Mem. Fac. Educ., Shimane Univ. (Nat. Sci.), Vol. 6, pp. 9–21, December 1972.

SEQUENTIAL PATTERN RECOGNITION SYSTEMS I.

ON THE DIVERGENCE AND CHERNOFF'S DISTANCE FOR THE FEATURE SELECTION AND ORDERING

Minoru YABUUCHI*

Abstract: In sequential pattern recognition systems, the selection and ordering of effective features from a given set of feature measurements is an important problem. The purpose of this paper is to discuss the efficiency of the divergence and Chernoff's distance (particularly, Bhattacharyya's distance) as a criterion of feature selection and ordering. After these probabilistic measures were reviewed, a very simple multiclass pattern recognition was simulated as an example of the application of these. In this situation, to maximize the expected divergence and the expected Bhattacharyya's distance was adopted as a criterion of feature selection and ordering. The relations of these measures to the number of features and to the mean probability of misrecognition were obtained.

1. INTRODUCTION

The class of pattern recognition systems to be considered in this paper, the so-called *statistical pattern recognition systems*, is characterized as follows: 1) information about the patterns is stored in the form of feature measurements and 2) the system's decision is based on the methods of statistical decision theory.[e. g., 1)] In the statistical recognition systems which a number of studies had been performed, all the feature measurements were processed by the systems at one stage. This procedure is also called as fixed-sample size decision procedure.

In this procedure, the cost of feature measurements has not been taken into consideration[2)]: 1) if the number of feature measurements is insufficient, it will not be able to give satisfactory results in correct classification and 2) on the other hand, an arbitorarily large number of features to be measured is impractical.

*Depertment of Psychology, Faculty of Education, Shimane University, Matue City, Japan.

Hence, recent studies in the design of pattern recognition systems have applied sequential decision procedures to this class of recognition problems to minimize misrecognition and average number of feature measurements.

In *sequential pattern recognition systems*, the selection and ordering of effective features from a given set of feature measurements is an important problem. Unfortunately, it is often difficult to obtain an analytical expression for the probability of misrecognition and even if one can be obtained it will usually be complicated to permit numerical computation.[3]

Hence, certain probabilistic distance measures, such as *divergence* and *Chernoff's distance*, which are easy to evaluate, are used for the selection of effective feature measurements.

The purpose of this series is to apply the divergence and Chernoff's distance to hand-drawn simple geometrical figures, and examine the functional relations between these and the number of featune measurements and the probability of misrecognition.

2. STATISTICAL PATTERN RECOGNITION SYSTEMS

Statistical pattern recognition systems consist of two parts[2]: a *feature extractor*, which generates a set of feature measurements of the input sample to be recognized, and a *classifier*, which performs the function of classification. A simplified block diagram of a pattern recognition system is shown in Fig. 1.

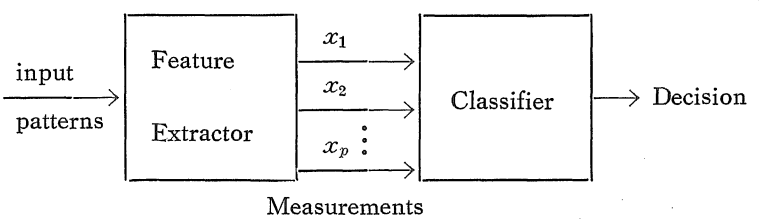


Fig. 1. A Pattern Recognition System

The feature extractor generates p feature measurements, $x_1, \ldots, x_p$, of each input sample. We will write these as a column vector

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

This quantity is the input to the classifier; it is the argument of the discriminant function $D(\boldsymbol{x})$. The discriminant function $D_j(\boldsymbol{x})$ associated with pattern class ω_j, $j=1,\ldots,m$, is such that if the input pattern represented by the feature vector $\boldsymbol{x}$ is in class ω_i, denoted as $\boldsymbol{x}\sim\omega_i$, the value of $D_i(\boldsymbol{x})$ must be the largest. That is, for all $\boldsymbol{x}\sim\omega_i$,

$$D_i(\boldsymbol{x})>D_j(\boldsymbol{x}),\ i,j=1,\ldots,m,\quad i\neq j \tag{1}$$

For the case that the multivariate probability density function of the feature vector $\boldsymbol{x}$, $P(\boldsymbol{x}|\omega_i)$, $i=1,\ldots,m$, is a multivariate normal density function, the computational algorithm of these processings has been shown by the author,[4] for example.

