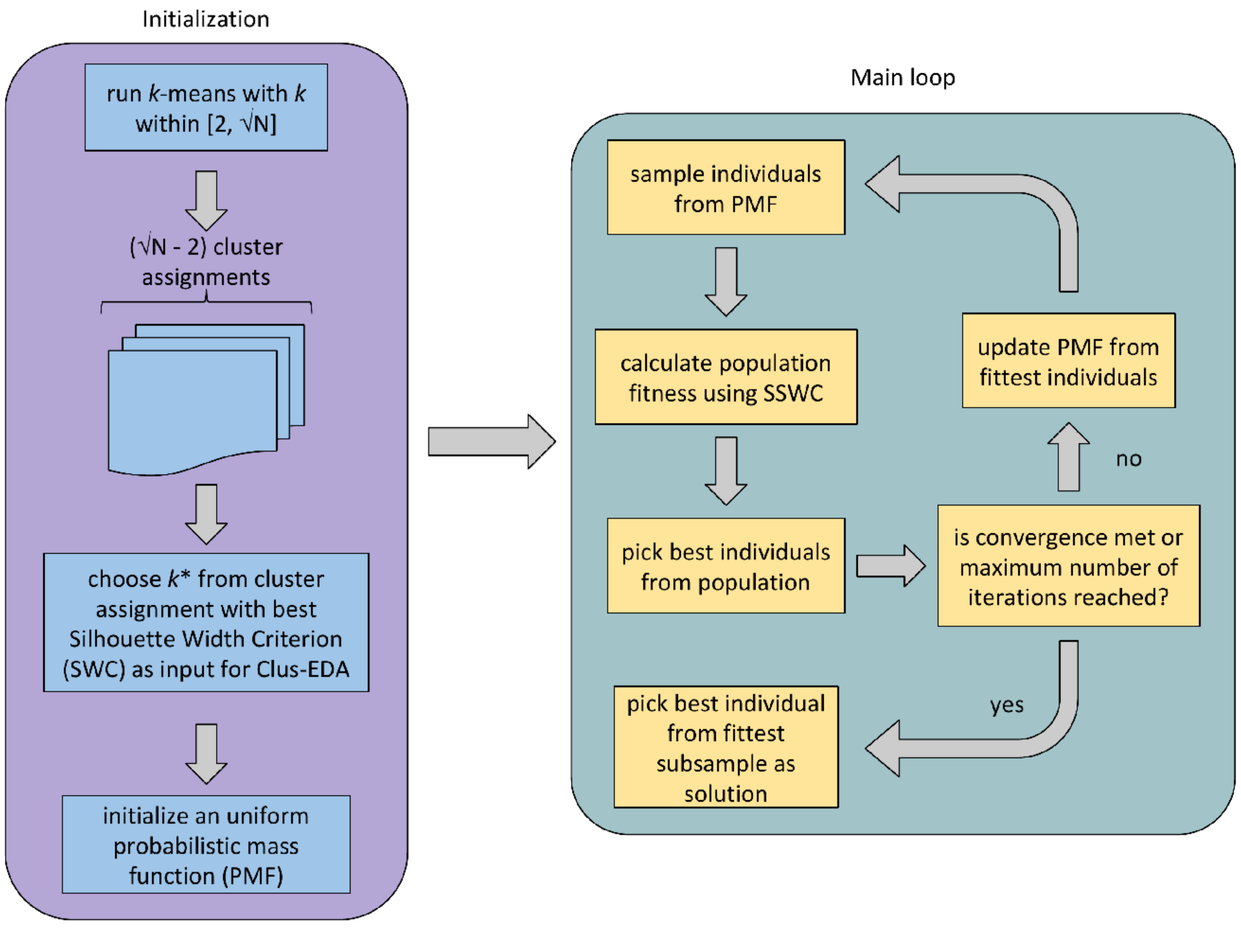
Medoid-based Data Clustering with Estimation of Distribution Algorithms

Henry E. L. Cagnini, Rodrigo C. Barros, Christian V. Quevedo
Faculdade de Informática - Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, RS, Brazil
{henry.cagnini, christian.quevedo}@acad.pucrs.br , rodrigo.barros@pucrs.br

Márcio P. Basgalupp
ICT-UNIFESP - Instituto de Ciência e Tecnologia da Universidade Federal de SP
S. J. dos Campos, SP, Brazil
basgalupp@unifesp.br

