1 to perfect agreement and values close or lower than 0 for agreement obtained by chance.

The first two indices considered in Table I, namely R and AR, are based on counting pairs of objects on which two partitions agree/disagree [7], [13]. AR is one of the most commonly used variations of the Rand index, and takes into account agreements arising by chance given a hypergeometric distribution [13]. In the case of AR, the lower bound, \(-\kappa\), depends on the exact data partitioning.

The NMI and AMI indices are based on the concepts of entropy and mutual information [6], [14]. Mutual information (MI), in the clustering context, refers to how much we can reduce the uncertainty about the cluster of an object picked at random, on average, given that we know its class in the ground truth \( V \). NMI and AMI are both normalization versions of MI. For example, AMI uses the expected mutual information \( E(MI) \) to make the adjustment for chance in the mutual information [14].

In this work, we investigate the FM version as defined in [12]. FM combines the notions of precision and recall from information retrieval, where each cluster \( U_i \) of the partition being validated is treated as if it were the result of a query and each class \( V_j \) of the ground truth as if it were the desired set of objects for a query. One characteristic of this measure is that it disregards the unmatched portions of the clusters.

Like FM, B3 is based on the notions of precision and recall. However, differently from FM, in calculating these values, B3 considers pairs of objects [4]. In order to do so, these quantities are computed separately for each object \( x_l \). More specifically, looking at the equation at Table I, the precision of \( x_l \), \( Pr(x_l) \), indicates the proportion of objects in the same cluster as \( x_l \) that shares with it the same class (including \( x_l \) itself); and the recall, \( Re(x_l) \), indicates the proportion of objects belonging to the same class of \( x_l \) that shares the same cluster with it. It is important to point out that, among the evaluation indices investigated by Amigó et al. [4], B3 was the only one capable of coping with the “rag bag” constraint (see Section III).

In the following, we propose a variation of the Rand index aimed specifically at assessing partitions generated from highly imbalanced data sets.

Normalized class size Rand index

The new index is based on a normalization of the agreement quantities, as calculated in the Rand index, by dividing them by the maximum possible agreement values for their respective class. By resorting to the notion of contingency table as previously discussed, let \( a \) be the number of objects in a pair that are placed in the same cluster in \( U \) and in the same class in \( V \); \( b \) be the number of objects in a pair that
Table I

EXTERNAL VALIDATION INDICES.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>$\binom{n}{2} + 2 \sum_{i=1}^{R} \frac{\binom{c_i}{2}}{n_i} - \sum_{i=1}^{R} \frac{\binom{n_i}{2}}{2} - \sum_{j=1}^{C} \binom{n_j}{2}$</td>
<td>[0,1]</td>
</tr>
<tr>
<td>AR</td>
<td>$\frac{1}{R} \sum_{i=1}^{R} \sum_{j=1}^{C} \binom{n_{ij}}{2} - \binom{n_i}{2} \sum_{j=1}^{C} \binom{n_j}{2}$</td>
<td>[-\infty,1]</td>
</tr>
<tr>
<td>NMI</td>
<td>$\frac{\sum_{i=1}^{R} \sum_{j=1}^{C} p(i,j) \log_2 \frac{p(i,j)}{p(i)p(j)}}{\sum_{i=1}^{R} p(i) \log_2 p(i)} - \frac{\sum_{j=1}^{C} p(j) \log_2 p(j)}{\sum_{j=1}^{C} p(j)}$</td>
<td>[0,1]</td>
</tr>
<tr>
<td>AMI</td>
<td>$\max \left{ \sum_{i=1}^{R} \sum_{j=1}^{C} p(i,j) \log_2 \frac{p(i,j)}{p(i)p(j)} - \frac{\sum_{j=1}^{C} p(j) \log_2 p(j)}{\sum_{j=1}^{C} p(j)} \right}$</td>
<td>[0,1]</td>
</tr>
<tr>
<td>FM</td>
<td>$\sum_{i=1}^{R} \frac{n_{ij}}{n} \cdot \max_{i,j} \left{ \left[ 2 \cdot \frac{n_{ij}}{n_i} \cdot \frac{n_{ij}}{n_j} \right] \left[ \frac{n_{ij}}{n_i} + \frac{n_{ij}}{n_j} \right] \right}$</td>
<td>[0,1]</td>
</tr>
<tr>
<td>B3</td>
<td>$\frac{2 \sum_{l=1}^{C} \text{Pr}(x_l) \sum_{l=1}^{C} \text{Re}(x_l)}{n \cdot \left( \sum_{l=1}^{C} \text{Pr}(x_l) + \sum_{l=1}^{C} \text{Re}(x_l) \right)}$</td>
<td>[0,1]</td>
</tr>
</tbody>
</table>

