AITION: A Scalable Platform for Interactive Data Mining

Harry Dimitropoulos, Herald Kllapi, Omiros Metaxas, Nikolas Oikonomidis, Eva Sitaridi, Manolis M. Tsangaris, and Yannis Ioannidis

MaDgIK Lab, Dept. of Informatics & Telecommunications,
University of Athens, Ilissia GR15784, Greece

Abstract. AITION is a scalable, user-friendly, and interactive data mining (DM) platform, designed for analyzing large heterogeneous datasets. Implementing state-of-the-art machine learning algorithms, it successfully utilizes generative Probabilistic Graphical Models (PGMs) providing an integrated framework targeting feature selection, Knowledge Discovery (KD), and decision support. At the same time, it offers advanced capabilities for multi-scale data distribution representation, analysis & simulation, as well as, for identification and modelling of variable associations.

AITION is built on top of Athena Distributed Processing (ADP) engine, a next generation data-flow language engine, capable of supporting large-scale KD on a variety of distributed platforms, such as, ad-hoc clusters, grids, or clouds. On the front end, it offers an interactive visual interface that allows users to explore the results of the KD process. The end result is that users not only understand the process that led to a statistical conclusion, but also the impact of that conclusion on their hypotheses.

In the proposed demonstration, we will show AITION in action at various stages of the knowledge discovery process, showcasing its key features regarding interactivity and scalability against a variety of problems.

1 AITION Description

PGMs are a popular and well-studied framework for compact representation of a joint probability distribution over a large number of interdependent variables, as well as, for efficient reasoning about such a distribution [5,6]. AITION (Fig. 1) is one of the latest and most advanced systems in this area. Developed as part of an EC project [1], AITION implements state-of-the-art algorithms & techniques (exact or approximate) for Bayesian Network (BN) Structure & Parameter Learning, Markov Blanket induction, and real-time inference. Furthermore, ontologies and *a-priori* knowledge can be incorporated with the BN, defining topological constraints, in order to automate causal discovery & feature selection and provide semantic modelling under uncertainty. This way, AITION presents a rich 'natural' framework for imposing structure and prior knowledge, providing the domain expert with the ability to seed the learning algorithm with knowledge about the problem at hand.

A. Ailamaki and S. Bowers (Eds.): SSDBM 2012, LNCS 7338, pp. 646–651, 2012.

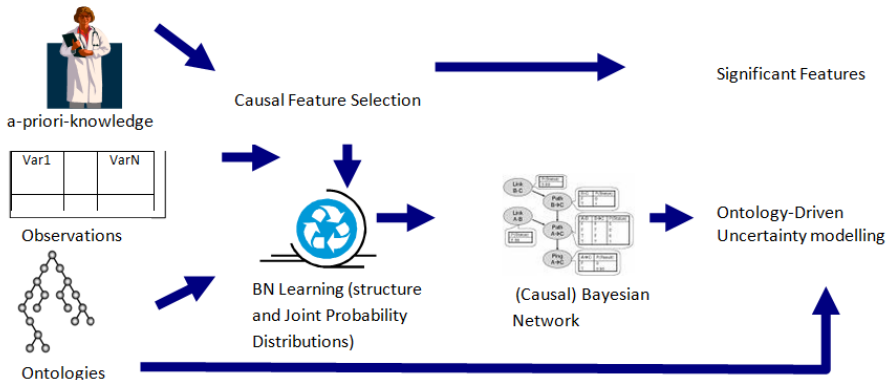

Fig. 1. Illustration of the AITION Framework

2 KDD Workflow

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately, understandable patterns in data [2]. In other words, KDD is the process of discovering useful information and patterns in data, converting raw data into useful information. The KDD workflow consists of a series of transformation steps starting from data pre-processing to model building, reasoning and knowledge extraction. AITION supports the whole KDD workflow, as shown in Figure 2.

The goal of the pre-processing is to validate, curate and transform (e.g. discretize) raw data to facilitate the application of a Data Mining (DM) algorithm. The model building step is related to the construction of a BN based on data & prior knowledge and consists of two subtasks: (a) Structure learning (or qualitative analysis), where the goal is to build a directed acyclic graph (DAG) encoding the assertions of conditional independence between variables & (b) Parameter learning (quantitative analysis), assessing conditional probability distributions. Finally, the goal of inference in a BN is to answer queries about unobserved variables, given values of some observed variables targeting either the most probable configuration - Maximum a Posteriori (MAP) - or estimating Posterior Marginal Densities.

3 System Architecture

The AITION system consists of several components as seen in Fig. 3, including the *User Interface (UI)* and the *backend*. The heart of the backend is the *ADP Engine* [7] (providing distributed query processing) and a *Relational Database* (for storing original data and knowledge models).

