Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

A primal–dual method for SVM training

Samia Djemai [a,*], Belkacem Brahmi [a,b], Mohand Ouamer Bibi [a,b]

[a] LaMOS Research Unit, University of Bejaia, Targa Ouzemmour, 06000 Bejaia, Algeria
[b] Department of Operations Research, University of Bejaia, Targa Ouzemmour, 06000 Bejaia, Algeria

ARTICLE INFO

Article history:
Received 18 July 2015
Received in revised form
31 December 2015
Accepted 2 January 2016

Keywords:
Support Vector Machines (SVM)
Convex quadratic problem
Adaptive method
Suboptimality estimate

ABSTRACT

Training support vector machines (SVM) consists of solving a convex quadratic problem (QP) with one linear equality and box constraints. In this paper, we solve this QP by a primal–dual approach that combines the adaptive method with an interior point method. To initialize the algorithm, a procedure of an interior point method is used to construct an initial support. The proposed approach provides an efficient implementation of a new algorithm that exploits the advantage of the adaptive method for training SVM problems. It is based on the principle of the support and the suboptimality estimate. Experimental results confirm the efficiency of our approach over state-of-the-art SVM algorithms such as SMO, LIBSVM and SVMLight for medium-sized problems.

1. Introduction

During more than ten years, several researches on machine learning have been focused on Support Vector Machines [1–4]. This concept has a high generalization performance that is due to its strong theoretical foundation based upon statistical learning theory [1]. The idea of the SVM is based on the notions of large margin and kernel functions. It aims to find a hyperplane that separates, in a best way, the examples of a sample into several classes.

The SVM has shown excellent performance for a wide class of applications, such as information processing [5], handwritten digit recognition [6], face recognition [7], financial engineering [8], database analysis [9] and bioinformatics [10].

Training an SVM classifier is reduced to a convex quadratic minimization problem with one linear equality constraint and box constraints. Several approaches have been proposed to solve this kind of problems. The most efficient ones are the decomposition methods and the active set methods. The basic idea of these methods is similar. It consists of splitting the original problem into subproblems that are easier to solve. However, the first approach, also called working set [11], operates by optimizing a fixed subset of the variables per iteration to minimize the training time. This technique has the advantage of dealing with large datasets. Another advantage of the decomposition methods consists of their limited use of memory. Efficiently implemented decomposition algorithms have been introduced such as SMO (Sequential Minimal Optimization) described in [12], it is an extreme case of the decomposition methods which solves a two-variables problem at each iteration. There is also LIBSVM [13] that is considered as an SMO-type decomposition method, it modifies a fixed size subset of variables per iteration. In addition to those, SVMlight [14] is based on the decomposition idea introduced in [15], it has the same principle of fixed size working set and it uses the heuristic of shrinking. However, since only few components are changed in each iteration, then the decomposition methods suffer from the slow convergence.

The second approach is shown to be suitable when the Hessian is dense [16,17] and the solution is sparse [18], that is generally the case of the SVM problems. The active set method is also recommended for both small and medium-sized problems [19]. Its advantage is the ability of performing incremental/decremental training [20], improved accuracy [16] and improved stability and convergence [18]. However, from existing active set implementations, in the SVM-QP introduced in [16], the subproblem solved in each iteration can be singular and thus the convergence is not always ensured. Another active set algorithm is SVM-RSQP described in [17]. It maintains the non-singularity of the inner problem compared to the dual active set method presented in [21] for SVM which does not need to compute the inverse of a matrix.

The simplex method [22] solves the quadratic problems and it is used for SVM training [16]. The revised simplex method [23] is also used to improve the training of the SVM problem [17].

The adaptive method [24–26] is intermediate between active set and interior point methods. It belongs to the class of primal–dual methods which are based on the primal and dual information for solving convex quadratic problems with bounded variables. It takes into account the specificities of the constraints and thus

* Corresponding author.
E-mail address: samia_djemai@hotmail.fr (S. Djemai).

http://dx.doi.org/10.1016/j.neucom.2016.01.103
0925-2312/

treats them as they arise, without trying to transform them. This avoids to extend the dimensions of the problem and therefore preserves space in the memory. The principle of this method is to use an adaptive metric that changes all non-optimal indices at the same time, thus generalizing that of the simplex method. Indeed, the adaptive method is based on the notion of the support which is a more general concept than the basis concept in linear programming. Therefore, the feasible solution and the support can be constructed independently from each other compared to a basic feasible solution. We notice that a support feasible solution can be an extreme point, a boundary point or an interior point contrary to a basic feasible solution.

In this paper, we propose to solve the SVM problems by using the adaptive method of quadratic programming. The application of this method is new. By adapting this method to the structure of SVM problems, we produce an effective new approach. The deduced algorithm is iterative and it consists of two phases. The first phase is inspired from the affine scaling method that is one of the simplest interior point methods [27]. It deals with the construction of an initial support. The second phase is based on the application of methods [26,28,29] for determining an optimal solution. Each iteration of this phase consists of computing a descent direction and a step along this direction to improve the value of the objective function. If the current feasible solution is not optimal, we change the support in such a way that the new support matrix is nonsingular.

