L1-norm based discriminant manifold learning for multi-label image classification

Jiafeng Cheng1, Siyang Deng1, Wei Xia1, Yang Liu†

†State Key Laboratory of Integrated Services Networks, Xidian University, Shaanxi, People’s Republic of China
E-mail: yangli@xidian.edu.cn

Abstract: Recently, L1-norm based robust discriminant feature extraction technique has been attracted much attention in dimensionality reduction. However, most existing approaches solve the column vectors of the optimal projection matrix one by one with a greedy strategy. Moreover, they are not suitable for solving the multi-label image classification. To solve these problems, the authors give a model named L1-norm based discriminant manifold learning in this study. An iterative non-greedy algorithm is proposed to solve the objective and the obtained optimal projection matrix necessarily best optimise the corresponding trace ratio objective function, which is the essential criterion function for general supervised dimensionality reduction. They also analyse the convergence of the authors’ proposed algorithm in detail. Extensive experiments on some databases illustrate the effectiveness of their proposed method.

1 Introduction
Dimensionality reduction has been an active topic in the field of pattern recognition and machine learning. It tries to seek a low-dimensional subspace where the data are suitable for different tasks such as clustering or classification. Principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2] are two of the most representative subspace learning methods. PCA extracts the most expressive features in which the variance of samples are maximised. LDA encodes discriminant information and learns an optimal matrix by minimising the within-class scatter while maximising the between-class scatter in the low-dimensional subspace. Generally, dimensionality reduction methods can be roughly divided into two forms: ratio trace and trace ratio. For ratio trace form, the most representative approaches include Fisherface [2], regularised discriminant analysis (RDA) [3] and tensor LDA [4]. This kind of algorithm needs to compute the inverse of the matrix, resulting in computational cost, while trace ratio form does not. Trace ratio form aims to simultaneously seek all the projection vectors by maximising the criterion function. For example, Guo et al. [5] transformed trace ratio form to an equivalent trace difference problem and provided an iterative bisection way to solve the optimal projection matrix.

Since squared Euclidean distance excessively emphasises the large distance [6–8], the aforementioned approaches are prone to the presence of outliers, which are usually defined as the points that deviate significantly from the rest of data [9]. This results in a sensitivity of approaches to noise and outliers. To handle this problem, L1-norm based subspace learning has been considered to be capable of obtaining robust projection vectors. Two of the most representative methods are L1-PCA [10] and PCA-L1 [11], which employ L1-norm to measure reconstruction error and variance in the low-dimensional space, respectively. Motivated by the impressive results of L1-norm PCA, L1-norm discriminant analysis has attracted much attention in machine learning [12–14], where LDA-L1 [13] and kernel LDA-L1 [14] are two of the most representative methods, which employ L1-norm as the distance metric to calculate between-class and within-class scatters in the linear and nonlinear criterion functions, respectively.

Although some signs of progress have been made in L1-norm-based discriminant analysis to improve the classification performance, most of them have two disadvantages. (i) They generally solve the optimal projections by a greedy strategy. Thus, the obtained optimal projection matrix does not necessarily best optimise the corresponding trace ratio problem, which is the essential criterion function for general dimensionality reduction [15]. (ii) Most of L1-norm based methods are only suitable for solving single-label classification rather than multi-label classification problems.

To solve these problems, we propose a model named L1-norm based Discriminant Manifold Learning (L1-DML). In addition, an iterative non-greedy algorithm is given to solve the objective. A detailed analysis of the convergence of the algorithm is also given. Finally, we point out that our proposed algorithm can also be used to solve multi-label classification problems.

2 Related work
2.1 Multi-label linear discriminant analysis

Given the training data matrix \( X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n} \), where \( m \) is the dimensionality of training data and \( n \) denotes the number of training samples. The label matrix is

\[
Y = [y_1, y_2, \ldots, y_c] \in \mathbb{R}^{n \times c},
\]

where \( c \) is the number of classes. \( y_{ij} = 1 \) if \( x_i \) belongs to the jth \( (j = 1, 2, \ldots, c) \) class, and \( y_{ij} = 0 \) otherwise. Wang et al. [16] proposed a multi-label linear discriminant analysis (MLDA). In this model, the inter-class scatter \( S_b \) and intra-class scatter \( S_w \) are defined as follows:

\[
S_b = \sum_{i=1}^{c} \sum_{j=1}^{n} y_{ij}(m_i - m)(m_i - m)^T \tag{1}
\]

\[
S_w = \sum_{i=1}^{c} \sum_{j=1}^{n} y_{ij}(x_{ij} - m_i)(x_{ij} - m_i)^T \tag{2}
\]

where \( m_i \) is the mean of class \( i \), and \( m \) is the multi-label global mean, which is defined as follows:

\[
m = \frac{\sum_{i=1}^{n} y_{i1}x_{i1} + \sum_{j=1}^{c} \sum_{i=1}^{n} y_{ij}x_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij}} \tag{3}
\]

