Pattern Recognition Letters 37 (2014) 4–14

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Pattern classification and clustering: A review of partially supervised learning approaches

Friedhelm Schwenker [a,*], Edmondo Trentin [b]

[a] Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany
[b] Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche, Università di Siena, Via Roma 56, 53100 Siena, Italy

ARTICLE INFO

Article history:
Available online 30 October 2013

Keywords:
Partially supervised learning
Semi-supervised learning
Active learning
Transductive learning
Multi-view learning
Neural network

ABSTRACT

The paper categorizes and reviews the state-of-the-art approaches to the partially supervised learning (PSL) task. Special emphasis is put on the fields of pattern recognition and clustering involving partially (or, weakly) labeled data sets. The major instances of PSL techniques are categorized into the following taxonomy: (i) active learning for training set design, where the learning algorithm has control over the training data; (ii) learning from fuzzy labels, whenever multiple and discordant human experts are involved in the (complex) data labeling process; (iii) semi-supervised learning (SSL) in pattern classification (further sorted out into: self-training, SSL with generative models, semi-supervised support vector machines; SSL with graphs); (iv) SSL in data clustering, using additional constraints to incorporate expert knowledge into the clustering process; (v) PSL in ensembles and learning by disagreement; (vi) PSL in artificial neural networks. In addition to providing the reader with the general background and categorization of the area, the paper aims at pointing out the main issues which are still open, motivating the ongoing investigations in PSL research.

1. Introduction

The development of robust pattern classifiers from a limited training set $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of observations (i.e., feature vectors) $\mathbf{x}_i \in X$, represented in a proper feature space X, has long been one of the most relevant and challenging tasks in machine learning and statistical pattern recognition (Jain et al., 2000). *Supervised learning* and *unsupervised learning* are the two major directions of traditional machine learning.

In the supervised framework, any given generic observation (or, pattern) $\mathbf{x} \in \mathcal{T}$ is uniquely associated with a corresponding target label $y \in Y$. It is assumed that X is a real-valued vector space (i.e., $X \subseteq \mathbb{R}^d$), and that $Y = \{y_1, \ldots, y_L\}$ is the set of L (different) class labels reflecting the ground truth of the classification problem at hand. Intervention from human experts is needed in order to label the training set correctly. During an initial phase of data collection and data annotation, a supervised training set

$$\mathcal{S} = \{(\mathbf{x}_i, y_i) \,|\, \mathbf{x}_i \in \mathbb{R}^d, y_i \in Y,\ i = 1, \ldots, m\}$$

is thus prepared. It is assumed that the data in $\mathcal{S}$ are independently drawn from some (unknown, yet identical) probability distribution defined on $\mathbb{R}^d \times Y$ (i.i.d. assumption) (Bishop, 2006). Subsequently, $\mathcal{S}$ is fed into a pre-selected supervised learning algorithm aimed at training a classifier C, that is a mapping $C : \mathbb{R}^d \to Y$. This algorithm is expected to exploit the information encapsulated within both the feature vectors and the corresponding class labels. Besides the training algorithm, a hypothesis space has to be fixed, as well (e.g., the space of multivariate polynomials of maximal degree p, or the set of two-layer artificial neural networks with p logistic hidden neurons). The hypothesis space consists of all the potential candidate classifiers C which may be the eventual outcome of the computation of the learning algorithm on the training set (Alpaydin, 2010).

As we say, data annotation is an additional, expensive, and error-prone preparation process. Individual data have to be carefully inspected (by one, or even more domain experts) in order to pinpoint somewhat reliable class labels for the training patterns. Instances of the difficulties involved in the process are found in areas such as bioinformatics, speech processing, or affective computing, where the exact class labels may not even be explicitly observable. Although annotating data might be extremely difficult and time consuming (or, sometimes, even impossible), supervised learning is still far the most prominent branch of machine learning and pattern recognition.

In the unsupervised learning framework a variety of methods and algorithms can be found in the literature. Major instances are represented by data clustering, density estimation, and dimensionality reduction (just to mention a few). The goal of the learning process is usually defined through an objective function, where the learning schemes use the observations without prior knowledge of

* Corresponding author. Tel.: +49 731 502 4159; fax: +49 731 502 4156.
E-mail address: friedhelm.schwenker@uni-ulm.de (F. Schwenker).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.017

the class labels $y \in Y$. In a typical unsupervised learning scenario the training set is defined as

$$\mathcal{U} = \{\mathbf{u}_i \,|\, \mathbf{u}_i \in \mathbb{R}^d, i = 1, \ldots, M\}$$

where the data $\mathbf{u}_i$ are independently drawn from an identical probability distribution over $\mathbb{R}^d$. Clearly, the lack of any prior expert knowledge renders unsupervised learning a particularly complex machine learning/pattern recognition task (Jain, 2010). In particular, the absence of target class labels during the training phase prevents the machine from resulting in a (more or less reliable) classifier. Indeed, from a general standpoint the data set $\mathcal{U}$ could not even involve any classification task at all. All the learning algorithm can do is analyzing the data, in an attempt to capture either probabilistic (e.g., the probability density function) or geometric/topological (e.g., some distance/similarity measure, or a partitioning of the data into homogeneous clusters) information describing the nature of the data distribution.

