Clustering Criteria in Multiobjective Data Clustering

Julia Handl[1] and Joshua Knowles[2]

[1] Manchester Business School, University of Manchester, UK
julia.handl@mbs.ac.uk
[2] School of Computer Science, University of Manchester, UK
j.knowles@manchester.ac.uk

Abstract. We consider the choice of clustering criteria for use in multiobjective data clustering. We evaluate four different pairs of criteria, three employed in recent evolutionary algorithms for multiobjective clustering, and one from Delattre and Hansen's seminal exact bicriterion method. The criteria pairs are tested here within a single multiobjective evolutionary algorithm and representation scheme to isolate their effects from other considerations. Results on a range of data sets reveal significant performance differences, which can be understood in relation to certain types of challenging cluster structure, and the mathematical form of the criteria. A performance advantage is generally found for those methods that make limited use of cluster centroids and assess partitionings based on aggregate measures of the location of all data points.

1 Introduction

Multiobjective clustering algorithms frame the data clustering problem as a multiobjective optimization problem in which a partitioning is optimized with respect to a number of conflicting criteria. This can be seen as a step beyond traditional clustering techniques, which commonly optimize a single criterion only [11]. It is also a step beyond techniques for internal cluster validation [9], which typically consider combinations of criteria, but usually do so by combining criteria in a linear or non-linear form.

The use of multiple objectives in data clustering has two key advantages. First, the framework of multiobjective optimization provides a natural way of defining a good partitioning: An exact definition of the clustering problem is elusive, but, loosely, a good partitioning can be described as one that meets at least the following two criteria: (i) data points within the same cluster are similar; while (ii) data points within different clusters are dissimilar. Second, single criteria for clustering are biased with respect to the number of clusters (i.e., the criteria naturally increase or decrease for partitionings with a larger number of clusters). One of the consequences of this is that the large majority of single-objective algorithms require the number of clusters to be specified as an input parameter. Multiobjective approaches to data clustering can tackle this issue in a novel way: by selecting two criteria that have opposite biases with respect to the number of clusters, these techniques are able to counter-balance this bias. In principle, multiobjective algorithms are therefore capable of exploring a range of solutions

C.A. Coello Coello et al. (Eds.): PPSN 2012, Part II, LNCS 7492, pp. 32–41, 2012.

with different number of clusters, which can support the user in identifying the most appropriate number of clusters [7].

Previous research on multiobjective clustering [4,7,8] has shown that bicriterion clustering methods often outperform their single-objective counterparts: an algorithm that optimizes two objectives, X and Y, simultaneously, will usually generate certain solutions that are better than the solutions generated by an algorithm that optimizes X or Y only. However, little research (if any) has been done to compare the different choices of (pairs of) criteria in terms of their conceptual aims, or their empirical performance, in bicriterion clustering. In this manuscript, we investigate this issue by comparing four pairs of clustering criteria that have been proposed in previous work on multiobjective clustering. We discuss the conceptual similarities and differences between these choices, and provide empirical results on the use of the criteria in an existing multiobjective evolutionary algorithm for data clustering.

2 Background and Methods

The principle of multiobjective data clustering was first introduced in 1980, when Delattre and Hansen described an exact algorithm for bicriterion clustering [4]. This algorithm was able to identify the set of partitionings corresponding to an optimal trade-off between two objectives, namely the split and the diameter of a partitioning. Given computational resources at the time (as well as the algorithm's reliance on graph colouring [10]), the method was evaluated on small data sets with tens of data items only. More recently, the idea of multiobjective clustering has been extended to a wider set of clustering criteria [1,2,7,8,9,12], which have been optimized using heuristic approaches to multiobjective optimization, principally evolutionary algorithms (EMO).

Here, based on this previous work, four versions of multiobjective clustering were implemented that differed solely in the clustering criteria used. An existing multiobjective clustering algorithm was used as the basis of the implementation, that is the underlying multiobjective evolutionary algorithm (PESA-II, [3]), as well as the encoding, variation operators, initialization and parameter settings are consistent with those described in [8]. The pairs of objectives used within the different versions are as follows.

