5. Inductive Semi-supervised Learning Methods for Natural Language Processing

Anoop Sarkar and Gholamreza Haffari, Simon Fraser University

Supervised machine learning methods which learn from labelled (or annotated) data are now widely used in many different areas of Computational Linguistics and Natural Language Processing. There are widespread data annotation endeavours but they face problems: there are a large number of languages and annotation is expensive, while at the same time raw text data is plentiful. Semi-supervised learning methods aim to close this gap.

The last 6-7 years have seen a surge of interest in semi-supervised methods in the machine learning and NLP communities focused on the one hand on analysing the situations in which unlabelled data can be useful, and on the other hand, providing feasible learning algorithms.

This recent research has resulted in a wide variety of interesting methods which are different with respect to the assumptions they make about the learning task. In this tutorial, we survey recent semi-supervised learning methods, discuss assumptions behind various approaches, and show how some of these methods have been applied to NLP tasks.

5.1 Tutorial Outline

1. Introduction
   - Spectrum of fully supervised to unsupervised learning, clustering vs. classifiers or model-based learning
   - Inductive vs. Transductive learning
   - Generative vs. Discriminative learning
2. Mixtures of Generative Models
   - Analysis
   - Stable Mixing of Labelled and Unlabelled data
   - Text Classification by EM
3. Multiple view Learning
   - Co-training algorithm
   - Yarowsky algorithm
   - Co-EM algorithm
   - Co-Boost algorithm
   - Agreement Boost algorithm
   - Multi-task Learning
4. Semi-supervised Learning for Structured Labels (Discriminative models)
   - Simple case: Random Walk
   - Potential extension to Structured SVM
5. NLP tasks and semi-supervised learning
   - Using EM-based methods to combine labelled and unlabelled data
   - When does it work? Some negative examples of semi-supervised learning in NLP
   - Examples of various NLP tasks amenable to semi-supervised learning: chunking, parsing, word-sense disambiguation, etc.
   - Semi-supervised methods proposed within NLP and their relation to machine learning methods covered in this tutorial
   - Semi-supervised learning for structured models relevant for NLP such as sequence learning and parsing
   - Semi-supervised learning for domain adaptation in NLP

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 307–308,
New York City, June 2006.

5.2 Target Audience

The target audience is expected to be researchers in computational linguistics and natural language processing who wish to explore methods that will possibly allow learning from smaller size labelled datasets by exploiting unlabelled data. In particular those who are interested in NLP research into new languages or domains for which resources do not currently exist, or in novel NLP tasks that do not have existing large amounts of annotated data. We assume some familiarity with commonly used supervised learning methods in NLP.

Anoop Sarkar is an Assistant Professor in the School of Computing Science at Simon Fraser University. His research has been focused on machine learning algorithms applied to the study of natural language. He is especially interested in algorithms that combine labeled and unlabeled data and learn new information with weak supervision. Anoop received his PhD from the Department of Computer and Information Science at the University of Pennsylvania, with Prof. Aravind Joshi was his advisor. His PhD dissertation was entitled Combining Labeled and Unlabeled Data in Statistical Natural Language Parsing. A full list of papers is available at http://www.cs.sfu.ca/ anoop. His email address is anoop@cs.sfu.ca

Gholamreza Haffari is a second year PhD student in the School of Computing Science at Simon Fraser University. He is working under the supervision of Prof. Sarkar towards a thesis on semi-supervised learning for structured models in NLP. His home page is http://www.cs.sfu.ca/ ghaffar1, and his email address is ghaffar1@cs.sfu.ca