WEIGHTED PAIRWISE SCATTER TO IMPROVE LINEAR DISCRIMINANT ANALYSIS

Yongxin Li, Yuqing Gao, Hakan Erdogan

IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY

ABSTRACT

Linear Discriminant Analysis (LDA) aims to transform an original feature space to a lower dimensional space with as little loss in discrimination as possible. We introduce a novel LDA matrix computation that incorporates confusability information between classes into the transform. Our goal is to improve discrimination in LDA. In conventional LDA, a between class covariance matrix that is based on the scatter of class means around the global mean is used. By rewriting the between class covariance expression in a more revealing way, we unveil that each class pair is considered equally confusable in the conventional LDA. We introduce a weighting factor for each pairwise scatter that enables to integrate the confusability information into the between class covariance matrix. There are many possibilities to choose the weighting factors. We consider few of them that depend on Euclidean and Kullback-Leibler distances between classes when a single Gaussian approximation is used for each class. The method combined with speaker cluster based transformation decreases the error rate by about relative 10% on a large vocabulary speech recognition task using IBM's speech recognition engine.

1. INTRODUCTION

In order to reduce computation and to decrease the effects of "the curse of dimensionality", it is common to apply linear discriminant analysis (LDA) for statistical pattern classification tasks. The LDA transform attempts to reduce dimension with minimal loss in discrimination information. LDA is used for speech recognition as a part of the front-end processing, because the computational complexity in speech recognition highly depends on the dimension of the feature space. On the other hand, feature spaces of higher dimension enable the acoustic model to carry more discriminant information. In speech recognition, feature space dimension can be increased by extending the feature vector to include a range of neighboring frame data. Doing this will increase discrimination but computation becomes impractical. Applying LDA to the extended feature vector is very necessary.

LDA has been used successfully in speech recognition, mixed results were obtained [1, 2, 3]. There are a lot of recent development, for instance [4]. In LDA, there are some implicit assumptions as analyzed in Kumar's work [3]. Basically, LDA formulation assumes that each class has equal within class covariance. Kumar formulated LDA in a maximum likelihood framework and generalize it to cases where the within class covariances are different. This generalization is called Heteroscedastic Linear Discriminant Analysis (HDA).

In this paper, we exploit another implicit assumption of LDA. The between class covariance matrix in LDA assumes that each class is equally confusable with all other classes. This might not be obvious from the formula but we rewrite the expression to reveal this property. The goal of this work is to remove this assumption of LDA to achieve higher classification performance and decoding accuracy. The resulting transformation is called Weighted Pairwise Scatter Linear Discriminant Analysis (WPS-LDA) transform.

In Section 2, we analyze conventional LDA formulation, particularly focusing on the between class scatter matrix. We point out a weakness of the conventional between class scatter matrix and illustrate it with an example. In Section 3, we define a generalization of conventional between class scatter matrix as a sum of weighted pairwise scatter matrices. This expression is the same as conventional between class scatter matrix if a uniform weight is used. The necessity of using non-uniform weights for better discrimination is discussed. We also consider appropriate weight choices. It is explained how this could remedy the weakness of conventional between class scatter matrix discussed in Section 2. In Section 4, we discuss briefly cluster based LDA transformations and its integration with WPS-LDA. In Section 5, we discuss the application of WPS-LDA to IBM speech recognition system. Experimental results using various weights based on Euclidean and Kullback-Leibler distances are presented and compared. Finally, in section 6, we discuss the restrictions of WPS-LDA and possibilities for further improvement.

2. THE CONVENTIONAL LDA

The LDA problem is formulated as follows. Let $x \in \mathbb{R}^n$ be a feature vector. We seek to find a transformation $y = \theta x$, $\theta : \mathbb{R}^n \to \mathbb{R}^p$ with $p < n$, such that in the transformed space, minimum loss of discrimination occurs. In practice, p is much smaller than n.

