Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Multiview self-learning

Ali Fakeri-Tabrizi [a,*], Massih-Reza Amini [b], Cyril Goutte [c], Nicolas Usunier [d]

[a] Université Pierre and Marie Curie (Paris 6), France
[b] Université Joseph Fourier (Grenoble 1), France
[c] National Research Council of Canada, Canada
[d] Université Technologique de Compiègne, France

ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form
24 October 2014
Accepted 17 December 2014
Communicated by Tao Mei

Keywords:
Multiview learning
Self-learning
Image annotation
Multilingual document categorization

ABSTRACT

In many applications, observations are available with different views. This is, for example, the case with image-text classification, multilingual document classification or document classification on the web. In addition, unlabeled multiview examples can be easily acquired, but assigning labels to these examples is usually a time consuming task. We describe a multiview self-learning strategy which trains different voting classifiers on different views. The margin distributions over the unlabeled training data, obtained with each view-specific classifier are then used to estimate an upper-bound on their transductive Bayes error. Minimizing this upper-bound provides an automatic margin-threshold which is used to assign pseudo-labels to unlabeled examples. Final class labels are then assigned to these examples, by taking a vote on the pool of the previous pseudo-labels. New view-specific classifiers are then trained using the labeled and pseudo-labeled training data. We consider applications to image-text classification and to multilingual document classification. We present experimental results on the NUS-WIDE collection and on Reuters RCV1-RCV2 which show that despite its simplicity, our approach is competitive with other state-of-the-art techniques.

1. Introduction

Most photo sharing sites like Flickr[1] give their users the opportunity to manually assign labels to images. These labels provide descriptive keywords for the image, and play an important role in the organization of the image collection, since they can be used to browse or search the collection of images. Their usage is however limited to the small portion of images that are manually labeled. Automating the annotation process is mandatory to extend the categorization to the entire, ever-growing image collection. Automatic image annotation has two key properties: (1) for many categories, we may have a very limited number of labeled examples, but a very large number of additional, unlabeled images are available, and (2) images can naturally be represented in several distinct feature spaces, including *visual* feature spaces such as bag-of-visual words obtained from SIFT [20] descriptors or color histograms, as well as *textual* features such as a bag-of-words representation of surrounding text.

Note that these two properties are not unique to images. We can also view a multilingual text corpus as a collection of documents with different views corresponding to the different languages. And similarly, annotating texts usually requires expensive manual inspection, whereas we can acquire large amounts of unannotated textual material from the web or other large collections.

In this paper, we formalize these problems within the framework of multiview, semi-supervised learning. Our approach relies on combining the various possible representations—also called *views*—of each example. Each feature space is used to train a classifier on the labeled training set, resulting in one classifier per view. We then adopt a consensus-based self-learning algorithm similar to [3], to carry out semi-supervised learning: each *view* of an *unlabeled* example is pseudo-labeled if the corresponding single-view classifier of that view is confident about the class label for that example, and the examples for which the majority of the view-specific classifiers predict the same class label are pseudo-labeled and added to the initial training set. The view-specific classifiers are then re-trained, and the procedure is repeated until convergence is achieved. The measure of confidence used to assign pseudo-labels to unlabeled single-view examples is related to an automatic margin threshold which is computed at each iteration by minimizing the upper-bound of the transductive Bayes error. At test time, examples are labeled according to the majority vote of all single-view classifiers.

We report on experiments carried out in two contexts. We address the problem of text-image classification using the NUS-WIDE dataset [10], and the problem of multilingual document categorization using a multiview text categorization collection extracted from Reuters RCV1-RCV2 [3]. The experimental results show that our multiview

* Corresponding author.
E-mail address: Ali.Fakeri-Tabrizi@nominum.com (A. Fakeri-Tabrizi).
[1] www.flickr.com

http://dx.doi.org/10.1016/j.neucom.2014.12.041
0925-2312/

semi-supervised approach significantly outperforms various supervised and semi-supervised state-of-the-art baselines.

The main contribution of this paper is the extension of the self-training algorithm to multi-view, with an arbitrary number of views, using a confidence-based self-labeling approach with majority voting where the confidence level is strictly controlled in a principled way through an upper bound on the transductive error.

The following section briefly reviews the state of the art in multiview learning and self-learning. In Section 3, we then present the multiview self-learning algorithm that we propose. Section 4 describes our experimental setup and results.

2. State of the art

We position our approach in the context of related work on multiview learning (Section 2.1) and self-learning (Section 2.2).

