Set Matching Measures for External Cluster Validity

Mohammad Rezaei, Pasi Fränti, *Senior Member, IEEE*

Abstract— Comparing two clustering results of a data set is a challenging task in cluster analysis. Many external validity measures have been proposed in the literature. A good measure should be invariant to the changes of data size, cluster size and number of clusters. We give an overview of existing set matching indexes and analyze their properties. Set matching measures are based on matching clusters from two clusterings. We analyze the measures in three parts: 1. cluster similarity 2. matching 3. overall measurement. Correction for chance is also investigated and we prove that normalized mutual information and variation of information are intrinsically corrected. We propose a new scheme of experiments based on synthetic data for evaluation of an external validity index. Accordingly, popular external indexes are evaluated and compared when applied to clusterings of different data size, cluster size and number of clusters. The experiments show that set matching measures are clearly better than the other tested. Based on the analytical comparisons, we introduce a new index called Pair Sets Index (PSI).

Index Terms— Clustering, External validity index, Cluster validation, Comparing clusterings, Normalization, correction for chance, adjustment for chance

1 INTRODUCTION

As a basic tool, *clustering* or cluster analysis partitions a set of unlabeled data objects into meaningful groups. A huge number of clustering techniques have been developed in different application fields [1]. Different algorithms or even one algorithm with different parameters can result in different partitions for the same data set. A question therefore arises that which partition best fits with the data set. Cluster validity indexes have been commonly used to address this problem [2], [3], [4], [5], [6], [7], [8], [9]. They are classified into *internal* and *external indexes* of which the former are based on information intrinsic to data while the latter measure the similarity between two clustering results of one data set. We focus on external validity indexes in this paper.

External validity indexes are used actively in searching for good clustering solutions, for example in ensemble clustering [10], [11], [12], [13], where the goal is to aggregate a set of clustering partitions. They have been used in genetic algorithms [14] to measure genetic diversity in a population. In [11], external indexes are used for comparing the results of multiple runs to study the stability of k-means. To evaluate internal validity indexes, a framework is introduced in [15] by using external indexes on ground-truth partition. Using these indexes we can identify those algorithms that generate similar partitions irrespective of data [1]. The indexes can also be used for determining the number of clusters for a data set [16], [17], [18].

External validity indexes measure how well the results of a clustering match the ground truth (if available) or another clustering [19], [20]. Several external validation measures have been studied in [7], [8], [9], [19], [20], [21], [22]. They can be categorized into *pair-counting*, *information theoretic* and *set matching* measures.

Pair-counting measures include *rand index*, *adjusted rand index*, *Jaccard coefficient, Fowlkes-Mallows index and several others* [9], [23]. They are based on counting the pairs of objects in the data set on which two different partitions agree or disagree. For instance, if two objects in one cluster in the first partition place also in the same cluster in the second partition, it is considered as an agreement. Most of the existing external validity indexes are classified in this group.

Information theoretic indexes such as *entropy*, *Mutual Information* and *variation of information* have also been used in comparing clusterings [9], [24], [25]. Mutual information measures the information that two clusterings share. Since there is no upper bound for mutual information, normalization is needed for easier interpretation and comparison [10]. A systematic study of this group of indexes, including several existing popular measures and recently proposed ones has been performed in [9].

Set matching indexes such as *F measure* [26], *criterion H* [27] and *Van Dongen* [28] are based on pairing similar clusters in two partitions. According to [24], existing indexes in this group suffer from the problem that clusters having no pair are not involved in comparison. The unmatched part of two paired clusters is also not taken into account. Taking use of the tight connection between partitions and centroids, cluster-level similarity indexes such as *Centroid Index* [20] and *Centroid Ratio* [29] employ the representatives of the clusters instead of point-level partitions. However, cluster-level indexes lack point-level information.

- *M. Rezaei is with the School of Computing, University of Eastern Finland, E-mail: rezaei@ cs.uef.fi.*
- *P. Fränti is with the School of Computing, University of Eastern Finland, FI 80110, E-mail: franti@cs.uef.fi.*

Comparison of different external validity indexes regarding to their properties have been reported in [7], [8], [9], [21], [24], [26]. Normalization and correction for chance, as desirable properties, keep the range of an index fixed in [-1, 1] or [0, 1] and make the index values comparable across different data sets. More specifically, correction for chance adjusts the index for randomness by transforming its expected value to zero. The importance of index normalization on data with imbalanced cluster distribution is discussed in [7], [26]. It is shown that the values of normalized measures are more spread in [0, 1], and have a wider range than unnormalized ones. According to [9] and [21], correction for chance is preferable when the number of data points is relatively small compared with the number of clusters. Other properties include sensitivity of an index to data size, cluster size imbalance and number of clusters. The effect of cluster size imbalance on a range of external validity indexes is analyzed in [26] and it is shown that normalization should be applied. Otherwise, an index is mostly affected by big clusters and does not detect changes in small clusters. Metric properties have been also discussed for external validity indexes and several researchers prefer metric because of the theoretical properties that exist on metric spaces [9], [21], [22], [24].

In this paper, we study set matching validity indexes by introducing and analyzing three components of the indexes: cluster similarity, matching and overall measurement. We also investigate correction for chance and show that normalized mutual information, variation of information and their adjusted forms are equivalent. We propose a new similarity index called Pair Sets Index (PSI) according to careful analysis and comparisons. Simplified form of PSI is also shown to be metric. Another contribution of the paper is to propose a new way of experiments for evaluating external indexes. The behavior of an index in comparison of clusterings with cluster size imbalance, different data size and number of clusters is extracted and analyzed systematically. We show by these experiments that set matching indexes clearly outperform other popular indexes.

2 Problem Definition

Given a data set $X \in R^d$ with N objects in a d-dimensional space, the problem of clustering is to group the data set into K clusters [14]. Given two sets of partitions $P=\{P_1, P_2,\ldots,P_K\}$ of K clusters and $G=\{G_1, G_2,\ldots,G_{K'}\}$ of K' clusters, an external validity index measures the similarity between P and G. A contingency table of P and G is a matrix where n_{ij} is the number of objects that are both in clusters P_i and G_j: $n_{ij}=|P_i \cap G_j|$, see Table 1. The sizes of clusters P_i and G_j are n_i and m_j, respectively.

An external validity index needs to satisfy several properties to be consistent and comparable for different data sets and clusterings structures.

Normalization transforms the index within a fixed range, for example [0, 1], which makes the comparison easier for data sets with different size and structure. Normalization is the most commonly agreed property in the clustering community [9]. To transform a dissimilarity index I_d to the range of [0, 1], normalization is performed as:

$$I_d^n(P,G) = \frac{I_d - \min(I_d)}{\max(I_d) - \min(I_d)} \tag{1}$$

where $min(I_d)$ and $max(I_d)$ are the minimum and maximum values of I_d.

TABLE 1
Contingency Table for Two Clustering P and G

	G_1	G_2	...	G_j	...	$G_{K'}$	Σ
P_1	n_{11}	n_{12}	...	n_{1j}	...	$n_{1K'}$	n_1
P_2	n_{21}	n_{22}	...	n_{2j}	...	$n_{2K'}$	n_2
...	...	...	...	...	...	...	...
P_i	n_{i1}	n_{i2}	...	n_{ij}	...	$n_{iK'}$	n_i
...	...	...	...	...	...	...	...
P_K	n_{K1}	n_{K2}	...	n_{Kj}	...	$n_{KK'}$	n_K
Σ	m_1	m_2	...	m_j	...	$m_{K'}$	N

The index values are expected to be constant when different random clusterings are compared with a ground truth [30]. A random partition is created by selecting random number of clusters of random size. The similarity between the random partition and the ground truth originates merely by chance. Take an example of rand index: the value of the index for two random partitions is not a constant, and is in a narrow range of [0.5, 1] instead of [0, 1]. By *correction for chance* or *adjustment*, the expected value of a similarity index is transformed to zero [21], [30]. Adjustment and normalization can be performed jointly as follows:

$$\text{Dissimilarity:} \quad I_d^{adj}(P,G) = \frac{I_d - \min(I_d)}{E(I_d) - \min(I_d)}$$
$$\text{Similarity:} \quad I_s^{adj}(P,G) = \frac{I_s - E(I_s)}{\max(I_s) - E(I_s)} \tag{2}$$

where the minimum of a similarity index (maximum of a dissimilarity index) is estimated by expected value $E(I_s)$.

Metric property has been also considered. Although a similarity / dissimilarity measure can be effective without being a metric [31], it is sometimes preferred. Considering dissimilarity index I_d and partitions P_1, P_2 and P_3, the metric properties require [22], [32]:

1. Non-negativity: $I_d(P_1,P_2) \geq 0$
2. Reflexivity: $I_d(P_1,P_2)=0$ if and only if $P_1=P_2$
3. Symmetry: $I_d(P_1,P_2)=I_d(P_2,P_1)$
4. Triangular inequality: $I_d(P_1,P_2)+I_d(P_2,P_3) \geq I_d(P_1,P_3)$

A similarity metric satisfies the following [32]:

1. Limited Range: $I_s(P_1,P_2) \leq I_0 < \infty$
2. Reflexivity: $I_s(P_1,P_2)= I_0$ if and only if $P_1=P_2$
3. Symmetry: $I_s(P_1,P_2)=I_s(P_2,P_1)$
4. Triangular inequality:
$$I_s(P_1,P_2) \times I_s(P_2,P_3) \leq I_s(P_1,P_3) \times (I_s(P_1,P_2)+I_s(P_2,P_3))$$

The triangular inequality for a similarity index I_s is derived according to the corresponding inequality for a dissimilarity index which is defined as c / I_s (c>0). However, other forms of the inequality are possible by defining other dissimilarities such as $\max(I_s)-I_s$. It is trivial to show

that if c/I_s (or $\max(I_s)-I_s$) is a dissimilarity metric, I_s is a similarity metric as well [32]. Hence, the metric properties for a similarity index can be checked for its corresponding dissimilarity.

Cluster size imbalance signifies that a data set can include clusters with big difference in their sizes. Some researchers argue that clusters with bigger sizes have more importance than smaller ones but in this paper we assume that each cluster has the same importance independent of its size. Invariance on the size of clusters is therefore another desired property of an index. Size of the data set should not affect on the index either.

An index should be independent on the number of clusters. Some indexes such as Rand Index give higher similarity for partitions with more clusters [22]. The index should also be applicable for comparing two clusterings with different number of clusters.

