Three Case Studies Using Agglomerative Clustering

Rodrigo C. Camargos[1(✉)] and Maria do Carmo Nicoletti[1,2]

[1] Faculdade Campo Limpo Paulista (FACCAMP), Campo Limpo Paulista, SP, Brazil
{rodrigocamargos,carmo}@cc.faccamp.br
[2] Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brazil

Abstract. Finding a data clustering in a data set is a challenging task since algorithms usually depend on the adopted inter-cluster distance as well as the employed definition of cluster diameter. The work described in this paper approaches a well-known agglomerative clustering algorithm named AGNES (*Agglomerative Nesting*), in regards to its performance on three case studies namely, datasets formed by clusters of different sizes, uneven inter-cluster distances and diameters. Clustering results are evaluated using three well-known indexes, Dunn, Davies-Bouldin and Rand. Results obtained with K-means were used for comparison purposes. The experiments were conducted divided into three case studies. Their results suggest that AGNES and K-means have similar performance as far as identifying clusters with different sizes and inter-cluster distances, however, AGNES obtained the best results when dealing with clusters having both, different sizes and diameters.

Keywords: Unsupervised machine learning · Agglomerative clustering · Case studies in clustering

1 Introduction

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) focused mainly on investigations and proposals of new formalisms and computational algorithms, aimed at proving theoretical support as well as implementing automatic learning by computers. Over the past few decades, many ideas on how to enable automatic learning have been presented, discussed and implemented. Among the many ways ML can be implemented, the so called clustering algorithms are a particular group of (unsupervised) algorithms which aim at organizing sets of data points into groups, having data exploratory processes in sight. In the literature one can find numerous works proposing new clustering algorithms as well as different clustering taxonomies. Although taxonomies aim at organizing such algorithms into categories, due to the fact that usually they adopt different criteria for grouping them, they are not necessarily, compatible to each other (see, for instance, those suggested in [1–4]).

As pointed out in [5], efficient clustering techniques are considered a challenge, mainly due to the fact that there is no external supervision, which implies knowing nothing about the internal structure of point sets (such as spatial distribution, volume, density, geometric shapes of clusters, etc.). In such a scenery automatic learning becomes an exploratory task, aiming at identifying which are the groups of data points that are statistically separable (or not), which are the most obvious clusters and how they relate to what is aimed at to

A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_7

discriminate, in an attempt to expose the underlying structure of the data, based only on their descriptions, generally given as a vector of attribute values.

The main focus of this work is an empirical research and evaluation of the AGNES (*AGglomerative NESting*) algorithm [3], taking into account data point sets with different characteristics, cluster sizes and inter-cluster distances. For the experiments a collection of data points was artificially created. Seven data point sets were used for evaluating the performance of AGNES, having the K-means algorithm [6] as baseline. The experiments were organized into three case studies, depending on the characteristics of the data point sets.

The remainder of this paper has four more sections. Section 2 comments on the main characteristics of the agglomerative approach and gives a high-level pseudocode of the AGNES algorithm, which has been implemented and used in the experiments. Section 3 describes the data used in the experiments, organized as three case studies. Section 4 briefly introduces the validation indices Dunn, Davies-Bouldin and Rand, used for evaluating the experiment results, followed by the set of experiments related to each case study, discussing their results and presenting some comparative analysis. Section 5 resumes the work done and highlights some conclusions.

2 AGNES (*AGglomerative NESting*) Algorithm

For a given data set containing N data points to be clustered, agglomerative hierarchical clustering algorithms usually start with N clusters (each single data point is a cluster of its own); the algorithm goes on by merging two individual clusters into a larger cluster, until a single cluster, containing all the N data points, is obtained. Obviously, the algorithm can have another stopping criteria, such as that of ending when a clustering containing a user-defined number of clusters (k) is obtained.

Figure 1 presents a high level pseudocode of the AGNES algorithm which, at each iteration, chooses two clusters to be merged, based on the shortest Euclidean distance between the clusters formed so far. The many clustering agglomerative algorithms found in the literature can be organized taking into account the way the inter-cluster distance is defined i.e., what definition is used to compute the shortest distance between all pairs of clusters. Among the various ways, three are particularly popular and are defined next. Given two clusters X and Y, let d(x,y) denotes the distance between two data points.

