Extending Information-Theoretic Validity Indices for Fuzzy Clustering

Yang Lei, James C. Bezdek, *Life Fellow, IEEE*, Jeffrey Chan,
Nguyen Xuan Vinh, Simone Romano, and James Bailey

***Abstract*—Previously, eight popular information-theoretic based cluster validity indices have been generalized and tested for probabilistic partitions built by the *expectation-maximization* (EM) algorithm for the Gaussian mixture model. But the analysis was limited to probabilistic clusters and there were limited explanations for differences in the performance of the indices. In this paper, we extend the tests to partitions found by *fuzzy c-Means* (FCM) and provide further explanations and insights about the performance of these indices. Of the eight generalized indices, we advocate a normalized version of the soft mutual information cluster validity index ($\mathrm{NMI}_{\mathrm{sM}}$) as the best overall choice, as it outperforms the other seven indices for both FCM and EM according to our tests on synthetic and real data. The superiority of $\mathrm{NMI}_{\mathrm{sM}}$ is most pronounced for datasets with overlapped and/or varying sized clusters. Finally, we provide a theoretical analysis which helps explain the superior performance of $\mathrm{NMI}_{\mathrm{sM}}$ compared to the other three normalizations of soft mutual information.**

***Index Terms*—Soft Cluster Validity, External Validity Indices, Fuzzy c-Means, Mutual Information.**

I. Introduction

CLUSTERING attempts to divide data representing objects into several groups, so that objects in the same group are similar whereas the objects in different groups are dissimilar. *Cluster validity indices* (CVIs) are used to evaluate the quality of clusterings (partitions) generated by clustering algorithms. There have been a large number of CVIs proposed, which are either internal or external [1]. They are distinguished by whether or not external information is used during the validation procedure. The CVIs studied in this article are external validity measures.

Most external validity indices compare two crisp partitions [1]. However, partitions can also be soft, i.e., fuzzy, probabilistic or possibilistic [2]. One approach to evaluating soft partitions is to "harden" them to crisp partitions by assigning each object to the cluster with highest membership (fuzzy partitions), posterior probability (probabilistic partitions), or typicality (possibilistic partitions). Then they are evaluated with crisp external validity indices. However, hardening may cause loss of information[3], as an infinite number of different soft partitions can be converted to the same crisp partition. Hence, several methods have been proposed for generalizing some crisp CVIs to non-crisp cases [2], [3], [4], [5]. A method reported in [2] can be used to generalize any CVI which is a function of the standard contingency table (see Table I), to soft indices. Subsequently the generalized soft indices can be utilized for comparing two partitions of any type. All of these papers compared fuzzy generalizations of some pair-counting based external CVIs, and tested them on fuzzy partitions. But none of these papers discussed soft generalizations of information-theoretic based CVIs.

Information-theoretic measures form a fundamental class of measures for comparing pairs of crisp partitions and have been shown to outperform other classes of comparison measures in certain common scenarios [6], [7], [8]. However, the CVIs discussed in those papers are designed for comparing crisp partitions and cannot compare soft ones. Therefore, the authors of [9] used the method developed in [2] to generalize eight *information-theoretic CVIs* (IT-CVIs) discussed in [8]. Their study demonstrated the effectiveness of the generalized soft indices on probabilistic clusters found by the *expectation-maximization* (EM) algorithm applied to the *Gaussian mixture decomposition* (GMD) problem. However, [9] only provided brief explanations and insights about the performance of the measures, hence it is difficult to know why any particular index should be selected under different circumstances. In addition, [9] is limited to soft partitions generated by the EM algorithm, so we do not know how these indices perform on different types of soft partitions. And the effectiveness of the generalized measures were only demonstrated on relatively small datasets (up to 1000 data objects), so we do not know if they are still effective on large datasets.

In this paper, we extend the analysis and validation studies of the eight IT-CVIs to fuzzy partitions generated by another popular algorithm, *fuzzy c-means* (FCM) [10]. In addition, we provide a theoretical analysis that partially explains the performance of the measures. And we test and demonstrate the effectiveness of the generalized indices on relatively large datasets [1]. Our contributions can be summarized as follows: (i) We demonstrate that the generalized information-theoretic indices can be effective on fuzzy partitions generated by FCM via experimental evaluation; (ii) We test and demonstrate the effectiveness of the generalized measures on relatively large datasets; (iii) We analyze the experimental results and recommend a normalized version of the soft mutual information cluster validity index ($\mathrm{NMI}_{\mathrm{sM}}$) as the IT-CVI which generally performs better than the other seven soft information-theoretic measures for FCM partitions generated from datasets with

Yang Lei, James C. Bezdek, Nguyen Xuan Vinh, Simone Romano and James Bailey are with the Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia, 3010. E-mail: yalei@student.unimelb.edu.au, {jbezdek, vinh.nguyen, simone.romano, baileyj} @unimelb.edu.au.

Jeffrey Chan is with the School of Science (Computer Science and Information Technology), RMIT University, Victoria, Australia, 3000. He conducted part of the work while at the University of Melbourne. Email: jeffrey.chan@rmit.edu.au.

Manuscript received ***; revised ***.

