Generalized Information Theoretic Cluster Validity Indices for Soft Clusterings

Yang Lei, James C. Bezdek, Jeffrey Chan, Nguyen Xuan Vinh, Simone Romano and James Bailey
Department of Computing and Information Systems,
The University of Melbourne, Victoria, Australia
yalei@student.unimelb.edu.au, jcbezdek@gmail.com, {jeffrey.chan, vinh.nguyen, simone.romano, baileyj}@unimelb.edu.au

I. Abstract

There have been a large number of external validity indices proposed for cluster validity. One such class of cluster comparison indices is the information theoretic measures, due to their strong mathematical foundation and their ability to detect non-linear relationships. However, they are devised for evaluating crisp (hard) partitions. In this paper, we generalize eight information theoretic crisp indices to soft clusterings, so that they can be used with partitions of any type (i.e., crisp or soft, with soft including fuzzy, probabilistic and possibilistic cases). We present experimental results to demonstrate the effectiveness of the generalized information theoretic indices.

II. Introduction

Clustering is one of the most important unsupervised techniques. It aims to divide the data objects into several groups, so that the objects in the same group are similar whereas the objects in different groups are dissimilar. Clustering validation, which evaluates the goodness of a clustering, is a challenging task. Unlike supervised learning, which can be evaluated according to the given ground truth labels, for the validation of clustering, an unsupervised learning task, there is no good standard solution. On the other hand, clustering validation is a crucial task. It helps users to select the appropriate clustering parameters, like the number of clusters, or even the appropriate clustering model and algorithm. It can be used to compare clusterings as part of finding a consensus which can reduce the bias and errors of the individual clusterings [1]. There have been a large number of clustering validity measures proposed, which can be generally classified into two categories, internal clustering validation and external clustering validation [2]. They are distinguished in terms of whether or not external information is used during the validation procedure. In this paper, we focus on the external validation measures.

Most external validity indices compare two crisp partitions [2]. However, partitions can also be soft, i.e., fuzzy, probabilistic or possibilistic partitions [3]. Soft partitions are usually converted to crisp partitions by assigning each object unequivocally to the cluster with highest membership (fuzzy partitions), probability (probabilistic partitions), or typicality (possibilistic partitions). Then they are evaluated by employing the crisp external validity indices. However, this kind of conversion may cause loss of information [4]. For example, different soft partitions may be converted to the same crisp partition. Several methods have been proposed for generalizing some of the crisp indices to non-crisp cases [3]–[6]. The most general one of these methods in [3] can be utilized to generalize crisp indices, which are functions of the standard pair-based contingency table (see Table I), to soft indices. Subsequently the generalized soft indices can be utilized for comparing two partitions of any type.

Information theoretic measures form a fundamental class of measures for comparing pairs of crisp partitions. They have drawn considerable attention in recent years [1], [7], [8], due to their strong mathematical foundations and their ability to detect non-linear relationships. However, they are designed for comparing crisp clusterings and cannot be used to compare soft ones. Therefore, in this paper, we generalize eight information theoretic crisp cluster validity indices (including six similarity indices and two distance indices) to the soft case using the technique proposed in [3]. To our knowledge, this is the first work that generalizes *information theoretic* crisp indices to soft indices, i.e., *comparing soft clusterings*.

Mixture model based clustering methods have been widely used and have proved be useful in many real applications, e.g., image segmentation [9], document clustering [10], and information retrieval [11]. This class of approaches assume the data comes from a mixture of probability distributions (components), each probability distribution representing a cluster. One of these methods, the well-known expectation maximization (EM) algorithm has been successfully used for clustering [12]. In this paper, we employ the EM algorithm with Gaussian mixture to generate the soft partitions (probabilistic partitions). We employ the Gaussian mixture with EM as it is well understood, mathematically easy to work with, and has been shown to produce good results in many instances [13]. We test the effectiveness of the eight generalized soft information theoretic indices, in terms of their ability for indicating the correct number of components in synthetic datasets generated from various Gaussian mixtures, and real world datasets. Here, the "correct" number of components refers either to the known number of components in the mixture from which Gaussian clusters are drawn, or the number of classes in labeled ground truth partitions of real data. In this paper, our objective is to show that the generalized information theoretic measures of validity can be useful for choosing the best number of components.

Our contributions can be described as follows:

- We generalize eight information theoretic crisp cluster validity indices to soft indices.
- We demonstrate that the generalized information theoretic indices can be useful for choosing the number of components via experimental evaluation.

- We analyze the experimental results and recommend indices for different scenarios.

III. Technique for Soft Generalization

In this section, we introduce the concept of soft clustering and the technique that we used to generalize the crisp information theoretic indices to soft indices.

Let $O = \{o_1, \dots, o_n\}$ denote n objects, each of them associated with a vector $\mathbf{x_i} \in \Re^p$ in the case of numeric data. There are four types of class labels we can associate with each object: *crisp*, *fuzzy*, *probabilistic* and *possibilistic*. Let c denote the number of classes, $1 < c < n$, we define three sets of *label vectors* in $\Re^c$ as follows:

$$N_{pc} = \{\mathbf{y} \in \Re^c : \forall i\, y_i \in [0,1], \exists j\, y_j > 0\} \tag{1a}$$

