Logic, Probability and Computation: Foundations and Issues of Statistical Relational AI

David Poole

Department of Computer Science,
University of British Columbia,
Vancouver, BC, V6T 1Z4, Canada
poole@cs.ubc.ca
http://cs.ubc.ca/~poole/

Abstract. Over the last 25 years there has been considerable body of research into combinations of predicate logic and probability forming what has become known as (perhaps misleadingly) statistical relational artificial intelligence (StaR-AI). I overview the foundations of the area, give some research problems, proposed solutions, outstanding issues, and clear up some misconceptions that have arisen. I discuss representations, semantics, inference and learning, and provide some references to the literature. This is intended to be an overview of foundations, not a survey of research results.

Keywords: statistical relational learning, relational probabilistic models, inductive logic programming, independent choice logic, parametrized random variables

1 Introduction

Over the last 25 years there has been a considerable body of research into combining logic and probability, evolving into what has come to be called *statistical relational AI*. Rather than giving a survey, I will motivate the issues from the bottom-up, trying to justify some choices that have been made. Laying bare the foundations will hopefully inspire others to join us in exploring the frontiers and unexplored areas.

One of the barriers to understanding this area is that it builds from multiple traditions, which often use the same vocabulary to mean different things. Common terms such as "variable", "domain", "relation", and "parameter" have come to have accepted meanings in mathematics, computing, logic and probability, but their meanings in each of these areas is different enough to cause confusion.

Both predicate logic (e.g., the first-order predicate calculus) and Bayesian probability calculus can be seen as extending the propositional calculus, one by adding relations, individuals and quantified variables, the other by allowing for measures over possible worlds and conditional queries. Relational probabilistic models[1], which form the basis of statistical relational AI can be seen as combinations of probability and predicate calculus to allow for individuals and relations as well as probabilities.

To understand the needs for such a combination, consider learning from the two datasets in Figure 1 (from [25]). Dataset (a) is the sort used in traditional supervised and unsupervised learning. Standard textbook supervised learning algorithms can learn a decision tree, a neural network, or a support vector machine to predict $UserAction$. A belief network learning algorithm can be used to learn a representation of the distribution over the features. Dataset (b), from

Example	*Author*	*Thread*	*Length*	*WhereRead*	*UserAction*
e_1	*known*	*new*	*long*	*home*	*skips*
e_2	*unknown*	*new*	*short*	*work*	*reads*
e_3	*unknown*	*follow_up*	*long*	*work*	*skips*
e_4	*known*	*follow_up*	*long*	*home*	*skips*
...	...	...	...	...	...

(a)

Individual	Property	Value
joe	*likes*	*resort_14*
joe	*dislikes*	*resort_35*
...	...	...
resort_14	*type*	*resort*
resort_14	*near*	*beach_18*
beach_18	*type*	*beach*
beach_18	*covered_in*	*ws*
ws	*type*	*sand*
ws	*color*	*white*
...	...	...

(b)

Fig. 1. Two datasets

which we may want to predict what Joe likes, is different. Many of the values in the table are meaningless names that can't be used directly in supervised learning. Instead, it is the relationship among the individuals in the world that provides the generalizations from which to learn. Learning from such datasets has been studied under the umbrella of inductive logic programming (ILP) [12, 10] mainly because logic programs provide a good representation for the generalizations required to make predictions. ILP is one of the foundations of StaR-AI, as it provides a toolbox of techniques for structure learning.

[1] Here we use this term in the broad sense, meaning any models that combine relations and probabilities.

One confusion about the area stems from the term "relational"; after all most of the datasets are, or can be, stored in relational databases. The techniques of relational probabilistic models are applicable to cases where the values in the database are names of individuals and it is the properties of the individuals and the relationship between the individuals that are modelled. It is sometimes also called multi-relational learning, as it is the interrelations that are important. This is a misnomer because, as can be seen in Figure 1 (b), it not multiple relations that cause problems (and provide opportunities to exploit structure), as a single triple relation can store any relational database (in a so-called triple-store).

The term statistical relational AI, comes from not only having probabilities and relations, but that the models are derived from data and prior knowledge.

