Bi-Parameter Space Partition for Cost-Sensitive SVM

Bin Gu*†, Victor S. Sheng‡, Shuo Li§†,
*School of Computer & Software, Nanjing University of Information Science & Technology, China
†Department of Computer Science, University of Western Ontario, Canada
‡Department of Computer Science, University of Central Arkansas, Arkansas
§GE HealthCare, Canada
jsgubin@nuist.edu.cn, ssheng@uca.edu, Shuo.Li@ge.com

Abstract

Model selection is an important problem of cost-sensitive SVM (CS-SVM). Although using solution path to find global optimal parameters is a powerful method for model selection, it is a challenge to extend the framework to solve two regularization parameters of CS-SVM simultaneously. To overcome this challenge, we make three main steps in this paper. (**i**) A critical-regions-based bi-parameter space partition algorithm is proposed to present all piecewise linearities of CS-SVM. (**ii**) An invariant-regions-based bi-parameter space partition algorithm is further proposed to compute empirical errors for all parameter pairs. (**iii**) The global optimal solutions for K-fold cross validation are computed by superposing K invariant region based bi-parameter space partitions into one. The three steps constitute the model selection of CS-SVM which can find global optimal parameter pairs in K-fold cross validation. Experimental results on seven normal datsets and four imbalanced datasets, show that our proposed method has better generalization ability and than various kinds of grid search methods, however, with less running time.

1 Introduction

Ever since Vapnik's influential work in statistical learning theory [Vapnik and Vapnik, 1998], Support Vector Machines (SVMs) have been successfully applied to a lot of classification problems due to its good generalization performance. However, in many real-world classification problems such as medical diagnosis [Park *et al.*, 2011], object recognition [Zhang and Zhou, 2010], business decision making [Cui *et al.*, 2012], and so on, the costs of different types of mistakes are naturally unequal. Cost sensitive learning [Sheng and Ling, 2006] takes the unequal misclassification costs into consideration, which has also been deemed as a good solution to class-imbalance learning where the class distribution is highly imbalanced [Elkan, 2001]. There have been several cost-sensitive SVMs, such as the boundary movement [Shawe-Taylor, 1999], biased penalty ($2C$-SVM [Schölkopf and Smola, 2002] and 2ν-SVM [Davenport *et al.*, 2010]), cost sensitive hinge loss [Masnadi-Shirazi and Vasconcelos, 2010], and so on. In this paper, we focus on the most popular one ($2C$-SVM[1]) of them.

Given a training set $\mathcal{S} = \{(x_1, y_1), \cdots, (x_l, y_l)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$, $2C$-SVM introduces two cost parameters C_+ and C_- to denote the costs of false negative and false positive respectively, and considers the following primal formulation:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\langle w, w\rangle + C_+ \sum_{i\in\mathcal{S}^+} \xi_i + C_- \sum_{i\in\mathcal{S}^-} \xi_i \qquad (1) \\ s.t. \quad & y_i\left(\langle w, \phi(x_i)\rangle + b\right) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, l \end{aligned}$$

where $\phi(x_i)$ denotes a fixed feature-space transformation, $\mathcal{S}^+ = \{(x_i, y_i) : y_i = +1\}$, and $\mathcal{S}^- = \{(x_i, y_i) : y_i = -1\}$. The dual problem of (1) is

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - \sum_{i\in\mathcal{S}} \alpha_i, \quad s.t. \sum_{i\in\mathcal{S}} y_i\alpha_i = 0, \qquad (2) \\ & 0 \leq \alpha_i \leq \frac{C_+ + C_- + y_i(C_+ - C_-)}{2}, i = 1, \cdots, l \end{aligned}$$

where Q is a positive semidefinite matrix with $Q_{ij} = y_i y_j K(x_i, x_j)$ and $K(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$. It is obviously noted that how one tunes the cost parameter pair (C_+, C_-) to achieve optimal generalization performance (it is also called the problem of model selection) is a central problem of CS-SVM.

A general approach to tackle this problem is to specify some candidate parameter values, and then apply cross validation (CV) to select the best choices. A typical implementation for this approach is grid search [Mao *et al.*, 2014]. However, extensive explorating the optimal parameter values is seldom pursued, because there exist double-sided difficulties. 1) It requires to train the classifier many times under different parameter settings. 2) And testing it on the validation dataset for each parameter setting.

To overcome the first difficulty, solution path algorithms were proposed for many learning models, such as C-SVM [Hastie *et al.*, 2004], ϵ-SVR [Gunter and Zhu, 2007], quantile regression [Rosset, 2009] and so on, to fit the entire solutions for every value of the parameter. It should be noted that there

[1] Actually, 2ν-SVM is equivalent to $2C$-SVM as proved in [Davenport *et al.*, 2010]. For the sake of convenience, we do not distinguish the names of $2C$-SVM and CS-SVM hereafter unless explicitly mentioned.

are also several works involving the bi-parametric solution path. Wang et al. [Wang *et al.*, 2008] works with respect to only one parameter of ϵ-SVR while the other parameter is fixed. Bach et al. [Bach *et al.*, 2006] search the space (C_+, C_-) of $2C$-SVM by using a large number of parallel one-parametric solution paths. Rosset's model [Rosset, 2009] follows a large number of one-parametric solution paths simultaneously. Essentially, they all follow one-parametric solution paths in bi-parameter space in different ways, and none of them can explore all solutions for every parameter pair. To address the second difficulty, a global search method [Yang and Ong, 2011] was recently proposed based on the solution path. It can find the global optimal parameters in K-fold CV for C-SVM [Yang and Ong, 2011] and ν-SVM [Gu *et al.*, 2012]. The power of the method is proved by theoretical and empirical analysis for model selection. Therefore, it is highly desirable to design an extension version for CV on the bi-parametric problem (e.g. CS-SVM) based on fitting all solutions for each parameter pair.

The contributions of this paper can be summarized as follows. (**i**) We propose a bi-parameter space partition (BPSP) algorithm, which can fit all solutions for every parameter pair (C_+, C_-). To the best of our knowledge, it is the first such contribution. (**ii**) Based on the bi-parameter space partition, we propose a K-fold cross validation algorithm for computing the global optimum parameter pairs of CS-SVM. Experimental results demonstrate that the method has better generalization ability than various kinds of grid search methods, however, with less running time.

2 CS-SVM and KKT conditions

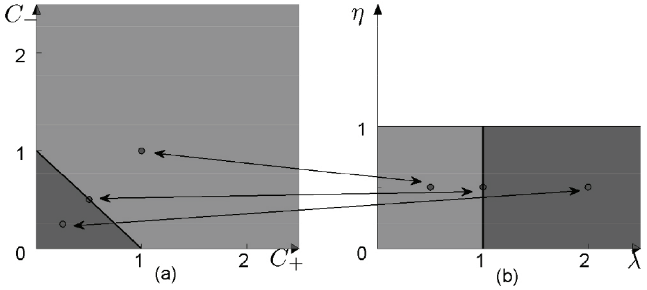


Figure 1: The corresponding relation between the (C_+, C_-) and (λ, η) coordinate systems.

