Understanding information theoretic measures for comparing clusterings

Hanneke van der Hoef¹ · Matthijs J. Warrens¹

Received: 4 March 2018 / Accepted: 14 November 2018 / Published online: 4 December 2018
© The Author(s) 2018

Abstract

Many external validity indices for comparing different clusterings of the same set of objects are overall measures: they quantify similarity between clusterings for all clusters simultaneously. Because a single number only provides a general notion of what is going on, the values of such overall indices (usually between 0 and 1) are often difficult to interpret. In this paper, we show that a class of normalizations of the mutual information can be decomposed into indices that contain information on the level of individual clusters. The decompositions (1) reveal that overall measures can be interpreted as summary statistics of information reflected in the individual clusters, (2) specify how these overall indices are related to individual clusters, and (3) show that the overall indices are affected by cluster size imbalance. We recommend to use measures for individual clusters since they provide much more detailed information than a single overall number.

Keywords Cluster analysis · Cluster validation · External validity indices · Information theory · Normalized mutual information · Shannon entropy

1 Introduction

Nowadays, we group almost everything: patient records, genes, web pages, behavioral problems, diseases, cancer tissue and so on (Kumar 2004). With the rapid technological growth, leading to a substantial increase in both the volume and variety of data, data analysis techniques are required to help analyze these data (Jain 2010).

Communicated by Alfonso Iodice D'Enza.

✉ Hanneke van der Hoef
h.van.der.hoef@rug.nl

Matthijs J. Warrens
m.j.warrens@rug.nl

¹ Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

A useful technique for this purpose is cluster analysis. Cluster analysis is a generic term for the collection of statistical techniques used to divide unlabeled data into groups (Kumar 2004). Cluster analysis is an exploratory and unsupervised classification technique, meaning that it does not require a known grouping structure a priori (Hennig et al 2015; Jain 2010; Kumar 2004). To help avoid confusion in the following sections, we would like to stress beforehand the terminology used in this paper. We use the word cluster for the unit of objects (e.g., patients, students, customers, animals, genes) that are placed together in a group. The words clustering and partition are used for a set of clusters that result from a cluster analysis.

Over the past 65 years, numerous clustering methods and algorithms have been developed (Jain 2010). All these different clustering methods generally obtain different clusterings of the same data set. As there is no ‘best’ clustering algorithm that dominates over all other algorithms, the question arises which clustering best fits the data set. A central topic in cluster analysis, therefore, is clustering validation: how to assess the quality of a clustering (Meilă 2015)? To help tackle this question, a large number of both internal and external validity indices have been proposed (Rendón et al 2011). Internal indices assess a clustering by characteristics using the data alone. External validity indices, on the other hand, use a priori information to assess the quality of a clustering (Jain 2010; Meilă 2015). External validity indices are commonly used to assess the similarity between clusterings, for example, clusterings obtained by different methods on the same data set (Pfitzner et al 2009).

External validity indices can be categorized into three approaches: (1) pair-counting (2) set-matching and (3) information theory (Meilă 2015; Vinh et al 2010). Most indices belong to the first approach, which is based on counting pairs of objects placed in identical and different clusters. Commonly used indices based on the pair-counting approach are the Rand index (Rand 1971) and the adjusted Rand index (Hubert and Arabie 1985; Steinley 2004; Warrens 2008b; Steinley et al 2016).

The second category is based on pairs of clusters instead of pairs of points (Meilă 2015). A central issue in the set-matching approach is ‘the problem of matching’ (Meilă 2007). Indices within this approach are problematic when two clusterings result in a different number of clusters, as it puts entire clusters outside consideration (Vinh et al 2010). Even with an equal number of clusters, these indices only assess the matched parts of each cluster, leaving the unmatched parts outside consideration (Meilă 2007; Vinh et al 2010). Examples of set-matching indices are the misclassification error (Steinley 2004), F-measure (Larsen and Aone 1999) and Van Dongen index (Van Dongen 2000).

A third class of indices is based on concepts from information theory (Cover and Thomas 1991). Information theoretic indices assess the difference in shared information between two partitions. Recently, information theoretic indices received increasing attention due to their strong mathematical foundation, ability to detect nonlinear similarities and applicability to soft clustering (Lei et al 2016; Vinh et al 2010). Commonly used information theoretic indices are the variation of information (Meilă 2007) and several normalizations of the mutual information (Amelio and Pizzuti 2016; Pfitzner et al 2009).

Just as a ‘best’ clustering method cannot be defined out of context, no ‘best’ validity criterion for comparing different clusterings can be defined that is appropriate

for all situations (Meilă 2015). To provide more insight into the wide variety of proposed indices, several authors have studied properties of these indices. Indices based on the pair-counting approach have been studied extensively over the past two decades (Albatineh et al 2006; Albatineh and Niewiadomska-Bugaj 2011; Baulieu 1989; Milligan 1996; Milligan and Cooper 1986; Steinley 2004; Warrens 2008a). In the past, information theoretic indices received less attention (Pfitzner et al 2009; Vinh et al 2010; Yao et al 1999) but they gained more attention recently (Amelio and Pizzuti 2016; Kvalseth 2017; Zhang 2015).

Since most of these validity indices are overall measures aimed to quantify agreement between two clusterings for all clusters simultaneously, they only give a general notion of what is going on. Often, their value (usually between 0 and 1) is hard to interpret. Usually, a value of 1 indicates perfect agreement, whereas a value of 0 indicates statistical independence of the two clusterings. Yet, prior studies that investigated validity indices did not provide much insight into how values between 0 and 1 should be interpreted. It is, therefore, desirable to perform more in-depth studies of overall indices to gain a more fundamental understanding of how their values between 0 and 1 may be interpreted.

