Chapter 12
Graph-Based Semi-Supervised Learning With Big Data

Prithish Banerjee
West Virginia University, USA

Mark Vere Culp
West Virginia University, USA

Kenneth Jospeh Ryan
West Virginia University, USA

George Michailidis
University of Florida, USA

ABSTRACT

This chapter presents some popular graph-based semi-supervised approaches. These techniques apply to classification and regression problems and can be extended to big data problems using recently developed anchor graph enhancements. The background necessary for understanding this Chapter includes linear algebra and optimization. No prior knowledge in methods of machine learning is necessary. An empirical demonstration of the techniques for these methods is also provided on real data set benchmarks.

1. INTRODUCTION

Automation and learning in the era of "Big Data" are the cornerstones of modern machine learning methods. The main idea is to predict new data points given a sequence of 'training' points. In many cases, these approaches are viewed as adapting to the prediction problem at hand by effectively emphasizing predictive characteristics within the training points and ignoring (or down weighting) other less meaningful noise within the data. This is all done on-the-fly in real time, so there is also the need for the automation of this type of learning process. This ability is often viewed as a learning paradigm and has deep roots within statistics and computer science (Hastie et al., 2009). In order to do this task, one must

DOI: 10.4018/978-1-7998-2460-2.ch012

have methods that are (i) computationally efficient (e.g., all the parameters can be quickly estimated from the training points) and (ii) well-grounded in theory. Machine learning is the field attributed to providing data driven algorithms and models for exploring the data to make these predictions in real applications. Machine learning approaches tend to show promise in several practical applications including but not limited to those listed below.

- **Cybernetics and System Science:** Artificial intelligence (AI) and machine learning are some of modern research methods used in the field of cybernetics and system science. Automated biometrics recognition systems provide a clear example of how machine learning methods paired with AI help advance this important field. The goal is to uniquely identify a person in a fully automated fashion based on their biometric traits such as fingerprint, iris, and facial image match scores or other biometric modalities (Jain et al., 2004). In movies, such identification of the suspect is usually shown instantaneously, but this task in reality is daunting primarily due to the quality and sheer volume of the biometric data that must be processed in order to form a match. Calibrating uncertainty of matches and providing probabilistic feedback in real time on big data are a direct application of machine learning and are already having a profound practical impact on this field (Kung et al., 2005; Palaniappan & Mandic, 2007).
- **Speech Recognition:** This problem involves identification of certain dialects and languages for communication. The data typically consist of different speech recordings that are quantified into a matrix by a linguistics expert (Deng et al., 2013).
- **Text Categorization:** Filtering out spam emails, categorizing user messages, and recommending internet articles are some of the tasks that one hopes computationally efficient algorithm can achieve (Sebastiani, 2002). Another pertinent and seemingly simpler problem is that of determining whether or not a text message is 'interesting.' Individuals cannot manually perform this relevant task in real time given the volume of information available at a given time point, so machine learning has gained traction in this content area.
- **Neuroscience:** Mapping out the network of dendrons, exons, and cell bodies is a non-trivial and time-consuming process (Lao et al., 2004; Richiardi et al., 2013), but is necessary to better understand the functioning of the brain. Machine learning approaches have had a significant impact on this challenging and practical problem.

This Chapter focuses on semi-supervised learning from a machine learning point-of-view with graphs. Semi-supervised learning in general is widely regarded as a compromise between unsupervised and supervised learning. Elements of these two extreme learning paradigms are summarized below.

- **Unsupervised Learning:** Suppose an $n \times p$ data matrix $\mathbf{X}$ is generated by some application. Each row is an observation, and each column is a variable. For example, the rows often represent different text documents in text categorization, and a column represents some common numerical summary, e.g., the number of times a keyword appears in a document. The goal in unsupervised learning is to hunt for patterns within the data that are informative about the application domain. In the text example, the documents could be papers about climate change that are published in a well-known journal, and the researcher's goal may be to determine what word frequencies scientists use most often to describe the current-state of climate change. Different methods tend to dig for patterns within the data and usually involve some form of clustering (Tryon, 1939; Everitt and