2000 International Symposium on Nonlinear
Theory and its Applications (NOLTA2000)
Dresden, Germany, Sep. 17–21, 2000

One-parameter family of nonlinear dynamics for online linear discriminant analysis

Kazuyuki Hiraoka, Ken-ichi Hidai,
Hiroshi Mizoguchi, Taketoshi Mishima, and Shuji Yoshizawa

Dept. of Information and Computer Sciences, Saitama University, 255 Shimo-okubo, Urawa, 338-8570, Japan
email: hira@me.ics.saitama-u.ac.jp

Abstract— A novel variation of online linear discriminant analysis (OLDA) is proposed based on an one-parameter family of nonlinear dynamics. Both previously proposed dynamics for OLDA and Oja's dynamics for principal component analysis are special cases in this family. With regard to this family, fixed-points and their stability are analyzed. The dependence of discrimination performance on the parameter is also studied experimentally.

I. Introduction

Linear discriminant analysis (LDA) has been applied extensively, e.g. for face identification [8][9]. However, LDA is poor at adaptivity. This is because LDA is a batch learning algorithm. Indeed, we have to recalculate the discrimination matrix A every time when we add new data to update the identification system. This calculation is heavy for high dimensional data such as face images.

Recently, the authors have proposed an *online* version of LDA [11], which is referred to online LDA (OLDA). By OLDA, the face identification system can be updated with low computational cost when new additional images are presented. Hence OLDA has the ability of adaptation to the change of environment. OLDA also has an advantage that huge matrices never appear in its calculation.

Though iterative algorithms have been proposed[1] for neural network based LDA [6][5], they are not sufficiently "online". Since those algorithms keep $n^2 \times n^2$ matrices when the image size is $n \times n$, they require $O(n^4)$ time for one step updating and $O(n^4)$ memory. In contrast to [6][5], our OLDA algorithm requires only $O(n^2ML)$ time for one step updating and $O(n^2M)$ memory, where M and L are the number of classes and features, respectively. Note that $n^2 \gg M > L$ in typical cases of face identification tasks.

In the present paper, we will propose a novel OLDA algorithm based on an one-parameter family of matrix dynamics. The first algorithm [11] is a special case of it. Moreover, Oja's dynamics for online principal component analysis (PCA) can be viewed as another special case of it.

II. One-parameter family of matrix dynamics

To determine the discrimination matrix A in LDA, we have to solve a generalized eigenvalue problem. Let B and W be the $N \times N$ variance matrices "between classes" and "within classes" respectively. Then, we have to find an $N \times L$ matrix A_{sol} and an $L \times L$ symmetric matrix Γ which satisfy

$$\begin{aligned} BA_{\mathrm{sol}} &= WA_{\mathrm{sol}}\Gamma, \\ A_{\mathrm{sol}}^T W A_{\mathrm{sol}} &= I, \end{aligned} \tag{1}$$

where I is the identity matrix.

In order to obtain the solution A_{sol} iteratively, the following matrix dynamics is discussed in the present study:

$$\begin{aligned} \frac{d}{dt}A(t) = BA(t) &- \alpha BA(t)A(t)^T WA(t) \\ &- (1-\alpha)WA(t)A(t)^T BA(t), \end{aligned} \tag{2}$$

where $\alpha \in \mathbf{R}$ is a parameter.

III. Main theorems

With regard to the one-parameter family of matrix dynamics (2), the following theorems are proved.

Theorem 1 *All the solutions of (1) are fixedpoints of (2).*

Theorem 2 *Suppose that rankB $\geq L$ and $\alpha = 0$. Then all the stable fixedpoints of (2) are solutions of (1).*

Theorem 3 *Suppose that rankB $\geq L$ and $\alpha > 0, \alpha \neq 1$. If A is a stable fixedpoint of (2) and A is not a solutions of (1), the eigenvalues $\omega_1, \cdots, \omega_L$ of A^TWA must satisfy $\omega_k = 1$ or $\omega_k \geq 1/\alpha$ ($k = 1, \cdots, L$).*

[1] The original algorithm in [5] is a batch learning. However, we can easily modify it to an iterative learning by replacing the calculation of the standard deviation in [5] with an iterative one.

Theorem 1 suggests that the dynamics (2) can be used to solve the generalized eigenvalue problem (1). Theorem 2 shows that there is no "spurious" solution when $\alpha = 0$. Theorem 3 shows that the spurious solutions exists only in a region which is "far" from O compared with the "true" solutions, when $0 < \alpha < 1$. This is because the true solutions satisfy $A^TWA = I$.

Thus, as for spurious solutions, smaller α is better. Especially, $\alpha = 0$ is a direct extension of Oja's dynamics for online PCA [1] and this dynamics has no spurious solution.