are placed in the same cluster in $\mathcal{U}$ and in different classes in $\mathcal{V}$; $c$ be the number of objects in a pair that are placed in the same class in $\mathcal{V}$ and in different clusters in $\mathcal{U}$; and $d$ be the number of objects in a pair that are placed in different clusters in $\mathcal{U}$ and different classes in $\mathcal{V}$. Moreover, let $\mathcal{V}_s$ denote the subset of classes of $\mathcal{V}$ that are singleton (i.e., have a single object) and let $C^*$ denote the cardinality of $\mathcal{V} \setminus \mathcal{V}_s$. The "normalized" versions of $a, b, c,$ and $d$ can be computed for comparing partitions $\mathcal{U}$ and $\mathcal{V} \setminus \mathcal{V}_s$ as follows (Eqs. (1)-(4)). One should notice that the largest number of true positive agreement pairs for a particular class $V_j$ amounts to $\binom{n_j}{2}$, whereas the maximum number of true negative cases between a pair of classes $V_j$ and $V_k$ is $n_j n_k$.

\[ a_N = \frac{1}{C^*} \sum_{i=1}^{C} \sum_{j=1}^{C^*} \binom{n_{ij}}{2} \binom{n_i}{2} \quad (1) \]
\[ b_N = \frac{1}{C^*} \sum_{j=1}^{C^*} \binom{n_j}{2} - \sum_{i=1}^{R} \binom{n_i}{2} \binom{n_j}{2} = 1 - a_N \quad (2) \]
\[ c_N = \frac{1}{C^*} \sum_{i=1}^{R} \sum_{k=j+1}^{C} \binom{n_{ij}}{2} + \sum_{i=1}^{R} \sum_{s=i+1}^{R} \binom{n_{ij}}{2} \binom{n_{sk}}{2} n_{j,k} \quad (3) \]
\[ d_N = \frac{1}{C^*} \sum_{i=1}^{R} \sum_{s=i+1}^{R} \sum_{j=1}^{C} \sum_{k=j+1}^{C} \binom{n_{ij}}{2} \binom{n_{jk}}{2} n_{sk} \quad (4) \]

The normalized class size Rand (NCR) index can then be defined as:

\[ NCR = \frac{a_N + d_N}{a_N + b_N + c_N + d_N}. \quad (5) \]

The index has values between 0 and 1, where 1 indicates perfect agreement.

III. DESIRABLE PROPERTIES FOR CLUSTERING EVALUATION MEASURES

Based on previous work [5], [8], Amigó et al. [4] characterized four formal constraints on external evaluation indices for clustering. To present these constraints, also referred to here as "desirable properties", we adopt the same methodology as in [4]. We will assume that, in reference to the ground truth $\mathcal{V}$, partition $\mathcal{U}_2$ is a better option than partition $\mathcal{U}_1$, according to some intuitive notion. Such a notion, which should be satisfied by any evaluation measure $f$, imposes that $f(\mathcal{U}_1) < f(\mathcal{U}_2)$. We will also use the examples in Table II as illustrations of the concepts presented.
The first desirable property, which is a very basic restriction, states that clusters should not mix objects belonging to different classes, that is, they should be homogeneous. We refer to such a property as cluster homogeneity. This is illustrated in the example given in the first row of Table II. Taking $\mathcal{V}$ as reference, one can observe that $\mathcal{U}_2$ has more homogeneous clusters than $\mathcal{U}_1$. So, $f(\mathcal{U}_1) < f(\mathcal{U}_2)$ should be true. Another basic property, which is complementary to cluster homogeneity, is cluster completeness. According to this property, objects belonging to the same class in $\mathcal{V}$ should be assigned to the same cluster. By observing the second row of Table II, one can notice that $\mathcal{U}_2$ splits the group of black circles into a fewer number of clusters than $\mathcal{U}_1$ does, so the score $f(\mathcal{U}_2)$ should be higher.

In some real-world problems, as claimed by [4], it could be useful to have a cluster representing a “rag bag” (or miscellaneous) of objects. Such a cluster would contain objects that should not be grouped with any other object — these objects could represent different classes of a single object (singletons). Based on this, they define a property called rag bag. According to this property, putting a heterogeneous object into a pure cluster ($\mathcal{U}_1$) should be penalized more than putting it into a rag bag ($\mathcal{U}_2$) — third row of Table II.