The main idea of this method is to change all the non-optimal indices at the same time, accelerating the convergence of the algorithm, while the most common approach to training SVM problems is to allow only a small number of variables to be changed.

Another particularity of our algorithm is that it uses the suboptimality estimate which allows us to stop the algorithm with a desired accuracy. This could be useful in practical applications. In addition, our algorithm starts with one index when the other methods begin with a subset of indices.

The proposed method is implemented in Matlab. We compare its performances with those of the popular SVM training algorithms, like SMO, LIBSVM and SVMLight, using several selected benchmark datasets from the UCI Machine Learning repository [30].

This paper is organized as follows: in Section 2, we present the modeling of classification problem by SVM as a convex quadratic minimization problem, then we define its parameters and present the KKT optimality conditions. In Section 3, we give some definitions and present the suggested algorithm. Experimental results for illustration purpose are presented in Section 4. Finally, in the last section, we conclude the paper and give some perspectives.

2. SVM classification

We are given a training example (x_i, y_i), $i \in I = \{1, 2, \ldots, n\}$, where $y_i \in \{-1, +1\}$ is the class of the example $x_i \in \Re^p$, n is the number of training examples and p is the number of features of each example. The binary SVM classification consists of solving the following constrained minimization problem:

$$\min_{w,\xi} \quad \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i, \tag{1a}$$

$$\text{subject to} \quad y_i [w' \phi(x_i) + b] \geq 1 - \xi_i, \tag{1b}$$

$$\xi_i \geq 0, \quad i \in I = \{1, 2, \ldots, n\}, \tag{1c}$$

where the nonlinear mapping ϕ projects each example $x_i \in \Re^p$ in a transformed feature space F with a larger dimension if the examples of the initial sample are not linearly separable by a simple hyperplane of equation $w'x + b = 0$. By this projection, the new data sample $(\phi(x_i), y_i)$, $i \in I$, becomes linearly separable by a hyperplane in the space F and the training sample may be separable by a nonlinear surface in the original input space. The real b is called the bias and $w \in F$ is the weight vector of the optimal hyperplane. The symbol (') represents the transposition operation. The surplus variable ξ_i represents the error associated with the margin of the ith example relative to the separating hyperplane. The regularization parameter C controls the misclassification of the sample, and it is used to penalize the variables ξ_i. A high value of C corresponds to assign a large penalty to errors.

Because of the possibility of a very large vector $w \in \mathcal{F}$, one usually solves the dual of the primal problem (1), that is equivalent to the following quadratic problem:

$$\min_{\alpha} L(\alpha) = \frac{1}{2} \alpha' D \alpha - e' \alpha, \tag{2a}$$

$$y' \alpha = 0, \tag{2b}$$

$$0 \leq \alpha_i \leq C, \quad i \in I, \tag{2c}$$

where $\alpha = \alpha(I) = (\alpha_i, i \in I) \in \mathbb{R}^n$ is the vector of Lagrange multipliers associated to the constraints (1b) of the primal problem. Here, e is an n-vector of ones and y is an n-vector, with $y = y(I) = (y_i, i \in I)$. The matrix D, defined by its elements $d_{ij} = y_i y_j k(x_i, x_j)$, is square of order n, symmetric and positive semidefinite, since the kernel function k, checking $k(x_i, x_j) = \phi(x_i)' \phi(x_j)$, respects the condition of Mercer [31].

2.1. KKT optimality conditions

The optimality conditions of Karush–Kuhn–Tucker (KKT) are crucial because they allow us to characterize the solutions and establish a strategy to construct algorithms for solving the problem. Since the primal problem (1) and its equivalent dual (2) are convex, then the KKT first order optimality conditions are both necessary and sufficient for the optimality of a feasible point α.

Using the Lagrangian duality theory [22], the Lagrange function associated with the dual problem is the following:

$$G(\alpha, b, s, \xi) = \frac{1}{2} \alpha' D \alpha - e' \alpha + b y' \alpha - s' \alpha + \xi' (\alpha - Ce), \tag{3}$$

where the nonnegative n-vector s represents the slack variables.

The KKT optimality conditions of the previous two problems are:

$$\frac{\partial G}{\partial \alpha} = D\alpha - e + by - s + \xi = 0, \tag{4a}$$

$$\frac{\partial G}{\partial b} = y' \alpha = 0, \tag{4b}$$

$$\xi_i (\alpha_i - C) = 0, \quad s_i \alpha_i = 0, \quad i \in I, \tag{4c}$$

$$s_i \geq 0, \quad \xi_i \geq 0, \quad i \in I, \tag{4d}$$

$$0 \leq \alpha_i \leq C, \quad i \in I. \tag{4e}$$

Please cite this article as: S. Djemai, et al., A primal–dual method for SVM training, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.01.103