Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures

Martijn Gösgens [1] **Alexey Tikhonov** [2] **Liudmila Prokhorenkova** [3 4 5]

Abstract

Many cluster similarity indices are used to evaluate clustering algorithms, and choosing the best one for a particular task remains an open problem. We demonstrate that this problem is crucial: there are many disagreements among the indices, these disagreements do affect which algorithms are preferred in applications, and this can lead to degraded performance in real-world systems. We propose a theoretical framework to tackle this problem: we develop a list of desirable properties and conduct an extensive theoretical analysis to verify which indices satisfy them. This allows for making an informed choice: given a particular application, one can first select properties that are desirable for the task and then identify indices satisfying these. Our work unifies and considerably extends existing attempts at analyzing cluster similarity indices: we introduce new properties, formalize existing ones, and mathematically prove or disprove each property for an extensive list of validation indices. This broader and more rigorous approach leads to recommendations that considerably differ from how validation indices are currently being chosen by practitioners. Some of the most popular indices are even shown to be dominated by previously overlooked ones.

1. Introduction

Clustering is an unsupervised machine learning problem, where the task is to group objects that are similar to each other. In network analysis, a related problem is called *community detection*, where groupings are based on relations between items (links), and the obtained clusters are expected to be densely interconnected. Clustering is used across various applications, including text mining, online advertisement, anomaly detection, and many others (Xu & Tian, 2015; Allahyari et al., 2017).

To measure the quality of a clustering algorithm, one can use either internal or external measures. *Internal measures* evaluate the consistency of the clustering result with the data being clustered, e.g., Silhouette, Hubert-Gamma, Dunn indices or modularity in network analysis (Newman & Girvan, 2004). Unfortunately, it is often unclear whether optimizing any of these measures would translate into improved quality in practical applications. *External* (cluster similarity) *measures* compare the candidate partition with a reference one (obtained, e.g., by human assessors). A comparison with such a gold standard partition, when it is available, is more reliable. There are many tasks where external evaluation is applicable: text clustering (Amigó et al., 2009), topic modeling (Virtanen & Girolami, 2019), Web categorization (Wibowo & Williams, 2002), face clustering (Wang et al., 2019), news aggregation (see Section 3), and others. Often, when there is no reference partition available, it is possible to let a group of experts annotate a subset of items and compare the algorithms on this subset.

Dozens of cluster similarity measures exist and which one should be used is a subject of debate (Lei et al., 2017). In this paper, we systematically analyze the problem of choosing the best cluster similarity index. We start with a series of experiments demonstrating the importance of the problem (Section 3). First, we construct simple examples showing the inconsistency of all pairs of different similarity indices. Then, we demonstrate that such disagreements often occur in practice when well-known clustering algorithms are applied to real datasets. Finally, we illustrate how an improper choice of a similarity index can affect the performance of production systems.

So, the question is: how to compare cluster similarity indices and decide which one is best for a particular application? Ideally, we would want to choose an index for which good similarity scores translate to good real-world performance. However, opportunities to experimentally perform such a *validation of validation indices* are rare, typically expensive, and do not generalize to other applications. In contrast, we

[1] Eindhoven University of Technology, Eindhoven, The Netherlands [2] Yandex, Berlin, Germany [3] Yandex, Moscow, Russia [4] Moscow Institute of Physics and Technology, Moscow, Russia [5] HSE University, Moscow, Russia. Correspondence to: Martijn Gösgens <research@martijngosgens.nl>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

suggest a theoretical approach: we formally define properties that are desirable across various applications, discuss their importance, and formally analyze which similarity indices satisfy them (Section 4). This theoretical framework allows practitioners to choose the best index based on relevant properties for their applications. In Section 5, we show how this choice can be made and discuss indices that are expected to be suitable across various applications.

Among the considered properties, *constant baseline* is arguably the most important and non-trivial one. Informally, a sensible index should not prefer one candidate partition over another just because it has too large or too small clusters. Constant baseline is a particular focus of the current research. We develop a rigorous theoretical framework for analyzing this property. In this respect, our work improves over the previous (mostly empirical) research on constant baseline of particular indices (Strehl, 2002; Albatineh et al., 2006; Vinh et al., 2009; 2010; Lei et al., 2017).

While the ideas discussed in the paper can be applied to all similarity indices, we provide an additional theoretical characterization of pair-counting ones (e.g., Rand and Jaccard), which gives an analytical background for further studies of pair-counting indices. We formally prove that among dozens of known indices, only two have all the properties except for being a distance: Correlation Coefficient and Sokal & Sneath's first index (Lei et al., 2017). Surprisingly, both indices are rarely used for cluster evaluation. Correlation Coefficient has the additional advantage of being easily convertible to a distance measure via the arccosine function. The obtained index has all the properties except *constant baseline*, which is still satisfied asymptotically.

To sum up, our main contributions are the following:

- We formally define properties that are desirable across various applications. We analyze an extensive list of cluster similarity indices and mathematically prove or disprove all properties for each of them (Tables 3, 4).
- We provide a methodology for choosing a suitable validation index for a particular application. In particular, we identify previously overlooked indices that dominate the most popular ones (Section 5).
- We formalize the notion of constant baseline and provide a framework for its analysis; for pair-counting indices, we introduce the notion of *asymptotic constant baseline* (Section 4.6). We also provide a definition for monotonicity that unifies and extends previous attempts; for pair-counting indices, we introduce a strengthening of monotonicity (Section 4.5).

We believe that our unified and extensive analysis provides a useful tool for researchers and practitioners because research outcomes and application performances are highly dependent on the validation index that is chosen.

Comparison with prior work While there are previous attempts to analyze cluster similarity indices, our work unifies and significantly extends them. In particular, Lei et al. (2017) only consider biases of pair-counting indices, Meilă (2007) analyzes properties of Variation of Information, and Vinh et al. (2010) analyze information-theoretic indices.