$$N_{fc} = \{\mathbf{y} \in N_{pc} : \sum_{i=1}^{c} y_i = 1\} \tag{1b}$$

$$N_{hc} = \{\mathbf{y} \in N_{fc} : \forall i\, y_i \in \{0,1\}\} \tag{1c}$$

Here, N_{hc} is the canonical (unit vector) basis of $\Re^c$. The i-th vertex of N_{hc}, i.e., $\mathbf{e}_i = (0, 0, \dots, \underbrace{1}_{i}, \dots, 0)^T$, is the crisp label for class i, $1 \le i \le c$. The set N_{fc} is a piece of a hyperplane, and is the convex hull of N_{hc}. For example, the vector $\mathbf{y} = (0.1, 0.2, 0.7)^T$ is a constrained label vector in N_{f3}; its entries between 0 and 1 and sum to 1. There are at least two interpretations for the elements of N_{fc}. If $\mathbf{y}$ comes from a method such as maximum likelihood estimation in mixture decomposition, $\mathbf{y}$ is a (usually posterior) *probabilistic* label, and y_i is interpreted as the probability that, given $\mathbf{x}$, it is generated from the class or component i of the mixture [14]. On the other hand, if $\mathbf{y}$ is a label vector for some $\mathbf{x} \in \Re^p$ generated by, say, the fuzzy c-means clustering model [15], $\mathbf{y}$ is a *fuzzy* label for $\mathbf{x}$, and p_i is interpreted as the membership of $\mathbf{x}$ in class i. An important point for this paper is that N_{fc} has the same structure for probabilistic and fuzzy labels. Finally, $N_{pc} = [0,1]^c \backslash \{\mathbf{0}\}$ is the unit (hyper)cube in $\Re^c$, *excluding the origin*. As an example, vectors such as $\mathbf{z} = (0.7, 0.3, 0.6)^T$ in N_{p3} are called *possibilistic* label vectors, and in this case, z_i is interpreted as the possibility that $\mathbf{x}$ is generated from class i. Labels in N_{pc} are produced, e.g., by possibilistic clustering algorithms [16]. Note that $N_{hc} \subset N_{fc} \subset N_{pc}$.

Clustering in unlabeled data is the assignment of one of three types of labels to each object in O. We define a partition of X on n objects as a $c \times n$ matrix $U = [\mathbf{U}_1 \dots \mathbf{U}_k \dots \mathbf{U}_n] = [u_{ik}]$, where $\mathbf{U}_k$ denotes the k-th column of U and u_{ik} indicates the degrees of membership of object k belongs to cluster i. The label vectors in equations (1) can be used to define three types of c-partitions:

$$M_{pcn} = \{U \in \Re^{cn} : \forall k\, \mathbf{U}_k \in N_{pc}, \forall i \sum_{k=1}^{n} u_{ik} > 0\} \tag{2a}$$

$$M_{fcn} = \{U \in M_{pcn} : \forall k\, \mathbf{U}_k \in N_{fc}\} \tag{2b}$$

$$M_{hcn} = \{U \in M_{fcn} : \forall k\, \mathbf{U}_k \in N_{hc}\} \tag{2c}$$

where M_{pcn} (2a) are possibilistic c-partitions, M_{fcn} (2b) are fuzzy or probabilistic c-partitions, and M_{hcn} (2c) are crisp (hard) c-partitions. For convenience, we call the set $M_{pcn} \setminus M_{hcn}$ as the *soft c-partitions* of O, which contains the fuzzy, probabilistic and possibilistic c-partitions and excludes the crisp c-partitions.

TABLE I. Contingency table and formulas used to compare crisp partitions U and V

			Partition V $\mathbf{V_j}$ = row j of V					
	Class		$\mathbf{v_1}$	$\mathbf{v_2}$	$\dots$	$\mathbf{v_c}$		Sums
Partition U $\mathbf{U}_i$ = row i of U	$\mathbf{U}_1$	$N =$	n_{11}	n_{12}	$\dots$	n_{1c}	$= UV^T$	$n_{1\bullet}$
	$\mathbf{U}_2$		n_{21}	n_{22}	$\dots$	n_{2c}		$n_{2\bullet}$
	$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$\mathbf{U}_r$		n_{r1}	n_{r2}	$\dots$	n_{rc}		$n_{r\bullet}$
	Sums		$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet c}$		$n_{\bullet\bullet} = n$

The traditional external cluster validity indices are designed for comparing two crisp partitions [2], among which there are a number of popular indices [3] that are built upon the standard pair-based contingency table. Let $U \in M_{hrn}$ and $V \in M_{hcn}$, the $r \times c$ contingency table of two crisp partitions U and V is shown in Table I. For soft partitions, work by Anderson et al. in [3] proposes the formation of the contingency table by the product $N = UV^T$. For crisp partitions, this formation reduces to the regular contingency table. Based on this formation, Anderson et al. propose generalizations of crisp indices for use with soft partitions (fuzzy, probabilistic or possibilistic partitions). These soft generalizations are applicable for any index that depends only on the entries of the contingency table and can be described as follows:

$$N^* = \phi UV^T = \Big[n / \sum_{i=1}^{r} n_{i\bullet}\Big] UV^T \tag{3}$$

where ϕ is a scaling factor, used to normalize the possibilistic indices to the range $[0,1]$ and $n_{i\bullet} = \sum_{j=1}^{c} nij$ (see Table I). Note that in the cases of crisp, fuzzy or probabilistic partitions, $\phi = 1$. In this work, we do not discuss the possibilistic case and leave it for future work. Crisp indices that are based solely on the entries in the contingency table N can be generalized by using the generalized contingency table N^*.

