The containt of this work will be presented at the conference:
"Model-Based Clustering and Classification (MBC2)", Catania, September 2016

An abstract is just given below... in this page

Evaluation of predictive clustering quality

Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols

Predictive clustering [1] is a new supervised learning framework derived from traditional clustering. These algorithms start by identifying pure clusters (in terms of classes) that have a high probability density. Based on the information given by the clusters, these algorithms can predict the class of new instances. Compared to supervised classification, predictive clustering can discover the internal structure of the target class. It thus allows users to find the different reasons behind the same prediction: two heterogeneous instances could have the same predicted label.

By its nature, predictive clustering incorporates the characteristics of both supervised classification and clustering. Thus, in the evaluation of predictive clustering results, three points should be taken into account: a high intra-cluster similarity, a low inter-cluster similarity and a good prediction rate. A predictive clustering quality criterion must balance these three points.

In this this work, we propose a new criterion for measuring the predictive clustering quality. This criterion calculates the compactness and the separability of clusters using a new supervised similarity measure. This measure exploits the information given by the target class in such way that two instances are considered similar if and only if a distance between them is small and they belong to the same class. And, they are considered heterogeneous if and only if a distance between them is large and they belong to different classes.

The obtained results from different simulated datasets show that the proposed criterion constantly gives the optimal number of clusters. To our knowledge, there is no analytic criterion in the state of the art that is able to measure the quality of the results generated by predictive clustering algorithms (the trade-off mentioned above) and therefore to compare with our suggested criterion. So, to compare our results, we use the well know unsupervised criterion (Davies-Bouldin) [2] and two supervised criteria (Adjusted Rand Index [3] and Variation of Information [4]) and we examine if our criterion find the good tradeoff.

Reference

[1] Bernard Z., Sašo D., Jan S., **Learning Predictive Clustering Rules**, In 4th International Workshop, KDID 2005, pp 234-250, 2006

[2] D. L. Davies and D. W. Bouldin: **Cluster Separation Measure**, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, No. 2, pp. 95-104, 1979

[3] Hubert, L. and Arabie, P: **Comparing partitions**. Journal of Classification, pp. 193–218, 1985

[4] Meila, Marina: **Comparing Clusterings by the Variation of Information**. Learning Theory and Kernel Machines, pp. 173–187, 2003