Self-weighted discriminative feature selection via adaptive redundancy minimization

Tong Wu a, Yicang Zhou a, Rui Zhang b, Yanni Xiao a, Feiping Nie b

a School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an, Shaanxi 710049, P.R. China
b School of Computer Science and Center for Optimal Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi’an, Shaanxi 710072, P.R. China

ARTICLE INFO

Article history:
Received 17 August 2017
Revised 23 October 2017
Accepted 25 November 2017
Available online 29 November 2017
Communicated by Dr. Chenping Hou

Keywords:
Self-adaptive weight
Linear discriminant analysis
Supervised feature selection
Adaptive correlation
Redundancy minimization

ABSTRACT

In this paper, a novel self-weighted orthogonal linear discriminant analysis (SOLDA) method is firstly proposed, such that optimal weight can be automatically achieved to balance both between-class and within-class scatter matrices. Since correlated features tend to have similar rankings, multiple adopted criteria might lead to the state that top ranked features are selected with large correlations, such that redundant information is brought about. To minimize associated redundancy, an original regularization term is introduced to the proposed SOLDA problem to penalize the high-correlated features. Different from other methods and techniques, we optimize redundancy matrix as a variable instead of setting it as a priori, such that correlations among all the features can be adaptively evaluated. Additionally, a brand new recursive method is derived to achieve the selection matrix heuristically, such that closed form solution can be obtained with holding the orthogonality. Consequently, self-weighted discriminative feature selection via adaptive redundancy minimization (SDFS-ARM) method can be summarized, such that non-redundant discriminative features could be selected correspondingly. Eventually, the effectiveness of the proposed SDFS-ARM method is further validated by the empirical results.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

How to choose relevant and informative features is the major concern of feature selection. To be more specific, feature selection is to pick out the features, which should be more representative of target object, such that crucial characteristics could be comprehended and grasped with much less cost of computation and identification. According to the criterion of utilizing the label information, feature selection methods can be categorized into unsupervised feature selection [1,2], semi-supervised feature selection [3], and supervised feature selection [4,5]. On the other hand, feature selection could also be classified into filter method [6,7], wrapper method [8,9], and embedded method [10–12]. These categories differ in the strategy, via which the learning algorithm is incorporated to evaluate the features. All in all, most existing feature selection approaches select features by ranking them, i.e., utilizing different criteria to compute feature scores, such that all the features could be sorted hierarchically. Although there are diverse feature selection methods, their core mechanism is similar, i.e., using different strategies to rank features.

In this paper, we focus on addressing the negative effect of correlated features on feature selection. As far as we know, correlated features frequently achieve similar rankings, due to the simple fact that they are considered to be equally important for feature selection. Consequently, top ranked features might be highly correlated to each other, such that these features would trigger redundant information, which may not be conducive to enhancing the efficiency of feature selection. Accordingly, how to select top non-redundant features is highlighted with many endeavors [9,13,14] being paid. In addition, redundancy matrix is often given in advance for minimization. However, it is inappropriate to set redundancy matrix as a priori, since it might be unlikely to estimate the correlation precisely via multiple choices of similarity functions even before pivotal features are clearly defined.

To tackle these two problems, an adaptive regularization term is introduced to penalize the high-correlated features. Specifically speaking, redundancy matrix is treated as an optimizing variable instead of setting it as a priori, such that correlations among all the features can be evaluated adaptively. Moreover, a novel self-weighted orthogonal linear discriminant analysis problem is proposed to jointly connect with the adaptive regularization term,
such that self-weighted discriminative feature selection via adaptive redundancy minimization (SDFS-ARM) problem can be formulated. Furthermore, we provide a new recursive method to solve the SDFS-ARM problem heuristically, such that closed form solution is obtained with holding the orthogonality. As a result, the optimal weight can be obtained automatically to balance both between-class and within-class scatters, while redundancy of the features can be minimized adaptively. In sum, SDFS-ARM method is proposed to select non-redundant discriminative features.