On the other hand, as for convergence speed, it is experimentally observed that positive α is superior. This will be discussed later.

IV. Derived OLDA algorithm

From the dynamics (2), following OLDA algorithm is derived.

At every time step $t = 1, 2, 3, \cdots$, a new pair $(\boldsymbol{x}(t), c(t))$ is presented, where $\boldsymbol{x}(t)$ is an N-dimensional data vector, $c(t) \in \{1, \cdots, M\}$ is the class of $\boldsymbol{x}(t)$, and M is the number of classes. Based on this pair, auxiliary variables and the $N \times L$ discrimination matrix A are updated as follows:

$$t^c(t) = t^c(t-1) + \delta(c, c(t)), \tag{3}$$

$$\bar{\boldsymbol{x}}(t) = \left(1 - \frac{1}{t}\right)\bar{\boldsymbol{x}}(t-1) + \frac{1}{t}\boldsymbol{x}(t), \tag{4}$$

$$\bar{\boldsymbol{x}}^c(t) = \begin{cases} \left(1 - \frac{1}{t^c(t)}\right)\bar{\boldsymbol{x}}^c(t-1) + \frac{1}{t^c(t)}\boldsymbol{x}(t) & (c = c(t)), \\ \bar{\boldsymbol{x}}^c(t-1) & (c \neq c(t)), \end{cases} \tag{5}$$

$$\boldsymbol{v}^c(t) = \bar{\boldsymbol{x}}^c(t) - \bar{\boldsymbol{x}}(t), \tag{6}$$

$$\boldsymbol{w}(t) = \boldsymbol{x}(t) - \bar{\boldsymbol{x}}^{c(t)}(t), \tag{7}$$

$$\boldsymbol{y}^c(t) = A(t-1)^T\boldsymbol{v}^c(t), \tag{8}$$

$$\boldsymbol{z}(t) = A(t-1)^T\boldsymbol{w}(t), \tag{9}$$

$$F(t) = \frac{1}{M}\sum_{c=1}^{M}\boldsymbol{v}^c(t)\boldsymbol{y}^c(t)^T, \tag{10}$$

$$\boldsymbol{g}(t) = \frac{1}{M}\sum_{c=1}^{M}\boldsymbol{y}^c(t)\left(\boldsymbol{y}^c(t)^T\boldsymbol{z}(t)\right), \tag{11}$$

$$\begin{aligned} A(t) = A(t-1) + \eta\Big[&F(t) - \alpha F(t)\boldsymbol{z}(t)\boldsymbol{z}(t)^T \\ &- (1-\alpha)\boldsymbol{w}(t)\boldsymbol{g}(t)^T - \alpha\epsilon F(t)\left(A(t-1)^TA(t-1)\right) \\ &- (1-\alpha)\epsilon A(t-1)\left(A(t-1)^TF(t)\right)\Big], \end{aligned} \tag{12}$$

where the parameter α is corresponding to α in (2), the learning coefficient η is a small positive number, and the regularization coefficient ϵ is also a small positive number. The number L of features is less than or equal to $\min(N, M-1)$. Note that the covariance matrix $W(t) = \frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{w}(\tau)\boldsymbol{w}(\tau)^T$ within classes is replaced with the instantaneous value $\boldsymbol{w}(t)\boldsymbol{w}(t)^T$. Such replacement is justified by the theory of stochastic approximation [7]. As for the initial values, $t^c(0) = 0$, $\bar{\boldsymbol{x}}(0)$ and $\bar{\boldsymbol{x}}^c(0)$ are arbitrary vectors, and $A(0)$ is an arbitrary matrix which satisfies $\mathrm{rank}A(0) = L$.

V. Experimental results

We have applied the above OLDA algorithm to face image identification for 4 persons. The performance is evaluated for test samples after $t = 400$ steps of iterations (Fig. 1).

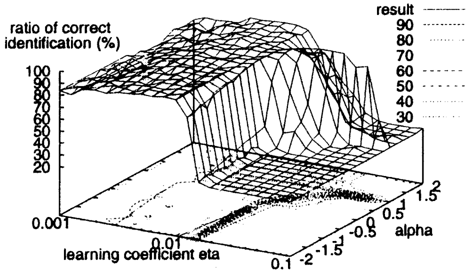

Figure 1: Performance of the proposed method for each η and α. Stability is improved around $\alpha = 1$.

It is experimentally shown that stability for large learning coefficient is improved around $\alpha = 1$. Moreover, the problem of spurious solutions turns out to be negligible if we take an initial matrix $A(0)$ near the zero matrix O.

