論文2002-39SP-2-16

Multimodal 데이터에 대한 분류 에러 예측 기법

(Error Estimation Based on the Bhattacharyya Distance for Classifying Multimodal Data)

崔 義 善 *, 金 在 熹 *, 李 哲 熙 *

(Euisun Choi, Jaehee Kim, and Chulhee Lee)

요 약

본 논문에서는 multimodal 특성을 갖는 데이터에 대하여 패턴 분류 시 Bhattacharyya distance에 기반한 에러 예측 기법을 제안한다. 제안한 방법은 multimodal 데이터에 대하여 분류 에러와 Bhattacharyya distance를 각각 실험적으로 구하고 이 둘 사이의 관계를 유추하여 에러의 예측 가능성을 조사한다. 본 논문에서는 분류 에러 및 Bhattacharyya distance를 구하기 위하여 multimodal 데이터의 확률 밀도 함수를 정규 분포 특성을 갖는 부클래스들의 조합으로 추정한다. 원격 탐사 데이터를 이용하여 실험한 결과, multimodal 데이터의 분류 에러와 Bhattacharyya distance 사이에 밀접한 관련이 있음이 확인되었으며, Bhattacharyya distance를 이용한 에러 예측 가능성을 보여주었다.

Abstract

In this paper, we propose an error estimation method based on the Bhattacharyya distance for multimodal data. First, we try to find the empirical relationship between the classification error and the Bhattacharyya distance. Then, we investigate the possibility to derive the error estimation equation based on the Bhattacharyya distance for multimodal data. We assume that the distribution of multimodal data can be approximated as a mixture of several Gaussian distributions. Experimental results with remotely sensed data showed that there exist strong relationships between the Bhattacharyya distance and the classification error and that it is possible to predict the classification error using the Bhattacharyya distance for multimodal data.

Key Word : error estimation, multimodal data, Bhattacharyya distance, classitication

Ⅰ. 서 론

분류 에러의 예측은 분류기의 성능을 분석하는 일반적인 수단으로 이용되며, 주어진 문제에 대한 적절한 분류기의 선택 및 설계 그리고 특징 추출 알고리즘 개발 등에 유용하다. 따라서, 에러 예측 문제는 패턴 인식 분야에 있어서 매우 중요한 연구주제로 다수의 연구자들에 의하여 폭넓게 연구되어 왔다[1-16]. 일반적으로 분류 에러의 직접적인 계산은 데이터의 확률 밀도 함수가 주어졌을 경우로 극히 제한되며, 대부분 다중 적분식과 같은 복잡한 수식을 포함하므로 실제 계산이 용이하지 못하다[1]. 예를 들면, Fukunaga는 두 개의 정규 분포 클래스에 대하여 Bayes 에러의 확률함수를 정의하고, 이로부터 분류 에러를 수학적으로 계산할 수 있는 이론적인 수식을 유도하였다[2]. 그러나 실제 패턴 인식 문제에 있어서 데이터의 확률 밀도 함수는 적절한 가정이나 유한한 샘플 데이터로부터 추정되며, 이를

* 正會員, 延世大學校 電氣·電子工學科
(Dept. Electrical and Electronic Eng., Yonsei University)
接受日字:2001年6月30日, 수정완료일:2002年2月26日

기반으로 에러를 예측하게 된다. 일반적으로 분류 에러 예측은 분류기의 형태에 따라 확률 분포의 모델 설정을 통한 parametric 접근방식과 데이터의 확률 분포에 의존하지 않는 non-parametric 접근방식으로 크게 대별된다[1, 3-4]. Fukunaga는 위 두 가지 접근방식에 대하여 유한한 샘플 데이터로부터 Bayes 에러를 추정하는 방법인 leave-one-out 방법을 제안하였는데, non-parametric 접근방식으로 Parzen 확률밀도함수 추정법을 바탕으로한 k-NN 분류를 사용하였다[4-6]. 또한 Heydorn은 Parzen 확률밀도함수 추정법과 Bhattacharyya 계수를 이용하여 분류 에러의 상위한계를 측정할 수 있는 방법을 제안한 바 있다[7]. 이러한 방법들은 확률 분포 추정에 사용되는 kernel 선택에 따라 예측 성능이 좌우되며 샘플 데이터의 크기에도 영향을 받는다[8-9]. 한편, 신경망을 이용한 Bayes 에러 예측 방법도 제안되었는데[10], 이 방법은 여러 개의 분류기를 조합하여 패턴 분류를 수행할 경우 결정 경계(decision boundary) 영역에서의 분류 오차 발생 확률이 감소한다는 사실에 근거한 방법이다[11]. 일반적으로 분류 에러의 예측은 비교적 복잡한 계산 과정을 포함하며, 실제 패턴 분류 문제에 적용할 경우 데이터의 분포특성, 분류기의 종류, 샘플의 크기 및 차원 등의 요소를 동시에 고려해야 한다[12-13]. 이러한 관점에서, 현재까지 제안된 기법들을 이용한 신뢰성 있는 Bayes 에러의 예측은 기대하기 어려우며[13-15], 특히 원격탐사 분야 등에서 종종 취급하는 multimodal 데이터의 경우 적절한 에러 예측 기법은 아직까지 부족한 실정이다. 본 논문에서는 Bhattacharyya distance를 이용하여 multimodal 데이터에 대한 분류 에러 예측 방법을 제안한다. 제안한 방법은 최근 발표된 정규분포 데이터에 대한 Bayes 에러 예측 방법[16]의 확장으로서, Bayes 에러와 Bhattacharyya distance와의 관계를 실험적으로 유추하고, 이를 바탕으로 multimodal 데이터에 대한 에러 예측 가능성을 조사한다.

