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

Masegosa, Andres; Martinez, Ana Maria; Ramos-López, Darío; Langseth, Helge; Nielsen, Thomas Dyhre; Salmerón, Antonio; Cabanas, 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.
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-López\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**

- **Sta\textsuperscript{4}t\textsuperscript{ary} data sets**
  - Weka
  - R Lbs.
  - Matlab

- **PGMs**
  - Elvira
  - Infer.net
  - Hugin
  - Vowpal Wabbit

- **AMIDST**

- **Data streams**
  - Apache SAMOA
  - MLlib
  - Spark/Flink

### Academic and Industrial partners

- NTNU
- Universidad de Almería
- Daimler
- Hugin Expert A/S
- CAJAMAR

### 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**
- **Big Data**
- **Modularity**
- **Open source**

### Code example

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

Amidst.eu

amidst.github.io/toolbox/

Use-case: Risk prediction in credit operations

Use-case: Correlated with unemployment rate

And much more...