SALSPARSE DISTANCE WEIGHTED DISCRIMINATION

Boxiang Wang & Hui Zou ∗
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA
{wang3660, zouxx019}@umn.edu

ABSTRACT

Distance weighted discrimination (DWD) was originally proposed to handle the data piling issue in the support vector machine. In this paper, we consider the sparse penalized DWD for high-dimensional classification. The state-of-the-art algorithm for solving the standard DWD is based on second-order cone programming, however such an algorithm does not work well for the sparse penalized DWD with high-dimensional data. In order to overcome the challenging computation difficulty, we develop a very efficient algorithm to compute the solution path of the sparse DWD at a given fine grid of regularization parameters. We implement the algorithm in a publicly available R package sdwd. We conduct extensive numerical experiments to demonstrate the computational efficiency and classification performance of our method.

Key words: High-dimensional classification, SVM, DWD.

1 INTRODUCTION

The support vector machine (SVM) (Vapnik, 1995) is a widely used modern classification method. In the standard binary classification problem, training dataset consists of \( n \) pairs, \( \{(x_i, y_i)\}_{i=1}^n \), where \( x_i \in \mathbb{R}^p \) and \( y_i \in \{-1, 1\} \). The linear SVM seeks a hyperplane \( \{x : \beta_0 + x^T \beta = 0\} \) which maximizes the smallest margin of all data points:

\[
\arg \max_{\beta_0, \beta} \min_i d_i,
\]

subject to

\[
d_i = y_i (\beta_0 + x_i^T \beta) + \eta_i \geq 0, \quad \eta_i \geq 0, \quad \forall i, \sum \eta_i \leq c, \quad ||\beta||^2 = 1,
\]

where \( d_i \) is defined as the margin of the \( i \)-th data point, \( \eta_i \)'s are slack variables introduced to ensure all margins non-negative, and \( c > 0 \) is a tuning parameter controlling the overlap. By using a kernel trick, the SVM can also produce nonlinear decision boundaries by fitting an optimal separating hyperplane in the extended kernel feature space.

Marron et al. (2007) noticed that when the SVM is applied on some data with \( n < p \), many data points lie on two hyperplanes parallel to the decision boundary. Marron et al. (2007) referred to this phenomenon as data piling and claimed that the data piling can “affect the generalization performance of SVM”. To overcome this issue, Marron et al. (2007) proposed a new method called the distance weighted discrimination (DWD), which finds a separating hyperplane minimizing the sum of the inverse margins of all data points,

\[
\arg \min_{\beta_0, \beta} \sum 1/d_i,
\]

subject to

\[
d_i = y_i (\beta_0 + x_i^T \beta) + \eta_i \geq 0, \quad \eta_i \geq 0, \quad \forall i, \sum \eta_i \leq c, \quad ||\beta||^2 = 1.
\]

Marron et al. (2007) asserted the DWD can avoid the data piling and thereby improve the generalizability. As for the computation of the DWD, Marron et al. (2007) observed that the DWD is an application of the second-order cone programming. The algorithm has been implemented in a Matlab implementation (Marron, 2013) and an R package DWD (Huang et al., 2012).

∗Boxiang Wang is a PhD candidate in Statistics at University of Minnesota. Hui Zou is a Professor of Statistics at University of Minnesota.
In this paper we focus on classification with high-dimensional data where the number of covariates is much larger than the sample size. The standard SVM and DWD are not suitable tools for high-dimensional classification for two reasons. First, based on the scientific hypothesis that only a few important variables affect the outcome, a good classifier for high-dimensional classification should have the ability to select important variables and discard irrelevant ones. However, the standard SVM and DWD use all variables and do not conduct variable selection. Second, because these two classifiers use all variables, they may have very poor classification performance. Owing to these two considerations, sparse classifiers are generally preferred for high-dimensional classification. In the literature, some penalties have been applied to the SVM to produce sparse SVMs such as the $\ell_1$ SVM (Bradley and Mangasarian, 1998; Zhu et al., 2004), the SCAD SVM (Zhang et al., 2006), and the elastic-net penalized SVM (Wang et al., 2006).

