Advances in Data Analysis and Classification (2023) 17:125–133
https://doi.org/10.1007/s11634-022-00491-w

REGULAR ARTICLE

Minimum adjusted Rand index for two clusterings of a given size

José E. Chacón[1] · **Ana I. Rastrojo**[2]

Received: 8 December 2020 / Revised: 23 November 2021 / Accepted: 24 January 2022 /
Published online: 9 February 2022

Abstract
The adjusted Rand index (ARI) is commonly used in cluster analysis to measure the degree of agreement between two data partitions. Since its introduction, exploring the situations of extreme agreement and disagreement under different circumstances has been a subject of interest, in order to achieve a better understanding of this index. Here, an explicit formula for the lowest possible value of the ARI for two clusterings of given sizes is shown, and moreover a specific pair of clusterings achieving such a bound is provided.

Keywords Adjusted Rand index · External clustering evaluation · Minimum agreement

Mathematics Subject Classification Primary 62H30 · Secondary 62H17

1 Introduction

The adjusted Rand index is one of the most commonly used similarity measures to compare two clusterings of a given set of objects. Indeed, it is the recommended criterion for external clustering evaluation in the seminal study of Milligan and Cooper (1986). Nevertheless, many other measures for external clustering evaluation were recently surveyed in Meilă (2016).

Initially, Rand (1971) considered a similarity index between two clusterings (the Rand index) defined as the proportion of object pairs that are either assigned to the same cluster in both clusterings or to different clusters in both clusterings. However, Morey

✉ José E. Chacón
jechacon@unex.es

[1] Departamento de Matemáticas, Universidad de Extremadura, E-06006 Badajoz, Spain

[2] Departamento de Matemáticas, IES Sierra La Calera, E-06150 Santa Marta de los Barros, Badajoz, Spain

and Agresti (1984) noted that such an index does not take into account the possible agreement by chance, and Hubert and Arabie (1985) introduced a corrected-for-chance version of the Rand index, which is usually known as the adjusted Rand index (ARI).

Exploring the situations of extreme agreement, as measured by the ARI, has been a subject of interest since the very inception of this index. Indeed, Hubert and Arabie (1985) posed the problem of finding the maximum ARI subject to given clustering marginals; i.e., when constrained to have fixed, given cluster sizes in each of the clusterings. Numerical algorithms to tackle this problem were developed initially by Messatfa (1992), and later by Brusco and Steinley (2008) and Steinley et al. (2015), and an explicit solution for clusterings of size two has been recently shown in Chacón (2021b).

A related but different problem concerns the obtention of lower bounds for the ARI of two clusterings of given sizes. According to Meilă (2016, p. 631), "the lower bound is usually hard to calculate". It should be noted that, due to the correction for chance, the ARI may take negative values for extremely discordant clusterings. This happens when the agreement between the two clusterings is less than the expected agreement when the clusters assignments are made at random, keeping the given marginals. Hence, finding the minimum possible ARI value allows quantifying how extreme is the discordance between two clusterings of given sizes.

Moreover, if the interest is to measure discordance instead of agreement, the ARI can be transformed into a semimetric by considering $\mathrm{ARD} = 1 - \mathrm{ARI}$ (Chacón 2021a). Thus, perfect agreement corresponds to null discordance, or $\mathrm{ARD} = 0$, and the case of less agreement than random assignment is related to values of $\mathrm{ARD} > 1$. However, in general, when dealing with semimetrics for measuring clustering disagreement it is useful to normalize them so that they take values in $[0, 1]$; for instance, Charon et al. (2006), Denœud (2008, Section 2.4) and Meilă (2016, Section 27.4) explored such a normalization for different distances between partitions. Therefore, obtaining the minimum ARI value makes it possible to define a normalized version of the ARD.

The main contribution of this paper is to find a lower bound for the ARI of two clusterings of given sizes, and to show that this bound is indeed the best possible one, since it is attained by an explicit pair of clusterings. More precise notation is introduced in Sect. 2, where in addition the main result is rigourously stated. Two numerical examples showing the possible applications of this result are presented in Sect. 3, and the proofs of the main result and another auxiliar lemma of independent interest are given in Sect. 5.