Compared to multiple LDA-based feature selection methods [15,16], SDFS-ARM is superior based on a brand new self-weighted model, which not only obtains the optimal weight automatically to leverage both between and within class scatters but also avoids the small sample problem of linear discriminate analysis (LDA). More importantly, the proposed SDFS-ARM introduces a novel regularization to adaptively reduce the redundancy among features. The technique is peculiar compared with the existing feature selection approaches.

Notations: Given an arbitrary sparse space $W \in \mathbb{R}^{d \times m}$, $w_i \in \mathbb{R}^{d \times 1}$ and $w^i \in \mathbb{R}^{1 \times m}$ are ith column and row of $W$, respectively. We introduce a brand new norm $\|w\|_{\Sigma_1}$ of $W$, i.e., $\|W\|_{\Sigma_1} = \sum_{i=1}^{m} \sqrt{\|w^i\|_{\Sigma_1}}$, where $\Sigma$ models the correlations of $w^i$. Particularly, this is not a standard matrix norm but we still name it as the norm for the convenience. Particularly, $\Sigma \geq 0$ represents that $\Sigma$ is a positive semi-definite matrix.

2. The proposed method

In this section, we at first propose a novel self-weighted orthogonal linear discriminant analysis problem. Additionally, an adaptive regularization term is further introduced to penalize the high-correlated features, since features with large correlations usually appear to have similar ranking scores. Consequently, not only the optimal weight can be achieved automatically to balance both between-class and within-class scatters, but also an adaptive redundancy matrix can be obtained simultaneously, such that non-redundant discriminative features are selected correspondingly.

2.1. Self-weighted orthogonal linear discriminant analysis

Given an input data matrix $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ with dimension $d$ and data number $n$, data point $x_i$, $(1 \leq i \leq n)$ is distributed into $c$ classes. In other words, each $x_i$ is assigned with a binary label vector $Y_i \in \{0, 1\}^c$. Note that binary label matrix $Y = [Y_1, \ldots, Y_n]^T \in \{0, 1\}^{n \times c}$ is composed of all the label vectors under data $X$, then total-class scatter matrix $S_T$, between-class scatter matrix $S_B$, and within-class scatter matrix $S_W$ can be represented as

$$
\begin{align*}
S_T &= XHX^T, \\
S_B &= XHY(Y^TY)^{-1}Y^THX^T, \\
S_W &= S_T - S_B = X(H - Y(Y^TY)^{-1}Y^T)HX^T,
\end{align*}
$$

(1)

where centering matrix $H = I - \frac{1}{n}11^T$ is idempotent, i.e., $H^2 = H$. According to Eq. (1), orthogonal linear discriminant analysis [17-19] could be demonstrated as a bi-objective optimization problem:

$$
\begin{align*}
\max_{W, W^T} & \text{Tr}(W^T S_B W) = \max_{W, W^T} \text{Tr}(W^TXH_X Y^T W) \\
\min_{W, W^T} & \text{Tr}(W^T S_W W) = \min_{W, W^T} \text{Tr}(W^TXL_X Y^T W),
\end{align*}
$$

(2)

where $L_X = HY(Y^TY)^{-1}Y^T$ and $L_X = H - HY(Y^TY)^{-1}Y^T$ with $W \in \mathbb{R}^{d \times m}$, $(m \leq d)$.

Based on the problem in (2), a novel self-weighted orthogonal linear discriminant analysis could be proposed to achieve the optimal weight automatically for leveraging both analysis in (2) as

$$
\begin{align*}
\max_{W, W^T} & \lambda \text{Tr}(W^T (S_B - \lambda S_W) W) \\
= & \max_{W, W^T} \text{Tr}(W^T X(L_0 - \lambda^2 L_w) X^T W) \\
\Rightarrow & \min_{W, W^T} \text{Tr}(W^T X L_X^{(\lambda^2)} X W),
\end{align*}
$$

(3)

where additional weight $\lambda$ is introduced as a multiplier in (3) with $L_X^{(\lambda^2)} = L_0 - \lambda^2 L_w$, such that Eq. (3) is a self-weighted model as further derived and described in (6).

