A Family of Two-Dimensional Benchmark Data Sets and Its Application to Comparing Different Cluster Validation Indices

Jorge M. Santos[1,2] and Mark Embrechts[3]

[1] ISEP, School of Engineering, Polytechnic of Porto - Dept. of Mathematics
[2] INEB, Biomedical Engineering Institute, Porto - Portugal
[3] Rensselaer Polytechnic Institute - Dept. Ind. Systems Eng., Troy, NY - USA

Abstract. There are two main objectives in this paper: the first one is to introduce a collection of two-dimensional benchmark data sets with a wide variety of clustering characteristics that are typical for real-world data sets. These simple 2-D data sets allow the user to easily evaluate clustering solutions from a variety of different clustering algorithms; the second one is to evaluate four different commonly used clustering validation indices by using these 2-D benchmark data sets. It is shown that even for simple 2-D data sets there is a large discrepancy on the ideal number of clusters suggested by traditional cluster validation indices. The performed experiments also suggest that the Dunn and the GAP statistic seems to be more robust cluster validation indices, even though they still fail to comply with common sense clustering solutions in more than 50% of the cases.

1 Introduction

The aim of clustering for a given data set is to identify different groups in such a way that (i) data within each group are similar to each other and (ii) different groups are dissimilar to each other. Clustering is a very complex and problem dependent task. There are several different clustering algorithms and it is well known that the resulting clusters are not unique: vastly different acceptable clustering solutions for the same data set can (and often will) occur. Two popular clustering algorithms will be applied in this paper to compare cluster validation indices: (i) the standard agglomerative average link hierarchical clustering with Euclidean distance and (ii) the K-means clustering. Hierarchical agglomerative clustering starts by assigning each data point to a single cluster and then enlarging the clusters by joining clusters based on cluster similarity measures (which can be based on different distance metrics (e.g., the Euclidean, and the Manhattan distance). The K-means algorithm on the other hand starts with a given number of cluster prototypes, then it iteratively assigns data points to these prototypes and updates them accordingly.

Because there isn't a known a priori solution for a clustering problem, cluster validation indices are often utilized (i) to evaluate and compare different proposed clustering solutions and (ii) to assess the cluster solution quality. There

J.F. Martínez-Trinidad et al. (Eds.): MCPR 2014, LNCS 8495, pp. 41–50, 2014.

is a variety of cluster validation indices. In this work four popular cluster validation indices will be applied as explained in Sect. 3. Cluster validation indices are typically used to determine the optimal number of clusters for a particular clustering method.

There are few two-dimensional benchmark data sets for clustering: in most cases such 2-D benchmark data are limited to two or more partially overlapping Gaussian blobs. An example of a set of 2-D data sets with a small set of different clustering situations can be found in [1]. In addition, there are some data sets for image segmentation with clustering solutions proposed by humans [2,3]. We introduce in this paper a benchmark data and make it available for benchmarking new clustering algorithms. 2-D data sets offer the advantage that it can easily be visualized to evaluate the proposed clustering solution.

For high-dimensional data sets one usually relies on cluster validation indices for assessing the clustering quality. It will be shown in this paper that these indices are a weak tool at best to assess cluster quality: even with very simple data sets such cluster validation indices often fail to identify the proper number of clusters using the most common clustering algorithms.

2 The Data Sets

The data sets presented in this work were first proposed by Santos et al. in 2005 [4] and are available via [5]. They consist of a set of 30 two-dimensional artificial data sets as depicted in Fig. 5 and they were specifically designed for assessing human clustering on two-dimensional data. All 30 data sets were manually constructed and it was attempted to create the different situations typically encountered in clustering-related tasks.

