Multimodal 분포 데이터를 위한 Bhattacharyya distance 기반 분류 에러예측 기법

최의선, 이철희
연세대학교 전기·컴퓨터 공학과
Tel: 02) 361-2779 Fax: 02) 312-4584

Estimation of Classification Error Based on the Bhattacharyya Distance for Data with Multimodal Distribution

Euisun Choi and Chulhee Lee
Dept. of Electrical and Computer Eng., Yonsei University
Email: acuaris@hdsp.yonsei.ac.kr

Abstract

In pattern classification, the Bhattacharyya distance has been used as a class separability measure and provides useful information for feature selection and extraction. In this paper, we propose a method to predict the classification error for multimodal data based on the Bhattacharyya distance. In our approach, we first approximate the pdf of multimodal distribution with a Gaussian mixture model and find the bhattacharyya distance and classification error. Exprimental results showed that there is a strong relationship between the Bhattacharyya distance and the classification error for multimodal data.

1. 서 론

패턴 분류 문제에 있어서 분류 에러의 예측은 다수의 연구자들에 의하여 폭넓게 연구되어온 분야로 분류기의 설계 및 성능 분석, 특징 추출 알고리즘 개발 등에 특히 유용하다. 최근, 정규 분포 데이터에 대하여 Bhattacharyya distance를 이용하여 가우시안 ML (maximum likelihood) 분류기의 분류 에러를 1~2%의 오차한계에서 예측할 수 있는 에러 예측 기법이 발표된 바 있다 [1]. Lee는 [1]에서 가능한 모든 클래스 조합들에 대하여 Bhattacharyya distance와 분류 에러와의 관계를 실험적으로 조사하였다. 본 논문에서는 일반적인 multimodal 분포를 갖는 데이터에 대하여 Bhattacharyya distance를 이용한 에러 예측 가능성을 조사하며, [1]에서 제안된 에러 예측 기법의 적용 가능성에 대해서도 분석한다. 일반적으로 multimodal 분포를 갖는 데이터의 경우 Bhattacharyya distance를 직접적으로 계산하는데 어려움이 있으므로 본 논문에서는 두 개 이상의 정규 분포를 이용하여 multimodal 분포를 합성 모델링하고 확률 밀도 함수를 추정한다. 그리고 이를 바탕으로 Bhattacharyya distance와 분류 에러와의 관계를 실험적으로 유도한다.

2. Bhattacharyya distance

Bhattacharyya distance는 클래스간 분리도 측정의 수단으로 패턴 분류 시 정규 분포 데이터의 경우 Bayes 에러의 상위한계와 하위한계를 제공한다 [2]. 그러나 그림 1에 볼 수 있듯이 상위한계와 하위한계의 간격이 넓어 실제 응용에는 적합하지 못하다.

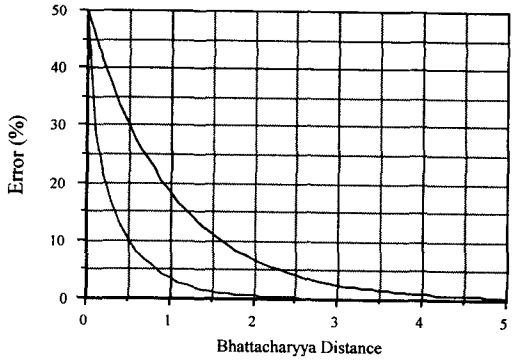

그림 1. Bayes 에러의 상위한계와 하위한계.

식 (1)과 같이 정의되는 Bhattacharyya distance는 데이터의 분포가 정규 분포인 경우 식 (2)와 같이 나타낼 수 있다.

$$\mu\left(\frac{1}{2}\right) = -\ln \int_{\varphi} [P(X|\omega_1)P(X|\omega_2)]^{1/2} dX \quad (1)$$

$$\mu\left(\frac{1}{2}\right) = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2}\ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}} \quad (2)$$

식 (1)에서 $P(X|\omega_i)$는 클래스 ω_i의 확률 밀도 함수이고 식 (2)의 M_i, Σ_i는 각각 클래스 ω_i의 평균 벡터와 공분산 행렬이다. 식 (2)에서 볼 수 있듯이 데이터의 분포가 정규 분포인 경우 Bhattacharyya distance는 직접 식 (2)를 이용하여 계산된다. [1]에서 제안된 에러 예측식은 정규 분포 데이터에 대하여 식 (2)의 Bhattacharyya distance와 분류 에러와의 관계로부터 유도되었으며 다음 식 (3)과 같다.