2.2. Self-weighted discriminative feature selection via adaptive redundancy minimization

Note that $\Sigma$ models the correlation among different features, then minimizing the term $\text{Tr}(W^T \Sigma W)$ is related to reducing the redundancy during feature selection. Due to the real-world fact that $\Sigma$ cannot be actually given in most cases, it is inappropriate to set $\Sigma$ as a priori in advance, no matter that similarity measure is utilized. To address this issue, we treat $\Sigma$ as an optimizing variable and try to adaptively evaluate the correlations of features directly from the input data $X$.

To further strengthen the robustness of the term $\text{Tr}(W^T \Sigma W)$, a novel regularization $\|W\|_{\Sigma_1}|_{+}$ is introduced to Eq. (3), such that redundancy triggered by high-correlated features can be adaptively mitigated. In sum, self-weighted discriminative feature selection via adaptive redundancy minimization (SDFS-ARM) problem could be proposed as

$$
\begin{align*}
\min_{W, \Sigma, \lambda} & \text{Tr}(W^T X (\lambda^2 L_w - \lambda L_0) X^T W) + \alpha \|W\|_{\Sigma_1} + \beta \text{Tr}(\Sigma) \\
= & \min_{W, \Sigma, \lambda} \text{Tr}(W^T X L_X^{(\lambda^2)} X W) + \alpha \sum_{i=1}^{m} \sqrt{\|w^i\|_{\Sigma_1}} + \beta \text{Tr}(\Sigma) \\
\text{s.t.} & W^T W = 1, \Sigma \geq 0, \text{Tr}(\Sigma^{-1}) \leq 1,
\end{align*}
$$

(4)

where both regularization $\text{Tr}(\Sigma)$ and constraint $\text{Tr}(\Sigma^{-1}) \leq 1$ are specifically added to prevent the potential trivial solution of $\Sigma$. Since Eq. (4) is difficult to solve directly, re-weighted counterpart of Eq. (4) is further utilized as

$$
\begin{align*}
\min_{W, \Sigma, \lambda} & \sum_{i=1}^{m} (w_i^T X L_X^{(\lambda^2)} X W_i + \alpha D_i w_i^T \Sigma W_i) + \lambda \text{Tr}(\Sigma) \\
= & \min_{W, \Sigma, \lambda} \text{Tr}(W^T X L_X^{(\lambda^2)} X^T W) + \alpha \text{Tr}(D^T \Sigma W) + \beta \text{Tr}(\Sigma) \\
\text{s.t.} & W^T W = 1, \Sigma \geq 0, \text{Tr}(\Sigma^{-1}) \leq 1,
\end{align*}
$$

(5)

where $D = \text{diag}(D_1, D_2, \ldots, D_m)$ is to be updated iteratively in the algorithm with the specific form as $D_i \leftarrow \frac{1}{2 \sqrt{\|w^i\|_{\Sigma_1}}}$. $\forall i$.

**Theorem 2.1.** Problem (5) is convex with respect to (w.r.t.) $\lambda$, $W$, and $\Sigma$.

**Proof.** It is easy to verify that both first and last terms of the objective function in (5) are convex w.r.t. all the variables. In particular, the second term of Eq. (5) could be reformulated into

$$
\begin{align*}
\text{Tr}(D^T \Sigma W) = \sum_{i=1}^{m} D_i w_i^T \Sigma W_i,
\end{align*}
$$

where $D_i w_i^T \Sigma W_i$ is known as the matrix fractional function and it has been proved by [20] to be convex function w.r.t. $w_i$ and $\Sigma$ if $D_i \Sigma \geq 0$, i.e., a positive semi-definite matrix. This requirement $D_i \Sigma \geq 0$ is simply satisfied due to the constraint in (5) as $\Sigma \geq 0$ and $D_i \geq \frac{1}{2 \sqrt{\|w^i\|_{\Sigma_1}}} > 0$. Besides, since convexity can be preserved through summation as also proved in [20], $\text{Tr}(D^T \Sigma W) = \sum_{i=1}^{m} D_i w_i^T \Sigma W_i$ is convex w.r.t. $\lambda$, $W$, and $\Sigma$. 

