Local Learning Multiple Probabilistic Linear Discriminant Analysis

Yi Yang[(✉)] and Jiasong Sun

Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing, China
{Yangyy,sunjiasong}@tsinghua.edu.cn

Abstract. Probabilistic Linear Discriminant Analysis (PLDA) has delivered impressive results in some challenging tasks, e.g. face recognition and speaker recognition. Similar with the most state-of-the-art machine learning techniques, PLDA tries to globally learn the model parameters over the whole training set. However, those globally-learnt PLDA parameters can hardly characterize all relevant information, especially for those data sets whose underlying feature-spaces are heterogeneous and abound in complex manifolds. PLDA has the data homogeneous assumption which could be interpreted by involved parameters estimated through the entire training dataset. Such a global learning idea has been proven ineffective in the case of the heterogeneous data. In this paper, we alleviate this assumption by separating the feature space and locally learning multiple PLDA models of each space. Various standard datasets are performed and the superiority of the proposed method over the original PLDA could be found. We complete this work by assigning a probability to measure which models the test individual data match. This probabilistic scoring approach could further integrate different recognition technologies including other kinds of biological characteristics recognition. We propose the novel log likelihood score in recognition part includes three steps to complete.

Keywords: Local learning · Probabilistic linear discriminant analysis · Clustering · Bayesian method · Fusion

1 Introduction

Probabilistic Linear Discriminant Analysis (PLDA) [1], as a probabilistic extension of LDA [2], has been demonstrated as an effective approach to learn the low-dimensional representation of feature by its excellent performance on both face recognition [1] and speaker recognition [3]. A generative model is adopted which incorporates both within-individual and between-individual variation. In the recognition stage, PLDA calculates the likelihood that the differences between face images are entirely due to within-individual variability.

Similar with the most state-of-the-art machine learning techniques, PLDA tries to globally learn the model parameters over the whole training set. Nevertheless, those globally-learnt PLDA parameters can hardly characterize all relevant information,

especially for those datasets whose underlying feature spaces are heterogeneous [4] and abound in complex manifolds [5]. Plenty of recent works have been presented to train the models locally rather than globally [6–8]. Observing these facts, we propose a novel approach to consider heterogeneous subtle data structures by locally learning the PLDA model parameters.

The rest of this paper is organized as follows. Some related work is briefly reviewed in Sect. 2. In Sect. 3, we simply review the PLDA algorithm. In Sect. 4, we propose a novel robust locally learning multiple PLDA method to overcome the non-linear subspace problem and extend this method by introducing individual clustering to deal with the trouble caused by noise distribution, and we give the log likelihood method to score this model. Experimental results on face recognition data as well as speaker recognition data are presented in Sect. 5, respectively, comparing our method with other methods, which is followed by the conclusions and future works of this paper.

2 Locally Learning Multiple PLDA Models with Clustering

Linear Discriminant Analysis (LDA) [6] is a powerful method for face recognition, yielding a linear transformation matrix on the original data space and subsequently projecting it into a low-dimensional feature space. The well-known Fisher criterion [2, 10] is adopted, meaning that the centroid of different classes is pushed away and the data from the same class are pulled closely to the great extent. This can be realized by maximizing between-class variation and minimizing the within-class variation. LDA has the small-sample problem and other improvements of LDA could not handle the situation on the large changes of light and posture which always are regarded as interference.

The details of this are rather involved and are presented in [1].

2.1 Locally Learning Multiple PLDA

Locally learning algorithm always attempts to find a locally mapping which projects individual features to an explicit point in each subspace. The capacity of a locally learning algorithm is decided by its optimal parameters. In this section we adopt nonlinear locally learning method to reduce the problem caused by dimension mismatch. Matching between the two individual features is based on the distance between the two mapped individual features. Assuming that the observed features were generated as follows:

$$y_i, c = w_c x_{i,c} + m_c \tag{1}$$

Where $y_{i,c}$ is the feature vector of the i-th person in the c-th PLDA models. w_c denotes the subspace projection mapping matrix in the c-th PLDA models. m_c is the bias vector in the c-th PLDA models.

The minimum mean-square error compression of data is the formulation for PCA that can be generalized to the case with missing values in a very straightforward manner [9]. By this method, the weighted cost function of Eq. 1 is:

$$\hat{y}_i, c = \arg\min_{w \in W}(w_c^T x_{i,j,c} + m_c) = \arg\min_{w \in W}(\Sigma_{k=1}^K w_{ck} x_{ck} + m_c) \tag{2}$$

Where W is defined as the subsets of weighted space. $\hat{y}_{i,c}$ is the least squares estimated value of $y_{i,c}$. And

$$K = \Sigma_{i \in O}(y_{i,c} - \hat{y}_{i,c})^2 \tag{3}$$

Where O is defined as the individual feature value space. This method is trying to find a constant approximation of $y_{i,c}$ as the desired output in each subspace. We have the same nonlinear locally learning weight in training part and recognition part.

