Towards Adversarial Reasoning in Statistical Relational Domains

Daniel Lowd and **Brenton Lessley** and **Mino De Raj**
Dept. of Computer and Information Science
University of Oregon
Eugene, Oregon 97403

Abstract

Statistical relational artificial intelligence combines first-order logic and probability in order to handle the complexity and uncertainty present in many real-world domains. However, many real-world domains also include multiple agents that cooperate or compete according to their diverse goals. In order to handle such domains, an autonomous agent must also consider the actions of other agents. In this paper, we show that existing statistical relational modeling and inference techniques can be readily adapted to certain adversarial or non-cooperative scenarios. We also discuss how learning methods can be adapted to be robust to the behavior of adversaries. Extending and applying these methods to real-world problems will extend the scope and impact of statistical relational artificial intelligence.

Introduction

Statistical relational models provide a powerful framework for artificial intelligence by combining first-order logic and probabilistic models. With some notable exceptions (e.g., (Poole 1997)), previous work in statistical relational artificial intelligence has mainly focused on representation, learning, and non-adversarial reasoning in statistical relational models. However, intelligent agents also need adversarial reasoning in order to be robust to worst-case situations or handle environments with multiple agents.

In this position paper, we discuss the problem of adversarial statistical relational reasoning and the related problem of adversarially robust statistical relational learning. We use the term "adversarial" informally to refer to both zero-sum games and general-sum non-cooperative games. We show that, when the domain is represented as a log-linear model such as a Markov logic network, certain adversarial reasoning tasks can be reduced to standard maximum a posteriori (MAP) inference. A similar transformation can convert learning methods that are based on MAP inference into adversarially robust learning methods.

Background

A number of statistical relational representations have been introduced over the years, either by extending probabilistic graphical models to handle relational domains or by extending logic programming to handle uncertainty. In this paper, we focus on the former case in general, and Markov logic networks (MLNs) (Domingos and Lowd 2009) in particular. An MLN uses first-order formulas as the features in a log-linear model. Together with a finite set of constants (and assuming known functions), this defines a probability distribution over possible worlds:

$$\log P(\mathcal{X} = \mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) - \log Z,$$

where $\mathcal{X} = \mathbf{x}$ is a possible world, consisting of the truth values of all ground atoms; $\phi_i(\mathbf{x})$ is the number of satisfied groundings of the ith formula; w_i is its weight; and Z is a normalization constant.

One notable extension of Markov logic is Markov logic decision networks (MLDNs), which adapt MLNs for the problem of utility maximization in uncertain relational domains (Nath and Domingos 2009). An MLDN distinguishes between evidence atoms, $\mathbf{e}$; action atoms controlled by agent, $\mathbf{a}$; and other state atoms, $\mathbf{x}$. An MLDN assigns a utility weight u_i to each formula. The utility of a world is the total utility weight of the satisfied formulas:

$$U(\mathbf{x}, \mathbf{a}, \mathbf{e}) = \mathbf{u}^T \phi(\mathbf{x}, \mathbf{a}, \mathbf{e})$$

The agent's goal is to select a configuration of the action atoms to maximize the expected utility, defined as follows:

$$E[U(\mathbf{x}, \mathbf{a}, \mathbf{e})|\mathbf{a}, \mathbf{e}] = \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{a}, \mathbf{e}) U(\mathbf{x}, \mathbf{a}, \mathbf{e})$$

This can naturally model many decision-theoretic problems, but it does not consider the actions of other rational agents in the environment.

Adversarial Relational Reasoning

MLDNs use an expected utility over many possible outcomes. In general, this expectation is very difficult to compute. We begin by considering the simpler case, in which there are no state atoms, and then extend this to include multiple agents. Efficiently incorporating state atoms with multiple agents remains an important direction for future work.