IV. Crisp Information Theoretic Indices and Soft Generalization

Information theoretic based measures are built upon fundamental concepts from information theory [17], and are a commonly used approach for crisp clustering comparison [7], [8]. This is because of their strong mathematical foundations and their ability to detect non-linear relationships. We first introduce some of the fundamental concepts. The information entropy of a discrete random variable $S = \{s_1, \dots, s_n\}$ is defined as:

$$H(S) = -\sum_{s \in S} p(s) \log p(s) \tag{4}$$

where $p(s)$ is the probability $p(S = s)$. The entropy is a measure of uncertainty of a random variable. Then, the *mutual information* (MI) between two random variables, S and T, is defined as follows:

$$I(S,T) = \sum_{s \in S} \sum_{t \in T} p(s,t) \log \frac{p(s,t)}{p(s)p(t)} \tag{5}$$

TABLE II. INFORMATION THEORETIC-BASED CLUSTER VALIDITY INDICES

Name	Expression	Range	Find
MI	I(U,V)	$[0, \min\{H(U), H(V)\}]$	Max
NMI$_{\text{joint}}$	$\frac{I(U,V)}{H(U,V)}$	[0,1]	Max
NMI$_{\text{max}}$	$\frac{I(U,V)}{\max\{H(U),H(V)\}}$	[0,1]	Max
NMI$_{\text{sum}}$	$\frac{2I(U,V)}{H(U)+H(V)}$	[0,1]	Max
NMI$_{\text{sqrt}}$	$\frac{I(U,V)}{\sqrt{H(U)H(V)}}$	[0,1]	Max
NMI$_{\text{min}}$	$\frac{I(U,V)}{\min\{H(U),H(V)\}}$	[0,1]	Max
Variation of Information (VI)	$H(U,V) - I(U,V)$	$[0, \log n]$	Min
Normalized VI (NVI*)	$1 - \frac{I(U,V)}{H(U,V)}$	[0,1]	Min

* NVI is the normalized distance measure equivalent to NMI$_{\text{joint}}$

where $p(s,t)$ is the joint probability $p(S = s, T = t)$. The MI measures the information shared between two variables. Intuitively, it tells us how similar these two variables are. Next, we will introduce some theoretic concepts in the clustering context.

Given one crisp partition $U = \{u_1, \ldots, u_c\}$ with c clusters on O, the entropy of U is $H(U) = -\sum_{i=1}^{c} p(u_i) \log p(u_i)$, where $p(u_i) = \frac{|u_i|}{n}$ indicates the probability of an object belonging to cluster u_i. Given two crisp partitions U and V, their entropies, joint entropy and *mutual information* (MI) can be defined according to the contingency table built upon U and V (Table I) respectively as [1]:

$$
\begin{aligned}
H(U) &= -\sum_{i=1}^{r} \frac{n_{i\bullet}}{n} \log \frac{n_{i\bullet}}{n}, \\
H(U,V) &= -\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n}, \\
I(U,V) &= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{n} \log \frac{n_{ij}/n}{n_{i\bullet} n_{\bullet j}/n^2}.
\end{aligned}
$$

Intuitively, the joint entropies measure how similar two clusterings are, by comparing the distribution of the cluster assignments of the points across the two clusterings. More detailed explanations of these concepts can be found in [1], [8].

Eight popular crisp information theoretic indices based on the above basic concepts are listed in Table II. The variants of *normalized mutual information* (NMI) (NMI$_{\{*\}}$ in Table II) are distinguished in terms of their different normalization factors, which all aim at scaling MI to range $[0, 1]$. The *variation of information* (VI) [8] has been proved to be a true metric on the space of clusterings. The normalized version of VI (NVI) ranges in $[0, 1]$.

In this paper, we generalize the above information theoretic concepts to compare soft clusterings. We define the entropy of a soft clustering U, as $H(U) = -\sum_{i=1}^{r} n_{i\bullet}/n \log(n_{i\bullet}/n)$, where $n_{i\bullet}$ is row sum from the generalized contingency table N^*. In the soft clustering setting, $n_{i\bullet}$ can be regarded as the probability that a point belongs to the i-th cluster. Similarly, we define the joint entropy of two soft clusterings, U and V, as $H(U,V) = -\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}/n \log(n_{ij}/n)$, where n_{ij} is taken from N^*, representing the joint probability that a point belongs to U_i and V_j. Finally, we define $I(U,V) = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}/n \log\left((n_{ij}/n)/(n_{i\bullet} n_{\bullet j}/n^2)\right)$. From this generalization, all the eight information theoretic indices can be computed from the generalized contingency table N^*.

V. EVALUATION OF THE SOFT COMPARISON INDICES

A. Synthetic data

We evaluate the eight generalized soft indices described in Table II on several synthetic and real world datasets.

There are many possible settings to generate the synthetic datasets for testing. It is popular to use Gaussian distributed data as synthetic data for clustering validation [1], [3]. In this paper we generate synthetic datasets from a population whose probability density function is a mixture of c bivariate normal distributions. We design the datasets in terms of two factors: overlapping and density (dens). Table III summarizes the model parameters for data generation, where data3 are the datasets with $c = 3$ and data5 are datasets with $c = 5$. The total number of objects for each dataset is $n = 1000$. The means of different components, $\{\mu_i\}$, were distributed on rays separated by equal angles according to the different number of clusters, centred equal distances from the origin at various radii r. For example, at $c = 3$, there are means, $\{\mu_i\}$, on lines separated by $120\,^\circ$, each r units from $(0,0)^T$. We vary the overlapping degrees of the clusters by varying the value of r with fixed priors $\pi_i = 1/c$. The density property of the datasets is considered by altering the prior of the first component π_1 (dens) with $\{c = 3, r = 5\}$ or $\{c = 5, r = 6\}$. After choosing the prior for the first component, then the other priors are $\pi_i = (1 - \pi_1)/(c - 1)$, where $1 < i \leq c$. The covariance matrices, $\{\Sigma_i\}$, are identity matrices. Thus, there are nine datasets for each c and 18 synthetic datasets in total. Figure 1 shows an example of scatter plots for four datasets, data3 with $\{r = 1, dens = 1/3\}$ and $\{r = 5, dens = 1/3\}$, and data5 with $\{r = 2, dens = 1/5\}$ and $\{r = 6, dens = 1/5\}$.

