Pattern Recognition Letters 37 (2014) 143–150

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Hinge loss bound approach for surrogate supervision multi-view learning

Gaole Jin, Raviv Raich *

School of EECS, Oregon State University, Corvallis, OR 97331-5501, United States

ARTICLE INFO

Article history:
Available online 22 June 2013

Keywords:
Multi-view learning
Surrogate supervision learning
Convex optimization

ABSTRACT

In multi-view learning, a classifier for different partitions (views) of the feature vector is commonly sought after. We consider the special case of surrogate supervision multi-view learning in which a classifier for one view is sought after, however, no labeled examples are available for that view. Instead, the training set consists of only labeled examples for the other view as well as unlabeled two-view data. While it is straightforward to train and test a classifier in the labeled view, it is challenging to perform the same task in the view where labels are unavailable. To solve this problem, we introduce an upper bound on the classical hinge loss (commonly used in support vector machines) that is well suited for the surrogate supervision multi-view learning setup. The bound only requires labeled examples from the other view and unlabeled examples of the two views. Using this approach, we introduce the surrogate supervision multi-class support vector machine (SSM–SVM). We evaluate the algorithm and compare it to other algorithms on a collection of datasets. We present an application of the algorithm to lip reading using audiovisual dataset.

1. Introduction

In multi-view learning a classifier for different partitions (views) of the feature vector is commonly sought after. Several problems are cast in the multi-view learning setting. Co-training is a semi-supervised multi-view learning technique directed at improving performance of a learning algorithm by expanding labeled training data using information from multiple views (Nigam and Ghani, 2000). For example, in Blum and Mitchell (1998) a set of labeled two-view examples and a set of unlabeled two-view examples are available. An assumption in Blum and Mitchell (1998) is that data from each of the views is sufficient for training an accurate classifier if labeled data are sufficient on both views. In transfer learning, the goal is to improve the classification performance on the target view using information from the auxiliary view (Pan et al., 2010). The regularized multi-task learning is a special case of transfer learning, where labeled data are available on both views (Evgeniou et al., 2006). The regularized multi-task learning algorithm transfers information from the auxiliary view to the target view to improve classification performance on the target view (Evgeniou et al., 2006).

In this work, we solve a novel multi-view learning problem – surrogate supervision multi-view learning (SSML) (Jin and Raich, 2012). In SSML, labeled data are available for only one view and unlabeled pairs are available for both views. The goal is to obtain a classifier on the view for which only unlabeled data are available. An important characteristic that distinguishes the SSML problem from other problems in the multi-view setting (e.g., co-training, multi-task leaning) is that in SSML the training of the desired view cannot be accomplished without information from the auxiliary view. Surrogate supervision multi-view learning can be applied in many areas. For example, classification of the same documents of different languages. Another example is the application to audiovisual data where video and audio are considered as two views. An intuitive solution to the SSML problem is to learn a classifier from one view and map it to the other view by learning the relationship between the two views through the unlabeled pairs. The canonical correlation analysis (CCA) technique can be used to obtain mapping from both views to a common representation space (Ngiam et al., 2011). Although this problem is fairly new, existing research in multi-view learning addresses similar issues. For example, the kernelized version of CCA (KCCA) is used in Vinokourov et al. (2003) and Li and Shawe-Taylor, 2006 to find the relationship between the same documents represented by different languages. One of the challenges with the aforementioned approach is that the components that are most correlated across views found by CCA are not necessarily optimal for classification. In Farquhar et al. (2005), the SVM-2 K algorithm combining the relationship (between views) learning stage and the classifier training stage into one is proposed. The experimental results show that the SVM-2 K algorithm outperforms the KCCA + SVM method. Counter to the setting in our paper, Farquhar et al. (2005) assumes

* Corresponding author. Tel.: +1 541 737 9862.
E-mail addresses: jing@eecs.oregonstate.edu (G. Jin), raich@eecs.oregonstate.edu (R. Raich).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.06.008

that the labeled data are available on both views. In addition Farquhar et al. (2005) does not give a solution to the multi-label classification problem.

The key difference between the problems addressed by this work to common multi-view learning problem is the availability of labels. In Blum and Mitchell (1998) a standard multi-view approach is taken with the assumption that a good classifier can be obtained from either view and their focus is how to combine information from both views. However, the problem in this paper assumes that no labeled examples are available for the desired view, that is, without information from the other view where labeled examples are available the classifier learning task (for the label-free view) will never be achieved. As a consequence, algorithms such as SVM-2 k or co-training which require labeled examples from both views need to be modified. Another issue is that the tuning of such algorithms cannot be done to optimize performance with respect to the desired view classifier, since no labeled examples from this view are available.

The contributions in this paper are as follows: (1) an SSML bound for the SVM hinge loss is derived for the binary classification problem; (2) we provide theoretical performance guarantees for a binary classifier obtained by minimizing the proposed bound; (3) an SSML bound for the SVM hinge loss is derived for the multiclass classification problem; (4) the SSM–SVM algorithm is proposed to solve the multi-class SSML problem; (5) SVM-2 k and co-regularization are modified for the SSML scenario; and (6) numerical evaluations are provided to analyze the performance of the proposed SSM–SVM and compare it with the performance of other algorithms including CCA + SVM, label-transferred learning, C^4A, SVM-2 k, and co-regularization.

In Section 2, we introduce the formulation of the SSML problem. Furthermore, we present a novel upper bound replacing SVM hinge loss for standard classification with its SSML counterpart and include theoretical performance guarantees on the classifier obtained by minimizing the proposed bound. In Section 3, we present the SSM–SVM algorithm based on the upper bound. Numerical evaluation of the algorithm is given in Section 4. Finally, Section 5 summarizes the paper.