Therefore, the objective function in (5) are convex w.r.t. all the variables. □

To avoid the inconvenience of solving Eq. (4), we solve its convex equivalent counterpart in (5) instead. Moreover, we utilize the block-coordinate descent method, i.e., optimizing the variables alternately in order to further solve Eq. (5).

**Optimizing \( \lambda \) by fixing \( \Sigma \) and \( W \):**

Since the first term in Eq. (5) includes the weight variable \( \lambda \), the extreme value condition w.r.t. \( \lambda \) can be illustrated as

\[
\frac{\partial \text{Tr} (WD^T) \Sigma W^T}{\partial \lambda} = 0 \Rightarrow \lambda = \frac{\text{Tr}(WD^T) \Sigma W^T}{2\text{Tr}(WD^T) \Sigma W^T},
\]

such that the proposed SDFS-ARM problem in (4) is self-weighted by updating the weight \( \lambda \) derived above until convergence.

**Optimizing \( \Sigma \) by fixing \( W \) and \( \lambda \):**

Apparently, variable \( \Sigma \) is contained in the last two terms of Eq. (5). According to Cauchy-Schwartz inequality and constraint (\( \Sigma^{-1} \leq 1 \)), we could infer that

\[
\alpha \text{Tr}(DW^T \Sigma W^T) + \beta \text{Tr}(\Sigma) = \alpha \text{Tr}
\]

\[
\left( WD^T + \frac{\beta}{\alpha} I \right) \left( WD^T + \frac{\beta}{\alpha} I \right)^{\frac{1}{2}} \text{Tr}
\]

\[
\left( \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right) \times 1
\]

\[
\geq \alpha \text{Tr} \left( WD^T + \frac{\beta}{\alpha} I \right) \left( WD^T + \frac{\beta}{\alpha} I \right)^{\frac{1}{2}} \text{Tr} \left( \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right)
\]

\[
\geq \alpha \left( \text{Tr} \left( WD^T + \frac{\beta}{\alpha} I \right) \right)^2,
\]

where equality holds if and only if

\[
\gamma \left( WD^T + \frac{\beta}{\alpha} I \right) \Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}} \Rightarrow \gamma \left( WD^T + \frac{\beta}{\alpha} I \right) = \Sigma^{-1},
\]

with an arbitrary constant \( \gamma \). In addition, we have

\[
\text{Tr}(\Sigma^{-1}) \leq 1 \Rightarrow \gamma = \frac{1}{\text{Tr}(WD^T + \frac{\beta}{\alpha} I)^{\frac{1}{2}}}
\]

Therefore, we could eventually obtain an adaptive redundancy matrix \( \Sigma \) as

\[
\Sigma^{-1} = \frac{WD^T + \frac{\beta}{\alpha} I^{\frac{1}{2}}}{\text{Tr}(WD^T + \frac{\beta}{\alpha} I^{\frac{1}{2}})}
\]

\[
\Rightarrow \Sigma = \text{Tr}(WD^T + \frac{\beta}{\alpha} I^{\frac{1}{2}})WD^T + \frac{\beta}{\alpha} I^{\frac{1}{2}},
\]

which largely depends on the choice of \( W \).

**Optimizing \( W \) by fixing \( \Sigma \) and \( \lambda \):**

The Lagrangian function of Eq. (5) can be formulated as

\[
\mathcal{L} = \text{Tr}(W^T \Sigma^{-1} X W) + \alpha \text{Tr}(W^T \Sigma W) + \beta \text{Tr}(\Sigma) - \text{Tr}(\Omega W^T W - I),
\]

where \( \Omega = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_m) \). Accordingly, the Karush-Kuhn-Tucker (KKT) condition w.r.t. \( W \) can be represented as

\[
\frac{\partial \mathcal{L}}{\partial W} = 0 \Rightarrow X^T \Sigma^{-1} X W^T + \alpha \Sigma W = \Omega W,
\]

which cannot be solved directly. However, Eq. (9) could be decoupled column-wisely into the following closed form solutions:

\[
(\Sigma^{-\frac{1}{2}} X^T \Sigma^{-\frac{1}{2}}) W_{i, \cdot} + \alpha D_{i, j} = \omega_{i, j},
\]

