2000年度 大韓電子工學會 夏季綜合學術大會 論文集 第23卷 第1號 2000/6

다중 클래스 분포 문제에 대한 분류 정확도 분석

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

Analysis of Classification Accuracy for Multiclass Problems

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

Abstract

In this paper, we investigate the distribution of classification accuracies of multiclass problems in the feature space and analayze performances of the conventional feature extraction algorithms. In order to find the distribution of classification accuracies, we sample the feature space and compute the classification accuracy corresponding to each sampling point. Experimental results showed that there exist much better feature sets that the conventional feature extraction algorithms fail to find. In addition, the distribution of classification accuracies is useful for developing and evaluating the feature extraction algorithm.

1. 서 론

일반적으로 패턴 인식은 효율적인 패턴 분류를 위하여 특징 추출 단계를 포함한다 [1]. 패턴 분류 시 특징 추출은 입력 데이터로부터 분류에 필요한 정보들만을 추출하게 되며 고차원 데이터인 경우 저차원 데이터로 변환시켜 분류기의 복잡도를 감소시킨다. 이와 같은 이유로 다수의 특징 추출 알고리즘들이 제안되었으며, 실제 여러 패턴 인식 시스템에 응용되고 있다. 예를 들면, 주성분 분석(principal component analysis) 방법은 전체 클래스들의 공분산 행렬을 이용하여 특징 벡터를 추출하며, canonical analysis 기법은 각 클래스의 평균 벡터와 공분산 행렬에 가중치를 곱하여 구한 결정 함수를 사용하여 특징 벡터를 추출한다 [2]. 그러나 다중 클래스 패턴 분류 문제와 관련하여 이러한 방법들은 주어진 문제에 대하여 최적의 해를 제공한다는 보장이 없다. 또한, 두 클래스 모델을 기반으로한 기존의 특징 추출 알고리즘들은 다중 클래스 문제의 경우 결정함수를 확장 적용하는데, 이 경우 결정 함수의 확장에 따른 최적성 결여의 문제가 제기된다.

본 논문에서는 다중 클래스 패턴 분류 문제에 대하여 기존의 특징 추출 알고리즘들이 갖고 있는 근본 문제를 분류 정확도 관점에서 고찰하고, 최적 알고리즘 개발에 대한 방향을 제시한다. 이를 위해, 본 논문에서는 전체 특징 공간에 대한 패턴 분류 시 분류 정확도의 분포를 구하고, 기존 특징 추출 알고리즘들의 성능을 분석한다.

2. 분류 정확도 분포

선형 특징 추출은 일반적으로 식 (1)과 같이 N차원 데이터 X를 M차원 ($M<N$) 데이터 Y로 선형 변환하는 것으로 볼 수 있다 [3].

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_M \end{bmatrix} = A^T X = [\phi_1, \phi_2, \cdots, \phi_M]^T X \qquad (1)$$

여기서, A는 $N\times M$ 직교행렬이며 ϕ_i는 $N\times 1$ 특징 벡터이다. Y는 데이터 X가 선형 변환된 특징 공간의 벡터로 분류기의 입력으로 사용된다. 즉, 특징 추출을 통한 패턴 분류의 분류 정확도는 특징 벡터 ϕ_i에 의하여 값이 변하게 되므로 이러한 관계를 이용하여 전체 특징 벡터 공간 Φ_A에 대한 분류 정확도의 분포를 구할 수 있다. 특징 공간 Φ_A는 특징 벡터 집합 $\{\phi_1, \phi_2, \cdots, \phi_M\}$으로 표현되며 $N\times M$개의 변수들이 존재한다. 따라서 이러한 변수들로 정의된 특징 공간을 샘플링하고 샘플링된 각 점에 대하여 분류 정확도

를 산출하여 분포를 구할 수 있다. 그러나 이 경우, 데이터의 차원이 고차원일수록 계산이 복잡하므로 본 논문에서는 다중 클래스 패턴 분류 문제에 관한 분류 정확도의 분포를 이해하기 위하여 3차원 데이터를 고려하며 2개의 특징 벡터를 추출하는 경우로 제한한다.

