Knowledge Discovery from Sensor Data (SensorKDD)

Olufemi A. Omitaomu	Auroop R. Ganguly	Joao Gama
Oak Ridge National Laboratory	Oak Ridge National Laboratory	University of Porto
1, Bethel Valley Road	1, Bethel Valley Road	Rua de Ceuta
Oak Ridge, TN 37831, USA	Oak Ridge, TN 37831, USA	Porto, Portugal
omitaomuoa@ornl.gov	gangulyar@ornl.gov	jgama@fep.up.pt

Ranga Raju Vatsavai	Nitesh V. Chawla	Mohamed Medhat Gaber
Oak Ridge National Laboratory	University of Notre Dame	Monash University
1, Bethel Valley Road	Department of Computer Science and	900 Dandenong Road
Oak Ridge, TN 37831, USA	Engineering, IN, USA	Melbourne, Victoria, 3145, Australia
vatsavairr@ornl.gov	nchawla@cse.nd.edu	mohamed.gaber@infotech.monash.edu.au

ABSTRACT

Wide-area sensor infrastructures, remote sensors, RFIDs, and wireless sensor networks yield massive volumes of disparate, dynamic, and geographically distributed data. As such sensors are becoming ubiquitous, a set of broad requirements is beginning to emerge across high-priority applications including adaptability to climate change, electric grid monitoring, disaster preparedness and management, national or homeland security, and the management of critical infrastructures. The raw data from sensors need to be efficiently managed and transformed to usable information through data fusion, which in turn must be converted to predictive insights via knowledge discovery, ultimately facilitating automated or human-induced tactical decisions or strategic policy based on decision sciences and decision support systems.

Keeping in view the requirements of the emerging field of knowledge discovery from sensor data, we took initiative to develop a community of researchers with common interests and scientific goals, which culminated into the organization of SensorKDD series of workshops in conjunction with the prestigious ACM SIGKDD International Conference of Knowledge Discovery and Data Mining. In this report, we summarize events at the Third ACM-SIGKDD International Workshop on Knowledge Discovery form Sensor Data (SensorKDD 2009).

1. INTRODUCTION

As wide-area sensor infrastructures, remote sensors, RFIDs, and wireless sensor networks are becoming ubiquitous, the challenges for the *knowledge discovery* community are expected to be immense. On the one hand, dynamic data streams or events require real-time analysis methodologies and systems, while on the other hand centralized processing through high end computing is also required for generating offline predictive insights, which in turn can facilitate real-time analysis. The online and real-time knowledge discovery imply immediate opportunities as well as intriguing short- and long-term challenges for practitioners and researchers in knowledge discovery. The opportunities would be to develop new data mining approaches and adapt traditional and emerging knowledge discovery methodologies to the requirements of the emerging problems. In addition, emerging societal problems require knowledge discovery solutions that are designed to investigate anomalies, hotspots, changes, extremes and nonlinear processes, and departures from the normal. The theme for the 2009 SensorKDD workshop is around three inter-related global priorities: *climate change, energy assurance, and infrastructural impacts*. The workshop brings together researchers from academia, government, and the industry working in various aspects of knowledge discovery from sensor data.

1.1 Motivation

The expected ubiquity of sensors in the near future, combined with the critical roles they are expected to play in high priority application solutions, point to an era of unprecedented growth and opportunities. The requirements described earlier imply immediate opportunities as well as intriguing short- and long-term challenges for practitioners and researchers in knowledge discovery. In addition, the knowledge discovery and data mining (KDD) community would be called upon, again and again, as partners with domain experts to solve critical application solutions in business and government, as well as in the domain sciences and engineering.

The main motivation for this workshop stems from the increasing need for a forum to exchange ideas and recent research results, and to facilitate collaboration and dialog between academia, government, and industrial stakeholders. Based on the positive feedback from the previous workshop attendees and our own experiences and interactions with the government agencies such as the United States Department of Energy, United States Department of Homeland Security, United States Department of Defense, and involvement with numerous projects on knowledge discovery from sensor data, we strongly believe in the continuation of this workshop. We also believe that the ACM SIGKDD conference is the right forum to organize this workshop as it brings the KDD community together in this important area to establish a much needed leadership position in research and practice in the near term, as well as in the long term.