Monotonicity is another needed property. It states that the similarity of two clusterings monotonically decreases as their difference increases.

Once the above desired properties are met, then it ensures that the index values for different data sets are on the same scale and comparable. For instance, if an index gives 90% and 70% similarities, 90% should represent higher similarity. However, this is true only if the index is independent on data set and its clustering structure.

3 Pair-Counting and Information Theoretic Indexes

Pair-counting measures count the pairs of points on which the two clusterings agree or disagree. Four values are defined: *a* represents the number of pairs that are in the same cluster both in *P* and *G*; *b* represents the number of pairs that are in the same cluster in P but in different clusters in *G*; *c* represents the number of pairs that are in different clusters in *P* but in the same cluster in *G*; *d* represents the number of pairs that are in different clusters both in *P* and *G*. Values *a* and *d* count the agreements while *b* and *c* the disagreements. Examples of each case are illustrated in Fig. 1. The values of *a*, *b*, *c* and *d* can be calculated from the contingency table [30] as follows:

$$
\begin{aligned}
a &= \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1) \\
b &= \frac{1}{2}\Big(\sum_{j=1}^{K'} m_j^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\Big) \\
c &= \frac{1}{2}\Big(\sum_{i=1}^{K} n_i^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\Big) \\
d &= \frac{1}{2}\Big(N^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \Big(\sum_{i=1}^{K} n_i^2 + \sum_{j=1}^{K'} m_j^2\Big)\Big)
\end{aligned}
\tag{3}
$$

Some of the popular indexes are listed in Table 2. *Rand index* (RI) is a well-known pair-counting measure. For random partitions, the similarity between two clusterings is desired to be close to zero. However, the expected value of rand index for random partitions is 0.5 and the index is within a narrow range of [0.5, 1] according to [11], [12], [30]. Hence, a corrected-for-chance version called *adjusted rand index* (ARI) was introduced in [30] which is upper bounded by one and lower bounded by zero. The expected value of the rand index is estimated using hyper-geometric distribution assumption in which the size and number of clusters are fixed [30].

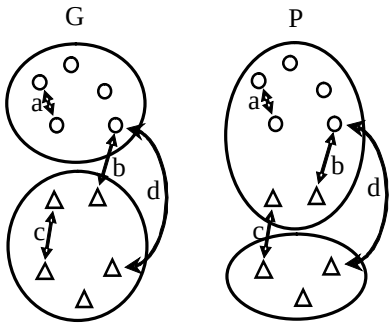


Fig. 1. The principle of pair-counting measures.

Existing information theoretic measures employ the concept of entropy [25] to compare two partitions. Entropy is measured by the average number of bits needed to store or communicate data. The entropy of clustering *P* with *K* clusters is defined as:

$$H(P) = -\sum_{i=1}^{K} p(P_i)\log p(P_i) \tag{4}$$

where $p(P_i)=n_i / N$ is the estimated probability of the cluster P_i.

Having clustering *G* and the joint distribution $p(P,G)$, the average number of bits for *P* is derived by conditional entropy [19] as follows:

$$H(P|G) = \sum_{i=1}^{K}\sum_{j=1}^{K'} p(P_i,G_j)\log p(P_i|G_j) \tag{5}$$

where the probability $p(P_i,G_j)$ can be estimated from the contingency table as n_{ij}/N.

Mutual information (MI) [9], [10] is derived from conditional entropy and represents the similarity of two clusterings [22]. If we choose a random object in the data set, knowing its cluster in *G*, mutual information measures the reduction in uncertainty of the object's cluster in *P* [22], [24]. Mutual information is defined formally as follows:

$$MI(P,G) = H(P) - H(P|G) = H(P) + H(G) - H(P,G) \tag{6}$$

In terms of probabilities, it is:

$$MI(P,G) = \sum_{i=1}^{K}\sum_{j=1}^{K'} p(P_i,G_j)\log\frac{p(P_i,G_j)}{p(P_i)p(G_j)} \tag{7}$$

Variation of Information (VI) [24] is complement of the mutual information, see Fig. 2, and is calculated by summing up the conditional entropies $H(P|G)$ and $H(G|P)$, see (8). Normalization of MI and VI is discussed in section 5.

$$
\begin{aligned}
VI(P,G) &= H(P|G) + H(G|P) = \\
&H(P) + H(G) - 2MI(P,G) = \\
&2H(P,G) - H(P) - H(G)
\end{aligned}
\tag{8}
$$

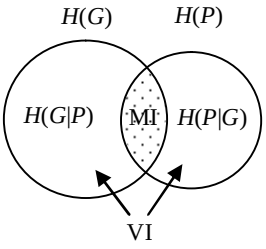


Fig. 2. Mutual information (MI) and variation of information(VI).

4 Set Matching Indexes

Set matching indexes are based on matching entire clusters. Similar clusters are first found either by pairing or matching, and their similarity is then measured using set matching methods. We classify the set matching indexes into two types: point-level and cluster-level.

Point-level indexes consider the intersection of paired clusters in two clusterings. Purity is an example of this group and it assumes one of the clusterings as ground truth [33]. Accuracy defined in [34] is equivalent (exactly the same) to Purity. Some authors use terms such as classification accuracy [35] or classification error [9] with refereeing to accuracy in [34] but this is not correct because they have other definitions in classification problem. F measure (FM) [26], Criterion H (CH) [27] and normalized Van Dongen (NVD) [28] are other set matching measures.

Cluster-level indexes include Centroid Index (CI) [20] and Centroid Ratio (CR) [29]. They use only cluster prototypes in contrast to point-level indexes which employ the labels of all objects in resulting partitions. Cluster level indexes are fast to calculate [20], and they provide clear interpretation about the differences in cluster-level structure. For example, CI=1, demonstrates one difference in the global allocation of the two clusterings. However, they do not measure partial cluster differences. Centroid Similarity Index (CSI) was introduced in [20] to extend CI to a point-level measure.

Set matching measures involve three design questions:

1. How to match the clusters
2. How to measure the similarity of two clusters
3. How to calculate overall similarity

Normalization and correction for chance (if applied) are also essential parts of the overall similarity derivation. We next give a detailed analysis of all these questions including the normalization.

1. *Similarity of two clusters*

Let P_i and G_j be two clusters in P and G respectively. Most of the set matching measures use $|P_i \cap G_j|$ to calculate the similarity of the two sets. For example, in Fig. 5, clusters G_1 and P_1 are more similar than G_2 and P_2 since the number of shared objects is 6 and 4 respectively. CH, NVD, CSI and Purity use this measure. Many other ways to measure similarity of two sets exist in literature and any of them can be employed for calculating the similarity of two clusters. Among the 76 methods listed in [36], we mention three popular ones: Jaccard [37], Sorensen-Dice [38] and Braun-Banquet [36].

TABLE 2
External Validity Indexes

Pair-counting measures	
Rand index [30]	$RI = \frac{a+d}{N(N-1)/2}$
Adjusted rand index [30]	$ARI = \frac{RI - E(RI)}{1 - E(RI)}$
Information theoretic measures	
Mutual information [25]	$MI = \sum_{i=1}^{K}\sum_{j=1}^{K'} \frac{n_{ij}}{N} \log \frac{N n_{ij}}{n_i m_j}$
Normalized Mutual Information type 1 [25]	$NMI = \frac{MI(P,G)}{(H(P)+H(G))/2}$
Normalized Mutual Information type 2 [25]	$NMI = \frac{MI(P,G)}{\sqrt{H(P)\times H(G)}}$
Normalized Variation of Information [7]	$NVI = \frac{H(P)+H(G)-2MI(P,G)}{H(P)+H(G)}$
Set matching measures	
F measure [26]	$FM = \frac{1}{N}\sum_{i=1}^{K} n_i \max_j \frac{2n_{ij}}{n_i + m_j}$
Criterion H [27]	$H = 1 - \frac{1}{N}\max_j \sum_{i=1}^{K} n_{ij}$
Normalized Van Dongen [28]	$NVD = \frac{2N - \sum_{i=1}^{K}\max_{j=1}^{K'} n_{ij} - \sum_{j=1}^{K'}\max_{i=1}^{K} n_{ij}}{2N}$
Purity [33]	$Purity = \frac{1}{N}\sum_{i=1}^{K} \max_{\pi} n_{i,\pi}(i)$
CI [20]	$CI_1(P,G) = \sum_{i=1}^{K'} orphan(G_i)$ $CI_2(P,G) = \max(CI_1(P,G), CI_1(G,P))$
CSI [20]	$CSI = \frac{\sum_{i=1}^{K} n_{ij} + \sum_{j=1}^{K'} n_{ji}}{2N}$ i, j: indexes of matched clusters
CR [29]	$CR = 1 - \sum_{i=1}^{K} \gamma_i / K$ $\gamma_i = \begin{cases} 1 & unstable\ pair \\ 0 & stable\ pair \end{cases}$
PSI	$\begin{cases} \frac{S - E(S)}{\max(K,K') - E(S)} & S \geq E(S), \max(K,K') > 1 \\ 0 & S < E(S) \\ 1 & K = K' = 1 \end{cases}$ $S = \sum_{i=1}^{\min(K,K')} \frac{n_{ij}}{\max(n_i, m_j)}$ i, j: indexes of paired clusters

$$J = \frac{|P_i \cap G_j|}{|P_i \cup G_j|} \quad (9)$$

$$SD = \frac{2|P_i \cap G_j|}{|P_i| + |G_j|} \quad (10)$$

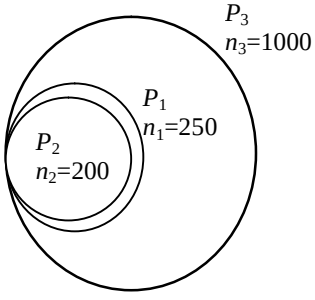
$$BB = \frac{|P_i \cap G_j|}{\max(|P_i|, |G_j|)} \quad (11)$$

These measures are in the range of [0, 1]. Distance forms of J and SD are defined as (1-J) and (1-SD) where the former is a true metric but the latter does not satisfy triangular inequality. In order to make the measure independent on cluster size, these measures normalize the number of shared objects $|P_i \cap G_j|$ according to the size of clusters in three different ways.

For example, consider the three clusters in Fig. 3 where we want to find out the more similar cluster to P_1 from P_2 or P_3. Similarity of P_1 and P_2 should be much higher than the similarity of P_1 and P_3 even though P_1 and P_3 share more objects. J, SD and BB give more intuitive similarity values than intersection. When comparing P_1 and P_3, the similarity 0.25 of J and BB is better than the 0.4 of SD. It is trivial to show that J≤BB≤SD for any two sets.

