Annals of Telecommunications (2020) 75:491–492
https://doi.org/10.1007/s12243-020-00811-1

EDITORIAL

IoT data stream analytics

Albert Bifet[1,2] · João Gama[3]

Published online: 16 September 2020

1 Introduction

The volume of IoT data is rapidly increasing due to the development of the technology of information and communication. This data comes mostly in the form of streams. Learning from this ever-growing amount of data requires flexible learning models that self-adapt over time. Traditional one shot memory-based learning methods trained offline from a static historic data cannot cope with evolving data streams. This is because firstly, it is not feasible to store all incoming data over time and secondly the generated models become quickly obsolete due to data distribution changes, also known as "concept drift." The basic assumption of offline learning is that data is generated by a stationary process and the learning models are consistent with future data. However, in multiple applications like IoT, web mining, social networks, network monitoring, sensor networks, telecommunications, financial forecasting, etc., data samples arrive continuously as unlimited streams often at high speed. Moreover, the phenomena generating these data streams may evolve over time. In this case, the environment in which the system or the phenomenon generated the data is considered to be dynamic, evolving, or non-stationary.

Learning methods used to learn from data generated by dynamically evolving and potentially non-stationary processes must take into account many constraints: (pseudo) real-time processing, high-velocity, and dynamic multiform change such as concept drift and novelty. In addition in data streams scenarios, the number of classes is often unknown in advance. Therefore, new classes can appear at any time and they must be detected, and the predictor structure must be updated.

It is worthwhile to emphasize that streams are very often generated by distributed sources, especially with the advent of Internet of Things, and, therefore, processing them centrally may not be efficient, particularly if the infrastructure is large and complex. Scalable and decentralized learning algorithms are potentially more suitable and efficient.

This special issue aims at discussing the problem of learning from IoT data streams generated by evolving non-stationary processes. It centers on the advances of techniques, methods, and tools that are dedicated to manage, exploit, and interpret data streams in non-stationary environments. In particular, it focuses on the problems of modeling, prediction, and classification based on learning from data streams.

2 The selected papers

We received several submissions of high interest. The review process helped to select the best ones, guaranteeing the quality of the form and the content and ensuring the scientific rigor and technical correctness. Hereafter, we provide a summary of each paper in this special issue.

The first paper entitled *Regularized and incremental decision trees for data streams* presents a regularization schema for Hoeffding trees, one of the most popular algorithms for decision tree induction over data streams. The main motivation for the regularization is that as new data become available, the tree grows leading to unnecessary complexity and making the tree less interpretable. The new regularization schema checks whether a new split is justifiable and whether it is worthy to use a new splitting attribute.

The second paper entitled *Discovering Locations and Habits from Human Mobility Data* presents a novel method for analyzing human mobility and habit data based on density clustering methods and Gaussian mixed models.

✉ Albert Bifet
albert.bifet@telecom-paris.fr

João Gama
jgama@fep.up.pt

[1] LTCI, Télécom Paris, IP Paris, Paris, France

[2] University of Waikato, Hamilton, New Zealand

[3] Laboratory of Artificial Intelligence and Decision Support, and Faculty of Economics, University of Porto, Porto, Portugal

The paper presents a set of experiments on three real-world datasets: a GPS trajectory collection, a GSM telecom dataset, and a Google location history dataset.

The third paper entitled *Bi-directional Online Transfer Learning: A Frame- work* presents an online transfer learning framework that tracks drift in each domain and uses past knowledge to enhance the prediction capability. The main contribution of the paper consists of the BOTL framework that employs a meta-learner trained with a subset of past knowledge in the source domain as well as the current target classifier and a new drift detection method.

The fourth paper *Resource Management for Model Learning at Entity Level* focuses on problems where streams are composed of data from multiple entities. The authors propose a resource management approach for model learning in that context. Two methods are proposed. The first uses Lossy IoT Data Stream Counting to identify models that should be kept in main memory or not, whereas the second is a naive model that predicts according to a majority label scheme.

The fifth paper entitled *Process mining on machine event logs for profiling abnormal behavior and root cause analysis* proposes a novel methodology for detecting abnormal machine behavior, using process mining, in order to perform cause root analysis on logs created by the machines. The idea is relevant since it tries to understand when a machine will start malfunctioning. The sixth paper *Profiling High Leverage Points for Detecting Anomalous Users in Telecom Data Networks* presents a novel approach for detecting anomalous users in a massive network of calls. It applies clustering on high leverage points to characterize differences between them. The paper also proposes new sociometric features that have a unique range of values for anomalous user profiles. Additionally, we present an analysis that shows the correlations and patterns in the data.

The seventh paper entitled *Interconnect Bypass Fraud Detection: a Case Study* presents a solution for a problem called interconnect bypass fraud, the most effective fraud tactic in the telecommunication domain. The fraudsters explore the forwarding of international calls using low-cost IP connections to increase their profits. This type of fraud is characterized by the occurrence of a burst of calls. The paper explores the use of a new fast forgetting mechanism for the Lossy Counting algorithm. This technique tries to capture as soon as possible abnormal behaviors on routing phone calls.

Finally, the last paper entitled *Active Feature Acquisition on Data Streams under Feature Drift* proposes an active feature acquisition strategy for data streams with feature drift. The main idea of the paper is interesting and important for the data stream domain.

3 Final words

We wish that these articles give the readers a taste of the main trends and current research topics and provide them with an opportunity to explore and collaborate in the related fields. Finally, we would like to thank all the authors and reviewers for the quality of their contribution which made this special issue possible.