In this work we consider sparse penalized DWD for high dimensional classification. Compared to the standard DWD, the sparse DWD is computationally more challenging and requires a different computing algorithm. To this end, we derive an efficient algorithm to solve the sparse DWD by combining majorization-minimization principle and coordinate-descent. We have implemented the algorithm in an R package `sdwd`. To give a quick demonstration here, we use the prostate cancer data (Singh et al., 2002) as an example. The left panel of Figure 1 depicts the solution paths of the elastic-net DWD, and `sdwd` only took 0.453 second to compute the whole solution path. We observed that the timing of the sparse SVM was about 290 times larger than that of the sparse DWD.

### 2 Sparse DWD

In this section we present several sparse penalized DWDs. We first propose an $\ell_1$ DWD:

\[
\left( \hat{\beta}_0(\text{lasso}), \hat{\beta}(\text{lasso}) \right) = \arg \min_{\beta_0, \beta} \frac{1}{n} \sum_{i=1}^{n} V \left( y_i (\beta_0 + x_i^T \beta) \right) + \lambda_1 \| \beta \|_1. \tag{2.1}
\]

where the loss function is given by

\[
V(u) = \begin{cases} 
1 - u, & \text{if } u \leq 1/2, \\
1/(4u), & \text{if } u > 1/2.
\end{cases}
\]

Similar to the $\ell_1$ SVM, we replace the $\ell_2$ norm penalty with the $\ell_1$ norm penalty to achieve sparsity in the DWD classifier. The lasso penalized DWD classification rule is $\text{Sign}(\hat{\beta}_0(\text{lasso}) + x^T \hat{\beta}(\text{lasso}))$.

Besides the $\ell_1$ norm penalty, we also consider the elastic-net penalty (Zou and Hastie, 2005). It is now well-known that the elastic-net often outperforms the lasso ($\ell_1$ norm penalty) in prediction. Wang et al. (2006) studied the elastic-net penalized SVM (DrSVM) and showed that the DrSVM performs better than the $\ell_1$ norm SVM. Similarly, we propose the elastic-net penalized DWD:

\[
\left( \hat{\beta}_0(\text{enet}), \hat{\beta}(\text{enet}) \right) = \arg \min_{\beta_0, \beta} \frac{1}{n} \sum_{i=1}^{n} V \left( y_i (\beta_0 + x_i^T \beta) \right) + \sum_{j=1}^{p} \left( \lambda_1 |\beta_j| + \frac{\lambda_2}{2} \beta_j^2 \right). \tag{2.2}
\]

The elastic-net penalized DWD classification rule is $\text{Sign}(\hat{\beta}_0(\text{enet}) + x^T \hat{\beta}(\text{enet}))$. In our paper, we also present the adaptive elastic-net DWD, which produces estimators with the oracle properties.
Table 1: The mean mis-classification percentage and timings (in seconds) for four benchmark datasets. All the timings include the five-folder cross validation. For each data, the methods with the best prediction accuracy are marked by black boxes.