[1] The code and more detailed information about this work are available at https://sites.google.com/site/yldatascience/home/tfs2016.

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

			Partition V $\mathbf{v_j}$ = row j of V					
	Class		$\mathbf{v_1}$	$\mathbf{v_2}$	$\dots$	$\mathbf{v_r}$		Sums
Partition U $\mathbf{u}_i$ = row i of U	$\mathbf{u}_1$	$N =$	n_{11}	n_{12}	$\dots$	n_{1r}	$= UV^T$	$n_{1\bullet}$
	$\mathbf{u}_2$		n_{21}	n_{22}	$\dots$	n_{2r}		$n_{2\bullet}$
	$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$\mathbf{u}_c$		n_{c1}	n_{c2}	$\dots$	n_{cr}		$n_{c\bullet}$
	Sums		$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet r}$		$n_{\bullet\bullet} = n$

overlapping and/or different sized clusters; (iv) We prove a theorem which helps explain why the soft NMI$_{\text{sM}}$ performs better than three normalized versions of soft mutual information, namely, NMI$_{\text{sj}}$, NMI$_{\text{ss}}$ and NMI$_{\text{sr}}$.

II. Background

A. Technique for Soft Generalization

A partition of X on n objects is 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 degree of membership of object k in cluster i. There are three types of c-partitions: $M_{pcn} = \{U \in \Re^{cn} | \forall i, k, u_{ik} \in [0,1], \forall i \sum_{k=1}^{n} u_{ik} > 0\}$, $M_{fcn} = \{U \in M_{pcn} | \forall k \sum_{i=1}^{c} u_{ik} = 1\}$, and $M_{hcn} = \{U \in M_{fcn} | \forall i, k, u_{ik} \in \{0,1\}\}$, where M_{pcn} are possibilistic c-partitions, M_{fcn} are fuzzy or probabilistic c-partitions, and M_{hcn} are crisp (hard) c-partitions. For convenience, we call the set $(M_{pcn} - M_{hcn})$ the *soft c-partitions* of O.

There are a number of popular indices [2] that are based on the entries of the standard contingency table. Let $U \in M_{hcn}$ and $V \in M_{hrn}$: the $c \times r$ contingency table of U and V is shown in Table I. Anderson et al. [2] observed that the contingency table could be constructed as the product $N = UV^T$. For crisp partitions, this formation reduces to the regular contingency table. *Any* comparison index that depends only on the entries of the contingency matrix can be generalized using the following equation:

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

where ϕ is a scaling factor that is needed in the possibilistic case, $n_{i\bullet} = \sum_{j=1}^{r} n_{ij}$ (see Table I). For crisp, fuzzy or probabilistic partitions, $\phi = 1$, the case of interest here.

B. Soft Generalization of Information-Theoretic Indices

Information-theoretic based measures are built upon fundamental concepts from information theory [11], and are a commonly used approach for crisp clustering comparison [6], [7]. Given a crisp partition U that partitions n objects into c subsets $\{u_1, \dots, u_c\}$, the (Shannon) entropy of U is $H(U) = -\sum_{i=1}^{c} p(u_i) \log p(u_i)$, where $p(u_i) = |u_i|/n$, indicates the probability of an object belonging to cluster u_i, and $|u_i| = n_i$ is the number of objects in cluster i. Note that $H(U)$ is different from $PE(U) = -(\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log_a(u_{ik}))/n$, where $a \in (1, \infty)$, the *partition entropy* of U [10]. Given two crisp partitions U and V, their *joint entropy* (JE) and *mutual information* (MI) can be defined according to the contingency table built upon U and V (Table I) respectively as [8]: $JE(U,V) = -\sum_{i=1}^{c} \sum_{j=1}^{r} (n_{ij}/n) \log(n_{ij}/n)$, $MI(U,V) = \sum_{i=1}^{c} \sum_{j=1}^{r} (n_{ij}/n) \log \frac{n_{ij}/n}{n_{i\bullet} n_{\bullet j}/n^2}$. Intuitively, the MI between two partitions measures how much information they share. The more information they share, the more similar they are, which results in a larger MI. More detailed explanations of these concepts can be found in [7], [8]. Eight popular crisp, external *information-theoretic cluster validity indices* (IT-CVIs) based on information-theoretic concepts are listed in Table II. The normalized version of VI (NVI) is equivalent to NMI$_{\text{j}}$. Thus, Table II contains seven independent IT-CVIs. We restrict our attention to the performance of NMI$_{\text{j}}$.

TABLE II: Information-theoretic cluster validity indices.