FM [22] uses *precision* and *recall* concepts by measuring n_{ij}/n_i and n_{ij}/m_j respectively. The criterion 2×precision×recall/(precision+recall) would be equivalent to SD but it avoids the normalization by cluster size using n_i×SD instead of SD.

Cluster-level indexes provide binary result (0 or 1), indicating whether the clusters have 1:1 match (CI), or the pair of clusters is unstable (CR). Table 3 lists the criteria for set matching indexes.



	P_1, P_2	P_1, P_3
J	0.80	0.25
SD	0.89	0.40
BB	0.80	0.25

Fig. 3. The effect of cluster size on cluster comparison

2. *Matching*

For every cluster, we need to find the pair to which the similarity is measured. Three cases are considered: 1. optimal pairing 2. greedy pairing 3. matching. Matching is performed based on nearest neighbor mapping so that any cluster in *P* is matched to a cluster in *G* with maximal similarity. Several clusters can be matched with the same cluster in the other clustering. Pairing is a special case of matching in which clusters are only allowed to be matched once. FM, NVD, Purity, CI and CSI employ matching whereas CH and CR use greedy pairing. We will use optimal pairing.

TABLE 3
CRITERIA FOR SIMILARITY OF TWO CLUSTERS

	Similarity criteria
FM	$\|P_i\|$×SD
H	$\|P_i \cap G_j\|$
NVD	$\|P_i \cap G_j\|$
Purity	$\|P_i \cap G_j\|$
PSI	BB
CI	0/1 (mapped or unmapped)
CSI	$\|P_i \cap G_j\|$
CR	0/1 (stable or unstable)

Matching results, in general, is not symmetric when finding pairs for clusters of *P* from *G* and vise versa. To make the index symmetric, the similarity results in both directions are usually combined, see NVD, CI and CSI equations in Table 2. FM and Purity assume that we compare a clustering against ground truth and they therefore consider matching in one direction only. Matching criterion in NVD and Purity is the number of shared objects; CI and CSI are based on similarity of prototypes.

Pairing problem, however, is not trivial to solve and different algorithms have been proposed to find approximate or optimal solution. The pairing can be seen as a matching problem in weighted bipartite graph where the nodes represent the clusters, see Fig. 4. Greedy pairing is mostly used with time complexity of $O(N^2)$. Two most similar clusters are iteratively matched and excluded. Instead of greedy pairing, we apply here Hungarian algorithm which finds the optimal solution with time complexity $O(N^3)$ where *N* is the maximum number of clusters in *P* and *G*.

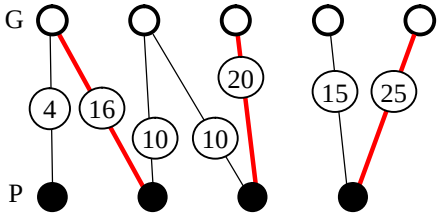


Fig. 4. Pairing clusters to maximize overall similarity. The thick lines show the optimal pairing where overall similarity according to number of shared objects would be (25+20+16)=61.

Fig. 5 demonstrates the matching from *G* to *P* based on the number of shared objects where P_2 remains unmatched. The matching from *P* to *G* will be different and the same as greedy pairing based on number of shared objects, resulting to (P_1,G_1), (P_2,G_2) and (P_3, G_3).

Fig. 6 shows matching in CI when there is different number of clusters. We assume that the objects are in 2-*D* Euclidean space; the centroids have been shown with crosses signs. In matching *P* to *G*, one orphan centroid is produced that indicates one difference in global allocation. NVD results the same matching as CI in this example. In general, if a cluster P_i has more shared objects with G_j than G_k, the probability that its centroid is also closer to

G_j is higher. Although, this is not always true as it depends on the distribution of data among clusters. It anyway implies that the matching using intersection criterion and centroid distance are expected to produce the same result.

Fig. 7 demonstrates the results with too few (above) and too many clusters (below) compared to another with the same clustering problem or to the correct clustering. In this example, both matching and pairing are performed

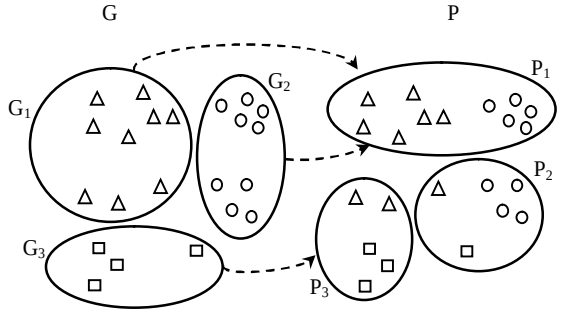


Fig. 5. Matching clusters based on maximum shared objects. Cluster P_2 remains unmatched. In pairing process of CH, G_2 is paired with P_2 after excluding G_1 and P_1 as the first pair.

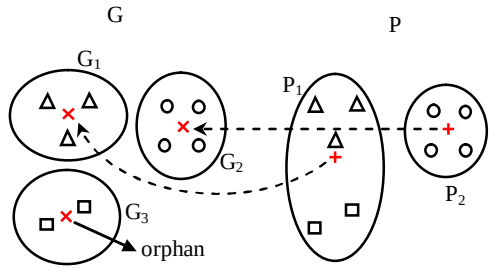


Fig. 6. Matching centroids from *P* to *G* based on nearest neighbor mapping used in CI and CSI; One orphan centroids shows one difference in global allocation.

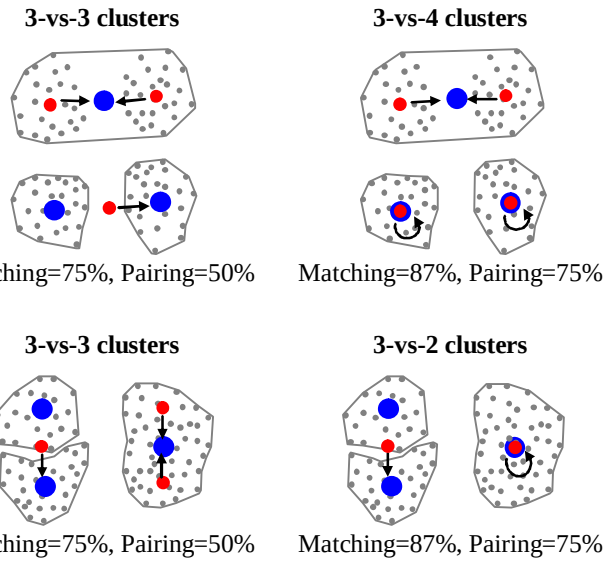


Fig. 7. Matching and pairing when too few (above) and too many (below) clusters exist. Arrows show matching from red to blue centroids. Pairing would use only part of those arrows because each cluster can be matched only once.

TABLE 4
SUMMARIZATION OF MATCHING METHODS OF INDEXES

	Pairing/ Matching	Matching criterion	Algorithm
FM	Matching	SD	One-way
CH	Pairing	$\|P_i \cap G_j\|$	Greedy
NVD	Matching	$\|P_i \cap G_j\|$	Two-way
Purity	Matching	$\|P_i \cap G_j\|$	One-way
PSI	Pairing	BB	Optimal
CI	Matching	Centroid distance	Two-way
CSI	Matching	Centroid distance	Two-way
CR	Pairing	Centroid distance	Greedy

based on number of shared objects. Matching results always higher values than pairing because in pairing some centroids remain unpaired. Pairing is more sensitive to differences in clustering structure. The result is also lower with 3-vs-3 than when comparing to the correct number of clusters (3-vs-4 and 3-vs-2). In comparing two clustering with different number of clusters, unpaired clusters indicate a disagreement on the number of clusters, which is an advantage of pairing. Table 4 summarizes the matching methods for several indexes.

3. *Overall similarity*

Overall similarity is obtained by summing up the similarities of all the matched clusters. The upper bound of overall similarity for CH is *N* (total number of objects) which is used for normalization, see Table 2. To remove the asymmetry effect of matching, NVD and CSI use 2*N* because of two-way matching, see Table 2. If we define the distance form of CSI and Purity as (1-CSI) and (1-Purity), NVD, CH, Purity and CSI are all equivalent if their matching results are the same. In fact, if matching in NVD and CSI is symmetric (*K*=*K*'), they would equal to CH and we can write:

$$NVD = 1 - \frac{\sum_{i=1}^{K} n_{ij} + \sum_{j=1}^{K'} n_{ji}}{2N} = 1 - \frac{2\sum_{i=1}^{K} n_{ij}}{2N} = 1 - \frac{\sum_{i=1}^{K} n_{ij}}{N} = CH = 1 - Purity = 1 - CSI \quad (12)$$

The overall dissimilarity of CI equals the number of zero mapped centroids of *G*. Since CI is not symmetric, CI_2 is defined as max(CI(*P*,*G*), CI(*G*,*P*)) [20]. Centroid index represents the number of differences in global allocations and it is in the range of [0, *K*-1] where *K* is the maximum number of clusters in the two clusterings. At least one non-zero mapped centroid exists and the upper bound therefore becomes *K*-1.

Centroid ratio (CR) defines the concept of (un)stable centroids. Consider a paired centroid C_i and C'_j with distance D_{ij} from clusterings *P* and *G*, respectively. Assume that the distances of C_i to the nearest centroid in *P* and C'_j to the nearest centroid in *G* are D_i and D_j. Then, if $D_{ij}^2 / (D_i \times D_j) > 1$, the pair is considered unstable. The overall similarity is defined based on the number of unstable

pairs [29], see Table 2. Table 5 summarizes the overall similarity derivation for the above mentioned indexes.

TABLE 5
OVERALL SIMILARITY DERIVATION

	Total summation	Range	Normalization
FM	similarity of matched clusters	[0, 1]	N
CH	Shared objects	[0, 1]	N
NVD	Shared objects in both directions	[0, 1]	$2N$
Purity	Shared objects in one direction	[0, 1]	N
PSI	Normalized similarity of paired clusters	[0, 1]	K
CI	Orphan clusters	[0, K-1]	-
CSI	Shared objects in both directions	[0, 1]	$2N$
CR	Unstable clusters	[0, 1]	K

5 CORRECTION FOR CHANCE

Normalization makes comparisons easier for different data sets. Correction for chance removes the similarity of two clusterings which merely originates by chance [21].