In this paper, we consider a class of information theoretic measures. All indices in this class are commonly used normalizations of the mutual information (Kvalseth 1987; Pfitzner et al 2009; Vinh et al 2010). The goal of the paper is to gain insight into what the values of the overall measures may reflect. To achieve this goal, we decompose the overall measures into indices that contain information on the individual clusters of the partitions, and we analyze the relationships between the overall indices, the indices for individual clusters and their associated weights in the decompositions.

The presented decompositions also provide insight into a phenomenon that has been observed earlier in the classification literature: sensitivity to cluster size imbalance of overall measures (De Souto et al 2012; Rezaei and Fränti 2016). Cluster size imbalance basically means that at least one of the partitions has clusters of varying sizes. If an overall measure is sensitive to cluster size imbalance this generally means that its value tends to reflect the degree of agreement between large clusters. The analyses presented in this paper provide new theoretical insight into how this phenomenon actually works for a class of information theoretic indices. This is investigated by studying the weights in the decompositions of the overall measures.

The paper is organized as follows. In Sect. 2, we introduce the notation and we define the indices. In Sect. 3, we present decompositions of two asymmetric indices that are the building blocks of our class of information theoretic indices. In addition, we show that each asymmetric index can be further decomposed into indices that contain information on individual clusters. In Sect. 4, we study properties of the weights that are used in the decompositions of the asymmetric indices. The analysis presented in this section shows that cluster size imbalance is quite a complicated concept for information theoretic indices. How the overall measures are affected by what is going on the cluster level depends on the particular combination of cluster sizes in the partitions. The various relationships between the indices and weights presented in Sects. 2, 3 and 4 are illustrated in Sect. 5 with artificial examples and a real world example. Finally, Sect. 6 contains a discussion.

2 Normalized mutual information

Suppose the data are scores of N objects on k variables. Let $U = \{U_1, U_2, \dots, U_I\}$ and $V = \{V_1, V_2, \dots, V_J\}$ be two partitions of the N objects in, respectively, I and J clusters. One partition could be a reference partition that purports to represent the true cluster structure of the objects, while the second partition may have been obtained with a clustering method that is being evaluated. Furthermore, let $P = \{p_{ij}\}$ be a matching table of size $I \times J$ where p_{ij} indicates the proportion of objects (with respect to N) placed in cluster U_i of the first partition and in cluster V_j of the second partition. The cluster sizes in the partitions are reflected in the row and column totals of P , denoted by p_{i+} and p_{+j} , respectively.

The Shannon entropy (Shannon 1948) of partition U is given by

$$H(U) := - \sum_{i=1}^I p_{i+} \log p_{i+}, \quad (1)$$

in which $\log$ denotes the base 2 logarithm as is common use in information theory, and $p_{i+} \log p_{i+} = 0$ if $p_{i+} = 0$. The entropy of partition U is a measure of the amount of randomness of a partition. It is always non-negative and has value 0 if all objects are in one cluster of the partition, i.e., $p_{i+} = 1$ for some i . The entropy of partition V is defined analogously:

$$H(V) := - \sum_{j=1}^J p_{+j} \log p_{+j}. \quad (2)$$

The mutual information of clusterings U and V is then defined as

$$I(U;V) := \sum_{i=1}^I \sum_{j=1}^J p_{ij} \log \frac{p_{ij}}{p_{i+}p_{+j}}. \quad (3)$$

The mutual information quantifies how much information the two partitions have in common (Pfutzner et al 2009). Mutual information is occasionally referred to as ‘correlation measure’ in information theory (Malvestuto 1986). It is always non-negative and has a value of 0 if and only if the partitions are statistically independent, i.e. $p_{ij} = p_{i+}p_{+j}$ for all i and j . Higher values of mutual information indicate more shared information.

To facilitate interpretation and comparison, various normalizations of (3) have been proposed such that the maximum value of the normalized index is equal to unity. In Table 1, different commonly used normalizations of the mutual information are presented. The upper two indices are called asymmetric versions of the normalized mutual information, since they normalize using $H(U)$ and $H(V)$, respectively.

The top index of Table 1 is given by

Table 1 Different normalizations of the mutual information between U and V

Index	Formula	Source
R	$\frac{I(U;V)}{H(U)}$	Kvalseth (1987), Malvestuto (1986), Quinlan (1986)
C	$\frac{I(U;V)}{H(V)}$	Kvalseth (1987), Malvestuto (1986), Quinlan (1986)
$NMI^{\min}$	$\frac{I(U;V)}{\max(H(U), H(V))}$	Horibe (1985), Kvalseth (1987)
$NMI^{\max}$	$\frac{I(U;V)}{\min(H(U), H(V))}$	Kvalseth (1987), Strehl and Ghosh (2002)
NMI^{sqrt}	$\frac{I(U;V)}{\sqrt{H(U)H(V)}}$	Kvalseth (1987), Strehl and Ghosh (2002)
NMI^{sum}	$\frac{2I(U;V)}{H(U) + H(V)}$	Danon et al (2005), Kvalseth (1987), Malvestuto (1986)

$$R = \frac{I(U;V)}{H(U)} = \frac{\sum_{i=1}^I \sum_{j=1}^J p_{ij} \log \frac{p_{ij}}{p_{i+} p_{+j}}}{-\sum_{i=1}^I p_{i+} \log p_{i+}}. \quad (4)$$