Name	Expression	Range	Find
MI	MI(U,V)	$[0, \min\{H(U), H(V)\}]$	Max
NMI$_{\text{j}}$	$MI(U,V)/JE(U,V)$	[0,1]	Max
NMI$_{\text{M}}$	$MI(U,V)/\max\{H(U), H(V)\}$	[0,1]	Max
NMI$_{\text{s}}$	$2MI(U,V)/(H(U) + H(V))$	[0,1]	Max
NMI$_{\text{r}}$	$MI(U,V)/\sqrt{H(U)H(V)}$	[0,1]	Max
NMI$_{\text{m}}$	$MI(U,V)/\min\{H(U), H(V)\}$	[0,1]	Max
Variation of Information (VI)	$JE(U,V) - MI(U,V)$	$[0, \log n]$	Min
Normalized VI (NVI*)	$1 - (MI(U,V)/JE(U,V))$	[0,1]	Min

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

The authors of [9] used equation (1) to generalize the indices in Table II for use with EM partitions. In particular, the entropy of a soft clustering U, is $H(U) = -\sum_{i=1}^{c} (n_{i\bullet}/n) \log(n_{i\bullet}/n)$, where $n_{i\bullet}$ is the row sum of the i-th row from the generalized contingency table N^*. Similarly, we define the joint entropy and mutual information of two soft partitions, $JE(U,V)$ and $MI(U,V)$, by taking n_{ij} and $n_{\bullet j} = \sum_{i=1}^{c} n_{ij}$ from N^*. Now soft versions of the eight IT-CVIs listed in Table II can be computed from the generalized contingency table N^* and are denoted as MI$_{\text{s}}$, NMI$_{\text{sj}}$, NMI$_{\text{sM}}$, NMI$_{\text{ss}}$, NMI$_{\text{sr}}$, NMI$_{\text{sm}}$, VI$_{\text{s}}$ and NVI$_{\text{s}}$.

III. Evaluation Methodology

A. Implementation and Settings

We modified the *fcm* function from the MATLAB Fuzzy Logic Toolbox to accommodate our initialization and termination criteria. *Initialization:* We randomly draw c distinct points from the data X as the initial cluster centers. The fuzzifier for FCM is $m = 2$ and the model norm is Euclidean. *Termination:* FCM is terminated when the difference between two successive estimates of the cluster centers, $\|\mathbf{W}_{t+1} - \mathbf{W}_t\|_\infty < \varepsilon$, where $\mathbf{W}_t = \{\mathbf{w}_1, \dots, \mathbf{w}_c\}$, and $\varepsilon = 10^{-3}$; the maximum number of iterations is 100.

B. Datasets

1) Synthetic Data: In this paper, we use 25 synthetic datasets, which contain five ground truth clusters, sampled from mixtures of two-dimensional Gaussian distributions. The covariance matrices for all clusters are identity matrices. We have tested on datasets possessed various attributes, e.g., shapes of clusters, the amount of overlap between clusters , cluster sizes (i.e., the number of samples in each cluster) and sample sizes. Three of these properties showed the largest impact on the CVIs, namely, the amount of overlap between clusters, cluster sizes and sample sizes. Hence, in the rest of the paper we focus on these three attributes. We generated four groups of datasets called groups G1, G2, G3 and G4. Among these four groups, the number of samples of each dataset in groups G1, G2 and G3 is $n = 1000$ and the sample size of datasets in group G4 are varied. More details:

Varying cluster overlap with equal sized clusters (G1). There are five equal sized clusters in the first group of datasets. We vary the overlap between two of the clusters by moving the mean of Cluster5 (μ_5) towards Cluster3 (μ_3) while keeping the other threes clusters' means fixed. The means of Cluster1, Cluster2, Cluster3 and Cluster4 are $\mu_1 = [2, 0], \mu_2 = [2, 13], \mu_3 = [13, 10], \mu_4 = [8, 17]$, respectively. The mean of the fifth cluster (Cluster5) is $\mu_5 = [13, 3 + i]$, where $i = 1, \ldots, 5$. That is, we increase the amount of overlap between Cluster3 and Cluster5 by moving Cluster5 up towards Cluster3 in the y direction. Thus, we generate five datasets, 'Ovp#', where # $\in \{1, \ldots, 5\}$.

Varying cluster sizes without overlapping clusters (G2). The second group of datasets are generated by varying the cluster sizes. For each dataset, the five clusters are well separated, i.e., non-overlapping, with fixed means of $\mu_1 = [2, 0], \mu_2 = [2, 13], \mu_3 = [13, 10], \mu_4 = [8, 17]$, and $\mu_5 = [13, 3]$. The size of Cluster5 is $n_5 = 100{*}i$, where $i = 1, \ldots, 6$, and $1/4th$ of the remaining $n - n_5$ objects are drawn for each of the other four clusters. Finally, we generated six datasets, 'Dens#', where # $\in \{1, \ldots, 6\}$.

Varying cluster sizes with overlapping clusters (G3). For the first two groups, we test the influence of a single factor (overlapping or cluster size) on the success of the generalized measures. However, real-world datasets are often more complicated and contain both overlapping and different sized clusters. To mimic this type of structure, we generated a third type of datasets. For each dataset in G3, the means of the five clusters are $\mu_1 = [2, 0], \mu_2 = [2, 13], \mu_3 = [13, 10], \mu_4 = [8, 17]$, and $\mu_5 = [13, 8]$. The means of Cluster3 and Cluster5 are close to each other and these clusters tend to be overlapping. We vary the sizes of Cluster5, based on $n_5 = 100 * i$, where $i = 1, \ldots, 6$ and $1/4th$ of the remaining $n - n_5$ objects are drawn for each of the other clusters. Thus, we generate six datasets, 'OvpDens#', where # $\in \{1, \ldots, 6\}$.