We reformulate the primal formulation of $2C$-SVM with $\lambda = \frac{1}{C_+ + C_-}$ and $\eta = \frac{C_+}{C_+ + C_-}$, as presented in (3), and name it (λ, η)-SVM. Fig. 1 shows the relation between the (C_+, C_-) and (λ, η) coordinate systems. Specifically, the region of $C_+ > 0$, $C_- > 0$, and $C_+ + C_- \geq 1$ corresponds the region of $0 < \lambda \leq 1$ and $0 \leq \eta \leq 1$. Thus, the whole region of (C_+, C_-) can be explored in 1.5 square units by searching the region of $(0, 1] \times [0, 1]$ in the (λ, η) coordinate system, and the lower triangle region of $[0, 1] \times [0, 1]$ in the (C_+, C_-) coordinate system.

$$\min_{w,b,\xi} \quad \frac{\lambda}{2}\langle w, w\rangle + \eta \sum_{i\in\mathcal{S}^+} \xi_i + (1-\eta)\sum_{i\in\mathcal{S}^-} \xi_i \tag{3}$$
$$s.t. \quad y_i\left(\langle w, \phi(x_i)\rangle + b\right) \geq 1 - \xi_i, \;\; \xi_i \geq 0, i = 1, \cdots, l$$

The corresponding dual of (3) is

$$\min_{\alpha} \quad \frac{1}{2\lambda}\alpha^T Q\alpha - \sum_{i\in\mathcal{S}} \alpha_i, \quad s.t. \; \sum_{i\in\mathcal{S}} y_i\alpha_i = 0, \tag{4}$$
$$0 \leq \alpha_i \leq \frac{1 - y_i + 2y_i\eta}{2}, \quad i = 1, \cdots, l$$

Letting $g_i = \frac{1}{\lambda}\left(\sum_{j\in\mathcal{S}} \alpha_j Q_{ij} + y_i b''\right) - 1$, from the KKT theorem [Boyd and Vandenberghe, 2009], we obtain the following KKT conditions for (4):

$$\forall i \in \mathcal{S}: \quad \begin{cases} g_i > 0 \; for \; \alpha_i = 0 \\ g_i = 0 \; for \; 0 \leq \alpha_i \leq \frac{1-y_i+2y_i\eta}{2} \\ g_i < 0 \; for \; \alpha_i = \frac{1-y_i+2y_i\eta}{2} \end{cases} \tag{5}$$
$$\sum_{i\in\mathcal{S}} y_i\alpha_i = 0 \tag{6}$$

where $b' = \lambda b''$, b'' is the Lagrangian multiplier corresponding to the equality constraint in (4). According to the value of g_i, a training sample set $\mathcal{S}$ is partitioned as $\pi(\lambda, \eta) = (\mathcal{M}(\lambda, \eta), \mathcal{E}(\lambda, \eta), \mathcal{R}(\lambda, \eta))$, where $\mathcal{M}(\lambda, \eta) = \{i : \; g_i = 0, \; 0 \leq \alpha_i \leq \frac{1-y_i+2y_i\eta}{2}\}$, $\mathcal{E}(\lambda, \eta) = \{i : \; g_i < 0, \; \alpha_i = \frac{1-y_i+2y_i\eta}{2}\}$; $\mathcal{R}(\lambda, \eta) = \{i : \; g_i > 0, \; \alpha_i = 0\}$.

Similar to (5)-(6), we can give the KKT conditions for (2). Accordingly, the set $\mathcal{S}$ has the partition $\pi(C_+, C_-) = (\mathcal{M}(C_+, C_-), \mathcal{E}(C_+, C_-), \mathcal{R}(C_+, C_-))$.

3 BPSP using Critical Regions

3.1 Detecting the Critical Convex Polygon Region

Given a partition $\pi(\lambda_0, \eta_0)$, we have the critical region $\mathcal{CR}(\lambda_0, \eta_0) = \{(\lambda, \eta) \in (0, 1] \times [0, 1] : \pi(\lambda, \eta) = \pi(\lambda_0, \eta_0)\}$ induced by the bi-parametric piecewise linear solution. Theorem 1 shows that $\mathcal{CR}(\lambda_0, \eta_0)$ is a convex polygon region.

Theorem 1. *The set $\mathcal{CR}(\lambda_0, \eta_0)$ is a convex set and its closure is a convex polygon region.*

When adjusting λ and η, the weights of the samples in $\mathcal{M}$ and the variable b' should also be adjusted accordingly. From $g_i = 0$, $\forall i \in \mathcal{M}$, and the equality constraint (6), and let $\widetilde{g}_i = \lambda(g_i + 1)$, we have the following linear system:

$$\begin{aligned} \Delta\widetilde{g}_i \;\; &\overset{\text{def}}{=} \;\; \sum_{j\in\mathcal{M}} Q_{ij}\Delta\alpha_j + y_i\Delta b' + \sum_{j\in\mathcal{E}} y_j Q_{ij}\Delta\eta \\ &= \;\; \Delta\lambda, \;\; \forall i \in \mathcal{M} \end{aligned} \tag{7}$$
$$\sum_{j\in\mathcal{M}} y_j\Delta\alpha_j + \sum_{j\in\mathcal{E}} \Delta\eta = 0 \tag{8}$$

If $\mathbf{1}_\mathcal{M}$ defined as the $|\mathcal{M}|$-dimensional column vector with all ones, and let $\mathbf{y}_\mathcal{M} = [y_1, \cdots, y_{|\mathcal{M}|}]^T$, the linear system

(7)-(8) can be rewritten as:

$$\begin{aligned}\underbrace{\left[\begin{array}{cc} 0 & \mathbf{y}_{\mathcal{M}}^T \\ \mathbf{y}_{\mathcal{M}} & Q_{\mathcal{M}\mathcal{M}} \end{array}\right]}_{\widetilde{Q}} \left[\begin{array}{c} \Delta b' \\ \Delta \alpha_{\mathcal{M}} \end{array}\right] & \qquad (9) \\ = \left[\begin{array}{cc} 0 & -|\mathcal{E}| \\ \mathbf{1}_{\mathcal{M}} & -\sum_{j\in\mathcal{E}} y_j Q_{\mathcal{M}j} \end{array}\right] \left[\begin{array}{c} \Delta\lambda \\ \Delta\eta \end{array}\right] &\end{aligned}$$