1.2 Previous SensorKDD Workshops

The previous two workshops – SensorKDD'07 and SensorKDD'08 – held in conjunction with the 13th and 14th ACM

SIGKDD Conference on Knowledge Discovery and Data Mining respectively attracted several participants as well as many high quality papers and presentations. The 2007 workshop was attended by more than seventy registered participants. The workshop program includes presentations by authors of six accepted full papers and four invited speakers. The invited speakers were Prof. Pedro Domingos of the University of Washington, Prof. Joydeep Ghosh of the University of Texas, Austin, Prof. Hillol Kargupta of the University of Maryland, Baltimore County, and Dr. Brian Worley of the Oak Ridge National Laboratory (ORNL). There were also poster presentations by authors of six accepted short papers. The extended versions of papers presented at the workshop were developed into a book titled *Knowledge Discovery from Sensor Data* [1], the first book published in this specific discipline.

The 2008 workshop was attended by more than 60 registered participants. There were presentations by authors of seven accepted full papers and six accepted short papers; the workshop program also include presentations by two invited speakers – Prof. Jiawei Han of the University of Illinois at Urbana-Champaign and Dr. Kendra Moore of the Defense Advanced Research Projects Agency. The extended versions of papers presented at the 2008 workshop are scheduled for publication as *Springer's LNCS post-proceedings* in 2009.

2. SUMMARY OF THE 2009 WORKSHOP

The highlights of the SensorKDD 2009 workshop are oral presentations of accepted papers, presentations by invited speakers, and oral presentations by the winners of the first SensorKDD Cup, similar in spirit to the KDD Cup, but in line with the workshop theme.

Based on a minimum of two reviews per paper, we selected *six full* paper and *eight short* papers. In addition to the oral presentations of accepted papers, the workshop featured three invited speakers: Dr. Aurelie Lozano of IBM T.J. Watson Research Center, Mr. Alessandro Donati of the European Space Agency, Darmstadt, Germany, and Prof. Carlos Guestrin of Carnegie Mellon University. For the SensorKDD Cup, we challenged contestants with intriguing problems along with geographically distributed and dynamic data from sensors and model simulations. The datasets were provided and/or compiled by leveraging the Oak Ridge National Laboratory resources. Two winners were selected to present their results at the workshop. All papers presented at the workshop are available in [2].

2.1 Full Research Papers

Below is a list of accepted full papers and their respective authors:

- Handling Outliers and Concept Drift in Online Mass Flow Prediction in CFB Boilers by *J. Bakker, M. Pechenizkiy, I. Zliobaite, A. Ivannikov, and T. Karkkainen (Best Paper Award Winner).*

- An Exploration of Climate Data Using Complex Networks by *Karsten Steinhaeuser, Nitesh V. Chawla, and Auroop R. Ganguly (1st Best Student Paper Award Winner).*

- A Comparison of SNOTEL and AMSR-E Snow Water Equivalent Datasets in Western U.S. Watersheds by *Cody L. Moser, Oubeidillah Aziz, Glenn A. Tootle, Venkat Lakshmi, and Greg Kerr.*

- EDISKCO: Energy Efficient Distributed in-Sensor-Network K-center Clustering with Outliers by *Marwan Hassani, Emmanuel Muller, and Thomas Seidl.*

- Phenological Event Detection from Multitemporal Image Data by *Ranga Raju Vatsavai.*

- Mining in a Mobile Environment by *Sean McRoskey, James Notwell, Nitesh V. Chawla, and Christian Poellabauer (2nd Best Student Paper Award Winner).*

2.2 Short Research Papers

The following is a list of accepted short papers and their respective authors:

- On the Identification of Intra-seasonal Changes in the Indian Summer Monsoon by *Shivam Tripathi and Rao S. Govindaraju.*

- Reduction of Ground-Based Sensor Sites for Spatio-Temporal Analysis of Aerosols by *Vladan Radosavljevic, Slobodan Vucetic, and Zoran Obradovic.*

