Pattern Recognition Letters 95 (2017) 51–57

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Max-Margin feature selection

Yamuna Prasad[a,*], Dinesh Khandelwal[a], K.K. Biswas[a]

Department of Computer Science & Engineering, Indian Institute of Technology Delhi, New Delhi, 110016, India

ARTICLE INFO

Article history:
Received 14 June 2016
Available online 20 April 2017

Keywords:
Feature selection
One class SVM
Max-Margin

ABSTRACT

Many machine learning applications such as in vision, biology and social networking deal with data in high dimensions. Feature selection is typically employed to select a subset of features which improves generalization accuracy as well as reduces the computational cost of learning the model. One of the criteria used for feature selection is to jointly minimize the redundancy and maximize the relevance of the selected features. In this paper, we formulate the task of feature selection as a one class SVM problem in a space where features correspond to the data points and instances correspond to the dimensions. The goal is to look for a representative subset of the features (support vectors) which describes the boundary for the region where the set of the features (data points) exists. This leads to a joint optimization of relevance and redundancy in a principled max-margin framework. Additionally, our formulation enables us to leverage existing techniques for optimizing the SVM objective resulting in highly computationally efficient solutions for the task of feature selection. Specifically, we employ the dual coordinate descent algorithm (Hsieh et al., 2008), originally proposed for SVMs, for our formulation. We use a sparse representation to deal with data in very high dimensions. Experiments on seven publicly available benchmark datasets from a variety of domains show that our approach results in orders of magnitude faster solutions even while retaining the same level of accuracy compared to the state of the art feature selection techniques.

© 2017 Published by Elsevier B.V.

1. Introduction

Many machine learning problems in vision, biology, social networking and several other domains need to deal with very high dimensional data. Many of these attributes may not be relevant for the final prediction task and act as noise during the learning process. A number of feature selection methods have already been proposed in the literature to deal with this problem. These can be broadly categorized into filter based, wrapper based and embedded methods.

In filter based methods, features (or subset of the features) are ranked based on their statistical importance and are oblivious to the classifier being used [7,11]. Wrapper based methods select subset of features heuristically and classification accuracy is used to estimate the goodness of the selected subset [10]. These methods typically result in good accuracy while incur high computational cost because of the need to train the classifier multiple number of times. In the embedded methods, feature selection criteria is directly incorporated in the objective function of the classifier [16,17]. Many filter and wrapper based methods fail on very high dimensional datasets due to their high time and memory requirements, and also because of inapplicability on sparse datasets [7,17].

In the literature, various max-margin formulation had been developed for many applications [1,6,9,19]. Recently, we have proposed a hard margin primal formulation for feature selection using quadratic program (QP) slover [12]. This approach jointly minimizes redundancy and maximizes relevance in a max-margin framework. We have formulated the task of feature selection as a one class SVM problem [14] in the dual space where features correspond to the data points and instances correspond to the dimensions. The goal is to search for a representative subset of the features (support vectors) which describes the boundary for the region in which the set of the features (data points) lies. This is equivalent to searching for a hyperplane which maximally separates the data points from the origin [14].

In this paper, we have extended the hard-margin formulation to develop a general soft-margin framework for feature selection. We have also modified the primal and dual formulations. We present the dual objective as unconstrained optimization problem. We employ the Dual Coordinate Descent (DCD) algorithm [8] for solving our formulation. The DCD algorithm simultaneously uses the information in the primal as well as in the dual to come up with a very fast solver for the SVM objective. In order to apply DCD approach, our formulation has been appropriately modified by including an

* Corresponding author.
E-mail address: yprasad@cse.iitd.ac.in (Y. Prasad).

http://dx.doi.org/10.1016/j.patrec.2017.04.011
0167-8655/© 2017 Published by Elsevier B.V.

additional term in the dual objective, which can be seen as a regularizer on the feature weights. The strength of this regularizer can be tuned to control the sparsity of the selected features weights. We adapt the liblinear implementation [3] for our proposed framework so that our approach is scalable to data in very high dimensions. We also show that the Quadratic Programming Feature Selection (QPFS) [13] falls out as a special case of our formulation in the dual space when using a hard margin.

Experiments on seven publicly available datasets from a vision, biology and Natural Language Processing (NLP) domains show that our approach results in orders of magnitude faster solutions compared to the state of the art techniques while retaining the same level of accuracy.

The rest of the paper is organized as follows. We describe our proposed max-margin formulation for feature selection (MMFS) including the dual coordinate descent approach in Section 2. We present our experimental evaluation in Section 4. We conclude our work in Section 5.

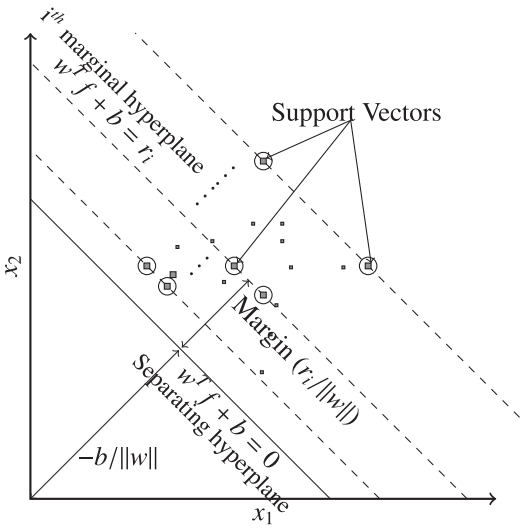

Fig. 1. Feature representation in sample space. The diagram is conceptual only.