Let $R = \widetilde{Q}^{-1}$, the linear relationship between $\left[\Delta b' \ \ \Delta\alpha_{\mathcal{M}}^T\right]^T$ and $[\Delta\lambda \ \ \Delta\eta]^T$ can be obtained as follows:

$$\begin{aligned}\left[\begin{array}{c} \Delta b' \\ \Delta \alpha_{\mathcal{M}} \end{array}\right] &= R \left[\begin{array}{cc} 0 & -|\mathcal{E}| \\ \mathbf{1}_{\mathcal{M}} & -\sum_{j\in\mathcal{E}} y_j Q_{\mathcal{M}j} \end{array}\right] \left[\begin{array}{c} \Delta\lambda \\ \Delta\eta \end{array}\right] \\ &\overset{\text{def}}{=} \left[\begin{array}{cc} \beta_{b'}^{\lambda} & \beta_{b'}^{\eta} \\ \beta_{\mathcal{M}}^{\lambda} & \beta_{\mathcal{M}}^{\eta} \end{array}\right] \left[\begin{array}{c} \Delta\lambda \\ \Delta\eta \end{array}\right] \qquad (10)\end{aligned}$$

Substituting (10) into (7), we can get the linear relationship between $\Delta\widetilde{g}_i$ ($\forall i \in \mathcal{S}$) and $[\Delta\lambda \ \ \Delta\eta]^T$ as follows:

$$\begin{aligned}\Delta\widetilde{g}_i &= \sum_{j\in\mathcal{M}} Q_{ij}\left(\beta_j^{\lambda}\Delta\lambda + \beta_j^{\eta}\Delta\eta\right) \\ &\quad + y_i\left(\beta_{b'}^{\lambda}\Delta\lambda + \beta_{b'}^{\eta}\Delta\eta\right) + \sum_{j\in\mathcal{E}} y_j Q_{ij}\Delta\eta \\ &\overset{\text{def}}{=} \gamma_i^{\lambda}\Delta\lambda + \gamma_i^{\eta}\Delta\eta \qquad (11)\end{aligned}$$

When adjusting λ and η, meanwhile keeping all the samples satisfying the KKT conditions, the following constraints should be kept:

$$\begin{aligned}0 \leq \alpha(\lambda_0,\eta_0)_i + \beta_i^{\lambda}(\lambda-\lambda_0) + \beta_i^{\eta}(\eta-\eta_0) & \qquad (12) \\ \leq \frac{1-y_i+2y_i\eta}{2}, \forall i \in \mathcal{M} &\end{aligned}$$

$$\widetilde{g}(\lambda_0,\eta_0)_i + \gamma_i^{\lambda}(\lambda-\lambda_0) + \gamma_i^{\eta}(\eta-\eta_0) \leq \lambda, \forall i \in \mathcal{E} \qquad (13)$$

$$\widetilde{g}(\lambda_0,\eta_0)_i + \gamma_i^{\lambda}(\lambda-\lambda_0) + \gamma_i^{\eta}(\eta-\eta_0) \geq \lambda, \forall i \in \mathcal{R} \qquad (14)$$

Obviously, the set of above inequalities is a convex polygon region. The compact representation can be obtained after removing redundant inequalities from (12)-(14) by the vertex enumeration algorithm [Avis and Fukuda, 1992]. Here, we use $\mathcal{CR}(\lambda_0,\eta_0)$ to denote the compact representation of (12)-(14).

Similarly, given a partition $\pi(C_+^0, C_-^0)$, we can obtain the critical region $\mathcal{CR}(C_+^0, C_-^0)$ as (12)-(14). Obviously, it is also a convex polygon region.

3.2 Critical-Regions-Based BPSP Algorithm

This section tries to find all critical convex polygon regions in $(0,1] \times [0,1]$ for the (λ,η) parameter space, and the lower triangle region of $[0,1] \times [0,1]$ for the (C_+, C_-) parameter space. It means all solutions of CS-SVM would be obtained.

An intuitive idea to find all the regions is using a progressive construction method. Before designing this progressive construction algorithm, there are three problems which should be answered. (i) How do we give an initial solution of the first critical convex polygon region for (λ,η)-SVM and $2C$-SVM? (ii) How do we handle the problem of overlapped critical convex polygon regions? (iii) How do we find the next critical convex polygon regions based on the current one? Our answers to the three problems are as follows, which derive a recursive bi-parameter space partition algorithm (i.e., Algorithm 1).

Initialization A simple strategy for initialization is directly using the SMO technology [Cai and Cherkassky, 2012] or other quadratic programming solvers to find solution for a parameter pair in $(0,1] \times [0,1]$ for the (λ,η) parameter space, or the lower triangle region of $[0,1] \times [0,1]$ for the (C_+, C_-) parameter space, respectively. Here, a method without requiring any numerical solver is presented in Lemma 1 and 2, which can directly give the solutions for the parameter pairs of (λ,η)-SVM and $2C$-SVM, respectively, under some conditions.

Lemma 1. *All the* $\alpha_i = \frac{1-y_i+2y_i\eta}{2}$*, which is the optimal solution of the minimization problem (4) with* $\eta = \frac{|\mathcal{S}^-|}{|\mathcal{S}|}$ *and* $\lambda \geq \frac{1}{2}\left(\max_{i\in\mathcal{S}^+}\sum_{j\in\mathcal{S}}\alpha_j Q_{ij} + \max_{i\in\mathcal{S}^-}\sum_{j\in\mathcal{S}}\alpha_j Q_{ij}\right)$.

Lemma 2. *All the* $\alpha_i = \frac{C_+ + C_- + y_i(C_+ - C_-)}{2}$*, which is the optimal solution of the minimization problem (2) with* $\frac{C_+}{C_-} = \frac{|\mathcal{S}^-|}{|\mathcal{S}^+|}$ *and* $C_+ + C_- \leq \frac{2}{\max_{i\in\mathcal{S}^+} h_i + \max_{i\in\mathcal{S}^-} h_i}$*, where* $h_i = \sum_{j\in\mathcal{S}^+}\frac{|\mathcal{S}^-|}{|\mathcal{S}|}Q_{ij} + \sum_{j\in\mathcal{S}^-}\frac{|\mathcal{S}^+|}{|\mathcal{S}|}Q_{ij}$.

Partitioning the Parameter Space The minimization problem (4) or (2) can not be guaranteed to be strict convex in many real-world problems. There exists the phenomenon of overlapped critical convex polygon regions (see Fig. 2 (a)). This makes it difficult to find all critical convex polygon regions by a progressive construction method. A parameter space partition method is presented by Theorem 2 [Borrelli, 2003], where A and b are issued from the compact representation of inequalities (12)-(14), which can be computed by the vertex enumeration algorithm [Avis and Fukuda, 1992]. m is the number of inequalities in the compact representation. (ρ, ϱ) is the shorthand implying (λ,η) and (C_+, C_-) hereafter.

