The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Trace Ratio Optimization with Feature Correlation Mining for Multiclass Discriminant Analysis

Forough Rezaei Boroujeni,[1] Sen Wang,[1] Zhihui Li,[2] Nicholas West,[3]
Bela Stantic,[1] Lina Yao,[4] Guodong Long[5]
[1]School of Information and Communication Technology, Griffith University, Gold Coast, Australia
[2]Beijing Etrol Technologies Company Ltd., Beijing 100095, China
[3]Immunology Research Group, Griffith Health Centre, Griffith University, Gold Coast, Australia
[4]School of Computer Science and Engineering, The University of New South Wales, Australia
[5]Centre for Aritificial Intelligence (CAI), University of Technology Sydney, Australia
forough.rezaeiboroujeni@griffithuni.edu.au,{sen.wang, n.west, b.stantic}@griffith.edu.au, zhihuilics@gmail.com,
lina.yao@unsw.edu.au, Guodong.long@uts.edu.au

Abstract

Fisher's linear discriminant analysis is a widely accepted dimensionality reduction method, which aims to find a transformation matrix to convert feature space to a smaller space by maximising the between-class scatter matrix while minimising the within-class scatter matrix. Although the fast and easy process of finding the transformation matrix has made this method attractive, overemphasizing the large class distances makes the criterion of this method suboptimal. In this case, the close class pairs tend to overlap in the subspace. Despite different weighting methods having been developed to overcome this problem, there is still a room to improve this issue. In this work, we study a weighted trace ratio by maximising the harmonic mean of the multiple objective reciprocals. To further improve the performance, we enforce the $\ell_{2,1}$-norm to the developed objective function. Additionally, we propose an iterative algorithm to optimise this objective function. The proposed method avoids the domination problem of the largest objective, and guarantees that no objectives will be too small. This method can be more beneficial if the number of classes is large. The extensive experiments on different datasets show the effectiveness of our proposed method when compared with four state-of-the-art methods.

1 Introduction

High dimensional samples make the learning tasks more complex and computationally demanding. To confront these problems, data dimensionality reduction methods are used to reduce the representation of a dataset since the reduced data delivers greater accuracy and faster learning. Dimensionality reduction algorithms can be classified as *feature selection* and *feature extraction*. Feature selection techniques reduce the representation of a dataset through a reduction in the number of attributes which can learn faster with higher accuracy compared to the initial dataset (Kantardzic 2011). In contrast, feature extraction algorithms transform or project the original data onto a smaller dataset which is more compact and of stronger discriminating power (Nie et al. 2010; 2008; Chang et al. 2014; Chang and Yang 2016; Canedo and Marono 2014; Han, Pei, and Kamber 2011).

Fisher's linear discriminant analysis (FLDA) is a well-known supervised feature extraction method which was proposed by Fisher (1936) for binary classification, and then extended to the multiclass scenario by Rao (1948). The objective of FLDA is to find a transformation matrix from an n-dimensional feature space to a d-dimensional space by maximising the between-class scatter matrix while minimising the within-class scatter matrix. The fast and easy procedure for identifying the transformation matrix has made this method attractive. However, FLDA may encounter the small sample size problem. This problem happens when the number of features is far larger than the number of training samples. In this instance, the within-class scatter matrix may be singular, and the use of LDA may be impossible. Furthermore, FLDA can identify a subspace with the dimensions of at most $c-1$, wherein c is the class number. Overemphasizing the large class distances is another drawback of this method, and this drawback makes the criterion of FLDA suboptimal. In this case, the close class pairs tend to overlap in the subspace, and this is known as the class separation problem.

To deal with the class separation problem, many weighting methods such as the approximate pairwise accuracy criterion (aPAC) (Loog 1999), have been proposed. Loog et al. (2001) employed an approximation of the Bayes error for pairs of classes in order to reduce the merging of close class pairs. Harmonic mean for subspace selection (HMSS) (Bian and Tao 2008) and geometric mean-based subspace selection (GMSS) (Tao et al. 2009), are other weighting methods to reduce the class separation problem. However, these weighting methods still cannot be guaranteed a solution to the class separation problem (Hu et al. 2014).

In order to guarantee the separation of all class pairs, Max-Min distance analysis (MMDA) was introduced by Bian and Tao (2011). MMDA directly maximises the minimum pairwise between-class distance in the low-dimensional subspace. Inspired by the idea of the max-min, several dimensionality reduction methods have been proposed (Hu et al. 2014; Shao and Sang 2012; 2017; Su et al. 2015). However, applying these methods is not feasible when datasets are high dimensional, and when the

number of classes is very large.

Taking all the discussed challenges into consideration, we propose a $\ell_{2,1}$-norm regularised framework based on the trace ratio optimisation for discriminant analysis in this work. Compared with the existing trace ratio methods which aim to maximise the arithmetic mean of multiple objectives, our proposed method aims to maximise the weighted harmonic mean of the objectives. The proposed method avoids the domination problem of the largest objective and guarantees that all objectives will not be too small. More importantly, we employ a $\ell_{2,1}$-norm regularisation term in the framework to exploit correlation between features. Adjusting the regularisation parameter can effectively improve the performance of the proposed method. To seek the optimal solution for the proposed objective function, an efficient iterative algorithm is proposed. Compared with several existing methods in the literature such as MMDA, the process of finding projection matrix is fast and this method converges after a few iterations. In line with the literature for experimental evaluation, we employ 12 image and biological datasets to examine the effectiveness of the proposed method.