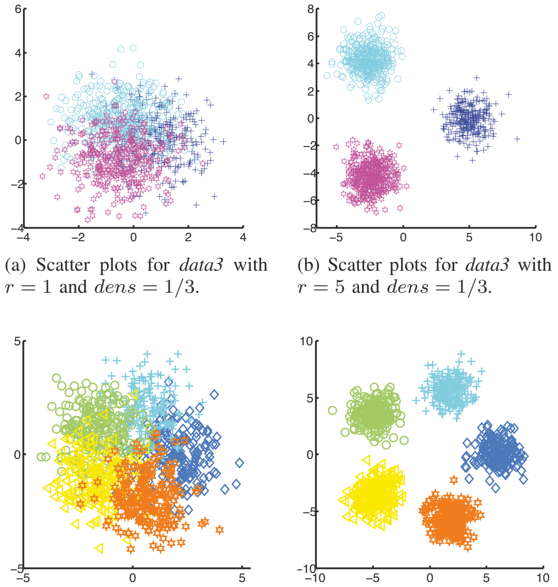
B. Real world data

Datasets from the UCI machine learning repository [18] are a typical benchmark for evaluating validity measures. They also have known classes which provide the ‘true’ number of classes. We use seven real datasets from the UCI repository and their details are shown in Table IV, where n, d and c correspond to the number of objects, features and classes, respectively.

C. Computing protocols

We modified a MATLAB implementation of the EM algorithm[1], according to our initialization and termination criteria which are described as follows.

[1] http://www.dcorney.com/ClusteringMatlab.html



(a) Scatter plots for *data3* with $r = 1$ and $dens = 1/3$.

(b) Scatter plots for *data3* with $r = 5$ and $dens = 1/3$.

(c) Scatter plots for *data5* with $r = 2$ and $dens = 1/5$.

(d) Scatter plots for *data5* with $r = 6$ and $dens = 1/5$.

Fig. 1. Four scatter plots for data3 and data5 datasets.

Initialization of EM: we randomly draw c points from the data X as the initial means of clusters. The initial covariance matrices for c clusters are diagonal, where the i^{th} element on the diagonal is the variance of the i^{th} feature vector of X; the initial prior probabilities are $1/c$.

Termination of EM: We define the termination criterion by considering the difference between two successive estimates of the means, $\|M_{t+1} - M_t\|_\infty < \varepsilon$, where $M_t = \{\mathbf{m_1}, \ldots, \mathbf{m_c}\}$, and $\varepsilon = 10^{-3}$; the maximum number of iterations is 100.

D. Experimental Design

We test the effectiveness of the eight generalized soft indices[2] by considering their ability to estimate the number of clusters (components) of the test datasets. Also, we compare these generalized soft indices with one non-information theoretic index, the soft version of the adjusted Rand Index (ARI) [3], to provide a comparison with the class of pair-based measures. The general idea is to run the EM algorithm over each dataset to generate a set of partitions with different number of clusters. Then, each of the eight generalized soft indices is computed for these partitions. The number of clusters with the partition obtaining the best results is considered as the predicted 'true' number of clusters, c_{pre}, for that particular dataset. Let c_{true} indicates the number of known clusters in the Gaussian datasets, or the number of labeled classes in the real world datasets. If $c_{pre} = c_{true}$, then we believe the prediction of this index on this dataset is successful. However, sometimes the number of 'true' clusters, c_{true}, may not correspond to the number of "apparently best" clusters, c_{pre}, found by a computational clustering algorithm. A possible reason is that the clustering algorithm failed to detect the underlying substructure of the data, rather than the inability of the index. For example, an algorithm, which is designed to look for spherical clusters, cannot detect elongated clusters. We will observe this phenomenon in several of the following results. More specifically, we run the EM algorithm on each dataset to generate a set of soft partitions with the number of clusters ranging from 2 to $2 \times c_{true}$ for synthetic datasets, and ranging from 2 to $3 \times c_{true}$ for real datasets. In order to reduce the influence of the random initialization for the EM algorithm, we generate 100 partitions for each c, and evaluate these soft indices based on these partitions.

We designed two sets of experiments for the evaluation. In the first set of experiments, we compare the generated partitions (soft partitions) against the ground truth cluster labels for synthetic datasets or class labels for real world datasets (crisp partitions). In detail, for the 100 partitions generated by the EM algorithm with respect to each c, we keep track of the percentage of the successful predictions (success rate) achieved by each index (e.g., Figure 3a). The success means that $c_{pre} = c_{true}$. Alternatively, as another measure, we also compute the average values over 100 partitions for each index with respect to each c (e.g., Figure 6b).

