Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Fusion 7 (2006) 264–275
INFORMATION FUSION
www.elsevier.com/locate/inffus

Moderate diversity for better cluster ensembles

Stefan T. Hadjitodorov [a,*], Ludmila I. Kuncheva [b], Ludmila P. Todorova [a]

[a] Center for Biomedical Engineering (CLBME), Bulgarian Academy of Sciences, "Acad G. Bonchev" Str., block 105, Sofia 1113, Bulgaria
[b] School of Informatics, University of Wales—Bangor, Bangor, Gwynedd LL57 1UT, United Kingdom

Received 21 September 2004; received in revised form 29 January 2005; accepted 29 January 2005
Available online 3 March 2005

Abstract

Adjusted Rand index is used to measure diversity in cluster ensembles and a diversity measure is subsequently proposed. Although the measure was found to be related to the quality of the ensemble, this relationship appeared to be non-monotonic. In some cases, ensembles which exhibited a moderate level of diversity gave a more accurate clustering. Based on this, a procedure for building a cluster ensemble of a chosen type is proposed (assuming that an ensemble relies on one or more random parameters): generate a small random population of cluster ensembles, calculate the diversity of each ensemble and select the ensemble corresponding to the median diversity. We demonstrate the advantages of both our measure and procedure on 5 data sets and carry out statistical comparisons involving two diversity measures for cluster ensembles from the recent literature. An experiment with 9 data sets was also carried out to examine how the diversity-based selection procedure fares on ensembles of various sizes. For these experiments the classification accuracy was used as the performance criterion. The results suggest that selection by median diversity is no worse and in some cases is better than building and holding on to one ensemble.

Keywords: Pattern recognition; Machine learning; Multiple classifiers; Cluster ensembles; Diversity measures, Adjusted Rand index

1. Introduction

Cluster ensembles emerged recently as a coherent stream out of the multiple classifier systems area [12,27,28,8–10,6,23,1,14,11]. They are deemed to be better than single clustering algorithms for discovering complex or noisy structures in the data. The strongest argument in favour of cluster ensembles is as follows. It is known that the current off-the-shelf clustering methods may suggest very different structures in the same data. This is the result of the different clustering criteria being optimized. There is no layman guide to choosing a clustering method for a given data set and so an inexperienced user runs the risk of picking an inappropriate clustering method. There is no ground truth against which the result can be matched, therefore there is no critique to the user's choice. Cluster ensembles provide a more universal solution in that various structures and shapes of clusters present in data may be discovered by the same ensemble method, and the solution is less dependent upon the chosen ensemble type [27].

Let $\mathbf{Z}$ be a data set and let $P = \{P_1, \ldots, P_L\}$ be a set of partitions on $\mathbf{Z}$. Each partition is obtained by applying a clustering algorithm on $\mathbf{Z}$ or a subset of it. We assume that the partitions are generated by varying a random parameter of the clustering algorithm, for example starting the algorithm from L random initializations. The clustering algorithm (or run) which produces P_i will be called here an "ensemble member" or "clusterer". The clusterers may be versions of the same clustering algorithm or different clustering algorithms. For simplicity, the same notation, P_i, will be used both

* Corresponding author. Tel.: +35 929875819; fax: +35 929816629.
E-mail address: sthadj@argo.bas.bg (S.T. Hadjitodorov).

1566-2535/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2005.01.008

for the clusterer and for the corresponding partition. The goal is to find a single (resultant) partition, P^*, based on the information contained in the set P.

The "accuracy" of a clustering algorithm (or a cluster ensemble) is measured by the match between the partition produced and some known ground-truth partition. A reliable ground-truth partition is seldom available, so most experimental studies employ generated data with pre-specified cluster structure. From the many matching indices suggested in the literature [4,5,16,26], we chose the adjusted Rand index [16] because of the following properties: (1) it has a fixed value of 0 if the two compared partitions are formed independently from one another; (2) in our preliminary experiments, this index was found to have a greater sensitivity to pick out good partitions compared to other indices.

Diversity within an ensemble is of vital importance for its success. An ensemble of identical clusterers or classifiers will not outperform the individual ensemble members. However, finding a sensible quantitative measure of diversity in classifier ensembles has been notoriously hard [19–21]. Diversity in *cluster* ensembles is considered here. A diversity measure is proposed and its relationship with the accuracy of the ensemble is demonstrated. Based on the results, a procedure is suggested for selecting a cluster ensemble from a small population of ensembles. The proposed diversity measure as well as the match index for the ensemble accuracy are based on the Adjusted Rand Index.

The rest of the paper is organized as follows. Section 2 introduces cluster ensembles. Section 3 contains the proposed diversity measure together with some results on its relationship with the ensemble accuracy. At the end of this section we list the steps of our proposed methodology for selecting a cluster ensemble from a small population. Section 4 offers the results from a statistical comparison of the proposed diversity measure with two other measures due to Fern and Brodley [6] and Greene et al. [13]. Section 5 contains an experiment with 9 data sets looking into the relationship between the performance of the proposed selection method and the ensemble size. Section 6 concludes the study.

2. Cluster ensembles

There are various ways to build a cluster ensemble:

- Use different subsets of features (overlapping or disjoint), called feature-distributed clustering in [13,27,28].
- Use different clustering algorithms within the ensemble [15]. Such ensembles are called heterogeneous or hybrid. Ensembles with the same clustering method obtained by varying a random parameter will be called homogeneous.
- Vary a random parameter of the clustering algorithm. For example, run the k-means clustering method from different initializations or generate L random projections the data on a low-dimensional space and run k-means for each projection [6,29].
- Use different a data set for each ensemble member, e.g. re-sampling with or without replacement [3,7,10,22,23], called object-distributed clustering [27,28].

Any combination of the above construction heuristics is also a possible construction method. Once $P_1, \ldots, P_L$ are constructed, the resultant partition P^* has to be found.

The *direct approach* (re-labeling) seeks correspondence between the cluster labels across the partitions and fuses the clusters of the same label [7,27,28,32]. Note that the labels that we assign to the clusters in the individual partitions are arbitrary. Thus two identical partitions might have permuted labels and be perceived as different. Suppose that the correspondence between the partitions has been solved. Then a voting between the clusterers would be straightforward: just count the number of votes for the respective cluster. For c clusters, there are $c!$ permutations of the labels and an exhaustive experiment might not be feasible for large c.

The *feature-based approach* treats the output of each clusterer as a categorical feature. The collection of L features can be regarded as an "intermediate feature space" and another clustering algorithm can be run on it. A mixture model for this case is proposed in [30].