Index (4) is a normalization of the mutual information that is frequently used in cluster analysis research (Malvestuto 1986; Quinlan 1986; Kvalseth 1987). It can be used to assess how well the clusters of the first partition U match the clusters of the second partition V (Malvestuto 1986). The index takes on values in the unit interval. We have $R = 1$ if no two objects from different clusters of U are put together in a cluster of V . In other words, each cluster of V only contains objects from a single cluster of U . Furthermore, we have $R = 0$ if the partitions are statistically

Table 2 Two examples of matching tables

First partition	Second partition					
	V_1	V_2	V_3	V_4	V_5	Total
(a)						
U_1	0.30	0	0	0	0	0.30
U_2	0	0.20	0.10	0	0	0.30
U_3	0	0	0	0.30	0.10	0.40
Total	0.30	0.20	0.10	0.30	0.10	1
(b)						
U_1	0.02	0.03	0.04	0.05	0.06	0.20
U_2	0.04	0.06	0.08	0.10	0.12	0.40
U_3	0.04	0.06	0.08	0.10	0.12	0.40
Total	0.10	0.15	0.20	0.25	0.30	1

independent, i.e., $p_{ij} = p_{i+}p_{+j}$ for all i and j . In general, higher values of index (4) imply higher similarity between U and V .

To illustrate the extreme values of index (4), consider the two matching tables in Table 2. Each matching table has size 3×5 . We have $R = 1$ for Table 2a (upper panel), since no two objects from different clusters of U are put together in a cluster of V . For example, the objects in U_2 are matched to clusters V_2 and V_3 . For both V_2 and V_3 , it can be seen that these clusters contain only objects from U_2 and no objects from U_1 or U_3 . Furthermore, we have $R = 0$ for Table 2b) (lower panel) because the two partitions are statistically independent.

The second index from the top of Table 1 is given by

$$C = \frac{I(U;V)}{H(V)} = \frac{\sum_{i=1}^I \sum_{j=1}^J p_{ij} \log \frac{p_{ij}}{p_{i+}p_{+j}}}{-\sum_{j=1}^J p_{+j} \log p_{+j}}. \quad (5)$$

Index (5) can be used to assess how well the clusters of partition V match with the clusters of the first partition U (Malvestuto 1986). We have $C = 1$, if no two objects from different clusters of V are put together in a cluster of U . In other words, each cluster of U only contains objects from a single cluster of V . Furthermore, we have $C = 0$ if the two partitions are statistically independent. For example, for Table 2a, we have $C = 0.72$, since objects from V_2 and V_3 are put together in U_2 and objects from V_4 and V_5 are put together in U_3 , and for Table 2b we have $C = 0$ since the partitions are statistically independent.

The bottom four indices in Table 1 normalize the index (1) using generalized means of indices (4) and (5). More precisely, the denominators of the indices are, from top to bottom, the minimum, maximum, geometric mean and harmonic mean of (4) and (5). This means that their values lie somewhere between the values of (4) and (5). Compared to the arithmetic mean, the harmonic and geometric means put more emphasis on the lowest of the two values. Thus, to understand all six indices in Table 1, it is instrumental to first understand asymmetric indices (4) and (5). To enhance our understanding of these two indices, they are decomposed into chunks of information on individual clusters in the next section.

3 Decompositions

Index (4) can be decomposed into indices for individual clusters. First, define for a cluster $U_i \in U$ the normalized weight

$$u_i := \frac{-p_{i+} \log p_{i+}}{H(U)} = \frac{-p_{i+} \log p_{i+}}{-\sum_{i=1}^I p_{i+} \log p_{i+}}. \quad (6)$$

The weight in (6) is the part of the entropy of partition U that is associated with cluster U_i divided by the total entropy of partition U . The weight in (6) is normalized

in the sense that if we add the weights associated with all the clusters of U , the sum is equal to unity. We study the weight in more detail in Sect. 4 below. The definition in (6) (and (8) below) makes the comparison between different weight scenarios easier.

Next, define for a cluster $U_i \in U$ the index

$$R_i := \frac{\sum_{j=1}^J p_{ij} \log \frac{p_{ij}}{p_{i+}p_{+j}}}{-p_{i+} \log p_{i+}}. \quad (7)$$

The numerator of (7) consists of the part of the mutual information between partitions U and V that is associated with cluster U_i only. Furthermore, the denominator of (7) is the part of the entropy of partition U that is associated with cluster U_i .

Index (7) can be used to assess how well cluster U_i matches to the clusters of partition V . The index takes on values in the unit interval. We have $R_i = 1$ if objects from U_i are in clusters of V that contain no objects from other clusters of U , i.e., we have $p_{ij} = p_{+j}$ if $p_{ij} > 0$ for all j . Furthermore, we have $R_i = 0$ if $p_{ij} = p_{i+}p_{+j}$ for all j . This is the case if the objects of cluster U_i are randomly assigned (in accordance with the p_{+j} 's) to the clusters of partition V .

Analogously, define for $V_j \in V$ the normalized weight

$$v_j := \frac{-p_{+j} \log p_{+j}}{H(V)} = \frac{-p_{+j} \log p_{+j}}{\sum_{j=1}^J -p_{+j} \log p_{+j}}, \quad (8)$$

and the index

$$C_j := \frac{\sum_{i=1}^I p_{ij} \log \frac{p_{ij}}{p_{i+}p_{+j}}}{-p_{+j} \log p_{+j}}. \quad (9)$$

The numerator of (9) consists of the part of the mutual information between partitions U and V that is associated with cluster V_j only, whereas the denominator of (9) is the part of the entropy of partition V that is associated with cluster V_j . Index (9) has properties analogous to index (7).

