Pattern Recognition Letters 37 (2014) 63–77

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Boosting for multiclass semi-supervised learning

Jafar Tanha *, Maarten van Someren [1], Hamideh Afsarmanesh [1]

Informatics Institute, University of Amsterdam, Science Park 904, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 21 October 2013

Keywords:
Semi-supervised learning
Boosting
Multiclass classification

ABSTRACT

We present an algorithm for multiclass semi-supervised learning, which is learning from a limited amount of labeled data and plenty of unlabeled data. Existing semi-supervised learning algorithms use approaches such as one-versus-all to convert the multiclass problem to several binary classification problems, which is not optimal. We propose a multiclass semi-supervised boosting algorithm that solves multiclass classification problems directly. The algorithm is based on a novel multiclass loss function consisting of the margin cost on labeled data and two regularization terms on labeled and unlabeled data. Experimental results on a number of benchmark and real-world datasets show that the proposed algorithm performs better than the state-of-the-art boosting algorithms for multiclass semi-supervised learning, such as SemiBoost (Mallapragada et al., 2009) and RegBoost (Chen and Wang, 2011).

1. Introduction

Supervised learning methods are effective when there are sufficient labeled instances. In many applications, such as object detection, document and web-page categorization, labeled instances however are difficult, expensive, or time consuming to obtain because they require empirical research or experienced human annotators. Semi-supervised learning algorithms use not only the labeled data but also the unlabeled data to build a classifier. The goal of semi-supervised learning is to combine the information in the unlabeled examples with the explicit classification information of labeled examples for improving the classification performance (Chapelle et al., 2006).

However, most existing semi-supervised methods were designed for binary classification problems (Bennett et al., 2002; Belkin et al., 2006; Mallapragada et al., 2009). To solve the multiclass classification problem two main approaches have been proposed. The first is to convert the multiclass problem into a set of binary classification problems. Examples of this approach include one-vs-all, one-vs-one, and error-correcting output code (Dietterich and Bakiri, 1995). This approach can have various problems such as imbalanced class distributions, increased complexity, no guarantee to obtain an optimal joint classifier or probability estimation, and different scales for the outputs of generated binary classifiers which complicates combining them, see Jin and Zhang (2007) and Saberian and Vasconcelos (2011). The second approach is to use a multiclass classifier directly. Although a number of methods have been proposed for multiclass supervised learning, for example Mukherjee and Schapire (2013), Zhu et al. (2009) and Saberian and Vasconcelos (2011), they are not able to handle multiclass semi-supervised learning, which is the aim of this study.

In this paper we present a boosting algorithm for multiclass semi-supervised learning, named Multi-SemiAdaBoost (MSAB). Unlike many semi-supervised learning algorithms that are extensions of specific base classifiers to semi-supervised learning, such as Semi-Supervised SVM (Bennett and Demiriz, 1999), low density separation (Chapelle et al., 2006), Transductive SVM (Joachims, 1999), and LapSVM (Belkin et al., 2006), MSAB can boost any base classifier. It minimizes both the empirical error on labeled data and the inconsistency over labeled and unlabeled data based on both cluster and manifold assumption (Chapelle et al., 2006). This generalizes the SemiBoost (Mallapragada et al., 2009) and the RegBoost (Chen and Wang, 2011) algorithms from binary to multiclass classification using a coding scheme for the multiclass classification problem (Zhu et al., 2009). Our proposed method uses the margin on labeled data, the similarity among labeled and unlabeled data, and the similarity among unlabeled data in an exponential loss function. We give a formal definition of this loss function and derive functions for the weights of classifiers and unlabeled data by minimizing an upper bound on the objective function. We then compare the performance of the algorithm to (a) binary algorithms with smoothness regularizer used in the one-versus-all scheme to handle multiclass classification (RegBoost and SemiBoost), (b) a boosting multiclass semi-supervised learner without smoothness regularizer (Song et al., 2011 and Bennett et al., 2002), and (c) the supervised multiclass AdaBoost learning algorithm (Zhu et al., 2009), which is trained only on labeled data, to evaluate the effect of using unlabeled data. The results of the experiments on the benchmark UCI datasets show

* Corresponding author. Tel.: +31 20 5258028; fax: +31 20 5257490.
E-mail addresses: J.Tanha@uva.nl (J. Tanha), M.W.vanSomeren@uva.nl (M. van Someren), H.Afsarmanesh@uva.nl (H. Afsarmanesh).
[1] Tel.: +31 20 5257512; fax: +31 20 5257490.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.008

that MSAB outperforms the other boosting algorithms and gives the best results. We also present a variation of the MSAB algorithm to show the effect of boosting on labeled data. We further apply the MSAB method on real-world application problems, text classification and bird behavior recognition, to show how MSAB can exploit information from unlabeled data to improve the classification accuracy.

This paper is organized as follows: Section 2 addresses the related work on semi-supervised learning. Sections 3 and 4 formalize the setting and the loss function. Section 5 derives the weights for the boosting algorithm. The variation of the proposed algorithm is presented in Section 6. Sections 7 and 8 present the experiments and the results, the results of experiments on real-world applications are presented in Sections 9 and 10, and finally Section 11 draws the main conclusions.

2. Related work

Methods for semi-supervised learning vary in the underlying assumptions. Two main semi-supervised assumptions are the manifold and cluster assumption (Chapelle et al., 2006).

