Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 4031626, 5 pages
https://doi.org/10.1155/2021/4031626

Research Article

Different Testing Results on SVM with Double Penalty Parameters

Chengkuan Yao,[1] **Liyong Cao**,[1] **Jianhua Xu**,[2] **and Mingya Yang**[3]

[1]*Department of Common Basic, Anqing Medical College, Anqing 246052, China*
[2]*Institute of Computer Science, Nanjing Normal University, Nanjing 210023, China*
[3]*Electrical and Mechanical Information Department, Anhui Vocational College of Press and Publishing, Hefei 230601, China*

Correspondence should be addressed to Mingya Yang; 346301992@qq.com

Received 28 July 2021; Revised 27 September 2021; Accepted 17 November 2021; Published 18 December 2021

Academic Editor: Javier Martinez Torres

The Support Vector Machine proposed by Vapnik is a generalized linear classifier which makes binary classification of data based on the supervised learning. SVM has been rapidly developed and has derived a series of improved and extended algorithms, which have been applied in pattern recognition, image recognition, etc. Among the many improved algorithms, the technique of regulating the ratio of two penalty parameters according to the ratio of the sample quantities of the two classes has been widely accepted. However, the technique has not been verified in the way of rigorous mathematical proof. The experiments based on USPS sets in the study were designed to test the accuracy of the theory. The optimal parameters of the USPS sets were found through the grid-scanning method, which showed that the theory is not accurate in any case because there is absolutely no linear relationship between ratios of penalty parameters and sample sizes.

1. Introduction

In the mid-1990s, the research team led by Vapnik proposed the advanced Support Vector Machine (SVM) [1–3]. By using a nonlinear mapping from lower-dimensional space to higher-dimensional space, the SVM seeks a hyperplane with the best classifying performance. Based on statistical learning theory and empirical risk minimization, SVM solving the optimization problem with the dual theory has become a valuable algorithm in the field of artificial intelligence.

The original SVM only had one penalty parameter. Cortes and Vapnik [3] proposed a new kind of SVM with two penalty parameters of C^+ and C^-. Chew et al. [4, 5] put forward a new idea that by using the quantities of two classes of samples to adjust C^+ and C^-, SVM has preferable classifying accuracy, which has been accepted widely. This theory, however, has not been proved mathematically. Furthermore the theory was derived from experiences described as "a rule of thumb" [4]. A number of experiments were designed to test the theory in this paper. The experiments were conducted on the dataset of USPS which is a standard handwriting database and comes from the United States Postal Service. The USPS contains ten categories of samples which are the 10 figures including 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The USPS is often used for testing in the field of machine learning.

Through the way of grid scanning for parameters, the relationships between optimal parameters were revealed when SVM achieved the optimal performance. The experimental results did not show that the optimal parameters of C^+ and C^- have the relationship with the class sizes, which was proposed and applied by Chew et al. [4, 5].

2. Support Vector Machine Algorithm

The initial SVM is to solve the quadratic programming as follows:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t. } y_i\left[(\mathbf{w}\cdot\mathbf{x}_i)+b\right] \geq 1-\xi_i, \xi_i \geq 0, i=1,\ldots,l, \tag{1}$$

where $C>0$ is the penalty parameter and $\xi_i \geq 0$ is the slack variable for the i-th data vector and $\mathbf{w}$ and b are the normal vector and the bias of the hyperplane, respectively. Figure 1 shows the example of SVM in two dimensions.

The kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ was introduced into formula (1) to replace the inner product operation, and the dual form was obtained as follows [1]:

$$\max \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k\left(\mathbf{x}_i,\mathbf{x}_j\right)$$
$$\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i = 0, 0\leq\alpha_i\leq C, i=1,\ldots,l, \tag{2}$$

where α_i are the Lagrange multipliers. The SVM with high classification accuracy should have the most samples correctly classified with $\alpha_i = 0$; meanwhile, the SVM should have the least samples misclassified with $\alpha_i > 0$.