(1) *Single Linkage*, the distance between two clusters X and Y is the shortest distance between two data points, $x \in X$ and $y \in Y$, formally represented by Eq. (1).

$$d_{SL}(X, Y) = \min_{x \in X,\, y \in Y} d(x, y) \tag{1}$$

(2) *Complete Linkage*, the distance between two clusters X and Y is the farthest distance between two data points, $x \in X$ and $y \in Y$, formally represented by Eq. (2).

$$d_{CL}(X, Y) = \max_{x \in X,\, y \in Y} d(x, y) \tag{2}$$

(3) *Average Linkage* or *UPGMA*, where the distance between two clusters is the mean distance between data points of each cluster in one cluster to every point in the other cluster, formally represented by Eq. (3).

$$d_{AL}(X, Y) = 1/(|X| \times |Y|) \times (\sum_{x \in X} \sum_{y \in Y} d(x, y)) \quad (3)$$

```
procedure AGNES (X,K,AC_t)
Input: X = {P_1, P_2,...,P_N} % dataset with N patterns
Output: AC_t
begin
t ← 0
AC_t ← {{P1}, {P2},...,{PN}} % initial clustering
Nro_C ← N                    %initial number of clusters
CI ← N+1        % cluster index for new created clusters
repeat
  t ← t + 1
  among all possible pairs of clusters {C_r,C_s} in AC_t-1,
  find one {C_i,C_j} such that g(C_i,C_j) = min_r,s g(C_r,C_s),
  where g is a dissimilarity function (distance)
  New_C ← C_i ∪ C_j % create a new cluster
  Nro_C  ← Nro_C − 1
  C_CI  ← New_G
  CI ← CI +1
  AC_t  ← (AC_t-1 − {C_i,C_j}) ∪ {C_CI}
until Nro_C ≤ K
end
```

Fig. 1. A customized version of AGNES pseudocode, based on [1, 3].

For the experiments described in Sect. 4, AGNES was implemented using the UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*), for inter-cluster distance, where all pair-wise distances contribute equally.

3 Data Description

The experiments described in Sect. 4 had three different focuses of empirical investigations: (1) sizes of clusters; (2) inter-cluster distances and (3) diameter of clusters. For addressing each focus, a corresponding case study was conducted, having two-dimensional point sets prepared according to the particular focus intended. For all the experiments a total of 7 synthetic sets of data points were created.

Case Study I (CS-I:Squares) uses three point sets created having their clusters at different distances between themselves, (a) Square1, (b) Square3 and (c) Square5. The only difference between the three point sets Square1, Square3 and Square5 is the degree of overlap between the four clusters that define each point set. In Square1, the clusters touch each other but hardly overlap, whereas in Square5 the overlap is such that there is little density difference when moving from one cluster to the next, as can be seen in

Fig. 2. Point sets Square1, Square3 and Square5 can be described as square arrangements of four clusters of equal size and spread, each cluster being a Gaussian distribution around a central point; they differ from each other only in relation to the inter-cluster distances. The number of clusters, the size of the clusters and the average and standard deviation vectors of each cluster were previously defined for both point sets: Square (CS-I) and Sizes (CS-II). Square and Sizes were created, based on the ones used in [7], having in mind to investigate the sensitivity of an agglomerative clustering algorithm to the inter-cluster distance as well as the increase of the overlapping between clusters.

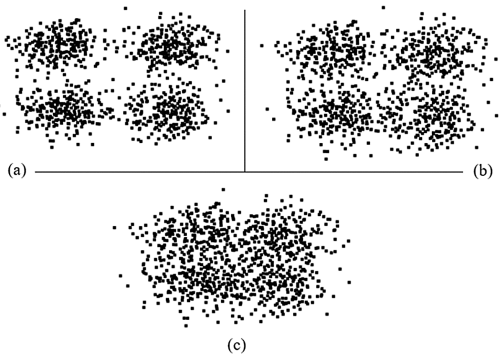

Fig. 2. Case *Study I* - Point sets formed by clusters with different inter-cluster distances (a) Square1, (b) Square3 and (c) Square5.