2. Hinge loss upper bound for SSML

In this section, we introduce the setting of SSML. We then proceed with the derivation of the hinge loss upper bound for the binary-class case and extend the result to the multi-class case. For the binary classification case, we present theoretical performance guarantees on the classifier obtained by minimizing the upper bound.

2.1. Surrogate supervision multi-view learning

In a two-view learning scenario (a special case of multi-view learning), data can be represented as a set of triplets: $\{(x_i, z_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$, $z_i \in \mathcal{Z}$ are the two views, and $y_i \in \mathcal{Y} = \{1, 2, \ldots, K\}$ is the label. *Surrogate supervision multi-view learning* deals with the case where labels y_i are never directly provided for z_i. Instead, we are given two independent sets of data $\{(x_i, z_i)\}_{i=1}^m$ and $\{(x_i, y_i)\}_{i=m+1}^n$ and are interested in learning a classifier for y given z. Note that the two training sets are independent and hence not even a single triplet is available for training. The surrogate supervision multi-view learning is different from other standard multi-view learning settings (e.g. in Farquhar et al., 2005; Evgeniou et al., 2006, an assumption of the latter is that the labeled data are available on both views). One challenge of the SSML setting is to obtain the mapping form $\mathcal{Z}$ to $\mathcal{Y}$ without a single example of the form (z_i, y_i). Additionally, cross-validation to determine the unlabeled view classifier parameters which minimize the empirical risk cannot be performed. The calculation of the risk requires labels from the desired view which are unavailable in the SSML setting.

2.2. Bounding the hinge loss: the two-class case

We start with the binary-class case. In classification, the goal is to minimize the following classification error objective with respect to $g(\cdot)$:

$$E_{z,y}\left[\frac{1}{2}|g(z) - y|\right], \tag{1}$$

where $g(\cdot) : \mathcal{X} \to \mathcal{Y}$ is a decision function mapping feature space $\mathcal{X}$ to a label in $\mathcal{Y} = \{1, -1\}$. A common approach (e.g., in SVMs) is to replace the 0–1 loss in (1) with a hinge loss:

$$E_{z,y}[(1 - g(z)y)_+], \tag{2}$$

where $(t)_+ = \max\{0, t\}$. In SVM, a classifier is obtained by minimizing the regularized sample based objective: $\frac{1}{n}\sum_{i=1}^n [(1 - g(z_i)y_i)_+] + Pen(g)$, where $Pen(g)$ denotes a regularization term. For example, in a linear SVM $g(z) = w^T z$, the regularization term is $Pen(g) = \frac{\lambda}{2}\|w\|^2$. In the SSML scenario, labeled examples are only available for samples from $\mathcal{X}$. In the absence of examples of the type (z_i, y_i), one cannot compute directly the classifier which minimizes (2) or its regularized sample-based alternative.

Naturally, in SSML, we can only deal with objectives that are based on samples of the type (x_i, y_i) and (x_i, z_i) or equivalently objectives that require the joint distributions of (x, y) and (x, z). This leads us to considering an upper bound approach to a surrogate objective. Consider the following upper bound to (2):

$$E_{z,y}[(1 - g(z)y)_+] \leqslant E_{x,y}[(1 - h(x)y)_+] + E_{x,z}[|h(x) - g(z)|], \tag{3}$$

where $h(\cdot) : \mathcal{X} \to \mathcal{Y}$ is a classifier mapping feature space $\mathcal{X}$ to a surrogate objective on $\mathcal{Y}$. For a proof of (3), we refer the reader to Appendix A. The right-hand side (RHS) of (3) consists of two terms. The first is a hinge-loss for the classifier $h(\cdot)$ measuring how well $h(\cdot)$ can predict the label y, while the second term measures how close are the predictions of the two classifiers $g(\cdot)$ and $h(\cdot)$. Note that the LHS cannot be empirically evaluated since no labeled examples of the type (z, y) are available. However, the RHS which requires only pairs of the type (x, y) and (x, z) can be empirically evaluated. Hence the RHS can be used as a realizable bound for the SVM hinge-loss on the LHS of (3). In other words, the objective on the RHS of (3), promotes a classifier $h(\cdot)$ on $\mathcal{X}$ which can simultaneously predict y and can be well-approximated by a classifier $g(\cdot)$ on $\mathcal{Z}$. Note that since the bound holds for any $h(\cdot)$, the bound can be tightened by minimizing the RHS w.r.t. $h(\cdot)$. This bound suggests the replacement of the hinge loss in one view with the hinge loss in the other view plus a multi-view classifier mismatch term. Next, we present a theorem which examines the performance achieved by using the RHS of (3) as a surrogate for the original SVM hinge loss.

Theorem 1. *Let $h(x)$ and $g(z)$ be classifiers $\mathcal{X} \to \mathcal{Y}$ and $\mathcal{Z} \to \mathcal{Y}$, respectively. Define the correlation distance between classifiers $h(x)$ and $g(z)$ as $d(g, h) = E[|h(x) - g(z)|]$. We denote $\tilde{g}$ and $\tilde{h}$ as the minimizers of the RHS of (3). Similarly, we denote g^* as the minimizer of the SVM hinge loss term in (2). The loss achieved by $\tilde{g}$, $E[(1 - \tilde{g}(z, y))_+]$ is bounded using the following two-sided inequality:*

$$E[(1 - g^*(z)y)_+] \leqslant E[(1 - \tilde{g}(z)y)_+] \leqslant \delta + E[(1 - g^*(z)y)_+] \tag{4}$$

where $\delta = \max_h \min_g E[|g(z) - h(x)|] + \max_g \min_h E[|g(z) - h(x)|]$.