An index is normalized using its lower and upper bounds as in (1). Correction for chance can be jointly performed with normalization according to (2). Some indexes do not have fixed lower or upper bounds. For example, several upper bounds have been proposed to normalize MI [9], [25].

In comparison of two clusterings *P* and *G*, the number and size of clusters are known. To consider the effect of random partitioning, the objects of clustering *P* are distributed randomly in clusters of *G* and the expected similarity value is calculated. This is called hyper-geometric distribution assumption and was first used for deriving ARI [30].

The measures in the pair-counting class as listed in [23] are in the ranges of [0,1], [-1, 1], [0.5, 1] or [-0.25, 0.25] that further clarifies the necessity of normalization. Since all the indexes are defined based on values *a*, *b*, *c*, and *d* in (3), the upper and lower bounds are simple to derive. Many of them become equivalent after applying correction for chance [21]. ARI is the most well-known and widely used index of this group [9].

In set matching measures, the overall similarity is derived either by summing up the number of shared objects or the similarities of the matched clusters. For example, NVD, CH, Purity and CSI sum up the number of shared objects and use the total number of objects for normalization. The similarity index proposed by Larsen and Aone [39] is calculated by summing up the normalized similarities (in the range of [0, 1]) of the matched clusters. In this case, the overall similarity is normalized for each cluster individually.

Both MI and VI are metric but they are not bounded to a fixed range [22]. Mutual information of clusterings *P* and *G* is lower bounded by zero. Geometric or arithmetic mean of entropies as an upper bound can be an option for normalization (type 1 and 2 in Table 2) [22], [25], [10]. In [25] $min(H(P), H(G))$ and $max(H(P), H(G))$ are also used for normalization. An upper bound for VI is $H(P)+H(G)$, which means that clusterings *P* and *G* do not share any information [7]. The upper bound can therefore be used for normalization of VI. To derive adjusted mutual information according to (2), obtaining the expected value $E(\mathrm{MI})$ is the key issue. An analytical formula for the expected value of mutual information is derived in [21] under the assumptions of hyper-geometric model of randomness. In [9], upper bounds for the expected value are given, and shown that, under certain assumptions, the adjusted MI measures derived based on different upper bounds become equivalent to the normalized MI measures.

We prove next that the adjusted forms of mutual information (AMI) and variation of information (AVI_s) are equivalent to their normalized forms (NMI, NVI_s) when the summation of the entropies $H(P)+H(G)$ is used for normalization.

Theorem 1. *Under hyper-geometric distribution assumption:*

$$AVI_s = NVI_s = AMI = NMI \tag{13}$$

where NVI_s and AVI_s denote the similarity form of NVI and AVI (1-NVI and 1-AVI) respectively.

Proof. See Appendix A.

6 PAIR SETS INDEX

In this section, we present a new set matching based measure called Pair Sets Index (PSI), which is designed so that the properties discussed in section 2 are all satisfied. The components of the proposed index are known but some of them are new in this context, and the overall combination is novel. In specific, PSI contains optimal pairing of the clusters (new), set matching measure using BB (new), the overall similarity measure in (14) (used also by CR), and the correction for chance (used by pair-counting and information theoretic methods only).

6.1 Similarity Measure

Given clusterings *P* and *G*, the first step is to find the pairs of clusters in two partitions. Pairing clusters in *P* and *G* is done by maximizing total similarity which is defined as:

$$S(P,G) = \sum_i S_{ij} \tag{14}$$

where S_{ij} denotes the similarity between clusters P_i and G_j and is calculated as from Braun-Banquet formula [36] as follows:

$$S_{ij} = \frac{n_{ij}}{\max(|P_i|,|G_j|)} \tag{15}$$

Here n_{ij} is the number of shared objects in the two clusters and $|P_i|$ and $|G_j|$ denote their sizes.

The corresponding distance variant is defined as $D_{ij}=1-S_{ij}$. Pairing clusters is solved as an *assignment problem* in a

bipartite graph, see Fig. 4, by minimizing the total distance. We use Hungarian algorithm to find the perfect matching in this assignment problem [40].

6.2 Correction for Chance

In this section, we describe the process of correction for chance for the proposed similarity measure in (14) and derivation of the final formula for the Pair Sets Index (PSI).

Obtaining the expected value is the key point to derive the adjusted version of the index. To derive the expected value, consider a random shuffling of P as P' under hyper-geometric distribution assumption where the number and size of the clusters in P' and P are the same. The objects of cluster G_j are distributed randomly in the clusters in P'. A larger cluster in P' gets more objects from G_j. Therefore, the number of shared objects of clusters G_j and P'_i is proportional to the size of P'_i. The number of objects of G_j (m_j) that places in P'_i (n_i) is $m_j \times (n_i / N)$, which is the number of shared objects between these two clusters when random partition P' is assumed.

Theorem 2. The maximum total similarity in (14) is achieved when the largest cluster in P' is paired with the largest one in G, and recursively the same applies to the rest of the clusters. Applying this greedy pairing, the expected value is:

$$E = \sum_{i=1}^{\min(K,K')} \frac{m_i \times (n_i / N)}{\max(m_i, n_i)} \tag{16}$$

where the size of clusters in P' is $n_1 > n_2 > \ldots > n_K$ and in G is $m_1 > m_2 > \ldots > m_{K'}$.

Proof. See Appendix B.

Next, we show that $E \leq 1$. Assuming $m_i = n_i, \forall i$, the summation in (16) is $(n_1+n_2+\ldots+n_{Kmin}) / N \leq 1$. Suppose that $n_i \geq m_i$, $\exists i$, the summation then becomes:

$$E=(n_1+n_2+\ldots+m_i+\ldots+n_{Kmin})/N$$
$$\leq (n_1+n_2+\ldots+n_{Kmin})/N \leq 1 \tag{17}$$

Therefore, it is always true that $E \leq 1$. Applying the results to (2), the adjusted index becomes:

$$PSI = \begin{cases} \dfrac{S-E}{\max(K,K')-E} & S \geq E, \max(K,K') > 1 \\ 0 & S < E \\ 1 & K = K' = 1 \end{cases} \tag{18}$$

where S is the total similarity from (14). In random partitioning S=E, PSI=0 and in a perfect match S=K, PSI=1. If there is a disagreement on the number of clusters, K≠K', max(K, K') is taken in (18) to achieve a lower similarity that reflects the disagreement. The expected value is not necessarily the minimum value of the similarity. If S<E, we consider PSI=0 because this case corresponds to a very low agreement of the two partitions.

Distance variant of PSI is defined as 1-PSI:

$$PSI_d = \begin{cases} \dfrac{\max(K,K')-S}{\max(K,K')-E} & S \geq E, \max(K,K') > 1 \\ 1 & S < E \\ 0 & K = K' = 1 \end{cases} \tag{19}$$

Value of E depends on the similarity between the structures of two clusterings in terms of number and size of clusters. If the structures are close to each other, $E \approx 1$. Accordingly, simplified variant of PSI is defined by assuming $E=1$:

$$PSI^* = \begin{cases} \dfrac{S-1}{\max(K,K')-1} & S \geq 1, \max(K,K') > 1 \\ 0 & S < 1 \\ 1 & K = K' = 1 \end{cases} \tag{20}$$

6.3 Metric Properties of PSI

The proposed index is normalized in the range of [0, 1] and corrected for chance. In this section, we prove metric property of the distance form of PSI in (19).

Nonnegative: In (19), where $S \geq 1$, $max(K, K') > 1$, since $E \leq 1$, $max(K, K')-E$ is always larger than or equal to 1. The total similarity S equals to $max(K, K')$ only in a perfect match. In all other situations it is less than $max(K, K')$, hence $max(K, K')-S \geq 0$ holds. Therefore, it is true that:

$$PSI_d \geq 0 \tag{21}$$

Symmetric: The similarity of two clusters according to (15) is symmetric. The pairs of clusters are found according to the maximum matching which does not depend on whether we compare P to G or vice versa. Therefore, the total similarity in (14) is symmetric. To derive the expected value of the similarity in (16), we take two largest clusters in P and G as the best match. This action is also independent on the direction of the comparison. According to (18), when the similarity S and its expected value E are symmetric, the whole index is also symmetric:

$$PSI_d(P,G) = PSI_d(G,P) \tag{22}$$

Reflexivity: If $P=G$, the total similarity according to (14) and (15) is $max(K,K')=K$, and therefore $PSI_d=0$. On the other hand, if $PSI_d=0$, it follows that $S=max(K, K')$. This may happen only if the number of clusters is the same and the similarity of every two paired clusters according to (15) is 1. The similarity of two clusters is 1 if and only if they are exactly the same. Therefore, all clusters in P and G must be equal, and accordingly, $P=G$:

$$PSI_d(P,G) = 0 \ \ if\ and\ only\ if \ \ P = G \tag{23}$$

Triangular inequality: In Appendix C, we prove the triangular inequality for the simplified form of PSI in (20). The simplified form is therefore proven to be metric. Experiments for clustering with different structures indicate that the triangular inequality in most cases holds for the original form of PSI as well. However, the term E in the denumerator in (18) makes it difficult to prove in general.

6.4 Other Properties

a) Normalized to the number of clusters

The proposed validity index has low dependency on the number of clusters and this dependency decreases as the number of clusters increases. In (18), the similarity is normalized by $max(K, K')-E$. Because of E, the index is not independent on the number of clusters. However, since $E \leq 1$ and when $max(K, K')$ increases, the impact of E de-

creases.

b) Imbalanced clusters

One important advantage of the proposed index is its independency on the size of clusters because each cluster, either small or large, has equal impact on the similarity value. For example, suppose that in two clusterings, there are two perfect pairs where in one pair the clusters are large and in the other one they are small. Both of them increase the total similarity by the same amount.

7 Experiments

We next evaluate the external validity indexes based on their performance on partitions. To investigate different properties of an index, a variety of partitions should be considered. We provide comparisons with artificially generated partitions to demonstrate whether an index meets the required properties. We also study the effect of dimensionality and cluster overlap.

7.1 Selected Indexes and Artificial Partitions

We compare the proposed index to the state-of-art external indexes. Since all adjusted indexes in the pair-counting group behave similar [23], we use only ARI as the most popular one. Variation of information and mutual information are two representing measures in the information theoretic group. Since NVI_s=AVI_s=NMI=AMI, only NMI is used in the experiments. The performance of arithmetic and geometric mean for normalization of NMI is the same, we therefore employ arithmetic mean only. The normalized Van Dongen criterion, Criterion H and Purity are chosen in the set matching group. The matching in Centroid similarity index depends on the centroids, and therefore, we need real datasets to calculate centroids. However, as we discussed in section 4.2, the results of matching is most likely similar to NVD. We therefore use this assumption in the following, and in these experiments NVD=CSI.