The remainder of the paper is organised as follows. In Section 2, several well-known dimensionality reduction methods are briefly explained and the concept of trace ratio is discussed. Based on this concept, the foundation of the proposed regularised trace ratio method is described in Section 3. The converging process of the proposed method, the effect of adjusting the regularisation parameter, and the effectiveness of this method compared with four state-of-the-art methods are examined in Section 4. Section 5 concludes the paper and the experiments.

2 Related Work

In this section, four state-of-the-art methods, namely FLDA (Rao 1948), aPAC (Loog, Duin, and Haeb-Umbach 2001), HMSS (Bian and Tao 2008) and GMSS (Tao et al. 2009) are explained.

2.1 Linear Discriminant Analysis and Extensions

Suppose $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ is the n training data points, and each data point x_i belongs to one of the classes $\{l_1, l_2, \ldots, l_c\}$. FLDA aims to find a projection matrix $W \in \mathbb{R}^{d\times m}(m < d)$ to transform the d-dimensional data point x_i to a m-dimensional data point throughout $W^T x_i$. The criterion to find this matrix is by maximising the ratio between the between-class scatter matrix S_b and within-class scatter matrix S_w. In FLDA, the S_w and S_b are defined as follows:

$$S_w = \sum_{k=1}^{c} \sum_{x_i \in l_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T \tag{1}$$

$$S_b = \sum_{k=1}^{c} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T \tag{2}$$

where n_k is the number of data points belonging to the class l_k, $\bar{x}_k = \frac{1}{n_k}\sum_{x_i \in l_k} x_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. In LDA, if S_w is a non-singular and invertible matrix, W is composed of the at most $c-1$ eigenvectors of $S_W^{-1}S_B$.

In some versions of LDA, the problem of finding a projection matrix is formulated based on the trace ratio (TR) problem. After $W^T X$ transformation and in the lower dimensional subspace, we have the following results:

$$Tr(W^T S_w W) = \sum_{k=1}^{c} \sum_{x_i \in l_k} \left\| W^T (x_i - \bar{x}_k) \right\|_2^2 \tag{3}$$

$$Tr(W^T S_b W) = \sum_{k=1}^{c} n_k \left\| W^T (\bar{x}_k - \bar{x}) \right\|_2^2 \tag{4}$$

The $Tr(W^T S_w W)$ (Eq. 3) measures the Euclidean distances within the same class while $Tr(W^T S_b W)$ (Eq. 4) measures the Euclidean distances between different classes. A common way to maximise the discriminative power under the projection matrix W is to use the ratio of $Tr(W^T S_b W)$ and $Tr(W^T S_w W)$ as the objective function in which the obtained W is an orthogonal projection matrix. The orthogonal constrained trace ratio problem is:

$$\max_{W^T W = I} \frac{Tr(W^T S_b W)}{Tr(W^T S_w W)} \tag{5}$$

This problem does not have a close-form global optimal solution and is approximated by solving a ratio trace problem $Tr\big((W^T S_b W)/(W^T S_w W)\big)$. However, the Eq. 5 has been well studied recently, and the global optimal solution can be efficiently obtained by an iterative algorithm with quadratic convergence. Trace ratio formulation has been used widely in both supervised and unsupervised learning. For instance, Wang et al. (2014) proposed an unsupervised feature selection method using unsupervised trace ratio with $\ell_{2,1}$-norm regularisation term. Extensively empirical results show the trace ratio objective outperforms the traditional ratio trace objective.

2.2 Approximate Pairwise Accuracy Criterion

The approximate pairwise accuracy criterion (aPAC) (Loog, Duin, and Haeb-Umbach 2001) is a weighting scheme in which the approximation of Bayes errors is assigned to individual class pairs. This method attempts to improve upon LDA by reducing the merging of class pairs. In this method, Loog et al. used S_b decomposition (Eq. 6) to rewrite the Fisher criterion as Eq. 7:

$$S_b = \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} \frac{n_k n_j}{n^2} (\bar{x}_j - \bar{x}_k)(\bar{x}_j - \bar{x}_k)^T \tag{6}$$

$$\max_W \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} \frac{n_k n_j}{n^2} Tr\Big(\frac{W^T S_{jk} W}{W^T S_w W}\Big) \tag{7}$$

where $S_{jk} = (\bar{x}_j - \bar{x}_k)(\bar{x}_j - \bar{x}_k)^T$. This equation then was modified by introducing a weighting function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$:

$$\max_W \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} \frac{n_k n_j}{n^2} f(\Delta_{jk}) Tr\Big(\frac{W^T S_{jk} W}{W^T S_w W}\Big) \tag{8}$$

where $\Delta_{jk} := \sqrt{(\bar{x}_j - \bar{x}_k)^T S_W^{-1} (\bar{x}_j - \bar{x}_k)}$; and $f(\Delta_{jk}) = 1/\Delta_{jk}^2 erf\big(\Delta_{jk}/2\sqrt{2}\big)$ in which erf is the error function. The optimal solution of Eq. 8 can be calculated by eigenvalue decomposition as LDA.

