1999 年 度 大 韓 電 子 工 學 會 夏 季 綜 合 學 術 大 會 論 文 集 第 22 卷 第 1 號 99/6

Bhattacharyya Distance에 기반한 다중클래스 문제에 대한 피춰 추출 기법

최의선, 이철희
연세대학교 전기 · 컴퓨터공학과
Tel: (02)361-2779, E-mail: chulhee@yonsei.ac.kr

Feature Extraction Method based on Bhattacharyya Distance for Multiclass Problems

Euisun Choi, Chulhee Lee
Dept. of Electrical and Computer Engineering, Yonsei Univ.
134 Shinchon-Dong Seodaemoon-Gu Seoul, Korea

Abstract

In this paper, we propose a feature extraction method based on Bhattacharyya distance for multiclass problems. The Bhattacharyya distance provides a valuable information in determining the effectiveness of a feature set and has been used as separability measure for feature selection. Recently, a feature extraction algorithm has been proposed for two normally distributed classes based on Bhattacharyya distance. In this paper, we propose to expand the previous approach to multiclass cases. Experiment results show that the proposed method compares favorably with the conventional methods.

1. 서 론

피춰 추출은 효율적인 데이터 처리를 위해 고차원의 데이터를 저차원의 데이터로 변환시키는 과정으로 볼 수 있다. 최근 패턴 분류 문제와 관련하여 두 개의 정규분포 클래스 데이터에 대해 Bhattacharyya distance에 기반한 피춰 추출 기법이 제안되었다. 이 기법은 전체 피춰공간을 탐색하여 Bhattacharyya distance를 계산하고, 이를 이용하여 예측된 분류 에러(classification error)가 최소가 되는 피춰 벡터를 구하는 방법으로서 데이터의 분포가 정규분포임을 가정하는 경우 기존의 피춰 추출 알고리즘에 비하여 우수한 성능을 보여준다 [1]. 본 논문에서는 상기 피춰 추출 방법을 다중 클래스 문제로 확장 적용할 수 있는 방법을 제안한다.

2. 연구배경 및 이론

2.1 Bhattacharyya distance와 분류 에러

Bhattacharyya distance는 두 클래스들간의 분리도 측정에 사용되는 수단으로 데이터가 가우시안 분포를 가지는 경우 Bhattacharyya distance는 다음 식과 같다 [2].

클래스1: $N(M_1, \Sigma_1)$ 클래스2: $N(M_2, \Sigma_2)$

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

대각화된 좌표계에서 위의 식은 다음과 같이 나타낼 수 있다.

클래스1: $N(0, I)$ 클래스2: $N(M, \Sigma)$

$$\mu = \frac{1}{2}\sum_{l=1}^{n}\left\{\frac{m_l^2}{2(1+\lambda_l)} + \ln\left(\frac{1+\lambda_l}{2}\right) - \frac{1}{2}\ln\lambda_l\right\}$$

여기서 $M=[m_1, m_2, \cdots, m_n]^t$이며, Σ는 $\lambda_1, \lambda_2, \cdots, \lambda_n$를 대각원소로 가지는 대각행렬이다. 즉 μ는 m_i와 λ_i의 함수로서, m_i와 λ_i를 각각 샘플링하여 분류 에러 ε와 Bhattacharyya distance μ와의 관계를 유도하면 다음의 식 (1)과 같은 에러 예측식을 얻을 수 있다.

$$\varepsilon = 40.219 - 70.019*\mu + 63.578*\mu^2 - 32.766*\mu^3 + 8.7172*\mu^4 - 0.91875*\mu^5 \quad (1)$$

식 (1)은 가우시안 최대우도 분류기의 에러를 1-2%의 오차한계에서 예측한다 [3]. 그림 1은 Bhattacharyya distance와 분류에러와의 관계를 보여준다.

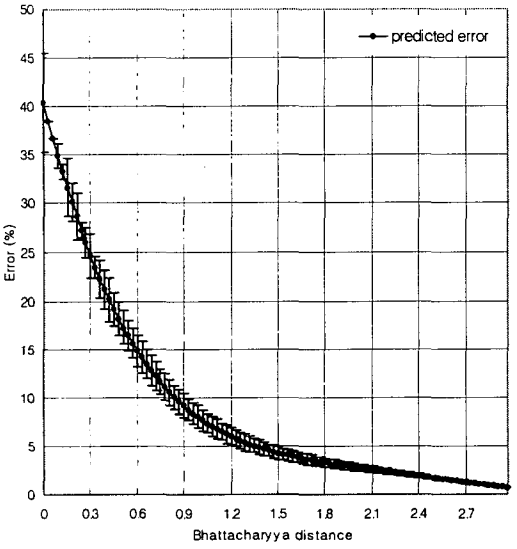

그림 1. Bhattacharyya distance와 분류에러와의 관계.

