A Panorama of Imminent Doctoral Research in Data Mining

Aparna S. Varde
Dept. of Computer Science
Montclair State University
Montclair, NJ, USA
vardea@montclair.edu

Nikolaj Tatti
HIIT, Dept. of Information and Computer Science
Aalto University, Finland
nikolaj.tatti@aalto.fi

ABSTRACT

As databases head towards data streams, discovering knowledge from the data poses challenges. The need to process and mine data is affected by the big data wave, advances in Web technology and other factors. The advent of the cloud with related technologies provides further momentum to enhance knowledge discovery. Data mining is also of interest to research communities outside computer science as there is a need to harvest data from various domains incorporating domain-specific factors. These and other issues motivate PhD students to pursue core and applied research in data mining along with stream data management, big data, cloud computing, Web knowledge discovery, domain-specific techniques and more. A PhD forum on data mining provides an excellent platform for doctoral students to present imminent research and get valuable feedback from experts. In addition it gives them the opportunity to disseminate the results of their work among fellow researchers, and publish their novel contributions at an early stage. IEEE ICDM hosts such a PhD forum for doctoral students with a data mining focus. This article describes the content of the work presented at the ICDM 2013 PhD forum. It also gives a brief overview of the organization of this forum. The article thus provides a panorama of recent doctoral student work in data mining that would be of interest to researchers in database management and related areas.

1. INTRODUCTION

The IEEE International Conference on Data Mining (ICDM) hosted its third PhD forum on December 7, 2013. This was after two PhD forums in 2011 [http://webdocs.cs.ualberta.ca/~icdm2011/phd-forum.php], and again the following year in 2012 [http://icdm2012.ua.ac.be/content/phd-forum]. The ICDM 2013 PhD forum [1] was held in conjunction with the main ICDM conference at Dallas, Texas from December 7 to 10, 2013. The PhD forum was co-chaired by Dr. Aristides Gionis, Associate Professor in the Department of Information and Computer Science at Aalto University, Finland, and Dr. Aparna Varde, Associate Professor in the Department of Computer Science at Montclair State University, USA. Dr. Nikolaj Tatti, a Post-Doctoral Researcher in Computer Science at Aalto University, Finland, served as a session chair in the forum.

The keynote talk was given by Dr. Jilles Vreeken from Max Planck Institute for Informatics, Saarbrucken, Germany. Dr. Vreeken is a Senior Researcher in their Databases and Information Systems Group and an Independent Research Group Leader of their Exploratory Data Analysis Group. His talk titled "Don't Panic: The Grad Student's guide to a PhD in Data Mining" [1] served as a motivating and humorous account of the road ahead for an early PhD student. He provided statistical data on the usefulness of a PhD, compared it with other advanced degrees and emphasized the milestones and challenges along the PhD path, with helpful career advice. He also stressed the importance of choosing the right advisor, research areas and core dissertation focus. This involved keeping in mind the students' passion for research as well as the job market in order to tap the practical angle.

This PhD forum attracted around 25 research paper submissions from various parts of the world. Among these, five submissions were selected as full papers and four as short papers. All papers were provided with oral and poster presentations, the full papers with a time slot of 20 minutes and short papers with 10 minutes followed by an interactive question-answer session for each paper. During this, useful feedback was offered by the audience comprising established research professionals and other PhD students. The PC members comprised a team of experts in data mining from academia and industry. There were 19 PC members from 10 countries across the globe.

A highlight of this forum was that despite unexpected weather conditions termed as a statistical outlier in Dallas during December, 100% of the speakers

attended and presented their work. This was a sheer example of the inspiration that the PhD forum provided to the doctoral students at the IEEE conference ICDM.

The work of the students presented at this forum spanned various areas including stream data mining, knowledge discovery from big data and domain-specific issues. The papers focused on topics such as MapReduce for classification, anomaly detection, data stream clustering, mining health records, financial news quantification, local distance metrics, discrete pattern mining, dynamically evolving concepts and time-sensitive route-planning.

