Pattern Recognition 55 (2016) 261–272

Contents lists available at ScienceDirect

Pattern Recognition

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

Incremental p-margin algorithm for classification with arbitrary norm

Saulo Moraes Villela, Saul de Castro Leite, Raul Fonseca Neto *

Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Received 12 September 2015
Received in revised form
9 January 2016
Accepted 19 January 2016
Available online 29 January 2016

Keywords:
Large margin classifiers
p-Norm
Perceptron algorithms
Binary classification

ABSTRACT

This paper presents a new algorithm to approximate large margin solutions in binary classification problems with arbitrary q-norm or p-margin, where p and q are Holder conjugates. We begin by presenting the online fixed p-margin perceptron algorithm (FMP_p) that solves linearly separable classification problems in primal variables and consists of a generalization of the fixed margin perceptron algorithm (FMP). This algorithm is combined with an incremental margin strategy called IMA_p, which computes an approximation of the maximal p-margin. To achieve this goal, IMA_p executes FMP_p several times with increasing p-margin values. One of the main advantages of this approach is its flexibility, which allows the use of different p-norms in the same primal formulation. For non-linearly separable problems, FMP_p can be used with a soft margin in primal variables. The incremental learning strategy always guarantees a good approximation of the optimal p-margin and avoids the use of linear or higher order programming methods. IMA_p was tested in different datasets obtaining similar results when compared to classical L_1 and L_∞ linear programming formulations. Also, the algorithm was compared to $ALMA_p$ and presents superior results.

1. Introduction

One of the main problems in machine learning consists in determining a linear function able to discriminate a set of vectors in input space belonging to two different classes. A critical issue in this scenario is achieving a high generalization performance. In this sense, the development of maximal margin classifiers, such as Support Vector Machines (SVMs), was a significant contribution to the field [1]. In the context of linearly separable problems, SVMs determine a hyperplane that maximizes the margin between the two classes. This problem has been originally formulated as a quadratic optimization problem, where the objective is to minimize the Euclidean norm of the normal vector. For large datasets this solution may be inefficient in terms of computational costs. Also, it is known that the p-norm minimization, for $p=1$ or $p=\infty$, results in a linear programming problem with a much lower computational cost and similar generalization performance. Indeed, formulations based on linear programming have been developed as an option for SVMs, providing scalability and lower computational costs when compared with quadratic programming solutions [2–4]. However, in order to obtain the maximal margin hyperplane, these formulations are designed specifically for either L_1 or L_∞ norms. In addition, they are usually solved in batch mode, since they require the use of linear programming solvers.

In recent years, considerable attention has been given to finding efficient online algorithms to construct large margin classifiers that avoid the complexity of quadratic programming [5–7]. Some of these are motivated by the fact that, usually, an adequate approximation of the maximal margin is sufficient to have good generalization performance. For instance, Gentile [5] proposes a p-norm formulation and introduces the Approximate Large Margin Algorithm ($ALMA_p$), which can be used to find an α-approximation to the maximal p-margin for $p \geq 2$. In addition, Leite and Fonseca Neto [7] present the Incremental Margin Algorithm (IMA), which is able to obtain a large L_2 margin solution by successively solving classification problems with increasing margin sizes, using the Fixed Margin Perceptron (FMP). In terms of theoretical bound, it has been shown in [7] that IMA has the same upper bound as $ALMA_2$ for the number of updates needed to obtain an α-approximation to the maximal margin, i.e., $O\left(\frac{R^2}{\alpha^2\gamma^{*2}}\right)$. However, computational experiments demonstrate that IMA outperforms $ALMA_2$ in both accuracy and computational efficiency [7].