VI. Discussion

The reason of the slow convergence for $\alpha \leq 0$ is not clearly understood at now. This point must be studied further. A key will be the fact that $\boldsymbol{w}(t)\boldsymbol{g}(t)^T$ has much larger variance than $F(t)\boldsymbol{z}(t)\boldsymbol{z}(t)^T$ in (12).

Another unclear point of the present study is the question whether the dynamics (2) can be derived from potential functions. The authors have only partial results. For the case $\alpha = 1/2$, (2) is derived as

$$da_{ij}/dt = (1/2)(\partial\phi(A)/\partial a_{ij}), \tag{13}$$

$$\phi(A) = \mathrm{Tr}\left[A^TBA\left(I - \frac{1}{2}A^TWA\right)\right]. \tag{14}$$

For the case $\alpha = 0$, (2) is derived similarly by another potential function

$$\psi(A) = \log\det[A^TBA] - \mathrm{Tr}[A^TWA] \tag{15}$$

together with a metric

$$\langle \Delta A_1, \Delta A_2 \rangle_A = \mathrm{Tr}[\Delta A_1^T \Delta A_2 (A^T BA)^{-1}]. \qquad (16)$$

Note that Oja's dynamics [1] is a special case of (2) with $\alpha = 0$ and $W = I$.

VII. Conclusion

A novel variation of online linear discriminant analysis (OLDA) is proposed based on an one-parameter family of nonlinear dynamics. Both previously proposed dynamics for OLDA and Oja's dynamics for principal component analysis are special cases in this family.

It is theoretically shown that (a) when the parameter $\alpha = 0$, no "spurious solution" exists, and (b) when $0 < \alpha < 1$, spurious solutions exists only in a region which is "far" from O compared with the "true" solutions. On the other hand, it is experimentally shown that stability for large learning coefficient is improved around $\alpha = 1$.

Acknowledgements

The authors wish to thank Mr. Morishita for his large help on simulations. This work has been partly supported by CREST of JST (Japan Science and Technology) 279102 and JSPS (Japan Society for the Promotion of Science) 12750203.

References

[1] E. Oja, "A simplified neuron model as a principal component analyzer", J. Math. Biol., Vol. 15, pp. 267–273, 1982.

[2] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network", Neural Networks, Vol. 2, No. 6, pp. 459–473, 1989.

[3] K. Hornik and C.-M. Kuan, "Convergence analysis of local feature extraction algorithms", Neural Networks, Vol. 5, No. 2, pp. 229–240, 1992.

[4] W.-Y. Yan, U. Helmke, and J. B. Moore, "Global analysis of Oja's flow for neural networks", IEEE Tr. NN, Vol. 5, No. 5, pp. 674–683, 1994.

[5] J. Mao and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection", IEEE Tr. NN, Vol. 6, No. 2, pp. 296–317, 1995.

[6] C. Chatterjee and V. P. Roychowdhury, "On self-organizing algorithms and networks for class-separability features", IEEE Tr. NN, Vol. 8, No. 3, pp. 663–678, 1997.

[7] M. B. Nevel'son and R. Z. Has'minskiĭ, *Stochastic Approximation and Sequential Estimation*, Nauka, 1968 (in Russian; translated into Japanese by T. Kitagawa and K. Tajima, 1983).

[8] T. Kurita and S. Hayamizu, "Gesture Recognition using HLAC Features of PARCOR Images and HMM based Recognizer", Proc. 3rd int. conference on automatic face and gesture recognition, April, 1998, pp. 422-427.

[9] W. Zhao, et al., "Discriminant Analysis of Principal Components for Face Recognition", Proc. 3rd int. conference on automatic face and gesture recognition, 1998, pp. 422-427. pp. 336-341.

[10] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, New York, 1989.

[11] K. Hiraoka and M. Hamahira, "On Successive Learning Type Algorithm for Linear Discriminant Analysis", IEICE Technical Report, NC99-73, pp. 85–92, 1999 (in Japanese).

[12] Kazuyuki Hiraoka, Shuji Yoshizawa, Ken-ichi Hidai, Masashi Hamahira, Hiroshi Mizoguchi, and Taketoshi Mishima, "Convergence Analysis of Online Linear Discriminant Analysis", International Joint Conference on Neural Networks (IJCNN), 2000, *to appear*.

