Robust 1-Norm Soft Margin Smooth Support Vector Machine

Li-Jen Chien, Yuh-Jye Lee, Zhi-Peng Kao, and Chih-Cheng Chang

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, 106 Taiwan
{D8815002,yuh-jye,M9515040,M9415011}@mail.ntust.edu.tw

Abstract. Based on studies and experiments on the loss term of SVMs, we argue that 1-norm measurement is better than 2-norm measurement for outlier resistance. Thus, we modify the previous 2-norm soft margin smooth support vector machine ($SSVM_2$) to propose a new 1-norm soft margin smooth support vector machine ($SSVM_1$). Both SSVMs can be solved in primal form without a sophisticated optimization solver. We also propose a heuristic method for outlier filtering which costs little in training process and improves the ability of outlier resistance a lot. The experimental results show that $SSVM_1$ with outlier filtering heuristic performs well not only on the clean, but also the polluted synthetic and benchmark UCI datasets.

Keywords: classification, outlier resistance, robustness, smooth technique, support vector machine.

1 Introduction

Support vector machines (SVMs) have been proven to be one of the promising learning algorithms for classification [6]. The standard SVMs have *loss* + *penalty* terms measured by 1-norm or 2-norm measurements. The "loss" part measures the quality of model fitting and the "penalty" part controls the model complexity. In this study, our purpose is to improve original 2-norm soft margin smooth support vector machine ($SSVM_2$) [9] with robust strategies. First, we find out that the measurement of the 2-norm loss term will amplify the effect of outliers much more than the measurement of the 1-norm loss term in training process. We argue that the 1-norm loss term is better than the 2-norm loss term for outlier resistance. From this robustness point of view, we modify the previous framework in $SSVM_2$ to a new 1-norm soft margin smooth support vector machine ($SSVM_1$). We show that $SSVM_1$ can remedy the drawback of $SSVM_2$ for outlier effect and improve outlier resistance as well.

Although SVMs have the advantage of being robust for outlier effect [15], there are still some violent cases that will mislead SVM classifiers to lose their generalization ability for prediction. For example, the classification results will be very dissimilar if the difference between the total sum of the hinge losses and

the total sum of the misclassification losses is too large. Hence secondly in this study, based on the design of Newton-Armijo iterations in SSVMs, we propose a heuristic method to filter outliers among Newton-Armijo iterations of the training process and make SSVMs be more robust while encountering datasets with extreme outliers. Our method differs with other methods by truncating hinge loss [10]. It can directly and effectively drop the effect of the outliers.

The rest of the paper is organized as follows: In Section 2, we show how outliers have a great impact on SVMs. Following the idea of SSVM_2, we propose the SSVM_1 in Section 3. In Section 4, we describe how to design the heuristic method for outlier filtering. The numerical results and comparisons are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Review on Soft Margin SVMs and Outlier Effect

We first introduce the standard 1-norm soft margin SVM (SVM_1) and the standard 2-norm soft margin SVM (SVM_2). Then, we argue that the SVM_1 is more robust than the SVM_2 in outlier resistance by observing their primal and Wolfe dual formulations.

Consider the binary problem of classifying m points in the n-dimensional real space R^n, represented by an $m \times n$ matrix A. According to membership of each point $A_i \in R^{n\times 1}$ in the classes +1 or -1, D is an $m \times m$ diagonal matrix with ones or minus ones along its diagonal. Sometimes, we will take the notation y_i as the class label of A_i and the notation x_i as $A_i^\top$ for convenience. The standard 1-norm soft margin and 2-norm soft margin support vector machines are given by the following optimization problems.
1-norm soft margin SVM (SVM_1):

$$\begin{aligned} \min_{(w,b,\xi)\in R^{(n+1+m)}} \quad & \tfrac{1}{2}\|w\|_2^2 + C\|\xi\|_1 \\ subject\ to: \quad & D(Aw + \mathbf{1}b) + \xi \geq \mathbf{1} \\ & \xi \geq \mathbf{0}. \end{aligned} \tag{1}$$

2-norm soft margin SVM (SVM_2):

$$\begin{aligned} \min_{(w,b,\xi)\in R^{(n+1+m)}} \quad & \tfrac{1}{2}\|w\|_2^2 + \tfrac{C}{2}\|\xi\|_2^2 \\ subject\ to: \quad & D(Aw + \mathbf{1}b) + \xi \geq \mathbf{1} \\ & \xi \geq \mathbf{0}. \end{aligned} \tag{2}$$

The SVMs try to minimize not only the *penalty term* but also the *loss term* in the object function. In the SVM_2 (2), the 2-norm loss term will amplify the outlier effect much more as compared to the 1-norm loss term in the SVM_1 (1). The convex quadratic programs [3] of (1) and (2) can also be transformed into the following Wolfe dual problems by the Lagrangian theory [6].