Amigó et al. (2009) consider properties desirable for text clustering and mostly focus on monotonicity. Most importantly, Amigó et al. (2009) do not consider constant baseline (the absence of preference towards specific cluster sizes), which we found to be extremely important. In contrast, the problem of indices favoring clusterings with smaller or larger clusters has been identified by, e.g., Albatineh et al. (2006); Lei et al. (2017); Vinh et al. (2009; 2010). This problem is typically addressed by modifying a particular index (or family of indices) such that the obtained measure does not suffer from this problem. However, as we show in this paper, these modifications often lead to other important properties not being satisfied. We refer to Appendix A for a more detailed comparison to related research.

In the current paper, we introduce new properties, formalize existing ones, and mathematically prove or disprove each property for an extensive list of validation indices. This broader and more rigorous approach leads to conclusions that considerably differ from how validation indices are currently being chosen.

2. Cluster Similarity Indices

We consider clustering n elements numbered from 1 to n, so that a clustering can be represented by a partition of $\{1, \dots, n\}$ into disjoint subsets. Capital letters A, B, C will be used to name the clusterings, and we will represent them as $A = \{A_1, \dots, A_{k_A}\}$, where A_i is the set of elements belonging to i-th cluster. If a pair of elements $v, w \in V$ lie in the same cluster in A, we refer to them as an *intra-cluster pair* of A, while *inter-cluster pair* will be used otherwise. The total number of pairs is denoted by $N = \binom{n}{2}$. The value that an index V assigns to the similarity between partitions A and B will be denoted by $V(A, B)$. We now define some of the indices used throughout the paper. A more comprehensive list, together with formal definitions, is given in Appendices B.1, B.2.

Pair-counting indices consider clusterings to be similar if they agree on many pairs. Formally, let $\vec{A}$ be the N-dimensional vector indexed by the set of element-pairs, where the entry corresponding to (v, w) equals 1 if (v, w) is an intra-cluster pair and 0 otherwise. Let M_{AB} be the $N \times 2$ matrix that results from concatenating the two (column-) vectors $\vec{A}$ and $\vec{B}$. Each row of M_{AB} is either $11, 10, 01$, or 00. Let the pair-counts $N_{11}, N_{10}, N_{01}, N_{00}$ denote the number of occurrences for each of these rows in M_{AB}.

Definition 1. *A* pair-counting index *is a similarity index that can be expressed as a function of the pair-counts* $N_{11}, N_{10}, N_{01}, N_{00}$.

Some popular pair-counting indices are *Rand* and *Jaccard*:

$$\mathrm{R} = \frac{N_{11} + N_{00}}{N_{11} + N_{10} + N_{01} + N_{00}}, \quad \mathrm{J} = \frac{N_{11}}{N_{11} + N_{10} + N_{01}}.$$

Adjusted Rand (AR) is an adaptation of Rand ensuring that when B is random, we have $\mathrm{AR}(A, B) = 0$ in expectation. A less widely used index is the Pearson *Correlation Coefficient* (CC) between the binary incidence vectors $\vec{A}$ and $\vec{B}$.[1] Another index, which we discuss further in more details, is the *Correlation Distance* $\mathrm{CD}(A, B) := \frac{1}{\pi} \arccos \mathrm{CC}(A, B)$. In Appendix B.2, we formally define 27 known pair-counting indices and only mention those of particular interest throughout the main text.

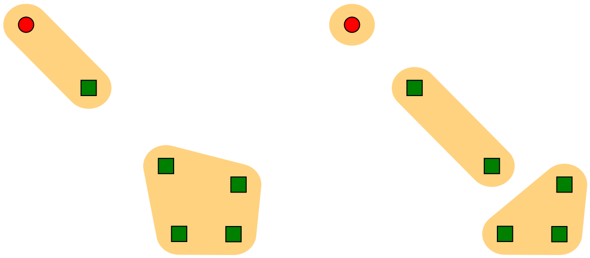
Information-theoretic indices consider clusterings similar if they share a lot of information, i.e., if little information is needed to transform one clustering into the other. Formally, let $H(A) := H(|A_1|/n, \dots, |A_{k_A}|/n)$ be the Shannon entropy of the cluster-label distribution of A. Similarly, the joint entropy $H(A, B)$ is defined as the entropy of the distribution with probabilities $(p_{ij})_{i \in [k_A], j \in [k_B]}$, where $p_{ij} = |A_i \cap B_j|/n$. Then, the mutual information of two clusterings can be defined as $M(A, B) = H(A) + H(B) - H(A, B)$. There are multiple ways of normalizing the mutual information:

$$\mathrm{NMI}(A, B) = \frac{M(A, B)}{(H(A) + H(B))/2},$$
$$\mathrm{NMI}_{\max}(A, B) = \frac{M(A, B)}{\max\{H(A), H(B)\}}.$$

NMI is known to be biased towards smaller clusters, and several modifications try to mitigate this bias: *Adjusted Mutual Information* (AMI) and *Standardized Mutual Information* (SMI) subtract the expected mutual information from $M(A, B)$ and normalize the obtained value (Vinh et al., 2009), while *Fair NMI* (FNMI) multiplies NMI by a penalty factor $e^{-|k_A - k_B|/k_A}$ (Amelio & Pizzuti, 2015).

3. Motivating Experiments

Evidently, many different cluster similarity indices are used by researchers and practitioners. A natural question is: *how to choose the best one?* Before trying to answer this question, it is important to understand whether the problem is relevant. Indeed, if the indices are very similar to each other and agree in most practical applications, then one can safely take any index. In this section, we demonstrate that this is not the case, and that the choice matters.