2. Proposed Max-Margin framework

The key objective in feature selection is to select a subset of features which are highly relevant (that is high predictive accuracy) and non-redundant (that is uncorrelated). Relevance is captured either using an explicit metric (such as the correlation between a feature and the target variable) or implicitly using the classifier accuracy on the subset of features being selected. Redundancy is captured using metrics such as correlation coefficient or mutual information. Most of the existing feature selection methods rely on a pairwise notion of similarity to capture redundancy [11,13,18].

We try to answer the question "Is there a principled approach to jointly capturing the relevance as well redundancy amongst the features?". To do this, we flip around the problem and examine the space where features themselves become the first class objects. In particular, we analyze the space where "features" represent the data points and "instances" represent the dimensions. Which boundary could describe well the set of features lying in this space? Locating the desired boundary is similar to one class SVM formulation [14]. This equivalently can be formulated as the problem of searching for a hyperplane which maximally separates the features (data points) from the origin in the appropriate kernel space over the features. In order to incorporate feature relevance, we construct a set of parallel marginal hyperplanes, one hyperplane for each feature. The margin of each separating hyperplane captures the relevance of the corresponding feature. Greater the relevance, higher the margin required (a greater margin increases the chances of a feature being a support vector). Redundancy among the features is captured implicitly in our framework. The support vectors which lie on respective margin boundaries constitute the desired subset of features to be selected. This leads to a principled max-margin framework for feature selection. The proposed formulation for MMFS is presented hereafter.

2.1. Formulation

Let X represent the data matrix where each row vector ${x_i}^T$ ($i \in 1 \dots M$) denotes an instance and each column vector f_j ($j \in 1 \dots N$) denotes a feature vector. We will use ϕ to denote a feature map such that the dot product between the data points can be computed via a kernel $k(u, v) = \phi(u)^T \phi(v)$, which can be interpreted as the similarity of u and v. We will use Y to denote the vector of class labels y_i's ($i \in 1 \dots M$). Based on the above notations, we present the following hard margin for feature selection in the primal:

$$\begin{aligned} &\min_{w,b} \frac{1}{2} w^T w + b \\ &\text{subject to} \quad w^T \phi(f_i) + b \geq r_i, \forall i = 1, \dots, N; \end{aligned} \tag{1}$$

where, w represents a vector normal to the separating hyperplane(s)[1] and b represents the bias term. r_i captures the relevance for the ith feature. The equation of the separating hyperplane is given by $w^T \phi(f_i) + b = 0$ with the distance of the hyperplane from the origin being $-b$. In the hard margin formulation, a single outlier can determine the boundary which makes it overly sensitive to noise in the features. In order to handle the noise in the features, we propose following soft margin formulation.

$$\begin{aligned} &\min_{w,b} \quad \frac{1}{2} w^T w + b + C \sum_{i=1}^{N} \xi_i \\ &\text{subject to} \quad w^T \phi(f_i) + b \geq r_i - \xi_i, \;\; \xi_i \geq 0, \quad \forall i = 1, \dots, N; \end{aligned} \tag{2}$$

where, ξ_i's represent slack variables and C represents trade-off between the margin width and sum of slack variables. Note that in this formulation the objective function is similar to the one class SVM [14]. However, the constraints are very much different as our formulation includes the relevance of the features (r). The choice of ϕ determines the kind of similarity (correlation) to be captured among the features. The set of support vectors obtained after optimizing this problem i.e. $\{f_i \mid w^T \phi(f_i) + b = r_i\}$ and the margin violators $\{f_i | \xi_i > 0\}$ constitute the set of features to be selected. In the dual space, this translates to those features being selected for which $0 < \alpha_i \leq C$ where α_i is the Lagrange multiplier for f_i. We will refer to our approach as Max-Margin Feature Selection (MMFS). Note that when dealing with hard margin (no noise) case and the term involving C disappears (since this enforces $\xi_i = 0, \forall i$).

Fig. 1 illustrates the intuition behind our proposed framework in the linear dot product space (with hard margin). In the figure, $w^T f + b = 0$ represents the separating hyperplane. The distance of this hyperplane from the origin is given by $-b/||w||$. The first term in the objective of Eq. (2) tries to minimize $w^T w$ i.e. maximize $1/||w||$. The second term in the objective tries to minimize b i.e. maximize $-b$. Hence, the overall objective tries to push the plane away from the origin. The ith dashed plane represents the margin

[1] All the separating hyperplanes are parallel to each other in our framework.

Download English Version:

https://daneshyari.com/en/article/4970020

Download Persian Version:

https://daneshyari.com/article/4970020

Daneshyari.com