Published in final edited form as:
IEEE Signal Process Lett. 2013 March 7; 20(5): 431–434. doi:10.1109/LSP.2013.2250281.

Trace Ratio Linear Discriminant Analysis for Medical Diagnosis: A Case Study of Dementia

Mingbo Zhao,
Electrical Engineering Department, City University of Hong Kong, Kowloon, Hong Kong SAR

Rosa H. M. Chan,
Electrical Engineering Department, City University of Hong Kong, Kowloon, Hong Kong SAR

Peng Tang,
Electrical Engineering Department, City University of Hong Kong, Kowloon, Hong Kong SAR

Tommy W. S. Chow, and
Electrical Engineering Department, City University of Hong Kong, Kowloon, Hong Kong SAR

Savio W. H. Wong
Psychological Studies, Hong Kong Institute of Education, N.T., Hong Kong SAR

Mingbo Zhao: mzhao4@cityu.edu.hk; Rosa H. M. Chan: rosachan@cityu.edu.hk; Peng Tang: rollegg@gmail.com; Tommy W. S. Chow: eetchow@cityu.edu.hk; Savio W. H. Wong: savio@ied.edu.hk

Abstract

Dementia is one of the most common neurological disorders among the elderly. Identifying those who are of high risk suffering dementia is important to the administration of early treatment in order to slow down the progression of dementia symptoms. However, to achieve accurate classification, significant amount of subject feature information are involved. Hence identification of demented subjects can be transformed into a pattern recognition problem with high-dimensional nonlinear datasets. In this paper, we introduce trace ratio linear discriminant analysis (TR-LDA) for dementia diagnosis. An improved ITR algorithm (iITR) is developed to solve the TR-LDA problem. This novel method can be integrated with advanced missing value imputation method and utilized for the analysis of the nonlinear datasets in many real-world medical diagnosis problems. Finally, extensive simulations are conducted to show the effectiveness of the proposed method. The results demonstrate that our method can achieve higher accuracies for identifying the demented patients than other state-of-art algorithms.

Index Terms

Dimensionality reduction; feature extraction; medical diagnosis

I. Introduction

Dementia, which causes a progressive decline in cognitive functions, is one of the most common neurolog-population, its prevalence is expected to increase [1]. However, there exists considerable regional variation in diagnosis practice because of the differences in available resources even within a country, e.g. lack of trained general practitioners and/or

The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kjersti Engan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

time to administer and analyze full cognitive function assessments. For example, it was approximated that only a third of people who were actually suffering dementia in the US ever received a formal medical diagnosis. Thus, limited patients suffering dementia are offered appropriate medical treatment or care, which can potentially slow down the progression of symptoms. To separate probably or possibly demented patients from normal subjects, a large amount of data with features for describing symptoms are currently required [1]. In that way, the identification of demented subjects can be transformed into a pattern recognition problem with a high-dimensional dataset.

But dealing with high-dimensional data has always been a major problem in pattern recognition. Hence finding a low-dimensional representation of high-dimensional data, namely dimensionality reduction is thus of great practical importance. Among the dimensionality reduction methods, linear discriminant analysis (LDA) [10] is the most popular method, which is to find the optimal low-dimensional presentation by maximizing the between-class scatter matrix while minimizing the within-class scatter matrix. Several variants of LDA have been proposed during the past decades, and trace ratio LDA (TR-LDA) is one of the most widely used variants [2], [11], [12]. TR-LDA is based on the trace ratio criterion, which can directly reflect Euclidean distances between data points of inter- and intra-classes. In addition, the optimal projection obtained by TR-LDA is orthogonal. As described in [2], when evaluating the similarities between data points based on Euclidean distance, the orthogonal projection can preserve such similarities without any change. Thus, TR-LDA tends to perform empirically better than the classical LDA and other variants of LDA in many problems.

In this paper, improved ITR algorithm (iITR), an efficient algorithm is proposed for solving TR-LDA problem, which can handle nominal attributes and missing values in many real-world medical diagnosis problems. To validate the effectiveness of the proposed method to assist medical screening, the performance of TR-LDA with iITR and other state-of-art dimensionality reduction methods will be compared here by a case study in the screening of demented subjects using only demographic data, medical history, and behavioral attributes, without the use of cognitive function assessment data. In our current study, results show that TR-LDA method can assist the identification of demented patients with higher accuracies even with less training data comparing to other state-of-art dimensionality reduction methods. The proposed dimensionality reduction method can be incorporated into computational screening program to identify probable or possible patients such that general practitioners can refer these subjects to specialists for full diagnosis.