The dual formulation of SVM_1:

$$\begin{aligned} \min_{\alpha \in R^m} \ & \tfrac{1}{2}\alpha^\top DAA^\top D\alpha - \mathbf{1}^\top \alpha \\ subject\ to: \ & \mathbf{1}^\top D\alpha = 0, \\ & 0 \le \alpha_i \le C,\ i = 1, 2, ..., m. \end{aligned} \tag{3}$$

The dual formulation of SVM_2:

$$\begin{aligned} \min_{\alpha \in R^m} \ & \tfrac{1}{2}\alpha^\top D(AA^\top + \tfrac{I}{C})D\alpha - \mathbf{1}^\top \alpha \\ subject\ to: \ & \mathbf{1}^\top D\alpha = 0, \\ & 0 \le \alpha_i,\ i = 1, 2, ..., m. \end{aligned} \tag{4}$$

In the dual form of SVM_2 (4), the constraint, $0 \le \alpha_i$, is a big cause of the outlier effect, where $\alpha_i = C\xi_i$ (by the optimality conditions). It means that the upper bound of α_i depending on the variable ξ_i is unlimited, and the normal vector, $w = A^\top D\alpha$, will be affected by the unrestricted α consecutively. In the SVM_1 (3), however, the maximum value of α_i could not exceed the constant value C due to the constraint, $0 \le \alpha_i \le C$. According to these observations, we argue that the SVM_1 is more robust than the SVM_2 in outlier resistance. Hence, we develop SSVM_1, which will be introduced in next section.

3 1-Norm Soft Margin Smooth SVM (SSVM_1)

Similar to the framework of SSVM_2 [9], the classification problem (1) is reformulated as follows:

$$\begin{aligned} \min_{(w,b,\xi) \in R^{n+1+m}} \ & \tfrac{1}{2}(\|w\|_2^2 + b^2) + C\|\xi\|_1 \\ subject\ to: \ & D(Aw + \mathbf{1}b) + \xi \ge \mathbf{1} \\ & \xi \ge \mathbf{0}. \end{aligned} \tag{5}$$

In the solution of problem (5), ξ is given by

$$\xi = (\mathbf{1} - D(Aw + \mathbf{1}b))_+, \tag{6}$$

where $(\cdot)_+$ is defined by $max\{\cdot, 0\}$. Namely, if $1 - D_{ii}(A_i w + b) \le 0$, then $\xi_i = 0$. Thus, this ξ in the objective function of problem (5) is replaced by $(\mathbf{1} - D(Aw + \mathbf{1}b))_+$ so that problem (5) can be converted into an unconstrained optimization problem as follows:

$$\min_{(w,b) \in R^{n+1}} \ \tfrac{1}{2}(\|w\|_2^2 + b^2) + C\left\|(\mathbf{1} - D(Aw + \mathbf{1}b))_+\right\|_1. \tag{7}$$

The problem is a strongly convex minimization problem without any constraint. Thus, problem (7) has a unique solution. Obviously, the objective function in problem (7) is not twice differentiable, so the Newton method can not be applied to solve this problem. Therefore, SSVM_2 employs a smoothing function [5] to replace the original plus function. The smoothing function is given by

$p(x, \alpha)$, the integral of the sigmoid function $\frac{1}{1+e^{-\alpha x}}$ of neural networks [11], that is,

$$p(x, \alpha) = x + \frac{1}{\alpha} \log(1 + e^{-\alpha x}) \text{ for } \alpha > 0, \tag{8}$$

where α is a smoothing parameter to adjust the degree of approximation. Note that if the value of α increases, the $p(x, \alpha)$ will approximate the plus function more accurately. Next, the $p(x, \alpha)$ is taken into problem (7) to replace the plus function as follows:

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}(\|w\|_2^2 + b^2) + C \left\| p(\mathbf{1} - D(Aw + \mathbf{1}b), \alpha) \right\|_1 . \tag{9}$$

By taking the advantage of the twice differentiability of the objective functions on problem (9), a prescribed quadratically convergent Newton-Armijo algorithm [3] can be introduced to solve this problem. Hence, the smoothing problem can be solved without a sophisticated optimization solver.

