EDITORIAL

Special Issue on the Best Papers of SDM'09

DOI:10.1002/sam.10056
Published online 17 November 2009 in Wiley InterScience (www.interscience.wiley.com).

This special issue contains extended papers of the best studies selected from submissions to the Ninth SIAM International Conference on Data Mining (SDM'09) that was held in Sparks, Nevada, April 30 to May 2, 2009 (http://www.siam.org/meetings/sdm09/). SDM'09 continues the tradition of the SDM conference series to focus on the theory and application of data mining with complex data in science, engineering, biomedicine, social sciences, and business. In particular, the program of SDM'09 conference covered the following variety of important and emerging data mining topics: Frequent Patterns and Relational Data Mining, Biomedical Applications, Data Reduction and Feature Selection, Data Stream Mining, Clustering and Unsupervised Learning, Privacy and Social Issues, Mining Spatial and Spatio-temporal Data, Probabilistic and Statistical Methods, Temporal Data Mining, Mining Graph and Semi-structured Data, Supervised Learning, and Web Search and Text Mining Applications. SDM'09 also included an additional category designated to data mining papers that may not fit any of clearly definable areas.

This year the conference drew a record number of 351 submissions. The submissions were from all over the world including Australia, Bangladesh, Belgium, Brazil, Canada, China, Finland, Germany, Greece, Hong Kong, India, Iran, Israel, Italy, Japan, Korea, Netherlands, New Zealand, Pakistan, Singapore, Spain, Sweden, Switzerland, Taiwan, United Kingdom, and USA. This truly reflected the international character of this conference. Each submitted paper was reviewed by at least three members of the international program committee. Area chairs then initiated discussion on papers with discrepant scores. SDM'09 continued the tradition started at SDM'08 and sought author feedback for a subset of papers where there was a need for clarification of some technical issues. Area chairs next provided their recommendations to the program co-chairs who integrated and refined these recommendations across all technical areas. Combining quality-based selection with time and space constraints imposed by the conference duration and available space, 55 papers were selected for oral presentation in 11 sessions over 2 days and additional 50 papers were presented in a poster session. The creative work of all the authors, the extensive efforts of the program committee members and external reviewers, and the superb organization and leadership of area chairs have resulted in an outstanding set of papers that will surely exert influence and promote excellence in data mining for many years to come. Out of the accepted papers, extended versions of seven papers were selected for this special issue on the Best Papers of SDM'09. We briefly introduce these papers below.

By understanding to what extent a particular set of examples represents the current concept the learning algorithm can incorporate the examples at hand to a larger or smaller extent in the generation of a classier. To facilitate such an improved learning in *Adaptive Concept Drift Detection*, Anton Dries and Ulrich Ruckert propose three novel drift detection tests, whose test statistics are dynamically adapted to match the actual data at hand. Their first method is based on a density estimation technique on a binary representation of the data. The second method measures the average margin of a linear classifier induced by a 1-norm support vector machine (SVM), whereas the third one is based on the average error rate of a linear classifier generated by a SVM. In precision-recall analysis, they compare these new approaches with the maximum mean discrepancy method, the StreamKrimp system and the multivariate Wald–Wolfowitz test. The obtained results provide evidence that the new methods are able to detect concept drifts reliably and are not too sensitive to noise in most cases. This study received the best paper award at SDM'09.

An event in time series is characterized by an interval of measurements that differs significantly from a baseline. A challenge in identifying these events is that one must distinguish between events that could have occurred by chance and events that are statistically significant. In *Discovering Arbitrary Event Types in Time Series*, Dan Preston, Pavlos Protopapas, and Carla Brodley propose an effective probabilistic method for finding the areas of significance by calculating the significance of an arbitrary-sized sliding window in a cost-effective way. In addition to an evaluation on synthetic time series of varying sizes, lengths, and different noise characteristics, their method is successfully applied to a large astronomical survey where its noise independence is demonstrated and where it was able to recover

all known events in the top ranked 1% and also to identify several events of interest that generally fail traditional tests and several events that were previously unidentified. This study received the runner up award at SDM'09.

One of the central goals of data mining is to find interesting patterns in data. The paper *Efficient Discovery of Interesting Patterns Based on Strong Closedness*, by Mario Boley, Tamas Horvath, and Stefan Wrobel, goes beyond the classic notion of interestingness (often measured by some type of frequency) and suggests finding patterns that are long but not necessarily very frequent. Their proposed measure is based on the strength of closedness of patterns. Closedness is a standard notion in formal concept analysis. The authors precisely define the notion of Δ-closed sets and demonstrate that by using strongly closed sets. They show that it is possible to arrive at meaningful and stable result sets containing long patterns, without using support threshold parameters, and the closure operator that can be computed efficiently. An empirical evaluation demonstrates that the proposed approach can find long, interesting patterns that are difficult to identify with frequency-based approaches, and that the selected patterns are robust against noise and/or dynamic changes.

