Modelling Relational Statistics With Bayes Nets

Oliver Schulte, Hassan Khosravi, Arthur Kirkpatrick, Tianxiang Gao,Yuke Zhu

School of Computing Science

Simon Fraser University, Vancouver, Canada

Project Website: http://www.cs.sfu.ca/~oschulte/jbn/

Introduction: Class-Level Queries.

Classic AI research distinguished two types of probabilistic relational queries (Halpern 1990, Bacchus 1990).

Relational Query

Class-level queries
- Relational Statistics
- Concern class proportions
- Type 1 probabilities

Instance-level queries
- Concern individuals, Ground facts
- Type 2 probabilities

Query	Reference Class
What is the percentage of flying birds?	Birds
What is the percentage of friendship pairs where both are women?	Pairs of Friends
What is the percentage of A grades awarded to highly intelligence students?	Student-course pairs

Query
Given that Tweety is a bird, what is the probability that Tweety flies?
Given that Sam and Hilary are friends, and given the genders of their other friends, what is the probability that Sam and Hilary are both women?
Given the grades of Jack in other courses, and that he is highly intelligent, what is the probability that he gets an A in CMPT 310?

Applications

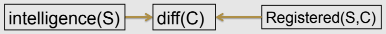
- 1^{st}-order rule or pattern learning (e.g., "intelligent students take difficulty courses").
- Strategic Planning (e.g., "increase SAT requirements to decrease student attrition").
- Query Optimization (Getoor, Taskar, Koller 2001). Choose optimal SQL query evaluation order.
- Please try our demo!

Related Work

Class-Level	Instance-Level
Statistical Relational Model	PRMs, MLNs, LBNs, RDNs,...
Parametrized Bayes Net + new random selection semantics	Parametrized Bayes Net + grounding semantics (Poole 2003)

Random Selection Semantics

- Adapted from Halpern 1990.
- A functor is a function or predicate symbol.
- A population variable *X,Y,…* ranges over a **population** or domain. X randomly select a member of its population.
- A functor node f(X), g(X,Y) is a function of a random variable ➔ also a random variable.



Example: P(int(S)=hi, diff(C) = hi, Reg(S,C) = T)= 20%

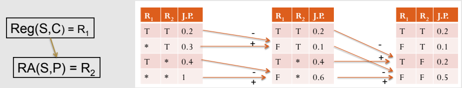
Means "if we randomly select a student and a course, there is a 20% probability that the student is highly intelligent, the course is highly difficult, and the student takes the course."

Parameter Learning

- **Database probability** of a formula = $\frac{\#\,\text{satisfying groundings}}{\#\,\text{possible groundings}}$
- Bayes net parameters = Conditional database probabilities Maximizes the random selection pseudo-likelihood (Schulte 2011).
- How to compute sufficient statistics for **negated relations**? e.g., number of U.S. users who are **not** friends?
  - For single relation, solved by Getoor et al. (2007).
  - General case: **New application** of the fast Mobius transform (Kennes and Smits 1990).

The Inverse Fast Mobius Transform

- Update equation $P(\sigma, \mathbf{R}, R = F) := P(\sigma, \mathbf{R}) - P(\sigma, \mathbf{R}, R = T)$
- Construct table of **joint** probabilities. * means "value unspecified".
- Order relationship variables, change * to False using update equation.



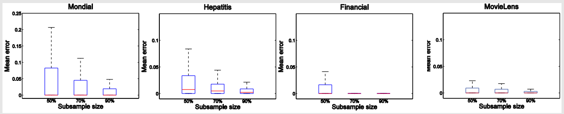
Experiments (more in paper)

- Structure Learning: Learn-and-Join Method (Khosravi et al. 2010, Schulte and Khosravi 2012).
- Parameter Learning: database probabilities.
- **Runtime** (seconds): compare Mobius transform (**IMT**) with constructing **complement** tables using SQL.

Database	Parameters	#tuples	Complement	IMT	Ratio
Mondial	1618	814	157	7	22
Hepatitis	1987	12,447	18,246	77	237
Financial	10926	17,912	228,114	14,821	15
MovieLens	326	82,623	2,070	50	41

- **Inference Performance** on Random queries.
- Train on whole database as in Getoor et al. 2001.
- Also good performance learning on subsamples, evaluating parameter estimates directly (please see paper).

Figure 4. Query Performance: Absolute difference between estimated vs. true probability. The median observation is the red center line and the box comprises 75% of the observed values. The whisker indicates the maximum acceptable value (1.5 IQR upper). Number of queries/average inference time per query: Mondial, 506/0.08sec; MovieLens, 546/0.05sec; Hepatitis, 489/0.1sec; Financial, 140/0.02sec.



Conclusion

- Parametrized Bayes nets support class-level inferences with a new **random selection semantics**.
- Mobius Transform ➔ fast and scalable parameter learning even with negated relationships.
- Excellent empirical performance on class-level quries.

References

1. Getoor, Lise, Friedman, Nir, Koller, Daphne, Pfeffer, Avi, and Taskar, Benjamin. Probabilistic relational models. In Introduction to Statistical Relational Learning
2. Poole, David. First-order probabilistic inference. In IJCAI, pp. 985–991, 2003.
3. Halpern, Joseph Y. An analysis of first-order logics of probability. Artificial Intelligence, 46(3):311–350, 1990.
4. Kennes, Robert and Smets, Philippe. Computational aspects of the Mobius transformation. In UAI, pp. 401– 416, 1990.
5. H. Khosravi, O. Schulte, T. Man, X. Xu, and B. Bina, Structure learning for Markov logic networks with many descriptive attributes, in AAAI, 2010, pp. 487–493.
6. O. Schulte and H. Khosravi. Learning graphical models for relational data via lattice search. Machine Learning, 2012. DOI: 10.1007/s10994-012-5289-4.
7. Schulte, Oliver. A tractable pseudo-likelihood function for Bayes nets applied to relational data. In SIAM SDM, pp. 462–473, 2011.