The *hypergraph* approach [27,28] organizes the L partitions into a hypergraph and uses methods for hypergraph partitioning to obtain the ensemble result.

Finally, the *pairwise approach* (also co-association approach) avoids the correspondence task altogether by using a coincidence matrix between all pairs of objects. The matrices for the partitions are then combined and a final clustering is derived thereof.

This study is based on the pairwise approach whose generic algorithm is detailed in Fig. 1. In the traditional implementation of this algorithm (voting k-means [8], evidence accumulation algorithm [9]), the following choices are made: (i) The clusterers are various runs of the k-means algorithm (see for details [2]). (ii) The same number of overproduced clusters, c, is assigned to each clusterer. (iii) The final consensus matrix, $\mathbf{M}$, has an (i, j)th entry as follows:

$$\mathbf{m}_{ij} = \frac{1}{L}\sum_{k=1}^{L} m_{i,j}^{k},$$

i.e., it contains the proportion out of L clusterers which have put objects i and j in the same cluster.

PAIRWISE CLUSTER ENSEMBLE ALGORITHM

1. Initialization. Given is a data set $\mathbf{Z}$ with N elements. Pick the ensemble size L and the number of clusters c. (Note 1: Usually c is larger than the suspected number of clusters so there is "overproduction" of clusters. Note 2: Each clusterer P_k could be assigned a specific number of overproduced clusters, c_k. Note 3: The rule of thumb is $c = \sqrt{N}$.)

2. Building the ensemble members. Generate L partitions of $\mathbf{Z}$ with c clusters in each partition.

3. Finding the resultant partition. For each partition, P_k, form a co-association matrix $M^{(k)}$ [1,8,9,10], termed sometimes connectivity matrix [23] or similarity matrix [27]. The L co-association matrices are of size N by N, and the (i,j)-entry in matrix $M^{(k)}$ is

$$m^k_{i,j} = \begin{cases} 1, & \text{if objects } i \text{ and } j \text{ are in the same cluster in partition } P_k \\ 0, & \text{otherwise} \end{cases}$$

Form a final co-association matrix $\mathbf{M}$ (consensus matrix) $M^{(k)}$, $k = 1,\ldots, L$, and derive the resultant partition P^* using this matrix. (Note: The number of clusters may be pre-specified or found through further analysis of $\mathbf{M}$).

Fig. 1. The generic pairwise cluster ensemble algorithm.

The resultant partition, P^*, is traditionally found in one of two ways. In the first way, $\mathbf{M}$ is "cut" with a pre-specified threshold, θ. All entries greater than θ are set to 1 and the remaining entries are set to 0. The new matrix is treated as the co-association matrix and the respective partition is derived thereof (a choice of $\theta = 0.5$ will correspond to majority voting between the clusterers as to the joint membership of objects i and j). In the second way, the entries of $\mathbf{M}$ are treated as "similarities" and another clustering algorithm is run on them. The common choice is the *single-link algorithm* (see [2]). In fact, cutting $\mathbf{M}$ at a certain threshold is equivalent to running the single link algorithm and cutting the dendrogram obtained from the hierarchical clustering at similarity θ. Viewed in this context, cluster ensemble is a type of stacked clustering, where layers of similarity matrices are generated and clustering algorithms are subsequently applied on them. Our pilot experiments showed slightly better results when a new clustering procedure was applied on the consensus matrix $\mathbf{M}$ used as *data*. This corresponds to a method recently proposed in pattern recognition whereby similarities are treated as new features [24,25]. If not stated otherwise, the experiments in this paper are carried out with a single link clustering using the consensus matrix, $\mathbf{M}$, as the data.

3. Diversity measures for cluster ensembles

The adjusted Rand index needed for both diversity and accuracy of the ensemble is calculated as follows [16]. Let A and B be two partitions on a data set $\mathbf{Z}$ with N objects. Let A have c_A clusters and B have c_B clusters. Denote by

- N_{ij} the number of objects in cluster i in partition A and in cluster j in partition B.
- $N_{.j}$ the number of objects in cluster j in partition B.
- $N_{i.}$ the number of objects in cluster i in partition A.

The adjusted Rand index is

$$t_1 = \sum_{i=1}^{c_A} \binom{N_{i.}}{2}, \quad t_2 = \sum_{i=1}^{c_B} \binom{N_{.j}}{2}, \quad t_3 = \frac{2t_1 t_2}{N(N-1)},$$

$$\mathrm{ar}(A,B) = \frac{\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1+t_2) - t_3},$$

where $\binom{a}{b}$ is the binomial coefficient. If we fix the number of clusters c_A and c_B, and the number of objects in each cluster, and draw A and B randomly (generalized hypergeometric distribution), the adjuster Rand index ar(A, B) should be zero. Values of ar(A, B) close to zero will indicate that by observing A, nothing can be predicted about B, and vice versa.

The accuracy of a clusterer, say P_i, is taken to be ar(P_i, P^{T}), where P^{T} is the known true partition of the data. Respectively, the accuracy of the ensemble is calculated as ar(P^*, P^{T}). Note that 'ar' is used as a measure of *accuracy*. In fact, Adjusted Rand Index measures the degree of departure from the assumption that the two compared clustering results have occurred by chance. A value of zero of this index will mean that the two partitions have been generated completely independently of one another. Thus a value of zero does not mean that there is no match between the labels! One distinguishing feature of our study is that we do not impose the correct (known) number of clusters onto the algorithm either at the stage of building the individual clusterers, or at the aggregation stage. Thus our clustering may end up with e.g. 2, 3, 4 or 5 clusters for a problem with 2 classes. Calculation of a classification error will be suitable when the number of clusters is set equal to the number of classes.

Two approaches to measuring the ensemble diversity are considered—pairwise and non-pairwise. In the pairwise approach, using the adjusted Rand index, the ensemble diversity is

$$D_P = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} (1 - \mathrm{ar}(P_i, P_j)).$$

Note that 'ar' gives similarity between partitions, therefore $1 - \mathrm{ar}$ would be the pairwise diversity. Pairwise diversity in cluster ensembles is discussed in [6]. Instead of the adjusted Rand index, the Normalized Mutual Information (NMI) is used. We have studied NMI in our previous experiments but chose here 'ar' for the reasons stated above. In any case, the two indices have very similar behaviour.

The non-pairwise approach can be subdivided into two: *group diversity* and *individual diversity*.

