Mining Internet of Things (IoT) Big Data Streams

Albert Bifet
LTCI, CNRS, Télécom ParisTech
Université Paris-Saclay
75634 Paris Cedex 13, FRANCE
albert.bifet@telecom-paristech.fr

Abstract

Big Data and the Internet of Things (IoT) have the potential to fundamentally shift the way we interact with our surroundings. 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 stream mining. Dealing with this setting, MOA is a software framework with classification, regression, and frequent pattern methods, and the new APACHE SAMOA is a distributed streaming software for mining IoT data streams.

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 precomputed 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.

IoT data streams are closely related to Big Data. *Big Data* is a new term used to identify the datasets that due to their large size, we can not manage them with the typical data mining software tools. Instead of defining "Big Data" as datasets of a concrete large size, for example in the order of magnitude of petabytes, the definition is related to the fact that the dataset is too big to be managed without using new algorithms or technologies. There is need for new algorithms, and new tools to deal with all of this data. Doug Laney (Laney, 2001) was the first to mention the 3 V's of Big Data management: Volume, Variety and Velocity.

2 MOA

Massive **O**nline **A**nalysis (MOA) (Bifet et al., 2010) is a software environment for implementing algorithms and running experiments for online learning from evolving data streams. MOA includes a collection of offline and online methods as well as tools for evaluation. In particular, it implements boosting, bagging, and Hoeffding Trees, all with and without Naïve Bayes classifiers at the leaves. Also it implements regression, and frequent pattern methods. MOA supports bidirectional interaction with WEKA, the Waikato Environment for Knowledge Analysis, and is released under the GNU GPL license.

3 APACHE SAMOA

APACHE SAMOA (SCALABLE ADVANCED MASSIVE ONLINE ANALYSIS) is a platform for mining big data streams (Morales and Bifet, 2015). As most of the rest of the big data ecosystem, it is written in Java.

APACHE SAMOA is both a framework and a library. As a framework, it allows the algorithm developer to abstract from the underlying execution engine, and therefore reuse their code on different engines. It features a pluggable architecture that allows it to run on several distributed stream processing engines such as Storm, S4, and Samza. This capability is achieved by designing a minimal API that captures the essence of modern DSPEs. This API also allows to easily write new bindings to port APACHE SAMOA to new execution engines. APACHE SAMOA takes care of hiding the differences of the underlying DSPEs in terms of API and deployment.

As a library, APACHE SAMOA contains implementations of state-of-the-art algorithms for distributed machine learning on streams. For classification, APACHE SAMOA provides a Vertical Hoeffding Tree (VHT), a distributed streaming version of a decision tree. For clustering, it includes an algorithm based on CluStream. For regression, HAMR, a distributed implementation of Adaptive Model Rules. The library also includes meta-algorithms such as bagging and boosting.

The platform is intended to be useful for both research and real world deployments.

3.1 High Level Architecture

We identify three types of APACHE SAMOA users:

1. Platform users, who use available ML algorithms without implementing new ones.

2. ML developers, who develop new ML algorithms on top of APACHE SAMOA and want to be isolated from changes in the underlying SPEs.

3. Platform developers, who extend APACHE SAMOA to integrate more DSPEs into APACHE SAMOA.

4 Conclusions

Mining Internet of Things (IoT) Big Data streams is a challenging task, that needs new tools to perform the most common machine learning algorithms such as classification, clustering, and regression.

APACHE SAMOA is is a platform for mining big data streams, and it is already available and can be found online at `http://www.samoa-project.net`. The website includes a wiki, an API reference, and a developer's manual. Several examples of how the software can be used are also available.

Acknowledgments

The presented work has been done in collaboration with Gianmarco De Francisci Morales, Nicolas Kourtellis, Bernhard Pfahringer, Geoff Holmes, Richard Kirkby, and all the contributors to MOA and APACHE SAMOA.

References

Gianmarco De Francisci Morales and Albert Bifet. 2015. SAMOA: Scalable Advanced Massive Online Analysis. *Journal of Machine Learning Research*, 16:149–153.

Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. 2010. MOA: Massive Online Analysis. *Journal of Machine Learning Research*, 11:1601–1604, August.

Doug Laney. 2001. 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note, February 6.*