Experimental Study on Multiple LDA Classifier Combination for High Dimensional Data Classification

Xiaogang Wang and Xiaoou Tang

Department of Information Engineering
The Chinese University of Hong Kong
{xgwang1,xtang}@ie.cuhk.edu.hk

Abstract. Multiple classifier systems provide an effective way to improve pattern recognition performance. In this paper, we use multiple classifier combination to improve LDA for high dimensional data classification. When dealing with the high dimensional data, LDA often suffers from the small sample size problem and the constructed classifier is biased and unstable. Although some approaches, such as PCA+LDA and Null Space LDA, have been proposed to address this problem, they are all at cost of discarding some useful discriminative information. We propose an approach to generate multiple Principal Space LDA and Null Space LDA classifiers by random sampling on the feature vector and training set. The two kinds of complementary classifiers are integrated to preserve all the discriminative information in the feature space.

1 Introduction

Multiple classifier combination is an effective way to improve pattern recognition performance. Random subspace [4] and bagging [5] are two popular techniques to combine weak classifiers into a powerful decision rule. In the random subspace method, a set of low dimensional subspaces are generated by randomly sampling from the high dimensional feature vector and multiple classifiers constructed in the random subspaces are combined in the final decision. In bagging, random independent bootstrap replicates are generated by sampling the training set. A classifier is constructed from each replicate, and the results of all the classifiers are finally integrated. Based on the two random sampling techniques, we propose an approach using multiple LDA classifier combination for high dimensional data classification.

Linear Discriminant Analysis (LDA) is a popular feature extraction technique for data classification. It determines a set of projection vectors maximizing the between-class scatter matrix (S_b) and minimizing the within-class scatter matrix (S_w) in the projective feature space. But when dealing with the high dimensional data, LDA often suffers from the small sample size problem. When there are not enough training samples, S_w is not well estimated and may become singular [3].

To address this problem, a two-stage PCA+LDA approach [1] is proposed. The high dimensional data is first projected to a low dimensional PCA subspace, in which S_w is non-singular, and then LDA is performed. We call it Principal Space LDA.

F. Roli, J. Kittler, and T. Windeatt (Eds.): MCS 2004, LNCS 3077, pp. 344–353, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The eigenvectors with small eigenvalues removed from the PCA subspace may also encode some information helpful for recognition. Their removal may introduce a loss of discriminative information.

Chen et. al. [2] suggested that the null space spanned by the eigenvectors of S_w with zero eigenvalues contains the most discriminative information. However, as explained in [2], with the existence of noise, when the training sample number is large, the null space of S_w becomes small, so much discriminative information outside this null space will be lost.

Some random sampling based LDA classification approaches can be found in [7][8]. Different from the previous work, our method simultaneously samples on the feature space and training samples, and takes advantage of the discriminative information in both the principal and null spaces of S_w. We also explain that both Principal Space LDA (P-LDA) and Null Space LDA (N-LDA) encounter the overfitting problem, but for different reasons. So we will improve them in different ways accordingly. A more detailed description on the algorithm can be found in [9][10]. In this paper, we make an extensive experimental study on the XM2VTS database [12].

2 LDA for High Dimensional Data Classification

Two conventional LDA approaches, PCA+LDA and N-LDA are briefly reviewed in this section. The high dimensional data is represented as a vector $\vec{x}$ with length *N*. The training set contains *M* samples belonging to *L* classes.

2.1 PCA+LDA

Principal Component Analysis (PCA) computes a set of eigenvectors of the ensemble covariance matrix *C* of the training set. Eigenvectors are sorted by eigenvalues, which represent the variance of data distribution. There are at most *M-1* eigenvectors with non-zero eigenvalues. Normally *K* eigenvectors, $U = [\vec{u}_1, \ldots \vec{u}_K]$, with the largest eigenvalues, are selected to span the PCA subspace. Low dimensional features are extracted by projecting the high dimensional data $\vec{x}$ into the PCA subspace,

$$\vec{w} = U^T (\vec{x} - \vec{m}). \tag{1}$$

where $\vec{m}$ is the mean of the training set.

