Pattern Recognition 44 (2011) 2305–2313

Contents lists available at ScienceDirect

Pattern Recognition

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

A fast quasi-Newton method for semi-supervised SVM

I. Sathish Reddy *, Shirish Shevade, M.N. Murty

Computer Science and Automation Department, Indian Institute of Science, Bangalore 560012, India

ARTICLE INFO

Available online 8 September 2010

Keywords:
Semi-supervised learning
Support vector machines
Quasi-Newton methods
Nonconvex optimization

ABSTRACT

Due to its wide applicability, semi-supervised learning is an attractive method for using unlabeled data in classification. In this work, we present a semi-supervised support vector classifier that is designed using quasi-Newton method for nonsmooth convex functions. The proposed algorithm is suitable in dealing with very large number of examples and features. Numerical experiments on various benchmark datasets showed that the proposed algorithm is fast and gives improved generalization performance over the existing methods. Further, a non-linear semi-supervised SVM has been proposed based on a multiple label switching scheme. This non-linear semi-supervised SVM is found to converge faster and it is found to improve generalization performance on several benchmark datasets.

1. Introduction

In the last few years, remarkable research work has been done in supervised learning. Most of these learning models apply the inductive inference concept, where prediction function, derived only from the labeled input data, is used to predict the label of any unlabeled object. A well-known two-class classification technique is based on the *support vector machine* (SVM) Burges [1], where a SVM solution corresponds to the maximum margin between the two classes of labeled objects under consideration. But in many applications, labeled data are scarce, manual labeling for the purposes of training SVMs is often a slow, expensive, and error-prone process. On the other hand, in many applications of machine learning and data mining, abundant amounts of unlabeled data can be cheaply and automatically collected. Some examples are text processing, web categorization, medical diagnosis, and bioinformatics. In spite of its natural and pervasive need, solutions to the problem of utilizing unlabeled data with labeled examples have only recently emerged in machine learning literature. Using both labeled and unlabeled data for the purpose of learning is called semi-supervised learning. An interested reader can refer to Zhu [19] for a nice review on semi-supervised learning.

A major body of work in *semi-supervised SVMs* (S^3VM) is based on the following idea Chapelle et al. [2]: solve the standard SVM problem while treating the unknown labels as additional optimization variables. By maximizing the margin in the presence of unlabeled data, one learns a decision boundary that traverses through low data-density regions while respecting labels in the input space. In other words, this approach implements the cluster assumption for semi-supervised learning-that is, points in a data cluster likely to have same class labels.

S^3VM might seem to be the perfect semi-supervised algorithm, since it combines the powerful regularization of SVMs with a direct implementation of cluster assumption. However, its main drawback is that the objective function is nonconvex and thus is difficult to optimize. Due to this reason, a wide spectrum of techniques have been applied to solve the nonconvex optimization problem associated with S^3VMs, for example, local combinatorial search Joachims [3]; gradient descent Chapelle et al. [2]; continuation techniques Chapelle et al. [15]; convex–concave procedures Fung et al. [9]; branch-and-bound algorithms Bennett et al. [8].

In this work, we propose to use the S^3VM to solve semi-supervised classification problems. In particular, we adopt the model described in Joachims [3], focusing on the specific features of the optimization problem to be solved, which can be formulated as a nonsmooth nonconvex minimization problem. To tackle this problem, we use a quasi-Newton method described in Yu et al. [6]. The main contributions made by us and reported in this work are:

1. We outline an implementation of a variant of S^3VM Joachims [3] designed for linear semi-supervised classification on large datasets. As compared to state-of-the-art large scale semi-supervised learning techniques described in Sindhwani et al. [5,10], our method effectively exploits data sparsity and linearity of the problem to provide superior scalability. The improved generalization performance and training speed turn the proposed scheme into a feasible tool for large scale applications.
2. We outline an implementation of a variant of S^3VM Joachims [3]; Collobert et al. [10] designed for non-linear semi-supervised

* Corresponding author.
E-mail addresses: sathish.indurthi@gmail.com (I.S. Reddy), shirish@csa.iisc.ernet.in (S. Shevade), mnm@csa.iisc.ernet.in (M.N. Murty).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.002

classification on the datasets, where it is difficult to find a linear decision boundary in the input space.
3. We conducted an experimental study on many binary classification tasks with several thousands of examples and features. This study clearly shows the usefulness of our algorithm for large scale semi-supervised classification.
4. In summary, the work is carried out with an efficient scheme for semi-supervised learning based on both linear and non-linear SVMs.

This paper is organized as follows: Section 2 analyzes the S^{3}VM objective function and studies its characteristics. In Section 3 we describe the quasi-Newton method for nonsmooth convex functions. In Section 4 we present S^{3}VM implementation using quasi-Newton method. Section 5 compares our work with other recent efforts in this area. Experimental results are reported in Section 6. Section 7 contains some useful concluding comments.