In the second set of experiments, we do not consider the ground truth labels and use the *consensus index* (CI) [1] to evaluate the eight generalized soft indices. For some of the datasets, the gold standard number of clusters c_{true} may not be the number of clusters that a typical algorithm (or human) would determine. For example, in data3 with $r = 1$ and $dens = 1/3$ (Figure 1a), visually there only appears to be one cluster, as the three generated clusters are highly overlapping. Hence, as an additional measure to give insights about the performance of these measures in these scenarios, we introduce the CI measure which does not use the ground truth labels. The CI is built on the idea of consensus clustering [7] which aims to produce a robust and high quality representative clustering by considering a set of partitions generated from the same dataset. The empirical observation according to work [1] is that, in regard to the set $\mathcal{U}_c$ of candidate partitions for a particular value of c, when the specified number of clusters coincides with the true number of clusters c_{true}, $\mathcal{U}_c$ has a tendency to be less diverse. Based on this observation, CI, built upon a suitable clustering similarity(distance) comparison measure, is used to quantify the diversity of $\mathcal{U}_c$. The definition of CI is described as follows: suppose a set of L clustering solutions (crisp or soft), $\mathcal{U}_c = \{U_1, U_2, \ldots, U_L\}$ have been generated, each with c clusters. The *consensus index* (CI) of $\mathcal{U}_c$ is defined as:

$$CI(\mathcal{U}_c) = \frac{\sum_{i<j} AM(U_i, U_j)}{L(L-1)/2} \tag{6}$$

where the *agreement measure* (AM) is a suitable clustering comparison index (similarity index or distance index). In this paper, we used the six max-optimal similarity indices and two min-optimal distance indices listed in Table II as the AM. Thus, the CI quantifies the average pairwise agreement in $\mathcal{U}_c$. The optimal number of clusters, c_{pre}, is chosen as the number with the maximum CI (as AM is a similarity index, or the minimum CI as AM is a distance index), i.e.,

[2] as NVI is equivalent to the $\text{NMI}_{\text{joint}}$, we just show and analyze the performance of $\text{NMI}_{\text{joint}}$.

TABLE III. SYNTHETIC DATASETS INFORMATION

c	Priors $\{\pi_i\}$	Means $\{\mu_i\}$	$\{\Sigma_i\}$	n	Name
3	$\pi_1 = 1/6$ $\pi_1 = 1/3$ $\pi_1 = 1/2$ $\pi_1 = 2/3$ $\pi_1 = 5/6$	$r = 1$ $r = 2$ $r = 3$ $r = 4$ $r = 5$	I	1000	data3
5	$\pi_1 = 1/10$ $\pi_1 = 1/5$ $\pi_1 = 3/10$ $\pi_1 = 2/5$ $\pi_1 = 1/2$	$r = 2$ $r = 3$ $r = 4$ $r = 5$ $r = 6$	I	1000	data5

TABLE IV. REAL WORLD DATASETS INFORMATION

Dataset	n	d	c
sonar	208	60	2
pima-diabetes	768	8	2
heart-statlog	270	13	2
haberman	306	3	2
wine	178	13	3
vehicle	846	18	4
iris	150	4	3

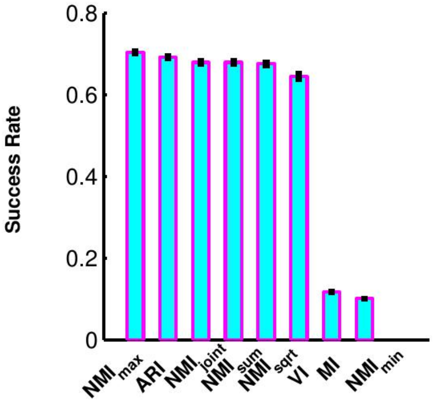


Fig. 2. Overall success rates on synthetic datasets. The error bars indicate standard deviation. These indices are shown in descending order in terms of their success rates.

$c_{pre} = \arg\max_{c=2\ldots c_{max}} CI(\mathcal{U}_c)$, where $c_{max} = 2 \times c_{true}$ for synthetic datasets, $c_{max} = 3 \times c_{true}$ for real world datasets. Specifically, we compute the CI values for the $L = 100$ generated soft partitions by the EM algorithm with respect to each value of c.

E. Simulation results - Success rate

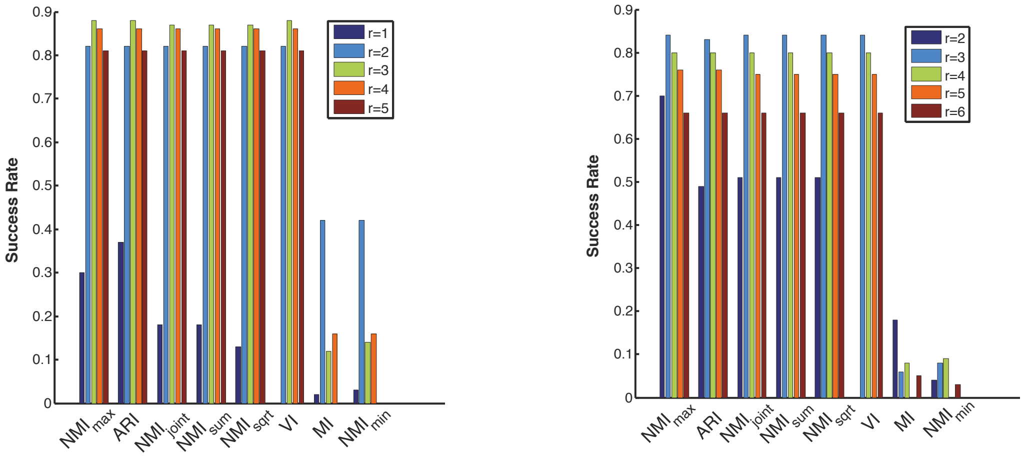
In this set of experiments, we present and analyze the experimental results with regard to the success rates of these indices on synthetic and real world datasets.

