Robust Online LDA by Adaptive Tuning of Learning Coefficient

Soichiro Morishita, Kazuyuki Hiraoka, Hiroshi Mizoguchi, and Taketoshi Mishima

Saitama University.
Dept. of Information and Computer Sciences, 255 Shimo-Okubo, Saitama, 338-8570, JAPAN
Tel: +81-48-858-3723, Fax: +81-48-858-3723
E-mail: mori@me.ics.saitama-u.ac.jp

Abstract: As a useful method of pattern recognition such as face identification, Linear Discriminant Analysis(LDA) is popular. A shortcoming of LDA is that it can not follow a changing situation, and the authors have proposed Online LDA(OLDA) to overcome it. However, OLDA is too sensitive about setting of the parameter called learning coefficient. Thus, towards robust OLDA, we propose an adaptive tuning of learning coefficient, and do some experiments applying the method with a task of person identification. As a result, we confirmed that presented method is more robust than conventional OLDA with regard to setting of parameters.

1. Introduction

As a useful method of pattern recognition such as face identification, Linear Discriminant Analysis(LDA) is popular. In short, LDA is the method which finds a proper matrix (linear transformation) for pattern recognition. This matrix is called the discriminant matrix.

Since LDA is batch algorithm, it can not follow a changing situation. To overcome this shortcoming of LDA, the authors have proposed Online LDA(OLDA) [1]. Actually, OLDA is applied to face identification, and person identification from binary silhouette image of full-length body [2].

It is pointed out that OLDA is too sensitive about setting of the parameter called learning coefficient. An inappropriate setting of learning coefficient causes divergence of the discriminant matrix. Hence the purpose of this paper is to prevent OLDA from divergence. We propose an adaptive tuning of learning coefficient.

2. OLDA algorithm

2.1 Basic Form

In learning phase, a new pair $(\boldsymbol{x}(t), c(t))$ is presented, at every time step $t = 1, 2, 3, \cdots$; 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 are updated as follows:

$$
\begin{aligned}
t^c(t) &= t^c(t-1) + \delta(c, c(t)), \\
\bar{\boldsymbol{x}}(t) &= \left(1 - \frac{1}{t}\right) \bar{\boldsymbol{x}}(t-1) + \frac{1}{t}\boldsymbol{x}(t), \\
\bar{\boldsymbol{x}}^c(t) &= \\
&\begin{cases} \left(1 - \frac{1}{t^c(t)}\right) \bar{\boldsymbol{x}}^c(t-1) + \frac{1}{t^c(t)}\boldsymbol{x}(t) & (c = c(t)), \\ \bar{\boldsymbol{x}}^c(t-1) & (c \neq c(t)), \end{cases} \\
\boldsymbol{v}^c(t) &= \bar{\boldsymbol{x}}^c(t) - \bar{\boldsymbol{x}}(t), \\
\boldsymbol{w}(t) &= \boldsymbol{x}(t) - \bar{\boldsymbol{x}}^{c(t)}(t), \\
B(t) &= \frac{1}{M}\sum_{c=1}^{M} \boldsymbol{v}^c(t)\boldsymbol{v}^c(t)^T,
\end{aligned}
$$

where $c = 1, \cdots, M$ and $\delta(c, c(t)) = 1\,(c = c(t)),\ 0\,(c \neq c(t))$. Then $N \times L$ discriminant matrix A is updated:

$$
\begin{aligned}
A(t) = A(t-1) + \eta\Big[& B(t)A(t-1) \\
& - \alpha B(t)\, A(t-1)A(t-1)^T \boldsymbol{w}(t)\boldsymbol{w}(t)^T A(t-1) \\
& - (1-\alpha)\, \boldsymbol{w}(t)\boldsymbol{w}(t)^T A(t-1)A(t-1)^T B(t)A(t-1)\Big],
\end{aligned} \tag{1}
$$

where the learning coefficient η is a small positive number, and $\alpha \in \boldsymbol{R}$ is a parameter.

In identification phase, we identify the class of new vector $\boldsymbol{x}(t)$ with a standard procedure using this discrimination matrix A. See [1] for details.

2.2 Variations

In the past, several variations of OLDA have been proposed [3].