In the test setup, we consider a ground-truth partition *G*, for example with 3000 objects, 1000 objects in each cluster, see Fig. 8, where light grey, grey and black represent the three clusters. In practice, we make an array of the length 3000 with values 1, 2 and 3 representing cluster labels of data. In this case, the first 1000 objects (light grey) have value 1. The partition *P* to be compared with is varied in different ways. The order of the data objects in the two partitions remains the same.

Fig. 8. Two partitions with 3000 objects.

The partitions in the experiment are considered in several aspects: random partitions, the impact of cluster size imbalance, number of clusters and consistency when the error increases in the partitions.

7.2 Random Partitions

Consider partition *P* which consists of random labels as shown in Fig. 9. We conduct experiments for different number of clusters from *K*=1 to 20 in *P*. The indexes NMI, ARI and PSI give values close to zero independent on the number of clusters. The values of the other three indexes are not zero because they are not corrected for chance, see Fig. 10. Normalized mutual information gives zero in this case which shows that NMI has the same performance as adjusted mutual information. This result further verifies our claim in (13).

Fig. 9. Clustering *P* represents a random partition with two clusters.

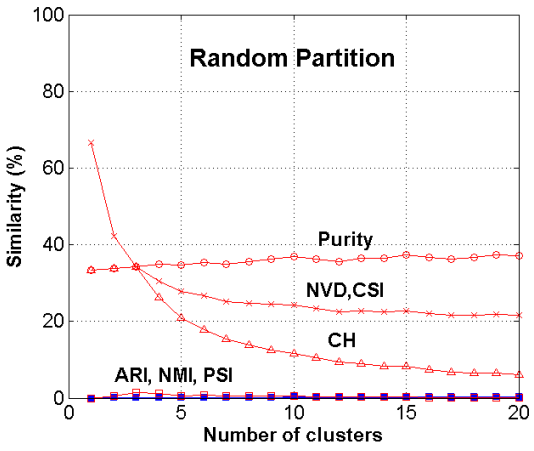


Fig. 10. Random partitioning with different number of clusters in *P* from *K*=1 to 20.

7.3 Monotonicity

We change the partition *P* linearly in three ways and study the response of the indexes.

First we enlarge the first (light grey) cluster in *P* in steps of 50 objects until only one cluster remains, see Fig. 11. Second, we enlarge the grey cluster in the same way, see Fig. 13, and third, we change part of the labels in all clusters of *P* and keep the cluster sizes unchanged, see Fig. 15. In Fig.12, NMI, ARI and NVD have very clear knee points when the light grey cluster reaches 2000 objects because at this point the number of clusters decreases by 1. For NMI and ARI, the index values increase when the cluster size approaches to 2000. In this situation, there are still three clusters and the results indicate that NMI and ARI ignore relatively small clusters and put more weights on large clusters. When the light grey cluster size is 2000, there is a local maximum when the number of clusters changes from three to two. NVD is constant between 1500 to 2000, and 2500 to 3000. The asymmetric matching of clusters in *NVD* causes the problem. Suppose that the size of the grey cluster (*x*) in *P* is less than 500. After matching *P* to *G*, the number of shared objects is 1000+*x*+1000 and *G* to *P* where both light grey and grey clusters in *G* are matched with the light grey one in *P*, the

number of shared objects is 1000+(1000-*x*)+1000. Summing up, the number of shared objects in two directions is 5000 which is independent of *x*. Therefore, when the size of the first cluster is between 1500 and 2000, the similarity remains a constant 5000 / 6000=0.83.

The proposed PSI has near linear dependency on the size of the light grey cluster. The indexes CH and Purity have good linear behavior but including an offset by 33% because they are not corrected for chance. If we made them corrected, the same issues as with the other indexes would appear. Note that Purity does not compare two clusterings in both directions. If we compare *G* to *P* instead of *P* to *G*, the results is different and without linear behavior.

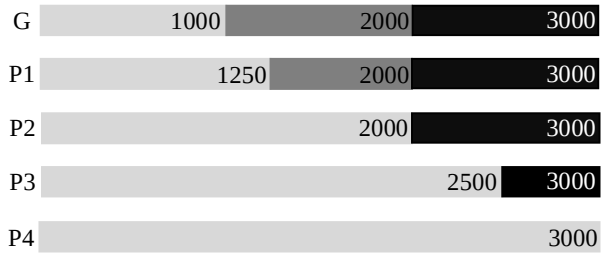


Fig. 11. Enlarging the first (light grey) cluster in steps of 50 objects by moving the objects from the other two clusters.

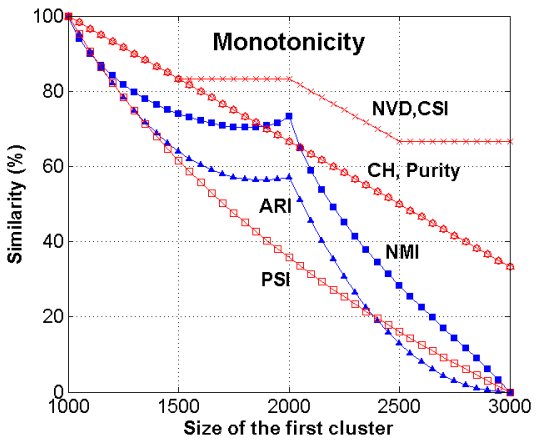


Fig. 12. Increasing the size of the first cluster.

We repeat the experiment by enlarging the size of the second cluster. The difference to the previous case is that the number of clusters remains 3 until the second cluster contains all the objects. The results in Fig. 14 show better performance for NMI and ARI compared to the previous case. The reason is that this time there is no change in the number of clusters in *P*. The same arguments for NVD, CH and CA are valid as for the previous case. The knee point for NVD is where the size of the biggest cluster becomes more than 2000 (compare P_2 and *G* in Fig. 13) and all three clusters of *G* are matched to the grey cluster of *P*. Interesting observation is that PSI results the same curves in both of the cases, which indicates that it depends less on the number and size of clusters than the other indexes.

Next, we change part of the labels in all clusters of *P*. At each step, 50 more objects will be wrongly labeled in each cluster until all objects in *G* are equally distributed among the three clusters in *P*, see Fig. 15.

Fig. 13. Enlarging the second (grey) cluster in steps of 50 objects as in Fig. 11.

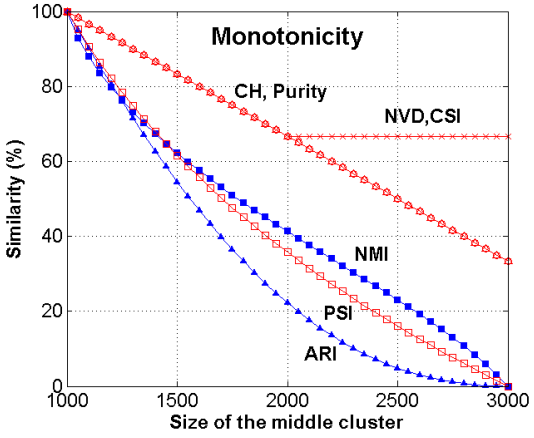


Fig. 14. Increasing the size of the second cluster until it contains all data objects.

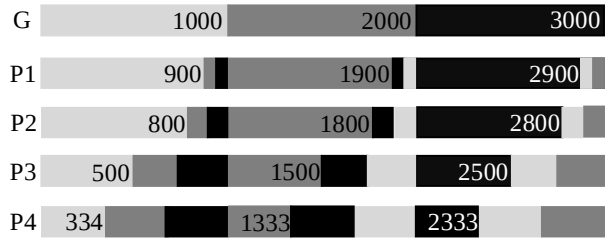


Fig. 15. Increasing the number of incorrectly labeled objects.

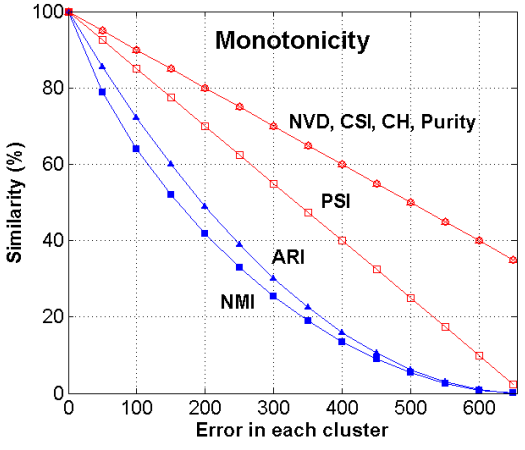


Fig. 16. Increasing the error of each cluster in *P*.

The similarity values of PSI and NVD, CH and Purity decrease linearly but NVD has higher similarity values than PSI, see Fig. 16. Since NVD, CH and Purity are not corrected for chance and are biased to random partitions, they have a higher lower bound. If we made them corrected, they would lose the linearity. The results of NVD, CH and Purity are exactly the same because the matching for all cases in this experiment is the same, which further verifies our claim in (12). Both NMI and ARI have decreasing curves and their values are always lower than those of the set matching indexes. One reason is that NMI and ARI consider also the unmatched parts of clusters.

7.4 Cluster Size Imbalance

In this experiment, we study the impact of cluster size. In Fig. 17, we consider sets of partitions where P_1 and P_2 have 200 objects (20%) wrongly labeled in the first two clusters. The size of the third cluster is decreased from 2000 to 50 in steps of 50.

Since the labels of the first two clusters remain exactly the same, the only difference originates from the size of the third cluster. We assumed that the clusters with different sizes have the same importance, and therefore, the results should be independent of the size of the third cluster. As shown in Fig. 18, all indexes except PSI are affected by the cluster size imbalance. For example, the similarity value of ARI is much lower (66%) when the size becomes 50 than when it is 2000 (91%). The results indicate that most indexes are affected more by the larger clusters. NVD, CSI, CH and Purity values are higher and in a narrower range, which indicates better performance of set matching indexes. Since matching results for NVD, CH and Purity are the same, their results are also the same, see (12). The proposed PSI is the one that copes best with the cluster size imbalance.

