An Experimental Study of Partition Quality Indices in Clustering

Céline Robardet[1], Fabien Feschet[1], and Nicolas Nicoloyannis[2]

[1] LASS - UMR 5823 - bât 101 - Université Lyon 1 - Villeurbanne France
[2] ERIC - bât L - Université Lyon 2 - Bron France

Abstract. We present a preliminary study to define a comparison protocol to evaluate different quality measures used in supervised and unsupervised clustering as objective functions. We first define an order on the set of partitions to capture the common notion of a good partition towards the knowing of the ideal one. We demonstrate the efficiency of this approach by providing several experiments.

keywords. Unsupervised clustering, partitions ordering, partition quality indices.

1 Introduction

Unsupervised clustering aims at organizing a data set by grouping objects into clusters to discover at best their relations. This clustering can conduct to overlapping groups, however it is simpler to look for partitions. The clustering should gather the couple of objects the most similar and should separate the couple the most dissimilar. Most clustering algorithms can be described through the concept of similarity and the optimization procedure. A lot of algorithms [8,2] use an objective function to operationally express a compromise between the intra-cluster proximity and the inter-cluster farness. It is also possible to use the EM algorithm [2] where two objects are closed if they can be considered as a realization of the same random variable. Another important family of methods can be pointed out under the name of conceptual clustering algorithms [6,5]. They have been originally constructed with nominal variables, but extensions to other type of data are available [7]. The particularity of those methods is that they aim to build understandable partitions. Those algorithms rely on non parametric probabilistic measures to define clusters and two objects are closed if they have the same value on *most* of the variables. Clusters are such that their probability vectors have the greatest entropy on each variable. Conversely to supervised learning where a class variable is known on a training data set, there are no references in unsupervised clustering. Thus, beside clusters construction, one might take special care to the relevance of the discovered data organization. It is necessary to check the validity of the obtained partition. In this article, we present an evaluation of some similarity measures used in supervised and unsupervised clustering as objective functions. We study their capability to discern,

D.A. Zighed, J. Komorowski, and J. Żytkow (Eds.): PKDD 2000, LNAI 1910, pp. 599–604, 2000.

in an ideal case, the best partition among all partitions. We are not concern with clusters validity. We only compare several objective functions regarding their discrimination capabilities. The variability of the similarity measures on the set of partitions must be necessarily as high as possible to ensure a non random choice in the set of measure equivalent partitions as it can be found in ISAAC [11] for instance. Even if the search strategy is important in the cluster construction, similarity measures must be sufficiently discriminant.

The organisation of the paper is the following. In the section one, we present the measures we have evaluated. Then, we present our strategy to have a meaningful comparison. Then, some results are given and discussed and some concluding remarks are given.

2 Objective Functions

We choose to evaluate only objective functions which require nominal variables, and not based on any kind of metrics. Thus we only study the behavior of an objective function and not a structuring of a data space. In the following we used the functions φ_β defined on $[0;1]$ such that $\varphi_\beta(x) = \frac{2^{\beta-1}}{2^{\beta-1}-1} x\left(1-x\right)^{\beta-1}$. This permits us to introduce generalized entropy [12] in the measures.

The Category Utility Function
This function is the one used in the well known conceptual clustering algorithm COBWEB [6] and in other related systems like ITERATE [1] and called category utility. It is a trade-off between intra-class similarity and inter-class dissimilarity,

$$CU = \frac{\sum_k P\left(C_k\right) \sum_i \sum_j \left[P\left(A_i = V_{ij} \mid C_k\right)^2 - P\left(A_i = V_{ij}\right)^2\right]}{K}$$

CU rewards the clusters which most reduce the collective impurity over all variables. This function has a form closed to the GINI index used in supervised clustering. Indeed, CU is the weighted average of the GINI index: $CU = \frac{\sum_k P(C_k) \times \text{GINI}}{K}$ in order to make this index independent of the number of clusters.

Quinlan's Gain Ratio
Other objective functions used in supervised clustering can be adapted for unsupervised clustering [5]. The adapted Quinlan's Gain ratio does not depend on the number of clusters and is given by,

$$\sum_i \frac{\sum_k P(C_k) \sum_j \left[\varphi_\beta(P(A_i = V_{ij}|C_k)) - \varphi_\beta(P(A_i = V_{ij}))\right]}{-\sum_k \varphi_\beta(P(C_k))}$$

3 Comparing Measures through Partitions Ordering

Since we compare measures on the set of partitions, simply taking values on the same partitions does not lead to a meaningful comparison. Indeed, the resulting

values on two different partitions can not be compared until the partitions are compared and ordered towards the clustering objectives. Thus, we consider data sets which are generated as the expression of a partitioning PI called the ideal case. Our comparisons aim at discovering if, among all the partitions, PI is measured as the best one by the different similarity indices and if those measures are sufficiently discriminating on the whole set of partitions. However, having PI, a natural choice to compare two partitions is to create a distance d such that $d(PI, P)$ permits to order the partitions P in reference to the ideal case PI. Of course, two different partitions can be at the same distance of PI. In order to have a significant measure, we use the two constituents of a partition: the objects and the vectors of variables values taken by the objects. We then build two distances, $\mu_{\mathcal{O}}$ and $\mu_{\mathcal{V}}$, such that the whole distance is constructed from both measures.

