Adaptive Learning and Mining for Data Streams and Frequent Patterns

Albert Bifet
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, Catalonia
abifet@lsi.upc.edu

Advisors: Ricard Gavaldà and José L. Balcázar
Ph.D. Dissertation Committee: Joost Kok, Joao Gama, Minos Garofalakis, Christian Borgelt, and Lluis A. Belanche.

ABSTRACT

This thesis is devoted to the design of data mining algorithms for evolving data streams and for the extraction of closed frequent trees. First, we deal with each of these tasks separately, and then we deal with them together, developing classification methods for data streams containing items that are trees.

In the data stream model, data arrive at high speed, and the algorithms that must process them have very strict constraints of space and time. In the first part of this thesis we propose and illustrate a framework for developing algorithms that can adaptively learn from data streams that change over time. Our methods are based on using change detectors and estimator modules at the right places. We propose an adaptive sliding window algorithm `ADWIN` for detecting change and keeping updated statistics from a data stream, and use it as a black-box in place or counters or accumulators in algorithms initially not designed for drifting data. Since `ADWIN` has rigorous performance guarantees, this opens the possibility of extending such guarantees to learning and mining algorithms. We test our methodology with several learning methods as Naïve Bayes, clustering, decision trees and ensemble methods. We build an experimental framework for data stream mining with concept drift, based on the MOA framework, similar to WEKA, so that it will be easy for researchers to run experimental data stream benchmarks.

Trees are connected acyclic graphs and they are studied as link-based structures in many cases. In the second part of this thesis, we describe a rather formal study of trees from the point of view of closure-based mining. Moreover, we present efficient algorithms for subtree testing and for mining ordered and unordered frequent closed trees. We include an analysis of the extraction of association rules of full confidence out of the closed sets of trees, and we have found there an interesting phenomenon: rules whose propositional counterpart is nontrivial are, however, always implicitly true in trees due to the peculiar combinatorics of the structures.

And finally, using these results on evolving data streams mining and closed frequent tree mining, we present high performance algorithms for mining closed unlabeled rooted trees adaptively from data streams that change over time. We introduce a general methodology to identify closed patterns in a data stream, using Galois Lattice Theory. Using this methodology, we then develop an incremental one, a sliding-window based one, and finally one that mines closed trees adaptively from data streams. We use these methods to develop classification methods for tree data streams.

1. EVOLVING DATA STREAM MINING.

We have proposed and illustrated a method for developing algorithms that can adaptively learn from data streams that change over time. Our methods are based on using change detectors and estimator modules at the right places; we choose implementations with theoretical guarantees in order to extend such guarantees to the resulting adaptive learning algorithm. We have proposed an adaptive sliding window algorithm (`ADWIN`) for detecting change and keeping updated statistics from a data stream, and use it as a black-box in place or counters or accumulators in algorithms initially not designed for drifting data. Since `ADWIN` has rigorous performance guarantees, this opens the possibility of extending such guarantees to the resulting learning algorithm.

A main advantage of our methods is that they require no guess about how fast or how often the stream will change; other methods typically have several user-defined parameters to this effect. The main contributions on evolving data streams are the following:

1 give a unified framework for data mining with time change detection that includes most of previous works in the literature.

2 design more efficient, accurate and parameter-free methods to detect change, maintain sets of examples and compute statistics.

3 prove that the framework and the methods are useful, efficient and easy to use, using them to build versions

of classical algorithms that work on the data stream settings:

- Classification :
  - Naïve Bayes
  - Decision Trees
  - Ensemble Methods

4 build an experimental framework for data streams similar to the WEKA framework, so that it will be easy for researchers to run experimental data stream benchmarks.

2. CLOSED FREQUENT TREE MINING.

We have described a rather formal study of trees from the point of view of closure-based mining. Progressing beyond the plain standard support-based definition of a closed tree, we have developed a rationale (in the form of the study of the operation of intersection on trees, both in combinatorial and algorithmic terms) for defining a closure operator, not on trees but on sets of trees, and we have indicated the most natural definition for such an operator; we have provided a mathematical study that characterizes closed trees, defined through the plain support-based notion, in terms of our closure operator, plus the guarantee that this structuring of closed trees gives us the ability to find the support of any frequent tree. Our study has provided us, therefore, with a better understanding of the closure operator that stands behind the standard support-based notion of closure, as well as basic algorithmics on the data type.