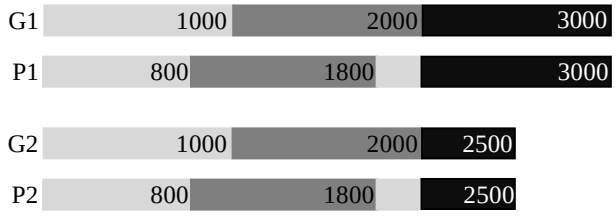


Fig. 17. The effect of cluster size imbalance; same error in the first two clusters and no error in the third clusters.

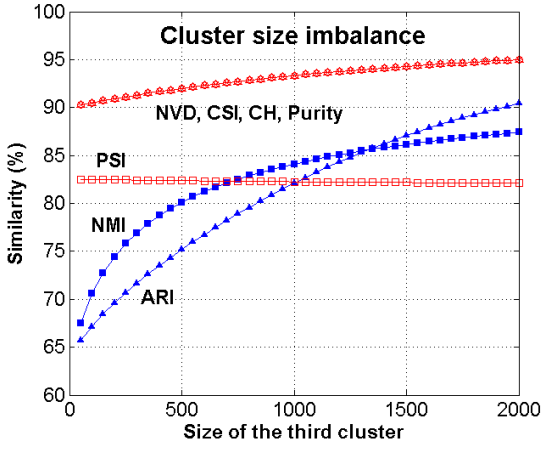


Fig. 18. The effect of cluster size imbalance on the indexes; the partitions contain two clusters with the fixed size and error and the size of the third cluster decreases in steps of 50.

7.5 Number of Clusters

We study the effect of the number of clusters by wrongly labeling 200 objects in each cluster and then varying the number of clusters as shown in Fig. 19. The size of clusters is fixed.

The indexes have similar trend on increasing the number of clusters except non-adjusted set matching indexes (NVD, CSI, CH and Purity), see Fig. 20. When increasing the number of clusters, the similarity values rise from as low as 25% up to 80%. However, the impact is much more significant for the small number of clusters from two to four. PSI has better performance than NMI and ARI, but only NVD, CSI, CH and Purity are completely independent on the number of clusters. Considering NVD equation in Table 2 and the same percentage of error across clusters in this experiment, it is trivial to show that NVD is independent on the number of clusters. Since matching results for NVD, CSI, CH and Purity are the same, their results are also the same, see (12). In this experiment, we see that correction for chance has bad effect as it makes the index dependent on the number of clusters. Overall, set matching indexes show better performance than the representatives from pair-counting and information theoretic indexes.

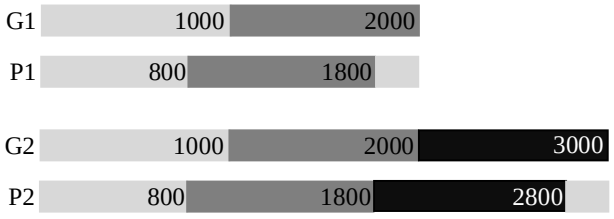


Fig. 19. There are 200 objects wrongly labeled in each cluster and the number of clusters varies.

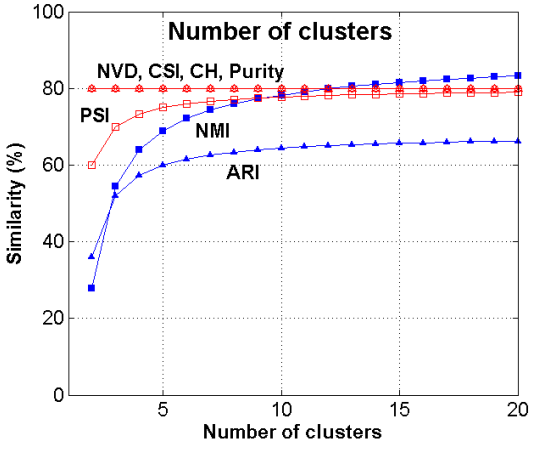


Fig. 20. The effect of number of clusters (K=2 to 20), while the size and error of each cluster are fixed.

7.6 Overlap of clusters

We use a series of data sets (called M2), all containing two clusters (1000 points each) in 8-dimensional space but with varying cluster overlap. The points were generated by Gaussian distribution with the same (constant) variance. The overlap was created by moving one of the clusters closer to the other step by step. The amount of overlap is measured by how many points in a cluster are closer to the centroid of the other cluster than to the centroid of its own cluster.

We cluster these datasets by random swap algorithm [41] and compare the result against ground truth partitions. Fig. 21 shows that all NVD, CSI, CH, Purity, and PSI react as expected. NVD approximately equals to the amount of the overlap, but is lower limited by 0.50. For example, with 15% overlap we expect to have 0.85 similarity. On the other hand, PSI applies correction for chance. Expected similarity of random partition into two clusters is 0.50, and corrected similarity 1-(overlap/0.50), accordingly. With 15% overlap, the expected similarity would be 0.70. The results of PSI are near optimal response (dashed black line).

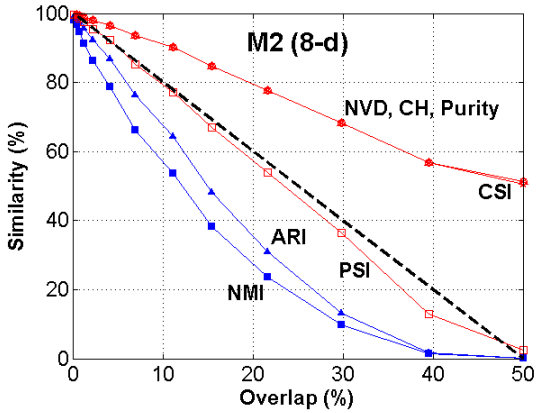


Fig. 21. Effect of the overlap on the similarity measures.

7.7 Dimensionality of data

We used the same M2 data sets but this time we fix the overlap to 15% and vary the dimensionality from 1 to 512. The results in Fig. 22 show that all the methods are invariant to the dimensionality up to a limit (about 256). Decrease of the index values is caused by over-optimization of the clustering algorithm: with high-dimensional data, it can optimize MSE better than would be with the ground truth partition. Otherwise, NVD, CSI, CH, Purity, and PSI again perform as expected with this overlap: NVD gives 0.85 (without) and PSI gives 0.70 (with correction for chance).

7.8 Applications

We study next how the four indexes (ARI, NMI, 1-NVD, PSI) perform with applications. We perform three experiments with the following hypotheses.

In the first experiment, we cluster the dataset Unbalance, see Fig. 23, to k=8 clusters by the following algorithms: random swap (RS) with 5,000 iterations [41], agglomerative clustering with ward criterion (AC), k-means

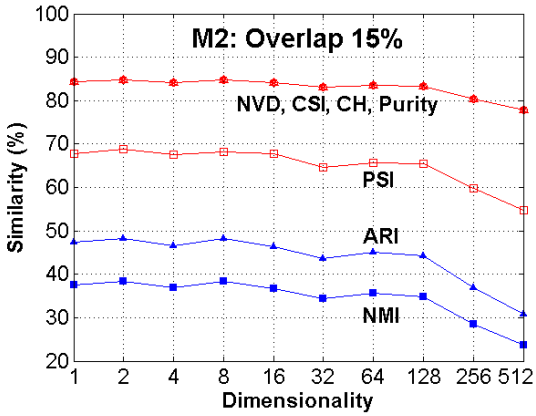


Fig. 22. Effect of dimensionality on the similarity measures.

(KM), and single link (SL). All these methods aim at minimizing total squared error except the single link.

The clusterings are then compared with the known ground truth in Table 6. The result of PSI corresponds best to the expectations: RS and AC are both good at optimizing the structure of the data whereas AC tends to make more point-wise errors at the partition borders. KM detects the dense cluster (2000 points) on the top, but it breaks the two other dense clusters into six smaller subclusters, and merges the five smaller ones (each 100

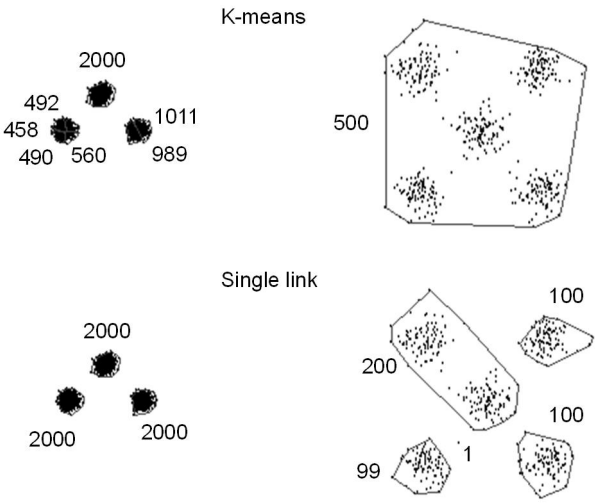


Fig. 23. Clustering results of the data set Unbalance using k-means (above), and single link (below).

TABLE 6
CLUSTERING OF UNBALANCE BY FOUR ALGORITHMS

Algorithms	External indexes			
	ARI	NMI	NVD	PSI
RS	1.00	1.00	1.00	1.00
AC	1.00	1.00	1.00	1.00
SL	1.00	0.99	0.99	0.78
KM	0.66	0.77	0.78	0.18

points) into one cluster, see Fig. 23 (top). All indexes react to these errors but only PSI recognizes that this clustering is off very low quality. SL finds all clusters correctly except that it merges two small ones leaving one orphan point as its tiny cluster, see Fig. 23 (below). Only PSI reacts strongly enough to this situation.

In the second experiment, we take ground truth clusters of the well-known *Yeast* data set (UCI), and then remove the smallest clusters one by one, see Fig. 24. The results in Table 7 show that only PSI provides significant differences due to the cluster removal, mainly because it treats all clusters of equal importance independent of their size.

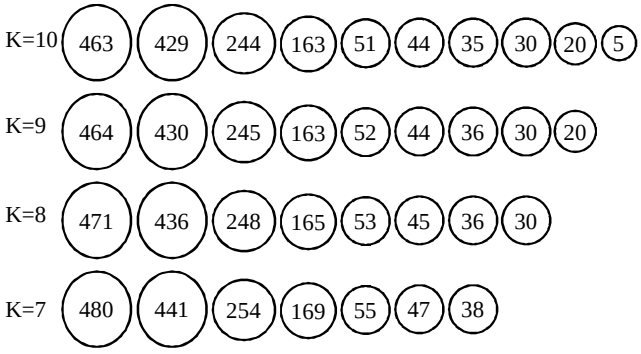


Fig. 24. Removing small clusters one by one and distributing their objects in the other clusters.