LDA tries to find a set of projecting vectors W maximizing the ratio of determinant of S_b and the determinant of S_w,

$$W = \arg\max \left| \frac{W^T S_b W}{W^T S_w W} \right|. \tag{2}$$

W can be computed from the eigenvectors of $S_w^{-1} S_b$ [6]. The rank of S_w is at most *M-L*. But when the training set is small and *M-L* is smaller than the vector length *N*, S_w may become singular and it is difficult to compute S_w^{-1}.

In the two-stage PCA+LDA approach [1], the data vector is first projected to a PCA subspace spanned by the *M-L* largest eigenvectors. LDA is then performed in the *M-L* dimensional subspace, such that S_w is nonsingular. But in many cases, *M-L* dimensionality is still too high for the training set. So the LDA classifier is often biased and unstable. Furthermore, much discriminative information outside the PCA subspace is discarded.

2.2 Null Space LDA

Chen et. al. [2] suggested that the null space of S_w also contains much discriminative information. It is possible to find some projection vectors W satisfying $W^T S_w W = 0$ and $W^T S_b W \neq 0$, thus the Fisher criteria in Eq. (2) definitely reaches its maximum value. The rank of S_w, $r(S_w)$, is bounded by $\min(M-L, N)$. Because of the existence of noise, $r(S_w)$ is almost equal to this bound. The dimension of the null space is $\max(0, N-M+L)$. As shown by experiments in [2], when the training sample number is large, the null space of S_w becomes small, thus much discriminative information outside this null space will be lost.

3 Multiple LDA Classifier Combination for High Dimensional Data Classification

Both P-LDA and N-LDA face the same problem: the constructed classifier is unstable and much discriminative information is discarded. But they are caused by different reasons. So we design different random sampling algorithms to improve the two LDA methods, and combine them in a multiple classifier structure.

3.1 Using Random Subspace to Improve P-LDA

In P-LDA, overfitting happens when the training set is relatively small compared to the high dimensionality of the feature vector. In order to construct a stable LDA classifier, we sample a small subset of features to reduce discrepancy between the training set size and the feature vector length. Using such a random sampling method, we construct a multiple number of stable LDA classifiers, and combine them into a powerful classifier covering the entire feature space without losing discriminative information.

We first apply PCA to the training set. All the eigenvectors with zero eigenvalues are removed, since all the training samples have zero projections on them. The *M-1* eigenvectors $U_0 = \{\bar{u}_1, \ldots, \bar{u}_{M-1}\}$ with positive eigenvalues are retained as candidates to construct random subspaces. Then, *K* random subspaces are generated. The dimen-

sion of random subspace is determined by the training set to make the LDA classifier stable. In each random subspace, the first N_0 dimensions are fixed as the largest eigenvectors, and the remaining N_1 dimensions are randomly selected from $\{\bar{u}_{M-N_0-1},\ldots,\bar{u}_{M-1}\}$. The N_0 largest eigenvectors encode much data structural information. If they are not included in the random subspace, the accuracy of LDA classifiers may be too low. Our approach guarantees that the LDA classifier in each random subspace has satisfactory accuracy. The N_1 random dimensions cover most of the remaining small eigenvectors. So the ensemble classifiers also have a certain degree of error diversity.

3.2 Using Bagging to Improve N-LDA

In N-LDA, the overfitting problem happens when the training sample number is large, since the null space will be too small. It can be alleviated by bagging. In bagging, random independent bootstrap replicates are generated by sampling the training set, so each replicate has a smaller number of training samples. We Generate *K* replicates by randomly sampling the training set. A N-LDA classifier is constructed from each replicate and the multiple classifiers are combined using a fusion rule.