Varying data sizes with equal sized, non-overlapping clusters (G4). We vary the number of samples in the data to test the influence of data size on the soft CVIs. To facilitate the comparison, we fix the other two factors, i.e., cluster overlap and cluster sizes while generating the datasets. Specifically, for each dataset, the five clusters are well separated, with fixed means of $\mu_1 = [2, 0], \mu_2 = [2, 13], \mu_3 = [20, 13], \mu_4 = [11, 20]$, and $\mu_5 = [15, 5]$. The five clusters are equal sized, i.e., $n_1 = n_2 = n_3 = n_4 = n_4 = n_5 = n/5$. The sizes of the datasets are $n = \{100, 500, 1000, 5000, 10^4, 5 \times 10^4, 10^5, 5 \times 10^5, 10^6\}$. Thus, we generate nine datasets, 'NSize#', where # $\in \{1, \ldots, 9\}$. Please note that dataset 'Ovp5' in G1 is actually same as 'OvpDens2' in G3. Thus, we have 25 datasets overall instead of 26.

2) Real-World Data: Datasets from the UCI machine learning repository [12] are often benchmarks for evaluating external validity measures [4], [5]. These datasets have ground truth partitions provided by physically labeled subsets. We use 10 real-world datasets: nine are from the UCI repository and one large dataset *MNIST*, which is a collection of handwritten digits [13]. Parameters of the datasets are shown in Table III, where n, p and c_{GT} correspond to the number of objects, features and classes, respectively.

TABLE III: Real-world datasets: n = number of points, p = number of dimensions and C_{GT} = number of ground truth classes.

Dataset	n	p	c_{GT}
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
Zoo	101	17	7
Vertebral-Column	310	6	3
MNIST	70000	784	10

C. Experimental Design

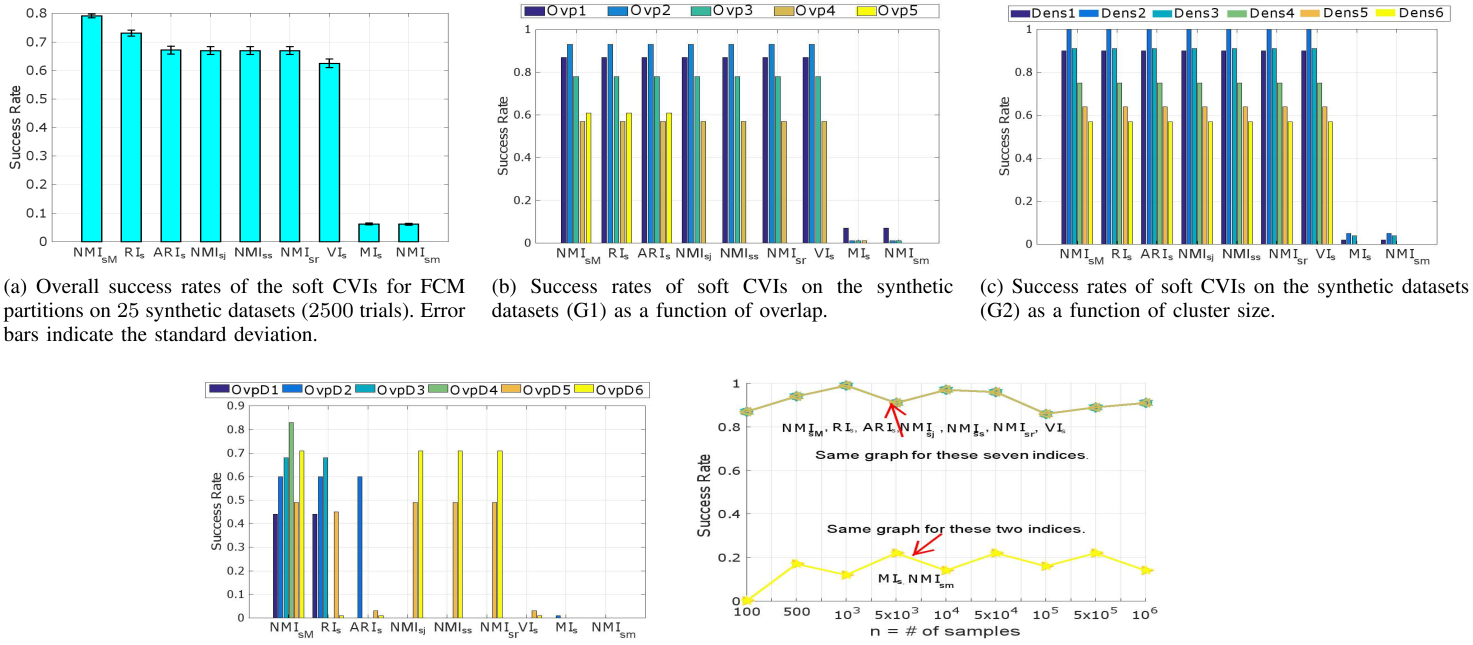
We test the effectiveness of the generalized soft indices by testing their ability to estimate the number of labeled clusters for synthetic datasets or classes for real-world datasets. In order to provide a baseline, we include two other soft CVIs that are not information-theoretic in nature, namely soft versions [2] of the *Rand Index* (RI) and the *adjusted Rand Index* (ARI), Hubert and Arabie version [14]). We denote these as the RI_{s} and $\mathrm{ARI}_{\mathrm{s}}$, respectively.