Case Study II (CS-II: Sizes) also uses three point sets, each created having four clusters. The differences among the three point sets rely on the sizes (number of data points) of their corresponding clusters, (a) Sizes1, (b) Sizes3 and (c) Sizes5, as shown in Fig. 3. The three point sets have been created based on the Square1 (Fig. 2(a)), where there were changes in the relative cluster sizes such that the ratio of the smaller to its immediate larger cluster was 2, 6 and 10, respectively. It is important to notice, though, that the spread of the clusters has been kept constant. Sizes has been created for investigating the sensibility of the algorithm to clusters of different sizes.

Case Study III (CS-III:Aggregation). The seventh point set, named *Aggregation*, (shown in Fig. 4) consists of 7 spherical-shaped clusters, (labeled 1, 3, 4, 5, 6 and 7 in Fig. 4) and 1 non-spherical cluster (labeled 2 in Fig. 4). The point set has been created based on the one used in [8] for experiments related to the clustering aggregation problem. Such set of points has characteristics that are known to create difficulties for many agglomerative algorithms, such as narrow "bridges" between clusters, uneven-sized clusters, clusters with different diameters, etc. On the one hand, agglomerative clustering algorithms based on the nearest neighbor, such as *single linkage*, tend to join groups that touch each other, such as the clusters 3 and 6 as well as 4 and 7 in Fig. 4. On the other hand, agglomerative clustering algorithms based on furthest neighbor, such as *complete linkage*, tend to break large clusters (such as cluster 4 in Fig. 4), based on

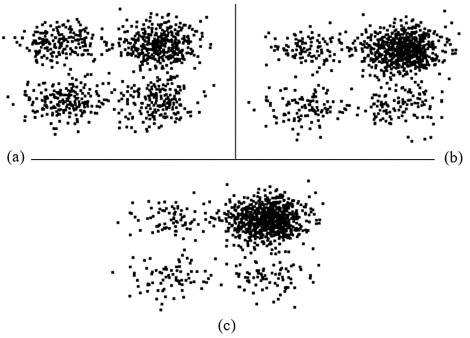

Fig. 3. Case *Study II* - Point sets used in the experiments, formed by clusters of different sizes (number of points). (a) Sizes1, (b) Sizes3 and (c) Sizes5.

the diameters of the small ones, such as clusters 1, 5 and 7. Table 1 presents a summary of the point sets involved in the experiments, describing their basic characteristics.

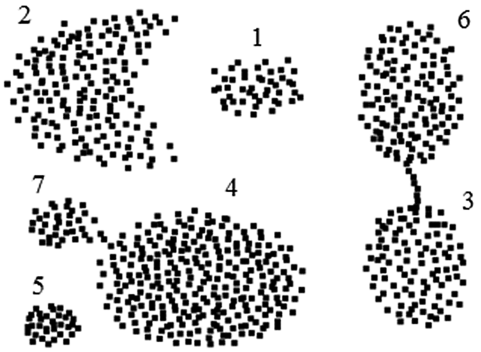

Fig. 4. Case *Study III* - Point set (*Aggregation*) formed by clusters with different diameters. Each cluster has been labeled for reference.

Table 1. Summary of the seven synthetic point sets. #NP: no. of points (size), #NC: no. of clusters. (*)No. of points taking into account the label ordering in the *Aggregation* data.

Point sets	#NP	#NC	Sizes of groups
Square1	1000	4	250-250-250-250
Square3	1000	4	250-250-250-250
Square5	1000	4	250-250-250-250
Sizes1	1000	4	400-200-200-200
Sizes3	1000	4	667-111-111-111
Sizes5	1000	4	769-77-77-77
Aggregation	788	7	45-170-102-273-34-130-34(*)

4 Validation Indices, Experiments, Results and Analysis

This section presents the empirical results obtained by AGNES algorithm [3], considering the three case studies described in the previous section, involving seven data point sets. For the experiments described in this section, AGNES and K-means [6] were implemented in C# and run under a Microsoft Windows platform. The K-Means algorithm was used for comparative purposes only. The results from running both algorithms using the same input data are presented in the following three tables, where the best results are bold faced. To make a fair comparison the number of clusters (K) supplied by the user, to both algorithms, was the same. It is important to mention that all instances in the seven point sets were previously labeled (i.e., has a class attribute associated) in order to allow for conducting external validation. The inherent separation easily perceived between the clusters in the point sets was employed for class assignment.