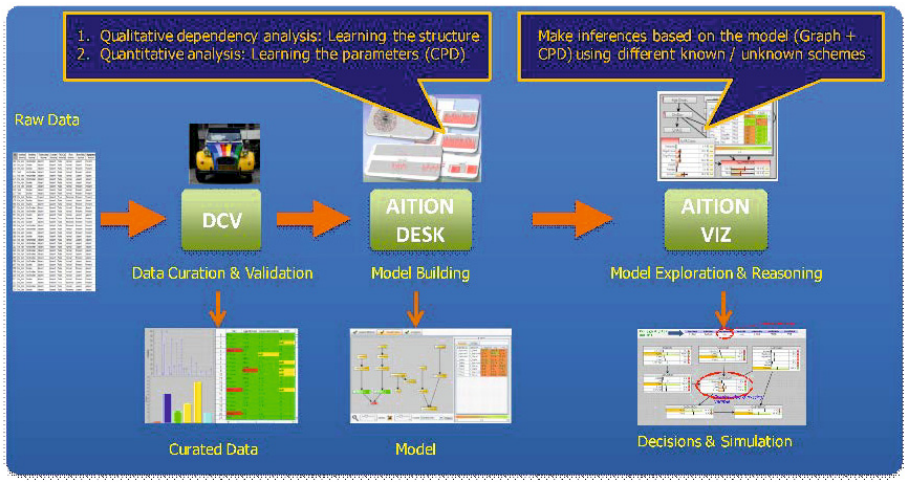

Fig. 2. The KDD workflow

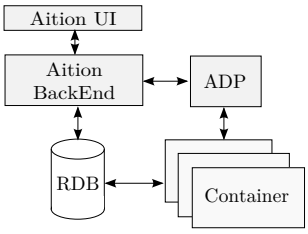

Fig. 3. AITION System Architecture

A collection of DM algorithms, most of them from WEKA [8], have been adopted and run as ADP operators, giving us the opportunity to express them as ADP queries. The optimizer facilitates "optimal" execution using all available resources, or by meeting certain cost-performance objectives. AITION applications need no modification to run over grids, ad-hoc clusters or cloud platforms.

The AITION UI Engine is a *thick client* connecting to the backend and managing all user interaction. It enables the user to execute the DM workflow. It also provides visualisation and analysis of the Bayesian knowledge models, utilizing the GraphViz toolkit of AT&T Research [3].

4 Demonstration Overview

In the demonstration, we will show AITION in a typical data mining session, as a sequence of steps: examining data samples, building a knowledge model from them, testing its validity, and finally, visualizing & exploring the end result performing a set of interesting inference scenarios.

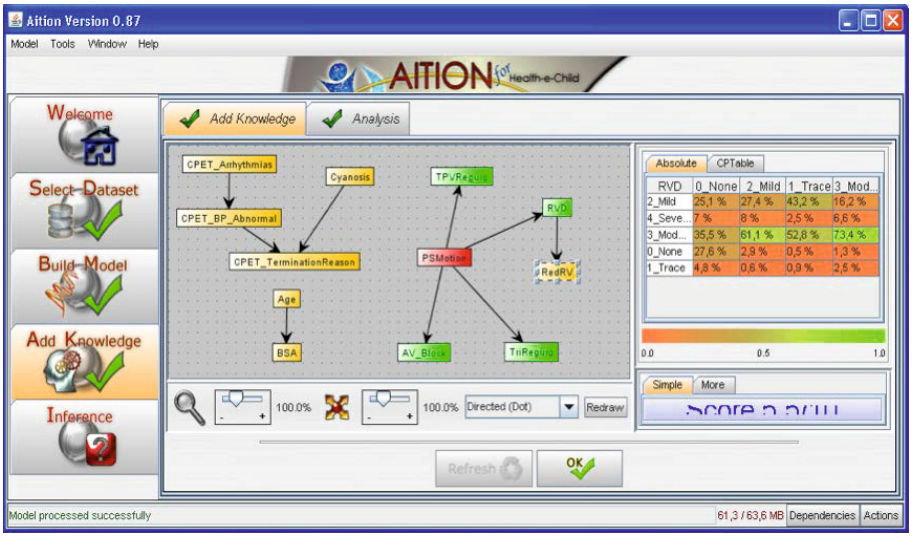

Fig. 4. A typical screen of the model building process

In more detail, we will start with an introductory example illustrating the basic notions of a KDD flow based on Bayesian Networks, as well as, familiarize the audience with the AITION workspace. Afterwards, we will show a real-world case from the medical domain, focusing on AITION's knowledge extraction and reasoning capabilities. Based on those two examples, we will cover some of the key aspects of the system, including:

Model Building: To learn the structure of the graph, AITION first performs a *qualitative dependency analysis* of the data; a repetitive process, in order to generate and evaluate several models in parallel using different training parameters. The user can then inspect the resulting graph (where nodes correspond to data features and the links/edges connecting the nodes indicate that there are probability relationships between them) and modify it (e.g., by adding or removing edges between nodes), before the next stage of *quantitative dependency analysis*, where AITION learns the parameters of the model (the conditional probability distribution). A typical screen of the model building process is show in Figure 4.