This paper extends the results presented by Leite and Fonseca Neto [7] by first introducing the Fixed p-Margin Perceptron (FMP_p), which solves binary classification problems for any given feasible fixed p-margin, for *any* p (i.e., for $p \geq 1$, including $p=\infty$). The FMP_p algorithm is then coupled with IMA to produce the Incremental p-Margin Algorithm (IMA_p), which is able to compute an α-approximation to the maximal p-margin by solving

* Corresponding author. Tel.: +55 32 2102 3311.
E-mail addresses: saulo.moraes@ufjf.edu.br (S.M. Villela), saul.leite@ufjf.edu.br (S.C. Leite), raulfonseca.neto@ufjf.edu.br (R. Fonseca Neto).

http://dx.doi.org/10.1016/j.patcog.2016.01.016
0031-3203/

successive classification problems using FMP_p with increasing p-margins values. For $p=2$, IMA_2 and FMP_2 are equivalent to IMA and FMP presented in [7]. Based on the theoretical bound on the number of updates for IMA_2 [7], we propose in this paper a decaying rule for the learning rate among the successive calls to FMP_p, which improves the efficiency of the method. We show through numerical experiments that IMA_p produces superior results to $ALMA_p$ and has the advantage that it can be interrupted any time after the first call to FMP_p. The correctness of the IMA_p is demonstrated by comparing the results for L_1 and L_∞ norms with the exact solution obtained from the linear programming formulations. This paper also introduces a novel strategy to allow soft margins in the primal variables for non-linearly separable datasets and for data with outliers. Flexible margins are very useful in the p-norm setting since no dual formulation is possible for $p \neq 2$, and hence, the kernel trick cannot be used. The soft margin concept was initially proposed by Cortes and Vapnik [8] and considers the linear penalty of the slack variables. In contrast, FMP_p adopts the quadratic penalty of slack variables as considered by Schölkopf and Smola [9].

As an illustration of the applicability of the proposed method, we consider the problem of feature selection [10]. In this sense, the algorithm performs the L_1 norm minimization in order to obtain a sparse solution for the normal vector and a separating hyperplane that reflects the maximization of the L_∞ margin [11]. The use of a large margin classifier with a built-in regularization technique that constraints the magnitudes of the components of the normal vector generating sparsity has been employed as an alternative to combinatorial search that demands large computational effort [12–14]. To that effect, IMA_p can be coupled with a greedy backward elimination method, such as Recursive Feature Elimination (RFE) algorithm [15], enabling the elimination of a large number of features at a time. Such approach, which combines RFE with an online classification algorithm has been proposed previously by Gentile [16], where the author couples $ALMA_p$ with RFE algorithm and adopts for p the value $\max\{2, \ln_f\}$, where f is the cardinality of the current set of features.

In a different context, we mention the development of kernel based classifiers in the dual space that adopt a Bayesian learning approach and obtain another form of sparseness, which is related to the number of support vectors in the final solution [17–19]. These methods have two main advantages: they reduce the computational effort in predicting new samples and increase the generalization performance [17]. This approach is often based on regularization through the minimization of some L_p norm in order to control the complexity of the solutions. However, the objective of this approach is different from what is considered in this paper. Here, we are concerned with the problem in primal variables and therefore the sparseness obtained in minimizing the L_1 norm is with respect to the normal vector of the solution, which can be useful for feature selection.

The remaining of this paper is structured as follows. Section 2 describes a flexible formulation for the binary classification problem with arbitrary norm and introduces some preliminary concepts that will be used throughout this work. Section 3 presents the development of the FMP_p algorithm which can be used to obtain a separating hyperplane given a fixed geometric p-margin. Next, Section 4 introduces the incremental strategy used to obtain an α-approximation of the maximal p-margin solution. In Section 5, we revised the special formulations of linear programming developed for L_1 and L_∞ norms. Section 6 contains the computational experiments and results. Finally, in Section 7, some final considerations and conclusions are presented.