Throughout the paper, we adopt the following notation: We denote by $\|\cdot\|$ the Euclidean norm in $\Re^d$ and by a^Tb or $a\cdot b$ the inner product of the vectors a and b. Moreover, the subdifferential of a convex function f at any point a is denoted by $\partial f(a)$. We recall that the subdifferential of a convex function f at point a is the set of the *subgradients* of f at a, that is, the set of vectors $g\in\Re^d$ satisfying the *subgradient inequality*,

$$f(b)\geq f(a)+g^T(b-a)\quad \forall b\in\Re^d.$$

2. Semi-supervised SVMs

We consider the problem of binary classification. The training set consists of l labeled examples $\{x_i,y_i\}_{i=1}^{l}, y_i\in\{-1,+1\}$, and u unlabeled examples $\{x_i\}_{i=l+1}^{n}$, with $n=l+u$, typically, $l\ll u$ and $x_i\in\Re^d$. Our goal is to construct a classifier that utilizes unlabeled data and gives better generalization performance.

S^{3}VM appends an additional term in the SVM objective function whose role is to find a hyperplane far away from both the labeled and the unlabeled points. Variants of this idea have appeared in the literature Joachims [3], Bennett et al. [8], and Fung et al. [9]. Since the formulation in Joachims [3] appears to be the most natural extension of standard SVMs among these methods, we will focus on developing its large scale implementation.

2.1. Linear case

The following optimization problem is setup for standard S^{3}VM:

$$J(w)\triangleq\min_{w,\{y_i\}_{i=l+1}^{n}}\frac{\lambda}{2}\|w\|^2+\frac{1}{l}\sum_{i=1}^{l}l_1(y_i,w^Tx_i)+\frac{\lambda'}{u}\sum_{i=l+1}^{n}l_2(|w^Tx_i|)$$

$$\text{s.t.}\frac{1}{u}\sum_{i=l+1}^{n}\max(0,y_i)=r\ \text{ where } y_i=\text{sign}(w^Tx_i),\tag{1}$$

where l_1 is a loss function of w which measures the discrepancy between y_i and the predictions arising from w via w^Tx_i. A loss function commonly used for binary classification is the hinge loss with $y_i\in\{-1,+1\}$ and the corresponding loss function is depicted in Fig. 1(a). l_2 is a margin penalty function involving the unlabeled data, and λ,λ' are positive weight parameters. In Chapelle et al. [2], Joachims [3,8], Fung et al. [9], and Chapelle et al. [15] the loss function l_2 is chosen as

$$l_1(y_i,w^Tx_i):=\max(0,1-y_i(w^Tx_i)),\tag{2a}$$

$$l_2(t)=\begin{cases}1-|t| & \text{for } -1<t<1,\\ 0 & \text{otherwise}\end{cases}\tag{2b}$$

and the proposed algorithms therein are characterized by different approaches to solve problem (1). The algorithm proposed in Joachims [3] learns first the SVM classifier by using the labeled dataset. Using this classifier, unlabeled patterns are labeled. Then, the current solution is improved by switching the labels of some unlabeled samples, selected on the basis of appropriate heuristic techniques.

In Bennett et al. [8], the authors formulate the semi-supervised SVM problem as a mixed integer program. Since they introduce a binary variable for each unlabeled point, the problem can be difficult to solve for a large number of unlabeled data. To avoid this difficulty, in Fung et al. [9], a concave minimization problem is tackled, and a stationary point is found by solving successive linear programs.

In Eq. (1), the S^{3}VM seeks a hyperplane w and the labels of the unlabeled examples, so that the SVM objective function is minimized, subject to the constraint that a fraction r of the unlabeled examples be classified positive. But, the main drawback of the objective function in (1) is that, it is not differentiable, see Fig. 1(a) and moreover, due to the third term involving the unlabeled points, it is even nonconvex, see Fig. 1(b).

Methods for solving problem (1) are reported in Chapelle et al. [2,15], where in Chapelle et al. [2] the authors perform a standard gradient descent method on a smooth approximation of the objective function. In Chapelle et al. [15], the authors use a deterministic annealing approach on a smoothed version of objective function. Our approach is, indeed, to adopt some recently proposed methods Yu et al. [6], Belloni [11], Teo et al. [12] of the subgradient type, which are capable to cope with nonsmoothness and nonconvex part is dealt by using multiple label switching procedure Sindhwani et al. [5].

In the literature, some approaches introducing nonconvex loss functions have been proposed. In particular, in Collobert et al.

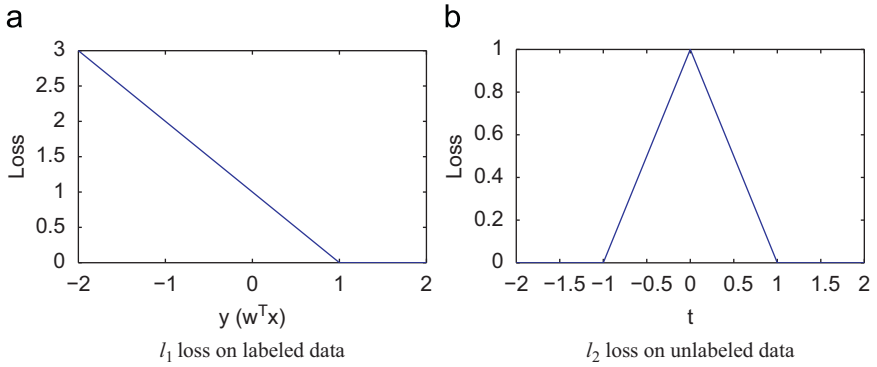

Fig. 1. Characteristics of loss functions (defined in (2a) and (2b)). (a) l_1 loss on labeled data. (b) l_2 loss on unlabeled data.