앞으로 본 논문의 구성은 제Ⅱ장에서 Bhattacharyya distance 및 이와 관련된 정규 분포 데이터에 대한 에러 예측 기법을 소개하고 제Ⅲ장에서 Bhattacharyya distance를 이용한 multimodal 데이터에 대한 에러 예측 방법을 기술한다. 제Ⅳ장에서는 실험 결과 및 정규 분포 데이터에 대한 기존 에러 예측 기법과 비교 분석하며, 마지막으로 제Ⅴ장에서 결론을 맺는다.

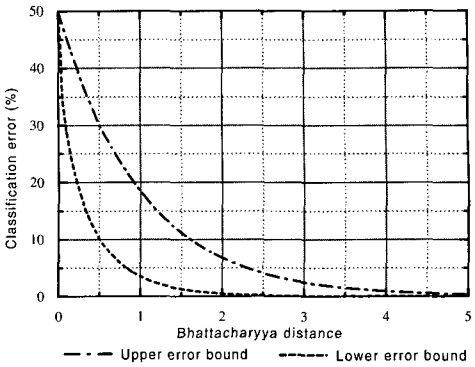


그림 1. Bhattacharyya distance와 Bayes 에러
Fig. 1. Relationship between the Bhattacharyya distance and the classification error.

Ⅱ. 연구 배경

1. Bhattacharyya distance

두 클래스간 분리도(separability)를 측정하는 통계적 수단인 Bhattacharyya distance는 특징 추출(feature extraction) 및 특징 선택(feature selection) 분야 등에 응용되어 왔으며, 분류 에러와도 밀접한 관련이 있는 것으로 알려졌다[1, 17]. Bhattacharyya distance는 두 개의 클래스에 대하여 다음 식과 같이 정의된다.

$$b = -\ln \int_{\Omega} [P(X|\omega_1)P(X|\omega_2)]^{1/2}\, dX \qquad (1)$$

여기서, $P(X|\omega_i)$는 클래스 ω_i ($i=1, 2$)의 확률밀도함수이고, Ω는 확률 분포상에서 정의되는 랜덤 변수 X의 영역이다. 특히, 클래스의 확률 분포를 정규분포로 가정할 경우 식 (1)은 다음과 같은 식으로 나타낼 수 있다[1].

$$b = \frac{1}{8}(\mu_2-\mu_1)^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}} \qquad (2)$$

여기서, μ_i와 Σ_i는 각각 클래스 ω_i의 평균 벡터와 공분산 행렬이다. 이 경우 식 (1)에서와 같이 적분 계산 과정은 필요로 하지 않는다. Bhattacharyya distance는 두 클래스의 분포가 정규분포이고 동일한 선행 확률 ($P(\omega_1)=P(\omega_2)$)을 가질 경우 다음 식 (3)과 같이 Bayes 분류 에러의 상위 한계와 하위 한계를 제공한다 [1].

$$\frac{1}{2}(1-\sqrt{1-e^{-2b}}) \le \varepsilon \le \frac{1}{2}e^{-b} \qquad (3)$$

여기서, ε 는 분류 에러이고 b는 Bhattacharyya distance이다. 그러나 그림 1에서 보여주듯이 상위 한계와 하위 한계의 간격이 비교적 크므로 실제적인 응용에는 적합하지 못하다.

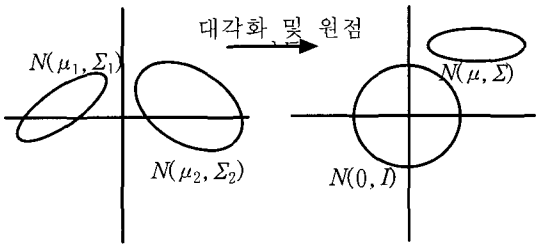


그림 2. 정규분포 클래스의 대각화(2차원 예)
Fig. 2. Diagonalization and translation of two Gaussian classes.(2 dimension)