The results obtained were evaluated using two internal validation indexes: the Dunn's index (D) [9] and the Davies-Bouldin index (DB) [10]. To quantify the number of data points incorrectly assigned (using the original classes previously assigned, and part of the description of each data point in all the seven point sets), the Rand index (R) [11] external validation index was also used.

Consider the following notation. S - point set clustered into clustering C_1, C_2,…, C_{NC}; |S| - number of data points in S; C_i - ith cluster of the clustering; n_i - number of data points in C_i; E_x, E_y - two data points, and let the distance between the two data points, E_x, E_y, be represented as dist(E_x, E_y). The Dunn's index (D) is defined by Eq. (4).

$$\begin{gathered} D = \min_i \left\{ \min_j (A/B) \right\}, \text{where} \\ A = \min_{Ex \in Ci, Ex \in Cj} \operatorname{dist}\left(E_x, E_y\right) \text{ and } B = \max_k \{ \max_{Ex,Ey \in Ck} \operatorname{dist}\left(E_x, E_y\right) \} \end{gathered} \tag{4}$$

For defining the Davies-Bouldin index (DB) as in [10], first consider s_i be a measure of dispersion of a cluster C_i (*i.e.*, a measure of its spread around its mean vector) and let $d(C_i,C_j) = d_{ij}$ be the dissimilarity between two clusters, using an appropriate dissimilarity measure (*e.g.*, distance). Consider the similarity index R_{ij}, between C_i and C_j be given by: $R_{ij} = (s_i + s_j)/d_{ij}$, provided that d_{ij} is symmetric. Let R_i be defined as $R_i = \max_{j=1,..,NC, j \neq i} R_{i,j}$, i = 1,…, NC. Then, the Davies-Bouldin index DB is defined by Eq. (5).

$$DB = 1/NC \times \sum_{i=1,\ldots NC} R_i \tag{5}$$

In data clustering the value of the Rand index [11] can be approached as a measure of the similarity between two data clusterings. For the experiments presented next, one of the clusterings will be the one induced by the clustering algorithm and the other, the one provided externally, by previously assigning a class to each data point, in each of the seven point sets considered. So, in a general setup, given the notation previously introduced and considering that one of the clustering of the point set S is given as $O = \{O_1, O_2, \ldots, O_{NO}\}$ and the other as $C = \{C_1, C_2, \ldots, C_{NC}\}$, the Rand index is composed by the following values:

(i) a: number of pairs of points in S that are in the same set in O and in the same set in C;
(ii) b: number of pairs of points in S that are in different sets in O and in different sets in C;
(iii) c: number of pairs of points in S that are in the same set in O and in different sets in C and
(iv) d: number of pairs of points in S that are in different sets in O and in the same set in C. So the Rand index (R) is given by Eq. (6).

$$R = (a + b)/(a + b + c + d) \tag{6}$$

Intuitively (a + b) can be thought of as the number of agreements between O and C and (c + d) as the number of disagreements between O and C.

In the following tables, (+) and (−) report the best and the worst result of K-Means, respectively. As far as the point sets of *CS-I:Squares* are concerned, the experiment results shown in Table 2 indicate that both algorithms, AGNES and K-Means had similar performance. It is important to mention, however, that AGNES performed slightly better than K-means when taking into account the K-Means worst-case results. Nevertheless, considering the CS-I results, AGNES and K-Means algorithms share similar performances when considering clusters with uneven-sizes and different inter-cluster distances.

Table 2. Clustering results for case study *CS-I:Squares* – Point set in Fig. 2.

Clustering algorithm	Point set	D	DB	R
AGNES	Square1	**0.06**	**0.42**	**0.974**
AGNES	Square3	0.03	0.55	0.934
AGNES	Square5	0.02	0.79	0.789
K-means (+)	Square1	**0.05**	**0.40**	**0.980**
K-means (−)	Square1	0.05	0.42	0.980
K-means (+)	Square3	0.01	0.52	0.950
K-means (−)	Square3	0.01	0.52	0.947
K-means (+)	Square5	0.01	0.61	0.887
K-means (−)	Square5	0.01	0.58	0.880

The numbers obtained from experiments in *CS-II:Sizes* shown in Table 3 suggest that, as the size (i.e., number of points) of the four clusters change, AGNES and the K-Means performances also change.