MOCK [8]: The first method uses the objectives employed in the multiobjective evolutionary clustering algorithm MOCK (Multiobjective clustering with automatic k-determination). The first of these, overall deviation, measures the compactness of the clusters in the partitioning. It is given as:

$$\text{(Min.)} \quad \sum_{c_k \in C} \sum_{i \in c_k} d(i, \mu_k),$$

where C is the given set of clusters, μ_k is the centroid of cluster c_k and $d(,)$ is a distance measure defined between data points. The second objective, connectivity, assesses to what extent data points that are close neighbours are found in the same cluster. It is given as:

$$\text{(Min.)} \quad \sum_{c_k \in C} \sum_{i \in c_k} \sum_{l \in 1..L} \delta(i, l),$$

where L is a parameter specifying the number of neighbours to use (here, the default $L = 20$ is used), and $\delta(i, l)$ is a function which is 0 when data item i and its lth nearest neighbour are in the same cluster and $1/l$ otherwise.

DH [10]: The second method employs the clustering criteria used in Delattre and Hansen's seminal biclustering algorithm. The first objective is the complete link clustering criterion, which minimizes the largest cluster diameter observed in a partitioning. The objective is formally given as

$$(\text{Min.}) \quad \max_{c_k \in C} \max_{i,j \in c_k} d(i,j),$$

where C is the given partitioning of the data. The second objective is the single link clustering criterion, which maximizes the minimum split (distance) between clusters present in a partitioning. This is given as

$$(\text{Max.}) \quad \min_{c_k \in C, c_l \in C, l \neq k} \min_{i \in c_k, j \in c_l} d(i,j).$$

BMM1 [1]: The third pair of objectives is taken from a multiobjective evolutionary algorithm originally designed for fuzzy clustering. For the case of crisp partitioning (considered here), the clustering objectives used simplify to the within-cluster sum of squares, and the minimum distance observed between cluster centroids. Formally, the within-cluster sum of squares is given as

$$(\text{Min.}) \quad \sum_{c_k \in C} \sum_{i \in c_k} d(i, \mu_k)^2,$$

where μ_k is the cluster centroid of cluster c_k. Evidently, this is very similar to the measure of overall deviation defined above with the difference that the distance values are here squared. The minimum distance between cluster centroids is given as

$$(\text{Max.}) \quad \min_{c_k \in C, c_l \in C, l \neq k} d(\mu_k, \mu_l),$$

where μ_k and μ_l are the cluster centroids of cluster c_k and c_l, respectively.

BMM2 [12]: The fourth method also uses the intra-cluster sum of squares (see above) as its first objective. The second objective is the summed pairwise distance between cluster centroids. Formally, this is given as

$$(\text{Max.}) \quad \sum_{c_k \in C, c_l \in C, l \neq k} d(\mu_k, \mu_l),$$

where μ_k and μ_l are the cluster centroids of cluster c_k and c_l, respectively.

3 Conceptual Characteristics

Key similarities and differences between MOCK, DH, BMM1 and BMM2 are summarized in Table 1 and are discussed in this section.

Table 1. Characteristics of the different clustering criteria: (i) Computational complexity associated with evaluating a partitioning of N data points in D dimensions into K clusters; L gives the number of neighbours used in MOCK's connectivity measure; (ii) Resolution of the criteria (the extent to which information about all data points is taken into account); (iii) Use of cluster centroids

	Complexity	Resolution	Centroids
Overall deviation (MOCK)	$\Theta(DN)$	Complete	Yes
Maximum diameter (DH)	$\Theta(N^2)$	Partial	No
Within-cluster sum of squares (BMM1, BMM2)	$\Theta(DN)$	Complete	Yes
Connectivity (MOCK)	$\Theta(LN)$	Complete	No
Minimum split (DH)	$\Theta(N^2)$	Partial	No
Minimum centroid distance (BMM1)	$\Theta(DK^2)$	Partial	Yes
Sum of centroid distances (BMM2)	$\Theta(DK^2)$	Complete	Yes

3.1 Similarities between the Objectives

There are some clear similarities in the way clustering objectives have been combined in the techniques considered. In all four cases, the pair of objectives has been selected to assess both of the key properties of a good partitioning (see Introduction): that (i) data points within the same cluster are similar; while (ii) data points within different clusters are dissimilar.

In MOCK, homogeneity of clusters is assessed using the measure of overall deviation. A similar role is played by the maximum diameter criterion in Delattre and Hansen's method and by the within-cluster sum of squares in Bandyopadhyay et al.'s methods (methods BMM1 and BMM2).

In MOCK, separation between clusters is considered implicitly through the measure of connectivity, which penalizes data points whose nearest neighbours do not reside in the same cluster. In Delattre and Hansen's technique the distance between clusters is assessed using the criterion of minimum split, which identifies the closest pair of data points that are not in the same cluster. Finally, Bandyopadhyay et al. measure cluster distance based on the distance of cluster representatives, either considering the entire set of cluster centres (method BMM2 [12]) or the minimum distance only (method BMM1 [1]).

