Diversifying Heuristics for Cluster Ensembles

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

[a] *School of Informatics, University of Wales, Bangor, LL57 1UT, UK*

[b] *CLBME, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria*

Abstract

Cluster ensembles are deemed to be better than single clustering algorithms for discovering complex or noisy structures in data. We consider different heuristics to introduce diversity in cluster ensembles and study their individual and combined effect on the ensemble accuracy. Our experiments with three artificial and three real data sets, and 12 ensemble types, showed that the most successful diversifying heuristic was the random choice of the number of clusters for each ensemble member.

Key words: Pattern recognition, multiple classifier systems, cluster ensembles, diversity

1 Introduction

Selecting a good clustering algorithm is more difficult than selecting a good classifier. The difficulty comes from the fact that in clustering there is no supervision, i.e., data have no labels against which to match the partition obtained through the clustering algorithm. Therefore, instead of running the risk of picking an unsuitable clustering algorithm, a cluster ensemble can be used [24]. The presumption is that even a basic off-the-shelf cluster ensemble will outperform a randomly chosen clustering algorithm. The question then becomes whether we can guide the selection of a cluster ensemble.

The interest in cluster ensembles has been growing in the past few years [1,5, 7–11,13,19,24,25]. The aim of combining several partitions into a single one is to improve the quality and robustness of the result.

* Corresponding author: Telephone: +44 1248 383661; Fax: + 44 1248 361429
Email address: `l.i.kuncheva@bangor.ac.uk` (Ludmila I. Kuncheva).

It is generally agreed that diverse *classifier* ensembles fare better than non-diverse ensembles but it is also accepted now that the relationship between diversity and accuracy at a close-up is not straightforward [17]. Thus the known relationship is too coarse to be useful for the ensemble design and can only be perceived as a guide statement. Here we are interested in *cluster ensembles.* We investigate the effect of various design heuristics on the ensemble diversity and accuracy. Fern and Brodley [5] give an example showing that more diverse ensembles offer larger improvement on the individual accuracy [1] than less diverse ensembles. Greene et al. [12] conclude that diversity among the ensemble members is necessary but not sufficient for a good result, and the strategy for combining the partitions plays an important role as well. We draw upon these studies and try to quantify diversity in cluster ensembles in two ways: diversity by design and obtained diversity. Then we look into the relationship between these diversities and the ensemble accuracy.

We are interested in the following questions

(1) How do the ensembles based on different heuristics compare to one another? Which is the best heuristic or combination of heuristics?
(2) Is ensemble diversity related to the ensemble accuracy?
(3) Is the level of diversity-by-design matched by the obtained diversity of the ensemble?

The rest of the paper is organized as follows. Section 2 suggests heuristics for building diverse cluster ensembles, explains the main ensemble algorithm and explains the measure of similarity/diversity between partitions used in this study. The choice of data sets and the experimental set-up are detailed in Section 3. Section 4 contains the results and a discussion thereupon and Section 5 concludes the study.

2 Diversity in cluster ensembles

2.1 Heuristics for building 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, or a 'clusterer'. The aim is to find a resultant partition P^* which best represents the structure of $\mathbf{Z}$.

Cluster ensembles can be built in different ways, among which

[1] By "accuracy" of a clustering algorithm we assume the degree of match between some known cluster labels and the labels produced by the algorithm.

- Use **different subsets of features** (overlapping or disjoint), called feature-distributed clustering in [12, 24].
- Use **different clustering algorithms** within the ensemble [14]. We shall call such ensembles heterogeneous or hybrid.
- **Randomize the clustering algorithm.** Some clustering algorithms rely on random choices. For example, k-means can be started from different initializations which may lead to different partitions of the same data. The classical hierarchical algorithms (single link, mean link, complete link, etc.) are deterministic and will need external randomization.
- **"Weaken" the clustering algorithm.** Clustering methods are judged by their stability, i.e. methods which are not too sensitive to data perturbations are perceived to be better. In cluster ensembles it is important to have diversity, so weaker clusterers have been considered. For example, we can run just one iteration of k-means by initializing the cluster centroids randomly and assigning the data points to these centroids [12]. Randomly projecting the data on a low-dimensional space and running k-means in it is another possibility being explored [5, 26].
- Use **different a data set** for each ensemble member, e.g. resampling with or without replacement [3, 6, 9, 18, 19].