The three validation indexes used to evaluate the quality of clusterings induced by AGNES suggest different outcomes. The values of the D index implies that AGNES achieved its best performance on the Sizes5 data point set; however, the DB and the R indexes suggest that the best results were obtained when running AGNES on Sizes1. Taking into account only clustering results from AGNES, its best performance was achieved having Size1 as input. Comparing AGNES and K-means results obtained in CS-II:Sizes, it is obvious that K-Means had the best performance as far as the values of indexes DB and R are concerned.

Table 3. Clustering results for case study *CS-II:Sizes* – Point set in Fig. 3.

Clustering algorithm	Point set	D	DB	R
AGNES	Sizes1	0.04	**0.43**	**0.974**
AGNES	Sizes3	0.03	0.51	0.936
AGNES	Sizes5	**0.05**	0.44	0.963
K-means (+)	Sizes1	0.03	**0.40**	**0.982**
K-means (−)	Sizes1	0.03	0.47	0.972
K-means (+)	Sizes3	0.03	0.43	0.972
K-means (−)	Sizes3	0.03	0.47	0.972
K-means (+)	Sizes5	0.02	0.46	0.966
K-means (−)	Sizes5	0.00	0.69	0.499

Case Study III (*CS-III:Aggregation*) focuses on clustering experiments having, as input, the *Aggregation* point set, shown in Fig. 4. The values of the three validation indexes applied to the clustering results obtained from several experiments using *Aggregation* as input are presented in Table 4.

The experiments were conducted following a different methodology than the one used in the previous two case studies. It was decided to initially consider *Aggregation* as having only clusters 1 and 6 and then, gradually grow the point set by adding to it some of its clusters, until reaching the 7 total clusters it effectively has. By doing so it was expected that the experiment would allow to investigate what point set configuration would have more impact on AGNES performance.

The *Aggregation* point set is initially formed by clusters 1 and 6; both clusters are different with regard to their sizes, diameters and shapes, but are well separated and their inter-cluster distance promotes a good performance of both, AGNES and K-means. The values of the Rand index in Table 4 suggest that both algorithms correctly identified clusters 1 and 6. Next, the clustering results shown in Table 4 are from running AGNES and K-means on the previous *Aggregation* point set, added with cluster 3. Such addition did not decrease AGNES performance, in spite of the "bridge" between clusters 3 and 6. The K-Means, however, failed to identify some points from cluster 3. In sequence, the *Aggregation* was modified again, by replacing cluster 3 and 6 with cluster 2. In spite of cluster 1 and 2 having different shapes, which eventually could create difficulties for some clustering algorithms, their differences in shape were not substantial enough and AGNES identify both correctly. The K-Means, however, could not fully identify points from cluster 2 due to their distance from the cluster's centroid.

The next modification of *Aggregation*, at this point having cluster 1 and 2, was done by adding to it the cluster 6. AGNES was able to correctly detect the three clusters (1, 2 and 6) while K-Means failed again for identifying points in cluster 2. *Aggregation* suffered another addition, this time of cluster 3. With this configuration both algorithms, AGNES and K-Means, did not identify correctly the four clusters; yet, AGNES was the one that came closer. Next, by adding cluster 4 to the previous *Aggregation*, the K-Means still had trouble to fully detect all clusters, while AGNES recovered its optimal performance, since with the addition of cluster 4, the average distance between the five clusters changed. Moreover, it can be easily noticed that, by adding cluster 5 to *Aggregation*,

Table 4. Clustering results for case study *CS-III:Aggregation* – Point set in Fig. 4.