Symmetric algorithm This is the case when $\alpha = 0.5$ in (1).

Asymmetric algorithm This is the case when $\alpha \neq 0.5$ in (1). In particular, the case when $\alpha = 0$ is a direct extension of Oja's dynamics for online PCA [4]. We let $\alpha = 0$ as Asymmetric algorithm in this article.

Fast algorithm The updating rule of this algorithm is slightly different from (1) as follows:

$$A(t) = A(t-1) + \eta\Big[B(t)A(t-1) - \alpha A(t-1)\,A(t-1)^T B(t)w(t)w(t)^T A(t-1) - (1-\alpha)A(t-1)\,A(t-1)^T w(t)w(t)^T B(t)A(t-1)\Big].$$

We can obtain fast convergence by this algorithm[5].

3. Towards robust OLDA

When we use conventional OLDA, we must preset learning coefficient η, and fix it throughout learning. However, there is no reason to set η so. Updating η in accordance with a 'scale' of A, and leading η to a pertinence value, we can obtain better performance. For this purpose, first, we focus on the fact that the following equation holds when discrimination matrix A has converged to a desirable value [6, 7]:

$$E[A^T w(t)w(t)^T A] = I. \tag{2}$$

In particular,

$$\frac{1}{L}E[w^T(t)AA^T w(t)] = 1 \tag{3}$$

is implied from (2). Secondly, as a measure of 'scale' of A, we introduce a variable $\rho(t)$ which is an approximation of left side of (3). Expectation $E[\,\cdot\,]$ in left side of (3) is replaced with weighted time-average in $\rho(t)$. This $\rho(t)$ is updated by

$$\rho(t+1) = \rho(t) + \eta_\rho\left(\frac{1}{L}w(t)^T A(t)A(t)^T w(t) - \rho(t)\right),$$

where η_ρ is a small positive number. The initial value of ρ is set $\rho(0) = 1$. Finally, at every time t, we set learning coefficient η as

$$\eta(t) = \frac{\eta_0}{\rho(t)}, \tag{4}$$

where η_0 is a small positive number. We will refer to the parameter η_0 as base learning coefficient. With normalization (4) based on $\rho(t)$ (a measure of 'scale' of A), we can let η proper at all times.

Note that conventional OLDA is a special case of the presented algorithm. when $\eta_\rho = 0$, ρ is not updated and $\eta(t) = \eta_0$. In the above sense, η_0 corresponds to conventional parameter η.

4. Experimental results

In order to verify the usability of the presented method, we did some experiments applying the method with a task of person identification. In these experiments, we evaluated the ratio of correct identification. We call its average of 100 trials *discrimination ratio.*

There are 4 persons, A, B, C, and D, and we prepared 100 images for each person. Thus, the total number of sample images is 400. Each image is reduced to 10×10 pixels. The same images are used for both learning and evaluation of identification performance.

Let t be the step of learning. One image is presented at every step.

A new person is added at every 100 steps. During $t = 1 \cdots 100$, only person A's images are presented, During $t = 101 \cdots 200$, only person B's images are presented, and identification of A and B is required. These procedures are repeated until $t = 400$ (Table 4.).

We applied presented method with three variation of OLDA: Symmetric algorithm, Asymmetric algorithm, and Fast algorithm.

Table 1. Procedure of experiments. Only the images of new person are presented for learning, while identification of all the past persons are required in evaluation.

t	learning	identification
$1 \cdots 100$	A	A
$101 \cdots 200$	B	A,B
$201 \cdots 300$	C	A,B,C
$301 \cdots 400$	D	A,B,C,D

Table 2 shows setting of the experiments.

The result of the experiments is shown in Fig 1, 2, and 3.

