A Java Toolbox for Analysis of Massive Data Streams using Probabilistic Graphical Models

Masegosa, Andres; Martinez, Ana M.; Ramos-López, Darío; Langseth, Helge; Nielsen, Thomas Dyhre; Salmerón, Antonio; Cabanas de Paz, Rafael; Madsen, Anders Læsø

Publication date:
2016

Link to publication from Aalborg University

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal?

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: November 09, 2018
A Java Toolbox for Analysis of Massive Data Streams using Probabilistic Graphical Models

Andrés R. Masegosa\textsuperscript{1}, Ana M. Martínez\textsuperscript{2}, Darío Ramos-Lopez\textsuperscript{3}, Helge Langseth\textsuperscript{1},
Thomas D. Nielsen\textsuperscript{2}, Antonio Salmerón\textsuperscript{3}, Rafael Cabañas\textsuperscript{2} & Anders L. Madsen\textsuperscript{2,4}

\textsuperscript{1} Department of Computer and Information Science, NTNU, Norway
\textsuperscript{2} Department of Computer Science, Aalborg University, Denmark
\textsuperscript{3} Department of Mathematics, University of Almería, Spain
\textsuperscript{4} Hugin Expert A/S, Aalborg, Denmark

Presentation

Data mining frameworks

- Stationary data sets
  - Weka
  - R Libs
  - Matlab
- PGMs
  - AMIDST
- Data streams
  - MOA
  - Apache SAMOA
  - MLlib
  - Apache Spark/Flink
  - MOA
  - Elvira
  - Infer.net
  - Hugin
  - Weka
  - R
  - Libs
  - Matlab
  - Apache SAMOA
  - Vowpal Wabbit

Academic and Industrial partners

Description

- **Analysis of big data streams**: A complete collection of algorithms for inference and learning of both static and dynamic Bayesian networks from streaming data. Existing software systems for PGMs only focus on stationary datasets.

- **Distributed parallel algorithms**: AMIDST provides parallel multi-core and distributed implementations of Bayesian parameter learning, using streaming variational Bayes and variational message passing.

Main Features

- Java 8 based
- Latent variable models
- Integration

Code example

```
import java8.
importLatentVariable.
importIntegration.

// Learn hidden naive Bayes model from data stream

// Load dataset
Dataset dataset = DatasetLoader.openFromFile("datasets/credit/income.arff");

// Use a 25% sub-sampling
SAMOAItrajectoriesAlgorithm algorithm = new SAMOAAlgorithm();

// Fix the size of the window
parameterLearningAlgorithm = parameterLearningAlgorithm.getWindowsSize(10);

// Compute the output
parameterLearningAlgorithm.setOutput(true);

// Use the data which is going to be used for the parameters
parameterLearningAlgorithm.setDataStream(data);

// Perform the learning
parameterLearningAlgorithm.runLearning();

// Print the model
System.out.println(bnModel.toString());
```

Use-case: Risk prediction in credit operations

- Concept drift
- Correlated with unemployment rate

And much more...

amidst.eu
amidst.github.io/toolbox/

AMIDST project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209.