Generalized Oja's Rule for Linear Discriminant Analysis with Fisher Criterion

Jose C. Principe, Dongxin Xu, Chuan Wang

Computational NeuroEngineering Laboratory
CSE 447 Electrical and Computer Engineering Department
University of Florida, Gainesville, FL 32611, USA
xu@synapse.cnel.ufl.edu

ABSTRACT

On-line learning rules for both Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) with Fisher criterion are analyzed under the same framework, and a generalized Oja's rule for both is derived. For the LDA problem, the relationship between the Fisher criterion and the criterion of minimizing Mean Square Error (MSE) is discussed. The experiments show that the convergence speed of the generalized Oja's rule as an adaptive method for Fisher Criterion is much faster than that of gradient descent method for MSE criterion.

1. INTRODUCTION

PCA (also known as KL Transform) and LDA are both standard statistical tools for data analysis. PCA focuses on representation while LDA concentrates on discrimination. Both have extensive applications in data compression, feature extraction, pattern recognition, etc. ([1], [2], [5], [6], [7], [8]), and have been well studied ([1], [2]). Actually, in spite of the difference, they are quite similar: the closed form solutions to both problems are the eigenvectors of the input covariance matrix (for PCA) ([2]) or a more complex matrix (for LDA, known as the generalized eigenvector problem) ([1]). Oja did pioneering work by relating PCA with a single layer linear network and providing an on-line local learning rule, known as Oja's rule ([4]). Other researches extended the rule to cover multiple components (e.g. [5], [6], [9]). However, to the best of our knowledge, no one has extended Oja's rule to LDA before. Although [7] compares PCA and LDA, and proposes a two-layer PCA network with two step training for LDA, it is actually the application of PCA to LDA and does not provide insight into the essence of these two similar problems. Recently, Chattterjee and Roychowdhury ([10]) independently proposed a two-layer linear network for LDA which uses supervised learning scheme. Although the algorithm they derived is similar to Oja's or Sanger's rule, the connection between them is not clearly revealed. In this paper, within the unsupervised learning scheme, a unified point of view is given under the framework of gradient descent or ascent learning on related cost functions. Oja's rule is generalized so as to be applicable to both PCA and LDA problems.

Previous work has shown that an alternate solution (optimal projection) to LDA (Fisher Criterion) can be obtained by minimizing the MSE criterion when the desired output is designed in a specific way ([1], [2], [3]). We argue that the generalized Oja's rule solution to Fisher criterion and the gradient method solution to MSE criterion have quite different properties because the performance surface of both criteria differs greatly. It is easy to show that the Fisher criterion actually looks for a line in the data space while the MSE criterion searches for a point. The equivalence just means that the optimal point of the MSE criterion is located in the optimal line of Fisher criterion. Intuitively, searching for the position of a line is much easier than searching for a point. This has been supported by the results of several experiments where the convergence speed of the generalized Oja's rule is much faster because a larger learning step size can be selected without producing divergence in the adaptation.

The implications of this result for pattern recognition are important. Old arguments state that minimization of the MSE criterion has no direct relationship to the goal of discrimination between data clusters with arbitrary distributions. We use it because it is easy to derive on-line learning algorithms even for nonlinear networks (i.e. backpropagation). Comparatively, the Fisher criterion has a more direct relation to classification, because it provides the best linear projection for discrimination. But it is not as popular as the MSE criterion because it requires classes with different means, and there is no efficient on-line learning algorithm. In this paper, we do provide an efficient on-line adaptation rule to implement the Fisher criterion for the two class case. We also shed light on how the two criteria are related, in what sense they are different and which one is better. In the following sections, we will summarize the major points of our work. The details can be found in [11].

2. Oja's rule for PCA & LDA

The one component PCA algorithm and the two class

LDA (Fisher criterion) can be described as follows:

PCA: $X = (x_1, x_2, \ldots, x_p)$ represents a set of data with p samples, where $x_i \in R^n$. Without loss of generality, the data set is assumed to be zero-mean. The problem is to find a vector $w \in R^n$ which maximizes (1):

$$J = \frac{w^T S w}{w^T w} \qquad \text{(EQ 1)}$$

where w^T is the transpose and $S = XX^T = \sum_i x_i x_i^T$ is the data scatter.