3차원 공간에서 두 개의 특징 벡터 ϕ_1과 ϕ_2를 추출하는 경우 원래의 데이터 X는 식 (1)에 의하여 다음과 같이 선형 변환된다.

$$y_1 = \phi_1^T X$$
$$y_2 = \phi_2^T X \quad (2)$$

여기서 선형 변환된 데이터 Y는 서로 직교하는 두 개의 특징 벡터 ϕ_1과 ϕ_2로 정의되는 평면에 투영된 값으로 볼 수 있다. 한편, 본 논문에서 실험을 위하여 사용한 가우시안 최대우도 분류기 (Gaussian Maximum likelihood classifier)는 불특이(nonsingular) 선형 변환의 경우 불변성(invariant)을 가지므로 동일 평면을 형성하는 모든 ϕ_1, ϕ_2 조합에 대하여 동일한 분류 정확도를 보인다 [3]. 따라서 이러한 평면에 수직한 단위벡터 V_n을 고려할 경우 특징 벡터 ϕ_1과 ϕ_2는 벡터 V_n에 의하여 결정된다. 그림 1은 벡터 V_n을 구 좌표계(spherical coordinate system)에서 표현한 것이다. 그림에서 볼 수 있듯이 V_n은 두 개의 각 φ, θ에 의하여 정의된다.

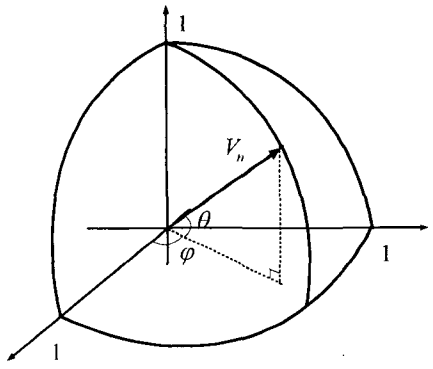

그림 1. 구 좌표계에서의 벡터 V_n.

본 논문에서는 3차원 데이터에 대하여 두 개의 특징 벡터 추출 시 다중 클래스 분포 문제에 대한 분류 정확도의 분포를 구하기 위하여 전체 특징 공간 Φ_A를 φ와 θ를 이용하여 샘플링하고, 샘플링된 각점에 대한 가우시안 최대우도 분류기의 분류 정확도를 계산한다.

3. 실험결과 및 고찰

본 논문에서는 다중 클래스 문제에 관한 분류 정확도의 분포를 추정하기 위하여 표 1의 실제 원격탐사 데이터 [4]에서 임의로 4개의 클래스를 선택하였으며 선택된 클래스의 통계치를 이용하여 정규 분포를 갖는 클래스 데이터를 발생시켰다. 각 클래스의 샘플 수는 1000이며 패턴 분류 시 가우시안 최대우도 분류기를 사용하여 분류 정확도를 계산하였다. 특징 공간 샘플링 과정에서 φ와 θ는 각각 -90° ~90° , 0° ~180° 의 범위에서 샘플링하였으며 스텝 크기는 1(degree)로 하였다.

표 1. 실험에 사용한 원격 탐사 데이터.

No. class	Species	Date	No. of sample
1	WINTER WHEAT	770308	691
2	WINTER WHEAT	770626	677
3	WINTER WHEAT	771018	660
4	WINTER WHEAT	770503	657
5	SUMMER FALLOW	770626	643
6	SPRING WHEAT	780726	515
7	SPRING WHEAT	780602	515
8	SPRING WHEAT	780515	474
9	SPRING WHEAT	780921	469
10	SPRING WHEAT	780816	464
11	SPRING WHEAT	780709	454
12	SPRING WHEAT	781026	441