II. Trace Ratio Linear Discriminant Analysis

A. Review of Linear Discriminant Analysis

LDA uses the within-class scatter matrix S_w to evaluate the compactness within each class and between-class scatter matrix S_b to evaluate the separability of different classes. The goal of LDA is to find a linear transformation matrix $W \in R^{D\times d}$, for which the between-class scatter matrix is maximized, while the within-class scatter matrix is minimized. Let $X = \{x_1, x_2, \dots X_l\} \in R^{D\times l}$ be the training set, each x_i belongs to a class $c_i = \{1, 2, \dots c\}$. Let l_i be the number of data points in the ith class and l be the number of data points in all classes. Then, the between-class scatter matrix S_b, within-class scatter matrix S_w, and total-class scatter matrix S_t are defined as follows:

$$\begin{aligned} S_t&=\sum_{i=1}^{c}\sum_{x\in c_i}(x-\mu)(x-\mu)^T\\ S_w&=\sum_{i=1}^{c}\sum_{x\in c_i}(x-\mu_i)(x-\mu_i)^T \\ S_b&=\sum_{i=1}^{c}l_i(\mu_i-\mu)(\mu_i-\mu)^T \end{aligned}\quad(1)$$

where $\mu_i = 1/l_i \Sigma_{x_j\in c_i} x_j$ is the mean of the data points in the *i*th class, and $\mu=1/l\sum_{i=1}^{l}x_i$ is the mean of the data points in all classes. The original formulation of LDA, called Fisher LDA [10], can only deal with binary classification. Two optimization criteria can be used to extend Fisher LDA to solve the multi-class classification problem. The first one is in the ratio trace form (we refer it as LDA):

$$W^*=\arg\max_W Tr\left\{\left(W^TS_wW\right)^{-1}W^TS_bW\right\}\quad(2)$$

and the second one is in the trace ratio form (we refer it as TR-LDA):

$$W^*=\arg\max_{W^TW=I}\frac{Tr\left(W^TS_bW\right)}{Tr\left(W^TS_wW\right)}\quad(3)$$

The optimal solution of LDA can be formed by the top eigenvectors of $S_w^{-1}S_b$. On the other hand, the optimization problem of TR-LDA in (3) has no close-form solution and has to calculate it by an Iterative Trace Ratio method (ITR) [7]. Specifically, if W_t denotes the solution at the *tth* iteration, then at the $(t + 1)th$ solution, W_{t+1} can be formed by the top eigenvectors of $S_b - \lambda_t S_w$, where $\lambda_t=Tr(W_t^TS_bW_t)/Tr(W_t^TS_wW_t)$. This procedure can be proved to converge to the globally optimal solution given any initialization W_0 [2].

B. A More Efficient Algorithm for Solving the TR Problem

Though the ITR algorithm works well for solving the TR problem, it has its own drawback. The ITR algorithm method has chosen d eigenvectors corresponding to the d largest eigenvalues of $S_b - \lambda^* S_w$ to form W^*. These eigenvectors can only maximize the trace difference value $Tr(W^T(S_b - \lambda^* S_w) W)$, but these eigenvectors cannot maximize trace ratio value $TR(W_t^TS_bW_t)/Tr(W_t^TS_wW_t)$. Thus, how to find eigenvectors to maximize the trace ratio value is an important question. Motivated by this issue, we then, in this subsection, propose a more efficient algorithm, called improved ITR algorithm (iITR), which can solve this problem.

Given any initial λ_t, by performing the eigen-decomposition of $S_b - \lambda_t S_w$, we can obtain the D eigenvectors of $S_b \lambda_t S_w$. The problem is then to choose the d eigenvectors $W_t = \{w_{i_1}, w_{i_2}, \ldots, w_{i_D}\}$ maximizing $\sum_{k=1}^{d}w_{i_k}^TS_bw_{i_k}/\sum_{k=1}^{d}w_{i_k}^TS_ww_{i_k}$, where $i = \{i_1, i_2, \ldots, i_d\}$ is a certain permutation chosen from $\{1, 2, \ldots, D\}$. Here, if we define $f = \{f_1, f_2, \ldots, f_D\} \in R^{1\times D}$, $g = \{g_1, g_2, \ldots, g_D\} \in R^{1\times D}$ with each element satisfying $f_i=w_i^TS_bw_i$ and $g_i=w_i^Ts_ww_i$, the above problem can be converted to find the optimal selection vector $b = \{b_1, b_2, \ldots b_D\} \in R^{1\times D}$ as:

$$\begin{aligned} b^*=\arg\max_b\frac{f_1b_1+f_2b_2+\cdots+f_Db_D}{g_1b_1+g_2b_2+\cdots+g_Db_D}.\\ subject\,to\,b_i\in\{0,1\},b1^T=d\end{aligned}\quad(4)$$

Note that the above problem is a linear fractional programming (LFP) problem [4], [9], [14]. It can be solved by Dinkelbach's algorithm which is a general algorithm for optimizing $\gamma = \Phi(b)/\Psi(b)$ with $\Psi(b) > 0$. In Dinkelbach's algorithm, it converts the problem to a sequence of sub-problems for optimizing $\Phi(b) - \gamma\Psi(b)$. Hence in our case, by initializing $\gamma_0 = \lambda_t$ and let f, g be defined as above, the optimal selection vector b^* can then be obtained by iteratively solving the following sub-problem:

$$\left\{\begin{array}{c}\gamma_0=\lambda_t\\ f,g\end{array}\right\}\rightarrow\left\{\begin{array}{c}b^k=\arg\max_b b(f-\gamma_k g)^T\\ subject\,to\,b_i^k\in\{0,1\},b^k1^T=d\\ \gamma_{k+1}=\frac{b^kf^T}{b^kg^T}\end{array}\right\}\\ \rightarrow\left\{\begin{array}{c}b^*:b^k=b^{k-1}\\ \gamma^*=\frac{b^*f^T}{b^*g^T}\end{array}\right\} \quad (5)$$

After b^* is obtained, we can output W_t by choosing the d eigenvectors with $b_i^*=1$. The basic steps of the algorithm are listed in Table I.

C. Convergence Analysis of iITR Algorithm

Here the convergence of the proposed iITR algorithm is also analyzed. In fact, as pointed in [3], [13], the algorithm of TR-LDA is Newton method, hence the convergence rate is quadratic and the very fast convergence of the algorithm of TR-LDA is theoretically guaranteed. It has been rigorously proved that for ITR algorithm, given any initial λ_t, the updated λ_{t+1} satisfying 1) $\lambda_{t+1}^{ITR} \geq \lambda_t$ and 2) $\lambda_{t+1}^{ITR} \leq \lambda^*$. Hence we only need to prove that for the proposed iITR algorithm, the updated λ_{t+1}^{iITR} is no smaller than λ_{t+1}^{ITR}. Following (5), this can be equivalent to prove that given the initial $\gamma_0 = \lambda_t$, the updated γ_{k+1} satisfying i) $\gamma_{k+1} \geq \gamma_k$ and ii) $\gamma_{k+1} \leq \gamma^*$. We next prove the two inequalities.

Proof—Let $h(\gamma_k) = \max_b b(f - \gamma_k g)^T$, since $\gamma_{k+1} = b^k f^T / b^k g^T$, we have $b^k f^T - \gamma_{k+1} b^k g^T = 0 \rightarrow b^k (f - \gamma_{k+1} g)^T = 0$. In addition, since $b^{k+1} = \arg\max_b b(f - \gamma_{k+1} g)^T$, it follows $h(\gamma_{k+1}) = b^{k+1} (f - \gamma_{k+1} g)^T \geq b^k (f - \gamma_{k+1} g)^T = 0$. This indicates that $h(\gamma_{k+1}) \geq 0$. We then have $h(\gamma_{k+1}) \geq 0 \rightarrow b^{k+1} f^T / b^{k+1} g^T \geq \gamma_{k+1} \rightarrow \gamma_{k+2} \geq \gamma_{k+1}$. By simply performing the notation substitution, i.e. $k + 1 \rightarrow k$, we thus prove the first inequality $\gamma_{k+1} \geq \gamma_k$. We next prove the second inequality. Recall that $\gamma^* = \max_b bf^T / bg^T = b^* f^T / b^* g^T$, it follows $b^* f^T - \gamma^* b^* g^T = 0 \rightarrow b^* (f_T - \gamma^* g)^T = 0$. Since $h(\gamma^*) = \max b(f - \gamma^* g)^T = b^*(f - \gamma^* g) = 0$, it can be rewritten as $h(\gamma^*) = h(\gamma^{k+1}) + (\gamma^{k+1} - \gamma^*) g^T = 0$. Note that $h(\gamma_{k+1}) \geq 0$ and g is a semi-positive vector, the equality can only holds as $\gamma^{k+1} \leq \gamma^*$, hence we prove the second inequality, i.e. $\gamma_{k+1} \leq \gamma^*$.