Moreover, we can obtain the unconstrained nonlinear smooth SVM_1 by applying the kernel trick [12] on problem (9) as follows:

$$\min_{(u,b)\in R^{m+1}} \frac{1}{2}(\|u\|_2^2 + b^2) + C \left\| p(\mathbf{1} - D(K(A, A^\top)u + \mathbf{1}b), \alpha) \right\|_1 . \tag{10}$$

The nonlinear separating surface is defined by the optimal solution u and b of (10):

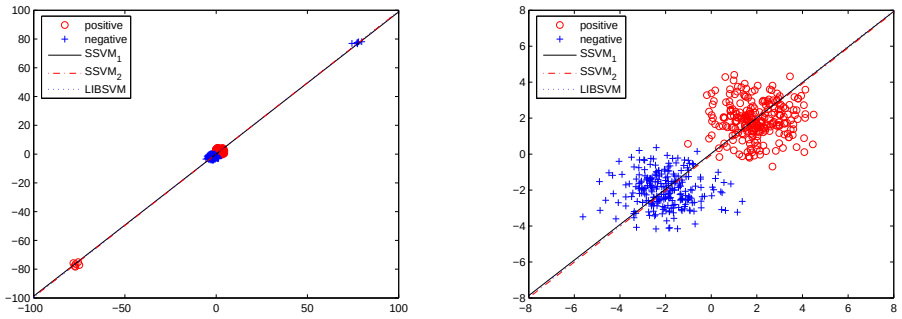
$$K(A, x)u + b = 0. \tag{11}$$

The computational complexity for solving (10) is $\mathcal{O}((m+1)^3)$. To conquer the computation difficulty caused by a big full kernel matrix $K(A, A^\top)$, we introduce the reduced kernel technique [8] to replace it by $K(A, \bar{A}^\top)$. The key idea of the reduced kernel technique is to randomly select a small portion of data and to generate a thin rectangular kernel matrix to replace the full kernel matrix. The reduced kernel method constructs a compressed model and cuts down the computational cost from $\mathcal{O}(m^3)$ to $\mathcal{O}(\bar{m}^3)$. It has been shown that the solution of reduced kernel matrix approximates the solution of full kernel matrix well.

4 A Heuristic Method for Outlier Filtering

So far, SSVM_1 has been developed for better outlier resistance, but there are some violent cases that are still easy to mislead either 1-norm soft margin SVMs or 2-norm soft margin SVMs to lose their generalization ability. We present a violent case in Fig. 1. It shows that no matter the 1-norm soft margin SVMs (SSVM_1 and LIBSVM [4]) or the 2-norm soft margin SVM (SSVM_2), all of them cannot separate the major parts of positive and negative examples. Why all of the SVMs lose their generalization ability in this case is that they pay too much effort to minimize the *loss term* and sacrifice for minimizing the *penalty term* because of these extreme outliers.

To rescue the SVMs from such the violent case, we prescribe a heuristic method to filter out the extreme outliers, which makes SVMs be more balanced to minimize both *penalty term* and *loss term* at the same time. Our strategy is



(a) entire dataset (b) centrally zoom in

Fig. 1. (Synthetic Dataset: a normal distribution, mean = 2 and -2, the standard deviation = 1) The outlier ratio is 0.025 (outliers are on the upper-right and lower-left corners in (a)). For the outliers, the outlier difference from the mean of their groups is set to be 75 times the standard deviation. All classifiers are seriously affected by these outliers.

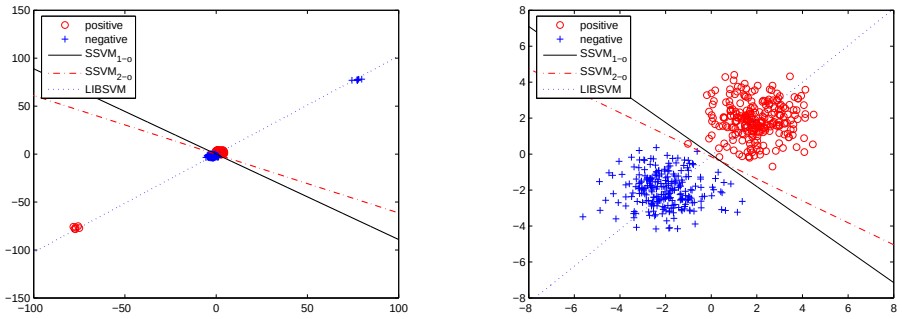
to continue removing the training input with a large ξ_i in each Newton iteration and make sure that the removed number is still smaller than the outlier ratio, which is given by the intuition of users or data collectors. In implementation, the removal is arranged to distribute fairly in several iterations according to the setting outlier ratio.

