Supplemental Materials of ISR

Tingjin Luo, Chenping Hou, *Member, IEEE,* Feiping Nie, Hong Tao, and Dongyun Yi

In this file, we provide the detailed proof of Proposition 1 and Proposition 2 in the paper "Semi-supervised Feature Selection via Insensitive Sparse Regression and Its Application in Video Semantic Recognition".

Lemma 1. *When $0 < q \leq 2$, for any nonzero vectors $\mathbf{a}$ and $\mathbf{a}_k$, the following inequality holds:*

$$\|\mathbf{a}\|_2^q / \|\mathbf{a}_k\|_2^q - \frac{q}{2}\|\mathbf{a}\|_2^2 \Big/ \|\mathbf{a}_k\|_2^2 \leq 1 - \frac{q}{2}. \tag{1}$$

Proof. Denote $\phi(t) = t^q - \frac{q}{2}t^2 + \frac{q}{2} - 1, t > 0$. By mathematic analysis, we know that $t = 1$ is the maximum point and $\phi(1) = 0$. Thus, when $t > 0$ and $0 < q \leq 2$, $\phi(t) \leq 0$. Let $t^* = \|\mathbf{a}\|_2^2 \Big/ \|\mathbf{a}_k\|_2^2$, then $\phi(t^*) \leq 0$, that is to say

$$\|\mathbf{a}\|_2^q / \|\mathbf{a}_k\|_2^q - \frac{q}{2}\|\mathbf{a}\|_2^2 \Big/ \|\mathbf{a}_k\|_2^2 + \frac{q}{2} - 1 \leq 0. \tag{2}$$

Therefore, we can arrive at Lemma 1. □

I. Proof of Proposition 1

Proposition 1. *The procedures of ISR shown in Table 1 will monotonically decrease the objective function of Eq. (3) in each step.*

$$J(\mathbf{W}, \mathbf{b}, \tilde{\mathbf{F}}, \mathbf{D}) = Tr\left[\left(\mathbf{W}^T\mathbf{X} + \mathbf{b}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}^T\mathbf{X} + \mathbf{b}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}\left(\mathbf{W}^T\mathbf{X} + \mathbf{b}\mathbf{1}^T\right)\right] + \gamma Tr(\mathbf{W}^T\mathbf{D}\mathbf{W}). \tag{3}$$

Proof. According to Lemma 1, we have

$$\left\|\mathbf{w}_{(k+1)}^i\right\|_2^q \Big/ \left\|\mathbf{w}_{(k)}^i\right\|_2^q - \frac{q}{2}\left\|\mathbf{w}_{(k+1)}^i\right\|_2^2 \Big/ \left\|\mathbf{w}_{(k)}^i\right\|_2^2 \leq 1 - \frac{q}{2} \tag{4}$$

The above Equation is equivalent to the following inequality

$$\left\|\mathbf{w}_{(k+1)}^i\right\|_2^q - \frac{q}{2}\frac{\left\|\mathbf{w}_{(k+1)}^i\right\|_2^2}{\left\|\mathbf{w}_{(k)}^i\right\|_2^{2-q}} \leq \left\|\mathbf{w}_{(k)}^i\right\|_2^q - \frac{q}{2}\frac{\left\|\mathbf{w}_{(k)}^i\right\|_2^2}{\left\|\mathbf{w}_{(k)}^i\right\|_2^{2-q}}. \tag{5}$$

Thus the following inequality holds:

$$\sum_{i=1}^d \left(\left\|\mathbf{w}_{(k+1)}^i\right\|_2^q - \tfrac{q}{2}\left\|\mathbf{w}_{(k+1)}^i\right\|_2^2 \Big/ \left\|\mathbf{w}_{(k)}^i\right\|_2^{2-q}\right) \leq \sum_{i=1}^d \left(\left\|\mathbf{w}_{(k)}^i\right\|_2^q - \tfrac{q}{2}\left\|\mathbf{w}_{(k)}^i\right\|_2^2 \Big/ \left\|\mathbf{w}_{(k)}^i\right\|_2^{2-q}\right) \tag{6}$$

Assume that we have derived D and $\tilde{\mathbf{F}}$ as $\mathbf{D}_{(k)}$ and $\tilde{\mathbf{F}}_{(k)}$ in the k-th step. In the $(k+1)$-th iteration, we fix $\mathbf{D}$ and $\tilde{\mathbf{F}}$ as $\mathbf{D}_{(k)}$ and $\tilde{\mathbf{F}}_{(k)}$ and then optimize $\mathbf{W}$ and $\mathbf{b}$.