When monitoring massive data streams in real time it is typically impossible to store all data due to its large size. Therefore, in such situations, the aim is to collect in single pass sufficient summary information about data streams to allow subsequent analysis. This problem is particularly challenging when streams consist of multidimensional data where one must compute more complex multidimensional and correlated aggregates. In *Time-decayed Correlated Aggregates over Data Streams*, Graham Cormode, Srikanta Tirthapura, and Bojian Xu develop space lower bounds for approximating time-decayed correlated sum (DCS) aggregates on a data stream and propose the first streaming algorithm for estimating the DCS of a stream using limited memory. This result is an important step toward better understanding which multidimensional queries can be answered on massive data streams using limited memory and computation.

In *A Family of Large Margin Linear Classifiers and Its Application in Dynamic Environments*, Jianqiang Shen and Thomas G. Dietterich study learning algorithms that are able to efficiently handle large-scale datasets and rapidly adapt to changes in the set of categories, their definitions, and their relative frequencies. Toward this objective, they propose three online algorithms that learn efficiently by combining large margin training with regularization methods that enable rapid adaptation to nonstationary environments. For new algorithms, authors prove error bounds with respect to an optimal online algorithm and show that these algorithms have some interesting characteristics that make them especially appropriate in dynamic environments. In particular, the proposed online learning algorithms exhibit feature selection ability, they naturally shrink the influence of the old instances and put more weight on the more recent ones, and they learn a sparse model that ignores or downweights irrelevant features.

In *Topic Modeling for OLAP on Multidimensional Text Databases: Topic Cube and Its Applications*, Duo Zhang, ChengXiang Zhai, Jiawei Han, Ashok Srivastava, and Nikunj Oza address the problem of explosive growth of textual information and suggest to simultaneously analyze both structured data records and unstructured text data. They propose a new data model, Topic Cube, to combine online analytical processing (OLAP) and probabilistic topic modeling. Topic Cube extends the traditional data cube to cope with a topic hierarchy and store probabilistic content measures of text documents obtained through a probabilistic topic model. A main challenge to Topic Cube is processing efficiency. This is addressed by designing two heuristic aggregations to speed up the iterative EM algorithm for estimating topic models. The proposed aggregations leverage the models learned on component data cells to choose a good starting point for iteration. In addition to experimental results and analysis, the authors also suggest potential uses of Topic Cube.

Competence-conscious Associative Classification by Adriano Veloso, Mohammed Zaki, Wagner Meira Jr, and Marcos Goncalves starts with an interesting finding that associative classifiers produced by different metrics may provide conflicting prediction performance, and the best metric to use is data dependent and rarely known in the classifier design phase. The authors phrase this uncertainty of the optimal match as a dilemma and set out to address it with an idea to learn each metric's domain of competence. They investigate stacking-based meta-learning methods that use training data to learn the domain of competence for each metric, and thus propose to create competence-conscious associative classifiers. Using various datasets and evaluation measures, the authors report experimental findings with respect to accuracy. This work should be very helpful in solving real-world problems where accuracy is a key performance measure.

The best papers selected for this special issue are representatives of many excellent papers published at SDM'09. The readers are encouraged to visit the nicely organized, online, open access proceedings at http://www.siam.org/proceedings/datamining/2009/dm09.php. All proceedings of previous SDM conferences (2001–2009) are available cost free at http://www.siam.org/proceedings/. We as the program chairs of SDM'09 are certain that many of SDM'09 papers will become influential in the near future in data mining research and applications. We take this opportunity to thank all the program committee members and external reviewers for their expert help in the challenging task

of reviewing, discussing, and recommending papers. We wholeheartedly appreciate the altruistic help from the area chairs, James Bailey, Ramana Davuluri, Guozhu Dong, Jennifer Dy, Minos Garofalakis, Joydeep Ghosh, Marina Meila, Dino Pedreschi, Jian Pei, Shashi Shekhar, Ashok Srivastava, Brani Vidakovic, Haixun Wang, Takashi Washio, Stefan Wrobel, Jianping Zhang, and Zhi-Hua Zhou who handled the reviewing process with great care and insight. We are grateful to the Best Paper Committee for their help and to Haesun Park, Srinivasan Parthasarathy, and Chandrika Kamath for offering guidance on many program-related issues.

Finally, we wish you all enjoy reading these fascinating and inspiring papers as we have.

Zoran Obradovic[1] and Huan Liu[2]

[1]*Information Science and Technology Center and Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA.* [2]*Computer Science and Engineering, School of Engineering, Arizona State University, Tempe, AZ, USA*