Assume $\{x_i\}_{1 \le i \le N}$ are N training feature vectors each labeled as belonging to a class $l_i \in \{1 \dots K\}$. Let $N_k = \sum_{l_i=k} 1$ be the number of training vectors in class k. Then, $\sum_{k=1}^{K} N_k = N$ is the total number of training samples. We define the following entities:

$$\Sigma_k = \frac{1}{N_k} \sum_{l_i=k} (x_i - \mu_k)(x_i - \mu_k)^T = \frac{1}{N_k} \sum_{l_i=k} x_i x_i^T - \mu_k \mu_k^T,$$

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T - \mu \mu^T,$$

where μ_k is the sample mean for class k, μ is the global sample mean, Σ_k is the covariance matrix for class k and Σ is the total covariance matrix. In some literature, for instance [5], scatter matrices are used. Essentially they are equivalent to covariance matrices except by a factor. For

example, the total scatter matrix [5] $T = N\Sigma$.

$$\begin{aligned} T &= \sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T \\ &= \sum_{i=1}^{N}\Big((x_i-\mu_{l_i})(x_i-\mu_{l_i})^T+(\mu_{l_i}-\mu)(\mu_{l_i}-\mu)^T\Big) \\ &= \sum_{i=1}^{K}N_k\Sigma_k+\sum_{i=1}^{K}N_k(\mu_k-\mu)(\mu_k-\mu)^T \end{aligned}$$

In classical LDA, the first term $W=\sum_{i=1}^{K}N_k\Sigma_k$ is called within class scatter matrix and the second term $B=\sum_{i=1}^{K}N_k(\mu_k-\mu)(\mu_k-\mu)^T$ is called between class scatter matrix. If θ is a linear projection, then in the new feature space, the within class scatter and between class scatter become $\theta W\theta^T$ and $\theta B\theta^T$ respectively.

It is popularly accepted that the between class scatter carries the discriminant information. The idea of LDA is to maximize in some sense the ratio of between class and within class scatter matrices after transformation. This will enable to choose a transform that keeps the most discriminative information while reducing the dimension. Precisely, we want to maximize the objective function

$$\max_{\theta}\frac{|\theta B\theta^T|}{|\theta W\theta^T|} \tag{1}$$

Fortunately, as well known, there is a close solution to this optimization problem. The columns of the optimum θ are the relative generalized eigenvectors corresponding to the first p maximal magnitude eigenvalues of the equation

$$Bv=\lambda Wv. \tag{2}$$

The following form of the between class covariance matrix is taken for granted in the literature.

$$B=\sum_{i=1}^{K}N_k(\mu_k-\mu)(\mu_k-\mu)^T. \tag{3}$$

This is a measure of how distributed the means of each class is from the center. Intuitively, it is better to have a "bigger" value of B since it shows that the classes are more spread out in the transformed space, thus easier to discriminate them. From this expression, it is not clear how the classes are discriminated from each other pairwise. To illustrate this point, we consider an example.

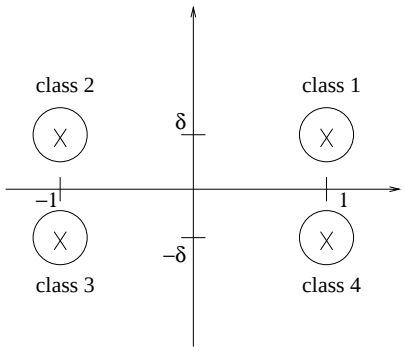

Figure 1: Illustration of example 1.

Example 1: Suppose there are four classes in $\mathbb{R}^2$, each of them have same number of feature vectors and equal variance. Let their means be:

$$\mu_1=(1,\delta) \qquad \mu_2=(-1,\delta)$$
$$\mu_3=(-1,-\delta) \qquad \mu_4=(1,-\delta)$$

This problem is illustrated in Figure 1. In this case, the between class scatter matrix is

$$\frac{1}{4}B=\begin{pmatrix}1&0\\0&\delta^2\end{pmatrix}\longrightarrow\begin{pmatrix}1&0\\0&0\end{pmatrix}$$

When $\delta\to 0$, the between class scatter matrix does not contain any discrimination in the vertical direction. Only discrimination is in the horizontal direction. We can say that (it will be clearer in next section) the between class scatter matrix is dominated by the covariance of the class pairs other than $(1,4)$ and $(2,3)$. Obviously, regarding the classification problem, the covariance of class pairs $(1,4)$ and $(2,3)$ are more important than others since they are more confusable, yet the conventional between class scatter obviously does not capture this information. Therefore, we conclude that the canonical method does not accurately represent the desired discrimination information.