III. Identifying Demented Patients via TR-LDA

A. Data Descriptions

The proposed method will be used to screen the demented subjects which meet the criteria for dementia in accordance with standard criteria for dementia of the Alzheimer's type or other non-Alzheimer's demented disorders in their first visits to Alzheimer disease Centers (ADCs) throughout the United States. Data from 289 demented subjects and 9611 controls collected by approximately 29 ADCs from 2005 to 2011 are studied. These data are organized and made available by the US National Alzheimer's Coordinating Center (NACC). Among the demented patients studied, 97% of them were classified as probable or possible Alzheimer's disease (AD) patients. Those with dementia and with neither probable AD nor possible AD have other types of dementia such as Dementia with Lewy Bodies, and Frontotemporal Lobar Degeneration. 5 nominal, 142 ordinal, and 9 numerical attributes of the subjects are included in the study. These attributes include demographic data, medical history, and behavioral attributes, with 5% being missing values. To make the classification problem more difficult, no cognitive assessment variable, such as Mini-Mental State Examination score, is included as attribute.

B. Prediction Stage

The next step is to apply TR-LDA in identifying demented patients from normal persons. Note that NACC dataset includes nominal attributes and missing values. It should be transferred to a numerical data before performing dimensionality reduction. To handle this problem, we use kernel method to map NACC dataset to a high-dimensional Hilbert space. We then use the data in such space to perform dimensionality reduction. The kernel function used is the radial basis function (RBF) defined by $K_{ij} = \exp(-\|x_i - x_j\|^2/\sigma^2)$. Here, to construct the kernel function, we use VDM (Value difference Metric) [8] to calculated the distance between x_i and x_j instead of only relying Euclidean distance. In detail, given two samples x_i and x_j, suppose the first j attributes of them are nominal, the following k ones are numeric and normalized to [0,1], and the remaining $D - j - k$ ones are missing if either x_i or x_j lacks the values in these attributes, the distance between x_i and x_j can be calculated by:

$$D(x_i, x_j)=\left(\sum_{h=1}^{j} VDM(x_{h,i}, x_{h,j})+\sum_{h=j+1}^{j+k} |x_{h,i}-x_{h,j}|^2\right)^{\frac{1}{2}} \quad (6)$$

Here, the VDM distance between two values z_1 and z_2 on nominal attribute Z can be calculated by:

$$VDM(z_1, z_2)\sum_{k=1}^{c}\left|\frac{N_{Z,z_1,k}}{N_{Z,z_1}}-\frac{N_{Z,z_2,k}}{N_{Z,z_2}}\right|^2 \quad (7)$$

where $N_{Z,z}$ denotes the number of training examples holding value z on Z, $N_{Z,z,k}$ denotes the number of training examples belonging to the kth class and holding value z on Z, c denotes the number of classes. Hence after we define the distance as in (6), we can either use it to construct the kernel function or to train a nearest neighbor classifier for evaluating the accuracies of test set.

IV. Simulations

This simulation aims at differentiating normal persons from demented persons by using TR-LDA and compares it with other state-of-the-art methods such as PCA, LPP, MMC and LDA. In this simulation, we randomly choose 500, 1000 and 2000 samples in AD data as training set and the remaining as test set. The data is preliminarily processed with KPCA operator to eliminate the null space of training set [7]. Then, each method uses the training set in the reduced output space to train a nearest neighborhood classifier to classify the demented and non-demented persons in test set.

The average accuracies over 20 random splits under different dimensionalities are in Table II and Fig. 2. As shown in Table II, the classification accuracies of all methods change greatly with the increase in the number of labeled samples. Another important observation is that the supervised methods such as LPP [6], MMC [5], LDA [10], TR-LDA outperform the unsupervised methods such as PCA and LPP. Among all the supervised methods, the proposed TR-LDA performs the best due to the trace ratio criterion. We also compare the convergence between ITR and iITR algorithms as in Fig. 1. From Fig. 1, we can see both algorithms can converge to the optimal trace ratio value. The iITR algorithm converges faster than ITR algorithm due to reason as in Section II-C.

