Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization

Ludmila I. Kuncheva, *Member*, *IEEE*, and Dmitry P. Vetrov

Abstract—Many clustering algorithms, including cluster ensembles, rely on a random component. Stability of the results across different runs is considered to be an asset of the algorithm. The cluster ensembles considered here are based on k-means clusterers. Each clusterer is assigned a random target number of clusters, k and is started from a random initialization. Here, we use 10 artificial and 10 real data sets to study ensemble stability with respect to random k, and random initialization. The data sets were chosen to have a small number of clusters (two to seven) and a moderate number of data points (up to a few hundred). Pairwise stability is defined as the adjusted Rand index between pairs of clusterers in the ensemble, averaged across all pairs. Nonpairwise stability is defined as the entropy of the consensus matrix of the ensemble. An experimental comparison with the stability of the standard k-means algorithm was carried out for k from 2 to 20. The results revealed that ensembles are generally more stable, markedly so for larger k. To establish whether stability can serve as a cluster validity index, we first looked at the relationship between stability and accuracy with respect to the number of clusters, k. We found that such a relationship strongly depends on the data set, varying from almost perfect positive correlation (0.97, for the glass data) to almost perfect negative correlation (-0.93, for the crabs data). We propose a new combined stability index to be the sum of the pairwise individual and ensemble stabilities. This index was found to correlate better with the ensemble accuracy. Following the hypothesis that a point of stability of a clustering algorithm corresponds to a structure found in the data, we used the stability measures to pick the number of clusters. The combined stability index gave best results.

Index Terms—Clustering, cluster ensembles, stability and diversity, cluster validity.

1 Introduction

Cluster ensembles have been introduced as a more accurate alternative to individual clustering algorithms. Many published studies have demonstrated the advantages of such ensembles over single clusterers in discovering clusters of arbitrary shape and size [12], [14], [32]. Two major themes in this literature are combination methods of the ensemble votes and diversifying heuristics for building the ensemble.

Here, we are interested in the stability of cluster ensembles. The stability of a clustering algorithm with respect to small perturbations of the data (e.g., data subsampling or resampling, small variations in the feature values), or the parameters of the algorithm (e.g., random initialization) is a desirable quality [29]. On the other hand, ensembles benefit from diverse clusterers [8], [16], [17]. This paper carries out an experimental study to examine whether cluster ensembles give more stable results than single clustering methods. In doing so, we also look for a cluster validity index which can help us to identify the "best" number of clusters. Not every clustering algorithm, be it an ensemble or a single clusterer, will be able to discover the true structure in the data. Therefore, there might be an optimal number of clusters for the considered algorithm which is not necessarily the true number of clusters. High correlation between stability and a suitable measure of accuracy of the clustering algorithm is paramount for finding this optimal number of clusters.

In this study, we are looking for answers to the following questions:

1. Are ensembles more stable than individual clusterers?
2. Is ensemble stability related to ensemble accuracy?
3. How good is ensemble stability as a cluster validity measure?

The rest of the paper is organized as follows: Cluster ensembles are briefly introduced in Section 2. Section 3 details the stability measures evaluated in this study and discusses their application as cluster validity indices. Section 4 describes the data sets, the experimental protocol, and the results. Section 5 contains our discussion and conclusions.

2 Cluster Ensembles

Let $P_1, \ldots, P_L$ be a set of partitions of a data set $\mathbf{Z}$, each one obtained from applying a clustering algorithm. The aim is to find a resultant partition P^* which best represents the structure of $\mathbf{Z}$. We can think of the L partitions as the decisions of an ensemble of clusterers with P^* being the combined decision of the ensemble.

- *L.I. Kuncheva is with the School of Informatics, University of Wales, Bangor, Bangor, Gwynedd LL57 1UT, UK. E-mail: l.i.kuncheva@bangor.ac.uk.*
- *D.P. Vetrov is with the Dorodnicyn Computing Centre, Russian Academy of Sciences, Vavilova str. 40, room 472, 119991 Moscow, Russian Federation. E-mail: vetrovd@yandex.ru.*

Manuscript received 29 July 2005; revised 5 Apr. 2006; accepted 10 Apr. 2006; published online 14 Sept. 2006.
Recommended for acceptance by J. Buhmann.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0407-0705.

The two major issues are how to build diverse yet accurate individual clusterers and how to combine their decisions. Various heuristics have been proposed in the literature for building the ensemble members. Among these are random initializations of the clustering algorithm, subsampling, or resampling the data [5], [8], [9], [13], [16], [26], [27], applying different types of clustering algorithms [1], [16], [18], [37], using subsets of features [16], [32], "weakening" the clustering algorithm [16], [34], projecting the data in random affine subspaces [8], [34], etc. One of the most successful heuristics has been randomly choosing the number of clusters assigned to each clusterer in the ensemble [12], [13], [16], [17], [21].

We can construct the resultant partition P^* following several approaches (called "consensus functions"): the direct approach (relabeling of P_i and finding P^* which has the best match with all $P_i, i = 1, \ldots, L$) [9], [32], [37], the feature-based approach (treating outputs from the clusterers as L categorical features and building a clusterer thereupon) [35], the hyper-graph approach (constructing a hyper-graph representing the total output from the clusterers and cutting the redundant edges) [32], and the pairwise approach [1], [8], [10], [11], [13], [27]. We implemented the pairwise approach because it has been a popular choice despite its comparatively large computational demand. As cluster ensembles are relatively new offspring of the multiple classifier systems area, to facilitate reproducibility of our results, we detail the generic pairwise cluster ensemble algorithm below:

1. Given is a data set $\mathbf{Z}$ with N elements. Pick the ensemble size L and the number of clusters k. Usually, k is larger than the suspected number of clusters so there is "overproduction" of clusters.[1]
2. Generate L partitions of $\mathbf{Z}$ with k clusters in each partition.
3. Form a coassociation matrix for each partition, $M^{(s)} = \{m_{ij}^{(s)}\}$, of size $N \times N$, $s = 1, \ldots, L$, where

$$m_{ij}^{(s)} = \begin{cases} 1, & \text{if } \mathbf{z}_i \text{ and } \mathbf{z}_j \text{ are in the same cluster in partition } s \\ 0, & \text{if } \mathbf{z}_i \text{ and } \mathbf{z}_j \text{ are in different clusters in partition } s. \end{cases}$$

4. Form a final coassociation matrix $\mathbf{M}$ (consensus matrix) from $M^{(s)}$, $s = 1, \ldots, L$, and derive the final clustering using this matrix. A typical choice for $\mathbf{M}$ is

$$\mathbf{M} = \frac{1}{L}\left(M^{(1)} + M^{(2)} + \ldots + M^{(L)}\right).$$

The consensus matrix $\mathbf{M}$ can be regarded as a similarity matrix between the points of $\mathbf{Z}$. Therefore, it can be used with any clustering algorithm which operates directly upon a similarity matrix. The output is taken to be the ensemble partition P^*. The name "pairwise" comes from relating pairs of objects to find P^*. Viewed in this context, a cluster ensemble is a type of *stacked clustering* whereby we can generate layers of similarity matrices and apply clustering algorithms on them. In this study, we use k-means as the base clusterer and single linkage as the consensus function, interpreting $\mathbf{M}$ as similarity. The target number of clusters for each clusterer is picked randomly between 2 and a chosen value $K_{\max}$ (here, $K_{\max} = 20$).

By "accuracy" of a clustering algorithm, we shall assume the similarity of the obtained clustering to a known labeling of the data. Such labeling is available in a clear form for artificially generated data sets. In order to use real data sets with known class labels, we have to make the convenient assumption that classes correspond to clusters in data. This may be true, partly or completely, for some real data sets, but is by no means guaranteed. Many authors have used real benchmark data sets with known class labels to evaluate clustering algorithms and we will follow this tradition here.

