Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Weighted doubly regularized support vector machine and its application to microarray classification with noise ☆

Juntao Li [a,*], Yadi Wang [a], Yimin Cao [a], Cunshuan Xu [b]

[a] School of Mathematics and Information Science, Henan Normal University, Xinxiang 453007, PR China
[b] State Key Laboratory Cultivation Base for Cell Differentiation Regulation, Henan Normal University, Xinxiang 453007, PR China

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form
29 April 2015
Accepted 2 August 2015
Communicated by M. Chetty

Keywords:
Support vector machine
Double weights
Gene selection
Microarray classification
Solution path algorithm

ABSTRACT

A weighted doubly regularized support vector machine was proposed and its solution path algorithm was developed. By using both the distance information between classes and within each class, a double-weighted mechanism was presented, based on which the weighted doubly regularized support vector machine was proposed. A staircase function between two model parameters along the solution path direction was proposed, the multiple parameter selection problem was transformed into single parameter problem and the corresponding solution path algorithm was developed. The proposed support vector machine can adaptively identify the important genes in groups, thus encouraging an adaptive grouping effect. The experiment results on leukaemia, colon cancer, lung cancer data sets demonstrated that the proposed method can effectively select genes and reduce the influence of noise.

1. Introduction

Microarray-based-classification of patient samples has attracted much attention in cancer diagnosis and treatment [1,2]. It is well known that the number of samples generated by the microarray technology is less, and each sample contains thousands of genes, i.e., the number of genes is much larger than the number of samples. Hence, selecting the genes related to the classification is an important challenge in microarray classification, besides achieving good accuracy.

Support vector machine [2–5] and its extensions have been widely applied to microarray classification and feature selection. In order to automatically select features in the process of the classification, lasso [6], 1-norm support vector machine [7] and sparse logistic regression with Bayesian regularization [8] have been developed. In order to automatically select genes in the multi-classification, Student and Fujarewicz proposed a stable feature selection method for multiclass microarray data [9], You et al. proposed the feature selection method for multi-class classification [10]. In order to select genes in groups, elastic net [11], doubly regularized support vector machine [12] and hybrid huberized support vector machine [13] have been developed. In order to encourage an adaptive grouping effect, adaptive huberized support vector machine [14], improved elastic net [15], and partly adaptive elastic net [16] have been developed. Especially, Xu et al. proposed $L_{1/2}$ regularization methods [17,18] by using $L_{1/2}$ penalty.

Due to the low complexity in parameter selection, the solution path algorithm has attracted much attention in machine learning [7,12,13,19–23]. Efron et al. seeked the piecewise linear regularization parameter solution path of lasso by using a linear connection, which was the least angle regression [19]. Motivated by this idea, the solution algorithms for the learning machines emerged [7,12,13,20–23]. Hastie et al. developed the solution path algorithm for the standard support vector machine [20]. Zhu et al. proposed the regularization solution path algorithm for 1-norm support vector machine [7]. Wang et al. developed the solution algorithms for doubly regularized support vector machine [12] and the hybrid huberized support vector machines [13]. In order to further improve the solving speed, Friedman et al. developed a fast regularization solution path algorithm (glmnet) for generalized linear models via coordinate descent [21]. Yuan et al. improved this glmnet algorithm and applied it to solve L_1-regularized logistic regression [22]. Following the same idea, Yang et al. developed an efficient algorithm for computing the hybrid huberized support vector machine and its generalizations [23].

The doubly regularized methods, such as elastic net, doubly support vector machine and huberized support vector machine,

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail addresses: juntaolimail@126.com (J. Li), wangyadime@126.com (Y. Wang), cao_yimin@126.com (Y. Cao), cellkeylab@126.com (C. Xu).

http://dx.doi.org/10.1016/j.neucom.2015.08.002
0925-2312/

can automatically select genes within each group at the same time of classification. Thus they can be successfully applied to microarray classification, handwritten digit recognition and so on. However, there are two model parameters in these methods. To develop the solution algorithm, one parameter must be fixed in advance. Hence, the solution path algorithm is piecewise linear with respect to single (another) parameter. Moreover, classification accuracy can be largely affected by the noise. For the first problem, Li et al. transformed the multiple parameter selection problem to single parameter problem along the solution path direction [14]. For the second problem, Zhang [24] proposed an evaluation for amplifier performance based on feature double weighted support vector machine which can improve the parameter test precision. Lin and Wang proposed a fuzzy weighted mechanism for reducing the influence of noise points [25]. Although the fuzzy support vector machine provided a good idea to deal with noise, this method only considered information within each class and did not use the information between classes. In order to deal with these problems, this paper is devoted to developing the double weighted mechanism by using both the distance information between classes and within each class. In addition, a regularization model with partly elastic net penalty is proposed and its corresponding regularization solution path algorithm is developed.