Any combination of the above can be used as well.

We can construct the resultant partition following several approaches (called "consensus functions"): direct (re-labeling) [6, 24, 29], feature-based approach [27], hypergraph approach [24] and the *pairwise approach* [1,5,7–9,19]. We implemented the pairwise approach because it has been a popular choice despite its comaprativaly large computational complexity. The general version of the pairwise cluster ensemble algorithm is outlined below.

(1) Given is a data set $\mathbf{Z}$ with N elements. Pick the ensemble size L and the number of clusters c. Usually c is larger than the suspected number of clusters so there is "overproduction" of clusters.

(2) Generate L partitions of $\mathbf{Z}$ with c clusters in each partition.

(3) Form a co-association matrix for each partition, $M^{(k)} = \left\{m_{ij}^{(k)}\right\}$, of size $N \times N$, $k = 1, \ldots, L$, where

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

(4) Form a final co-association matrix $\mathbf{M}$ (consensus matrix) from $M^{(k)}$, $k = 1, \ldots, L$, 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 on $\mathbf{Z}$. Therefore, it can be used with any clustering algorithm which operates directly upon a similarity matrix. Viewed in this context, cluster ensemble is a type of *stacked clustering* whereby we can generate layers of similarity matrices and apply clustering algorithms on them. In our preliminary studies we found that the results were slightly better if we used $\mathbf{M}$ as a new feature space and ran the single link clustering on that. This idea is not novel in the pattern recognition and machine learning communities; treating similarities between objects as the new feature space has been studied recently by Pękalska et al. [20–22]

2.2 Diversity/Similarity between partitions

Various measures of similarity between two partitions have been proposed in the literature. In our preliminary experiments we considered six indices: Rand index [23], Jaccard index [4], adjusted Rand index [15], Correlation index [4], the mutual information index used in [9, 24, 25] and the entropy [12]. Based on the results, we chose the adjusted Rand index to measure both diversity between clusterers and the ensemble accuracy.

Consider partitions A and B and their confusion matrix, where the rows correspond to the clusters in A and the columns correspond to the clusters in B. Denote by N_{ij} the (i, j)th entry in this confusion matrix, where N_{ij} is the number of objects in both cluster i of partition A and cluster j in partition B. Denote by $N_{i.}$ the sum of all columns for row i; thus $N_{i.}$ is the number of objects in cluster i of partition A. Define $N_{.j}$ to be the sum of all rows for column i, i.e. $N_{.j}$ is the number of objects in cluster j in partition B. Suppose that the two partitions A and B are drawn randomly with a fixed number of clusters and a fixed number of objects in each cluster (generalized hypergeometric distribution). There is no requirement that the number of clusters in A and B should be the same. Let c_A be the number of clusters in A and c_B be the number of clusters in B. The expected value of the adjusted Rand index for this case is zero. The adjusted Rand index, ar, is calculated from the values N_{ij} of the confusion matrix for the two partitions as follows

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

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

where $\binom{a}{b}$ is the binomial coefficient $\frac{a!}{b!(a-b)!}$.

four-gauss easy-doughnut difficult-doughnut

Fig. 1. The three artificial data sets

We will use ar in two roles. First, the accuracy of an ensemble will be measured as $ar(P^*, P^t)$, where P^* is the ensemble decision and P^t is the partition defined by the true cluster labels. Second, ar measures the similarity between two partitions, $1 - (ar(P_i, P^*)$ will be used to measure the difference (diversity) between an individual partition P_i and the ensemble output. Our previous experiments led us to a measure for the ensemble quality defined as

$$D(P_1, \ldots, P_L) = \frac{1}{2}\left(\text{mean}(ar(P_i, P*) + \text{std}(ar(P_i, P*))\right) \tag{3}$$