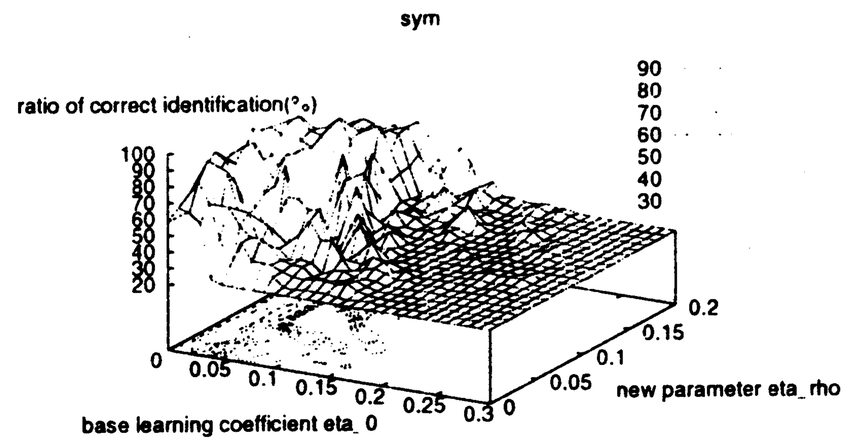

Figure 1. *Discrimination ratio for each base learning coefficient η_0 and new parameter η_ρ on Symmetric Algorithm. The case when $\eta_\rho = 0$, it is equivalent to conventional OLDA. The case when $\eta_\rho > 0$, it corresponds to presented method.*

Table 2. *Setting of the experiments*

task	person identification from face images
algorithm	Symmetric / Asymmetric / Fast
data vector $\boldsymbol{x}(t)$	face images under various illumination conditions (front view, 256 level gray scale, normalized to $[-1,+1]$)
size of $\boldsymbol{x}(t)$ number of classes to be identified number of features	$N = 10 \times 10 = 100$ (pixels) $M = 1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ (persons) $L = M - 1$ (= number of columns in A)
initial values of elements in A base learning coefficient new parameter	random values from the uniform distribution on $[-0.001,+0.001]$ $\eta_0 = 0.00 \cdots 0.30$ $\eta_\rho = 0.00 \cdots 0.20$
procedure of learning procedure of evaluation number of face images	Face images for learning is presented in a random order The ratio of the correct identification is evaluated for face images which are different from the face images for learning. 100(images per person) $\times$ 4(persons) = 400
number of trials	100 independent trials with different random seeds

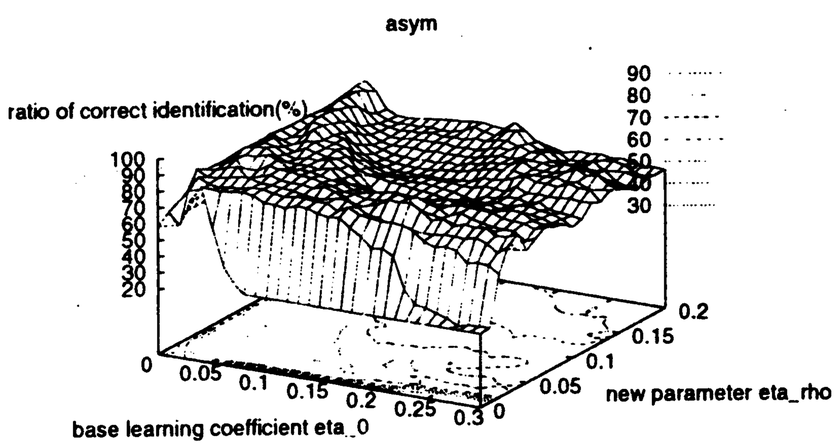

Figure 2. *Discrimination ratio for each η_0 and η_ρ on Asymmetric Algorithm.*

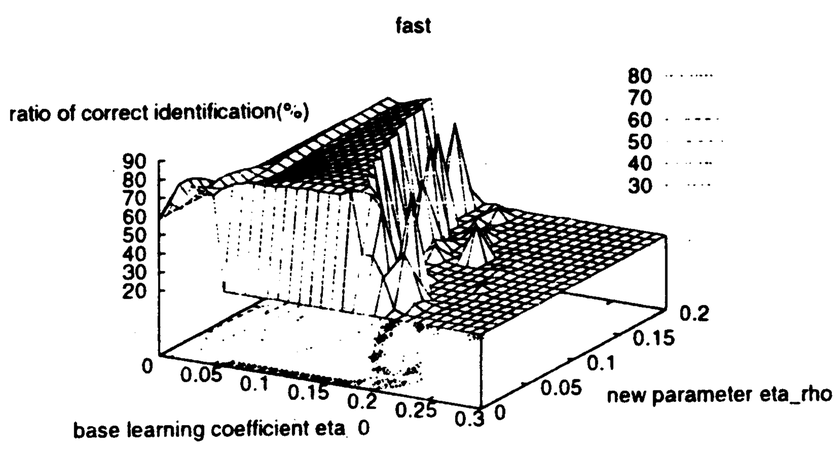

Figure 3. *Discrimination ratio for each η_0 and η_ρ on Fast Algorithm.*

We discuss the results of these experiments focusing on three points.