Theorem 2. *Let* $\mathcal{X} \subseteq \mathbb{R}^2$ *be a convex polygon region, and* $\mathcal{R}_0 = \{(\rho,\varrho) \in \mathcal{X} : A\left[\rho \ \ \varrho\right]^T \leq b\}$ *be a convex polygon subregion of* $\mathcal{X}$*, where* $A \in \mathbb{R}^{m\times 2}$*,* $b \in \mathbb{R}^{m\times 1}$*,* $\mathcal{R}_0 \neq \emptyset$*. Also let*

$$\begin{aligned}\mathcal{R}_i &= \left\{(\rho,\varrho) \in \mathcal{X} \,\middle|\, \begin{array}{l} A_i\left[\rho \ \ \varrho\right]^T > b_i \\ A_j\left[\rho \ \ \varrho\right]^T \leq b_j, \ \ \forall j < i \end{array}\right\}, \\ & \quad \forall i = 1, \cdots, m\end{aligned}$$

then $\{\mathcal{R}_0, \mathcal{R}_1, \cdots, \mathcal{R}_m\}$ *is a partition of* $\mathcal{X}$*, i.e.,* $\bigcup_{i=0}^m \mathcal{R}_i = \mathcal{X}$*, and* $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$*,* $\forall i \neq j$*,* $i, j \in \{0, 1, \cdots, m\}$*.*

Theorem 2 defines a partition procedure which consists of considering one by one the inequalities of $\mathcal{R}_0$. See Fig. 2 (b), the four inequalities of $\mathcal{R}_0$ induce four disjoint subregions of $\mathcal{X}$ (i.e., $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, and $\mathcal{R}_4$), respectively, and $\bigcup_{i=0}^4 \mathcal{R}_i = \mathcal{X}$. Obviously, this partition method can be used to handle the problem of overlapped critical convex polygon regions.

Computing Solution for a Parameter Pair in $\mathcal{R}_i$ For each convex subregion $\mathcal{R}_i$, similar to the above Initialization, we need to find the solution for a parameter pair (ρ_i, ϱ_i) in $\mathcal{R}_i$, and compute the corresponding $\mathcal{CR}(\rho_i, \varrho_i)$, then partition $\mathcal{R}_i$ based on $\mathcal{CR}(\rho_i, \varrho_i) \cap \mathcal{R}_i$. Repeat the above steps until the full parameter space are partitioned with critical convex polygon regions (see Fig. 2c and 2d). Obviously, how to find the solution for a parameter pair in $\mathcal{R}_i$ is the key to compute the next critical convex polygon region. A simple strategy is using the SMO technology [Cai and Cherkassky, 2012] or other quadratic programming solvers similar to the Initialization. Instead, Theorem 3 allows us to compute α and $\widetilde{g}$ ($\overline{g}$) for a parameter pair in the subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ directly.

Theorem 3. *Supposing $\mathcal{X} \subseteq \mathbb{R}^2$ is a convex polygon region, $\mathcal{CR}(\lambda_0, \eta_0) \cap \mathcal{X} \overset{\text{def}}{=} \mathcal{R}_0$ or $\mathcal{CR}(C_+^0, C_-^0) \cap \mathcal{X} \overset{\text{def}}{=} \mathcal{R}_0$, $\mathcal{R}_0$ has the partition π, and $\{\mathcal{R}_0, \mathcal{R}_1, \cdots, \mathcal{R}_m\}$ is a partition of $\mathcal{X}$ as Theorem 2. $\forall i \in \{1, \cdots, m\}$, if $\mathcal{R}_i \neq \emptyset$, the i-th inequality $A_i \left[\rho \;\; \varrho\right]^T \leq b_i$ of $\mathcal{CR}$ only corresponds to the t-th sample of $\mathcal{S}$,*

1. *from the left part of (12), there will exist a subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ with the partition $\pi = (\mathcal{M} \setminus \{t\}, \mathcal{E}, \mathcal{R} \cup \{t\})$;*
2. *from the right part of (12), there will exist a subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ with the partition $\pi = (\mathcal{M} \setminus \{t\}, \mathcal{E} \cup \{t\}, \mathcal{R})$;*
3. *from (13), there will exist a subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ with the partition $\pi = (\mathcal{M} \cup \{t\}, \mathcal{E} \setminus \{t\}, \mathcal{R})$;*
4. *and from (14), there will exist a subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ with the partition $\pi = (\mathcal{M} \cup \{t\}, \mathcal{E}, \mathcal{R} \setminus \{t\})$.*

The partition π for the subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ is given by Theorem 3. Thus, we can update the inverse matrix R corresponding to the extended kernel matrix $\widetilde{Q}$ in time $\mathcal{O}(|\mathcal{M}|^2)$ as the method described in [Laskov *et al.*, 2006], and compute the linear relationships between $\Delta b'$ ($\Delta b''$), $\Delta\alpha_{\mathcal{M}}^T$, $\Delta\widetilde{g}$ ($\Delta\overline{g}$) and $[\Delta\rho \;\; \Delta\varrho]$ as (10)-(11). Further, $\alpha(\rho_i, \varrho_i)$ and $\widetilde{g}(\lambda_i, \eta_i)$ ($\overline{g}(C_+, C_-)$), where (ρ_i, ϱ_i) is a parameter pair in the subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$ with the partition π, can be computed by (12)-(14) directly.

Algorithm 1 $\mathcal{CR}$-BPSP ($\mathcal{CR}$s-based BPSP algorithm)

```
Input: $\alpha(\rho_0, \varrho_0)$, $\widetilde{g}(\lambda_0, \eta_0)$ ($\overline{g}(C_+^0, C_-^0)$), $\pi(\rho_0, \varrho_0)$, a convex polygon region $\mathcal{X}$ with $(\rho_0, \varrho_0) \in \mathcal{X}$.
Output: $\mathcal{P}$ (a partition of $\mathcal{X}$ in a nested set structure).
Detect $\mathcal{CR}(\rho_0, \varrho_0)$ according to (12)-(14).
Let $\mathcal{R}_0 := \mathcal{CR}(\rho_0, \varrho_0) \cap \mathcal{X}$, and $\mathcal{P} := \{\mathcal{R}_0\}$.
Partition the parameter space $\mathcal{X}$ with $\{\mathcal{R}_0, \mathcal{R}_1, \cdots, \mathcal{R}_m\}$ (c.f. Theorem 2).
while $i \leq m$ & $\mathcal{R}_i \neq \emptyset$ do
    Update $\pi$, $R$, for the subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$.
    Compute $\alpha$ and $\widetilde{g}$ ($\overline{g}$) for a parameter pair $(\rho_i, \varrho_i)$ in the subregion of $\mathcal{R}_i$ adjacent to $\mathcal{R}_0$.
    $\mathcal{P}_i$:=$\mathcal{CR}$-BPSP($\alpha(\rho_i, \varrho_i)$, $\widetilde{g}(\lambda_0, \eta_0)$ ($\overline{g}(C_+^0, C_-^0)$), $\pi(\rho_i, \varrho_i)$, $\mathcal{R}_i$). {$\mathcal{P}_i$ is the partition of $\mathcal{R}_i$.}
    Update $\mathcal{P} := \mathcal{P} \cup \{\mathcal{P}_i\}$.
    $i := i + 1$.
end while
```

