Concurrent Data Structures for Big Data Streaming

To deal with the vast amount of data that are produced every day from all kinds of devices, we need fast and efficient data processing algorithms that operate in real time. *Data streams* are an algorithmic abstraction to support real-time analytics. *Classification* is one of the principal algorithms in big stream data mining. *Decision tree learners* are the most popular category of classifiers in settings where black-box classifiers are not desirable, for example, in health-related applications [1].

The project aims at developing a library of decision trees and their ensembles (*forests* [2]). As a first step, we plan to construct a multi-threaded version of the classical Hoeffding tree [3], in order to speed up its performance. The concurrent implementation should allow for various optimizations and be independent of which classifier is used at the leaves of the tree.

The next step is to adapt the designed algorithms to more elaborated trees (e.g., for CVFDT [4]), by supporting concurrent deletions and replacements of entire branches. The project also aims to study whether ensembles of Hoeffding trees can be implemented efficiently in a *distributed* setting, e.g., a cluster/cloud environment with multiple nodes. For example, we can start with considering a hybrid approach where a concurrent version of a Hoeffding tree is employed in each of the cluster's nodes.

This is a joint project of Télécom Paris and the University of Crete.

Contact

Albert Bifet
http://albertbifet.com/
albert@albertbifet.com
INFRES, LTCI, Télécom Paris
19 place Marguerite Perey 91120 Palaiseau, FRANCE

Panagiota Fatourou
http://users.ics.forth.gr/~faturu/
faturu@ics.forth.gr
FORTH, University of Crete
Voutes Campus GR-70013 Heraklion Crete Island, Greece

Petr Kuznetsov
http://www.infres.enst.fr/~kuznetso/
petr.kuznetsov@telecom-paris.fr
INFRES, LTCI, Télécom Paris
19 place Marguerite Perey 91120 Palaiseau, FRANCE

References

[1] A. Bifet, R. G. G. Holmes, and B. Pfahringer. *Machine Learning for Data Streams.* MIT Press, 2017. ISBN 9780262037792.

[2] L. E. B. Ferreira, H. M. Gomes, A. Bifet, and L. S. Oliveira. Adaptive random forests with resampling for imbalanced data streams. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pages 1–6, 2019.

[3] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[4] G. Hulten, L. Spencer, and P. M. Domingos. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA, August 26-29, 2001*, pages 97–106, 2001.