Huberized Multiclass Support Vector Machine for Microarray Classification

LI Jun-Tao[1] JIA Ying-Min[1]

Abstract This paper proposes a new multiclass support vector machine (SVM) for simultaneous gene selection and microarray classification. Combining the huberized hinge loss function and the elastic net penalty, the proposed SVM can perform automatic gene selection and encourages a grouping effect. The coefficient paths of the proposed SVM are shown to be piecewise linear with respect to the single regularization parameter, based on which the solution path algorithm is developed with low computational complexity. Experiments performed on the leukemia data set are provided to verify the obtained results.

Key words Gene selection, grouping effect, microarray classification, solution path, support vector machine (SVM)

DOI 10.3724/SP.J.1004/2010.00399

The DNA microarray technology[1−2] is a powerful tool for biological and medical research. To classify microarray gene expression data, support vector machine (SVM) has attracted considerable attention[3−12]. It is well known that SVM and its various extensions have been successfully applied to two-class classification of microarray data[3−6]. In recent years, a few attempts have been made to generalize the SVM to multiclass problems[7−12]. In particular, multicategory support vector machine (MSVM)[8] that considers all of the classes at once has been developed. However, this method selects genes by using the marginal criterion[1], which tends to yield redundant genes. To perform joint classification and gene selection, L_1-norm MSVM[9] and sup-norm MSVM[10] have also been developed. Unfortunately, they cannot reveal the mutual information among genes.

The group lasso[13−14] has been developed to select relevant variables in groups for two-class problems. However, rare results on extending this method to multiclass problems have been reported since it is difficult to construct gene clusters in advance. When used as two-class classifiers, the elastic net penalized methods[4−5, 15] appear to perform well on microarray data and encourage a grouping effect. The key feature of these methods is the use of the elastic net penalty which not only retains the benefits of the L_1 norm penalty but also tends to generate similar coefficients for highly correlated variables. Taking into account the advantages of the elastic net penalty, this paper is devoted to extending it to multiclass problem and further developing a new MSVM. This is still a challenging problem due to the following two facts:

1) Gene selection of multiclass problems becomes more complex than the binary case. This is because that the MSVM requires to estimate multiple discriminating functions, among which each function has its own subset of important predictors. Moreover, gene selection is a necessary demand for microarray classification, especially grouped gene selection. Hence, the grouping effect for multiclass gene selection should be encouraged.

2) The normal solving algorithms for MSVM, such as quadratic programming[7−8] and linear programming[9−10], fail to work since multiple penalties are required simultaneously. Furthermore, the piecewise linear regularization path algorithms[16−17]used for parameter selection are not suitable due to multiple classes and multiple tuning parameters. Hence, new solving algorithm and parameter tuning method should be developed.

This paper attempts to deal with the aforementioned difficulties by developing new statistical learning tool. To this end, we first propose the huberized multiclass support vector machine (HMSVM). Second, the grouping effect of HMSVM described by using 2-norm is presented. Third, we prove that the coefficients of the HMSVM are piecewise linear with respect to the single regularization parameter and give their concrete form. Next, an efficient regularized solution algorithm is developed to compute the optimal coefficients. Finally, we apply our method to leukemia classification and achieve promising results.

1 Problem statement

Assume that the n training pairs $\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, n\}$ are independently and identically distributed according to an unknown probability distribution $P(\boldsymbol{x}, y)$. For microarray gene expression data with multiple cancer types, $\boldsymbol{x}_i$ represents the i-th sample and y_i represents the tumor type, which can be coded as $\{1, \cdots, K\}$. Let $\boldsymbol{f} = (f_1, \cdots, f_K)$ denote the decision function vector, where $k = 1, \cdots, K$. A popular multiclass classifier[10] can be defined as

$$\phi(\boldsymbol{x}) = \arg \max_{k=1,\cdots,K} f_k(\boldsymbol{x}) \tag{1}$$

The K-class classification problem is to learn the decision function vector and hence accurately predict the cancer type of a new sample.

There are some popular machine learning methods for the K-class classification problem, e.g., MSVM[8], 1-norm MSVM[9], and sup-norm MSVM[10]. However, all these methods cannot reveal the mutual information among genes. This paper aims to deal with this problem by developing a new MSVM. In the following, we present our notation. Similar to sup-norm MSVM[10], we use the linear decision functions $f_k(\boldsymbol{x}) = b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}$, $k = 1, 2, \cdots, K$ to build the classifier, and let $\boldsymbol{b} = (b_1, \cdots, b_K)^{\mathrm{T}}$ denote the bias vector, w denote the coefficient matrix, $X = (\boldsymbol{x}_1; \boldsymbol{x}_2; \cdots; \boldsymbol{x}_n) = \left(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \cdots, \boldsymbol{x}_{(p)}\right)$ denote the model matrix, where $\boldsymbol{x}_{(j)} = (x_{1j}, \cdots, x_{nj})^{\mathrm{T}}$, $j = 1, \cdots, p$. We use $\boldsymbol{w}_k = (w_{k1}, \cdots, w_{kp})^{\mathrm{T}}$ and $\boldsymbol{w}_{(j)} = (w_{1j}, \cdots, w_{Kj})^{\mathrm{T}}$ to represent the k-th row vector and the j-th column vector of w, respectively.

Manuscript received January 6, 2009; accepted March 18, 2009
Supported by National Basic Research Program of China (973 Program) (2005CB321902), National Natural Science Foundation of China (90916024, 60727002, 60774003), the Ph. D. Programs Foundation of Ministry of Education of China (20030006003), and the Commission on Science, Technology, and Industry for National Defense (A2120061303)
1. The Seventh Research Division, Beihang University, Beijing 100191, P. R. China

2 Main results

2.1 Huberized multiclass support vector machine

The elastic net penalty has been successfully applied to binary classification problem[4−5,15]. Taking into account the advantages of the elastic net penalty, we extend it to K-class classification problem, i.e.,

$$J(\boldsymbol{b}, w) = \lambda_2 \sum_{k=1}^{K}\sum_{j=1}^{p} w_{kj}^2 + \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p} |w_{kj}| \tag{2}$$

Similar to binary hybrid huberized SVM[4], we substitute

$$L_{ki} = \begin{cases} 0, & \text{if } b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i < -1 \\ 1 + b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i - \dfrac{\delta}{2}, & \text{if } b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i \geq -1 + \delta \\ \dfrac{(1 + b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i)^2}{2\delta}, & \text{otherwise} \end{cases}$$

for hinge loss function $[b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i + 1]_+$, where L_{ki} is the abbreviation of $L_{ki}(b_k, \boldsymbol{w}_k, \boldsymbol{x}_i)$. Combining the elastic net penalty (2) with the huberized hinge loss function L_{ki}, we propose the following HMSVM

$$\arg\min_{\boldsymbol{b}, w} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik} L_{ki} + \lambda_2 \sum_{k=1}^{K}\sum_{j=1}^{p} w_{kj}^2 + \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p} |w_{kj}| \tag{3}$$

subjected to the sum-to-zero constraint[4]

