IoT Big Data Stream Mining

Gianmarco De Francisci Morales
Qatar Computing Research Institute
gdfm@acm.org

Albert Bifet
LTCI, CNRS
Télécom ParisTech
Université Paris-Saclay
albert.bifet@telecom-paristech.fr

Latifur Khan
Department of Computer Science
University of Texas at Dallas
lkhan@utdallas.edu

Joao Gama
LIAAD-INESC TEC
Faculty of Economics
University of Porto
jgama@fep.up.pt

Wei Fan
Baidu Research
Big Data Laboratory,
Sunnyvale, CA
fanwei03@baidu.com

ABSTRACT

The challenge of deriving insights from the Internet of Things (IoT) has been recognized as one of the most exciting and key opportunities for both academia and industry. Advanced analysis of big data streams from sensors and devices is bound to become a key area of data mining research as the number of applications requiring such processing increases. Dealing with the evolution over time of such data streams, i.e., with concepts that drift or change completely, is one of the core issues in IoT stream mining. This tutorial is a gentle introduction to mining IoT big data streams. The first part introduces data stream learners for classification, regression, clustering, and frequent pattern mining. The second part deals with scalability issues inherent in IoT applications, and discusses how to mine data streams on distributed engines such as Spark, Flink, Storm, and Samza.

CCS Concepts

•**Information systems** → **Data stream mining;**

Keywords

IoT, Big Data, Data Streams, Data Science

1. INTRODUCTION

The Internet of Things (IoT), the large network of physical devices that extends beyond the typical computer networks, will be creating a huge quantity of Big Data streams in real time in the next future. The realization of IoT depends on being able to gain the insights hidden in the vast and growing seas of data available. Since current approaches don't scale to Internet of Things (IoT) volumes, new systems with novel mining techniques are necessary due to the velocity, but also variety, and variability, of such data.

This IoT setting is challenging, and needs algorithms that use an extremely small amount (iota) of time and memory resources, and that are able to adapt to changes and not to stop learning. These algorithms should be distributed and run on top of Big Data infrastructures. How to do this accurately in real time is the main challenge for IoT analytics systems in the near future.

In the IoT data stream model, data arrives at high speed, and algorithms that process it must do so under very strict constraints of space and time. Consequently, data streams pose several challenges for data mining algorithm design. First, algorithms must work within limited resources (time and memory). Second, they must deal with data whose nature or distribution changes over time. We need to deal with resources in an efficient and low-cost way. In data stream mining, we are interested in three main dimensions:

- accuracy
- amount of space (computer memory) necessary
- time required to learn from training examples and to predict

These dimensions are typically interdependent: adjusting the time and space used by an algorithm can influence its accuracy. By storing more pre-computed information, such as look up tables, an algorithm can run faster at the expense of space. An algorithm can also run faster by processing less information, either by stopping early or storing less, thus having less data to process. The more time an algorithm has, the more likely it is that accuracy can be increased.

The outline of the tutorial is the following:

- IoT Fundamentals and IoT Stream Mining Algorithms
  - Stream mining setting
  - Concept drift
  - Classification and Regression
  - Clustering
  - Frequent Pattern mining

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2945385

- Distributed Big Data Stream Mining
  - Distributed Stream Processing Engines
  - Classification
  - Regression

1.1 IoT Fundamentals and IoT Stream Mining Algorithms

In this part we present some basic concepts of IoT data stream mining and classification, regression, clustering and frequent pattern mining for IoT data streams. We will introduce some strategies to deal with concept drift, when it is present, and we will demonstrate basic algorithmic concepts about streams.

1.1.1 IoT Stream mining setting

We start giving some motivation and examples of IoT massive data streams that are continuously being generated. We show examples of how traditional mining methods can not deal with large amounts of data, to motivate the need for specific streaming methods. We give some notion of stream algorithmic complexity, and we show briefly some of the most frequently used approximation techniques in stream mining methods.

