Intelligent Data Analysis 13 (2009) 403–404
DOI 10.3233/IDA-2009-0372
IOS Press

Introduction

Knowledge discovery from data streams

João Gama[a], Auroop Ganguly[b], Olufemi Omitaomu[b], Raju Vatsavai[b] and Mohamed Gaber[c]
[a]*LIAAD-University of Porto, Portugal*
[b]*Oak Ridge National Laboratory, TN, US*
[c]*Monash University, Australia*

Wide-area sensor infrastructures, remote sensors, and wireless sensor networks yield massive volumes of disparate, dynamic, and geographically distributed data. As sensors are becoming ubiquitous, a set of broad requirements is beginning to emerge across high-priority applications including disaster preparedness and management, adaptability to climate change, 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. The challenges for the Knowledge Discovery community are immense. Sensors produce dynamic data streams or events requiring real-time analysis methodologies and systems. Moreover, in most of the cases these streams are distributed in space, requiring spatio-temporal knowledge discovery solutions.

All these aspects are of increasing importance to the research community, as new algorithms are needed to process this continuously flow of data in reasonable time. Learning from data streams require algorithms that process examples in constant time and memory, usually scanning data once. Moreover, if the process is not strictly stationary (as most of real world applications), the target concept could gradually change over time. This is an incremental task that requires incremental learning algorithms that take drift into account.

For this special issue of Intelligent Data Analysis we selected 5 papers from the accepted papers of the Fourth International Workshop on Knowledge Discovery from Data Streams, an associated workshop of the 18th European Conference on Machine Learning (ECML) and the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), co-located in Warsaw, Poland, 2007 and the First ACM SIGKDD Knowledge Discovery from Sensor Data – SensorKDD07, co-located with the Knowledge Discovery and Data Mining (KDD) 2007 conference organized by the American Computing Machinery (ACM).

The selected papers cover a large spectrum in the research of Knowledge Discovery from Data Streams that goes from recommendation algorithms, Clustering, Drifting Concepts and Frequent pattern mining. The common concept in all the papers is that learning occurs while data continuously flows.

The first paper *Novelty Detection with Application to Data Streams* by Spinosa, Carvalho and Gama, presents and evaluates a new approach to novelty detection from data streams. The ability to detect novel concepts is an important aspect of a machine learning system, essential when dealing with non-stationary distributions. The approach presented here intends to take novelty detection beyond one-class

classification, by detecting emerging cohesive and representative clusters of examples, and further by merging similar concepts. The proposed technique goes in the direction of constructing a class structure that aims at reproducing the real one in an unsupervised continuous learning fashion.

The paper *Context-Aware Adaptive Data Stream Mining* by Haghighia, Zaslavskya, Krishnaswamya, Gaber, and Lokeb presents a general approach for context-aware adaptive mining of data streams that aims to dynamically and autonomously adjust data stream mining parameters according to changes in context and situations. Adaptation of data stream processing to variations of data rates and availability of resources is crucial for consistency and continuity of running applications. The authors perform real-time analysis of data streams generated from sensors that is under-pinned using context-aware adaptation.

The paper *Anomaly Detection using Manifold Embedding and its Applications in Transportation Corridors* by Agovic, Banerjee, Ganguly, and Protopopescu studies the problem of detection anomalous cargo based on sensor readings in truck weigh stations. The paper shows the relevance of appropriate feature representation for anomaly detection methods in high dimensionality and noisy domains as in transportation corridors. The authors experimentally show the usefulness of manifold embedding methods for feature representation in these problems.

The paper *Spatio-Temporal Sensor Graphs* (*STSG*)*: A Data Model for the Discovery of Spatio-Temporal Patterns* by George, Kang and Shekhar presents a new data model called Spatio-Temporal Sensor Graphs (STSG), which is designed to model sensor data on a graph by allowing the edges and nodes to be modeled as time series of measurement data. Case studies illustrate the ability of the STSG model to find patterns like hotspots in sensor data.

Finally, the paper, *A System for Analysis and Prediction of Electricity-Load Streams* by Rodrigues and Gama, presents novel methodological adaptations for data streams in the context of a real-world application domain, specifically, electricity load forecasting based on dynamic sensor information. Incremental algorithms are developed or utilized for clustering and change detection, learning of neural networks for predicting at multiple lead times, and improving predictive accuracy based on Kalman filters.

In summary, the five selected papers represent some of the latest research in an emerging field with rapid and exciting growth. Learning from Sensor Data poses new challenges to the data mining community. The present issue reports the current state of the research in Knowledge Discovery from Data Streams.

The editors would like to thank Intelligent Data Analysis journal, the Editor-in-Chief Professor Fazel Famili, and the authors who submitted their work. A special word to the anonymous reviewers for their collaboration. This work was developed under the auspices of KdUbiq-WG3 and project Adaptive Leaning Systems II (POSC/EIA/55340/2004).