3.3 Integrating Random Subspace and Bagging for LDA Based Classification

While P-LDA is computed from the principal subspace of S_w, in which $W^T S_w W \neq 0$, N-LDA is computed from its orthogonal subspace in which $W^T S_w W = 0$. Both of them discard some discriminative information. Fortunately, the information retained by the two kinds of classifiers complements each other. So we combine them to construct the final classifier. Many methods on combining multiple classifiers have been proposed [11]. In this paper, we use two simple fusion rules: majority voting and sum rule. More complex combination algorithms may further improve the system performance.

4 Experiments

We apply the random sampling based LDA approach to face recognition and make a extensive experimental study on the XM2VTS face database [12]. There are 295 people, and each person has four frontal face images taken in four different sessions. In our experiments, two face images of each class are selected for training, and the remaining two for testing. In preprocessing, the face image is normalized by translation, rotation, and scaling, such that the centers of two eyes are in fixed positions. A 46 by 81 mask removes most of the background. So the face data dimension is $46\times 81 = 3726$. We adopt the recognition test protocol used in FERET [13]. All the face classes in the reference set are ranked. We measure the percentage of the "correct answer in top 1 match".

4.1 Random Subspace LDA

We first compare random subspace LDA with the conventional PCA+LDA approach. Table 1 reports the accuracy of a single P-LDA classifier constructed from PCA subspace with different dimension. Since there are 590 face images of 295 classes in the training set, there are 589 eigenfaces with non-zero eigenvalues. According to [1], the PCA subspace dimension should be *M-L=295*. However, the result shows that the accuracy is only 79% using a single P-LDA classifier constructed from 295 eigenfaces, because this dimension is too high for this training set and S_w cannot be well estimated. We observe that P-LDA classifier has the best accuracy 92.9% when the PCA subspace dimension is set at 100. So for this training set 100 seems to be a suitable dimension to construct a stable P-LDA classifier. In the following experiments, we choose 100 as the dimension of random subspaces to construct the multiple P-LDA classifiers.

First, we generate the random subspaces by randomly selecting 100 eigenfaces from 589 eigenfaces with nonzero eigenvalues. The result of combining 20 P-LDA classifiers using majority voting is shown in Figure 1. The accuracy of each individual P-LDA classifier is low, between 50% and 70%. Using majority voting, the weak classifiers are greatly enforced, and 87% accuracy is achieved. This shows that P-LDA classifiers constructed from different random subspaces are complementary of each other. In Table 2, as we increase the classifier number *K*, the accuracy of the combined classifier improves, and even becomes better than the highest accuracy in Table 1. Although increasing classifier number and using more complex combining rules may further improve the performance, it will increase the system burden.

Table 1. Recognition accuracy of PCA+LDA classifier constructed from PCA subspace with different dimension.

Dim	30	50	70	100	150	200	250	295
Accuracy	0.870	0.925	0.927	0.929	0.898	0.864	0.820	0.792

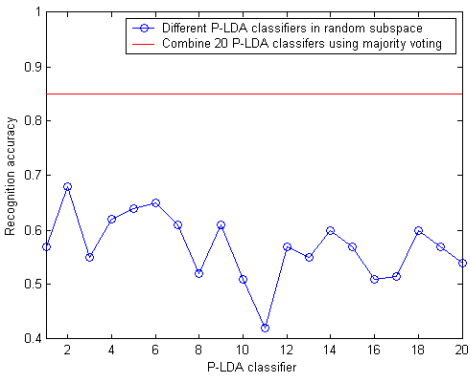

Fig. 1. Recognition accuracy of combing 20 P-LDA classifiers constructed from random subspaces using majority voting. Each random subspace randomly selects 100 eigenfaces from 589 eigenfaces with non-zero eigenvalues.

Table 2. Accuracy of combining different number (K) of P-LDA classifiers constructed from random subspaces using majority voting. Each random subspace randomly selects 100 eigenfaces from 589 eigenfaces with non-zero eigenvalues.