(10)

where \( W_{i, \cdot} \in \mathbb{R}^{d \times 1} \) is the \( i \)th column of orthogonal matrix \( W = [w_1, w_2, \ldots, w_m] \) with unit \( \ell_2 \)-norm, i.e., \( \|w_i\|_2 = \|w_{i, \cdot}\|_2 = 1 \). Although \( w_i \) can be individually solved by (10), the orthogonality, i.e., \( \langle w_i, w_j \rangle = 0, (i \neq j) \) could not be guaranteed to hold any more. Therefore, we will try to propose a novel recursive method to achieve the subspace \( W \) heuristically with holding the orthogonality.

**General recursive regression problem can be represented as**

\[
\min_{W} \sum_{l=1}^{m} f(w_l | X, Y) = \min_{W} \sum_{l=1}^{m} f(w_l | X, Y) \text{ s.t. } \langle w_l, w_j \rangle = 0, (i \neq j)
\]

(11)

with regression function \( f \), input data \( X \), and binary label matrix \( Y \). Due to the fact that

\[
\min_{W} \sum_{l=1}^{m} f(w_l | X, Y) \geq \sum_{l=1}^{m} \min_{w_l} f(w_l | X, Y)
\]

**Table 1** Information of the selected benchmark datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data No.</th>
<th>Class No.</th>
<th>Feature No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>840</td>
<td>120</td>
<td>768</td>
</tr>
<tr>
<td>COIL200</td>
<td>1440</td>
<td>20</td>
<td>1024</td>
</tr>
<tr>
<td>COIL200</td>
<td>7200</td>
<td>100</td>
<td>1024</td>
</tr>
<tr>
<td>FEI</td>
<td>2800</td>
<td>200</td>
<td>1024</td>
</tr>
<tr>
<td>IMM</td>
<td>240</td>
<td>40</td>
<td>1024</td>
</tr>
<tr>
<td>USPS</td>
<td>9298</td>
<td>10</td>
<td>256</td>
</tr>
</tbody>
</table>
(a) Convergent curve  (b) Adaptive Correlation $\Sigma$

**Fig. 1.** Convergence and adaptive correlation of the proposed SDFS-ARM are performed on the IMM dataset.

(a) $\alpha$ testing(KNN) (b) $\alpha$ testing(SVM) (c) $\beta$ testing(KNN) (d) $\beta$ testing(SVM)

(e) $\alpha$ testing(KNN) (f) $\alpha$ testing(SVM) (g) $\beta$ testing(KNN) (h) $\beta$ testing(SVM)

**Fig. 2.** The comparisons of classification accuracy are performed via tuning different values of the parameters, i.e., $\alpha$ and $\beta$ under datasets AR and USPS. (a)-(d): USPS. (e)-(h): AR. (a)-(b)&(e)-(f): $\alpha$ testing. (c)-(d)&(g)-(h): $\beta$ testing.

with orthogonal constraint $w_i^T w_j = 0, (i \neq j)$, then the recursive regression problem (11) degenerates to a series of sub-problems as

\[
\begin{align*}
\min_{w_1} & f(w_1 | X, Y) \\
\min_{w_2} & f(w_2 | X, Y) \\
\vdots \\
\min_{w_M} & f(w_M | X, Y) \\
\text{subject to} & \quad w_i^T w_j = 0, (i \neq j).
\end{align*}
\]

(12)

subject to $w_i^T w_j = 0, (i \neq j)$. To approximate to the problem (11), we provide the following proposition to achieve the heuristic solution of problem (12).