And then the optimization problem can be reformulated as:

$$\left(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}\right) = \arg\min_{\mathbf{W},\mathbf{b}} J(\mathbf{W}, \mathbf{b}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k)}) \tag{7}$$

When $\tilde{\mathbf{F}}$ and $\mathbf{D}$ are fixed in Eq. (3), the formulation in Eq. (3) is the regularized least square regression model. In other word, there is a closed solution for Eq. (3) in this case. So we have the following inequality:

$$J(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k)}) \leq J(\mathbf{W}_{(k)}, \mathbf{b}_{(k)}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k)}), \tag{8}$$

which presents that

$$\begin{aligned} &Tr\left[\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}_{(k)}\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)\right] + \gamma Tr\left(\mathbf{W}_{(k+1)}^T\mathbf{D}_{(k)}\mathbf{W}_{(k+1)}\right) \\ &\leq Tr\left[\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}_{(k)}\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)\right] + \gamma Tr(\mathbf{W}_{(k)}^T\mathbf{D}_{(k)}\mathbf{W}_{(k)}). \end{aligned} \tag{9}$$

That is to say,

$$\begin{aligned} &Tr\left[\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}_{(k)}\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)\right] + \gamma\left(\sum_{i=1}^{d}\frac{q}{2}\frac{\left\|\mathbf{w}_{(k+1)}^i\right\|_2^2}{\left\|\mathbf{w}_{(k)}^i\right\|_2^{2-q}}\right) \\ &\leq Tr\left[\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}_{(k)}\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)\right] + \gamma\sum_{i=1}^{d}\frac{q}{2}\frac{\left\|\mathbf{w}_{(k)}^i\right\|_2^2}{\left\|\mathbf{w}_{(k)}^i\right\|_2^{2-q}} \end{aligned} \tag{10}$$

where $\mathbf{w}_{(k+1)}^i$ and $\mathbf{w}_{(k)}^i$ are the i-th row of the matrix $\mathbf{W}_{(k+1)}$ and $\mathbf{W}_{(k)}$ respectively. Combining Eq. (6) and Eq. (10), we can get

$$\begin{aligned} &Tr\left[\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}_{(k)}\left(\mathbf{W}_{(k+1)}^T\mathbf{X} + \mathbf{b}_{(k+1)}\mathbf{1}^T\right)\right] + \gamma\sum_{i=1}^{d}\left\|\mathbf{w}_{(k+1)}^i\right\|_2^q \\ &\leq Tr\left[\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)\mathbf{S}\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)^T\right] - 2Tr\left[\tilde{\mathbf{F}}_{(k)}\left(\mathbf{W}_{(k)}^T\mathbf{X} + \mathbf{b}_{(k)}\mathbf{1}^T\right)\right] + \gamma\sum_{i=1}^{d}\left\|\mathbf{w}_{(k)}^i\right\|_2^q. \end{aligned} \tag{11}$$

That is to say, the equation is equivalent to the following inequality

$$J(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k+1)}) \leq J(\mathbf{W}_{(k)}, \mathbf{b}_{(k)}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k)}). \tag{12}$$

And then we will update $\tilde{\mathbf{F}}_{(k)}$ with $\mathbf{W}_{(k+1)}$ and $\mathbf{b}_{(k+1)}$. Then we will prove that the procedure of updating $\tilde{\mathbf{F}}_{(k)}$ also decreases the objective function value. In other words, we want to proof the following inequality holds:

$$J(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}, \tilde{\mathbf{F}}_{(k+1)}, \mathbf{D}_{(k+1)}) \leq J(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k+1)}). \tag{13}$$

Similar with [1] and [2], we define two auxiliary functions:

$$\begin{aligned} &\mathbf{h} : \mathbb{R}^{C\times n} \mapsto \mathbb{R}_+^n, \mathbf{h}(\tilde{\mathbf{X}}) = \left[\left\|\tilde{\mathbf{X}}_1\right\|_2^p, ..., \left\|\tilde{\mathbf{X}}_n\right\|_2^p\right]^T, \tilde{\mathbf{X}} = \left[\tilde{\mathbf{X}}_1, .., \tilde{\mathbf{X}}_n\right], \tilde{\mathbf{X}}_i \in \mathbb{R}^C \\ &g : \mathbb{R}_+^n \mapsto \mathbb{R}_+, g\left(\mathbf{u}\right) = \sum_{i=1}^{n}\lambda_i \min\left(u_i, \varepsilon\right) \end{aligned} \tag{14}$$