From above example, we can see that if there are some classes much closer relatively as compared to others, the between class scatter matrix mostly ignores the discriminatory information between the classes that are close to each other.

3. WEIGHTED PAIRWISE SCATTER LDA

The discussion in section 2 leads us to define a general between class scatter matrix that is equal to the sum of weighted "pairwise scatter" matrices.

$$B_w=\frac{1}{2N}\sum_{k,l=1}^{K}w_{kl}N_kN_l(\mu_k-\mu_l)(\mu_k-\mu_l)^T \tag{4}$$

where $\{w_{kl}\}$ is a set of weights. w_{kl} is a non-negative weight assigned to class pair (k,l). w_{kl} represents how important it is to discriminate class k from class l.

At first glance, there does not seem to be much relation between equations (3) and (4). Let us assume uniform weights for each class pair, $w_{kl}=1$. In other words, each pairwise scatter contribute equally to the between class scatter matrix.

$$\begin{aligned} & B_{\text{uniform}} \\ =\ & \frac{1}{2N}\sum_{k,l=1}^{K}N_kN_l(\mu_k-\mu_l)(\mu_k-\mu_l)^T \\ =\ & \frac{1}{2N}\sum_{k,l=1}^{K}N_kN_l(\mu_k-\mu+\mu-\mu_l)(\mu_k-\mu+\mu-\mu_l)^T \\ =\ & \sum_{k=1}^{K}N_k(\mu_k-\mu)(\mu_k-\mu)^T=B \end{aligned}$$

It is interesting that if we use uniform weight, the new between class scatter matrix (4) is exactly the same as conventional between class scatter matrix. Therefore, it turns out that the definition B_w is a generalization of conventional between class scatter matrix.

With this new definition of between class scatter matrix, it is easy to understand why canonical between class scatter matrix ignores the information about the pairs of classes which are close to each other in Example 1. In the expression of B_{uniform}, we just sum up the pairwise scatters $(\mu_k - \mu_l)(\mu_k - \mu_l)^T$. Obviously, it is in favor of those class pairs (k, l) with large $(\mu_k - \mu_l)$ because later we search for a relative eigenvector of $Bv = \lambda Wv$ with relatively bigger eigenvalue. It is unlikely that the contribution from pair (k', l') can compete with the contribution from pair (k, l) if $\mu_{k'} - \mu_{l'}$ is much smaller compared with $\mu_k - \mu_l$. If such a situation happens, we will lose the discriminant information between classes k' and l'. In fact, what is desired is the opposite effect, that is the classes that are closer (or more confusable) should be weighted more for maximum discrimination.

3.1. Normalization Weight based on Euclidean Distance

In order to keep enough discriminant information, we need to adjust the weights. A natural candidate is a normalization weight equal to the square of the inverse of the Euclidean distance between class means.

$$w_{kl} = \frac{1}{\|\mu_k - \mu_l\|^2} = \frac{1}{(\mu_k - \mu_l)^T(\mu_k - \mu_l)}$$

We weight the classes which have their means closer to each other more than the ones which have means farther. In this sense, more confusable classes are weighted more and less confusable classes are weighted less. According to the normalization weight, the between class scatter matrix is

$$B_{\text{norm}} = \frac{1}{2N} \sum_{k,l=1}^{K} N_k N_l \frac{(\mu_k - \mu_l)(\mu_k - \mu_l)^T}{(\mu_k - \mu_l)^T(\mu_k - \mu_l)} \qquad (5)$$

Next, we apply the above scatter to Example 1.

Example 1 (revisited) :
For the problem defined in example 1, the between class covariance computed using (5) is now as follows:

$$\frac{1}{4}B_{\text{norm}} = \begin{pmatrix} 1 + \frac{1}{1+\delta^2} & 0 \\ 0 & 1 + \frac{\delta^2}{1+\delta^2} \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

So, using the new between covariance matrix (5), no matter how close the pairs $(1, 4)$ and $(2, 3)$ are, we take their spread into account in computing the WPS-LDA. This is very desirable regarding the classification problem.