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{b} = 0, \qquad \mathbf{1}^{\mathrm{T}}\boldsymbol{w}_{(j)} = 0 \tag{4}$$

where λ_1, λ_2 are regularization parameters, $a_{ik} = I(y_i \neq k)$, and $I(\cdot)$ represents the indicator function. Substituting (4) into (3), we have the following unconstrained convex optimization problem:

$$\arg\min_{\boldsymbol{b}_-, w_-} \bar{L}(\lambda_1, \lambda_2, \boldsymbol{b}_-, w_-) \tag{5}$$

where $\boldsymbol{b}_-$, w_- denote the vector and matrix formed by the first $K-1$ rows of $\boldsymbol{b}$ and w, and $\bar{L}$ is defined as

$$\begin{aligned} \bar{L} = \frac{1}{n}\Bigg(&\sum_{i=1}^{n} a_{iK} L_K\left(-\sum_{k'=1}^{K-1} b_{k'}, -\sum_{k'=1}^{K-1}\boldsymbol{w}_{k'}, \boldsymbol{x}_i\right) + \\ &\sum_{i=1}^{n}\sum_{k=1}^{K-1} a_{ik} L_k\Bigg) + \lambda_2\left(\sum_{k=1}^{K-1}\sum_{j=1}^{p} w_{kj}^2 + \sum_{j=1}^{p}\Big(\sum_{k'=1}^{K-1} w_{k'j}\Big)^2\right) + \\ &\lambda_1\left(\sum_{k=1}^{K-1}\sum_{j=1}^{p}|w_{kj}| + \sum_{j=1}^{p}\left|\sum_{k'=1}^{K-1} w_{k'j}\right|\right) \end{aligned} \tag{6}$$

Define each region of the k-th classifier as $\mathcal{L}_k = \{i : b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i < -1\}$, $\mathcal{E}_k = \{i : -1 \leq b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i < -1 + \delta\}$, $\mathcal{R}_k = \{i : b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i \geq -1 + \delta\}$ for $k = 1, \cdots, K$, define the indices for non-zero w_{kj} as $\mathcal{A}_k = \{j : w_{kj} \neq 0,\ j = 1, 2, \cdots, p\}$ for $k = 1, \cdots, K-1$, and define the indices for non-zero $\sum_{k'=1}^{K-1} w_{k'j}$ as $\bar{\mathcal{A}} = \{j : \sum_{k'=1}^{K-1} w_{k'j} \neq 0,\ j = 1, 2, \cdots, p\}$. If we continuously decrease λ_1 or λ_2 or both of them, some of the sets of $\mathcal{L}_k$, $\mathcal{E}_k$, $\mathcal{R}_k$, $\mathcal{A}_k \bigcap \bar{\mathcal{A}}$ will change. We call this an event, and four types of events may occur: 1) A training point reaches the boundary between $\mathcal{L}_k$ and $\mathcal{E}_k$; 2) A training point reaches the boundary between $\mathcal{R}_k$ and $\mathcal{E}_k$; 3) An index j leaves $\mathcal{A}_k \bigcap \bar{\mathcal{A}}$; 4) An index j joins $\mathcal{A}_k \bigcap \bar{\mathcal{A}}$.

2.2 The grouping effect

Since the correlations between genes sharing the same biological "pathway" can be high, the correlated and relevant genes should be selected or removed together. From the statistical point of view, this can be described as a grouping effect. In the following, we prove that the HMSVM can encourage a grouping effect described by using 2-norm.

Theorem 1. Let $\hat{\boldsymbol{b}}$ and $\hat{w}$ denote the optimal solution of the huberized multiclass support machine. If $\boldsymbol{x}_{(m)}$ and $\boldsymbol{x}_{(l)}$ are normalized, then

$$\left\|\hat{\boldsymbol{w}}_{(m)} - \hat{\boldsymbol{w}}_{(l)}\right\|_2 \leq \frac{\sqrt{K}}{\sqrt{n}\lambda_2}\sqrt{2(1-\rho)} \tag{7}$$

where $\rho = \boldsymbol{x}_{(m)}^{\mathrm{T}}\boldsymbol{x}_{(l)} = \sum_{i=1}^{n} x_{im}x_{il}$.

Proof. From the sum-to-zero constraint, it is easy to know that $\hat{\boldsymbol{b}}_-$ and $\hat{w}_-$ formed by the first $K-1$ rows of $\hat{\boldsymbol{b}}$ and $\hat{w}$ will be the optimal solution of unconstrained convex optimization problem (5). For $k = 1, 2, \cdots, K-1$, $j' = 1, 2, \cdots, p$ and any given $1 \leq m, l \leq p$, we construct the following vector and matrix

$$\boldsymbol{b}_-^* = \hat{\boldsymbol{b}}_-$$

$$w_{kj'}^* = \begin{cases} \dfrac{1}{2}(\hat{w}_{km} + \hat{w}_{kl}), & \text{if } j' = m \text{ or } j' = l \\ \hat{w}_{kj'}, & \text{otherwise} \end{cases}$$

By the definition of $\hat{\boldsymbol{b}}_-$, $\hat{w}_-$, $\boldsymbol{b}_-^*$, and w_-^*, we have

$$0 \leq \bar{L}(\lambda_1, \lambda_2, \boldsymbol{b}_-^*, w_-^*) - \bar{L}(\lambda_1, \lambda_2, \hat{\boldsymbol{b}}_-, \hat{w}_-) \tag{8}$$

Note that $|L_{ki}(b_k^*, \boldsymbol{w}_k^*, \boldsymbol{x}_i) - L_{ki}(\hat{b}_k, \hat{\boldsymbol{w}}_k, \boldsymbol{x}_i)| \leq |(\boldsymbol{w}_k^* - \hat{\boldsymbol{w}}_k)^{\mathrm{T}}\boldsymbol{x}_i|$, $|\sum_{k=1}^{K-1}(\hat{w}_{km} - \hat{w}_{kl})(x_{im} - x_{il})| = |(\hat{w}_{Km} - \hat{w}_{Kl})(x_{im} - x_{il})|$. Hence,

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1} a_{ik}\left[L_{ki}(b_k^*, \boldsymbol{w}_k^*, \boldsymbol{x}_i) - L_{ki}(\hat{b}_k, \hat{\boldsymbol{w}}_k, \boldsymbol{x}_i)\right] + \\ &\frac{1}{n}\sum_{i=1}^{n} a_{iK}\left[L_{Ki}\left(-\sum_{k'=1}^{K-1} b_{k'}^*, -\sum_{k'=1}^{K-1}\boldsymbol{w}_{k'}^*, \boldsymbol{x}_i\right) - \right. \\ &\left. L_{Ki}\left(-\sum_{k'=1}^{K-1}\hat{b}_{k'}, -\sum_{k'=1}^{K-1}\hat{\boldsymbol{w}}_{k'}, \boldsymbol{x}_i\right)\right] \leq \\ &\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|(\boldsymbol{w}_k^* - \hat{\boldsymbol{w}}_k)^{\mathrm{T}}\boldsymbol{x}_i\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\sum_{k=1}^{K-1}(\boldsymbol{w}_k^* - \hat{\boldsymbol{w}}_k)^{\mathrm{T}}\boldsymbol{x}_i\right| = \\ &\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}|(\hat{w}_{km} - \hat{w}_{kl})(x_{im} - x_{il})| + \\ &\frac{1}{2n}\sum_{i=1}^{n}\left|\sum_{k=1}^{K-1}(\hat{w}_{km} - \hat{w}_{kl})(x_{im} - x_{il})\right| = \\ &\frac{1}{2n}\sum_{i=1}^{n}|x_{im} - x_{il}| \cdot \sum_{k=1}^{K}|\hat{w}_{km} - \hat{w}_{kl}| = \\ &\frac{1}{2n}\left\|\boldsymbol{x}_{(m)} - \boldsymbol{x}_{(l)}\right\|_1 \cdot \left\|\hat{\boldsymbol{w}}_{(m)} - \hat{\boldsymbol{w}}_{(l)}\right\|_1 \end{aligned} \tag{9}$$