(a) FNMI, R, AR, J, D, W, FMeasure, BCubed

(b) NMI, $\mathrm{NMI}_{\max}$, VI, AMI, S&S, CC, CD

Figure 1. Inconsistency of indices: shapes denote the reference partition, captions indicate indices favoring the candidate.

Table 1. Inconsistency on real-world clustering datasets, %

	NMI	**VI**	**AR**	**S&S1**	**CC**
NMI	–	40.3	15.7	20.1	18.5
VI		–	37.6	36.0	37.2
AR			–	11.7	8.3
S&S1				–	3.6
CC					–

First, we illustrate the inconsistency of all indices. We say that two indices V_1 and V_2 are inconsistent for a triplet of partitions (A, B_1, B_2) if $V_1(A, B_1) > V_1(A, B_2)$ but $V_2(A, B_1) < V_2(A, B_2)$. We took 15 popular cluster similarity measures and constructed just four triplets such that each pair of indices is inconsistent for at least one triplet. One such triplet is shown in Figure 1: for this simple example, about half of the indices prefer the left candidate, while the others prefer the right one. Other examples can be found in Appendix F.1.

Thus, we see that the indices differ. But can this affect conclusions obtained in experiments on real data? To check that, we ran 8 well-known clustering algorithms (Scikit-learn, 2020) on 16 real-world datasets from the UCI machine learning repository (Dua & Graff, 2017). Each dataset, together with a pair of algorithms, gives a triplet of partitions (A, B_1, B_2), where A is a reference partition and B_1, B_2 are provided by two algorithms. For a given pair of indices and all such triplets, we look at whether the indices are consistent. Table 1 shows the relative inconsistency for several popular indices.[2] The inconsistency rate is significant: e.g., popular measures Adjusted Rand and Variation of Information disagree in almost 40% of the cases. Importantly, the best agreeing indices are S&S and CC, which satisfy most of our properties, as shown in the next section.

[1] Spearman and Pearson correlation are equal when comparing binary vectors. Kendall rank correlation for binary vectors coincides with the Hubert index that is linearly equivalent to Rand.

[2] The extended table together with a detailed description of the experimental setup and more analysis is given in Appendix F.2.

Table 2. Comparing algorithms according to different indices

	A_1	A_2
NMI	0.9479	**0.9482**
FNMI	**0.9304**	0.8722
AMI	**0.7815**	0.7533
VI	0.5662	**0.5503**
R	**0.9915**	0.9901
AR	0.5999	**0.6213**
J	0.4329	**0.4556**
S&S	0.8004	**0.8262**
CC	0.6004	**0.6371**

To demonstrate that the choice of similarity index may affect the final performance in a real production scenario, we conducted an experiment within a major news aggregator system. The system groups news articles to *events* and shows the list of most important events to users. For grouping, a clustering algorithm is used, and the quality of this algorithm affects the user experience: merging different clusters may lead to not showing an important event, while too much splitting may cause duplicate events. When comparing several candidate clustering algorithms, it is important to determine which one is the best for the system. Online experiments are expensive and can be used only for the best candidates. Thus, we need a tool for an offline comparison. For this purpose, we manually created a reference partition on a small fraction of news articles to evaluate the candidates. We performed such an offline comparison for two candidate algorithms A_1 and A_2 and observed that different indices preferred different algorithms (see Table 2). In particular, well-known FNMI, AMI, and Rand prefer A_1 that disagrees with most of the indices. Then, we launched an online user experiment and verified that the candidate A_2 is better for the system according to user preferences. This shows the importance of choosing the right index for offline comparisons. See Appendix F.3 for a more detailed description of this experiment.

4. Analysis of Cluster Similarity Indices

In this section, we motivate and formally define properties that are desirable for cluster similarity indices. We start with simple and intuitive ones that can be useful in some applications but not always necessary. Then, we discuss more complicated properties, ending with *constant baseline*, which is extremely important but least trivial. In Tables 3 and 4, indices of particular interest are listed along with the properties satisfied. In Appendix C, we give the proofs for all entries of these tables. For pair-counting indices we perform a more detailed analysis and define additional properties. For such indices, we interchangeably use the notation $V(A, B)$ and $V(N_{11}, N_{10}, N_{01}, N_{00})$.

Some of the indices have slight variants that are essentially the same. For example, the Hubert index (Hubert, 1977) is a linear transformation of the Rand index: $H = 2R - 1$. All the properties defined in this paper are invariant under linear transformations and interchanging A and B. Hence, we define the following linear equivalence relation on similarity indices and check the properties for at most one representative of each equivalence class.

Definition 2. *Similarity indices V_1 and V_2 are* linearly equivalent *if there exists a nonconstant linear function f such that either $V_1(A, B) = f(V_2(A, B))$ or $V_1(A, B) = f(V_2(B, A))$.*

This allows us to conveniently restrict to indices for which higher numerical values indicate higher similarity of partitions. Appendix Table 2 in lists equivalences among indices.

4.1. Property 1: Maximal Agreement

The numerical value that an index assigns to a similarity must be easily interpretable. In particular, it should be easy to see whether the candidate clustering is maximally similar to (i.e., coincides with) the reference clustering. Formally, we require that $V(A, A) = c_{\max}$ is constant and either a strict upper bound for $V(A, B)$ for all $A \neq B$. The equivalence from Definition 2 allows us to assume that $V(A, A)$ is a maximum w.l.o.g. This property is easy to check, and it is satisfied by almost all indices, except for SMI and Wallace.

