논문 2006-43CI-1-2

Prototype Reduction Schemes와 Mahalanobis 거리를 이용한 Relational Discriminant Analysis

(Relational Discriminant Analysis Using Prototype Reduction Schemes and Mahalanobis Distances)

김 상 운*

(Sang-Woon Kim)

요 약

RDA(Relational Discriminant Analysis)는 패턴의 특징벡터 대신에 학습 패턴을 대표하는 프로토타입들과의 비유사도 벡터에 기반하여 식별기를 설계하는 방법이다. 따라서 RDA 식별기의 성능은 프로토타입을 선택하는 방법과 비유사도를 측정하는 방법에 따라 결정된다. 본 논문에서는 PRS(Prototype Reduction Schemes)를 이용하여 프로토타입을 추출한 다음, 샘플 벡터들 간의 마할라노비스 거리에 의한 상관행렬로 RDA의 식별성능을 향상시키는 방법을 제안한다. 인공 데이터 및 실-생활 데이터를 대상으로 실험한 결과, 제안한 방법의 식별성능이 기존의 방법에 비하여 개선되었음을 확인하였다.

Abstract

RDA(Relational Discriminant Analysis) is a way of finding classifiers based on the dissimilarity measures among the prototypes extracted from feature vectors instead of the feature vectors themselves. Therefore, the accuracy of the RDA classifier is dependent on the methods of selecting prototypes and measuring proximities. In this paper we propose to utilize PRS(Prototype Reduction Schemes) and Mahalanobis distances to devise a method of increasing classification accuracies. Our experimental results demonstrate that the proposed mechanism increases the classification accuracy compared with the conventional approaches for samples involving real-life data sets as well as artificial data sets.

Keywords : RDA(Relational Discriminant Analysis), PRS(Prototype Reduction Schemes), 비유사도, 마할라노비스거리

I. 서 론

통계적 패턴인식(statistical pattern recognition)에서의 식별은 패턴을 특징공간에 사상한 다음, 이 공간에서 클래스를 결정할 수 있는 식별기를 학습하는 방법이다[1]. 즉, 사상된 특징공간에서 패턴들간의 유사성이나 군집성에 기반한 식별이다. 그러나 패턴벡터의 차원이 고차원이 되면 특징들 사이에 정보가 중복되고, 처리시간이 증가하는 차원의 유해로움(the curse of dimensionality)[1]이라는 문제가 발생한다. 특히, 데이터 마이닝이나 멀티미디어 검색 등의 응용에서는 패턴차원은 고차원이 되는 반면, 학습을 위한 샘플 수는 충분치 못한 경우가 많다. 즉, 패턴 샘플의 수가 패턴 차원에 비하여 적을 경우 변동행렬(scatter matrix)이 특이행렬이 되기 때문에 변별력이 떨어지는 희소성 문제(undersampled problem)[2]가 발생한다. 본 논문에서는 희소성 문제를 해결하기 위해 RDA를 적용하는 방법을 검토하고, PRS와 마할라노비스 거리를 이용하여 식별

* 정회원, 명지대학교 컴퓨터공학과
(Dept of Computer Science & Engineering, Myongji University)
접수일자 : 2005년6월10일 수정완료일 : 2006년1월3일

성능을 향상시킨 실험결과를 보고한다.

RDA(Relational Discriminant Analysis)[3]는 입력패턴을 특징벡터 대신에 학습패턴에서 추출한 프로토타입(prototype)들과의 비유사도(dissimilarity) 또는 유사도(similarity)를 이용하는 식별법(dissimilarity-based classification)이다. 즉, n개의 학습 패턴 집합 T에서 $m(\leq n)$개의 프로토타입을 선택한다면, 입력 패턴과 프로토타입들과의 비유사도는 m차원 벡터가 되며 (앞으로, 이 벡터를 비유사도 벡터라 한다), 이 벡터들로 구성되는 공간에서 식별을 수행할 수 있다. 이와 같이, RDA는 주어진 특징공간과는 다른 공간인 비유사도 벡터공간에서 비유사도만으로 식별하기 때문에 무특징 분류(featureless classification)라고도 한다.

최근, RDA의 식별특성 및 기존의 식별법들과의 관계를 규명하려는 연구가 활발히 진행되고 있다. Duin 등[3][4]은 비유사도 벡터공간과 주어진 특징공간에서 다양한 데이터를 대상으로 k-NN 식별기, 선형 식별기, 2차 식별기 들의 성능을 비교 분석하였다. 특히, 그들은 문헌 [4]에서 RLNC(Regularized Linear Normal density-based Classifier)나 RQNC(Regularized Quadratic Normal density-based Classifier) 식별기가 k-NN 식별기보다 우수함을 보고하였다. 또한, Horikawa[5]는 데이터의 분포가 클래스별로 다를 경우, 패턴 차원이 증가함에 따라 비유사도 공간에서의 k-NN 성능이 증가한다는 컴퓨터 실험 결과를 보고하고 있다.