2 Motivation

Artificial intelligence (AI) is the study of computational agents that act intelligently [25]. The basic argument for probability as a foundation of AI is that agents that act under uncertainty are gambling, and probability is the calculus of gambling in that agents who don't use probability will lose to those that do use it [33]. While there are a number of interpretations of probability, the most suitable is a Bayesian or subjective view of probability: our agents do not encounter generic events, but have to make decisions in particular circumstances, and only have access to their beliefs.

In probability theory, possible worlds are described in terms of so-called *random variables* (although they are neither random nor variable). A random variable has a value in every world. We can either define random variables in terms of worlds or define worlds in terms of random variables. A random variable having a particular value is a proposition. Probability is defined in terms of a non-negative measure over sets of possible worlds that follow some very intuitive axioms.

In Bayesian probability, we make explicit assumptions and the conclusions are logical consequences of the specified knowledge and assumptions. One particular explicit assumption is the assumption of conditional independence. A Bayesian network [14] is an acyclic directed graphical model of probabilistic dependence that encapsulates the independence: a variable is conditionally independent of other variables (those that are not its descendants in the graph) given its parents in the graph. This has turned out to be a very useful assumption in practice. Undirected graphical models encapsulate the assumption that a variable is independent of other variables given its neighbours.

These motivations for probability (and similar motivations for utility) do not depend on non-relational representations.

3 Representation

Statistical relational models are typically defined in terms of parametrized random variables [20] which are often drawn in terms of plates [3]. A parametrized random variable corresponds to a predicate or a function symbol in logic. It can include logical variables (which form the parameters). In the following examples, we will write logical variables (which denote individuals) in upper case, and constants, function and predicate symbols in lower case. We assume that the logical variables are typed, where the domain of the type, the set of individuals of the type, is called the population.

Parametrized random variables are best described in terms of an example. Consider the case of diagnosing students' performance in adding multi-digit numbers of the form

$$\begin{array}{rrrr} & & x_1 & x_0 \\ + & & y_1 & y_0 \\ \hline & z_2 & z_1 & z_0 \end{array}$$

A student, given values for the x's and the y's, provides values for the z's.

Whether a student gets the correct answer for z_i depends on x_i, y_i, the value carried in and whether she knows addition. Whether a student gets the correct carry depends on the previous x, y and carry, and whether she knowns how to carry. This dependency can be seen in Figure 2. Here $x(D, P)$ is a parametrized

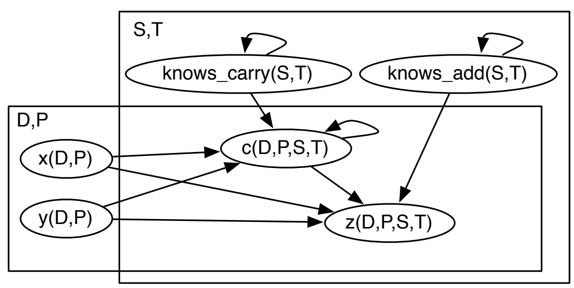

Fig. 2. Belief network with plates for multidigit addition

random variable. There is a random variable for each digit D and each problem P. A ground instance, such as $x(d_3, problem_{57})$, is a random variable that may represent the third digit of problem 57. Similarly, there is a z-variable for each digit D, problem P, student S, and time T. The plate notation can be read as duplicating the random variable for each tuple of individual the plate is parametrized by.

The basic principle used by all methods is that of *parameter sharing*: the instances of the parametrized random created by substituting constants for logical variables share the same probabilistic parameters. The various languages differ in how to specify the conditional probabilities of the variables variable given its parents, or the other parameters of the probabilistic model.

The first such languages (e.g., [8]), described the conditional probabilities directly in term of tables, and require a combination function (such as noisy-and or noisy-or) when there is a random variable parametrized by a logical variable that is a parent of a random variable that is not parametrized by the logical variable. Tables with combination functions turn out to be not a very flexible representation as they cannot represent the subtleties involved in how one random variable can depend on others.

In the above example, $c(D, P, S, T)$ depends, in part, on $c(D-1, P, S, T)$, that is, on the carry from the previous digit (and there is some other case for the first digit). A more complex example is to determine the probability that two authors are collaborators, which depends on whether they have written papers in common, or even whether they have written papers apart from each other.