LDA: $X_1 = (x_{11}, x_{12}, \ldots, x_{1p_1})$, $X_2 = (x_{21}, x_{22}, \ldots, x_{2p_2})$, are two sets of data for class 1 and class 2 respectively, $x_{ij} \in R^n$. Let m_1 and m_2 be the means for X_1 and X_2. Let $m = m_1 - m_2$, $X = (X_1 - m_1 u_1, X_2 - m_2 u_2)$, where $u_1 = (1, 1 \ldots 1)_{1 \times p_1}$, $u_2 = (1, 1 \ldots 1)_{1 \times p_2}$. The problem is to find a vector $w \in R^n$ which minimizes (2):

$$J = \frac{w^T S w}{w^T S_B w} = \frac{w^T S w}{w^T m m^T w} = \frac{w^T S w}{(w^T m)^2} \qquad \text{(EQ 2)}$$

where $S_B = mm^T$ is the between-class scatter, and $S = XX^T = \sum_i x_i x_i^T$ is the within-class scatter. Note that the usual Fisher Criterion is the maximization of the inverse of (2). The reason for using (2) will be clear in the following.

From (1) and (2), it is obvious that the norm of w: $\|w\|$, is irrelevant to both criteria. This implies we can keep $w^T w = 1$ for (1) and $w^T m = 1$ for (2). Another consequence of this fact is that the gradients for both criteria are always perpendicular to vector w. (3) gives the gradient of (2) where a factor of 2 is ignored,

$$\nabla_w J = \frac{1}{\|w\|^2} \sum_i y_i \left(x_i - \frac{y_i}{\|w\|^2} w \right) \qquad \text{(EQ 3)}$$

If we keep $w^T w = 1$, (3) will become (4) which is the same as Oja's rule,

$$\Delta w = \sum_i y_i (x_i - y_i w) \qquad \text{(EQ 4)}$$

Usually $w^T w \neq 1$, so the updating of weights in Oja's rule can be decomposed into two components:

$$\Delta w = \Delta w_w^{\perp} + \Delta w_w \qquad \text{(EQ 5)}$$

where $\Delta w_w^{\perp} = \|w\|^2 \nabla_w J$ is the gradient component which is perpendicular to w and $\Delta w_w = ((1 - \|w\|^2) / \|w\|^2) (\sum y_i^2) w$ is the component which is along the direction of w. Now, let's introduce the idea of "Base Vector" which serves as the basic measurement for the problem of concern. For PCA, the major issue is signal representation. Thus, the base in PCA is the normalized vector w ($\|w\| = 1$). So, yw is just the projected version of datum x on the Base Vector w and $x - yw$ is the difference between x and its projected version. As Fig. 1 (a) shows, the gradient component in Oja's rule forces the projected version of data towards the data themselves. Fig.1 (b) and (c) show that when $\|w\| \neq 1$, Oja's rule has the component Δw_w which works as the negative feedback about the norm of w, forcing $\|w\|$ to be 1. This simply explains why in Oja's rule w will converge to unit length even without normalization. Unfortunately, when Oja's rule is applied to the minor component (the eigenvector with smallest eigenvalue) analysis, the feedback becomes positive. This is the reason why Oja's rule is unstable for minor components. Any modification to Oja's rule which keeps the gradient component unchanged but reverses the sign of the feedback will work for minor component. [9] gives an example. Actually, explicit normalization $w^T w = 1$ will also achieve the same goal.

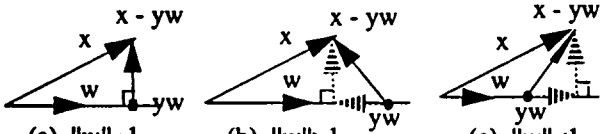

Figure 1. Illustration of Oja's Rule for PCA

Similarly, we can get the gradient for LDA as (6):

$$\nabla_w J = \frac{1}{(w^T m)^2} \sum_i y_i \left(x_i - \frac{y_i}{w^T m} m \right)$$
$$\Delta w = \sum_i y_i (x_i - y_i m) \qquad \text{(EQ 6)}$$

Note the similarity of between (6) and (4), where the only

difference is the substitution on w by m, the mean difference vector. For classification it is reasonable to use the mean difference as the basic measurement. Now, unifying (4) and (6), we get the generalized rule for both PCA and LDA as (7),

$$\Delta w = \sum_i y_i (x_i - y_i b) \qquad \text{(EQ 7)}$$
$$w^T b = 1$$

where b is the Base Vector which is w for PCA and m for LDA. For the major component of PCA, the explicit normalization $w^T b = 1$ is not necessary. But for both the minor component algorithm and LDA with Fisher Criterion, the explicit normalization is necessary for stability.

