Pattern Recognition 47 (2014) 2153–2164

Contents lists available at ScienceDirect

Pattern Recognition

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

Primal explicit max margin feature selection for nonlinear support vector machines

Aditya Tayal [a,*], Thomas F. Coleman [b], Yuying Li [a]

[a] Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1
[b] Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form
11 August 2013
Accepted 1 January 2014
Available online 15 January 2014

Keywords:
Feature selection
Nonlinear
Embedded
Support vector machine
Non-convex optimization
Trust-region method
Alternating optimization

ABSTRACT

Embedding feature selection in nonlinear support vector machines (SVMs) leads to a challenging non-convex minimization problem, which can be prone to suboptimal solutions. This paper develops an effective algorithm to directly solve the embedded feature selection primal problem. We use a trust-region method, which is better suited for non-convex optimization compared to line-search methods, and guarantees convergence to a minimizer. We devise an alternating optimization approach to tackle the problem efficiently, breaking it down into a convex subproblem, corresponding to standard SVM optimization, and a non-convex subproblem for feature selection. Importantly, we show that a straightforward alternating optimization approach can be susceptible to saddle point solutions. We propose a novel technique, which shares an explicit margin variable to overcome saddle point convergence and improve solution quality. Experiment results show our method outperforms the state-of-the-art embedded SVM feature selection method, as well as other leading filter and wrapper approaches.

1. Introduction

Feature selection has become a significant research focus in statistical machine learning and data mining communities. As increasingly more data is available, problems with hundreds and thousands of features have become common. Some examples include text processing of internet documents, gene micro-array analysis, combinatorial chemistry, economic forecasting and context based collaborative filtering. However, irrelevant and redundant features reduce the effectiveness of data mining and may detract from the quality and accuracy of the resulting model. The goal of feature selection is to identify the most relevant subset of input features for the learning task, improving generalization error and model interpretability.

In this paper, we focus on feature selection for nonlinear support vector machine (SVM) classification. SVM is based on the principle of maximum-margin separation, which achieves the goal of structural risk minimization by minimizing a generalization bound on model complexity and training error concurrently [1,2]. The model is obtained by solving a convex quadratic programming problem. Linear SVM models can be extended to the nonlinear ones by transforming the input features using a set of nonlinear basis functions. An important advantage of the SVM is that the transformation can be done implicitly using the "kernel trick", thereby allowing even infinite-dimensional feature expansions [3]. Empirically, SVMs have performed extremely well in diverse domains [e.g. see [4,5]].

Determining the optimal set of input features is in general NP-hard, requiring an exhaustive search of all possible subsets. Practical alternatives can be grouped into filter, wrapper, and embedded techniques [6]. In addition, there are a class of Bayesian approaches which tackle the problem by incorporating sparsity inducing priors [7–11].

Filter methods operate independent of the SVM classifier to score features according to how useful they are in predicting the output. Relief [12,13] is a popular multivariate nonlinear filter that has successfully been used as a preprocessing step for SVMs [14]. Wrapper methods, on the other hand, use the SVM classifier to guide the search in the space of all possible subsets. For instance the most common wrapper, recursive feature elimination, greedily removes the worst (or adds the best) feature according to the loss (or gain) of the SVM classifier at each iteration [15]. Finally, embedded approaches incorporate the feature selection criterion in the SVM objective itself. Embedded methods can offer significant advantages over filters and wrappers, since they tightly couple feature selection with SVM learning, simultaneously searching over the feature and model space.

For linear SVMs, several embedded feature selection methods have been proposed. The general idea is to incorporate sparse

* Corresponding author. Tel.: +1 647 606 3808; fax: +1 647 800 6713.
E-mail addresses: amtayal@uwaterloo.ca (A. Tayal), tfcoleman@uwaterloo.ca (T.F. Coleman), yuying@uwaterloo.ca (Y. Li).

0031-3203/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2014.01.003

regularization of the primal weight vector [16–21]. However, similar techniques cannot be readily applied to nonlinear SVM classifiers, since the weight vector is not explicitly formed. Sparse regularization of the dual variables (support vectors) lead to a reduction in the number of kernel functions needed to generate the nonlinear surface, but does not result in a reduction of input features [19]. Recently, supervised sparse dimension reduction techniques have also been applied under nonlinear manifolds with success [22].

Embedding feature selection in a nonlinear SVM requires optimizing over additional parameters in the kernel function. This can be viewed as an instance of Generalized Multiple Kernel Learning (GMKL) [23], which offers the state-of-the-art solution for embedded nonlinear feature selection. In general, the resulting problem is non-convex. The algorithm proposed by Varma and Babu [23] to solve GMKL is based on gradient descent, i.e. line-search along the negative gradient. Hence, it uses a first-order convex approximation at each iterate, which can fail to find a minimizer when the problem is non-convex. In contrast, trust-region algorithms are better suited for non-convex optimization. At each iterate they solve non-convex second-order approximations with guaranteed convergence to a minimizer.

This paper develops an effective algorithm to solve the non-convex optimization problem that results from embedding feature selection in nonlinear SVMs. Our contributions in this paper are as follows:

1. We invoke the Representor Theorem to formulate a primal embedded feature selection SVM problem and use a smoothed hinge loss function to obtain a simpler bound constrained problem. We solve the resulting non-convex problem using a generalized trust-region algorithm for bound constrained minimization.
2. To improve efficiency we propose a two-block alternating optimization scheme, in which we iteratively solve (a) the standard SVM problem and (b) a smaller non-convex feature selection problem. Importantly, we propose a novel alternate optimization method by sharing a single perspective variable. We establish mathematical conditions under which this perspective variable sharing the AO method avoids saddle points. For SVM feature selection, the perspective variable explicitly represents the margin. We provide computational evidence to illustrate that this helps avoid suboptimal local solutions. Moreover, by focusing on maximizing margin in the feature selection problem—a critical quantity for generalization error—we are able to further improve solution quality.
3. We compare our methods to GMKL and other leading nonlinear feature selectors, and show that our approach improves results.