Similar to the traditional LDA, MLDA aims to search for an optimal \( W \) for discrimination by maximising the trace ratio between the inter-class and intra-class scatters.
Equation (4) uses the square of the L2 norm as the distance metric. However, this metric has two disadvantages: First, it is susceptible to outliers or noises. Second, since the square of the L2 norm expands the inter-class distances that are far apart, and at the same time weakens the inter-class distances that are close together, the samples cannot be effectively classified.

3 L1-norm based discriminant manifold learning (L1-DML)

In this section, we propose a new model named L1-norm based discriminant manifold learning (L1-DML). Then, a non-greedy iterative algorithm is given to solve our model. Finally, we analyse the convergence of the algorithm in detail.

3.1 Objective function

First, the L1 norm is more robust to noise and outliers than the square of the L2 norm [6, 17, 18]. Secondly, compared with the square of the L2 norm, the L1 norm can enlarge the distance between the classes closer to each other and weaken the influence of the distance between the classes farther apart, thus effectively improving the classification accuracy. According to the above analysis, we summarise the objective function as follows:

\[ W_{opt} = \arg \max_{W} \frac{\text{tr}(W^T S_1 W)}{\text{tr}(W^T S_2 W)} \tag{4} \]

3.2 Non-greedy iterative algorithm

Most of the existing algorithms usually adopt the greedy strategy to solve the trace ratio problem, i.e., solve each vector in the projection matrix \( W \) sequentially. One disadvantage of this strategy is the final projection matrix does not effectively maximise or minimise the objective function. To solve this problem, we propose a non-greedy iterative algorithm to solve the objective function. Motivated by the authors of [15, 19], (5) can be approximately equivalent to solving the trace difference objective function, which is

\[ G(W, \lambda) = \arg \max_{W} H(W) - \lambda M(W) \tag{6} \]

where

\[ H(W) = \sum_{i=1}^{c} \sum_{j=1}^{n} y_{ij} \left( W^T (m_i - m) \right)^T \tag{7} \]

\[ M(W) = \sum_{i=1}^{c} \sum_{j=1}^{n} y_{ij} \left( W^T (x_j - m) \right)^T \tag{8} \]

Equation (6) includes two unknown variables, i.e., \( W \) and \( \lambda \). We alternatively update \( W \) (while fixing \( \lambda \)) and \( \lambda \) (while fixing \( W \)). Specifically, assume that in the kth iteration, we solve \( \lambda^k \) by the following equation with \( W^{k-1} \) which is obtained in the \((k-1)\)th iteration:

\[ \lambda^k = \frac{H(W^{k-1})}{M(W^{k-1})} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} y_{ij} (W^{k-1})^T (m_i - m)}{\sum_{i=1}^{c} \sum_{j=1}^{n} y_{ij} (W^{k-1})^T (x_j - m)} \tag{9} \]

Then, we update \( W \) by the following equation:
where \( \mathbf{w}_p \) and \( \mathbf{w}_p^{k-1} \) denote the \( p \)th \((p = 1, \ldots, d) \) column of \( \mathbf{W} \) and \( \mathbf{W}^{k-1} \), respectively, and \( \text{sign()} \) is a polarity function.

**BD** and \( \mathbf{A} \) are defined as follows:

\[
\mathbf{B} = \{ m_1, m_2, \ldots, m_m \}
\]

\[
\mathbf{D} = \text{diag} \left( \frac{y_{j1}}{\sum_{j=1}^{n} y_{j1}}, \ldots, \frac{y_{jn}}{\sum_{j=1}^{n} y_{jn}} \right)
\]

\[
\mathbf{A} = \sum_{j=1}^{c} \sum_{j=1}^{c} y_{ji}(x_{j} - m_j)^T
\]

Then the following equation is an auxiliary function for \( F(\mathbf{W}) = \mathbf{H}(\mathbf{W}) - \lambda^k \mathbf{N}(\mathbf{W}, \mathbf{W}^{k-1}) \)

\[
L(\mathbf{W}, \mathbf{W}^{k-1}) = K(\mathbf{W}, \mathbf{W}^{k-1}) - \lambda^k \mathbf{N}(\mathbf{W}, \mathbf{W}^{k-1})
\]

