ProbLog: A Probabilistic Prolog and its Application in Link Discovery*

Luc De Raedt, Katholieke Universiteit Leuven
Angelika Kimmig, PhD Student at University of Freiburg
Hannu Toivonen, University of Helsinki

Motivated by the real-life application of mining large biological networks where edges are labeled with probabilities, we introduce ProbLog, a probabilistic extension of Prolog. A ProbLog program defines a distribution over logic programs by specifying for each clause the probability that it belongs to a randomly sampled program, and these probabilities are mutually independent. The semantics of ProbLog is then defined by the success probability of a query, which corresponds to the probability that the query succeeds in a randomly sampled program.

ProbLog is closely related to other probabilistic logics which have been developed over the past two decades to face the general need of combining deductive abilities with reasoning about uncertainty. Prominent examples include e.g. PHA [3], PRISM [5], SLPs [2], MLNs [4] and probabilistic Datalog (pD) [1]. All those frameworks attach probabilities to logical formulae, most often definite clauses (or some variant thereof). In addition, they often impose various constraints on probabilities which facilitate the computation of the probability of queries by essentially excluding the possibility that certain combinations of facts are simultaneously true. On the other hand, it seems that there are – despite a great potential – still only few real-life applications of these representations. We believe that one reason for this might well be that the assumptions are too strong and sometimes hard to manage by the user. ProbLog therefore assumes that all probabilities are mutually independent, which holds for many applications as e.g. link discovery in network structures with probabilistic edges.

Given a ProbLog program and a query, the inference task is to compute the success probability of the query, i.e. the probability that the query succeeds in a randomly sampled non-probabilistic subprogram of the ProbLog program. We show that the success probability can be computed as the probability of a monotone DNF formula of binary random variables. Motivated by the recent advances in using binary decision diagrams (BDDs) for dealing with Boolean functions, we develop a BDD approach that can effectively solve reasonably complex queries. Furthermore, we introduce an approximation algorithm for computing the success probability. To demonstrate the practical usefulness of the approach, we report on experiments using this algorithm in networks of biological concepts (genes, proteins, phenotypes, etc.) extracted from large public databases. Obviously, it is straightforward to transfer ProbLog to other link and network mining domains.

*To appear at IJCAI 2007

References

[1] Norbert Fuhr. Probabilistic datalog: Implementing logical information retrieval for advanced applications. *Journal of the American Society of Information Science*, 51(2):95–110, 2000.

[2] S. H Muggleton. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming*. IOS Press, 1996.

[3] D. Poole. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64:81–129, 1993.

[4] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[5] T. Sato and Y. Kameya. Parameter learning of logic programs for symbolic-statistical modeling. *Journal of Artificial Intelligence Research*, 15:391–454, 2001.