V. Conclusion

Dementia is one of the most common neurological disorders among the elderly. Identification of demented patients from normal subjects can be transformed into a pattern recognition problem with high-dimensional nonlinear datasets. In this paper, we introduce trace ratio linear discriminant analysis (TR-LDA) for dementia diagnosis and propose an improved ITR algorithm (iITR) to solve the TR-LDA problem. The new proposed algorithm can handle nominal attributes and missing values in many real-world medical diagnosis problems. Finally, extensive simulations are presented to show the effectiveness of the proposed algorithms. The results demonstrate that our proposed algorithm can achieve higher accuracies for identifying the demented patients than other state-of-art algorithms.

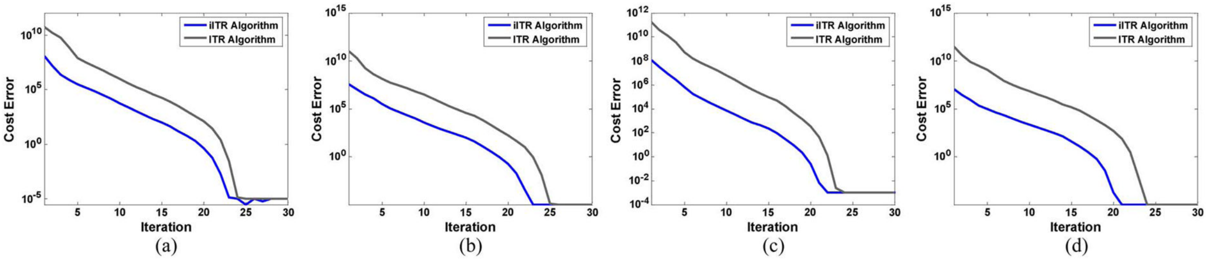
Acknowledgments

The authors thank L. M. Besser for database support.

The NACC database was supported by NIA Grant UO1 AG016976.

References

1. Beekly DL, et al. The national alzheimer's coordinating center (NACC) database: The uniform data set. Alzheimer Dis Assoc Disord. 2007; 21(3):249–258. [PubMed: 17804958]
2. Wang H, Yan S, Xu D, Tang X, Huang T. Trace ratio vs. ratio trace for dimensionality reduction. Proc CVPR. 2007
3. Jia Y, Nie F, Zhang C. Trace ratio problem revisited. IEEE Trans Neural Netw. 2009; 20(4):729–735. [PubMed: 19304481]
4. Zhou L, Wang L, Shen CH. Feature selection with redundancy-constrained class separability. IEEE Trans Neural Netw. 2010; 21(5)
5. Li H, Jiang T. Efficient and robust feature extraction by maximum margin criterion. IEEE Trans Neural Netw. 2006; 17(1):157–165. [PubMed: 16526484]
6. He X, Yan S, Hu Y, Niyogi P, Zhang H. Face recognition using Laplacianfaces. IEEE Trans Patt Anal Mach Intell. 2005; 27(3):328–340.
7. Zhang C, Nie F, Xiang S. A general kernelization framework for learning algorithms based on kernel PCA. Neurocomputing. 2010; 73(4–6):959–967.
8. Stanfill C, Waltz D. Toward memory-based reasoning. Commun ACM. 1986; 29(12)
9. Matsui, T.; Saruwatari, Y.; Shigeno, M. An Analysis of Dinkelbach's Algorithm for 0–1 Fractional Programming Problems. Dept. Math. Eng. Inf. Phys., Univ; Tokyo, Japan: 1992. METR92-14
10. Fukuaga, K. Introduction to Statistical Pattern Classification. New York, NY, USA: Academic; 1990.
11. Nie F, Xiang S, Zhang C. Neighborhood MinMax projections. IJCAI. 2007
12. Xiang S, Nie F, Zhang C. Learning a Mahalanobis distance metric for data clustering and classification. Patt Recognit. 2008; 41(12):3600–3612.
13. Nie F, Xiang S, Jia Y, Zhang C. Semi-supervised orthogonal discriminant analysis via label propagation. Patt Recognit. 2009; 42(11):2615–2627.
14. Nie F, Xiang S, Jia Y, Zhang C, Yan S. Trace ratio criterion for feature selection. AAAI. 2008