**Proposition 2.2.** For a linear learning model $L$, we suppose that the output projection vector $w = X\delta$, where $X$ is the input data and $\delta$ is the associated linear combination. In addition, an orthonormal projection vector $v$ has been learned based on the same learning model $L$, i.e., $v^Tv = 1$. If another projection vector $\tilde{w}$ is learned from the following reconstruction residue as $\tilde{X} = X - vv^T X$ under the same learning model $L$, then $\tilde{w}$ is orthogonal to $v$.

**Proof.** The output projection $w$ can be represented by $w = X\delta$. Likewise, we have $\tilde{w} = \tilde{X}\delta$ under certain $\delta$. Thus, we could infer that

$$<\tilde{w}, v> = \delta^T \tilde{X}^T v = \delta^T (X^T - Xvv^T)v = 0,$$

which denotes that $\tilde{w}$ is orthogonal to $v$. □
Table 2
The comparisons of classification accuracy and deviation are performed under selected 100 top features for 2 benchmark datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>COIL20</th>
<th>FEI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COIL20</td>
<td>FEI</td>
</tr>
<tr>
<td></td>
<td>(K-NN)</td>
<td>(SVM)</td>
</tr>
<tr>
<td>T-test</td>
<td>78.25 ± 2.18</td>
<td>78.95 ± 2.36</td>
</tr>
<tr>
<td>mRMR</td>
<td>82.67 ± 2.02</td>
<td>84.55 ± 1.52</td>
</tr>
<tr>
<td>TRC-FS</td>
<td>83.99 ± 1.41</td>
<td>84.97 ± 1.08</td>
</tr>
<tr>
<td>RALM-FS</td>
<td>84.38 ± 1.62</td>
<td>85.30 ± 1.62</td>
</tr>
<tr>
<td>CR-FS</td>
<td>84.79 ± 2.23</td>
<td>85.81 ± 2.29</td>
</tr>
<tr>
<td>SDFS-ARM(our)</td>
<td>86.24 ± 1.68</td>
<td>87.09 ± 1.37</td>
</tr>
</tbody>
</table>

Fig. 3. The comparisons of classification accuracy are performed via increasing number of features under datasets COIL20 and IMM. (a)-(b): COIL20, (c)-(d): IMM.

Equipped with Proposition 2.2, the recursive regression method can be proposed in Algorithm 1 to solve the problem (12) heuristically. In fact, the recursive regression serves as a greedy method, via which an optimization problem is decomposed into a series of subproblems. In other words, how to solve the subproblem \( \min_{w_i} f(X, Y) \) takes priority. Actually, the recursive regression is similar with the coordinate blocking method in the way of different initialization of \( X \) for each iteration.

Via applying KKT condition derived in (10), Eq. (5) w.r.t. subspace \( W \) can be reformulated into the following recursive regression form as

\[
\min_{w_i \in W} \sum_{i=1}^{m} \left( w_i^T \left( X^T X + \alpha D_n \Sigma \right) w_i \right) \quad \text{s.t.} \quad w_i^T w_i = 0, \quad (i \neq j)
\]

which could be easily solved via the proposed recursive method in Algorithm 1. Accordingly, SDFS-ARM method is proposed in Algorithm 2, such that \( D, \lambda, \Sigma, \) and \( W \) can be iteratively updated. Since matrix \( D \) is not a variable, we optimize Eq. (5) to get \( \lambda, \Sigma, W \) and then re-calculate Eq. (5), where \( D \) just serves as a transitional matrix.

3. Experiments

In this section, we compare the proposed SDFS-ARM feature selection algorithm with other five state-of-the-art feature selection algorithms including T-test [21], mRMR [9], TRC-FS [5], RALM-FS [22], and CR-FS [23]. In our experiments, we employ 50% of the input data as training set and set the residual part as test set. Additionally, classification accuracies on the test set are recorded and compared via two classifiers as K-NN [24] and linear SVM [25]. Moreover, classification accuracy is calculated by

\[
\text{Classification Accuracy} = \frac{N_{\text{correct}}}{N_{\text{test}}} \times 100\%,
\]

where \( N_{\text{correct}} \) and \( N_{\text{test}} \) represent the numbers of correctly classified test samples and whole test samples, respectively.