The rest of the paper is organized as follows. Section 2 formulates the embedded feature selection problem. Section 3 describes the bound constrained trust-region approach to solve the problem in the full feature and model space. Section 4 develops the explicit margin alternating optimization approach. Section 5 compares our approach with other nonlinear feature selection methods on several datasets and we conclude with a discussion in Section 6.

2. Feature selection in nonlinear SVMs

We start by describing the embedded feature selection problem for nonlinear SVMs. We motivate and explain the formulation with respect to margin-based generalization bounds.

Consider a set of n training points, $\mathbf{x}_i \in \mathbb{R}^d$, and corresponding class labels, $y_i \in \{+1,-1\}$, $i=1,\ldots,n$. Each component of $\mathbf{x}_i$ is an input feature. In classical SVM, proposed by [1], a linear classifier $(\mathbf{w},b)$ is learned by maximizing the geometric margin, defined as $\gamma = \min_i y_i(\mathbf{w}^T\mathbf{x}_i+b)/\|\mathbf{w}\|$, where $\|\cdot\|$ denotes 2-norm. Since the decision hyperplane associated with $(\mathbf{w},b)$ does not change upon rescaling to $(\lambda\mathbf{w},\lambda b)$, for $\lambda \in \mathbb{R}^+$, the function output at the margin (functional margin) is fixed to 1; geometric margin is given by $\gamma = 1/\|\mathbf{w}\|$, and the norm of the weight vector is minimized. Thus in the standard setting, SVM results in the following convex quadratic programming problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i+b) \geq 1-\xi_i, \quad i=1,\ldots,n, \\ & \xi_i \geq 0, \quad i=1,\ldots,n. \end{aligned} \tag{1}$$

Here, ξ_i's are the margin violations, and C is a penalty controlling the trade-off between empirical error and (implicitly computed) geometric margin.

To obtain a non-linear decision function, the kernel trick [3] is used by defining a kernel function, $K(\mathbf{x},\mathbf{x}') \equiv \phi(\mathbf{x})^T\phi(\mathbf{x}')$, where $K:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ and $\phi:\mathbb{R}^d\to F$ is a non-linear map from input features to a (potentially infinite dimensional) derived feature space. A kernel function, satisfying Mercer's condition [24,25], directly computes the inner product of two vectors in a derived feature space, without the need to explicitly determine the feature mapping. Conventionally, the kernel is used in the dual of problem 1, where all occurrences of data appear inside an inner product. However, we can also formulate the primal problem in the derived feature space by expressing the weight vector as a linear combination of mapped data points, $\mathbf{w} = \sum_{i=1}^{n} y_i u_i \phi(\mathbf{x}_i)$, due to Representor Theorem [26]. We denote the coefficients as u_i, and not α_i as used in the standard SVM literature, in order to distinguish them from the typical Lagrange multiplier interpretation. Substituting this form in (1) leads to the following primal non-linear SVM problem:

$$\begin{aligned} \min_{\mathbf{u},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j u_i u_j K(\mathbf{x}_i,\mathbf{x}_j) + C\sum_{i=1}^{n}\xi_i, \\ \text{s.t.} \quad & y_i\left(\sum_{j=1}^{n} y_j u_j K(\mathbf{x}_i,\mathbf{x}_j)+b\right) \geq 1-\xi_i, \quad i=1,\ldots,n, \\ & \xi_i \geq 0, \quad i=1,\ldots,n. \end{aligned} \tag{2}$$

The geometric margin in the derived feature space is given by

$$\gamma = \frac{1}{\sqrt{\sum_{i,j=1}^{n} y_i y_j u_i u_j K(\mathbf{x}_i,\mathbf{x}_j)}}.$$

The dual of problem (2) reveals that the primal variable u_i is equivalent to the standard SVM dual Lagrange multiplier α_i, i.e. $u_i = \alpha_i$, when the kernel matrix is non-singular. If the kernel matrix is singular, then the coefficient expansion u_i is not unique (even though the decision function is) and solving (2) will produce one of the possible expansions, of which α_i is also a minimizer.

The maximum margin classifier is motivated by theoretical bounds on the generalization error. Specifically, Ref. [2] shows that generalization error for n points is bounded by

$$err \leq \frac{c}{n}\left[\left(\frac{R^2}{\gamma^2}+\|\boldsymbol{\xi}\|^2\right)\log^2 n + \log\frac{1}{\delta}\right], \tag{3}$$

for some constant c with probability $1-\delta$, where γ is the geometric margin of the classifier. The key expression, on which generalization depends, is $R^2/\gamma^2+\|\boldsymbol{\xi}\|^2$, where ξ is the margin slack vector (normalized by γ), and R is the radius of the ball that encloses the set of points in the derived feature space, $\{\phi(\mathbf{x}_i)\}_{i=1}^{n}$. For a fixed dataset and kernel choice, R is constant, and thus maximizing the margin while reducing margin violations minimizes the upper bound in (3). Although the generalization bound suggests using a 2-norm penalty on margin violations, a 1-norm penalty is preferred for classification tasks, since it is a better approximation to a step penalty [2].