How to answer the most frequently asked questions in clustering with predictive clustering trees

Saso Dzeroski, Jozef Stefan Institute, Ljubljana, Slovenia

Predictive clustering trees (PCTs) [1] are an extension of decision trees and in particular regression trees. While regression trees split the data into subsets (clusters) to reduce the variance of a single target variable, PCTs try to reduce the variance along a set of target variables (i.e., to reduce intra-cluster/within cluster variance). Many variants of PCTs have been developed, some of them within HBP (incl. SGA2 [2]), most of them to address different tasks of predicting structured outputs, such as multi-label classification [2]. However, PCTs can be also used for semi-supervised learning and for fully unsupervised learning, i.e., for hierarchical clustering.

What is the optimal number of clusters? PCTs support the paradigm of constrained clustering, where we can impose different requirements on the results of clustering. We can specify the number of clusters in advance, but we can also specify alternative criteria which will do so indirectly. These can include constraints on the size of clusters (number of examples in them) and on cluster membership (must-link/cannot-link constraints). The criterion that is probably the most sound is the one based on statistical significance: In the construction of PCTs, the f-test can be employed to decide whether a particular split in the tree reduces variance significantly (and hence should be applied). This indirectly determines the number of clusters.

How should clusters be validated? By their very nature, PCTs allow for both clustering and prediction. The tests that appear in the tree use descriptive variables, while the variance calculation uses the clustering variables. In supervised learning (predictive modeling), these two sets of variables are disjoint, while in unsupervised learning (clustering) they (fully) overlap. PCTs can be evaluated just as any predictive data mining method, by applying a distance between actual and predicted values of the variables used for clustering, and performing hold-out or cross-validation: The results indicate the validity of the clusters.

What is the appropriate distance to use? PCTs allow the use of different distances according to which variance is measured. Despite the fact that the distance to be used in clustering is part of the specification of the task of clustering, the question of selecting the 'most appropriate' distance is often encountered. The evaluation of PCTs in predictive mode offers us an answer to this question. We can consider the overfitting score for a given PCT, which is the ratio between testing error (on unseen data) and training (resubstitution) error: The lower, the better. One possible answer is to use the distance measure that most reduces overfitting.

References:

[1] J. Struyf, S. Džeroski. (2010). Constrained predictive clustering. In S. Džeroski, and, B. Goethals, and P. Panov, editors. Inductive Databases and Constraint-based Data Mining, pp. 155-175. Springer, Berlin.

[2] STEPIŠNIK PERDIH, Tomaž, KOCEV, Dragi, DŽEROSKI, Sašo. (2019) Option predictive clustering trees for multi-label classification. Acta Polytechnica Hungarica – Journal of Applied Sciences. Accepted for publication.