ON THE CLASSIFICATION OF PATTERNS BY THE KARHUNEN-LOEVE ORTHOGONAL SYSTEM WITHOUT SUPERVISOR

by

Shingo Tomita and Shoichi Noguchi
(Research Institute of Electrical Communication, Tohoku University, Sendai, Japan)

Summary

In this paper we show how the pattern samples generated from the unknown asymmetric finite mixture distribution are dichotomized by the Karhunen-Loeve orthogonal system in the optimal way.

We introduce the concept of the difference of features of two categories and prove that decreasing the error probability of dichotomy is equal to increasing the difference of features of two categories. And it can be shown that the maximum difference of features obtained by the KL system is equivalent to that which is obtained by the Bayes solution.

Moreover it can be proved that there exists a dichotomy which converges to the Bayes soJution by increasing the number of samples.

1. Introduction

Let us explain the outline of the Karhunen-Loeve orthogonal system. Let a pattern on the N-dimensional space RN be X and a normal and orthogonal system on R_N be a. Then a pattern X is expanded by a as follows:

$$X = \xi_1^X(a)\ a_1 + \xi_2^X(a)\ a_2 + \ldots + \xi_N^X(a)\ a_N,$$

where

$$\xi_\nu^X(a) = a_\nu^T X,$$

$$a_i^T a_j = \delta_{ij} \text{ and } a = (a_1, a_2, \ldots, a_N)\ .$$

Let the mean square of coefficients of X be

$$\xi_\nu(a) = E\ [\xi_\nu^X(a)^2\ ]$$

$(\nu = 1,2,..,N)$, where $E[*] = \int * dG(X)$ is the distribution of patterns. Suppose that all $\xi_\nu(a)$'s are ordered as follows:

$$\xi_1\ (a) \geqq \xi_2\ (a) \geqq .. \geqq \xi_N(a).$$

The sum, $\xi(a,\ m) = \sum_{\nu=1}^m \xi_\nu(a)$, is defined as a quantity of features extracted by $a_1,\ a_2,..,a_m$. Let the autocorrelation of a set of patterns X be $E[XX^T]$, then an eigenvalue λ_ν and the eigenvector t_ν corresponding to λ_ν is obtained by the equation $E[XX^T]\ t_\nu = \lambda_\nu t_\nu$. Let the normal and orthogonal system be $t = (t_1,\ t_2,\ ..,\ t_N)$, then $\lambda_\nu = \xi_\nu(t)$ and $\xi(t,m) = \text{Max}\{\ \xi\ (a,m)\ |\ \forall a\ \}$ are proved respectively.

The above normal and orthogonal system is called the Karhunen-Loève orthogonal system or simply the KL system.(1)

2. Classification without supervisor

Let an asymmetric finite mixture distribution of patterns be $G(X) = PF\ (X\,|\,w_1) + qF\ (X\,|\,w_2)$, where $q+p = 1, p,q > 0$, and w_1 and w_2 are the given two categories. Let the two vetcors and two matrices be defined as follows:

$$M = \frac{1}{2}\{E\ [X\,|\,w_1] + E[X\,|\,w_2]\}\ ,$$

$$\alpha = \frac{1}{2}\{E\ [X\,|\,w_1] - E\ [X\,|\,w_2]\}\ ,$$

$$\Lambda = \frac{1}{2}\{E\ [(X-E\ [X\,|\,w_1\ ]\ )(X-E[X|w_1])^T|w_1] + E\ [(X-E[X|w_2])(X-E[X|w_2])^T|w_2]\}$$

$$\Sigma = \frac{1}{2}\{\ E[(X-E[X\,|\,w_1])\ (X-E[X|w_1])^T\,|\,w_1] - E[(X-E[X\,|\,w_2])\ (X-E[X\,|\,w_2])^T\,|\,w_2]\}$$

where $E[*\,|\,w_i] = \int * \ dF(X\,|\,w_i)$. Then,

$$E[XX^T|\ w_1] = E + D,\ E[XX^T\,|\,w_2] = E-D,$$

where $E = \Sigma + MM^T + \alpha\alpha^T,\ D = \Lambda + \alpha M^T + M\alpha^T.$