Note that the outlier filtering process is also embedded in SSVM_2 to compare with SSVM_1 in experiments. We denote $\text{SSVM}_{1\text{-}o}$ and $\text{SSVM}_{2\text{-}o}$ to represent the SSVM_1 and SSVM_2 with filtering strategy. In order to see the power of the heuristic filtering method, we test $\text{SSVM}_{1\text{-}o}$ and $\text{SSVM}_{2\text{-}o}$ on the identical synthetic dataset in Fig.1 again. Fig. 2 shows that $\text{SSVM}_{1\text{-}o}$ and $\text{SSVM}_{2\text{-}o}$ indeed remedy the previous classification results of SSVM_1 and SSVM_2 in Fig. 1, and they are superior to LIBSVM without outlier filtering mechanism.

5 Numerical Results

All codes of SSVMs are written in Matlab [14]. In experiments, we test the SSVM_2, SSVM_1, LIBSVM [4], $\text{SSVM}_{2\text{-}o}$ and $\text{SSVM}_{1\text{-}o}$ on ten publicly available binary class datasets from the UCI Machine Learning Repository [2] and CBCL datasets: `Wisconsin Prognostic Breast Cancer Database` [13], `Ionosphere`, `BUPA Liver`, `Pima Indians`, `Cleveland Heart Problem`, `WDBC`, `Image`, `Singleton`, `Waveform` and `CBCL Face Database` [1]. We perform 10-fold cross-validation on each dataset in order to evaluate how well each SVM generalizes to future data.

We train all of the classifiers by Gaussian (RBF) kernel, which is defined as $K(A, A^\top)_{ij} = e^{-\gamma\|A_i - A_j\|_2^2}$, $i, j = 1, 2, 3 \ldots m$. To build up a satisfied SVM model, we need to search a good pair of *Gaussian kernel width parameter* γ and *regularization parameter* C. A well developed model selection method is nested



(a) entire dataset (b) centrally zoom in

Fig. 2. $SSVM_{1\text{-}o}$ and $SSVM_{2\text{-}o}$ have successfully remedied the classification results of $SSVM_1$ and $SSVM_2$ in Fig. 1. LIBSVM is still affected by the outliers a lot.

uniform designs (UDs) [7], which is applied in experiments. In [7], the results by using the nested-UDs are usually good enough with much less computational cost as compared to the grid search for parameters tuning. For the large-scale datasets (`CBCL Face Database`, `Image`, `Singleton` and `Waveform`), we apply the reduced kernel technique (1% from the columns of the full kernel) to the SSVMs except for LIBSVM.

Since the specificity and the sensitivity of the tests are not unusual for all the methods, on the limit of space we just report the average training and testing correctness of 10-fold cross-validation in Table 1. In the part (b) of Table 1, we try to pollute the datasets by replacing 10% outlier training samples into each dataset. The experiments show that $SSVM_{1\text{-}o}$ performs very well in dealing with the problems with outliers.

6 Conclusions

In this paper, we argue that 1-norm soft margin SVMs have better outlier resistance than 2-norm soft margin SVMs, so we develop $SSVM_1$ by modifying the previous framework in $SSVM_2$. To strengthen the robustness of $SSVM_1$ in some violent cases, we also propose the heuristic method for outlier filtering. From experiments, we see that the 1-norm soft margin SVMs do have better robustness, and the heuristic filtering method, which costs little in training process, improves the outlier resistance a lot.

References

1. CBCL Face Database #1. MIT Center For Biological and Computation Learning. `http://cbcl.mit.edu/software-datasets/FaceData2.html`.

2. A. Asuncion and D. J. Newman. UCI Machine Learning Repository, 2007. `http://www.ics.uci.edu/~mlearn/MLRepository.html`
University of California, Irvine, School of Information and Computer Sciences.
3. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, second edition, 1999.
4. C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
5. C. Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5(2):97–138, 1996.
6. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.
7. C. M. Huang, Y. J. Lee, D. K. J. Lin, and S. Y. Huang. Model selection for support vector machines via uniform design. *A special issue on Machine Learning and Robust Data Mining of Computational Statistics and Data Analysis*, 52:335–346, 2007.
8. Y. J. Lee and S. Y. Huang. Reduced support vector machines: a statistical theory. *IEEE Transactions on Neural Networks*, 18:1–13, 2007.
9. Y. J. Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications*, 20:5–22, 2001.
10. Y. Liu and Y. Wu. Robust truncated-hinge-loss support vector machines. *Journal of the American Statistical Association*, 102:974–983, 2007.
11. O. L. Mangasarian. Mathematical Programming in Neural Networks. *ORSA Journal on Computing*, 5(4):349–360, 1993.
12. O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Shuurmans, editors, *Advance in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press.
13. O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.
14. MATLAB. *User's Guide*. The MathWorks, Inc., Natick, MA 01760, 1994–2001.
15. H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

Table 1. Numerical comparisons of nonlinear SVMs on the original and polluted data problems.

