Journal of Classification (2019) 36:1–2
https://doi.org/10.1007/s00357-019-09334-4

Editorial: Journal of Classification Vol 36-1

Douglas L. Steinley [1]

Published online: 22 May 2019

The first issue of 2019 has eight papers that cover a wide range of topics, from mixture modeling to traditional clustering and classification methods. The first paper is from Flynt and Dean on growth mixture modeling. Specifically, they look at the prospect of incorporating variable selection in the growth mixture model—which is a somewhat radical idea as it is usually assumed that all the measurement occasions are required to accurately estimate the growth trajectories; however, the authors show the selecting specific measurement occasions that separate the clusters well can lead to improved recovery and reduced error. I imagine that this paper will serve as a springboard for incorporating other variable selection methods, such as those described in Steinley and Brusco (2008), into growth mixture modeling.

The second article, by Cristina Tortora, Brian Franczak, Ryan Browne, and Paul McNicholas, introduces a mixture model for a new set of distributions—multiple scaled generalized hyperbolic distributions. This is unique in that the multivariate generalized hyperbolic distribution is *not* a special case of the multiple scaled generalized hyperbolic distribution. Accordingly, the authors develop a coalesced generalized hyperbolic distribution. While this is impressive, Tortora et al. also highlight the importance of cluster convexity and develop a special case of this new distribution that is guaranteed to be convex—a consideration that is not usually taken up when these methods are introduced. I believe the added flexibility of the general model developed in this paper will result in a greater ability to probe the boundaries—likely expanding them—of the capabilities of mixture models.

From mixture modeling, we move to traditional cluster analysis with Zdeněk Šulc and Hana Řezanková comparing similarity measures on categorical variables when used in the context of hierarchical clustering. Beyond comparing 11 existing similarity measures for categorical variables, the authors introduce two new similarity measures (variable entropy and variable mutability) for nominal variables. This algorithmic comparison of similarity coefficients extends some of the theoretical work that has been conducted by Albatineh et al. (2006) and Warrens (2008) to illustrate the comparative performance of similarity measures in the context of hierarchical clustering.

Kensuke Tanioka and Hiroshi Yadohisa provide a non-negative matrix factorization to aid in the interpretation of the cluster structure in a lower dimensional space. Additionally, the lower

✉ Douglas L. Steinley
steinleyd@missouri.edu

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

dimensional representation results in orthogonal components, resulting in so-called simple structure. This simple structure allows for direct interpretation of the relationship between the variables and the cluster structure. The authors note that there remain issues concerning choosing the number of clusters and the potential to get stuck in locally optimal solutions; however, the present manuscript lays some foundational results for fertile ground of research in the application of non-negative matrix factorization in the context of cluster analysis.

In the fifth article, Manazhy Rashmi and Praveen Sankaran extend the classic Isomap algorithm of Silva and Tenenbaum (2002) for nonlinear data reduction. Specifically, the authors improve on a variation of Isomap, Landmark Isomap (L-Isomap), that was developed for complex data structures. While the Isomap algorithm requires the computation of numerous geodesics, the L-Isomap algorithm nominates a small number of observations to serve as so-called landmarks that can serve as reference points to compute geodesic distances to. Rashmi and Sankaran note that the nature in which the landmarks are chosen matters, introducing a clustering algorithm for determining the landmarks, resulting in a principled approach that results in improved overall performance of the data reduction.

Md. Matiur Rahaman and Md. Nurul Haque Mollah take on what is often overlooked in practice the adaptation of theoretical approaches to the conditions of the real-world, where applications will be conducted. In this case, the authors consider the situation when outliers exist in the test and training data sets in supervised classification. To address the bias that was introduced by the inclusion of outliers, Rahaman and Mollah develop a robust Bayes classifier that reduces to a maximum likelihood estimator as a special case when the tuning parameter goes to zero.

In the penultimate paper, Hossein Baloochian and Hamid Reza Ghaffray address the situation of classifying observations when there are more than two classes. The authors propose decomposing the multiclass classification problem into a set of binary classification problems, and then recombining the solutions from the binary problems to obtain the final solution. The proposed recombination results in a decision tree with K terminal nodes (where K is the number of classes). Conveniently, the proposed method is agnostic with regard to classification, allowing the use of methods designed for two-class classification (such as logistic regression or support vector machines) to be scaled up to multiclass classification problems.

The final article of the first issue is by Maryam Abaszade and Sohrab Effati and provides a method for classifying random variables based on support vector machines. The authors incorporate a set of probabilistic constraints that results in better performance than the standard support vector machine when attempting to obtain the optimal separating hyperplane.

Douglas Steinley, Editor

References

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification, 23*, 301–313.

Silva, V. D., & Tenenbaum, J. B. (2002). Global versus local methods in nonlinear dimensionality reduction. In *Advances in neural information processing systems* (pp. 705–712).

Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika, 73*, 125–144.

Warrens, M. J. (2008). *Similarity coefficients for binary data. Ph.D. thesis*. University of Leiden.