Once we get a new between class scatter matrix, we solve equation (1) with the new between class scatter matrix. The rest remains the same. We need to solve the generalized eigenvector problem $B_w v = \lambda W v$ to compute the WPS-LDA projection matrix θ.

3.2. Other Weights based on Euclidean Distance

It is possible to use other weights w_{kl}. For instance, to emphasize the discriminant information for those classes close to each other, we can use the square of the previous weights:

$$w_{kl} = \frac{1}{((\mu_k - \mu_l)^T(\mu_k - \mu_l))^2}.$$

With these weights, the between class scatter matrix of Example 1 is

$$4B_w = \begin{pmatrix} 1 + \frac{1}{(1+\delta^2)^2} & 0 \\ 0 & \frac{1}{\delta^2} + \frac{\delta^2}{(1+\delta^2)^2} \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 \\ 0 & +\infty \end{pmatrix}$$

So the closer are the class pairs $(1, 4)$ and $(2, 3)$, the more we take them into account. It sounds reasonable in theory, but it could be unstable in practice.

Actually, any decreasing function of a distance measure can be applied as the weight, *i.e.* $w_{kl} = f(d(k, l))$ where $d(k, l)$ is a metric between two classes k and l and $f(\cdot)$ is a monotonically decreasing function in $\mathbb{R}^+$. For example, in the above choices we used $d(k, l) = ||\mu_k - \mu_l||$ and $f(t) = 1/t^2$ or $f(t) = 1/t^4$. It is not clear what function $f(\cdot)$ is the most appropriate one. This might require some experimentation and might differ from system to system.

3.3. Kullback-Leibler Distance

The weights we introduced above do not consider the within class covariances of each class in computing the distance between them. Obviously, the variance can be a factor in discriminating two classes. Thus, it makes sense to use a distance measure that incorporates the covariance. When each class is assumed to be normally distributed, we can compute KL distance or divergence between them and use it in the weights.

$$w_{kl} = f(D(P_k||P_l)),$$

where P_k represents the Gaussian distribution for class k and $D(P||Q)$ is the KL distance between two Gaussians and $f(\cdot)$ represents a monotonically decreasing function. We have used diagonal covariances for computational simplicity.

4. SPEAKER CLUSTER BASED TRANSFORMATIONS

In most speech recognition systems, it is advantageous to cluster data into distinct groups and build a different speech recognition system based on each cluster especially when there is enough acoustic data for all. A typical clustering is done by separating the training speakers into male and female clusters and training two sets of HMMs corresponding to each. During decoding, appropriate model is determined and used for the test speaker.

It is possible to generate different LDA matrices based on different clusters (or superclusters) as well. This is reasonable since the acoustic characteristics of male and female speech vary widely.

We performed WPS-LDA separately on male and female clusters. This approach yielded more reduction in the word error rate as shown in the next section.

5. IMPLEMENTATION AND RESULTS

We implemented LDA and WPS-LDA on around 300K sentences of training data composed of read continuous speech, spelling, isolated speech and spontaneous speech recorded using headset microphone for dictation. 9 frames of Melcepstra were concatenated to form the initial feature vector and LDA is performed on this data to reduce the feature dimension to 40. HMM models were trained with 3547 context dependent phone states and a total of 42672 Gaussian mixture components representing state output distributions.

The baseline LDA system and the new method WPS-LDA were evaluated using various testsets containing continuous read speech by native and non-native speakers, spontaneous speech, spelling and teen speech data. The word

error rate results are shown in Tables 1 and 2. Here the columns represent different LDA and systems used. LDA is the baseline LDA system with no speaker clustering. WLDA(1) is the WPS-LDA system with the reciprocal of the square of the Euclidean distances used as the weights (*i.e.* $f(t) = 1/t^2$). For WLDA(2), a different weight function $f(t) = 1/t^4$ is used. LDA-C4 is a 4-cluster system used with one LDA matrix. Four clusters correspond to male, female, teen male and teen female clusters. Four different sets of HMMs are trained, one for each cluster, but one LDA matrix is used. WLDA-C4 is same as LDA-C4 but WPS-LDA with $f(t) = 1/t^2$ and again Euclidean distance measure is used. And finally WLDA2-C4 represents a system with two WPS-LDA transformations and 4 HMM clusters. In this case, two different WPS-LDA matrices are found for male and female speakers separately.

