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):

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.

Downloaded from vbn.aau.dk on: August 18, 2024
A Java Toolbox for Analysis of Massive Data Streams using Probabilistic Graphical Models

Andrés R. Masegosa¹, Ana M. Martínez², Darío Ramos-López³, 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
MOA
Apache SAMOA
MLib/Apache Spark/Flink
Vowpal Wabbit

Data streams

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
import net.pacheco.parametrification.ParameterLearningAlgorithm;
import net.pacheco.parametrification.methods.ParametersToBayesianNetwork;
import net.pacheco.testcases.DataStreamLoader;
import net.pacheco.testcases.DataStreamLoader.DataSet;
import net.pacheco.testcases.DataStreamLoader.DatasetType;

public class Main { // We print the model
    public static void main(String[] args) {
        DataSet simulated = DataSetType.SIMULATED_DATASTREAM.openFromFile("datasets/simulated/WasteIncineratorSample.arff");
        // We set the number of cores
        parameterLearningAlgorithm.setNumberOfCores(50);
        // We fix the size of the window
        parameterLearningAlgorithm.setWindowSize(100);
        // We can open the data stream using the static class DataStreamLoader
        // We perform the learning
        ParameterLearningAlgorithm.getLearntBayesianNetwork();
        // We print the model
        System.out.println("The model is: "+model.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.