2.2 피춰 추출: 두 클래스 경우(FEBHA2)

두 개 클래스의 패턴 분류 시 유용한 피춰를 추출하기 위하여 먼저 N차원 유클리드 벡터공간에서 직교 기저집합 $\Phi_N=\{\phi_1, \phi_2, \cdots, \phi_N\}$의 단위 벡터 ϕ_1을 초기 피춰 벡터로 가정하여 분류 에러를 구한다. 이 경우 분류에러는 직접 패턴 분류를 수행하지 않고 식 (1)을 이용하여 예측함으로써 계산량을 크게 줄일 수 있다. 그 다음 피춰 벡터 ϕ_1을 식 (2)와 같이 조금씩 이동시킨다.

$$\phi_1^i = \phi_1 + a\phi_i \quad (i=2, \cdots, N) \quad (2)$$

여기서 a는 스텝 크기를 나타내는 상수이며 이동된 피춰 벡터 ϕ_1^i에 따른 에러의 변화율(gradient)은 다음의 식 (3)과 같은 방법으로 계산된다.

$$r_i = \frac{\Delta\varepsilon}{a} \approx \frac{\varepsilon(\phi_1^i) - \varepsilon(\phi_1)}{a} \quad (3)$$

여기서 $\varepsilon(\phi_1^i)$와 $\varepsilon(\phi_1)$는 각각 피춰 벡터 ϕ_1^i과 ϕ_1을 패턴분류에 사용할 경우 예측되는 분류 에러이다. 이와 같은 과정을 ϕ_i $(i=2, \cdots, N)$ 벡터들에 대해 반복하여 에러의 변화율 r_i $(i=2, \cdots, N)$를 계산한다. 최종적으로 피춰 벡터 ϕ_1은 다음의 식 (4)와 같이 예측되는 분류에러가 가장 크게 감소하는 방향으로 갱신된다.

$$\phi_{1,updated} = \phi_1 + \beta\sum_{i=2}^{N}(-r_i\phi_i) \quad (4)$$

여기서 β는 상수이다. 식 (2), (3), (4)의 과정을 갱신된 피춰 벡터 $\phi_{1,updated}$을 사용할 경우 예측되는 분류 에러의 변화율이 문턱치보다 작아질 때까지 반복하여 최종적으로 피춰 벡터를 구한다. 추가로 피춰 벡터를 구할 경우 이미 구해진 피춰 벡터와 함께 최소의 분류 에러를 얻을 수 있는 피춰 벡터를 위와 동일한 과정을 통해 구한다 [1, 4].

3. 다중 클래스 경우의 피춰 추출

본 논문에서는 Bhattacharyya distance에 기반한 두 클래스 경우의 피춰 추출 기법을 다중 클래스 경우의 문제로 확장 적용할 수 있는 방법을 제안한다. 제안한 방법은 크게 두 단계의 피춰 추출 과정을 거치게 된다. 1단계는 가능한 모든 두 개의 클래스 조합에 대하여 Bhattacharyya distance에 기반한 두 클래스 경우의 피춰 추출 기법(FEBHA2)을 이용하여 피춰 벡터를 추출한다. 즉 클래스의 수가 l이고 N차원 데이터라고 가정하면 이 경우 가능한 클래스 조합의 수는 ${}_lC_2 = l(l-1)/2$가 되며 각각의 클래스 조합에 대하여 앞에서 기술한 FEBHA2 방법을 적용하여 피춰 벡터를 추출 한다. 이 때 추출할 피춰 벡터의 수는 추출된 피춰 벡터를 이용하여 패턴 분류를 수행할 경우 예측되는 분류정확도(classification accuracy)가 피춰 추출을 하지 않은 원래의 고차원의 데이터를 사용하여 예측한

분류정확도의 η배가 되도록 제한한다. 여기서 η은 1보다 작은 양의 상수이다. 예를 들어, 임의의 클래스 조합 $\Omega_k(\omega_i, \omega_j)$에 대하여 피쳐 벡터를 추출하는 경우 다음과 같다.

