Knowledge Discovery from Sensor Data (SensorKDD)

Varun Chandola[1], Olufemi A. Omitaomu[1], Auroop R. Ganguly[1], Ranga R. Vatsavai[1], Nitesh V. Chawla[2], Joao Gama[3], Mohamed M. Gaber[4]

[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA

[2]University of Notre Dame, Department of Computer Science & Engineering, IN, USA

[3]University of Porto, Rua de Ceuta, Porto, Portugal

[4]University of Portsmouth, Hampshire, UK

{chandolav, omitaomuoa, gangulyar, vatsavairr}@ornl.gov, nchawla@cse.nd.edu, jgama@fep.up.pt, mohamed.gaber@port.ac.uk

ABSTRACT

Sensor data is being collected at an unprecedented rate across a variety of domains from a broad spectrum of sources, such as wide-area sensor infrastructures, remote sensing instruments, RFIDs, and wireless sensor networks. With the recent proliferation of smart-phones, and similar GPS enabled mobile devices, collection of sensor data is no longer limited to scientific communities, but has reached general public. With massive volumes of such disparate, dynamic, and geographically distributed data available, many high-priority applications have been identified that involve analysis of such data to solve real world problems such as understanding climate change and its impacts, electric grid monitoring, disaster preparedness and management, national or homeland security, and the management of critical infrastructures.

Given the unique characteristics of sensor data, particularly its spatiotemporal nature and presence of constraints associated with the data collection and computational resources, there have been many research efforts to analyze the sensor data which build upon the general research in the data mining community but are significantly different in terms of how they address the specific challenges encountered when dealing with sensor data. In particular, the raw data from sensors needs 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 Fourth ACM-SIGKDD International Workshop on Knowledge Discovery form Sensor Data (SensorKDD 2010).

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. According to the data mining and domain experts present at the NSF-sponsored Next Generation Data Mining Summit (NGDM '09) held in October 2009, "finding the next generation of solutions to these challenges is critical to sustain our world and civilization" [1]. Some of the organizers of the SensorKDD workshop were part of the summit. The 4th International Workshop on Knowledge Discovery from Sensor Data (SensorKDD-2010) is a step in bringing researchers together to address these challenges and moving toward the development of the next generation data mining solutions require to address these challenges.

The theme for the 2010 SensorKDD workshop is around four focus areas: offline and online knowledge discovery from sensor data, decision and policy support, theoretical models for sensor and streaming data, and case studies in national and global priority applications. 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 motivation for the SensorKDD-workshop in conjunction with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining 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. The expected ubiquity of sensors in the 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 first workshop was organized in 2007. 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 Homeland Security, United States Department of Defense, and involvement with numerous projects on knowledge discovery from sensor data, has encouraged us to continue this workshop. We believe that the ACM SIGKDD conference provides 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 three workshops – SensorKDD'07, SensorKDD'08, SensorKDD'09 – held in conjunction with the 13th, 14th, and 15th 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.

The 2009 workshop was attended by several registered participants. There were presentations by authors of eight accepted full papers, eight accepted short papers, two entries for the SensorKDD-2009 cup, and three invited speakers. The invited speakers were Prof. Carlos Guestrin of Carnegie Mellon University, Dr. Aurelie Lozano of IBM T.J. Watson Research Center, and Mr. Alessandro Donati of the European Space Agency. The extended versions of papers presented at the 2009 workshop are scheduled for publication as Springer's LNCS post-proceedings in 2010. The best paper, two best student papers, and two SensorKDD-2009 cup entries were awarded certificates and cash prizes. The prizes were donated by the Computational Sciences and Engineering Division of the Oak Ridge National Laboratory and Cooperating Objects Network of Excellence of the European Union. The workshop was partially sponsored by the Geographic Information Science and Technology Group of CSED at ORNL and Cooperating Objects Network of Excellence (CONET).

2. SUMMARY OF THE 2010 WORKSHOP

The highlights of the SensorKDD 2010 workshop were oral presentations of accepted papers and an invited talk.

The workshop featured one invited speaker: Prof Philip S. Yu of University of Illinois at Chicago, that talks about Mining Data Streams. Prof. Yu refers in the abstract of his talk: "The problem of streaming data has become increasingly importance in recent years. The ubiquitous presence of data streams in a number of practical domains has generated a lot of research in this area. Sensor network is a typical example of data stream applications. Problems such as data mining which have been widely studied for traditional data sets cannot be easily solved for the data stream domain. This is because the large volume of data arriving in a stream renders most algorithms to inefficient as most mining algorithms require multiple scans of data which is unrealistic for streaming data. More importantly, the characteristics of the data stream can change over time and the evolving pattern needs to be captured. Furthermore, the stream data can often be noisy as in sensor data streams. This talk will provide an overview, discuss the issues and focus on how to mine uncertain data streams and massive graph streams."

In addition to the Invited speaker, the PC selected, based on a minimum of two reviews per paper, five full papers and six short papers. The eleven papers spanned a broad spectrum of applications using sensor data. Interestingly, one of the most popular applications targeted by the authors was *smart and energy aware* buildings, which follows the general interest in sustainability across the world. Papers by Chikhaoui et al., Chen et al., Rashidi & Cook, and Marwah et al., all focused on problems associated with monitoring energy consumption in homes and data centers. Another new area that attracted interest among authors as well as the workshop audience was in the area of human activity monitoring using sensors attached to the human body. Papers by Kwapisz et al. and Srinivasan et al. applied machine learning methods to predict the activities of individuals using sensor data. In particular, the paper by Kwapisz et al. showed how the activity of an individual can be monitored using data collected using a smartphone (*Iphone etc.*), which highlights the possibilities for knowledge discovery research from data collected by such widely used devices. In addition, there were several papers that dealt with classical sensor network related problems, such as distributed and embedded computing, as described in papers by Djuric & Vucetic and Rodriguez et al., among others. The data mining methods used by the authors also broadly encompass all key areas including clustering [Rodriguez et al., Hassani et al.], classification [Kwapisz et al., Srinivasan et al., Chen et al., Rashidi & Cook, Djuric & Vucetic], frequent pattern mining [Chikhouai et al.], and anomaly detection [Marwah et al., Jiang et al.]. All papers presented at the workshop are available in [4].

2.1 Full Research Papers

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

- Activity Recognition using Cell Phone Accelerometers
  Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore
- A New Algorithm Based on Sequential Pattern Mining for Person Identification in Ubiquitous Environments
  Belkacem Chikhaoui, Shengrui Wang, and Helene Pigot

- Activity Recognition Using Actigraph Sensor
  Raghavendiran Srinivasan, Chao Chen, and Diane Cook
- Network Comprehension by Clustering Streaming Sensors
  Pedro Pereira Rodrigues, Joao Gama, Joao Araujo, and Luis Lopes
- Energy Prediction Based on Resident's Activity
  Chao Chen, Barnan Das, and Diane Cook

2.2 Short Research Papers

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

- Multi Home Transfer Learning for Resident Activity Discovery and Recognition
  Parisa Rashidi and Diane Cook
- Using Semantic Annotation for Knowledge Extraction from Geographically Distributed and Heterogeneous Sensor Data
  Alexandra Moraru, Carolina Fortuna, and Dunja Mladenic
- Random Kernel Perceptron on ATTiny2313 Microcontroller
  Nemanja Djuric and Slobodan Vucetic
- Anomalous Thermal Behavior Detection in Data Centers using Hierarchical PCA
  Manish Marwah, Ratnesh Sharma, Wilfredo Lugo, and Lola Bautista
- Self-Organizing Energy Aware Clustering of Nodes in Sensor Networks using Relevant Attributes
  Marwan Hassani, Emmanuel Muller, Pascal Spaus, Adriola Faqolli, Themis Palpanas, and Thomas Seidl
- Anomaly Localization by Joint Sparse PCA in Wireless Sensor Networks
  Ruoyi Jiang, Hongliang Fei, and Jun Huan

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 "Lecture Notes in Computer Science" series.

4. SPONSORSHIP

The SensorKDD'10 workshop was sponsored by the Geographic Information Science and Technology (GIST) Group at Oak Ridge National Laboratory within the Computational Sciences and Engineering Division at the Oak Ridge National Laboratory.

5. ACKNOWLEDGMENTS

We would like to thank our sponsors for their kind donations. We would like to thank the authors of all submitted papers and presenters. Their innovation and creativity has resulted in a strong technical program. We are highly indebted to the program committee members, whose reviews ensured the development of a competitive and strong technical program. The program committee listed in alphabetical order of last names are: Adedeji B. Badiru, Budhendra L. Bhaduri, Eric Auriol, Albert Bifet, Michaela Black, Jose del Campo-Avila, Andre Carvalho, Sanjay Chawla, Diane Cook, Alfredo Cuzzocrea, Jing (David) Dai, Christie Ezeife, David J. Erickson III, Yi Fang, Francisco Ferrer, James H. Garrett, Joydeep Ghosh, Bryan L. Gorman, Sara Graves, Ray Hickey, Forrest Hoffman, Luke (Jun) Huan, Volkan Isler, Vandana Janeja, Yu (Cathy) Jiao, Ralf Klinkenberg, Miroslav Kubat, Vipin Kumar, Mark Last, Chang-Tien Lu, Elaine Parros Machado de Sousa, Sameep Mehta, Laurent Mignet, S. Muthu Muthukrishnan, George Ostrouchov, Guangzhi Qu, Rahul Ramachandran, Pedro Rodrigues, Josep Roure, Bernhard Seeger, Cyrus Shahabi, Shashi Shekhar, Mallikarjun Shankar, Lucio Soibelman, Alexandre Sorokine, Eduardo J. Spinosa, Karsten Steinhaeuser, Nithya Vijayakumar, Gary Weiss, Peng Xu, Eiko Yoneki, and Philip S. Yu.

We would like to thank our invited speakers, Professor Philip S. Yu from University of Illinois at Chicago, who, despite his busy schedules, readily agreed and delivered highly motivating and informative talk. We would like to thank, Dr. Brian Worley, Director, Computational Sciences and Engineering Division (CSED), Oak Ridge National Laboratory (ORNL), for his encouragement, support, and continued patronage of SensorKDD workshop series, and Dr. Budhendra Bhaduri, Group Leader, Geographic Information Science and Technology, CSED, ORNL, for his enthusiastic support and sponsorship.

This workshop report was compiled by Dr. Varun Chandola of the Computational Sciences and Engineering Division at Oak Ridge National Laboratory. The workshop report has been co-authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains, and the publisher by accepting the article for publication, acknowledges that the United States Government retains, a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

6. REFERENCES

[1] Bhat, C., Ganguly A.R., Gehrke, J., Giannella, C., McGranaghan, M., and Melby, P. (2009). National Science Foundation Summit on the Next Generation of Data Mining for Dealing with Energy, Greenhouse Emissions, and Transportation Challenges (NGDM '09), Committee Report, November 2009 – With contributions from Olufemi A. Omitaomu (Unpublished).

[2] 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.

[3] 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.

[4] Olufemi A. Omitaomu, Varun Chandola, Auroop R. Ganguly, Joao Gama, Ranga Raju Vatsavai, Nitesh V. Chawla, and Mohamed Medhat Gaber (2010). *Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data*. Washington, D. C.

About the Workshop Organizers:

Dr. Varun Chandola is a post doctoral research associate within the Computational Sciences and Engineering division of the Oak Ridge National Laboratory since 2009. His current responsibilities include applying spatio-temporal knowledge discovery and machine learning techniques to detect and analyze unexpected changes in crop growing patterns. Dr. Chandola received his PhD in Computer Science from the University of Minnesota. His dissertation was titled "Anomaly detection for symbolic sequences and time series data". His areas of expertise include data mining, time series data analysis, machine learning, and algorithm development. His research focus is in the area of anomaly detection applied to spatiotemporal data. He has applied his research, with significant success, to varying application domains, such as aviation safety, cyber intrusion detection, tax fraud analysis, cardiac health monitoring, and click fraud analysis.

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. He received a PhD in information engineering from the University of Tennessee. His research interests include machine learning and data mining, signal processing, time series analysis, and spatial data modeling with applications to smart electrical grids, smart communities, renewable energy, and energy storage. His research has also been applied to other domains including transportation security, process monitoring and control, and crime data analysis. He previously worked as a data analyst with Mobil Exploration and Production Company for more than five years.

Dr. Auroop R. Ganguly is a senior research scientist within the Computational Sciences and Engineering division of the Oak Ridge National Laboratory in Oak Ridge, TN. In addition, he is a joint faculty with Civil and Environmental Engineering and an adjunct professor with Industrial and Information Engineering at the University of Tennessee in Knoxville. His research interests are climate extremes and uncertainty as well as interdisciplinary computational data sciences. Prior to ORNL, he has several years of professional experience in the private industry and academia. He obtained a PhD from Civil and Environmental Engineering at the Massachusetts Institute of Technology.

Dr. Joao Gama is a senior 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, in conjunction with ECML-PKDD, and ACM-SAC. He recently published the book Knowledge Discovery from Data Streams CRC Press.

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 senior lecturer at University of Portsmouth, UK. He has published more than 60 papers. He 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.