Here 'std' denotes the standard deviation. We note that D may not be a proper *diversity* measure because it includes $ar(P_i, P*)$ with a positive sign, i.e., the closer the partition to the ensemble decision, the higher D is. This counterintuitive choice was dictated by the dependency that we found between the "proper" diversity ($1 -$ mean $ar(.,.)$) and the ensemble accuracy. This study was a part of a larger project where we examined various diversity measures, pairwise and non-pairwise, two of which were taken from the recent literature [5, 12]. It truned out that if we combined the *similiratiy* to the ensemble decision and the *scatter* about the mean similarity, as in (3), we could see a pattern of relationship. Although we admit that this relationship is not strong, the other measures studied showed even weaker relationship.

3 The experiment

3.1 Data sets

Figure 1 shows three artificial data sets called four-gauss, easy-doughnut and difficult-doughnut, respectively. 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.

Three real data sets from UCI Machine Learning Repository Database [2] were also chosen for this experiment: glass (9 features, 6 classes, 214 cases), iris (4 features, 3 classes, 150 cases) and wine (13 features, 3 classes, 178 cases). These data sets have often been picked for evaluating cluster ensembles, e.g.,

[2] http://www.ics.uci.edu/∼mlearn/MLRepository.html

in [16, 28], because they are relatively small, features are continuous-valued and there are no missing values. We 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.

3.2 Ensemble construction and diversity by design

Two clustering procedures were chosen for the experiments: the classical k-means and the mean link (average link). k-means has been the most popular choice for the base algorithm in cluster ensembles. Apart from being an intuitive choice, k-means has been shown to be a kind of a "center-stage" algorithm on the landscape of a large spectrum of clustering algorithms [16]. On the other hand, mean link has been found to be similar by performance to single link in that both algorithms are sensitive to outliers. This instability is not necessarily a drawback because it may be a basis for diversity when each ensemble member is given a subsample of the data to cluster. Indeed, our experiments show that mean link creates more diverse ensembles than k-means does, which, curiously, are either much better or much worse than the k-means ensembles.

To have a base for comparison, we ran the two chosen clustering procedures as single clusterers using the following protocol.

- *k-means.* The number of clusters, c, was varied from 2 to 10. For each c we ran the clustering algorithm from 10 different initializations and chose the labeling with the minimum sum-of-squares criterion, J_e [2]. To determine the final number of clusters we took the minimum of the Xie-Beni index, $u_{XB}(c)$, across $c = 2, \dots, 10$

$$u_{XB}(c) = \frac{\sum_{j=1}^{c} \sum_{\mathbf{z} \in C_j} ||\mathbf{z} - \mathbf{v}_j||^2}{N(\min_{j \neq l} ||\mathbf{v}_j - \mathbf{v}_l||^2)} \tag{4}$$

  where $\mathbf{v}_j$ is the centroid of cluster C_j, $j = 1, \dots, c$, and $\mathbf{z} \in \mathbf{Z}$.
- *mean link.* We built all the partitions from N down to 1 cluster. The largest "jump" in the distance at which two clusters were merged was found and this determined the final number of clusters. If this number appeared to be too large, we conjectured that there is no reasonable structure in the data and reassigned the final number of clusters to 1. We used a threshold of 80% of the total size of the data set, N. If the obtained number of clusters was greater than the threshold, we abstained from identifying a structure and labeled all the points as cluster 1.

The ensembles were built according to the pairwise cluster ensemble procedure in Section 2. When k-means was used as the base clusterer, we started each ensemble member from a different initialization. Since mean link is a deterministic algorithm and does not depend on an initialization, we used a random subsample from the data set for each ensemble member (sampling without replacement). The length of the subsample was chosen randomly between $\frac{N}{2}$ and N.

The ensemble output was derived from the consensus matrix, $\mathbf{M}$, by running a single link on it. The final number of clusters was decided based on the largest jump of the distance criterion as in the mean like procedure explained above.

All our ensembles consisted of $L = 25$ members.