앞에서 설명한 바와 같이, RDA의 식별성능은 프로토타입을 선택하는 방법과 비유사도를 측정하는 방법에 따라 결정된다. 지금까지 수행된 Duin팀의 연구에서는 학습패턴에서 프로토타입을 랜덤하게 선택하였고, 유클리드 거리 (Euclidean Distance: ED)로 비유사도를 측정하였다. RDA에서 n개의 학습패턴에서 m개의 프로토타입을 선택하는 이유는 비유사도의 차원을 가능한 한 최소의 것으로 줄이기 위해서 이다. 그런데 실제의 식별은 비유사도 공간에서 이루어지기 때문에 학습패턴의 정보가 충실히 반영되도록 비유사도 공간을 작성하여야 한다. 또, 비유사도를 측정하는 방법도 패턴구조를 효과적으로 반영할 수 있어야 한다.

본 논문에서는 PRS(Prototype Reduction Schemes)[6]~[9]를 이용하여 학습패턴의 프로토타입을 추출하고, 비유사도 측정방법으로 분산구조를 고려한 마할라노비스 거리(Mahalanobis Distance: MD)를 이용하는 방법을 제안한다. 여기서 PRS란 식별성능을 희생시키지 않는 범위내에서 주어진 학습패턴의 수를 최소한의 것으로 줄이는 방법이다. Hart[10]가 CNN(the Condensed Nearest Neighbour rule)을 제안한 이후, PNN(the Prototypes for Nearest Neighbour classifiers)[7] 등 성능을 개선한 다양한 기법들이 개발되었다[7]~[11]. 최근에 Kim and Oommen[8]은 CNN, PNN 등과 같은 기존의 PRS로 초기 프로토타입을 선정한 다음, LVQ3 타입의 알고리즘을 이용하여 최적의 위치로 조정하는 HYB (the Hybridized technique)[8]를 발표하였다.

본 논문에서 제안한 방법을 평가하기 위해, 인공적으로 생성한 데이터와 UCL machine learning database[12]에서 내려 받은 실-생활 벤취마크 데이터를 이용한다. 실험을 통해, 프로토타입을 랜덤하게 선택한 다음 유클리드 거리(ED)로 유사도를 측정하는 기존의 방법과 PRS을 이용하여 프로토타입을 추출하여 마할라노비스 거리(MD)로 비유사도를 측정하는 제안 방법의 식별 성능을 비교한다. 본 논문의 구성은 다음과 같다. 먼저, 제 Ⅱ장과 제 Ⅲ장에서 RDA와 PRS를 간단히 소개하고, 제 Ⅳ장에서는 실험 데이터와 실험 방법 및 실험결과를 설명한다. 그리고 제 Ⅴ장에서 결론을 맺는다.

Ⅱ. RDA의 개요

RDA는 입력패턴을 패턴의 특징대신에 학습패턴에서 추출한 프로토타입들과의 비유사도 조합을 특징으로 이용하는 식별법이다[3]~[5]. n개의 p차원 벡터로 이루어진 학습패턴을 $T=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}\in R^p$이라 한다. 여기서 T는 클래스를 미리 알고 있는 데이터 집합으로, T를 c개의 부분집합 $T_1,\cdots,T_c$로 나누면 $T=\cup_{k=1}^{c}T_k$가 되고, $T_i\cap T_j=\Phi,\ \forall i\neq j$가 된다. 여기서 문제는 T를 이용하여 새로운 입력패턴 $\boldsymbol{z}$를 식별하는 식별기를 설계하는 일이다. 먼저, 각각의 부분집합

$$T_i=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n_i}\},\ n=\sum_{i=1}^{c}n_i \tag{1}$$

에서 추출한 프로토타입 벡터 집합을

$$Y_i=\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{m_i}\},\ m=\sum_{i=1}^{c}m_i \tag{2}$$

이라 한다. 여기서 두 벡터 $\boldsymbol{x}_i$, $\boldsymbol{y}_j$의 비유사도를 $d(\boldsymbol{x}_i,\boldsymbol{y}_j)$라고 하면, 두 패턴집합 T와 Y의 비유사도로