K	20	40	60	80	100	120	140	160
Accuracy	0.871	0.907	0.917	0.922	0.937	0.932	0.939	0.939

Table 3. Recognition accuracy of P-LDA classifiers constructed from different parts of eigenface sequence which has been sorted by eigenvalues. The first row is the index of eigenfaces spanning the subspace from which LDA classifier is constructed, and the second row is the recognition accuracy.

Index	1-100	101-200	201-300	301-400	401-500	501-589	vote
Accuracy	0.929	0.514	0.378	0.148	0.06	0.04	0.613

Table 4. Recognition accuracy of combining P-LDA classifiers using different number (K) of random subspaces (sum rule). In each random subspace, the first 50 dimensions are fixed as the 50 largest eignfaces, and another 50 dimensions are randomly selected from the remaining 593 eigenfaces with positive eigenvalues. We run ten times on the same training set and testing set, and record the accuracy means and variances.

K	5	10	15	20	25	30
Mean	0.954	0.958	0.959	0.961	0.961	0.962
Variance	0.0133	0.0127	0.0094	0.0101	0.0068	0.0049

Some largest eigenfaces encode much face structural information. If they are not included in the random subspace, the individual LDA classifier is poor. This can be further proved in Table 3, in which six LDA classifiers are constructed based on different parts of eigenface sequence. The first row is the index of eigenfaces spanning the subspace. Using only the eigenfaces with small eigenvalues, the recognition accuracy of LDA classifier is poor. But it doesn't mean these eigenfaces are not useful for recognition.

A better approach to improve the performance of the combined classifier is to increase the accuracy of each individual weak classifier. To improve the accuracy of each individual P-LDA classifier, as illustrated in Section 3.1, in each random subspace, we fix the first 50 basis as the 50 largest eigenfaces, and randomly select another 50 basis from the remaining 539 eigenfaces. As shown in Figure 2, individual P-LDA classifiers are improved significantly. They are similar to the LDA classifier based on the first 100 eigenfaces. These classifiers are also complementary of each other, so much better accuracy (96%) is achieved when they are combined. The recognition performance of using different number of random subspaces is shown in Table 4. We run 10 times on the same training set and testing set, recording the accuracy means and variances. Using more random subspaces, the accuracy is higher and more stable.

We also apply random subspace to N-LDA. Similar to the method in Section 3.1, the random subspaces with dimension D ($295<D<590$) are generated from PCA subspace and a N-LDA classifier is constructed from each random subspace. As shown in Figure 3, there is no improvement in recognition performance. When the random

subspace dimensionality D is low, the null space dimension (D-295) is small, so the recognition accuracy drops greatly. Random subspace further reduces the null space dimension and deteriorates the overfitting problem of N-LDA.

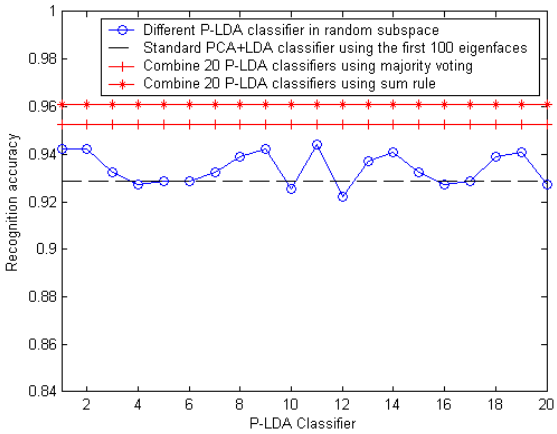

Fig. 2. Recognition accuracy of combing 20 P-LDA classifiers constructed from random subspaces. For each 100 dimensional random subspace, the first 50 dimensions are fixed as the 50 largest eigenfaces, and another 50 dimensions are randomly selected from the remaining 539 eigenfaces with non-zero eigenvalues.