1) Synthetic datasets: First, to gain an overall comparison of the measures across the synthetic datasets, we show the overall success rates for the eight generalized soft indices, including seven information theoretic ones and ARI, in Figure 2. The overall success rate for an index is computed as the total number of successes across the 18 synthetic datasets divided by the total number of partitions, i.e., 18 × 100. The indices are sorted in descending order in terms of their success rates. The graph shows that the first six soft indices, NMI_{max}, ARI, $\text{NMI}_{\text{joint}}$, NMI_{sum}, NMI_{sqrt} and VI perform similarly well and achieved a success rate of around 70%. In contrast the soft MI and NMI_{min} do not perform well and only have a success rate of around 10%. We hypothesize the reason for this is that MI monotonically increases with the number of clusters c [1]. Hence, MI tends to favour clusterings with more clusters. For NMI_{min}, we found that $H(U)$, the entropy of the generated soft partitions, increases as c increases which is because the distribution of clusters is more balanced. The entropy of the ground truth labels $H(V)$ is constant q. At some c, $H(U) > H(V)$, and subsequently, $NMI_{min}(U, V) = MI(U, V)/H(V) = MI/q$, which means NMI_{min} has became equivalent to the scaled version of MI and has the same deficiencies as it. Next we show more detailed experimental results for synthetic datasets with respect to the overlapping and density factors. For the convenience of comparison, we will keep the order of the indices shown in Figure 2 in the following graphs.

We show the results on data3 and data5 with various overlapping settings in Figure 3. From these graphs, we find out that the performance of these indices on these datasets is consistent with their overall performance shown in Figure 2. The first six generalized soft indices all have similar and good success rates. In contrast, the soft MI and NMI_{min} perform poorly. This observation suggests that the comparison of the eight generalized soft indices is not affected by the overlapping factor of the datasets. Furthermore, we can observe that for the datasets containing clusters with higher overlapping degree ($r = 1$ with data3, $r = 2$ with data5), the success rates of the first six generalized soft indices are relatively low compared to the datasets with less overlap, which is not surprising. This is because the quality of the partitions generated by the EM algorithm on these higher overlap datasets has poor quality. The clusters in datasets with $r = 1$ for data3, or $r = 2$ for data5, are highly overlapped and the scatter plots of these two datasets are just like a big dense cluster (Figure 1a and Figure 1c). With the decreasing overlap (increasing r), the first six indices work better and have similar performance on these datasets.

Next, we present the results on the datasets with various density settings in Figure 4. Firstly, we can find out that the general performance of these indices on the datasets with various density settings conform with their overall performance ranking shown in Figures 2 and 3. Thus, the results suggest that density of the datasets also does not affect the relative success rate ranking of these indices. In addition, increasing the imbalance of the clusters in the datasets (increasing the density of the first cluster), decreases success rates of these indices. This reflects the fact that EM partitions on these imbalanced datasets are of poor quality.

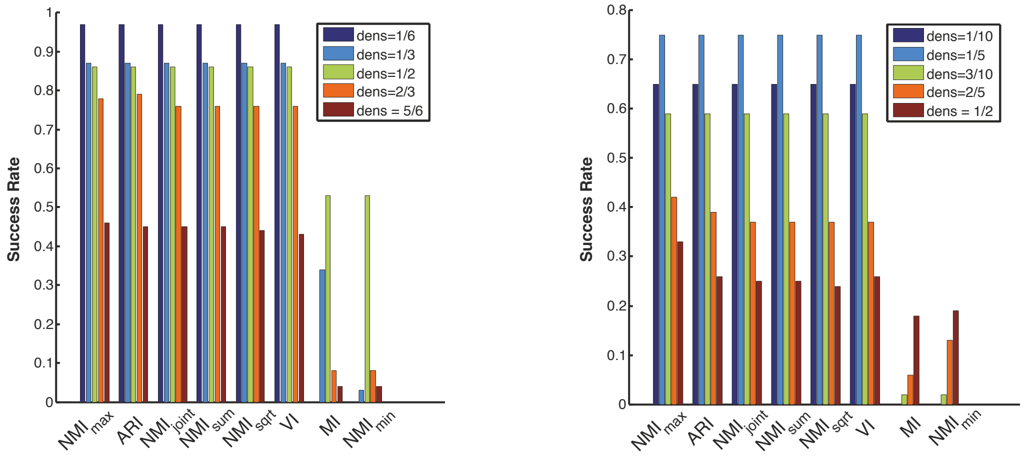
2) Real world datasets: The experimental results on the seven real world datasets are shown in Figure 5. We first show the overall success rates over all the real datasets in Figure 5a. The most striking observation from this graph is that VI works very well compared to all the other seven indices. In addition, ARI behaves worse than the information theoretic soft indices. We hypothesize that this is because the information theoretic measures are good at distinguishing



(a) Success rates on data3 with various overlapping settings.

(b) Success rates on data5 with various overlapping settings.

Fig. 3. Success rates of generalized soft indices on synthetic datasets with various overlapping settings.



(a) Success rates on data3 with various density settings.

(b) Success rates on data5 with various density settings.

