Review on Heteroscedastic Discriminant Analysis

Bohumir Jelinek

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762 USA
email: jelinek@isip.mstate.edu

ABSTRACT

Discriminant feature spaces are attractive way to improve the word error rate performance of the speech recognition systems. Heteroscedastic discriminant analysis (HDA) is a generalized method for the feature space transformation that does not impose the equal within-class covariance assumptions required by the standard linear discriminant analysis (LDA). It will be shown that the combination of HDA with the maximum likelihood linear transformation (MLLT) leads to the increased classification accuracy even though HDA alone actually degrades word recognition performance.

Theoretical review of the mentioned techniques will be provided and contribution of the reference and the reviewed article will be evaluated in this paper.

1. Introduction

We always use some kind of preprocessing to get the significant features of the speech signal. The most systems currently use the MEL frequency cepstral coefficients supported by their derivatives and acceleration values. Each speech class have its own Gaussian mixture distributions to model its own specific feature distribution.

Principal component analysis (PCA) is a technique which is able to find the feature components that contribute to the data representation most significantly. Less representative components are discarded and that leads to the dimensionality reduction.

For the speech recognition application we prefer the LDA, because LDA looks for the components that are most significant for the discrimination between data in different classes, while it still provides the benefit of the dimensionality reduction.

The reviewed paper provides a comparison between the results of the discriminative techniques imposing the diagonal variance modelling constraint and the full variance modelling case in the projected space.

The authors proved that the combination of the HDA and MLLT provides a slight improvement in the classification performance, but they have not explained the failure of the alone HDA projection.

2. Article Review

Basic feature analysis techniques used in most systems are outlined in the Section 1 of the article. Short history of the HDA applications development is reviewed.

In the Subsecion 2.1 authors provide basic equations of the LDA. The objective function of the LDA is given and the solution method is shown.

Subsection 2.2 introduces objective function of the heteroscedastic extension. The gradient method for the optimization is suggested because the analytical solution for stationary points is not available.

Subsection 2.3 derives a likelihood interpretation of the HDA objective function. It is shown that HDA feature transformation can be interpreted as a constrained ML projection, with the constraint given by the maximization of between-class scatter volume. Authors then provide an objective function for a combined HDA and MLLT projection called diagonal HDA (DHDA).

The results of the conducted experiments are outlined in the Subsection 4.

3. Multiple Discriminant Analysis

LDA applies a linear feature space transformation

$$\boldsymbol{y} = \boldsymbol{W}^t \boldsymbol{x} \tag{1}$$

where **x** is *d*-dimensional input sample vector belonging to one of the *c* classes (categories), ***y*** is transformed sample vector in the *c-1* dimensional projected space and **W** is projection matrix.

The goal of LDA is to find a projection matrix **W** resulting in the best possible separation of the classes in the projected space.

For two classes problem in two dimensional space the transformation (1) means just a projection of the sampled vectors onto the line perpendicular to the projection vector **W**. In this case **y** is a scalar value. It can be seen that the direction of the projection vector is important for the discriminative properties of the transformed space.

We define two scatter matrices, between-class scatter matrix $\mathbf{S}_\mathrm{b}$ and within-class scatter matrix $\mathbf{S}_\mathrm{i}$ that are used to find the best direction of the projection vector **W**. Within-class scatter matrix $\mathbf{S}_\mathrm{i}$ is the measure of variability of the samples in one class and it is defined as

$$\boldsymbol{S}_i = \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t \tag{2}$$

where $\mathbf{m}_\mathrm{i}$ is the mean of the class C_i

The summation of within-class scatter matrices $\mathbf{S}_\mathrm{i}$ defines total within-class variability $\mathbf{S}_\mathrm{w}$

$$\boldsymbol{S}_W = \sum_{i=1}^{c} \boldsymbol{S}_i \tag{3}$$

If *n* denotes the total number of samples and n_i the number of samples belonging to the class C_i, then the total mean vector is defined by

$$\boldsymbol{m} = \frac{1}{n} \sum_{\boldsymbol{x}} \boldsymbol{x} \tag{4}$$

and the mean vector for individual class is

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x} \tag{5}$$

Between-class scatter matrix $\mathbf{S}_\mathrm{b}$ is given by

$$\boldsymbol{S}_b = \sum_{i=1}^{c} n_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^t \tag{6}$$

and it provides a measure of between-class variability.