Property 1′: Minimal Agreement The maximal agreement property makes the upper range of the index interpretable. Similarly, a numerical value for low agreement would make the lower range interpretable. A minimal agreement is not well defined for general partitions: it is unclear which partition is most dissimilar to a given one. However, by Lemma 1 in Appendix B.3, pair-counting indices form a subclass of graph similarity indices. For a graph with edge-set E, it is clear that the most dissimilar graph is its complement (i.e., with edge-set E^C). Comparing a graph to its complement results in pair-counts $N_{11} = N_{00} = 0$ and $N_{10} + N_{01} = N$. This motivates the following definition:

Definition 3. *A pair-counting index V has the* minimal agreement property *if there exists a constant $c_{\min}$ so that $V(N_{11}, N_{10}, N_{01}, N_{00}) \geq c_{\min}$ with equality if and only if $N_{11} = N_{00} = 0$.*

This property is satisfied by Rand, Correlation Coefficient, and Sokal&Sneath, while it is violated by Jaccard, Wallace, and Dice. Adjusted Rand does not have this property since substituting $N_{11} = N_{00} = 0$ gives the non-constant $\mathrm{AR}(0, N_{10}, N_{01}, 0) = -\frac{N_{10}N_{01}}{\frac{1}{2}N^2 - N_{10}N_{01}}$.

4.2. Property 3: Symmetry

Similarity is intuitively understood as a symmetric concept. Therefore, a good similarity index is expected to be symmetric, i.e., $V(A, B) = V(B, A)$ for all partitions A, B.[3] Tables 3 and 4 show that most indices are symmetric. The asymmetric ones are precision and recall (Wallace) and FNMI (Amelio & Pizzuti, 2015), which is a product of NMI and an asymmetric penalty factor.

4.3. Property 4: Linear Complexity

For clustering tasks on large datasets, running time is crucial, and algorithms with superlinear time can be infeasible. In these cases, a validation index with superlinear running time would be a significant bottleneck. Furthermore, computationally heavy indices also tend to be complicated and hard to interpret intuitively. We say that an index has *linear complexity* when its worst-case running time is $O(n)$. In Appendix C.2, we prove that any pair-counting index has $O(n)$ complexity. Many general indices have this property as well, except for SMI and AMI.

4.4. Property 4. Distance

For some applications, a distance-interpretation of dissimilarity may be desirable: whenever A is similar to B and B is similar to C, then A should also be somewhat similar to C. For example, assume that the reference clustering (e.g., labeled by experts) is an approximation of the ground truth. In such situations, it may be reasonable to argue that the reference clustering is at most a distance ε from the true one, so that the triangle inequality bounds the dissimilarity of the candidate clustering to the unknown true clustering.

A function d is a distance metric if it satisfies three distance axioms: 1) symmetry ($d(A, B) = d(B, A)$); 2) positive-definiteness ($d(A, B) \geq 0$ with equality iff $A = B$); 3) the triangle inequality ($d(A, C) \leq d(A, B) + d(B, C)$). We say that V is linearly transformable to a distance metric if there exists a linearly equivalent index that satisfies these three distance axioms. Note that all three axioms are invariant under rescaling of d. We have already imposed symmetry as a separate property, and positive-definiteness is equivalent to the maximal agreement property. Therefore, whenever V has these two properties, it satisfies the distance property iff $d(A, B) = c_{\max} - V(A, B)$ satisfies the triangle inequality, for $c_{\max}$ as defined in Section 4.1.

Examples of popular indices having this property are Variation of Information and the Mirkin metric. In Vinh et al. (2010), it is proved that when Mutual Information is normalized by the maximum of entropies, the resulting NMI is equivalent to a distance metric. A proof that the Jaccard index is equivalent to a distance is given in Kosub (2019). See Appendix C.1 for all the proofs.

Correlation Distance Among all the considered indices, there are two pair-counting ones having all the properties except for being a distance: Sokal&Sneath and Correlation Coefficient. However, the correlation coefficient can be transformed to a distance metric via a non-linear transformation. We define Correlation Distance (CD) as $\mathrm{CD}(A, B) := \frac{1}{\pi} \arccos \mathrm{CC}(A, B)$, where CC is the Pearson correlation coefficient and the factor $1/\pi$ scales the index to $[0, 1]$. To the best of our knowledge, this Correlation Distance has never before been used as a similarity index for comparing clusterings throughout the literature.

Theorem 1. *The Correlation Distance is indeed a distance.*

Proof. A proof of this is given in (Van Dongen & Enright, 2012). We give an alternative proof that allows for a geometric interpretation. First, we map each partition A to an N-dimensional vector on the unit sphere by

$$\vec{u}(A) := \begin{cases} \frac{1}{\sqrt{N}}\mathbf{1} & \text{if } k_A = 1, \\ \frac{\vec{A} - \frac{m_A}{N}\mathbf{1}}{\|\vec{A} - \frac{m_A}{N}\mathbf{1}\|} & \text{if } 1 < k_A < n, \\ -\frac{1}{\sqrt{N}}\mathbf{1} & \text{if } k_A = n, \end{cases}$$

where $\mathbf{1}$ is the N-dimensional all-one vector, $\vec{A}$ is the binary vector representation of a partition introduced in Section 2, and $m_A = N_{11} + N_{10}$ is the number of intra-community pairs of A. Straightforward computation gives $\|\vec{A} - \frac{m_A}{N}\mathbf{1}\| = \sqrt{m_A(N - m_A)/N}$, and standard inner product

$$\begin{aligned} \langle \vec{A} - \tfrac{m_A}{N}\mathbf{1}, \vec{B} - \tfrac{m_B}{N}\mathbf{1} \rangle &= N_{11} - \frac{m_A m_B}{N} \\ &= \frac{N_{11}N_{00} - N_{10}N_{01}}{N}, \end{aligned}$$

so that the inner product indeed corresponds to CC:

$$\begin{aligned} \langle \vec{u}(A), \vec{u}(B) \rangle &= \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{m_A(N - m_A)m_B(N - m_B)}} \\ &= \mathrm{CC}(A, B). \end{aligned}$$

It is a well-known fact that the inner product of two vectors of unit length corresponds to the cosine of their angle. Hence, taking the arccosine gives us the angle. The angle between unit vectors corresponds to the distance along the unit hypersphere. As $\vec{u}$ is an injection from the set of partitions to points on the unit sphere, we may conclude that this index is indeed a distance on the set of partitions. □

In Section 4.6, we show that the distance property of Correlation Distance is achieved at the cost of not having the exact constant baseline, though it is still satisfied asymptotically.