2.2   Novel Log Likelihood Score

First, the probability of test individual data belonging to one of multiple separating space is defined as:

$$P(t = k) = \frac{\Sigma_{k=0}^K P(c = k)P(t = k|c = k)}{\Sigma_{k=0}^K P(t = k|c = k)} \tag{4}$$

Where $P(t = k)$ is probabilities of test individual data belongs to the k th space. $P(c = k)$ is probabilities of train individual data c belongs to the k th space. $P(t = k|c = k)$ is the conditional probability of above. $k = 0, \ldots, K$ and K is the total number of space. And the classified test individual data is projected to same subspace as in training part:

$$\hat{y}_{i,c} = w_c^T x_{i,j,c} + m_c \tag{5}$$

Second, the novel log likelihood score is computed by the probability of classified test individual data which generated from certain c-th PLDA models is the summing of all variables:

$$P(\hat{y}_c|\theta_c) = \Sigma_{c=1}^K \int \int P(\hat{y}_c, h_{i,c}, w_{i,j,c}, c)dh_{i,c}dw_{i,j,c} \tag{6}$$

Finally, the decision fusion is to combine multiple PLDAs θ_c by Bayes criterion into one PLDA model whose matching likelihood of $\hat{y}_i$ and $\hat{y}_j$ is:

$$P(\hat{y}_i, \hat{y}_j|\theta_c) = \prod_{c=1}^K P(\hat{y}_{i,c}, \hat{y}_{j,c}|\theta_c) \tag{7}$$

And the log likelihood score is defined as:

$$L_\theta = \frac{\Sigma_{c=1}^{K}[P(t=c)P(\hat{y}_i, \hat{y}_j|\theta_c)]}{log\Sigma_{c=1}^{K}[P(t=c)(1-P(\hat{y}_i, \hat{y}_j)|\theta_c)]} \tag{8}$$

Where L_θ calculate the ratio of log likelihood with the one PLDA model that two individual test data match to that of two individual test data do not match. i and j separately represents the i-th individual and the j-th individual of test data. $\Sigma_{c=1}^{K}[P(t=c)P(\hat{y}_i, \hat{y}_j|\theta_c)]$ and $\Sigma_{c=1}^{K}[P(t=c)(1-P(\hat{y}_i, \hat{y}_j)|\theta_c)]$ separately represents the matching likelihood and the not matching likelihood with the c-th PLDA model.

3   Experimental Results

3.1   Data Preprocessing and Experimental Setup

We performed experiments on three standard corpora: TIMIT and PIE. In TIMIT there are 48 possible phonetic classes for the training, which are later merged into 39 classes for the performance evaluation. The sizes of the training, testing, and development (for parameter tuning) sets are around 140, 7, and 15 thousand, which are the common practices in the speech society [9], based on which we also generate acoustic feature vectors. Since PLDA is a classification algorithm, the PIE dataset, as a face dataset, is used, which consists of more than 40,000 of faces and authors in [5] suggested a representative portion of this corpus. We choose to use 30 samples to train the models and the remainders to test. Commonly, the experiments on PIE were repeated 10 times of random data splitting and the average results are to be reported.