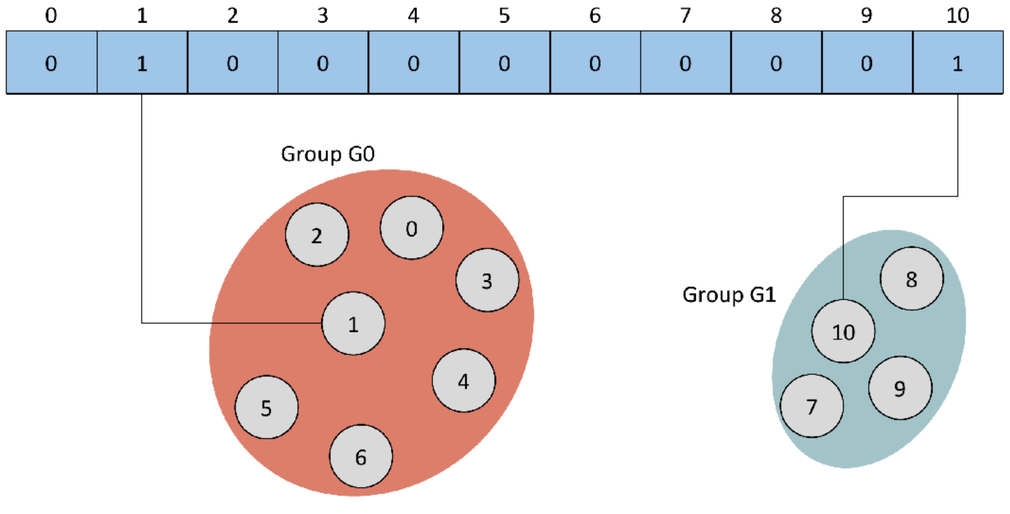
Abstract

Data clustering is the machine learning task that aims at arranging data into groups (clusters) of objects according to a similarity criterion. From an optimisation perspective, it is a particular kind of NP-hard grouping problem, thus attracting much attention from the evolutionary computation community. In this paper, we propose a novel data clustering algorithm based on a univariate estimation of distribution algorithm, namely Clus-EDA. It employs a medoid-based representation in which the cluster prototypes necessarily coincide with objects from the dataset. We compare Clus-EDA with both traditional non-evolutionary clustering algorithms such as k-means and hierarchical agglomerative clustering, and also with an evolutionary algorithm for clustering, in artificial and synthetic datasets. Our results show that Clus-EDA often outperforms the baseline algorithms with regard to distinct cluster validity criteria.



Individuals

An individual is a partition produced by the sampling process: once the objects which will be medoids are chosen, non-medoid objects are assigned to the group of its closest medoid.



Fitness function

We use the Simplified Silhouette Width Criterion (SSWC) [1] as fitness function for assigning how good each individual (solution) is:

$$SSWC = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{b(i),a(i)\}}$$

where b(i) is the distance to the closest different-cluster medoid and a(i) the distance to the closest same-cluster medoid.

Probabilistic Mass Function

The probability of an object becoming a medoid is initially uniform over all objects, and is set to k^*/N: k^* is the number of groups from the partition generated by k-means with best SSWC, and N the number of objects.

	object 0	1	2	3	4	5	...	N
chance of becoming medoid:	k*/N	k*/N	k*/N	k*/N	k*/N	k*/N	k*/N	k*/N
chance of not becoming medoid:	1 - (k*/N)	1 - (k*/N)	1 - (k*/N)	1 - (k*/N)	1 - (k*/N)	1 - (k*/N)	1 - (k*/N)	1 - (k*/N)

Baseline Algorithms

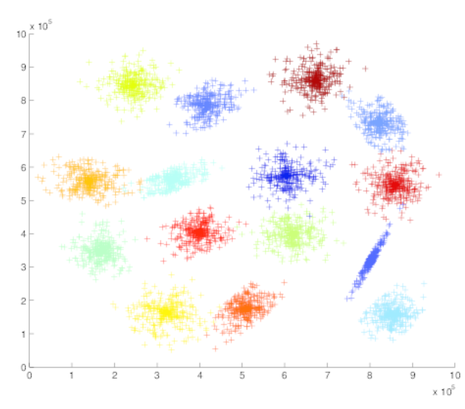
We compare our algorithm with another three: k-means [3], UPGMA [4], and F-EAC [5], which is a mutation-based EA (no crossover is performed whatsoever), with specialised mutation operators for the clustering task.

Evaluation Measures

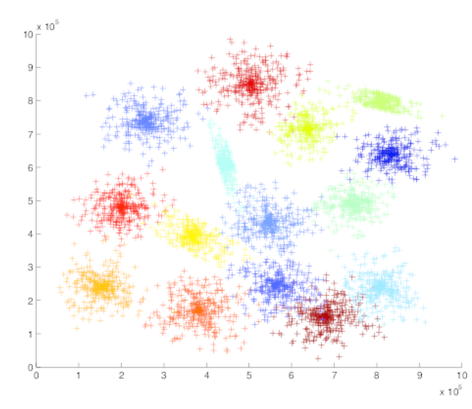
We make use of three measures for validating the quality of produced partitions: Silhouette Width Criterion, [2] Davies-Bouldin Index [8] and Adjusted Rand Index [9].