3.1 Comparing Clusters

A Distance Between Two Clusters Taken on the Objects
Marczewski and Steinhaus [10] present a distance for the comparison of two sets. Let $\mathcal{P}(X)$ be the class of all subsets of the finite set X. Then[1],

$$\forall C_k, C_{k'} \in \mathcal{P}(X),\ \mu_{\mathcal{O}}(C_k, C_{k'}) = \begin{cases} \frac{|C_k \Delta C_{k'}|}{|C_k \cup C_{k'}|} \text{ if } |C_k \cup C_{k'}| > 0 \\ 0 \qquad \text{otherwise} \end{cases}$$

A Distance Taken on the Variables
On the same principle, we define a distance between two probabilistic vectors, $\forall C_k, C_{k'} \in \mathcal{P}(X)$,

$$\mu_{\mathcal{V}}(C_k, C_{k'}) = \frac{1}{m\ p} \sum_{i=1}^{p} \sum_{j=1}^{m} |P(A_i = V_{ij} \mid C_k) - P(A_i = V_{ij} \mid C_{k'})|$$

3.2 Distance between Partitions

An Hausdorff Like Distance
Following Karonski and Palka [9], the previous distances can be extended to the case of comparing two partitions P_1 and P_2 given by $P_i = \{C_{i1}, \ldots, C_{iI_i}\}$. Given a measure μ between two sets, we can construct an Hausdorff like distance:

$$\mathcal{D}_{\mu}(P_1, P_2) = \frac{1}{2}\left[\max_{i \in I_1} \min_{j \in I_2} \mu(C_{1i}, C_{2j}) + \max_{j \in I_2} \min_{i \in I_1} \mu(C_{1i}, C_{2j})\right]$$

In order that the distance take into account the two previous measures, we use the following distance,

$$\mathcal{D}_{\min-\max}(P_1, P_2) = \sqrt{\mathcal{D}^2_{\mu_{\mathcal{O}}}(P_1, P_2) + \mathcal{D}^2_{\mu_{\mathcal{V}}}(P_1, P_2)}$$

[1] with $A \Delta B = A \cup B - 2A \cap B$ the symmetric difference

With the min-max distance, we hope to obtain a good discrimination of the partitions regarding to both aspects taken in account in conceptual clustering.

An Improved Measure
Through experiments, we observed that the distance $\mathcal{D}_{\min-\max}$ has some drawbacks. This measure is not very sensitive to minor changes because of its principle of worth case. That is why we propose another distance based on the search of the matching of maximal cardinality and minimal weight inside a bipartite graph [4]. The graph $G = (X_1 \cup X_2, E)$ associated to two partitions, has the clusters of the two partitions as nodes and all edges between X_1 and X_2 exist and are weighted by either $\mu_{\mathcal{O}}$ or $\mu_{\mathcal{V}}$. Notice that we can restrict the number of edges by selecting those whose weights are sufficiently low. The value of the matching permits to order the partitions. It decreases when partitions closer to PI, in the sense of the used measure, are taken and by definition is zero when PI is taken.

4 The Results

In the following study we reduce our investigations to boolean variables.The results obtained on such variables can not be extrapolated to nominal variables without bias, but a such constraint is necessary to reduce the number of parameters. We use an artificial dataset given by a diagonal block matrix, each block containing only 1 and the sizes of the blocks vary. This also defines PI. The partitions compared to PI are either the whole set, when the combinatory permits (8 objects and 8 variables), or random selected partitions otherwise (60 objects and 15 variables). In case of a random selection, 30000 partitions are sampled. Finally, we introduce some noise in order to evaluate the noise resistance of the measures. It is generated by random permutations of 1 and 0 in the boolean matrix. To measure the influence of β, we also study different values in $\{0.5, 1, 2, 3\}$.

Comparison of the Two Indexes
On the figure 1, CU function is plotted relatively to both measures in the exhaustive case. As expected, the matching index is more discriminant than the min-max distance. This is an important result which confirm the sensitivity of the matching measure. Notice that we observed similar results for all other measures and noise levels.

