Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Scaling-Up MAP and Marginal MAP Inference in Markov Logic

Somdeb Sarkhel
Department of Computer Science
The University of Texas at Dallas
somdeb.sarkhel@utdallas.edu

Introduction

Markov Logic Networks (Domingos and Lowd 2009) (MLNs) use a few weighted first-order logic formulas to represent large probabilistic graphical models and are ideally suited for representing both relational and probabilistic knowledge in a wide variety of application domains such as, NLP, computer vision, and robotics. However, inference in them is hard because the graphical models can be extremely large, having millions of variables and features (potentials). Therefore, several lifted inference algorithms that exploit relational structure and operate at the compact first-order level, have been developed in recent years. However, the focus of much of existing research on lifted inference is on marginal inference while algorithms for MAP and marginal MAP inference are far less advanced. The aim of the proposed thesis is to fill this void, by developing next generation inference algorithms for MAP and marginal MAP inference.

Progress to Date

Fast Scalable MAP Inference

A key inference task in MLNs used in computer vision and NLP is maximum-a-posteriori (MAP) inference which is defined as the task of finding the maximum probability assignment. One can solve this task by grounding the MLN, yielding a Markov network and then using propositional inference algorithms, such as those developed in the graphical models literature, over the Markov network. To scale these algorithms further and to take advantage of relational structure, one can lift them by looking into each of their various steps and checking whether symmetries can be exploited in order to improve its computational efficiency. However, this requires significant modifications to be made to the propositional algorithm. This is time consuming, as one has to lift decades of advances in propositional inference.

To circumvent this problem, we (Sarkhel et al. 2014b) advocate using the "lifting as pre-processing" paradigm. The key idea is to apply lifted inference as pre-processing step and construct a Markov network that is lifted in the sense that its size can be much smaller than the ground Markov network and a complete assignment to its variables may represent several complete assignments in the ground Markov network. We observe that for non-shared MLN s(i.e., MLNs where first-order formulas have no shared terms), one of the MAP state can be found using 'uniform assignment', i.e., *all* ground atoms of a predicate are either $true$ or $false$. We proposed to solve the MAP problem on these MLNs by first, creating a Markov network defined over only the predicates of the MLN and then using existing solvers to solve them. Our experiments on both synthetic and real-world MLNs demonstrated the scalability of our approach.

Unfortunately, this approach does not use existing research on lifted inference to the fullest extent and is efficient only for non-shared MLNs. Hence, in our next paper (Sarkhel et al. 2014a), we propose a novel lifted MAP inference approach which is also based on the "lifting as preprocessing" paradigm and is at least as powerful as probabilistic theorem proving (Gogate and Domingos 2011), an advanced lifted inference algorithm. The key idea in our approach is to reduce the lifted MAP inference problem to an equivalent *Integer Polynomial Program* (IPP). Each variable in the IPP potentially refers to an assignment to a large number of ground atoms in the original MLN. Hence, the size of the search space of the generated IPP can be significantly smaller than the ground Markov network. To solve the IPP generated from the MLN we convert it to an equivalent Integer Linear Program (ILP) using a classic conversion method outlined in (Watters 1967). A desirable characteristic of our reduction is that we can use any off-the-shelf ILP solver to get exact or approximate solution to the original problem. Experimental results show that our approach is superior to existing approaches in terms of scalability and accuracy.

Although the aforementioned two approaches are sound, in a vast majority of cases, they, like many other lifted inference algorithms, are unable to identify several useful symmetries (lifting rules are sound but not complete), either because the symmetries are approximate or because the symmetries are domain-specific and do not belong to a known type. In such cases, lifted inference algorithms partially ground the MLN (namely ground only a few first-order atoms) and search for a solution in this much larger partially propositionalized space. In our recent paper (Sarkhel, Singla, and Gogate 2015), we propose a principled, approximate approach for solving this *partial grounding* problem. Our approach is straight-forward: partition the ground atoms into groups and force the inference algorithm to treat all

atoms in each group as indistinguishable (symmetric). We prove that our proposed approach yields a consistent assignment that is a lower-bound on the MAP value and show that the quality of the MAP solution can be improved systematically by refining the partitions. We also show how to further improve the complexity of our refinement procedure by exploiting the *exchangeability* property of successive refinements. Specifically, we show that the exchangeable refinements can be arranged on a lattice, which can then be searched via a heuristic search procedure to yield an efficient any-time, any-space algorithm for MAP inference. We demonstrate experimentally that our method is highly scalable and yields close to optimal solutions in a fraction of the time as compared to existing approaches.