3.2   Results

Figures 1 and 2 shows the individual verification results by four methods on TIMIT and PIE datasets. Table 1. In Fig. 1(a) the PLDA (C = 1, S = 60) method received the highest score and PLDA (C = 1, S = 50) has the second score which indicates that separating space improving the original PLDA under different PLDA subspace projection dimension. Also we can know the nonlinear locally learning will promote the performance under three conditions (C = 1, 2, 3). But with the reduced dimension of nonlinear locally learning, the correct rate has a significant decline under all the PLDA subspace projection dimension. In Figs. 2(b) and 2(c) the PLDA (S = 60) also has the highest score but PLDA (S = 50) has the second score. At the same time we can observe the better results on each nonlinear locally learning method being used. Compared with TIMIT data, Fig. 2 shows the face recognition corpus PIE result with seven methods. Same as TIMIT, The PLDA (S = 120) has the highest score and PLDA (S = 100) has the second score. And it presents the reduced dimension of nonlinear locally learning does make worse influence on PIE data.
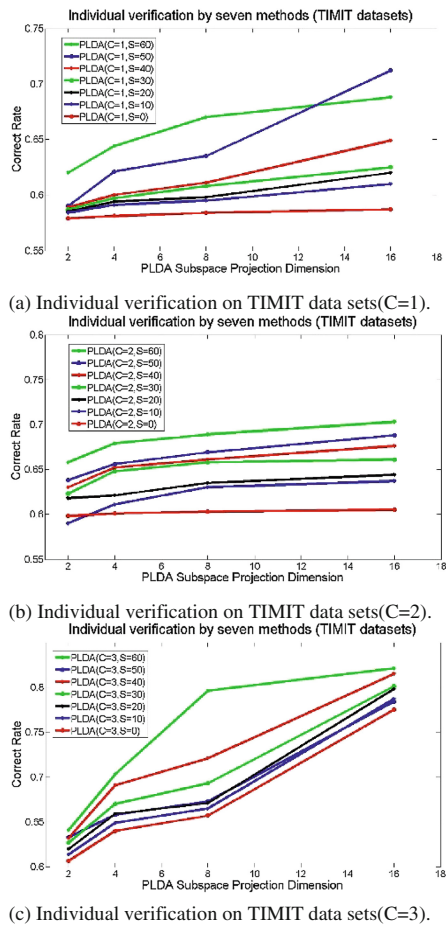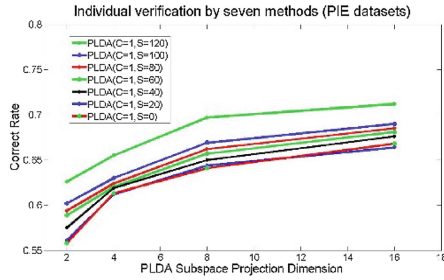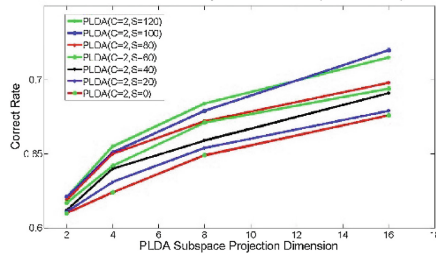
(a) Individual verification on TIMIT data sets(C=1).

(b) Individual verification on TIMIT data sets(C=2).

(c) Individual verification on TIMIT data sets(C=3).

Fig. 1. Individual verification by seven methods on TIMIT data sets.

Table 1. Conditions with their parameters for example

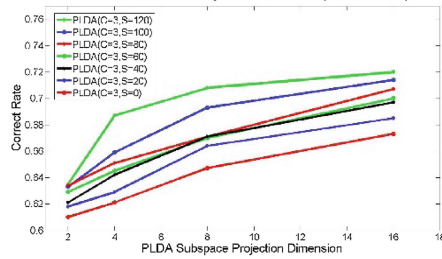| Methods	Separating	Subspace dimension
PLDA(C = 1,S = 0)	1	NONE
PLDA(C = 1,S = 60)	1	60
PLDA(C = 2,S = 0)	2	NONE
PLDA(C = 2,S = 120)	2	120
PLDA(C = 3,S = 0)	3	NONE
PLDA(C = 3,S = 60)	3	60
PLDA(C = 3,S = 0)	3	NONE
PLDA(C = 3,S = 120)	3	120

(a) Individual verification on PIE data sets(C=1).

(b) Individual verification on PIE data sets(C=2).

(c) Individual verification on PIE data sets(C=3).

Fig. 2. Individual verification by seven methods on PIE data sets.

4   Conclusions

In this paper, we have presented one approach to generate multiple PLDA models based on feature space separating which enables us to obtain better results than original single PLDA model. We have also shown that our approach is robust both on speaker recognition and on face recognition standard corpus without other prior information. And a new probabilistic scoring approach is proposed to achieve soft decision based on feature space separating and locally learning multiple PLDA models. Combining other biometric individual components with our model is a promising approach to many recognition tasks.

References

1. Prince, S.J.D., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: 11th International Conference on Computer Vision 2007, ICCV 2007, pp. 1–8. IEEE, 14–21 October 2007
2. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
3. Senoussaoui, M., Kenny, P., Brummer, N., Dumouchel, E.d.V.P.: Mixture of plda models in i-vector space for gender independent speaker recognition. In: Interspeech 2011, pp. 1–19. IEEE (2011)
4. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Commun. **26**, 283–297 (1998)
5. He, X., Niyogi, P.: Locality preserving projections. In: Proceedings of Neural Information Processing Systems, vol. 16, Vancouver, British Columbia (2003)
6. Kim, T., Kittler, J.: Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 318–327 (2005)
7. Liu, Y., Liu, Y., Chan, K.: Tensor-based locally maximum margin classifier for image and video classification. Comput. Vis. Image Underst. **115**(3), 300–309 (2011)
8. Mahanta, M., Aghaei, A., Plataniotis, K., Pasupathy, S.: Heteroscedastic linear feature extraction based on sufficiency conditions. Pattern Recognit. **45**, 821–830 (2012)
9. Ilin, A., Raiko, T.: Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. **11**, 1957–2000 (2011)
10. Halberstadt, A.: Heterogeneous acoustic measurements and multiple classifiers for speech recognition, Ph.D. thesis, MIT (1998)