**Proof:**

\[
\mathbf{H}(\mathbf{W}) = \sum_{j=1}^{c} \sum_{j=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_{j} - m_j) \right\|^2
\]

\[
\geq \sum_{j=1}^{c} \sum_{j=1}^{c} y_{ji} \text{sign}(\mathbf{w}_p^{k-1})^T(m_j - m_i)^T \mathbf{w}_p
\]

\[
= \sum_{i=1}^{p} (\mathbf{w}_p^T \mathbf{B} \text{sign}(\mathbf{w}_p^{k-1}) \mathbf{w}_p^{k-1})
\]

\[
= K(\mathbf{W}, \mathbf{W}^{k-1})
\]

According to Lemma 1, we have

\[
M(\mathbf{W}) = \sum_{j=1}^{c} \sum_{j=1}^{c} \frac{y_{ji}}{\left\| \mathbf{W}^T(x_{j} - m_j) \right\|^2}
\]

\[
\leq \frac{1}{2} \sum_{i=1}^{p} \sum_{i=1}^{c} \sum_{i=1}^{c} \frac{y_{ji}}{\left\| \mathbf{w}_p^{k-1} \right\|^2} (x_{j} - m_j)^T \mathbf{w}_p
\]

\[
= \frac{1}{2} \sum_{i=1}^{p} \sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \left\| \mathbf{w}_p^{k-1} \right\|^2 (x_{j} - m_j)
\]

\[
N(\mathbf{W}, \mathbf{W}^{k-1}) = \sum_{i=1}^{p} (\mathbf{w}_p^T \mathbf{A}_p \mathbf{w}_p + (\mathbf{w}_p^{k-1})^T \mathbf{A}_p \mathbf{w}_p^{k-1})
\]

\[
\text{Proof:}
\]

\[
\mathbf{H}(\mathbf{W}) = \sum_{j=1}^{c} \sum_{j=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_{j} - m_j) \right\|^2
\]

\[
\geq \sum_{j=1}^{c} \sum_{j=1}^{c} y_{ji} \text{sign}(\mathbf{w}_p^{k-1})^T(m_j - m_i)^T \mathbf{w}_p
\]

\[
= \sum_{i=1}^{p} (\mathbf{w}_p^T \mathbf{B} \text{sign}(\mathbf{w}_p^{k-1}) \mathbf{w}_p^{k-1})
\]

\[
= K(\mathbf{W}, \mathbf{W}^{k-1})
\]

1. Calculate \( \lambda^k \) by (9), and the sub-gradient \( G(\mathbf{W}) = \partial L(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}) \) by (25).
2. Calculate \( \mathbf{W}^k = \mathbf{P}(\mathbf{W}^{k-1} + \beta G(\mathbf{W}^k)) \) by (26).
3. Calculate \( F(\mathbf{W}) = H(\mathbf{W}) - \lambda^k M(\mathbf{W}) \). If \( F(\mathbf{W}^k) \geq 0 \), go to step 4; else \( m = m + 1 \), and go to step 2.
4. \( k = k + 1 \), return step 1 until \( (H(\mathbf{W}^k)/M(\mathbf{W}^k)) \) converges.

**Output:** \( \mathbf{W}^k \)

Now we consider how to find a \( \mathbf{W}^* \) satisfying \( L(\mathbf{W}^*, \mathbf{W}^{k-1}) \geq 0 \). It can be easily solved by a projected sub-gradient method with Armijo line search [13]. The sub-gradient of \( L(\mathbf{W}, \mathbf{W}^{k-1}) \) at \( \mathbf{W} \) is

\[
\partial L(\mathbf{W}, \mathbf{W}^{k-1}) = \left[ \sum_{i=1}^{c} B_i \text{sign}(\mathbf{w}_p^{k-1})(m_j - m_i) \right]
\]

\[
- \beta \mathbf{D} \mathbf{w}
\]

Note that, for any matrix \( \mathbf{W} \), the following operator (21) can project it onto an orthogonal cone. This guarantees the orthotropic constraint of the projection matrix

\[
P(\mathbf{W}) = \mathbf{W}(\mathbf{W}^T \mathbf{W})^{- \frac{1}{2}}
\]

We summarise the pseudo-code of solving the objective function (5), i.e. L1-DML in Algorithm 1.

**3.3 Convergence analysis**

Before analysing the convergence of Algorithm 1, we first introduce the following theorems.