Note that $g\left(\cdot\right)$ is a concave function and we say that a vector $\mathbf{s} \in \mathbb{R}^n$ is a sub-gradient of g at $\mathbf{v} \in \mathbb{R}_+^n$, if for all vector $\mathbf{u} \in \mathbb{R}_+^n$, the inequality holds:

$$g\left(\mathbf{u}\right) \leq g\left(\mathbf{v}\right) + \left\langle \mathbf{s}, \mathbf{u} - \mathbf{v}\right\rangle, \tag{15}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. Define $\tilde{\mathbf{X}}_j = \mathbf{W}^T\mathbf{X} + \mathbf{b}\mathbf{1}^T - \mathbf{t}_j\mathbf{1}^T = \left[\tilde{\mathbf{X}}_{j1},..,\tilde{\mathbf{X}}_{jn}\right], \tilde{\mathbf{X}}_{ji} \in \mathbb{R}^C, \mathbf{h}(\tilde{\mathbf{X}}_j) = \left[\left\|\tilde{\mathbf{X}}_{j1}\right\|_2^p,...,\left\|\tilde{\mathbf{X}}_{jn}\right\|_2^p\right]^T$, $\mathbf{u} = \mathbf{h}(\tilde{\mathbf{X}}_j), g_j(\mathbf{u}) = g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right) = \sum\limits_{i=1}^{n} F_{ij}\min(u_i,\varepsilon)$. Note that the form of $g_j(\cdot)$ is consistent with the form of $g(\cdot)$.

According to the above defined functions, the original formulation of ISR can be equivalently rewritten as follows:

$$\min_{\mathbf{W},\mathbf{b}} \sum_{j=1}^{C} g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right) + \gamma R(\mathbf{W}), R(\mathbf{W}) = Tr(\mathbf{W}^T\mathbf{D}\mathbf{W}). \tag{16}$$

Based on the definition of the sub-gradient for a concave function given above, we can obtain an upper bound of $g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right)$ using a locally linear approximation at $\mathbf{h}(\hat{\mathbf{X}}_j^{(k)})$:

$$g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right) \le g_j\left(\mathbf{h}(\hat{\mathbf{X}}_j^{(k)})\right) + \left\langle \mathbf{s}_j^{(k)}, \mathbf{h}(\tilde{\mathbf{X}}_j) - \mathbf{h}(\hat{\mathbf{X}}_j^{(k)})\right\rangle, \tag{17}$$

where $\mathbf{s}_j^{(k)}$ is a sub-gradient of $g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right)$ at $\mathbf{h}(\hat{\mathbf{X}}_j^{(k)})$. It can be shown that a sub-gradient of $g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right)$ at $\mathbf{h}(\hat{\mathbf{X}}_j^{(k)})$ is

$$\mathbf{s}_j^{(k)} = \frac{p}{2}\left[F_{1j}\left\|\hat{\mathbf{X}}_{j1}^{(k)}\right\|_2^{p-2} Ind\left(\left\|\hat{\mathbf{X}}_{j1}^{(k)}\right\|_2^p, \varepsilon\right), ..., F_{nj}\left\|\hat{\mathbf{X}}_{jn}^{(k)}\right\|_2^{p-2} Ind\left(\left\|\hat{\mathbf{X}}_{jn}^{(k)}\right\|_2^p, \varepsilon\right)\right]^T. \tag{18}$$

Furthermore, we can obtain an upper bound of the objective function in Eq.(16), if the solution $\hat{\mathbf{X}}_j^{(k)}$ at the k-th iteration is available:

$$\forall \mathbf{W} \in \mathbb{R}^{d\times C}, \mathbf{b} \in \mathbb{R}^C : \sum_{j=1}^{C} g_j\left(\mathbf{h}(\tilde{\mathbf{X}}_j)\right) + \gamma R(\mathbf{W}) \le \sum_{j=1}^{C}\left[g_j\left(\mathbf{h}(\hat{\mathbf{X}}_j^{(k)})\right) + \left\langle \mathbf{s}_j^{(k)}, \mathbf{h}(\tilde{\mathbf{X}}_j) - \mathbf{h}(\hat{\mathbf{X}}_j^{(k)})\right\rangle\right] + \gamma R(\mathbf{W}). \tag{19}$$