We have the following decomposition for index (4). Index (4) is a weighted average of the indices in (7) using the u_i 's in (6) as weights:

$$R = \sum_{i=1}^I u_i R_i. \quad (10)$$

Since R is a weighted average of the R_i values, the overall R value lies somewhere between the minimum and maximum of the R_i values. Equation (10) shows that the overall R value is largely determined by the R_i values of clusters with high u_i values. The overall R value will be high if R_i values corresponding to high u_i values are

themselves high. Vice versa, the overall R value will be low if R_i values corresponding to high u_i values are low.

We have an analogous decomposition for index (5). Index (5) is a weighted average of the indices in (9) using the v_j 's in (8) as weights:

$$C = \sum_{j=1}^J v_j C_j. \quad (11)$$

Since C is a weighted average of the C_j values, the overall C value lies somewhere between the minimum and maximum of the C_j values. Equation (11) shows that the overall C value is largely determined by the C_j values of the clusters with high v_j values. The overall C value will be high if C_j values corresponding to high v_j values are themselves high. Vice versa, the overall C value will be low if C_j values corresponding to high v_j values are low.

Decompositions (10) and (11) show that the indices in Table 1 are functions of the R_i 's and C_j 's corresponding to individual clusters. For example, the bottom index of Table 1 is a weighted average of the R_i 's and C_j 's, using the u_i 's and v_j 's as weights:

$$\text{NMI}^{\text{sum}} = \frac{2I(U;V)}{H(U) + H(V)} = \frac{H(U) \sum_{i=1}^I u_i R_i + H(V) \sum_{j=1}^J v_j C_j}{H(U) \sum_{i=1}^I u_i + H(V) \sum_{j=1}^J v_j}. \quad (12)$$

The values of the indices in Table 1 are largely determined by the R_i values and C_j values of clusters with high u_i values and v_j values. The normalized weights in (6) and (8), and how they act in decompositions (10) and (11), are further studied in the next section.

4 Weights

Decompositions (10) and (11) show that the contribution of the cluster indices to the overall measures R and C depends on the normalized weights in (6) and (8). Since the weights are functions of the relative cluster sizes, R and C are sensitive to some form of cluster size imbalance. In this section, the particular form of cluster size imbalance is further explored.

To enhance our understanding of the weights (6) and (8), we define the function $f(p) := -p \log p$ with $p \in [0, 1]$ and $f(0) = 0$. Since the second derivative $f''(p) = -1/(p \ln 2)$ is always negative on $(0, 1)$, function $f(p)$ is a concave function of p , which has a maximum. Since the first derivative is $f'(p) = -(\ln p + 1)/\ln 2$, and since $f'(p) = 0$ if and only if $p = 1/e \approx 0.368$, the maximum of 0.531 is obtained at approximately $p = 0.368$.

Figure 1 is a plot of the function $f(p)$ on the unit interval. The figure shows that the function is concave and slightly skewed to the right. In sum, R_i values and C_j values of clusters with relative size approximately equal to 0.368 receive maximum

Fig. 1 Plot of $f(p)$ on the unit interval

weight in overall indices (4) and (5). Clusters of which the relative size is smaller or larger than $p = 0.368$ receive lower weights.

For different number of clusters, we may distinguish between a number of different scenarios, in which the weighting influences the index values differently. We will discuss these scenarios briefly below. To begin with, we encounter a special situation if a partition has precisely two clusters. If the first cluster has relative size p , the associated normalized weight is given by

$$\frac{-p \log p}{-p \log p - (1 - p) \log(1 - p)}. \quad (13)$$

Table 3 presents various relative cluster sizes and corresponding normalized weights for a partition with two clusters. Close inspection of the table reveals that with two clusters the smallest cluster (i.e., cluster 1 in Table 3) always has the highest normalized weight. The clusters have the same weight only if they have the same size, i.e., if $p_1 = 0.50$ and $p_2 = 0.50$. The weights tend to be more different when the cluster sizes are also more different. For example, if the small cluster has size $p = 0.05$, its corresponding weight is three times the weight of the large cluster. Thus, the R_i value (C_j value) of the small cluster contributes three times as much to the overall R value (C value) than the R_i value (C_j value) of the large cluster. Furthermore, if the small cluster has size $p = 0.15$, its corresponding weight is twice the weight of the large cluster. Thus, the R_i value (C_j value) of the small cluster contributes two times

Table 3 Cluster sizes and normalized weights for a partition with two clusters

Relative cluster size (p)		Normalized weight	
Cluster 1	Cluster 2	Cluster 1	Cluster 2
0.05	0.95	0.75	0.25
0.10	0.90	0.71	0.29
0.15	0.85	0.67	0.33
0.20	0.80	0.64	0.36
0.25	0.75	0.62	0.38
0.30	0.70	0.59	0.41
0.35	0.65	0.57	0.43
0.40	0.60	0.54	0.46
0.45	0.55	0.52	0.48
0.50	0.50	0.50	0.50

as much to the overall R value (C value) than the R_i value (C_j value) of the large cluster.