3.2 Differences between the Objectives

Despite these clear similarities, there are also some fundamental differences between the criteria considered.

One defining characteristic of a clustering criterion is the extent to which its calculation takes into account the cluster assignment of all data points within a data set. This can be most easily understood using the examples of the within-cluster sum of squares and the maximum diameter criterion. The within-cluster sum of squares is calculated as the sum of the distances of all data items to their cluster centre. A change in the cluster assignment of any single data item will therefore usually result in a change to the value of the criterion. In contrast, the maximum diameter of a partitioning is defined as the largest dissimilarity observed between data items that reside in the same cluster. This means that changes in the cluster assignment of individual data items will often have no effect on the value of the criterion, provided that the maximum diameter remains unchanged.

A second defining characteristic is the presence or absence of the concept of cluster centroids in the calculation. Methods that use a cluster centroid make certain implicit assumptions on the shape of the surrounding clusters: it is clear that the definition of a cluster centroid makes relatively little sense for a nonconvex cluster. Out of the objectives discussed, overall deviation, within-cluster sum of squares and the measures of cluster dissimilarity in BMM1 and BMM2 all rely on the definition of a cluster centre. On the other hand, MOCK's connectivity measure, as well as both of Delattre and Hansen's clustering criteria, make no such assumptions on the presence of a centroid and the shape of the underlying cluster.

3.3 Computational Complexity

A further significant difference between the clustering criteria is their computational complexity.

As discussed above, Delattre and Hansen's measure of cluster homogeneity does not make use of a cluster centroid. This comes at the expense of quadratic complexity, as all pairwise dissimilarities between data items need to be considered. In contrast, methods of cluster homogeneity that do utilize a centroid (i.e., overall deviation in MOCK, within-cluster sum of squares in BMM1 and BMM2) have linear complexity.

For measures of cluster separation, the differences in complexity are even more significant. Again, Delattre and Hansen's is the computationally most expensive: the identification of the minimum split requires the pairwise comparison of all data items, resulting in quadratic complexity. MOCK's objective (connectivity) ranks second in complexity: it requires the one-off calculation of N sorted lists of length N (complexity $\Theta(N \times N \log N)$), but has linear complexity for all further evaluations. The objectives in BMM1 and BMM2 have a complexity of only $\Theta(DK^2)$, where K is the number of clusters in the partitioning.

4 Empirical Performance Analysis

4.1 Experimental Setup

For the empirical comparison of the four methods, a benchmark set of Gaussian Clusters in 2, 10 and 100 dimensions was used. This benchmark set has been described previously [8], and the data and results are summarized as supplementary material [6]. In addition, eight two-dimensional data sets available at http://cs.joensuu.fi/sipu/datasets/, were used; these are summarized in Table 2. These feature a variety of challenging cluster properties which we discuss in the next section. The Euclidean distance measure was used for all data sets.

The results returned by each method were evaluated by monitoring the size of the non-dominated set, their quality with respect to the set of eight clustering criteria, as

Table 2. Two-dimensional data sets (also see http://cs.joensuu.fi/sipu/datasets/)

Name	N	D	K	Name	N	D	K
Jain	373	2	2	Compound	399	2	6
Aggregation	788	2	8	Path-based	300	2	3
Flame	244	2	2	r15	600	2	15
Spiral	312	2	3	d31	3100	2	31

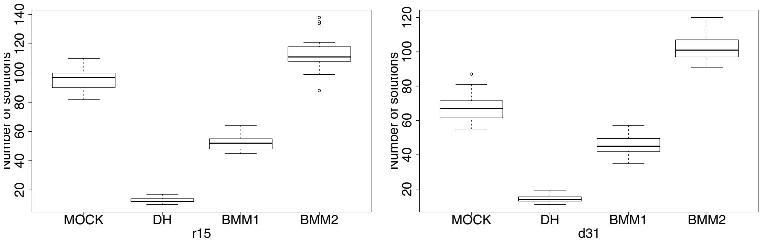

Fig. 1. Size of the solution sets of non-dominated solutions. Representative results over 21 runs for data sets r15 and d31. The data shows a general trend with an ordering of the solutions sets S returned by the methods as $|S_{BMM2}| > |S_{MOCK}| > |S_{BMM1}| > |S_{DH}|$.

well as their accuracy with respect to the known class labels for the data. The latter was assessed using the Adjusted Rand Index (AR, [5]), which is an established external technique of cluster validation that can be used to compare a clustering to a set of known true class labels. It is normalized with respect to the number of clusters in the partitioning, and is therefore well suited for the comparison of partitionings with different numbers of clusters as done in this work [9]. It takes values between 0 and 1, with 1 indicating a partitioning that accurately matches all known class labels.