The measure proposed in [13] can be branded as "group diversity". The mutual information of the consensus matrix **M** is calculated by regarding each entry in the consensus matrix as a probability distribution. The random variable in each cell of the matrix has two values: "Yes" (meaning that objects i and j belong in the same cluster), an estimate of Pr(Yes) being $\mathbf{m}_{ij}$, and "No" with Pr(No) = $1 - \mathbf{m}_{ij}$. The entropy of this distribution is

$$H_{ij} = -(\mathbf{m}_{ij}\log_2(\mathbf{m}_{ij}) + (1 - \mathbf{m}_{ij})\log_2(1 - \mathbf{m}_{ij})),$$

and the overall measure of diversity is the averaged entropy across the consensus matrix **M**,

$$H = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} H_{ij}.$$

The larger the entropy, the more diverse the ensemble is. If all clusterers gave the same partition, then **M** would contain only 0s and 1s, and the entropy would be 0. (Note that by convention $0 \cdot \log(0) = 0$.)

For the *individual diversity* subgroup, the ensemble decision is derived and each clusterer is assigned a diversity value measuring its difference from the ensemble decision. Recalling that P^* denotes the resultant clustering (the ensemble decision), the individual diversity of clusterer P_i is $1 - \mathrm{ar}(P_i, P^*)$. To obtain an overall measure of diversity we may simply take the average of the L individual diversities,

$$D_{np-1} = \frac{1}{L} \sum_{i=1}^{L} (1 - \mathrm{ar}(P_i, P^*)).$$

In our previous studies [18] we found that ensembles that exhibit a larger spread of individual diversities are generally better than ensembles with a smaller spread. Therefore, we choose as the second non-pairwise diversity measure, D_{np-2} the standard deviation of the individual diversities

$$D_{np-2} = \sqrt{\frac{1}{(L-1)} \sum_{i=1}^{L} (1 - \mathrm{ar}(P_i, P^*) - D_{np-1})^2}.$$

It turned out that the spread alone was not strongly related to the ensemble accuracy either. Therefore a third non-pairwise diversity measure, D_{np-3}, is proposed here based on the following intuition. Since it is believed that the ensemble decision is close to the true labeling of the data, the accuracy of the individual clusterers may be estimated based on how close they are to the ensemble decision. Thus larger values of $1 - D_{np-1}$ should be preferred. On the other hand, variability within the ensemble can be estimated by the spread of the individual diversities. Large variability will be indicated by larger values of D_{np-2} The simplest compromise measure, D_{np-3} can be devised as

$$D_{np-3} = \tfrac{1}{2}(1 - D_{np-1} + D_{np-2}).$$

Another compromise measure can be constructed using the coefficient of variation [1]

$$D_{np-4} = \frac{D_{np-2}}{D_{np-1}}.$$

The goal is to find a measure related to the quality of the ensemble so that we can pick from a set of ensembles the one that is most likely to be good. Figs. 2 and 3 show the relationship between the six diversity measures: D_p, H, D_{np-1}, D_{np-2}, D_{np-3} and D_{np-4} and the ensemble accuracy, calculated through the adjusted Rand index for two data sets used in the experiments reported in the following section. Fig. 2 shows the six plots for an artificial data set consisting of 4 Gaussian clusters, called "four-gauss". Fig. 3 shows the plots with the UCI wine data (http://www.ics.uci.edu/~mlearn/MLRepository. html). Each plot contains the accuracies of 100 ensembles against their diversity. The ensembles in Fig. 2 were built on 100 different data sets of 4 Gaussian clusters sampled from the same distribution. The 100 ensembles in Fig. 3 were built by varying the random parameters of the ensemble (explained later in the experimental section). The solid line shows the ensemble accuracy and the line with the dot markers shows the averaged individual accuracy.

The behaviour of the measures except for D_{np-1} and D_{np-3} is rather erratic. On the other hand, D_{np-1} shows an unexpected pattern. The four-gauss data has a clear-cut structure, however the accuracy of the ensemble is inversely related with its diversity D_{np-1}. As seen in Fig. 2(c), more diverse ensembles are less accurate than less diverse ensembles. We could attribute this phenomenon to the intuition that more diversity would be associated with many clusterers not getting right the clustering structure and therefore having lower individual accuracy. It is interesting to observe that the average accuracy of the individual clusterers (dot marker) shows no substantial increase or decrease. We can say that, although D_{np-1} is a valid measure of diversity by design, its behaviour with respect to accuracy is to some extent counterintuitive.

Figs. 2 and 3 display the two typical patterns that we found in our experiments. For the sets with a clear-cut cluster structure, such as four-gauss, the proposed indices D_{np-3} and D_{np-4} have roughly a monotonically increasing relationship with the ensemble accuracy. For other data sets, an example of which is the wine data set, there is a "cup" pattern, showing that *moderate* values of the two indices are associated with higher ensemble accuracy.

[1] Other combinations have been tried as well, e.g. ones that included both pairwise and non-pairwise indices (D_p, and the standard deviation of pairwise diversities). The results were similar or worse to the ones reported here.

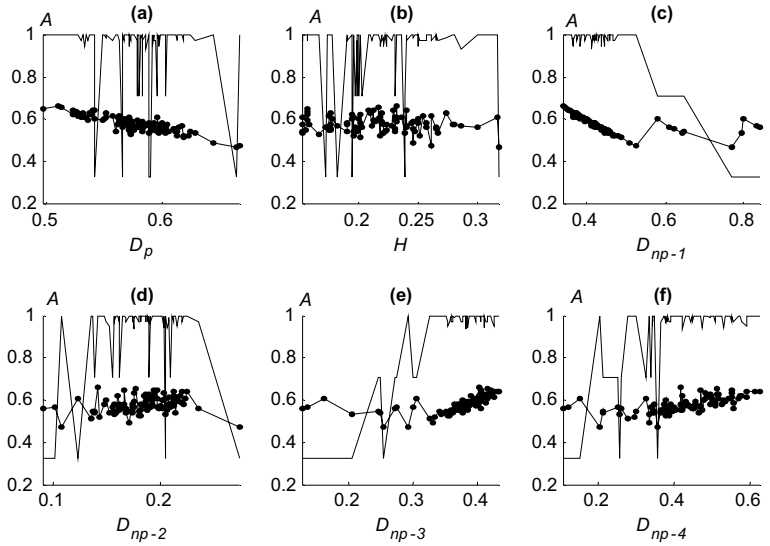

Fig. 2. Ensemble accuracy, A, versus 6 diversity measures for the four-gauss data. The bottom curve in each plot is the averaged individual accuracy (dot marker).