2.3 HMSS and GMSS

HMSS (Bian and Tao 2008) and GMSS (Tao et al. 2009) are weighting methods to reduce the class separation problem. These methods represent better performance than FLDA by assigning large weights on close class pairs. HMSS and GMSS are defined as Eq. 9 and Eq. 10 respectively:

$$\max_{W^TW=I} -\sum_{1\le i<j\le c} q_i q_j \big(tr\big((W^T S_w W)^{-1}(W^T D_{ij} W)\big)\big)^{-1} \quad (9)$$

$$\max_{W^TW=I} \prod_{1\le i<j\le c} \big[tr\big((W^T S_w W)^{-1}(W^T D_{ij} W)\big)\big]^{\frac{q_i q_j}{\sum_{1\le m<n\le c} q_m q_n}} \quad (10)$$

where $D_{ij} = (\overline{x}_i - \overline{x}_j)(\overline{x}_i - \overline{x}_j)^T$. These weighting methods still cannot be guaranteed a solution to the class separation problem.

3 Trace Ratio with $\ell_{2,1}$-norm

For the binary class problem, according to the definition of Eq. 1, the within-class scatter matrix S_w^{jk} for the j-th and k-th class is defined as follows:

$$S_w^{jk} = \sum_{h\in\{j,k\}} \sum_{x_i\in l_h} (x_i - \bar{x}_h)(x_i - \bar{x}_h)^T \quad (11)$$

According to the definition of Eq. 2, the between-class scatter matrix S_b^{jk} for the j-th and k-th class is as follows:

$$S_b^{jk} = \sum_{h\in\{j,k\}} n_h (\bar{x}_h - \bar{x}_{jk})(\bar{x}_h - \bar{x}_{jk})^T \quad (12)$$

where $\bar{x}_{jk} = \frac{1}{n_j+n_k}(\sum_{x_i\in l_j} x_i + \sum_{x_i\in l_k} x_i)$ is the mean of all the data from class j and k. So the orthogonal constrained trace ratio problem in the binary class case is:

$$\max_{W^TW=I} \frac{Tr(W^T S_b^{jk} W)}{Tr(W^T S_w^{jk} W)} \quad (13)$$

According to the definitions of S_w, S_w^{jk}, S_b, and S_b^{jk}, Eq. 14 and 15 can be easily verified:

$$S_w = \frac{1}{c-1} \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} S_w^{jk} \quad (14)$$

and

$$S_b = \frac{1}{n} \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} (n_j + n_k) S_b^{jk}. \quad (15)$$

To obtain the Eq. 15, the $(n_j + n_k)S_b^{jk}$ term in the right side of this equation should be expanded based on Eq. 12. By substituting of $(n_j\bar{x}_j + n_k\bar{x}_k)(n_j\bar{x}_j + n_k\bar{x}_k)^T$ instead of $\sum_{h\in\{j,k\}} x_h$, the expansion of the right side of Eq. 15 is equal to the expansion of Eq. 2.

In the traditional trace ratio problem, the sum of between-class distances of all binary classes is maximised and the sum of within-class distances of all binary classes is minimised simultaneously. As a result, the ratios of between-class distances and within-class distances are not explicitly maximised for all the binary classes. Therefore, there might be binary classes that totally overlap, particularly, for datasets with the large number of classes. To overcome this problem, maximising the weighted sum of ratios of between-class and within-class distances for all the binary classes is proposed as:

$$\max_{W^TW=I} \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} (n_j + n_k) \frac{Tr(W^T S_b^{jk} W)}{Tr(W^T S_w^{jk} W)} \quad (16)$$

To control the capacity of W, making it more suitable for feature selection, a regularisation term is added to Eq. 16 as following:

$$\max_{W^TW=I} \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} (n_j + n_k) \frac{Tr(W^T S_b^{jk} W)}{Tr(W^T S_w^{jk} W)} + \alpha\|W\|_{2,1} \quad (17)$$

where α is a regularization parameter and $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$ norm, which can select correlated features to improve the performance. In this equation, $\|W\|_{2,1} = Tr(W^T D W)$, where D is a diagonal matrix with diagonal entries defined as:

$$D^{ii} = \frac{1}{2\|w^i\|_2}, \quad (18)$$

where w^i denotes the i-th row of W. Therefore, D is:

$$D = \begin{bmatrix} \frac{1}{2\|w^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|w^d\|_2} \end{bmatrix} \quad (19)$$

Eq. 17 is the objective function we are using in this paper.

3.1 Weighted Harmonic Mean of Trace Ratio

A simple way to maximise multiple objectives $J_1, J_2, J_3 \ldots$, is maximising the weighted sum of all the objectives as:

$$\max_x \sum_i p_i \cdot J_i(x), \quad (20)$$

where p_i is the weight of the objective J_i. However, in (20), the largest objective J_k could dominate the sum of the objectives, making some other objectives very small. As a result, maximising the arithmetic mean is not a good choice for maximising the multiple objectives. To confront this problem, maximising the minimal of the objectives, as $\max_x \min_i J_i(x)$, can be a solution. Consequently, maximising the weighted harmonic mean of the objectives, as