The 30 benchmark data sets can be divided in different groups according to specific characteristics such as

- Connectedness - Probably the most basic feature leading someone to join points into clusters whenever connecting paths are perceived. This feature is valued in the data set of Fig. 1a when a human "sees" one cluster instead of two.
- Structuring direction - This feature leads us to "see" the two arms of the cross in Fig. 1b instead of only one cluster. Humans are good at perceiving structuring directions in data set graphs, independently of those directions being straight or curved lines.
- Structuring density - This feature leads us to "see" two clusters in Fig. 1c instead of only one.
- Structuring morphology - This feature leads us to "see" two clusters in Fig. 1d instead of only one, deciding differently of the similar Fig. 1a. The reason is that, contrary to Fig. 1a, we now identify the bulging out wart of Fig. 1d with a known form.

A set of manual clustering experiments were performed with control groups of adults and children in order to grasp the clustering process and divide the results

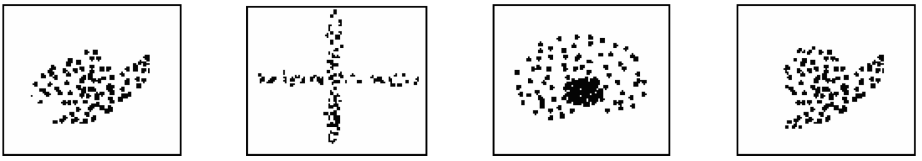


Fig. 1. Clustering features: a) connectedness; b) structuring direction; c) structuring density; d) structuring morphology

according to different kinds of data sets such as data sets (i) with well-separated clusters; (ii) with different density clusters; (iii) with crossing clusters; (iv) with nested-clusters; (v) with spiral-shaped clusters.

Experimental results showed that solutions proposed by adults are more consistent, exhibiting fewer solutions for each data set than the ones proposed by children (6-7 years). A more detailed analysis on children solutions revealed that a large percentage of them usually build clusters with a small number of points. It seems that they pay particular attention to small groups. An example of such behavior is shown in Fig. 2. In this case we may consider as outliers points belonging to extremely small clusters. Adults produce more consistent solutions but also very different. An example of four different solutions proposed for a data set with different point densities is depicted in Fig. 3(a-d).

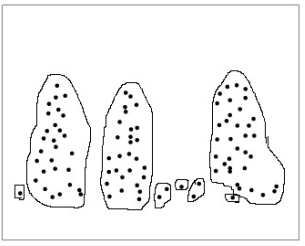
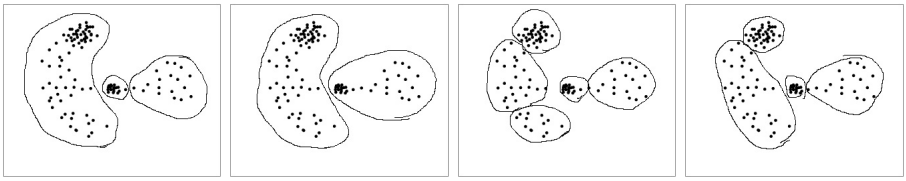


Fig. 2. One clustering solution proposed by children for 'Anthills'

It is well known that there is no unique solution for clustering and the results of the experiments reported in [4] show exactly that.

The 2-D benchmark data sets can also be used for evaluating the performance of clustering algorithms by comparing their solutions with the ones proposed by humans in [4] or with solutions proposed by a specific person. Although the benchmark data sets are relatively small, they can be easily transformed to big data sets by generating additional data points scattered around the original data points. Similarly for the benchmarking of outlier detection algorithms it would be straightforward to produce some outliers for these data sets. In addition we colored in plausible clusters that can be interpreted as different classes for



(a) a (b) b (c) c (d) d

Fig. 3. Four different clustering solutions proposed by adults for 'Anchor' (a-d)

benchmarking supervised learning algorithms. For other possible solutions we refer to [4].

3 Cluster Validation Indices

In this work we chose to use some of the most common cluster validation indices like Davies-Bouldin, Dunn and Silhouette indices but also the GAP statistics, not so well known but that proved to be a very interesting index. In the following we will briefly present these indices.