Reasoning and Visualization: An interactive workspace enables the user to perform *reasoning* using *inference* in graphs. *A-posteriori* probabilities can be computed for a specific node given some evidence: e.g., in a medical application, we can perform diagnostic, predictive, and inter-causal inference. The inference capabilities of AITION are highly interactive, including the ability to *perturbate* the values of a selected node and visually see the degree by which the other nodes in the graph are affected. A typical screen from this analysis is shown in Figure 5, where we set a value to a specific node (*RVD*), and marginal distributions for all related nodes are estimated. Finally, given a pre-computed model, the

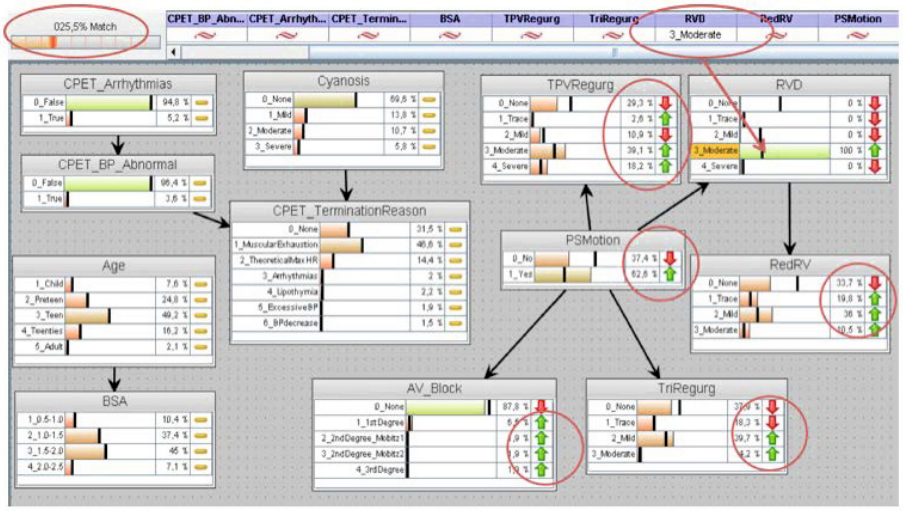

Fig. 5. A snapshot of a Knowledge Model for a medical problem during Reasoning. By setting the value of node *RVD* to Moderate, the marginal distributions for all related nodes are estimated.

user can load another set of instances (a test dataset) to perform classification, decision support, or predict missing values.

5 Conclusion and Future Work

We have demonstrated AITION applied on different domains. Solving these problems required some of its key features, including the parallel processing aspect in order to compute an appropriate PGM, and its visualization in order to make both the model & the DM process better understood by a non-technical audience. We plan to adopt more advanced algorithms for model learning & inference, while also enhancing the analytical capabilities of the tool, including the automatic generation of reports.

We also plan to further extend AITION incorporating advanced Statistical Relational Learning (SRL) and Graph Mining techniques. This way, we will create a comprehensive reasoning and simulation framework able to provide multi-scale and multi-entity predictive models. SRL [4] is an emerging area of research at the intersection of machine learning, graph mining, relational data mining, and inductive logic programming, aiming at combining statistical learning and probabilistic reasoning within logical or relational (frame-based) representations. Implementing this framework, we will be able to represent complex situations involving a variety of entities/objects, as well as, relations between them; something not possible using the simpler propositional or feature vector based representations.

References

1. http://www.health-e-child.org (2010)
2. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1–34 (1996)
3. Gansner, E.R., North, S.C.: An open graph visualization system and its applications to software engineering. Softw., Pract. Exper. 30(11), 1203–1233 (2000)
4. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. MIT Press (2007)
5. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT Press (2009)
6. Pearl, J.: Probabilistic Reasoning in Intelligent Systems, 2nd revised edn. Morgan Kaufmann, San Mateo (1988)
7. Tsangaris, M.M., Kakaletris, G., Kllapi, H., Papanikos, G., Pentaris, F., Polydoras, P., Sitaridi, E., Stoumpos, V., Ioannidis, Y.E.: Dataflow processing and optimization on grid and cloud infrastructures. IEEE Data Eng. Bull. 32(1), 67–74 (2009)
8. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques, 2nd edn. Elsevier, Morgan Kaufman, Amsterdam (2005)