Doctoral Dissertations 2014-current

Scaling MCMC Inference and Belief Propagation to Large, Dense Graphical Models

Sameer Singh, *University of Massachusetts - Amherst*

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

Year Degree Awarded
Spring 2014

First Advisor
Andrew McCallum

Second Advisor
Benjamin Marlin

Third Advisor
David Jensen

Keywords
machine learning, graphical models, markov chain monte carlo, belief propagation, approximate inference, distributed computing, parallel computing

Subject Categories

Artificial Intelligence and Robotics | Other Statistics and Probability

Abstract

With the physical constraints of semiconductor-based electronics becoming increasingly limiting in the past decade, single-core CPUs have given way to multi-core and distributed computing platforms. At the same time, access to large data collections is progressively becoming commonplace due to the lowering cost of storage and bandwidth. Traditional machine learning paradigms that have been designed to operate sequentially on single processor architectures seem destined to become obsolete in this world of multi-core, multi-node systems and massive data sets. Inference for graphical models is one such example for which most existing algorithms are sequential in nature and are difficult to scale using parallel computations. Further, modeling large datasets leads to an escalation in the number of variables, factors, domains, and the density of the models, all of which have a substantial impact on the computational and storage complexity of inference. To achieve scalability, existing techniques impose strict independence assumptions on the model, resulting in tractable inference at the expense of expressiveness, and therefore of accuracy and utility, of the model.

Motivated by the need to scale inference to large, dense graphical models, in this thesis we explore approximations to Markov chain Monte Carlo (MCMC) and belief propagation (BP) that induce dynamic sparsity in the model to utilize parallelism. In particular, since computations over some factors, variables, and values are more important than over others at different stages of inference, proposed approximations that prioritize and parallelize such computations facilitate efficient inference. First, we show that a synchronously distributed MCMC algorithm that uses dynamic partitioning of the model achieves scalable inference. We then identify bottlenecks in the synchronous architecture, and demonstrate that a collection of MCMC techniques that use asynchronous updates are able to address these drawbacks. For large domains and high-order factors, we find that dynamically inducing sparsity in variable domains, results in scalable belief propagation that enables joint inference. We also show that formulating distributed BP and joint inference as generalized BP on cluster graphs, and by using cluster message approximations, provides significantly lower communication cost and running time. With these tools for inference in hand, we are able to tackle entity tagging, relation extraction, entity resolution, cross-document coreference, joint inference, and other information extraction tasks over large text corpora.

Recommended Citation