To represent such examples, it is useful to be able to specify how the logical variables interact, as is done in logic programs. The independent choice logic (ICL) [18, 22] (originally called probabilistic Horn abduction [15, 17]) allows for arbitrary (acyclic) logic programs (including negation as failure) to be used to represent the dependency. The conditional probability tables are represented as independent probabilistic inputs to the logic program. A logic program that represents the above example is in Chapter 14 of [25]. This idea also forms the foundation for Prism [29, 30], which has concentrated on learning, and for Problog [4], a project to build an efficient and flexible language.

There is also work on undirected models, exemplified by Markov logic networks [26], which have a similar notion of parametrized random variables, but the probabilities are represented as weights of first-order clauses. Such models have the advantage that they can represent cyclic dependencies, but there is no local interpretation of the parameters, as probabilistic inference relies on a global normalization.

4 Inference

Inference in these models refers to computing the posterior distribution of some variables given some evidence.

A standard way to carry out inference in such models is to try to generate and ground as few of the parametrized random variables as possible. In the ICL, the relevant ground instances can be carried out using abduction [16]. More recently, there has been work on lifted probabilistic inference [20, 5, 31, 11], where the idea is to carry out probabilistic reasoning at the lifted level, without grounding out the parametrized random variables. Instead, we count how many of the probabilities we need, and when we need to multiply a number of identical probabilities, we can take the probability to the power of the number of individuals. Lifted inference turns out to be a very difficult problem, as the possible interactions between parametrized random variables can be very complicated.

5 Learning

The work on learning in relational probabilistic models has followed two, quite different, paths.

From a Bayesian point of view, learning is just a case of inference: we condition on all of the observations (all of the data), and determine the posterior distribution over some hypotheses or any query of interest. Starting from the work of Buntine [3], there has been considerable work in using relational models for Bayesian learning [9]. This work uses parametrized random variables (or the equivalent plates) and the probabilistic parameters are real-valued random variables (perhaps parametrized). Dealing with real-valued variables requires sophisticated reasoning techniques often in terms of MCMC and stochastic processes. Although these methods use relational probabilistic models for learning, the representations learned are typically not relational probabilistic models.

There is a separate body of work about learning relational probabilistic models [29, 7]. These typically use non-Bayesian techniques, to find the most likely models given the data (whereas the Bayesian technique is to average over all models). What is important about learning is that we want to learn general theories that can be learned before the agent know the individuals, and so before the agent knows the random variables.

It is still an open challenge to bring these two threads together, mainly because of the difficulty of inference in these complex models.

6 Actions

There is also a large body of work on representing actions. The initial work in this area was on representations, in terms of the event calculus [18] or the situation calculus [19, 1][2]. This is challenging because to plan, an agent needs to be concerned about what information will be available for future decision. These models combined perception, action and utility to form first-order variants of fully-observable and partially-observable Markov decision processes.

Later work has concentrated on how to do planning with such representations either for the fully observable case [2, 27] or the partially observable case [35, 28]. The promise of being able to carry out lifted inference much more efficiently is slowly being realized. There is also work on relational reinforcement learning [32, 34], where an agent learns what to do before knowing what individuals will be encountered, and so before it knows what random variables exist.

7 Identity and Existence Uncertainty

The previously outlined work assumes that an agent knows which individuals exist and can identify them. The problem of knowing whether two descriptions refer to the same individual is known as identity uncertainty [13]. This arises

[2] These two papers are interesting because they make the opposite design decisions on almost all of the design choices. For example, whether an agent knowns what situation it is in, and whether a situation implies what is true: we can't have both for a non-omniscient agent.

in citation matching when we need to distinguish whether two references refer to the same paper and in record linkage, where the aim is to determine if two hospital records refer to the same person (e.g., whether the current patient who is requesting drugs been at the hospital before). To solve this, we have the hypotheses of which terms refer to which individuals, which becomes combinatorially difficult.

The problem of knowing whether some individual exists is known as existence uncertainty [21]. This is challenging because when existence is false, there is no individual to refer to, and when existence is true, there may be many individuals that fit a description. We may have to know which individual a description is referring to. In general, determining the probability of an observation requires knowing the protocol for how observations were made. For example, if an agent considers a house and declares that there is a green room, the probability of this observation depends on what protocol they were using: did they go looking for a green room, did they report the colour of the first room found, did they report the type of the first green thing found, or did they report on the colour of the first thing they perceived?

