Novel Data Stream Pattern Mining
Report on the StreamKDD'10 Workshop

Margaret H. Dunham
Southern Methodist University
Dallas, TX, USA
mhd@lyle.smu.edu

Michael Hahsler
Southern Methodist University
Dallas, TX, USA
mhahsler@lyle.smu.edu

Myra Spiliopoulou
Otto-von-Guericke-University
Magdeburg, Germany
myra@iti.cs.uni-magdeburg.de

ABSTRACT

This report summarizes the First International Workshop on Novel Data Stream Pattern Mining held at the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, on July 25 2010 in Washington, DC.

1. INTRODUCTION

Data stream mining has gained in importance over the last years because it is indispensable for many real applications such as prediction and evolution of weather phenomena; security and anomaly detection in networks; evaluating satellite data; and mining health monitoring streams. Stream mining algorithms must take into account the unique properties of stream data, including temporal ordering, concept drifts and shifts, demand for scalability and that the data are unbounded.

This workshop brought together scholars working in different areas of learning on streams, including sensor data and news streams. Most of the contributions were on unsupervised learning with clustering methods. Issues addressed included the detection of outliers and anomalies, evolutionary clustering and incremental clustering, learning in subspaces of the complete feature space and learning with exploitation of context, deriving models from text streams and visualizing them.

The workshop received 15 submissions, from which 8 papers were accepted after a desk evaluation and a review process by two to three reviewers. The best paper of the workshop (see below) was selected for inclusion in this issue, as an extended version. All other workshop papers are available in the ACM Digital Library.

The workshop was well attended with lively discussions and question sessions.

In this report we discuss the works presented at the workshop, grouped into the following thematic areas: data mining on multiple data streams (section 2), detection of concept drift and outliers (section 3), and stream clustering (section 4).

2. MULTIPLE DATA STREAMS

Dr. Le Gruenwald, IIS Program Director at NSF and Professor at the University of Oklahoma delivered the Keynote speech for the workshop. MDS involves examining disparate heterogeneous data sources with different characteristics and timing constraints. After introducing the concept of multiple data streams (MDS), she discussed some of the complex research issues associated with mining of MDS.

Much work as been devoted to MDS within the database community, but there is still too little research on mining these streams. Dr. Grunewald urged data mining researchers to start to address some of these complex problems.

In fact, MDS is the subject of the Best Paper of the workshop. In "Fully Decentralized Computation of Aggregates over Data Streams" [2], Beccetti et al proposed a novel technique to perform aggregate operations on multiple data streams. The environment examined is one where many different nodes each receive data from a data stream and need to combine them for delivery to a collecting agent or coordinating node. To this purpose, each node keeps a summary, sketch, of the stream data it has seen. Sketch updates are propagated to neighbor nodes when the sketch values are deemed to be sufficiently different from the last propagation. In this manner, information concerning aggregation values ultimate migrates to all nodes in the network.

3. CONCEPT DRIFT AND OUTLIER DETECTION

Data streams usually exhibit concept drifts that occur slowly, or shifts that emerge suddenly. A crucial problem is distinguishing between a change of the concept and the appearance of outliers.

At StreamKDD, we had three papers devoted to this subject. In [3] the authors outlined the approach they are currently working on for the detection of outliers: they use a density based approach to create an outlier score for an incoming point and decide whether a data point is an outlier, before the next one arrives. The authors of [4] propose CALDS, a context based learning technique that detects concept drift by exploiting context. The authors focus on context recurrence within the stream. To address this problem they propose to learn multiple models and apply the most fitting one when the context changes. New models are created if the arriving data are too dissimilar to all existing models.

The third paper in this area examined the prediction of anomalies in the trajectories of ships [5]. For this trajectory surveillance application, the authors use conformal prediction.

4. DATA STREAM CLUSTERING

As a maxim, data stream clustering should be online, incremental, and single pass. Three StreamKDD papers investigated different aspects of this topic.

