Journal of Universal Computer Science, vol. 11, no. 8 (2005), 1349-1352
submitted: 25/7/05, accepted: 29/7/05, appeared: 28/8/05

Data Streams
J.UCS Special Issue

Jesús S. Aguilar–Ruiz
Department of Computer Science
University of Seville, Spain
aguilar@lsi.us.es

João Gama
LIACC–University of Porto, Portugal
jgama@fep.up.pt

1 Introduction to Data Streams

Many sources produce data continuously. Examples include customer click streams, telephone record calls, large sets of web pages, multimedia and scientific data, sets of retail chain transactions, etc. These sources are called data streams. Examples of applications include network monitoring, web applications, sensor networks, telecommunications data management, financial applications, etc. In this applications it is not feasible to load the arriving data into a traditional database management system (DBMS). Traditional DBMS are not designed to directly support the continuous queries required in these applications, in which data cannot be modelled as persistent tables, but rather as transient data streams.

In [1] the authors enumerate relevant issues of the data stream model:

– The data elements in the stream arrive online.

– The system has no control over the order in which data elements arrive, either within a data stream or across data streams.

– Data streams are potentially unbound in size.

– Once an element from a data stream has been processed it is discarded or archived. It cannot be retrieved easily unless it is explicitly stored in memory, which is small with relation to the size of the data streams.

From the point of view of a data stream management system several research issues emerge:

– Approximate query processing techniques to evaluate queries that require unbounded amount of memory.

– Sliding window query processing both as an approximation technique and as an option in the query language.

– Sampling to handle situations where the flow rate of the input stream is faster than the query processor.

– The meaning and implementation of blocking operators (e.g. aggregation and sorting) in the presence of open–ended streams.

Solutions to these problems require new sampling and randomize techniques, and new approximate and incremental algorithms. Data mining offers several algorithms for these problems, and learning from data streams pose new problems to data mining. In [2] the authors present desirable properties for learning in data streams: incrementality, online learning, constant time to process each example, single scan over the training set, take drift into account. In this context is highly unlikely the assumption that the examples are randomly generated according to a stationary probability distribution. At least in complex systems and for large time periods, we should expect changes in the distribution of the examples. A natural approach for this *incremental tasks* are *adaptive learning algorithms*, incremental learning algorithms that take into account *concept drift*. P. Domingos and G. Hulten [2] identified desirable properties of learning systems that are able to mine continuous, high–volume, open–ended data streams as they arrive.

– Require small constant time per data example.

– Use fix amount of main memory, irrespective of the total number of examples.

– Built a decision model using a single scan over the training data.

– Any time model.

– Independence from the order of the examples.

– Ability to deal with concept drift. For stationary data, ability to produce decision models that are nearly identical to the ones we would obtain using a batch learner.

These ideas were the basis for the *Very Fast Machine Learning* toolkit[1]. VFML contains tools for learning decision trees, for learning the structure of belief nets (so–called Bayesian networks), and for clustering.

[1] http://www.cs.washington.edu/dm/vfml/

2 Overview of the Special Issue

Nowadays all these problems are the main stream of several research projects, international workshops, PhD thesis, etc. For this special issue we have selected the best papers from international workshops organized in conjunction with the European Conference on Machine Learning, European Conference on Principles and Practice of Knowledge Discovery in Databases and the Symposium of Applied Computing. Overall, they present the most relevant directions of research in data streams.

In *Learning Decision Trees from Dynamic Data Streams*, the authors presents a system for induction of forest of functional trees from data streams able to detect concept drift. They present Ultra Fast Forest of Trees (UFFT) system, that is an incremental algorithm, which works online, processing each example in constant time, and performing a single scan over the training examples. It uses analytical techniques to choose the splitting criteria, and the information gain to estimate the merit of each possible splitting–test. UFFT implements mechanisms to detect and react to concept drift by monitoring the online error of naive Bayes installed in inner nodes of the trees.

In *Network Attack Scenarios Extraction and Categorization by Mining IDS Alert Streams* the authors present a complete system to extract and categorize attack scenarios from Intrusion Detection System alerts, a semantic vector space model based on First–order Logic. The system aggregates distributed alerts into streams of alerts, reducing unneeded computations, and applies a modified Case Grammar to unify heterogeneous alerts. The attack scenarios are then resolved into attack semantic space vectors by an attack ontology and alert contexts. Text categorization is then applied to categorize intrusion stages.

In *Semantic Preprocessing of Web Request Streams for Web Usage Mining* the authors focus on two main tasks, semantic outlier detection from online Web request streams and segmentation (or sessionization) of them. They thereby exploit semantic technologies to infer the relationships among Web requests.

In *Evaluating Trigger Conditions on Streaming Time Series with User-given Quality Requirements* the authors propose a framework for designing trigger condition evaluation system that considers user–specified quality requirements. This framework uses statistical analysis to derive the likelihood of a condition to be true at a time position. By using this likelihood and the associated confidence (due to finite sampling), they estimated the quality of approximate answers.

In *Online Mining Changes of Items over Continuous Append-only and Dynamic Data Streams* the authors propose new algorithms, called MFC–append and MFC–dynamic, for mining frequent frequency changed items, vibrated frequency–changed items, and stable frequency changed items over continuous append–only and dynamic data streams, respectively. A new summary data structure, called Change–Sketch, is developed to store the frequency changes between two data streams as fast as possible.

In *Incremental Rule Learning and Border Examples Selection from Numerical Data Streams* the authors propose FACIL, an incremental rule learning algorithm with partial instance memory based on moderate generalization that may store updated border examples to avoid unnecessary revisions when virtual drifts are present in data. Consistent rules classify new test examples by covering and inconsistent rules classify them by distance as the nearest neighbour algorithm. In addition, the algorithm provides an implicit forgetting heuristic so that these examples are removed when they do not describe a decision boundary.

Finally, in *Resource-aware Mining of Data Streams* the author presents a novel adaptable approach that can cope with the challenging inherent features of data streams, paying attention to the data stream rate with respect to the available resources. Results in a resource constrained environment show the applicability and scalability of this clustering based approach.

3 Final Words

In summary, the seven papers selected represent some of the latest research in an emerging field with rapid and exciting growth. Data Streams pose new challenges both for database technology and data mining. This special issue is a contribution to the present state–of–art of the research.

The editors would like to thank the authors and the anonymous reviewers for their collaboration.

References

1. Babcock B., Babu S., Datar M., Motwani R., and Widom J. Models and issues in data stream systems. In Phokion G. Kolaitis, editor, *Proceedings of the 21nd Symposium on Principles of Database Systems*, pages 1–16. ACM Press, 2002.
2. Geoff Hulten and Pedro Domingos. Catching up with the data: research issues in mining data streams. In *Proc. of Workshop on Research issues in Data Mining and Knowledge Discovery*, 2001.