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

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

Use-case: Risk prediction in credit operations

Use-case: Risk prediction in credit operations

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.