Then, we have presented efficient algorithms for subtree testing and for mining ordered and unordered frequent closed trees. A number of variants have suggested themselves for further study: we have evaluated the behavior of our algorithms if we take into account labels, a case where our algorithm does not fare as well as in the unlabeled case. The sequential form of the representation we use, where the number-encoded depth furnishes the two-dimensional information, is key in the fast processing of the data.

And finally, we include an analysis of the extraction of association rules of full confidence out of the closed sets of trees, along the same lines as the corresponding process on itemsets, and we have found there an interesting phenomenon that does not appear if other combinatorial structures are analyzed: rules whose propositional counterpart is nontrivial are, however, always implicitly true in trees due to the peculiar combinatorics of the structures. That study is not yet finished since we have powerful heuristics to treat those implicit rules but wish to obtain a full mathematical characterization.

3. EVOLVING TREE DATA STREAMS MINING.

Using the previous work done in evolving data streams mining and closed frequent tree mining, we have presented efficient algorithms for mining ordered and unordered frequent unlabeled closed trees on evolving data streams.

If the distribution of the tree dataset is stationary, the best method to use is IncTreeNat, as we do not need to delete any past transaction. If the distribution may evolve, then a sliding window method is more appropiate. If we know which is the right size of the sliding window, then we can use WinTreeNat, otherwise AdaTreeNat would be a better choice, since it does not need the window size parameter.

And finally, we have presented a scheme for XML classification based on our methods, that efficiently selects a reduced number of attributes, and achieves higher accuracy (even more in the more selective case in which we keep only attributes corresponding to maximal trees).

4. REFERENCES

[1] Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *SIAM International Conference on Data Mining*, 2007.

[2] Albert Bifet and Ricard Gavaldà. Kalman filters and adaptive windows for learning in data streams. In *Discovery Science*, pages 29–40, 2006.

[3] Albert Bifet and Ricard Gavaldà. Adaptive parameter-free learning from evolving data streams. In *IDA*, 2009.

[4] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Richard Kirkby, and Ricard Gavaldà. New ensemble methods for evolving data streams. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[5] José L. Balcázar, Albert Bifet, and Antoni Lozano. Intersection algorithms and a closure operator on unordered trees. In *MLG 2006, 4th International Workshop on Mining and Learning with Graphs*, 2006.

[6] José L. Balcázar, Albert Bifet, and Antoni Lozano. Mining frequent closed unordered trees through natural representations. In *ICCS 2007, 15th International Conference on Conceptual Structures*, pages 347–359, 2007.

[7] José L. Balcázar, Albert Bifet, and Antoni Lozano. Subtree testing and closed tree mining through natural representations. In *DEXA Workshops*, pages 499–503, 2007.

[8] José L. Balcázar, Albert Bifet, and Antoni Lozano. Closed and maximal tree mining using natural representations. In *MLG 2007, 5th International Workshop on Mining and Learning with Graphs*, 2007.

[9] José L. Balcázar, Albert Bifet, and Antoni Lozano. Mining implications from lattices of closed trees. In *Extraction et gestion des connaissances (EGC'2008)*, pages 373–384, 2008.

[10] José L. Balcázar, Albert Bifet, and Antoni Lozano. Mining frequent closed rooted trees. In *Machine Learning Journal*, 2009..

[11] Albert Bifet and Ricard Gavaldà. Mining adaptively frequent closed unlabeled rooted trees in data streams. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[12] Albert Bifet and Ricard Gavaldà. Adaptive XML Tree Classification on evolving data streams. In *MLG 2009, 7th International Workshop on Mining and Learning with Graphs*, 2009.

[13] Albert Bifet and Ricard Gavaldà. Adaptive XML Tree Classification on evolving data streams. In *ECML-PKDD*, 2009.