4 BPSP using Invariant Regions

4.1 Invariant-Regions-Based BPSP Algorithm

Based on the above bi-parameter space partition, the decision function of CS-SVM can be obtained as $f(\lambda, \eta)(x) = \frac{1}{\lambda}\left(\sum_{j \in \mathcal{S}} \alpha_j(\lambda, \eta) y_j K(x_j, x) + b'(\lambda, \eta)\right)$ for all (λ, η) in $(0, 1] \times [0, 1]$, and $f(C_+, C_-)(x) = \sum_{j \in \mathcal{S}} \alpha_j(C_+, C_-) y_j K(x_j, x) + b''(C_+, C_-)$ for all (C_+, C_-) in the lower triangle region of $[0, 1] \times [0, 1]$. Given a validation set $\mathcal{V} = \{(\widetilde{x}_1, \widetilde{y}_1), \cdots, (\widetilde{x}_n, \widetilde{y}_n)\}$, and assuming $C(-, +)$ and $C(+, -)$ are the misclassification costs of false negative and false positive respectively (no costs for the true positive and the true negative), the empirical error on the validation set can be computed as $\mathcal{C}(\rho, \varrho) = \frac{1}{n}\sum_{i=1}^{n} C\left(sign\left(f(\rho, \varrho)(\widetilde{x}_i)\right), \widetilde{y}_i\right)$. To select the parameter pairs with the lowest empirical error, we need to investigate empirical errors for all parameter pairs.

Detecting the Invariant Convex Polygon Region According to the sign of $f(\widetilde{x}_i)$, the validation set $\mathcal{V}$ can be partitioned as:

$$\begin{aligned} \widetilde{\pi}(\rho, \varrho) &= \{\{i \in \mathcal{V} : f(\rho, \varrho)(\widetilde{x}_i)) \geq 0\}, \\ &\quad\ \{i \in \mathcal{V} : f(\rho, \varrho)(\widetilde{x}_i)) < 0\}\} \\ &\overset{\text{def}}{=} \{\mathcal{I}_+(\rho, \varrho), \mathcal{I}_-(\rho, \varrho)\} \qquad (15) \end{aligned}$$

Based on the partition, we have the invariant region $\mathcal{IR}(\rho_0, \varrho_0) = \{(\rho, \varrho) \in \mathcal{CR}(\rho_0, \varrho_0) : \widetilde{\pi}(\rho, \varrho) = \widetilde{\pi}(\rho_0, \varrho_0)\}$, in which the empirical error obviously remains unchanged. Theorem 4 shows that $\mathcal{IR}(\rho_0, \varrho_0)$ is also a convex polygon region. Thus, we can compute all empirical errors though finding invariant convex polygon regions in the two parameter spaces.

Theorem 4. *The sets $\mathcal{IR}(\rho_0, \varrho_0)$ is a convex set and its closure is a convex polygon region.*

$\forall(\lambda, \eta) \in \mathcal{CR}(\lambda_0, \eta_0)$, according to (10), we can get the linear relationship between $\Delta f(\widetilde{x}_i)$ and $[\Delta\lambda \;\; \Delta\eta]$ as follows:

$$\begin{aligned} \Delta f(\widetilde{x}_i) &= \sum_{j \in \mathcal{M}} y_j K(x_j, \widetilde{x}_i)\left(\beta_j^\lambda \Delta\lambda + \beta_j^\eta \Delta\eta\right) \\ &\quad + \left(\beta_{b'}^\lambda \Delta\lambda + \beta_{b'}^\eta \Delta\eta\right) + \sum_{j \in \mathcal{E}} K(x_j, \widetilde{x}_i)\Delta\eta \\ &\overset{\text{def}}{=} \widetilde{\gamma}_i^\lambda \Delta\lambda + \widetilde{\gamma}_i^\eta \Delta\eta \qquad (16) \end{aligned}$$

Combining (16) with the constraint of $\widetilde{\pi}(\lambda_0, \eta_0)$, we can get the following constraints:

$\forall i \in \mathcal{I}_+(\lambda_0, \eta_0)$:

$$f(\lambda_0, \eta_0)(\widetilde{x}_i) + \widetilde{\gamma}_i^\lambda(\lambda - \lambda_0) + \widetilde{\gamma}_i^\eta(\eta - \eta_0) \geq 0 \qquad (17)$$

$\forall i \in \mathcal{I}_-(\lambda_0, \eta_0)$:

$$f(\lambda_0, \eta_0)(\widetilde{x}_i) + \widetilde{\gamma}_i^\lambda(\lambda - \lambda_0) + \widetilde{\gamma}_i^\eta(\eta - \eta_0) < 0 \qquad (18)$$

Obviously, the closure of inequalities (17)-(18) is a convex polygon region, and the compact representation is denoted $\mathcal{IR}(\lambda_0, \eta_0)$. The same analysis can be extended to $\mathcal{IR}(C_+^0, C_-^0)$.

Partitioning Each $\mathcal{CR}$ with $\mathcal{IR}$s To find all invariant convex polygon regions in the whole parameter space, we use the strategy of divide and conquer (i.e., find all invariant convex polygon regions for each critical convex polygon region).

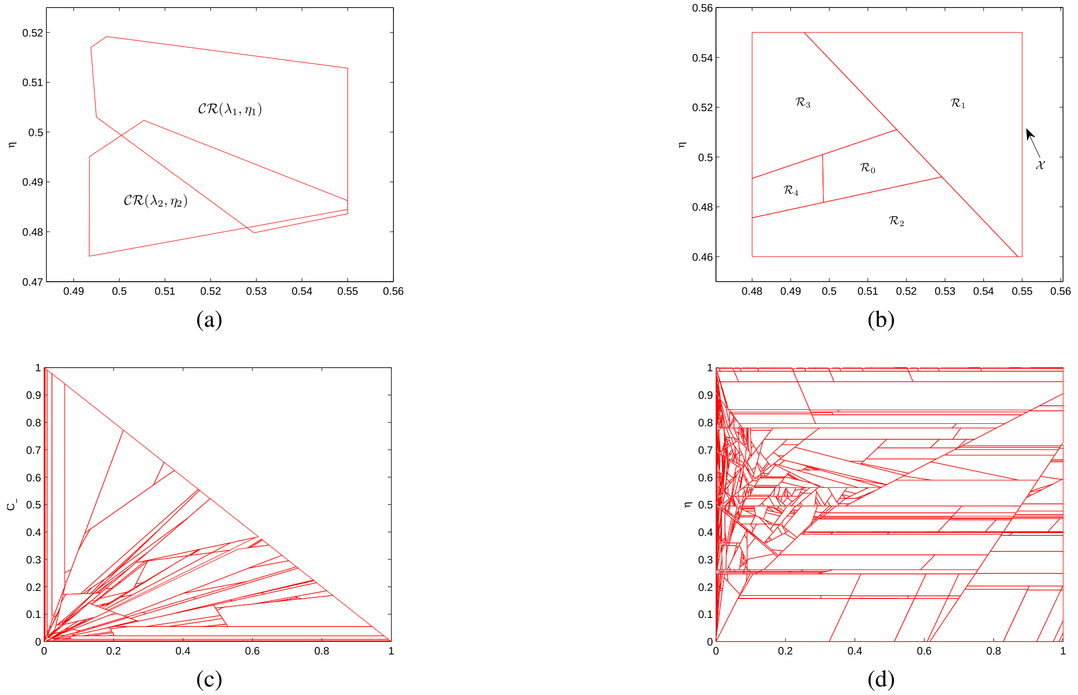


Figure 2: (a): Two overlapped $\mathcal{CR}$s ($\lambda_1 = \lambda_2 = 0.5$, $\eta_1 = 0.51$, and $\eta_2 = 0.49$). (b): Partitioning the parameter space $\mathcal{X}$ based on Theorem 2. (c): Partitioning the lower triangle region of $[0, 1] \times [0, 1]$ for (C_+, C_-) through $\mathcal{CR}$s. (d): Partitioning $(0, 1] \times [0, 1]$ for (λ, η) through $\mathcal{CR}$s.

