Guest Editorial

Knowledge discovery from data streams

João Gama[a] and Jesus Aguilar-Ruiz[b]

[a]*LIACC-University of Porto, Portugal*
E-mail: jgama@fep.up.pt
[b]*School of Engineering, Pablo de Olavide University, Seville, Spain*
E-mail: jsagurui@upo.es

The rapid expansion in information science and technology in general and the complexity and volume of data in particular have introduced new challenges for the research community. New data are incessantly added to databases. In many cases, we need to extract useful knowledge from this continuous stream of data. Examples include customer click streams, telephone records, large sets of web pages, multimedia data, and sets of retail chain transactions. These sources of continuous data are called data streams.

Standard machine learning algorithms work on static data. 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 flow continuously, eventually at high speed rate.

Data streams are increasingly important in the research community, as new algorithms are needed to process these streaming 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 takes 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 is becoming a consolidated topic.

For this special issue of Intelligent Data Analysis we selected five works from the accepted papers for the Second International Workshop on Knowledge Discovery from Data Streams, a workshop associated to the 16th European Conference on Machine Learning (ECML) and the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), co-located in Porto, Portugal, 2005.

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 flow.

In the first paper, *An Ensemble Classifier for Drifting Concepts*, the author proposes a multiple model approach for detecting and reacting to concept drift. The method uses a boosting-like weighting schema to train classifiers.

Algorithms for some critical applications involving Data Streams provide faster answers using approximate solutions. Approximate solutions require estimates that bound the admissible error. Two papers focus on this topic. The paper *Statistical Supports for Mining Sequential Patterns and Improving the Incremental Update Process on Data Streams* studies bounds for mining sequential patterns. The paper *Approximate Mining Frequent Patterns on Streams* introduces new algorithms for approximate mining frequent itemsets from streams of transactions using a limited amount of memory.

The paper *Collaborative Filtering on Data Streams* presents an online recommendation algorithm that maintains decision models from the incoming data stream. The system is able to adapt itself to changes of the user's interest.

An emergent area in data streams involves analysing data emanating from distributed and heterogeneous sources in the form of a continuous stream with mobile and/or embedded devices. This is the topic of the last paper *A Fuzzy Approach for Interpretation and Application of Ubiquitous Data Stream Clustering*. The authors present a new approach to automate the annotation of results obtained from ubiquitous data stream clustering to facilitate their interpretation and use, enabling real-time and mobile decision making.

In summary, the five selected papers represent some of the latest research in an emerging field with rapid and exciting growth. Knowledge Discovery from Data Streams poses new challenges both for database technology and data mining. The present issue reports the current state of the research in the field. We hope it can contribute to disseminate and promote the research in the area.