3. THE DIVERGENCE and CHERNOFF'S DISTANCE

3. 1. The Divergence

divergence and normal variables with unequal covariance matrices—— Assume that we are dealing with a feature extractor which for any ω_i, maps inputs of pattern class ω_i into the multivariate normal density with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Thus, when the input is from pattern class ω_i, the measurement vector $\boldsymbol{x}$ has density,

$$P(\boldsymbol{x}|\omega_i)=(2\pi)^{-p/2}|\boldsymbol{\Sigma}_i|^{-1/2}\ \exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\} \tag{2}$$

where $\boldsymbol{\Sigma}_i^{-1}$ is the matrix inverse of $\boldsymbol{\Sigma}_i$, and $(\boldsymbol{x}-\boldsymbol{\mu}_i)^t$ is the transpose of the column vector $(\boldsymbol{x}-\boldsymbol{\mu}_i)$.

Then, the likelihood function u_{ij} is expressed as

$$\begin{aligned}u_{ij}&=\log\frac{P(\boldsymbol{x}|\omega_i)}{P(\boldsymbol{x}|\omega_j)}=\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|}\\&\quad-\tfrac{1}{2}\ \mathrm{tr}\ \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^t+\tfrac{1}{2}\ \mathrm{tr}\ \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)(\boldsymbol{x}-\boldsymbol{\mu}_j)^t\end{aligned} \tag{3}$$

where tr $\boldsymbol{A}$ is the trace of matrix $\boldsymbol{A}$.

On the assumption that the input is from pattern class ω_i, the expected value of u_{ij}[5]

$$E[u_{ij}|i] = \int P(\boldsymbol{x}|\omega_i) \log \frac{P(\boldsymbol{x}|\omega_i)}{P(\boldsymbol{x}|\omega_j)}\, d\boldsymbol{x}$$

$$= \tfrac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} + \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}_i(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1})$$

$$+ \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \qquad (4)$$

Whereas, if the input is from pattern class ω_j, the corresponding value is

$$E[u_{ij}|j] = \int P(\boldsymbol{x}|\omega_j) \log \frac{P(\boldsymbol{x}|\omega_i)}{P(\boldsymbol{x}|\omega_j)}\, d\boldsymbol{x}$$

$$= \tfrac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} + \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1})$$

$$- \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \qquad (5)$$

Then, the divergence $J(\omega_i, \omega_j)$ between the classes ω_i and ω_j is defined as

$$J(\omega_i, \omega_j) = \int [P(\boldsymbol{x}|\omega_i) - P(\boldsymbol{x}|\omega_j)] \log \frac{P(\boldsymbol{x}|\omega_i)}{P(\boldsymbol{x}|\omega_j)}\, d\boldsymbol{x}$$

$$= \tfrac{1}{2} \operatorname{tr} (\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1})$$

$$+ \tfrac{1}{2} \operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \qquad (6)$$

Certain properties of divergence may be noted.[6]

1°) $J(\omega_i, \omega_j) \geqslant 0$

2°) $J(\omega_i, \omega_j) = 0 \iff P(\boldsymbol{x}|\omega_i) = P(\boldsymbol{x}|\omega_j)$

3°) $J(\omega_i, \omega_j) = J(\omega_j, \omega_i)$

4°) If the features, $t_1, \ldots, t_p$, have statistically independent outputs, then, $J(\omega_i, \omega_j|t_1, \ldots, t_p) = \sum_{k=1}^{p} J(\omega_i, \omega_j|t_k)$

