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.
**Data mining frameworks**

- **AMIDST**
- **Weka**
- **R Libs**
- **Matlab**

**PGMs**

- **Apache SAMOA**
- **MOA**
- **Apache Spark/Flink**

**Stationary data sets**

- ** stationary data**
- **Elvira**
- **Infer.net**
- **Hugin**

**Data streams**

- **Apache SAMOA**
- **Vowpal Wabbit**

**Academic and Industrial partners**

- **NTNU**
- **UNIVERSIDAD DE ALMERÍA**
- **DAIMLER**

**Presentation**

**Data mining frameworks**

- **AMIDST**
- **Weka**
- **R Libs**
- **Matlab**

**PGMs**

- **Apache SAMOA**
- **MOA**
- **Apache Spark/Flink**

**Stationary data sets**

- ** stationary data**
- **Elvira**
- **Infer.net**
- **Hugin**

**Data streams**

- **Apache SAMOA**
- **Vowpal Wabbit**

**Academic and Industrial partners**

- **NTNU**
- **UNIVERSIDAD DE ALMERÍA**
- **DAIMLER**

**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**

```
// We create a 2D array for the parameter estimation algorithm
// We fix the 2D structure
parametersEstimationAlgorithm = newParametersEstimationAlgorithm();
// We fix the size of the window
parametersEstimationAlgorithm.setParameter(windowSize); // 100

// We set the data which is going to be used for the learning
parametersEstimationAlgorithm.setDataStream(data);
// We perform the learning
parametersEstimationAlgorithm.runLearning();

// We get the model
BayesianNetwork bnModel = parametersEstimationAlgorithm.getLearnedBayesianNetwork();
// We print the model
System.out.println(bnModel.toString());
```

**Use-case: Risk prediction in credit operations**

- **Concept drift**
- **Correlated with Unemployment Rate**

```
// Run the code
amidst.eu/amidst.github.io/toolbox/
```

**And much more...**

```
amidst.eu
amidst.github.io/toolbox/
```