Advanced Counting for MLN Inference

The main computational bottleneck in MLN inference, which specifically affects sampling-based and local-search based inference algorithms such as Gibbs sampling and MaxWalkSAT, is counting the true groundings of a first-order formula f given a world ω. Existing MLN systems solve this counting problem using the following naïve generate-and-test approach: generate each possible grounding and test whether it is true in ω. This naïve approach is the main reason for their poor scalability. We proposed a novel, practical approach (Venugopal, Sarkhel, and Gogate 2015) for solving the aforementioned counting problem. The key advantages of our approach are that it never grounds the full MLN and in most cases is orders of magnitude better than the "generate-and-test" approach. Specifically, we encode each formula (f) as a CSP ($\mathcal{C}$) such that the number of solutions to $\mathcal{C}$ can be directly used to count the satisfied groundings of f. This allows us to leverage several years of advances in CSPs and graphical model inference and use virtually any inference algorithm along with its associated guarantees to efficiently solve the counting problem. Our experiments clearly show that our new algorithms are several orders of magnitude more scalable than existing systems.

Approximate Counting for Weight Learning

The main reason for the poor scalability of existing weight learning systems in MLNs is the high polynomial complexity of algorithms used for computing the gradient. Specifically, a sub-step in gradient computation is computing the number of groundings of a first-order formula that evaluate to $true$ given a truth assignment to all the ground predicates. Exact algorithms for solving this counting problem have high polynomial complexity. Even our advanced exact algorithm, mentioned in the previous sub-section, fails to scale-up to large domains. To scale-up MLN weight learning, in our recent paper (Sarkhel et al. 2015), we propose new objective functions for weight learning that approximates well known functions such as likelihood, pseudo-likelihood and contrastive divergence by using *approximate* instead of exact approaches for solving the aforementioned counting problem. Our experiments on large datasets demonstrate that our approach is both accurate and scalable compared to state-of-the-art MLN systems like Alchemy and Tuffy.

Future Work

Dual-decomposition formulation of Lifted MAP

A popular approach for approximate MAP inference in graphical models is to use the solution of the linear-programming relaxation of the integer linear programming formulation of the MAP problem. Often the Lagrangian dual of this LP relaxation can be solved efficiently, as the problem can be decomposed into multiple smaller 'slave' problems. This *dual decomposition* formulation, uses message-passing algorithms to achieve a tighter upper bound of the original MAP problem. We plan to formulate our lifted MAP algorithm with this architecture.

Lifted Marginal MAP

A major benefit of probabilistic graphical models is their ability to use latent variables h to explain co-dependency among the primary variables x. However, in these models the most common inference method is the *marginal MAP* inference, which finds the optimal estimate of the sub-model over just x (by summing out h). This max-sum-product problem is significantly harder than the MAP problem and has been ignored for MLNs. We plan to combine our lifted dual-decomposition formulation mentioned in the previous sub-section with Lifted Belief Propagation (Singla and Domingos 2008), to obtain a lifted marginal MAP algorithm. We plan to apply our lifted formulation to learn discriminative models with (marginalized) latent variables.

References

Domingos, P., and Lowd, D. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan & Claypool.

Gogate, V., and Domingos, P. 2011. Probabilistic Theorem Proving. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*.

Sarkhel, S.; Venugopal, D.; Singla, P.; and Gogate, V. 2014a. An Integer Polynomial Programming Based Framework for Lifted MAP Inference. In *Advances in Neural Information Processing Systems*.

Sarkhel, S.; Venugopal, D.; Singla, P.; and Gogate, V. 2014b. Lifted MAP Inference for Markov Logic Networks. In *Proceedings of the Seventeenth AISTATS Conference*.

Sarkhel, S.; Singla, P.; and Gogate, V. 2015. Fast Lifted MAP Inference via Partitioning. In *Advances in Neural Information Processing Systems*.

Sarkhel, S.; Venugopal, D.; Pham, T; Singla, P and Gogate, V. 2015. Scalable Training of Markov Logic Networks using Approximate Counting. In *Thirtieth AAAI Conference*.

Singla, P., and Domingos, P. 2008. Lifted First-Order Belief Propagation. In *Proceedings of the Twenty-Third AAAI Conference* .

Venugopal, D.; Sarkhel, S.; and Gogate, V. 2015. Just Count the Satisfied Groundings: Scalable Local-Search and Sampling Based Inference in MLNs. In *Twenty-Ninth AAAI Conference*.

Watters, L. J. 1967. Reduction of Integer Polynomial Programming Problems to Zero-One Linear Programming Problems. *Operations Research* .