SPARSE 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, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. The linear SVM seeks a hyperplane $\{\mathbf{x} : \beta_0 + \mathbf{x}^T\boldsymbol{\beta} = 0\}$ which maximizes the smallest margin of all data points:

$$\begin{aligned}&\arg\max_{\beta_0,\boldsymbol{\beta}} \ \min_i d_i,\\ &\text{subject to } \ d_i = y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) + \eta_i \geq 0, \ \eta_i \geq 0, \ \forall i, \ \sum \eta_i \leq c, \ ||\boldsymbol{\beta}||_2^2 = 1,\end{aligned}$$

where d_i is defined as the *margin* of the ith data point, η_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 pilling* and claimed that the data pilling 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,

$$\begin{aligned}&\arg\min_{\beta_0,\boldsymbol{\beta}} \ \sum 1/d_i,\\ &\text{subject to } \ d_i = y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) + \eta_i \geq 0, \ \eta_i \geq 0, \ \forall i, \ \sum \eta_i \leq c, \ ||\boldsymbol{\beta}||_2^2 = 1.\end{aligned}$$

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.

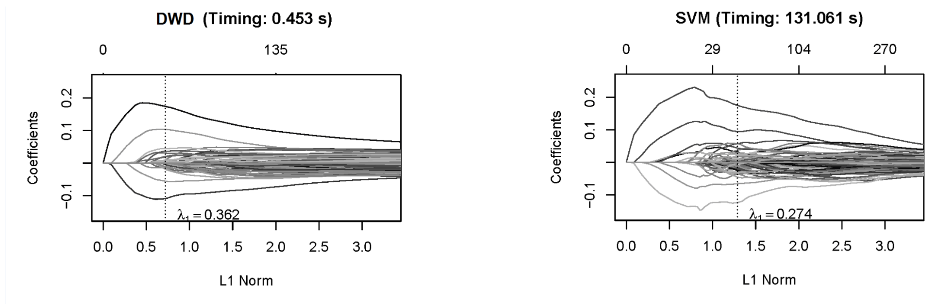

Figure 1: The solution paths for the prostate data ($n = 102$, $p = 6033$) using the elastic-net DWD and the elastic-net SVM. In every method, λ_2 is fixed to be 1. The dashed vertical lines indicate the λ_1 selected by the five-folder ross validation. Both timings are averaged over 10 runs.

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 ℓ_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 ℓ_1 DWD:

$$\left(\hat{\beta}_0(\text{lasso}), \hat{\boldsymbol{\beta}}(\text{lasso})\right) = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} V\left(y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})\right) + \lambda_1 ||\boldsymbol{\beta}||_1. \tag{2.1}$$

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 ℓ_1 SVM, we replace the ℓ_2 norm penalty with the ℓ_1 norm penalty to achieve sparsity in the DWD classifier. The lasso penalized DWD classification rule is Sign$(\hat{\beta}_0(\text{lasso}) + \mathbf{x}^T \hat{\boldsymbol{\beta}}(\text{lasso}))$.

Besides the ℓ_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 (ℓ_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 ℓ_1 norm SVM. Similarly, we propose the elastic-net penalized DWD:

$$\left(\hat{\beta}_0(\text{enet}), \hat{\boldsymbol{\beta}}(\text{enet})\right) = \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} V(y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})) + \sum_{j=1}^{p} \left(\lambda_1 |\beta_j| + \frac{\lambda_2}{2} \beta_j^2\right). \tag{2.2}$$

The elastic-net penalized DWD classification rule is Sign$(\hat{\beta}_0(\text{enet}) + \mathbf{x}^T \hat{\boldsymbol{\beta}}(\text{enet}))$. 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.