$$\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left(|w^*_{kj}|-|\hat{w}_{kj}|\right)+\sum_{j=1}^{p}\left(\left|\sum_{k'=1}^{K-1}w^*_{k'j}\right|-\left|\sum_{k'=1}^{K-1}\hat{w}_{k'j}\right|\right)=$$
$$\sum_{k=1}^{K-1}\left(2\left|\frac{\hat{w}_{k'm}+\hat{w}_{k'l}}{2}\right|-|\hat{w}_{km}|-|\hat{w}_{kl}|\right)+$$
$$2\left|\sum_{k'=1}^{K-1}\frac{\hat{w}_{k'm}+\hat{w}_{k'l}}{2}\right|-\left|\sum_{k'=1}^{K-1}\hat{w}_{k'm}\right|-\left|\sum_{k'=1}^{K-1}\hat{w}_{k'l}\right|\le 0 \tag{10}$$

$$\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left(w^{*2}_{kj}-\hat{w}^2_{kj}\right)+\sum_{j=1}^{p}\left[(\sum_{k'=1}^{K-1}w^*_{k'j})^2-(\sum_{k'=1}^{K-1}\hat{w}_{k'j})^2\right]=$$
$$-\frac{1}{2}\sum_{k=1}^{K-1}(\hat{w}_{km}-\hat{w}_{kl})^2-\frac{1}{2}\left(\sum_{k=1}^{K-1}(\hat{w}_{km}-\hat{w}_{kl})\right)^2=$$
$$-\frac{1}{2}\left\|\hat{\boldsymbol{w}}_{(m)}-\hat{\boldsymbol{w}}_{(l)}\right\|_2^2 \tag{11}$$

From (8) $\sim$ (11), we have

$$0\le\frac{1}{n\lambda_2}\left\|\boldsymbol{x}_{(m)}-\boldsymbol{x}_{(l)}\right\|_1\cdot\left\|\hat{\boldsymbol{w}}_{(m)}-\hat{\boldsymbol{w}}_{(l)}\right\|_1-$$
$$\left\|\hat{\boldsymbol{w}}_{(m)}-\hat{\boldsymbol{w}}_{(l)}\right\|_2^2\le\frac{\sqrt{K}}{n\lambda_2}\left\|\boldsymbol{x}_{(m)}-\boldsymbol{x}_{(l)}\right\|_1\times \tag{12}$$
$$\left\|\hat{\boldsymbol{w}}_{(m)}-\hat{\boldsymbol{w}}_{(l)}\right\|_2-\left\|\hat{\boldsymbol{w}}_{(m)}-\hat{\boldsymbol{w}}_{(l)}\right\|_2^2$$

which is equivalent to

$$\left\|\hat{\boldsymbol{w}}_{(m)}-\hat{\boldsymbol{w}}_{(l)}\right\|_2\le\frac{\sqrt{K}}{n\lambda_2}\sum_{i=1}^{n}|x_{im}-x_{il}| \tag{13}$$

If $\boldsymbol{x}_{(m)}$ and $\boldsymbol{x}_{(l)}$ are normalized, then we have

$$\frac{\sqrt{K}}{n\lambda_2}\left\|\boldsymbol{x}^{\mathrm{T}}_{(m)}-\boldsymbol{x}^{\mathrm{T}}_{(l)}\right\|_1\le\frac{\sqrt{K}}{n\lambda_2}\cdot\sqrt{n}\left\|\boldsymbol{x}_{(m)}-\boldsymbol{x}_{(l)}\right\|_2=$$
$$\frac{\sqrt{K}}{\sqrt{n}\lambda_2}\sqrt{2-2\boldsymbol{x}^{\mathrm{T}}_{(m)}\boldsymbol{x}_{(l)}}=\frac{\sqrt{K}}{\sqrt{n}\lambda_2}\sqrt{2(1-\rho)} \tag{14}$$

Substituting (14) into (13) yields (7). □

For the highly correlated predictors $\boldsymbol{x}_{(m)}$ and $\boldsymbol{x}_{(l)}$ ($\rho=1$), HMSVM tends to assign the same coefficient vectors to them. This means that the highly correlated genes tend to be selected or removed together. According to the terminology of statistical learning[15], we claim that HMSVM exhibits a grouping effect. For $K=2$, it can be easily obtained that

$$\left|\hat{w}_{(1m)}-\hat{w}_{(1l)}\right|\le\frac{1}{\sqrt{n}\lambda_2}\sqrt{2(1-\rho)}$$

This coincides with the known results of binary SVMs[4−5].

2.3 The piecewise linear regularization solution

Similar to the hybrid huberized SVM[4], we continuously decrease λ_1 and use the superscript l to index the sets immediately after the l-th event has occurred. Let b^l_k, w^l_{kj}, λ^l_1, λ^l_2 be the values of these parameters at the point of entry, m_k be the cardinality of set $\mathcal{A}^l_k$. For $\lambda^l_1\ge\lambda_1>\lambda^{l+1}_1$, let

$$\lambda^l_2=\beta-\frac{\beta-\alpha}{\ln(\mathrm{e}+\lambda^l_1)} \tag{15}$$

where $0<\alpha<\beta$ are given constants.

Let $\mathbf{1}_{K-1}$ denote the $(K-1)$-th-order matrix whose elements are all ones, and I_p denote the p-th-order identity matrix. Let $\tilde{\boldsymbol{b}}_-=(\hat{b}_1-\hat{b}^l_1,\hat{b}_2-\hat{b}^l_2,\cdots,\hat{b}_{K-1}-\hat{b}^l_{K-1})^{\mathrm{T}}$, $\tilde{\boldsymbol{w}}_k=(\hat{w}_{kj_1}-\hat{w}^l_{kj_1},\hat{w}_{kj_2}-\hat{w}^l_{kj_2},\cdots,\hat{w}_{kj_{m_k}}-\hat{w}^l_{kj_{m_k}})^{\mathrm{T}}$, $\tilde{\boldsymbol{w}}_-=(\tilde{\boldsymbol{w}}^{\mathrm{T}}_1,\tilde{\boldsymbol{w}}^{\mathrm{T}}_2,\cdots,\tilde{\boldsymbol{w}}^{\mathrm{T}}_{K-1})^{\mathrm{T}}$, $A_{11}=\sum_{i\in\mathcal{E}^l_K}a_{iK}\mathbf{1}_{K-1}+\mathrm{diag}\{\sum_{i\in\mathcal{E}^l_1}a_{i1},\cdots,\sum_{i\in\mathcal{E}^l_{K-1}}a_{iK-1}\}$, $A_{22_1}=(\tilde{A}_{kk'})_{(m_1+m_2+\cdots+m_{K-1})\times(m_1+m_2+\cdots+m_{K-1})}+\mathrm{diag}\{\tilde{\tilde{A}}_{11},\tilde{\tilde{A}}_{22},\cdots,\tilde{\tilde{A}}_{K-1K-1}\}$,

$$\bar{A}_{22_2}=\begin{bmatrix}2I_p & I_p & \cdots & I_p\\ I_p & 2I_p & \cdots & I_p\\ \vdots & \vdots & \ddots & \vdots\\ I_p & I_p & \cdots & 2I_p\end{bmatrix}$$

$$A_{12}=\begin{bmatrix}\bar{X}^{\mathrm{T}}_1+\bar{\bar{X}}^{\mathrm{T}}_1 & \bar{X}^{\mathrm{T}}_2 & \cdots & \bar{X}^{\mathrm{T}}_{K-1}\\ \bar{X}^{\mathrm{T}}_1 & \bar{X}^{\mathrm{T}}_2+\bar{\bar{X}}^{\mathrm{T}}_2 & \cdots & \bar{X}^{\mathrm{T}}_{K-1}\\ \vdots & \vdots & \ddots & \vdots\\ \bar{X}^{\mathrm{T}}_1 & \bar{X}^{\mathrm{T}}_2 & \cdots & \bar{X}^{\mathrm{T}}_{K-1}+\bar{\bar{X}}^{\mathrm{T}}_{K-1}\end{bmatrix}$$