Furthermore, if a partition has three or more clusters, but still a small number of clusters, a number of different situations may occur. This may be illustrated with a partition that has precisely three clusters. Table 4 presents various relative cluster sizes and associated normalized weights for a partition with three clusters. In contrast to the case of two clusters, a number of different combinations of weights may occur with three clusters.

First of all, the smallest cluster may have the highest weight (upper panel of Table 4). In the upper panel, the two smallest clusters (clusters 1 and 2) have the same size. The numbers show that when the sizes of the small clusters and the large cluster differ substantially, the small clusters have a higher weight. A special situation occurs when the two small clusters both have a relative size that is half the size

Table 4 Cluster sizes and normalized weights for a partition with three clusters

Relative cluster size (p)			Normalized weight		
Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
0.05	0.05	0.90	0.38	0.38	0.24
0.10	0.10	0.80	0.36	0.36	0.28
0.15	0.15	0.70	0.35	0.35	0.30
0.20	0.20	0.60	0.34	0.34	0.32
0.25	0.25	0.50	0.33	0.33	0.33
0.05	0.30	0.65	0.19	0.46	0.35
0.10	0.30	0.60	0.26	0.40	0.34
0.15	0.35	0.50	0.28	0.37	0.35
0.10	0.45	0.45	0.24	0.38	0.38
0.20	0.40	0.40	0.31	0.35	0.35
0.30	0.35	0.35	0.33	0.34	0.34

of the large cluster (last row of upper panel). In that case, the two small clusters and the large cluster receive equal weights.

Secondly, the medium-sized cluster may have the highest weight (middle panel of Table 4). The middle panel of Table 4 presents three examples. In each example, the medium-sized cluster (cluster 2) receives the highest weight. The difference in weights is most obvious in the second example (second row of the middle panel), where the weight of the medium-sized cluster equals the sum of the other two cluster weights.

Finally, the largest cluster may have the largest weight (bottom panel of Table 4). In the bottom panel, the two largest clusters (clusters 2 and 3) have the same size. Moreover, the numbers show that when the sizes of the small cluster and the large clusters differ substantially, the weights also differ substantially.

If all relative cluster sizes are smaller than $p = 0.368$, the largest cluster will also have the highest weight. In this case, the R_i values (C_j values) of the larger clusters contribute (much) more to the overall R value (C value) than the R_i values (C_j values) of the smaller clusters. All cluster sizes can be smaller than $p = 0.368$ if a partition consists of at least three clusters. Furthermore, if the number of clusters (possibly with different sizes) is quite large, e.g., with big data, then it is likely that all relative cluster sizes are smaller than $p = 0.368$.

Next, a partition may consist of one large cluster together with a few or many small clusters. For example, suppose we have a cluster of size $n = 1000$ and two clusters of size $n = 40$. In this case, we have $p = 0.926$ for the large cluster and $p = 0.037$ for both small clusters. The large cluster has a normalized weight of size 0.226, whereas both small clusters have a weight of 0.387. In this case, the R_i values (C_j values) of each small cluster contribute almost twice as much to the overall R value (C value) compared to the R_i values (C_j values) of the large cluster.

Finally, a partition may consist of several large clusters together with many small clusters. For example, suppose there are three clusters of size $n = 1000$ and 50 clusters of size $n = 20$. In this case, we have $p = 0.25$ for the large clusters and $p = 0.005$ for the small clusters. The large clusters each have a normalized weight of size 0.128, while each small cluster has a weight of 0.0098. In this case, the R_i value (C_j value) of a large cluster contributes 13 times more to the overall R value (C value) than the R_i value (C_j value) of a small cluster. Hence, the overall R value (C value) to a large extent assessing how well the large clusters of the partitions match.

5 Numerical examples

In this section, we illustrate with numerical examples the relationships between the overall indices, cluster information and the normalized weights as presented in the previous sections. To do this, we use the matching tables presented in Table 2, the matching tables presented in Table 5, and a real world example presented in Table 7. Table 6 presents for each matching table of Tables 2 and 5 the corresponding values of overall indices (4) and (5), the indices for individual clusters (7) and (9), and the corresponding weights (6) and (8).

Table 5 Two more examples of matching tables

First partition	Second partition			
	V_1	V_2	V_3	Total
(a)				
U_1	0.96	0	0	0.96
U_2	0	0.01	0.01	0.02
U_3	0	0.01	0.01	0.02
Total	0.96	0.02	0.02	1
(b)				
U_1	0.24	0.24	0	0.48
U_2	0.24	0.24	0	0.48
U_3	0	0	0.04	0.04
Total	0.48	0.48	0.04	1

Table 6 Values of indices and normalized weights for the four matching tables in Tables 2 and 5

Index	Table 2a		Table 2b		Table 5a		Table 5b	
	Value	Weight	Value	Weight	Value	Weight	Value	Weight
R	1		0		0.86		0.20	
C	0.72		0		0.86		0.20	
R_1	1	0.33	0	0.30	1	0.20	0.06	0.42
R_2	1	0.33	0	0.35	0.82	0.40	0.06	0.42
R_3	1	0.34	0	0.35	0.82	0.40	1	0.16
C_1	1	0.24	0	0.15	1	0.20	0.06	0.42
C_2	0.75	0.21	0	0.18	0.82	0.40	0.06	0.42
C_3	0.52	0.15	0	0.21	0.82	0.40	1	0.16
C_4	0.76	0.24	0	0.22				
C_5	0.40	0.15	0	0.23				

