Meta-learning System for Automated Clustering

Sergey Muravyov and Andrey Filchenkov

ITMO University, Computer Technology Lab,
Kronverksky pr. 49, 197101 St.Petersburg, Russia
{smuravyov,afilchenkov}@corp.ifmo.ru

Introduction. Clustering is the most common unsupervised learning task. The number of the clustering algorithms is constantly increasing, which raises a problem of clustering algorithm selection. Selection of proper features for clustering is also important, and several algorithms are designed to solve this problem. Thus, the question arises if we can simultaneously choose both feature selection (FS) algorithm and clustering algorithm given a dataset. To make it even more complicated, no standard method to assess clustering quality exists. To cope with that four approaches for clustering quality assessment may be distinguished. In [1], a method for clustering quality evaluation based on human assessments was proposed. It was also shown there that none of existing cluster validity index (CVI) suits for all problems. Thus, another question arises if we can select a proper CVI given a dataset.

This paper aims to answer both questions above. We propose to use a system consisting of two parts. The first part predicts the best CVI given a dataset using approach presented in [1]. The second part predicts the best clustering algorithm and the best FS algorithm for a given dataset and a clustering performance measure. The second part does not depend on a performance measure and can work without the first part.

Related works. In comparison to supervised learning tasks, only a few papers are devoted to creating meta-learning systems for clustering. One of the latest is [2], the authors of which suggested a set of meta-features that are used for selection of a clustering algorithm based on an average ranking of several clustering validity indexes based on inter-/intra-cluster measures. In earlier papers [3–5], the authors used external efficacy measures, which is not applicable in real life. We found no other works on this topic, neither did the authors of [6].

Meta-learning system for clustering. We propose a nested system consisting of three components. The first component **CVIsel** predicts the best CVI for a given dataset. The second component **Clusel** predicts the best clustering algorithm for a given dataset and a performance measure. The last component **FSsel** predicts the best FS algorithm given a dataset, a performance measure and a clustering algorithm.

CVIsel component is taught using training datasets labelled according to the approach proposed in [1]. The main idea is to ask human assessors to evaluate several partitions of a dataset, to rank them, and to compare with how each CVI ranks these partitions. The more similar resulting ranks are the better CVI is. Another approach is to ask human assessors to simply mark partitions as adequate or not and then to evaluate how high CVI higher inadequate partitions.

CVIsel uses the labeled set of dataset to learn a meta-classifier that predicts the best CVI for a given dataset and it predicts if this CVI is adequate or not.

Clusel component is learned using training datasets labelled according to performance of clustering algorithms with respect to a chosen performance measure. It works in the same way as **CVIsel** does.

FSsel component works in a slightly different way. Application of a FS algorithm produces a new dataset that may have other properties than the original one. This means that if we also chose from several CVIs, the new dataset may have another best CVI and another best clustering algorithm. If two different FS algorithms produce two datasets that have different best CVIs then these two algorithms are incomparable. This is why we evaluate each FS algorithm for each CVI. **FSsel** uses the labeled set of datasets and prediction of adequacy of each CVI for resulting dataset to predict the best FSsel with respect to each CVI predicted to be adequate.

We 19 meta-features for clustering that are based on distances between objects proposed in [2]. We also generated 60 landmarks obtained by running of 12 different clustering algorithms with different parameters by each CVI. We applied FS for each described meta-classifier resulting into various meta-features for each of them.

Experiments. We took 200 real clustering dataset from different sources, including UCI[1] and KEEL[2]. We choose the five best CVIs in terms of their rank from the 19 that were compared in [1]. These CVIs are: OS-index, Symmetric index, GD41, GD33 and GD43. We used six clustering algorithms: k-Means [7], X-Means [8], EM [9], DBSCAN [10], FarthestFirst [11], and Hierarchical [12]. We used four FS algorithms for clustering [13]: spectral FS; Laplasian score, localized FS based on scatter separability, and multi-cluster FS.

We used F_1-measure and leave-one-out cross-validation to estimate meta-classifier performance. The results are presented in Table 1. As we can see, each component has achieved more than 80% of F_1-measure for multi-class classification tasks using Random Forest.

Table 1. Values of F-measure of different components for predicting CVI, clustering algorithm and mean values for predicting FS algorithm

Component	kNN	Multilayer Perceptron	Naive Bayes	Bayesian Network	**Random Forest**	Random Trees
CVIsel	.680	.690	.700	.725	**.855**	.805
Clusel	.645	.660	.715	.730	**.805**	.775
FSsel (average)	.590	.611	.690	.797	**.818**	.808

Conclusion. We proposed a full model selection system for clustering. The system contains three components that predict appropriate CVI, clustering al-

[1] `https://archive.ics.uci.edu/ml/datasets.html`
[2] `http://sci2s.ugr.es/keel/datasets.php`

gorithm and FS algorithm for a given unlabeled dataset. The system was tested on real life dataset and got satisfactory results. We plan to improve the accuracy of the system and explore if we can overcome the constraint on FS algorithm incomparability.

Acknowledgments. This work was financially supported The Russian Science Foundation, Agreement 17-71-30029, and Russian Foundation for Basic Research, Grant 16-37-60115 mol_a_dk and by The Russian Science Foundation, Agreement 17-71-30029.

References

1. Filchenkov, A., Muravyov, S., Parfenov, V.: Towards cluster validity index evaluation and selection. In: Artificial Intelligence and Natural Language Conference (AINL), IEEE. pp. 1–8. IEEE (2016)
2. Ferrari, D.G., de Castro, L.N.: Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. Information Sciences 301, 181–194 (2015)
3. de Souto, M., Prudêncio, R., Soares, R., de Araujo, D., Costa, I., Ludermir, T., Schliep, A.: Ranking and selecting clustering algorithms using a meta-learning approach. In: Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. pp. 3729–3735 (June 2008)
4. Soares, R.G., Ludermir, T.B., De Carvalho, F.A.: An analysis of meta-learning techniques for ranking clustering algorithms applied to artificial data. In: Artificial Neural Networks–ICANN 2009, pp. 131–140. Springer (2009)
5. Ferrari, D.G., de Castro, L.N.: Clustering algorithm recommendation: a meta-learning approach. In: Swarm, Evolutionary, and Memetic Computing, pp. 143–150. Springer (2012)
6. Van Craenendonck, T., Blockeel, H.: Using internal validity measures to compare clustering algorithms. In: AutoML Workshop at ICML 2015. pp. 1–8 (2015)
7. Vassilvitskii, S., Arthur, D.: k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035 (2006)
8. Pelleg, D., Moore, A.W., et al.: X-means: Extending k-means with efficient estimation of the number of clusters. In: ICML. pp. 727–734 (2000)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)
10. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. vol. 96, p. 226231 (1996)
11. Rosenkrantz, D.J., Stearns, R.E., Lewis, II, P.M.: An analysis of several heuristics for the traveling salesman problem. SIAM journal on computing 6(3), 563–581 (1977)
12. Rokach, L., Maimon, O.: Clustering methods. In: Data mining and knowledge discovery handbook, pp. 321–352. Springer (2005)
13. Alelyani, S., Tang, J., Liu, H.: Feature selection for clustering: A review. Data Clustering: Algorithms and Applications 29, 30–55 (2013)