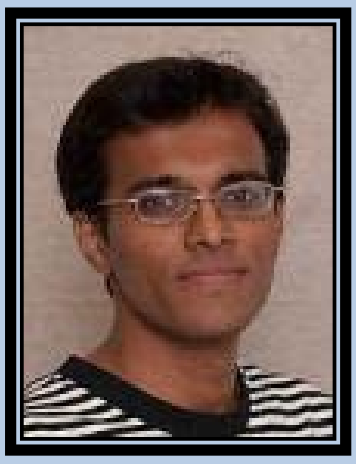
Thesis Defense

Large-scale Structured Learning

Candidate:

Siddharth Gopal



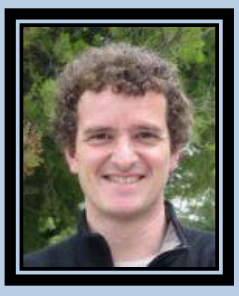
Committee:

Yiming Yang (Chair)

Jaime Carbonell

Andrew Moore

Thomas Hofmann
(ETH Zurich)



Abstract

In this thesis we study large-scale structured learning in the context of supervised, unsupervised and semi-supervised learning. In the first part of the thesis, we focus on how to harness external supervision about the structural dependencies between class-labels to improve classification performance. We propose two frameworks in this context (a) A Hierarchical Bayesian model that can exploit hierarchical dependencies using Bayesian priors and (b) A non-Bayesian Risk minimization framework that can exploit hierarchical and graphical dependencies. For both frameworks we develop fast inference and training methods that can scale to hundreds of thousands of classes in a matter of several hours. We also develop scalable training procedures for well-studied conditional models in the presence of large number of outcomes. In the second part of this thesis, we focus on automatically generating structures for organizing data. Using the von Mises-Fisher distribution as the building block we propose three Bayesian models that can recover flat, hierarchical as well as temporal structures from data on unit-spheres. Our experiments on multiple datasets showed that our proposed models are better alternatives to existing topic models based on multinomial and Gaussian distributions in their ability to recover ground-truth clusters. In the third part of the thesis, the semisupervised setting, we focus on how to expand a set of human-specified classes (or clusters) in a manner that is consistent with user expectations. We propose two new frameworks in this regard, a probabilistic framework and a constrained optimization framework, both of which rely on learning a transformation of the data such that the clusters in the transformed space match user expectations better. Our extensive evaluation on several application domains showed significant improvement in clustering performance over other competing methods.

http://www.cs.cmu.edu/~sgopal1/Thesis.pdf

GHC 6501 | Friday, August 29 | 9:30am