[3] In some applications, A and B may have different roles (e.g., reference and candidate partitions), and an asymmetric index may be suitable if there are different consequences of making false positives or false negatives.

Table 3. Requirements for general similarity indices

	Max. agreement	Symmetry	Distance	Lin. complexity	Monotonicity	Const. baseline
NMI	✓	✓	✗	✓	✓	✗
NMI$_{\max}$	✓	✓	✓	✓	✗	✗
FNMI	✓	✗	✗	✓	✗	✗
VI	✓	✓	✓	✓	✓	✗
SMI	✗	✓	✗	✗	✗	✓
FMeasure	✓	✓	✗	✓	✗	✗
BCubed	✓	✓	✗	✓	✓	✗
AMI	✓	✓	✗	✗	✓	✓

Table 4. Requirements for pair-counting indices[4]

	Max. agreement	Min. agreement	Symmetry	Distance	Lin. complexity	Monotonicity	Strong monotonicity	Const. baseline	As. const. baseline	Type of bias
R	✓	✓	✓	✓	✓	✓	✓	✗	✗	↗↘
AR	✓	✗	✓	✗	✓	✓	✗	✓	✓	
J	✓	✗	✓	✓	✓	✓	✗	✗	✗	↘
W	✗	✗	✗	✗	✓	✗	✗	✗	✗	↘
D	✓	✗	✓	✗	✓	✓	✗	✗	✗	↘
CC	✓	✓	✓	✗	✓	✓	✓	✓	✓	
S&S1	✓	✓	✓	✗	✓	✓	✓	✓	✓	
CD	✓	✓	✓	✓	✓	✓	✓	✗	✓	

4.5. Property 5: Monotonicity

When one clustering is changed such that it resembles the other clustering more, the similarity score ought to improve. Hence, we require an index to be monotone w.r.t. changes that increase the similarity. This can be formalized via the following definition.

Definition 4. *For clusterings A and B, we say that B' is an A*-consistent improvement of *B iff $B \neq B'$ and all pairs of elements agreeing in A and B also agree in A and B'.*

This leads to the following monotonicity property.

Definition 5. *An index V satisfies the* monotonicity property *if for every two clusterings A, B with $1 < k_A < n$ and any B' that is an A-consistent improvement of B, it holds that $V(A, B') > V(A, B)$ and $V(B', A) > V(B, A)$.*

The trivial cases $k_A = 1$ and $k_A = n$ were excluded to avoid inconsistencies with the constant baseline property defined in Section 4.6. To look at monotonicity from a different perspective, we define the following operations:

- **Perfect split**: B' is a perfect split of B (w.r.t. A) if B' is obtained from B by splitting a single cluster B_1 into two clusters B'_1, B'_2 such that no two elements of the same cluster of A are in different parts of this split, i.e., for all i, $A_i \cap B_1$ is a subset of either B'_1 or B'_2.
- **Perfect merge**: We say that B' is a perfect merge of B (w.r.t. A) if there exists some A_i and $B_1, B_2 \subset A_i$ such that B' is obtained by merging B_1, B_2 into B'_1.

The following theorem gives an alternative definition of monotonicity and is proven in Appendix E.1.

Theorem 2. *B' is an A-consistent improvement of B iff B' can be obtained from B by a sequence of perfect splits and perfect merges.*

Note that this monotonicity is a stronger form of the first two constraints defined in (Amigó et al., 2009): *Cluster Homogeneity* is a weaker form of our monotonicity w.r.t. perfect splits, while *Cluster Equivalence* is equivalent to our monotonicity w.r.t. perfect merges.

Monotonicity is a critical property that should be satisfied by any sensible index. Surprisingly, not all indices satisfy this: we have found counterexamples that prove that SMI, FNMI, and Wallace do not have the monotonicity property. Furthermore, for NMI, whether monotonicity is satisfied depends on the normalization: the normalization by the average of the entropies has monotonicity, while the normalization by the maximum of the entropies does not.

Property 5′. Strong Monotonicity For pair-counting indices, we can define a stronger monotonicity property in terms of pair-counts.

Definition 6. *A pair-counting index V satisfies* strong monotonicity *if it is increasing in N_{11}, N_{00} when $N_{10} + N_{01} > 0$, and decreasing in N_{10}, N_{01} when $N_{11} + N_{00} > 0$.*

Note that the conditions $N_{10} + N_{01} > 0$ and $N_{11} + N_{00} > 0$ are needed to avoid contradicting maximal and minimal agreement respectively. This property is stronger than monotonicity as it additionally allows for comparing similarities across different settings: we could compare the similarity between A_1, B_1 on n_1 elements with the similarity between A_2, B_2 on n_2 elements, even when $n_1 \neq n_2$. This ability to compare similarity scores across different numbers of elements is similar to the *Few data points* property of SMI (Romano et al., 2014) that allows its scale to have a similar interpretation across different settings.

We found several examples of indices that have Property 5 while not satisfying Property 5′. Jaccard and Dice indices are constant w.r.t. N_{00}, so they are not strongly monotone.

[4] All known pair-counting indices excluded from this table do not satisfy either constant baseline, symmetry, or maximal agreement.