To further explain the generalized rule, (7) can be rewritten as (8):

$$\Delta w = \sum_i y_i [\, (x_i - y_i e_w) + y_i (e_w - b)\,] \qquad \text{(EQ 8)}$$
$$where, \qquad e_w = w / \| w \|$$

There are 2 terms, the first one $x_i - y_i e_w$ is the difference between the input sample and its projected version; the second term $e_w - b$ is the difference between the normalized vector e_w and the base vector b. So, there are two kind of forces to push the vector w during adaptation. The first force comes from the zero-mean data, and the second force comes from the base vector. For the PCA problem, the second term will disappear because the base vector is just the normalized vector itself. For the LDA problem, the base vector is the mean difference. So, in this case, the data will repel their projected version (gradient descent) while the mean difference will attract the normalized vector towards itself. Since e_w is unit length, a weighting factor y_i is attached to the second term to balance the contributions.

3. IMPLEMENTATION

Figure 2 shows the network to solve the two-class LDA problem with Fisher criterion.

It is a linear network with n inputs and a single output. From (6), the on-line weight adjustment is computed as

$$\Delta w = y (x - ym) \qquad \text{(EQ 9)}$$

From (4), we can see that the structure of the algorithm is very similar to Oja's rule. The major difference is the need to pre-compute the mean difference vector m between the data clusters. We also must remember that there is a required normalization for the weights as shown in (7). So the algorithm can be described as follows:

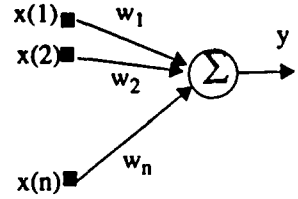

Figure 2. Network for 2 class LDA

1. Initialize parameter set w
2. calculate the mean of each of two classes data, make the data zero-mean with respect to their class means, calculate the mean difference m.
3. get $y_m = w^T m$ *, then normalize w by* $w = w / y_m$
4. for the input data x_i *do (batch learning)*

$$y_i = w^T x_i$$
$$\Delta w = \sum y_i (x_i - y_i m)$$

5. update w by $w = w - \eta \Delta w$ *, where* η *is the learning step size.*
6. if finish, then exit; otherwise, go to 3.

4. RELATIONSHIP BETWEEN FISHER & MSE

Fig. 3 shows a simple two class classification problem in 2D space. For these two classes the Fisher discriminant is a line as shown in the figure. The MSE solution using the desired response as defined in the literature [1], is shown as a point near the origin of the input space. The optimal MSE point exists in the Fisher discriminant line.

Although the solutions (the optimal projection) to both criteria are equivalent, their performance surfaces are quite different, as shown in Fig 3. The MSE performance surface is a paraboloid, while the Fisher surface is a set of two triangular shaped folds that meet at the origin (the Fisher surface plot was clipped for better visualization).

Therefore, the two algorithms are expected to have very different properties. Fig. 4 shows the learning curves for both methods with different step sizes η (vertical axis represents the angle between the optimal Fisher line and the current vector, the horizontal axis is the iteration index). For small step sizes both methods have similar adaptation speeds. However, for η=0.1 the MSE is divergent, while the Fisher method is able to find the optimal line in 5 iterations. The MSE never converged faster than the Fisher method. For η=0.5 both methods diverge. This means that the Fisher criterion converged one order of magnitude faster than the MSE for this problem. Similar speed improvements were obtained for other problems.

Although the previous research ([1], [2], [3]) has shown that the solution to the MSE criterion is equivalent to that of the Fisher criterion for a specific choice of the desired output, the relationship between the optimal values for the criteria was never given. In [11], we have proven that there is a very simple relation shown as (10):

$$J_f^o = \frac{n_t^4}{n_1^2 n_2^2} \frac{1}{J_m^o} - \frac{n_t}{n_1 n_2} \qquad \text{(EQ 10)}$$

where J_f^o and J_m^o are the optimal values for the Fisher and MSE criteria respectively, n_t is the number of total samples, n_1 is the number of samples of class 1, n_2 is the number of samples of class 2. Notice that (10) doesn't depend on the input data at all.

