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

Citation for published version (APA):
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

- Stationary data sets
  - Weka
  - RLibs
  - Matlab
- PGMs
  - AMIDST
  - Sta4onary data sets
  - Data streams
- 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
- Big Data
- Modularity
- Open source

Code example

```java
// We create a 2D array to store the input data
double[][] inputArray = new double[20][2];

// We create the model
BayesianNetwork bnModel = new BayesianNetwork();

// We perform inference
InferenceEngine inferenceEngine = bnModel.createInferenceEngine();
// We run the inference
inferenceEngine.runInference();

// We 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.