where $\bar{\bar{\boldsymbol{X}}}^{\mathrm{T}}_k=(\sum_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{k1}},\cdots,\sum_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{km_k}})$, $\bar{\boldsymbol{X}}^{\mathrm{T}}_k=(\sum_{i\in\mathcal{E}^l_K}a_{iK}x_{ij_{k1}},\cdots,\sum_{i\in\mathcal{E}^l_K}a_{iK}x_{ij_{km_k}})$, $\tilde{\boldsymbol{X}}^{\mathrm{T}}(j,\ k')=(\sum_{i\in\mathcal{E}^l_K}a_{iK}x_{ij_{k1}},\cdots,\sum_{i\in\mathcal{E}^l_K}a_{iK}x_{ij}x_{ij_{k'1}},\cdots,\sum_{i\in\mathcal{E}^l_K}a_{iK}x_{ij}x_{ij_{k'm_{k'}}})$,

$$\tilde{\tilde{A}}_{kk}=\begin{bmatrix}\sum\limits_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{k1}}x_{ij_{k1}} & \cdots & \sum\limits_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{k1}}x_{ij_{km_k}}\\ \sum\limits_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{k2}}x_{ij_{k1}} & \cdots & \sum\limits_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{k2}}x_{ij_{km_k}}\\ \vdots & \ddots & \vdots\\ \sum\limits_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{km_k}}x_{ij_{k1}} & \cdots & \sum\limits_{i\in\mathcal{E}^l_k}a_{ik}x_{ij_{km_k}}x_{ij_{km_k}}\end{bmatrix}$$

and $m_k\times m'_k$ submatrix of $\tilde{A}$

$$\tilde{A}_{kk'}=\begin{bmatrix}\tilde{\boldsymbol{X}}^{\mathrm{T}}(j=j_{k1},k'=k')\\ \tilde{\boldsymbol{X}}^{\mathrm{T}}(j=j_{k2},k'=k')\\ \vdots\\ \tilde{\boldsymbol{X}}^{\mathrm{T}}(j=j_{km_k},k'=k')\end{bmatrix}$$

for $k=1,\cdots,K-1$, $k'=1,\cdots,K-1$. Let $A_{22}=\frac{1}{n\delta}A_{22_1}+2\lambda_2A_{22_2}$, where A_{22_2} is constructed by orderly selecting the j_{k1}-th, j_{k2}-th, $\cdots$, j_{km_k}-th rows and columns of the k-th p rows and columns in $\bar{A}_{22_2}$.

Theorem 2. If the regularization parameters satisfy (15), then the coefficient paths of HMSVM are piecewise linear with respect to the regularization parameter λ_1, i.e.,

$$\begin{cases}\hat{b}_k=\hat{b}^l_k+(\lambda_1-\lambda^l_1)\bar{a}_{k0}\\ \hat{w}_{kj}=\hat{w}^l_{kj}+(\lambda_1-\lambda^l_1)\bar{a}_{kj},\quad\text{for } j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l\end{cases} \tag{16}$$

hold for $\lambda^l_1\ge\lambda_1>\lambda^{l+1}_1$, $k=1,\cdots,K-1$, and $j=1,\cdots,p$, where $\bar{a}_{k0}$ and $\bar{a}_{1j}$ are respectively the k-th and $(K+j)$-th elements of vector $A^{-1}_l1^\alpha$, $\bar{a}_{kj}$ is the $(K-1+m_1+\cdots+m_{k-1}+k+j)$-th elements of vector $A^{-1}_l1^\alpha$ for $k\ge2$, and A_l and 1^α are defined as

$$A_l=\begin{bmatrix}A_{11} & A_{12}\\ A^{\mathrm{T}}_{12} & A_{22}\end{bmatrix}$$

$$1^\alpha=\begin{bmatrix}0_{K-1}\\ \mathrm{sgn}((A_{22_2}-I)\hat{\tilde{\boldsymbol{w}}}^l_-)+\mathrm{sgn}(\hat{\tilde{\boldsymbol{w}}}^l_-)\end{bmatrix}$$

Proof. Note that (5) is an unconstrained convex optimization problem. Hence, for $b_k \neq 0, w_{kj} \neq 0, \sum_{k=1}^{K-1} b_k \neq 0$, and $\sum_{k=1}^{K-1} w_{kj} \neq 0$, we have

$$\begin{cases} \dfrac{\partial \bar{L}(\lambda_1, \lambda_2, \boldsymbol{b}_-, w_-)}{\partial b_k}\Big|_{\boldsymbol{b}_- = \hat{\boldsymbol{b}}_-, w_- = \hat{w}_-} = 0 \\ \dfrac{\partial \bar{L}(\lambda_1, \lambda_2, \boldsymbol{b}_-, w_-)}{\partial w_{kj}}\Big|_{\boldsymbol{b}_- = \hat{\boldsymbol{b}}_-, w_- = \hat{w}_-} = 0 \end{cases} \tag{17}$$

where $k = 1, \cdots, K-1$, $j = 1, \cdots, p$. Note that sets $\mathcal{E}_k^l$, $\mathcal{R}_k^l$, $\mathcal{L}_k^l$, $\mathcal{A}_k^l$, $\bar{\mathcal{A}}$ and the regularization parameter λ_2 will not change for $\lambda_1^l \geq \lambda_1 > \lambda_1^{l+1}$. Hence, we have

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} \frac{1 + \hat{b}_k + \hat{\boldsymbol{w}}_k^{\mathrm{T}} \boldsymbol{x}_i}{n\delta} + \\ &\quad \sum_{i \in \mathcal{E}_K^l} \frac{a_{iK}}{n\delta} \left(-1 + \sum_{k'=1}^{K-1} (\hat{b}_{k'} + \hat{\boldsymbol{w}}_{k'}^{\mathrm{T}} \boldsymbol{x}_i) \right) + \\ &\quad \frac{1}{n} \left(\sum_{i \in \mathcal{R}_k^l} a_{ik} - \sum_{i \in \mathcal{R}_K^l} a_{iK} \right) = 0 \end{aligned} \tag{18}$$

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} \frac{1 + \hat{b}_k + \hat{\boldsymbol{w}}_k^{\mathrm{T}} \boldsymbol{x}_i}{n\delta} + 2\lambda_2^l \left(\hat{w}_{kj} + \sum_{k'=1}^{K-1} \hat{w}_{k'j} \right) + \\ &\quad \sum_{i \in \mathcal{E}_K^l} \frac{a_{iK} x_{ij}}{n\delta} \left(-1 + \sum_{k'=1}^{K-1} (\hat{b}_{k'} + \hat{\boldsymbol{w}}_{k'}^{\mathrm{T}} \boldsymbol{x}_i) \right) + \\ &\quad \frac{1}{n} \left(\sum_{i \in \mathcal{R}_k^l} a_{ik} x_{ij} - \sum_{i \in \mathcal{R}_K^l} a_{iK} x_{ij} \right) + \\ &\quad \lambda_1 \left(\operatorname{sgn}(\hat{w}_{kj}) + \operatorname{sgn} \left(\sum_{k'=1}^{K-1} \hat{w}_{k'j} \right) \right) = 0 \end{aligned} \tag{19}$$

where $k = 1, \cdots, K-1$, $j \in \mathcal{A}_k^l \bigcap \bar{\mathcal{A}}^l$. Analogously, for $\lambda_1 = \lambda_1^l$, we also have