Since $\mathbf{h}(\hat{\mathbf{X}}_j^{(k)}), j = 1, ..., C$ and γ are constant with respect to $\mathbf{W}$ and $\mathbf{b}$, we have

$$\left(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}\right) = \arg\min_{\mathbf{W},\mathbf{b}} \sum_{j=1}^{C}\left[g_j\left(\mathbf{h}(\hat{\mathbf{X}}_j^{(k)})\right) + \left\langle \mathbf{s}_j^{(k)}, \mathbf{h}(\tilde{\mathbf{X}}_j) - \mathbf{h}(\hat{\mathbf{X}}_j^{(k)})\right\rangle\right] + \gamma R(\mathbf{W}) = \sum_{j=1}^{C}\left\langle \mathbf{s}_j^{(k)}, \mathbf{h}(\tilde{\mathbf{X}}_j)\right\rangle + \gamma R(\mathbf{W}), \tag{20}$$

which obtains the next iterative solution by minimizing the upper bound of the objective function in Eq. (19) . Thus, in the viewpoint of the locally linear approximation, we can understand Algorithm 1 in Table 1 as follows: the original formulation of ISR is non-convex and is difficult to solve; and then the proposed method minimizes an upper bound in each step, which is convex and can be solved efficiently. It is closely related to concave convex procedure (CCCP) [3]. In addition, we can easily verify that the objective function value decreases monotonically as follows:

$$\begin{aligned}
\sum_{j=1}^{C} g_j\left(\mathbf{h}\left(\hat{\mathbf{X}}_j^{(k+1)}\right)\right) + \gamma R\left(\hat{\mathbf{W}}_{(k+1)}\right) &\le \sum_{j=1}^{C}\left[g_j\left(\mathbf{h}\left(\hat{\mathbf{X}}_j^{(k)}\right)\right) + \left\langle \mathbf{s}_j^{(k)}, \mathbf{h}\left(\hat{\mathbf{X}}_j^{(k+1)}\right) - \mathbf{h}\left(\hat{\mathbf{X}}_j^{(k)}\right)\right\rangle\right] + \gamma R\left(\hat{\mathbf{W}}_{(k+1)}\right) \\
&\le \sum_{j=1}^{C}\left[g_j\left(\mathbf{h}\left(\hat{\mathbf{X}}_j^{(k)}\right)\right) + \left\langle \mathbf{s}_j^{(k)}, \mathbf{h}\left(\hat{\mathbf{X}}_j^{(k)}\right) - \mathbf{h}\left(\hat{\mathbf{X}}_j^{(k)}\right)\right\rangle\right] + \gamma R\left(\hat{\mathbf{W}}_{(k)}\right) \\
&\le \sum_{j=1}^{C}\left[g_j\left(\mathbf{h}\left(\hat{\mathbf{X}}_j^{(k)}\right)\right)\right] + \gamma R\left(\hat{\mathbf{W}}_{(k)}\right).
\end{aligned} \tag{21}$$

According to the definition of $\tilde{\mathbf{F}}$, we know that $\mathbf{s}_j^{(k)}$ is equivalent to $\tilde{\mathbf{F}}_j^{(k)}$. Apparently, the inequality in Eq. (13) holds. Finally, combining Eq. (12) and Eq. (13), we can arrive at

$$J(\mathbf{W}_{(k+1)}, \mathbf{b}_{(k+1)}, \tilde{\mathbf{F}}_{(k+1)}, \mathbf{D}_{(k+1)}) \leq J(\mathbf{W}_{(k)}, \mathbf{b}_{(k)}, \tilde{\mathbf{F}}_{(k)}, \mathbf{D}_{(k)}). \tag{22}$$

This inequality indicates that the objective function in Eq. (3) will monotonically decrease in each iteration. An import issue we should mention is that a monotonic decrease of the objective function value does not guarantee the convergence of the algorithm, even if the objective function is strictly convex and continuously differentiable. □

II. Proof of Proposition 2

Proposition 2. *The procedures of ISR shown in Table 1 will monotonically decrease the objective function of Eq. (3) in each step and then converge to the local optimum of the problem (23).*

$$\min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{n}\sum_{j=1}^{C} F_{ij} \min(\left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right\|_2^p, \varepsilon) + \gamma \left\|\mathbf{W}\right\|_{2,q}^q \tag{23}$$

Proof. Based on the conclusion of proposition 1, we adopt the similar strategy in the conference paper [4] to prove it to converge to the local optimum of the problem (23).

In each iteration of ISR in Table 1, we find the optimal solution to the problem (3). Thus the converged solution of ISR satisfies the KKT condition of problem (3). Taking the derivative w.r.t. $\mathbf{W}$ and $\mathbf{b}$ respectively and setting them to zero, we get the KKT conditions of the problem of Eq. (3) as Eq. (24) and Eq. (25) in the main parts of this paper.

$$\frac{\partial J(\mathbf{W}, \mathbf{b}, \tilde{\mathbf{F}}, \mathbf{D})}{\partial \mathbf{W}} = \gamma \mathbf{D}\mathbf{W} + \mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{W} + \mathbf{X}(\mathbf{S}\mathbf{1}\mathbf{b}^T - \tilde{\mathbf{F}}) = \mathbf{0}. \tag{24}$$

$$\frac{\partial J(\mathbf{W}, \mathbf{b}, \tilde{\mathbf{F}}, \mathbf{D})}{\partial \mathbf{b}} = \mathbf{W}^T\mathbf{X}\mathbf{S}\mathbf{1} + \mathbf{1}^T\mathbf{S}\mathbf{1}\mathbf{b} - \tilde{\mathbf{F}}^T\mathbf{1} = \mathbf{0} \tag{25}$$