2. Bhattacharyya distance와 에러 예측

최근, Bhattacharyya distance와 Bayes 에러와의 관계를 실험적으로 연구하여 1-2%의 오차 한계에서 에러를 예측할 수 있는 에러 예측식이 제안되었다[16]. 이 방법은 먼저 N차원 공간에서 정규 분포를 가지는 두 개의 클래스 $N(\mu_1, \Sigma_1)$, $N(\mu_2, \Sigma_2)$를 그림 2와 같이 각각 $N(0, I)$, $N(\mu, \Sigma)$으로 선형 변환하는 대각화(diagonalization)과정을 수행한다. 여기서 I는 단위행렬이고 $\mu = [m_1, m_2, \ldots, m_N]^T$ 이며 Σ는 $\lambda_1, \lambda_2, \ldots, \lambda_N$ 을 대각원소로 가지는 대각행렬이다. 대각화된 좌표계에서 Bhattacharyya distance는 다음의 식 (4)와 같이 나타낼 수 있다.

$$b = \frac{1}{2}\sum_{l=1}^{N}\left\{\frac{m_l^2}{2(1+\lambda_l)} + \ln\left(\frac{1+\lambda_l}{2}\right) - \frac{1}{2}\ln\lambda_l\right\} \quad (4)$$

Lee는 [16]에서 $m_1, m_2, \ldots, m_N$ 과 $\lambda_1, \lambda_2, \ldots, \lambda_N$으로 구성되는 N차원 $m-\lambda$ 공간을 정의하고, 정의된 $m-\lambda$ 공간에 대하여 Bhattacharyya distance의 성질에 기초한 샘플링 기법을 도입하여 약 1억 6천만 개의 클래스 조합을 생성하였다. 그리고 생성된 각 클래스 조합에 대하여 실제 패턴 분류 에러와 Bhattacharyya distance 사이의 관계를 조사하여 다음의 식 (5)와 같은 에러 예측식을 유도하였다.

$$\varepsilon = 40.219 - 70.019 * b + 63.578 * b^2 - 32.766 * b^3 + 8.7172 * b^4 - 0.91875 * b^5 \quad (5)$$

식 (5)는 정규분포를 가지는 두 클래스에 대하여 식 (2)의 Bhattacharyya distance를 이용하여 1-2%의 오차한계에서 에러 예측이 가능하다[16]. 그림 3은 식 (5)의 에러 예측식과 표준 편차, 그리고 Bayes에러의 상위 한계 및 하위 한계를 도시한 그림이다. 그림 3에서 볼 수 있듯이 비교적 표준 편차가 작으며, 상위 한계 및 하위 한계와 비교해 볼 때 간격이 작아 비교적 신뢰성있는 에러 예측이 가능하다.

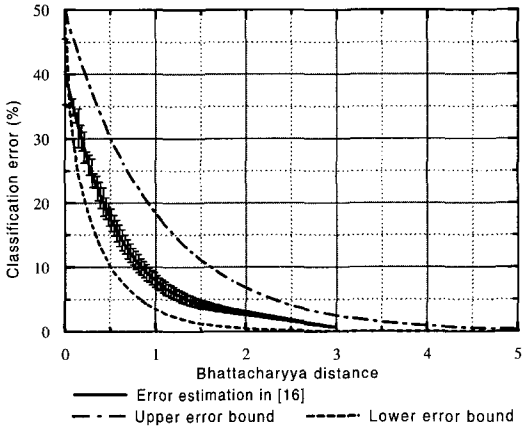


그림 3. Bhattacharyya distance와 Bayes 에러 예측[16]
Fig. 3. Error estimation and Bhattacharyya distance[16].

본 논문에서는 multimodal 분포 데이터에 대한 에러 예측을 위하여, 위에서 언급된 방법과 유사한 접근 방식을 적용하여 Bhattacharyya distance와 분류 에러와의 관계 규명을 시도하며, 특히[16]에서 소개된 에러 예측식 (5)의 multimodal 분포 데이터에 대한 적용 가능성을 조사한다.

Ⅲ. Multimodal 분포 데이터에 대한 에러 예측

본 논문에서는 패턴 분류 시 Bayes 분류기를 사용한다. Bayes 분류기는 두 클래스의 조건부 확률밀도함수 $P(X|\omega_1)$, $P(X|\omega_2)$로부터 정의되는 결정기준함수 (discrimination criterion function) $h(X) = -\ln(P(X|\omega_1)/P(X|\omega_2))$를 가지며, 다음 식 (6)과 같이 분류를 수행한다[1].

$$\text{If } h(X) < t, \text{ decide } \omega_1 \quad \text{else } \omega_2 \quad (6)$$

여기서 $t=\ln(P(\omega_1)/P(\omega_2))$이다. 특히, 두 클래스의 분포가 정규분포일 경우 결정기준함수는 다음 식과 같이 나타낼 수 있다.

$$h(X)=\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1) - \frac{1}{2}(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)+\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \quad (7)$$