3.1. Datasets with initialization

Firstly, six real world datasets are used to validate the effectiveness of the proposed SDFS-ARM feature selection algorithm as AR[1],

COIL_{20}^2, COIL_{100}^3, FEI^4, IMM^5, and USPS^6. More specifics of each dataset are listed in Table 1. Additionally, reduced dimensionality is chosen as m = c − 1 in the experiment to ensure a fair comparison Fig. 1.

3.2. Parameter sensitivity

Secondly, we utilize the datasets AR and USPS to demonstrate the parameter sensitivity of the proposed SDFS-ARM method in Algorithm 2. Specifically speaking, we divide the illustration into two parts as α testing and β testing in Fig. 2. In other words, we tune α with fixing β in α testing and vice versa in β testing. From Fig. 2, we could draw the following conclusions:

1) From Fig. 2, we notice that classification accuracies fluctuate according to the values of α. Particularly, better results are achieved, when α ∈ [10^{-1}, 10^0].

2) From Fig. 2, we could observe that variations on the values of β have little influence on the classification outcomes.

Therefore, we tune α between the interval [10^{-1}, 10^0], while fixing β = 0.7α thereafter.

3.3. Classification comparisons

Finally, the classification comparisons of feature selection approaches previously mentioned are performed under both K-NN [24] and linear SVM [25] classifiers. In Table 2, average classification accuracies and deviations of selected 100 top features are recorded under 2 datasets as COIL_{100} and FEI. In Fig. 3, classification accuracies are compared via increasing number of features under the datasets COIL_{20} and IMM. From Table 2 and Fig. 3, we could conclude that:

1) From Table 2, the proposed SDFS-ARM method performs consistently better than the rest of the feature selection approaches.

2) From Fig. 3, the proposed SDFS-ARM method has better classification results regarding most feature numbers.

4. Conclusion

In this paper, a novel self-weighted orthogonal linear discriminant analysis (SOLDA) problem is firstly introduced, such that optimal weight can be obtained to balance both between-class and within-class scatter matrices automatically. To further minimize the redundancy among features, an original regularization is introduced to the proposed SOLDA problem to penalize the high-correlated features. With treating the redundancy matrix as an optimizing variable instead of presetting it as a priori, correlations among all the features can be evaluated adaptively. As a result, a brand new self-weighted discriminative feature selection via adaptive redundancy minimization (SDFS-ARM) method is proposed, such that associated non-redundant discriminative features could be selected.

References


T. Wu received her M.S. degree in applied mathematics from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2014. She is currently pursuing the Ph.D. degree in applied mathematics with School of Mathematics and Statistics, Xian Jiaotong University, Xi’an 710049, Shaanxi, P. R. China.

Yicang Zhou is currently a full professor in School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an 710049, Shaanxi, P. R. China.
Rui Zhang received his first M.S. degree in applied mathematics from Northwest University, Xi’an, Shaanxi, China, in 2011, and his second M.S. degree in applied mathematics from Georgia Southern University, Statesboro, GA, USA, in 2013. He is currently pursuing the Ph.D. degree in computer science with Northwestern Polytechnical University, Xi’an 710072, Shaanxi, P. R. China.

Yanni Xiao received Ph.D. from Academy of Mathematics and System Sciences, Chinese Academy of Sciences in 2001. She had been working as a Postdoc in the University of Liverpool, UK from 2003 to 2006. She is a professor in applied mathematics in School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an 710049, Shaanxi, P. R. China.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, China, in 2009. He is currently a full professor at Northwestern Polytechnical University, Xi’an, China. His research interests are machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He has published more than 100 papers in the prestigious journals and conferences like TIP, TPAMI, TNNLS, TKDE, ICML, NIPS, KDD, etc. He is currently serving as an associate editor or a PC member for several prestigious journals and conferences in the related fields.