2. The binary linear classification problem with p-Norm

Let $Z=\{z_i=(x_i,y_i): i\in\{1,\ldots,m\}\}$ be a training set composed of points $x_i\in\mathbb{R}^d$ and labels $y_i\in\{-1,+1\}$. In addition, let Z^+ and Z^- be defined as the sets $\{(x_i,y_i)\in Z: y_i=+1\}$ and $\{(x_i,y_i)\in Z: y_i=-1\}$, respectively. A binary linear classification problem consists of finding a hyperplane, which is given by its normal vector $w\in\mathbb{R}^d$ and a constant $b\in\mathbb{R}$, such that the points in Z^+ and Z^- lie separated in the two half spaces generated by it. That is, we look for (w,b) such that:

$$y_i(w\cdot x_i+b)\geq 0,\ \text{for all}\ (x_i,y_i)\in Z.$$

Clearly, this hyperplane may not exist for some training sets Z. When it exists, Z is usually called linearly separable. We suppose that Z is linearly separable throughout the paper, unless otherwise stated.

We say that Z accepts a *margin* $\gamma\geq 0$ when there is a hyperplane $\mathcal{H}:=\{x\in\mathbb{R}^d: w\cdot x+b=0\}$ such that:

$$y_i(w\cdot x_i+b)\geq\gamma,\ \text{for all}\ (x_i,y_i)\in Z.$$

In this case, we define two additional hyperplanes parallel to $\mathcal{H}$, given by $\mathcal{H}^+:=\{x\in\mathbb{R}^d: w\cdot x+(b-\gamma)=0\}$ and $\mathcal{H}^-:=\{x\in\mathbb{R}^d: w\cdot x+(b+\gamma)=0\}$. The distance between these two parallel hyperplanes under a p-norm is given by [20]:

$$\text{dist}_p(\mathcal{H}^-,\mathcal{H}^+)=\frac{(-(b-\gamma)+(b+\gamma))}{\|w\|_q}=\frac{2\gamma}{\|w\|_q},$$

where $\|\cdot\|_q$ is the conjugated norm, where p and q satisfy $1/p+1/q=1$. Let $\gamma_g^p:=\text{dist}_p(\mathcal{H}^-,\mathcal{H}^+)/2$, we call this γ_g^p the *geometric p-margin* between the hyperplanes $\mathcal{H}^+$ and $\mathcal{H}^-$. In this way, we say that Z accepts a *geometric p-margin* $\gamma_g^p\geq 0$ when there exists a hyperplane with (w,b) such that:

$$y_i(w\cdot x_i+b)\geq\gamma_g^p\|w\|_q,\ \text{for all}\ (x_i,y_i)\in Z.$$

3. Fixed p-margin perceptron – FMP_p

Given a fixed p-margin γ_f and a training set Z, which accepts γ_f as geometric p-margin, consider the problem of finding a separating hyperplane (w,b) such that:

$$y_i(w\cdot x_i+b)\geq\gamma_f\|w\|_q,\ \text{for all}(x_i,y_i)\in Z. \tag{1}$$

For that, let us define the following error function $J^q:\mathbb{R}^{d+1}\to\mathbb{R}$ which is given by:

$$J^q(w,b):=\sum_{(x_i,y_i)\in\mathcal{M}}\gamma_f\|w\|_q-y_i(w\cdot x_i+b),$$

where $\mathcal{M}$ is a subset of Z that violates Eq. (1) for the choice of data (w,b), that is:

$$\mathcal{M}:=\{(x_i,y_i)\in Z: y_i(w\cdot x_i+b)<\gamma_f\|w\|_q\}.$$

Using the online stochastic gradient approach, the minimization process begins with a initial value (w^0,b^0), usually $(0,0)$. At each iteration t of the algorithm, a single pair $z_i=(x_i,y_i)$ is chosen and verified against (w^t,b^t). If this pair is a mistake, that is, if $y_i(w^t\cdot x_i+b^t)<\gamma_f\|w^t\|_q$, then a new normal vector w^{t+1} and constant b^{t+1} are constructed using the gradient of J^q. In this way, taking the partial derivatives of J^q with respect to w_j, $j\in\{1,\ldots,d\}$,