3 Stability Measures and Cluster Validity

The stability of a clustering algorithm with respect to small perturbations of data and also different initializations is a desirable quality of the algorithm. Cluster ensembles, on the other hand, enforce and exploit some instability so that the ensemble is comprised of diverse clusterers. Although built upon unstable components, the ensemble is expected to be more accurate and robust than the individual clustering method. Here, we look at the stability of the ensemble.

3.1 Pairwise and Nonpairwise Stability

We consider two approaches to measuring the stability of a set of clusterers, $P_1, \ldots, P_L$: pairwise and nonpairwise.[2]

In the pairwise approach, the match between each of the $L(L-1)/2$ pairs of clusterers is calculated and the stability index is obtained as the averaged degree of match across the pairs. Let $S(P_i, P_j)$ be the degree of match (agreement or stability) between partitions P_i and P_j. The pairwise stability index $\mathcal{S}_p$ is

$$\mathcal{S}_p = \frac{2}{L(L-1)} \sum_{\substack{1 \le i,j \le L, \\ i<j}} S(P_i, P_j). \tag{1}$$

There are many indices evaluating the match between two partitions, fromamong which we selected the adjusted Rand index [19], [29]. This index takes value 1 if the partitions are identical and has an expected value of 0 if they are drawn independently of one another, regardless of the number of clusters.

Let A and B be partitions of Z with k_A and k_B clusters, respectively. Let n_i be the number of objects in cluster i in partition A and m_j be the number of objects in cluster j in partition B. Denote by n_{ij} the number of objects which belong simultaneously to cluster i in partition A and cluster j in partition B. The adjusted Rand index is calculated as

$$AR(A,B) = \frac{\sum_{i=1}^{k_A}\sum_{j=1}^{k_B}\binom{n_{ij}}{2} - t_3}{\frac{1}{2}(t_1+t_2) - t_3}, \tag{2}$$

where

$$t_1 = \sum_{i=1}^{k_A}\binom{n_i}{2},\ t_2 = \sum_{j=1}^{k_B}\binom{m_j}{2},\ \text{and}\ t_3 = \frac{2t_1t_2}{N(N-1)}.$$

We will use the adjusted Rand index (2) to calculate the pairwise stability, $\mathcal{S}_p$, in (1) and also to evaluate the accuracy of the clustering algorithm with respect to the

1. Note that, although k is fixed for all ensemble members in the original algorithm, in the version which we use later on, k is chosen randomly for each ensemble member. This induces diversity in the ensemble, and has been found to be one of the most useful cluster ensemble heuristics [12], [13], [16], [17], [21].

2. The pairwise approach to measuring stability refers to pairs of *clusterers* and should not be confused with the pairwise method for constructing the ensemble.

known true partition P^{true} as $AR(P^*, P^{\text{true}})$ for the ensemble and $AR(P_i, P^{\text{true}})$ for the ith individual clusterer.

In the nonpairwise approach, the consensus matrix $\mathbf{M}$ is analyzed. If all the clusterers agree on joining objects i and j in the same cluster, then $m_{ij} = 1$. If all clusterers agree that objects i and j are in different clusters, then $m_{ij} = 0$. Only if there is disagreement on joint membership of the two objects, will m_{ij} be between 0 and 1. In the case of the largest disagreement, where i and j are in the same clusters in exactly $L/2$ of the partitions $P_1, \ldots, P_L$, $m_{ij} = 0.5$. It seems natural to measure the disagreement between the clusterers as the averaged entropy of the cells of $\mathbf{M}$ (recall that $\mathbf{M}$ is of size $N \times N$, where N is the number of objects in the data set, $\mathbf{Z}$)[3]

$$H(\mathbf{M}) = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\big(m_{ij}\log(m_{ij}) + (1-m_{ij})\log(1-m_{ij})\big). \tag{3}$$

Entropy has been used as a measure of diversity of cluster ensembles by Greene et al. [16]. In the same vein, Monti et al. [27] propose looking at the "contrast" of the distribution of the values m_{ij}. We shall use as the nonpairwise stability index $\mathcal{S}_{np} = -H(\mathbf{M})$.

3.2 Using Stability as a Cluster Validity Index

Finding a suitable number of clusters is an ill-posed problem of crucial relevance in cluster analysis [15], [20]. Various solution paths being explored can be roughly grouped into two: approaches based on geometrical properties of the clusters (compactness, isolation, within and between-cluster dispersion, etc.) and approaches based on the concept of stability of the clustering. Within the first approach, the indices by Calinski-Harabasz and Krzanowski-Lai have been repeatedly chosen as benchmarks [7], [25], [28], [33]. The Gap statistic by Tibshirani et al. [33] has been shown to be very accurate for finding the true number of clusters, while simultaneously testing for existence of a structure in data. The stability approach is based on the idea that the correct number of clusters is a point of stability for the clustering algorithm. In other words, the true number of clusters is sought as the value for which the partitions obtained through data perturbation are highly similar to one another. Different cross-validation protocols can be used, the two most widely explored being 2-fold cross-validation [15], [31] and bootstrap resampling or subsampling [7], [9], [22], [23], [24], [27], [28].

The problem is ill-posed because there is no rigorous definition of what a cluster is. Validity measures are based on geometrical properties of the clusters. Thus, each validity measure will favor a specific shape of clusters and will not be useful if clusters are of very different shapes. If we are looking for the *true* number of clusters with a particular validity measure, we need to assume what shape the clusters are likely to have. There might be clusters of very different shapes in the same data set and there might generally be no information on the shape of the clusters in real data sets.

Different clustering algorithms may produce differently shaped clusters. It makes sense to couple a measure of cluster validity with a particular clustering algorithm. Thus, if the measure indicates that the data is likely to contain k hyper-spherical clusters, k-means can be used to find the

3. Assume $0\log(0) = 0$.

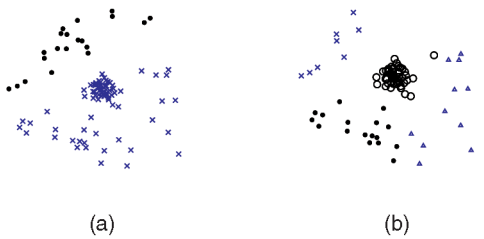

Fig. 1. Difficult doughnut data set (contains 10 more noise dimensions) clustered by k-means in: (a) two clusters and (b) four clusters.

labels. In this case, the number of clusters found by such measures does not have to be the true number of clusters. Knowing the true number of clusters and trying to enforce it upon k-means may lead to very poor results. Fig. 1 illustrates this point on a data set called "difficult doughnut" (used later in the experiment). There are two clusters in this data set, the outer ring and the Gaussian within, which are impossible to find by the standard k-means algorithm. Any attempt to arrive at $k = 2$ clusters (Fig. 1a) will give intuitively worse results than clustering in larger k, where the outer ring is broken into subclusters (Fig. 1b).

The stability-based validity indices are not bound to the clustering method used for partitioning the perturbed data. More importantly, there is no implied guess on the clusters' shape and size. This makes stability-based indices more adequate for using with cluster ensembles, knowing that the main claim of cluster ensembles is exactly that the obtained clusters can be of any shape and size. The problem here is that the assumption that stability corresponds to high accuracy may not always hold.

Here, we take the stability route and assume that ensemble stability corresponds to high ensemble accuracy. Note that by *ensemble stability* we shall mean the stability of the ensemble decision, not stability among the clusterers within the ensemble. The ensemble stability will be used as a validity index and compared to the results obtained through the stability of single clusterers.

This study differs from the previous works that use stability for validating clustering results by the chosen source of variability. We evaluated stability of k-means and ensembles thereof across different initializations, while the previous works have used data resampling/subsampling. For a single k-means algorithm, this choice amounts to evaluating by Monte Carlo simulations the landscape of the sum-of-squared-error criterion J_e [4] for a given k. A landscape with a single minimum (leading to the same partition) will correspond to high stability. The hypothesis is that this scenario indicates a true cluster structure in the data. If there are multiple minima but they are such that their corresponding partitions are similar to one another, again stability for the respective k will be large. On the other hand, if the multiple minima of the criterion function lead to very different partitions, stability will be low and, according to our hypothesis, the plausibility of this structure will be low. Cluster ensembles optimize a different criterion function, in most cases not explicitly defined. We note that we do not use the information about the "depth" of the minima nor do the other methods based on stability. For the individual clusterers, this depth is the value of the criterion, albeit not comparable across different k. For ensembles, defining and interpreting such a criterion value is not straightforward.