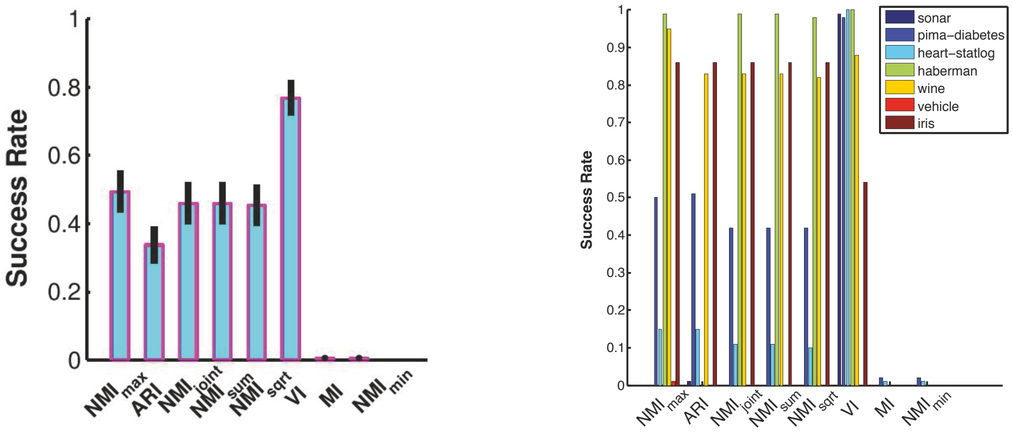
Fig. 4. Success rates of generalized soft indices on synthetic datasets with various density settings.

clusters with non-linear relationships [1], and real datasets are more likely to have these. Comparing Figures 2 and 5a shows that the two worst indices (MI and NMI_{min}) are $6 - 7$ times less reliable than the six good ones for the synthetic datasets, but are essentially not very useful for the real datasets. The more detailed results with respect to these indices on different real world datasets are shown in Figure 5b. The first six indices generally work well on the three datasets 'haberman', 'wine' and 'iris', except ARI perform poorly on the 'haberman' dataset. VI performs well on these datasets, as well as the three other datasets 'sonar', 'pima-diabetes' and 'heart-statlog'. None of the indices responds well to the 'vehicle' data, possibly because the labeled subsets do not form computationally recognizable clusters for the EM algorithm in the 18 dimensional feature space and are likely to require appropriate feature selection before clustering.

F. Simulation results - Consensus Index

In this set of experiments, we employ the *consensus index* (CI) coupled with the eight generalized soft indices as AM for testing the effectiveness of these soft indices. For further confirmation of the effectiveness of these measures, we present another set of experimental results coupled with CI, which are the average results of these generalized soft indices over 100 partitions for each c. Notice that the CI experiments compare pairwise generated soft partitions without the help of ground truth labels, and the average results compare generated soft partitions against the ground truth labels. For brevity, we discuss only a few representative results.

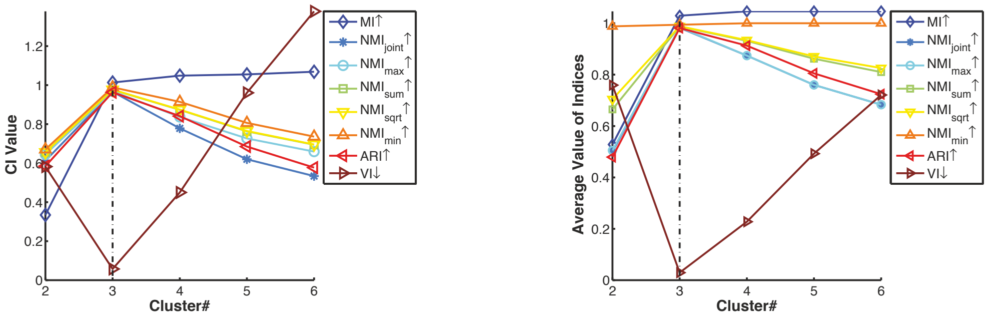
In Figure 6a, we show the CI values for the data3 with parameter setting $\{r = 5, dens = 1/6\}$, with c ranging from 2 to 6. The up-arrow (↑) (down-arrow (↓)) besides each index in the legends means that a larger (smaller) value of that index



(a) Overall success rates on real world datasets. The error bars indicate the standard deviation.

(b) Success rates on different real datasets.

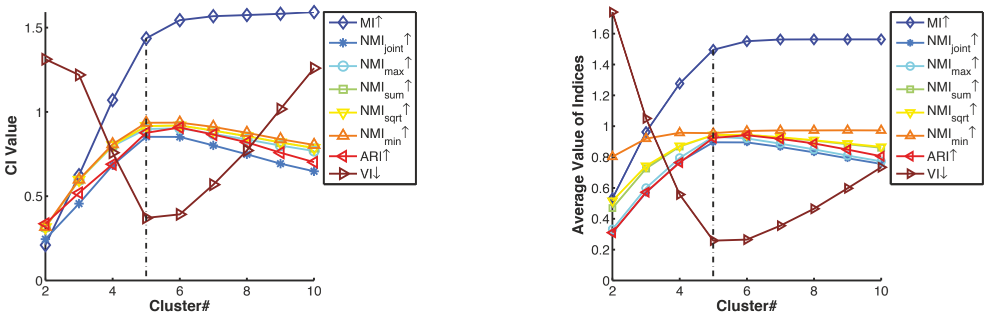
Fig. 5. Success Rates on real world datasets.



(a) The CI results on data3 with $dens = 1/6$ and $r = 5$.

(b) The average results on data3 with $dens = 1/6$ and $r = 5$.

Fig. 6. The CI results and average results on data3 with $dens = 1/6$ and $r = 5$. The dashed line on the x axis indicates the true number of clusters in the cluster labels. The up-arrow (↑) (down-arrow (↓)) besides each index in the legends means that a larger (smaller) value of that index indicates a 'better' partition.