The manifold assumption is that the (high-dimensional) data lie on a low-dimensional manifold. Many graph-based semi-supervised learning methods are based on this assumption, for example Markov random walks (Jaakkola, 2002), Label propagation (Zhu and Ghahramani, 2002), and Local and Global consistency (Zhou et al., 2004). These methods build a graph based on the pairwise similarity between examples (labeled and unlabeled). The goal is then to estimate a function on the graph that minimizes the loss on labeled examples and is smooth on the whole graph. In other words, graph-based methods learn the manifold structure of the feature space with labeled and unlabeled data and use this information to improve the performance of the supervised learning algorithm, see Fig. 1. Manifold regularization (Belkin et al., 2006) is another discriminative learning algorithm based on the manifold assumption. It constructs a large-margin classifier on the data while minimizing the corresponding inconsistency between data using the similarity matrix. Our proposed method is closely related to the graph-based approaches in the sense that it uses the pairwise similarity. We utilize the inconsistency measure as in graph-based methods using an exponential form of it.

The cluster assumption says that the data space consists of a number of clusters and if points are in the same cluster, then they likely belong to the same class. The cluster assumption thus says that the decision boundary should lie in a low-density region (Chapelle et al., 2006). In other words, the cluster assumption emphasizes that the examples being in very dense regions of the feature space are likely to share the same class label, see Fig. 1. Many successful semi-supervised learning methods are based on this assumption, such as Semi-Supervised SVM (Bennett and Demiriz, 1999), low density separation (Chapelle et al., 2006), and Transductive SVM (Joachims, 1999). These methods basically extend the SVM classifier to semi-supervised learning, and are not easily extensible to non-margin based learners such as decision trees (Mallapragada et al., 2009).

Another dimension of the semi-supervised learning methods – especially for those that are based on cluster assumption – is whether they perform iterative or additive improvement. Iterative improvement algorithms (e.g. self-training Rosenberg et al., 2005 and co-training Blum and Mitchell, 1998) replace their hypothesis at each iteration by a new one. Additive algorithms add a new component to a linear combination of classifiers as in boosting algorithms (Freund and Schapire, 1996). Boosting is one of the most successful ensemble methods for supervised learning. It has been extended to semi-supervised learning, e.g. MarginBoost (dAlché Buc et al., 2002) and ASSEMBLE (Bennett et al., 2002). MarginBoost and ASSEMBLE use a base classifier to predict class labels of the unlabeled examples, the "pseudo-labels". A sample of the pseudo-labeled data is then used in the next iteration. The main difficulty in this approach is how to assign pseudo-labels to the unlabeled data and then how to sample from them. These algorithms typically attempt to minimize the margin cost of the labeled data and a cost associated with the "pseudo-labels" of the unlabeled data. The pseudo-labels and also the associated cost depend strongly on the classifier predictions. Therefore, this type of algorithm cannot effectively exploit information from the unlabeled data and the final decision boundary will be very close to that of the initial classifier, see Chen and Wang (2011) and Mallapragada et al. (2009).

To solve the above problem, a recent approach is to use a smoothness regularizer based on the cluster and the manifold assumptions. The idea is to not use the "pseudo-margin" from the predictions of the base learner directly. Instead, beside the margin on the labeled data, also the "consistency" is maximized. Consistency is a form of smoothness. Class labels are consistent if they are equal for data that are similar. For this the computation of the weights for data and for classifiers in the boosting algorithm must be modified. This is done in SemiBoost (Mallapragada et al., 2009) and RegBoost (Chen and Wang, 2011). Experiments show

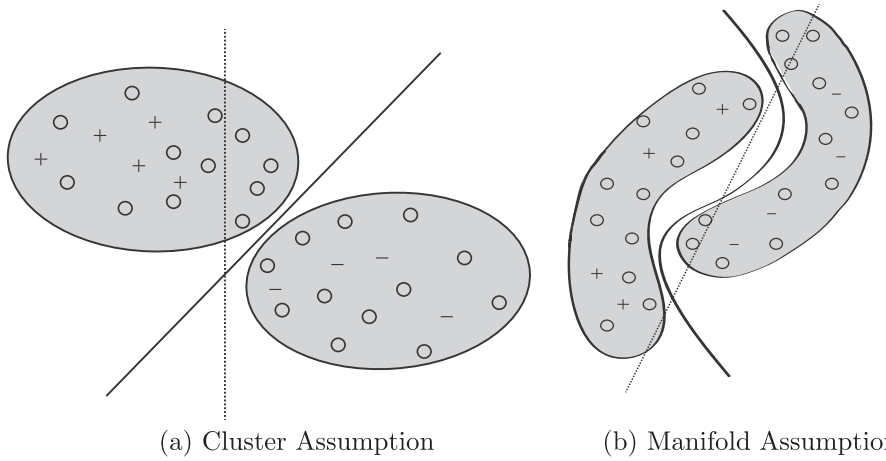

Fig. 1. The positive and negative signs show labeled examples from two different classes. The circles depict the unlabeled examples. The dashed line for decision boundary is obtained by only training on labeled examples and usually crosses the dense regions of the feature space. These decision boundaries are moved to regions with lower density (solid line) using unlabeled data.