Suppose that Rank $E[XX^T] = N$, then there exists a matrix S such that $SE[XX^T]\,S^T = I$, where I is an unit matrix. Let all the patterns generated from the distrubution G(X) transform by S, and denote them by the same symbol $\{X\}$, then $E[XX^T \mid w_1] = I + 2qD$, $E[XX^T \mid w_2] = I-2pD$.
The two quantities of features corresponding to w_1 and w_2 extracted by the KL system are proved to be the eigenvalues of $E[XX^T \mid w_1]$ and $E[XX^T \mid w_2]$, respectively, so $\lambda_\nu^{(1)}$ and $\lambda_\nu^{(2)}$ are two quantities of features corresponding to w_1 and w_2. $\lambda_\nu^{(i)}$ is obtained by the following equation:
$E[XX^T \mid w_i]\, t_\nu^{(i)} = \lambda_\nu^{(i)}\, t_\nu^{(i)}$ $(i=1,2;\ \nu=1, 2, .., N)$.
Let $Dt_\nu = \lambda_\nu t_\nu$ $(\nu=1, 2, ..., N)$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$.
then $\lambda_\nu = \dfrac{\lambda_\nu^{(1)}-1}{2q} = \dfrac{1-\lambda_\nu^{(2)}}{2p}$ by the following relations :
$t_\nu = t_\nu^{(1)} = t_\nu^{(2)}$ $(\nu = 1, 2, .., N)$.
In order to classify a set of given patterns into two true categories, we introduce a parameter to evaluate the extent of the correct classification and adopt the parameter ρ such that $\rho = \sqrt{\sum_{\nu=1}^{N} (\lambda_\nu^{(1)} - \lambda_\nu^{(2)})^2}$. ρ is defined as the difference of features. Although the mixture distrubution G(X) is given, ρ is not determined uniquely in general. But if G(X) is identifiable ρ is determined uniquely. So we obtain the following theorm.

Theorem 1

If the asymmetric finite mixture distribution G(X) is identifiable, then the difference of features ρ is determined uniquely by G(X).
Let w_1^* and w_2^* be two sets of patterns which are decided by some dichotomy, then the conditional probability $P(w_j \mid w_i^*)$ satisfies the following formula:
$P(w_1 \mid w_1^*) + P(w_2 \mid w_1^*) = P(w_1 \mid w_2^*) + P(w_2 \mid w_2^*) = 1$.
Moreover let p^* and q^* be defined as follows:

$p^* = P(w_1 \mid w_1^*) + P(w_2 \mid w_2^*)$,

$q^* = P(w_2 \mid w_1^*) + P(w_1 \mid w_2^*)$. If $p^* > 1$ then q^* is defined as an evaluated error value, if $p^* < 1$, then p^* is defined as an evaluated error value.
For a set of pattern w_i^*, $P(X \mid w_i^*) = P(X \mid w_1)\,P(w_1 \mid w_i^*) + P(X \mid w_2)\,P(w_2 \mid w_i^*)$, then we have $F(X \mid w_i^*) = F(X \mid w_1)\,p(w_1 \mid w_i^*) + F(X \mid w_2)\,p(w_2 \mid w_i^*)$.
Consequently, $E[XX^T \mid w_i^*] = E[XX^T \mid w_1]\,p(w_1 \mid w_i^*) + E[XX^T \mid w_2]\,p(w_2 \mid w_i^*)$ ....(1)
For two sets of patterns w_1^* and w_2^*, the difference of features is defined as follows :

$$\rho^* = \sqrt{\sum_{\nu=1}^{N} (\lambda_\nu^{*(1)} - \lambda_\nu^{*(2)})^2},$$