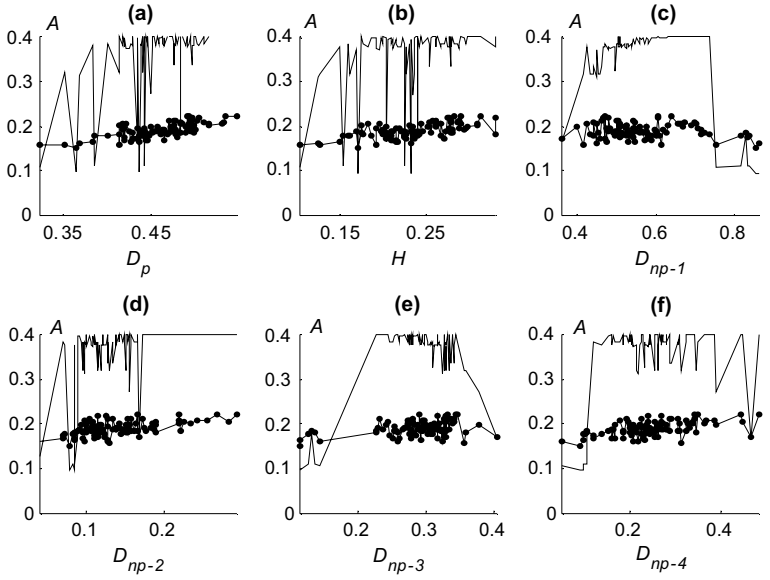

Fig. 3. Ensemble accuracy, A, versus 6 diversity measures for the wine data. The bottom curve in each plot is the averaged individual accuracy (dot marker).

To ease the interpretation of Figs. 2 and 3, the correlation between the 6 diversity measures and the ensemble accuracy has been calculated and displayed in Table 1. Since the correlation coefficient is a measure of linear dependency, the "cup" pattern in Fig. 3(c) and (e) will not stand out.

Fig. 4 shows the two subplots (e) from Figs. 2 and 3 and a fitted polynomial curve of degree 3 (the thick line) of the ensemble accuracy as a function of D_{np-3}. The monotonic and the cup pattern can be seen from the interpolation. The problem is that it is not known in advance whether our data will exhibit one or the other. The two patterns suggest that a compromise can be sought in the medium value of the diversity. Therefore we suggest the following simple procedure for building cluster ensembles:

1. Generate K ensembles varying the random parameter(s) of the clustering algorithm.
2. Calculate diversity using a chosen diversity measure (we prefer D_{np-3} or D_{np-1} for reasons stated above and the explanations later in the experimental section).
3. Find the median of the diversity values and pick the corresponding ensemble.

Table 1
Correlation coefficients between the 6 diversity measures and the ensemble accuracy for the examples in Figs. 2 and 3

Data set	D_p	H	D_{np-1}	D_{np-2}	D_{np-3}	D_{np-4}	Individual average
Four-gauss	−0.1313	−0.0329	−0.9263	0.2356	0.8701	0.6171	0.2245
Wine	0.4456	0.3378	−0.3796	0.4006	0.5394	0.3953	0.4095

Shown also is the correlation between the individual average and the ensemble accuracy.

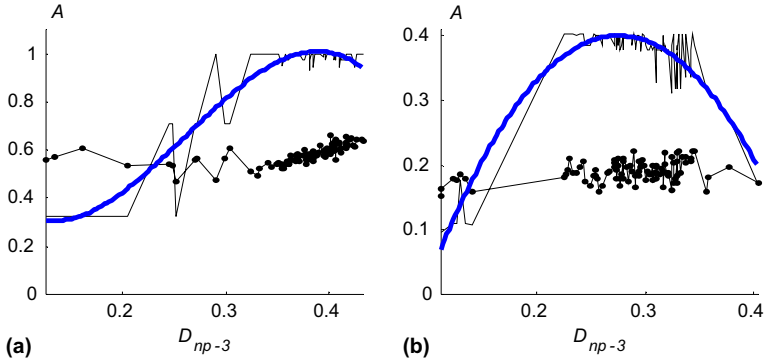

Fig. 4. Fitted polynomial of degree 3 for the ensemble accuracy versus D_{np-3} for the four-gauss and wine data. The bottom curve in each plot is the averaged individual accuracy (dot marker).

Our hypothesis is that ensembles selected through median diversity will fare better than randomly selected ensembles or ensembles selected through maximum diversity.

4. Experiments

Seven types of homogeneous ensembles were constructed as summarized in Table 2. Two most common types of clusterers were used: the k-means and the mean link method (average link, average linkage). All ensemble consisted of $L = \mathbf{25}$ clusterers. The parameters that we varied were:

- the number of overproduced clusters, c. The value was either fixed at $c = 20$ (ensembles 1 and 5) or chosen randomly for each ensemble member in the range from 2 to 22;
- the initialization of k-means for ensembles 1, 2, 3 and 4;
- the sample submitted for clustering to each ensemble member. In ensemble models 1, 2 and 4, the whole data set $\mathbf{Z}$ was submitted to each clusterer. In the remaining ensembles, a random sub-sample of $\mathbf{Z}$ was submitted to each clusterer with size between $N/2$ and N;
- the noise injection. For building ensembles 4 and 7 we altered the data for each ensemble member by adding a Gaussian noise with mean 0 and standard deviation 0.1.

All ensemble types were applied to 5 data sets, summarized in Table 3. For each ensemble type and each data set, 100 ensembles were generated in order to measure diversity and its relationship with the ensemble accuracy. For the three artificial data sets, each ensemble was built on a different data set generated from the respective distribution. Fig. 5 shows an example of three such data sets.