The themes of stream data mining and domain-specific data mining seemed to be prevalent among many of the presentations. This indicated the enthusiasm among current PhD students to explore streaming data in addition to the traditional databases addressing challenges such as infinite length, scalability, feature evolution and concept drift in continuous data streams. The fact that domain-specific data mining formed the theme of several papers indicated that data mining research has spread across multiple disciplines and that applied research in data mining has acquired significance in addition to core research. The theme of big data also seemed to be noticed among some papers. This implied that knowledge discovery from data of the order of terabytes and more in conjunction with classical data mining techniques has gained importance among data mining researchers.

Based on the papers presented at the PhD Forum, the rest of this article is organized as follows. We outline a survey of the papers from the forum in Section 2. This is placed in the categories of core research and applied research in data mining respectively based on the primary contributions of the papers. Section 3 presents a discussion including a list of open issues that provide the scope for further research in data mining and related areas. Section 4 gives the conclusions along with the motivation to organize more such events. The acknowledgments and references appear thereafter.

2. SURVEY OF PAPERS

We divide the papers into two categories, namely, “Core Research” and “Applied Research”, respectively, based on their major impact. The Core Research papers are those that have their primary focus on contributing to data mining techniques. The papers in the Applied Research area are the ones that propose novel adaptations of data mining addressing the challenges therein leading to significant contributions.

2.1 Core Research in Data Mining

Anomaly Detection with Clustering: The issue of detecting anomalies is an important aspect of computer security and was addressed in the paper by Mustafa et al. The authors proposed learning techniques in the area of clustering for host-based anomaly detection [1]. In one technique CMN (clustering with Markov network), they clustered benign or secure data in the training phase and from each cluster, built an individual Markov network for modeling benign behavior. In the testing phase each Markov network found the probability of every testing instance. If this probability as calculated from many Markov networks was low, the concerned point was classified as malicious or insecure. Other techniques proposed were CMN-OS (clustering with Markov networks with outlying subspace) and CLP (Clustered Label Propagation). In their experimental evaluation they proved that these approaches were not very sensitive to noise and outperformed other state-of-the-art methods for anomaly detection.

Multi Density Clustering for Streams: Another paper that dealt with clustering was on MuDi-Stream by Amini et al. [1], a multi density-based clustering algorithm to handle streaming data with noise. Streams are continuous and need to be processed with limited time and memory. Also, data of varying densities needs to be clustered. Both these needs were addressed in the proposed algorithm MuDi-Stream which performed clustering in online and offline phases. The online phase developed core-mini-clusters with a new proposed core distance based on number of data points around the core. The offline phase conducted clustering on the core-mini-clusters with a density-based method. Since the algorithm had different core distances for different clusters, it covered some multi-density environments.

MapReduce for Stream Classification: The work by Haque and Khan [1] further delved into the issue of stream data mining. The infinite length and evolving nature of data streams poses challenges in mining. These challenges were addressed in this paper by a multi-tiered ensemble-based method. Several AdaBoost ensembles were built for each numeric feature upon receiving each chunk of the data stream. This was expected to cause scalability problems for huge data chunks and/or too many numeric attributes. Thus, the authors exploited the parallelism of MapReduce in order to propose two approaches to build ensembles such that they enhanced scalability, yet also maintained accuracy. Experiments on benchmark datasets indicated that these approaches were indeed efficient, scalable and accurate.

Tracking Evolving Data Streams: The mining of data streams is affected by factors such as feature evolution, concept drift and novel class emergence. This requires rapid labeling and tracking of such dynamically evolving streams, which was addressed in the paper by Parker and Khan [1]. Feature evolution consists of features getting added, removed or altered in ranges. Concept drift implies that the concepts defining class labels can change over the span of the data. Previously unknown classes can also appear in the data stream which is called novel class emergence. The authors developed approaches for tracking and labeling these streams while adhering to the required constraints of the continuous data. They developed an adaptive supervised ensemble to predict instance labels and a stream clustering approach to monitor new characteristics defining the concepts and new classes that emerge accordingly. Thus, new classes with unknown labels were not treated as noise but instead appropriately labeled. They tested the accuracy and efficiency of their data stream mining approaches by comparison with baseline methods on benchmark data streams.