If there are no state atoms, then maximizing the utility is a combinatorial optimization problem:

$$\mathbf{a}_o = \arg\max_{\mathbf{a}} U(\mathbf{a}, \mathbf{e}) = \arg\max_{\mathbf{a}} \mathbf{u}^T \phi(\mathbf{a}, \mathbf{e}).$$

This problem has the same form as standard maximum a posteriori (MAP) inference in a log-linear model and can be solved with the same (approximate) inference algorithms.

Now we extend this model to consist of two agents: s ("self") with utility weights $\mathbf{u}_s$, and o ("opponent") with utility weights $\mathbf{u}_o$. We model the game as follows: First, Agent s selects an action, $\mathbf{a}_s$. Agent o observes this action and selects a response, $\mathbf{a}_o$. Agents s and o then receive rewards $\mathbf{u}_s^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e})$ and $\mathbf{u}_o^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e})$, respectively. Assuming that o is rational, the optimal action of s is given by the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{a}_s}{\text{maximize}}\ \mathbf{u}_s^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) \\ &\quad \text{where}\ \ \mathbf{a}_o = \arg\max_{\mathbf{a}_o} \mathbf{u}_o^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}). \end{aligned}$$

This corresponds to a Stackelberg game where s is the leader. For mathematical convenience, we assume the $\arg\max$ is unique; if not, we can break ties arbitrarily. We can equivalently state that $\mathbf{a}_o$ has higher utility than all alternate actions, $\mathbf{a}'_o$:

$$\begin{aligned} \underset{\mathbf{a}_s, \mathbf{a}_o}{\text{maximize}} \quad & \mathbf{u}_s^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) \\ \text{s.t.} \quad & \mathbf{u}_o^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) \geq \mathbf{u}_o^T \phi(\mathbf{a}'_o, \mathbf{a}_s, \mathbf{e}) \quad \forall \mathbf{a}'_o \neq \mathbf{a}_o. \end{aligned}$$

Solving this optimization problem in either form is typically intractable: the first is a bilevel optimization problem, and the second involves an exponential number of constraints – one for each of the exponentially many possible actions $\mathbf{a}_o$. We can make the second form more tractable by reducing the number of constraints. For example, suppose we only consider a single alternate action $\mathbf{a}'_o$. Taking the Lagrangian relaxation,

$$\begin{aligned} \underset{\mathbf{a}_s, \mathbf{a}_o}{\text{maximize}} \min_{\lambda \geq 0} \quad & \mathbf{u}_s^T \phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) + \\ & \lambda \mathbf{u}_o^T (\phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) - \phi(\mathbf{a}'_o, \mathbf{a}_s, \mathbf{e})). \quad (1) \end{aligned}$$

For a fixed value of λ, Equation 1 can be represented as a single MAP inference problem:

$$\underset{\mathbf{a}_s, \mathbf{a}_o}{\text{maximize}} \quad \mathbf{w}'^T \phi'(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) \quad (2)$$

where $\mathbf{w}' = [(\mathbf{u}_s - \lambda \mathbf{u}_o); \lambda \mathbf{u}_o]$ and $\phi'(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}) = [\phi(\mathbf{a}_o, \mathbf{a}_s, \mathbf{e}); \phi(\mathbf{a}'_o, \mathbf{a}_s, \mathbf{e})]$. We can search over different values of λ as necessary to obtain tighter bounds.

To summarize, this optimization problem finds actions $\mathbf{a}_s$ to maximize the utility of agent s, assuming that agent o will respond with an action $\mathbf{a}_o$ that has higher utility than reference action $\mathbf{a}'_o$. If most of the opponent's high-utility actions are similar, then this could be a reasonable approximation. If not, then we can iteratively introduce more reference actions and Lagrange multipliers, obtaining a better and better approximation of the true objective.

Example: Consider a webspam domain, in which the spammer's goal is to create a set of webpages with many advertisements and have them appear among the results of popular search engines. We wish to predict the actions of spammers in order to understand their behavior, develop better defenses, and evaluate the robustness of these defenses.