All the training samples in Figure 1 could be divided into three cases:

(1) Nonsupport vectors (NSVs), which could be correctly classified and not in H_1 and H_2, can satisfy the following formula:

$$y_i\left[(\mathbf{w}\cdot\mathbf{x}_i)+b\right]-1+\xi_i>0, \quad \alpha_i=0, \mu_i=C, \xi_i=0. \tag{3}$$

(2) Support vectors (SVs), which could be correctly classified but located on H_1 and H_2, can satisfy the following formula:

$$y_i\left[(\mathbf{w}\cdot\mathbf{x}_i)+b\right]-1+\xi_i=0, \quad \xi_i=0, \mu_i\neq 0, 0<\alpha_i<C. \tag{4}$$

(3) Bounded support vectors (BSVs), which are misclassified, can satisfy the following formula:

$$y_i\left[(\mathbf{w}\cdot\mathbf{x}_i)+b\right]-1+\xi_i=0, \quad \xi_i\neq 0, \mu_i=0, \alpha_i=C. \tag{5}$$

3. Improved SVM Algorithm

3.1. SVM with Double Penalty Parameters. Two penalty parameters, namely, C^+ and C^-, were introduced by Osuna et al. [6]. The optimization problem is minimized, taking the form

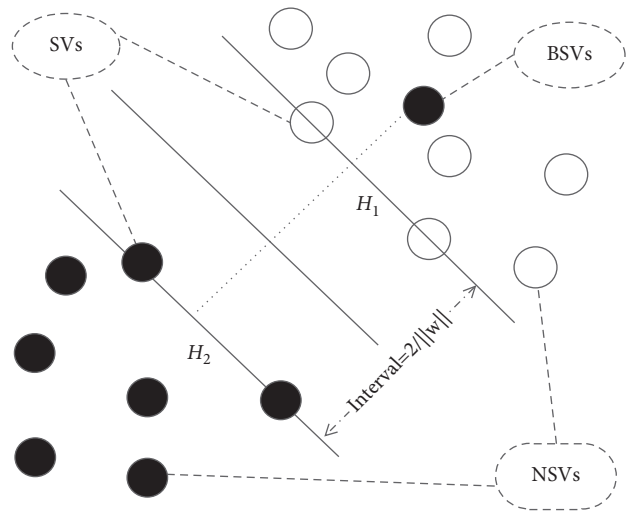

Figure 1: Example of SVM in two dimensions.

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C^+\left(\sum_{i:y_i=+1}\xi_i\right) + C^-\left(\sum_{i:y_i=-1}\xi_i\right)$$
$$\text{s.t. } y_i\left[(\mathbf{w}\cdot\mathbf{x}_i)+b\right]\geq 1-\xi_i, \quad \xi_i\geq 0, i=1,\ldots,l, \tag{6}$$

where C^+ and C^- are the error penalties for the positive ($y_i=+1$) and the negative ($y_i=-1$) vectors, respectively. The dual form of (6) is

$$\max \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k\left(\mathbf{x}_i,\mathbf{x}_j\right)$$
$$\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i = 0$$
$$0\leq\alpha_i\leq C^+, \text{if } y_i=+1$$
$$0\leq\alpha_i\leq C^-, \text{if } y_i=-1. \tag{7}$$

Chew et al. raised a point [4] that in order to avoid to overlearning of SVM, the BSVs should greatly outnumber the SVs in the SVM. That is to say, the misclassified samples should far exceed the samples on the boundary line. In other words, in the SVM with the best performance, there are enough BSVs to avoid overlearning and the SVs could be ignored because of their minuscule amount.

In the view of Chew et al. [4], the error rates of the two classes of samples can be expressed as B_+/N^+ and B_-/N^-, where B_+ *and* B_- are the numbers of the misclassified samples (BSVs) of the two classes and $N^+$and N^- are the numbers of samples of the two classes, respectively. Thus, the constraint in formula (7) can be repressed as

$$\begin{aligned} \sum \alpha_i &\approx B_+ \cdot C^+, \quad \text{if } (y_i = +1), \\ \sum \alpha_i &\approx B_- \cdot C^-, \quad \text{if } (y_i = -1), \\ B_+ \cdot C^+ &\approx B_- \cdot C^-. \end{aligned} \tag{8}$$

Setting the ratios of the error rate between the positive class and the negative class,

$$\begin{aligned} &\frac{B_+}{N^+} : \frac{B_-}{N^-}, \\ &\frac{1}{(C^+ N^+)} : \frac{1}{(C^- N^-)}, \\ &C^- N^- : C^+ N^+. \end{aligned} \tag{9}$$