A small error in a large cluster is preferable to a large number of small errors in small clusters. This is what is known as the cluster size versus quantity constraint. Indeed, if a small class were scattered into small clusters, it is likely that those resulting clusters would be interpreted as noise. As a consequence, the small class could not be discovered from the partition. Such a property is illustrated at the fourth row of Table II.

By presenting these four properties, Amigó et al. [4] do not claim that they should be considered as exhaustive. For example, another desirable property is the correction for randomness, which for instance is present at the adjusted Rand index.

On the other hand, some properties are domain specific, as the rag bag. Although it is a reasonable constraint in some sense, the fact is that all but one (B3) of the clustering evaluation indices investigated by the authors failed in properly satisfying it.

In our case, we are concerned with highly imbalanced data sets. In these data sets, almost all the instances are labeled as one class, whereas far fewer examples are labeled as the other classes. In such a context, we propose here a novel property, named as the class size imbalance effect. The motivation for the definition of such a property is that, for highly imbalanced data sets, the distribution of objects among the actual classes of the data is so heterogeneous that it could introduce some artifacts in the calculation of the evaluation measure used to compare the quality of two alternative partitions, bringing about misleading results. In the same way as B3 stands for the rag bag constraint, the normalized class size Rand index was conceived to target the class size imbalance effect.

More specifically, for highly imbalanced data sets, the misallocation of an object in the clusters should take into account whether the object belongs to a large or a small class, as well as whether it has been placed into a large or a small cluster. This situation is illustrated at the last row of Table II. Taking $\mathcal{V}$ as reference, for partition $\mathcal{U}_1$, by putting only one object of the class of black squares (i.e., the minority class) together with the objects of the cluster with black circles (i.e., the majority cluster), it leads to a 50% misplacement of the elements of the minority class. In contrast, looking at $\mathcal{U}_2$, by putting two objects of the class of black circles (i.e., the majority class) in the cluster with black squares (i.e., the minority cluster), one is misplacing only 10% of the elements of the majority class.

Table III shows the values yielded by the seven external evaluation indices for the partitions that we used to illustrate the five constraints (Table II). With the exception of the partitions at the last row of Table II, which regards the property of class size imbalance, all the other partitions were taken from the example presented in [4].

As one can observe, the basic properties (homogeneity and completeness) are satisfied by the majority of the indices analyzed, exception given to FM. Such a measure does not satisfy the completeness property, since, by definition, each class is evaluated taking into account the cluster that has

<table>
<thead>
<tr>
<th>Prop.</th>
<th>$\mathcal{V}$</th>
<th>$\mathcal{U}_1$</th>
<th>$\mathcal{U}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hom.</td>
<td><img src="image1" alt="Homogeneity Example" /></td>
<td><img src="image2" alt="Homogeneity Example" /></td>
<td><img src="image3" alt="Homogeneity Example" /></td>
</tr>
<tr>
<td>Comp.</td>
<td><img src="image4" alt="Completeness Example" /></td>
<td><img src="image5" alt="Completeness Example" /></td>
<td><img src="image6" alt="Completeness Example" /></td>
</tr>
<tr>
<td>R.B.</td>
<td><img src="image7" alt="Rag Bag Example" /></td>
<td><img src="image8" alt="Rag Bag Example" /></td>
<td><img src="image9" alt="Rag Bag Example" /></td>
</tr>
<tr>
<td>S × Q</td>
<td><img src="image10" alt="Size x Quantity Example" /></td>
<td><img src="image11" alt="Size x Quantity Example" /></td>
<td><img src="image12" alt="Size x Quantity Example" /></td>
</tr>
<tr>
<td>Imb.</td>
<td><img src="image13" alt="Imbalance Example" /></td>
<td><img src="image14" alt="Imbalance Example" /></td>
<td><img src="image15" alt="Imbalance Example" /></td>
</tr>
</tbody>
</table>
more objects belonging to it. Thus, changes in other clusters are not detected [4], [5], [8].