$$\varepsilon = 40.129 - 70.019*\mu + 63.578*\mu^2 - 32.766*\mu^3 + 8.7172*\mu^4 - 0.91875*\mu^5 \quad (3)$$

본 논문에서는 데이터의 분포가 multimodal 분포인 경우 Bhattacharyya distance를 계산하기 위하여 확률 밀도 함수 $P(X|\omega)$를 식 (4)과 같이 두 개 이상의 정규 분포를 이용하여 합성 모델링한다 [3, 4].

$$P(X|\omega_i) = \frac{1}{N}\sum_{k=1}^{N}(2\pi)^{-n/2}|\Sigma_i^k|^{-1/2} \times \exp\left[-\frac{1}{2}(X - M_i^k)^T \Sigma_i^{k-1}(X - M_i^k)\right] \quad (4)$$

여기서 M_i^k, Σ_i^k는 각각 multimodal 분포의 확률 밀도 함수 $P(X|\omega_i)$를 모델링하기 위하여 사용한 정규 분포의 평균 벡터와 공분산 행렬이며 N은 정규 분포의 개수, n은 데이터 X의 차원수이다. 이 경우 multimodal 분포 데이터에 대한 식 (1)의 Bhattacharyya distance는 식 (4)의 확률 밀도 함수를 이용하여 계산 가능하다. 그림 2는 2차원 공간에서 두 개의 정규 분포를 사용하여 추정된 bimodal 분포 데이터의 확률 밀도 함수를 나타낸다. 본 논문에서는 Bhattacharyya distance를 이용하여 multimodal 분포 데이터에 대한 패턴 분류 시 에러 예측 가능성을 조사하기 위하여 식 (4)의 확률 밀도 함수 추정에 따른 Bhattacharyya distance와 패턴 분류 시 분류 에러와의 관계를 구한다. 패턴 분류의 경우 본 논문에서는 다음 식 (5)와 같이 최대우도 판정(Maximum likelihood test)을 통하여 패턴 분류를 수행한다.

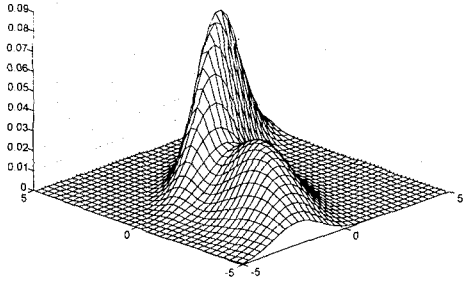

그림 2. 두 개의 정규 분포로 합성 모델링된 bimodal 분포의 확률 밀도 함수.

$$\text{Decide } \omega_1 \text{ if } h(X) < t \quad (4)$$

여기서, $h(X) = -\ln \frac{P(X|\omega_1)}{P(X|\omega_2)}$, $t = \ln \frac{P(\omega_1)}{P(\omega_2)}$ 이다.

3. 실험 및 고찰

본 논문에서는 multimodal 분포를 갖는 데이터에 대하여 Bhattacharyya distance와 패턴 분류 시 분류 에러와의 관계를 조사하기 위하여 총 1225개의 두 클래스 조합들에 대하여 각각 Bhattacharyya distance 및 분류 에러를 조사하였다. 데이터의 차원 수는 2이며 bimodal 분포를 갖는 경우로 제한하였다. 실험에 사용한 데이터는 표 1의 실제 원격탐사된 데이터 [5]의 통계치를 이용하여 발생 시킨 정규 분포 데이터로 샘플 수는 500 개이다.

표 1. 실험에 사용된 원격탐사 데이터 정보.

Species	Data	No. of sample
WINTER WHEAT	770308	691
WINTER WHEAT	770626	677
WINTER WHEAT	771018	660
SPRING WHEAT	780602	515
SPRING WHEAT	780726	515
SPRING WHEAT	780921	469
SPRING WHEAT	780709	454
SUMMER FALLOW	770626	643
SUMMER FALLOW	760928	411
GRAIN SORGHUM	770308	279
GRAIN SORGHUM	760928	277
OATS	780602	259
PASTURE	780921	225
PASTURE	781026	217
NATIVE GRASS PAS	780602	209

먼저 표 1의 데이터로부터 임의로 두 개의 클래스를 추출하여 bimodal 분포를 갖는 클래스 100개를 합성하였으며, 이와같이 생성된 bimodal 클래스 데이터로부터 두 클래스 조합 1225개에 대하여 식 (3)을 이용하여 확률 밀도 함수를 추정하고 Bhattacharyya distance 및 분류 에러를 조사하였다. 표 2는 1225개 클래스 조합들에 대하여 Bhattacharyya distance를 측정한 결과이다. 표 2에서 볼 수 있듯이 실험에 사용한 데이터의 97% 이상이 3.0 이하의 Bhattacharyya distance를 나타냈다.