The general idea is to run FCM on each dataset to generate a set of partitions with different numbers of clusters. Then, each of the nine generalized soft indices is computed on every partition, where the comparison matrix V in equation (1) is the ground truth partition of the data. The number of clusters associated with the computed partition U obtaining the best result is c_{pre}, for that particular dataset. Let c_{true} be the number of known clusters in the synthetic datasets, and let c_{GT} denote the number of labeled classes in the real-world datasets. If $c_{pre} = c_{true}$ for the synthetic data, or $c_{pre} = c_{GT}$ for the real-world data, then we declare the prediction of this index on this dataset a success. We ran FCM on each dataset with the number of clusters c ranging from 2 to $3 \times c_{true}$ and $3 \times c_{GT}$ for the synthetic datasets and real-world datasets respectively. In order to reduce the influence of random initialization for FCM, we generate 100 partitions for each c, and evaluate the nine soft indices on each of the 100 partitions, so that we can make a histogram depicting the percentage of successes for each index over the 100 trials.

IV. Experimental Results

A. FCM Tests with the Synthetic Gaussian Datasets

The overall success rate for an index is the total number of successes across the 25 datasets divided by the total number of partitions, i.e., 25×100. *The indices are sorted in descending order by their success rates, displayed from left to right in Figure 1a.* In general, Figure 1a shows that $\mathrm{NMI}_{\mathrm{sM}}$ performs best among the nine soft CVIs on these synthetic datasets, having a success rate of approximately 80%. RI_{s}, $\mathrm{ARI}_{\mathrm{s}}$, $\mathrm{NMI}_{\mathrm{sj}}$, $\mathrm{NMI}_{\mathrm{ss}}$, $\mathrm{NMI}_{\mathrm{sr}}$ and VI_{s} achieve a success rate of $62 - 72\%$. In contrast, MI_{s} and $\mathrm{NMI}_{\mathrm{sm}}$ perform poorly, having a success



(a) Overall success rates of the soft CVIs for FCM partitions on 25 synthetic datasets (2500 trials). Error bars indicate the standard deviation.

(b) Success rates of soft CVIs on the synthetic datasets (G1) as a function of overlap.

(c) Success rates of soft CVIs on the synthetic datasets (G2) as a function of cluster size.

(d) Success rates of soft CVIs on the synthetic datasets (G3) as a function of cluster size with overlapping clusters.

(e) Success rates of soft CVIs on the synthetic datasets (G4) as a function of sample size.

Fig. 1: Success rates of soft CVIs for FCM partitions on 25 synthetic datasets.

rate of about 5%. A possible reason for this is that MI tends to monotonically increase with the number of clusters [8]. Hence, $\mathrm{MI_s}$ is likely to favour partitions with more clusters. For $\mathrm{NMI_{sm}}$, if the sizes of the discovered clusters are more equally distributed, the entropy of the generated soft partition, $H(U)$, increases with the number of clusters c. This effect may occur with FCM as FCM tends to favour clusters that are evenly sized. Note that the entropy of the ground truth labels $H(V) = q$ is constant. At some c, $H(U) > H(V)$, and subsequently, $NMI_{sm}(U,V) = MI_s(U,V)/H(V) = MI_s/q$, so $\mathrm{NMI_{sm}}$ becomes equivalent to the scaled version of $\mathrm{MI_s}$ and has the same deficiency.

1) Results on datasets with overlapping clusters (G1): The results on G1 appear in Figure 1b. The first seven CVIs perform similarly when evaluating FCM partitions for the first four datasets, but behave differently on the dataset Ovp5, which has the most overlapping clusters in G1. A missing vertical bar means that there were no successes for the given index on a particular dataset. For Ovp5, only $\mathrm{NMI_{sM}}$, $\mathrm{RI_s}$ and $\mathrm{ARI_s}$ perform relatively well with success rates of about 60% while the rest of the soft measures perform poorly (nearly 0%). This suggests that the efficacy of $\mathrm{NMI_{sj}}$, $\mathrm{NMI_{ss}}$, $\mathrm{NMI_{sr}}$ and $\mathrm{VI_s}$ to evaluate FCM partitions is more severely affected by overlap, while $\mathrm{NMI_{sM}}$, $\mathrm{RI_s}$ and $\mathrm{ARI_s}$ are more robust to this factor. At the other extreme, Figure 1b shows that $\mathrm{MI_s}$ and $\mathrm{NMI_{sm}}$ are inadequate for the G1 datasets.

2) Results on datasets with different sized clusters (G2): The bar chart in Figure 1c shows the results on G2. The first seven measures provide identical evaluations for all six datasets in G2. As with the first tests, $\mathrm{MI_s}$ and $\mathrm{NMI_{sm}}$ show poor performance. This indicates that the first seven CVIs are not influenced much by datasets containing different sized, non-overlapping clusters. Compared to the results on the first group of datasets, it seems that FCM, in common with most other clustering algorithms, has more difficulty finding partitions that match the ground truth when there is overlap than it does on the well-separated clusters.