All three sets were generated in 2-D (as plotted) and then 10 more dimensions of uniform random noise were appended to each data set. A total of 100 points were generated from each distribution. The noise is bound to introduce diversity perhaps both helpful and harmful to the clustering. Usually noise is being artificially injected into data for the purpose of simulating reality, i.e. exactly for the purpose of creating diversity. We felt that the 2-D

Table 2
Summary of the design of the 7 ensembles types

Number	1	2	3	4	5	6	7
Type of the base clusterer	k-Means	k-Means	k-Means	k-Means	Mean link	Mean link	Mean link
Number of overproduced clusters, c	20	Random	Random	Random	20	Random	Random
Sample size for the base clusterer	Whole	Whole	Random	Whole	Random	Random	Random
Noise added	No	No	No	Yes	No	No	Yes

Table 3
Data sets

Name	Type	Number of objects, N	Number of features	Number of classes (supposed clusters)
Four-gauss	Artificial	100	12	4
Easy-doughnut	Artificial	100	12	2
Difficult-doughnut	Artificial	100	12	2
Glass	Real (UCI)	214	9	6
Wine	Real (UCI)	178	13	3

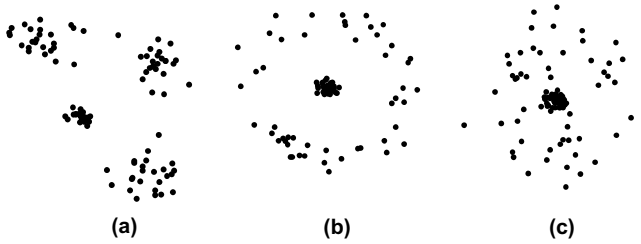

Fig. 5. Artificial data sets: (a) four-gauss; (b) easy-doughnut; (c) difficult-doughnut.

problems, on their own, will be too easy for a single algorithm and there will be no need for an ensemble at all. For example, a popular choice of a data set elsewhere is the 2-spirals data. The two spirals are perfectly distinguishable by the single linkage, so the benefit from using an ensemble (and studying it) becomes unclear.

The two real data sets from UCI have often been picked for evaluating cluster ensembles, e.g., in [17,31], because they are relatively small, features are continuous-valued and there are no missing values. Note that the correspondence between the known labels and the labels obtained by clustering is not necessarily a good measure of the quality of the clustering method because the class labels may not correspond to natural groups in the data. Nevertheless, experiments with real-life (labeled) data have been reported in most studies on clustering, so here we follow this tradition.

The 6 diversity measures were calculated for the 7 ensemble types (100 ensembles of each type) for each of the 5 data sets. Table 4 shows the 35 ensemble accuracies averaged across 100 realizations. The ensemble model used for Fig. 2 was 2 and for Fig. 3, 6. These two models were chosen for the illustration because they had the best accuracies for the respective data sets.

To find out whether the ensemble selection method works, the following experiment was carried out. For each data set and for each ensemble model, out of the 100 ensembles, 15 were randomly selected and the median and the maximum diversity were found within the selection. The corresponding ensembles were identified and their accuracies were stored. This procedure was repeated 100 times for each combination of ensemble model and data set. An example of the format of the obtained results is shown in Table 5. The accuracies are the averages over the 100 runs. Table 5 is based entirely on ensemble 2, therefore it is interesting to compare the results there with column 2 in Table 4. The selection procedures using each of the non-pairwise measures, D_{np-1} to D_{np-4} improves on the previous values of the ensemble accuracy for all data sets. Both D_p and H improve on the ensemble accuracies for the two "easier" data sets, four-gauss and easy-doughnut. However, for the other three data sets, the selection procedure using these diversity measures is not very successful.

Table 4
Ensemble accuracies (ar(P^*, P^T)) for the 7 ensemble models and the 5 data sets, averaged across 100 realizations

Data set/ensemble model	1	2	3	4	5	6	7
Four-gauss	0.8574	**0.9410**	0.8997	0.8813	0.4604	0.7304	0.6695
Easy-doughnut	0.8660	0.8643	0.8285	0.8288	**0.9460**	0.7749	0.5465
Difficult-doughnut	0.2631	0.5344	0.3906	0.5041	**0.6514**	0.3551	0.2076
Glass	0.1885	0.1843	0.1536	0.1329	**0.2516**	0.1767	0.1824
Wine	0.1892	0.2374	0.2721	0.2119	0.1179	**0.3623**	0.3535

The largest value for each data set is shown in boldface.

Table 5
Ensemble accuracies (ar(P_{med}, P^T)) for ensemble model 2 and the 5 data sets, averaged across 100 runs (samples of 15 ensembles and selection)

Data set/measure	D_p	H	D_{np-1}	D_{np-2}	D_{np-3}	D_{np-4}
Four-gauss	0.9427	0.9485	0.9872	0.9730	**0.9882**	0.9877
Easy-doughnut	0.8983	0.8897	**0.9363**	0.8937	0.9143	0.8709
Difficult-doughnut	0.4769	0.4987	**0.7856**	0.6627	0.7303	0.7719
Glass	0.1752	0.1857	0.2150	0.2113	**0.2176**	0.2102
Wine	0.2432	0.2201	0.2786	0.2717	**0.2924**	0.2505

The largest value for each data set is shown in boldface.

Since displaying all the results is not feasible, statistical comparisons between the obtained accuracies were carried out. Each of the 6 diversity measures has 2 attached results, one for the median selection and one for the maximum selection.

Following the proposed procedure, the ensemble can be selected using any of the 6 measures and either the median or the maximum selection strategy. To be able to refer to these selection choices, they will be called "competitors". The question is which of the 12 competitors will give the best ensemble. The original ensemble accuracy (Table 4) can be considered as the base design and added as a competitor as well. If an ensemble selection method is successful, then the obtained accuracy will be greater than the base accuracy and the difference will be statistically significant.

The statistical tests were carried out at level of significance 0.05. Let $P_i^{(1)}$ and $P_i^{(2)}$, $i = 1, \ldots, 100$, be the ensemble accuracies for competitors 1 and 2, for a fixed ensemble model and data set. The 100 differences $d_i = P_i^{(1)} - P_i^{(2)}$ were formed, and the mean and the standard deviation of d were calculated. The 95% confidence interval was constructed and checked whether it contained the 0. If yes, then the obtained difference between competitors 1 and 2 was marked as not statistically significant. If the zero was outside the confidence interval, to the left, then competitor 1 was said to be better than competitor 2. If the 0 was outside, to the right, then competitor 2 was said to be better than competitor 1. There are $5 \times 7 = 35$ comparisons between each pair of competitors. Table 6 contains the results from the statistical tests. Entry (i, j) in the table shows the number of comparisons (out of 35) where competitor i has been better than competitor j.

Table 6 shows that D_{np-1} and D_{np-3} are the best indices. While D_{np-3} finds good ensembles for both median values of diversity and large values of diversity, D_{np-1} fails with the maximum-selection method. This once again confirms the counterintuitive behaviour of the index in that its values indicating large diversity correspond to poor ensembles. To get an overall view on the comparative results, we calculate two resultant values for each index, one for each selection method. We sum up the times this index has been better than the other indices (sum of the respective row in Table 6). Fig. 6 depicts the resultant numbers for the 13 indices. To evaluate the robustness of this result with respect to the number of selected ensembles (set to 15 here), Table 7 shows the resultant numbers for the 13 indices for samples of sizes 5, 15 and 25.