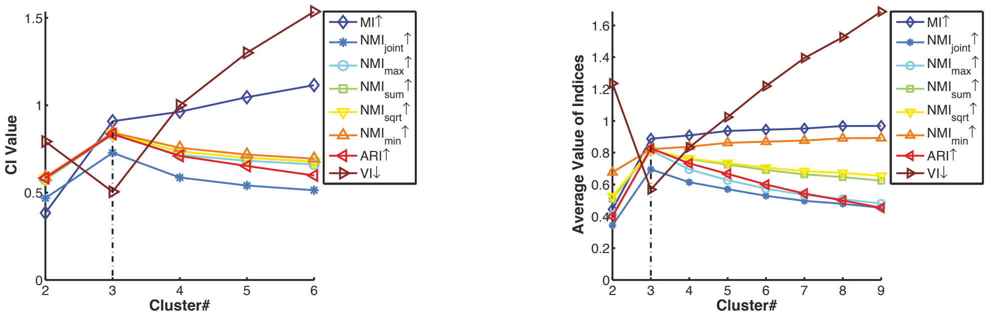


(a) The CI results on data5 with $dens = 1/5$ and $r = 5$.

(b) The average results on data5 with $dens = 1/5$ and $r = 5$.

Fig. 7. The CI results and average results on data5 with $dens = 1/5$ and $r = 5$. The dashed line on the x axis indicates the true number of clusters in the cluster labels. The up-arrow (↑) (down-arrow (↓)) besides each index in the legends means that a larger (smaller) value of that index indicates a 'better' partition.

indicates a 'better' partition. As we can see, except MI, all the variants of the NMI indices and ARI achieve maximum values, and VI get the minimum values at the correct number of clusters, i.e., $c_{pre} = c_{true} = 3$. The average results of the eight generalized soft indices are presented in Figure 6b. We can see that all the generalized indices, except MI and NMI$_{\text{min}}$,



(a) The CI results on wine dataset.

(b) The average results on wine dataset.

Fig. 8. The CI results and average results on wine dataset. The dashed line on the x axis indicates the true number of classes in the class labels. The up-arrow (↑) (down-arrow (↓)) besides each index in the legends means that a larger (smaller) value of that index indicates a 'better' partition.

find the correct number of components of the data.

Figures 7a and 7b show the CI values and the average results on data5 with parameter setting $\{r = 5, dens = 1/5\}$, respectively. The different observation from that in Figure 6 is that, some of these indices (e.g., NMI$_{\text{joint}}$ and VI) may be confused with the number of clusters at $c_{pre} = 5$ or $c_{pre} = 6$, and ARI slightly prefer $c_{pre} = 6$, while $c_{true} = 5$. However, we can tell from these two graphs that both of these two sets of experiments show this problem. Thus, we may hypothesize that the partitions generated by the EM algorithm on this dataset with $c = 5$ and $c = 6$ are ambiguous. To sum up, these two sets of experiments, that is, CI (without ground truth labels) and average results (compare against the ground truth labels), show that these indices work well.

The CI values on the real dataset wine are shown in Figure 8a. Similar to the results shown in Figure 6a, all variants of the NMI, as well as ARI and VI, successfully discover the right number of clusters. In Figure 8b, all the indices except the MI and NMI$_{\text{min}}$, also are able to find the right number of clusters in the data.

VI. Conclusion

In this paper, we generalized eight well known crisp information-theoretic indices to compare soft clusterings. We tested the soft generalizations on probabilistic clusters found by the EM algorithm in 18 synthetic sets of Gaussian clusters and seven real world data sets with labeled classes, and also compared them with one non-information theoretic index ARI. Overall, six of the eight soft indices return average success rates in the range of $50 - 70\%$ and they have higher success rates than ARI on the real datasets. Our numerical experiments suggest that the soft VI index is perhaps the best of the eight over both kinds of data; and that MI and NMI$_{\text{min}}$ should almost certainly be avoided. Our next effort will be towards expanding both theory and tests of soft information-theoretic indices in the direction of the other popular approach to soft clustering-viz., fuzzy clustering.

References

[1] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 9999, pp. 2837–2854, 2010.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988.

[3] D. T. Anderson, J. C. Bezdek, M. Popescu, and J. M. Keller, "Comparing fuzzy, probabilistic, and possibilistic partitions," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 5, pp. 906–918, 2010.

[4] R. J. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, 2007.

[5] E. Hüllermeier and M. Rifqi, "A fuzzy variant of the rand index for comparing clustering structures." in *IFSA/EUSFLAT Conf.*, 2009, pp. 1294–1298.

[6] R. K. Brouwer, "Extending the rand, adjusted rand and jaccard indices to fuzzy partitions," *Journal of Intelligent Information Systems*, vol. 32, no. 3, pp. 213–235, 2009.

[7] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[8] M. Meilă, "Comparing clusteringsan information based distance," *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[9] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, "A spatially constrained mixture model for image segmentation," *IEEE Transactions on Neural Networks*, vol. 16, no. 2, pp. 494–498, 2005.

[10] S. Zhong and J. Ghosh, "Generative model-based document clustering: a comparative study," *Knowledge and Information Systems*, vol. 8, no. 3, pp. 374–384, 2005.

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.

[12] J. C. Bezdek, W. Li, Y. Attikiouzel, and M. Windham, "A geometric approach to cluster validity for normal mixtures," *Soft Computing*, vol. 1, no. 4, pp. 166–179, 1997.

[13] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., 2005.

[14] D. M. Titterington, A. F. Smith, U. E. Makov *et al.*, *Statistical analysis of finite mixture distributions*. Wiley New York, 1985, vol. 7.

[15] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[16] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[18] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml