Intelligent Data Analysis 12 (2008) 251–252
IOS Press

Introduction

Knowledge discovery from data streams

João Gama[a], Jesus Aguilar-Ruiz[b] and Ralf Klinkenberg[c]
[a]*LIAAD-University of Porto, Porto, Portugal*
[b]*Polytechnic Pablo de Olavide University, Seville, Spain*
[c]*University of Dortmund, Dortmund, Germany*

Traditional pratice in machine learning algorithms involve fixed data sets and static models. Most of the times, all the data is loaded into memory and the learning task is solved by performing multiple scans over the training data. These assumptions fail with the advent of new application areas, like ubiquitous computing, sensor networks, e-commerce, etc., where data flows continuously, eventually at high speed rate. Other examples include scientific data, customer click streams, telephone records, large sets of web pages, multimedia data, sets of retail chain transactions, etc. These sources of continuous data are called data streams.

Data streams are increasingly important in the research community, as new algorithms are needed to process this streaming data in reasonable time. Learning from data streams require algorithms that process examples in constant time and memory, usualy scaning 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.

Many researchers coming from different areas (data mining, machine learning, OLAP, databases, etc.) are designing new approaches or adapting some of the traditional algorithms to data streams. The number of researchers in this field is also growing considerably, and, in many conferences, data streams are becoming a consolidated topic.

For this special issue of Intelligent Data Analysis we selected 4 papers from the accepted papers for the Fourth International Workshop on Knowledge Discovery from Data Streams, an associated workshop of the 17th European Conference on Machine Learning (ECML) and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), co-located in Berlin, Germany, 2006.

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.

In the first paper, *Schema Matching on Streams with Accuracy Guarantees* by S. Jaroszewicz, L. Ivantysynova, and T. Scheffer, address the problem of matching imperfectly documented schemas from data streams. The paper *Modeling Dynamic Substate Chains among Massive States* by V. Nguyen, and T. Washio, proposes a framework for handling high-dimensional data from large-scale transactional data wharehouses. The paper *Mining Frequent Items in a Stream Using Flexible Windows* by T. Calders, N.

Dexters, and B. Goethals study the problem of finding frequent items in a continuous stream of itemsets, introducing a new frequency measure, based on a flexible window length. Finaly, the paper *Improving the Performance of an Incremental Algorithm Driven by Error Margins* by J. del Campo-Ávila, G. Ramos-Jiménez, J. Gama and R. Morales-Bueno discuss relevant issues for incremental classification learning, a relevant topic of data stream learning algorithms.

In summary, the four selected papers represent some of the latest research in an emerging field with rapid and exciting growth. Learning from Data Streams 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).