When the following equation is true, the SVM has the best classification performance:

$$\frac{C^+}{C^-} = \frac{N^-}{N^+}. \tag{10}$$

3.2. υ-SVM. Schölkopf et al. [7] put forward the υ-SVM algorithm with the parameter υ to replace C and ξ' in the original SVM. The υ-SVM is shown as follows:

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{w}\|^2 - \upsilon\rho + \frac{1}{l}\sum_i \xi_i \\ &\text{s.t.} \quad y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq \rho - \xi_i \\ &\qquad \xi_i \geq 0, i = 1, \ldots, l, \rho \geq 0, \end{aligned} \tag{11}$$

where ρ is the location of the margin and the $2\rho/\|\mathbf{w}\|$ is the width of the margin. When minimizing the classifying error, the maximum width of the margin could be obtained. Chew et al. improved the υ-SVM [5] and made the same conclusion that the optimal parameters of υ-SVM satisfy the relationship of equation (10).

Equation (10) has been widely recognized and put into use in the two-class sample sets, and the results have been shown with the optimal classification performance [8–10].

4. Hypothesis Testing

There is a lack of rigorous mathematical derivations and proofs for formula (10). Furthermore, the SVs which could be correctly classified (located on H_1 and H_2 in Figure 1) are ignored, which bring about the only usage of the symbol "≈" instead of "=" in formula (8). In addition, formula (10) only shows the ratio of C^+/C^- and not the absolute value of C^+ and C^- that actually should be specified.

In the following experiment, it will be verified whether the optimal parameters of SVM satisfy equation (10).

4.1. Methods and Steps of the Testing. The experiment was divided into two main steps. In the first step, ten two-class data sets were constructed, which were tested by different parameters in the second part.

4.1.1. Establishment of Two-Class Data Sets. The USPS handwritten digital dataset is often used for algorithm testing in the field of pattern recognition and machine learning. There are 10-class samples in the USPS, which are the 10 figures formed from 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The USPS has 256 attributions. The sample size of the training set and the testing set are 7291 and 2007, respectively [11].

The USPS was transformed into 10 two-class data sets described as USPS-0, USPS-1, USPS-2, . . ., USPS-9. The USPS-0 set contains two class samples, which are the digitals of 0 and all the non-0 digitals formed from digitals of 1, 2, 3, 4, 5, 6, 7, 8, and 9. The USPS-1 set contains two class samples, which are the digitals of 1 and all the non-1 digitals formed from digitals of 0, 2, 3, 4, 5, 6, 7, 8, 9, ... The USPS-9 set contains two class samples, which are the digitals of 9 and all the non-9 digitals formed from digitals of 0, 1, 2, 3, 4, 5, 6, 7, and 8.

4.1.2. Testing Procedure. The method of grid scanning for parameters was adopted to try to find out the optimal parameters of the ten two-class datasets. The polynomial kernel function was employed, and the software package of LIBSVM [12] was used to carry out the experiment.

The steps are as follows:

(1) The order of the polynomial $d = \{2, 3, 4, 5, 6\}$ was specified; in other words, there were five major cycles with five different values of order of the polynomial.

(2) The penalty parameters C^+ and $C^- = [100 \ldots 10000]$, and the step width was 100. To be more specific, the value of C^+ was 100, 200, 300, . . ., 1000, 1100, . . ., 2000, 2100, . . ., 3000, and 3100, . . ., 10000 in order, so was the value of C^-.

(3) We performed three rounds of cyclic scanning for the three parameters, namely, d, C^+, and C^-. That is to say, each of the ten two-class datasets was tested to get the testing precision more than 50,000 times with different combinations of the three parameters.

When the highest accuracy was achieved, the parameter is the optimal parameter. If there were multiple optimal parameters, the following two rules should be abided to choose the optimal parameters:

(1) Choosing $\min(C^+ + C^-)$, which was the inflection point of parameters

(2) If there are still multiple equal $\min(C^+ + C^-)s$, choosing $\min d$ which was the lowest order of polynomial

4.2. Testing Result. After performing grid scanning on the ten two-class data sets, the optimal parameters are obtained as shown in Table 1. The first column on the left side in the table is the category identifier of the 10 data sets. In the first row, d is the order of polynomial; N^+ and N^- are the sizes of samples of the two classes, respectively; and C^+ and C^- are the penalty parameters of the two categories.

