A METHOD TO VALIDATE CLUSTERING PARTITIONS

Luca Frigau [1], Giulia Contu[1], Marco Ortu[1] and Andrea Carta[1]

[1] Department of Economics and Business Sciences, University of Cagliari, (e-mail: {frigau, giulia.contu, marco.ortu, andrea.carta88} @unica.it)

ABSTRACT: To evaluate the performance of clustering algorithms is challenging because typically the true classes are unknown. In this paper we propose a new cluster validity method that combines internal and relative criteria and employs Machine Learning algorithms to produce a relative validity ranking of partitions obtained from different clustering algorithms. Compared to other methods, the proposed approach considers the features' structure explicitly, can handle high-dimensional data, and can be applied to various clustering algorithms. The method has been tested on a simulated benchmark dataset, demonstrating its ability to rank correctly 11 classical clustering algorithms.

KEYWORDS: cluster validity, machine learning, simulation.

1 Introduction

Statistical learning methods can be categorized as supervised or unsupervised, according to on the availability of an associated response variable. In supervised methods the goodness of the estimated model is computed by comparing the prediction with the response variable, whereas in unsupervised methods the evaluation of their performance is very challenging because typically the true classes are unknown (Hastie *et al.*, 2009). Cluster analysis is an unsupervised method that deals with grouping a collection of objects into homogeneous clusters without having any information about the class of any object (Hennig *et al.*, 2015). There are several clustering algorithms available, none of which can be considered universally "best" in all circumstances. Therefore, it is common practice to compare the performance of several algorithms. The evaluation of a clustering algorithm's results is called cluster validity, which can be investigated through three main approaches: external, internal, and relative criteria. External criteria compare the obtained partition to externally known results, while internal criteria use only inherent quantities and features of the dataset, such as the proximity matrix. Relative criteria compare a set of defined partitions based on a pre-specified criterion. This paper proposes a cluster validity method that combines internal and relative criteria, inspired by the validation of gray-level thresholding image segmentation algorithms. The proposed method employs Machine Learning algorithms to produce a relative validity ranking of partitions obtained from different clustering algorithms according to a predefined validity criterion. The goodness of the method's fit is evaluated through tests on a simulated clustering benchmark dataset.

The paper is structured as follows: Section 2 describes the proposed cluster validity approach's methodology. In Section 3, a simulation study is performed. Finally, Section 4 contains some concluding remarks and a discussion of future work.

2 Validation method

The aim of a clustering algorithm is to split observations into subsets based on a reasonable pattern in the data. The assigned classes express information about the pattern identified by the algorithm in the data, allowing to measure how much the identified pattern corresponds to the features' structure. As the true pattern is unknown, the quality of the identified pattern cannot be assessed absolutely, but it can be assessed relatively.

To evaluate the coherence between the assigned classes and the features' structure, Machine Learning algorithms (ML) are employed, using the classes as the response variable and the features as independent variables. The performance of ML, indicated as ρ, serves as a relative proxy for the reliability of the output of the clustering algorithm. The performance of ML can be measured by several indexes, such as accuracy, specificity, sensitivity, etc., according to which aspect the analyst wants to focus on. Particularly, ρ does not indicate how well the partition corresponds to the pattern in the data, but it indicates the clustering algorithm output's quality compared to that of other algorithms. Therefore, by ordering the different ρ obtained in each partition, it is possible to rank the clustering algorithms according to their capability to cluster the objects on the basis of the pattern in the data.

Compared to other cluster validity approaches (Arbelaitz *et al.*, 2013), the proposed method has some advantages. For example, external criteria approaches require externally known results, which may not always be available or applicable to the problem at hand. Internal criteria approaches only use quantities and features inherent to the data set and may not provide an accurate assessment of the clustering output's quality. Relative criteria approaches compare different partitions based on a pre-specified criterion, but they do not consider the features' structure explicitly. The proposed method, on the other hand, uses Machine Learning algorithms to evaluate the coherence between the assigned classes and the features' structure and produces a relative validity ranking that takes this coherence into account. Moreover, the proposed method has the potentiality to be further developed to handle high-dimensional data and to be applied to various clustering algorithms, making it a versatile and robust method for cluster validity assessment.