Discrete Pattern Mining with Matrices: Jiang and Health [1] considered binary matrix factorization where the goal was to approximate a binary matrix as a product of two binary matrices. They argued that a natural approach to force the product to be binary was to force one of the matrices have only one per column. They claimed that this made the problem equivalent to clustering and adopted the standard Lloyd's algorithm for discovering such matrices. They tested their method on synthetic and image datasets, and also used the approximate matrices for mining patterns.

2.2 Applied Research in Data Mining

Discriminative Metric for Applications: Mu and Ding [1] considered learning a local discriminative distance metric for real-world applications. More specifically, they focused on discovering Mahalanobis distance that simultaneously minimized distances of neighboring points belonging to the same class while maximizing the distances of neighboring points belonging to a different class. Using these local distances, the authors constructed a kNN-style classifier and applied it to crater prediction from images, crime prediction, and accelerometer data.

Financial News Quantification: Minev [1] presented his ongoing work on the topic of quantifying financial news. The author considered announcements from Federal Reserve from which he extracted features using Natural Language Processing techniques. Once these features were extracted, the goal was to predict the movement of S&P500, a major stock index, within a day of the published announcement.

Time-sensitive Route Planning: Hsieh et al. [1] proposed a framework for recommending tourist routes. In their approach they constructed a score of a route by taking into account the popularity of the place, the most popular visit time, and transition times between the locations. Given a query, a starting point, the authors proposed a heuristic algorithm, and conducted experiments with several baseline algorithms. The authors also conducted a user study and assessed the users' satisfaction for the proposed routes.

EHR Mining: Lo et al. [1] studied how measuring adverse drug reactions can be discovered in electronic health records (EHR). A standard way of detecting such drug reactions was through a spontaneous reporting system. While discovering correlations between drug and symptoms was straightforward in this system as the symptom and the drug is reported in the same report, it was less trivial when electronic health records were used since the data was collected over a period of time. The authors designed a method for discovering such correlations and tested their method on a synthetic electronic health records dataset where they obtained effective results.

3. DISCUSSION

The PhD Forum in ICDM 2013 was organized for the third time and was a highly successful event after two similar events in 2011 and 2012 respectively. It involved presentations from PhD students on a range of topics in data mining and related areas. These included core research in areas such as learning techniques, big data mining and knowledge discovery algorithms as well as applied research on various topics such as electronic health records, route planning, financial news and other real world applications.

Some of the problems discussed at the forum presented the scope for further research in areas such as stream data mining and domain-specific problems. A few potential areas for future work were gathered from the presentation of the papers and the question-answer sessions that occurred thereafter. These are listed herewith as follows:

- Scientific data mining taking into account sensitive information with aspects such as security
- Discovery of knowledge from sensitive data by cloud mining approaches

- Enhancement of stream data mining by integration of clustering and classification
- Advances in big data management and mining with respect to streaming data
- Techniques to address privacy issues in mining electronic health records
- Adaptation of route planning and other location-specific approaches to mobile devices
- Multilingual processing in knowledge discovery for applications such as financial news

These and other related topics could lead to more interesting findings in data mining research.

4. CONCLUSIONS

This article provides a panorama of the research by upcoming doctoral students in data mining. It would be of interest to students and professionals in data mining, databases and related areas. More details of this imminent doctoral research can be found in the respective papers in the proceedings of the ICDM 2013 PhD Forum.

We hope that some of the open research issues emerging from the papers herewith lead to further research via the PhD students' dissertation sub-problems. This would also promote more interactions among data mining researchers through future work in the concerned areas.

We aspire that the PhD Forum remains an annual event at ICDM and is even more successful in the forthcoming years. This would further serve as the motivation to continue organizing events of this nature in various existing and upcoming data mining and database conferences.

5. ACKNOWLEDGMENTS

The authors thank Dr. Aristides Gionis for co-chairing the PhD forum and ICDM 2013 organizers Dr. Diane Cook, Dr. Bhavani Thuraisingham and Dr. Xingdong Wu for hosting this event. We also thank Dr. Jilles Vreeken for giving an encouraging keynote talk. We express our gratitude towards all the PC members for reviewing papers. Finally, we thank the doctoral students for presenting their work at the PhD forum and making it an exciting event.

6. REFERENCES

[1] Aristides Gionis, Aparna Varde eds., ICDM PhD Forum, http://icdm2013.rutgers.edu/phd-forum, 2013.