The Ideal Case
Given PI, we first study the influence of β using Quinlan gain ratio (see the figure 2 on the left). The effect of the parameter β is to spread out the values of the measures thus having more variations with β different to one. Using $\beta = 0.5$ leads to the measure with smaller variations and thus to the less discrimant one. The others choice give similar behavior with only differences in the scale. Thus, we choose to keep the original version of the gain ratio for comparison purposes. However, some experiments should be extended to better precise the influences of a high value of β. Having chosen a β for Quinlan gain ratio, we can make a comparison with the CU measure (see the figure 2 on the right). The variations

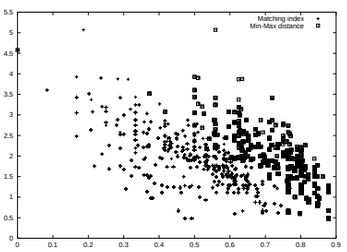

Fig. 1. Relative performance of the two orderings (x is distances, y is CU)

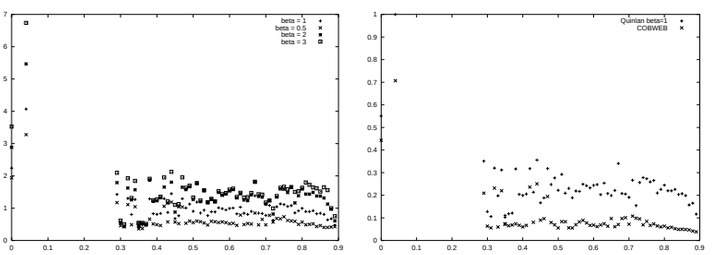

Fig. 2. (x is order, y is measure): (left) influence of β. (right) CU vs Quinlan.

of CU are very small so that nearly all partitions seem to be similar for that measure, except the extremal one. Following these preliminary results, Quinlan measure seems to be a better measure than CU even if more experiments are necessary to conclude.

The Noisy Ideal Case
In order to simulate a real case, we introduce some noise in the boolean matrix (see the figure 3). Following our experiments, Quinlan measure appears to be more noise resistant than CU. With a 5 percent noise level, it behaves like in the ideal case. When noise increases, some partitions take aberrant values (see the figure 3 (left)). However, this measure remains regular when CU becomes very perturbated (see the figure 3 (right)).

5 Conclusion

In this article, we have presented a new ordering of partitions to objectively compare the behavior of different quality measures which can be used in unsupervised learning. Through our experiment protocol, we have first established that our index has a better discrimant power than previous one of Marczewski and Steinhaus. Secondly, we shown than Quinlan gain ratio is noise resistant

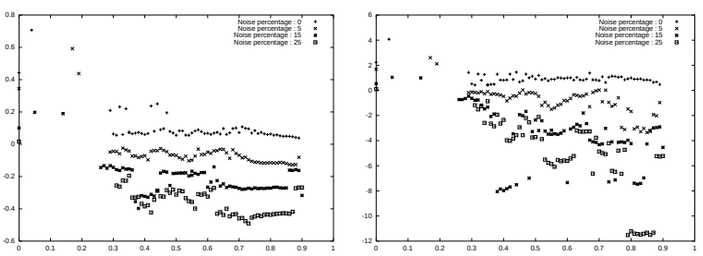

Fig. 3. Noise influence (x is order, y is measure): (left) Quinlan - (right) CU

and more discrimant than the other functions. When generalizing it with φ_β, it appeared that a bad choice was $\beta = 0.5$. Other values were not significantly different in our experiments. Due to lake of space, we have not reported here all our results but they all confirm the conclusions done in this paper. Let us also report that we also studied an adapted version of the Mantaras function [3] which behaved, in this set of experiments, like Quinlan gain ratio.

References

1. G. Biswas, J. Weinberg, and C. Li. Iterate: a conceptual clustering method for knowledge discovery in databases. Technical report, CS Departement, Vanderbilt university, Nashville, 1995.
2. G. Celeux and E. Diday et al. *Classification automatique des données*. Dunod, 1988.
3. R. López de Màntaras. A distance based attribute selection measure for decision tree induction. *Machine Learning*, 6:81–92, 1991.
4. J. Edmonds. Maximum matching and a polyhedron with 0-1 vertices. *Res. Nat. Bureau Standards*, 69B(1-2):125–130, 1965.
5. Doug Fisher. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4:147–180, 1996.
6. Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.
7. W. Iba and P. Langley. Unsupervised learning of probabilistic concept hierarchies. Technical report, Inst. for the study of learning and expertise, Pablo Alto, 1999.
8. A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood cliffs, New Jersey, 1988.
9. M. Karonski and Z. Palka. On marczewski-steinhaus type distance between hypergraphs. *Zastosowania Mat. Appli. Mathematicae*, 16:47–57, 1977.
10. E. Marczewski and H. Steinhaus. On a certain distance of sets and the corresponding distance of functions. *Colloquium Mathematicum*, 6:319–327, 1958.
11. L. Talavera and J. Béjar. Efficient construction of comprehensible hierarchical clusterings. In J. M. Żytkow and M. Quafafou, editors, *2nd PKDD*, volume 1510, pages 93–101. Springer-Verlag, 1998.
12. L. Wehenkel. On uncertainty measures used for decision tree induction. In *Info. Proc. and Manag. of Uncertainty*, pages 413–418, 1996.