3.1 Davies-Bouldin Index

The Davies-Bouldin (DB) cluster validation index [6] is a measure that helps to estimate the ideal number of clusters (K) in a dataset and it is based on the average ratio between the within cluster scatter (S_l) for all clusters and the distance D between two clusters. The DB Index is computed by:

$$DB = \frac{1}{N}\sum_{l=1}^{N} R_l \tag{1}$$

with

$$R_l = \underset{l \neq m}{max} R_{lm} \tag{2}$$

$$R_{lm} = \frac{S_l + S_m}{D_{lm}} \tag{3}$$

$$D_{lm} = \|\mathbf{v}_l - \mathbf{v}_m\| = \sqrt{\sum_{k=1}^{N} |v_{kl} - v_{km}|^2} \tag{4}$$

$$S_l = \sqrt{\frac{1}{N_l}\sum_{m=1}^{N_l} |\mathbf{x}_m - \mathbf{v}_l|^2} \tag{5}$$

where N is the number of clusters for which we are computing the DB index, S_l is the within cluster scatter for cluster l, $\mathbf{v}_l$ is the n dimensional cluster centroid for

cluster l, $v_i l$ is the ith component of $\mathbf{v}_l$, $\mathbf{x}_l$ representes an individual data point in cluster l, and N_l and N_m are the number of data in clusters l and m respectively. D_{lm} is the distance between vectors which are chosen as characteristic of clusters l and m. In formulas 4 and 5 we use the Euclidian distance but the original ones use the Minkowski distance.

A lower DB index indicates a better cluster quality.

3.2 Dunn Index

The Dunn index [7] is simpler but similar to the Davies-Bouldin index because it is also based on relations between within-cluster distance and between-clusters distance. The within-cluster distance Δl is computed as the maximum distance between two points in the same cluster l and the between-clusters distance $\delta(l, m)$ between cluster l and cluster m is computed as the smallest distance between two points in different clusters (l and m). The Dunn index is defined as the minimum ratio between the between-clusters distance, $\delta(l, m)$, and the within-cluster distance Δl,

$$Dunn = \frac{\delta(l, m)}{\Delta l} \tag{6}$$

with

$$\Delta l = \underset{x,y \in l}{max}(d(x, y)) \tag{7}$$

$$\delta(l, m) = \underset{x \in l, y \in m}{min}(d(x, y)) \tag{8}$$

A higher Dunn index indicates a better cluster quality.

3.3 Cluster Silhuette Width Index

The cluster silhouette width (SHW) index [8] is a measure that compares cluster tightness and cluster separation and it is based on the silhouette width for each sample, the average silhouette width for each cluster, and the overall average silhouette width for all the data. The optimal number of clusters maximizes the cluster silhouette width index:

$$SHW = \frac{1}{N} \sum_{l=1}^{N} \frac{1}{N_l} \sum_{i=1}^{N_l} s_i \tag{9}$$

with

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

$$a_i = \frac{1}{N_A} \sum_{\substack{x_i, x_j \in A \\ i \neq j}} d(x_i, x_j)$$

$$b_i = \underset{A \neq C}{\text{minimum}}\ d(x_i, C)$$

where N is the number of clusters, N_l is the number of data in cluster l, a_i is the average within-cluster distance, b_i is the minimal average between-cluster distance between two clusters, $d(x_i, x_j)$ is the dissimilarity between objects x_i and x_j of the same cluster, and $d(i, C)$ is the average dissimilarity of object x_i to all objects of C. The silhouette value varies from $-1 \leq S \leq 1$ and a value close to unity for a sample indicates that the sample is 'well-clustered', a value close to -1 is indicative of a possible misclassification and a value close to 0 means that that sample could also be assigned to another cluster.

A higher SHW index indicates a better cluster quality.