$\max_x \frac{1}{\sum_i p_i \cdot \frac{1}{J_i(x)}}$, can be used to maximise the multiple objectives. This problem can be converted into:

$$\min_x \sum_i p_i \cdot \frac{1}{J_i(x)}$$

We are minimising the weighted sum of the multiple objective reciprocals here to guarantee that no objective is too small. Therefore, maximising the harmonic mean is a good solution for maximising multiple objectives. Based on these analyses, we propose to maximise the harmonic mean of the trace ratios to overcome the problem raised in (16). Similarly, we will minimise the following objective function:

$$\min_{W^TW=I} \sum_{k=1}^{c-1} \sum_{j=k+1}^{c} (n_j + n_k) \frac{Tr(W^T S_w^{jk} W)}{Tr(W^T S_b^{jk} W)} + \alpha Tr(W^T D W) \tag{21}$$

The main difference of Eq. 21 with our previous work (Li et al. 2017) is that we incorporate a $\ell_{2,1}$-norm into the framework. For notation simplicity, we rewrite the Eq. 21 as Eq. 22 and continue the further analysis using this equation:

$$\min_{W^TW=I} \sum_{i=1}^{\acute{c}} \frac{Tr(W^T A_i W)}{Tr(W^T B_i W)} + \alpha Tr(W^T D W) \tag{22}$$

where: $\acute{c} = \frac{c(c-1)}{2}$, $B_i = S_b^{jk}$, and $A_i = (n_j + n_k) S_w^{jk}$, in which j and k are i and $i+1$ respectively. The Lagrangian function of (22) is:

$$f(W, \Lambda) = \sum_{i=1}^{\acute{c}} \frac{Tr(W^T A_i W)}{Tr(W^T B_i W)} + \alpha Tr(W^T D W) - Tr(\Lambda(W^T W - I)) \tag{23}$$

By taking the derivative of Eq. 23 with respect to W, we have:

$$\left(\sum_{i=1}^{\acute{c}} \frac{1}{Tr(W^T B_i W)} \left(A_i - \frac{Tr(W^T A_i W)}{Tr(W^T B_i W)} B_i \right) + \alpha D \right) \times W = W\Lambda \tag{24}$$

which can be rewritten as $MW = W\Lambda$, where the matrix M is:

$$M = \sum_{i=1}^{\acute{c}} \frac{1}{Tr(W^T B_i W)} \left(A_i - \frac{Tr(W^T A_i W)}{Tr(W^T B_i W)} B_i \right) + \alpha D \tag{25}$$

Note that M in Eq. 25 is dependent on W, which is unknown. To seek the optimal solution for W, we propose the iterative Algorithm 1. In the rest of this paper, we call our proposed trace ratio with $\ell_{2,1}$-norm regularisation as TRLN. In each iteration of Algorithm 1, the value of Eq. 22 is decreased until the algorithm converges. The proof is similar to (Li et al. 2017; Nie et al. 2010) and thus we omit it due to limited space. It is worth noting that Eq. 24 is the KKT condition of Eq. 22, and consequently is a local optimal solution to the problem (22).

Algorithm 1: TRLN

Input: Training data $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, and m

Initialize $W \in \mathbb{R}^{d \times m}$ such that $W^TW = I$;
calculate S_w^{jk} according to Eq. 11;
calculate S_b^{jk} according to Eq. 12;
calculate D according to Eq. 19 ;
while *not converge* **do**
    1. calculate M according to Eq. 25;
    2. update W, which is formed by the m eigenvectors of M corresponding to the m smallest eigenvalues;
    3. update D;
end

Output: $W \in \mathbb{R}^{d \times m}$

4 Experiments and Discussion

In this section, the converging process of the proposed algorithm and the effect of the regularisation parameter on the predictive performance of learning algorithms are discussed. In line with the literature, we compare the performance of the proposed method with four well-known methods to examine the effectiveness of the TRLN. In our experiments, the following datasets are considered:

- **Biology datasets** including Obesity, CLL_SUB_111 (Haslinger et al. 2004), GLIOMA (Nutt et al. 2003), TOX_171 from the Arizona State University (ASU) datasets repository[1], lung (Bhattacharjee et al. 2001) and Carcinom (Su et al. 2001). Obesity is a high-dimensional data collected within our institution, which contains 114 obese and non-obese people with 47,233 biomarkers per person. CLL_SUB_111 contains gene expression for 100 genetically well-characterized B-CLL samples, and 11 healthy control samples. GLIOMA has 50 high-grade glioma samples in which glioma is defined as cancer of the brain. Lung has 203 samples from snap-frozen lung tumors and normal lung. Carcinom contains 174 samples for carcinomas of the prostate, breast, lung, ovary, colorectum, kidney, liver, pancreas, bladder/ureter, and gastroesophagus.
- **Image datasets** including pixraw10P from the ASU datasets repository, Yale (Belhumeur, Hespanha, and Kriegman 1997), COIL20 (Nene et al. 1996), YaleB (Georghiades, Belhumeur, and Kriegman 2001) and ORL (Samaria and Harter 1994). Yale contains 165 gray-scale images of 15 individuals (11 images for each subject). COIL20 has 20 objects and 72 gray-scale images for each object under various poses (total of 1,440 images). YaleB is the extended Yale face database B and contains 2,414 images for 38 individuals. ORL has 400 gray-scale images from 40 distinct subjects (10 images for each subject). We also use orlraws10P in our experiments which is extracted from the ORL dataset and contains 10 subjects.

The characteristics of these data are presented in Table 1.