Clustering algorithm	Point set	D	DB	R
AGNES	Clusters 1 and 6	0.35	0.48	1
K-means (+)	Clusters 1 and 6	0.35	0.48	1
K-means (−)	Clusters 1 and 6	0.35	0.48	1
AGNES	Clusters 1, 3 and 6	0.04	0.33	**1**
K-means (+)	Clusters 1, 3 and 6	0.04	**0.32**	0.998
K-means (−)	Clusters 1, 3 and 6	0.04	**0.32**	0.998
AGNES	Clusters 1 and 2	**0.04**	0.33	**1**
K-means (+)	Clusters 1 and 2	**0.04**	**0.32**	0.998
K-means (−)	Clusters 1 and 2	0.08	0.54	0.972
AGNES	Clusters 1, 2 and 6	**0.33**	**0.33**	**1**
K-means (+)	Clusters 1, 2 and 6	0.08	0.34	0.993
K-means (−)	Clusters 1, 2 and 6	0.08	0.35	0.989
AGNES	Clusters 1, 2, 3 and 6	0.04	0.41	**0.993**
K-means (+)	Clusters 1, 2, 3 and 6	0.04	**0.36**	0.989
K-means (−)	Clusters 1, 2, 3 and 6	0.04	**0.36**	0.989
AGNES	Clusters 1, 2, 3, 4 and 6	**0.04**	0.39	**1**
K-means (+)	Clusters 1, 2, 3, 4 and 6	0.03	**0.38**	0.997
K-means (−)	Clusters 1, 2, 3, 4 and 6	0.02	0.51	0.900
AGNES	Clusters 1, 2, 3, 4, 5 and 6	**0.04**	**0.34**	**0.998**
K-means (+)	Clusters 1, 2, 3, 4, 5 and 6	0.03	0.41	0.928
K-means (−)	Clusters 1, 2, 3, 4, 5 and 6	0.03	0.49	0.889
AGNES	All 7 clusters	**0.04**	**0.33**	**0.998**
K-means (+)	All 7 clusters	0.03	0.44	0.927
K-means (−)	All 7 clusters	0.03	0.50	0.919

the average cluster distances computed by AGNES was negatively affected. In spite of that, AGNES still had a much better performance than K-means. Finally, *Aggregation* was restored to its original seven clusters, as in Fig. 4. Notice that the addition of cluster 7 did not decrease AGNES performance and the clustering it induced was very close to the optimal result, as confirmed by its Rand index of 0.998.

5 Conclusions

This paper addresses the use of the agglomerative hierarchical clustering algorithm AGNES, in unsupervised tasks involving 7 point sets, grouped into three case studies: the *CS-I:Squares*, which focuses on inter-cluster distance, the *CS-II:Sizes*, with focus on clusters of different sizes and the *CS-III:Aggregation*, involving mainly different shapes and diameters. As far as both case studies, *CS-I:Squares* and *CS-II:Sizes*, are concerned, AGNES and K-Means results were similarly evaluated; however, in the *CS-I:Squares* AGNES had a slightly better performance than K-means when taking into account K-Means worst-case results. In the third case study, the *CS-III:Aggregation*, in

most experiments, the best results were obtained with AGNES. In spite of AGNES being more strongly affected by the inter-cluster distance than any of the other chosen characteristics, such as size, diameter or shapes (except for elongated clusters where *single linkage* based algorithms usually have better performance), the algorithm was still very robust considering a combination of all the chosen characteristics.

Acknowledgments. The authors would like to thank CAPES, CNPq and FACCAMP.

References

1. Theodorides, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Elsevier, Amsterdam (2009)
2. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
3. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Hoboken (1990)
4. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann-Elsevier, Amsterdan (2012)
5. Aggarwal, C.C., Reddy, C.K. (eds.): Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, Boca Raton (2013)
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, vol. 1, pp. 281–297 (1967)
7. Handl, J., Knowles, J.: Evolutionary multiobjective clustering. In: Yao, X., et al. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 1081–1091. Springer, Heidelberg (2004). doi: 10.1007/978-3-540-30217-9_109
8. Gionis, A., Manilla, H., Tsaparas, P.: Clustering aggregation. ACM Trans. Knowl. Disc. Data **1**, 30 p. (2007). Article 4
9. Dunn, J.: Well separated clusters and optimal fuzzy partitions. J. Cybern. **4**, 95–104 (1974)
10. Davies, D.L., Boudin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**, 224–227 (1979)
11. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)