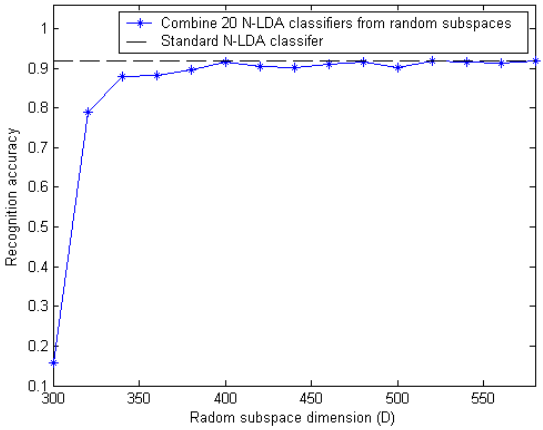

Fig. 3. Recognition accuracy of combining 20 N-LDA classifiers from random subspaces with different dimensions using majority voting.

4.2 Bagging LDA

Figure 4 reports the performance of bagging based N-LDA. We generate 20 replicates and each replicate contains 300 training samples. The individual N-LDA classi-

fier constructed from each replicate is less effective than the original classifier trained on the full training set. This is because that some intra-class variations are not included in each replicate. However, when the multiple classifiers are combined, the accuracy is significantly improved, and becomes much better than the standard N-LDA. Table 5 reports performance of bagging based N-LDA using different number of replicates, but fixing training sample number in each replicate as 300. As similar in Table 4, it is more stable using a relatively large number of replicates. In Figure 5, we fix the bagging replicates number as 20, but change the training sample number contained in the replicates from 100 to 500. The best performance is achieved using proper moderate training sample number in each replicate. When the training sample number in each replicate is too small, the null space cannot effectively remove the intra-class variation. When the training sample number in each replicate is too large, the null space dimension is too small to contain enough discriminative information, and different replicates are similar.

Table 5. Recognition accuracy of combining N-LDA classifiers using different number (K) of bagging replicates (sum rule). We run ten times on the same training set and testing set, and record the accuracy means and variances.

K	5	10	15	20	25	30
Mean	0.929	0.934	0.942	0.956	0.951	0.961
Variance	0.0120	0.0109	0.097	0.009	0.036	0.027

We also study using bagging to improve P-LDA classifiers. The PCA subspace is spanned by the 100 largest eigenfaces and 20 replicates are generated. The accuracies with the replicate containing different number of people are shown in Figure 6. As expected, the combined classifier shows no improvement over the original P-LDA classifier. In each replicate, the P-LDA classifier is constructed from an even smaller number of training samples. It deteriorates the small sample size problem.

4.3 Integrating Random Subspace and Bagging Based LDA

Integrating the multiple P-LDA classifiers generated by random subspace and N-LDA classifiers generated by bagging, the recognition accuracy can be further improved. We combine 10 P-LDA classifiers constructed from random subspaces and 10 N-LDA classifiers constructed from bagging replicates, and set an even better result as shown in Table 6.

Table 6. Compare random sampling based LDA with conventional LDA approaches. R-LDA (1): random subspace based LDA; R-LDA (2): bagging based N-LDA; R-LDA (3): integrating random subspace and bagging based LDA

PCA+LDA	N-LDA	R-LDA (1)	R-LDA (2)	R-LDA (3)
0.929	0.919	0.961	0.956	0.976

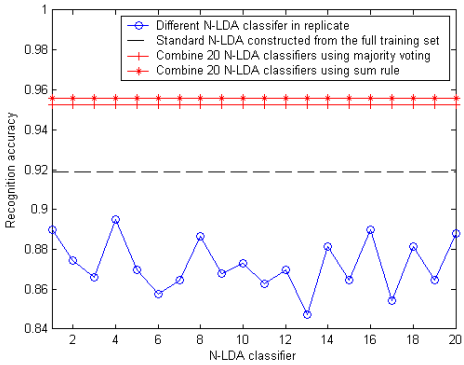

Fig. 4. Recognition accuracy of combining 20 N-LDA classifiers constructed from bagging replicates.