식 (7)의 결정기준함수를 갖는 분류기를 가우시안 ML 분류기(Gaussian Maximum likelihood classifier)라고 하며, 원격 탐사 등과 같은 패턴 인식 분야에서 널리 사용되는 대표적인 분류기 중의 하나이다[18]. 그러나 본 논문에서 고려하는 데이터는 multimodal 특성을 갖는 데이터이므로, 패턴 분류 시 식 (7)의 직접적인 사용은 적합하지 못하며, 또한 식 (2)를 이용한 Bhattacharyya distance의 계산도 불가능하다. 본 논문에서는 multimodal 특성을 갖는 두 클래스의 Bhattacharyya distance 및 패턴 분류를 위하여 식 (8)과 같이 정규분포를 갖는 몇 개의 부클래스(subclass)들의 조합으로 multimodal 클래스의 확률밀도함수를 추정하는 혼합 정규 분포 모델링(Gaussian mixture modeling) 기법을 적용한다[19-20].

$$P(X|\omega_i)=\sum_{j=1}^{K_i}\alpha_i^j\,\mathrm{N}(\mu_i^j,\Sigma_i^j)(X) \quad (8)$$

여기서 $\mathrm{N}(\mu_i^j,\Sigma_i^j)(X)$는 multimodal 클래스 ω_i를 구성하는 K_i개의 정규분포 부클래스 중 j 번째 부클래스의 확률 밀도 함수이며, μ_i^j와 Σ_i^j는 각각 평균 벡터와 공분산 행렬이다. α_i^j는 다음의 식을 만족하는 정규화 상수이다.

$$\sum_{j=1}^{K_i}\alpha_i^j=1 \quad (9)$$

본 논문에서는 $\alpha_i^j=1/K_i$으로 설정하였다. 그림 4는 임의의 두 개의 정규 분포 부클래스들을 이용하여 모델링한 2차원 multimodal 클래스의 확률 밀도 함수를 도시한 예이다. 본 논문에서는 식 (8)과 같이 임의의 다차원 정규분포 부클래스들을 이용하여 multimodal 두 클래스 조합들을 생성하고, 식 (6)에서 정의된 Bayes 결정기준함수를 사용하여 패턴 분류를 수행하고 분류 에러를 구한다. 또한, 생성된 multimodal 두 클래스 조합들에 대하여 식 (1)을 이용하여Bhattacharyya distance를 추정하고 분류 에러와의 관계를 조사한다.

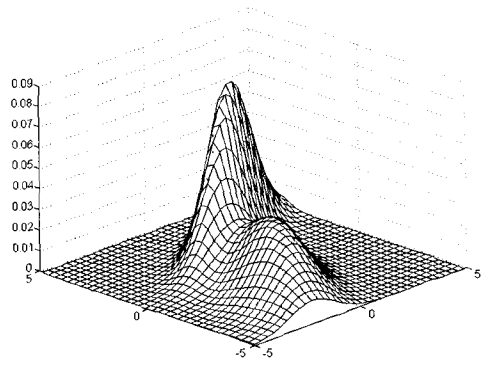


그림 4. 혼합 정규 분포(Gaussian mixture) 모델링을 이용한 multimodal 클래스의 확률 분포의 예 (2차원, K=2, α=0.5)

Fig. 4. Example of the distribution of multimodal class using the Gaussian mixture model.(2 dimension, K=2, α=0.5)

한편, 식 (1)의 경우 다차원 공간에서 다중 적분 계산이 요구되는데 본 논문에서는 Newton-Cotes 방법에 기초한 식 (10)과 같은 비교적 간단한 방법인 rectangular 수치 적분법을 이용하여 근사화시킨다[21-22]. 식 (10)은 3차원 수치 적분의 형태로, 차원이 증가할 경우 계산양은 차원 수의 지수승에 비례하게 된다.

$$I=\int_a^b\int_c^d\int_e^f f(x,y,z)\,dx\,dy\,dz \simeq h^3\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l}f_{i,j,k} \quad (10)$$

여기서 $n=(b-a)/h$, $m=(d-c)/h$, $l=(f-e)/h$ 이다. h는 다차원 공간의 샘플링 격자 간격으로 본 논문에서는 0.1을 사용하였다.

Ⅳ. 실험 결과

본 논문에서는 Multimodal 데이터에 대한 Bhattacharyya distance와 분류 에러와의 관계를 조사하기 위하여 다중 스펙트럴 데이터인 FSS(Field Spectrometer System) 원격 탐사 데이터로부터 총 4248개의 multimodal 두 클래스 조합들을 생성하였다. 표 1은 FSS 시스템의 주요사양을 나타낸다[23]. 먼저 multimodal 클래스 생성을 위하여 FSS 데이터로부터 샘플 수가 비교적 큰 40개의 클래스들을 선택하고, 선택된 각 클래스들의 통계치(평균 벡터, 공분산 행렬)를 추정하여 샘플 수