1단계:

$$FEBHA2\langle \Omega_k(\omega_i, \omega_j)\rangle = \{\phi_1^k, \phi_2^k, \ldots, \phi_{M(\eta)}^k\} \quad (5)$$

$$\begin{aligned} where \quad & \phi_i^k: N\times 1 \ vector \\ & M(\eta) < N \\ & k = 1,2,\cdots, n = l(l-1)/2 \\ & 1 \le i < j \le l \end{aligned}$$

그 다음 1단계에서 구해진 각 클래스 조합들에 대한 피쳐 벡터들을 이용하여 다음의 2단계 과정을 수행한다.

2단계:

$$\begin{aligned} C_k &= \Phi_{M(\eta)}^k W_{M(\eta)}^k \Phi_{M(\eta)}^{k\,T} \\ &= \sum_{i=1}^{M(\eta)} w_i^k \phi_i^k \phi_i^{k\,T} \end{aligned} \quad (6)$$

$$where \quad \Phi_{M(\eta)}^k = [\phi_1 \ \phi_2 \ \ldots \ \phi_{M(\eta)}] \ (N\times M)$$

$$W_{M(\eta)}^k = \begin{bmatrix} w_1^k & & & \\ & w_2^k & & \\ & & \ddots & \\ & & & w_{M(\eta)}^k \end{bmatrix}$$

여기서 C_k는 $N\times M$행렬이고 w_i^k는 가중치로서 피쳐 벡터의 증가에 따른 예측된 분류정확도의 증가분이다.

$$w_i^k = acc(\Phi_i^k) - acc(\Phi_{i-1}^k), \quad i = 1,2,\ldots, M(\eta)$$
$$acc(\Phi_0^k) = acc(\phi_0^k) = 0$$

최종적으로 다중 클래스 $\omega_i \ (i = 1,2,\ldots, l)$에 대한 피쳐 벡터를 추출하기 위해 다음의 식 (7)과 같이 행렬 C의 고유치를 구하여 고유치가 큰 순서에 해당하는 고유벡터를 피쳐 벡터로 선택한다.

$$\begin{aligned} C &= \sum_{k=1}^{n} C_k \quad , n = l(l-1)/2 \\ eigen\ value\{C\} &= \lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_M > \cdots \\ eigen\ vector\{C\} &= \{\phi_1, \phi_2, \phi_3, \cdots, \phi_M, \cdots\} \end{aligned} \quad (7)$$

4. 실험 결과

본 논문에서 제안한 피쳐 추출 방법의 성능을 평가하기 위하여 기존의 다중 클래스 문제에 대한 피쳐 추출 알고리즘인 canonical analysis방법 및 주성분 분석 방법(principal component analysis) [5]과 비교하였다. 실험에 사용된 분류기는 가우시안 최대우도 분류기이며 실제 원격 탐사된 데이터를 이용하였다 [6]. 표 1은 실험에 사용된 클래스 데이터의 정보를 보여준다.

표 1. 클래스 정보.

Group	Species	Date	No. of sample (train)
1	SPRING WHEAT	78.06.02	515(300)
	SPRING WHEAT	78.07.26	515(300)
	SUMMER FALLOW	77.06.26	643(300)
	WINTER WHEAT	77.10.18	660(300)
	WINTER WHEAT	77.03.08	691(300)
2	SPRING WHEAT	78.06.02	515(300)
	SPRING WHEAT	78.07.26	515(300)
	WINTER WHEAT	77.05.03	657(300)
	WINTER WHEAT	77.10.18	660(300)
	WINTER WHEAT	77.03.08	691(300)
3	SPRING WHEAT	78.05.15	474(300)
	SPRING WHEAT	78.06.02	515(300)
	WINTER WHEAT	77.10.18	660(300)
	WINTER WHEAT	77.06.26	677(300)
	WINTER WHEAT	77.03.08	691(300)
4	SPRING WHEAT	78.05.15	474(300)
	SUMMER FALLOW	77.06.26	643(300)
	WINTER WHEAT	77.05.03	657(300)
	WINTER WHEAT	77.10.18	660(300)
	WINTER WHEAT	77.06.26	677(300)
5	SPRING WHEAT	78.07.26	515(300)
	SUMMER FALLOW	77.06.26	643(300)
	WINTER WHEAT	77.05.03	657(300)
	WINTER WHEAT	77.10.18	660(300)
	WINTER WHEAT	77.06.26	677(300)
6	SPRING WHEAT	78.06.02	515(300)
	SUMMER FALLOW	77.06.26	643(300)
	WINTER WHEAT	77.05.03	657(300)
	WINTER WHEAT	77.06.26	677(300)
	WINTER WHEAT	77.03.08	691(300)

1단계 실험 과정에서 사용한 α와 β의 값은 각각 0.05와 2.0이며 문턱치는 0.0001이다. 클래스의 수(l)는 5개이며 데이터의 차원 수(N)는 8이다. 그림 2는 표 1의 그룹 1, 2 ,3에 대한 통계치를 사용하여 각 클래스의 샘플 수가 1000개인 정규분포 데이터를 발생시켜 실험한 결과이며 그림 3은 그룹 4, 5, 6의 실제 데이터를 이용한 실험 결과이다. 그림에서 볼 수 있듯이 대체적으로 기존 피쳐 추출 알고리즘에 비하여 향상된 성능을 나타냈다.