4.2 Results

A set of 21 runs was obtained for each combination of data set and pair of objectives. Each run generated a set of non-dominated solutions, which was then analyzed with respect to the performance measures discussed above. Full results are available as supplementary material [6]. In the following, we will show selected results obtained by the methods, with the aim of highlighting key strengths and limitations of the four combinations of objectives used.

In terms of the size of the solution sets returned by the methods, we find that BMM2 and MOCK return respectively the most and second most non-dominated partitionings. DH returns the least solutions, followed by BMM1. We postulate that this ordering is due to the different levels of resolution of the objectives used. As discussed in the previous section, measures with partial resolution (such as minimum split) are calculated based on the position of a few, extreme data points only. Consequently, many different

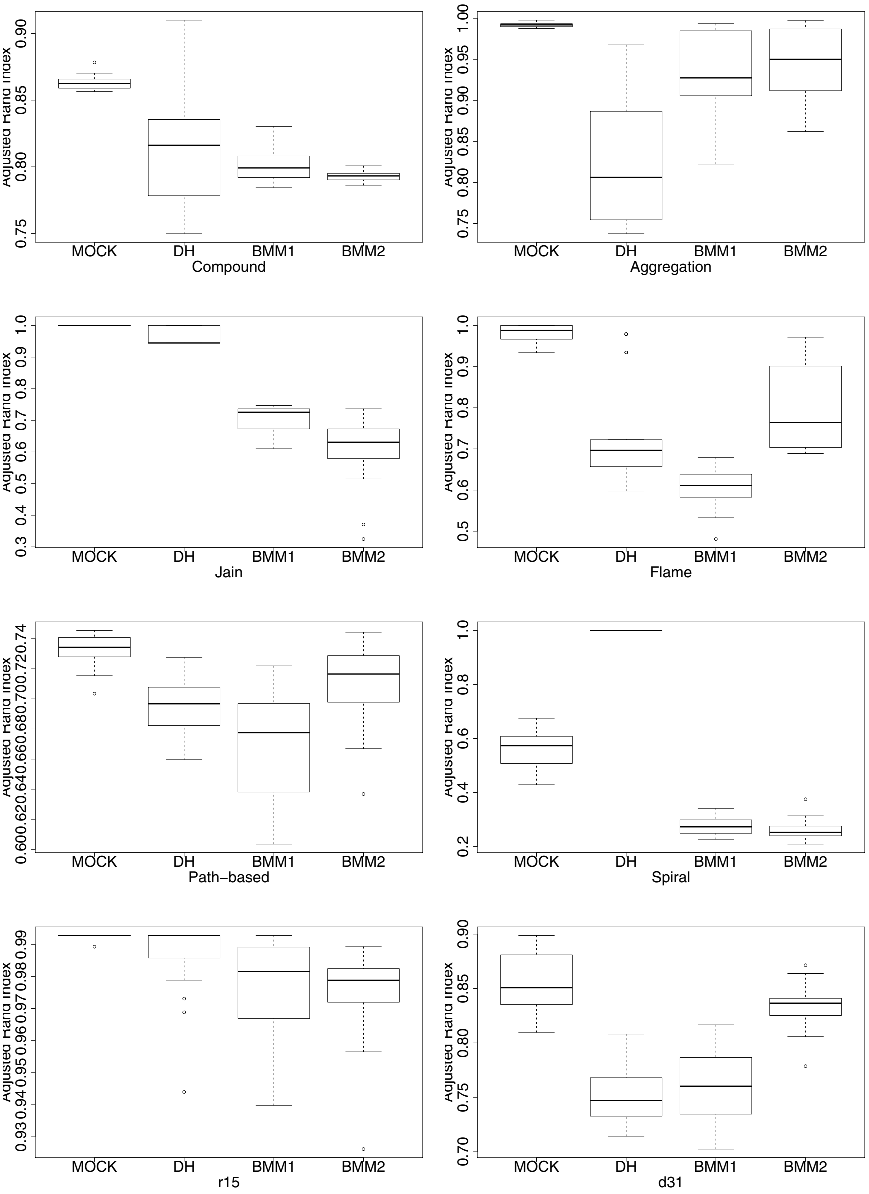

Fig. 2. Best solutions (as judged by the Adjusted Rand Index) identified by the four methods on the two-dimensional data sets. Results are over 21 runs.

partitionings will result in the same objective value such that plateaus are introduced into the search space. Our results indicate that this also reduces the number of Pareto optimal solutions. Figure 1 shows representative results for data sets r15 and d31.