1000개인 정규 분포 클래스들을 발생시켰다. 이렇게 생성된 각 클래스들은 multimodal 클래스 생성 시 부클래스로 사용된다. 표 2는 40개 클래스 데이터에 대한 정보를 보여주며 괄호 안의 숫자는 각 클래스 데이터의 초기 샘플 수이다. 본 논문에서는 표 2의 정규 분포 클래스들을 임의로 선택하여 차원 수와 mode 수를 변화시켜 4248개의 다차원 multimodal 두 클래스 조합들을 생성하였다. 여기서 mode 수는 선택된 정규 분포 클래스들의 개수에 해당한다. 표 3은 생성된 multimodal 데이터의 차원과 mode 수에 따른 두 클래스 조합들의 개수를 보여준다. 차원 수와 mode 수는 최대 5로 제한하였으며, 각 조합의 두 multimodal 클래스들은 서로 같은 mode 수($K_1 = K_2$)를 가지는 것으로 가정하였다. 표 3의 multimodal 두 클래스 조합 4248개에 대하여 Bhattacharyya distance와 분류 에러를 각각 구하고 비교한 결과의 일부를 그림 5에 나타냈다. 그림 5는 mode 수가 4인 경우(K=4) 에 대하여 각 차원별로 분류 에러와 Bhattacharyya distance사이의 관계를 점으로 도시한 그래프로, Bayes 에러의 상위 및 하위 한계와 식 (5)의 에러 예측 식에 해당하는 곡선과도 비교하여 나타냈다. 그림 5(a)-(c)에서 각 점들은 Bayes 에러의 상위 및 하위 한계와 비교해 볼 때, 비교적 좁은 영역에 분포되어 있으며, 특히 식 (5)의 에러 예측 곡선 위에 대부분 밀집되어 있는데, 이는 Bhattacharyya distance와 multimodal 데이터의 분류 에러 사이에 밀접한 관련이 있음을 보여준다. 한편, 그림 5(d)-(e)의 경우는 식 (1)의 다중 적분의 근사화 과정에서 발생한 오차로 인해 Bhattacharyya distance가 제대로 추정되지 않았음을 간접적으로 보여준다. 표 4는 전체 클래스 조합 4248개 데이터에 대하여 Bhattacharyya distance 각 구간에 따른 분류 에러들의 평균, 표준편차, 최대값 및 최소값을 보여준다. 또한 각 구간에 해당하는 클래스 조합들의 개수도 표시하였는데, 표 4에서 전체 데이터의 99%가 3.0 미만의 Bhattacharyya distance를 나타냈다. 3.0 이상의 Bhattacharyya distance의 경우 그림 1에서 볼 수 있듯이 Bayes 에러의 상위한계와 하위한계의 간격이 충분히 작으므로 실제 에러 예측에서는 고려하지 않는다.

표 4에서 Bhattacharyya distance가 0.8 이상인 경우 에러 예측의 신뢰도에 해당하는 표준 편차값이 2% 미만인 것으로 나타났다. 이는 기존 정규분포 데이터에 대한 Bayes 에러의 상위 및 하위 한계가 10~15%의 예측 오차를 가지는 것과 비교해 볼 때, 본 논문에서 제안한 에러 예측 방법이 multimodal 데이터의 에러 예측에 적합함을 의미한다.

표 1. FSS 주요 사양
Table 1. Parameters of FSS.

No. bands	60 bands
Spectral Coverage	0.4 - 2.4 μm
Altitude	60 m
IFOV(ground)	25 m

표 2. 원격 탐사된 40개 클래스 데이터의 정보
Table 2. Information of remotely sensed 40 classes.

No.	Species	Date	No.	Species	Date
1	WINTER WHEAT (691)	770308	21	PASTURE (225)	780921
2	WINTER WHEAT (677)	770626	22	WINTER WHEAT (223)	780515
3	WINTER WHEAT (660)	771018	23	NATIVE GRASS PAS(208)	780726
4	WINTER WHEAT (657)	770503	24	PASTURE (217)	781026
5	SUMMER FALLOW (643)	770626	25	SUMMER FALLOW (216)	780816
6	SPRING WHEAT (515)	780726	26	NATIVE GRASS PAS(209)	780602
7	SPRING WHEAT (515)	780602	27	NATIVE GRASS PAS(212)	780816
8	SPRING WHEAT (474)	780515	28	SUMMER FALLOW (209)	770503
9	SPRING WHEAT (469)	780921	29	SUMMER FALLOW (200)	780726
10	SPRING WHEAT (464)	780816	30	NATIVE GRASS PAS(196)	780515
11	SPRING WHEAT (454)	780709	31	SUMMER FALLOW (190)	780709
12	SPRING WHEA (441)	781026	32	NATIVE GRASS PAS(183)	771018
13	SUMMER FALLOW (411)	760928	33	OATS (182)	780921
14	WINTER WHEAT (393)	781026	34	OATS (173)	780726
15	SPRING WHEAT (313)	771018	35	NATIVE GRASS PAS(170)	780709
16	WINTER WHEAT (292)	770920	36	OATS (165)	780816
17	WINTER WHEAT (292)	780921	37	OATS (163)	780515
18	GRAIN SORGHUM (279)	770308	38	OATS (159)	780709
19	GRAIN SORGHUM (277)	760928	39	OATS (161)	771018
20	OATS (259)	780602	40	GRAIN SORGHUM (157)	770626

표 3. 차원 수와 mode 수에 따른 multimodal 두 클래스 조합들의 개수
Table 3. Number of class pairs according to the number of dimensionalities and modes.