2 Notation and main result

A clustering of a set $\mathcal{X}$ of n objects is a partition of $\mathcal{X}$ into non-empty, disjoint and exhaustive classes, called clusters. The number of such classes is known as the size of the clustering. Given two clusterings $\mathscr{C} = \{C_1, \ldots, C_r\}$ and $\mathscr{D} = \{D_1, \ldots, D_s\}$, of sizes r and s, respectively, all the information regarding their concordance is registered in the $r \times s$ matrix $\mathbf{N}$ whose (i, j)th element n_{ij} records the cardinality of $C_i \cap D_j$. This matrix is usually known as confusion matrix or contingency table. Its row-wise and column-wise totals, $(n_{1+}, \ldots, n_{r+})$ and $(n_{+1}, \ldots, n_{+s})$, with $n_{i+} = \sum_{j=1}^{s} n_{ij}$ and $n_{+j} = \sum_{i=1}^{r} n_{ij}$, give an account of the cluster sizes in $\mathscr{C}$ and $\mathscr{D}$, respectively,

and are commonly referred to as the marginals, or marginal clustering distributions. Note that all cluster sizes must be strictly greater than zero in order to respect the assumptions on the clustering sizes, which implies that $n \geq \max\{r, s\}$.

The Rand index is a summary statistic for $\mathbf{N}$, based on inspecting the behaviour of object pairs across the two clusterings. There are four possible types of object pairs, formed by taking into account if: a) both objects belong to the same cluster in both clusterings, b) they belong to the same cluster in $\mathscr{C}$ but to different clusters in $\mathscr{D}$, c) they belong to different clusters in $\mathscr{C}$ but to the same cluster in $\mathscr{D}$, and d) they belong to different clusters in both clusterings. The cardinalities of each of these categories will be denoted a, b, c and d, respectively. They can be easily expressed in terms of the entries of $\mathbf{N}$ and its marginals; for instance, Hubert and Arabie (1985) noted that

$$
\begin{aligned}
a &= \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}^2\right) - n}{2}, \\
b &= \frac{\sum_{i=1}^{r} n_{i+}^2 - \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}^2}{2}, \\
c &= \frac{\sum_{j=1}^{s} n_{+j}^2 - \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}^2}{2}, \\
d &= \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}^2\right) + n^2 - \sum_{i=1}^{r} n_{i+}^2 - \sum_{j=1}^{s} n_{+j}^2}{2}.
\end{aligned}
$$

With this notation, the Rand index is obtained as $\mathrm{RI} = (a+d)/(a+b+c+d) = (a+d)/N$ where $N = a+b+c+d = \binom{n}{2} = n(n-1)/2$ is the total number of pairs of objects from $\mathcal{X}$. It takes values in $[0, 1]$, with 1 corresponding to perfect agreement between the clusterings and 0 attained for the comparison of the two so-called trivial clusterings: one with all the n objects in a single cluster, and the other one with n clusters with a single object in each of them (see Albatineh et al. 2006).

One of the drawbacks of the Rand index is that it does not take into account the possibility of agreement by chance between the two clusterings (Morey and Agresti 1984). Hence, Hubert and Arabie (1985) obtained $\mathbb{E}[\mathrm{RI}]$, the expected value of this index when the partitions are made at random, but keeping the same marginal clustering distributions, and suggested to alternatively use the ARI, a corrected-for-chance version of the Rand index defined by $\mathrm{ARI} = (\mathrm{RI} - \mathbb{E}[\mathrm{RI}])/(1 - \mathbb{E}[\mathrm{RI}])$. Steinley (2004) provided a concise formula for the ARI, which reads as follows:

$$
\mathrm{ARI} = \frac{N(a+d) - \{(a+b)(a+c) + (c+d)(b+d)\}}{N^2 - \{(a+b)(a+c) + (c+d)(b+d)\}}.
$$

A careful inspection shows that the ARI is undefined if and only if $r = s = 1$ or $r = s = n$ (see the details in Sect. 5 below). The first case occurs in the degenerate situation where both of the two compared clusterings have only one cluster (none of them really involves a partition), whereas the second case corresponds to another degenerate scenario, where both clusterings only comprise singleton groups. It is assumed henceforth that none of these two degenerate situations holds.