8 Ontologies and Semantic Science

Data that are reliable and people care about, particularly in the sciences, are being reported using the vocabulary defined in formal ontologies [6]. The next stage in this line of research is to represent scientific hypotheses that also refer to formal ontologies and are able to make probabilistic predictions that can be judged against data [23]. This work combines all of the issues of relational probabilistic modelling as well as the problems of describing the world at multiple level of abstraction and detail, and handling multiple heterogenous data sets. It also requires new ways to think about ontologies [24], and new ways to think about the relationships beween data, hypotheses and decisions.

9 Conclusions

Real agents need to deal with their uncertainty and reason about individuals and relations. They need to learn how the world works before they have encountered all the individuals they need to reason about. If we accept these premises, then we need to get serious about relational probabilistic models. There is a growing community under the umbrella of statistical relational learning that is tackling the problems of decision making with models that refer to individuals and relations. While there have been considerable advances in the last two decades, there are more than enough problems to go around!

References

1. Bacchus, F., Halpern, J.Y., Levesque, H.J.: Reasoning about noisy sensors and effectors in the situation calculus. Artificial Intelligence 111(1–2), 171–208 (1999), http://www.lpaig.uwaterloo.ca/\~{}fbacchus/on-line.html

2. Boutilier, C., Reiter, R., Price, B.: Symbolic dynamic programming for first-order MDPs. In: Proc. 17th International Joint Conf. Artificial Intelligence (IJCAI-01) (2001)
3. Buntine, W.L.: Operations for learning with graphical models. Journal of Artificial Intelligence Research 2, 159–225 (1994)
4. De Raedt, L., Kimmig, A., Toivonen, H.: ProbLog: A probabilistic Prolog and its application in link discovery. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007). pp. 2462–2467 (2007)
5. de Salvo Braz, R., Amir, E., Roth, D.: Lifted first-order probabilistic inference. In: Getoor, L., Taskar, B. (eds.) Introduction to Statistical Relational Learning. M.I.T. Press (2007), `http://www.cs.uiuc.edu/~eyal/papers/BrazRothAmir_SRL07.pdf`
6. Fox, P., McGuinness, D., Middleton, D., Cinquini, L., Darnell, J., Garcia, J., West, P., Benedict, J., Solomon, S.: Semantically-enabled large-scale science data repositories. In: 5th International Semantic Web Conference (ISWC06). Lecture Notes in Computer Science, vol. 4273, pp. 792–805. Springer-Verlag (2006), `http://www.ksl.stanford.edu/KSL_Abstracts/KSL-06-19.html`
7. Getoor, L., Friedman, N., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: Dzeroski, S., Lavrac, N. (eds.) Relational Data Mining, pp. 307–337. Springer-Verlag (2001)
8. Horsch, M., Poole, D.: A dynamic approach to probabilistic inference using Bayesian networks. In: Proc. Sixth Conference on Uncertainty in AI. pp. 155–161. Boston (Jul 1990)
9. Jordan, M.I.: Bayesian nonparametric learning: Expressive priors for intelligent systems. In: Dechter, R., Geffner, H., Halpern, J.Y. (eds.) Heuristics, Probability and Causality: A Tribute to Judea Pearl, pp. 167–186. College Publications (2010)
10. Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, NY (1994)
11. Milch, B., Zettlemoyer, L.S., Kersting, K., Haimes, M., Kaelbling, L.P.: Lifted probabilistic inference with counting formulas. In: Proceedings of the Twenty Third Conference on Artificial Intelligence (AAAI) (2008), `http://people.csail.mit.edu/lpk/papers/mzkhk-aaai08.pdf`
12. Muggleton, S., De Raedt, L.: Inductive logic programming: Theory and methods. Journal of Logic Programming 19,20, 629–679 (1994)
13. Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity uncertainty and citation matching. In: NIPS. vol. 15 (2003)
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA (1988)
15. Poole, D.: Representing diagnostic knowledge for probabilistic Horn abduction. In: Proc. 12th International Joint Conf. on Artificial Intelligence (IJCAI-91). pp. 1129–1135. Sydney (1991)
16. Poole, D.: Logic programming, abduction and probability: A top-down anytime algorithm for computing prior and posterior probabilities. New Generation Computing 11(3–4), 377–400 (1993)
17. Poole, D.: Probabilistic Horn abduction and Bayesian networks. Artificial Intelligence 64(1), 81–129 (1993)
18. Poole, D.: The independent choice logic for modelling multiple agents under uncertainty. Artificial Intelligence 94, 7–56 (1997), `http://cs.ubc.ca/~poole/abstracts/icl.html`, special issue on economic principles of multi-agent systems
19. Poole, D.: Decision theory, the situation calculus and conditional plans. Electronic Transactions on Artificial Intelligence 2(1–2) (1998), `http://www.etaij.org`