3.4 GAP Statistic

The GAP statistic was introduced in [9] and is used to estimate the number of clusters in a data set by comparing the cluster dispersion obtained for a clustering algorithm against the cluster dispersion obtained with the same algorithm for a uniformly random distribution of the same number of data with the same number of attributes. Cluster dispersion is algorithmic dependent. For hierarchical clustering, cluster dispersion is defined as:

$$Gap(k) = \mathbf{E}\{\log(W_k)\} \tag{10}$$

with

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{11}$$

where $\mathbf{E}\{.\}$ is the expectation operator for the reference uniformly random distribution, k is the number of clusters, W_k is the pooled within-cluster sum of squares around the cluster means, and D_r is the sum of all pairwise distances in cluster r. The GAP statistic is a function of the number of clusters in a given dataset, meaning that one usually computes it for a different number of clusters k and take as the possible number of clusters the one for which there is a sudden jump in the difference of the cluster dispersion between the actual data and the random gauge data.

4 Experiments

We performed an exhaustive number of experiments with the 30 data sets using both K-means [10] and hierarchical clustering algorithms [11]. In both cases Euclidean distance-based similarity measures were used. The purpose of these experiments is to demonstrate that (even for simple 2-D data sets) different cluster validation indices often result in different assessments for the optimal number of clusters. In these experiments the suggested number of clusters for both clustering algorithms was determined for each of the four cluster validation indices. We show in Fig. 4 the clustering indexes results obtained with K-means (left) and the hierarchical clustering (right) on the 'Citroen' data set. As one can see, there is a considerable difference on the results obtained for both clustering

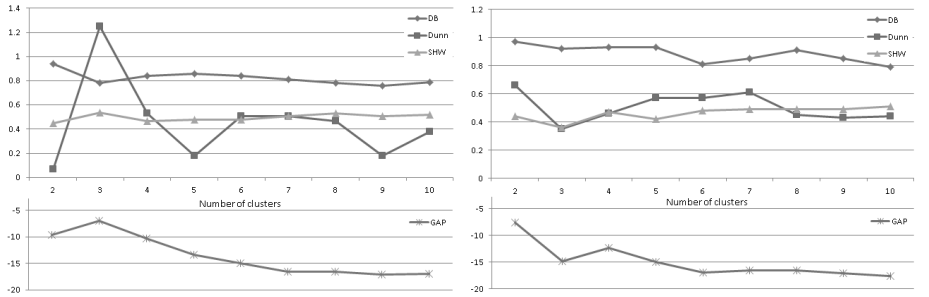


Fig. 4. Davies-Bouldin, Dunn, Silhuette and GAP indices values for data set Citroen for both K-means (a) and hierarchical (b) clustering algorithms

algorithms and also very different values for the different clustering indices. In this case, the suggested number of clusters for K-means clustering was DB=9, Dunn=3, SHW=3, GAP=3 and for hierarchical clustering was DB=10, Dunn=2, SHW=10, GAP=2.

We did not attempt to reconcile the results of obtained number of clusters with the 'real' number of clusters (as suggested by human evaluations). Only in data sets with well-separated and globular shaped clusters one can observe a match between the number of clusters obtained from cluster validation indices and human assessments.

5 Results

Results of the performed experiments with 29 data sets are presented in Table 1. We do not present any results for data set 'One' because there is only 1 cluster

