

Chapter 8.1

Learning with Partial Supervision

Abdelhamid Bouchachia
University of Klagenfurt, Austria

INTRODUCTION

Recently the field of machine learning, pattern recognition, and data mining has witnessed a new research stream that is *learning with partial supervision*-LPS- (known also as *semi-supervised learning*). This learning scheme is motivated by the fact that the process of acquiring the labeling information of data could be quite costly and sometimes prone to mislabeling. The general spectrum of learning from data is envisioned in Figure 1. As shown, in many situations, the data is neither perfectly nor completely labeled.

LPS aims at using available labeled samples in order to guide the process of building classification and clustering machineries and help boost their accuracy. Basically, LPS is a combination of two learning paradigms: supervised and unsu-

pervised where the former deals exclusively with labeled data and the latter is concerned with unlabeled data. Hence, the following questions:

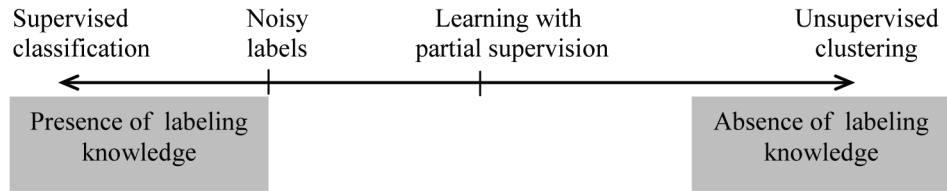
- Can we improve supervised learning with unlabeled data?
- Can we guide unsupervised learning by incorporating few labeled samples?

Typical LPS applications are medical diagnosis (Bouchachia & Pedrycz, 2006a), facial expression recognition (Cohen et al., 2004), text classification (Nigam et al., 2000), protein classification (Weston et al., 2003), and several natural language processing applications such as word sense disambiguation (Niu et al., 2005), and text chunking (Ando & Zhang, 2005).

Because LPS is still a young but active research field, it lacks a survey outlining the existing approaches and research trends. In this chapter,

DOI: 10.4018/978-1-60960-818-7.ch8.1

Figure 1. Learning from data spectrum



we will take a step towards an overview. We will discuss (i) the background of LPS, (iii) the main focus of our LPS research and explain the underlying assumptions behind LPS, and (iv) future directions and challenges of LPS research.

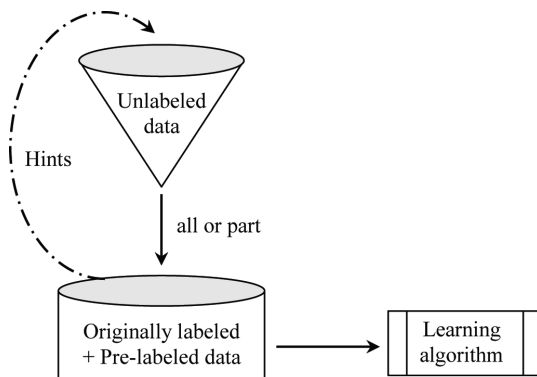
BACKGROUND

LPS is about devising algorithms that combine labeled and unlabeled data in a symbiotic way in order to boost classification accuracy. The scenario is portrayed in Figure 2 showing that the combination can mainly be done in two ways: active/passive pre-labeling, or via ‘pure’ LPS (Figure 4). We try to draw a clear picture about these schemes by means of an up-to-date taxonomy of methods.

Active and Passive Pre-Labeling

Pre-labeling aims at assigning a label to unlabeled samples (called queries). These samples are used

Figure 2. Combining labeled and unlabeled data



together with the originally labeled samples to train a fully supervised classifier (Figure 3). “Passive” pre-labeling means that pre-labeling is done automatically and is referred to as selective sampling or self-training. It has been extensively discussed and consists of first training a classifier before using it to label the unlabeled data (for more details see, Bouchachia, 2007). Various algorithms are used to perform selective sampling, such as multilayer perceptron (Verikas et al., 2001), self-organizing maps (Dara et al., 2002), and clustering techniques (Bouchachia, 2005a). On the other hand, in active learning, queries are sequentially submitted to an oracle for labeling. Different models have been applied; such as neural networks inversion (Baum, 1991), decision trees (Wiratunga et al., 2003), and query by committee (Freund & Shapire, 1997).

Pure LPS

This paradigm corresponds to LPS where the labeled and unlabeled data are treated at the same time. There are several LPS approaches including generative, ensemble learning, graph-based, boundary classification, and semi-supervised clustering models (Figure 4). We briefly outline the concepts behind these approaches without claiming any exhaustiveness due to space limitation.

Much LPS research has focused on *generative approaches* (Nigam et al., 2000)) which seek to optimize a probabilistic model defined by some characteristics such as the cluster center and standard deviation. Usually such models rely on an iterative process to estimate their characteristics. A frequently used method is *Expectation-Maximization* (EM) (Lu et al., 2007).

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/learning-partial-supervision/56231

Related Content

Toward an Agent Framework that Assists the Individuals with Mild Memory Impairment

Hung-Hsuan Huang, Hiroki Matsushita, Kyoji Kawagoe and Kiyoshi Yasuda (2014). *International Journal of Software Science and Computational Intelligence* (pp. 17-29).

www.irma-international.org/article/toward-an-agent-framework-that-assists-the-individuals-with-mild-memory-impairment/114094

Data Security and Privacy-Preserving in Edge Computing: Cryptography and Trust Management Systems

Manoranjini J. and Anbuchelian S. (2021). *Cases on Edge Computing and Analytics* (pp. 188-202).

www.irma-international.org/chapter/data-security-and-privacy-preserving-in-edge-computing/271712

Estimating which Object Type a Sensor Node is Attached to in Ubiquitous Sensor Environment

Takuya Maekawa, Yutaka Yanagisawa and Takeshi Okadome (2010). *International Journal of Software Science and Computational Intelligence* (pp. 86-101).

www.irma-international.org/article/estimating-object-type-sensor-node/39107

Theoretical Framework and Denotatum-Based Models of Knowledge Creation for Monitoring and Evaluating R&D Program Implementation

Igor Zatsman and Pavel Buntman (2013). *International Journal of Software Science and Computational Intelligence* (pp. 15-31).

www.irma-international.org/article/theoretical-framework-and-denotatum-based-models-of-knowledge-creation-for-monitoring-and-evaluating-rd-program-implementation/88989

Impact of Heavy Metal Stress (Ni and Co) on Maize (Zea mays): Exploring Biochemical and Morphological Responses for Climate Resilience and Sustainability

Mahmoud Oudghiri, Mohamed El Bakkali, Yassine Mouniane, Nagla Abid, Samah Bouhassoun, Fatima-ezzahra Jaayefar, Fath Alah Elwahab, Issam El-Khadir, Ahmed Chriqui and Mohammed Ibriz (2026). *Advances in Computational Intelligence for Climate Change Security and Sustainability* (pp. 127-152).

www.irma-international.org/chapter/impact-of-heavy-metal-stress-ni-and-co-on-maize-zea-mays/386125