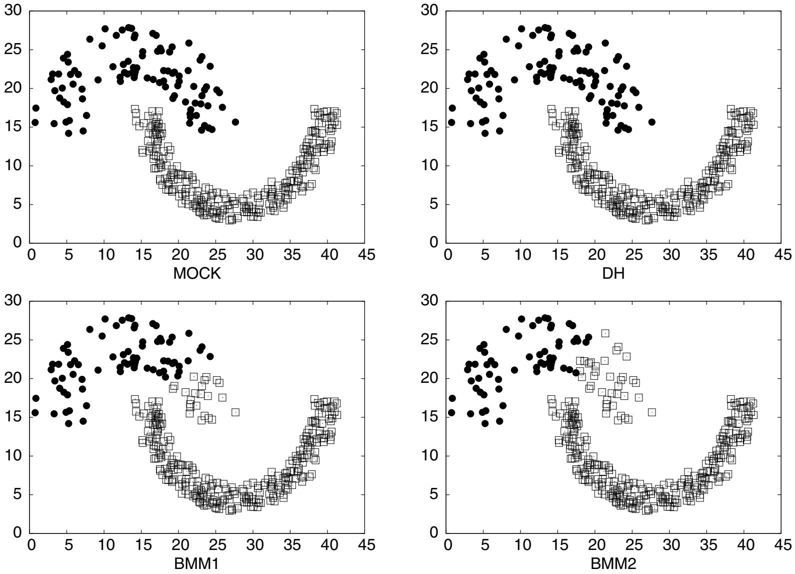

Fig. 3. Illustrative example: Non-spherical clusters. Best two cluster solution returned on the Jain data set by the first run of each method. The nonconvex shape of the clusters introduces problems for methods BMM1 and BMM2, which implicitly assume a convex shape through the use of cluster centroids in both objectives.

Internal validation of clustering results, based on the eight different criteria of clustering quality, indicates that, as expected, all of the four methods outperform their contestant techniques at optimizing their individual pair of objectives (results not shown).

Figure 2 summarizes the results of external cluster validation (based on the Adjusted Rand Index) for the two-dimensional data sets. From these data, it is clear that the choice of objectives has significant impact on the quality of the best solutions returned. This is further confirmed by the results obtained for the Gaussian data sets (see supplementary material [6]). Out of the four methods tested, MOCK shows the best peak performance for the majority of the data sets. The performance differences observed can be understood in more detail by considering the objectives' performance with respect to cluster structures that pose challenges. Using the clustering results obtained for the Jain, Flame and Spiral data, Figures 3 to 5 highlight the effects of nonconvex clusters, chaining between clusters [4] and highly elongated clusters. Key observations from this analysis are limitations of Bandyopadhyay et al.'s techniques with respect to unequally sized and nonconvex clusters (a direct consequence of the use of cluster centroids in both objectives), limitations of Delattre and Hansen's technique with respect to chaining / overlap between cluster, and limitations of MOCK's connectivity measure for extremely elongated clusters (which may be overcome through adjustment of the parameter L in the connectivity measure).

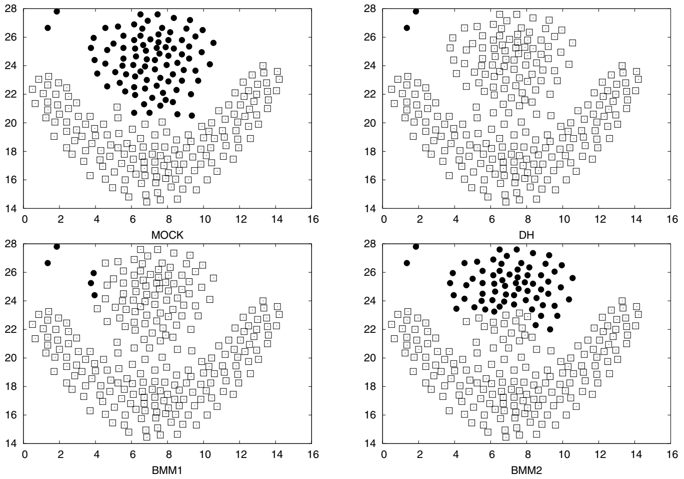

Fig. 4. Illustrative example: Non-spherical clusters with chaining. Best two cluster solution returned on the Flame data set by the first run of each method. The chaining between clusters poses problems for method DH. As the objectives used in DH do not consider the location of all data points, they are more sensitive to this type of noise. The presence of non-spherical clusters makes this data set problematic for methods BMM1 and BMM2.