A more interesting example is the Adjusted Rand index, which may become strictly larger if we only increase N_{10}.

4.6. Property 6. Constant Baseline

This property is arguably the most significant: it is less intuitive than the other ones and may lead to unexpected consequences in practice. Informally, a good similarity index should not give a preference to a candidate clustering B over another clustering C just because B has many or few clusters. This intuition can be formalized using random partitions: assume that we have some reference clustering A and two random partitions B and C. While intuitively both random guesses are equally bad approximations of A, it has been known throughout the literature (Albatineh et al., 2006; Vinh et al., 2009; 2010; Romano et al., 2014) that some indices tend to give higher scores for random guesses with a larger number of clusters. Ideally, we want the similarity value of a random candidate w.r.t. the reference partition to have a fixed expected value c_{base} (independent of A or the sizes of B). However, this does require a careful formalization of random candidates.

Definition 7. *We say that a distribution over clusterings $\mathcal{B}$ is* element-symmetric *if for every two clusterings B and B' that have the same cluster-sizes, $\mathcal{B}$ returns B and B' with equal probabilities.*

This allows us to define the constant baseline property.

Definition 8. *An index V satisfies the* constant baseline property *if there exists a constant c_{base} so that, for any clustering A with $1 < k_A < n$ and every element-symmetric distribution $\mathcal{B}$, it holds that $\mathbf{E}_{B\sim\mathcal{B}}[V(A,B)] = c_{base}$.*

In the definition, we have excluded the cases where A is a trivial clustering consisting of either 1 or n clusters. Including them would cause contradictions with maximal agreement whenever we choose $\mathcal{B}$ as the (element-symmetric) distribution that returns A with probability 1. In Appendix D.1, we prove that to verify whether an index satisfies Definition 8, it suffices to check whether it holds for distributions $\mathcal{B}$ that are uniform over clusterings with fixed cluster sizes. From this equivalence, it will also follow that Definition 8 is indeed symmetric. Note that the formulation in terms of element-symmetric distributions allows for a wide range of clustering distributions. For example, the cluster sizes could be drawn from a power-law distribution, which is often observed in practice (Arenas et al., 2004; Clauset et al., 2004).

Constant baseline is extremely important in many practical applications: if an index violates this property, then its optimization may lead to undesirably biased results. For instance, if a biased index is used to choose the best algorithm among several candidates, then it is likely that the decision will be biased towards those who produce too large or too small clusters. This problem is often attributed to NMI (Vinh et al., 2009; Romano et al., 2014), but we found that almost all indices suffer from it. The only indices that satisfy the constant baseline property are Adjusted Rand index, Correlation Coefficient, SMI, and AMI with $c_{\text{base}} = 0$ and Sokal&Sneath with $c_{\text{base}} = {}^1\!/_2$. Interestingly, out of these five indices, three were *specifically designed* to satisfy this property, which made them less intuitive and resulted in other important properties being violated.

The only condition under which the constant baseline property can be safely ignored is knowing in advance *all cluster sizes*. In this case, bias towards particular cluster sizes would not affect decisions. However, we are not aware of any practical application where such an assumption can be made. Note that knowing only the number of clusters is insufficient. We illustrate this in Appendix D.4, where we also show that the bias of indices violating the constant baseline is easy to identify empirically.

Property 6′: Asymptotic Constant Baseline For pair-counting indices, a deeper analysis of the constant baseline property is possible. Let $m_A = N_{11} + N_{10}$, $m_B = N_{11} + N_{01}$ be the number of intra-cluster pairs of A and B, respectively. If the distribution $\mathcal{B}$ is uniform over clusterings with given sizes, then m_A and m_B are both constant. Furthermore, the pair-counts N_{10}, N_{01}, N_{00} are functions of N, m_A, m_B, N_{11}. Hence, to find the expected value of the index, we need to inspect it as a function of a single random variable N_{11}. For a random pair, the probability that it is an intra-cluster pair of both clusterings is $m_A m_B/N^2$, so the expected values of the pair-counts are

$$\begin{aligned} \overline{N_{11}} &:= \frac{m_A m_B}{N}, & \overline{N_{10}} &:= m_A - \overline{N_{11}}, \\ \overline{N_{01}} &:= m_B - \overline{N_{11}}, & \overline{N_{00}} &:= N - m_A - m_B + \overline{N_{11}}. \end{aligned} \tag{1}$$

We can use these values to define a weaker variant of constant baseline.

Definition 9. *A pair-counting index V has an* asymptotic constant baseline *if there exists a constant c_{base} so that $V\left(\overline{N_{11}}, \overline{N_{10}}, \overline{N_{01}}, \overline{N_{00}}\right) = c_{base}$ for all $m_A, m_B \in (0, N)$.*

In contrast to Definition 8, asymptotic constant baseline is very easy to verify: one can substitute the values from (1) to the index and check whether the obtained value is constant. Another important observation is that under mild assumptions $V(N_{11}, N_{10}, N_{01}, N_{00})$ converges in probability to $V\left(\overline{N_{11}}, \overline{N_{10}}, \overline{N_{01}}, \overline{N_{00}}\right)$ as n grows which justifies the usage of the name *asymptotic constant baseline*, see Appendix D.2 for more details.

Note that the non-linear transformation of Correlation Coefficient to Correlation Distance makes the latter one violate the constant baseline property. CD does, however, still have the asymptotic constant baseline at $1/2$ and we prove in Appendix E.2 that the expectation in Definition 8 is very close to this value.[5]

Biases of Cluster Similarity Indices Given the fact that there are so many biased indices, one may be interested in what kind of candidates they favor. While it is unclear how to formalize this concept for general validation indices, we can do this for pair-counting ones by analyzing them in terms of a single variable: the number of *inter-cluster pairs*. This value characterizes the granularity of a clustering: it is high when the clustering consists of many small clusters while it is low if it consists of a few large clusters.