The results in Tables 6 and 7 and Fig. 6 show that median selection is better than the maximum selection on all 6 indices. In other words, large diversity is not a recipe for a good cluster ensemble. This was observed not only for the proposed indices D_{np-1} to D_{np-4} but also for the indices D_p and H suggested in [6,13]. From the proposed set we favour D_{np-1} and D_{np-3}. Although D_{np-1} achieved a slightly better resultant number (total number of favourable statistical comparisons), its behaviour is to some extent counterintuitive. The maximum selection method with D_{np-1} shows that ensembles with large values of the index (large diversity) are the worst, so much so, that a random selection of an ensemble would be better. On the other hand, D_{np-3} gives stable results for both selection methods, better for the median selection. Table 7 shows that the pattern of relationship between the indices does not change across different sample sizes of selected ensembles.

Table 6
Statistical significance of the differences between the "competitors"

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	**0**	0	1	2	3	2	1	18	14	35	11	12	28
2	3	**0**	6	2	3	2	2	19	17	35	10	12	29
3	2	0	**0**	2	4	2	2	18	17	35	12	12	28
4	29	26	28	**0**	16	5	14	29	26	35	18	16	31
5	19	15	18	1	**0**	4	7	21	22	35	15	14	29
6	26	22	25	2	11	**0**	15	29	25	35	16	16	30
7	18	15	18	4	8	6	**0**	22	25	35	18	16	30
8	5	5	4	3	4	4	5	**0**	7	35	8	10	27
9	6	4	4	2	3	2	4	9	**0**	34	9	10	28
10	0	0	0	0	0	0	0	0	0	**0**	0	0	5
11	17	15	17	11	12	11	15	20	20	33	**0**	13	27
12	19	19	20	12	17	13	15	21	21	30	11	**0**	27
13	4	1	2	1	2	1	3	2	2	22	1	5	**0**

Entry (i, j) in the table shows the number of comparisons (out of 35) where competitor i has been better than competitor j.

Key:
1 Base accuracy (equivalent to randomly chosen ensemble)
2 D_p, selection by median
3 H, selection by median
4 D_{np-1}, selection by median
5 D_{np-2}, selection by median
6 D_{np-3}, selection by median
7 D_{np-4}, selection by median
8 D_p, selection by maximum
9 H, selection by maximum
10 D_{np-1}, selection by maximum
11 D_{np-2}, selection by maximum
12 D_{np-3}, selection by maximum
13 D_{np-4}, selection by maximum

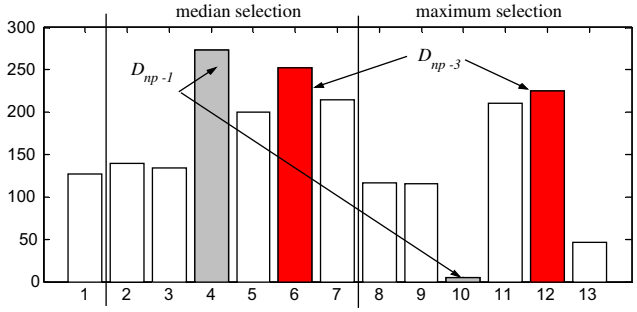

Fig. 6. Total numbers of statistically significant differences in favour of each method. The numbers on the x-axis correspond to these in Table 6.

Table 7
Total number of statistically significant differences in favour of each method for sample sizes 5, 15 and 25

Method	1	2	3	4	5	6	7	8	9	10	11	12	13
Size = 5	91	**111**	**103**	**219**	**165**	**215**	**193**	80	96	13	208	236	43
Size = 15	127	**140**	**134**	**273**	**200**	**252**	**215**	117	115	5	211	225	46
Size = 25	128	**158**	**143**	**264**	**193**	**257**	**223**	128	123	6	226	222	45

The median-selection results are highlighted in bold. Methods are numbered as in Table 6.

5. Relationship between diversity-selection procedure and the ensemble size

Our final set of experiments seeks to find out how the proposed selection methods behave for various ensemble sizes. The following set-up was used:

- Ensemble method 2 was employed as the one with the best performance among the studied ensembles.
- To make the results more easily understandable, the *classification accuracy* is shown as the performance criterion. The classification accuracy is calculated as the proportion of the correctly labeled objects. Each cluster is labeled with the class most represented within. This labeling guarantees the largest possible classification accuracy. For this criterion to work, the number of clusters must be the same as the number of true classes. Otherwise, the trivial partition where each point is a cluster on its own will be the most accurate partition! Therefore the target number of clusters for the combination (consensus) function was set to be equal to the (known) number of classes.
- The consensus function was k-means clustering using the consensus matrix as the input data (as explained in Section 2). Our choice was based on a small pilot set of experiments which showed this consensus function to be superior to the one used before for the current set-up.
- Four additional real data sets were included; a summary is given in Table 8.
- The ensemble size was varied from 5 to 100 with a step of 5.

Table 8
Additional data sets

Name	Type	Number of objects, N	Number of features	Number of classes (supposed clusters)
Iris	Real (UCI)	150	4	3
Segmentation	Real (UCI)	210	19	7
Soybean (small)	Real (UCI)	47	35	4
Contractions	Real[a]	96	34	2

[a] Personal communication from Dr. Fernando Vialriño, Computer Vision Centre, Barcelona, Spain.

- Each point on the "select-and-choose" curves is an average of 100 selections followed by a choice.

Figs. 7(A)–(B) display the results for the 9 data sets. The x-axis is the ensemble size and the y-axis is the classification accuracy. The four lines shown correspond to single k-means (no ensemble), single ensemble, ensemble selected through median diversity and ensemble selected through maximum diversity, both using D_{np-3}.

There is nothing intriguing in Fig. 7(A). The ensembles' accuracy quickly shoots to 100%, and selection is not relevant at all. The accuracy for the four-gauss data varies slightly but this is only within a fraction of a percent. The high accuracy is due to the fact that here we supply the correct number of clusters to the combiner while in the experiments described in Section 4 the models were disadvantaged by having to guess the number of clusters.