2.1. Multiview learning

Multiview learning deals with observations that can be described in several representation spaces, such that each representation space may be used to build a predictor. These may be naturally different views on the same object, such as poses in object recognition, or different language versions of the same document. They may also be different sets of features, obtained from different means, such as color-based features vs. descriptive text for an image. The goal of multiview learning is to combine predictors over each view (called *view-specific* predictors) in order to improve the overall performance beyond that of predictors trained on each view separately, or on trivial combinations of views. The first successful multiview learning technique was [8]'s co-training algorithm, which iteratively labels unlabeled examples based on predictors trained in different views. A related approach is co-regularization [25] where the view-specific predictors are constrained to produce similar predictions. Other notable multiview techniques are multiple kernel learning [5], and techniques relying on (kernel) Canonical Correlation Analysis [17], multiview Fisher Discriminant Analysis [12], graph-based semi-supervised learning [36], hypergraph learning [35], spectral embedding [33] and click-through-based cross-view learning [23]. Note that although co-training and co-regularization have different theoretical backgrounds and motivation, empirical evidence shows that view-specific classifiers trained by the iterative co-training algorithm tend to agree on the pool of unlabeled data. The pseudo-labeling method of co-training can thus be seen as an iterative method for increasing the agreement between predictors. This issue will be at the core of our approach (Section 3.1).

The co-training algorithm has inspired some methods where a single input feature space is split or sampled in various ways in order to build an ensemble of models [37,32]. These are not multiview learning approaches in the sense of this paper as the "views" they create are artificial and usually redundant. They clearly differ from an approach such as ours where observations are naturally described from different representations.

2.2. Self-learning

Self-learning (or self-training) has been one of the first successful semi-supervised learning techniques [38]. The general idea is to automatically label unlabeled examples on which the model is sufficiently confident, and gradually add that to the labeled training data. This has been particularly effective in many Natural Language Processing tasks and used successfully, for example, on word sense disambiguation [34], entity classification [11] or statistical machine translation [30].

Other semi-supervised learning approaches typically rely on various assumptions. The *cluster assumption* states that examples within a given cluster are likely to be of the same class. This is particularly suited to generative mixture models [22]. *Low density separation* assumes that high-density regions do not contain the decision boundary [9], which leads to propagating class labels in high density regions. The *manifold assumption* holds when high dimensional data lie on a low-dimensional manifold [7]. In such cases the curse of dimensionality is avoided by operating in a low-dimensional space [6].

In the single view setting, one common problem with self-learning is that estimating labeling confidence may be difficult and errors in the self-labeling may impact system performance later on [29]. In co-training [8], examples are alternatively labeled on each view and used to train the classifier on the other view, in order to avoid biasing the model with its own prediction mistakes. It is also possible to use the multiple views by pooling prediction information from the multiple views and use consensus as a proxy for confidence in order to improve the self-labeling process. Consensus across multiple views has been exploited and shown to be effective, for example in ranking multilingual documents [31].

The approach presented in the next section extends this idea by exploring different ways to exploit the multiple predictions obtained on the different views.

3. Model

We first describe our semi-supervised, multiview approach for binary classification inspired by consensus-based algorithms. Notations used in the paper are given in Table 1.

3.1. Multiview self-learning

Denoting V the number of available views, each object is a tuple

$$\mathbf{x} \stackrel{\text{def}}{=} (x^1, \ldots, x^V)$$

where $x^v \in \mathcal{X}_v$ is the v-th view of example $\mathbf{x}$ and $\mathcal{X}_v$ is the corresponding input space. These can be various visual representations of an image (e.g. a bag-of-visual-words representation based on SIFT descriptors) or bag-of-words representations for textual content.

The data consists of a set of labeled training examples $\mathcal{Z}_\ell = \{(\mathbf{x}_i, y_i) | i \in \{1, \ldots, l\}\}$, with binary labels $y_i \in \{-1; +1\}$, and a set of unlabeled training data $X_{lu} = \{\mathbf{x}_i | i \in \{l+1, \ldots, l+u\}\}$. Our goal is to obtain V binary classifiers

$$\{h_v : \mathcal{X}_v \to \{-1, 1\} | v \in \{1, \ldots, V\}\}$$

Note that, in this section, we focus on binary classification.

Table 1
Notations.

Notation	Description
V	Number of views
$(x^v)_{v=1}^V$	V views of an observation $\mathbf{x} \stackrel{\text{def}}{=} (x^1, \ldots, x^V)$
$\mathcal{X}_v$	Input space corresponding to the v-th view
$\mathcal{Z}_\ell$	Labeled training set of size l
X_{lu}	Unlabeled training set of size u
$\mathcal{H}_v$	Hypothesis space associated with the v-th view
Q^v	Probability distribution over $\mathcal{H}_v$
B_{Q^v}	Bayes classifier for view v
G_{Q^v}	Gibbs classifier for view v
$R_u(\cdot)$	Transductive risk
$m_{Q^v}(x^v)$	Unsigned margin of the v-th view of example x^v

Please cite this article as: A. Fakeri-Tabrizi, et al., Multiview self-learning, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.12.041