표 2. 1225 클래스 조합에 대한 Bhattacharyya distance 측정 결과.

$\mu(1/2)$	No. pairs of 2 classes
0 ~ 0.5	577
0.5 ~ 1.0	413
1.0 ~ 1.5	113
1.5 ~ 2.0	50
2.0 ~ 2.5	31
2.5 ~ 3.0	15
3.0 ~ 3.5	10
3.5 ~ 7.0	16
계	1225

표 3은 Bhattacharyya distance 구간에 따른 분류 에러의 평균값과 표준 편차, 최대값 및 최소값을 나타내고 있다.

표 3. Bhattacharyya distance 구간에 따른 분류에러.

$\mu(1/2)$	Avg. (%)	Std. (%)	Max. (%)	Min. (%)
0 ~ 0.5	24.83	5.69	50.0	14.64
0.5 ~ 1.0	13.77	3.55	26.3	7.35
1.0 ~ 1.5	6.25	1.45	9.95	3.90
1.5 ~ 2.0	3.3	0.88	5.6	2.1
2.0 ~ 2.5	1.89	0.5	3.25	1.05
2.5 ~ 3.0	0.9	0.35	1.6	0.4
3.0 ~	0.29	0.26	0.8	0.0

그림 3은 1225개의 클래스 조합들에 대하여 조사한 분류 에러와 Bhattacharyya distance를 도시한 그림으로 실선은 [1]에서 제안된 에러 예측식 (3)을 적용한 그래프이다. 그림 3은 그림 1과 비교하여 볼 때, Bhattacharyya distance를 이용하여 분류 에러의 예측이 가능함을 보여준다. 또한 그림 3에서 multimodal 분포 데이터에 대한 Bhattacharyya distance 와 분류 에러와의 관계가 [1]에서 제안된 에러 예측식과 밀접한 유사성을 보이는데 이는 식 (3)의 에러 예측식을 사용하여 multimodal 분포 데이터에 대한 에러 예측이 가능함을 시사한다.

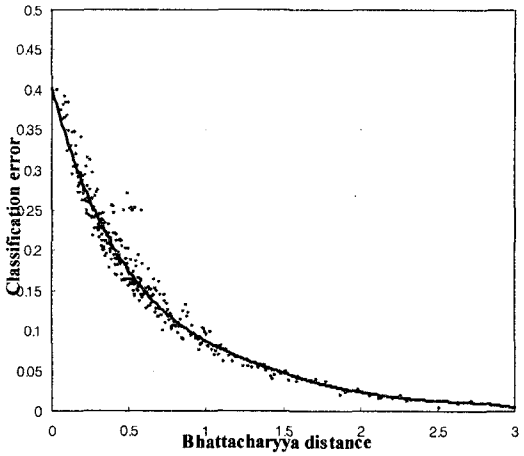

그림 3. Bhattacharyya distance 와 분류 에러와의 관계.

4. 결 론

본 논문에서는 multimodal 분포를 가지는 데이터에 대하여 Bhattacharyya distance를 이용한 에러 예측 가능성에 대하여 조사하였다. 이를 위해 두 개 이상의 정규 분포를 이용하여 multimodal 분포 데이터의 확률 밀도 함수를 추정하고 Bhattacharyya distance와 분류 에러와의 관계를 분석하였다. 실험 결과 [1]에서 제안된 에러 예측식을 통하여 multimodal 분포 데이터에 대한 패턴 분류 시 에러 예측이 가능함을 확인하였다.

참 고 문 헌

[1] C. Lee and D. A. Landgrebe, "Error estimation of the Gaussian ML classifier," *IEEE International Symp. on Information Theory*, pp. 535-535, 1997.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, New York: Academic Press, 1990.

[3] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[4] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, 1984.

[5] 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.