Evolutionary clustering was examined in [6], focusing on the changes encountered in a sequence of clusters over text data, such as news feeds. Remarkably, the core idea is in combining stream clustering with frequent itemset mining over the feature space. In [7], the authors propose a clustering approach that adapts density based clustering to subspace clustering using a preference weighted distance function, by extending the algorithm PREDECON [9] into an incremental method.

The last workshop paper proposed a visualization technique for news streams [8], based on heat maps, whereby heat models the frequency of occurrence of articles in different clusters over time.

5. CONCLUSION

The first StreamKDD workshop at the ACM SIGKDD International Conference in 2010 has been motivated by research advances on data streams that were reported in the earlier KDD conferences, adjoint workshops and tutorials. The intention of StreamKDD'10 has been to bring the researchers who contribute to the field together for the sharing of ideas and the identification of new research challenges. The contributing authors, the evaluating reviewers, and the engaged audience made StreamKDD'10 a success.

6. ACKNOWLEDGMENTS

We would like to thank all authors who submitted papers to this workshop. Their work and their presentations were the key to the workshop's success. We are grateful to the members of the Program Committee for their thorough and dedicated work: *Sanjay Chawla, João Gama, Le Gruenwald, Eamonn Keogh, Latifur Khan, Chi-Hoon Lee, Mohammad Masud, Tamer Özsu, Spiros Papadimitriou, Thomas Seidl, Dimitris Tasoulis.*

We are also indebted to the KDD'10 Workshop Chairs for their support in realizing this workshop.

The organization of this workshop was supported in part by the National Science Foundation under Grant No. IIS-0948893.

7. REFERENCES

[1] Wu, W. and Gruenwald, L. 2010. Research issues in mining multiple data streams. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 56-60.

[2] Becchetti, L., Bordino, I., Leonardi, S., and Rosen, A. 2010. Fully decentralized computation of aggregates over data streams. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 1-9.

[3] Assent, I., Kranen, P., Baldauf, C., and Seidl, T. 2010. Detecting outliers on arbitrary data streams using anytime approaches. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 10-15.

[4] Gomes, J. B., Menasalvas, E., and Sousa, P. A. 2010. CALDS: context-aware learning from data streams. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 16-24.

[5] Laxhammar, R. and Falkman, G. 2010. Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 47-55.

[6] Shankar, R., Kiran, G. V., and Pudi, V. 2010. Evolutionary clustering using frequent itemsets. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 25-30.

[7] Kriegel, H., Kröger, P., Ntoutsi, I., and Zimek, A. 2010. Towards subspace clustering on dynamic data: an incremental version of PreDeCon. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 31-38.

[8] Krstajić, M., Bertini, E., Mansmann, F., and Keim, D. A. 2010. Visual analysis of news streams with article threads. In Proceedings of the First international Workshop on Novel Data Stream Pattern Mining Techniques (Washington, D.C., July 25, 2010). StreamKDD '10. ACM, New York, NY, 39-46.

[9] Bohm, C., Kailing, K., Kriegel, H. P., and Kroger, P. 2004. Density Connected Clustering with Local Subspace Preferences, Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM-04), 27-34.

About the authors:

Margaret H. Dunham is a Professor of Computer Science and Engineering at the Lyle School of Engineering, Southern Methodist University, Dallas, TX. She is a Distinguished ACM speaker and co-director of the school's Intelligent Data Analysis (IDA) research group.

(http://lyle.smu.edu/~mhd/)

Michael Hahsler is a Visiting Assistant Professor of Computer Science at the Lyle School of Engineering, Southern Methodist University, Dallas, TX. He is also a co-director of the school's Intelligent Data Analysis (IDA) research group. His research interests are data stream clustering, association rule mining, recommender systems and data visualization.

(http://michael.hahsler.net)

Myra Spiliopoulou is a Professor of Business Information Systems at the Faculty of Computer Science, Otto-von-Guericke University, Magdeburg, Germany and Head of the "Knowledge Management and Discovery" group. Her research interests are on data mining, text mining and web mining for dynamic environments, including the analysis of stream data for social platforms and business applications.

(http://omen.cs.uni-magdeburg.de/itikmd)