5. CONCLUSION

The connection for the two class case between the gradient method and Oja's rule is established in this paper which enables the extension of Oja's rule to LDA with Fisher criterion. For the multiple class case, further extensions have been made which will be presented in a later paper. Although the optimal solutions to both Fisher and MSE criteria for LDA are the same (except for a scale factor), the gradient adaptation for both criteria are quite different because of the difference of the performance surface. For the two-class case, the experiments have shown the effectiveness (faster convergence) of the generalized Oja's rule as the adaptation solution to Fisher criterion. We are investigating the effectiveness of the solution for the multiple class case.

ACKNOWLEDGEMENT

This work was partially supported by NSF grant ECS-9510715 and ONR grant N00014-94-1-0858.

REFERENCES

[1] Richard O. Duda, Peter E. Hart "Pattern Classification and Scene Analysis" John Wiley & Sons Inc. 1973.
[2] Christopher M. Bishop "Neural Networks for Pattern Recognition". Clarendon Press, Oxford. 1995.
[3] J. S. Koford, G. G. Groner "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier". IEEE Trans. on Information Theory. Vol. 12. No. 1. Jan. 1966. 42-50.
[4] Erkki Oja "A simplified neuron model as a principal component analyzer" Journal of Mathematical Biology 15. 267-273.
[5] Terence D. Sanger "Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network" Neural Network, Vol 2. pp 459-473. 1989.
[6] S. Y. Kung, K. I. Diamantaras, J. S. Taur "Adaptive Principal Component EXtraction (APEX) and Applications" IEEE Trans. on Signal Processing. Vol. 42, No. 5 May 1994.
[7] Jianchang Mao and Anil K. Jain "Artificial Neural Networks for Feature Extraction and Multivariate Data Projection" IEEE Trans. on Neural Network. Vol 6. No. 2. March 1995.
[8] T. Hastie, A. Buja, R. Tibshirani "Penalized Discriminant Analysis" Tech. Report. AT&T Bell Labs. May 1994
[9] Erkki Oja "Principal Components, Minor Components, and Linear Neural Networks" Neural Networks, Vol 5, No. 6 pp. 927-935, 1992
[10] Chanchal Chatterjee and Vwani Roychowdhury "Self-Organization with Hetero-Associative Networks for Linear Discriminant Analysis", World Congress on Neural Networks, San Diego, California, Sept., 1996.
[11] Dongxin Xu "Generalized Framework of Gradient Method for PCA and LDA". Tech. Rept. CNEL Dept. of ECE. University of Florida. May 1996.

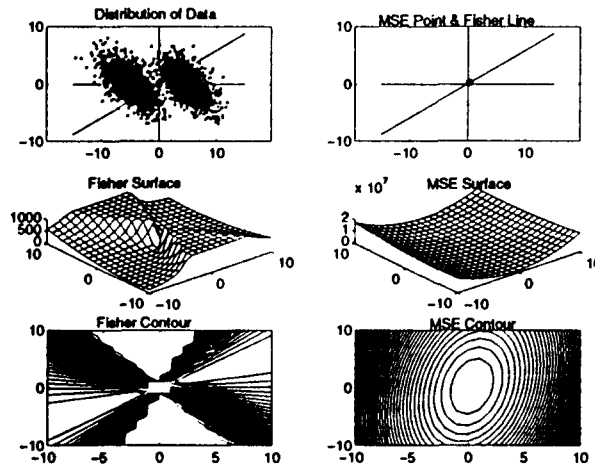

Fig. 3: Comparison of Fisher criterion and MSE for a simple two class problem.

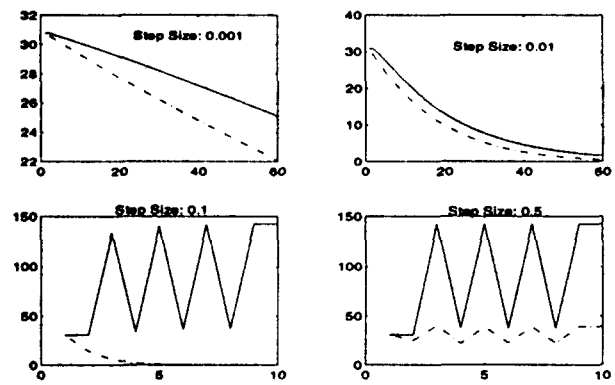

Fig. 4: Learning curves for the classification problem. MSE--solid line; Fisher--dashed line.