1.1.2 Concept drift

We discuss the problem of evolving data over time We discuss the problem of evolving data over time and define concept drift and emerging novel class (concept evolution). We discuss why and when concept drift happens. We outline the most representative approaches to handle concept drift, concept evolution and, in detail, some change detection methods. We also discuss evaluation challenges of adaptive learning methods, and the most common evaluation methodologies.

1.1.3 Classification and Regression

We start by presenting classification algorithms. We show the basic ones, such as the majority class, Naive Bayes, perceptron, and then we motivate the use of more advanced ones, such as decision trees and stochastic gradient descent learners. We give some insights on ensemble methods, as they have several advantages over single classifier methods: they are easy to scale and parallelize, they can adapt to change quickly by pruning under-performing parts of the ensemble, and they therefore usually also generate more accurate concept descriptions.

1.1.4 Clustering

We present recent methods on stream clustering as StreamKM++, CluStream, ClusTree, or Den-Stream.. We discuss cluster evaluation measures. A common classification of these measures is the separation into so called internal measures and external measures. Internal measures only consider the cluster properties, e.g. distances between points within one cluster or between two different clusters. External evaluation measures compare a given clusterings to separately given ground truth.

1.1.5 Frequent Pattern mining

We present recent methods to deal with structured data as itemsets, sequences, trees and graphs.

1.2 IoT Distributed Big Data Stream Mining

In this part we focus on open source software tools for distributed processing used nowadays as Spark, Flink, Storm, Samza, and how to do data stream mining with them.

2. INSTRUCTORS

Gianmarco De Francisci Morales is a Scientist at QCRI. Previously he worked as a Visiting Scientist at Aalto University in Helsinki, as a Research Scientist at Yahoo Labs in Barcelona, and as a Research Associate at ISTI-CNR in Pisa. He is one of the lead developers of Apache SAMOA, an open-source platform for mining big data streams. He co-organizes the workshop series on Social News on the Web (SNOW), co-located with the WWW conference.

Albert Bifet is Associate Professor at Telecom ParisTech and Honorary Research Associate at the WEKA Machine Learning Group at University of Waikato. Previously he worked at Huawei Noah's Ark Lab in Hong Kong, Yahoo Labs in Barcelona, University of Waikato and UPC BarcelonaTech. He is one of the leaders of MOA and Apache SAMOA software environments for implementing algorithms and running experiments for online learning from evolving data streams.

Latifur Khan is a full Professor (tenured) in the Computer Science department at the University of Texas at Dallas where he has been teaching and conducting research since September 2000. He has received prestigious awards including the IEEE Technical Achievement Award for Intelligence and Security Informatics. Dr. Khan is an ACM Distinguished Scientist and a Senior Member of IEEE. He has chaired several conferences and serves (or has served) as associate editor on multiple editorial boards including IEEE Transactions on Knowledge and Data Engineering (TKDE) journal.

Joao Gama Joao Gama is Associate Professor at the Faculty of Economy of the University of Porto and a senior researcher and vice-director of LIAAD, a group belonging to INESC TEC. He served as Co-Program chair of ECML'2005, DS'2009, ADMA'2009, IDA' 2011, and ECM-PKDD'2015. He served as track chair on Data Streams with ACM SAC from 2007 till 2016. He is author of several books in Data Mining (in Portuguese) and authored a monograph on Knowledge Discovery from Data Streams.

Wei Fan is the Deputy Head at Baidu Research Big Data Lab. His main research interests and experiences are in various areas of data mining and database systems, such as, stream computing, high performance computing, extremely skewed distribution, cost-sensitive learning, risk analysis, ensemble methods, easy-to use nonparametric methods, graph mining, predictive feature discovery, feature selection, sample selection bias, transfer learning, time series analysis, bioinformatics, social network analysis, novel applications and commercial data mining systems. He received 2010 IBM Outstanding Technical Achievement Award for his contribution to IBM Infosphere Streams.