However, when we investigate properties such as rag bag and class size imbalance, which are more domain specific, only B3 and NCR, respectively, were able to indicate the correct partition, viz. $U_2$. Furthermore, one should notice that this sort of asymmetry in terms of error measurement due to class size imbalance is not captured well by the “cluster size versus quantity” constraint, for its focus is on the relation between cluster fragmentation and cluster sizes. Ideally, a good evaluation measure should be resilient to this sort of asymmetry, rewarding partitions that somehow avoid the migration of objects belonging to underrepresented groups. In the next section, we turn our analysis to the specific context of highly imbalanced data sets.

Table III
VALUES OF MEASURES FOR PARTITIONS IN TABLE II. BOLD VALUES INDICATE CASES WHERE THE CONSTRAINT WAS MET.

<table>
<thead>
<tr>
<th>Index</th>
<th>H</th>
<th>C</th>
<th>RB</th>
<th>SQ</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>R($U_1$)</td>
<td>0.799</td>
<td>0.703</td>
<td>0.722</td>
<td>0.949</td>
<td>0.935</td>
</tr>
<tr>
<td>R($U_2$)</td>
<td>0.413</td>
<td>0.244</td>
<td>0.400</td>
<td>0.804</td>
<td>0.861</td>
</tr>
<tr>
<td>AR($U_2$)</td>
<td>0.450</td>
<td>0.311</td>
<td>0.400</td>
<td>0.804</td>
<td>0.753</td>
</tr>
<tr>
<td>NCR($U_1$)</td>
<td>0.727</td>
<td>0.533</td>
<td>0.767</td>
<td>0.600</td>
<td>0.750</td>
</tr>
<tr>
<td>NCR($U_2$)</td>
<td>0.733</td>
<td>0.557</td>
<td>0.667</td>
<td>0.960</td>
<td>0.952</td>
</tr>
<tr>
<td>NMI($U_1$)</td>
<td>0.672</td>
<td>0.366</td>
<td>0.659</td>
<td>0.884</td>
<td>0.839</td>
</tr>
<tr>
<td>NMI($U_2$)</td>
<td>0.732</td>
<td>0.605</td>
<td>0.659</td>
<td>0.942</td>
<td>0.768</td>
</tr>
<tr>
<td>AMI($U_1$)</td>
<td>0.371</td>
<td>0.320</td>
<td>0.229</td>
<td>0.513</td>
<td>0.749</td>
</tr>
<tr>
<td>AMI($U_2$)</td>
<td>0.441</td>
<td>0.382</td>
<td>0.229</td>
<td>0.777</td>
<td>0.663</td>
</tr>
<tr>
<td>FM($U_1$)</td>
<td>0.693</td>
<td>0.625</td>
<td>0.610</td>
<td>0.795</td>
<td>0.956</td>
</tr>
<tr>
<td>FM($U_2$)</td>
<td>0.722</td>
<td>0.625</td>
<td>0.630</td>
<td>0.957</td>
<td>0.934</td>
</tr>
<tr>
<td>B3($U_1$)</td>
<td>0.714</td>
<td>0.704</td>
<td>0.585</td>
<td>0.818</td>
<td>0.944</td>
</tr>
<tr>
<td>B3($U_2$)</td>
<td>0.778</td>
<td>0.723</td>
<td>0.714</td>
<td>0.934</td>
<td>0.891</td>
</tr>
</tbody>
</table>

IV. EXPERIMENTS AND RESULTS

A. Experimental Setting

We analyze here three data sets with the following class distributions: dataset1 {139, 17, 21, 20}, dataset2 {42, 9, 11}, and dataset3 {17, 40, 5}. They present typical structures of cancer gene expression data, that is, they have a highly skewed class distribution. See [15] for more details. For example, for the case of dataset1, the most populated class (majority class) has 139 instances, whereas the less populated one (minority class) has only 17 instances. In this specific case, when compared to the majority class, all the other classes have significantly fewer examples.

In order to assess the impact of the class size imbalance on the evaluation indices, we employ the previous data sets. More specifically, let $\mathcal{V}$ be the ground truth (or actual class structure). From $\mathcal{V}$, we generate the following hypothetical clustering solutions:

1) We move 10% of the elements of the largest class (majority class) to the smallest class (minority class) ($S_1$).

2) We move 60% of the elements of the smallest class to the largest class ($S_2$).

3) We merge the largest and smallest classes ($S_3$).

4) We move 60% of the elements of the smallest class to the second smallest class ($S_4$).

5) We merge the two smallest classes ($S_5$).