- OcVFDT: One-class Very Fast Decision Tree for One-class Classification of Data Streams by *Chen Li, Yang Zhang, and Xue Li.*

- A Frequent Pattern Based Framework for Event Detection in Sensor Network Stream Data by *Li Wan, Jianxin Liao, and Xiaomin Zhu.*

- Supervised Clustering via Principal Component Analysis in a Retrieval Application by *Esteban Garcia-Cuesta, Ines M. Galvan, and Antonio J. de Castro.*

- A Novel Measure for Validating Clustering Results Applied to Road Traffic by *Yosr Naija and Kaouther Blibech Sinaoui.*

- SkyTree: Scalable Skyline Computation for Sensor Data by *Jongwuk Lee and Seung-won Hwang.*

- Clustering of Power Quality Event Data Collected via Monitoring Systems Installed on the Electricity Network by *Mennan Guder, Nihan Kesim Cicekli, Ozgul Salor, and Isik Cadirci.*

2.3 SensorKDD'09 Cup Winners

The titles and authors of the SensorKDD'09 Cup winners are:

- Change Detection in Rainfall and Temperature Patterns over India by *Shivam Tripathi and Rao S. Govindaraju.*

- Anomaly Detection and Spatio-Temporal Analysis of Global Climate System by *Mahashweta Das and Srinivasan Parthasarathy.*

3. CONCLUSIONS

Extracting knowledge and emerging patterns from sensor data is a nontrivial task. The challenges for the knowledge discovery community are expected to be immense. As evidenced from the participation and quality of submissions to the previous three SensorKDD workshops, it is clear that the "Knowledge Discovery from Sensor Data or SensorKDD" is a growing area and an important specialty (sub-area) within knowledge discovery. The

SensorKDD workshop is proven to be an attractive forum for the researchers from academia, industry and government, to exchange ideas, initiate collaborations and lay foundation to the future of this important and growing area. The workshop witnessed lively participation from all quarters, generated interesting discussions immediately after each presentation and as well as at the end of the workshop. All participants agreed for continued patronage for the SensorKDD workshop. In addition to the ACM workshop proceedings, extended papers will be published as post workshop proceedings in Springer's well-known "Lecture Notes in Computer Science" series.

4. SPONSORSHIP

The SensorKDD'09 workshop was sponsored by the Geographic Information Science and Technology (GIST) Group at Oak Ridge National Laboratory, the Computational Sciences and Engineering (CSE) Division at the Oak Ridge National Laboratory, and Cooperating Objects Network of Excellence (CONET).

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Auroop R. Ganguly, Joao Gama, Olufemi A. Omitaomu, Mohamed M. Gaber, and Ranga Raju Vatsavai (2009). *Knowledge Discovery from Sensor Data*. New York, NY: CRC Press, January.

[2] Olufemi A. Omitaomu, Auroop R. Ganguly, Joao Gama, Ranga Raju Vatsavai, Nitesh V. Chawla, and Mohamed Medhat Gaber (2009). *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*. Paris, France: ACM Digital Library. Available online: http://portal.acm.org/toc.cfm?id=1601966&coll=GUIDE&dl=GUIDE&type=proceeding&idx=SERIES939&part=series&WantType=Proceedings&title=KDD&CFID=49782169&CFTOKEN=54456709.

About the Workshop Organizers:

Dr. Olufemi A. Omitaomu is a research scientist in the Computational Sciences and Engineering Division at the Oak Ridge National Laboratory. He is also an adjunct assistant professor at the University of Tennessee, Knoxville. His research interests include streaming and real-time data mining, infrastructure modeling and analysis, machine learning, and signal processing. He received Ph.D. in information engineering from the University of Tennessee. He has published in top peer-reviewed journals and conferences; co-organized and co-chaired workshop and sessions at professional conferences including the ACM Workshop on Knowledge Discovery from Sensor Data held in conjunction with ACM SIGKDD 2007 and ACM SIGKDD 2008. He previously worked as a data analyst with Mobil Exploration and Production Company for more than five years.

