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.
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
- R Libs
- Matlab

PGMs
- Elvira
- Infer.net
- Hugin
- AMIDST
- Apache SAMOA

Data sets
- MOA
- Apache SAMOA

Data streams
- MLlib
- Apache Spark/Flink
- MOA
- Elvira
- Infer.net
- Hugin
- Weka
- R Libs
- 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

Big Data  Modularity  Open source

Code example

Learn hidden naive Bayes model from data stream

```java
public class HNBModelExample {
    HNBModel bnModel;
    ParameterLearningAlgorithm parameterLearningAlgorithm;
    DataStreamLoader dataStreamLoader;

    public static void main(String[] args) {
        // We create a HNB Model
        bnModel = new HNBModel();
        // We fix the 3D structure
        ParameterLearningAlgorithm parameterLearningAlgorithm = new ParameterLearningAlgorithm(bnModel, dataStreamLoader);
        // We fix the size of the window
        parameterLearningAlgorithm.setWindowSize(100);
        // We perform the learning
        parameterLearningAlgorithm.setDAG(DAGGenerator.getHiddenNaiveBayesStructure(dataStreamLoader));
        parameterLearningAlgorithm.setOutput(true);
        parameterLearningAlgorithm.setWindowsSize(100);
        parameterLearningAlgorithm.setDAG(DAGGenerator.getHiddenNaiveBayesStructure(dataStreamLoader));
        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.