It can be seen that both speaker clustering and WPS-LDA reduces the WER. The gains in doing WPS-LDA is more for the clustered system than the speaker independent system, probably due to less variation in acoustic features corresponding to same HMM states due to less speaker variation which enables better separation of acoustic classes.

Testset	LDA	WLDA(1)	WLDA(2)	#words
BOS	11.55	10.60	10.98	11180
NRR	14.83	14.39	14.26	9060
EBOS	9.59	9.58	9.41	11167
ENRR	11.45	11.10	11.12	18093
KID	23.50	22.71	23.00	23457
SPO	25.48	25.50	25.42	28846
AVERAGE	18.31	17.93	17.98	101803

Table 1: Comparison of word error rates for various testsets for speaker independent systems. BOS, NRR, EBOS, ENRR are native speaker read speech testsets, KID is teenager read speech, SPO is spontaneous speech testsets.

Testset	LDA-C4	WLDA-C4	WLDA2-C4	#words
BOS	10.82	9.96	10.08	11180
NRR	13.19	13.25	12.31	9060
EBOS	9.48	9.18	8.77	11167
ENRR	10.67	10.30	9.79	18093
KID	20.26	18.86	18.51	23457
SPO	23.02	22.26	21.17	28846
AVE1	16.49	15.76	15.17	101803
LMNEW	14.59	14.03	13.69	21155
NBOS	25.90	23.74	23.77	16489
SPELL	4.11	3.61	3.31	6076
AVE2	16.76	15.89	15.43	145523

Table 2: Comparison of word error rates for various testsets for speaker clustered systems . In addition to the testsets in Table 1, we used LMNEW (native speaker read speech), NBOS (nonnative read speech) and SPELL (spelling) testsets. AVE1 error rate is over the testsets same as in Table 1. AVE2 error rate is average error rate over all testsets.

We also tried using KL distance based weights. The class distributions were assumed to be diagonal covariance Gaussians and KL distance between them was used in the weights. $f(t) = 1/t^2$ as before. The results for female speaker cluster is shown in Table 3. KL distance did not result in better performance than the Euclidean distance.

Testset female	WLDA	WLDA-KL	#words
BOS	11.22	11.63	6704
NRR	16.36	15.98	3624
EBOS	6.03	6.17	5571
ENRR	8.23	8.40	9038
KID	17.21	17.47	12287
SPO	23.31	25.11	15403
LMNEW	16.92	16.73	8462
NBOS	23.02	22.51	3354
SPELL	3.83	4.24	3631
AVERAGE	17.51	18.23	68074

Table 3: Comparison of word error rates for various testsets for *female cluster only* using Euclidean and KL distance measures in the weights.

6. CONCLUSION AND FUTURE WORK

We show that weighted pairwise scatters in LDA improve WER in a large vocabulary continuous speech recognition test. Among different weights, the normalization weight based on Euclidean distance is simple and works best. KL distance does not help much even though it utilizes within class covariances. WPS-LDA transformations help reduce the error rate for both speaker independent system and speaker clustered system. But, it helps more in the speaker clustered system.

It might be possible to improve the performance of WPS-LDA by considering N-best confusability between classes and modeling classes with a Gaussian mixture distribution instead of single Gaussian distribution.

7. REFERENCES

[1] R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, volume 1, pp. 13–16, 1992.

[2] E. G. Schukat-Talamazzini, J. Hornegger, and H. Niemann, "Optimal linear feature space transformations for semi-continuous hidden Markov models," in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, pp. 369–72, 1995.

[3] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–97, 1998.

[4] G. Saon, M. Padmanabhan, R. A. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, 2000.

[5] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, John Wiley & Sons, New York, 1973.