3) Results on datasets with different sized and overlapping clusters (G3): The success rates of the CVIs for FCM partitions generated from G3 are shown in Figure 1d. There are some significant differences between the graphs in Figures 1b, 1c and the chart in Figure 1d, which corresponds to this set of tests. Specifically, $\mathrm{NMI_{sM}}$ is the only index in the experiments with G3 that successfully recovers a positive fraction of the 100 trials for each of the six datasets. The other eight indices have relatively poor performance. In particular, note the dropoff in performance by the soft $\mathrm{ARI_s}$, which did well for G1 and G2, but is quite ineffective here. These results suggest that only $\mathrm{NMI_{sM}}$ has (relatively) consistent good success rates for more complicated datasets, like those in this group of experiments.

4) Results on datasets with different data sizes (G4): Different from the previous bar graphs, we use a line graph to show the trend of success rates of all nine indices with increasing data sizes in Figure 1e. The x axis indicates the number of samples in the datasets. The y axis represents the success rates. There are only two graphs in Figure 1e: the upper graph has seven coincident plots which correspond to the indices $\mathrm{NMI_{sM}}$, $\mathrm{RI_s}$, $\mathrm{ARI_s}$, $\mathrm{NMI_{sj}}$, $\mathrm{NMI_{ss}}$, $\mathrm{NMI_{sr}}$ and $\mathrm{VI_s}$. The lower graph shows two coincident plots which represent $\mathrm{MI_s}$ and $\mathrm{NMI_{sm}}$. Please recall that all these datasets contain five well separated, equal sized clusters, so FCM is expected to find partitions similar to the ground truth on these datasets. We can draw several conclusions from this graph: (i) The first seven soft CVIs work well on all these datasets, while the last two, i.e., MI$_{\mathrm{s}}$ and NMI$_{\mathrm{sm}}$, work poorly . Apparently the first seven measures identify the right number of clusters when there are reasonable FCM partitions, while MI$_{\mathrm{s}}$ and NMI$_{\mathrm{sm}}$ do not; (ii) The size of the dataset does not impact the performance of these measures very much, i.e., the success rates of the first seven measures are consistently high and the success rates of MI$_{\mathrm{s}}$ and NMI$_{\mathrm{sm}}$ are always low.

Summary of experimental results on the synthetic datasets: NMI$_{\mathrm{sM}}$ performs better than the other eight soft CVIs. This suggests that NMI$_{\mathrm{sM}}$ might be preferred for detecting the right number of clusters when validating FCM partitions. We point out that NMI$_{\mathrm{sM}}$ performs better than the other three variants of NMI, i.e., NMI$_{\mathrm{sj}}$, NMI$_{\mathrm{ss}}$ and NMI$_{\mathrm{sr}}$ in certain scenarios, even though they are all based on soft mutual information but have different normalizations. Because of its very poor performance, we do not include NMI$_{\mathrm{sm}}$ in the comparison discussion. We will discuss this further in Section V.

B. FCM Tests with Real-World Datasets

The success rates of these indices on each real-world dataset are summarized in Table IV. The highlighted entries in the table show that NMI$_{\mathrm{sM}}$ performs better than the other measures. The last row of Table IV shows the column sums. The higher the number, the greater the overall success on these 10 datasets: a perfect score would be 10. NMI$_{\mathrm{sM}}$, with a score of 5.9, is clearly superior to the other eight indices. The RI$_{\mathrm{s}}$ comes in second, with a sum of 5. The last two columns, MI$_{\mathrm{s}}$ and NMI$_{\mathrm{sm}}$ are tied for last place at 2.06. Note that the indices are shown in the same order as in Figure 1a. Most of the indices keep the same ranking (column sums: higher scores to lowers scores from left to right) as they had on the synthetic datasets (Figure 1a), but NMI$_{\mathrm{sr}}$ is out of rank order and is not as effective as NMI$_{\mathrm{sj}}$ and NMI$_{\mathrm{ss}}$ in this set of experiments.

V. Theoretical Analysis

Our experiments suggest that NMI$_{\mathrm{sM}}$ has superior performance to NMI$_{\mathrm{sj}}$, NMI$_{\mathrm{ss}}$ and NMI$_{\mathrm{sr}}$ (Figures 1b and 1d). In this section, we provide a theoretical explanation which enables us to explain why NMI$_{\mathrm{sM}}$ outperforms the other three variants of NMI in certain situations. Please find the related proofs in the Appendix. First, we define two measures of change in the computation of NMI:

Definition 1. *Let $V \in M_{hrn}$ be a crisp reference partition (ground truth), $r \geq 3$. Let $U' \in M_{f(r-k)n}$ and $U^* \in M_{frn}$ be two soft partitions on n objects with $r - k$ and r clusters respectively. The relative change in MI$_s$ with respect to U' on moving from U' to U^* (note that the number of clusters increases by k, from $(r-k)$ to r) is*

$$\alpha = \big(MI(U^*,V) - MI(U',V)\big)/MI(U',V) \qquad (2)$$

Let NMI_* denote any of the three normalizations $\{NMI_{sj}, NMI_{ss}, NMI_{sr}\}$ of MI$_s$, and let $B_*(U,V)$ denote the denominators (normalization factors as shown in Table II) of $\{NMI_{sj}, NMI_{ss}, NMI_{sr}\}$. The relative change in the denominator of any of these CVIs with respect to U' on moving from U' to U^* is

$$\beta = \big(B_*(U^*,V) - B_*(U',V)\big)/(B_*(U',V)) \qquad (3)$$

Theorem 1. *Let $V \in M_{hrn}$ be a crisp reference partition (ground truth), $r \geq 3$. Let $U' \in M_{f(r-k)n}$ and $U^* \in M_{frn}$ be two soft partitions on n objects with $r-k$ and r clusters respectively, where $(r-k) \geq 2$ and $k \geq 1$. Let NMI_* denote any of the three normalizations $\{NMI_{sj}, NMI_{ss}, NMI_{sr}\}$ of MI_s. If $MI(U^*,V) > MI(U',V)$ and $H(V) \geq H(U^*), H(U')$, then **(A)** $NMI_{sM}(U^*,V) > NMI_{sM}(U',V)$, and **(B)** $NMI_*(U^*,V) = \big((1+\alpha)/(1+\beta)\big)NMI_*(U',V)$.*

Equation (A) shows that when $H(V) \geq H(U^*), H(U')$, and the number of clusters in the soft partition increases from $r-k$ in U' to r in U^*, that when MI also increases, i.e., $MI(U^\star,V) > MI(U',V)$, then $NMI_{sM}(U^*,V) > NMI_{sM}(U',V)$. In contrast, the other three forms of normalized MI$_{\mathrm{s}}$ depend on relative changes of both their numerators and denominators. i.e., if $\alpha > \beta$, then $NMI_*(U^*,V) > NMI_*(U',V)$; if $\alpha = \beta$, then $NMI_*(U^*,V) = NMI_*(U',V)$; if $\alpha < \beta$, then $NMI_*(U^*,V) < NMI_*(U',V)$. Thus, when $MI(U^*,V) > MI(U',V)$, NMI$_{\mathrm{sM}}$ will favour U^* (r clusters, matching the number of clusters in the reference partition V) over U', which has $r-k$ clusters. But for the other three measures NMI_*, α and β are sensitive to changes from U' to U^*, and hence can fluctuate easily, making these three measures unstable and hence, their performance more uncertain. Next, we discuss a specific case for Theorem 1, when the ground truth is balanced.

Definition 2. *Let $U \in M_{fcn}$ be any crisp, fuzzy or probabilistic partition of n objects with c clusters. Then U is **balanced** if and only if $\sum_{k=1}^{n} u_{ik} = n/c$, $1 \leq i \leq c$.*

In other words, each of the c clusters in U is allocated the same amount of membership. When U is crisp, this is equivalent to saying that each of the c crisp clusters has the same number of objects in it. The importance of this concept is contained in the following well know result.

Proposition 1. *Let $U \in M_{fcn}$ be any crisp, fuzzy or probabilistic partition with $c > 1$. The entropy $H(U) = -\sum_{i=1}^{c} p(u_i) \log p(u_i)$, where $p(u_i) = (\sum_{k=1}^{n} u_{ik})/n$, is maximum if and only if U is balanced. The maximum entropy of U is $\max_{U \in M_{fcn}}\{H(U)\} = \log c$.*

Now we are in a position to show why NMI$_{\mathrm{sM}}$ is the best normalization of the mutual information when the reference partition is balanced (datasets G1 in our experiments).

Corollary 1. *Let $V \in M_{hrn}$ be a crisp, balanced reference partition. If $MI(U^*,V) > MI(U',V)$, then statements (A) and (B) in Theorem 1 hold.*

In summary, the better performance of NMI$_{\mathrm{sM}}$ when compared to the other three NMI measures, is due to their denominators (normalization factors) having different sensitivity to changes in the number of clusters in candidate partitions. NMI$_{\mathrm{sM}}$ is more robust to the changes, while the other three indices suffer from sensitivity to α and β in equations 2 and 3.

VI. Conclusions

This paper has presented an organized study of eight IT-CVIs for FCM partitions on 25 synthetic and 10 real-world

TABLE IV: Success rate (% of successes in 100 trials) of nine indices for FCM on 10 real-world datasets. The highlighted *numbers* indicate success rates above 85%. The highlighted *datasets* have at least one rate above 85%.