Appendix: Proofs of theorems

Proof of theorem 1: Trivial. ■

Lemma 1 *Let $J_B = A^T BA$ and $J_W = A^T WA$. Then $J_B J_W = J_W J_B = J_B$ if A is a fixedpoint of (2).*

Proof: Multiplying A^T from left of (2), we obtain

$$\alpha J_B(I - J_W) + (1-\alpha)(I - J_W)J_B = O. \qquad (17)$$

Then, by adding (17) and the transpose of (17), we obtain

$$J_B(I - J_W) + (I - J_W)J_B = O. \qquad (18)$$

This means $J_B(I - J_W) = (I - J_W)J_B = O$, because both J_B and $I - J_W$ are symmetric. Thus $J_B J_W = J_W J_B = J_B$. ■

Corollary 1 *By rotating the coordinate system in $\mathbf{R}^L$, we can assume*

$$J_B = diag(\beta_1, \cdots, \beta_L), \qquad (19)$$

$$J_W = diag(\omega_1, \cdots, \omega_L), \qquad (20)$$

$$\beta_1 \geq \cdots \geq \beta_K > 0, \qquad (21)$$

$$\beta_{K+1} = \cdots = \beta_L = 0, \qquad (22)$$

$$\omega_1 = \cdots = \omega_K = 1, \qquad (23)$$

$$\omega_{K+1} \geq \cdots \geq \omega_L \geq 0, \qquad (24)$$

without loss of generality.

Lemma 2 *The following conditions are equivalent when $\alpha \neq 1$:*

1. *A is a fixedpoint of (2).*
2. *A satisfies*

$$BA = WAJ_B, \qquad (25)$$

$$J_B = J_W J_B = J_B J_W. \qquad (26)$$

Proof: ($\Leftarrow$) Trivial. ($\Rightarrow$) Let $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_L)$. When A is a fixedpoint, lemma 1 and its corollary hold. Then, from (2), $(BA - WAJ_B)(I - \alpha J_W) = O$. Thus, $B\boldsymbol{a}_k = \beta_k W \boldsymbol{a}_k$ if $\omega_k \neq 1/\alpha$. Moreover, $B\boldsymbol{a}_k = \beta_k W \boldsymbol{a}_k$ also holds if $\omega_k = 1/\alpha \neq 1$, because $\beta_k = 0$ in this case. Note that $B\boldsymbol{a}_k = 0$ and $\beta_k = \boldsymbol{a}_k^T B \boldsymbol{a}_k = 0$ are equivalent since B is symmetric.

■

Proof of theorem 2 and 3: Let $\breve{A}$ be a "spurious" solution, namely, $\breve{A}$ is a fixedpoint of (2) and $\breve{A}$ is not a solution of (1). Then lemma 1 and its corollary hold.

Now we can take a basis $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_N\}$ which satisfies $B\boldsymbol{e}_k = \beta_k W \boldsymbol{e}_k$, $\breve{A} = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_L)\mathrm{diag}(\sqrt{\omega_1}, \cdots, \sqrt{\omega_L})$, and $\boldsymbol{e}_k^T W \boldsymbol{e}_{k'} = 1 (k = k'), 0 (k \neq k')$. Suppose that $\omega_1 = \cdots = \omega_K = 1$ and $\omega_{K+1}, \cdots, \omega_L \neq 1$. For simplicity of description, we only show here a discussion for the case that $\beta_1, \cdots, \beta_N$ are different each other except that $\beta_{K+1} = \cdots \beta_L = 0$.

Let $\mathcal{A} = \{\breve{A}C \mid C \text{ is an arbitrary } L \times L \text{ matrix}\}$. Note that there exists a neighborhood U of $\breve{A}$ such that all the fixedpoints in U belong to $\mathcal{A}$. In order to observe the distance between $A(t)$ and $\mathcal{A}$, we define the metric $\langle F, G \rangle \equiv \mathrm{Tr} F^T G$.

Let $\Delta A(t) = A(t) - \breve{A}$ and $\Delta A(0) = (\boldsymbol{0}, \cdots, \boldsymbol{0}, \epsilon \boldsymbol{e}_H)$, where H satisfies $L + 1 \leq H \leq N$ and $\beta_H > 0$. The existence of such H is guaranteed from assumptions. Since $\Delta \breve{A}^T W A(0) = O$, $\Delta \dot{A} \equiv d\Delta A/dt|_{t=0} = \beta_H(1 - \alpha\omega_L) W \Delta A(0) + o(\epsilon)$. Hence $\langle \breve{A}, \Delta \dot{A} \rangle \approx 0$ for all $\breve{A} \in \mathcal{A}$, namely, $\Delta \dot{A}$ is orthogonal to $\mathcal{A}$.

Moreover, $d\langle \Delta A(t), \Delta A(t) \rangle / dt|_{t=0} \approx \beta_H (1 - \alpha\omega_L)\epsilon^2 > 0$ for sufficiently small ϵ if $\alpha = 0$ or $\omega_L < 1/\alpha$. Thus the theorems are proved. ■