Editorial

This is the second installment of *Machine Learning* to include selected papers on unsupervised learning. The editorial prefacing the first installment of unsupervised learning papers (Fisher, 2001) highlighted association-rule learning, belief network learning, and clustering as the primary forms of unsupervised learning, as well as describing relationships between these three unsupervised learning paradigms. The topics addressed in the seven papers of the first special-issue volume included empirical comparisons of model-based clustering algorithms, fast discovery of patterns in sequential data, discovering rules in first-order logic, comparisons of unsupervised and supervised learning for classification, and unsupervised approaches to document organization.

The papers of this second issue represent additional research in unsupervised learning concerning the form of data, the control strategies used to search for knowledge structures, and the tasks to which unsupervised learning can be applied.

Cadez, Smyth, McLachlan, and McLaren describe an EM-based unsupervised method for modeling binned (and truncated) data. Binned data (or multivariate histograms) arise when measuring instruments have limited precision and in other contexts. The paper extends previous work with EM to fit maximum-likelihood finite-mixture models to binned univariate data, to the multivariate case. Numerical integration techniques and subsampling are used to manage computational costs. The approach is motivated by and evaluated on anemia diagnosis data, and evaluated on simulated data as well. This evaluation shows that the EM-based approach for binned data compares favorably with a baseline approach that uses the full data (no binning), as well as accurately capturing the underlying PDF of simulated data.

Dubnov, El-Yaniv, Gdalyahu, Schneidman, Tishby, and Yona report a novel (dis)similarity-based pairwise clustering algorithm that iteratively (re)normalizes the (dis)similarities between pairs of data points. Empirical results and some theoretical treatment indicate that this iterative transformation of the data consistently converges to a stable data representation that expresses the cluster membership of each datum. The paper also describes the use of this basic partitional method to recursively decompose the data into a hierarchical clustering. Cross-validation is used to terminate recursive decomposition so as to mitigate the problem of overfitting. The algorithm is robust in the face of noise and is tested on artificial and natural datasets. The authors point to scaleup to large datasets as a major issue for further research.

Peña, Lozano, and Larrañaga describe a paradigm of learning *recursive Bayesian multinets*, which represent a recursive decomposition of data by a decision tree, with a Bayesian network at each leaf to model the subset of data covered by that leaf. A recursive Bayesian multinet naturally represents varying conditional independencies within data subsets. The paper highlights relationships between clustering and learning of Bayesian networks, two major paradigms of unsupervised learning.

Ramoni, Sebastiani, and Cohen describe a model-based agglomerative method of clustering sequential data, notably time-series data, with the goal of identifying groups of time series with similar behavior. The representation of a base datum as a sequence of states or events is relatively little studied in the unsupervised learning literature. A Markov chain in the form of a transition matrix is induced for each sequence, and clustering occurs over these Markov chains. The authors evaluate their algorithm on artificial time series, as well as sensory data from a mobil robot, and compare their approach to an EM-based alternative from the literature.

Collectively, the papers of this volume and the previous special volume sample the great variety of work in unsupervised learning, and they advance research in important directions. I thank the authors and reviewers for their efforts in bringing this special collection to fruition.

Doug Fisher

Department of Electrical Engineering and Computer Science,
Vanderbilt University
dfisher@vuse.vanderbilt.edu

Reference

Fisher, D. (2001). Editorial: Special issue on unsupervised learning. *Machine Learning*, 42, 5–7.