Moving a step froward from the traditional learning frameworks, it is easy to see that a somewhat intermediate scenario occurs under all practical circumstances where a classification problem is faced relying on a data set $\mathcal{T}$ whose data are only partially labeled, such that $\mathcal{T} = \mathcal{S} \cup \mathcal{U}$ for a proper, labeled subset $\mathcal{S}$ and its unlabeled counterpart $\mathcal{U}$. While classic unsupervised techniques do not lead to any classifier C in this setup, practitioners can still rely on regular supervised classifiers trained over $\mathcal{S}$. Unfortunately, in so doing all the data in $\mathcal{U}$ would not be exploited, resulting in a waste of potentially useful additional information which could strengthen the very classifier. As a consequence, the framework of *partially supervised learning* (PSL) was introduced, having the form of a family of machine learning algorithms lying between supervised and unsupervised learning. Moreover, PSL can be seen as machine learning under weak supervision, for instance learning with a fuzzy teacher (or, with fuzzy rewards).

1.1. Prominent directions of PSL research

In practical PSL applications, after collecting the raw data, several questions arise concerning the following data processing steps:

1. How many data shall be labeled, and how do we select the (possibly small) subset of informative patterns that will be labeled?
2. How do we combine and exploit both labeled and unlabeled data within a unifying, effective training scheme?
3. How many human experts should be involved in the (robust) labeling process, and how will labels be represented in case some of the experts mutually disagree?
4. How can the machine deal with soft/fuzzy labels or multiple labels in a PSL scenario?

In an attempt to put forward plausible answers to these and further questions, several prominent directions of research have been developed so far by the community in the PSL area, including: *active learning*, general *semi-supervised learning* (SSL) (further classified into *semi-supervised classification* and *semi-supervised clustering*), SSL with graphs, PSL in ensembles and multiple classifier systems. Furthermore, a variety of PSL approaches have been investigated in the broad realms of artificial neural networks, deep learning architectures and support vector machines.

In active learning, also known as *selective sampling* or *instance selection*, it is assumed that the learning algorithm can select the most informative input training data from the pool of unlabeled examples, and a human expert is asked to add label information to the selected examples (Settles, 2009). Popular algorithms are *uncertainty sampling* and *query by committee* sampling. The former trains a single classifier and then query the unlabeled example on which the classifier is least confident (Lewis and Catlett, 1994); the latter constructs multiple classifiers and then queries the unlabeled example on which the classifiers disagree the most (Freund et al., 1997).

In semi-supervised learning the basic idea is to take advantage of unlabeled data during a supervised learning procedure (known as *semi-supervised classification*), or to incorporate some type of prior information of data points such as class labels, or constraints on pairs of patterns as "must-link" or "cannot-link" (know as *semi-supervised clustering*). In contrast to active learning, an annotator is not involved in the processing cycle. *Transductive learning* is a special case of semi-supervised classification introduced in Vapnik (1995), where the test data set is known in advance, and the goal is to optimize the classification performance on the test set itself. Recent research on SSL concentrates, in addition to semi-supervised classification (Blum and Mitchell, 1998; Nigam et al., 2000; Zhou and Li, 2005; Li and Zhou, 2007; Peng et al., 2009) and semi-supervised clustering (such as constrained and seeded k-means clustering) (Wagstaff et al., 2001; Basu et al., 2002, 2004; Chu et al., 2009; Soleymani Baghshah and Bagheri Shouraki, 2010), on semi-supervised dimensionality reduction (Zhou et al., 2007; Kalakech et al., 2011), semi-supervised non-negative matrix factorization (Lee et al., 2010), semi-supervised manifold regularization (Belkin et al., 2006), or semi-supervised regression (Zhou and Li, 2005).

Other relevant branches of PSL encompass investigations of SSL with generative models (Nigam et al., 2000; Nigam, 2001), SSL with graphs (Blum and Chawla, 2001; Zhou et al., 2004; Zhu et al., 2003; Kulis et al., 2009), multi-view learning (including co-training) (Blum and Mitchell, 1998), PSL in ensembles/multiple classifiers (including learning by disagreement) (Zhou and Li, 2010), and PSL in neural networks and kernel machines. All these research directions are surveyed in the following sections.

1.2. Organization of the paper

We made every effort in trying and categorizing the different approaches to PSL in a suitable taxonomy. This resulted in the following organization of the paper. Sections 2 reviews active learning, including uncertainty sampling and query by committee. Next, learning from a fuzzy teacher is introduced in Section 3, embracing (amongst others) fuzzy nearest prototype, fuzzy learning vector quantization, and fuzzy-input fuzzy-output support vector machines. Both active learning and fuzzy learning paradigms are basically supervised learning schemes. SSL for classification is then discussed in Section 4, according to the following sub-topics: self-training, SSL with generative models, semi-supervised support vector machines and transductive learning, and SSL with graphs. In Section 5, SSL is surveyed in the context of unsupervised cluster analysis (including must-link/cannot-link strategies). PSL in multiple classifier systems/ensembles is reviewed in Section 6 (covering, amongst others: query by committee, learning by disagreement, multi-view learning and co-training, democratic co-learning, tri-training, etc.), while Section 7 covers a number of partially supervised approaches to artificial neural networks (multilayer perceptrons, deep architectures, radial basis function networks, self-organizing maps, and ad hoc architectures). Finally, conclusions are drawn in Section 8.

2. Active learning

The key idea behind active learning is that the learning algorithm is allowed to build a labeled training set $\mathcal{S}^* \subset \mathcal{S}$ autonomously. Starting from a small subset of labeled data, say $\mathcal{S}^0 \subset \mathcal{S}$,