4.1 dependence on conventional parameter η_0

On Asymmetric and Fast algorithms, the range of allowable value of η_0 is expanded. Namely, the presented method gives more robustness to conventional OLDA. It is effective especially on Fast algorithm. As for Symmetric algorithm, the improvement is small.

4.2 dependence on new parameter η_ρ

Though discrimination ratio tends to decrease for too large η_ρ, it's dependency is not strong compared with that of η_0 in conventional methods. This means that 'rough' turning is sufficient for the presented method, while 'strict' tuning is necessary for conventional method.

4.3 decrease of peak discrimination ratio

Expect for narrow ($0.00 < \eta_0 \leq 0.03$ for Symmetric Algorithm, $0.00 < \eta_0 \leq 0.04$ for Asymmetric Algorithm, and $0.01 < \eta_0 \leq 0.05$ for Asymmetric Algorithm), the presented method show a performance advantage obliviously. Within these narrow ranges, decrease of peak performance is observed in some cases. On Symmetric algorithm, discrimination ratio on presented method is higher than that on conventional OLDA. On Asymmetric algorithm, the difference of discrimination ratio is little. On Fast algorithm, the discrimination ratio on presented method tends to decrease. In the worst case, the discrimination ratio is decreased from 80%(conventional) to 69%(presented).

Fig. 4 shows learning curve of each value of η_ρ on Fast algorithm. As the value which conventional OLDA works properly, we observe the case $\eta_0 = 0.02$. When value of η_ρ is large, decrease of discrimination ratio tends to be large [1]. This decreasing of discrimination ratio is

[1] There is no remarkable difference of discrimination ratio with conventional OLDA and presented method on Symmetric and Asymmetric algorithm.

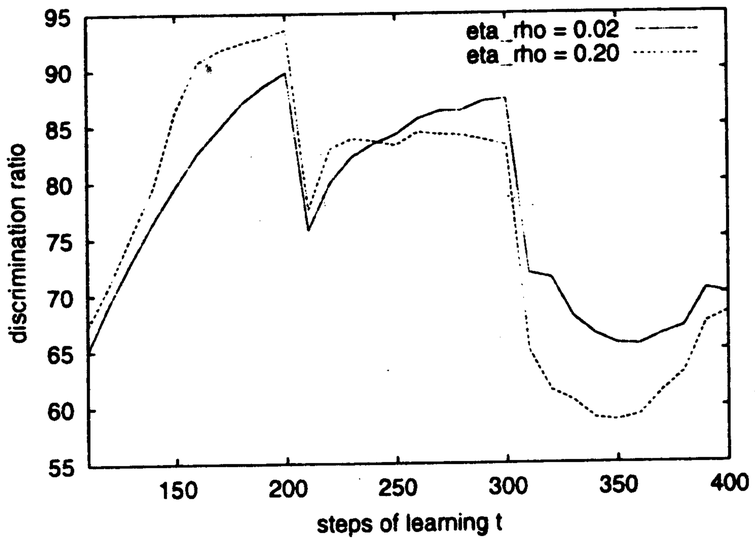

Figure 4. *Learning curves for* $\eta_\rho = 0.2$ *and* $\eta_\rho = 0.02$. *Learning coefficient* η_0 *is fixed to* 0.02.

