Machine Learning, 56, 5–7, 2004

Introduction: Special Issue on Theoretical Advances in Data Clustering

This special issue is devoted to papers that advance the state of the art in the theory of clustering. Three of the papers in this issue describe algorithms for classical clustering objectives. The remaining five papers explore issues such as the design of new clustering objectives, clustering large datasets, generalizing the objects to be clustered, and clustering to improve learning. In the rest of this introduction, we describe these research trends in the context of the papers that appear in this special issue.

Theoretical underpinnings of classical clustering objectives

Among the most widely studied clustering objectives is the *squared error distortion* or *k-Median-squared* objective which is: Given a set of n points in a metric space and given k, find a set of k centers (that are themselves points in the metric space) such that the sum of squared distances from points to nearest centers is minimized. As it is unlikely that the optimum solution to the squared error distortion objective can be found in polynomial time (the problem is NP-hard), interest has shifted towards approximation algorithms. Such algorithms guarantee that the ratio of the algorithm's distortion to the optimum distortion is boundably small.

For the squared error distortion objective, P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay describe an algorithm that finds a solution whose distortion is at most twice the distortion of the optimum solution in the case that the points to be clustered fall in a Euclidean space. The authors obtain an approximation guarantee by applying the Singular Value Decomposition (SVD) to the matrix of points.

Related to the k-Median-squared objective is the k-Median objective where the goal is to find a collection of k centers such that the sum of distances from points to nearest centers is minimized. In many cases, approximation algorithms for k-Median also apply to k-Median-squared, and thus the k-Median objective, being even simpler to describe, has been studied more widely in the theoretical community. R. Mettu and G. Plaxton give an algorithm that, with high probability, finds a constant-factor approximation to the optimum solution in $O(nk)$ time. Also, an accompanying lower bound demonstrates that any randomized algorithm requires $\Omega(nk)$ time in order to find a constant factor approximation with even a small probability. A. Meyerson, L. O'Callaghan, and S. Plotkin improve the running time under the assumption that no cluster is very small in the optimum solution. Specifically, their algorithm finds a constant-factor approximation to the optimum solution in time that depends polynomially on k and $1/\delta$, assuming that each cluster has $\Omega(n\delta/k)$ points and δ is a given confidence parameter.

Design of new clustering objectives

Beyond k-Median, new clustering objectives have also recently been introduced. These objectives tend to capture some of the constraints that arise more typically in practice. Among the more natural and conceptually clean formulations is the objective given by A. Blum, N. Bansal, and S. Chawla. In this problem one is given information about each pair of objects to be clustered in the form of "+" if the two objects should belong to the same cluster and "−" if they should belong to different clusters. The goal of correlation clustering is to partition the points so as to minimize the number of disagreements with the "+"/"−" labels. The authors give a constant-factor approximation to the problem of minimizing disagreements, in addition to other results.

N. Mishra, D. Ron, and R. Swaminathan address the problem of finding conjunctive cluster descriptions. For example, a cluster of web documents may be described by the conjunction of words that is common to that cluster. A new formulation of the clustering problem is given that differs from previous approaches in that clusters may overlap, not all points are clustered, and a point may be assigned to a cluster even if it only satisfies most of the attributes in the conjunction. Algorithms are given to identify a collection of well-separated, descriptive conjunctive cluster descriptions.

Clustering large datasets

Many of the papers in this special issue acknowledge that modern datasets, like the web, are very large, and thus efficient algorithms are essential. The paper by P. Drineas, A. Frieze, R. Kannan, and S. Vempala demonstrates that the SVD of an appropriately chosen random submatrix provides an approximation of the SVD of the entire matrix. The consequence is a fast randomized algorithm that can be applied to very large datasets. As previously mentioned, the paper by R. Mettu and G. Plaxton gives a linear time algorithm for obtaining a constant factor approximation to the k-Median problem. In addition, the paper by A. Meyerson, L. O'Callaghan, and S. Plotkin finds an approximate k-Median clustering in sub-linear time, under the assumption that there are no small clusters. Finally, the paper by N. Mishra, D. Ron, and R. Swaminathan identifies a collection of conjunctive descriptions in time that depends on the number of attributes (and other parameters), but does not depend on the number of points to be clustered.

A. Borodin, R. Ostrovsky, and Y. Rabani consider graph clustering problems where the vertices correspond to the points to be clustered, and edges exist between two vertices if they are at least some specified distance apart from each other. The goal is to output the connected components (interpreted as clusters) of the graph. The authors describe algorithms that in subquadratic time produce approximate solutions to this (and other) problems.

Generalizing the objects to be clustered

A new research trend is to consider clustering objects that possess a richer structure than points in Euclidean space. One such generalization is the problem of clustering graphs

studied by B. Jain and F. Wysotzki. By way of example, consider the optical character recognition problem where one is given a collection of samples of the numbers 0–9, and the goal is to discover the clusters corresponding to each digit. Each sample character may be represented as a graph where the vertices correspond to pixel values that exceed a certain gray-level threshold and the weights on the edges correspond to the relative distance between the pixels. It can be shown that clusters of graphs allow for rotation, translation, and scale invariance. The algorithms given in this paper exploit the richer structure of a graph representation so as to derive more meaningful clusters.

Clustering to improve learning

In the semi-supervised setting there is typically a wealth of unlabeled data and a small amount of labeled data and the goal is to design a classifier that is a good predictor of the label of future data. Under the assumption that the data lies in a submanifold, M. Belkin and P. Niyogi use unlabeled data to discover a low-dimensional manifold on which the data lies and then subsequently develop classifiers on that manifold. Theoretical justification of the algorithms are provided and experiments are also given in various applications including handwritten digit recognition.

Acknowledgments

We thank the authors for their contributions, and the reviewers for their detailed and thoughtful reviews. We also thank Rob Holte for his advice and support.

Nina Mishra
HP Labs, Palo Alto, CA 94304, USA
Department of Computer Science, Stanford University, Palo Alto, CA 94304, USA
nina.mishra@cs.stanford.edu

Rajeev Motwani
Department of Computer Science, Stanford University, Palo Alto, CA 94305, USA
rajeev@cs.stanford.edu