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.

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

Andrés R. Masegosa¹, Ana M. Martínez², Darío Ramos-Lopez³, Helge Langseth¹, Thomas D. Nielsen², Antonio Salmerón³, Rafael Cabañas² & Anders L. Madsen²,⁴

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

Presentation

Data mining frameworks

- Static data sets
  - Weka
  - R Libs
  - Matlab
  - Elvira
  - Infer.net
  - Hugin
  - AMIDST

- Data streams
  - MOA
  - Apache SAMOA
  - Milib/Apache Spark/Flink
  - Vowpal Wabbit

PGMs

- Apache
- SAMOA
- Vowpal
- Wabbit

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

Java 8

Code example

```java
// We create a 2DFG object
SGM parameterLearningAlgorithm = new SGM();
// We fix the 2D structure
parameters = new SGMStructure.Random2DStructure(20, 2);
// We fix the size of the window
parameters = new SGMStructure.Random2DStructure(100);
// We can activate the output
parameters = new SGMStructure.Random2DStructure(true);
// We set the data which is going to be used for learning
parameters = new SGMStructure.Random2DStructure(data);
// We perform the learning
parameters = new SGMStructure.Random2DStructure(true);
// And we get the model
BayesianNetwork bnModel = new BayesianNetwork(parameters);
// We print the model
System.out.println(bnModel.toString());
```

Academic and Industrial partners

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

Use-case: Risk prediction in credit operations

Concept drift

Correlated with Unemployment Rate

And much more...

amidst.eu

amidst.github.io/toolbox/