Model-Driven Data Analysis

C. R. Ramakrishnan, I. V. Ramakrishnan, D. S. Warren
I/UCRC Center for Dynamic Data Analytics
Stony Brook University

Project Definition

• Analysts construct mental models from data
  – Models support understanding, prediction
  – Inferences drawn from a model should have explanations
• Probabilistic Tabled Logic Programming (PTLP): a framework to combine logical and statistical models
• This project: algorithms and system for incremental inferencing and learning in PTLP

Related Work

• Language support for statistical models (e.g. GMTK)
• Combination of Probabilistic and Relational Knowledge (e.g. PRMs, Plate Models)
• Combination of First-Order Logic and Markov Random Fields (e.g. Markov Logic Networks)
• XSB: a deductive system based on logic programming (Warren et al, since ’93).
• Incremental techniques for deduction [ICLP’06]
• Reasoning with noisy data on sensor networks [SenSys’08]
Models of Relationships

- Smoking behavior and friendships
  - A smoker’s friend is likely a smoker
    - \( \forall X.Y. \text{smoker}(X), \text{friend}(X,Y) \implies \text{smoker}(Y) \)
  - People having common friends become friends
    - \( \forall X,Y,Z. \text{friend}(X,Z), \text{friend}(Y,Z) \implies \text{friend}(X,Y) \)
- Note these formulas are “soft constraints”: they may not always hold
- Some formulas may express “hard constraints” which always hold:
  - \( \forall X,Y. \text{friend}(X,Y) \implies \text{friend}(Y,X) \)

PTLP

- Based on logic programming, that can perform deduction over general logic programs (LPs with negation in rules).
- Hard constraints are written as traditional rules (clauses) in a logic program
- Soft constraints are expressed by combining rules with random variables:
  - \( \text{smoker}(X), \text{friend}(X,Y), \text{rv}(s_{\text{true}}) \implies \text{smoker}(Y) \)
  - Distribution of Boolean random variable “s_{\text{true}}” governs the likelihood of this rule being true.
- Deduction in PTLP builds a proof, which explains how a conclusion was reached.

Inference and Learning in PTLP

- Inference is done via resolution
  - when a random variable is evaluated, the proofs branch
  - the probability of branches are based on the random variable’s distribution
  - Each proof thus has a probability; the probability of a conclusion is the probability that at least one of its proofs hold.
- Learning is done using EM
  - Underlying model is a multinomial distribution
Proposed approach to Incremental Inference

- For each conclusion, maintain its support, a data structure that has one step of explanation.
- When rules or facts (data) are deleted, propagate the deletion to supports.
- If key supports of a conclusion are deleted,
  - Check if any remaining supports are valid
  - remove the conclusion if all supports are invalid
  - propagate deletion of this conclusion (recursively).
- Recompute probabilities of conclusions whose supports have changed.
- Propagate changed probabilities via the support data structure.

Proposed approach to Incremental Learning

- Maintain support structures for the set of training examples.
- When a new example is added, update the counts computed for random variables used in the new example.
- Restart M-step based on the changed expected values.

Implication of Project

- Outcome:
  - A system providing cognitive support for data analysis
  - Efficient support for model construction and model refinement
- Possible Applications:
  - Add-on to CRM tools to support advanced data analysis
  - Analysis of vulnerabilities and risks in large networks
Deliverables

- First 6 months:
  - PTLP with incremental inference
- Second 6 months:
  - PTLP with incremental learning
- Knowledge transfer:
  - System prototype
  - Technical reports
- Longer term plans:
  - Support for complex statistical models
  - Logical abduction (analysis of what-if scenarios)
**NSF Industry/University Center for Dynamic Data Analytics (CDDA)**

**Project Summary**

<table>
<thead>
<tr>
<th><strong>Project Name:</strong></th>
<th>Model-Driven Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Investigators:</strong></td>
<td>C. R. Ramakrishnan, I. V. Ramakrishnan and David S. Warren</td>
</tr>
<tr>
<td><strong>Description:</strong></td>
<td>Analysis of complex data in a variety of domains involves the analyst constructing mental models to capture relationships in the data. We propose to use Probabilistic Tabled Logic Programming, a framework that combines statistical and logical inference, to support the construction and analysis of models from data. At its core, PTLP uses traditional logic programming notation (e.g. Horn clause formulas) to specify hard constraints or logical relationships in the model. It extends this notation with explicit random variables to specify soft constraints or statistical relationships in the model. PLTP’s inference technique is based on building proof structures using traditional reasolution, adding statistical inference on top of the proof structures. PTLP’s has an in-built learning technique to learn the distribution parameters of random variables used in the model from training data. Each conclusion inferred in PTLP has an associated proof, which serves as an explanation: an important feedback for the analyst constructing models from data. In this project, we will add incremental inference and learning techniques to PTLP, thereby making it suitable for supporting model refinement and evolution. Model-based analysis using PTLP can be applied for complex data analysis problems such as vulnerability and risk analysis of networked systems. We envision a PTLP to be an useful add-on to CRM tools to support advanced data analysis.</td>
</tr>
</tbody>
</table>
| **Development Plan:** | - Months 1-4: development and integration of incremental inference algorithm  
- Months 5-8: development and integration of incremental learning algorithm  
- Months 5-11: Performance evaluation and tuning  
- Month 12: Prototype release and completion of technical reports |
| **Related Work Elsewhere:** | - Statistical relational learning: Combination of relational models with probabilistic models  
- Languages for programming over statistical models  
- No work on incremental inference or learning |
| **How Ours Is Different:** | - Supports combination of hard and soft constraints  
- Provides explanations for each inferred conclusion to help model refinement  
- Incremental inference and learning |
| **Related Work in Center:** | - Visual analytics |
| **Milestones:** | -2010-2011: Focus on incremental inference and learning  
-beyond 2011: Extend work to support complex statistical models (e.g. regression models) and logical reasoning (e.g. what-if scenarios). |
| **Deliverables:** | - Technical demonstration along with a technical report resulting in a publication. |
| **Budget:** | $75,000 |
| **Potential Benefits to Member Companies:** | - Introduction of tool support for advanced data analysis. |