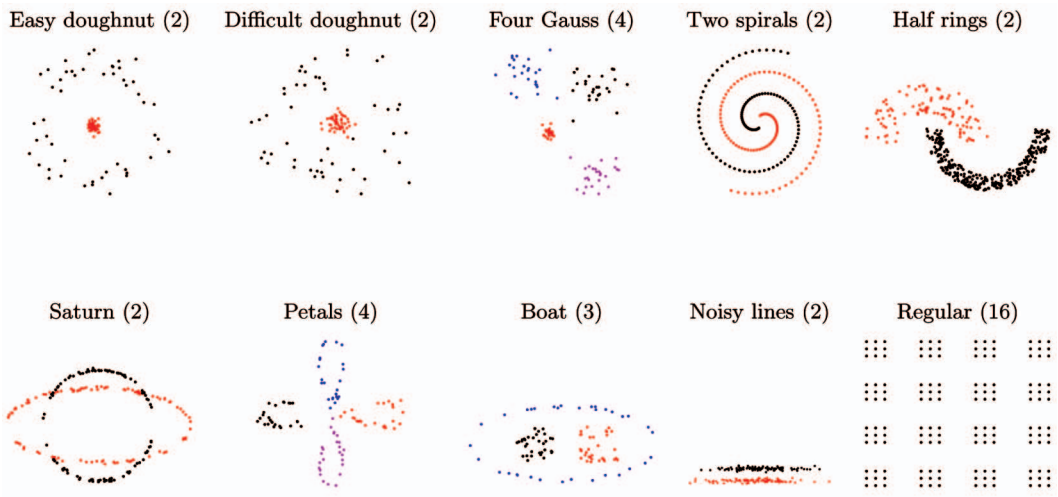

Fig. 2. Ten artificial data sets used in this study. The first three data sets were generated with 10 additional noise features. The number of clusters is given in parentheses.

4 The Experiment

4.1 The Data Sets

Ten artificial and 10 real data sets were selected for this study. The artificial data sets are shown in Fig. 2. These are all created in two dimensions and are meant to present different degree of challenge to the clustering algorithm. Ten dimensions of uniform random noise were appended to each of the first three data sets (easy doughnut, difficult doughnut, and four gauss), while the other seven data sets were kept as two-dimensional.

The 10 real data sets are described in Table 1.

4.2 Experimental Protocol

The ensembles studied here consist of $L = 25$ clusterers, where each clusterer is assigned a random number of clusters between two and $K_{\max}$ ($K_{\max} = 20$ was chosen). The consensus matrix $\mathbf{M}$ is calculated for each ensemble and fed to the single linkage clustering algorithm. The ensemble decision is obtained by stopping the single linkage at a predefined number of clusters, k. For each data set, we built 100 such ensembles. Denote by $P^*(k, j)$ the resultant partition by ensemble j, $j = 1, \ldots, 100$, for number of clusters k. The following statistics were calculated for k:

TABLE 1
Characteristics of the 10 Real Data Sets Used in This Study

dataset	Classes (c)	Objects (N)	Features (n)	Source
contractions	2	98	27	[36]
crabs	2	200	7	[30]
glass	7	214	9	UCI [2]
ionosphere	2	351	0	UCI [2]
iris	3	150	0	UCI [2]
respiratory	2	85	17	(private)
segmentation	7	210	19	UCI [2]
soybean-small	4	47	35	UCI [2]
thyroid	3	215	5	UCI [2]
wine	3	178	13	UCI [2]

Note: Data sets contractions and respiratory are explained in the Appendix.

1. Average ensemble accuracy

$$\mathcal{A}^e(k) = \frac{1}{100}\sum_{j=1}^{100} AR\big(P^*(k,j), P^{\text{true}}\big),$$

where $AR(.,.)$ is the adjusted Rand index.

2. Total ensemble accuracy

$$\mathcal{A}^t(k) = AR(P^*(k), P^{\text{true}}),$$

where $P^*(k)$ is the decision of the entire ensemble of the pooled 2,500 clusterers.

3. Individual accuracy

$$\mathcal{A}^i(k) = \frac{1}{|I_k|}\sum_{j\in I_k} AR(P_j(k), P^{\text{true}}),$$

where $I_k \subset \{1, 2, \ldots, 2,500\}$ is the index set of all clusterers within the set of 2,500 which clustered in k, and $|I_k|$ is the cardinality of I_k (approximately $2,500/(K_{\max} - 1)$). $P_j(k)$ denotes the partition produced by clusterer j.

4. Pairwise ensemble stability[4]

$$\mathcal{S}_p^e(k) = \frac{2}{100 \times 99}\sum_{1\le i,j\le 100,\ i<j} AR(P^*(k,i), P^*(k,j)).$$

5. Pairwise individual stability

$$\mathcal{S}_p^i(k) = \frac{2}{|I_k|(|I_k|-1)}\sum_{i,j\in|I_k|,i<j} AR(P_i(k), P_j(k)).$$

For the adjusted Rand, the maximum value of 1 is obtained for identical partitions and values around 0 are obtained for independent partitions (negative values are possible).

The nonpairwise measures based on entropy should be normalized before calculating correlations

4. Recall that the pairwise stability index for an ensemble is the averaged Adjusted Rand index (AR) across all pairs of clusterers (Section 3). The nonpairwise stability index is based on the entropy of the consensus matrix $\mathbf{M}$.

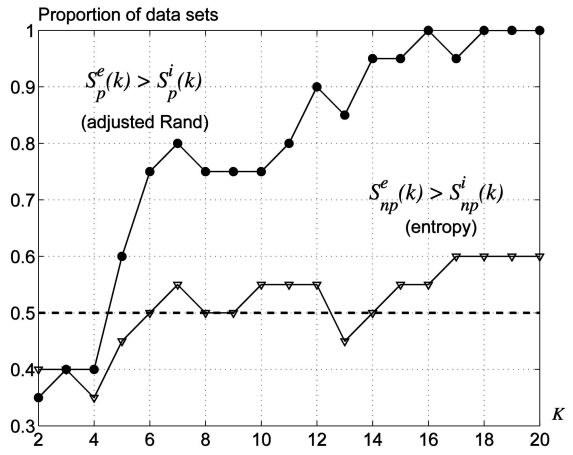

Fig. 3. Proportion of the data sets for which ensemble stability exceeds individual stability for the pairwise ($\mathcal{S}_p$) and the nonpairwise ($\mathcal{S}_{np}$) measures.

or using these measures to select number of clusters. The minimum value of 0 is obtained when all partitions are the same. However, the maximum value of entropy for a given k will depend on k. For example, suppose that $k_2 > k_1$ and the calculated entropies of the respective consensus matrices are such that $H(k_2) < H(k_1)$. This could be either because the clustering method is more stable for k_2 or because the maximum achievable entropy is lower and the method is more unstable for k_2. To eliminate this effect, some form of normalization is needed. For the asymptotic case where $L \to \infty$ and $N \to \infty$, the maximum entropy of the consensus matrix for k clusters will be

$$H_{\max}(k) = -\left(\frac{1}{k}\right)\log\left(\frac{1}{k}\right).$$

The nonpairwise stability measures are then:

6. Nonpairwise ensemble stability

$$\mathcal{S}_{np}^e(k) = -H(\mathbf{M}^e(k))/H_{\max}(k),$$

where $\mathbf{M}^e(k)$ is the consensus matrix obtained from the 100 ensemble outputs $P^*(k,j)$, $j = 1,\ldots,100$ and

7. Nonpairwise individual stability

$$\mathcal{S}_{np}^i(k) = -H(\mathbf{M}^i(k))/H_{\max}(k),$$

where $\mathbf{M}^i(k)$ is the consensus matrix obtained from the partitions $P_j(k)$, $j \in I_k$.