No. pairs of two multimodal classes					
	No. Gaussian components (K)				
	2	3	4	5	Total
1	300	300	300	300	1200
2	300	300	300	300	1200
3	300	300	300	300	1200
4	100	100	100	100	400
5	100	50	50	48	248
Total	1100	1050	1050	1048	4248

표 4. 전체 클래스 조합들에 대한 Bhattacharyya distance 구간별 분류 에러 통계

Table 4. Statistics of the classification errors according to the ranges of the Bhattacharyya distance.

Classification error (%)						
b	Avg.	Std.	Max.	Min.	No. pairs	Propotion
$b < 0.1$	39.72	3.70	49.37	30.23	889	20.9
$0.1 \le b < 0.2$	31.36	2.77	40.01	23.84	591	13.9
$0.2 \le b < 0.4$	24.67	3.01	34.88	17.95	941	22.2
$0.4 \le b < 0.6$	18.71	2.59	27.33	12.89	662	15.6
$0.6 \le b < 0.8$	14.43	2.55	25.15	9.46	407	9.6
$0.8 \le b < 1.0$	10.73	1.87	18.05	7.29	255	6.0
$1.0 \le b < 1.2$	8.15	1.53	16.67	5.78	145	3.4
$1.2 \le b < 1.4$	6.23	1.12	12.54	4.58	92	2.2
$1.4 \le b < 1.6$	4.68	0.61	6.28	3.43	91	2.1
$1.6 \le b < 1.8$	3.65	0.56	4.83	2.68	59	1.4
$1.8 \le b < 2.0$	2.88	0.50	3.90	1.98	23	0.5
$2.0 \le b < 2.2$	2.14	0.33	3.08	1.50	23	0.5
$2.2 \le b < 2.4$	1.77	0.40	2.78	1.09	21	0.5
$2.4 \le b < 2.6$	1.35	0.27	1.93	0.88	13	0.3
$b \ge 2.6$	0.45	0.46	1.45	0	36	0.8

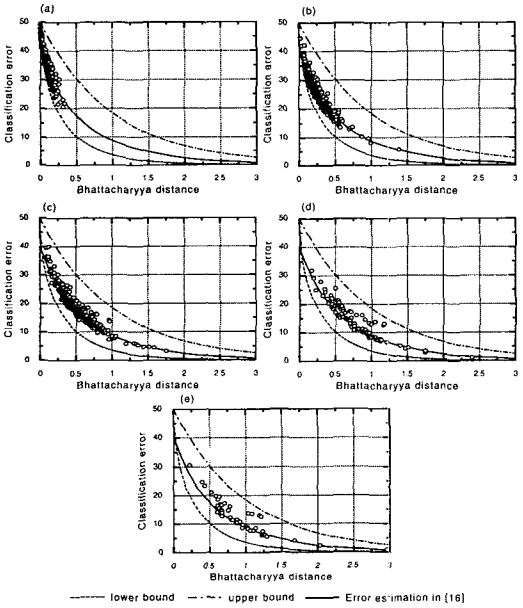


그림 5. 분류 에러와 Bhattacharyya distance와의 관계 (a) N=1, K=4 (b) N=2, K=4 (c) N=3, K=4 (d) N=4, K=4 (e) N=5, K=4

Fig. 5. Relationship between the classification error and the Bhattacharyya distance. (a) N=1, K=4 (b) N=2, K=4 (c) N=3, K=4 (d) N=4, K=4 (e) N=5, K=4

그림 6은 표 4의 multimodal 데이터에 대한 Bhattacharyya distance의 각 구간별 에러의 평균값들과 표준

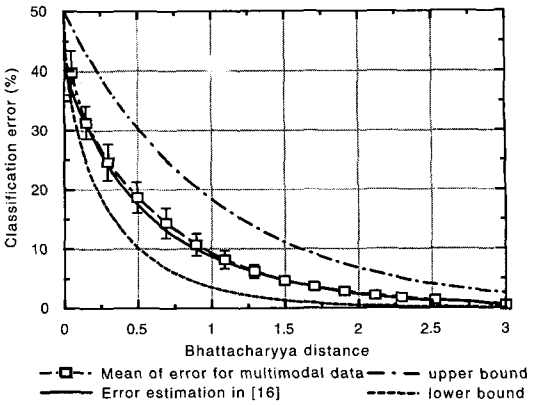


그림 6. Multimodal 데이터의 에러 예측 통계

Fig. 6. Statistics of the estimated classification error for multimodal data.

