Theoretical foundations of clustering

Cascade evaluation

L. Candillier[1,2], I. Tellier[1], F. Torre[1], O. Bousquet[2]
[1]GRAppA, Lille 3 University / [2]Pertinence, Paris

1. The evaluation problem
   - Proposition
   - Methodology
2. Experiments
3. Conclusion

I. The evaluation problem

Difficult to evaluate clustering results : there may be many relevant and different way to group together some given data objects

Existing methods

- artificial datasets : specific generated distributions, no generalization to real data
- supervised datasets : other relevant grouping may be possible
- expert : no comparison possible, no generalization to other datasets
- internal criteria : predefined notion of what is a good clustering (distance)

Proposition

The goal of clustering is to help to apprehend a given dataset : add new and useful information

$\Rightarrow$ consider a dataset with classes information

$\Rightarrow$ enrich the dataset with new information coming from the clustering results

$\Rightarrow$ measure if this new information help to improve the results of a supervised algorithm

Questions

1. do the information captured by clustering algorithms improve the results of supervised algorithms ?
2. which information shall we transmit from clustering to supervised algorithm ?
3. does the improvement give us a way to evaluate the clustering results ?

Cascade evaluation

Inspired by [Gama and Brazdil, 2000]

Being given a dataset with classes information

1. learning 1
   - supervised learning on the initial dataset
2. learning 2
   - clustering on the dataset without using the classes information
   - create new attributes from the results of clustering
   - add these new attributes to the initial dataset
   - use this new dataset for supervised learning
3. compare the results of both supervised learning

Evaluation methodology

- Attributes added to a data point
  - associated cluster (categorical)
  - center of associated cluster (categorical/numerical)
  - weights on dimensions for the associated cluster (numerical)
- Change the parameters of the algorithm. ex : $K \in [2..10]$
- C4.5 as the supervised learner : gives a way to evaluate the importance of the new attributes, fast, able to manage categorical and numerical attributes

Evaluation methodology

- tests on different datasets
- 5 cross-validations with dataset cut into 2 equal parts
- compute the balanced error rate of the supervised algorithm with and without the information added from the clustering
- number of wins of each
- number of significant wins of each (5x2cv [Dietterich, 1998])
- wilcoxon signed rank test : do the differences be significant on the set of problems ?
- mean balanced error rate

II. Experiments

Numerical datasets of UCI repository [Blake and Merz, 1998]

Comparing clusterings

- K-means
- LAC [Domeniconi et al., 2004]
- SSC = EM with the assumption that the dimensions follow gaussian independent distributions [Candillier et al., 2005]
- SuSE = SSC + hard feature selection

$K \in [2..10]$

Comparing clusterings : error rate

	C4.5 alone	C4.5 + K-means	C4.5 + LAC	C4.5 + SSC	C4.5 + SuSE
glass	0.326	0.357	0.370	0.404	0.349
iono	0.141	0.142	0.131	0.098	0.112
iris	0.073	0.067	0.037	0.051	0.047
pima	0.310	0.321	0.321	0.308	0.300
sonar	0.310	0.300	0.288	0.288	0.272
vowel	0.295	0.250	0.264	0.241	0.222
wdbc	0.059	0.046	0.039	0.051	0.031
wine	0.087	0.104	0.096	0.027	0.036

Significance of the improvement

1 - pvalue associated to the 5x2cv-F test [Alpaydin, 1999]

	C4.5 + K-means	C4.5 + LAC	C4.5 + SSC	C4.5 + SuSE
glass	0.33	0.57	0.24	0.33
iono	0.32	0.62	0.02	0.09
iris	0.81	0.65	0.43	0.22
pima	0.43	0.50	0.53	0.27
sonar	0.57	0.39	0.33	0.09
vowel	0.33	0.44	0.23	0.03
wdbc	0.25	0.04	0.63	0.02
wine	0.55	0.60	0.01	0.01

Summary

	C4.5 alone	C4.5 + Kmeans	C4.5 + LAC	C4.5 + SSC	C4.5 + SuSE
no wins	-	4/4	5/3	7/1	7/1
sign wins	-	0/0	1/0	2/0	3/0
wilcoxon	-	0	0.84	1.40	2.24
av perf	0.200	0.198	0.193	0.183	0.171

Conclusion

New evaluation method for clustering algorithms

- objective and quantitative test of relevance
- allows to evaluate a given result, a given method, or to compare various ones
- improves supervised learning

Future work

- complementarity between algorithms : ex : C4.5 and SuSE : allows a test on various attributes at one node of the tree
- influence of the supervised algorithm to compare clustering algorithms ?
- which information to add from the results of clustering ?