The next three sections seek to answer the questions formulated in the Introduction:

1. Are ensembles more stable than individual clusterers? (Can we claim that $\mathcal{S}_p^e(k) \geq \mathcal{S}_p^i(k)$ and $\mathcal{S}_{np}^e(k) \geq \mathcal{S}_{np}^i(k)$? For what values of k does this hold?)
2. Is ensemble stability related to ensemble accuracy? (What is the correlation across k between $\mathcal{A}^e(k)$ on the one hand and $\mathcal{S}_p^e(k)$ or $\mathcal{S}_{np}^e(k)$ on the other hand?)
3. How good is ensemble stability as a cluster validity measure?

Fig. 4. Clusters found by the overwhelming majority of the k-means clusterers for $k = 2$ on the "noisy lines" data set.

4.3 Are Ensembles More Stable than Individual Clusterers?

Fig. 3 plots the proportion of the data sets (out of 20) for which $\mathcal{S}_p^e(k) \geq \mathcal{S}_p^i(k)$ (dot marker) and also the proportion for which $\mathcal{S}_{np}^e(k) \geq \mathcal{S}_{np}^i(k)$ (triangle marker), as a function of the number of clusters, k.

It appears that single clusterers tend to be slightly more stable for a small number of clusters, while ensembles are more stable for larger k. This tendency is more pronounced for the pairwise stability index. This suggests that if the number of clusters is decided by the maximum stability, ensembles will be likely to pick a larger number of clusters than will single clusterers.

We noticed that the individual stability is usually greater for a small number of clusters. However, greater stability does not necessarily mean greater accuracy. Consider for example the "noisy lines" data set. The individual stability for two clusters is almost perfect, $\mathcal{S}_p^i(2) = 0.9607$, but this is because all partitions agree on the wrong two clusters, as illustrated in Fig. 4. The low ensemble stability, $\mathcal{S}_p^e(2) = 0.3545$, suggests that the two clusters found by the individual k-means for $k = 2$ may not be the true clusters.

The fact that ensemble stability was lower than individual stability on more than half of the data sets for a small number of clusters requires further explanation. The reason for this seemingly anomalous result is that the ensembles were built using a *random* assignment of the number of clusters for each ensemble member. This number was varied between 2 and 20. Thus, an ensemble with a small number of target clusters might be composed of diverse and unstable individual clusterers. The natural ensemble tendency toward stabilization may not be sufficient to raise the stability of such ensembles to that of the individual clusterers for small k, as demonstrated by the example. This suggests that neither of the stability indices should be lightly ignored and that a combination of the two may be beneficial.

4.4 Is Ensemble Stability Related to Ensemble Accuracy?

Table 2 shows the Pearson correlation coefficients between ensemble accuracy $\mathcal{A}^e$ and the stability indices for the 20 data sets. The correlation coefficients are computed from the vector obtained by collecting the indices for $k = 2 \ldots K_{\max}$.

Shown in Table 3 are the correlations averaged across the 20 data sets between the two ensemble accuracy measures on the one hand and the stability indices.

Table 2 shows that, while, for some data sets, the correlation between ensemble accuracy and ensemble stability is almost perfect (e.g., difficult doughnut, regular and glass), for other data sets, strong negative correlation is observed (e.g., petals, crabs, and noisy-lines). It seems that both measures "fit" some data sets well and fail on others, not necessarily in conjunction with one another.

TABLE 2
Correlation between Stability Indices and Ensemble Accuracy $\mathcal{A}^e$

Dataset	$\mathcal{S}^i_{np}$	$\mathcal{S}^e_{np}$	$\mathcal{S}^i_p$	$\mathcal{S}^e_p$	$\mathcal{S}^*_p$
boat	-0.6701	-0.8748	-0.4794	-0.4159	-0.7344
difficult-doughnut	0.5486	0.8892	0.8509	0.9615	0.9746
easy-doughnut	-0.2926	0.9431	0.9618	0.8749	0.9525
four-gauss	0.5296	0.7673	0.5554	0.8269	0.7189
halfrings	0.9033	0.7395	0.9677	0.7702	0.9319
noisy-lines	-0.4395	-0.9096	-0.1037	-0.7363	-0.9576
petals	0.4193	-0.8258	0.7025	-0.7723	0.0557
regular	0.6191	0.9933	0.5248	0.9521	0.8847
saturn	0.5844	0.5513	0.3310	0.2053	0.4383
spirals	-0.1684	-0.008	-0.3489	0.5372	0.4001
contractions	-0.9013	-0.7495	-0.8803	0.9717	0.9532
crabs	0.7401	-0.7746	0.7367	-0.9258	-0.1571
glass	-0.9116	-0.1551	-0.8356	0.9651	0.6609
ionosphere	-0.6619	-0.9604	-0.493	0.6819	0.6270
iris	0.4931	0.5586	0.6753	0.5117	0.6385
respiratory	0.7982	-0.7617	0.8867	-0.5453	0.5802
segmentation	-0.0452	-0.4380	-0.1374	-0.3119	-0.4649
soybean	-0.1145	0.3724	0.5450	0.6252	0.5981
thyroid	-0.4727	0.3063	0.1929	0.7118	0.8841
wine	0.5648	-0.8414	0.6447	-0.3689	0.6824

The last columns in Tables 2 and 3 present the correlation with a new stability measure defined as

$$\mathcal{S}^*_p(k) = \mathcal{S}^i_p(k) + \mathcal{S}^e_p(k). \quad (4)$$

The rationale for this measure comes from the argument above about the counterintuitive finding that, for a small number of clusters, single clusterers appear to be more stable than cluster ensembles.[5] The final goal in devising a stability measure is to use it to guide the choice of a better ensemble. Thus, we would like to be able to relate it to the ensemble accuracy. The example in Fig. 4 shows that individual clusterers can be stable but incorrect in the case of a small number of target clusters, k. This means that high stability indicated by $\mathcal{S}_{p^i(k)}$ for small k may not be trusted to predict high accuracy of the clustering result. Instead of a stable single classifier, we can use an ensemble, but, according to Table 3, it seems that the ensemble stability alone is not a very good accuracy predictor either. The choice of the sum as a stability measure was based on the observation that coincidental failures did not happen too often. While ensemble stability slightly dominates individual stability in terms of correlation (Table 3), they rather complement one another, and there could be a benefit in combining the two. We tried the sum as the simplest way for such combination, without a theoretical ground as to why we should do so.

Table 4 gives the list of the data sets sorted by $Corr(\mathcal{S}^i_p, \mathcal{A}^i)$ and also $Corr(\mathcal{S}^*_p, \mathcal{A}^t)$. The maximum achievable accuracy (obtained in the experiment) for each data set is also shown. The sorted lists show that stability, both individual and combined, relates almost perfectly with the respective accuracy for some data sets and completely fails for other data sets. An interesting example in this table is the "regular" data set. It contains 16 clusters which could be identified by k-means for k = 16. Thus, the maximum accuracy is high, both for individual clusterers (0.846) and for the ensemble (1.000). However, while the individual $\mathcal{S}^i_p$ does not correlate very well with $\mathcal{A}^i$, the correlation between $\mathcal{S}^*_p$ and $\mathcal{A}^t$ is very high (0.902). This means that the ensemble will be much more likely to find the 16 clusters if k is picked by the maximum stability. Not only is the accuracy better but the chance of achieving it is better too, which demonstrates the advantage of using a cluster ensemble together with a stability measure.