Intuitively, if one has to compare the quality of the partitions from these five scenarios, the case in which only 10% of the elements of the majority class are misplaced ($S_1$) seems to be less damaging than the other scenarios ($S_2$ to $S_5$). Note also that it would not make sense to use real solutions from clustering methods, as we could not tell a priori which solution should be preferable. With these simulated data, we can control the generation of clustering solutions for exploring desirable and undesirable solutions.

B. Results

Table IV illustrates the results of the seven evaluation indices applied to the five different scenarios in the context of the three data sets.

As highlighted in bold, out of the seven indices analyzed, only NCR was able to indicate that the partition from scenario $S_1$ should be ranked better than the others ($S_2$ to $S_5$) in all three data sets.

FM was also able to discriminate the situation in dataset2. In particular, R, CR, NMI, AMI, and B3 always yielded a larger value for scenario $S_5$ than for $S_1$. That is, they somewhat “prefer” a partition in which two small classes are merged to have a 10% error in the majority class. Note that, for the three data sets analyzed, the number of objects moved from the majority class were, respectively, 14, 5, and 4. These values are all smaller than the number of objects in the respective minority class (respectively, 17, 9, and 5).

For some indices, the explanation for such behavior follows directly from their definition. CR and R are based on counting how many pairs of objects are in the same class and grouped correctly (a.k.a. true positives). For example, by removing 10% of the elements from the larger class with 139 objects in dataset1 (scenario $S_1$), we have $14 \times (139 - 14) = 1750$ false negative errors out of a total of 9591 positive ones. On the other hand, the error arising for joining the two smaller classes, a case of false positives (scenario $S_5$), amounts to only $17 \times 20 = 340$. That is, the combinatorics of pairs imposes a larger penalization from errors arising from larger classes than smaller ones.

Information theoretic indices, as $NMI$ and $AMI$, display the similar behavior of penalizing more errors in larger classes, but their formulation is more intricate and not straightforward to interpret.

Albeit capturing the desired behavior for dataset2, $FM$ produced very similar values for the distinct scenarios over the other data sets. As discussed in Section III, in the case of $FM$, each class is evaluated by taking into account
the cluster that has more objects belonging to it. Therefore, such a measure can present a poor performance when the match is not possible. Indeed, FM in general does not satisfy the completeness property (see Table III).

For B3, a probably explanation for its behavior may have to do with the “rag bag” property. To a certain extent, by merging (or exchanging items from) two small classes, B3 may “interpret” this situation as the formation of a rag bag, thus endorsing it. Conversely, when elements from a large class are moved to a smaller one, such measure, like R and CR, puts more emphasis on the counting of pairs of objects.

V. CONCLUSIONS

As far as we are aware of, this is the first study analyzing the class size imbalance effect on a range of external clustering validation indices. In this context, we propose NCR as an extension to the Rand index to cope better with this effect. Our preliminary empirical results reveal that not all the indices considered here, apart from NCR, have failed in scenarios involving different levels of class size imbalance. Note that, despite being based on few data sets, these conclusions are general, as the failure of an index to hold a particular desirable property in a single data set is already a strong assessment.

As future work, we shall conduct more systematic experiments to investigate more deeply the sensitivity levels displayed by the indices (especially, NCR) to small perturbations applied to the class size distributions. We also plan to characterize in more detail the trade-off that may exist among the desired properties; in particular, how the satisfaction of the class size imbalance property really impacts the satisfaction of the others.

ACKNOWLEDGMENTS

The authors would like to thank CNPq, CAPES, FACEPE and FAPESP for the financial support provided.

REFERENCES


<table>
<thead>
<tr>
<th>Index</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
<th>Dataset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
</tr>
<tr>
<td>R</td>
<td>0.897</td>
<td>0.917</td>
<td>0.878</td>
</tr>
<tr>
<td>AR</td>
<td>0.795</td>
<td>0.833</td>
<td>0.752</td>
</tr>
<tr>
<td>NCR 0.969</td>
<td>0.885</td>
<td>0.917</td>
<td>0.885</td>
</tr>
<tr>
<td>NMI</td>
<td>0.830</td>
<td>0.854</td>
<td>0.840</td>
</tr>
<tr>
<td>AMI</td>
<td>0.775</td>
<td>0.779</td>
<td>0.701</td>
</tr>
<tr>
<td>FM</td>
<td>0.937</td>
<td>0.932</td>
<td>0.890</td>
</tr>
<tr>
<td>B3</td>
<td>0.896</td>
<td>0.927</td>
<td>0.917</td>
</tr>
</tbody>
</table>