Evolutionary Multiobjective Clustering

J. Handl and J. Knowles

Supporting Material

1 Synthetic data

Here, we describe the two-dimensional synthetic data sets used to study the robustness of VIENNA towards different cluster properties, in particular non-spherically shaped, overlapping and unequally-sized clusters. Table 1 gives the definition of the benchmarks, and plots of one sample instance for each one of them can be found in the accompanying .pdf files.

All clusters (except for two in the Smile data set) are described by two-dimensional normal distributions $N(\vec{\mu}, \vec{\sigma})$. The number of clusters, the sizes of the individual clusters, and the mean vector $\vec{\mu}$ and vector of the standard deviation $\vec{\sigma}$ for each normal distribution are manually fixed. In each run of the experiments, a new set of data is sampled from these distributions.

The *Square* and *Sizes* data sets consist of four clusters, arranged in a square, which are generated by normal distributions with a standard deviation of 2 in both dimensions. In the *Square* data sets all clusters are of equal size (125 data items each), and Square1, Square3 and Square5 only differ by the distance between the individual clusters (i.e. the length of the edges of the square), which is 10, 8 and 6 respectively. They are employed in order to study the relative sensitivity of the algorithms to increasing overlap between clusters. In the *Sizes* data sets, edge length and standard deviation are kept constant, and, instead, the relative size of the individual clusters is varied. In particular, the ratio between the smallest and the largest cluster on the Sizes1, Sizes3 and Sizes5 data set is 2, 6 and 10 respectively. By this means we investigate the algorithms' sensitivity to unequally-sized clusters.

The last two of our synthetic data sets contain at least one non-spherical shaped cluster, making it difficult for methods based on minimizing variance. The Long1 data set consist of two horizontal, long elliptical Gaussian clusters, one positioned directly above the other (not overlapping), at a very small distance compared with the length of each cluster. The minimum variance solution on this clustering problem splits the two clusters down the middle, producing a result that is very far from the true cluster structure. The last data set, name Smile consists of four equally-sizes clusters – two eyes, a long curved smile, and a circular cluster enclosing them. The density of points in the eyes (which are generated by Normal Distributions) is much higher than in the surround or the smile, which are comparatively far more spread out and are geometrically constructed.

Table 1: Summary of the synthetic data sets. D is the dimensionality, C gives the number of clusters, and N_i gives the number of data elements for cluster i. The test sets are generated by either multidimensional Normal or Uniform Distributions $N(\vec{\mu}, \vec{\sigma})$, where $\vec{\mu}$ is the vector of means and $\vec{\sigma}$ is the vector of the standard deviations. Only for the Smile data set circles $C(\vec{\mu}, r, start..end)$ are additionally used, where $\vec{\mu}$ is the centre of the circle, r is its radius, and $start..end$ described the part of the circle that is actually drawn.

Name	C	N_i	D	Source
Square1	4	4×125	2	$N([0,0],[2,2])$, $N([10,10],[2,2])$, $N([0,10],[2,2])$, $N([10,0],[2,2])$
Square3	4	4×125	2	$N([0,0],[2,2])$, $N([8,8],[2,2])$, $N([0,8],[2,2])$, $N([8,0],[2,2])$
Square5	4	4×125	2	$N([0,0],[2,2])$, $N([6,6],[2,2])$, $N([0,6],[2,2])$, $N([6,0],[2,2])$
Sizes1	4	$200, 100, 100, 100$	2	$N([0,0],[2,2])$, $N([10,10],[2,2])$, $N([0,10],[2,2])$, $N([10,0],[2,2])$
Sizes3	4	$335, 55, 55, 55$	2	$N([0,0],[2,2])$, $N([10,10],[2,2])$, $N([0,10],[2,2])$, $N([10,0],[2,2])$
Sizes5	4	$386, 38, 38, 38$	2	$N([0,0],[2,2])$, $N([10,10],[2,2])$, $N([0,10],[2,2])$, $N([10,0],[2,2])$
Long1	2	2×250	2	$N([-0.3,0.7],[0.01,0.01])$, $N([-0.7,0.7],[0.01,0.01])$
Smile	4	4×125	2	$N([0,0],[1,0.1])$, $N([0,1],[1,0.1])$, $C([-0.5,0.5],0.5,0..2\pi)$, $C([-0.5,0.5],0.3,1.25\pi..1.75\pi)+U([0,0],[0.1,0.1]$

Table 2: Summary of the used real data sets from the UCI Machine Learning Repository. D is the dimensionality, C gives the number of clusters, and N_i gives the number of data elements for cluster i.

Name	C	N	N_i	D	**Type**
Iris	3	150	3×50	4	Continuous
Wine	3	178	59, 71, 48	13	Continuous
Zoo	7	101	41, 20, 5, 13, 4, 8, 10	16	Boolean
Wisconsin	2	699	458, 241	9	Integer
Yeast	10	1484	463, 429, 244, 163, 51 44, 37, 30, 20, 5	8	Continuous
Dermatology	6	366	112, 61, 72, 49, 52, 20	34	Integer
Digits	10	3498	363, 364, 364, 336, 364 335, 336, 364, 336, 336	16	Integer

2 Real data

Table 2 briefly describes the real data sets taken from the UCI Machine Learning Repository. A variety of different benchmarks has been chosen in order to account for different problem sizes and problem structures (such as a differing number of clusters, clusters of varying sizes, high dimensionality etc.).

3 Data Processing

Prior to clustering, the real data is normalised to mean 0 and standard deviation 1 in each dimension.

The distance functions used are as follows: For the synthetic data sets we use the Euclidean distance, which, for two D-dimensional data items i and j, is defined as

$$d_{euclidean}(i,j) = \sqrt{\sum_{k=0}^{D}(i_k - j_k)^2}$$

The advantage of using the Euclidean distance is the straightforward interpretation and visualisation of the data, which facilitates the derivation of

appropriate test sets and the analysis of the results.

For the real data benchmark set, we use a distance function based on the Cosine measure of similarity. It is given as

$$d_{cosine}(i,j) = 1.0 - 0.5 \cdot (1.0 + \frac{\sum_{k=0}^{D} i_k \cdot j_k}{(\sum_{k=0}^{D} i_k \cdot i_k)(\sum_{k=0}^{D} j_k \cdot j_k)})$$

Hence, we compute the Cosine of the two data vectors, translate and scale the resulting value to lie within the interval $[0, 1]$, and finally convert this similarity value to a dissimilarity value by subtracting it from 1.0.