where $\lambda_\nu^{*(i)}$ is the eigenvalue of the autocorrelation $E[XX^T \mid w_i^*]$ and $\lambda_\nu^{(1)}$ is ordered in the following way :

$$\lambda_1^{*(1)} \geq \lambda_2^{*(1)} \geq \ldots \geq \lambda_N^{*(1)} \text{ and } \lambda_\nu^{*(2)} \text{ is}$$

vice versa.
From the formula (1), we obtain the important relation between ρ and ρ^* such that $\rho^* = |1-q^*|\,\rho$.
If two dichotomies are executed, then two difference of features ρ^* and ρ^{**} are decided respectively. And let q^*, q^{**} be evaluated error probabilities, then $\rho^{**} - \rho^* = (q^* - q^{**})\rho$.

Theorem 2

Let the following condition be A. Condition A: G(X) is identifiable and ρ is not zero.
Then $\rho^{**} > \rho^*$ if and only if $q^{**} < q^*$ under the condition A.
If the dichotomy to get the maximum difference of features is equal to decreasing the evaluated error value q^* into zero, then $P(w_i \mid w_j^*) = \delta_{ij}$ or $1 - \delta_{ij}$.

Theorem 3

Let the condition A be satisfied. Then

$$P(w_i \mid w_j^*) = \delta_{ij}$$

if and only if $P(X \mid w_i) = P(X \mid w_j^*)$.
From the result of theorem 3, it is possible to estimate the true distributions corresponding to each given category, if it is possible to decrease the evaluated error value q^* into zero.
Let the ideal decision function be $d(X) = P(w_1 \mid X) - P(w_2 \mid X)$ and a decision function be $d^*(X) = P(w_1^* \mid X) - P(w_2^* \mid X)$ when w_1^* and w_2^* are given.
Suppose that the specified two sets of patterns are $\hat{w}_1$ and $\hat{w}_2$ which are obtained by some dichotomy, and let the difference of features and the decision function corresponding to $\hat{w}_1$ and $\hat{w}_2$ be $\hat{\varrho}$ and $\hat{d}(X)$, respectively.
If $\hat{d}(X) = \pm d(X)$, then $P(\hat{w}_1) \neq 0$, $P(\hat{w}_2) \neq 0$.
Then the following theorem is obtained.

Theorem 4

Suppose that the condition A is satisfied. Then $\hat{\varrho} = \varrho$ if and only if $\hat{d}(X) = \pm d(X)$.
Let the true distance of means between the two categories w_1 and w_2 be

$$r = \| E[X \mid w_1] - E[X \mid w_2] \| ,$$

and a distance of means between w_1^* and w_2^* be

$$r^* = \| E[X \mid w_1^*] - E[X \mid w_2^*] \| ,$$

where $\| \cdot \|$ is a norm of a vector.
Then $r^* = | 1 - q^* | \; r$.

Theorem 5

Suppose that the condition A is satisfied and $E[X \mid w_1] \neq E[X \mid w_2]$.
Then $\varrho^{**} > \varrho^*$ if and only if $r^{**} > r^*$.
Let the symmetric finite mixture distribution $G(X)$ be defined as the special form such that $G(X) = pF(X-\alpha) + qF(X+\alpha)$, $p+q = 1$, $p, q > 0$.
The two mean vectors corresponding to w_1 and w_2 are defined respectively as follows :
$E[X \mid w_1] = \int X dF(X-\alpha)$, $E[X \mid w_2] = \int X dF(X+\alpha)$,
where α is an unknown parameter.
Let the mean vector of w_i^* be $E[X \mid w_i^*]$, then
$E[X \mid w_i^*] = E[X \mid w_1] p(w_1 \mid w_i^*) + E[X \mid w_2] p(w_2 \mid w_i^*)$.
Suppose that the specified vector α^* is defined as follows :

$$\alpha^* = \frac{1}{2} \left(E[X \mid w_1^*] - E[X \mid w_2^*] \right).$$

Then from the theorem 5, the following corollary is obtained.[3]