Informally, we say that an index suffers from *PairDec* bias if it may favor less inter-cluster pairs. Similarly, *PairInc* bias means that an index may prefer more inter-cluster pairs. These biases can be formalized as follows.

Definition 10. *Let V be a pair-counting index and define $V^{(s)}(m_A, m_B) = V\left(\overline{N_{11}}, \overline{N_{10}}, \overline{N_{01}}, \overline{N_{00}}\right)$ for the expected pair-counts as defined in (1). We say that*

(i) V suffers from PairDec bias *if there are $m_A, m_B \in (0, N)$ such that $\frac{d}{dm_B}V^{(s)}(m_A, m_B) > 0$;*

(ii) V suffers from PairInc bias *if there are $m_A, m_B \in (0, N)$ such that $\frac{d}{dm_B}V^{(s)}(m_A, m_B) < 0$.*

Note that this definition does require $V^{(s)}$ to be differentiable in m_A and m_B. However, this is the case for all pair-counting indices in this work. Applying this definition to Jaccard $J^{(s)}(m_A, m_B) = \frac{m_A m_B}{N(m_A+m_B)-m_A m_B}$ and Rand $R^{(s)}(m_A, m_B) = 1 - (m_A + m_B)/N + 2m_A m_B/N^2$ immediately shows that Jaccard suffers from PairDec bias and Rand suffers from both biases. The direction of the monotonicity for the bias of Rand is determined by the condition $2m_A > N$. Performing the same for Wallace and Dice shows that both suffer from PairDec bias. Note that an index satisfying the asymptotic constant baseline property will not have any of these biases as $V^{(s)}(m_A, m_B) = c_{\text{base}}$.

While there have been previous attempts to characterize types of biases (Lei et al., 2017), they mostly rely on analyses based on the number of clusters. However, our analysis shows that the number of clusters is not the correct variable for such a characterization of pair-counting indices. While having many clusters often goes hand-in-hand with having many inter-cluster pairs, it is not always the case: if there are significant differences between the cluster sizes (e.g., one large cluster and many small clusters), then the clustering may consist of many clusters while having relatively few inter-cluster pairs. We discuss this in more detail in Appendix E.3. Additionally, Experiments shown in Figures 1 and 2 of the Appendix show that in such cases, most indices have a similar bias as if there were few clusters, which is consistent with our characterization of such biases in terms of the number of inter-cluster pairs.

5. Discussion and Conclusion

At this point, we better understand the theoretical properties of cluster similarity indices, so it is time to answer the question: *which index is the best?* Unfortunately, there is no simple answer, but we can make an informed decision. In this section, we sum up what we have learned, argue that there are indices that are strictly better alternatives than some widely used ones, and give practical advice on how to choose a suitable index for a given application.

Among all properties discussed in this paper, *monotonicity* is the most crucial one. Violating this property is a fatal problem: such indices can prefer candidates which are strictly worse than others. Hence, we advise against using the well-known NMI$_{max}$, FMeasure, FNMI, and SMI indices.

The *constant baseline* property is much less trivial but is equally important: it addresses the problem of preferring some partitions only because they have small or large clusters. This property is essential unless you know *all cluster sizes*. Since we are not aware of practical applications where all cluster sizes are known, we assume below that this is not the case.[6] This requirement is satisfied by just a few indices, so we are only left with AMI, Adjusted Rand (AR), Correlation Coefficient (CC), and Sokal&Sneath (S&S). Additionally, Correlation Distance (CD) satisfies constant baseline asymptotically and deviations from the exact constant baseline are extremely small (see Appendix E.2).

Let us note that among the remaining indices, AR is strictly dominated by CC and S&S since it does not have the minimum agreement and strong monotonicity. Also, similarly to AMI, AR is specifically created to have a constant baseline, which made this index more complex and less intuitive than other pair-counting indices. Hence, we are only left with four indices: AMI, S&S, CC, and CD.

According to their theoretical properties, all these indices are good, and any of them can be chosen. Figure 2 illustrates how a final decision can be made. First, one can decide whether the distance property is needed. For example, suppose one wants to cluster the algorithms by comparing the partitions provided by them. If one would want to use a metric clustering algorithm for this, the index would have to be a distance. In this case, CD would be the best choice. If the distance property is not needed, one could base the decision

[5] There is also another transformation of CC to a distance $\text{CD}' = \sqrt{2(1 - \text{CC})}$. However, it can be shown that CD$'$ approximates a constant baseline less well than CD.

[6] However, in applications where such an assumption holds, it can be reasonable to use, e.g., BCubed, Variation of Information, and NMI.

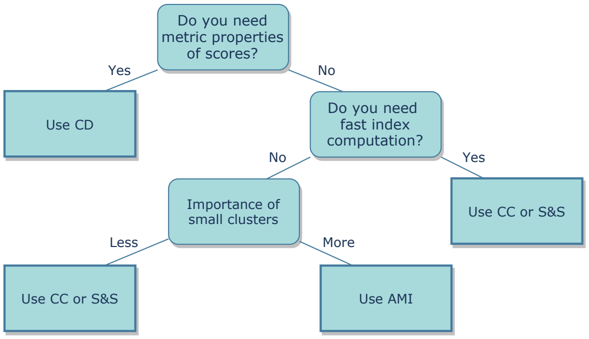

Do you need metric properties of scores?
Yes
No
Use CD
Do you need fast index computation?
No
Yes
Importance of small clusters
Use CC or S&S
Less
More
Use CC or S&S
Use AMI

Figure 2. Example of how one can make a decision among good cluster similarity indices.

on computational complexity. In many large-scale applications, using clustering algorithms with higher than linear running time is infeasible. Understandably, it is undesirable if the computation of a validation score takes longer than the actual clustering algorithm. Another example is multiple comparisons: choosing the best algorithm among many candidates (differing, e.g., by a parameter value). If fast computation is required, then AMI is not a proper choice, and one has to choose between CC and S&S. Otherwise, all three indices are suitable according to our formal constraints.