5°) Adding new features to a set of features never decreases the divergence : i. e.,
$J(\omega_i, \omega_j|t_1, \ldots, t_p) \leqslant J(\omega_i, \omega_j|t_1, \ldots, t_p, t_{p+1})$

divergence and normal variables with equal covariance matrix —— Consider now the divergence in the case of normal variables with equal covariance matrix, i. e., $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, $i = 1, \ldots, m$.
In this case, we obtain the likelihood function

$$u_{ij} = -\frac{1}{2}\ \mathrm{tr}\ \Sigma^{-1}(x-\mu_i)(x-\mu_i)^t + \frac{1}{2}\ \mathrm{tr}\ \Sigma^{-1}(x-\mu_j)(x-\mu_j)^t$$
$$= \mathrm{tr}\ \Sigma^{-1}(\mu_i-\mu_j)\,x^t - \frac{1}{2}\ \mathrm{tr}\ \Sigma^{-1}(\mu_i+\mu_j)(\mu_i-\mu_j)^t \tag{6}$$

the expected value of u_{ij} when input is from pattern class ω_i

$$E[u_{ij}|i] = \frac{1}{2}\ \mathrm{tr}\ \Sigma^{-1}(\mu_i-\mu_j)(\mu_i-\mu_j)^t \tag{7}$$

and, the expected value of u_{ij} when input is from pattern class ω_j

$$E[u_{ij}|j] = -\frac{1}{2}\ \mathrm{tr}\ \Sigma^{-1}(\mu_i-\mu_j)(\mu_i-\mu_j)^t \tag{8}$$

Substracting (7) from (8), the divergence is given by

$$J(\omega_i, \omega_j) = \mathrm{tr}\ \Sigma^{-1}(\mu_i-\mu_j)(\mu_i-\mu_j)^t \tag{9}$$

the probability of misrecognition and the divergence —— It is noted if $\Sigma^{-1} = I$, the identity matrix, then $J(\omega_i, \omega_j)$ represents the squared distance between μ_i and μ_j. If a fixed sample size or nonsequential Bayes decision rule is used for the classifier, then for $P(\omega_i) = P(\omega_j) = \frac{1}{2}$ (see for example[2]),

$$x \sim \omega_i, \text{ if } u_{ij} > 0,$$
$$x \sim \omega_j, \text{ if } u_{ij} < 0 \tag{10}$$

The probability of misrecognition is

$$P_e(i, j) = \frac{1}{2}\ P(u_{ij} > 0|\omega_j) + \frac{1}{2} P(u_{ij} < 0|\omega_i)$$

It is readily shown that in this case the error probabiliy is given by the following quantity[7]

$$P_e(i, j) = \int_{\frac{1}{2}\sqrt{J}}^{\infty} (2\pi)^{-1/2}\ \exp\{-\frac{1}{2}\ y^2\}\, dy \tag{11}$$

where $J = J(\omega_i, \omega_j)$.

Hence, the probability of misrecognition $P_e(i, j)$ is a monotonically decreasing function of $J(\omega_i, \omega_j)$. Therefore, features selected and ordered according to the magnitude of $J(\omega_i, \omega_j)$ will imply their corresponding discriminatory power between ω_i and ω_j.[2]

For the case of normal variables with unequal covarince matrices there is no simple function that relates divergence to the probability of misrecognition.

Marill and Green[8] have shown upper and lower bounds on probability of correct recognition as a function of divergence for normal variables with unequal covariance matrices, with the aid of a Monte-Carlo type computer program.

the expected divergence between any pair of pattern classes —— For more than two pattern classes, the criterion of maximizing the minimum divergence or the expected divergence between any pair of pattern classes has been proposed for signal detection[9] and pattern recognition. The expected divergence between any pair of pattern classes is given

$$J(\omega) = \sum_{i=1}^{m} \sum_{j=1}^{m} P(\omega_i)\ P(\omega_j)\ J(\omega_i, \omega_j) \tag{12}$$

Let

$$d^2 = \underset{i,j}{\text{Min}}\ J(\omega_i, \omega_i), \qquad i \neq j \tag{13}$$

then

$$J(\omega) \geqslant d^2 \left\{ 1 - \sum_{j=1}^{m} [P(\omega_j)]^2 \right\} \tag{14}$$