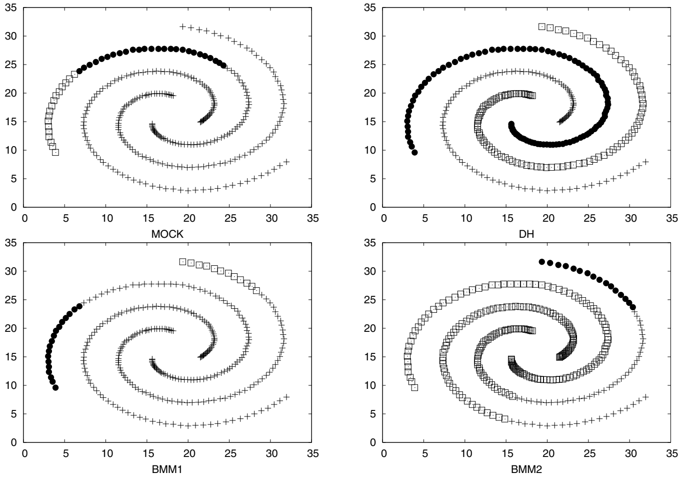

Fig. 5. Illustrative example: Highly elongated clusters. Best three-cluster solution returned on the Spiral data set by the first run of each method. Again, the non-spherical shape of the clusters is problematic for methods BMM1 and BMM2. MOCK also shows a poor performance on this data, as the clusters are so elongated that the connectivity measure ceases to work (some of the L nearest neighbours of each data point are located in another cluster).

5 Conclusion

This manuscript has focused on the comparison of four pairs of criteria for multiobjective clustering. One pair of criteria — from Delattre and Hansen's early bicriterion clustering algorithm — has not previously been evaluated except on very small data sets. The results show that, despite some conceptual similarities in the clustering criteria compared here, significant performance differences can be observed when they are employed within a multiobjective evolutionary algorithm for clustering. Overall, the pair of objectives employed in the multiobjective clustering algorithm MOCK emerges as the strongest combination. We offer two explanations for this result: (i) the limited use of cluster centroids in MOCK's objectives (use in one rather than both objectives) and (ii) the consideration of all data points in the calculation of both of MOCK's objectives. Here, results were all generated using PESA-II; future work may seek to generalize our findings to alternative metaheuristic or exact methods.

References

1. Bandyopadhyay, S., Mukhopadhyay, A., Maulik, U.: An improved algorithm for clustering gene expression data. Bioinformatics 23(21), 2859–2865 (2007)
2. Caballero, R., Laguna, M., Martí, R., Molina, J.: Scatter tabu search for multiobjective clustering problems. Journal of the Operational Research Society 62(11), 2034–2046 (2010)
3. Corne, D., Jerram, N., Knowles, J., Oates, M.: PESA-II: Region-based selection in evolutionary multiobjective optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2001, pp. 283–290. Morgan Kaufmann Publishers (2001)
4. Delattre, M., Hansen, P.: Bicriterion cluster analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 2(4), 277–291 (1980)
5. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17(2), 107–145 (2001)
6. Handl, J., Knowles, J.: `http://personalpages.manchester.ac.uk/mbs/julia.handl/moc.html`
7. Handl, J., Knowles, J.: Exploiting the Trade-off — The Benefits of Multiple Objectives in Data Clustering. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 547–560. Springer, Heidelberg (2005)
8. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. IEEE Transactions on Evolutionary Computation 11(1), 56–76 (2007)
9. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation for post-genomic data analysis. Bioinformatics 21(15), 3201–3212 (2005)
10. Hansen, P., Delattre, M.: Complete-link cluster analysis by graph coloring. Journal of the American Statistical Association, 397–403 (1978)
11. Jain, A., Murty, M., Flynn, P.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)
12. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes. IEEE Transactions on Evolutionary Computation 13(5), 991–1005 (2009)