$MO - Mine_{clust}$: A framework for multi-objective Clustering

Benjamin Fisset[1], Clarisse Dhaenens[1,2] and Laetitia Jourdan[1,2]

[1] Inria Lille - Nord Europe, DOLPHIN Project-team, 59650 Villeneuve dAscq, France
[2] Université Lille 1, LIFL, UMR CNRS 8022, 59650 Villeneuve dAscq cedex, France
benjamin.fisset@inria.fr, clarisse.dhaenens@lifl.fr,
laetitia.jourdan@lifl.fr

1 Clustering: A Multi-objective combinatorial problem

Clustering is a very common and popular datamining technique. In a context where data are described by a set of variables, clustering algorithms group similar data in clusters. Thus, elements in one cluster are similar among them and different from elements of the other clusters. This problem may be seen as a combinatorial optimization problem as soon as a criterion able to evaluate the quality of a given clustering can be found. In the literature, many criteria have been proposed and multi-objective models have been adopted. A clustering solution, that will assign each element to a given cluster, is considered good when elements of each cluster are very similar among them (low intra-cluster variance) and very different from the elements of the other clusters (high inter-cluster variance). This problem is by nature a multi-objective one. The multi-objective clustering allows to find a solution by using multi-objective approaches. A pareto approach evaluates each objective simultaneously for each clustering solution. As a result, the clustering solutions are stored in a collection of solutions where each one represents a different trade-off among objectives (Pareto front). The aim of this paper is to present $MO - Mine_{clust}$, a framework dedicated to multi-objective clustering. The genericity of this framework allows to adopt different models, taking into account several combinations of optimization criteria.

2 Experiments and Discussion

Regarding the numerous models (and in particular combinations of objective functions) able to deal with the multi-objective clustering, the objective of the proposed platform, $MO - Mine_{clust}$, is to identify the best combination of model/engine/parameters to a particular dataset in order to offer, to a non specialist, the ability of discovering the best clustering for his/her dataset. Such an approach requires the implementation of several components that can be combined.

Table 1 presents results obtained by the platform $MO - Mine_{clust}$ on two types of experimental data : handcrafted two-dimensional data sets and generated dataset [3]. Results obtained by the well-known Kmeans algorithm, as well as those obtained by the state-of-the-art MOCK (See in [2]) algorithm based on PESA II (See in [1]) are presented. The number of clusters computed as well as the average quality (in term of Adjusted Rand Index (ARI)) are presented and its standard deviation (Std.) is also indicated. For $MO-Mine_{clust}$. The average value of the ARI is an external criterion to select the best solution among the pareto front generated by each algorithm but it is not used within the algorithms as an optimization criterion.

Table 1. Compare performance between $MO - Mine_{clust}$ and MOCK.

Data sets	MOCK [2]		Kmeans		$MO - Mine_{clust}$		
Name	k	Av. ARI	k	Av. ARI	k	Av. ARI	Std
Square1	4.22	0.9622	4	0.9651	19.6	**0.9901**	0.013
Square4	4.32	0.7729	4	0.8048	4.4	**0.8196**	0.0225
Sizes5	4.2	0.976	3.92	0.9557	37.8	**0.9838**	0.005
Long1	2	**0.9998**	4.98	0.3562	2	**0.9998**	0.0001
Spiral	2	**1**	5.12	0.5502	2	**1**	0
2d-4c	4.12	**0.9893**	3.99	0.9143	4 .2	0.988	0.0002
2d-20c	19.94	0.9454	33.79	0.8633	19.2	**0.9832**	0.009
2d-40c	42.14	0.8654	42.36	0.692	19.4	**0.9835**	0.034
10d-4c	4.07	0.9962	3.99	0.9704	4.2	**0.9975**	0.001
10d-20c	20.26	**0.9981**	21.45	0.9820	20.2	0.9979	0.004
10d-40c	42.84	**0.9896**	43.48	0.9678	19.2	0.9859	0.01

In this table, we can observe that for the majority of the datasets, our approach improves the average ARI. The average relative percentage deviation on the other datasets is less than 0.1%. With a friedman test, we observe that the algorithms are different with a p-value of 0.001. Concerning the comparison between MOCK and $MO-Mine_{clust}$ the difference is also statistically significative and $MO - Mine_{clust}$ performs better in term of ARI.
Such an approach can be applied to all the classical knowledge extraction task.

References

1. D. Corne, N.R. Jerram, J. Knowles, and M. J. Oates. Pesa-II: Region-based selection in evolutionary multiobjective optimization. In *GECCO2001*, pages 283–290. Morgan Kaufmann Publishers, 2001.
2. J. Handl and J. D. Knowles. An evolutionary approach to multiobjective clustering. *IEEE Trans. Evolutionary Computation*, 11(1):56–76, 2007.

[3] http://personalpages.manchester.ac.uk/mbs/Julia.Handl/mock.html