Using the scatter matrices we define a criterion function *J*(**W**)

$$J(\boldsymbol{W}) = \frac{\left| \boldsymbol{W}^t \boldsymbol{S}_b \boldsymbol{W} \right|}{\left| \boldsymbol{W}^t \boldsymbol{S}_W \boldsymbol{W} \right|} \tag{7}$$

By maximizing this criterion function we get a projection vector that provides a maximum discrimination of samples in the projected space.

Solution to this maximization leads to eigenvalue problem. The columns of an optimal **W** are the eigenvectors corresponding to the largest eigenvalues in

$$\boldsymbol{S}_b \boldsymbol{w}_i = \lambda_i \boldsymbol{S}_w \boldsymbol{w}_i \tag{8}$$

The main purpose of multiple discriminant analysis is the dimensionality reduction. Another result of the feature space transformation can be a possibility to apply a classification technique that would not be feasible in the original space.

4. Extension to HDA

The heteroscedastic discriminant analysis is a generalization of LDA that handles unequal sample

covariance classes.

The objective of LDA is a reduction of the feature dimension by choosing a p dimensional subspace of the feature space and rejection of an (n-p) dimensional subspace. The implicit assumption is that (n-p) subspace does not carry any significant information for classification. For Gaussian modes it is significant to the assumption that the means and the variances in the rejected subspace are the same for all classes. We can consider HDA to be the constrained maximum likelihood projection where log likelihood of the samples in the projected space is maximized. The constraint is given by the maximization of the projected between-class scatter volume. The detailed and explanatory description of this approach can be found in Kumar and Andreou paper [5].

In the reviewed article authors generated heteroscedastic extension of LDA by introducing a modified objective function for HDA analysis. The modified objective function takes into account weighted contributions of the individual classes

$$\prod_{i=1}^{c}\left(\frac{\left|\boldsymbol{w}^t\boldsymbol{s}_b\boldsymbol{w}\right|}{\left|\boldsymbol{w}^t\Sigma_i\boldsymbol{w}\right|}\right)^{n_i} = \frac{\left|\boldsymbol{w}^t\boldsymbol{s}_b\boldsymbol{w}\right|^{n}}{\prod_{i=1}^{c}\left|\boldsymbol{w}^t\Sigma_i\boldsymbol{w}\right|^{n_i}} \tag{9}$$

where Σ_i is covariance matrix of the class C_i given by

$$\Sigma_i = \frac{1}{n_i}\boldsymbol{S}_i \tag{10}$$

By taking log of (9) we get a discriminant function:

$$H(\boldsymbol{W}) = \sum_{i=1}^{c} -n_i\log\left|\boldsymbol{W}^t\Sigma_i\boldsymbol{W}\right| + n\log\left|\boldsymbol{W}^t\boldsymbol{S}_b\boldsymbol{W}\right| \tag{11}$$

The HDA solution, similar to LDA solution, is invariant to linear transformations of the data in original space. Moreover, objective function *H* has another invariant property. It is invariant to subsequent projected feature space transformations.

We will maximize objective function (11) by the matrix differentiation

$$\frac{d}{d\boldsymbol{W}}H(\boldsymbol{W}) = \sum_{i=1}^{C} -2n_j\left(\boldsymbol{W}^t\Sigma_j\boldsymbol{W}\right)^{-1}\boldsymbol{W}^t\Sigma_j + n\left(\boldsymbol{W}^t\boldsymbol{S}_b\boldsymbol{W}\right)^{-1}\boldsymbol{W}^t\boldsymbol{S}_b \tag{12}$$

Since there is no analytical solution for the local minima, gradient descent method is used for the optimization of *H*.

5. Likelihood interpretation

If we assume single full covariance Gaussian model for each class, HDA transformation defined by $\boldsymbol{y} = \boldsymbol{W}^t\boldsymbol{x}$ provides us following expression for the sample log likelihood in the transformed space:

$$\sum_{i=1}^{C} -\frac{n_i}{2}\log\left|\boldsymbol{W}^t\Sigma_j\boldsymbol{W}\right| + C \tag{13}$$

We can see the similarity between this equation and the equation (11) suggesting that HDA is a constrained maximum likelihood projection if constraint is given by the maximization of the projected between-class scatter volume.