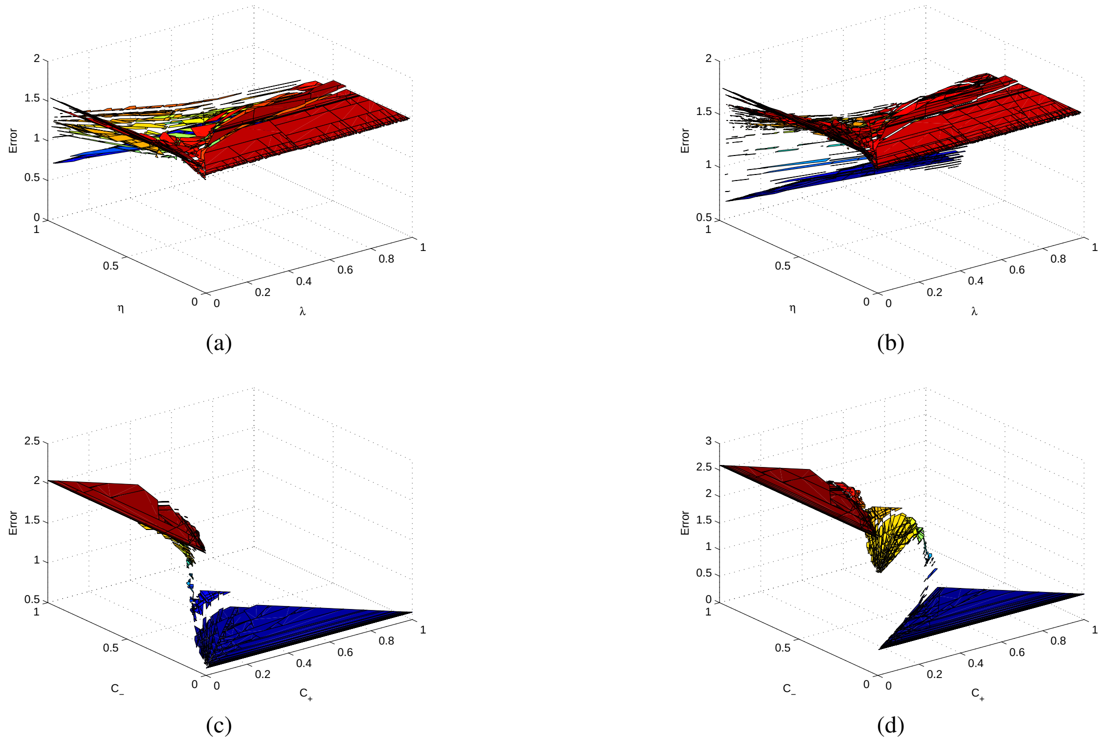


Figure 3: BPSP in 2-fold CV. (a)-(b): All parameter pairs of (λ, η) in $(0, 1] \times [0, 1]$. (c)-(d): All parameter pairs of (C_+, C_-) in the lower triangle region of $[0, 1] \times [0, 1]$. (a), (c): The results of the first fold. (b), (d): The results of 2-fold CV.

Thus, similar to Algorithm 1, a recursive procedure (called $\mathcal{IR}$-BPSP) can be designed to find all invariant convex polygon regions and compute the corresponding empirical errors for each critical convex polygon region. A nested set structure for the output of $\mathcal{IR}$-BPSP can be retained based on Theorem 2. The nested set structure will speed up finding the global optimal solution for K-fold CV in Section 4.2. Combining all results of the critical convex polygon regions based on the framework of Algorithm 1, we can obtain the empirical errors for all parameter pairs of (λ, η) in the region of $(0, 1] \times [0, 1]$ as shown in Fig. 3a, and the empirical errors for all parameter pairs of (C_+, C_-) in the lower triangle region of $[0, 1] \times [0, 1]$ as shown in Fig. 3c.

4.2 Computing the Superposition of K BPSPs

The validation set $\mathcal{V}$ is randomly partitioned into K equal size subsets. For each $k = 1, \cdots, K$, we fit the CS-SVM model with a parameter pair (ρ, ϱ) to the other $K - 1$ parts, which produces the decision function $f(\rho, \varrho)(x)$ and compute its empirical error in predicting the k part $\mathcal{C}^k(\rho, \varrho) = \frac{1}{|\mathcal{V}_k|} \sum_{i \in \mathcal{V}_k} C\left(sign\left(f^k(\rho, \varrho)(\widetilde{x}_i)\right), \widetilde{y}_i\right)$. This gives the CV error $\mathcal{CVC}(\rho, \varrho) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{C}^k(\rho, \varrho)$. The superposition of K invariant region partitions can be easily computed for selecting the best parameter pairs of (C_+, C_-) in $\mathbb{R}^+ \times \mathbb{R}^+$ (see Fig. 3b and 3d), which is omitted here.

5 Experiments

Design of Experiments We compare the generalization ability and runtime of BPSP with other three typical model selection methods of CS-SVM: (**1**) grid search (GS): a two-step grid search strategy is used for $2C$-SVM. The initial search is done on a 20×20 coarse grid linearly spaced in the region $\{(\log_2 C_+, \log_2 C_-)| - 9 \leq \log_2 C_+ \leq 10, -9 \leq \log_2 C_- \leq 10\}$, followed by a fine search on a 20×20 uniform grid linearly spaced by 0.1; (**2**) a hybrid method of one-parametric solution path searching on η and grid searching on λ (SP$_\eta$+GS$_\lambda$): λ is selected by a two-step grid search in the region $\{\log_2 \lambda| - 9 \leq \log_2 \lambda \leq 10\}$ with the granularity 1 and followed by 0.1; (**3**) a hybrid method of one-parametric solution path searching on λ and grid searching on η (SP$_\lambda$+GS$_\eta$): η is selected by a two-step grid search in the region $\{\eta | 0 \leq \eta \leq 1\}$ with the granularity 0.1 and followed by 0.01.

