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):
**Presentation**

Data mining frameworks

- **PGMs**
  - AMIDST
- **Stationary data sets**
  - Weka
  - R Libs
  - Matlab
- **Data streams**
  - Elvira
  - Infer.net
  - Vowpal Wabbit
  - Apache SAMOA

**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
DatasetReader data = new ArffSAXStream(xml); // data file name from file header
AMIDSTude model = new HiddenNaiveBayesStructuring(data, new DAGGenerator());
model.setModelType(HiddenNaiveBayesModelType.STATIC);
model.setParameterLearningAlgorithm(new SVB());
model.runLearning();
System.out.println(model.toBnxml());
```

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

- Concept drift
- Correlated with Unemployment Rate

**Academic and Industrial partners**

- **NTNU**, Norway
- **UNIVERSIDAD DE ALMERÍA**, Spain
- **DAIMLER**

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