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} \frac{1 + \hat{b}_k^l + \hat{\boldsymbol{w}}_k^{l\mathrm{T}} \boldsymbol{x}_i}{n\delta} + \frac{1}{n} \left(\sum_{i \in \mathcal{R}_k^l} a_{ik} - \sum_{i \in \mathcal{R}_K^l} a_{iK} \right) + \\ &\quad \sum_{i \in \mathcal{E}_K^l} \frac{a_{iK}}{n\delta} \left(-1 + \sum_{k'=1}^{K-1} \left(\hat{b}_{k'}^l + \hat{\boldsymbol{w}}_{k'}^{l\mathrm{T}} \boldsymbol{x}_i \right) \right) = 0 \end{aligned} \tag{20}$$

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} \frac{1 + \hat{b}_k^l + \hat{\boldsymbol{w}}_k^{l\mathrm{T}} \boldsymbol{x}_i}{n\delta} + 2\lambda_2^l \left(\hat{w}_{kj}^l + \sum_{k'=1}^{K-1} \hat{w}_{k'j}^l \right) + \\ &\quad \sum_{i \in \mathcal{E}_K^l} \frac{a_{iK} x_{ij}}{n\delta} \left(-1 + \sum_{k'=1}^{K-1} (\hat{b}_{k'}^l + \hat{\boldsymbol{w}}_{k'}^{l\mathrm{T}} \boldsymbol{x}_i) \right) + \\ &\quad \frac{1}{n} \left(\sum_{i \in \mathcal{R}_k^l} a_{ik} x_{ij} - \sum_{i \in \mathcal{R}_K^l} a_{iK} x_{ij} \right) + \\ &\quad \lambda_1^l \left(\operatorname{sgn}(\hat{w}_{kj}^l) + \operatorname{sgn} \left(\sum_{k'=1}^{K-1} \hat{w}_{k'j}^l \right) \right) = 0 \end{aligned} \tag{21}$$

Subtracting (20) from (18) gives

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} (\hat{b}_k - \hat{b}_k^l + (\hat{\boldsymbol{w}}_k - \hat{\boldsymbol{w}}_k^l)^{\mathrm{T}} \boldsymbol{x}_i) + \\ &\quad \sum_{i \in \mathcal{E}_K^l} a_{iK} \sum_{k'=1}^{K-1} [\hat{b}_{k'} - \hat{b}_{k'}^l + (\hat{\boldsymbol{w}}_{k'} - \hat{\boldsymbol{w}}_{k'}^l)^{\mathrm{T}} \boldsymbol{x}_i] = 0 \end{aligned} \tag{22}$$

Note that $\operatorname{sgn}(\hat{w}_{kj}) = \operatorname{sgn}(\hat{w}_{kj}^l)$ and $\operatorname{sgn}(\sum_{k'=1}^{K-1} \hat{w}_{k'j}) = \operatorname{sgn}(\sum_{k'=1}^{K-1} \hat{w}_{k'j}^l)$. Subtracting (21) from (19) gives

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} \frac{\hat{b}_k - \hat{b}_k^l + (\hat{\boldsymbol{w}}_k - \hat{\boldsymbol{w}}_k^l)^{\mathrm{T}} \boldsymbol{x}_i}{n\delta} + \\ &\quad \sum_{i \in \mathcal{E}_K^l} a_{iK} x_{ij} \sum_{k'=1}^{K-1} \frac{\hat{b}_{k'} - \hat{b}_{k'}^l + (\hat{\boldsymbol{w}}_{k'} - \hat{\boldsymbol{w}}_{k'}^l)^{\mathrm{T}} \boldsymbol{x}_i}{n\delta} + \\ &\quad 2\lambda_2^l (\hat{w}_{kj} - \hat{w}_{kj}^l + \sum_{k'=1}^{K-1} (\hat{w}_{k'j} - \hat{w}_{k'j}^l)) + \\ &\quad (\lambda_1 - \lambda_1^l) \left(\operatorname{sgn}(\hat{w}_{kj}^l) + \operatorname{sgn} \left(\sum_{k'=1}^{K-1} \hat{w}_{k'j}^l \right) \right) = 0 \end{aligned} \tag{23}$$

Note that

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} (\hat{b}_k - \hat{b}_k^l) + \sum_{i \in \mathcal{E}_K^l} a_{iK} \sum_{k'=1}^{K-1} (\hat{b}_{k'} - \hat{b}_{k'}^l) = \\ &\quad \sum_{i \in \mathcal{E}_k^l} a_{ik} (\hat{b}_k - \hat{b}_k^l) + \sum_{k'=1}^{K-1} \sum_{i \in \mathcal{E}_K^l} a_{iK} (\hat{b}_{k'} - \hat{b}_{k'}^l) \end{aligned}$$

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} (\hat{\boldsymbol{w}}_k - \hat{\boldsymbol{w}}_k^l)^{\mathrm{T}} \boldsymbol{x}_i + \sum_{i \in \mathcal{E}_K^l} a_{iK} \sum_{k'=1}^{K-1} (\hat{\boldsymbol{w}}_{k'} - \hat{\boldsymbol{w}}_{k'}^l)^{\mathrm{T}} \boldsymbol{x}_i = \\ &\quad \sum_{k'=1}^{K-1} \sum_{j' \in \mathcal{A}_{k'}^l} \sum_{i \in \mathcal{E}_K^l} a_{iK} x_{ij'} (\hat{w}_{k'j'} - \hat{w}_{k'j'}^l) + \\ &\quad \sum_{j \in \mathcal{A}_k^l} \sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} (\hat{w}_{kj} - \hat{w}_{kj}^l) \end{aligned}$$

Hence, system of linear equations in (22) can be rewritten as

$$A_{11} \tilde{\boldsymbol{b}}_- + A_{12} \tilde{\boldsymbol{w}}_- = 0 \tag{24}$$

Note also that

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} (\hat{b}_k - \hat{b}_k^l) + \sum_{i \in \mathcal{E}_K^l} a_{iK} x_{ij} \sum_{k'=1}^{K-1} (\hat{b}_{k'} - \hat{b}_{k'}^l) = \\ &\quad \sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} (\hat{b}_k - \hat{b}_k^l) + \sum_{k'=1}^{K-1} \sum_{i \in \mathcal{E}_K^l} a_{iK} x_{ij} (\hat{b}_{k'} - \hat{b}_{k'}^l) \end{aligned}$$

$$\begin{aligned} &\sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} (\hat{\boldsymbol{w}}_k - \hat{\boldsymbol{w}}_k^l)^{\mathrm{T}} \boldsymbol{x}_i + \sum_{i \in \mathcal{E}_K^l} a_{iK} x_{ij} \sum_{k'=1}^{K-1} (\hat{\boldsymbol{w}}_{k'} - \hat{\boldsymbol{w}}_{k'}^l)^{\mathrm{T}} \boldsymbol{x}_i = \\ &\quad \sum_{j'' \in \mathcal{A}_k^l} \sum_{i \in \mathcal{E}_k^l} a_{ik} x_{ij} x_{ij''} (\hat{w}_{kj''} - \hat{w}_{kj''}^l) + \\ &\quad \sum_{k'=1}^{K-1} \sum_{j' \in \mathcal{A}_{k'}^l} \sum_{i \in \mathcal{E}_K^l} a_{iK} x_{ij} x_{ij'} (\hat{w}_{k'j'} - \hat{w}_{k'j'}^l) \end{aligned}$$