Table 7 Matching table for the Zoo data set

First partition	Second partition				
	V_1	V_2	V_3	V_4	Total
U_1 = mammal	0.41	0	0	0	0.41
U_2 = bird	0	0	0.20	0	0.20
U_3 = reptile	0	0.01	0.04	0	0.05
U_4 = fish	0	0.13	0	0	0.13
U_5 = amphibian	0	0	0.04	0	0.04
U_6 = insect	0	0	0.08	0	0.08
U_7 = mollusc et al.	0	0	0.09	0.01	0.10
Total	0.41	0.14	0.45	0.01	1

For Table 2a we have $R = R_1 = R_2 = R_3 = 1$, since no two objects from different clusters of U are put together in a cluster of V . Similarly, we have $C_1 = 1$. However, the cluster indices C_2, C_3, C_4 and C_5 are lower than unity because clusters V_2 and V_3 are both matched to U_2 , while clusters V_4 and V_5 are both matched to U_3 . For Table 2b all indices have value zero because the two partitions are statistically independent.

In Table 5a, the partitions consist of one large cluster and two small clusters. Overall indices (4) and (5) have a value of 0.86, suggesting there is high, yet not perfect, similarity between the partitions on the overall cluster structure. Furthermore, the indices for individual clusters show that there is perfect agreement on the large cluster ($R_1 = C_1 = 1$), but that agreement on the small clusters is only substantial ($R_2 = C_2 = R_3 = C_3 = 0.82$).

Overall index (4) is a weighted average of indices R_1, R_2 and R_3 . The weights associated with R_2 and R_3 are twice as high (0.40) as the weight corresponding to R_1 (0.20). The value 0.86 of the average index is, therefore, closer to the value 0.82 than to the value 1. Similar properties hold for the overall index (5). The indices for individual clusters provide in general more information than the overall indices.

In Table 5b, the partitions consist of two large clusters and one small cluster. Overall indices (4) and (5) have a value of 0.20, suggesting there is low similarity between the partitions on the overall cluster structure. However, the indices for individual clusters show that there is perfect agreement on the small cluster ($R_3 = C_3 = 1$), and that similarity on the large clusters is rather poor ($R_1 = C_1 = R_2 = C_2 = 0.06$).

Overall index (4) is a weighted average of indices R_1, R_2 and R_3 . The weights associated with R_1 and R_2 are more than two and a half times higher (0.42) than the weight corresponding to R_3 (0.16). The value 0.20 of the average index is, therefore, closer to the value 0.06 than to the value 1. Similar properties hold for the overall index (5).

As a final example, we consider the Zoo data set, which is available from the UCI Machine Learning Repository (Newman et al 1998). This data set consists of $n = 101$ animals, classified into seven categories: mammal, bird, reptile, fish, amphibian, insect, and mollusc et al.. Sixteen characteristics are provided, of which fifteen are binary (1 = possesses, 0 = does not possess): hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, hoof, and horns. A sixteenth variable is categorical, the number of legs (0, 2, 4, 5, 6, 8). This variable has been dichotomized (1 = yes, 0 = no) for this example.

We applied hierarchical cluster analysis to the sixteen binary variables using the Manhattan distance and median linkage method. After inspecting the dendrogram, we chose to report the solution with 4 clusters. Table 7 presents the matching table between the 7-cluster reference partition and the 4-cluster solution. Inspection of Table 7 shows that the mammals are recovered perfectly by the trial partition (i.e. $U_1 = V_1$). The second cluster V_2 consists of all 13 fish, accompanied by one reptile (i.e., the seasnake). The third cluster V_3 is a mix of birds, reptiles, amphibian, insects, and mollusca. Finally, the fourth cluster consists of only one animal (i.e., scorpion).

Table 8 presents the corresponding values of overall indices (4) and (5), the indices for individual clusters (7) and (9), and the corresponding weights (6) and (8) for matching Table 7. We consider indices (4) and (7) first. Index (7) can be used to assess how well cluster U_i of the reference partition matches to the clusters of trial partition V . We have $R_i = 1$ if objects from U_i are in clusters of V that contain no objects from other clusters of U . The mammals in U_1 are recovered perfectly by the trial partition, which is reflected in $R_1 = 1$. All fish are put together in a single cluster, but cluster V_2 also contains one reptile. So the recovery of the fish is very good, yet not perfect, which is reflected in $R_4 = 0.96$. Moreover, the birds are also put together in the same cluster, but cluster V_3 also contains species from other categories. Because there are relatively many birds in V_3 , and since all birds are put together, we have $R_2 = 0.50$, which is a moderate value. The remaining R_i values are quite low (ranging from 0.18 to 0.37). Although all species of several categories have been assigned to the same cluster, there are relatively many animals from other categories in cluster V_3 as well.

Overall we have $R = 0.60$. This value is a weighted average of the R_i values, which range from 0.18 to 1. Because R combines information from seven clusters, an interpretation of its value is not straightforward. The R_i values and associated weights provide some insight. The third and fourth columns of Table 8 show that, for index (7), the largest clusters actually have higher associated normalized weights. Thus, for these data, the R_i values associated with the larger animal categories contribute more to $R = 0.60$ than the smaller categories. The three largest categories are the mammals, birds and fish. The value $R = 0.60$ is moderately high because the mammals and fish have high associated indices ($R_1 = 1$ and $R_4 = 0.96$).

