AITION: A Scalable Platform for Interactive Data Mining

Harry Dimitropoulos   Herald Kllapi   Omiros Metaxas
Nikolas Oikonomidis   Eva Sitaridi   Manolis M. Tsangaris

{harryd, herald, omiros, n.oikonomidis, evas, mmt}@di.uoa.gr
MaDgIK Lab, Dept. of Informatics & Telecom., Uni. 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. 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.

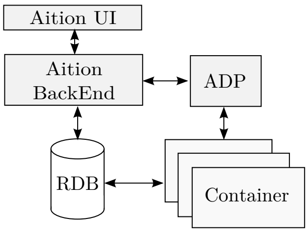


Figure 2: AITION System Architecture

2. SYSTEM ARCHITECTURE

The AITION system consists of several components as seen in Fig. 2, including the *User Interface (UI)* and the *backend*. The heart of the backend is the *ADP Engine* [3] (providing distributed query processing) and a *Relational Database* (for storing original data and knowledge models).

A collection of DM algorithms, most of them from WEKA [4], 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 [2].

3. DEMONSTRATION OVERVIEW

In the demonstration, we will show AITION running over the ADP system in a typical data mining session, as a sequence of steps: examining data samples, generating a knowledge model from them, testing its validity, and finally, visualizing & exploring the end result. 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 param-

HDMS'11, June 17–18, 2011, Athens, Greece.
.

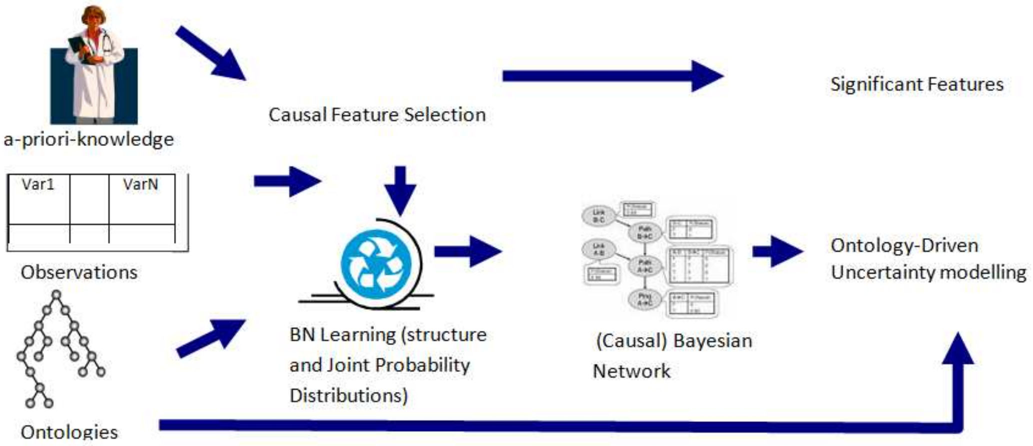
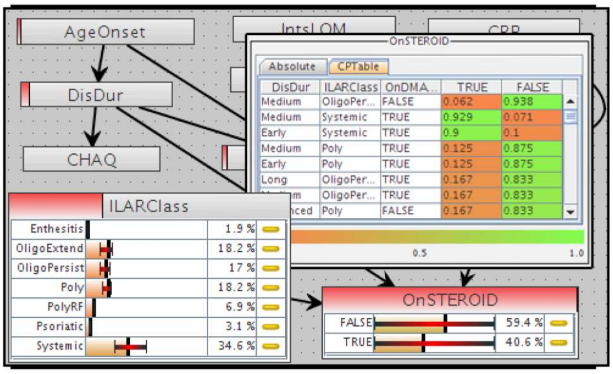

a-priori-knowledge
Causal Feature Selection
Significant Features
Var1
VarN
Observations
Ontologies
BN Learning (structure and Joint Probability Distributions)
(Causal) Bayesian Network
Ontology-Driven Uncertainty modelling

Figure 1: Illustration of the AITION Framework



DisDur	ILARClass	OnDMA...	TRUE	FALSE
Medium	OligoPer...	FALSE	0.062	0.938
Medium	Systemic	TRUE	0.929	0.071
Early	Systemic	TRUE	0.9	0.1
Medium	Poly	TRUE	0.125	0.875
Early	Poly	TRUE	0.125	0.875
Long	OligoPer...	TRUE	0.167	0.833
m	OligoPer...	TRUE	0.167	0.833
nced	Poly	FALSE	0.167	0.833

ILARClass

Enthesitis	1.9 %
OligoExtend	18.2 %
OligoPersist	17 %
Poly	18.2 %
PolyRF	6.9 %
Psoriatic	3.1 %
Systemic	34.6 %

OnSTEROID

FALSE	59.4 %
TRUE	40.6 %

Figure 3: A snapshot of a Knowledge Model for a medical problem. The strong dependency of *OnSTEROID* to *ILARClass* is shown.

eters. 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).

Mining & Visualization: Interactive tools enable 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 3, where a node is selected (*OnSTEROID*), the nodes affecting it most are shown, as well as the absolute or relative probabilities.

Finally, given a pre-computed model, the user can load another set of instances (a test dataset) to perform classification, decision support, or predict missing values.

4. CONCLUSION & 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 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.

5. ADDITIONAL AUTHORS

Additional authors: Yannis Ioannidis (University of Athens, email: `yannis@di.uoa.gr`).

6. REFERENCES

[1] www.health-e-child.org, 2010.

[2] E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Softw., Pract. Exper.*, 30(11):1203–1233, 2000.

[3] M. M. Tsangaris and more. Dataflow processing and optimization on grid and cloud infrastructures. *IEEE Data Eng. Bull.*, 32(1):67–74, 2009.

[4] I. H. Witten and E. Frank. *Data mining : practical machine learning tools and techniques*. Elsevier, Morgan Kaufman, Amsterdam [u.a.], 2. ed. edition, 2005.