<table>
<thead>
<tr>
<th></th>
<th>Arcene n = 100, p = 10000</th>
<th>Breast n = 42, p = 2283</th>
<th>LSVT n = 126, p = 309</th>
<th>Prostate n = 102, p = 6031</th>
</tr>
</thead>
<tbody>
<tr>
<td>error</td>
<td>time</td>
<td>error</td>
<td>time</td>
<td>error</td>
</tr>
<tr>
<td>enet DWD</td>
<td>34.43</td>
<td>123.41</td>
<td>26.50</td>
<td>58.40</td>
</tr>
<tr>
<td>(0.56)</td>
<td>(5.16)</td>
<td>(1.00)</td>
<td>(1.90)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>enet DWD</td>
<td>34.60</td>
<td>200.19</td>
<td>26.86</td>
<td>116.12</td>
</tr>
<tr>
<td>(0.57)</td>
<td>(9.24)</td>
<td>(1.00)</td>
<td>(3.78)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>enet logistic</td>
<td>34.16</td>
<td>211.18</td>
<td>24.67</td>
<td>145.35</td>
</tr>
<tr>
<td>(0.58)</td>
<td>(3.40)</td>
<td>(1.00)</td>
<td>(0.74)</td>
<td>(0.37)</td>
</tr>
<tr>
<td>enet logistic</td>
<td><strong>34.15</strong></td>
<td>393.03</td>
<td>25.12</td>
<td>290.31</td>
</tr>
<tr>
<td>(0.57)</td>
<td>(6.52)</td>
<td>(0.87)</td>
<td>(1.47)</td>
<td>(0.37)</td>
</tr>
<tr>
<td>enet SVM</td>
<td>35.10</td>
<td>7410.09</td>
<td><strong>23.95</strong></td>
<td>567.43</td>
</tr>
<tr>
<td>(0.67)</td>
<td>(1465.68)</td>
<td>(1.00)</td>
<td>(15.19)</td>
<td>(0.37)</td>
</tr>
</tbody>
</table>

3 Computation

In this section, we propose an intuitive but efficient algorithm for computing the solution paths of the sparse DWD. Our algorithm uses the generalized coordinate descent (GCD) proposed by [Yang and Zou (2013)]. The same algorithm solves all $\ell_1$, elastic-net, and adaptive elastic-net DWDs.

Without loss of generality, we assume that the variables $x_j$ are standardized. We fix $\lambda_1$ and $\lambda_2$ and let $u_i = y_i(\hat{\beta}_0 + x_i^T \hat{\beta})$. We focus on $\beta_j$’s first. For each $\beta_j$, we define the coordinate-wise update:

$$F(\beta_j | \tilde{\beta}_0) = \frac{1}{n} \sum_{i=1}^{n} V(u_i + y_i x_{ij} (\beta_j - \tilde{\beta}_j)) + p \lambda_1, \lambda_2 (\beta_j). \quad (3.1)$$

Then the standard coordinate descent algorithm suggests cyclically updating

$$\tilde{\beta}_j = \arg \min_{\beta_j} F(\beta_j | \tilde{\beta}_0, \hat{\beta}) \quad (3.2)$$

for each $j = 1, \ldots, p$. However, (3.2) does not have a closed-form solution. The GCD algorithm solves this issue by adopting the MM principle [De Leeuw and Heiser (1977) Lange et al. (2000) Hunter and Lange 2004]. We approximate the $F$ function by a quadratic function

$$Q(\beta_j | \tilde{\beta}_0, \hat{\beta}) = \frac{\sum_{i=1}^{n} V(u_i)}{n} + \frac{\sum_{i=1}^{n} V'(u_i) y_i x_{ij} (\beta_j - \tilde{\beta}_j)}{n} + 2(\beta_j - \tilde{\beta}_j)^2 + p \lambda_1, \lambda_2 (\beta_j). \quad (3.3)$$

Define $S(z, r) = \text{sign}(z)(|z| - r)_+$ [Donoho and Johnston (1994)]. We then update $\tilde{\beta}_j$ by $\tilde{\beta}_j^{\text{new}}$, the closed-form minimizer of (3.3).

We prove the algorithm enjoys the strict descent property and guarantees convergence to the correct solution satisfying the KKT condition. We have implemented the algorithm in an R package $\text{sdwd}$, where we exploit the warm-start, the strong rule, and the active set trick to accelerate the algorithm.

4 Real Data Examples

In this section we analyzed four benchmark data [Lichman (2013)]. We compared timings and prediction accuracy of elastic-net DWD, adaptive elastic-net DWD, elastic-net logistic regression, adaptive elastic-net logistic regression, and elastic-net SVM. Table summarizes the results. For the sparse DWD, we get the same message as Marron et al. (2007) concluded for the standard DWD: “it very often is competitive with the best of the others and sometimes is better.” We also notice that the computation of the sparse DWD is the fastest in almost all cases.
REFERENCES