$$A_{22_2}\tilde{\boldsymbol{w}}_- = \begin{bmatrix} \hat{w}_{1j_{11}} - \hat{w}^l_{1j_{11}} + \sum_{k'=1}^{K-1}(\hat{w}_{k'j_{11}} - \hat{w}^l_{k'j_{11}}) \\ \vdots \\ \hat{w}_{2j_{21}} - \hat{w}^l_{2j_{21}} + \sum_{k'=1}^{K-1}(\hat{w}_{k'j_{21}} - \hat{w}_{k'j_{21}}) \\ \vdots \\ A_{22\alpha} \end{bmatrix}$$

where $A_{22\alpha} = \sum_{k'=1}^{K-1}(\hat{w}_{k'j_{K-1m_{K-1}}} - \hat{w}^l_{k'j_{K-1m_{K-1}}}) + \hat{w}_{K-1j_{K-1m_{K-1}}} - \hat{w}^l_{K-1j_{K-1m_{K-1}}}$. Hence, system of linear equations in (23) can be rewritten as

$$\frac{1}{n\delta}A_{12}^{\mathrm{T}}\tilde{\boldsymbol{b}}_- + 2\lambda_2 A_{22_2}\tilde{\boldsymbol{w}}_- = (\lambda_1 - \lambda_1^l)(\mathrm{sgn}((A_{22_2} - I)\hat{\tilde{\boldsymbol{w}}}^l_-) + \mathrm{sgn}(\hat{\tilde{\boldsymbol{w}}}^l_-)) \tag{25}$$

If A_l has full rank, then (16) can be obtained by solving the combined systems of linear equations (24) and (25). □

It should be noted that if A_l does not have full rank, then the solution paths are not unique, and more care has to be taken[16]. According to the sum-to-zero constraint, it can be easily obtained that $\hat{b}_K = -\sum_{k=1}^{K-1}\hat{b}^l_k - (\lambda_1 - \lambda_1^l)\sum_{k=1}^{K-1}\bar{a}_{k0}$ and $\hat{w}_{Kj} = -\sum_{k=1}^{K-1}\hat{w}^l_{kj} - (\lambda_1 - \lambda_1^l)\sum_{k=1}^{K-1}\bar{a}_{kj}$. Substituting (16) into $f_k(\boldsymbol{x}) = b_k + \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}$ gives $f_k(\boldsymbol{x}_i) = f^l_k(\boldsymbol{x}_i) + (\lambda_1 - \lambda_1^l)(\bar{a}_{k0} + \sum_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l} x_{ij}\bar{a}_{kj})$ and $f_K(\boldsymbol{x}_i) = -\sum_{k=1}^{K-1} f^l_k(\boldsymbol{x}_i) - (\lambda_1 - \lambda_1^l)\sum_{k=1}^{K-1}(\bar{a}_{k0} + \sum_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l} x_{ij}\bar{a}_{kj})$, where $k = 1, \cdots, K-1$.

2.4 Algorithm

Similar to the binary solution path algorithms[4–6, 16], our algorithm starts from $\lambda_1 \to \infty$ and we let $\hat{w}^0 = 0_{K\times p}$, $\hat{\boldsymbol{b}}^0_- = \arg\min_{b_-}\frac{1}{n}(\sum_{i=1}^n\sum_{k=1}^{K-1} a_{ik}L_k(b_k, 0, 0) + \sum_{i=1}^n a_{ik}L_k(-\sum_{k'=1}^{K-1} b_{k'}, 0, 0))$, $\hat{b}^0_K = -\sum_{k=1}^{K-1}\hat{b}^0_k$. According to the obtained $\hat{b}^0_k$, we determine the sets $\mathcal{E}^0_k$, $\mathcal{L}^0_k$, and $\mathcal{R}^0_k$ for $k = 1, \cdots, K$. Furthermore, we let $\lambda_1^0 = \max_{j=1}^p \max_{k=1}^{K-1}(|\sum_{i\in\mathcal{E}^0_k}\frac{a_{ik}x_{ij}}{n\delta}(\hat{b}^0_k + \sum_{k'=1}^{K-1}\hat{b}^0_{k'})|)$, $\mathcal{A}^0_k = \arg\max_{j\in\{1,\cdots,p\}}(|\sum_{i\in\mathcal{E}^0_k}\frac{a_{ik}x_{ij}}{n\delta}(\hat{b}^0_k + \sum_{k'=1}^{K-1}\hat{b}^0_{k'})|)$, $\lambda_2^0 = \beta - (\beta-\alpha)/(\ln(\mathrm{e} + \lambda_1^0))$, and determine $\bar{\mathcal{A}}^0$ according to $\mathcal{A}^0_k$. This completes the initialization.

After the l-th event has occurred, the most important problem of our algorithm is to determine the step size for the event which will occur first. Note that the first event will occur when $f_k(\boldsymbol{x}_i)$ reaches -1. Hence, the step size for the first event can be determined by

$$d_1 = \min\{d_{11}, \quad d_{12}\} \tag{26}$$

where $d_{11} = \min_{k=1,\cdots,k-1}\min_{i\in\mathcal{E}^l_k\cup\mathcal{L}^l_k}|(-1 - f^l_k(\boldsymbol{x}_i)/(\bar{a}_{k0} + \sum_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l} x_{ij}\bar{a}_{kj})|$, $d_{12} = \min_{i\in\mathcal{E}^l_K\cup\mathcal{L}^l_K}|(-1 + \sum_{k=1}^{K-1} f^l_k(\boldsymbol{x}_i))/(\sum_{k=1}^{K-1}(\bar{a}_{k0} + \sum_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l} x_{ij}\bar{a}_{kj}))|$. Let $d_{21} = \min_{k=1,\cdots,k-1}\min_{i\in\mathcal{E}^l_k\cup\mathcal{R}^l_k}|(-1 + \delta - f^l_k(\boldsymbol{x}_i))|/|(\bar{a}_{k0} + \sum_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l} x_{ij}\bar{a}_{kj})|$, $d_{22} = \min_{i\in\mathcal{E}^l_K\cup\mathcal{R}^l_K}|(-1 + \delta + \sum_{k=1}^{K-1} f^l_k(\boldsymbol{x}_i))/(\sum_{k=1}^{K-1}(\bar{a}_{k0} + \sum_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l} x_{ij}\bar{a}_{kj}))|$. Analogously, the step size for the second event can be determined by

$$d_2 = \min\{d_{21}, \quad d_{22}\} \tag{27}$$

Note that the third event will occur when a non-zero parameter $\hat{w}_{kj}$ reduces to zero or the sum of the first $K-1$ parameters in $\hat{\boldsymbol{w}}_{(j)}$ becomes zero. Hence, the step size for the third event can be determined by

$$d_3 = \min_{j\in\mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l}\min_{k\in\{1\cdots,K\}}|\frac{\hat{w}^l_{kj}}{\bar{a}_{kj}}| \tag{28}$$

where $\hat{w}^l_{Kj} = \sum_{k'=1}^{K-1}\hat{w}^l_{k'j}$ and $\bar{a}_{Kj} = \sum_{k'=1}^{K-1}\bar{a}_{k'j}$.