Corollary 51 : $\varrho^* = \varrho$ if and only if $\alpha^* = \pm\alpha$.
Suppose that the distribution $G(X)$ satisfies the following condition B.
Condition B : For any small $\varepsilon > 0$, there exists a large positive number K such that

$$\left| \int_{-\infty}^{-K} X^T X \, dG(X) \right| < \varepsilon \quad \text{and} \quad \left| \int_{K}^{\infty} X^T X \, dG(X) \right| < \varepsilon ,$$

where $X^T = (x_1, x_2, \ldots, x_N)$.
Let a set of finite patterns be $\{X^{(\nu)}\}^{2n+1}$ or $\{X^{(-n)}, X^{(-n+1)}, \ldots, X^{(n)}\}$ and $\{X^{(\nu)}\}^{2n+1}$ satisfy the following condition.
Condition C : For any small $\varepsilon > 0$, there exists a large integer n such that

$$P\left\{ \left| x_i^{(\nu)} x_j^{(\nu)} - x_i^{(\nu-1)} x_j^{(\nu-1)} \right| < \frac{A}{2n+1} \right\} \geq 1 - \varepsilon$$

for some positive number A, where $X^{(-n)T} = (-K, -K, \ldots, -K)$, $X^{(n)T} = (K, K, \ldots, K)$.
The autocorrelation of $\{X^{(\nu)}\}^{2n+1}$ is defined as follows :

$$E[XX^T] = \frac{1}{2n+1} \sum_{\nu=-n}^{n} X^{(\nu)} X^{(\nu)T}.$$

$E[XX^T]$ is a symmetric matrix, so there exists a normal matrix $S^{(n)}$ such that $S^{(n)} E[XX^T]^{(n)} S^{(n)T} = I$.
Let all patterns $\{X^{(\nu)}\}^{2n+1}$ transform by $S^{(n)}$, and denote them by the same symbol $\{X^{(\nu)}\}^{2n+1}$, and two sets of patterns which are obtained by a dichotomy from $\{X^{(\nu)}\}^{2n+1}$ be $w_1^{*(n)}$ and $w_2^{*(n)}$ respectively.
The autocorrelation of a set of patterns w_i^* is defined as follows :

$$E[XX^T \mid w_i^{*(n)}] = \frac{1}{n_i} \sum_{\nu} X^{(\nu)} X^{(\nu)T},$$

where n_i^* is a number of $w_i^{*(n)}$ and $n_1^* + n_2^* = 2n + 1$.

Let the difference of features $w_1^{*(n)}$, $w_2^{*(n)}$ be $\varrho^{*(n)}$ and $w_1^{(n)}$, $w_2^{(n)}$ be $\varrho^{(n)}$, respectively, where $w_i^{(n)}$ satisfies the following equation :
$P(w_1^{(n)} \mid w_i^{*(n)}) + P(w_2^{(n)} \mid w_i^{*(n)}) = 1$.
Then the following theorm is obtained.

Theorem 6

Suppose that all the conditions A, B and C are satisfied.

Then

(i) $P\{\lim_{n\to\infty} \rho^{*(n)} = \rho\} = 1$ if and only if

$P\{\lim_{n\to\infty} d_n^*(X) = \pm d(X)\} = 1,$

where $d_n^*(X) = P(w_1^{*(n)} \mid X) - P(w_2^{*(n)} \mid X)$.

(ii) Let the maximum difference of features be defined as follows :

$$\rho_{max}^{(n)} = \mathrm{Max}\{\rho^{*(n)} \mid \forall w_1^{*(n)}, \forall w_2^{*(n)}\}.$$

Then $P\{\lim_{n\to\infty} \rho_{max}^{(n)} = \rho\} = 1$.