According to optimization theory, the Lagrangian function of problem of Eq. (23) is

$$J\left(\mathbf{W}, \mathbf{b}\right) = \sum_{i=1}^{n}\sum_{j=1}^{C} F_{ij} \min(\left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right\|_2^p, \varepsilon) + \gamma \left\|\mathbf{W}\right\|_{2,q}^q. \tag{26}$$

Taking the derivative w.r.t. $\mathbf{W}$ and $\mathbf{b}$ respectively and setting them to zero, we get the KKT condition of the problem of Eq. (26) as follows:

$$\begin{aligned}\frac{\partial J\left(\mathbf{W}, \mathbf{b}\right)}{\partial \mathbf{W}} =& \sum_{i=1}^{n}\sum_{j=1}^{C} F_{ij}\frac{p}{2}\left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right\|_2^{p-2}\mathbf{x}_i\left[\mathbf{x}_i^T\mathbf{W} - (\mathbf{b} - \mathbf{t}_j)^T\right] Ind(\left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right\|_2^p \leq \varepsilon)\\ &+ 2\gamma\sum_{i=1}^{d}\frac{q}{2}\left\|\mathbf{w}^i\right\|_2^{q-2}\mathbf{w}^i = 0\end{aligned} \tag{27}$$

$$\frac{\partial J\left(\mathbf{W}, \mathbf{b}\right)}{\partial \mathbf{b}} = \sum_{i=1}^{n}\sum_{j=1}^{C} F_{ij}\frac{p}{2}\left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right\|_2^{p-2}\left(\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right) Ind\left(\left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{t}_j\right\|_2^p \leq \varepsilon\right) = 0. \tag{28}$$

Using the matrix calculus, we can write the above Eq. (27) and Eq. (28) as following:

$$\begin{aligned}\frac{\partial J(\mathbf{W},\mathbf{b})}{\partial \mathbf{W}} &= \sum_{i=1}^{n}\sum_{j=1}^{C}\tilde{F}_{ij}\mathbf{x}_i\left[\mathbf{x}_i^T\mathbf{W}+(\mathbf{b}-\mathbf{t}_j)^T\right]+\gamma\sum_{i=1}^{d}D_{ii}\mathbf{w}^i = \gamma\mathbf{DW}+\sum_{i=1}^{n}\sum_{j=1}^{C}\tilde{F}_{ij}\left[\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}+\mathbf{x}_i\mathbf{b}^T-\mathbf{x}_i\mathbf{t}_j^T\right]\\ &= \gamma\mathbf{DW}+\sum_{i=1}^{n}\left(\sum_{j=1}^{C}\tilde{F}_{ij}\right)\left(\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}+\mathbf{x}_i\mathbf{b}^T\right)-\sum_{i=1}^{n}\mathbf{x}_i\left(\sum_{j=1}^{C}\tilde{F}_{ij}\mathbf{t}_j^T\right)\\ &= \gamma\mathbf{DW}+\sum_{i=1}^{n}S_{ii}\left(\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}+\mathbf{x}_i\mathbf{b}^T\right)-\sum_{i=1}^{n}\mathbf{x}_i\tilde{\mathbf{F}}^i, \tilde{\mathbf{F}}^i=\sum_{j=1}^{C}\tilde{F}_{ij}\mathbf{t}_j^T\\ &= \gamma\mathbf{DW}+\mathbf{XSX}^T\mathbf{W}+\mathbf{XS1b}^T-\mathbf{X}\tilde{\mathbf{F}}=0\end{aligned} \tag{29}$$

$$\begin{aligned}\frac{\partial J(\mathbf{W},\mathbf{b})}{\partial \mathbf{b}} &= \sum_{i=1}^{n}\sum_{j=1}^{C}\tilde{F}_{ij}\left(\mathbf{W}^T\mathbf{x}_i+\mathbf{b}-\mathbf{t}_j\right)=\sum_{i=1}^{n}\left(\sum_{j=1}^{C}\tilde{F}_{ij}\right)\left(\mathbf{W}^T\mathbf{x}_i+\mathbf{b}\right)-\sum_{i=1}^{n}\left(\sum_{j=1}^{C}\tilde{F}_{ij}\mathbf{t}_j\right)\\ &= \sum_{i=1}^{n}S_{ii}\left(\mathbf{W}^T\mathbf{x}_i+\mathbf{b}\right)-\sum_{i=1}^{n}\left(\tilde{\mathbf{F}}^i\right)^T=\mathbf{W}^T\mathbf{XS1}+\mathbf{1}^T\mathbf{S1b}-\tilde{\mathbf{F}}^T\mathbf{1}=0\end{aligned} \tag{30}$$