Implementation We implemented SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$, and our BPSP in MATLAB. To compare the run-time in the same platform, we did not directly modify the LIBSVM software package [Chang and Lin, 2011] as stated in [Davenport *et al.*, 2010], but implemented the SMO-type algorithm of $2C$-SVM in MATLAB. All experiments were performed on a 2.5-GHz Intel Core i5 machine with 8GB RAM and MATLAB 7.10 platform. $C(-, +)$ and $C(+, -)$ are the misclassification costs of false negative and false positive respectively. To investigate how the performance of an approach changes with different settings in misclassification cost, $C(-, +)$ was set to 2, 5, 10 for normal datasets of binary classification, and the class imbalance ratio for imbalanced datasets, respectively, while $C(+, -)$ was fixed at 1. Gaussian kernel $K(x_1, x_2) = \exp(-\kappa \|x_1 - x_2\|^2)$ was used in all the experiments with $\kappa \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, where the value of κ having the lowest CV error was adopted. For BPSP (or SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$), our implementation returns a center point from the region (or the line segment) with the minimum error.

Table 1: The results of 5-fold CV with GS, SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$ and BPSP (time was measured in minutes).

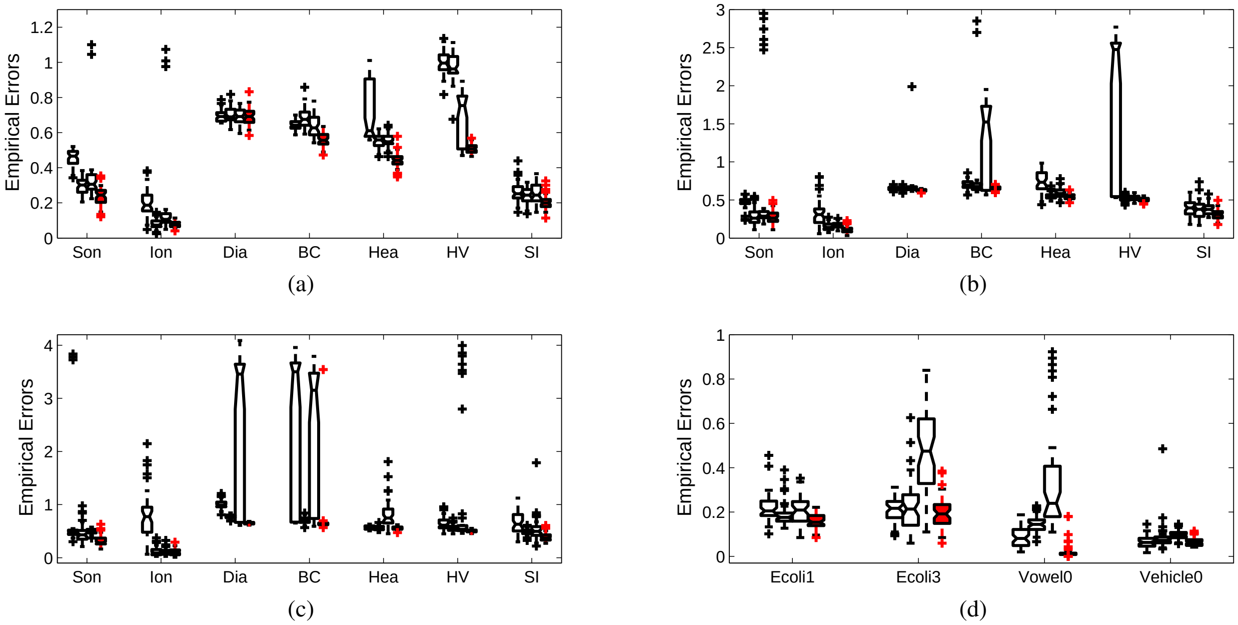
$C(+,-)$	Dataset	GS		SP$_\eta$+GS$_\lambda$		SP$_\lambda$+GS$_\eta$		BPSP	
		CV error	time	CV error	time	CV error	time	CV error	time
2	Son	0.4667	43	0.282	7.7	0.271	7.3	**0.2564**	**4.4**
	Ion	0.3623	73	0.0725	12.7	0.0857	13.3	**0.0435**	**9.7**
	Dia	0.6275	294	0.5948	9.5	0.606	10.2	**0.5752**	**5.5**
	BC	0.6593	229	0.6	9	0.611	9.82	**0.5642**	**7.1**
	Hea	0.52	59	0.464	9.2	0.478	9.1	**0.444**	**5.9**
	HV	0.463	176	0.45	5.3	0.462	5.8	**0.4417**	**4.9**
	SI	0.5017	86	0.2754	7.4	0.278	8.2	**0.2650**	**4.9**
5	Son	0.4872	51	0.3167	7	0.322	7.3	**0.3167**	**4.6**
	Ion	0.3768	65	0.1159	14	0.1324	16.3	**0.1159**	**10.3**
	Dia	0.6536	302	0.632	9.6	0.638	10.1	**0.632**	**5.9**
	BC	0.6741	227	0.6074	8	0.6222	8.3	**0.6074**	**6.7**
	Hea	0.537	57	0.463	6.8	0.485	8.5	**0.463**	**5.5**
	HV	0.6	164	0.55	5.6	0.493	7.2	**0.4417**	**5.2**
	SI	0.524	77	0.383	8.1	0.3795	8.5	**0.3562**	**5.6**
10	Son	0.564	46	0.4615	6.4	0.473	6.8	**0.4359**	**4.9**
	Ion	0.3823	77	0.2319	15.3	0.2425	16.1	**0.2319**	**9.9**
	Dia	0.6863	312	0.6601	9.2	0.672	9.8	**0.6601**	**5.6**
	BC	0.6815	219	0.6741	7.3	0.6741	7.2	**0.6626**	**6.9**
	Hea	0.556	69	0.556	5.6	0.562	5.9	**0.556**	**5.4**
	HV	0.5	169	0.5	4.9	0.5	6.3	**0.458**	**4.7**
	SI	0.536	81	0.4783	7.8	0.464	8.3	**0.4493**	**5.1**
ratio	Ecoli1	0.1722	65	0.117	12.2	0.124	13.1	**0.0833**	**8.8**
	Ecoli3	0.1905	76	0.0909	11.6	0.1102	12.3	**0.0595**	**9.3**
	Vowel0	0.1586	195	0.101	103	0.095	89	**0.0449**	**21**
	Vehicle0	0.472	262	0.1834	16d5	0.2092	134	**0.1024**	**26**

Datasets The sonar (Son), ionosphere (Ion), diabetes (Dia), breast cancer (BC), heart (Hea), and hill-valley (HV) datasets were obtained from the UCI benchmark repository [Bache and Lichman, 2013]. The spine image (SI) dataset collected by us is to diagnose degenerative disc disease depending on five image texture features quantified from magnetic resonance imaging. The dataset contains 350 records, where 157 are normal and 193 are abnormal. They are normal datasets for binary classification. Ecoli1, Ecoli3, Vowel0, and Vehicle0 are the imbalanced datasets from the KEEL-dataset repository[2]. Their class imbalance ratios are varying from 3.25 to 9.98.

We selected 30% from a dataset once as a validation set. The validation set was used with a 5-fold CV procedure to determine the optimal parameters. We then randomly partitioned each dataset into 75% for training and 25% for testing for many times. Each time, we removed the instances appearing in the validation set from the testing set to guarantee that the test set of each run is disjoint from the validation set.