**Theorem 3:** The objective function (5) has an upper bound.

**Proof:** According to \( \sqrt{a^2 + b^2 + c^2} \leq |a| + |b| + |c| \), for the numerator of the objective function (5), we have

\[
\sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_j - m_j) \right\|^2
\]

\[
\geq \left\| \sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \mathbf{W}^T(x_j - m_j) \right\|^2
\]

\[
\geq \left\| \sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \right\| \left\| \sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \mathbf{W}^T(x_j - m_j) \right\|_2
\]

\[
\sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_j - m_j) \right\|_2
\]

It is easy to know \( \sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_j - m_j) \right\| \) is a convex function and is always greater than zero, thus the numerator of (5) has an upper bound.

According to the Cauchy–Schwarz inequality

\[
[x, y] \leq \| x \| \| y \|
\]

for the denominator of (5), we have (see (28)) . The above formula shows that the denominator of (5) has a lower bound and is always greater than zero.

Combining (27) and (28), we have that the objective function (5) has an upper bound. □

**Theorem 4:** In Algorithm 1, for each iteration, if \( \mathbf{W}^k \) is the solution of \( F(\mathbf{W}) \geq 0 \) and satisfies \( \mathbf{W}^T \mathbf{W} = \mathbf{I}_d \), then we have \( J(\mathbf{W}^k) \geq J(\mathbf{W}^{k-1}) \), where

\[
J(\mathbf{W}^k) = \frac{\sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji}}{\left\| \mathbf{W}^T(x_j - m_j) \right\|_2}
\]

\[
\sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_j - m_j) \right\|_2
\]

\[
\sum_{i=1}^{c} \sum_{i=1}^{c} y_{ji} \left\| \mathbf{W}^T(x_j - m_j) \right\|_2
\]
Proof: Since $W^k$ is the solution of $F(W^k) \geq 0$, then we have

$$F(W^k) = \sum_{i=1}^c \sum_{j=1}^n y_{ij} \left\| (W^k)^T (x_j - m_i) \right\|_1 - x^k \geq 0$$

By simple algebraic steps, we have

$$J(W^k) = \sum_{i=1}^c \sum_{j=1}^n y_{ij} \left\| (W^k)^T (x_j - m_i) \right\|_1 \geq x^k$$

Substituting (9) into (31), we have

$$J(W^k) \geq \sum_{i=1}^c \sum_{j=1}^n y_{ij} \left\| (W^{k-1})^T (x_j - m_i) \right\|_1 = J(W^{k-1})$$

Theorem 3: indicates that our proposed Algorithm 1, which solves the objective function (5), monotonically increases in each iteration and will converge to a local optimum. 

4 Experiments

In this section, we validate our proposed method on a large dataset (Yahoo), whose detail information is illustrated in Table 1, where the Card column refers to the average number of labels of the samples in the corresponding dataset. We also compare the proposed method with related multi-label learning methods: MLkNN [20], BRkNN [21], LP [22], IBLR-ML [23], PPLS-MD [24], MDDM [25], MLDA [16] and LSMLDA [26].

By analysing the above experimental results, we can get the following conclusions:

(i) Table 2 shows that L1-DML performs best under the four metrics on the Yahoo dataset. This is probably because the proposed method L1-DML adopts the L1 norm to metric the similarity or difference between samples, which is robust to different sub-datasets. The MLkNN, BRkNN, IBLR-ML, and LP methods perform worse than other algorithms because MLkNN and BRkNN do not consider the importance of the labels. Although IBLR-ML solves this problem by using logistic regression, it treats class labels independently and does not consider the correlation between labels. In addition, the performance of MLkNN and IBLR-ML also depends on the value of the parameter $k$ of kNN, which is difficult to determine in practice.

(ii) Figs. 1 and 2 show that when a number of projection vectors are bigger than 15, L1-MLDA performs better than MDDM, MLDA and LSMLDA three algorithms under each index. Different from the other three algorithms, the performance of L1-MLDA is almost unaffected by the number of projection vectors. Specifically, when the projection vectors are bigger than 50, the performance of the two algorithms MLDA and LSMLDA that based on the L2 norm square metric is poor, while the performance of the L1-MLDA that based on the L1 norm metric has almost no change. This demonstrates that the algorithm proposed in this paper can maintain important attributes of the original sample after reducing dimensionality.

(iii) Fig. 3 shows that the L1-DML can converge to a local optimum when the number of iteration steps is less than 50. It is confirmed that the proposed non-greedy algorithm is single-increasing in each iteration. And eventually, it can converge to a local maximum.