	Arcene $n=100, p=10000$		Breast $n=42, p=22283$		LSVT $n=126, p=309$		Prostate $n=102, p=6033$	
	error	time	error	time	error	time	error	time
enet DWD	34.43	123.41	26.50	58.40	16.01	8.28	**10.22**	28.18
	(0.56)	(5.16)	(1.00)	(1.90)	(0.34)	(0.23)	(0.30)	(0.95)
aenet DWD	34.60	200.19	26.86	116.12	**15.92**	13.72	10.26	39.25
	(0.57)	(9.24)	(1.00)	(3.78)	(0.34)	(0.29)	(0.26)	(1.24)
enet logistic	34.16	211.18	24.67	145.35	16.96	10.73	10.65	102.19
	(0.58)	(3.40)	(1.00)	(0.74)	(0.37)	(0.18)	(0.29)	(1.56)
aenet logistic	**34.15**	393.03	25.12	290.31	16.93	17.02	10.75	189.44
	(0.57)	(6.52)	(0.87)	(1.47)	(0.37)	(0.29)	(0.29)	(2.84)
enet SVM	35.10	7410.09	**23.95**	567.43	16.27	63.10	10.56	2508.94
	(0.67)	(1465.68)	(1.00)	(15.19)	(0.37)	(0.77)	(0.36)	(0.77)

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 ℓ_1, elastic-net, and adaptive elastic-net DWDs.

Without loss of generality, we assume that the variables $\mathbf{x}_j$ are standardized. We fix λ_1 and λ_2 and let $u_i = y_i(\tilde{\beta}_0 + \mathbf{x}_i^T \tilde{\boldsymbol{\beta}})$. We focus on β_j's first. For each β_j, we define the coordinate-wise update:

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

Then the standard coordinate descent algorithm suggests cyclically updating

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

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

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

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), $\tilde{\beta}_j^{\text{new}} = S\left(M\tilde{\beta}_j - \frac{1}{n}\sum_{i=1}^n V'(u_i) y_i x_{ij}, \lambda_1\right)/(4 + \lambda_2)$.

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 `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 1 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

Bradley, P. and Mangasarian, O. (1998), "Feature selection via concave minimization and support vector machines," in *Machine Learning Proceedings of the Fifteenth International Conference (ICML'98)*, Citeseer, 82–90.

De Leeuw, J. and Heiser, W. (1977), "Convergence of correction matrix algorithms for multidimensional scaling", 735–752.

Donoho, D. L. and Johnston, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81(3), 425–455.

Huang, H., Lu, X., Liu, Y., Haaland, P., and Marron, J. (2012), "R/DWD: distance-weighted discrimination for classification, visualization and batch adjustment," *Bioinformatics*, 28(8), 1182–1183.

Hunter, D. and Lange, K. (2004), "A tutorial on MM algorithms," *The American Statistician*, 58(1), 30–37.

Lange, K., Hunter, D., and Yang, I.(2000), "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, 9(1), 1–20.

Lichman, M. (2013), "UCI Machine Learning Repository", `http://archive.ics.uci.edu/ml`, Irvine, CA: University of California, School of Information and Computer Science.

Marron, J., Todd, M., and Ahn, J. (2007), "Distance-weighted discrimination," *Journal of American Statistical Association*, 102(480), 1267–1271.

Marron, J.S. (2013), "Smoothing, functional data analysis, and distance weighted discrimination software," `http://www.unc.edu/~marron/marron_software.html`.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J. (2002), "Gene expression correlates of clinical prostate cancer behavior," *Cancer cell*, 1(2), 203–209.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Wang, L. Zhu, J., and Zou, H.(2006), "The doubly regularized support vector machine," *Statistica Sinica*, 16(2), 589–616.

Yang, Y. and Zou, H.(2013), "An efficient algorithm for computing the HHSVM and its generalizations," *Journal of Computational and Graphical Statistics*, 22(2), 396–415.

Zhang, H. H., Ahn, J., Lin, X., and Park, C. (2006), "Gene selection using support vector machines with nonconvex penalty," *Bioinformatics*, 22(1), 88–95.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004), "1-norm support vector machines," *The Annual Conference on Neural Information Processing Systems*, 16(1), 49–56.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic-net," *Journal of the Royal Statistical Society, Series B*, 67(2), 301–320.