Machine Learning, 42, 5–7, 2001

Introduction

This installment, and a forthcoming installment of *Machine Learning*, include selected papers on unsupervised learning of patterns from data. In contrast to supervised concept learning, unsupervised methods do not use class labels to constrain the search for patterns.

Learning association rules, learning belief networks, and clustering, are commonly-studied forms of unsupervised learning. The knowledge structures learned vary across these paradigms. Association rules amount to if-then rules: if a given conjunction of variable values is true of a datum, then a second conjunction is true of the datum with some probability. Belief networks are graphical models that also represent relationships between variables. Nodes in the network correspond to variables, and arcs between nodes express conditional (in)dependencies between variables that appear to exist in the data. Algorithms for learning belief networks encode the joint probability distribution of variable values observed in training data as a graphical model that is, ideally, sparse due to conditional independencies that are reflected in the data. Finally, clustering algorithms group objects into classes based on some measure of similarity (or distance) between objects, or an objective function that measures the quality of a set of clusters. Clustering algorithms can be classified as partitional methods, which partition the data, clumping methods that form overlapping clusters of objects, and hierarchical methods that recursively partition objects. Within machine learning, special attention is paid to characterizing clusters by concepts or models, which facilitate interpretation by human analysts.

Despite the differences, the paradigms share common characteristics. Notably, learning in each paradigm facilitates inference along multiple variables used to describe data, rather than or in addition to inference of a class label, which is the focus of inference in supervised contexts. Given the rich and flexible inferencing possibilities of learned knowledge structures and the limited feedback required by their associated learning algorithms, unsupervised methods of each paradigm are used for various hypothesis generation and knowledge discovery tasks, such as document organization to support browsing and retrieval activities by users with a variety of goals, and intelligent agent design and Web-site customization. Very recent overviews of unsupervised learning algorithms, their associated knowledge structures, and performance tasks from each of the paradigms can be found in the Handbook of Data Mining and Knowledge Discovery (in press).

Of course, the inference "potential" of learned knowledge structures varies. For example, association rules are typically intended to represent only the most common inter-variable relationships, which are presented in a form that is easily understood by human analysts. In contrast, a belief network is intended to represent a complete statistical model of the data. Clusterings, together with their cluster characterizations, vary in the extent that they statistically model the data, with hierarchical clusterings and overlapping clusterings offering richer modeling possibilities than partitional methods. In principle, association rules can be extracted from a belief network or clustering. Inversely, association rule learning can be viewed as clustering/clumping, since rules identify subsets of objects, albeit without covering all data, suggesting that tradeoffs used to constrain cluster formation might usefully serve to limit the number of discovered association rules to those that are frequent and "interesting."

The papers of this collection highlight important 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.

Meilă and Heckerman experimentally compare the three main approaches to *model-based* clustering. The model-based framework views the clustering task as constructing a belief network, typically where the clusters correspond to values of a single hidden variable, which renders the observed variables independent conditioned on cluster membership (i.e., a hidden variable value). Equivalently, this can be viewed as learning a mixture model, where clusters are identified with the mixture components. The algorithms examined by Meilă and Heckerman differ in the control strategies that they use to search for the best model structure. They also examine the performance implications of "soft" assignment, in which each object is probabilistically assigned to each cluster, versus "hard" assignment, in which each object is placed unambiguously into a single cluster. The authors compare the algorithms on a number of performance dimensions, including runtime and space complexity, and criteria related to inference accuracy on test data.

Zaki describes a new algorithm for discovering patterns from sequence data, in which events/features within a sequence are ordered (e.g., temporally). Zaki's algorithm for "sequence mining" shares important processing characteristics with association rule learners, but the problems of sequence mining and association rule learning also differ in some significant ways, thus motivating specialized sequence mining algorithms. In keeping with traditional data mining concerns, Zaki's algorithm discovers sequence patterns from a very limited number of database scans, while still making guarantees about the completeness of the discovered rule set.

Flach and Lachiche describe a new algorithm for discovery of rules in first-order logic, which has been little-studied relative to the research activity on unsupervised, propositional learners. Among other functionality, their system can be viewed as extending association rule learning to first-order representations. Importantly, their algorithm exploits an evaluation function that includes aspects of rule "interestingness" as well as rule coverage, and they show how this function can be used to effectively prune the search for the "best" rules.

Two papers compare supervised and unsupervised approaches to learning. While unsupervised methods are not guided by an overarching goal of predicting class labels, the knowledge structures formed by unsupervised approaches can be used for this task. Equivalently, supervised learning can build a classifier for any single variable (and variable combinations) that describes the data. Thus, a single knowledge structure built through unsupervised learning can be compared against multiple classifiers, each built through supervised learning, one classifier per variable. **Japkowicz** experimentally compares the performance obtained by supervised, backpropagation of a feedforward neural network for binary classification tasks, with an unsupervised, auto-associator, also trained through backpropagation, but with positive examples only. The experiments indicate that the auto-associator outperforms supervised backpropagation for predicting class membership in some circumstances, and characterizes the class of problems in which this is likely to happen.

Grove and Roth compare two unsupervised algorithms of the model-based paradigm, one based on EM and a second, original algorithm that uses covariation among the observed variables, against a well-known supervised approach. They find that when model assumptions are satisfied, the unsupervised approaches are competitive, if not superior to the selected supervised method, but if model assumptions are violated, then the unsupervised approaches prove fragile, perhaps more than expected, relative to the non-parametric supervised approach.

The final two papers deal with document organization via unsupervised learning, which has a rich history in information retrieval, but which has gained even greater importance with the advent of the World Wide Web. Documents are represented as vectors of their constituent words and counts, which are very sparse, as no document is likely to contain a significant proportion of words that is found throughout a document corpus. **Dhillon and Modha** explore a variation on the *k-means* clustering algorithm to organize document collections. Their characterization of the resulting clusters as *fractal-like* arguably speaks to the utility of recursive decomposition in the form of hierarchical clusterings of document collections. They show that cluster prototypes lead to intuitively pleasing and information preserving decompositions of a document space.

Hofmann describes the *aspect model*, based on a probabilistic variant of Latent Semantic Analysis, which can be viewed as a clustering of word-document *pairs* in the model-based framework. The aspect model considers documents as belonging generally to more than one cluster, and evidence across clusters is combined for purposes of document retrieval and other forms of inference. Hofmann also introduces a form of temperature-regulated EM, akin to simulated annealing, which demonstratably yields high quality models.

Collectively, these papers 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

Klösgen, J. & Żytkow, J. Eds. (in press), *Handbook of Data Mining and Knowledge Discovery*. New York: Oxford University Press.