Table 1: Sample sizes and the optimal parameters and accuracy of the 10 two-class data sets.

Category Identifier	d	N^+	N^-	C^+	C^-	Classification accuracy
USPS-0	5	644	6647	1200	200	0.9945
USPS-1	5	1194	6097	2600	1000	0.9945
USPS-2	4	1005	6286	400	200	0.9965
USPS-3	5	731	6560	2800	3000	0.9875
USPS-4	6	658	6633	600	200	0.9910
USPS-5	3	652	6639	300	400	0.9865
USPS-6	4	556	6735	200	100	0.9895
USPS-7	3	664	6627	2600	200	0.9955
USPS-8	5	645	6646	1200	800	0.9940
USPS-9	4	542	6749	1700	200	0.9880

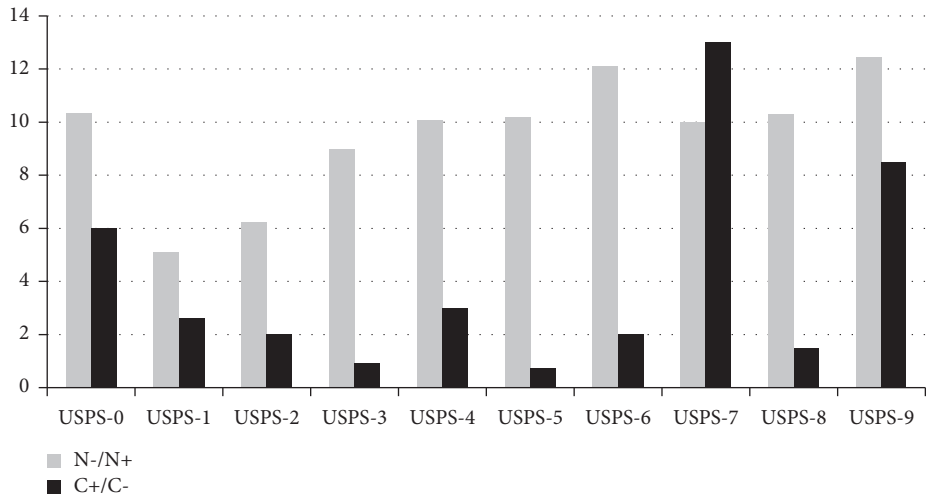

Figure 2: Histogram of C^+/C^- and N^-/N^+.

In order to find whether the optimal parameters in Table 1 satisfy formula (10) more intuitively, the histogram of C^+/C^- with N^-/N^+ is shown in Figure 2.

In Figure 2, there are nine datasets satisfying the relationship of $N^-/N^+ > C^+/C^-$. Meanwhile, only one dataset of USPS-7 meets the relationship of $N^-/N^+ < C^+/C^-$. The experiment came to the conclusion that the relationship between the sample size and the optimal parameters does not satisfy formula (10); in other words, the following formula is true:

$$\frac{C^+}{C^-} \neq \frac{N^-}{N^+}. \tag{12}$$

Formula (10) has not been proved in the mathematical way rigorously, but a kind of reasoning based on experience. Also, formula (10) is too simple to reveal the internal relations between the sample size and the optimal parameters.

The core of SVM is SVs [1–3] which could be correctly classified and located on H_1 and H_2 in Figure 1, and the classification hyperplane is mainly derived from the attributes of SVs. No matter the number of SVs is more or less, it cannot be ignored [4, 5] under any circumstances.

In fact, the number of the correctly classified support vectors (SVs) should exceed the number of bounded support vectors (BSVs) in the SVM with preferable performance. Obviously, as seen from Figure 1, the more the points on H_1 and H_2, the better the performance the SVM has. Meanwhile, the misclassified points should be as few as possible.

4.3. Analysis of the Means and Standard Deviation. In the grid scanning in Section 4.1.2, there are still many parameters with the same optimal value. That is to say, the optimal value is obtained at many points. In order to further verify the relationship between penalty parameters and sample sizes, parameters with the same optimal classification accuracy were statistically analyzed.