Figs. 7(B) and (C) show that selection by median is better than selection by maximum in 3 occasions (glass, wine and soybean data), by a clear margin in 2 of these cases (glass and wine data). On the other hand, selection by maximum is better than selection by median for iris, segmentation and contractions data sets. Note that the rate of improvement varies almost randomly with the number of clusterers in the ensembles (segmentation and contractions data), which suggests a certain instability and fluctuations related to this selection method. What is more important, selection by median is either similar or no worse than the ensemble itself (no selection), in some cases better. On the other hand, selection by maximum may be substantially worse than the ensemble (glass and wine). Since in real problems, there is no way of knowing which situation we are in, we recommend selection by median as the safer option.

Further experiments indicated that there is a great variability of the results depending upon the choice of the ensemble model and the consensus function. This suggests that it is not only the characteristics of the data set that determine which selection strategy is better. Opposite trends of the accuracy as a function of the ensemble size have been found as well. For the same data set, one ensemble model will improve with increasing the ensemble size while another model will deteriorate.

Why may ensembles not work well for real data? The results reveal several interesting phenomena. The con-

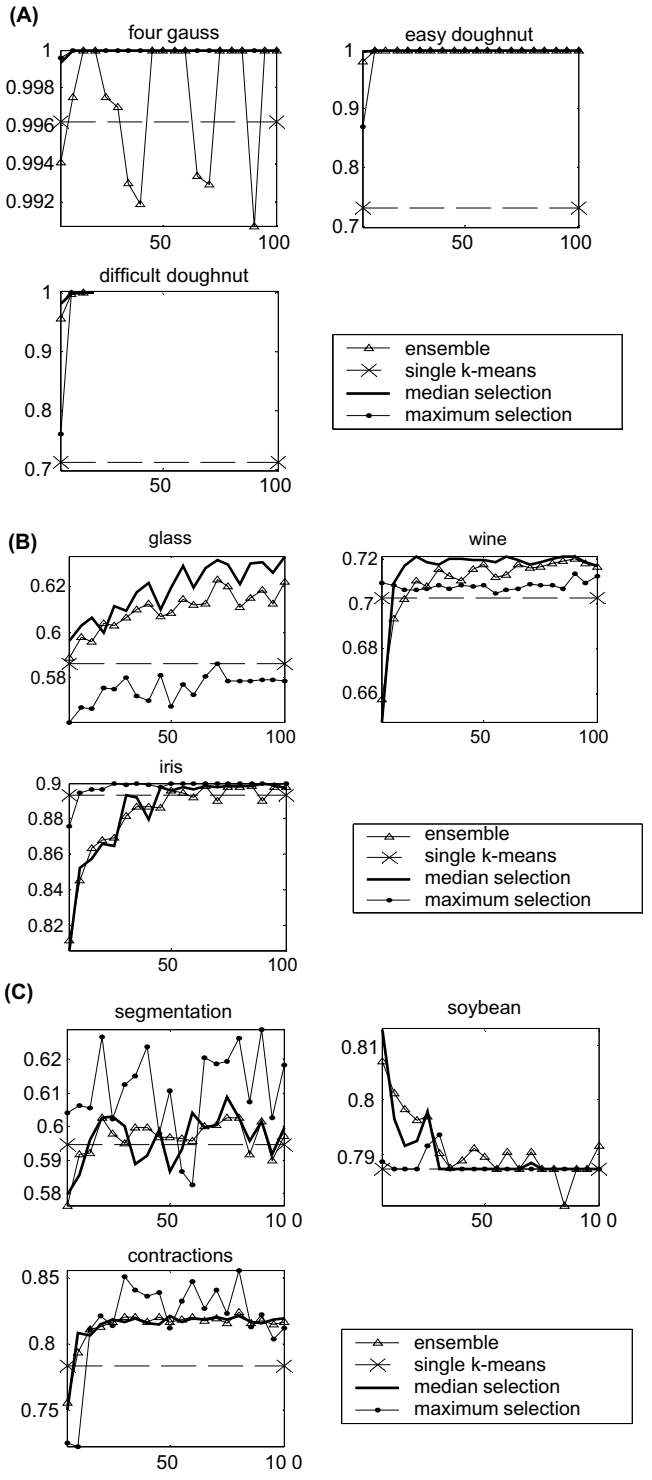

Fig. 7. (A) Classification accuracy versus ensemble size for glass, wine and iris data. (B) Classification accuracy versus ensemble size for glass, wine and iris data. (C) Classification accuracy versus ensemble size for segmentation, soybean and contractions data.

sensus function does have a great effect on the ensemble performance, as found by other authors. It was interesting to see that sometimes the size of the ensemble is actually inversely related to the ensemble accuracy. Moreover, the ensemble itself seems to be less accurate than a single k-mean clustering. The reason for this could be that the class labels do not correspond to natural clusters in data. This scenario is not unlikely to happen. Suppose that your problem is to label pixels in an image into classes "black" and "white". If you omit the labeling, the data set will consist of the coordinates of all the pixels in an image. No clustering procedure can give any meaningful result on this data; the only reasonable answer would be that there are no distinct clusters. If the success is judged by the classification accuracy, our choice of a best clustering procedure out of a pool would be completely random. The data sets in this study were taken from a benchmark repository for *classification*. If the classes corresponded to natural clusters in data, then classification would be easy and the sets would not be in the benchmark suite. Therefore the mismatches between clusters and class labels for some data sets are not surprising.

6. Conclusions

Since diversity in classifier and cluster ensembles is a loosely defined concept, there are many ways to specify and measure it. Four indices are proposed here for estimating diversity in cluster ensembles. They are based on an observation in our previous studies [18] that only an averaged disagreement measure is insufficient. The results in this study support selecting the ensemble with medium diversity from a randomly generated set of ensembles. Two averaged measures of disagreement for cluster ensembles were discussed in the recent literature, D_p in [6] and H in [13]. The difference between these measures and the ones proposed here is that the proposed measures take the ensemble decision as the point of reference and calculate diversity using the averaged deviation from this decision. In the experiments carried out the proposed measures compared favourably to D_p and H.

It was observed that D_{np-1} shows a counterintuitive behaviour in that large diversity leads to very poor ensembles. Based on our previous observations, the *spread* of the diversities was included in the measure. From the three such measures proposed here, D_{np-2} D_{np-3} and D_{np-4} the results favoured D_{np-3} as the one most related to the ensemble accuracy. Two typical patterns of diversity–accuracy relationship were found as shown in Fig. 4. One is almost monotonic—the larger the measure value, the higher the accuracy, while the other is shaped as a parabola with a maximum at about the middle of the diversity range. This led us to the idea of selecting from a set of randomly generated ensembles the one with the median diversity. The results show that

ensembles selected through D_{np-1} or D_{np-3} in this way usually are significantly better than a randomly chosen ensemble or ensembles chosen using D_p or H (either by median or maximum diversity).