TABLE 7
CLUSTERING OF YEAST

Clusters (K)	External indexes			
	ARI	NMI	NVD	PSI
9	1.00	0.99	1.00	0.88
8	0.99	0.97	0.98	0.74
7	0.97	0.93	0.97	0.60

In the third experiment, we study how well the indexes apply for the task of detecting the number of clusters for Unbalance data set that contains eight clusters. We use the stability-based approach in [42] as follows. Ten subsets are generated by random sampling (with the sampling rate 0.2) from the data set. Each subset is then clustered by random swap algorithm with different number of clusters in range $k \in [2, 20]$. The similarity between the clustering of each subset and the clustering of the fullest is calculated by an external index. The stability is then measured as the average index values for all the subsets. The hypothesis is that the correct number of clusters is the one with highest stability (highest index value).

The results in Fig. 25 show that all indexes are applicable to this task, and the bigger problems originate from other factors than the choice of the index. All the indexes show maximum stability with k=8, but the clustering results are also stable with k=2 and k=4. Overall, PSI performs most consistent especially for values $k \geq 5$. In the range of k=5..7, all the indexes except PSI fail to detect high instability in the 5 small-sized clusters.

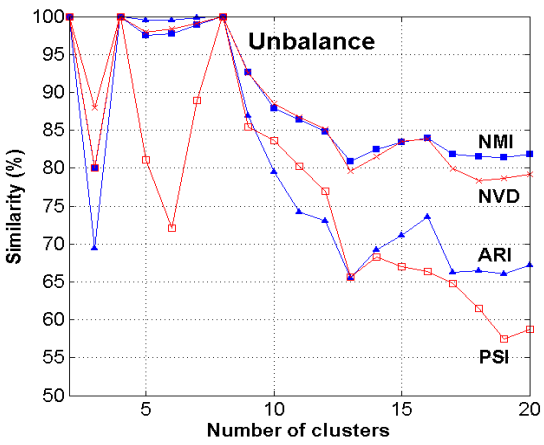


Fig. 25. Solving number of clusters based on stability of clusterings.

8 CONCLUSION

We have conducted a systematic study on existing set matching indexes by analyzing them in three different aspects: similarity measure of two clusters, matching the clusters, and the overall summation. We have shown that the difference between NVD, CH, Purity and CSI is only about their matching. If their matching result were the same, all these indexes would provide equivalent result. We have also pointed out that Purity and the measures cited as classification error or classification accuracy are equivalent.

We defined concrete requirements that an external index should meet, and introduced new arrangement of experiments based on synthetic data that can be used for systematic evaluation of any index according to these criteria. According to our experiments, set matching indexes perform better than the selected indexes of pair-counting and information theoretic indexes in many aspects such as cluster size imbalance, number of clusters and linear changes.

None of the existing set matching measures use correction for chance, and they also normalize the index across all data points. Based on these observations, we propose a new index called PSI that applies correction for chance, and performs normalization for each cluster separately. We show that the simplified form of PSI is a metric.

For the information theoretic measures, we have also shown that NMI=AMI=NVIs=AVIs under hypergeometric distribution assumption, which was also verified by our experiments.

ACKNOWLEDGMENT

This research has been supported by MOPIS project and partially by Nokia Foundation grant.

REFERENCES

[1] A.K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, 31(8), pp. 651–666, 2010.

[2] J. Handl, J. Knowles and D.B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, 21(15), pp. 3201-3212, 2005.

[3] G.W. Milligan and M.C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, 50(2), pp. 159–179, 1985.
[4] E. Dimitriadou, S. Dolnicar and A. Weingassel, "An examination of indexes for determining the number of clusters in binary data sets," *Psychometrika*, 67(1), pp. 137–160, 2002.
[5] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On clustering validation techniques," *J. Intelligent Information Systems*, 17(2-3), pp. 107–145, 2001.
[6] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(12), pp. 1650–1654, 2002.
[7] J. Wu, H. Xiong and J. Chen, "Adapting the right measures for k-means clustering," *15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'09)*, pp. 877–886, 2009.
[8] J. Wu, J. Chen, H. Xiong and M. Xie, "External validation measures for k-means clustering: A data distribution perspective," *Expert Systems with Applications*, 36(3), pp. 6050–6061, 2009.
[9] N.X. Vinh, J. Epps and J. Bailey, "Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance," *J. Machine Learning Research*, 11, pp. 2837–2854, 2010.
[10] A. Strehl, J. Ghosh and C. Cardie, "Cluster ensembles – A knowledge reuse framework for combining multiple partitions," *J. Machine Learning Research*, 3, pp. 583-617, 2003.
[11] L.I. Kuncheva and D.P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(11), pp. 1798–1808, 2006.
[12] S. Zhang, H. Wong and Y. Shen, "Generalized adjusted rand indices for cluster ensembles," *Pattern Recognition*, 45(6), pp. 2214–2226, 2012.
[13] L. I. Kuncheva, S. T. Hadjitodorov and L. P. Todorova, "Experimental comparison of cluster ensemble methods," *9th Int. Conf. on Information Fusion*, pp. 1-7, 2006.
[14] P. Fränti, J. Kivijärvi, T. Kaukoranta and O. Nevalainen, "Genetic algorithms for large scale clustering problems," *The Computer Journal*, 40 (9), pp. 547-554, 1997.
[15] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J.M. Pérez and J. Martín, "Towards a standard methodology to evaluate internal cluster validity indices," *Pattern Recognition Letters*, 32, pp. 505-515, 2011.
[16] M.H.C. Law and A.K. Jain, "Cluster validity by bootstrapping partitions," *Technical Report MSU-CSE-03-5*, Dept. of Computer Science and Engineering, MSU, Michigan, USA, 2003.
[17] M. Falasconi, A. Gutierrez, M. Pardo, G. Sberveglieri and S. Marco, "A stability based validity method for fuzzy clustering," *Pattern Recognition*, 43(4), pp. 1292-1305, 2010.
[18] D. Pascual, F. Pla, and J.S. Sanchez, "Cluster validation using information stability measures," *Pattern Recognition Letters*, 31(6), pp. 454-461, 2010.
[19] B.E. Dom, "An information-theoretic external cluster-validity measure," *Research Report RJ 10219*, IBM, 2001.
[20] P. Fränti, M. Rezaei and Q. Zhao, "Centroid index: cluster level similarity measure," *Pattern Recognition,* 47(9), pp. 3034-3045, 2014.
[21] N.X. Vinh, J. Epps and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?," *26th Int. Conf. on Machine Learning (ICML'09)*, pp. 1073-1080, 2009.
[22] S. Wagner, D. Wagner, "Comparing clusterings – an overview," *Technical Report, 2006-4*, Fakultät für Informatik, Universit¨at Karlsruhe (TH) , 2006.
[23] A.N. Albatineh, M. Niewiadomska-Bugaj and D. Mihalko, "On similarity indices and correction for chance agreement," *J. Classification*, 23(2), pp. 301-313, 2006.
[24] M. Meila, "Comparing clusterings – an information based distance," *J. Multivariate Analysis*, 98(5), pp. 873-895, 2007.
[25] T.O. Kvalseth, "Entropy and correlation: some comments," *IEEE Trans. Syst. Man Cybern.*, 17(3), pp. 517–519, 1987.
[26] M.C.P. de Souto, A.L.V. Coelho, K. Faceli, T.C. Sakata, V. Bonadia and I.G. Costa, "A comparison of external clustering evaluation indices in the context of imbalanced data sets," *2012 Brazilian Symposium on Neural Networks*, pp. 49-54, 2012.
[27] M. Meila and D. Heckerman, "An experimental comparison of model based clustering methods," *Machine Learning*, 41(1-2), pp. 9–29, 2001.
[28] S.V. Dongen, "Performance criteria for graph clustering and Markov cluster experiments," *Technical Report INSR0012*, Centrum voor Wiskunde en Informatica, 2000.
[29] Q. Zhao and P. Fränti, "Centroid ratio for a pairwise random swap clustering algorithm," *IEEE Trans. Knowledge and Data Engineering*, 26(5), pp. 1090-1101, 2014.
[30] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, pp. 193–218, 1985.
[31] A.A. Goshtasby, "Similarity and dissimilarity measures," in *Image Registration - Principles, Tools and Methods. Advances in Computer Vision and Pattern Recognition*, pp. 7-66, Springer London, 2012.
[32] S. Theodoridis, K. Koutroumbas, "Clustering: basic concepts," Pattern Recognition 4th edn, Academic Press, New York, pp. 595, 624, 2009.
[33] E. Rendon, I. Abundez, A. Arizmendi, E.M. Quiroz, "Internal versus external cluster validation indexes," *International Journal of Computers and Communications* 5(1), pp. 27-34, 2011.
[34] N. Nguyen and R. Caruana, "Consensus Clusterings," *IEEE Int. Conf. Data Mining*, pp. 607-612, 2007.
[35] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A Link-Based Approach to the Cluster Ensemble Problem," *IEEE Trans. Pattern Analysis and Machine Intelligence* 33(12) pp. 2396-2409, 2011.
[36] S. Choi, S. Cha and C. Tappert, "A survey of binary similarity and distance measures," *J. Systemics, Cybernetics and Informatics* 8(1), pp. 43-48, 2010.
[37] SB. Dalirsefat, A. Meyer and SZ. Mirhoseini, "Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori," *J. Insect Science*, 9, pp. 681-689, 2009.
[38] B. Sarker, "The resemblance coefficients in group technology: A survey and comparative study of relational metrics," *Computers and Industrial Engineering*, 30, pp. 103–116, 1996.
[39] B. Larsen, C. Aone, "Fast and effective text mining using linear time document clustering," *5th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pp. 16–22, 1999.
[40] H. W. Kuhn, "The Hungarian method for the assignment problem," *50 Years of Integer Programming 1958–2008*, pp. 29–47, 2010.
[41] P. Fränti and J. Kivijärvi, "Randomised local search algorithm for the clustering problem," *Pattern Analysis & Applications*, *3*(4), pp. 358-369, 2000.
[42] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural computation* 13(11), pp. 2573-2593, 2001.

Mohammad Rezaei received his BSc degree in Electronic engineering in 1996 and his M.Sc. degree in biomedical engineering in 2003 both from Amirkabir university of Technology, Tehran, Iran. Currently he is a PHD student in the university of Eastern Finland. His research interests include data clustering, multimedia processing, classification and retrieval.

