Cascade Evaluation of Clustering Algorithms

Laurent Candillier[1, 2], Isabelle Tellier[1], Fabien Torre[1], Olivier Bousquet[2]

Problem: evaluate clustering algorithms is difficult because many different relevant grouping of the data may exist for a given dataset

- use artificial data : evaluation only on the generated distributions, no generalization to real data
- use labeled datasets : other groupings can be more meaningful
- use an expert : no comparison possible, no generalization to other data
- use some internal criteria : pre-defined notions of the interest of clustering

Proposition: consider clustering as a pre-processing step for another task that we are able to evaluate

1. supervised learning on a labeled dataset
2. supervised learning on the same dataset enriched by clustering
   - clustering on the dataset without using the classes information
   - enrich the dataset from the results of clustering
   - supervised learning on the enriched dataset
3. compare the results of both learned classifiers

Base: *if the results of a supervised learning algorithm are improved when some extra-knowledge coming from clustering is added, then it means that clustering managed to capture some new meaningful and useful knowledge*

Experiments :

- 2 enrichment methods : add new attributes to the dataset or divide the dataset into different groups
- 4 supervised algorithms : C4.5, C5 boosted, DLG and SVM
- 5 clustering algorithms of increasing complexity : Random, K-means, LAC, SSC and SuSE
- 10 datasets coming from UCI Machine Learning Repository
- 5 two-fold cross validations and 5x2cv F-test to check the significance of the improvements
- 4 comparison measures : number of wins of the supervised algorithm when it is (or not) provided with information coming from clustering, number of significant wins, wilcoxon signed rank test (check the significance of the improvement over the different datasets), mean balanced error rate

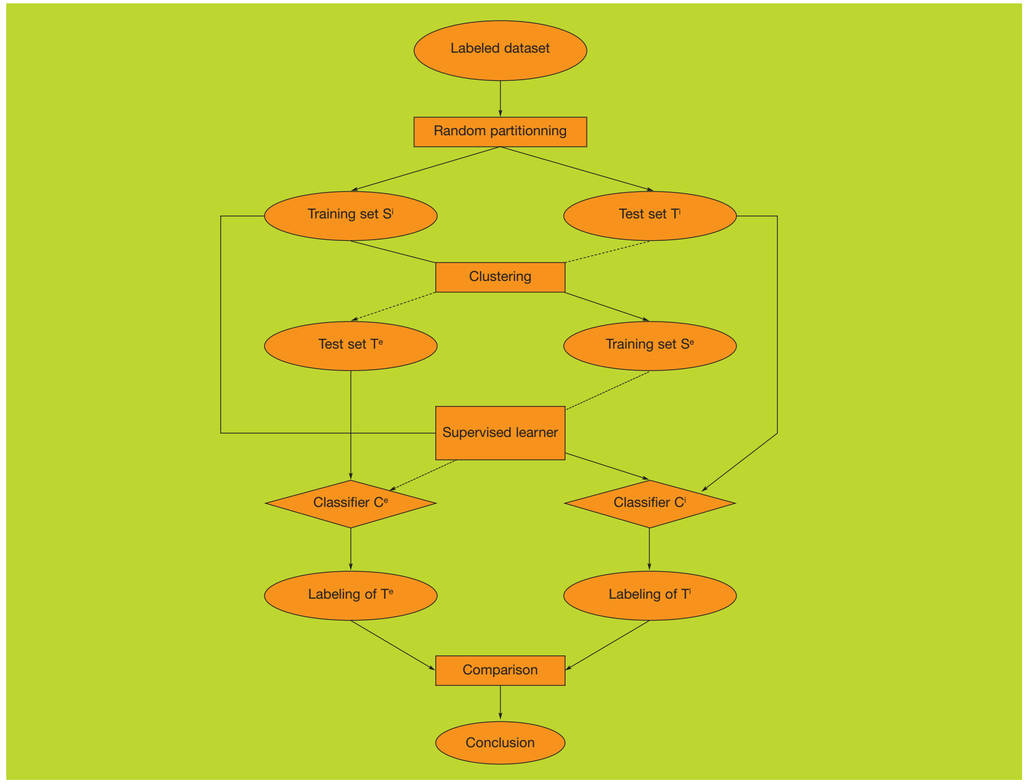
Results :

- the order in which the clustering methods are ranked remains the same no matter which supervised algorithm and which enrichment method are used
- clustering methods based on more complex models outperform methods based on simpler models

Conclusion :

- new objective and quantitative evaluation method that provides coherent results
- does not only evaluate the mapping between the class labels and the cluster labels

[1] Mostrare Project (INRIA Futurs, LIFL, GRAppA, Lille 3)
[2] Pertinence (Paris)



	C4.5 alone	C4.5 + Random	C4.5 + K-means	C4.5 + LAC	C4.5 + SSC	**C4.5 + SuSE**
number of wins	-	1/9	5/4	7/3	9/1	**9/1**
number of significant wins	-	0/1	0/0	1/0	2/0	**3/0**
wilcoxon signed rank test	-	-2,67	-0,05	1,31	1,83	**2,56**
mean balanced error rate	21,3%	24,3%	20,6%	20%	19,3%	**18,5%**

pertinence simple solutions for complex processes

lille 3 UNIVERSITE CHARLES-DE-GAULLE

lifL

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA FUTURS

http://www.grappa.univ-lille3.fr/~candillier