20. Poole, D.: First-order probabilistic inference. In: Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03). pp. 985–991. Acapulco, Mexico (2003)
21. Poole, D.: Logical generative models for probabilistic reasoning about existence, roles and identity. In: 22nd AAAI Conference on AI (AAAI-07) (July 2007), http://cs.ubc.ca/~poole/papers/AAAI07-Poole.pdf
22. Poole, D.: The independent choice logic and beyond. In: De Raedt, L., Frasconi, P., Kersting, K., Muggleton, S. (eds.) Probabilistic Inductive Logic Programming: Theory and Application. LNCS 4911, Springer Verlag (2008), http://cs.ubc.ca/~poole/papers/ICL-Beyond.pdf
23. Poole, D., Smyth, C., Sharma, R.: Semantic science: Ontologies, data and probabilistic theories. In: da Costa, P.C., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) Uncertainty Reasoning for the Semantic Web I. LNAI/LNCS, Springer (2008), http://cs.ubc.ca/~poole/papers/SemSciChapter2008.pdf
24. Poole, D., Smyth, C., Sharma, R.: Ontology design for scientific theories that make probabilistic predictions. IEEE Intelligent Systems 24(1), 27–36 (Jan/Feb 2009), http://www2.computer.org/portal/web/computingnow/2009/0209/x1poo
25. Poole, D.L., Mackworth, A.K.: Artificial Intelligence: foundations of computational agents. Cambridge University Press, New York, NY (2010), http://artint.info
26. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62, 107–136 (2006)
27. Sanner, S., Boutilier, C.: Approximate linear programming for first-order MDPs. In: Proceedings of the Twenty-first Conference on Uncertainty in Artificial Intelligence (UAI-05). pp. 509–517. Edinburgh (2005)
28. Sanner, S., Kersting, K.: Symbolic dynamic programming for first-order POMDPs. In: Proc. AAAI-2010 (2010)
29. Sato, T., Kameya, Y.: PRISM: A symbolic-statistical modeling language. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97). pp. 1330–1335 (1997)
30. Sato, T., Kameya, Y.: New advances in logic-based probabilistic modeling by PRISM. In: De Raedt, L., Frasconi, P., Kersting, K., Muggleton, S. (eds.) Probabilistic Inductive Logic Programming, vol. LNCS 4911, pp. 118–155. Springer (2008), http://www.springerlink.com/content/1235t75977x62038/
31. Singla, P., Domingos, P.: Lifted first-order belief propagation. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. pp. 1094–1099 (2008)
32. Tadepalli, P., Givan, R., Driessens, K.: Relational reinforcement learning: An overview. In: Proc. ICML Workshop on Relational Reinforcement Learning (2004)
33. Talbott, W.: Bayesian epistemology. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (Fall 2008), http://plato.stanford.edu/archives/fall2008/entries/epistemology-bayesian/
34. van Otterlo, M.: The Logic of Adaptive Behavior - Knowledge Representation and Algorithms for Adaptive Sequential Decision Making under Uncertainty in First-Order and Relational Domains. IOS Press (2009), http://people.cs.kuleuven.be/~martijn.vanotterlo/phdbook_vanOtterlo_v2010a.pdf
35. Wang, C., Khardon, R.: Relational partially observable MDPs. In: Proc. AAAI-2010 (2010)