Pasi Fränti received his MSc and PhD degrees from the University of Turku, 1991 and 1994 in Science. Since 2000, he has been a professor of Computer Science at the University of Eastern Finland. He has published 65 journals and 145 peer review conference papers, including 13 IEEE transaction papers. His research interests include clustering algorithms and location-based systems.

APPENDIX A

Proof of Theorem 1

First, we introduce a new way to derive the expected value of mutual information in case of random partitions and under hyper-geometric distribution assumption and then we use the expected value to prove (13). Consider a pair of clusters P_i and G_j. The probability that an object in P_i exists in G_j is m_j/N. Accordingly, the number of objects in both P_i and G_j is simplified as: $n_{ij}=n_i\times(m_j/N)$. Then, the expected value can be calculated according to (7) as:

$$E(MI)=E\left\{\sum_i\sum_j\frac{n_{ij}}{N}\log\left(\frac{N\times\left(n_i\times m_j/N\right)}{n_i\times m_j}\right)\right\}$$
$$=E\left\{\sum_i\sum_j\frac{n_{ij}}{N}\log(1)\right\}=0 \quad (24)$$

According to (2), AMI=NMI which confirms the result from [9]. Applying $max(MI)=(H(P)+H(G))/2$ as an option for normalization [22], [17], we can write:

$$AMI=NMI=\frac{2\times MI(P,G)}{H(P)+H(G)} \quad (25)$$

Since $E(H(P))=H(P)$ and $E(H(G))=H(G)$ under hyper-geometric distribution assumption, the expected value of VI (8) is derived as:

$$E(VI)=H(P)+H(G) \quad (26)$$

VI is a dissimilarity measure and $min(VI)=0$ when the two partitions are equal. Therefore, the adjusted variation of information according to (2) is:

$$AVI=\frac{VI}{H(P)+H(G)} \quad (27)$$

An upper bound for VI is $H(P)+H(G)$ and therefore (27) also represents the normalized variation of information. We simplify AVI_s and NVI_s using (8) as follows:

$$AVI_s=NVI_s=\frac{2\times MI(P,G)}{H(P)+H(G)} \quad (28)$$

From (25) and (28), we see that the adjusted mutual information and adjusted variation of information are equal to their normalized forms, and thus, theorem 1 is proven.

APPENDIX B

Proof of Theorem 2

Suppose that in a matching, m_1 is paired to $n_i<n_1$ and n_1 is paired to $m_j<m_1$ (case *a*). We show that if we change the matching so that m_1 is paired to n_1 and m_j is paired to n_i (case *b*), higher similarity is achieved. The total similarities for these two cases (*a* and *b*) are:

$$S_a=\frac{m_1\times(n_i/N))}{\max(m_1,n_i)}+\frac{m_j\times(n_1/N))}{\max(m_j,n_1)}$$
$$S_b=\frac{m_1\times(n_1/N))}{\max(m_1,n_1)}+\frac{m_j\times(n_i/N))}{\max(m_j,n_i)} \quad (29)$$

where S_a is the original pairing and S_b is the new pairing after changing the pairs for m_1 and m_j. Six different situations may happen:

1. $m_1>m_j>n_1>n_i$

$$\left[S_a=\frac{1}{N}(n_1+n_i)\right]=\left[S_b=\frac{1}{N}(n_1+n_i)\right]$$

2. $m_1>n_1>m_j>n_i$

$$\left[S_a=\frac{1}{N}(n_i+m_j)\right]<\left[S_b=\frac{1}{N}(n_1+n_i)\right]$$

3. $m_1>n_1>n_i>m_j$

$$\left[S_a=\frac{1}{N}(n_i+m_j)\right]<\left[S_b=\frac{1}{N}(n_1+m_j)\right]$$

4. $n_1>m_1>n_i>m_j$

$$\left[S_a=\frac{1}{N}(n_i+m_j)\right]<\left[S_b=\frac{1}{N}(m_1+m_j)\right]$$

5. $n_1>n_i>m_1>m_j$

$$\left[S_a=\frac{1}{N}(m_1+m_j)\right]=\left[S_b=\frac{1}{N}(m_1+m_j)\right]$$

6. $n_1>m_1>m_j>n_i$

$$\left[S_a=\frac{1}{N}(m_j+n_i)\right]<\left[S_b=\frac{1}{N}(m_1+n_i)\right] \quad (30)$$

Considering all the above situations, pairings (m_1, n_i) and (n_1, m_j) must be changed to (n_1, m_1) and (m_j, n_i) to achieve higher similarity. We can apply this proof recursively to all the smaller clusters as well. Hence, the two largest clusters must be always paired and then the next two largest and so on in order to achieve maximum total similarity with a random partition. This proves the theorem 2.

APPENDIX C

Triangular Inequality Proof for the Simplified form of PSI

Let P_1, P_2 and P_3 be three partitions with K_1, K_2 and K_3 clusters, and $K_{12}=max(K_1,K_2)$, $K_{23}=max(K_2,K_3)$, $K_{13}=max(K_1,K_3)$. Let n_i, n_j and n_k be the number of objects in clusters *i*, *j* and *k* in P_1, P_2 and P_3 respectively. We denote the number of shared objects between clusters by n_{ij}, n_{jk} and n_{ik}. The simplified distance form of PSI, for P_1 and P_2, according to (20) is:

$$D_{12}=\frac{K_{12}-S_{12}}{K_{12}-1}$$

Lemma. $D_{12}+D_{23}\geq D_{13}$ (31)

Proof. We define $D'_{12}=K_{12}-S_{12}$, $D'_{23}=K_{23}-S_{23}$ and $D'_{13}=K_{13}-S_{13}$ and prove first that: $D'_{12}+D'_{23}\geq D'_{13}$ which is equivalent to

$$K_{12}-S_{12}+K_{23}-S_{23}\geq K_{13}-S_{13} \quad (32)$$

We consider three possible situations and simplify (32):

(1) $K_1\geq K_{23}$: $S_{12}+S_{23}\leq K_{23}+S_{13}$
(2) $K_3\geq K_{12}$: $S_{12}+S_{23}\leq K_{12}+S_{13}$
(3) $K_2\geq K_{13}$: $S_{12}+S_{23}\leq K_2+(K_2-K_{13})+S_{13}$

In the case (3), since $K_2\geq K_{13}$, it is sufficient to prove $S_{12}+S_{23}\leq K_2+S_{13}$. Since $K_{23}\geq K_2$ and $K_{12}\geq K_2$, for all cases it is sufficient to prove:

$$S_{12}+S_{23}\le K_2+S_{13} \tag{33}$$

According to the definitions (14) and (15), we divide the inequality (33) into K_2 sub-inequalities by considering each cluster j in P_2 on the left. Each sub-inequality is of the form:

$$\frac{n_{ij}}{\max(n_i,n_j)}+\frac{n_{jk}}{\max(n_j,n_k)}\le 1+\frac{n_{ik}}{\max(n_i,n_k)} \tag{34}$$

Clusters *i* and *k* from P_1 and P_3 which are the pairs for cluster *j* are not necessarily a pair in comparing P_1 and P_3. Since S_{13} is derived according to perfect matching, we can consider another matching of P_1 and P_3 in which *i* and *k* are paired. If (33) holds in this case, it will also be true for S_{13} which is the maximum possible similarity.

If the cluster *j* has a pair cluster only in P_1 or P_3, it is trivial to prove (34). If it has pair clusters both in P_1 and P_3, and $n_{ij}+n_{jk}\le n_j$, proving (34) is trivial as well since the left side of the inequality is smaller than one. Note that if the clusters *i* and *k* do not have any shared objects, $n_{ij}+n_{jk}\le n_j$. So we prove (34) when $n_{ij}+n_{jk} > n_j$. Considering a minimum value for n_{ik} as $n_{ij}+n_{jk}-n_j$, we rewrite (34) as follows:

$$\frac{n_{ij}}{\max(n_i,n_j)}+\frac{n_{jk}}{\max(n_j,n_k)}\le 1+\frac{n_{ij}+n_{jk}-n_j}{\max(n_i,n_k)} \tag{35}$$

Three possible cases are:

(1) $n_j \ge max(n_i,n_k)$: By replacing $max(n_i,n_j)$ and $max(n_j,n_k)$ by n_j and after simplifications, we have:

$(n_{ij}+n_{jk}-n_j)(n_j-\max(n_i,n_k)) \ge 0$

which is always true in this case.

(2) $n_i \ge max(n_j,n_k)$: We replace $max(n_i,n_j)$ and $max(n_i,n_k)$ by n_i. Since $max(n_j,n_k) \ge n_j$, it is sufficient to prove (35) by replacing $max(n_j,n_k)$ by n_j. The equivalent inequality derived after simplification:

$(n_i-n_j)(n_j-n_{jk}) \ge 0$

is always true.

(3) $n_k \ge max(n_i,n_j)$: The same proof in the case (2) can be applied.

The lemma (31) can now be represented as:

$$\frac{K_{12}-S_{12}}{K_{12}-1}+\frac{K_{23}-S_{23}}{K_{23}-1}\ge\frac{K_{13}-S_{13}}{K_{13}-1} \tag{36}$$

We consider three possible cases:

(1) $K_1 \ge K_{23}$: It is sufficient to prove (36) if K_{23} in denumerator is replaced by K_1. So we simplify (36) as follows:

$$\frac{K_{12}-S_{12}}{K_1-1}+\frac{K_{23}-S_{23}}{K_1-1}\ge\frac{K_{13}-S_{13}}{K_1-1}$$

Since $K_1 \ge 2$, The denumerators can be canceled and the inequality is true according to (32).

(2) $K_3 \ge K_{12}$: The same inference as the case (1) can be performed by replacing K_{12} with K_3.

(3) $K_2 \ge K_{13}$: By simplifying (36), the following equivalent inequality is resulted:

$$S_{12}+S_{23}\le 2K_2-\frac{(K_{13}-S_{13})(K_2-1)}{K_{13}-1} \tag{37}$$

Using (32), it is sufficient to prove:

$$K_2+S_{13}\le 2K_2-\frac{(K_{13}-S_{13})(K_2-1)}{K_{13}-1}$$

After simplification we have:

$$S_{13}(K_2-K_{13})\ge(K_2-K_{13})$$

According to (14), $S_{13} \ge 0$, and therefore the above inequality is true.

According to the cases (1), (2) and (3), the inequalities (36) and consequently (31) hold, thus, the lemma is proven.