The spammer's opponent is the search engine's anti-spam component, which uses an MLN to label web pages as spam or non-spam, after they have been modified by the spammer. We view the MLN weights as utility weights, which makes the search engine's MAP labeling its maximum utility action. We can represent the spammer's utility as the number of spam web pages that are not detected by the search engine, minus a penalty for the number of words and links modified in order to disguise these web pages (representing the effort or cost of evading detection). To make the problem more tractable, we use the true spam/non-spam labels as the alternate reference action, $\mathbf{a}'_o$. Thus, the task is to find web page modifications, $\mathbf{a}_s$, and alternate labeling, $\mathbf{a}_o$, to maximize the spammer's utility, under the constraint that $\mathbf{a}_o$ is a "better" labeling than the true labels. Following Equation 2, this can be solved by standard MAP inference in an MLN.

Adversarial relational reasoning can also be used to develop adversarially robust learning methods. For example, suppose we wish to learn the parameters of a webspam classification system that will be robust to adaptive spammers. In standard max-margin learning, this is done by enforcing a margin between the true labels and all alternate labelings. With cutting plane methods, these margin constraints are added one at a time using MAP inference (Tsochantaridis et al. 2005). For adversarially robust learning, we instead enforce a margin between true and alternate labelings for *any* adversarially selected action. The most violated adversarial margin constraints can be found by solving a problem similar to Equation 2. In previous work (Torkamani and Lowd 2013), we have solved a special case of this problem, but the more general case has a lot of potential.

Ongoing Work and Future Directions

Adversarial reasoning is a core component of artificial intelligence that has mostly been neglected by research in statistical relational methods. In ongoing work, we are exploring how statistical relational models can detect spam and fraud in relational domains such as Twitter (Yang, Harkreader, and Gu 2011) and YouTube (O'Callaghan et al. 2012), how adversarial reasoning can automatically predict spammer strategies and model weaknesses, and how adversarial learning can construct provably robust models. As a first step, we are extending RockIt (Noessner, Niepert, and Stuckenschmidt 2013) to perform adversarial reasoning in any MLN with any linear utility model.

However, the general problem of adversarial and multi-agent reasoning in relational domains is much larger. To cope with the complexities of real-world environments, agents must handle rational adversaries, boundedly rational or irrational adversaries, multiple cooperating or competing adversaries, and random (non-adversarial) effects from the environments. Moving forward, we need to identify the most promising domains for exploration, extend statistical relational representations to efficiently model these additional factors, and scale up learning and inference methods to handle these problems.

Acknowledgments

We thank the anonymous reviewers for useful comments. This research was partly funded by ARO grant W911NF-08-1-0242. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO or the U.S. Government.

References

Domingos, P., and Lowd, D. 2009. *Markov Logic: An Interface Layer for AI*. San Rafael, CA: Morgan & Claypool.

Nath, A., and Domingos, P. 2009. A language for relational decision theory. In *Proceedings of the International Workshop on Statistical Relational Learning*.

Noessner, J.; Niepert, M.; and Stuckenschmidt, H. 2013. RockIt: Exploiting parallelism and symmetry for MAP inference in statistical relational models. In *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence*. Bellevue, WA: AAAI Press.

O'Callaghan, D.; Harrigan, M.; Carthy, J.; and Cunningham, P. 2012. Network analysis of recurring YouTube spam campaigns. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. AAAI Press.

Poole, D. 1997. The independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence* 94:7–56.

Torkamani, M. A., and Lowd, D. 2013. Convex adversarial collective classification. In *Proceedings of the Thirtieth International Conference on Machine Learning*. Atlanta, Georgia: JMLR: W&CP 28.

Tsochantaridis, I.; Joachims, T.; Hofmann, T.; and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* 6(2):1453–1484.

Yang, C.; Harkreader, R. C.; and Gu, G. 2011. Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In *Recent Advances in Intrusion Detection*, 318–337. Springer.