Hence

$$d^2 \leqslant \frac{J(\omega)}{1 - \sum_{j=1}^{m} [P(\omega_j)]^2} \tag{15}$$

The tightest upper bound of d^2 must occur when $1 - \sum_{j=1}^{m} [P(\omega_j)]^2$ is a maximum. This maximum is $1 - (1/m)$ which yields

$$d^2 = \frac{m\ J(\omega)}{m - 1} \tag{16}$$

3. 2. Chernoff's Distance

In the case of normal variables with equal covariance matrix Σ, if the features have statistically independent and $\Sigma = \sigma^2 I$, the divergence (9) is transformed as follows :

$$J(\omega_i\ \omega_j) = \frac{||\mu_i - \mu_j||^2}{\sigma^2} \tag{17}$$

where $||\mu_i - \mu_j||$ is the norm of $\mu_i - \mu_j$.

Then, adding the above result to properties of the divergence 1), 2), and 3), the divergence satisfies the requirement called metric axiom.[10]

On the other hand, when we deal with a feature extractor, which, for any ω_i, maps inputs of pattern class ω_i into any multivariate distribution other than normal density, the divergence does not necessarily satisfy the requirement of metric axiom, and it is difficult that the probability of misrecognition is expressed by the divergence in that case.

Then, in order to overcome this difficulty, we may introduce Chernoff's distance defined by the following equation,

$$Cher(\omega_i, \omega_j ; \lambda) = -\log \int P(\boldsymbol{x}|\omega_i)^{\lambda} P(\boldsymbol{x}|\omega_j)^{1-\lambda} d\boldsymbol{x}, \quad 0 < \lambda < 1 \qquad (18)$$

We may also write the distance,

$$Cher(\omega_i, \omega_j ; \lambda) = -\log E\left\{\left(\frac{P(\boldsymbol{x}|\omega_i)}{P(\boldsymbol{x}|\omega_j)}\right)^{\lambda} \middle| \omega_j\right\}, \quad 0 < \lambda < 1 \qquad (19)$$

certain properties of Chernoff's distance —— In the same manner as the divergence, certain properties of Chernoff's distance may be noted,[10]

1°) $\forall\, 0 < \lambda < 1 : Cher(\omega_i, \omega_j ; \lambda) \geqslant 0$

2°) $\forall\, 0 < \lambda < 1 : Cher(\omega_i, \omega_j ; \lambda) = 0 \Longleftrightarrow P(\boldsymbol{x}|w_i) = P(\boldsymbol{x}|\omega_j)$

3°) $\forall\, 0 < \lambda < 1,\ \lambda \neq 1/2 : Cher(\omega_i, \omega_j ; \lambda) \neq Cher(\omega_j, \omega_i ; \lambda)$

4°) If the features $t_1, \ldots, t_p$ have statistically independent outputs, then

$$\forall\, 0 < \lambda < 1 : Cher(\omega_i, \omega_j ; \lambda, t_1, \ldots, t_p) = \sum_{k=1}^{p} Cher(\omega_i, \omega_j ; \lambda, t_k)$$

5°) Adding new features to a set of features never decreases distance

$$\forall\, 0 < \lambda < 1 : Cher(\omega_i, \omega_j ; \lambda, t_1, \ldots, t_p) \leqslant Cher(\omega_i, \omega_j ; \lambda, t_1, \ldots, t_p, t_{p+1})$$

Particularly, when $\lambda = 1/2$, Chernoff's distanee is called *Bhattacharyya's distance.*

Chernoff's distance and normal variables with unequal covarisnce matrices—— In the case of normal variables with unequal covariance matrices, Chernoff's distance is expressed as

$$Cher\,(\omega_i, \omega_j\,;\,\lambda) = \tfrac{1}{2}\lambda(1-\lambda)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\{(1-\lambda)\boldsymbol{\Sigma}_i+\lambda\boldsymbol{\Sigma}_j\}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) + \tfrac{1}{2}\log\left(\frac{|(1-\lambda)\boldsymbol{\Sigma}_i+\lambda\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|^{1-\lambda}\,|\boldsymbol{\Sigma}_j|^{\lambda}}\right) \tag{20}$$