Experimental Results The CV errors are presented in Table 1 for 5-fold CV of each method. It is easily observed that BPSP obtains lowest CV error for all datasets and settings of $C(-, +)$. This is reasonable because GS and SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$ are points-based and lines-based grid search method, respectively, however, BPSP is a regions-based method which covers all candidate values in the bi-parameter space, and give the best choices from them. Noted

[2] http://sci2s.ugr.es/keel/imbalanced.php.



(a) (b) (c) (d)

Figure 4: The results of cost sensitive errors on the test sets, over 50 trials. The grouped boxes represent the results of GS, SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$, and BPSP (red color), from left to right on different datasets. The notched-boxes have lines at the lower, median, and upper quartile values. The whiskers are lines extended from each end of the box to the most extreme data value within $1.5\times$IQR (Interquartile Range) of the box. Outliers are data with values beyond the ends of the whiskers, which are displayed by plus signs. (a): $C(-,+) = 2$. (b): $C(-,+) = 5$. (c): $C(-,+) = 10$. (d): $C(-,+) =$ ratio, for imbalanced learning.

that SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$ and BPSP can have the same CV error for some datasets, because both of them find the optimal on these datasets. BPSP always find an optimal parameter pair, and SP$_\eta$+GS$_\lambda$, SP$_\lambda$+GS$_\eta$ can also find an optimal sometimes. Based on the optimal parameters in Table 1, the empirical errors on each dataset in different methods over 50 trials are presented in Figure 4 as $C(-,+) = 2$, 5, 10, and ratio, respectively. The results show that BPSP has better generalization ability than GS, SP$_\eta$+GS$_\lambda$, and SP$_\lambda$+GS$_\eta$ generally. Especially, BPSP has the best stability, because it returns a center point from the optimal region with the minimum error.

The empirical running time (in minutes) for different algorithms on each dataset is also presented in Table 1, which is the average result on the seven different values of κ. It is easy to find that GS method has the longest running time. Because SP$_\eta$+GS$_\lambda$ and SP$_\lambda$+GS$_\eta$ searche a large number of parallel one-parametric solution paths, we find that BPSP has the less running time than SP$_\eta$+GS$_\lambda$ and SP$_\lambda$+GS$_\eta$.

6 Conclusion

We proposed a bi-parameter space partition algorithm for CS-SVM which can fit all solutions for each parameter pair (C_+, C_-). Based on the space partition, a K-fold cross validation algorithm was proposed which can find the global optimum parameter pair. The experiments indicate that our method has better generalization ability than various kinds of grid search methods, however, with less running time. In future work, we plan to extend this framework to a more general formulation which can cover bi-parametric learning models, and even multi-parametric learning models [Mukhopadhyay *et al.*, 2014].

Acknowledgments

This work was supported by the NSF of China (61232016 and 61202137) and the U.S. NSF (IIS-1115417).

References

[Avis and Fukuda, 1992] David Avis and Komei Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, 8(1):295–313, 1992.

[Bach *et al.*, 2006] Francis R Bach, David Heckerman, and Eric Horvitz. Considering cost asymmetry in learning classifiers. *The Journal of Machine Learning Research*, 7:1713–1741, 2006.

[Bache and Lichman, 2013] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[Borrelli, 2003] Francesco Borrelli. Constrained optimal control of linear and hybrid systems. 2003.

[Boyd and Vandenberghe, 2009] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[Cai and Cherkassky, 2012] Feng Cai and Vladimir Cherkassky. Generalized smo algorithm for svm-based multitask learning. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(6):997–1003, 2012.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[Cui *et al.*, 2012] Geng Cui, Man Leung Wong, and Xiang Wan. Cost-sensitive learning via priority sampling to improve the return on marketing and crm investment. *Journal of Management Information Systems*, 29(1):341–374, 2012.

[Davenport *et al.*, 2010] Mark A Davenport, Richard G Baraniuk, and Clayton D Scott. Tuning support vector machines for minimax and neyman-pearson classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1888–1898, 2010.

[Elkan, 2001] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Citeseer, 2001.

[Gu *et al.*, 2012] Bin Gu, Jian-Dong Wang, Guan-Sheng Zheng, and Yue-Cheng Yu. Regularization path for ν-support vector classification. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(5):800–811, May 2012.

[Gunter and Zhu, 2007] Lacey Gunter and Ji Zhu. Efficient computation and model selection for the support vector regression. *Neural Computation*, 19(6):1633–1655, 2007.

[Hastie *et al.*, 2004] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. In *Journal of Machine Learning Research*, pages 1391–1415, 2004.

[Laskov *et al.*, 2006] Pavel Laskov, Christian Gehl, Stefan Krüger, and Klaus-Robert Müller. Incremental support vector learning: Analysis, implementation and applications. *The Journal of Machine Learning Research*, 7:1909–1936, 2006.

[Mao *et al.*, 2014] Wentao Mao, Xiaoxia Mu, Yanbin Zheng, and Guirong Yan. Leave-one-out cross-validation-based model selection for multi-input multi-output support vector machine. *Neural Computing and Applications*, 24(2):441–451, 2014.

[Masnadi-Shirazi and Vasconcelos, 2010] Hamed Masnadi-Shirazi and Nuno Vasconcelos. Risk minimization, probability elicitation, and cost-sensitive svms. In *ICML*, pages 759–766, 2010.

[Mukhopadhyay *et al.*, 2014] A Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C.A Coello Coello. A survey of multiobjective evolutionary algorithms for data mining: Part i. *Evolutionary Computation, IEEE Transactions on*, 18(1):4–19, Feb 2014.

[Park *et al.*, 2011] Yoon-Joo Park, Se-Hak Chun, and Byung-Chun Kim. Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis. *Artificial intelligence in medicine*, 51(2):133–145, 2011.

[Rosset, 2009] Saharon Rosset. Bi-level path following for cross validated solution of kernel quantile regression. *The Journal of Machine Learning Research*, 10:2473–2505, 2009.

[Schölkopf and Smola, 2002] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[Shawe-Taylor, 1999] Grigoris Karakoulas John Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference*, volume 11, page 253. MIT Press, 1999.

[Sheng and Ling, 2006] Victor S Sheng and Charles X Ling. Thresholding for making classifiers cost-sensitive. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 476. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[Vapnik and Vapnik, 1998] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.

[Wang *et al.*, 2008] Gang Wang, Dit-Yan Yeung, and Frederick H Lochovsky. A new solution path algorithm in support vector regression. *Neural Networks, IEEE Transactions on*, 19(10):1753–1767, 2008.

[Yang and Ong, 2011] Jian-Bo Yang and Chong-Jin Ong. Determination of global minima of some common validation functions in support vector machine. *Neural Networks, IEEE Transactions on*, 22(4):654–659, 2011.

[Zhang and Zhou, 2010] Yin Zhang and Zhi-Hua Zhou. Cost-sensitive face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1758–1769, 2010.