Let us discuss an (informal) criterion that may help to choose between AMI and pair-counting alternatives. Different indices may favor a different balance between errors in small and large clusters. In particular, all pair-counting indices give larger weights to errors in large clusters: misclassifying one element in a cluster of size k costs $k - 1$ incorrect pairs. It is known (empirically) that information-theoretic indices do not have this property and give a higher weight to small clusters (Amigó et al., 2009).[7] Amigó et al. (2009) argue that for their particular application (text clustering), it is desirable not to give a higher weight to large clusters. In contrast, there are applications where the opposite may hold. For instance, consider a system that groups user photos based on identity and shows these clusters to a user as a ranked list. In this case, a user is likely to investigate the largest clusters consisting of known people and would rarely spot an error in a small cluster. The same applies to any system that ranks the clusters, e.g., to news aggregators. Based on what is desirable for a particular application, one can choose between AMI and pair-counting CC and S&S.

The final decision between CC and S&S is hard to make since they are equally good in terms of their theoretical properties. Interestingly, although some works (Choi et al., 2010; Lei et al., 2017) list Pearson correlation as a cluster similarity index, it has not received attention that our results suggest it deserves, similarly to S&S. First, both indices are interpretable. CC is a correlation between the two incidence vectors, which is a very natural concept. S&S is the average of precision, recall (for binary classification of pairs) and their inverted counterparts, which can also be intuitively understood. Also, CC and S&S usually agree in practice: in Tables 1 and 3 we can see that they have the largest agreement. Hence, one can take any of these indices. Another option would be to check whether there are situations where these indices disagree and, if this happens, perform an experiment similar to what we did in Section 3 for news aggregation. While some properties listed in Tables 3 and 4 are not mentioned in the discussion above, they can be important for particular applications. For instance, maximum and minimum agreements are useful for interpretability, but they can also be essential if some operations are performed over the index values: e.g., averaging the scores of different algorithms. Symmetry can be necessary if there is no "gold standard" partition, but algorithms are compared only to each other.

Finally, let us remark that in an early version of this paper, we conjectured that the constant baseline and distance properties are mutually exclusive. This turns out to be true: in ongoing work, we prove an impossibility theorem: for pair-counting indices *monotonicity*, *distance*, and *constant baseline* cannot be simultaneously satisfied.

Acknowledgements

Most of this work was done while Martijn Gösgens was visiting Yandex and Moscow Institute of Physics and Technology (Russia). The work of Martijn Gösgens is supported by the Netherlands Organisation for Scientific Research (NWO) through the Gravitation NETWORKS grant no. 024.002.003. The work of Liudmila Prokhorenkova is supported by the Ministry of Education and Science of the Russian Federation in the framework of MegaGrant 075-15-2019-1926 and by the Russian President grant supporting leading scientific schools of the Russian Federation NSh-2540.2020.1. Furthermore, the authors would like to thank Nelly Litvak and Remco van der Hofstad for their helpful feedback and guidance throughout this project. We also thank Borislav Kozlovskii for the help with the news aggregator experiment.

[7] This is an interesting aspect that has not received much attention in our research since we believe that the desired balance between large and small clusters may differ per application and we are not aware of a proper formalization of this "level of balance" in a general form.

References

Albatineh, A. N., Niewiadomska-Bugaj, M., and Mihalko, D. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

Amelio, A. and Pizzuti, C. Is normalized mutual information a fair measure for comparing community detection methods? In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 1584–1585, 2015.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486, 2009.

Arenas, A., Danon, L., Diaz-Guilera, A., Gleiser, P. M., and Guimera, R. Community analysis in social networks. *The European Physical Journal B*, 38(2):373–380, 2004.

Choi, S.-S., Cha, S.-H., and Tappert, C. C. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

Clauset, A., Newman, M. E., and Moore, C. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Hubert, L. Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, 30(1):98–103, 1977.

Kosub, S. A note on the triangle inequality for the jaccard distance. *Pattern Recognition Letters*, 120:36–38, 2019.

Lei, Y., Bezdek, J. C., Romano, S., Vinh, N. X., Chan, J., and Bailey, J. Ground truth bias in external cluster validity indices. *Pattern Recognition*, 65:58–70, 2017.

Meilă, M. Comparing clusteringsan information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. *Physical review E*, 69 (2):026113, 2004.

Romano, S., Bailey, J., Nguyen, V., and Verspoor, K. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *International Conference on Machine Learning*, pp. 1143–1151, 2014.

Scikit-learn. Clustering algorithms. https://scikit-learn.org/stable/modules/clustering.html, 2020.

Strehl, A. *Relationship-based clustering and cluster ensembles for high-dimensional data mining*. PhD thesis, 2002.

Van Dongen, S. and Enright, A. J. Metric distances derived from cosine similarity and pearson and spearman correlations. *arXiv preprint arXiv:1208.3145*, 2012.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080, 2009.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

Virtanen, S. and Girolami, M. Precision-recall balanced topic modelling. In *Advances in Neural Information Processing Systems*, pp. 6750–6759, 2019.

Wang, Z., Zheng, L., Li, Y., and Wang, S. Linkage based face clustering via graph convolution network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1117–1125, 2019.

Wibowo, W. and Williams, H. E. Strategies for minimising errors in hierarchical web categorisation. In *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 525–531, 2002.

Xu, D. and Tian, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.