Conference Publication

Publication	
Title	Financial Data Analysis with PGMs Using AMIDST
Authors	R. Cabañas, A. M. Martínez, A. R. Masegosa, D. Ramos-López, A. Samerón, T. D. Nielsen, H. Langseth, and A. L. Madsen
Year	2016
DOI	https://doi.org/10.1109/ICDMW.2016.0185

Conference details	
Book title	Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on
Location	Barcelona, Spain

2016 IEEE 16th International Conference on Data Mining Workshops

Financial Data Analysis with PGMs using AMIDST

Rafael Cabañas*
Aalborg University
Email: rcabanas@cs.aau.dk

Ana M. Martínez*
Aalborg University
Email: ana@cs.aau.dk

Andrés R. Masegosa*
Norwegian University of Science and Technology
Email: andresrm@idi.ntnu.no

Darío Ramos-López*
University of Almería
Email: dramoslopez@ual.es

Antonio Salmerón
University of Almería
Email: antonio.salmeron@ual.es

Thomas D. Nielsen
Aalborg University
Email: tdn@cs.aau.dk

Helge Langseth
Norwegian University of Science and Technology
Email: helgel@idi.ntnu.no

Anders L. Madsen
Hugin Expert A/S and
Aalborg University
Email: anders@hugin.com

***Abstract*—The AMIDST Toolbox is an open source Java 8 library for scalable learning of probabilistic graphical models (PGMs) based on both batch and streaming data. An important application domain with streaming data characteristics is the banking sector, where we may want to monitor individual customers (based on their financial situation and behavior) as well as the general economic climate. Using a real financial data set from a Spanish bank, we have previously proposed and demonstrated a novel PGM framework for performing this type of data analysis with particular focus on concept drift. The framework is implemented in the AMIDST Toolbox, which was also used to conduct the reported analyses. In this paper, we provide an overview of the toolbox and illustrate with code examples how the toolbox can be used for setting up and performing analyses of this particular type.**

I. Introduction

The AMIDST Toolbox is an open source library implemented in Java 8 for scalable machine learning of *probabilistic graphical models* (PGMs). PGMs constitute a principled modeling framework for learning and reasoning under uncertainty, and are defined by two parts: first, a qualitative component in the form of graph representing independence relations between the variables (i.e., nodes) in the domain being modeled; secondly, a quantitative component consisting of a set of probability distributions quantifying the relations specified in the graph. In particular, we consider hybrid *Bayesian networks (BNs)* [6], [3], [4], where the qualitative part is a direct acyclic graph (DAG) and the quantitative part is a set of conditional probability distributions, one for each variable given its parents. We will also consider the dynamic extension of BNs, namely *dynamic Bayesian networks (DBNs)* in the form of 2-Timeslice BNs. The AMIDST Toolbox supports the specification of both BNs and DBNs, including models containing *latent variables*, thereby also facilitating a fully Bayesian approach for doing model learning [5].

The financial sector represents an important application domain, where data usually have a streaming nature, i.e., new data is continuously being generated. When analyzing this kind of data, even a modest updating frequency can produce huge volumes of data, thereby making efficient and accurate data analysis and prediction extremely difficult. To handle these big data streams, current systems for PGMs often employ simplistic solution techniques that, e.g., only consider the most recently generated data or only store the data at a much lower frequency than with which it is generated. By doing so, potentially valuable data, which could otherwise have contributed to an improved system accuracy, are being ignored. By contrast, the AMIDST Toolbox has the possibility of handling large data streams. To achieve that, the algorithms implemented in AMIDST are scalable and can leverage both multi-core and distributed architectures, the latter through integration with Apache Flink[1]. To the best of our knowledge, the AMIDST Toolbox is the first system to directly support scalable mining and analysis of data streams based on PGMs.

A salient aspect of streaming data in general, and financial data in particular, is that the domain being modeled is often *non-stationary*. That is, the distribution governing the data changes over time. This situation is known as *concept drift* [2] and if not carefully taken into account, the result can be a failure to capture and interpret intrinsic properties of the data. This detection is especially crucial when analyzing financial data as the economic cycles and changes in the general economic climate can undermine the learned models.

We have previously addressed the issue of concept drift detection in the financial domain based on real customer data from a Spanish bank [1]. The results reported in [1] were obtained using the AMIDST Toolbox, which also implements the concept drift detection framework that was proposed. In this paper we first provide an overview of the AMIDST Toolbox with focus on its main features and architecture. To illustrate the use and significance of the toolbox we next describe and provide complete code examples for setting up and performing the concept drift analyses reported in [1]. The code examples have been chosen to not only illustrate how to do financial data analysis using AMIDST, but also to provide insight into the more general toolbox design and functionality. The toolbox is open source and can be downloaded from http://www.amidsttoolbox.com.

*These four authors are considered as first authors and contributed equally.

[1] flink.apache.org

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.170