Graph-based Semi-Supervised Learning Algorithms for NLP

Amar Subramanya
Google Research
asubram@google.com

Partha Pratim Talukdar
Carnegie Mellon University
ppt@cs.cmu.edu

Abstract

While labeled data is expensive to prepare, ever increasing amounts of unlabeled linguistic data are becoming widely available. In order to adapt to this phenomenon, several semi-supervised learning (SSL) algorithms, which learn from labeled as well as unlabeled data, have been developed. In a separate line of work, researchers have started to realize that graphs provide a natural way to represent data in a variety of domains. Graph-based SSL algorithms, which bring together these two lines of work, have been shown to outperform the state-of-the-art in many applications in speech processing, computer vision and NLP. In particular, recent NLP research has successfully used graph-based SSL algorithms for PoS tagging (Subramanya et al., 2010), semantic parsing (Das and Smith, 2011), knowledge acquisition (Talukdar et al., 2008), sentiment analysis (Goldberg and Zhu, 2006) and text categorization (Subramanya and Bilmes, 2008).

Recognizing this promising and emerging area of research, this tutorial focuses on graph-based SSL algorithms (e.g., label propagation methods). The tutorial is intended to be a sequel to the ACL 2008 SSL tutorial, focusing exclusively on graph-based SSL methods and recent advances in this area, which were beyond the scope of the previous tutorial.

The tutorial is divided in two parts. In the first part, we will motivate the need for graph-based SSL methods, introduce some standard graph-based SSL algorithms, and discuss connections between these approaches. We will also discuss how linguistic data can be encoded as graphs and show how graph-based algorithms can be scaled to large amounts of data (e.g., web-scale data).

Part 2 of the tutorial will focus on how graph-based methods can be used to solve several critical NLP tasks, including basic problems such as PoS tagging, semantic parsing, and more downstream tasks such as text categorization, information acquisition, and sentiment analysis. We will conclude the tutorial with some exciting avenues for future work.

Familiarity with semi-supervised learning and graph-based methods will not be assumed, and the necessary background will be provided. Examples from NLP tasks will be used throughout the tutorial to convey the necessary concepts. At the end of this tutorial, the attendee will walk away with the following:

- An in-depth knowledge of the current state-of-the-art in graph-based SSL algorithms, and the ability to implement them.
- The ability to decide on the suitability of graph-based SSL methods for a problem.
- Familiarity with different NLP tasks where graph-based SSL methods have been successfully applied.

In addition to the above goals, we hope that this tutorial will better prepare the attendee to conduct exciting research at the intersection of NLP and other emerging areas with natural graph-structured data (e.g., Computation Social Science).

Please visit http://graph-ssl.wikidot.com/ for details.

References

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the ACL: Human Language Technologies*.

Andrew B. Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the Workshop on Graph Based Methods for NLP*.

Amarnag Subramanya and Jeff Bilmes. 2008. Soft-supervised text classification. In *EMNLP*.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Graph-based semi-supervised learning of structured tagging models. In *EMNLP*.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 2008. Weakly supervised acquisition of labeled class instances using graph random walks. In *EMNLP*.

Tutorial Abstracts of ACL 2012, page 6,
Jeju, Republic of Korea, 8 July 2012. ©2012 Association for Computational Linguistics