According to the definition of D_{ii} and $\tilde{\mathbf{F}}$ in the main procedures of ISR, we can see that Eq. (29) and Eq. (30) are the same as Eq. (24) and Eq. (25) when our algorithm is converged. Therefore, the converged solution of Algorithm in Table 1 satisfies Eq. (29) and Eq. (30), the KKT conditions of problem (23). Thus the converged solution of Algorithm in Table 1 is a local minimal solution to the problem (23). □

III. Additional Experiments

A. Convergence Results

According to the above analysis of two propositions, we know ISR monotonically decreases in each iteration. The above theorem only indicates that the objective function is nonincreasing and does not show whether $\mathbf{W}$ converges. Since $\mathbf{W}$ is used for feature selection, we need to present the convergence behavior of it. We measure the divergence between $\mathbf{W}_{(k+1)}$ and $\mathbf{W}_{(k)}$ by the following metric:

$$Error(\mathbf{W},k)=\sum_{i=1}^{d}\left|\left\|\mathbf{w}_{(k+1)}^i\right\|_2-\left\|\mathbf{w}_{(k)}^i\right\|_2\right|. \tag{31}$$

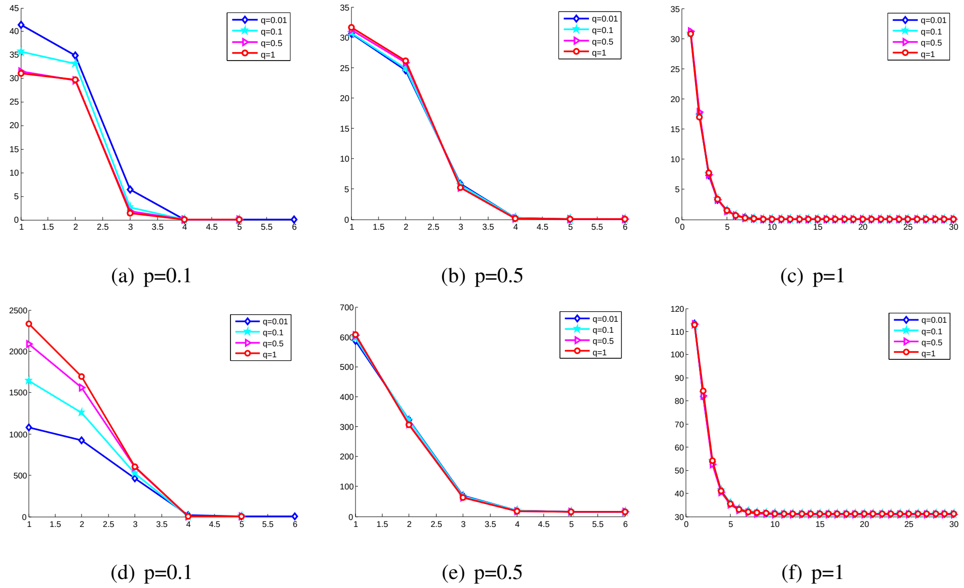
It will guarantee that the final feature results will not be changed drastically.

In this subsection, we provide some numerical results to validate the efficiency of our proposed method. We present the convergence behavior of ISR when $p = 0.1, 0.5, 1$ and $q = 0.01, 0.1, 0.5, 1$. We provide two kinds of results: the objective function value and the divergence between two consecutive W in Eq. (31). Two datasets Coil20[1] and Lung Cancer (LUNG)[2] are employed. The results are shown in Fig.1 and Fig.2, respectively.

As seen from Fig.1 and Fig.2, the objectives of ISR on LUNG and Coil20 are non-increasing during the iterations and they all converge to a fixed value. Additionally, on the two datasets, the divergences between two sequential W

