EDITORIAL

Special Issue on the Best Papers of SDM'08

DOI:10.1002/sam.10010
Published online in Wiley InterScience (www.interscience.wiley.com).

This special issue is devoted to the best papers submitted to the eighth SIAM International Conference on Data Mining (SDM'08) that was held in Atlanta, Georgia, during April 24–26, 2008 (http://www.siam.org/sdm08). It also includes a summary of the panel on "Perspectives on Research Directions and Trends for the Data Mining Research Community", which was held at the conclusion of SDM'08.

SDM'08 •continued a series of conferences whose focus was on the theory and application of data mining to complex datasets in science, engineering, biomedicine, social sciences, and business. These datasets challenge our abilities to analyze them because they are large and often noisy. Sophisticated, high-performance, and principled analysis techniques and algorithms, based on sound statistical foundations, are required to extract useful knowledge from them. Visualization is often critically important, tuning for performance is a significant challenge, and the appropriate levels of abstraction to allow end-users to exploit sophisticated techniques and to understand clearly both the constraints and the interpretation of results are still something of an open question. The SIAM data mining conference papers from the past conferences (2002–2008) are available online at http://www.siam.org/proceedings/.

In response to the call for papers, the conference received 282 papers from over 25 countries. Each paper was reviewed initially by at least three members of the international program committee. Area chairs then initiated discussion on papers with discrepant scores. For the first time, this year we sought author feedback for selected papers, mainly to clarify technical issues. Area chairs subsequently provided their recommendations to the program co-chairs, who then collated and refined these suggestions across all areas. In the end, 40 papers were selected to appear as full papers, and 37 papers were selected as short papers or posters. Out of the accepted papers, six papers were selected for this special issue on the best papers of SDM'08. The authors were given about two months to revise, extend, and improve upon the original submissions. A brief overview of these papers appears below.

The paper "*Global Correlation Clustering Based on the Hough-Transform*", by Elke Achtert, Christian Böhm, Jörn David, Peer Kröger, and Arthur Zimek, presents a novel approach to subspace clustering that is robust. The idea is to map each point in the original data space into a function in the Hough-transform space. Intersections of the curves in the Hough-space correspond to points that lie on the same manifold in the original space. The clustering task is then transformed into finding dense regions of "intersection" points in the Hough space. Unlike in the original space, clustering in the new space is insensitive to grid boundaries, and also tolerant to noise. Instead of exhaustively searching in the new space, which would have an exponential complexity, the authors propose an effective recursive exploration strategy that typically converges to the dense regions. Other advantages of the approach are that it does not need the number of clusters as input, and can also output hierarchical subspace clusters.

In the paper, "*A Scalable Local Algorithm for Distributed Multivariate Regression*", Kanishka Bhaduri and Hillol Kargupta present a local distributed algorithm for performing multivariate regression and monitoring the model in a P2P network. The algorithm can be used for distributed inference, data compaction, data modeling, and classification tasks in many emerging peer-to-peer applications for bioinformatics, astronomy, social networking, sensor networks, and web mining. The approach is scalable, decentralized, asynchronous, and inherently based on in-network computation.

''*Cluster Ensemble Selection*", by Xiaoli Z. Fern, and Wei Lin studies the ensemble selection problem for clustering. That is, given a set of clusterings, how to select a smaller, better ensemble clustering solution. The two issues of relevance include cluster quality and diversity. Instead of considering these issues in isolation they present new strategies that consider them simultaneously, and they show that this combined approach is much more effective than other alternatives.

Many clustering methods place assumptions on the distribution of the data that may or may not hold in practice. Most methods require at least an initial guess as to the appropriate number of clusters or classes. Such assumptions are particularly problematic for the knowledge discovery process. The paper "*The Relevant-Set Correlation Model for Data Clustering*", by Michael E. Houle introduces a

model for clustering, the Relevant-Set CorrelationRelevant-Set Correlation (RSC) model, that requires no direct knowledge of the nature or representation of the data. Instead, the RSC model relies solely on the existence of an oracle that accepts a query in the form of a reference to a data item, and returns a ranked set of references to items that are most relevant to the query. The effectiveness of the •RCS model is confirmed through experimentation.

Most clustering algorithms produce a single clustering for a given dataset even when the data can be clustered naturally in multiple ways. In the paper "*Simultaneous Unsupervised Learning of Disparate Clusterings*", Prateek Jain, Raghu Meka, and Inderjit S. Dhillon address the problem of uncovering disparate clusterings from the data in a totally unsupervised manner. They introduce two approaches. One is modeling the data as a sum of mixtures and associating each mixture with a clustering. The other is finding good clusterings of the data that are also decorrelated with one another. They demonstrate that their methods achieve remarkably higher accuracy than do the existing factorial learning as well as traditional clustering algorithms.

In their paper ''*Fast Monitoring Proximity and Centrality on Time-Evolving Bipartite Graphs",* Hanghang Tong, Spiros Papadimitriou, Philip S. Yu, and Christos Faloutsos tackle the important problem of proximity tracking in dynamic bipartite graphs such as author-conference, movie-user ratings, hub-authority graphs and so on. In the evolving graphs new links may arrive, old links may disappear, and link weights may change. They present random walk-based methods to compute the centrality of the nodes, as well as the proximity of a pair or set of nodes. They show how to incrementally update these measures efficiently in a dynamic setting, and show the effectiveness of their approach experimentally.

In closing, we believe that the excellent set of papers for this special issue will be valuable both to researchers and practitioners in data mining for many years to come. We would like to take this opportunity to thank all the program committee members and external reviewers for their help in the difficult task of reviewing and choosing papers for presentation, posters, and awards. We are especially grateful for the help of the area chairs, Naoki Abe, Chris Clifton, Inderjit S. Dhillon, Wei Fan, Lise Getoor, Bart Goethals, Vasant Honavar, Eamonn Keogh, Bamshad Mobasher, Zoran Obradovic, and Jian Pei, who handled the reviewing process with great care and insight. We are grateful to the Best Paper Committee members for the help in identifying the best papers; the Committee comprised of Jiawei Han, Johannes Gehrke, George Karypis, Vipin Kumar, and David Skillicorn. We are also grateful to Chid Apte, Haesun Park, Chandrika Kamath, and Vipin Kumar for offering guidance on various program-related issues.

Mohammed J. Zaki[1] and Ke Wang[2]
[1]Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY, USA [2]Computer Science Department, Simon Fraser University, Burnaby, BC, Canada