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):
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

¹Department of Computer and Information Science, NTNU, Norway
²Department of Computer Science, Aalborg University, Denmark
³Department of Mathematics, University of Almería, Spain
⁴Hugin Expert A/S, Aalborg, Denmark

Presentation

Data mining frameworks

<table>
<thead>
<tr>
<th>Stationary data sets</th>
<th>PGMs</th>
<th>Data streams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weka</td>
<td>Elvira</td>
<td>Apache SAMOA</td>
</tr>
<tr>
<td>R Libs</td>
<td>Infer.net</td>
<td>MOA</td>
</tr>
<tr>
<td>Matlab</td>
<td>AMIDST</td>
<td>Spark/Flink</td>
</tr>
<tr>
<td>Data mining frameworks</td>
<td>MOA</td>
<td>Flink</td>
</tr>
<tr>
<td>Elvira</td>
<td>Apache Spark/Flink</td>
<td>MOA</td>
</tr>
<tr>
<td>AMIDST</td>
<td>Weka</td>
<td>Apache SAMOA</td>
</tr>
<tr>
<td>Sta4onary</td>
<td>data</td>
<td>Data streams</td>
</tr>
<tr>
<td>sets</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 a hidden naive Bayes model from data stream

// AMIDST toolbox
DataStream stream = DataStreamLoader.openFromFile("datasets/simulated/WasteIncineratorSample.arff");

// Create a HMM object
HMM hmm = new HMM();

// Set the number of states
hmm.setNumberOfStates(2);

// Train the HMM
hmm.train(stream);

// Print the model
System.out.println(hmm.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.