5 Conclusion

In this paper, we propose an L1-norm based discriminant manifold learning framework for multi-label classifications. Different from most existing related algorithms, the proposed algorithm uses the L1 norm to metric the similarity or difference between samples. In addition, a non-greedy algorithm is given to solve the proposed model. The proposed algorithm simultaneously solves all the projection vectors and can best optimise the corresponding L1-norm based trace ratio objective function, which is the essential criterion function for general supervised dimensionality reduction. Compared with some existing L1-norm based algorithms, the proposed algorithm can obtain the large objective function value and have a local convergence. Experiments illustrate the superiority of our algorithm.

| Dataset | Samples | Dim | Labels | Card |
|---------|---------|-----|--------|------|
| arts    | 5000    | 462 | 26     | 1.636|
| education | 5000    | 550 | 33     | 1.461|
| entertainment | 5000    | 640 | 21     | 1.420|
| health  | 5000    | 612 | 32     | 1.662|
| recreation | 5000    | 606 | 22     | 1.423|
| reference | 5000    | 793 | 33     | 1.169|
| science | 5000    | 743 | 40     | 1.451|
| social  | 5000    | 1047| 39     | 1.283|

J. Eng., 2020, Vol. 2020 Iss. 13, pp. 664-669
This is an open access article published by the IET under the Creative Commons Attribution -NonCommercial License (http://creativecommons.org/licenses/by-nc/3.0/)
Acknowledgment

The authors thank the anonymous reviewers and AE for their constructive comments and suggestions, which improved the paper substantially. This work is supported in part by the National Natural Science Foundation of China under Grant 61906141, the National Natural Science Foundation of Shaanxi Province under Grant No. 2020QJ-317, China Postdoctoral Science Foundation (Grant No. 2019M655364), Open Project Program of the State Key Lab of CAD&CG (Grant No. A2018), Zhejiang University and the Fundamental Research Funds for the Central Universities.

Table 2 Experimental results (mean±std.) of the compared algorithms on Yahoo dataset

| Algorithms | Evaluation criterion       |
|-----------|-----------------------------|
|           | Hamming Loss (↓) | Ranking Loss (↓) | One-error (↓) | Average precision (↑) |
| MLkNN     | 0.045 ± 0.015 | 0.116 ± 0.044 | 0.562 ± 0.120 | 0.565 ± 0.094 |
| BRkNN     | 0.048 ± 0.024 | 0.332 ± 0.079 | 0.710 ± 0.119 | 0.405 ± 0.100 |
| LP        | 0.062 ± 0.037 | 0.581 ± 0.101 | 0.701 ± 0.122 | 0.244 ± 0.098 |
| IBLR-ML   | 0.045 ± 0.014 | 0.122 ± 0.042 | 0.560 ± 0.115 | 0.564 ± 0.090 |
| PPLS-MD   | $0.043±0.014$ | 0.106 ± 0.036 | 0.433 ± 0.096 | 0.654 ± 0.079 |
| MDDM      | 0.045 ± 0.015 | 0.115 ± 0.043 | 0.493 ± 0.100 | 0.598 ± 0.085 |
| MLDA      | 0.044 ± 0.014 | 0.104 ± 0.035 | 0.456 ± 0.083 | 0.633 ± 0.078 |
| LSLMDA    | 0.044 ± 0.013 | 0.104 ± 0.035 | 0.455 ± 0.086 | 0.631 ± 0.076 |
| L1-DML    | 0.040 ± 0.014 | 0.099 ± 0.038 | 0.397 ± 0.100 | 0.666 ± 0.085 |

For each evaluation criterion, ‘↑’ indicates ‘the bigger the better’ and ‘↓’ indicates ‘the smaller the better’.