[1] http://featureselection.asu.edu/datasets.php

Table 1: The characteristics of the datasets

Data sets	Sample	Feature	Class
Obesity	114	47233	2
CLL_SUB_111	111	11340	3
GLIOMA	50	4434	4
TOX_171	171	5748	4
lung	203	3312	5
pixraw10P	100	10000	10
orlraws10P	100	10304	10
Carcinom	174	9128	11
Yale	165	1024	15
COIL20	1440	1024	20
YaleB	2414	1024	38
ORL	400	1024	40

4.1 The Algorithm Convergence

In order to evaluate the convergence of the proposed algorithm, the values of m and α were altered within $\{c-1,\ldots,1\}$ and $\{0,10^{-3},\ldots,10^{1}\}$ respectively. The $\alpha=0$ means that the regularisation term is removed from Eq. 21. In the converging process, we consider the convergence threshold as 0.05 and the maximum number of iterations as 30. In all experiments reported in this paper, PCA was performed as a pre-processing step while 95% of the total variance was captured by reduced datasets. In order to prevent the singularity problem, a small term λI was added to the covariance matrix, where λ is 10^{-5}. All experiments were implemented on a PC with 3.2 GHz CPU frequency and 8 GB memory.

The experiments showed that the algorithm convergence does not depend on the value of α and the variation of m. Furthermore, not only the converging process of this algorithm happens after a few iterations, this process is also fast. Based on the experiments, the spending time for finding the transformation matrix for each α varies between 0.012 and 0.73 second for different datasets used in this study. Figure 1 illustrates this trend for $m=c-1$ for various datasets. The spending time for performing the TRLN is also illustrated.

4.2 The Effect of The Regularisation Parameter

To investigate the effect of adjusting the parameter α in our proposed method, this variable is altered within $\{0,10^{-3},10^{-2},\ldots,10^{1}\}$ for each dimension belonging to $\{c-1,\ldots,1\}$. $\alpha=0$ means that the $\ell_{2,1}$-norm regularisation which is the correlation exploitation, is removed from Eq. 21. The Eq. 21 with $\alpha=0$ is equivalent to the work introduced by Li et al. (2017). The 1-nearest neighbours classifier (1NN) was employed to examine the predictive performance (accuracy) of the TRLN. Reducing data, and measuring the predictive performance of the 1NN classifier were performed by 5-fold cross validation. In order to ensure the results are not biased towards the data sequence, the process of data reduction and classification are repeated 5 times after shuffling datasets, and the average of classification accuracies is reported as the predictive performance.

Table 2 summarises the effect of parameter α on average predictive performance of four datasets for the variation of

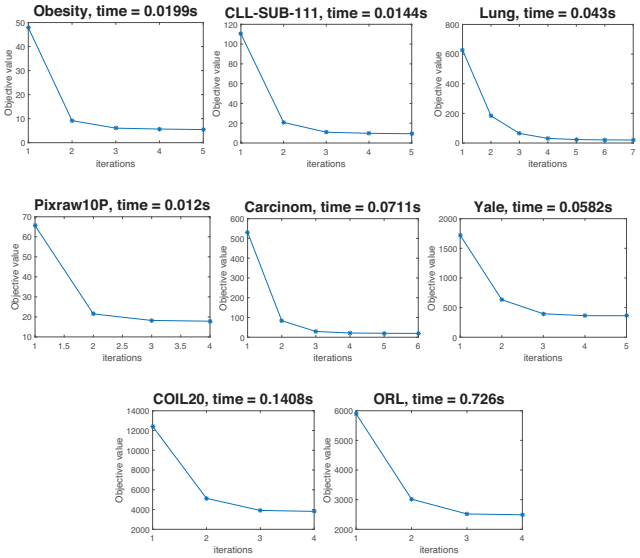


Figure 1: The converging trend of the TRLN for different datasets with $m=c-1$ and $\alpha=1$.

dimension. For each dimension, the maximum and the average of the predictive performance of 1NN classifier are illustrated. Furthermore, the best value of α leading to the maximum predictive performance, is also shown. Figure 2 illustrates the accuracy of 1NN applied to the orlraws10P data for the variation of α and m.

These results indicate the importance of adding $\ell_{2,1}$-norm to the TRLN algorithm, and the effect of regulating parameter α for each dimension. The almost 10% deviation of the 1NN performance for some dimensions, implies that adjusting α leads to better performance than simply excluding the regularisation term in the TRLN algorithm. The only excep-

Table 2: The maximum and the average accuracy of the 1NN classifier for the four datasets with respect to variation of α and m. The best α leading to the best predictive performance is also represented.