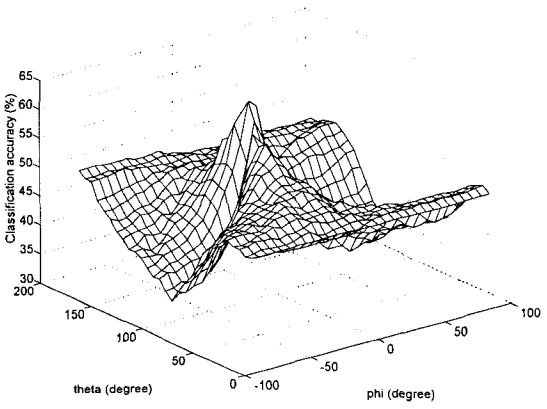

그림 2. 다중클래스 문제에 관한 분류 정확도의 분포.

그림 2는 φ와 θ를 이용하여 전체 특징 공간 Φ_A을 샘플링하고 샘플링된 각 점에 대하여 특징 벡터 ϕ_1과 ϕ_2를 구하여 분류 정확도를 계산한 후 얻은 다중 클래스 패턴 분류 문제의 분류 정확도 분포이다.

다음 실험에서는 이러한 분류 정확도의 분포를 이용하여 기존 특징 추출 알고리즘들의 성능을 평가하였다. 실험을 위해 사용한 알고리즘은 canonical analysis [2], 주성분 분석 방법 [2] 및 decision boundary feature extraction 방법 [5]으로 다중 클래스 패턴 분류 문제에 있어서 일반적으로 적용되는 방법들이다. 그림 3, 4는 분류 정확도의 분포와 함께 기존 특징 추출 알고리즘으로 구한 특징 벡터를 이용하여 패턴 분류를 수행하고 분류 정확도를 측정한 결과이다.

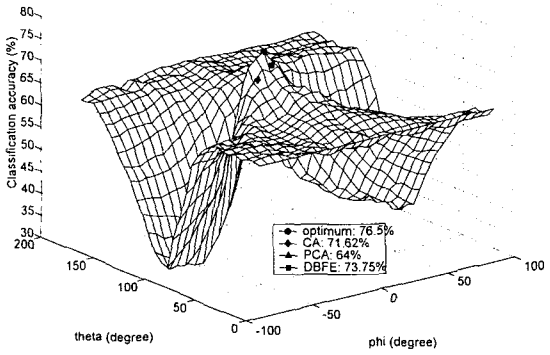

그림 3. 기존 특징 추출 알고리즘들의 성능 분석 I.

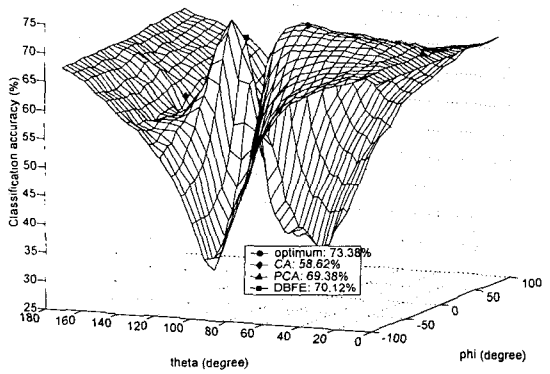

그림 4. 기존 특징 추출 알고리즘들의 성능 분석 II.