(a)

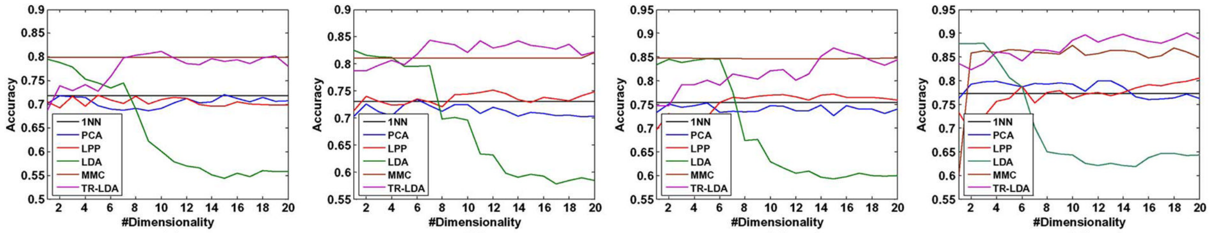
(b)

(c)

(d)

Fig. 1.

Convegence between ITR and iITR algorithms: (a) 500 samples; (b) 1000 samples; (c) 1500 samples; (d) 2000 samples.



(a) (b) (c) (d)

Fig. 2.

Average accuracies under different dimensionalities: (a) 500 samples; (b) 1000 samples; (c) 1500 samples; (d) 2000 samples.

TABLE I

iITR Algorithm for Solving the Trace Ratio Problem

```
1  Initialize $\lambda_0 = 0$.
2  Compute the eigen-decomposition of $S_b - \lambda_t S_w$ as $(S_b - \lambda_t S_w)\, w_i = \tau_i w_i$, where $w_i$ ($i = 1,2,\ldots D$) is the eigenvector of $S_b - \lambda_t S_w$.
3  Calculate $f_i = w_i^T S_b w_i$ and $g_i = w_i^T S_w w_i$ for $i \in \{1,2,\ldots,D\}$ and initialize $\gamma_0 = \lambda_t$ and $b^0 = \left[b_1^0, b_2^0, \ldots b_D^0\right]$ be a zero vector, iteratively solving the sub-problem of Eq. (5) until convergence:
     •  Sort $f_i - \gamma_k g_i$ and set $b_i^k = 1$ corresponding to the $d$ largest value of $f_i - \gamma_k g_i$, $b_i^k = 0$ otherwise.
     •  Update $\gamma_{i+1} = b^k f^T / b^k g^T$.
     •  If $b^k = b^{k-1}$, output $b^* = b^k$ and $\gamma^* = b^* f^T / b^* g^T$.
4  Form $W_t$ by choosing the $d$ eigenvectors of $w_i$, with $b_i^* = 1$ and Update $\lambda_{t+1} = \gamma^*$.
5  Iterate the steps (2–4) until $|\lambda_{t+1} - \lambda_t| < \varepsilon$. Output $W^*$.
```

TABLE II

The Average Accuracies Over 20 Random Splits

Method	**500 samples**		**1000 samples**		**1500 samples**		**2000 samples**	
	mean ± var	**dim**	**mean ± var**	**dim**	**mean ± var**	**dim**	**mean ± var**	**dim**
1NN	71.76 ± 3.89	-	73.09 ± 1.68	-	75.45 ± 1.26	-	77.32 ± 1.14	-
KPCA+1NN	72.08 ± 4.09	23	73.35 ± 2.14	25	75.26 ± 2.37	24	80.02 ± 2.05	23
KPCA+LPP+1NN	71.94 ± 3.65	24	75.09 ± 2.35	33	77.12 ± 2.23	30	80.55 ± 1.50	32
KPCA+MMC+1NN	79.56 ± 3.50	1	82.46 ± 2.00	1	84.63 ± 2.56	1	87.90 ± 1.22	1
KPCA+LDA+1NN	79.94 ± 3.86	1	82.09 ± 2.41	1	84.82 ± 2.29	1	87.46 ± 1.02	10
KPCA+TR-LDA+1NN	**81.12 ± 3.29**	**10**	**84.35 ± 2.10**	**7**	**86.83 ± 2.52**	**15**	**90.01 ± 1.25**	**15**