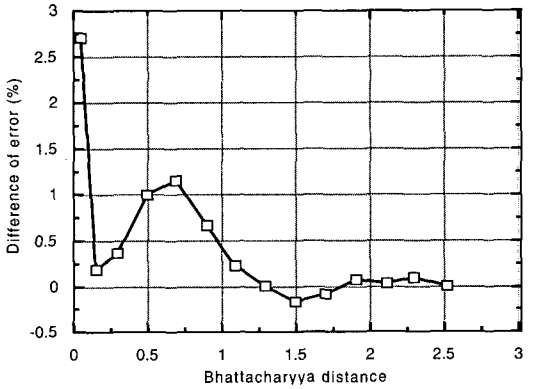


그림 7. Multimodal 데이터의 에러 평균값과 에러 예측식([16])과의 차이 비교

Fig. 7. Differences between the average of the estimated classification errors and the error estimation equation([16]).

편차를 에러 예측식 (5)와 비교한 그래프이다. 그림 6에서 볼 수 있듯이 multimodal 데이터에 대한 에러의 평균값 곡선은 식 (5)의 에러 예측 곡선과 유사하며, 에러 예측 식과 구간별 평균값 에러들과의 최대 차이는 $b < 0.05$에서 약 2.7%로 나타났다. 특히 그림 7에서 볼 수 있듯이 $b > 1.0$ 구간에서는 식 (5)의 에러 예측 곡선과의 차이가 0.5% 미만으로 나타났는데, 이는 multimodal 데이터의 에러 예측에 식 (5)의 적용 가능성을 시사하는 결과라고 볼 수 있다.

그림 8은 표 3의 전체 4248개 multimodal 클래스 조합들에 대하여 실험한 Bhattacharyya distance와 분류 에러 값들에 대하여, 다항 곡선 정합(polynomial curve fitting) 방법을 적용하여 얻은 식과 식 (5)의 에러 예측

식을 비교한 그래프이다. 그림 8에서 Bhattacharyya distance 구간 $b>2.6$인 경우 조합들의 개수가 충분하지 않았던 관계로 식 (5)와의 최대 차이는 약 1.5%였으며, 반면 나머지 구간에 대해서는 0.5% 미만으로 나타났다.

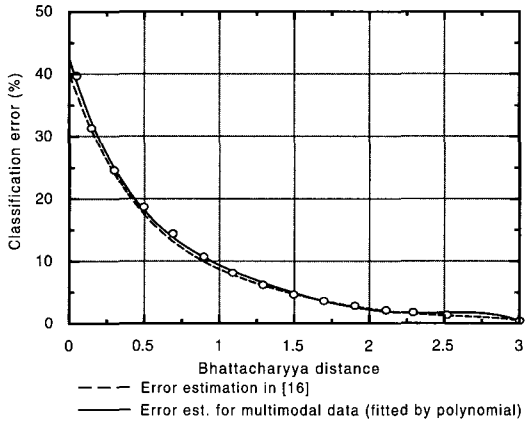


그림 8. Multimodal 데이터에 대한 에러 예측의 다항 곡선 정합 결과와 에러 예측식([16])과의 비교
Fig. 8. Differences between the polynomial curve fitted results and the error estimation equation([16]).

V. 결 론

본 논문에서는 multimodal 분포 특성을 갖는 데이터에 대한 패턴 분류 시 에러 예측을 위하여 Bhattacharyya distance와 분류 에러사이의 관계를 실험적으로 조사하고, 이를 바탕으로 Bhattacharyya distance를 이용하여 분류 에러를 예측할 수 있는 방법을 제안하였다. 혼합 정규 분포 가정을 사용하여 multimodal 데이터의 확률 분포 특성을 추정하고 실험한 결과, Bhattacharyya distance와 분류 에러사이의 연관성 및 Bhattacharyya distance를 이용한 에러 예측 가능성을 확인하였다. 특히 정규 분포 데이터의 에러 예측에 사용되는 에러 예측 식의 적용가능성을 추가로 확인하였다.

참 고 문 헌

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, New York : Academic Press, 1972.

[2] K. Fukunaga and T. F. Krile, "Calculation of Bayes' recognition error for two multivariate Gaussian distributions," IEEE Trans. Comp., vol. 18, no. 3, pp. 220~229, 1969.

[3] D. Hand, "Recent advances in error rate estimation," Pattern Recognition Lett., vol. 4, pp. 335~346, 1968.

[4] K. Fukunaga and D. Kessell, "Estimation of classification error," IEEE Trans. Comp., vol. 20, no. 12, pp. 1521~1527, 1971.

[5] K. Fukunaga and D. Hummels, "Leave-one-out procedures for nonparametric error estimates," IEEE Trans. Pattern Anal. & Machine Intelli., vol. 11, no. 4, pp. 421~423, 1989.

[6] L. Buturović, "Towards Bayes-optimal linear dimension reduction," IEEE Trans. Pattern Anal. & Machine Intelli., vol. 16, pp. 420~424, 1994.

