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

Andrés R. Masegosa 1, Ana M. Martínez 2, Dário Ramos-Lopez 3, Helge Langseth 1,
Thomas D. Nielsen 2, Antonio Salmerón 3, Rafael Cabañas 2 & Anders L. Madsen 2,4

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
- Libraries
- Matlab
- 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

Code example

```java
// We create a 2D output
SFMParameterEstimationAlgorithm outputs = new SFM();
// We fix the DAG structure
parameters = sfmParameterEstimationAlgorithm.getDAGStructure(data);
// We fix the size of the window
parameters = sfmParameterEstimationAlgorithm.setWindowSize(10);
// We can estimate the output
parameters = sfmParameterEstimationAlgorithm.runLearning();
// We set the data which is going to be used for testing the parameters
parameters = sfmParameterEstimationAlgorithm.setTestData(data);
// We perform the learning
parameters = sfmParameterEstimationAlgorithm.runLearning();
// And we get the model
BayesianNetwork bnModel = sfmParameterEstimationAlgorithm.getLearntBayesianNetwork();
// 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.