Chernoff's distance and normal varidbles with equal covariance matrix —— In the case of normal variables with eqal covariance matrix, equation (20) can be written as

$$Cher\,(\omega_i, \omega_j\,;\,\lambda) = \tfrac{1}{2}\lambda(1-\lambda)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \tag{21}$$

the probability misrecognition and Chernoff's distance —— The mean probability of misrecognition obtained with aid of Bayes' decision rule is

$$P_e \leqslant \sum_{1\leqslant i<j\leqslant m} P(\omega_i)^{\alpha_{ij}} P(\omega_j)^{1-\alpha_{ij}} \int P(\boldsymbol{x}\,|\,\omega_i)^{\alpha_{ij}} P(\boldsymbol{x}\,|\,\omega_j)^{1-\alpha_{ij}}\,d\boldsymbol{x},$$
$$i, j = 1, \ldots, m, \quad 0 \leqslant \alpha_{ij} \leqslant 1 \tag{22}$$

Then, from (18), (22), the mean probability of misrecognition is given by

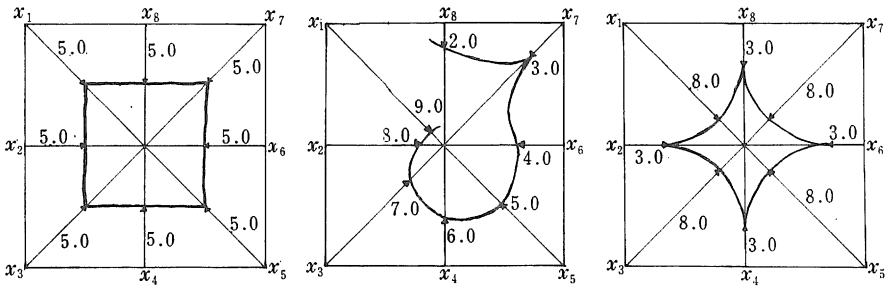
$$P_e \leqslant P(\omega_i)^{\lambda}\,P(\omega_j)^{1-\lambda}\exp\{-Cher\,(\omega_i, \omega_j\,;\,\lambda)\}, \quad 0 \leqslant \lambda \leqslant 1 \tag{23}$$

4. A NUMERICAL EXAMPLES

As an example of the application of the divergence or Chernoff's distance as a criterion of the feature selection and ordering, a very simple multiclass pattern recognition situation was simulated.

4. 1. Tentative Experimental Sitnation

Let's imagine the experimental situation asking subjects to draw three kinds of figures (let ω_A, ω_B, and ω_C denote three figure classes). These figures may have been as is seen in Fig. 2.



(a) Pattern Class w_A (b) Pattern Class w_B (c) Pattern Class w_C

Fig. 2. Tentative Pattern Samples and their Feature Measurements

Assume that eight features, $t_1, t_2, \ldots, t_8$ were selected for hand-drawn figures and the distance from the square to the edge of each figure (Fig. 2, see also e. g.,[2,4,11]) was measured and the mean vectors estimated were given in TABLE I. Again, assume that the features measurements are statistically independent and the covariance matrices of these classes are the same, $\Sigma_i = \Sigma$, $i = A, B, C$.

TABLE I

MEAN MEASUREMENT VECTORS BASED EIGHT TESTS

μ_A^t =	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
μ_B^t =	9.0	8.0	7.0	6.0	5.0	4.0	3.0	2.0
μ_C^t =	8.0	3.0	8.0	3.0	8.0	3.0	8.0	3.0

That is, in the case of normal variables with mean vectors in TABLE I and equal covariance matrix $\Sigma = \sigma^2 I$, it is the purpose of this section to examine the efficiency of the divergence and Chernoff's distance (Bhattacharyya's distance) as a criterion of feature selection and ordering, and the relation of these to the mean probability of misrecognition.

