Journal of Classification 27:1 (2010)
DOI: 10.1007/s00357-010-9048-6

Editorial

The first three articles in this issue all develop clustering methodology, but each originates from quite a different setting and uses a different conceptual framework. In the first article, Chiang and Mirkin present an experimental study of the classic unsolved problem of choosing the number of clusters K. The setting is general multivariate data, and the conceptual framework is within-sums-of-(squared)-distances partitioning. The paper introduces an interesting distinction between recovery of K itself, cluster membership recovery, and recovery of the centers characterizing the clusters. For the recovery of K, the good old Hartigan rule outperforms some well-known recent proposals, but for recovery of cluster membership and cluster centers the authors argue that their own version of "intelligent" K-means is the way to go.

The setting of the second article by Lee, Taddy and Gray is the selection of K points that are representative for a given regression situation, and they work in a mixture modeling framework with Bayesian estimation. A novel feature is that the means of the distributions in the mixture are constrained to coincide with a data point. The Bayesian approach allows them to formally encode desirable properties of the representative sample of points, such as spread across the domain and balance in cluster sizes. In the third article, Coppi, D'Urso and Giordani work in a setting of spatial time series, and use the fuzzy clustering framework, which shares features with both the within-sum-of-squares framework and the mixture modeling framework. They use penalty functions to deal with the dependencies of the trajectories in space and time, and show in a socio-demographic application that incorporating the dependency information leads to more reasonable clusterings.

After these clustering articles, Kauermann, Ormerod and Wand present a supervised classification algorithm built on semi-parametric logistic regression. Their algorithm performs fast fitting and variable selection by borrowing ideas from generalized linear mixed modeling. It compares favorably with well-known current algorithms which use generalized additive modeling and generalized cross-validation. In the last paper of this issue, Bennani Dosse and Ten Berge study a special form of Procrustes analysis, in which two matrices are optimally matched by rotation and dimension-wise rescaling. Their approach based on majorization gives more insight into the problem, and leads to computational schemes that are slightly less prone to local optima and saddle points than existing algorithms.

In the Book Review section, Pratte favorably discusses a text book on probability with **R**.

Willem J. Heiser