To enable visual evaluation of the relationship between accuracy and stability, Fig. 5 plots $\mathcal{S}^i_p$, $\mathcal{S}^e_p$, and $\frac{\mathcal{S}^*_p}{2}$, and ensemble accuracy $\mathcal{A}^t$ as functions of k for the thyroid and petal data sets. For the thyroid data, $\mathcal{S}^e_p$ matches the shape of $\mathcal{A}^t$ very well, whereas $\mathcal{S}^i_p$ does not. The combined measure exhibits a stronger correlation with $\mathcal{A}^t$ than either of the two measures does individually. The petal data set has a poor match between ensemble stability and accuracy, but a good match between $\mathcal{S}^i_p$ and $\mathcal{A}^t$. The combined stability measure is inferior to the individual measure, but reaches its maximum at the right number of clusters ($k = 4$). Thus, if we use one of $\mathcal{S}^i_p$ or $\mathcal{S}^e_p$, we would have a good predictor of accuracy on one of the data sets and a poor predictor on the other. If we use $\mathcal{S}^*_p$, we would have a reasonable predictor on both data sets.

As argued earlier, stability would measure the quality of a particular clustering method rather than a general property of the data set. According to Tables 2 and 3, the combined stability index, $\mathcal{S}^*_p$ fares better than both the individual and the ensemble stability indices. An interesting question here is whether stability-accuracy correlation is better when the data set is easy or difficult to cluster. To answer this question, Fig. 6 displays a scatterplot of 20 points corresponding to the data sets in the plane spanned by the maximum possible accuracy for each data set, i.e., $\max_k \mathcal{A}^t(k)$, and the correlation between $\mathcal{A}^t$ and $\mathcal{S}^*_p$ calculated across k.

With the exception of the petals data set, there is no point in the zone where $\max_k \mathcal{A}^t(k) > 0.6$ and Correlation $(\mathcal{A}^t, \mathcal{S}^*_p) < 0.5$. This suggests that if high accuracy is possible, the correlation will be reasonably strong. The exception is the petal data set where high accuracy is possible but the combined index $\mathcal{S}^*_p$ may not pick it up because it is not well related to accuracy. However, Fig. 5 shows that, even for this worst-case scenario, a good ensemble will be selected if we pick the ensemble with the maximum $\mathcal{S}^*_p$. In fact, this will be the ensemble also picked by the individual measure, $\mathcal{S}^i_p$, which exhibits much stronger correlation with accuracy for this data set.

On the other hand, high correlation does not guarantee high accuracy as the contractions data set demonstrates. We also note that there are no data sets for which $\max_k \mathcal{A}^t(k) < 0.25$ and Correlation $(\mathcal{A}^t, \mathcal{S}^*_p) > 0.5$.

In other words, if high accuracy is possible, it is likely that the stability index might work well for choosing a good ensemble. If high accuracy is not possible, applying the index will do no harm as the result will not be useful anyway.

4.5 How Good Is Ensemble Stability as a Cluster Validity Measure?

To answer this question, we consider the following ways for determining the number of clusters:

1. True k. We assume that there is an oracle to give the true number of clusters for each data set. With the reservations explained above, we assume that the number of clusters for the real data sets is equal to the number of classes.

5. We also tried a combined stability index between $\mathcal{S}^i_{np}$ and $\mathcal{S}^e_{np}$ but the results were worse and we do not show them here.

TABLE 3
Correlation between Stability Indices and Accuracy Averaged Across the 20 Data Sets

Accuracy measure	$\mathcal{S}^i_{np}$	$\mathcal{S}^e_{np}$	$\mathcal{S}^i_p$	$\mathcal{S}^e_p$	$\mathcal{S}^*_p$
$\mathcal{A}^i$ (individual)	0.3410	0.2651	0.4958	0.0942	0.3191
$\mathcal{A}^e$ (100 ensembles of 25 clusterers each)	0.0761	-0.0589	0.2649	0.2759	0.4333
$\mathcal{A}^t$ (one ensemble of 2500 clusterers)	0.1348	-0.0045	0.2754	0.2401	0.4174

2. k-total. Consider the whole ensemble of 2,500 clusterers. The consensus matrix for the ensemble is submitted as the similarity matrix to the single linkage procedure acting as consensus function. k-total is the number of clusters corresponding to the largest jump of the distance criterion function. This is a traditional way of choosing the number of clusters when using single linkage.
3. k-majority. Consider now the 100 ensembles of 25 clusterers each. The consensus matrix of each ensemble was submitted to single linkage in order to get the ensemble partition. The stopping k is again the number of clusters corresponding to the largest jump in the criterion for a particular ensemble. k-majority is the value most often selected among the 100 suggested ks.

TABLE 4
Sorted Correlations and Maximum Achievable Accuracy

Dataset	Corr $(\mathcal{S}^i_p, \mathcal{A}^i)$	max $\mathcal{A}^i$	Dataset	Corr $(\mathcal{S}^*_p, \mathcal{A}^t)$	max $\mathcal{A}^t$
contractions	0.977	0.331	difficult-doughnut	0.962	0.640
difficult-doughnut	0.965	0.446	halfrings	0.911	1.000
halfrings	0.934	0.582	regular	0.902	1.000
four-gauss	0.930	0.835	contractions	0.872	0.291
wine	0.841	0.366	thyroid	0.822	0.594
easy-doughnut	0.829	0.410	four-gauss	0.763	0.973
saturn	0.829	0.036	easy-doughnut	0.722	1.000
respiratory	0.824	0.116	wine	0.658	0.403
iris	0.779	0.673	glass	0.633	0.301
soybean	0.761	0.589	ionosphere	0.624	0.296
petals	0.610	0.904	soybean	0.538	0.937
regular	0.491	0.846	iris	0.504	0.713
glass	0.386	0.258	respiratory	0.432	0.100
spirals	0.328	0.059	spirals	0.410	0.426
ionosphere	0.265	0.208	saturn	0.335	0.060
boat	0.236	0.428	petals	0.005	0.894
thyroid	0.199	0.460	crabs	-0.081	0.044
segmentation	-0.284	0.378	segmentation	-0.298	0.495
crabs	-0.385	0.037	boat	-0.429	0.511
noisy-lines	-0.599	0.161	noisy-lines	-0.937	0.409

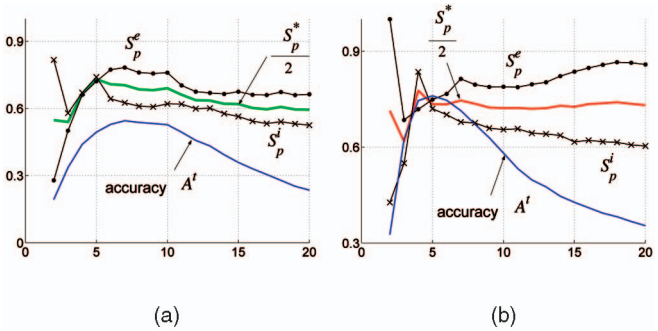

Fig. 5. Stability $\mathcal{S}^i_p$, $\mathcal{S}^e_p$, and $\frac{\mathcal{S}^*_p}{2}$, and ensemble accuracy $\mathcal{A}^t$ as functions of k for the thyroid and petal data sets.

4-8. The numbers of clusters obtained through the maxima of the five stability indices explored in this study.
9. Best k. The maximum of the ensemble accuracy is identified together with the corresponding k. This is again a type of oracle solution which will gauge the maximum achievable accuracy for a particular data set.

The number of clusters produced by an ensemble was further compared to an empirically set threshold. If the suggested number of clusters exceeded 80 percent of the number of points in the data, the number was reassigned to 1 and no cluster structure was reported. Also, since we limited the study to $K_{\max} = 20$ clusters, all numbers obtained for k-total and k-majority greater than 20 were reassigned to 20.

Table 5 shows the suggested number of clusters, k^*, and the corresponding ensemble accuracies, $\mathcal{A}^e(k^*)$, for the 20 data sets.

