Special Session on "Online Real-Time Strategies for Data Stream Mining" for 2017 IEEE 2017, Hawai

Plamen P. Angelov, Chin-Teng Lin, Mahardhika Pratama, Sundaram Suresh, Edwin Lughofer, Mukesh Prasad

Aims and Scope:

Data stream mining is among the most challenging research issues in the machine learning community. A large amount of data is generated in many today's real-world applications with fast rate, thereby resulting in the so-called data explosion. Furthermore, data streams do not follow static and predictable data distributions rather change overtime with no specific patterns. The two issues require significant innovation in the machine learning area, because conventional machine learning algorithms are inherent with the problems of offline working principle, slow learning speed, lack of adaptive mechanism and over-dependency on manual intervention. There is significantly growing research interest in the literature to propose an autonomous learning machine (ALMA), which is capable of self-evolving both their structure and parameters from data streams with minor manual intervention. The system should be computationally efficient to be scalable against large-scale applications and capable of dealing with online life-long learning environments. Nevertheless, research in the ALMA area is still at an infant stage and deserves in-depth study to deal with the issue of uncertainty in data streams:

1) *Concept Drifts*: Because of dynamic natures of real-world problems, concept change is among the most prevalent issue in data streams. A real-world problem may contain various concept drifts and a learning algorithm should not be constrained with how slow, rapid, gradual, recurrent, incremental or otherwise changes in data streams. Currently applied ALMAs in the literature are usually targeted to solve an abrupt concept drift but not yet mature to address other concept drifts in accordance to its rates and types.

2) *Data Processing*: The ALMAs usually adopt the single-pass learning mode, which discards directly a data sample once learned. Although this learning scenario has a low computational complexity and memory demand, which is independent from the number of training samples, it is not fully efficient to handle data streams, because it still learns all incoming samples regardless of their true contribution to the training process. As a matter of fact, training samples can be redundant to the training process and learning such samples are not recommended because it leads to the overfitting issue. In addition, the single-pass learning mode still requires training samples to be fully labelled and this issue causes expensive labelling cost.

3) *Curse of Dimensionality*: Vast majority of ALMAs assume that input attributes are pre-selected in the pre-processing step. This fact undermines the concept of online learner, which should put forward the notion of plug-and play system. That is, all learning components are embedded in a single training process without pre-and/or post-training steps. Furthermore, the feature selection plays important role to lower computational burden, while improving accuracy because it avoids the high variance problem.

4) *Data Representation*: Because of disagreement of expert knowledge, inaccurate measurement, and noisy environment, data may be biased from their true representation. This contradicts crisp and certain characteristics of existing ALMAs, which entails a precise parameter identification step.

This special session aims to bring together research works of online real-time strategies for

data stream mining. Special attention will be devoted to handle advanced issues of data stream mining.

Topics:

The main topics of this special session include, but are not limited to, the following:

- Online real-time unsupervised learning and clustering for large data streams
- Online real-time supervised classification and regression for large data streams
- Online real-time time-series modelling for large data streams
- Online real-time intelligent controller for large data streams
- Appropriate handling of data uncertainty in learning from large data streams
- Tools and techniques for data stream mining in uncertain environments
- Computational intelligence methods for big data analytics and huge data bases
- Techniques to address drifts and shifts in data streams
- On-line dynamic dimension reduction in high-dimensional streams
- Feature selection and extraction techniques for large data streams
- Sample selection and active learning for large data streams
- Robustness and safety aspect in learning from large data streams
- Practical applications of computational intelligence techniques for data stream mining
- Real world cases of uncertainties in large data stream

Special Session Organizers:

1. Prof. Plamen P. Angelov, Lancaster University, UK
2. Prof. Chin-Teng Lin, University of Technology, Sydney
3. Asst. Prof. Mahardhika Pratama, Nanyang Technological University, Singapore
4. Assoc. Prof. Sundaram Suresh, Nanyang Technological University, Singapore
5. Dr. Edwin Lughofer, Johannes Kepler University, Austria
6. Dr. Mukesh Prasad, University of Technology, Sydney, Australia

Paper Submission

A paper should be submitted through IEEE SSCI's submission central (http://www.ele.uri.edu/ieee-ssci2017/PaperSubmission.html). After logging into the submission system, you need to choose *Special Session on "Online Real-Time Strategies for Data Stream Mining"*. Paper submission deadline is

July 2, 2017

We look forward to receiving your high-quality submissions.