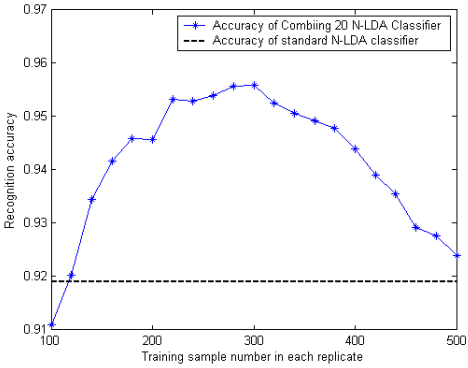

Fig. 5. Recognition accuracy of combing 20 N-LDA classifiers with different number of training samples contained in the bagging replicates (sum rule).

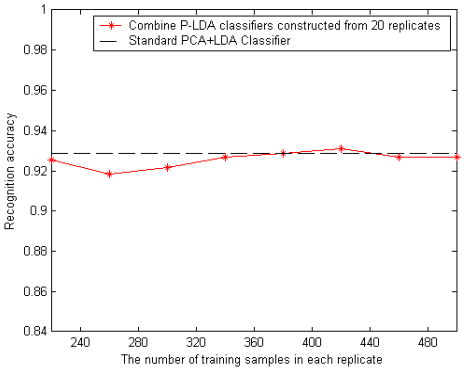

Fig. 6. Recognition accuracy of combining 20 P-LDA classifiers constructed from bagging replicates containing different number of training samples. The PCA space is spanned by 100 largest eigenfaces. The combining rule is majority voting.

5 Conclusion

Both P-LDA and N-LDA encounter the overfitting problems in when dealing with the high dimensional data classification, however, for different reasons. So we improve them using different random sampling approaches, sampling on feature for P-LDA and sampling on training samples for N-LDA. The two kinds of complementary classifiers are finally integrated in our system. The extensive experimental study on the XM2VTS database illustrates the effectiveness of our method and how it works.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK 4190/01E and CUHK 4224/03E).

References

1. P.N. Belhumeur, J. Hespanda, and D. Kiregeman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, *IEEE Trans. on PAMI,* Vol. 19, No. 7, pp. 711-720, July 1997.
2. L. Chen, H. Liao, M. Ko, J. Liin, and G. Yu, "A New LDA-based Face Recognition System Which can Solve the Samll Sample Size Problem", *Pattern Recognition*, Vol. 33, No. 10, pp. 1713-1726, Oct. 2000.
3. A. M. Martinez and A.C. Kak, “PCA versus LDA,” *IEEE Trans. on PAMI*, Vol. 23, No. 2, pp. 228-233, 2001.
4. T. Kam Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. on PAMI*, Vol. 20, No. 8, pp. 832-844, August 1998.
5. L. Breiman, “Bagging Predictors,” *Machine Learning*, Vol. 24, No. 2, pp. 123-140, 1996.
6. K. Fukunnaga, “Introduction to Statistical Pattern Recognition”, Academic Press, second edition, 1991.
7. M. Skurichina and R.P.W. Duin, “Bagging and the Random Subspace Method for Linear Classifiers,” Pattern Analysis and Application, Vol. 5, No. 2, pp. 121-135, 2002.
8. X. Lu and A.K. Jain, “Resampling for Face Recognition,” *in Proceedings of AVBPA03*, 2003.
9. X. Wang and X. Tang, “Random Sampling Based LDA for Face Recognition,” *in Proceedings of CVPR*, 2004.
10. X. Wang and X. Tang, “Random Subspace Based LDA for Face Recognition Integrating Multi-Features,” in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.
11. J. Kittler and F. Roli, (Eds): Multiple Classifier Systems.
12. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: The Extended M2VTS Database,” *Proceedings of International Conference on Audio- and Video-Based Person Authentication*, pp. 72-77, 1999.
13. P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET Evaluation,” in Face Recognition: From Theory to Applications, H. Wechsler, P. J. Phillips, V. Bruce, F.F. Soulie, and T. S. Huang, Eds., Berlin: Springer-Verlag, 1998.