For comparison, Table 6 displays the averaged *percentage achievement* of $\mathcal{A}^i$, $\mathcal{A}^e$, and $\mathcal{A}^t$ for the suggested k, as in Table 5. The percentage achievement of method X is the achieved accuracy $\mathcal{A}$ divided by the maximum possible $\mathcal{A}$ for this data set across all k multiplied by 100. Shown also are the average *ranks* of the eight methods for suggesting k. The best k was excluded from this comparison because it will always give the best solution and occupy the winning place anyway. The ranks were calculated so that, for each data set, the most accurate method received rank 1, the next best received rank 2, etc. Thus, the worst method will receive rank 8 for a particular data set. If there was a tie, the ranks were recalculated so that the tied methods altogether received the sum of the ranks for the places they would have if there was no tie. For example, if methods B, C, and D have the same

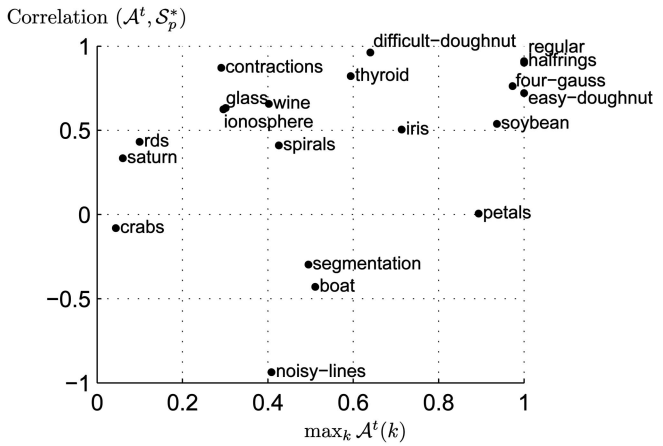

Fig. 6. Correlation between $\mathcal{A}^t$ and $\mathcal{S}^*_p$ versus $\max_k \mathcal{A}^t(k)$ for the 20 data sets.

TABLE 5
Suggested Number of Clusters, k^*, and the Corresponding Ensemble Accuracies, $\mathcal{A}^e(k^*)$

Dataset	true k	k-total	k-maj	$k\left(\mathcal{S}^i_{np}\right)$	$k\left(\mathcal{S}^e_{np}\right)$	$k\left(\mathcal{S}^i_{p}\right)$	$k\left(\mathcal{S}^e_{p}\right)$	$k\left(\mathcal{S}^*_{p}\right)$	best k
boat	3 (0.42)	2 (0.34)	2(0.34)	2 (0.34)	20 (0.28)	2 (0.34)	20(0.28)	2(0.34)	7(0.48)
contractions	2 (0.02)	12 (0.24)	2(0.02)	3 (0.04)	2 (0.02)	3 (0.04)	14(0.23)	13(0.24)	11(0.24)
crabs	2 (0.03)	2 (0.03)	20(0.01)	2 (0.03)	20 (0.01)	2 (0.03)	20(0.01)	4(0.03)	3(0.03)
difficult-d	2 (0.26)	4 (0.53)	2(0.26)	5 (0.56)	7 (0.58)	4 (0.53)	7(0.58)	7(0.58)	6(0.58)
easy-d	2 (0.66)	3 (0.74)	3(0.74)	2 (0.66)	3 (0.74)	2 (0.66)	3(0.74)	3(0.74)	3(0.74)
4-gauss	4 (0.90)	6 (0.97)	5(0.95)	5 (0.95)	6 (0.97)	5 (0.95)	6(0.97)	5(0.95)	6(0.97)
glass	6 (0.24)	6 (0.24)	2(0.02)	2 (0.02)	2 (0.02)	2 (0.02)	19(0.28)	6(0.24)	10(0.29)
halfrings	2 (0.98)	3 (0.81)	2(0.98)	2 (0.98)	2 (0.98)	2 (0.98)	2(0.98)	2(0.98)	2(0.98)
ionosphere	2(-0.01)	4(-0.02)	20(0.20)	2(-0.01)	2(-0.01)	2(-0.01)	20(0.20)	20(0.20)	20(0.20)
iris	3 (0.60)	2 (0.57)	2(0.57)	2 (0.57)	2 (0.57)	2 (0.57)	2(0.57)	2(0.57)	4(0.66)
noisys	2 (0.25)	2 (0.25)	2(0.25)	2 (0.25)	19 (0.12)	2 (0.25)	19(0.12)	20(0.12)	4(0.35)
petals	4 (0.75)	2 (0.33)	2(0.33)	4 (0.75)	2 (0.33)	4 (0.75)	2(0.33)	4(0.75)	5(0.76)
respiratory	2 (0.04)	3 (0.09)	3(0.09)	3 (0.09)	14 (0.03)	3 (0.09)	12(0.04)	3(0.09)	4(0.09)
regular	16 (1.00)	16 (1.00)	16(1.00)	4 (0.24)	16 (1.00)	4 (0.24)	16(1.00)	17(0.98)	16(1.00)
saturn	2 (0.02)	2 (0.02)	2(0.02)	10 (0.03)	20 (0.04)	8 (0.02)	4(0.02)	19(0.04)	17(0.04)
segment	7 (0.25)	2 (0.00)	2(0.00)	4 (0.12)	2 (0.00)	4 (0.12)	2(0.00)	2(0.00)	19(0.45)
soybean	4 (0.74)	3 (0.65)	3(0.65)	2 (0.48)	3 (0.65)	2 (0.48)	3(0.65)	3(0.65)	5(0.81)
spirals	2 (0.11)	2 (0.11)	20(0.13)	20 (0.13)	20 (0.13)	4 (0.14)	20(0.13)	20(0.13)	6(0.15)
thyroid	3 (0.34)	12 (0.46)	20(0.24)	2 (0.19)	2 (0.19)	2 (0.19)	7(0.55)	5(0.49)	7(0.55)
wine	3 (0.32)	6 (0.29)	2(0.29)	2 (0.29)	20 (0.13)	2 (0.29)	8(0.27)	2(0.29)	3(0.32)
	(0.40)	(0.38)	(0.35)	(0.33)	(0.34)	(0.33)	(0.40)	(0.42)	(0.48)

TABLE 6
Overall Accuracies and Ranks for the Chosen Number of Clusters

Accuracy	true k	k-total	k-maj	$k\left(\mathcal{S}^i_{np}\right)$	$k\left(\mathcal{S}^e_{np}\right)$	$k\left(\mathcal{S}^i_{p}\right)$	$k\left(\mathcal{S}^e_{p}\right)$	$k\left(\mathcal{S}^*_{p}\right)$
$\mathcal{A}^i$ (individual)	76	64	67	73	69	73	67	**79**
rank	1.875	2.850	2.675	2.500	2.875	2.525	2.875	**1.825**
$\mathcal{A}^e$ (ensemble,	72	75	66	66	58	65	75	**85**
$L = 25$) rank	2.125	2.400	2.875	2.650	3.225	2.625	2.325	**1.775**
$\mathcal{A}^t$ (ensemble,	67	72	61	63	50	62	69	**78**
$L = 2500$) rank	2.125	2.150	2.825	2.450	3.425	2.575	2.425	**2.025**

score, which is the second best after method A, then A will get rank 1 and each of B, C, and D will get rank 3. The ranks were averaged across the 20 data sets. Marked in boldface are the best results in each row.

The tables show that the combined stability index is the best cluster validity index among the eight compared ones, including the true number of clusters. As mentioned before, the true number of clusters may not be the optimal number for which a particular clustering algorithm will disclose, to its best potential, the structure in the data. Cluster ensembles often produce better results for a number of clusters different from true k. The combined stability index appeared to be able to identify, if not the optimal k, then a close rival. It should be noted, however, that, given this number of experiments the differences between the eight methods were not found to be statistically significant according to the Friedman Two-Way ANOVA.

One possible explanation for the lack of statistical significance of the differences is that the only parameter that is altered is the number of clusters, k. The clustering method is the same in all experiments, an ensemble of 25 clusterers. Because of this, multiple ties can be expected, as seen in Table 5, corresponding to the same ensemble accuracy. Thus, the total rankings of the methods are likely to be similar.