[1] http://www.cs.columbia.edu/CAVE/research/coil-20.html

[2] https://sites.google.com/site/feipingnie/resoure



(a) p=0.1

(b) p=0.5

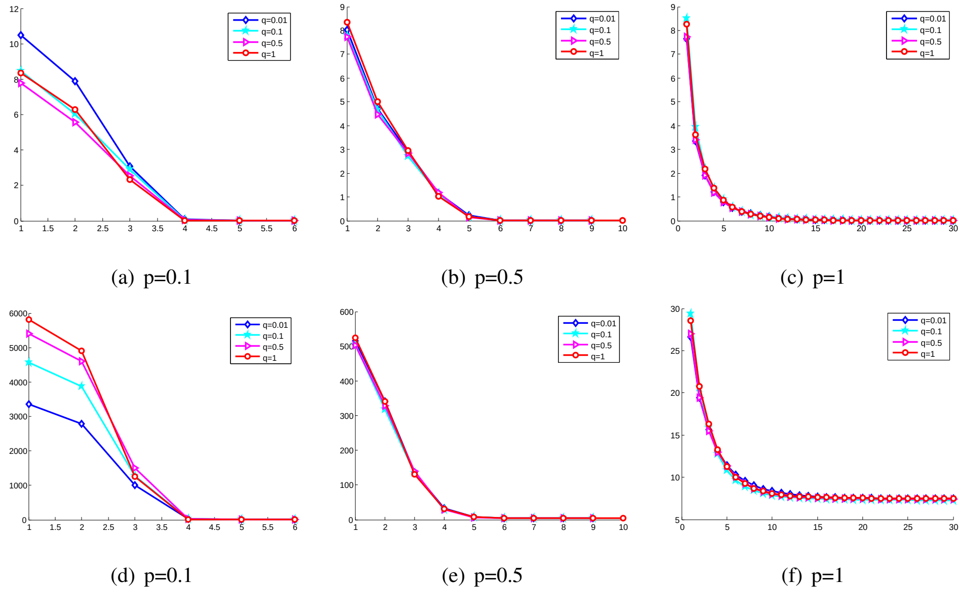
(c) p=1

(d) p=0.1

(e) p=0.5

(f) p=1

Fig. 1. Convergence behavior of ISR on data set COIL20 when $p = 0.1, 0.5, 1$. (a)-(c) is the objective value of ISR. (d)-(f) is divergence between two consecutive W.



(a) p=0.1

(b) p=0.5

(c) p=1

(d) p=0.1

(e) p=0.5

(f) p=1

Fig. 2. Convergence behavior of ISR on data set LUNG when $p = 0.1, 0.5, 1$. (a)-(c) is the objective value of ISR. (d)-(f) is divergence between two consecutive W.

all converge to zero, which illustrates that the final results will not be changed drastically. Moreover, ISR converges within 15 iterations on these two data sets for different p and q values. Therefore, with fast convergence speed, our proposed method scales well in practice.

TABLE I

STUDENT t-TEST RESULTS BETWEEN ISR AND OTHER APPROACHES FOR THE RESULTS OF SVM.

dataset	method	$s = 10$	$s = 20$	$s = 30$	$s = 40$	$s = 50$	$s = 60$	$s = 70$	$s = 80$	$s = 90$	$s = 100$
Umist	FisherScor	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	LapScor	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	RFS	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	MCFS	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.01)	B(–)	F(.01)	B(–)
	TRCFS	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	S^2FS^2R	F(.03)	W(.00)	W(.00)	W(.01)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
dataset	method	$s = 10$	$s = 20$	$s = 30$	$s = 40$	$s = 50$	$s = 60$	$s = 70$	$s = 80$	$s = 90$	$s = 100$
PIE	FisherScor	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	LapScor	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	RFS	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	MCFS	W(.00)	W(.00)	W(.00)	B(–)	B(–)	W(.01)	W(.03)	W(.00)	W(.00)	W(.00)
	TRCFS	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	S^2FS^2R	W(.00)	W(.00)	B(–)	B(–)	W(.00)	W(.00)	W(.00)	W(.03)	W(.00)	W(.00)