If there was a further indication about which of the two relationship patterns is applicable to a given data set or ensemble model, the median-selection strategy could be applied for the parabola pattern and maximum-selection strategy for the monotonic pattern, using D_{np-3} in both cases.

It should be noted that patterns of diversity–accuracy relationship tend to vary from one data set to another (Section 5), and also from one ensemble construction model to another (Section 4). These patterns also vary with respect to the combination method (consensus function), which is not shown in this study but was found in the course of our experiments. Thus we are cautious to generalize from this experiment because of the observed variability. It is important to concentrate effort in the future in finding which combinations of design heuristics, consensus functions and ensemble sizes are appropriate for which types of data.

Acknowledgements

This work was supported by research grant # 15035 under the European Joint Project scheme, Royal Society, UK.

References

[1] H. Ayad, M. Kamel, Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors, in: T. Windeatt, F. Roli, (Eds.), Proc. 4th International Workshop on Multiple Classifier Systems, MCS'03, Guildford, UK, 2003, Lecture Notes in Computer Science, vol. 2709, pp. 166–175.

[2] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley, New York, 2001.

[3] S. Dudoit, J. Fridlyand, Bagging to improve the accuracy of a clustering procedure, Bioinformatics 19 (9) (2003) 1090–1099.

[4] A. Ben-Hur, A. Elisseeff, I. Guyon, A stability based method for discovering structure in clustered data, in: Proc. Pacific Symposium on Biocomputing, 2002, pp. 6–17.

[5] E.B. Fawlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, Journal of the American Statistical Association 78 (383) (1983) 553–584.

[6] X.Z. Fern, C.E. Brodley, Random projection for high dimensional data clustering: a cluster ensemble approach, in: Proc. 20th International Conference on Machine Learning, ICML, Washington, DC, 2003, pp. 186–193.

[7] B. Fischer, J.M. Buhmann, Bagging for path-based clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (11) (2003) 1411–1415.

[8] A. Fred, Finding consistent clusters in data partitions, in: F. Roli, J. Kittler (Eds.), Proc. 2nd International Workshop on Multiple Classifier Systems, MCS'01, Cambridge, UK, Lecture Notes in Computer Science, vol. 2096, Springer-Verlag, 2001, pp. 309–318.

[9] A. Fred, A.K. Jain, Data clustering using evidence accumulation, in: Proc. 16th International Conference on Pattern Recognition, ICPR, Canada, 2002, pp. 276–280.

[10] A.L.N. Fred, A.K. Jain, Robust data clustering, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, USA, 2003, vol. II, pp. 128–136.

[11] V. Di Gesu, Integrated fuzzy clustering, Fuzzy Sets and Systems 68 (1994) 293–308.

[12] J. Ghosh, Multiclassifier systems: back to the future, in: F. Roli, J. Kittler (Eds.), Proc. 3d International Workshop on Multiple Classifier Systems, MCS'02, Cagliari, Italy, Lecture Notes in Computer Science, vol. 2364, Springer-Verlag, 2002, pp. 1–15.

[13] D. Greene, A. Tsymbal, N. Bolshakova, P. Cunningham, Ensemble clustering in medical diagnostics, in: R. Long et al. (Eds.), Proc. 17th IEEE Symp. on Computer-Based Medical Systems CBMS'2004, Bethesda, MD, National Library of Medicine/National Institutes of Health, IEEE CS Press, 2004, pp. 576–581.

[14] K. Hornik, Clustrer ensembles. http://www.imbe.med.uni-erlangen.de/links/EnsembleWS/talks/Hornik.pdf.

[15] X. Hu, I. Yoo, Cluster ensemble and its applications in gene expression analysis, in: Y.-P.P. Chen (Ed.), Proc. 2-nd Asia-Pacific Bioinformatics Conference (APB2004), Dunedin, New Zealand, 2004, pp. 297–302.

[16] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.

[17] A.K. Jain, A. Topchy, M.C.H. Law, J.M. Buhmann, Landscape of clustering algorithms, in: Proceedings of ICPR, Cambridge, UK, 2004, pp. 260–263.

[18] L.I. Kuncheva, S.T. Hadjitodorov, Using diversity in cluster ensembles, in: Proceedings of IEEE Int. Conf. on Systems, Man and Cybernetics, The Hague, The Netherlands, 2004, pp. 1214–1219.

[19] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles, Machine Learning 51 (2003) 181–207.

[20] L.I. Kuncheva, Diversity in multiple classifier systems (Editorial), Information Fusion 6 (1) (2005) 3–4.

[21] L.I. Kuncheva, That elusive diversity in classifier ensembles, in: Proc. IbPRIA 2003, Mallorca, Spain, Lecture Notes in Computer Science, vol. 2652, Springer-Verlag, 2003, pp. 1126–1138.

[22] B. Minaei, A. Topchy, W. Punch, Ensembles of partitions via data resampling, in: Proceedings of the International Conference on Information Technology on: Coding and Computing, ITCC'04, Las Vegas, NV, 2004, vol. 2, pp. 188–192.

[23] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: a resampling based method for class discovery and visualization of gene expression microarray data, Machine Learning 52 (2003) 91–118.

[24] E. Pekalska, R.P.W. Duin, Automatic pattern recognition by similarity representations, Electronics Letters 37 (3) (2001) 159–160.

[25] E. Pekalska, R.P.W. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognition Letters 23 (8) (2002) 943–956.

[26] W.M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 (1971) 846–850.

[27] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research 3 (2002) 583–618.

[28] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining partitionings, in: Proc. of 11-th National Conf. on Artificial Intelligence, NCAI, Edmonton, Alberta, Canada, 2002, pp. 93–98.

[29] A. Topchy, A.K. Jain, W. Punch, Combining multiple weak clusterings, in: Proceedings of IEEE Int. Conf. on Data Mining, Melbourne, Australia, 2003, pp. 331–338.
[30] A. Topchy, A.K. Jain, W. Punch, A mixture model for clustering ensembles, in: Proceedings of SIAM Conference on Data Mining, 2004, pp. 379–390.
[31] A. Topchy, B. Minaei, A.K. Jain, W. Punch, Adaptive clustering ensembles, in: Proceedings of ICPR, 2004, Cambridge, UK, 2004, pp. 272–275.
[32] A. Weingessel, E. Dimitriadou, K. Hornik, An ensemble method for clustering, Working paper, 2003. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/.