For $k = 1, \cdots, K-1$, $j = 1, \cdots, p$, we define $C_{kj} = \sum_{i=\mathcal{E}^l_K} a_{iK}x_{ij}(-1 + \sum_{k'=1}^{K-1}(\hat{b}_{k'} + \hat{\boldsymbol{w}}_{k'}^{\mathrm{T}}\boldsymbol{x}_i))/(n\delta) + \sum_{i\in\mathcal{E}^l_k} a_{ik}x_{ij}(1 + \hat{b}_k + \hat{\boldsymbol{w}}_k^{\mathrm{T}}\boldsymbol{x}_i)/(n\delta) + (\sum_{i\in\mathcal{R}^l_k} a_{ik}x_{ij} - \sum_{i\in\mathcal{R}^l_K} a_{iK}x_{ij})/n + 2\lambda_2^l(\hat{w}_{kj} + \sum_{k'=1}^{K-1}\hat{w}_{k'j})$. Note that $|C_{kj}| \le 2\lambda_1$, for $j \notin \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$; $C_{kj} = 0$ for $j \in \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$ and $\mathrm{sgn}(\hat{w}_{kj}) = -\mathrm{sgn}(\sum_{k'=1}^{K-1}\hat{w}_{k'j})$; $C_{kj} = -2\lambda_1\mathrm{sgn}(\hat{w}_{kj})$, for $j \in \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$ and $\mathrm{sgn}(\hat{w}_{kj}) = \mathrm{sgn}(\sum_{k'=1}^{K-1}\hat{w}_{k'j})$. Hence, after $|C_{kj}|$ ($j \notin \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$) meets the decreasing $2\lambda_1$ or reaches zero, the fourth event will occur if we keep moving λ_1 in the same direction. For $j \notin \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$, there are three cases: 1) $j \notin \mathcal{A}^l_k$ and $j \notin \bar{\mathcal{A}}^l$; 2) $j \in \mathcal{A}^l_k$ and $j \notin \bar{\mathcal{A}}^l$; 3) $j \notin \mathcal{A}^l_k$ and $j \in \bar{\mathcal{A}}^l$. Note that $\hat{w}_{kj} = 0$ and $\sum_{k'=1}^{K-1}\hat{w}_{k'j} = 0$ for $j \notin \mathcal{A}^l_k$ and $j \notin \bar{\mathcal{A}}^l$. Hence, for case 1), we have $C_{kj} = C^l_{kj} + C_{\mathcal{E}kj}(\lambda_1 - \lambda_1^l)$, where $C_{\mathcal{E}kj} = \sum_{i\in\mathcal{E}^l_k}\frac{a_{ik}x_{ij}}{n\delta}(\bar{a}_{k0} + \sum_{j''\in\mathcal{A}^l_k} x_{ij''}\bar{a}_{kj''}) + \sum_{i\in\mathcal{E}^l_K}\frac{a_{iK}x_{ij}}{n\delta}\sum_{k'=1}^{K-1}(\bar{a}_{k'0} + \sum_{j'\in\mathcal{A}^l_{k'}} x_{ij'}\bar{a}_{k'j'})$. The step size d_{411}, which makes $|C_{kj}|$ ($j \notin \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$) reach zero, can be calculated by

$$d_{411} = \left|\frac{C^l_{kj}}{C_{\mathcal{E}kj}}\right| \tag{29}$$

The step size d_{412}, which makes $|C_{kj}|$ ($j \notin \mathcal{A}^l_k\bigcap\bar{\mathcal{A}}^l$) meet the $2\lambda_1$, can be calculated by

$$|C^l_{kj} - C_{\mathcal{E}kj}d_{412}| = |C_{kj}| = 2\lambda_1 = 2(\lambda_1^l - d_{412}) \tag{30}$$

If $\mathrm{sgn}(C^l_{kj}) = \mathrm{sgn}(C_{\mathcal{E}kj})$ and $d_{412} \le |C^l_{kj}|/|C_{\mathcal{E}kj}|$, then

$$d_{412} = \frac{2\lambda_1^l - |C^l_{kj}|}{2 - |C_{\mathcal{E}kj}|}$$

If $\mathrm{sgn}(C^l_{kj}) = \mathrm{sgn}(C_{\mathcal{E}kj})$ and $d_{412} > |C^l_{kj}|/|C_{\mathcal{E}kj}|$, then

$$d_{412} = \frac{2\lambda_1^l + |C^l_{kj}|}{2 + |C_{\mathcal{E}kj}|}$$

Otherwise, we have

$$d_{412} = \frac{2\lambda_1^l - |C^l_{kj}|}{2 + |C_{\mathcal{E}kj}|}$$

It should be noted that the inequality $d_{412} \le |C^l_{kj}|/|C_{\mathcal{E}kj}|$ (or $d_{412} > |C^l_{kj}|/|C_{\mathcal{E}kj}|$) is equivalent to the computable inequality $\lambda_1^l|C_{\mathcal{E}kj}| \le |C^l_{kj}|$ (or $\lambda_1^l|C_{\mathcal{E}kj}| > |C^l_{kj}|$). Hence, the step size d_{41}, which determines the fourth event for case 1), can be determined by

$$d_{41} = \min_{j\notin\mathcal{A}^l_k, j\notin\bar{\mathcal{A}}^l}\min_{k=1,\cdots,K-1}\min\{d_{411}, d_{412}\} \tag{31}$$

Analogously, we can get the step sizes d_{42} and d_{43}, which determine the fourth event for case 2) and 3), respectively. It should be noted that the calculation procedure is the same as the case 1) except the different value of $C_{\mathcal{E}kj}$. Note that $\hat{w}_{kj} \ne 0$ and $\sum_{k'=1}^{K-1}\hat{w}_{k'j} = 0$ for

$j \in \mathcal{A}_k^l$ and $j \notin \bar{\mathcal{A}}^l$. Hence, for case 2), we have $C_{\mathcal{E}kj} = 2\lambda_2^l \bar{a}_{kj} + \sum_{i\in\mathcal{E}_k^l} \frac{a_{ik}x_{ij}}{n\delta}(\bar{a}_{k0} + \sum_{j''\in\mathcal{A}_k^l} x_{ij''}\bar{a}_{kj''}) + \sum_{i\in\mathcal{E}_K^l} \frac{a_{iK}x_{ij}}{n\delta} \sum_{k'=1}^{K-1}(\bar{a}_{k'0} + \sum_{j'\in\mathcal{A}_{k'}^l} x_{ij'}\bar{a}_{k'j'})$. Note also that $\hat{w}_{kj} = 0$ and $\sum_{k'=1}^{K-1}\hat{w}_{k'j} \neq 0$ for $j \notin \mathcal{A}_k^l$ and $j \in \bar{\mathcal{A}}^l$. Hence, for case 3), we have $C_{\mathcal{E}kj} = 2\lambda_2^l \sum_{k'=1}^{K-1}\bar{a}_{k'j} + \sum_{i\in\mathcal{E}_k^l} \frac{a_{ik}x_{ij}}{n\delta}(\bar{a}_{k0} + \sum_{j''\in\mathcal{A}_k^l} x_{ij''}\bar{a}_{kj''}) + \sum_{i\in\mathcal{E}_K^l} \frac{a_{iK}x_{ij}}{n\delta} \sum_{k'=1}^{K-1}(\bar{a}_{k'0} + \sum_{j'\in\mathcal{A}_{k'}^l} x_{ij'}\bar{a}_{k'j'})$. To sum up, the step size for the fourth event can be determined by

$$d_4 = \min\{d_{41}, d_{42}, d_{43}\} \tag{32}$$

The algorithm that computes the whole solution path $\hat{\boldsymbol{b}}, \hat{w}$ proceeds as follows:

Step 1. Initialization: calculate $\hat{w}^0$, $\hat{\boldsymbol{b}}^l$, λ_1^0, λ_2^0, $\mathcal{E}_k^0$, $\mathcal{L}_k^0$, $\mathcal{R}_k^0$, $\mathcal{A}_k^0$, and $\bar{\mathcal{A}}^0$, where $k = 1, \cdots, K$.

Step 2. Find λ_1^{l+1} and λ_2^{l+1}: let $\lambda_2^l = \beta - (\beta - \alpha)/(\ln(\mathrm{e} + \lambda_1^l))$, calculate $\hat{\boldsymbol{b}}$, $\hat{w}$ and $f_k(\boldsymbol{x}_i)$, and determine the step size $d = \min\{d_1, d_2, d_3, d_4\}$.