With the base clustering techniques and the combination method fixed, we tried the following diversifying heuristics

(1) Random number of overproduced clusters, c, for each ensemble member. We varied c in the interval $2 \leq c \leq 22$.
(2) Random subsample (Since this is the basic randomization heuristic for the mean link ensembles, random sampling without replacement was applied as an extra-heuristic to k-means ensembles only.)
(3) Noise injected in the data. For all data sets we used Gaussian noise with mean 0 and standard deviation 0.1.
(4) Hybridization, whereby 13 k-means clusterers and 12 mean link clusterers were pooled to make the ensemble.

The two single methods and the twelve ensembles in our experiments are summarized in Table 1.

We can form sequences of cluster ensembles with progressively larger diversity-by-design. For example, the following is a chain of ensemble types with increasing number of heuristics

$$\boxed{\;2\;} \rightarrow \boxed{\;3\;} \rightarrow \boxed{\;4\;} \rightarrow \boxed{\;11\;} \rightarrow \boxed{\;12\;} \rightarrow \boxed{\;13\;} \tag{5}$$

4 Results and discussion

Here we offer answers to the questions put up in the Introduction.

Table 1
Types of cluster ensembles

Ensemble number	Structure	Clustering method	Number of clusters, c	Sample size, N	Noise
$\boxed{1}$	single	k-means	N/A		
$\boxed{2}$	ensemble	k-means	20	whole	
$\boxed{3}$	ensemble	k-means	random	whole	
$\boxed{4}$	ensemble	k-means	random	random	
$\boxed{5}$	ensemble	k-means	random	whole	Y
$\boxed{6}$	single	mean link	N/A		
$\boxed{7}$	ensemble	mean link	20	random	
$\boxed{8}$	ensemble	mean link	random	random	
$\boxed{9}$	ensemble	mean link	random	random	Y
$\boxed{10}$	ensemble	hybrid	$\boxed{2}+\boxed{7}$		
$\boxed{11}$	ensemble	hybrid	$\boxed{3}+\boxed{7}$		
$\boxed{12}$	ensemble	hybrid	$\boxed{3}+\boxed{8}$		
$\boxed{13}$	ensemble	hybrid	$\boxed{4}+\boxed{8}$		
$\boxed{14}$	ensemble	hybrid	$\boxed{5}+\boxed{9}$		

4.1 How do the ensembles compare to one another?

We compare the (assumed to be) true labels with the labels obtained through the 14 clustering methods (Table 1). The *ar* index was calculated for the 14 methods and for the 6 data sets. Each value was averaged across 100 runs of the respective ensemble method. Table 2 shows the best and the worst ensembles for the 6 data sets, as well as the averaged number of clusters detected by the respective ensemble (or single clusterer). The number in the brackets next to the data set name is the (assumed) true number of clusters. The values of *ar* for the best ensembles are highlighted in boldface.

The first curious result from Table 2 is the behaviour of ensemble $\boxed{7}$ (mean link ensemble where base classifiers are built on random subsets of the data). While for three of the data sets $\boxed{7}$ was found to be the best ensemble (easy-doughnut, difficult-doughnut and glass) for the other three data sets (four-gauss, iris and wine) $\boxed{7}$ was the worst ensemble.

Table 2
The best and the worst ensembles for the 6 data sets

Data set	Best ensemble			Worst ensemble		
	ensemble	index	clusters	ensemble	index	clusters
four-gauss (4)	$\boxed{3}$	**0.9410**	3.90	$\boxed{7}$	0.4604	6.27
easy-doughnut (2)	$\boxed{7}$	**0.9460**	2.83	$\boxed{9}$	0.5465	8.04
difficult-doughnut(2)	$\boxed{7}$	**0.6514**	9.27	$\boxed{9}$	0.2521	10.19
glass (6)	$\boxed{7}$	**0.2516**	2.87	$\boxed{5}$	0.1329	3.63
iris (3)	$\boxed{11}$	**0.5755**	2.24	$\boxed{7}$	0.0947	2.94
wine (3)	$\boxed{1}$	**0.3693**	2.00	$\boxed{7}$	0.1179	8.40

Table 2 gives also an early inkling about the controversial role of diversity. The most diverse ensembles according to diversity-by-design are $\boxed{4}$, $\boxed{5}$, $\boxed{9}$, $\boxed{12}$, $\boxed{13}$ and $\boxed{14}$. None of these appeared to be among the best ensembles for any of the data sets. On the other hand, "diverse" ensembles $\boxed{9}$ and $\boxed{5}$ figure among the worst ensembles.