Note that, while the real data sets were chosen randomly, the artificial sets were designed with specific difficulties in mind. They do not represent a random sample from data sets which may occur in practice; they are, rather, special cases, some of them intentionally created to be impossible to solve with k-means. Hence, a statistical conclusion based on the current selection of data sets is not necessarily valid in the general case.

Finally, to show how close the decision by $k(\mathcal{S}^*_p)$ is to the maximum possible accuracy, Fig. 7 plots a bar graph with the maximum $\mathcal{A}^t$ for the data sets (gray), and the corresponding accuracy obtained for $k(\mathcal{S}^*_p)$ clusters. The combined stability index $\mathcal{S}^*_p$ gives close to optimal performance on a large majority of the data sets.

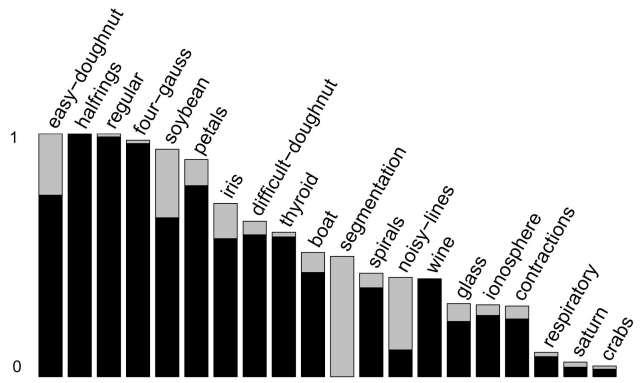

Fig. 7. Maximum possible accuracy (gray) and obtained accuracy using $k\left(\mathcal{S}^*_p\right)$ (black).

5 Summary and Conclusions

The stability of clustering algorithms relying on a random component is an important issue. High stability across different runs is considered to be an asset. We examined the stability of cluster ensembles consisting of k-means clusterers, each clusterer run with a random initialization and with a random assignment of k. The stability of the ensemble was evaluated and compared to the stability of the individual k-means for values of k from 2 to 20. The questions addressed by the experiment and the answers found are summarized below:

1. *Are ensembles more stable than individual clusterers*? Generally, yes. This is more clearly expressed for larger k (Fig. 3). We note, however, that the true number of clusters for the data sets in this study is relatively small, which means that the dominance between individual stability and ensemble stability around the true number of clusters is not clear-cut.
2. *Is ensemble stability related to ensemble accuracy*? We discovered an interesting phenomenon about the stability-accuracy relationship. While, for some data sets, $\mathcal{S}^e(k)$ and $\mathcal{A}^e(k)$ exhibited almost perfect positive correlation (0.97, for the glass data), for other data sets, almost perfect negative correlation was observed (-0.93, for the crabs data). Thus, we introduced a combined stability index, $\mathcal{S}^*_p$, aimed at preserving the chance of finding a suitable k. If we use only the ensemble stability index for cluster validity, we might miss a peak of stability around the true number of clusters detected by the individual stability index. An example of this phenomenon is the result for the boat data set in Table 5. The ensemble on its own suggests $k(\mathcal{S}^e_p) = 20$ clusters (accuracy 0.28). The combined index agrees with the individual index on two clusters (accuracy 0.34). An example of the opposite case is the ionosphere data set, where the initial peak of the individual stability at $k = 2$ (the true number) is not sufficient to pull up the combined index to reach maximum at $k = 2$. Even though the true number of clusters is not recovered by the combined index, the accuracy of the ensemble is better for $k = 20$ as chosen by the ensemble and, subsequently, by the combined index. In general, $\mathcal{S}^*(k)$ correlated reasonably with $\mathcal{A}^e(k)$ and $\mathcal{A}^t(k)$ although, again, strongly varying across data sets (Tables 2, 3, and 4). In reality, we will not have true labels and will not know which of the two situations we are in. The best option is to use $\mathcal{S}^*(k)$ as it has the fewer number of negative correlations compared to the other four stability indices.

   We looked further to single out the data sets with negative correlations. The scatterplot in Fig. 6 suggests that if high accuracy is possible, it is likely that the stability index might correlate well with the accuracy (points in the top right corner).
3. *How good is ensemble stability as a cluster validity measure?* Here, we followed a hypothesis strongly motivated and used for cluster validity in the relevant literature. This hypothesis states that a point of stability of a clustering algorithm corresponds to a structure found in the data. Therefore, we used the maximum stability measures to pick the number of clusters. Without an oracle, the next most widely used heuristic for selecting number of clusters is cutting the dendrogram of a hierarchical clustering algorithm at the largest jump of the distance criterion. We used this method in two variants: With an ensemble of 2,500 clusterers, and as the majority k of 100 ensembles of 25 clusterers each. The combined stability proposed here gave the best results compared to pairwise and nonpairwise individual and ensemble stabilities (Table 5). Curiously, a small improvement of the clustering accuracy was also observed when cluster ensembles were assigned k found through the combined stability index, $\mathcal{S}^*_p(k)$, compared to the known (assumed true) number of clusters.

   There are many open questions here. First, the findings of this study suggest a methodology for measuring cluster validity. As a large number of clusterers will be produced and previous studies suggest that large ensembles fare better [16], [26], we may use a large ensemble anyway. In this paper, we considered both $\mathcal{A}^e(k)$ (averaged across the 100 ensembles of 25 clusterers) and $\mathcal{A}^t(k)$ (for the whole ensemble of 2,500 clusterers). The overall results with the whole ensemble were slightly better, although, to verify this statistically, a number of large ensembles have to be constructed. The clustering procedure is then the following:

   a. Choose $K_{\max}$, the ensemble size L, and number of ensembles T.
   b. Generate $L \times T$ k-means clusterers with random k from 2 to $K_{\max}$.
   c. Group the clusterers randomly into T ensembles of L and evaluate $\mathcal{S}^*_p(k)$ using (1), for $k = 2, \ldots, K_{\max}$.
   d. Find $k^* = \arg\max_k\{\mathcal{S}^*_p(k)\}$.
   e. Pool the $L \times T$ clusterers together, calculate the consensus matrix $\mathbf{M}$, and feed it as a similarity matrix to a single linkage clusterer. Cut the dendrogram at k^* clusters and return the labeling P^*.

It is interesting to find out how stable and consistent the results would be for smaller L and T than considered here and probably for larger $K_{\max}$.

Another open question is whether findings similar to ours will hold for different types of base clusterers and consensus functions. We chose k-means as the base clusterer and single linkage as the consensus function because they are simple and efficient as found by many authors. Clearly, for some data sets used here, k-means and ensembles thereof (with the chosen number $K_{\max}$) were inadequate, e.g., crab, saturn, boat, respiratory, and two-spirals. Path-based clustering would have been a suitable alternative [9]. Without prior knowledge or at least a hypothesis about the type of clusters, we cannot predict which method will be more suitable. Therefore, experiments with k-means ensembles and path-based ensembles should be carried out on the whole variety of data sets, not only the ones which are known to have benefited from a particular clustering method. It will be interesting to keep the collection of data sets and extend the study to other clustering methods and ensembles as well.

In this paper, we only evaluated stability with respect to the intrinsic randomness of k-means and k-means ensembles. Many previous studies use resampling or subsampling

of the data set. A parallel can be drawn with stability estimation in supervised learning based on small alterations of the training data [3], [6]. Theoretical results for clustering methods and ensembles can be sought following this pattern. The stability indices considered here can be applied without change to ensembles of different structures, diversied approaches, and consensus functions. However, the answers to the three main questions offered here may not be valid for other ensemble methods. In other words, there may be ensemble types for which stability is a much better predictor of ensemble accuracy.

Finally, stability and the stability-plasticity dilemma for online clustering and online cluster-validity presents a challenging extension of this study.

APPENDIX

Below is a brief explanation of the two real data sets contractions and respiratory. The data can be downloaded from http://www.informatics.bangor.ac.uk/~kuncheva/patrec1.html.

