Cascade evaluation

Laurent Candillier[1,2], Isabelle Tellier[1], Fabien Torre[1], Olivier Bousquet[2]
[1] GRAppA - Charles de Gaulle University - Lille 3
candillier@grappa.univ-lille3.fr
[2] Pertinence - 32 rue des Jeûneurs -75002 Paris
olivier.bousquet@pertinence.com

The evaluation of the relevance of clustering algorithms is still an open issue. The main problem is that evaluating clustering results is subjective by nature. Indeed, there are often many relevant reasons to group together some given data objects. In practice, there are four main ways to measure the quality of clustering algorithms. But each of these methods has its limitations.

1. Use artificial datasets where the desired grouping is known. But the algorithms are thus evaluated only on the corresponding generated distributions, and results on artificial data can not necessarily be generalized to real data.

2. Use supervised databases, ignore the class labels for clustering, and check if the clusters found gather data points that belong to the same initial classes. But the classes of a supervised problem are not necessarily the classes that have to be found by a clustering algorithm because other grouping can also be meaningful.

3. Work with an expert that will evaluate the meaning of the clustering in a particular field. However, if it is possible for an expert to tell if a particular clustering has some meaning, it is much harder to tell if a given result is better than another one, and the interest of the method can not be generalized to various types of data.

4. Use some internal criterion, like the intra-cluster inertia and/or the inter-clusters separation. But such pre-defined criteria are also subjective by nature because they use some pre-defined notion of what is a good clustering. For example clusters separation is not always the best criterion to use: clusters that overlap may sometimes be more relevant.

In fact, what we want to evaluate is how well a given clustering algorithm is able to capture interesting, meaningful and usable information. Meaningful information in that case correspond to a new knowledge that is interesting to use for some purpose. We also expect the algorithm to be able to capture such interesting information on various types of datasets.

Based on these considerations, we propose a new methodolgy for clustering evaluation and clustering comparison. The main idea is to use the clustering results to enrich a given dataset, and check if this extra-knowledge is somewhat useful for the apprehension of the dataset. In particular, we conjecture that, if the results of a given supervised learning algorithm are improved when new information are added to a dataset, from the results of a clustering process, then it means that such information were somewhat useful. And thus, the clustering results can be regarded as relevant. Figure 1 summarizes the main steps of our proposed methodology.

An example of information that can be added from clustering results is the membership of the data points to the clusters, if the output of the clustering is a partition of the dataset. Another possible way to add information from clustering results is to associate to each data point a set of attributes representing the center of the cluster it belongs to.

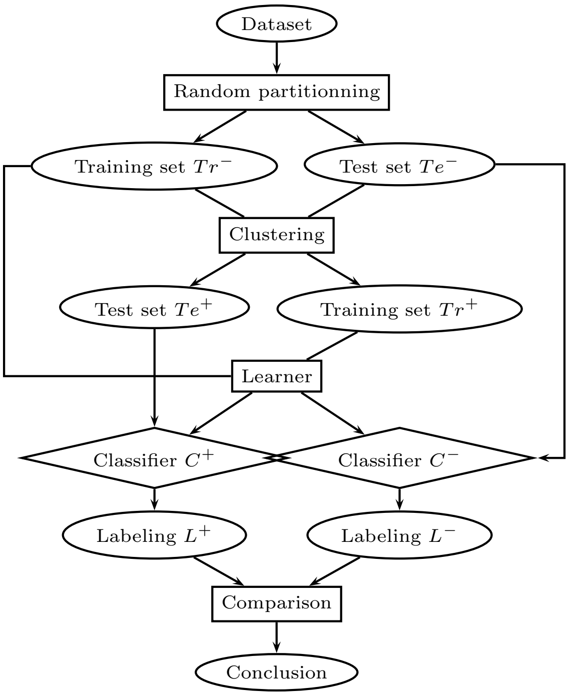

Figure 1: Methodology for evaluating the interest of adding to a dataset new information coming from a clustering algorithm.

This research is inspired by the works on classifier combinations, and in particular *cascade generalization* [Gama and Brazdil, 2000]. We conjecture that clustering algorithms can help supervised learners to specialize their treatments according to different specific areas in the input space. They can also help supervised learners fit more complex decision surfaces.

To evaluate the improvement in the results with or without the new information coming from the clustering process, we propose to test both methods on various independent databases. On each database, five 2-fold cross-validations can be performed, as proposed in [Dietterich, 1998]. Then various measures can be used to compare the results of both methods and evaluate if the new information coming from the clustering significantly improve the supervised learner.

To perform these comparisons, C4.5 [Quinlan, 1993] as the supervised learning algorithm is well suited because, as it performs feature selection during the learning process, and provides as output a tree where we can observe which features were selected, it is able to clearly point out whether or not the new information were helpful. It has also the advantage of being fast and to be able to manage discrete features, as well as continuous ones.

Our first experiments in the use of such a methodology to compare different clustering algorithms were conducted on various numerical databases coming from the UCI Machine Learning Repository [Blake and Merz, 1998]. We thus point out that clustering methods based on the use of probabilistic models outperform K-means based clustering methods. This result is not surprising since the models used by the former methods are richer than those used by K-means based methods. But we can thus observe that our method exhibits coherent results.

Besides, during our experiments, we observe that most of the time, the extra-knowledge added from the clustering results improves the results of C4.5. So our work also opens a new field of investigations concerning the improvement of the results of supervised learning algorithms by their combination with unsupervised learning methods.

References

[Blake and Merz, 1998] Blake, C. and Merz, C. (1998). UCI repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html].

[Dietterich, 1998] Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

[Gama and Brazdil, 2000] Gama, J. and Brazdil, P. (2000). Cascade generalization. *Machine Learning*, 41(3):315–343.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* KAUFM.