Next we calculated the *ranks* for the 14 clustering methods. We sorted the values of *ar* for each data set and assigned a rank of 14 to the best method and rank 1 to the worst method. Thus if one method was the best across all 6 data sets, it would get a total rank of $14 \times 6 = 84$. If there was a single worst method, it would get a rank of $1 \times 6 = 6$. Table 3 shows the 14 clustering methods sorted by their total rank. The highest and the lowest ranks are underlined, and the highest rank is shown in boldface.

Table 3
Ranks for the 14 clustering methods. The higher the rank, the better the method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
R	43	37	57.5	47	41	21.5	47	36	33	54	**68**	57	46	42

The performance results favor method $\boxed{11}$, a hybrid ensemble based on 13 k-means clusterers (started from different initializations and with random number of overproduced clusters between 2 and 22) and 12 mean-link clusterers (run on random subsamples of the data set of size between half and total size of the data). Ensembles based on k-means are generally better than ensembles based on mean link. We found that $\boxed{7}$ is a non-stable method which may give excellent or disastrous results depending on the data set.

The pattern that starts showing through is that there are "lucky" heuristics and others, which may induce diversity that does not materialize as an improvement on the performance. The best heuristic appears to be randomization of the number of overproduced clusters in the k-means ensembles. Hybridization of the ensembles also seems to be good in certain cases (ensembles $\boxed{11}$ and $\boxed{12}$).

4.2 Is diversity related to the ensemble accuracy?

Figure 2 shows 6 accuracy-diversity plots, one for each data set. On the x-axis is the obtained diversity D, (3), and on the y-axis is the accuracy of the ensemble measured by $ar(P^*, P^t)$. The ensemble methods are marked with their numbers. The encircled numbers correspond to ensembles based on k-means, the framed ensembles are based on mean link and the ensembles in gray boxes are the hybrid models. Each point on the plot is the average of the 100 runs with the respective model.

The plots demonstrate the large variability of the diversity-accuracy pattern or rather the lack of it.

First, there is no clear message indicating that largest diversity is best. The only data set for which the accuracy-diversity relationship follows the expected pattern is the wine data.

Interestingly, the four-gauss data also follows the pattern but only within the 'model streak'. In other words, there is a "large diversity - better ensembles" tendency separately for the k-means ensembles, for mean link ensembles and for the hybrid ensembles, but not when we pool them together. This means that accuracy-diversity relationship depends upon the way we construct the ensemble and may not have a simple generalization. For example, suppose that we have ensembles A, B and C to choose from for the four-gauss data set, and that $D(A) < D(B) < D(C)$. If we know that the construction method is the same, we may choose C. However, if we know that, say, B uses k-means whereas A and C use mean link, we should prefer B regardless of the diversity. The expected monotone relationship holds also for the mean-link ensembles for the iris data and for the k-means ensembles for the difficult doughnut data. As the results in Figure 2 are averaged across 100 runs, we expected a stronger relationship to appear in support of the general belief that more diverse ensemble fare better.