Step 3. If the generalized correlation reduces to zero or a pre-specified maximum iteration number is reached, then stop the algorithm.

Step 4. Otherwise, let $l = l + 1$, $\lambda_1^{l+1} = \lambda_1^l - d$, $\lambda_2^{l+1} = \beta - (\beta - \alpha)/(\ln(\mathrm{e} + \lambda_1^{l+1}))$ and update $\hat{\boldsymbol{b}}^l, \hat{w}^l$, $\mathcal{E}_k^l$, $\mathcal{L}_k^l$, $\mathcal{R}_k^l$, $\mathcal{A}_k^l$ and $\bar{\mathcal{A}}^l$. Then goto the Step 2.

3 Experiments on leukemia data

The aim of the leukemia benchmark[2] is to form a decision rule capable of distinguishing between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). This data set contains 38 samples for training and 34 samples for testing. The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7 129 genes (Affymetrix probes) are available. The original data set is available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. The identification of the two cancer types was based on their origins, lymphoid (lymph or lymphatic tissue related), and myeloid (bone marrow related), respectively. Similar to MSVM[8], we modify the leukemia data set as a three-class problem by dividing ALL into B-cell ALL and T-cell ALL. This study examines 72 samples from three types of acute leukemia with 38 samples in B-cell ALL, 9 samples in T-cell ALL, and 25 samples in AML.

In the first experiment, we train and test HMSVM using the original separation for the training and testing data. In the original training data set, there are 19 samples in B-cell ALL, 8 samples in T-cell ALL, and 11 samples in AML. The remaining 34 samples are used to test the prediction accuracy. To make the computation more manageable, we use the pre-processing steps proposed by Dudoit[1] and select the most significant 3 571 genes as the predictors. Let $\beta = 10$, $\alpha = 0.5$, $\delta = 0.05$. We compute the entire regularization solution paths according to the algorithm in Subsection 2.4. Fig. 1 shows the curve of correlation between the two regularization parameters. It is shown that λ_2 is the piecewise constant function with respect to λ_1. Fig. 2 shows the curve of prediction error. It is shown that the prediction error is piecewise linear with respect to λ_1 and the optimal mode is given when the regularization parameter λ_1 is selected in the interval [2.280905, 2.951294]. The corresponding prediction error is 0.02941176, i.e., only one sample in the test data set is misclassified.

In the second experiment, we compare HMSVM with several competitors for multiclass classification, including one-versus-all (OVA) classifier, L_2-norm MSVM and L_1-norm MSVM. We test their averaged prediction accuracy using a randomly splitting approach: the original training and testing data are combined together and we randomly split it into 38 and 34 samples for training and testing, respectively. The entire process is repeated 50 times and Table 1 gives the averaged test error and the averaged number of selected genes. In L_2-norm MSVM, 100 most important genes are selected to build the classifier by adopting the marginal criterion[1]. Although L_2-norm MSVM still works well by setting appropriate Gaussian kernel parameter σ and regularization parameter λ_2, it does not select genes in a satisfactory way. L_1-norm MSVM seems to overcome this disadvantage and improve the prediction accuracy. However, the number of the selected genes is upper bounded by the sample size and the selected genes vary largely with the randomly splitting training set. The HMSVM gives the best averaged test error and selects the moderate amount of genes. It is likely that the grouping effect in the process of automatic gene selection contributes much to the improved prediction accuracy.

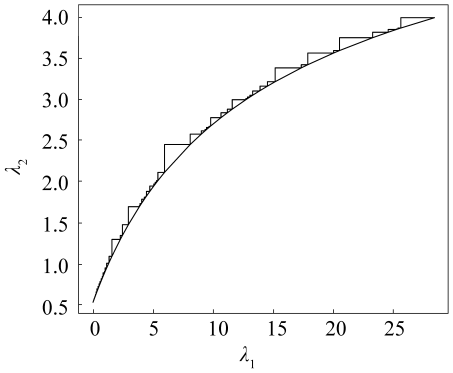

Fig. 1 Correlation between λ_2 and λ_1

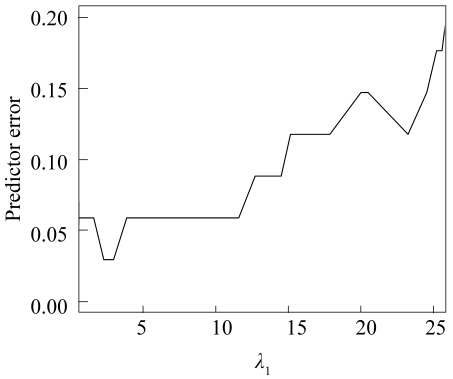

Fig. 2 Curve of the prediction error

4 Conclusions

The huberized multiclass SVM has been proposed in this paper, and the corresponding properties have been studied. It is shown that the HMSVM can encourage a grouping effect for multiclass gene selection. Based on a reasonable correlation of the two regularization parameters, the optimal coefficients are shown to be piecewise linear with re-

spect to the single regularization parameter and an efficient regularized path algorithm is developed. We have applied HMSVM to the leukemia data and achieved promising results.

Table 1 Classification results for the leukemia data

Method	Averaged test error (%)	Averaged number of selected genes
OVA	6.24	26.73
2-norm MSVM	4.20	100
1-norm MSVM	3.76	20.92
HMSVM	2.94	62.37

References

1 Dudoit S, Fridlyand J, Speed T P. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 2002, **97**(457): 77−87

2 Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, **286**(5439): 531−537

3 Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2002, **46**(1-3): 389−422

4 Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 2008, **24**(3): 412−419

5 Wang L, Zhu J, Zou H. The doubly regularized support vector machine. *Statistica Sinica*, 2006, **16**(2): 589−615

6 Zhu J, Rosset S, Hastie T, Tibshirani R. 1-norm support vector machines. *Advances in Neural Information Processing Systems*. New York: MIT Press, 2004. 49−56

7 Lee Y, Cui Z. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 2006, **16**(2): 391−409

8 Lee Y, Lin Y, Wahba G. Multicategory support vector machines: theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 2004, **99**(465): 67−81

9 Wang L F, Shen X T. On l_1-norm multi-class support vector machines: methodology and theory. *Journal of the American Statistical Association*, 2007, **102**(478): 583−594

10 Zhang H H, Liu Y, Wu Y, Zhu J. Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2008, **2**(1): 149−167

11 Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2001, **2**(3): 265−292

12 Li J, Jia Y, Du J, Yu F. Gene selection of multiple cancer types via huberized multi-class support vector machine. In: Proceedings of the 48th IEEE Conference on Decision and Control and the 28th Chinese Control Conference. Shanghai, China: IEEE, 2009. 1520−1525

13 Ma S, Song X, Huang J. Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics* [Online], available: http://www.biomedcentral.com/1471-2105/8/60, March 15, 2009

14 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 2006, **68**(1): 49−67

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

16 Hastie T, Rosset S, Tibshirani R, Zhu J. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 2004, **5**(12): 1391−1415

17 Rosset S, Zhu J. Piecewise linear regularized solution paths. *Annals of Statistics*, 2007, **35**(3): 1012−1030

LI Jun-Tao Ph. D. candidate at the Seven Research Division, Beihang University. His research interest covers intelligent control, statistical learning, data mining, and machine-learning-based bioinformatics. Corresponding author of this paper. E-mail: juntaolimail@yahoo.com.cn

JIA Ying-Min Professor at Beihang University. His research interest covers multivariable systems, robust control, adaptive control, and intelligent control and their applications in vehicle systems and industrial processes.
E-mail: ymjia@buaa.edu.cn