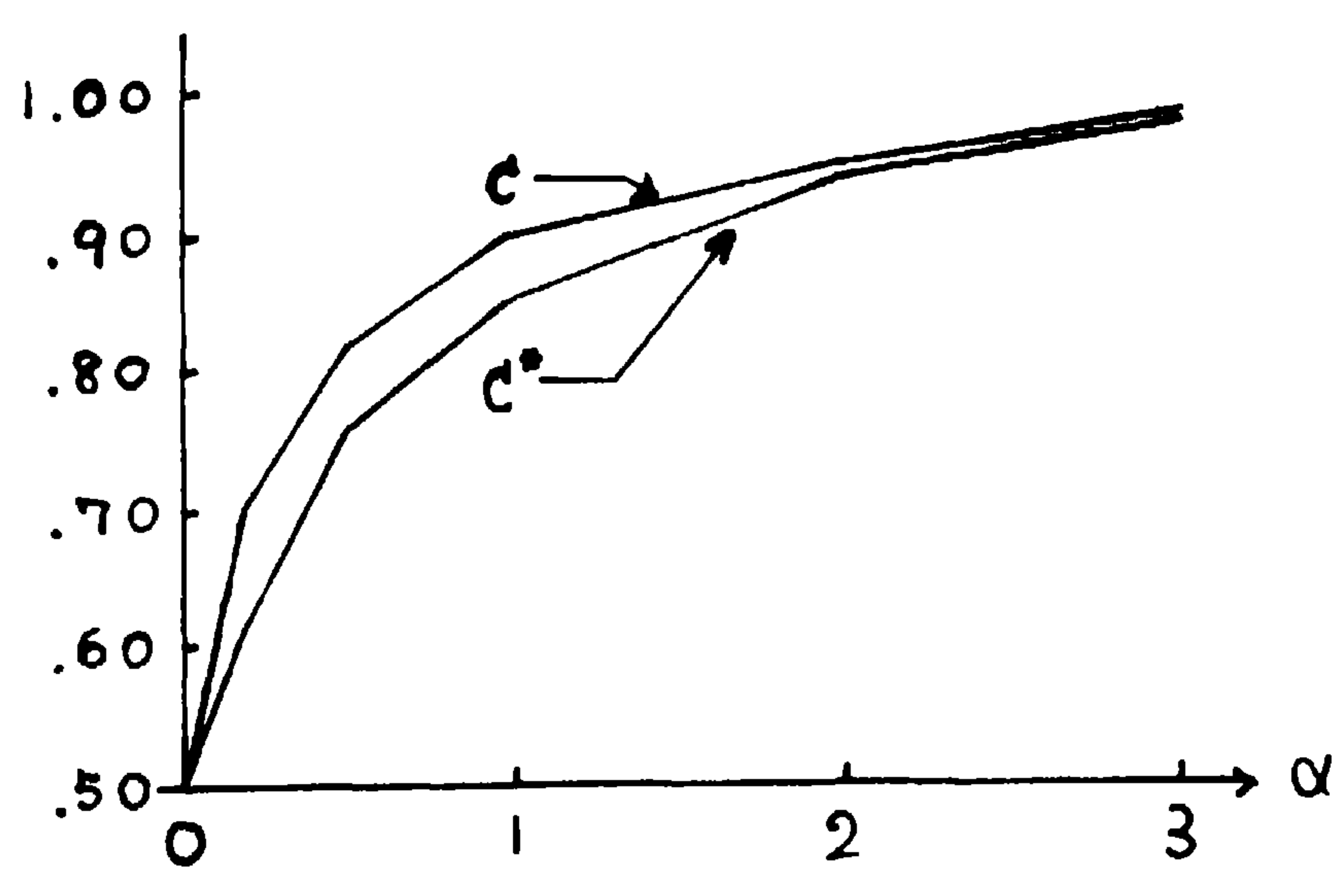
As the summary, we obtain the final conclusions as follows:

(A) If the asymmetric finite distribution of patterns is identifiable, increasing the difference of features is equivalent to decreasing the error probability of classification, and obtaining the maximum differnce of features is equivalent to estimating the true distrubutions of given categories.