These preliminaries allow us to formulate the main result of this paper, whose proof is deferred to Sect. 5.

Theorem 1 *The minimum ARI for two clusterings of an arbitrary number of objects, with given sizes r and s, respectively, can be explicitly written as*

$$\min \mathrm{ARI} = \left[1 - \frac{1}{2}\binom{r+s-1}{2}\left\{\binom{r}{2}^{-1} + \binom{s}{2}^{-1}\right\}\right]^{-1} \tag{1}$$

if $\min\{r, s\} \geq 2$ *and* $\min \mathrm{ARI} = 0$ *if* $\min\{r, s\} = 1$. *Such a minimum value is attained for a comparison of precisely* $n = r + s - 1$ *objects, in which the* $r \times s$ *contingency table* $\mathbf{N}$ *has exactly one row of ones, exactly one column of ones and all the remaining entries are zeroes.*

The expression for the minimum ARI given in Theorem 1 is equivalent to, but notably simpler than, the one previously conjectured in Chacón (2021a, p. 218).

If in addition the clustering sizes are allowed to vary, then it is easily seen that the minimum possible value of the ARI is $-1/2$, which corresponds to a 2×2 matrix with one entry equal to zero and all the remaining entries equal to one.

Furthermore, for $r = s \geq 2$, Eq. (1) simplifies to $-r/(3r-2)$, so it follows that for $r = s$ the range of possible ARI values approaches $[-1/3, 1]$ as r increases. Moreover, in order to get insight on the behaviour of the minimum ARI for large values of r and s it is useful to note that, by means of the simple first order approximation $\binom{r}{2} \sim r^2/2$, it is possible to express

$$\min \mathrm{ARI} \approx -\frac{2r^2s^2}{r^4 + 2r^3s + 2rs^3 + s^4}$$

as r and s increase.

3 Examples

3.1 A synthetic data example

Theorem 1 is useful to appreciate how extreme is the discordance between two distant clusterings of given sizes.

For instance, let us consider the example presented in Table 3 in Steinley (2004), which concerns the comparison of two partitions of $n = 13$ objects into $r = s = 5$ clusters. The 5×5 confusion matrix for this example is given by

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Chacón (2021a) noted that this example deals with two very distant clusterings. More precisely, it is easy to check that for this comparison we have $a = 0, b = c = 11$

and $d = 56$, so that $\text{ARI} = -242/1474 \simeq -0.164$. The fact that $\text{ARI} < 0$ already indicates that the agreement between these two partitions is less than the expected agreement if the label assignments would have made at random, so that supports the idea that the two clusterings are quite distant.

But one may wonder if two partitions with 5 clusters each can be made much more distant that these two, and that is precisely the question that Theorem 1 solves, since it shows that the minimum possible agreement for $r = s = 5$ is $\min \text{ARI} = -5/13 \simeq -0.385$. Thus, for two clusterings of size 5 the range of possible ARI values is $[-0.385, 1]$, so the value -0.164 for the partitions in this example is indeed quite close to the lower limit.

Moreover, Theorem 1 also shows that the lowest possible value of the ARI for two clusterings of size 5 is attained for the comparison of two clusterings of 9 objects whose confusion matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

or any other that can be obtained by permuting the rows and/or the columns of the former.

