Journal of Classification (2020) 37:1–3
https://doi.org/10.1007/s00357-020-09365-2

Editorial: Journal of Classification Vol. 37-1

Douglas L. Steinley [1]

Published online: 5 June 2020

The first issue of 2020 has fourteen articles that cover a broad set of topics. The first article, by Roman Hornung, develops ordinal forests as a type of random forest used for predicting ordinal response variables. As in the vein of standard random forest, Hornung develops a variable importance assessment to aid in understanding differential prediction rates for covariates. This is a welcome addition to span the standard regression trees (continuous data) and classification trees (binary data), and is likely to see use in the social and behavioral sciences where ordinal data are commonly seen.

Zaineb Degdia and Zied Elouedi also introduce a classification algorithm, with their contribution being based on the traditional Dendritic Cell Algorithm (DCA). The authors have modified the original DCA algorithm by incorporating a fuzziness in the classification boundaries and follow-up with a data cleaning step that refines the signal discovered by the machine learning algorithm. Introducing this type of refinement, that performs well when compared with extant supervised learning approaches, opens up potential alterations/additions to other methods potentially leading to an overall increase expected performance in several classes of algorithms.

The third article has a trio authors, Marek Smieja, Lukasz Struski, and Jacek Tabor, also working in the context of classification algorithms with a focus on how to handle missing data. The authors cleverly represent missing data in affine subspaces, allowing for general affine transformations (perhaps most excitedly dimensionality reduction). This clever representation allows for a straightforward and easy implementation, enhancing the likelihood that it will be adopted among applied users.

In the fourth article, Federico Fioravanti and Fernando Tohme develop a method for aggregating a set of subjective responses from individuals in a group as a way to partition the same individuals—addressing a specific type of social choice theory. The authors extend this classic problem to a type of "asymptotic" situation where there are an infinite number of voters.

The next article covers two binary optimization algorithms based on evolutionary processes for clustering algorithms as presented by Gehad Ismail Sayed, Ashraf Darwish, and Aboul Ella Hassanien. The interesting contribution of this paper is taking tried and true machine learning approaches for supervised settings and adapting them to unsupervised settings. A key

✉ Douglas L. Steinley
steinleyd@missouri.edu

[1] University of Missouri, Columbia, MO, USA

component of this approach is the attempt to identify the correct set of variables that are most likely to be related to the unknown cluster structure.

Volodymyr Melnykov and Semhar Michael introduce a novel method for clustering large datasets in the sixth paper. Specifically, the authors utilize the speed of k-means clustering for merging multiple clusters in a hierarchical fashion. The novel contribution of this approach is overcoming the tendency of k-means clustering to produce spherical clusters. In fact, the authors show that this merging can return highly nonlinear, novel shapes that are extremely difficult to obtain with many traditional clustering approaches (especially those that are based on distributional forms). In my opinion, this is a great contribution to the literature and a wonderful usage of k-means clustering.

In the seventh article, Matthieu Marbac, Mohammed Sedki, and Etenne Patin continue a long tradition in the *Journal of Classification* of developing a method for variable selection. The unique aspects of this particular approach are the fact that the dimensionality of the data is much greater than the number of observations, and the data are of mixed type. This is a clever methodological development that pushes the boundaries of variable selection in the context of model-based clustering.

Next, Renato Cordeiro de Amorim, Vladimir Makarenkov, and Boris Mirkin tackle the issue of potential mislabeled group data. The authors develop a clustering approach that extracts so-called "cores" of the labeled group that are most likely to be correctly labeled. Then, using an iterative approach, the labels of the remaining entries can be "corrected." The authors propose this as a potential pre-processing data correction step in supervised machine learning; at the very least, it could be used as a diagnostic tool for the potential explanation of observed poor performance in a supervised learning setting.

The ninth paper, Yunli Yang, Zhouwang Yang, and Baiyu Chen also address the problem of ordinal classification. However, rather than a machine learning algorithm as used earlier, the classification is achieved through pairwise comparisons. This process converts the ordinal classification into a regression framework that is disordered (e.g., ignores the ordinal structure) to conduct the analysis and then converts back into the ordinal classification setting.

The tenth paper has Dominique Fortin provides a review of clustering a dissimilarity. Specifically, the author connects the algebraic representation of the clustering with a graph theoretic visual representation, providing a one-to-one mapping between the two across a variety of constraints. While theoretical in nature, this presentation provides an insight that heightens the understanding of the clustering process.

Much like the prior paper, Francois Brucker, Pascal Prea, and Celia Chatel provide more theoretical relationships between dissimilarities and graph structures (specifically, hypergraphs) in the eleventh paper. Through a series of theorems, the authors provide several results regarding balanced structures and introduce a new class of dissimilarities—chordal dissimilarities, which they postulate will provide a good model for cluster analysis.

The twelfth paper concerns improving random forests by introducing tree weights for the individual decision trees prior to their averaging to get the final result for the random forest. Hieu Pham and Sigurdur Olafsson provide numerous results that show the Cesaro Average (taken from harmonic analysis) provides a weighting of trees that outperform the equal weighting of decision trees found in standard random forests. This initial work opens up several opportunities for refinement, including (but not limited to) the determination of the sequence in which the decision trees are averaged.

The final paper by Guang Ouyang, Dipak Dey, and Panpan Zhang develops a method for clustering social network data. First, the authors develop a measure based on cliques to assess

the quality of a clustering, and subsequently develop an algorithm for maximizing that measure. This is a refreshing departure from many community detection approaches that continue to have an overreliance on "modularity" as a measure for cluster structure. While further research is necessary to compare the two approaches, the variety in perspectives for clustering social networks is welcomed to help foster innovation in clustering this type of data.