(B) To obtain the maximum difference by some dichotomy is equivalent to getting the Bayes solution and to getting the maximum distance of means between two categories, if the two means of categories are not equal.

(C) There exists a dichotomy with probability one to get the maximum difference of features by incresing the number of samples and this dichotomy converges to the Bayes solution with probability one.

Fig. 1



α	3.0	2.0	1.0	0.75	0.5	0.25
Div	72.0	32.0	8.0	4.5	2.0	0.5
ε^*	0.012	0.057	0.103	0.350	0.599	0.863
ρ_{max}	1.978	1.910	1.594	1.391	1.104	0.735

Table 1

3. Results on a computer

Let a 2-dimensional normal distribution with the mean vector $M_\circ$ and the covariance matrix $\Sigma_\circ$ be $N(X:M_\circ, \Sigma_\circ)$ and the mixture of two normal distributions be $G(X) = pN(X:M_\circ+\alpha_\circ, I) + qN(X:M_\circ-\alpha_\circ, I)$, where $M_\circ^T = (1,1)$, $\alpha_\circ^T = (\alpha, 0)$, $\alpha > 0$ and I is an unit matrix of 2×2.

To compare our method with a classification with teacher, the decision function d(X) by the Bayes law is adopted as follows :

$d(X) = \log P(W_1 \mid X) - \log P(w_2 \mid X)$, where $P(X \mid w_i)$

$$= \frac{1}{2\pi} \left|\Sigma_i\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-E_i)^T \Sigma_i^{-1}(X-E_i)\right\},$$

$$E_1 = M_\circ + \alpha_\circ,\ E_2 = M_\circ - \alpha_\circ,\ \Sigma_1 = \Sigma_2 = I.$$

In this case, the decision function d(X) is reduced as follows :

$$d(X) = \log\frac{p}{q} + (X-M_\circ)^T\alpha_\circ + \alpha_\circ(X-M_\circ)^T \qquad (2)$$

The dichotomy is as follows : If $d(X) > 0$, then $X \in w_1$; if $d(X) < 0$, then $X \in w_2$. To simulate the mixture of normal distributions, random numbers are generated from the normal distribution with mean value zero and the variance value one. The uniform distribution with the interval [0,1] is used, and the classification based on theorem 6 and the formula (2) are executed on a digital computer.

The results on a computer are shown Fig. 1 and Table 1 in the case of $p = q = \frac{1}{2}$ and $M = 800$, where M is the number of samples. In Fig. 1 and Table 1, four symbols are as follows :

C : Classification with teacher based on d(X).

C^*: Classification without teacher based on the theorem 6.

Div: Divergence reported by Kullback(4).

ε^* : Absolute error rate.

$$\varepsilon^* = \mathrm{Min}\{\|\alpha^* - \alpha_\circ\| / \|\alpha_\circ\|,\ \|\alpha^* + \alpha_\circ\| / \|\alpha_\circ\|\}$$

4. Conclusion

Utlizing the concept of the Karhunen-Loeve orthogonal system which extracts the maximum difference of features of patterns, the algorithm of classification for patterns generated from an unknown asymmetric finite mixture distribution is obtained, and this one is proved to become the Bayes solution.

As a concreTfe example, random numbers generated from the 2-dimensional mixture of normal distributions are used and good results are obtained on a digital computer.

Reference

(1) S. Watanabe : Knowing and Guessing, Wily, 1969.

(2) K. Fukunage, etc. : Application of the Karhunen-Loeve expansion to feature selection and ordering, IEEE Trans. C-19 (1970).

(3) D.B. Cooper, etc. : On suitable condition for statistical pattern recognition without supervision, SIAM. Appl. Math. 17 (1969).

(4) S. Kullback : Information theory and statistics, Dover, 1968.