Dataset size (reduced ratio) m x n	10-fold training correctness, % 10-fold testing correctness, % Method				
	$SSVM_2$	$SSVM_1$	LIBSVM	$SSVM_{2\text{-}o}$	$SSVM_{1\text{-}o}$
WPBC	**88.69**	86.02	85.91	78.13	82.10
194 × 34	**81.67**	80.00	80.00	79.44	79.44
Ionosphere	96.78	**99.43**	**99.43**	96.66	98.45
351 × 34	96.18	**96.47**	**96.47**	95.59	95.88
BUPA	76.21	76.4	75.88	**76.50**	76.3
345 × 6	74.41	75.29	74.71	**75.59**	74.71
Pima Indians	77.95	77.62	77.88	**82.34**	77.76
768 × 8	**78.82**	**78.82**	78.42	78.29	78.55
Cleveland	**86.67**	85.54	84.53	84.34	85.47
296 × 13	84.14	**85.17**	84.48	84.14	84.48
WDBC	99.14	**99.24**	99.06	96.78	98.81
569 × 30	**98.21**	**98.21**	**98.21**	96.96	98.04
Face (r=0.01)	98.76	**98.82**	98.68	97.90	98.28
6977 × 361	98.29	**98.51**	98.38	97.84	98.05
Image (r=0.01)	**92.39**	91.52	91.67	90.49	90.54
2310 × 18	**92.16**	91.17	91.26	89.91	90.04
Singleton (r=0.01)	79.58	80.56	81.32	**81.98**	81.41
3175 × 60	79.11	79.68	**81.30**	81.17	79.87
Waveform (r=0.01)	91.52	91.86	91.47	91.94	**92.34**
5000 × 21	91.08	**91.38**	91.04	91.28	91.00

(a) The results on original data problems and the best values are emphasized in boldface. The outlier ratio parameters of $SSVM_{2\text{-}o}$ and $SSVM_{1\text{-}o}$ are set to 5%.

Dataset size (reduced ratio) m x n	10-fold training correctness, % 10-fold testing correctness, % Method				
	$SSVM_2$	$SSVM_1$	LIBSVM	$SSVM_{2\text{-}o}$	$SSVM_{1\text{-}o}$
WPBC	72.84	71.02	71.02	80.23	**81.93**
194 × 34	78.33	77.78	77.78	79.44	**80.00**
Ionosphere	87.03	88.58	85.49	84.42	**88.74**
351 × 34	92.35	**93.24**	92.94	92.06	**93.24**
BUPA	**73.05**	72.80	72.38	71.03	72.48
345 × 6	72.06	72.06	72.65	72.65	**73.82**
Pima Indians	69.93	72.30	72.47	**73.71**	72.88
768 × 8	75.00	76.71	76.84	**77.89**	77.24
Cleveland	79.25	80.67	80.04	78.50	**80.79**
296 × 13	84.83	84.83	84.14	84.83	**85.17**
WDBC	87.80	88.79	88.69	89.24	**89.55**
569 × 30	97.32	**97.68**	97.32	97.32	97.14
Face (r=0.01)	91.33	**90.89**	89.06	90.49	90.20
6977 × 361	93.39	93.74	93.29	93.97	**94.90**
Image (r=0.01)	82.02	81.99	81.30	82.71	**84.25**
2310 × 18	89.65	89.91	89.26	90.52	**91.95**
Singleton (r=0.01)	73.61	76.12	74.21	74.10	**77.82**
3175 × 60	78.45	80.89	78.45	78.58	**82.59**
Waveform (r=0.01)	83.19	83.36	83.21	**83.67**	83.61
5000 × 21	90.84	90.86	91.14	**91.18**	91.20

(b) The results on the data problems with 10% outlier pollution and the best values are emphasized in boldface. The outlier ratio parameters of $SSVM_{2\text{-}o}$ and $SSVM_{1\text{-}o}$ are set to 10%.