Let's consider a diagonal covariance modeling constraint in the target feature space. In this case we use a maximum likelihood linear transform. MLLT aims at minimizing the loss in likelihood between the full and diagonal covariance gaussian models

$$\begin{aligned}\hat{\Psi} &= \arg max \sum_{i=1}^{C} -\frac{n_i}{2}\langle \log\left|diag\left(\Psi^t\hat{\Sigma}_i\Psi\right)\right| - \log\left|\Psi^t\hat{\Sigma}_i\Psi\right|\rangle \\ &= \arg max \sum_{i=1}^{C} -\frac{n_i}{2}\log\left|diag\left(\Psi^t\boldsymbol{W}^t\Sigma_i\boldsymbol{W}\Psi\right)\right| + n\log|\Psi|\end{aligned} \tag{14}$$

Because of the mentioned invariant property, the objective function (11) is the same for both the composite $\Psi\boldsymbol{W}$ and original $\boldsymbol{W}$ transformations. The authors refer to this composite transform as the

maximum likelihood discriminant (MLD) projection.

By means of (11) and (14) we can combine the HDA and MLLT projection through the following objective function

$$G(\boldsymbol{W}) = \sum_{i=1}^{C} -n_i \log\left|diag(\boldsymbol{W}^t \Sigma_i \boldsymbol{W})\right| + n\log\left|\boldsymbol{W}^t \boldsymbol{S}_b \boldsymbol{W}\right| \quad (15)$$

The above approach is referred as the diagonal HDA (DHDA) projection. This approach provides different results than Kumar's HDA, because DHDA objective function directly maximize the between-class separation $\left|\boldsymbol{w}^t \boldsymbol{s}_b \boldsymbol{w}\right|$ in projected space, while Kumar's HDA assumes this is done implicitly by minimizing $\left|diag\langle \boldsymbol{W}^t \boldsymbol{S}_b \boldsymbol{W}\rangle\right|$.

6. Experimental results

The experiments were performed on voicemail and Switchboard database tasks. The baseline system used 39-dimensional frames. For the LDA and HDA versions, every 9 consecutive 24-dimensional cepstral vectors were spliced together forming 216 dimensional feature vectors. Subsequently, a 39x216 transformation $\boldsymbol{W}^t$ was computed using the objective function for LDA, HDA and DHDA, which projected a feature space down to 39 dimensions. After feature vectors were reclustered, an MLLT was computed to leading to LDA+MLLT and HDA+MLLT (MLD) feature space.

As already mentioned in Alphonso's article [1], the claimed advantage of using HDA were not confirmed by experiments. Application of HDA significantly decreased the baseline MFCC system performance.

The final improvement was achieved by MLLT application. The article has a represents a high knowledge of the research team, but does not provide enough explanatory information about used transformations.

Practical application introduces additional consideration - e.g. determination of the reduced feature space dimension [4].

7. Summary

The application of the suggested technique in large vocabulary continuous speech recognition area system seems promising. The recent improvements obtained at IBM research center claim 10-15% relative performance improvement.

8. Acknowledgments

Genuine affection to my wife Veronika whose encouragement contributed greatly to this review.

Thanks go also to my supervisor Dr. Joseph Picone for forcing me to do painful steps to increase the quality of my life.

References

[1] I. Alphonso, "Heteroscedastic Discriminant Analysis", Critical Review Paper, ECE 8990 - Special Topics in ECE - Pattern Recognition, Institute for Signal and Information Processing, Mississippi State University, 2001.

[2] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 5, pp. 179-190, 1983.

[3] R.E.Duda, P.E.Hart, D.G.Stork, Pattern Classification, John Wiley & Sons, 2001.

[4] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 2, pp. 661-664, Seattle 1998.

[5] N. Kumar, A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," Speech Communication, 26, pp. 283-297, 1998.

[6] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen, "Maximum likelihood discriminant feature spaces", Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. II1129-II1132, Beijing, China, 2000.