4. 2. The Expected Divergeuce and the Probability of Misrecognition

In the case of $\Sigma = 2I$, first, the expected divergence (12) between any pair of classes for eight individual tests were calculated (see TABLE II). Next, when (1) the successive features were unordered (i. e ; following the natural order $t_1, t_2, \ldots, t_8$), or (2) ordered (i. e ; following the criterion of maximizing the expected divergence) or (3) insuccessive features were ordered (i. e ; by minimizing the expected divergence), the expected divergence versus the number of features was obtained (see TABLE II). The probability of misrecognition corresponding with each expected divergence was obtained by (11). This probability is given in the right side of TABLE II.

TABELE II

THE EXPECTED DIVERGENCE AND THE PROBABILITY OF MISRECOGNITION FOR VARIOUS SETS OF TESTS: $\Sigma = 2I$

	The Expected Divergence	The Probability of Misrecognition*
Individual Tests	$J(\omega \mid t_1,) = 2.89$	.198
	$J(\omega \mid t_2,) = 4.22$	.152
	$J(\omega \mid t_3,) = 1.56$	.264
	$J(\omega \mid t_4,) = 1.56$	.264
	$J(\omega \mid t_5,) = 2.00$	.239
	$J(\omega \mid t_6,) = 0.67$	.341
	$J(\omega \mid t_7,) = 4.22$	.152
	$J(\omega \mid t_8) = 1.56$	.264
Natural Order Subsets	$J(\omega \mid t_1,) = 2.89$	.198
	$J(\omega \mid t_1, t_2,) = 7.11$	.090
	$J(\omega \mid t_1, t_2, t_3,) = 8.67$	.071
	$J(\omega \mid t_1, t_2, t_3, t_4,) = 10.22$	.055
	$J(\omega \mid t_1, t_2, t_3, t_4, t_5,) = 12.22$	.040
	$J(\omega \mid t_1, t_2, t_3, t_4, t_5, t_6,) = 12.89$	.036
	$J(\omega \mid t_1, t_2, t_3, t_4, t_5, t_6, t_7,) = 17.11$	.019
Effective Order Subsets	$J(\omega \mid t_2,) = 4.22$	.152
	(or $J(\omega \mid t_7,) = 4.22$)	(.152)
	$J(\omega \mid t_2, t_7,) = 8.44$	.072
	$J(\omega \mid t_1, t_2, t_7,) = 11.33$	.047
	$J(\omega \mid t_1, t_2, t_5, t_7,) = 13.33$	.034
	$J(\omega \mid t_1, t_2, t_3, t_5, t_7,) = 14.89$	.027
	(or $J(\omega \mid t_1, t_2, t_4, t_5, t_7,) = 14.89$)	(.027)
	(or $J(\omega \mid t_1, t_2, t_5, t_7, t_8) = 14.89$)	(.027)
	$J(\omega \mid t_1, t_2, t_3, t_4, t_5, t_7,) = 16.44$	.021
	(or $J(\omega \mid t_1, t_2, t_3, t_5, t_7, t_8) = 16.44$)	(.021)
	(or $J(\omega \mid t_1, t_2, t_4, t_5, t_7, t_8) = 16.44$)	(.021)
	$J(\omega \mid t_1, t_2, t_3, t_4, t_5, t_7, t_8) = 18.00$	.017
Ineffective Order Subsets	$J(\omega \mid t_6,) = 0.67$	.341
	$J(\omega \mid t_6, t_8) = 2.22$	.227
	(or $J(\omega \mid t_4, t_6,) = 2.22$)	(.227)
	(or $J(\omega \mid t_3, t_6,) = 2.22$)	(.227)
	$J(\omega \mid t_4, t_6, t_8) = 3.78$	.166
	(or $J(\omega \mid t_3, t_6, t_8) = 3.78$)	(.166)
	(or $J(\omega \mid t_3, t_4, t_6,) = 3.78$)	(.166)
	$J(\omega \mid t_3, t_4, t_6, t_8) = 5.33$	.123
	$J(\omega \mid t_3, t_4, t_5, t_6, t_8) = 7.33$	.087