Contractions. This data set comes from wireless capsule endoscopy [36]. The problem is to detect intestinal contractions in video images sent by a small capsule traveling along the intestinal tract. Contractions which are of interest to the physician constitute about 1 percent of the video time, therefore automatic labeling in preparation for further inspection is necessary. In a video sequence of nine frames, a contraction is represented as the lumen progressively closing and reopening. Twenty-seven features were extracted using basic image descriptors: mean intensity of each frame (nine features), hole size of each frame (nine features), and global contrast of each frame (nine features). Two classes are considered: contractions and noncontractions. The 98 objects (49 in each class) were manually selected to represent the most clear examples of the classes. Note that the prior probabilities for the two classes cannot be evaluated as the sample proportions.

Respiratory. The respiratory data set consists of the clinical records (17 features) for 85 newborn children with two types of respiratory distress syndrome (RDS): Hyaline Membrane Disease (HMD) and non-HMD. The two classes need urgent and completely different treatments, therefore an accurate RDS classification is crucial within the first few hours after delivery.

ACKNOWLEDGMENTS

This study was carried out as a part of INTAS project #YS04-83-2942.

REFERENCES

[1] H. Ayad and M. Kamel, "Finding Natural Clusters Using Multiclusterer Combiner Based on Shared Nearest Neighbors," *Proc. Fourth Int'l Workshop Multiple Classifier Systems,* 2003.

[2] C.L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases," 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[3] O. Bousquet and A. Elisseeff, "Stability and Generalization," *J. Machine Learning Research,* vol. 2, no. 3, pp. 499-526, 2002.

[4] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[5] S. Dudoit and J. Fridlyand, "Bagging to Improve the Accuracy of a Clustering Procedure," *Bioinformatics,* vol. 19, no. 9, pp. 1090-1099, 2003.

[6] A. Elisseeff, T. Evgeniou, and M. Pontil, "Stability of Randomized Learning Algorithms," *J. Machine Learning Research,* vol. 6, no. 1, pp. 55-79, 2005.

[7] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A Stability Based Method for Discovering Structure in Clustered Data," *Proc. Pacific Symp. Biocomputing,* pp. 6-17, 2002.

[8] X.Z. Fern and C.E. Brodley, "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach," *Proc. 20th Int'l Conf. Machine Learning,* pp. 186-193, 2003.

[9] B. Fischer and J.M. Buhmann, "Bagging for Path-Based Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 11, pp. 1411-1415, Nov. 2003.

[10] A. Fred, "Finding Consistent Clusters in Data Partitions," *Proc. Second Int'l Workshop Multiple Classifier Systems,* 2001.

[11] A. Fred and A.K. Jain, "Data Clustering Using Evidence Accumulation," *Proc. 16th Int'l Conf. Pattern Recognition,* pp. 276-280, 2002.

[12] A. Fred and A.K. Jain, "Combining Multiple Clusterungs Using Evidence Accumulation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 6, pp. 835-850, June 2005.

[13] A. Fred and A.K. Jain, "Robust Data Clustering," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2003.

[14] J. Ghosh, "Multiclassifier Systems: Back to the Future," *Proc. Third Int'l Workshop Multiple Classifier Systems,* 2002.

[15] A.D. Gordon, *Classification.* Boca Raton, Fla.: Chapman and Hall/CRC, 1999.

[16] D. Greene, A. Tsymbal, N. Bolshakova, and P. Cunningham, "Ensemble Clustering in Medical Diagnostics," Technical Report TCD-CS-2004-12, Dept. of Computer Science, Trinity College, Dublin, Ireland, 2004.

[17] S.T. Hadjitodorov, L.I. Kuncheva, and L.P. Todorova, "Moderate Diversity for Better Cluster Ensembles," *Information Fusion,* 2005.

[18] X. Hu and I. Yoo, "Cluster Ensemble and Its Applications in Gene Expression Analysis," *Proc. Second Asia-Pacific Bioinformatics Conf.,* 2004.

[19] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, pp. 193-218, 1985.

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

[21] L.I. Kuncheva and S.T. Hadjitodorov, "Using Diversity in Cluster Ensembles," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics,* 2004.

[22] T. Lange, V. Roth, M.L. Braun, and J.M. Buhmann, "Stability-Based Validation of Clustering Solutions," *Neural Computation,* vol. 16, pp. 1299-1323, 2004.

[23] M.H. Law and A.K. Jain, "Cluster Validity by Boostrapping Partitions," Technical Report MSU-CSE-03-5, Michigan State Univ., 2003.

[24] E. Levine and E. Domany, "Resampling Method for Unsupervised Estimation of Cluster Validity," *Neural Computation,* vol. 13, pp. 2573-2593, 2001.

[25] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 12, pp. 1650-1654, Dec. 2002.

[26] B. Minaei, A. Topchy, and W. Punch, "Ensembles of Partitions via Data Resampling," *Proc. Int'l Conf. Information Technology: Coding and Computing,* 2004.

[27] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning,* vol. 52, pp. 91-118, 2003.

[28] G. BelMufti, P. Bertrand, and L. ElMoubarki, "Determining the Number of Groups from Measures of Cluster Validity," *Proc. Int'l Symp. Applied Stochastic Models and Data Analysis,* pp. 404-414, 2005.

[29] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Statistical Assoc.,* vol. 66, pp. 846-850, 1971.

[30] B.D. Ripley, *Pattern Recognition and Neural Networks.* Cambridge: Univ. Press, 1996.

[31] V. Roth, T. Lange, M. Braun, and J. Buhmann, "A Resampling Approach to Cluster Validation," *Proc. Conf. Computational Statistics,* pp. 123-128, 2002.

[32] A. Strehl and J. Ghosh, "Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions," *J. Machine Learning Research*, vol. 3, pp. 583-618, 2002.
[33] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Data Set via Gap Statistic," *J. Royal Statistical Soc. B*, vol. 63, pp. 411-423, 2001.
[34] A. Topchy, A.K. Jain, and W. Punch, "Combining Multiple Weak Clusterings," *Proc. IEEE Int'l Conf. Data Mining*, pp. 331-338, 2003.
[35] A. Topchy, A.K. Jain, and W. Punch, "A Mixture Model for Clustering Ensembles," *Proc. SIAM Conf. Data Mining*, pp. 379-390, 2004.
[36] F. Vilarino, L.I. Kuncheva, and P.I. Radeva, "ROC Curves in Video Analysis Optimization in Intestinal Capsule Endoscopy," *Pattern Recognition Letters*, 2005.
[37] A. Weingessel, E. Dimitriadou, and K. Hornik, "An Ensemble Method for Clustering," 2003, working paper, http://www.ci.tuwien.ac.at/Conferences/DSC-2003.

Ludmila I. Kuncheva received the MSc degree from the Technical University, Sofia, in 1982 and the PhD degree from the Bulgarian Academy of Sciences in 1987. Until 1997, she worked at the Central Laboratory of Biomedical Engineering, Bulgarian Academy of Sciences, as a senior research associate. Dr. Kuncheva is currently a reader at the School of Informatics, University of Wales, Bangor, United Kingdom. Her interests include pattern recognition, classifier combination, diversity measures, and nearest neighbor classifiers. She has published more than 100 research papers and two books. She won the best paper award for 2006 in the *IEEE Transactions on Fuzzy Systems* and the Sage best Transaction paper award for 2003 across the *IEEE Transactions on Systems, Man, and Cybernetics A*, *B*, and *C*. She has served as an associate editor for *IEEE Transactions on Fuzzy Systems* and is currently an associate editor for the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. She is a member of the IEEE.

Dmitry P. Vetrov received the MSc degree from Moscow State University in 2003 and is currently a PhD student there. He also works as a mathematician at the Dorodnicyn Computing Center of the Russian Academy of Sciences. His areas of interests include machine learning, data-mining, and artificial intelligence. He is the author of 27 papers. In 2005, he won the scholarship of the President of Russian Federation.