Dim	Max	Mean	Best
Orlraws10P			
9	98.60	95.20	10
8	99.20	94.93	10
7	98.20	95.23	10
6	99.80	96.17	10
5	97.40	96.53	10
4	95.80	93.97	10
3	86.40	83.77	0.001
2	68.60	62.90	0.01
1	42.60	38.70	10
TOX_171			
3	96.96	86.82	10
2	72.98	69.43	0.001
1	44.33	41.93	1

Dim	Max	Mean	Best
Pixraw10P			
9	99.80	98.97	10
8	99.40	98.90	10
7	99.20	98.67	0.001
6	99.60	98.53	10
5	99.80	98.63	10
4	100.00	98.80	10
3	93.80	92.87	0.01
2	84.80	82.20	0.1
1	53.80	52.00	0
GLIOMA			
3	72.00	60.20	10
2	69.60	62.07	1
1	48.40	44.40	0.01

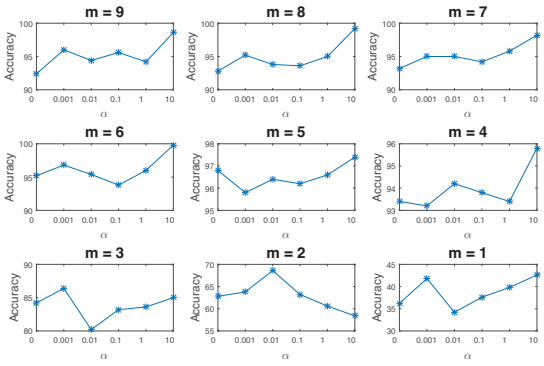


Figure 2: The effect of altering the regularisation parameter (α) along different dimensions for the orlraws10P dataset.

tion is the pixraw10P when $m = 1$. In this dimension, the performance of the algorithm for both cases of including and excluding the regularisation term is quite poor. Similar patterns and results were achieved for the other datasets used in this paper.

In the rest of this paper, we alter the value of α within $\{10^{-3}, \ldots, 10^{1}\}$ for each dimension as the performance gain over our earlier work (Li et al. 2017) demonstrated the effectiveness of $\ell_{2,1}$-norm regularisation for dimension reduction. To obtain the best value for the regularisation parameter that leads to highest average predictive performance, the proposed method is executed five times for each m. However, this process is fast as the algorithm converges after a few iterations.

4.3 The Effectiveness of The Proposed Method Compared with Previous Methods

In this section, the performance of the TRLN method is compared with four state-of-the-art methods, namely FLDA (Rao 1948), aPAC (Loog, Duin, and Haeb-Umbach 2001), HMSS (Bian and Tao 2008) and GMSS (Tao et al. 2009). The performance of a data reduction method should be evaluated with a classification algorithm. However, the predictive performance depends on the learning algorithm used. In order to obtain results as classifier-independent as possible, we employed two classifiers in this study, namely the 1NN classifier and the linear support vector machine (SVM). In the linear SVM, the box constraint parameter (C) remained constant for each dataset to perform a fair comparison between different methods. However, in order to select the best C, this parameter was altered within the range of $\{2^{-3}, \ldots, 2^{5}\}$. Afterwards, a C which led to the highest average predictive performance was considered as the best value for further experiments.

We followed the same procedure explained in section 4.2 for the classification, that is, the accuracy of the two classifiers was measured by 5-fold cross validation and for five times repetitions. The results of experiments are illustrated in Figure 3, Tables 3 and 4. Figure 3 represents the accuracy of the 1NN classifier for different datasets. The performance of the ORL dataset has been illustrated in two figures to make a better comparison of the five methods. The ORL(a) and ORL(b) show the accuracies associated with the dimensions belonging to $\{c-1, \ldots, 10\}$ and $\{10, \ldots, 1\}$ respectively. Tables 3 and 4 also show the best predictive performance of 1NN and linear SVM for the 12 datasets and the five methods. The dimension which leads to the highest accuracy is also presented in parentheses for each method and dataset in these tables.

Figure 3 shows that while the proposed method underperformed for small dimensions, it can significantly outperform the other methods for dimensions not much smaller than $c-1$. For example, at $m = c-1$, the predictive performance of 1NN obtained by applying TRLN to the CLL_SUB_111 and TOX_171 datasets, are higher than the other methods by almost 11% and 14% respectively. This figure changes to 9% and 15% by employing the linear SVM (Table 4). The two tables also emphasize the outstanding performance of TLRN compared with other methods with respect to achieved highest accuracy. For instance, for the TOX_171, CLL_SUB_111, Carcinom, Yale and GLIOMA datasets the predictive performance of the proposed method is higher than the other methods by almost 14%, 9%, 5%, 5%, and 4% respectively. The only exception is the COIL20 data in which the HMSS method has the higher 1NN and linear SVM performance than TRLN by at most 0.13%. The dimensions represented in parentheses show that the highest predictive performance for all methods is achieved for not very small values of m.