3 Simulation study

To test the effectiveness of our method in ranking clustering partitions based on their ability to accurately reflect the data pattern, we conducted a simulation study. Our assumption was that the greater the noise in the data, the poorer the partition obtained by clustering algorithms. Therefore, we expected our method to rank the partitions based on the level of noise in the data.

For each clustering algorithm, we selected the best partition identified by the indexes included in the `R` function of `clusterCrit::intCriteria` (Desgraupes, 2018) within the range of 10-25 clusters. We then varied the level of noise in the data from 0% to 100% by randomly changing the classes of the partition. For instance, a noise level of 0% meant that no noise was added, and the classes of the partition remained the same. A noise level of 50% indicated that the classes of half of the observations were randomly assigned, while the classes of the other half were kept the same. In this simulation, we considered Support Vector Machine (Steinwart & Christmann, 2008) as Machine Learning algorithm, and 11 classical clustering algorithms.

Figure 1 shows that with the lower level of noise in the data, the method obtains higher values of ρ. So considering a partition is better when ρ is high, the method ranked correctly the partitions from the best (obtained in the data with no noise) to the worst (obtained in the data with the highest level of noise), for each of the 11 clustering algorithms. In that way, it is possible to use the method to rank different partitions without knowing the "true" one.

4 Conclusions

Validation of clustering algorithm output is of high interest due to the lack of a response variable to supervise the analysis. We have illustrated how the use of a Machine Learning algorithms-based method could allow for the ranking of clustering algorithms based on the proximity of their partitions to the unknown "true" partition. Using a simulated dataset, we showed that the method can rank the clustering algorithms among 11 different scenarios characterized by

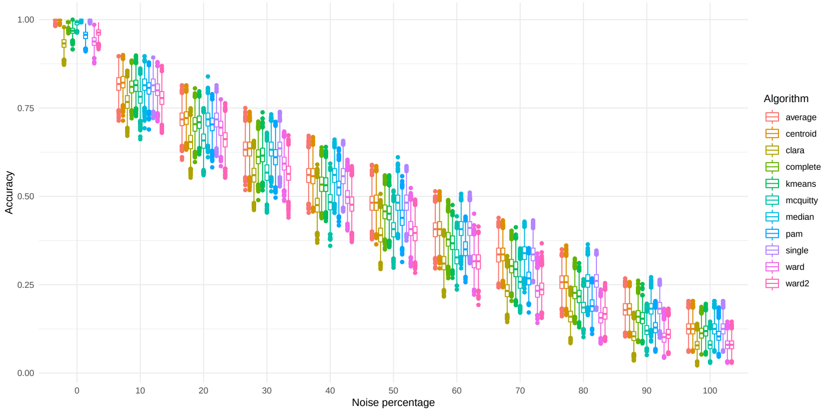

Figure 1. *Trend of performance of the validation method according to the level of noise added in the data.*

different noise levels. We believe that the proposed validation approach can enable the comparison of widely used clustering algorithms and help auditors choose the appropriate method for each situation.

As a potential extension, we are exploring the feasibility of applying the algorithm to big data scenario. In fact, many classical cluster validation indexes that already exist are characterized by high computational cost. Thus, it can be prohibitive to use them in big data scenarios.

References

Arbelaitz, Olatz, Gurrutxaga, Ibai, Muguerza, Javier, Pérez, Jesús M, & Perona, Iñigo. 2013. An extensive comparative study of cluster validity indices. *Pattern recognition*, **46**(1), 243–256.

Desgraupes, Bernard. 2018. *clusterCrit: Clustering Indices*. R package version 1.2.8.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome H, & Friedman, Jerome H. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hennig, Christian, Meila, Marina, Murtagh, Fionn, & Rocci, Roberto. 2015. *Handbook of cluster analysis*. CRC press.

Steinwart, Ingo, & Christmann, Andreas. 2008. *Support vector machines*. Springer Science & Business Media.