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

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

- PGMs
- AMIDST
- Stationary 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

Code example

Learn hidden naive Bayes model from data stream

```java
// We can open the data stream using the static class DataStreamLoader
DataStream<DataInstance> data = DataStreamLoader.openFromFile("datasets/simulated/UnemploymentRate/UnemploymentRate_2000-2014.csv");
// We set the data which is going to be used for learning the parameters
parameterLearningAlgorithm.setDataStream(data);
// We fix the size of the window
parameterLearningAlgorithm.setWindowsSize(100);
// We can activate the output
parameterLearningAlgorithm.setOutput(true);
```

Use-case: Risk prediction in credit operations

Concept drift

And much more...

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