	$J(\omega \mid t_1, \quad t_3, t_4, t_5, t_6, \quad t_8)$	$=10.22$	.055
	$J(\omega \mid t_1, \quad t_3, t_4, t_5, t_6, t_7, t_8)$	$=14.44$	.029
	(or $J(\omega \mid t_1, t_2, t_3, t_4, t_5, t_6, \quad t_8)$	$=14.44$)	(.029)
All Tests	$J(\omega \mid t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8)$	$=18.66$	.015

$$*Pe(i, j) = \int_{\frac{1}{2}\sqrt{J}}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}y^2\right\} dy$$

In the same manner, Fig. 3 shows the expected divergence and the expected Bhattacharyya's distance versus the number of feature measurements for each ordered condition when covariance matrix $\Sigma = I$.

For the successive order subsets, the expected divergence and the expected Bhattacharyya's distance versus the number of features, when covariance matrix $\Sigma = \sigma^2 I$, $\sigma^2 = \frac{1}{2}$, 1, 2, 3 and 4 respectively, is shown in Fig. 4.

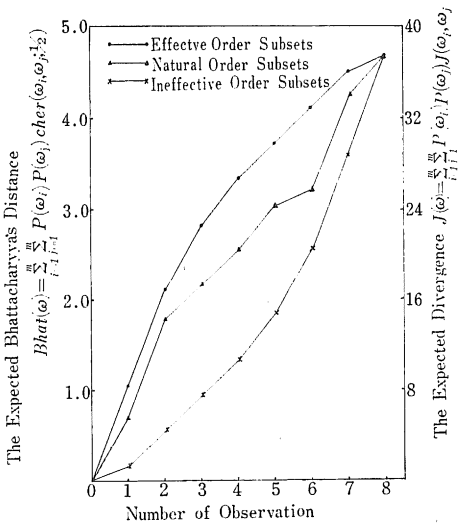


Fig. 3. The Expected Divergence and the Expected Bhattacharyya's Distance versus the Number of Feature Measurements for Each Ordered Condition (when $\Sigma = I$)

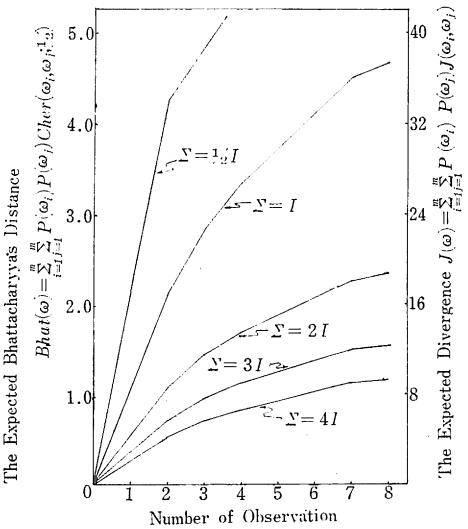


Fig. 4. The Expected Divergence and the Expected Bhattacharyya's Distance versus the Number of Feature Measure ments when Covariance Matrix $\Sigma = \sigma^2 I$

As we described, the relation between the mean probability of misrecognition and the expected Bhattacharyya's distance Bhat (ω) was given by (23). Fig. 5 shows upper bounds on the mean probability of misrecognition as a function of the expected Bhattacharyya's distance when covariance matrices are varied $\frac{1}{2}I$, I, $2I$, $3I$, and $4I$, respectively.