Next, we consider indices (5) and (9). Index (9) can be used to assess how well cluster V_j of the trial partition matches to the clusters of reference partition U . We have $C_j = 1$ if objects from V_j are in clusters of U that contain no objects from other

Table 8 Values of indices and associated normalized weights for Table 7 (Zoo data)

Index	Value	Relative cluster size (p)	Normal- ized weight
R	0.60		
C	0.95		
R_1	1	0.41	0.22
R_2	0.50	0.20	0.19
R_3	0.18	0.05	0.09
R_4	0.96	0.13	0.16
R_5	0.25	0.04	0.08
R_6	0.32	0.08	0.12
R_7	0.37	0.10	0.14
C_1	1	0.41	0.35
C_2	0.94	0.14	0.26
C_3	0.95	0.45	0.34
C_4	0.50	0.01	0.05

clusters of V . All mammals are put together in V_1 and V_1 does not contain other species, which is reflected in $C_1 = 1$. Furthermore, all fish are put together in V_2 , and cluster V_2 only contains one other animal, which is reflected in $C_2 = 0.94$. Moreover, all birds, amphibians and insects, and most of the reptiles and mollusca are put together in V_3 , which is reflected in $C_3 = 0.95$.

Overall we have $C = 0.95$. This value is a weighted average of the C_j values, which range from 0.50 to 1. Because C combines information from four clusters, its value is generally difficult to interpret. However, in this example three clusters have high C_j values, and this is more or less reflected in $C = 0.95$. For these data, the two large clusters V_1 and V_3 have the highest weights (0.35 and 0.34), although the weight corresponding to cluster V_2 is only a bit smaller (0.26). Interestingly, the weight associated with cluster V_1 is a bit higher than that of cluster V_3 , despite that V_1 is the slightly smaller cluster. This example illustrates that (with a few clusters) the relationships between the cluster information and an overall index may be rather complicated. Because the weight associated with cluster V_4 is rather small (0.05), the value $C = 0.95$ basically summarizes the C_j values corresponding to the first three clusters.

6 Discussion

Given that different clustering strategies generally result in different clusterings of the same data set, it is important to assess the similarity between different clusterings. For this purpose, external validity indices can be used. Yet, only a few research studies have focused on a thorough understanding of external validity indices. Especially, in the information theoretic approach, there is a lack of research offering a fundamental understanding of the behavior of such indices. As a result, users of cluster analysis generally provide an overall measure without proper insight into how values between 0 (no agreement) and 1 (perfect agreement) may be interpreted. There is a lot of room for fundamental research in this area.

This paper has focused on two commonly used asymmetric normalizations of the mutual information. The indices are actually the building blocks of a complete class of normalizations of the mutual information. We presented decompositions of both indices. The decompositions (1) show that the overall measures can be interpreted as summary statistics of information reflected in the individual clusters, (2) specify how these overall indices are related to individual clusters, and (3) show that the overall indices are affected by cluster size imbalance. The overall indices are functions, i.e., weighted averages, of the individual cluster indices. The values of the overall indices lie somewhere between the minimum and maximum of the values of the individual cluster indices.

In contrast to prior research (Pfitzner et al 2009; Vinh et al 2010), we found that normalizing the mutual information does not protect against sensitivity to cluster size imbalance. Instead, our findings are in line with De Souto et al. (2012). In this paper, by studying weights in detail, we made more precise in what way the two normalizations of the mutual information are sensitive to cluster size imbalance. We

provided mathematical proof for the optimal cluster size and weight, and we illustrated the consequences of cluster size imbalance by a graphical display and various numerical examples. More precisely, we showed that the relationship between the index value and the cluster sizes is rather complex when there is cluster size imbalance. Whether small, medium or large clusters have the biggest impact on the overall value depends on the particular combination of cluster sizes. Consequently, overall values do not have a universal or intuitive interpretation of the recovery of individual clusters. We purport to raise awareness that these overall measures are indeed affected by cluster size imbalance.

Because of the context dependency of the cluster size imbalance, we recommend researchers to examine and report the measures for the individual clusters in (7) and (9), since they provide more detailed information than a single overall number. When there is a large number of clusters, reporting all cluster indices is perhaps not feasible. One solution here is to report the distribution of the values of the cluster indices for each partition. Another solution for this case is to summarize the cluster indices by counting how many cluster indices have a value over a certain number that reflects high similarity (say 0.95) and all values below a certain number that indicates poor similarity (say 0.50).

Future work could focus on systematic experiments to examine how the cluster size imbalance property impacts real world data clusterings. In line with this, the tendency of information theoretic indices to overestimate the number of clusters could be investigated. Another interesting future direction is to consider a different weighting system, one that is insensitive to cluster size imbalance. A possibility here is to use unit weights for the observations regardless their cluster assignment. A fundamental understanding of validity indices facilitates more careful and proper use of such indices and accordingly supports the overarching aim of obtaining high-quality clusterings that best fit the data.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Albatineh AN, Niewiadomska-Bugaj M (2011) Correcting jaccard and other similarity indices for chance agreement in cluster analysis. *Adv Data Anal Classif* 5(3):179–200. <https://doi.org/10.1007/s11634-011-0090-y>
- Albatineh A, Niewiadomska-Bugaj M, Mihalko D (2006) On similarity indices and correction for chance agreement. *J Classif* 23(2):301–313. <https://doi.org/10.1007/s00357-006-0017-z>
- Amelio A, Pizzuti C (2016) Correction for closeness: adjusting normalized information measure for clustering comparison. *Comput Intell* 33(3):579–601. <https://doi.org/10.1111/coin.12100>
- Baulieu FB (1989) A classification of presence/absence based dissimilarity coefficients. *J Classif* 6(1):233–246. <https://doi.org/10.1007/BF01908601>
- Cover TM, Thomas JA (1991) Elements of information theory. In: Schilling D (ed) Wiley series in telecommunications. Wiley, New York, pp 12–49
- Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *J Stat Mech* 9:219–228. <https://doi.org/10.1088/1742-5468/2005/09/P09008>