References

[1] Jolliffe, I.: ‘Principal component analysis’ (Wiley Online Library, New York, USA, 2002)
[2] Fisher, R.A.: ‘The use of multiple measurements in taxonomic problems’, Ann. Eugen., 1936, 7 (2), pp. 179–188
[3] Friedman, J.H.: ‘Regularized discriminant analysis’, J. Am. Stat. Assoc., 1989, 84 (405), pp. 165–175
[4] Wang, S.-J., Yang, J., Zhang, N., et al.: ‘Tensor discriminant colour space for face recognition’, IEEE Trans Image Process, 2011, 20 (9), pp. 2490–2501
[5] Guo, Y.-F., Li, S.-J., Yang, J.-Y., et al.: ‘A generalized Foley-Sammon transform based on generalized fisher discriminant criterion and its application to face recognition’, Pattern Recognit. Lett., 2003, 24 (1), pp. 147–158
[6] De La Torre, F., Black, M.J.: ‘A framework for robust subspace learning’, Int. J. Comput. Vis., 2003, 54 (1–3), pp. 117–142
[7] Zhang, F., Yang, J., Qian, J., et al.: ‘Nuclear norm-based 2-dpca for extracting features from images’, IEEE Trans. Neural Netw. Learn. Syst., 2015, 26 (10), pp. 2247–2260
[8] Gao, Q., Xu, S., Chen, F., et al.: ‘R1-2DPCA and face recognition’, IEEE Trans. Cybern., 2019, 49 (4), pp. 1212–1223
[9] Ding, C., Zhou, D., He, X., et al.: ‘R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization’. Proc. of the 23rd Int. Conf. on Machine Learning, Pittsburgh, PA, USA., 2006, pp. 281–288
[10] Ke, Q., Kanade, T.: ‘Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA., 2005, pp. 739–746
[11] Kwak, N.: ‘Principal component analysis based on L1-norm maximization’, IEEE Trans. Pattern Anal. Mach. Intell., 2008, 30 (9), pp. 1672–1680
[12] Zhong, F., Zhang, J.: ‘Linear discriminant analysis based on l1-norm maximization’, IEEE Trans. Image Process., 2013, 22 (8), pp. 3018–3027
[13] Liu, Y., Gao, Q., Miao, S., et al.: ‘A non-greedy algorithm for l1-norm lda’, IEEE Trans. Neural Netw. Learn. Syst., 2014, 25 (4), pp. 793–805
[14] Wang, H., Yan, S., Xu, D., et al.: ‘Trace ratio vs. Ratio trace for dimensionality reduction’. Proc. IEEE Int. Conf. Computer Vision on Pattern Recognition, Minneapolis, MN, USA., 2007, pp. 1–8
[15] Wang, H., Ding, C., Huang, H.: ‘Multi-label linear discriminant analysis’. European Conf. on Computer Vision, San Francisco, CA, USA, 2016, pp. 126–139

Fig. 1 Hamming loss versus number of projection vectors on the Yahoo dataset

Fig. 2 Average precision versus number of projection vectors on the Yahoo dataset

Fig. 3 Convergence curve of our method on Yahoo dataset
[17] Liu, Y., Gao, Q., Gao, X., et al.: ‘L2, 1-norm discriminant manifold learning’, IEEE Access, 2018, 6, pp. 40723–40734

[18] Aanas, H., Fisker, R., Astrm, K., et al.: ‘Robust factorization’, IEEE Trans. Pattern Anal. Mach. Intell., 2002, 24 (9), pp. 1215–1225

[19] Gao, Q., Ma, L., Liu, Y., et al.: ‘Angle 2DPCA: a new formulation for 2DPCA’, IEEE Trans. Cybern., 2018, 48 (5), pp. 1672–1678

[20] Zhang, M.-L., Zhou, Z.-H.: ‘ML-kNN: a lazy learning approach to multi-label learning’, Pattern Recognit., 2007, 40 (7), pp. 2038–2048

[21] Tsoumakas, G., Vlahavas, I.: ‘Random k-label sets: an ensemble method for multilabel classification’, European Conf. on Machine Learning, Warsaw, Poland, 2007, pp. 406–417

[22] Tsoumakas, G., Katakis, I., Vlahavas, I.: ‘Mining multi-label data’, in Maimon, O., Rokach, L. (Eds.): Data mining and knowledge discovery handbook (Springer, Boston, MA, 2009, 2nd edn.)

[23] Cheng, W., Hüllermeier, E.: ‘Combining instance-based learning and logistic regression for multilabel classification’, Mach. Learn., 2009, 76, (2-3), pp. 211–225

[24] Liu, H., Ma, Z., Zhang, S., et al.: ‘Penalized partial least square discriminant analysis with l1-norm for multi-label data’, Pattern Recognit., 2015, 48, (5), pp. 1724–1733

[25] Zhang, Y., Zhou, Z.-H.: ‘Multilabel dimensionality reduction via dependence maximization’, ACM Trans. Know. Discov. Data (TKDD), 2010, 4, (3), pp. 1–21

[26] Shu, X., Xu, H., Tao, L.: ‘A least squares formulation of multi-label linear discriminant analysis’, Neurocomputing, 2015, 156, pp. 221–230