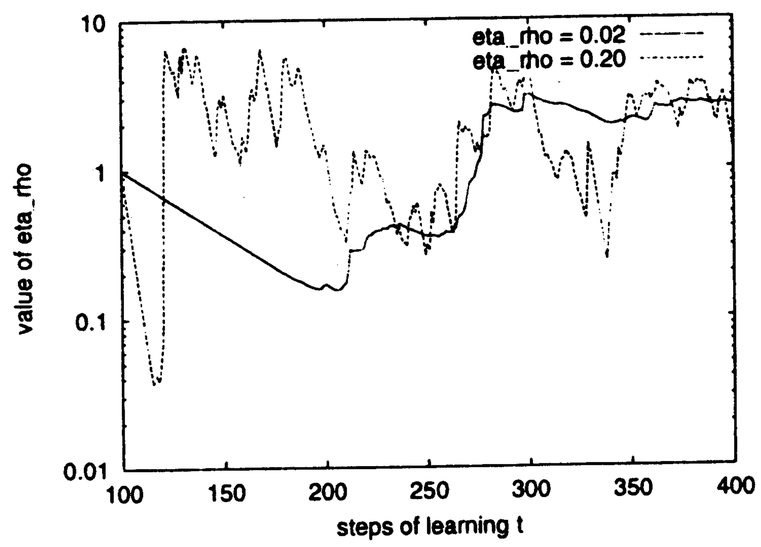

Figure 5. *Value of* ρ *for* $\eta_\rho = 0.2$ *and* $\eta_\rho = 0.02$ *in one trial. Learning coefficient* η_0 *is fixed to* 0.02.

likely caused by fluctuation of $\rho(t)$. Fig. 5 shows an example.

Although it is an undesirable side effect, It is small and acceptable compared with the merit(robustness) which covers it.

Table 3 shows summary of results mentioned above.

Table 3. *Summary of results*

Algorithm	Dependence on η_0	Dependence on η_ρ	Decreace of peak
Sym.	*Not so good*	*None*	*Up*
Asym.	*Good*	*A little*	*Even*
Fast.	*Good*	*Some*	*Down*

5. Conclusion

We pointed out weakness of conventional OLDA that it is too sensitive about setting of learning coefficient. Hence, we proposed the method which gives more robustness to OLDA by the adaptive tuning of learning coefficient. Additionally, we did some experiments applying the method with the task of face identification for three variations of OLDA. As a result, we confirmed that presented method prevents divergence of the discriminant matrix. Though decrease of discrimination ratio was observed as an undesirable side effect for some cases, it was small and acceptable compared with the merit that given by the presented method.

References

[1] K. Hiraoka, K. Hidai, M. Hamahira, H. Mizoguchi, T. Shigehara, and T. Mishima, "Derivation of Online Linear Discriminant Analysis and Its Application to Face Identification," Proc. 5th Robotics Symposia, Kobe, Japan, pp. 226–231, Mar. 2000.

[2] K. Hiraoka, S. Morishita, H. Mizoguchi, and T. Mishima, "Person Identification from Binary Silhouette Image of Full-length Body," Proc. ISIM2000, KyongSangNam-Do, Korea, pp. 23–26, Oct. 2000.

[3] K. Hiraoka, K. Hidai, M. Hamahira, H. Mizoguchi, T. Mishima, and S. Yoshizawa, "Comparison of Various Algorithms for Online LDA," Proc. 2000 JSME Conference on Robotics and Mechatronics, Kumamoto, Japan, 1A1-77-110(1)-(2) (CDROM), May 2000.

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

[5] K. Hiraoka, M. Hamahira, K. Hidai, H. Mizoguchi, T. Mishima, and S. Yoshizawa, "Fast algorithm for online linear discriminant analysis," Proc. The 2000 ITC-CSCC, Pusan, Korea, pp. 274–277, Jul. 2000.

[6] K. Hiraoka and M. Hamahira, "On Successive Learning Type Algorithm for Linear Discriminant Analysis: Proposal of Learning Algorithm and Proof of Local Convergence," IEICE Technical Report, Nagoya, Japan, NC99-73, pp. 85–92, Dec. 1999.

[7] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima, "Convergence Analysis of Online Linear Discriminant Analysis," Proc. IEEE-INNS-ENNS IJCNN, Como, Italy, III-387-391(CD-ROM), Jul. 2000.

[8] S. Morishita, K. Hiraoka, H. Mizoguchi, and T. Mishima, "Study on Automatic Setting Method of Learning Coefficient in Online LDA towards Robust Convergence," Proc. The 2000 Information and Systems Society Conferences of IEICE, Nagoya, Japan, p. 217, Oct. 2000.