Datasets

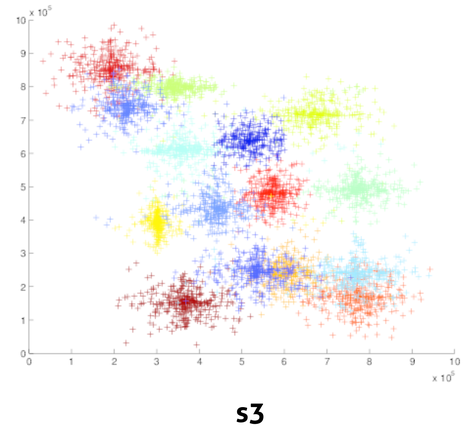
dataset	attributes	objects	groups	source
s1, s2, s3, s4	2	5000	15	[6]
sin1, sin2, sin3, sin4, sin5	20	400	6	[7]



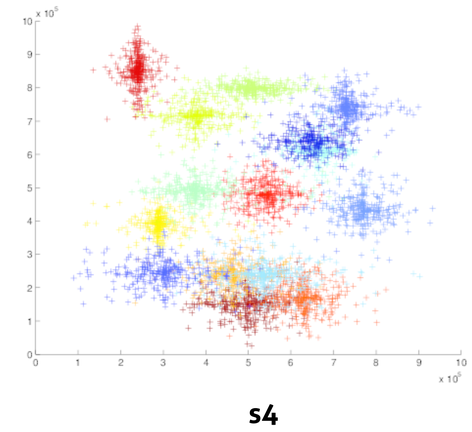
s1



s2



s3



s4

		k-means				UPGMA				F-EAC				Clus-EDA			
Dataset	k	$k*$	SWC	DB	ARI	$k*$	SWC	DB	ARI	$k*$	SWC	DB	ARI	$k*$	SWC	DB	ARI
s1	15	16.00	0.63	0.61	0.90	19.00	0.51	0.63	0.85	15.00	**0.71**	0.46	0.87	15.07	**0.71**	**0.42**	**0.99**
s2	15	14.00	0.61	**0.48**	0.89	15.00	0.52	0.68	0.91	15.00	**0.63**	0.57	0.87	15.07	0.62	0.53	**0.93**
s3	15	14.00	0.41	**0.69**	0.62	15.00	0.19	0.91	0.69	15.00	**0.49**	0.76	**0.86**	14.73	**0.49**	0.70	0.73
s4	15	17.00	0.47	**0.68**	0.64	18.00	0.10	0.98	0.61	15.00	**0.48**	0.77	**0.85**	15.20	0.47	0.73	0.65
sin1	6	5.00	0.63	0.53	0.67	6.00	**0.65**	0.71	**0.84**	6.00	**0.65**	0.60	0.67	6.00	**0.65**	**0.52**	**0.84**
sin2	6	6.00	0.54	1.11	0.43	5.00	**0.69**	0.62	**0.67**	5.00	**0.69**	0.48	0.55	5.00	**0.69**	**0.43**	**0.67**
sin3	6	4.00	0.45	0.89	0.44	5.00	**0.73**	0.47	**0.54**	4.00	0.72	0.44	0.44	4.00	0.72	**0.43**	0.54
sin4	6	6.00	0.51	1.06	0.64	8.00	**0.70**	0.91	0.83	6.93	0.69	0.62	0.67	6.00	0.69	**0.43**	**0.84**
sin5	6	6.00	0.55	0.78	0.65	5.00	**0.64**	0.70	0.67	8.00	**0.64**	0.57	0.67	6.00	0.63	**0.56**	**0.83**

References

[1] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro. Evolving clusters in gene-expression data. Information Sciences, 176(13):1898 - 1927, 2006.
[2] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(0):53 - 65, 1987.
[3] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281-297. California, USA, 1967.
[4] L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. John Willey & Sons, 1990.
[5] M. C. Naldi, R. J. G. B. Campello, E. R. Hruschka, and A. C. P. L. F. de Carvalho. Efficiency issues of evolutionary k-means. Applied Soft Computing Journal, 11(2):1938{1952, Mar. 2011.
[6] P. Fränti and O. Virmajoki. Iterative shrinking method for clustering problems. Pattern Recognition, 39(5):761 - 775, 2006.
[7] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner. Clustering gene-expression data with repeated measurements. Genome Biology, 4(5):R34, Apr. 2003.
[8] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2):224 - 227, 1979.
[9] Vendramin, L, Campelo and R.J, Hruschka, E.R. Relative clustering validity criteria: A comparative Overview. Statistical Analysis and Data Mining: The ASA Data Science Journal, vol.3-4, pp. 209-235, 2010.