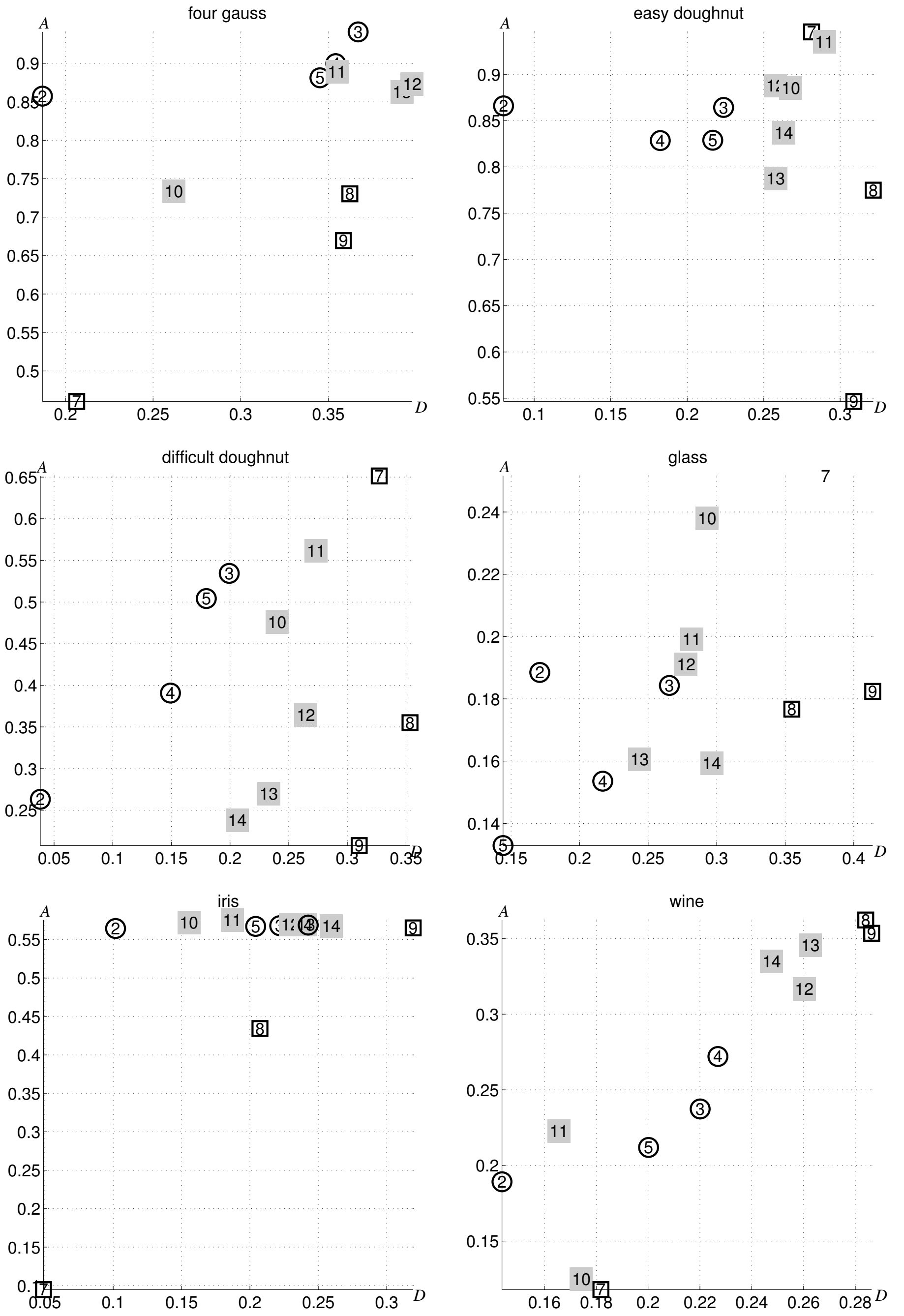

Fig. 2. Accuracy versus diversity. The ensemble methods are marked with their numbers. The encircled ensembles are based on k-means, the framed ensembles are based on mean link and the ensembles in gray boxes are the hybrid models.

4.3 Is the level of diversity-by-design matched by the obtained diversity of the ensemble?

Next we examine the relationship between diversity by design, accuracy and obtained diversity. To do this, we use the same type of plots as in Figure 2.