- De Souto M, Coelho A, Faceli K, Sakata T, Bonadia V, Costa I (2012) A comparison of external clustering evaluation indices in the context of imbalanced data sets. Brazilian Symposium on neural networks pp 49–54. <https://doi.org/10.1109/SBRN.2012.25>
- Hennig C, Meilă M, Murtagh F, Rocci R (2015) Handb Clust Anal. Chapman and Hall/CRC, New York
- Horibe Y (1985) Entropy and correlation. IEEE Trans Syst Man Cybern 15(5):641–642. <https://doi.org/10.1109/TSMC.1985.6313441>
- Hubert LJ, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218. <https://doi.org/10.1007/BF01908075>
- Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recogn Lett 31(8):651–660. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Kumar V (2004) Cluster analysis: basic concepts and algorithms. In: Tan P, Steinbach M, Kumar V (eds) Introduction to data mining. Pearson Education, New York, pp 487–568
- Kvalseth TO (1987) Entropy and correlation: some comments. IEEE Trans Syst Man Cybern 17(3):519–519. <https://doi.org/10.1109/TSMC.1987.4309069>
- Kvalseth TO (2017) On normalized mutual information: measure derivations and properties. Entropy 19:1–14. <https://doi.org/10.3390/e19110631>
- Larsen B, Aone C (1999) Fast and effective text mining using linear-time document clustering. In: Proceedings of the Fifth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '99, pp 16–22. 10.1145/312129.312186
- Lei Y, Bezdek JC, Chan J, Vinh NX, Romano S, Bailey J (2016) Extending information theoretic validity indices for fuzzy clustering. IEEE Trans Fuzzy Syst 25(4):1013–1018. <https://doi.org/10.1109/TFUZZ.2016.2584644>
- Malvestuto FM (1986) Statistical treatment of the information content of a database. Inf Syst 11(2):211–233. [https://doi.org/10.1016/0306-4379\(86\)90029-3](https://doi.org/10.1016/0306-4379(86)90029-3)
- Meilă M (2007) Comparing clusterings, an information based distance. J Multivar Anal 98(5):873–895. <https://doi.org/10.1016/j.jmva.2006.11.013>
- Meilă M (2015) Criteria for comparing clusterings. In: Hennig C, Meilă M, Murtagh F, Rocci R (eds) Handb Cluster Anal. Chapman and Hall, New York, pp 619–636
- Milligan GW (1996) Clustering validation: results and implications for applied analyses. In: Arabie P, Hubert L, De Soete G (eds) Clustering and classification. World Scientific, River Edge, pp 341–375. https://doi.org/10.1142/9789812832153_0010
- Milligan GW, Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivar Behav Res 21:441–458. https://doi.org/10.1207/s15327906mbr2104_5
- Newman D, Hettich S, Blake C, Merz C (1998) UCI Repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Pfritznr D, Leibbrandt R, Powers D (2009) Characterization and evaluation of similarity measures for pairs of clusterings. Knowl Inf Syst 19:361–394. <https://doi.org/10.1007/s10115-008-0150-6>
- Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106. <https://doi.org/10.1007/BF00116251>
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(3):846–850. <https://doi.org/10.2307/2284239>
- Rendón E, Abundez I, Arizmendi A, Quiroz E (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5(1):27–34
- Rezaei M, Fränti P (2016) Set matching measures for external cluster validity. IEEE Trans Knowl Data Eng 28(8):2173–2186. <https://doi.org/10.1109/TKDE.2016.2551240>
- Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):623–656. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Steinley D (2004) Properties of the Hubert-Arabie adjusted Rand index. Psychol Methods 9(3):386–396. <https://doi.org/10.1037/1082-989X.9.3.386>
- Steinley D, Brusco MJ, Hubert LJ (2016) The variance of the adjusted Rand index. Psychol Methods 21(2):261–272. <https://doi.org/10.1037/met0000049>
- Strehl A, Ghosh J (2002) Cluster ensembles. A knowledge reuse framework for combining partitionings. J Mach Learn Res 3:583–617. <https://doi.org/10.1037/met0000049>
- Van Dongen S (2000) Performance criteria for graph clustering and markov cluster experiments. Tech. rep, National Research Institute for mathematics and computer science
- Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clustering comparison: variants, properties, normalization and correction for chance. J Mach Learn Res 11:2837–2854

- Warrens MJ (2008a) On similarity coefficients for 2x2 tables and correction for chance. *Psychometrika* 73(3):487–502. <https://doi.org/10.1007/s11336-008-9059-y>
- Warrens MJ (2008b) On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. <https://doi.org/10.1007/s00357-008-9023-7>
- Yao YY, Wong SKM, Butz CJ (1999) On information theoretic measures of attribute importance. In: Zhong N, Zhou L (eds) *Methodologies for Knowledge Discovery and Data Mining. PAKDD 1999. Lecture Notes in Computer Science*, vol 1574, Springer, Berlin, Heidelberg, pp 487–502, https://doi.org/10.1007/3-540-48912-6_18
- Zhang P (2015) Evaluating accuracy of community detection using the relative normalized mutual information. *J Stat Mech* 11(P11):006. <https://doi.org/10.1088/1742-5468/2015/11/P11006>