FCM	$\mathrm{NMI}_{\mathrm{sM}}$	RI_{s}	$\mathrm{ARI}_{\mathrm{s}}$	$\mathrm{NMI}_{\mathrm{sj}}$	$\mathrm{NMI}_{\mathrm{ss}}$	$\mathrm{NMI}_{\mathrm{sr}}$	VI_{s}	MI_{s}	$\mathrm{NMI}_{\mathrm{sm}}$	Row Sums
Sonar	**0.95**	**1**	**1**	**0.93**	**0.93**	**0.93**	**1**	**0.87**	**0.87**	8.48
Pima-diabetes	**0.96**	**1**	**0.98**	**0.89**	**0.89**	**0.85**	**1**	0	0	6.57
Heart-statlog	**0.99**	**1**	**1**	**0.99**	**0.99**	**0.99**	**1**	**0.99**	**0.99**	8.94
Haberman	0	**1**	0	0	0	0	**1**	0	0	2
Wine	**1**	0	**1**	**1**	**1**	**1**	0	0.20	0.20	5.4
Vehicle	0	0	0	0	0	0	0	0	0	0
Iris	**1**	**1**	**1**	**1**	**1**	0	0	0	0	5
Zoo	0	0	0	0	0	0	0	0	0	0
Vertebral Column	**1**	0	0	0	0	0	0	0	0	1
MNIST	0	0	0	0	0	0	0	0	0	0
Column Sums	5.9	5	4.98	4.81	4.81	3.77	4	2.06	2.06	

datasets. We demonstrated that soft generalizations of the eight IT-CVIs are quite capable of identifying the "correct" number of clusters or classes from candidate partitions generated by FCM on these synthetic and real-world datasets. The results of this study, combined with previous computational results in [9], provide a reasonably strong empirical argument about the effectiveness of generalized IT-CVIs for both fuzzy and probabilistic cluster validity. In particular, $\mathrm{NMI}_{\mathrm{sM}}$ is superior to the other seven generalized IT-CVIs for both FCM and EM partitions on datasets with overlapped and/or various sized clusters. Finally, Theorem 1 provides a theoretical reason to expect better performance of $\mathrm{NMI}_{\mathrm{sM}}$ over the other three variants of NMI, i.e., $\mathrm{NMI}_{\mathrm{sj}}$, $\mathrm{NMI}_{\mathrm{ss}}$ and $\mathrm{NMI}_{\mathrm{sr}}$ in certain situations.

To the best of our knowledge, this is the first cluster validity study which demonstrates that the distribution of the ground truth subsets can bias the value of an external CVI. Our theorem covers a very specific case for one external CVI, but suggests a much richer question for further research: to what extent does the distribution of the ground truth partition affect any external cluster validity index? We have taken one step in this direction with some new results about ground truth bias in the Rand index using quadratic entropy which will be reported in a forthcoming paper.

Appendix

Proof for Theorem 1

Proof. (A) if $H(V) \geq H(U^*), H(U')$, then $\max\{H(U'), H(V)\} = \max\{H(U^*), H(V)\} = H(V)$. By hypothesis, $MI(U^*, V) > MI(U', V)$, so $NMI_{sM}(U^*, V) = MI(U^*, V)/\max\{H(U^*), H(V)\} = MI(U^*, V)/H(V) > MI(U', V)/H(V) = NMI_{sM}(U', V)$. This completes the proof of (A). (B) Rearranging equation (2) yields $MI(U^*, V) = (1+\alpha)MI(U', V)$. Similarly, rearranging equation (3) yields $B_*(U^*, V) = (1+\beta)B_*(U', V)$. Then for any of the three normalized forms of MI_{s} we have $NMI_*(U^*, V) = MI(U^*, V)/B_*(U^*, V) = ((1+\alpha)MI(U', V))/((1+\beta)B_*(U', V)) = \big((1+\alpha)/(1+\beta)\big)NMI_*(U', V)$. This completes the proof of (B). □

Proof for Proposition 1

Proof. Regard the row sums of U as c "events". Here are three well know facts from information theory [11]: (i) $0 \leq H(U) \leq \log c$; (ii) $H(U) = 0$ when exactly one of the $p(u_i)$'s is 1 and all the rest are zero; (iii) $H(U) = \log c$ if and only all of the events have the same probability $p(u_i) = 1/c, i = \{1, \ldots, c\}$. ($\Rightarrow$) Assume that $H(U) = \log c$. From the fact (iii), the only time this can happen is when U is balanced. ($\Leftarrow$) Assume that U is balanced. When U is balanced, its row sums are all equal by Definition 2, that is, the c "events" are all equally likely. Again by fact (iii), this guarantees that $H(U)$ is maximum, with value $H(U) = \log c$. □

Proof for Corollary 1

Proof. If V is balanced, according to Proposition 1, so $H(V) = \log r \geq H(U^*)$. Also, $H(V) = \log r > \log(r-k) \geq H(U')$, so $H(V) \geq H(U^*), H(U')$.

□

References

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

[2] 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.

[3] 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.

[4] 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.

[5] E. Hüllermeier, M. Rifqi, S. Henzgen, and R. Senge, "Comparing fuzzy partitions: A generalization of the rand index and related measures," *Fuzzy Systems, IEEE Transactions on*, vol. 20, no. 3, pp. 546–556, 2012.

[6] 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.

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

[8] 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. 11, pp. 2837–2854, 2010.

[9] Y. Lei, J. C. Bezdek, J. Chan, N. Xuan Vinh, S. Romano, and J. Bailey, "Generalized information theoretic cluster validity indices for soft clusterings," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 24–31.

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

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

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

[13] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," *Fuzzy Systems, IEEE Transactions on*, vol. 20, no. 6, pp. 1130–1146, 2012.

[14] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.