3.2 A real data example

While clustering comparisons are often made based on indices, some authors advocate the advantages of using distances as dissimilarity measures (see Meilă 2016, p. 620). Hence, as noted in the Introduction, another application of Theorem 1 is that it allows normalizing the dissimilarity measure $\text{ARD} = 1 - \text{ARI}$, dividing it by its maximum so that it takes values in [0, 1], which makes it easier to appreciate the relative closeness of two partitions with respect to a third one (see also Charon et al. 2006)

In this sense, let us consider the yeast data set introduced in Nakai and Kaneisa (1991, 1992), a version of which is publicly available at the UCI machine learning repository (https://archive.ics.uci.edu/). The data consist of 8 variables measured on $n = 1484$ proteins. An additional label variable is available, that classifies these proteins according to their cellular localization sites as CYT or ME3, which induces a partition that can be thus viewed as the ground truth. Then, the goal is to compare the partitions obtained by different clustering procedures against the true classification.

Gaussian mixture models (GMMs) and shifted asymmetric Laplace (SAL) mixture models were applied in Franczak et al. (2014) to cluster this data set. The reported fitted SAL mixture model has 2 clusters, with $\text{ARI} = 0.81$, whereas the fitted GMM has 3 clusters and $\text{ARI} = 0.56$. The higher value of the ARI already seems to indicate a better fit for the SAL mixture model but, in order to better appreciate the relative gains of this method over the GMM, it is useful to calculate the normalized ARD, which takes values in [0, 1]. By using Theorem 1, the normalized ARDs for the SAL mixture model fit and the GMM fit can be computed to be 0.13 and 0.33, respectively,

thus showing on a [0, 1] scale how the SAL mixture model fit is quite closer to the true classification than the GMM fit.

Moreover, after aggregating the results for 25 model fits with a fixed number of 2 clusters, based on random initializations with 70 percent of the true labels taken as known, the reported results yielded ARI values for the SAL and GMM clusterings against the ground truth of 0.86 and -0.08, respectively. This shows that extreme discordant results in terms of ARI can indeed occur in practice. And again, to better appreciate on a [0, 1] scale the differences in performance between the two methodologies, it is useful to note that the normalized ARDs for the SAL and GMM clusterings against the ground truth were 0.09 and 0.72, respectively, thus showing a considerably lower normalized dissimilarity for the SAL mixture model clustering against the GMM partition.

4 Conclusions

The adjusted Rand index is one of the most commonly used measures to evaluate the degree of agreement between two data partitions, but, even so, many of its properties are still not fully known. The main contribution of this paper is to provide a precise result showing its minimum value given the clustering sizes, and describing the clustering configurations for which that minimum is attained.

Apart from its theoretical interest, which has been put forward in some Machine Learning forums,[1,2] the usefulness of this result is illustrated via two practical applications: first, knowledge of this lower bound helps appreciating how extreme is the disagreement between two distant partitions; and second, such lower bound allows normalizing the ARD semimetric so that it takes values in [0, 1], thus facilitating the comparison of the dissimilarities of several clusterings with respect to a fixed reference partition, as noted in Charon et al. (2006). A real data example illustrating this situation is included in Sect. 3.2.

Different constraint levels could be imposed when investigating the minimum value of the ARI, and indeed two anonymous reviewers noted that it would be interesting to find the minimum ARI for a fixed sample size n, in addition to the given clustering sizes r and s. This represents a challenging open problem for further research.

5 Proofs

The proof of Theorem 1 makes use of the following result, which is of independent interest. Intuitively, it shows that if a certain amount is to be distributed among several parts, the configuration that yields the maximum sum of the part squares is that which accumulates the highest possible quantity in one of the parts and keeps the remaining ones to their minimum.

[1] *Adjusted Rand index bounds*, https://github.com/scikit-learn/scikit-learn/issues/8166.

[2] *Lower bound for adjusted Rand index?* https://stats.stackexchange.com/questions/254950.

Lemma 1 *Let $a_1 \geq a_2 \geq \cdots \geq a_p$ and $t \geq \sum_{i=1}^{p} a_i$ be real numbers and consider the region*

$$\mathcal{A} \equiv \mathcal{A}(t; a_1, \ldots, a_p) = \big\{(x_1, \ldots, x_p) \in \mathbb{R}^p : \ \textstyle\sum_{i=1}^{p} x_i = t \\ \text{and} \ \ x_i \geq a_i \quad \text{for all} \ \ i = 1, \ldots, p\big\}.$$

The maximum of $\sum_{i=1}^{p} x_i^2$ over $\mathcal{A}$ is attained for $x_1 = t - \sum_{i=2}^{p} a_i$, $x_2 = a_2, \ldots, x_p = a_p$. Hence, $\max\big\{\sum_{i=1}^{p} x_i^2 : (x_1, \ldots, x_p) \in \mathcal{A}\big\} = \big(t - \sum_{i=2}^{p} a_i\big)^2 + \sum_{i=2}^{p} a_i^2$.

Proof The result follows by noting that if $a \leq b$ then $a^2 + b^2 \leq (a - c)^2 + (b + c)^2$ for any $c \geq 0$. □

Now we are ready to prove the main result of the paper. First note that minimizing the ARI is equivalent to maximizing the semimetric ARD $= 1 -$ ARI introduced in Chacón (2021a), where it is also shown that it can be readily expressed as

$$\text{ARD} \equiv \text{ARD}(a, b, c, d) = \frac{N(b + c)}{(a + b)(b + d) + (a + c)(c + d)}. \tag{2}$$

This expression is undefined if the denominator is zero, and this happens if and only if its two summands simultaneously equal zero, a fact that is equivalent to having either $a = b = c = 0$ or $b = c = d = 0$. Albatineh et al. (2006) noted that $b = c = 0$ if and only if there is only one strictly positive entry in each row and column of $\mathbf{N}$; that is, if $r = s$ and $\mathbf{N}$ is a row (or column) permutation of a diagonal matrix. Moreover, they also observed that $a = 0$ if and only if $n_{ij} \in \{0, 1\}$ for all i, j, and that $d = 0$ if and only if $\min\{r, s\} = 1$. So $b = c = d = 0$ is equivalent to $r = s = 1$ and $a = b = c = 0$ is equivalent to $r = s = n$, and those are the only two situations in which the ARI is undefined, which are discarded in our subsequent analysis.

Proof of Theorem 1 It is clear that the roles of b and c in (2) are interchangeable, in the sense that $\text{ARD}(a, b, c, d) = \text{ARD}(a, c, b, d)$. The same is true for the roles of a and d. Moreover, ARD is clearly a decreasing function of a and d, so its maximum value is attained for the lowest possible values of a and d.

As noted above, $a = d = 0$ occurs when one of the clusterings consists of a single cluster (so that $d = 0$), and the contingency table with maximum ARD is a row or column vector of ones (so that $a = 0$). In this case, the resulting ARD $= 1$, so minimum ARI $= 0$.

On the other hand, if $\min\{r, s\} \geq 2$ then necessarily $d > 0$, but it is equally possible to have $a = 0$ if all the entries of $\mathbf{N}$ are just zeroes or ones, so this will be imposed henceforth. Notice that this yields $n \leq rs$, which means that the highest values of the ARD are achieved when the number of objects is small. For $a = 0$ we have $d = N - (b + c)$, and the ARD simplifies to

$$\text{ARD} = \frac{N(b + c)}{b^2 + c^2 + \{N - (b + c)\}(b + c)} = \frac{N}{N - 2bc/(b + c)}, \tag{3}$$

which is an increasing function of b and c. So, to maximize it, we must find the maximum possible values for b and c.

Since $a = 0$, it follows that $b = \left(\sum_{i=1}^{r} n_{i+}^2 - n\right)/2$ and $c = \left(\sum_{j=1}^{s} n_{+j}^2 - n\right)/2$. Hence, maximizing b is equivalent to maximizing the sum of the squared sizes of the clusters of $\mathscr{C}$, constrained to the facts that the total size is n and each cluster has size greater than or equal to one (because degenerate, empty clusters are not allowed). This is exactly the setting of Lemma 1 for $p = r$, $a_1 = \cdots = a_r = 1$ and $t = n$. So for $n \geq r$ (which is necessary to have r non-empty clusters in $\mathscr{C}$), the maximum value of b is attained when there is a cluster in $\mathscr{C}$ with $n - (r - 1)$ objects and the remaining $r - 1$ clusters have one object each, so that $\sum_{i=1}^{r} n_{i+}^2 = \{n - (r - 1)\}^2 + r - 1$.