Table 1. Clustering results for the 29 data sets

		Three	Kites	Citroen	Bermuda	Clock	Snake	Anthills	Birds	Anvils	Anchor	Cross	Starfish	Swarm	Helix	Eye	Hockey	Bean	Rings	Boomer.	Stamp	Layers	Sticks	Lips	Duck	Food	Epiglottis	Penguin	Sunset	Spiral	Hits
	Sep.	n	n	n	n	y	y	y	y	y	n	n	n	n	y	n	n	n	n	n	y	y	y	y	n	n	n	n	n	n	
	Clust.	3	2	2	4	3	2	3	3	2	4	2	5	6	2	2	2	2	2	4	3	3	2	2	2	4	2	2	3	2	
K-means	DB	3	3	9	8	3	2	7	3	2	5	5	12	8	3	6	4	5	2	8	6	10	2	6	11	12	7	9	4	8	7
	Dunn	3	2	3	3	4	2	2	2	2	3	2	5	3	2	2	2	3	2	2	8	3	2	6	2	2	8	9	3	3	14
	SHW	2	4	3	6	3	2	6	3	2	3	5	8	5	3	6	3	5	8	10	6	10	2	6	8	11	7	3	3	7	6
	GAP	3	2	3	9	4	2	6	3	2	5	1	1	6	3	3	4	1	2	2	1	2	5	2	3	2	2	3	3	3	10
Hierarch.	DB	3	3	10	8	3	2	9	3	2	5	5	12	8	3	8	4	4	9	10	7	9	2	9	12	12	7	10	5	7	6
	Dunn	3	2	2	2	3	2	2	2	2	6	2	6	2	5	2	4	4	2	2	6	2	2	7	3	2	8	10	3	3	11
	SHW	2	5	10	8	3	2	11	3	2	3	5	7	5	3	5	3	4	7	6	6	7	2	5	4	11	7	4	3	6	6
	GAP	3	2	2	3	3	2	2	3	2	4	4	7	3	3	3	5	3	2	3	3	3	2	2	4	2	2	4	4	3	14
	Hits	6	4	2	0	6	8	0	6	8	1	2	1	1	1	2	1	0	5	0	1	2	7	2	1	0	2	0	5	0	