그림에서 범례 기호 '•'는 분류 정확도 분포상에서 최대 분류 정확도를 나타낸다. 그림 3에서 기존 특징 추출 알고리즘인 canonical analysis 및 decision boundary feature extraction 방법으로 추출한 특징 벡터는 비교적 최적 특징 벡터와 유사하나 분류 정확도는 각각 71.62%, 73.75%로 최대 분류 정확도 76.5%와는 차이를 보이고 있다. 또한 주성분 분석 방법은 64%의 분류 정확도로 저조한 성능을 보였다. 한편 그림 4에서 기존 특징 추출 알고리즘들로 구한 특징 벡터들은 최적 특징 벡터와 큰 차이를 보이고 있는데, 다중 클래스 패턴 분류 문제에 있어서 최적 특징 벡터를 추출하는 하기 위한 방법으로 특징 공간에서 분류 정확도의 변화율에 근거한 특징 추출 방법을 고려할 수 있다. 그림 5, 6은 임의의 특징 벡터를 사용하여 분류 정확도의 변화율을 조사하고 steepest ascent gradient 알고리즘을 적용하여 특징을 추출하는 OPTFE(optimal feature extraction) 특징 추출 기법 [6, 7]과 기존 특징 추출 알고리즘들과의 성능을 비교한 그림이다. 그림 5에서 볼 수 있듯이 OPTFE 기법은 최적 특징 벡터를 성공적으로 추출하고 있는 반면, 기존 알고리즘들은 상대적으로 저조한 성능을 보이고 있다.

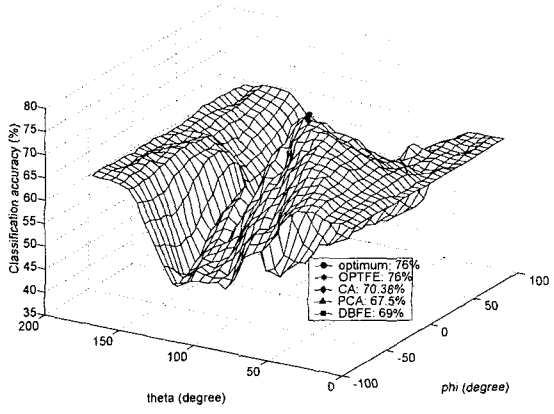

그림 5. 분류 정확도의 변화율을 이용한 특징 추출 I.

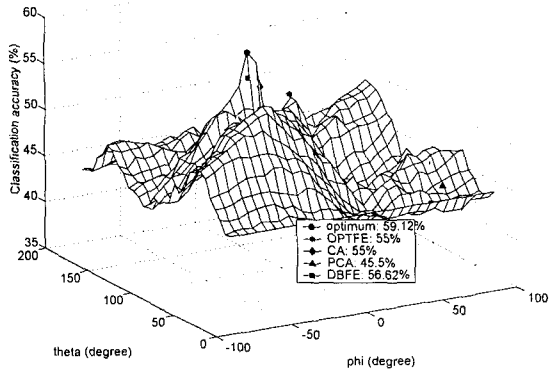

그림 6. 분류 정확도의 변화율을 이용한 특징 추출 II.

그러나 이와 같이 분류 정확도의 변화율에 근거한 특징 추출 방법은 그림 6과 같이 지역 극대에 도달하여 최적해를 구하기 어려운 문제점이 발생할 수 있다.

4. 결 론

본 논문에서는 다중 클래스 패턴 분류 문제의 분류 정확도 분포를 구하고 이를 바탕으로 기존 특징 추출 알고리즘들의 성능을 분석하였다. 실험 결과 기존 특징 추출 알고리즘들이 찾지 못하는 우수한 특징들이 존재함을 확인하였으며, 최적 특징 벡터를 추출하기 위한 해결책의 한 방법으로 분류 정확도의 변화율에 기초한 특징 추출 방법을 제시하였다. 본 논문에서 사용한 분류 정확도 분포는 패턴 분류 문제와 관련하여 특징 추출 알고리즘 개발 및 성능 평가에 유용할 것으로 기대된다.

참 고 문 헌

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

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

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

[4] 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.

[5] C. Lee and D.A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 15, no. 4, pp. 388-400, 1993.

[6] C. Lee and J. Hong, "Optimizing Feature Extraction for Multiclass cases," *IEEE Intl. Conf. on Systems Man and Cybernetics*, pp. 2545-2548, 1997.

[7] 최의선, 이철희 "다중 클래스 데이터를 위한 분류 오차 최소화 기반 특징추출 기법," 대한 전자공학회지, 제 37 권 제 2 호, 2000년 3월.