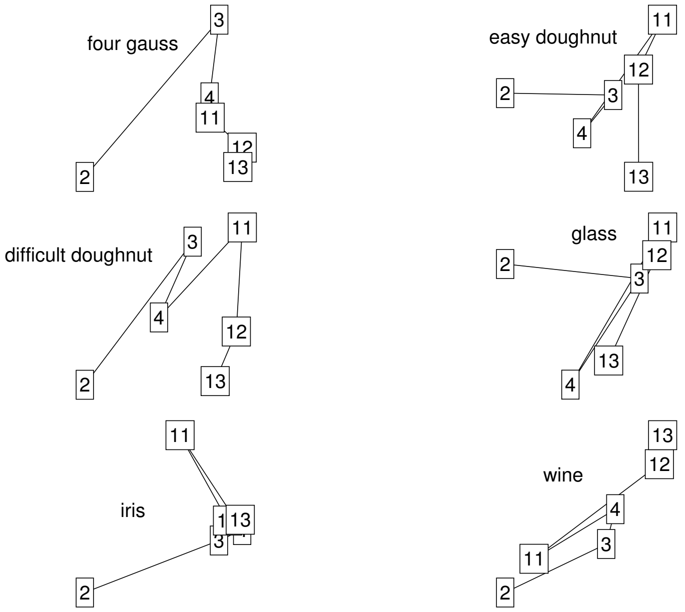

Fig. 3. Accuracy versus diversity-by-design. The ensemble methods are marked with their numbers.

The 6 plots corresponding to the data sets in this study are shown in Figure 3. For a better visual effect, we removed the coordinate axes. Depicted in each subplot are the ensembles in the chain (5). They are linked in increasing order of diversity-by-design, starting from $\boxed{2}$ and finishing with $\boxed{13}$. The x-coordinate is the diversity D, and the y-coordinate is the accuracy of the respective ensemble, exactly as in Figure 2.

If there was a link between diversity-by-design and accuracy, the plots would show all the ensembles aligned on a straight line with positive slope. None of the subplots shows such a pattern. The non-monotonicity of the x-axis indicates that diversity-by-design is not strongly related to the obtained diversity either. Therefore, including more design heuristics does not necessarily lead to more diverse or more accurate ensembles.

5 Conclusions

We studied 14 clustering methods: 2 individual methods and 12 ensemble methods based on them. Our task was to try to evaluate the design heuristics in terms of how diverse and how accurate ensembles they produce. Below we give brief answers to the four questions of this study.

1. How do the ensembles compare to one another? Which is the best heuristic or combination of heuristics?

The best ensemble method was found to be $\boxed{11}$, a hybrid ensemble based on 13 k-means clusterers (started from different initializations and with random number of overproduced clusters between 2 and 22) and 12 mean-link clusterers (run on random subsamples of the data set of size between the half and the total size of the data). Ensembles based on k-means are generally better than ensembles based on mean link. However, the simple mean link ensemble $\boxed{7}$ was found to be either a very good or a very bad choice depending on the data set. Thus mean link ensembles appear to be a bit of a gamble. In real life problems we will not have true labels to match our results against, and will have to rely on other clues to find out whether the data at hand is a good one or a bad one for a mean link ensemble. Ensemble $\boxed{3}$ was the second best in our study, based on only one "lucky heuristic" – the number of overproduced clusters, c, is randomly chosen for each ensemble member. A simple ensemble of type $\boxed{3}$ could be a good practical choice.

2. Is diversity related to the ensemble accuracy?

We found that there is a general pattern such that diverse ensembles tend to fare better but it does not hold for all data sets and all ensemble methods.

3. Is the level of diversity-by-design matched by the obtained diversity of the ensemble?

There is no clear-cut relationship between the two. For example, ensembles which are designed to have larger diversity, e.g., $\boxed{4}$, compared to $\boxed{3}$, actually have smaller values of obtained diversity for 4 out of the 6 data sets ($\boxed{4}$ is to the left of $\boxed{3}$ in the top four plots in Figure 3).

We note that the pairwise cluster ensemble paradigm does not scale well for large data sets. The co-association matrix $\mathbf{M}$ is of size $N \times N$, and running a single-link clustering on it may be too time-consuming. We chose the pairwise method because it has been a popular choice elsewhere. Also, the noise injection was not explored in depth, i.e. with different values of the noise variance. It may turn out that for some specific value of the variance, noise injections becomes one of the "lucky heuristics" too. Using different consensus functions for obtaining the final partition is another option which we have not considered in this study. Finally, our study was limited to $L = 25$ ensemble members and relatively small data sets which have been used as benchmark data sets in cluster ensembles.