Moreover, the fact that all $n_{ij} \in \{0, 1\}$ also implies that the maximum size of any cluster in $\mathscr{C}$ is s so, since the configuration maximizing b has a cluster with $n - (r - 1)$ objects, it must be $n - (r - 1) \leq s$. Thus, if we define $z = n - (r - 1)$ then the maximum value of b is $(z^2 - z)/2$. And that upper bound is maximized, among all the sample sizes that satisfy the constraints $1 \leq z \leq s$, by taking precisely $z = s$, that is, $n = r + s - 1$. Hence, the confusion matrix that maximizes b must have one row with all its entries equal to one, and each of the remaining rows having exactly one entry equal to one and all the rest equal to zero. In principle, the nonzero entries of the latter rows could be arbitrarily placed but, mimicking the above reasoning regarding b, the value of c is maximized when there is a column with all its entries equal to one, so the contingency table configuration that maximizes the ARD must be precisely the one announced in the statement of the theorem.

In addition, it is straightforward to check that the configuration that maximizes the ARD has $a = 0$, $b = \binom{s}{2}$, $c = \binom{r}{2}$ and $d = \binom{n}{2} - \binom{r}{2} - \binom{s}{2} = (r - 1)(s - 1)$ since $n = r + s - 1$. Hence, from (3) it follows that the maximum ARD is given by

$$\left[1 - 2\binom{r+s-1}{2}^{-1}\binom{r}{2}\binom{s}{2} \Big/ \left\{\binom{r}{2} + \binom{s}{2}\right\}\right]^{-1}$$

so that the minimum ARI is as stated in the theorem. □

Acknowledgements We thank the Associate Editor and two anonymous referees for their comments, which led to notable improvements in the paper. The first author acknowledges the support of the Spanish Ministerio de Economía y Competitividad Grant PID2019-109387GB-I00 and the Junta de Extremadura Grant GR18016.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Albatineh AN, Niewiadomska-Bugaj M, Mihalko D (2006) On similarity indices and correction for chance agreement. J Classif 23:301–313

Brusco MJ, Steinley D (2008) A binary integer program to maximize the agreement between partitions. J Classif 25:185–193

Chacón JE (2021a) A close-up comparison of the misclassification error distance and the adjusted Rand index for external clustering evaluation. Br J Math Stat Psychol 74:203–231

Chacón JE (2021b) Explicit agreement extremes for a 2×2 table with given marginals. J Classif 38:257–263

Charon I, Denœud L, Guénoche A, Hudry O (2006) Maximum transfer distance between partitions. J Classif 23:103–121

Denœud L (2008) Transfer distance between partitions. Adv Data Anal Classif 2:279–294

Franczak BC, Browne RP, McNicholas PD (2014) Mixtures of shifted asymmetric Laplace distributions. IEEE Trans Pattern Anal Mach Intell 36:1149–1157

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Meilă M (2016) Criteria for comparing clusterings. In: Hennig C, Meilă M, Murtagh F, Rocci R (eds) Handbook of cluster analysis. CRC Press, Boca Raton, pp 619–635

Messatfa H (1992) An algorithm to maximize the agreement between partitions. J Classif 9:5–15

Milligan GW, Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivar Behav Res 21:441–458

Morey LC, Agresti A (1984) The measurement of classification agreement: an adjustment of the Rand statistic for chance agreement. Educ Psychol Meas 44:33–37

Nakai K, Kaneisa M (1991) Expert system for predicting protein localization sites in gram-negative bacteria. Proteins 11:95–110

Nakai K, Kaneisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14:897–911

Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66:846–850

Steinley D (2004) Properties of the Hubert–Arabie adjusted Rand index. Psychol Methods 9:386–396

Steinley D, Hendrickson G, Brusco MJ (2015) A note on maximizing the agreement between partitions: a stepwise optimal algorithm and some properties. J Classif 32:114–126