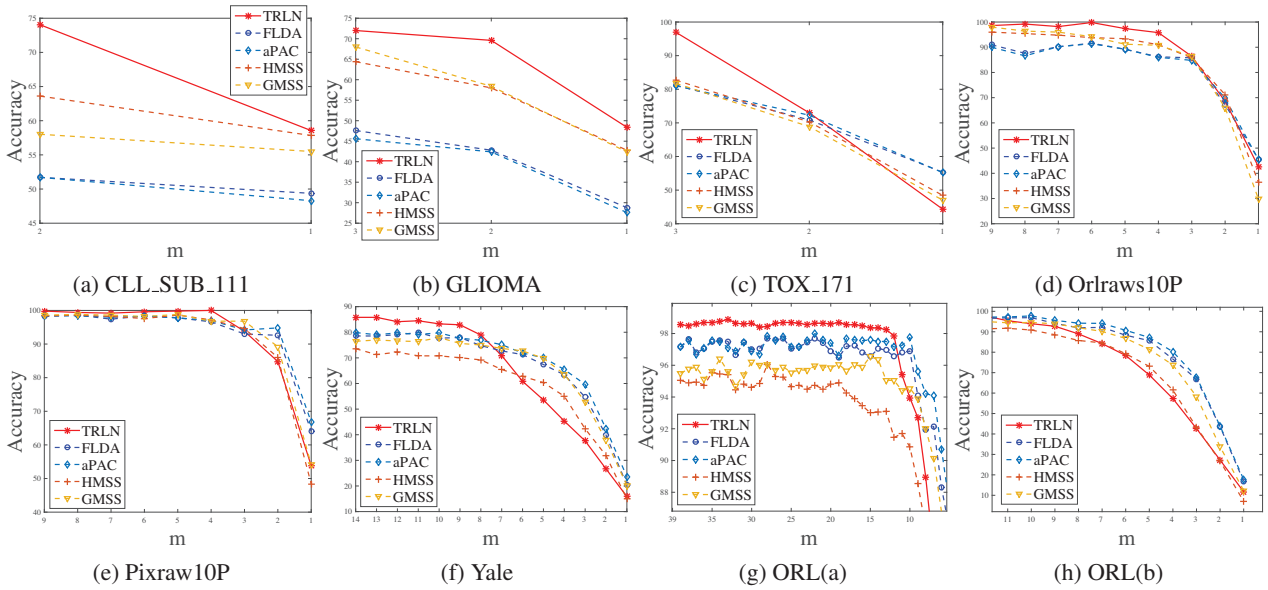
5 Conclusion

In this study, we propose a $\ell_{2,1}$-norm regularised weighted trace ratio to address the domination problem of the largest class distance leading to the overlapped close class pairs in the subspace. The proposed method TRLN maximises the harmonic mean of a new weighted sum of ratios of between-class and within-class distances for all the binary classes. Employing the harmonic mean guarantees that no objective will be too small. We also added a regularisation term to improve the performance of the trace ratio method. To evaluate the performance of the proposed method, we employed 12 various datasets.

The experiments showed that the algorithm converges after a few iterations. Moreover, the importance of adding the regularisation term and adjusting the regularisation parameter was comprehensively discussed. We showed that the higher predictive performance can be achieved by adjusting the regularisation parameter rather than simply excluding this term in the proposed method. The performance of the TRLN was also compared with the four well-known methods, namely FLDA, aPAC, HMSS and GMSS. To do so, we employed 1NN and linear SVM and applied them to the reduced subsets. The results showed that while the proposed method cannot compete with the other methods for small values of dimension, it significantly outperforms the four methods when the value of m is $c-1$ and near this value.

References

Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific



(a) CLL_SUB_111 (b) GLIOMA (c) TOX_171 (d) Orlraws10P

(e) Pixraw10P (f) Yale (g) ORL(a) (h) ORL(b)

Figure 3: The accuracy of 1NN classifier for different datasets with respect to variation of m

Table 3: The best predictive performance of 1NN obtained from TRLN, FLDA, aPAC, HMSS, and GMSS methods.

Data sets	**TRLN**	**FLDA**	**aPAC**	**HMSS**	**GMSS**
Obesity	**64.74±4.04**(1)	64.04±2.70(1)	64.04±2.70(1)	62.11±1.14(1)	57.19± 3.90(1)
Cll_sub_111	**74.05±3.62(2)**	51.71±1.86(2)	51.71±1.86(2)	63.04±2.07(2)	58.02±5.35(2)
Glioma	**72±5.09(3)**	47.6±6.23(3)	45.6±4.33(3)	64.4±5.89(3)	68±4.69(3)
TOX_171	**96.96±2.24(3)**	81.17±3.68(3)	81.05±4.15(3)	82.69±3.76(3)	81.75±1.5(3)
Lung	**95.17±0.41(4)**	74.58±3.05(4)	74.38±2.39(4)	93.50±1.79(3)	94.88±0.66(4)
Pixraw10P	**100±0(4)**	98.6±0.89(8)	98.4±0.89(8)	98.6±1.14(5)	98.8±0.84(5)
Orlraws10P	**99.8±0.45(6)**	91.6±2.51(6)	91.4±2.07(8)	96±1.58(9)	98±0.71(9)
Carcinom	**95.17±0.31(10)**	69.54±1.95(9)	69.54±1.95(9)	85.63±0.81(9)	89.43±0.87(10)
Yale	**85.82±2.76(13)**	79.88±1.63(11)	79.88±0.66(14)	73.58±2.62(14)	77.93±1.63(10)
COIL20	99.44±0.2(8)	99.07±0.14(15)	99.15±0.11(15)	**99.57±0.12(18)**	94.78±0.40(19)
YaleB	**93.29±0.34(29)**	92.55±0.37(19)	92.87±0.26(19)	82.40±0.91(15)	88.12±0.22(22)
ORL	**98.9±0.52(33)**	97.7±0.60(22)	98±0.39(22)	96.05±0.622(28)	96.55±0.65(15)

Table 4: The best predictive performance of the linear SVM obtained from TRLN, FLDA, aPAC, HMSS, and GMSS methods.