Acknowledgment

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

References

[1] H. Ayad and M. Kamel. Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors. In T. Windeatt and F. Roli, editors, *Proc. 4th International Workshop on Multiple Classifier Systems, MCS'03*, volume 2709 of *Lecture Notes in Computer Science*, pages 166–175, Guildford, UK, 2003. Springer-Verlag.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, NY, second edition, 2001.

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

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

[5] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proc. 20th International Conference on Machine Learning, ICML*, pages 186–193, Washington,DC, 2003.

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

[7] A. Fred. Finding consistent clusters in data partitions. In F. Roli and J. Kittler, editors, *Proc. 2nd International Workshop on Multiple Classifier Systems, MCS'01*, volume 2096 of *Lecture Notes in Computer Science*, pages 309–318, Cambridge, UK, 2001. Springer-Verlag.

[8] A. Fred and A.K. Jain. Data clustering using evidence accumulation. In *Proc. 16th International Conference on Pattern Recognition, ICPR*, pages 276–280, Canada, 2002.

[9] A.L.N. Fred and A.K. Jain. Robust data clustering. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, USA, 2003.

[10] V. Di Gesu. Integrated fuzzy clustering. *Fuzzy Sets and Systems*, 68:293–308, 1994.

[11] J. Ghosh. Multiclassifier systems: Back to the future. In F. Roli and J. Kittler, editors, *Proc. 3d International Workshop on Multiple Classifier Systems, MCS'02*, volume 2364 of *Lecture Notes in Computer Science*, pages 1–15, Cagliari, Italy, 2002. Springer-Verlag.

[12] D. Greene, A. Tsymbal, N. Bolshakova, and P. Cunningham. Ensemble clustering in medical diagnostics. Technical Report TCD-CS-2004-12, Department of Computer Science, Trinity College, Dublin, Ireland, 2004.

[13] K. Hornik. Clustrer ensembles.

[14] X. Hu and I. Yoo. Cluster ensemble and its applications in gene expression analysis. In *Proc. 2-nd Asia-Pacific Bioinformatics Conference (APB2004)*, Dunedin, New Zealand, 2004.

[15] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[16] A. K. Jain, A. Topchy, M. C. H. Law, and J. M. Buhmann. Landscape of clustering algorithms. In *Proceedings of ICPR, 2004, Cambridge, UK*, 2004.

[17] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.

[18] B. Minaei, A. Topchy, and W. Punch. Ensembles of partitions via data resampling. In *Proceedings of the International Conference on Information Technology: Coding and Computing, ITCC04*, Las Vegas, 2004.

[19] 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*, 52:91–118, 2003.

[20] E. Pękalska. *Dissimilarity Representations in Pattern Recognition.* PhD thesis, Delft University of Technology, The Netherlands, 2005.

[21] E. Pękalska and R. P. W. Duin. Automatic pattern recognition by similarity representations. *Electronic Letters*, 37(3):159–160, 2001.

[22] E. Pękalska, M. Skurichina, and R. P. W. Duin. Combining Fisher linear discriminant for dissimilarity representations. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 230–239, Cagliari, Italy, 2000. Springer.

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

[24] A. Strehl and J. Ghosh. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–618, 2002.

[25] A. Strehl and J. Ghosh. Cluster ensembles - A knowledge reuse framework for combining partitionings. In *Proceedings of AAAI.* AAAI/MIT Press, 2002.

[26] A. Topchy, A. K. Jain, and W. Punch. Combining multiple weak clusterings. In *Proceedings of IEEE Int Conf on Data Mining*, pages 331–338, Melbourne, 2003.

[27] A. Topchy, A. K. Jain, and W. Punch. A mixture model for clustering ensembles. In *Proceedings of SIAM Conference on Data Mining*, pages 379–390, 2004.

[28] A. Topchy, B. Minaei, A. K. Jain, and W. Punch. Adaptive clustering ensembles. In *Proceedings of ICPR, 2004, Cambridge, UK*, 2004.

[29] A. Weingessel, E. Dimitriadou, and K. Hornik. An ensemble method for clustering, 2003. Working paper, `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/`.