The means and the standard deviations of C^+/C^-s with the same optimal value were calculated, which are shown in Table 2. The second column on the left is the number of the same optimal values, which is presented as N(opt. C^+/C^-). Meanwhile, the means and the standard deviations of the optimal C^+/C^-s are presented as E(opt. C^+/C^-) and σ in the third and the right column.

Based on $N^- > N^+$ in Table 1, the following inequality is true:

$$\frac{N^-}{N^+} > 1. \tag{13}$$

If formula (10) is true, combined with formula (13), the following formula can be derived:

Table 2: Means and the standard deviations of the same optimal values of C^+/C^-s of the 10 two-class sets.

Category identifier	$N(\text{opt}.C^+/C^-)$	$E(\text{opt}.C^+/C^-)$	σ
USPS-0	40	4.04	2.38
USPS-1	98	0.45	2.26
USPS-2	31	0.27	3.59
USPS-3	70	2.01	1.40
USPS-4	22	3.98	2.74
USPS-5	145	0.56	3.45
USPS-6	51	0.31	2.89
USPS-7	96	2.77	2.98
USPS-8	155	1.69	1.84
USPS-9	57	3.14	2.13

$$\frac{C^+}{C^-} > 1. \tag{14}$$

However, there are four $E(\text{opt. } C^+/C^-)$s in Table 2, namely, USPS-1, USPS-2, USPS-5, and USPS-6, satisfying the following inequality:

$$\frac{C^+}{C^-} < 1. \tag{15}$$

In fact, formulas (14) and (15) contradict each other, which is another proof that formula (10) is not true at any condition.

5. Conclusion

The method of grid scanning for parameters was employed to find the optimal values, which was designed to reveal the relationship between the optimal parameters and the sample sizes. Since the parameters are infinite, it is impossible to test all of the parameter possibilities. Also, the optimal parameters were tested in a very wide range of thresholds, which used much more time.

From the results of the study, it is believed that the optimal parameters of C^+ and C^- by no means rely on the size of samples. To be more exact, there is absolutely no linear relationship between ratios of penalty parameters and sample sizes.

At present, all parameter optimization in machine learning is local optimization and the study in the paper is no exception. Optimization algorithms, such as gradient descent, Newton's method, and Quasi-Newton methods, could be used to find out the optimal parameters of SVM, which is an iterative process and certainly takes a lot of time. Therefore, finding the optimal parameters in limited and acceptable time must be very valuable, which is the new research direction worth exploring.

Data Availability

All data included in this study are available upon request by contact with the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the Professor Foundation of Anqing Medical College under the grant of Feng Guang.

References

[1] V. N. Vapnik, *Statistical Learning Theory*, Wiley, NewYork, NY, USA, 1998.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Conference on Computational Learning Theory*, pp. 144–152, Pittsburgh, PA, USA, July 1992.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] H. G. Chew, D. J. Crisp, R. E. Bogner, and C. C. Lim, "Target detection in radar imagery using support vector machines with training size biasing," *Southern Medical Journal*, vol. 90, no. 10, pp. 959–963, 2000.

[5] H. G. Chew, R. E. Bogner, and C. C. Lim, "Dual v-support vector machine with error rate and training size biasing, digital object identifier," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 1269–1272, Salt Lake City, UT, USA, May 2001.

[6] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: training and application," MIT AI lab, Cambridge, MA, USA, AIM1602, 1997.

[7] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[8] P. Xu and A. Chan, "Support vector machine for multi-class signal classification with unbalanced samples," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 1116–1119, Portland, OR, USA, July 2003.

[9] H. Kim and H. Park, "Protein secondary structure prediction based on an improved support vector machines approach," *Protein Engineering Design and Selection*, vol. 16, no. 8, pp. 553–560, 2003.

[10] K. Morik, P. Brochhausen, and T. Joachims, "Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring," in *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, June 1999.

[11] LIBSVM Data: Classification (Multi Class) [EB/OL], https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps.

[12] LIBSVM: A Library for Support Vector Machines [EB/OL], https://www.csie.ntu.edu.tw/~cjlin/libsvm/.