Alpaydin, E. (1999).
Combined 5x2cv F-test for comparing supervised classification learning algorithms.
Neural Computation, 11(8) :1885–1892.

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

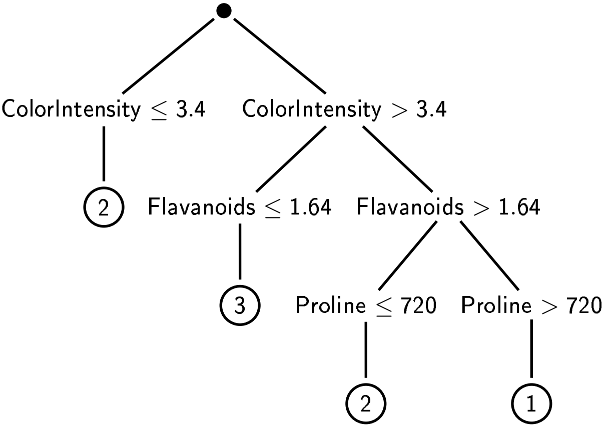
Candillier, L., Tellier, I., Torre, F., and Bousquet, O. (2005).
SSC : Statistical Subspace Clustering.
In Perner, P., editor, Machine Learning and Data Mining in Pattern Recognition, LNCS, pages 100–109, Leipzig, Germany. Springer Verlag.

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

Domeniconi, C., Papadopoulos, D., Gunopolos, D., and Ma, S. (2004).
Subspace clustering of high dimensional data.
In SIAM Int. Conf. on Data Mining.

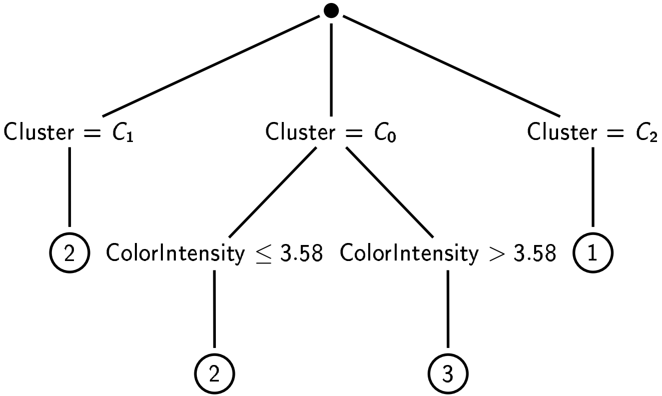
Gama, J. and Brazdil, P. (2000).
Cascade generalization.
Machine Learning, 41(3) :315–343.

C4.5 on wine



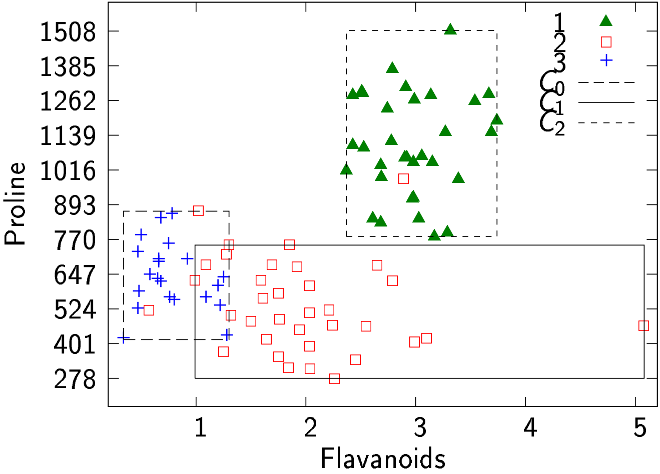
(178 data points, 13 attributes, 3 classes)

C4.5+SuSE on wine



$K = 3$

SuSE on wine for K=3



Diminution of the error on wine

method	C4.5	C4.5+SuSE
total number of errors	5	3
number of errors between classes 1 and 2	1	1
number of errors between classes 2 and 3	4	2

Kind of improvement

- allow supervised learning algorithms to specialize their treatments according to specific areas in the input space
- add new attributes of higher level
- allow to fit more complex decision surfaces

Other information added

- bounds of the rule of the associated cluster
- membership probability to the clusters
- membership probability to the associated cluster
- binary version of the membership
- binary version of the relevance of the dimensions
- only information for the best number of clusters (BIC)
- many mix