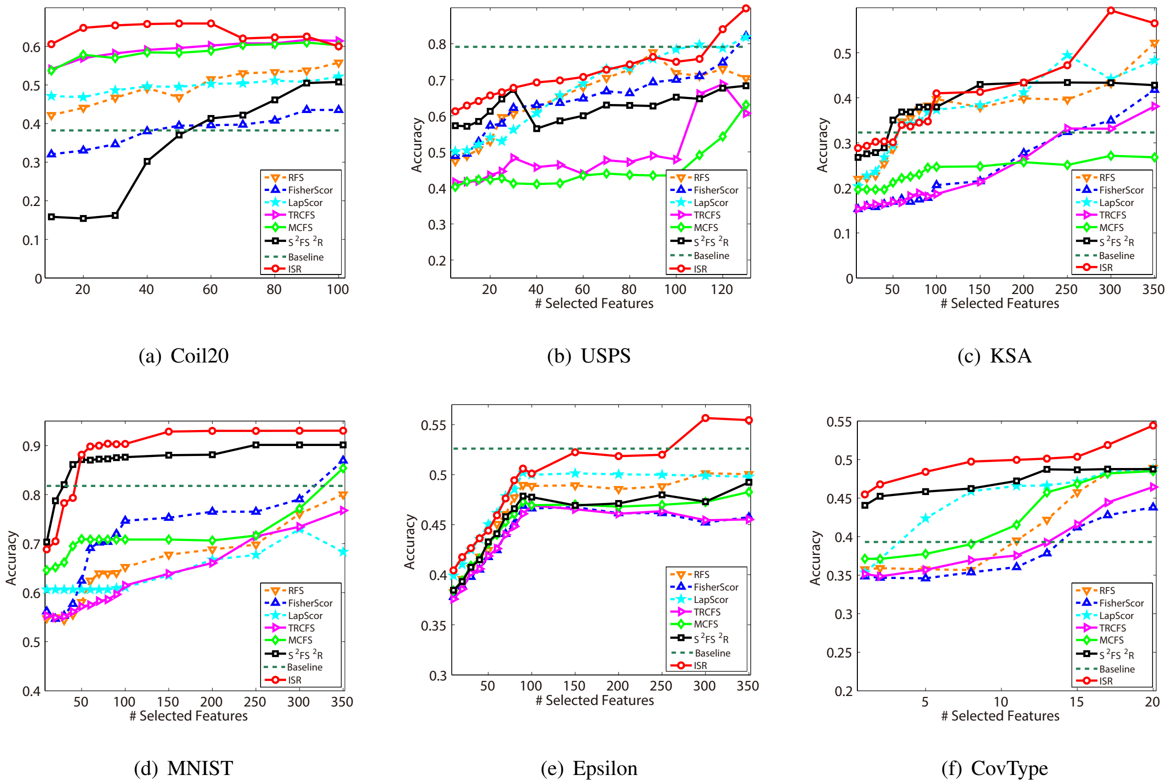
B. T-test of SVM Classification Results

In the main paper, we can see that the performance of ISR with SVM outperforms that of other methods. However, it seems that the experimental results of ISR does not obviously indicate that the performance of ISR is better than others' on Umist and PIE. To analyze the performance of ISR, we compare our method with other approaches by Students t-test. The experimental results are listed in Table I with a threshold of 0.05 statistical significance. In this table, the first letter means the comparing results and the value in brackets in the corresponding p-value. "W" means that ISR performs better than other approaches and "F" means that ISR fails. "B"means that we can not distinguish them in statistical view. In this case, we have not reported the p-value. The smaller p-value means the higher assurance of the conclusion.

From the statistical view, Table I illustrate that ISR achieves significantly better results comparing to the other algorithms in most cases. And the experimental results also show consistently that ISR can select features very efficiently and effectively for different classifiers.

C. Classification Results by Nearst Neiborhood Classifier

In this section, to evaluate the effect of different calsiifiers, we compare the classification accuracy of different methods by nearest neiborhood classifier (NNC). There are different number of sample sizes for different data sets. Thus we randomly select 3, 3, 10, 10, 15, 100, 100 and 100 instances with label information per class from Umist, Coil20, USPS, PIE, KSA, MNIST, Epsilon and CovType and 40% unlabeled samples as the training data, and the remained samples are used for testing. All the tests were repeated 20 times, and we then calculate the average classification accuracy. Since different data sets have different dimensions of features, we select various number of features according to the ranked feature indexes $\{r_1, r_2, \cdots, r_s\}$. Other parameters are determined by cross validation if necessary. The mean classification accuracy results are shown in Fig. 3.



(a) Coil20

(b) USPS

(c) KSA

(d) MNIST

(e) Epsilon

(f) CovType

Fig. 3. Classification accuracy of NNC with different number of selected features. The x-axis is the number of selected features and y-axis is classification accuracy.

As seen from the results in Fig. 3, the classification accuracies of different methods vary with the increase of the number of selected features. For all data sets except Umist and Coil20, all feature selection approaches achieve higher classification accuracy with more selected features. For Umist data set, the accuracy achieved by each method fluctuates within a certain range. With more features, the data can be characterized better. Generally, in most of cases, ISR outperforms all the other methods on all data sets for classification accuracy.

References

[1] P. Gong, J. Ye, and C.-s. Zhang, "Multi-stage multi-task feature learning," in *Advances in Neural Information Processing Systems*, 2012, pp. 1988–1996.

[2] P. Gong, J. Ye, and C. Zhang, "Multi-stage multi-task feature learning," *Journal of Machine Learning Research*, vol. 14, pp. 2979–3010, 2013. [Online]. Available: http://jmlr.org/papers/v14/gong13a.html

[3] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, Apr. 2003.

[4] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. ICML, 2014, pp. 1062–1070.