Data sets	**TRLN**	**FLDA**	**aPAC**	**HMSS**	**GMSS**
Obesity	**64.74±4.04**(1)	64.04±2.70(1)	64.04±2.70(1)	61.05±2.37(1)	57.19± 3.90(1)
Cll_sub_111	**74.23±1.97(2)**	52.07±2.33(2)	52.07±2.33(2)	65.04±2.66(2)	62.88±3.95(2)
Glioma	**72.4±7.4(3)**	48.4±6.84(3)	46±3.74(3)	67.2±5.93(3)	68.4±5.37(3)
Tox_171	**95.67±1.97(3)**	81.05±3.5(3)	80.94±3.57(3)	80.94±4.29(3)	79.42±2.32(3)
Lung	**95.37±0.44(4)**	73.6±2.57(4)	73.99±2.19(4)	94.29±1.13(4)	94.48±0.64(4)
Pixraw10P	**99.8±0.45(5)**	98.6±0.89(6)	98.4±0.89(8)	98.8±0.84(8)	99±0.71(6)
Orlraws10P	**99±0.71(8)**	91.8±2.17(6)	91.6±1.95(6)	96.4±1.52(9)	97.8±0.45(9)
Carcinom	**94.71±0.63(10)**	69.54±1.95(9)	69.54±1.95(9)	87.36±0.91(10)	89.08±1.68(10)
Yale	**84.24±1.96(14)**	79.39±3.18(13)	80±3.69(12)	76.36±0.61(14)	80.12±2.82(13)
COIL20	99.13±0.06(19)	98.88±0.06(16)	98.94±0.27(19)	**99.15±0.16(19)**	94.02±0.71(19)
Yaleb	**94.08±0.71(32)**	93.72±0.27(36)	93.76±.34(36)	93.27±0.29(36)	93.95±0.35(35)
ORL	**98.7±0.41(29)**	97.90±0.78(34)	98.05±0.21(22)	97.5±0.68(38)	97.5±0.77(15)

linear projection. *IEEE Transactions on pattern analysis and machine intelligence* 19(7):711–720.

Bhattacharjee, A.; Richards, W. G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Ladd, C.; Beheshti, J.; Bueno, R.;

Gillette, M.; et al. 2001. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* 98(24):13790–13795.

Bian, W., and Tao, D. 2008. Harmonic mean for subspace selection. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 1–4. IEEE.

Bian, W., and Tao, D. 2011. Max-min distance analysis by using sequential sdp relaxation for dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(5):1037–1050.

Canedo, V. B., and Marono, N. S. 2014. *Novel feature selection methods for high dimensional data*. Ph.D. Dissertation, Universidade da Coruña.

Chang, X., and Yang, Y. 2016. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE transactions on neural networks and learning systems*.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI Conference on Artificial Intelligence*, 1171–1177.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2):179–188.

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on pattern analysis and machine intelligence* 23(6):643–660.

Han, J.; Pei, J.; and Kamber, M. 2011. *Data mining: concepts and techniques*. Elsevier.

Haslinger, C.; Schweifer, N.; Stilgenbauer, S.; Do?hner, H.; Lichter, P.; Kraut, N.; Stratowa, C.; and Abseher, R. 2004. Microarray gene expression profiling of b-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and vh mutation status. *Journal of Clinical Oncology* 22(19):3937–3949.

Hu, J.; Deng, W.; Guo, J.; and Xu, Y. 2014. Max-k-min distance analysis for dimension reduction. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 726–731. IEEE.

Kantardzic, M. 2011. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Li, Z.; Nie, F.; Chang, X.; and Yang, Y. 2017. Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2100–2110.

Loog, M.; Duin, R. P. W.; and Haeb-Umbach, R. 2001. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(7):762–766.

Loog, M. 1999. Approximate pairwise accuracy criteria for multiclass linear dimension reduction: Generalisations of the fisher criterion. *WBBM Report Series 44*.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20).

Nie, F.; Xiang, S.; Jia, Y.; Zhang, C.; and Yan, S. 2008. Trace ratio criterion for feature selection. In *AAAI Conference on Artificial Intelligence*, volume 2, 671–676.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, 1813–1821.

Nutt, C. L.; Mani, D.; Betensky, R. A.; Tamayo, P.; Cairncross, J. G.; Ladd, C.; Pohl, U.; Hartmann, C.; McLaughlin, M. E.; Batchelor, T. T.; et al. 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer research* 63(7):1602–1607.

Rao, C. R. 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)* 10(2):159–203.

Samaria, F. S., and Harter, A. C. 1994. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, 138–142. IEEE.

Shao, G., and Sang, N. 2012. Fractional-step max-min distance analysis for dimension reduction, feature selection. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, 396–400. IEEE.

Shao, G., and Sang, N. 2017. Regularized max-min linear discriminant analysis, feature selection. *Pattern Recognition*.

Su, A. I.; Welsh, J. B.; Sapinoso, L. M.; Kern, S. G.; Dimitrov, P.; Lapp, H.; Schultz, P. G.; Powell, S. M.; Moskaluk, C. A.; Frierson, H. F.; et al. 2001. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research* 61(20):7388–7393.

Su, B.; Ding, X.; Liu, C.; and Wu, Y. 2015. Heteroscedastic max-min distance analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4539–4547.

Tao, D.; Li, X.; Wu, X.; and Maybank, S. J. 2009. Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2):260–274.

Wang, D.; Nie, F.; and Huang, H. 2014. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 306–321. Springer.