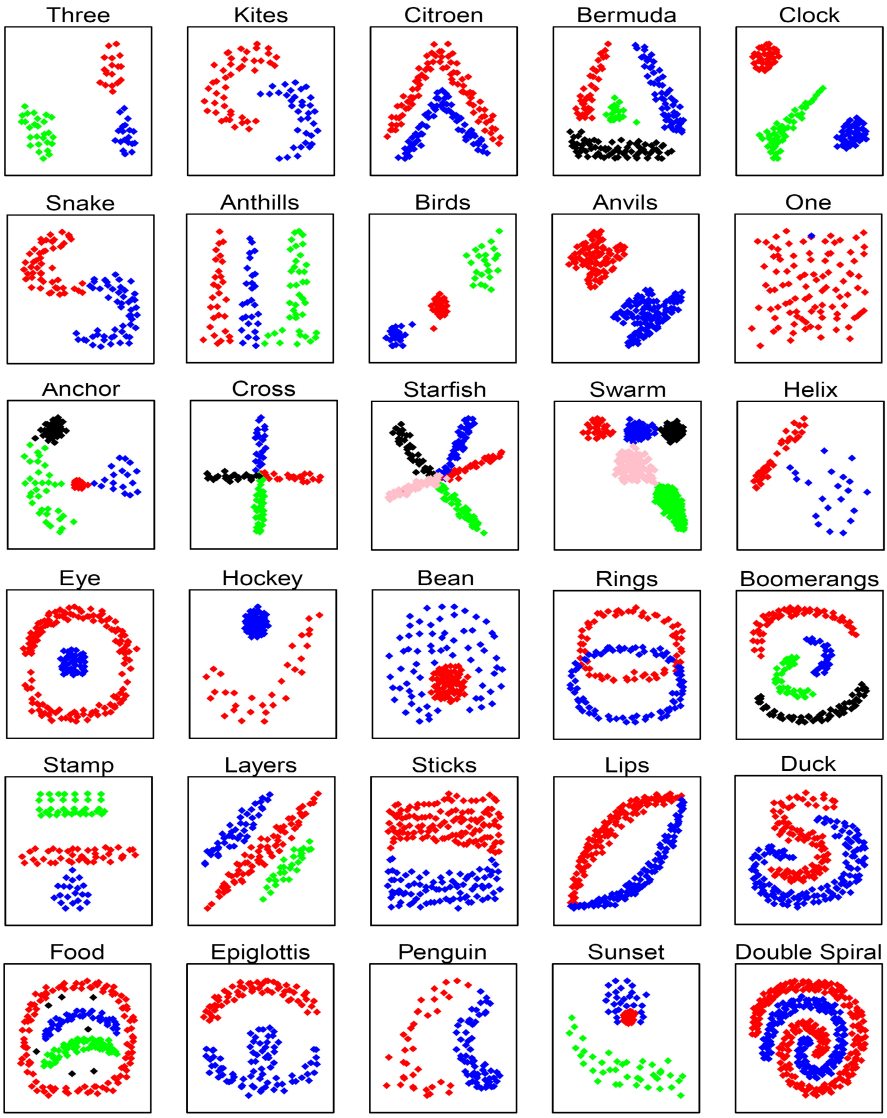


Fig. 5. The 30 Two-dimensional data sets

and several indices exhibit numerical difficulties to yield a single cluster as a valid assessment. The first row (Sep.) shows if the clusters are linearly separable meaning that the clustering problem is an easier one and the second row (Clust.) shows the number of suggested clusters. In the last row we present the number of times there is a coincidence between the number of clusters given by the cluster validation index and the number of suggested clusters. The last column shows

the number of times each cluster validation index gives the correct number of clusters.

As one can see, there is a significant difference in the number of clusters proposed by each clustering validation index for each data set and each clustering algorithm. The difference is related with the way these indices use the within cluster or between clusters distances. Different cluster shapes and structure also represent a problem for clustering algorithms. A closer view to the results may suggest that the Davies-Bouldin and the Cluster Silhuette Width Indices are the indices presenting the worst results and the Dunn and the GAP statistics indices being the ones presenting the best results. We can also see that the results are better for 'simple' data sets as those marked as linearly separable compared with those more 'complex' data sets where the results are quite bad. There are almost 50% of the data sets where the number of hits is zero or one meaning that almost every clustering algorithm and cluster validation index fail to give the correct number of clusters. One can see for example the case of the Bermuda data set that presents 4 distinct clusters and the results of the cluster validation indices are quite different none of them with the correct answer.

We must emphasize again that the purpose of these experiments was not to compare clustering solutions with those proposed in [4] but only to evaluate the results for the number of clusters. If we have done such a comparison we can assume that the results would have been much worst because apart from the bad results on the number of clusters one should add the expected bad results on the clustering solutions because the clustering algorithms used in the experiment would have failed on given the correct solution in several data sets.

6 Conclusions

We have presented in this work a novel set of two-dimensional benchmark data sets for evaluating clustering algorithms and clustering validation indices. Because 2-D data can be visualized, it is easy to assess the resulting clustering solutions and compare them with the 'natural' clustering provided by humans. The natural clustering solution were obtained from a previous experiment using human subjects but they can be substituted by the users own preferred labeling. The presented data sets can also be easily transformed in big data sets by adding data points randomly distributed in the neighborhood of each point.

The benchmark data sets were used to assess the performance of cluster validation indices with two different popular clustering algorithms. This study demonstrates that even for data sets that would be perceived as 'easy' to cluster, the optimal number of clusters as suggested by the cluster validation indices can vary widely.

References

1. Ultsch, A.: Clustering with som: U*c. In: Workshop on Self-Organizing Maps, pp. 75–82 (2005), www.uni-marburg.de/fb12/datenbionik/Daten

2. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision, vol. 2, pp. 416–423 (July 2001)
3. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2007)
4. Santos, J.M., Marques de Sá, J.: Human clustering on bi-dimensional data: An assessment. Technical Report 1, INEB - Instituto de Engenharia Biomédica, Porto, Portugal (October 2005)
5. Santos, J.M.: Bi-dimensioanl data sets, `http://www.dema.isep.ipp.pt/~jms/datasets`
6. Davies, D., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227 (1971)
7. Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4(1), 95–104 (1974)
8. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Computational and Applied Mathematics 20, 53–65 (1987)
9. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(2), 411–423 (2001)
10. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8), 651–666 (2010)
11. Xu, R., Wunsch, D.: Clustering. IEEE Press Series on Computational intelligence. IEEE (2008)