Dr. Auroop R. Ganguly is a research scientist within the Computational Sciences and Engineering division of the Oak Ridge National Laboratory since 2004. His research interests are climate change impacts, geoscience informatics, civil and environmental engineering, computational data sciences, and knowledge discovery. Prior to ORNL, he has more than five years of experience in the software industry, specifically Oracle Corporation and a best-of-breed company subsequently acquired by Oracle, and about a year in academia, specifically at the University of South Florida in Tampa. He has a PhD from the Civil and Environmental Engineering department of the Massachusetts Institute of Technology, several years of research experience with a group at the MIT Sloan School of Management, experience in private consulting, and a wide range of peer-reviewed publications spanning multiple disciplines. Currently, he is also an adjunct professor at the University of Tennessee in Knoxville.

Dr. Joao Gama is a researcher at LIAAD-INESC Porto LA, the Laboratory of Artificial Intelligence and Decision Support of the University of Porto. His main research interest is learning from Data Streams. He has published several articles in change detection, learning decision trees from data streams, hierarchical clustering from streams, among others. Editor of special issues on Data Streams in Intelligent Data Analysis, Journal of Universal Computer Science, and New Generation Computing. Co-chair of ECML 2005 Porto, Portugal 2005, Conference chair of Discovery Science 2009, and of a series of Workshops on Knowledge Discovery in Data Streams, ECML 2004, Pisa, Italy, ECML 2005, Porto, Portugal, ICML 2006, Pittsburg, US, ECML 2006 Berlin, Germany, and SAC2007, Korea. Together with Dr. Mohamed M. Gaber edited the book Learning from Data Streams Processing Techniques in Sensor Networks, published by Springer.

Dr. Ranga Raju Vatsavai has been conducting research in the area of spatiotemporal databases and data mining for the past 15 years. Before joining the Oak Ridge National Laboratory (ORNL) as a research scientist, he worked at IBM-Research (2004-06; IIT-Delhi campus), U of Minnesota (1999-2004; Twin-cities campus, MN), AT&T Labs (1998; Middletown, NJ), Center for Development of Advanced Computing (1995-98; C-DAC, U of Pune campus, India), and National Forest Data Management Center (1990-95; FRI Campus, Dehradun, India). He has published over thirty peer-reviewed articles and served on program committees of several international conferences (KDD, ICTAI, SSTDM). He was also involved in the design and development of several highly successful software systems (UMN-MapServer - a world leading open source WebGIS, *Miner - a spatiotemporal data mining workbench, EASI/PACE classification modules, and first parallel softcopy photogrammetry system for IRS-1C/1D satellites). His broad research interests are centered on spatial, spatiotemporal databases and data mining, and computational geoinformatics; in particular he is interested in statistical pattern recognition, semi-supervised learning, multiple classifier systems, time series analysis and forecasting, information retrieval, uncertainty and error handling.

Dr. Nitesh V. Chawla is an assistant professor at the University of Notre Dame. Dr. Chawla's research interests are broadly in the areas of data mining, machine learning, pattern recognition, and their applications. More specifically his research has focused on learning from massive datasets, distributed data mining/machine learning, ensemble techniques, cost/distribution sensitive learning, feature selection, and semi-supervised learning. His research has also focused on the inter-disciplinary applications such as intelligent scientific visualization, biometrics, bioinformatics, natural language processing, and customer analytics.

Dr. Mohamed Medhat Gaber is a research Fellow at Monash University, Australia. He has published more than 60 papers. Mohamed is the co-editor of the book: Learning from Data Streams: Processing Techniques in Sensor Networks, published by Springer in 2007. His research interests include data stream mining, wireless sensor networks and context-aware computing. Mohamed has served in the program committees of several international and local conferences and workshops in the area of data mining and context-aware computing. He was the co-chair of the IEEE International Workshop on Mining Evolving and Streaming Data held in conjunction with ICDM 2006, International Workshop on Knowledge Discovery from Ubiquitous Data Streams held in conjunction with ECML/PKDD 2007, and the First and Second International Workshop on Knowledge Discovery from Sensor Data held in conjunction with ACM SIGKDD 2007/2008.