This paper is organized as follows. Section 2 presents the preliminary of the problem. The weighted doubly regularized support vector machine is presented and its corresponding properties and solution path algorithm are studied in Section 3. Experimental results obtained on three data sets are provided in Section 4. Finally, we conclude the paper in Section 5.

2. Problem formulation and preliminaries

Given a training set $\{(x_i, y_i)\}_{i=1}^n$, where the input $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, x_{ip} is the expression levels of p genes of the i-th observation, and the output $y_i \in \{+1, -1\}$ which represents the class label corresponding to input x_i. For the microarray data, n and p represent the number of experiments and the number of genes, respectively. Restricted to the high experiment cost, only a few (less than one hundred) samples can be obtained with thousands of genes in one sample. Similar to [14–16], the notations are defined as follows: $y = (y_1, y_2, \ldots, y_n)^T$ is the response vector and $X = (x_{(1)}, x_{(2)}, \ldots, x_{(p)})$ is the model matrix, where $x_{(j)} = (x_{1j}, x_{2j}, \ldots, x_{nj})^T, j = 1, 2, \ldots, p$, are the predictors. Suppose that the response vector y is centered and the columns of X are standardized, i.e.,

$$\sum_{i=1}^{n} y_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1. \tag{1}$$

In order to select genes in groups in the process of classification, Zou and Hastie proposed the elastic net [11] by combining the squared error loss and the elastic net penalty. Applying the elastic net penalty to hinge loss function, Wang developed the doubly regularized support vector machine [12]:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} [1 - y_i(\beta_0 + x_i^T\beta)]_+ + \frac{\lambda_2}{2}\|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \tag{2}$$

where $[1 - y_i(\beta_0 + x_i^T\beta)]_+$ and $\frac{\lambda_2}{2}\|\beta\|_2^2 + \lambda_1 \|\beta\|_1$ represent hinge loss and the elastic net penalty, respectively. Both λ_1 and λ_2 are nonnegative regularization parameters. Given a new input x, its class label is predicted by $\text{sign}(\beta_0 + x^T\beta)$.

As a new regularized method, the doubly regularized support vector machine has shown great success in many situations due to the merits of the elastic net penalty. However, there are two model parameters in this method. To develop the solution algorithm, one parameter must be fixed in advance. Hence, the solution path algorithm is piecewise linear with respect to single (another) parameter. On the other hand, the doubly regularized support vector machine does not consider the influence of the noise which often occurs in many cases.

This paper is devoted to the aforementioned problems by developing new statistical learning model and algorithm. The following statistical terminologies are needed. True positive (TP) represents the number of predicted positives that are correct. False negative (FN) represents the number of predicted negatives that are incorrect. False positive (FP) represents the number of predicted positives that are incorrect. True negative (TN) represents the number of predicted negatives that are correct. According to [27], we explain the following statistical indicators (TPR, TNR, R, P and F).

The accuracy at positive class level is regard as true positive rate (TPR) or sensitivity which measures the proportion of actual positives which are correctly identified as such, i.e.,

$$TPR = \frac{TP}{TP + FN}. \tag{3}$$

The accuracy at negative class level is regarded as true negative rate (TNR) or specificity which measures the proportion of negatives which are correctly identified as such, i.e.,

$$TNR = \frac{TN}{TN + FP}. \tag{4}$$

Recall is the fraction of relevant instances that are retrieved, i.e.,

$$R = \frac{TP}{TP + FN}. \tag{5}$$

Precision is the fraction of retrieved instances that are relevant, i.e.,

$$P = \frac{TP}{TP + FP}. \tag{6}$$

The traditional F-Measure or balanced F-Score is the harmonic mean of precision and recall, i.e.,

$$F = \frac{PR}{P + R}. \tag{7}$$

There are several reasons that the F-Score can be criticized in particular circumstances due to its bias as an evaluation metric. This is also known as the F_1 measure, because recall and precision are evenly weighted.

3. Main results

3.1. Double-weighted mechanism

In order to reduce the influence of noise, Lin and Wang [25] proposed fuzzy weighted mechanism and developed the fuzzy support vector machines. Although the fuzzy support vector machines provided a good idea to deal with noise, this method only considered information within each class and did not use the information between classes. In order to effectively reduce the influence of the noise, in the following we will be devoted to constructing the double weighted mechanism.

Similar to [25], we construct the first weights by using the information within class.

Let m_+ be the means of class $+1$, i.e.,

$$m_+ = \frac{1}{l_+}\sum_{y_i = +1} x_i, \tag{8}$$

Please cite this article as: J. Li, et al., Weighted doubly regularized support vector machine and its application to microarray classification with noise, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.08.002