[7] R. P. Heydorn, "An upper bound estimate on classification error," IEEE Trans. Inform. Theory, vol. 14, pp. 783~784, 1968.

[8] G. Lugosi and M. Pawlak, "On the posterior-probability estimate of the error rate of nonparametric classification rules," IEEE Trans. Inform. Theory, vol. 40, no. 2, 1994.

[9] R. Lotlikar and R. Kothari, "Adaptive linear dimensionality reduction for classification," Pattern Recognition, vol. 33, pp. 185~194, 2000.

[10] K. Tumer and J. Ghosh, "Estimating the Bayes error rate through classifier combining," ICPR'96, vol. 2, pp. 695~699, 1996.

[11] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," Pattern Recognition, vol. 29, no. 2, pp. 341~348, 1996.

[12] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," IEEE Trans. Pattern Anal. & Machine Intelli., vol. 2, no. 3, 1980.

[13] H. Schulerud, "Bias of error in linear discriminant analysis caused by feature selection and sample size," ICPR 2000, vol. 2, pp. 372~377, 2000.

[14] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," IEEE Trans. Pattern Anal. & Machine Intelli., vol. 4, pp. 154~157, 1982.

[15] A. Antos, L. Devroye and L. Gyorfi, "Lower bounds for Bayes error estimation," IEEE Trans. Pattern Anal. & Machine Intelli., vol. 21, no. 7, pp. 643~645, 1999.

[16] C. Lee and E. Choi, "Bayes error evaluation of the Gaussian ML classifier," IEEE Trans. Geos. Remote Sensing, vol. 38, no. 3, pp. 1471~1475, 2000.

[17] T. Kaliath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Commun. Technol., vol. 15, pp. 52~60, 1967.

[18] J. A. Richards, Remote Sensing Digital Image Analysis, Springer-Verlag, 1993.

[19] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, NY, 1973.

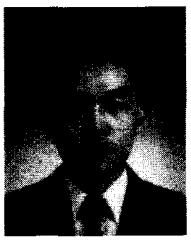
[20] T. Taxt, N. L. Hjort and L. Eikvil, "Statistical classification using a linear mixture of two multinormal probability densities," Pattern Recognition Letters, vol. 12, pp. 731~737, 1991.

[21] M. Sadiku and R. Kiem, "Newton-Cotes rules for triple integral," IEEE Proceedings. Southeastcon '90, vol. 2, pp. 471~475, 1990.

[22] W. H. Press and et al., Numerical Recipies in C, 2nd ed. pp. 161~164, Cambridge Press, NY, 1992.

[23] L. L. Biel and et al., "A Crops and Soils Data Base For Scene Radiation Research," Proc. Machine Process. of Remotely Sensed Data Symp., West Lafayette, Indiana, 1982.

저 자 소 개

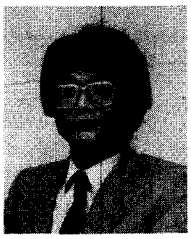


崔 義 善(正會員)

1998년 2월 : 연세대학교 전자공학과 졸업 (공학사). 1998년 3월~2000년 2월 : 연세대학교 전기·컴퓨터공학과 졸업(공학석사). 2000년 3월~현재 : 연세대학교 전기·전자공학과 박사과정. <주관심분야 : 패턴인식, 영상신호처리>

李 哲 熙(正會員)

1980년 3월~1984년 2월 : 서울대학교 전자공학과 졸업(공학사). 1984년 3월~1986년 2월 : 서울대학교 대학원 전자공학과 졸업(공학석사). 1986년 9월~1987년 3월 : Technical University of Denmark (Researcher). 1987년 8월~1992년 12월 : Purdue University Electrical Engineering(Ph. D.). 1993년 7월~1996년 8월 : National Institutes of Health, Maryland, USA.(Visiting Fellow). 1996년 9월~1999년 8월 : 연세대학교 기계전자공학부 조교수. 1999년 9월~현재 : 연세대학교 기계전자공학부 부교수. <주관심분야 : 영상신호처리, 음성신호처리, 패턴인식, 웨이블릿, 신경망>



金 在 熹(正會員)

1972년 3월~1979년 2월 : 연세대학교 전자공학(공학사). 1980년 8월~1982년 7월 : Case Western Reserve Univ. (Master). 1982년 8월~1984년 5월 : Case Western Reserve Univ. (Ph.D.). 1998년 5월~1999년 12월 : 국방부 정보정책 자문위원. 1984년 8월~현재 : 연세대학교 기계전자공학부 교수. 2001년 3월~현재 : 생체인식협의회 기술분과 위원장. 2002년 3월~현재 : 연세대학교 신호처리 연구센터 소장. 2002년 1월~현재 : 산업자원부 응용S/W 상임 평가 위원장. <주관심분야 : 인공지능, 영상 인식, 패턴인식, 전문가 시스템>