5. 결 론

본 논문에서는 Bhattacharyya distance에 기반한 두 클래스 경우의 피쳐 추출 기법을 다중 클래스 경우의

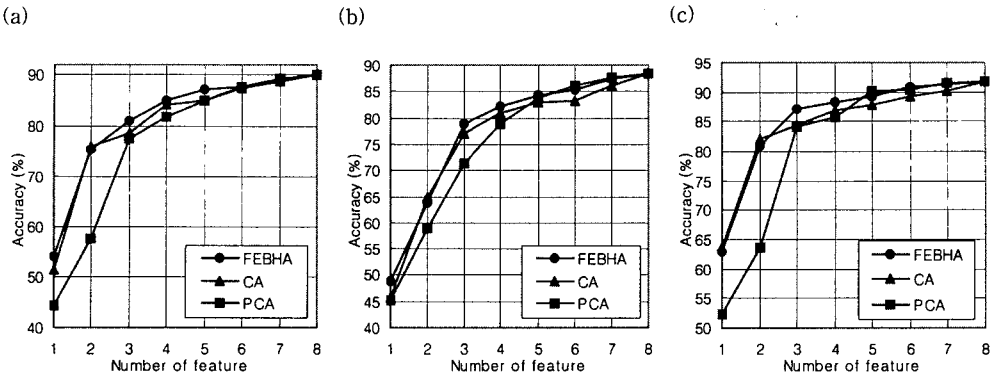

그림 2. 정규분포로 발생시킨 데이터에 대한 실험결과(No. sample: 1000). (a)그룹 1 (b)그룹 2 (c)그룹 3

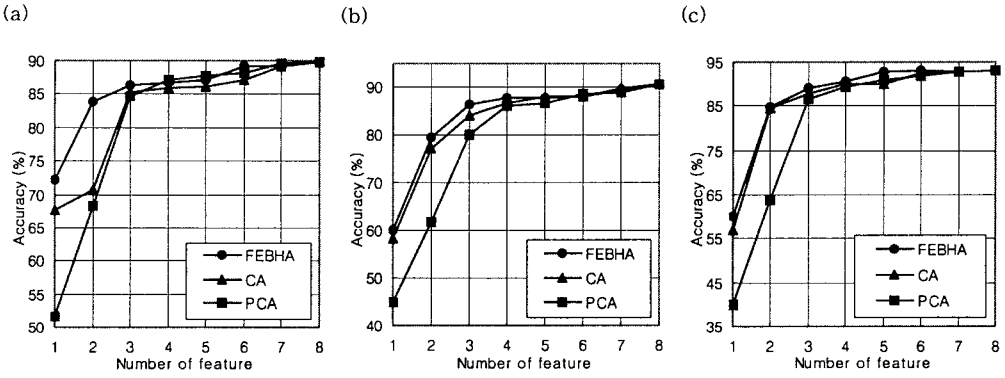

그림 3. 실제 데이터에 대한 실험결과. (a)그룹 4 (b)그룹 5 (c)그룹 6

문제로 확장 적용할 수 있는 방법을 제안하였다. 실험 결과 기존의 피춰 추출 알고리즘에 비하여 약간의 성능 향상이 있었다. 향후 2단계 피춰 추출 과정에서의 정보 손실을 최소화 할 수 있는 최적화 방안에 대한 연구가 요구된다.

참 고 문 헌

[1] C. Lee and D. Hong, "Feature extraction using the Bhattacharyya distance," *IEEE International Conf. on Systems, Man, and Cybernetics, Orlando, Florida,* 1997, pp. 2147-2150.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* NewYork: Academic Press, pp.225-226, 1990.

[3] C. Lee "Error estimation of the Gaussian ML classifer," in Proc. *IEEE International Symposium on Information Theory,* Ulm, Germany, Tune 1997.

[4] C. Lee, E. Choi and J. Kim, "Optimal Feature Extraction for Normally Distributed Data," SPIE AeroSense '98, Orlando, Florida, pp. 223-232, 1998.

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

[6] 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.