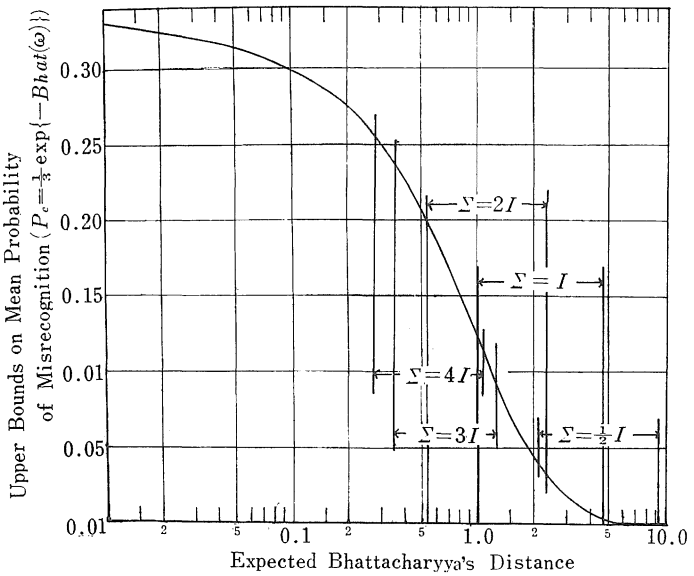


Fig. 5. Upper Bounds on the Mean Probility of Misrecognition as a Function of the Expected Bhattacharyya's Distance

5. DISCUSSION

As is evident from TABLE II and Fig. 3, although the expected divergence and Bhattacharyya's distance for all eight features is constant regardless of three ordered conditions, the first several measurements can be very effective in leading to correct recognitions when the measurements are ordered following the criterion of maximizing the expected value. That is, for example, the probability of misrecognition corresponding with the expected divergence using *effective four features* was 0.034, whereas the probability of misrecognition using *six features according to natural order* was 0.036, and the probability of misrecognition using *ineffective seven features* was 0.029 (Table II).

However, care must be taken that the probability of misrecognition defined by the equation (11) is given in the particular case of $P(\omega_i) = P(\omega_j) = \frac{1}{2}$ and the cost of misrecognition being equal. Fu, Min, and Li[11] have suggested that in a multiclass pattern recognition problem, the performance of classification can be expressed in terms of a weighted sum of all pairwise probability of misrecognition, i. e.,

$$e = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{m} P_e(i, j), \quad i \neq j \tag{24}$$

In order to minimizing this probability measure e, the criterion of maximizing the minimum divergence would be adopted rather than the criterion of maximizing the expected value. For example, when covariance matrix $\Sigma = 2I$, the probability of misrecognition corresponding with the expected divergence using only feature t_5 or only feature t_3 (or t_1) was 0.239 or 0.264, respectively, whereas, contorary above results, the weighted sum of all pairwise probability of misrecognition using only feature t_5 or only feature t_3 (or t_4) was 0.265 or 0.231, respectively.

In a multiclass pattern recognition problem, whether the criterion of maximizing the minimum divergence must be adopted or the criterion of maximizing the expected divergence must be adopted can not be known untill further study is done.

REFERENCES

1) Lewis II, P. M., The characteristic selection problem in recognition systems, *IRE Trans. Information Theory*, vol. IF-8, pp. 171-178, 1962.
2) Fu, K. S., *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, New York, 1968.
3) Babu, C. C., On the application of divergence for the extraction of features from imperfectly labeled patterns, *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-2, pp. 290-292, 1972.
4) Yabuuchi, M., Recognition of hand-drawn geometrical figures using multiple discriminant analysis, *Japanese J, Ergonomics*, vol. 8, pp. 128-134, 1972. (in Japanese)
5) Miyazawa, K., *An Introduction to Information and Decision Theory*, Iwanami, Tokyo, 1971. (in Japanese)
6) Kullback, S., *Information Theory and Statistics*, John Wiley & Sons, New York, 1958.
7) Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, New York, 1958.
8) Marill. T., and Green, D. M., On the effectiveness of receptors in recognition systems, *IEEE Trans. Information Theory*, vol. IT-9, pp. 11-17, 1963.
9) Grettenberg, T. L., Signal selection in communication and radar systems, *IEEE Trans. Information Theory*, vol. IT-9, pp. 265-275, 1963.
10) Uesaka, Y., *Theory of Pattern Recognition and Learning*. Sogo-Tosyo, Tokyo, 1971. (in Japanese)
11) Fu, K. S., Min, P. J., and Li, T. J., Feature selection in pattern recognition, *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-6, pp. 33-39, 1970.