이루어지는 행렬 $D(T, Y)$는 $n \times m$차원으로, 이 행렬의 열 벡터(column vector)는

$$(d(\boldsymbol{x}_i, \boldsymbol{y}_1), d(\boldsymbol{x}_i, \boldsymbol{y}_2), \cdots, d(\boldsymbol{x}_i, \boldsymbol{y}_m))^T,\ 1 \leq i \leq n \tag{3}$$

이 된다. 즉, $D(\boldsymbol{x}, Y)$는 벡터 x와 프로토타입 집합 Y와 비유사도를 측정한 값의 배열이 된다. 여기서 식 (3)의 열 벡터를 비유사도 벡터라 정의하고, $d(\boldsymbol{x})$로 표기한다. 벡터 $d(\boldsymbol{x})$는 p차원 특징 공간의 벡터 $\boldsymbol{x}$를 m차원 유사도 공간으로 사상시킨 벡터이다. 또한, 비유사도 공간은 특징 공간과는 무관한 새로운 공간으로, p가 매우 큰 고차원 응용의 경우 $m \ll p$이 되도록 m을 선택하면 희소성 문제를 해결할 수 있는 범위로 벡터차원을 축소할 수 있다. 따라서 RDA란, 특징공간의 p차원 벡터 $\boldsymbol{x}$ 대신에 비유사도 공간의 m차원벡터 $d(\boldsymbol{x})$를 대상으로 식별기를 학습하는 방법이다. 이 때, 학습패턴은 $\{d(\boldsymbol{x}_i)\}_{i=1}^n$이 되고, 식별규칙으로 k-NN 식별기를 비롯한 선형 식별기, 2차 식별기 등을 이용할 수 있다. 또, 식별할 벡터는 $d(\boldsymbol{z})$가 된다.

Ⅲ. RDA의 성능을 향상시키는 방법

통계적 패턴인식에서 이용하는 특징벡터 대신에 프로토타입과의 비유사도에 기반하여 클래스를 결정하는 RDA의 식별성능은 프로토타입의 선택방법과 비유사도를 측정하는 방법에 따라 결정된다. 이 장에서는 PRS를 이용하여 프로토타입을 추출한 다음, 마할라노비스 거리에 의한 상관행렬로 RDA의 성능을 향상시키는 방법을 설명한다.

가. PRS에 의한 프로토타입의 추출

PRS란 식별성능을 희생시키지 않는 범위내에서 주어진 학습패턴의 수를 최소한의 것으로 줄이는 방법이다. 지금까지 매우 다양한 PRS가 개발되어 있으며, 최근에 벤취마크 데이터를 대상으로 이들을 실험하여 PRS의 성능을 유형별로 비교하였다[9]. 각종 PRS에 대한 자세한 설명은 잘 알려져 있는 문헌[10][11]을 참조할 수 있다. 여기서는 간단한 예를 들어 프로토타입을 추출하는 과정을 설명한다.

그림 1은 7개의 벡터로 구성된 두 클래스 1차원 샘플 집합에서 3개의 프로토타입을 추출하는 과정이다. 여기서, 전체 7개의 샘플 중 $\{A, B, C, F, G\}$의 5개 샘플

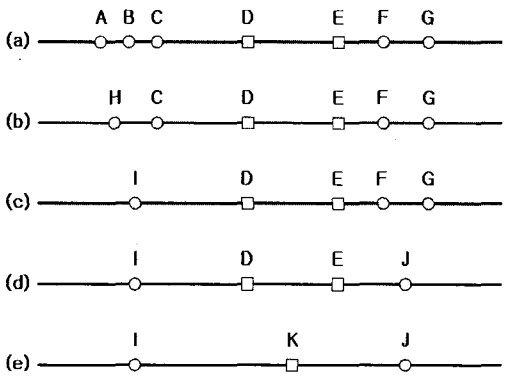

그림 1. 데이터집합에서 3개 프로토타입을 추출하는 예
Fig. 1. Example of extracting 3 prototypes from a dataset.

은 같은 클래스의 샘플로 원으로 표시하였고, $\{D, E\}$의 2개 샘플은 또 다른 클래스의 샘플로 정사각형으로 표기하였다.

먼저, 그림 1(a)에서 두 샘플 A와 B를 평균하여 하나의 샘플 H로 한다(그림 (b) 참조). 이 결과에서 다시 H와 C를 평균하면 I가 된다(그림 (c) 참조). 같은 방법으로, 샘플 F와 G를 J로 줄일 수 있고(그림 (d) 참조), D와 E를 K로 나타낼 수 있다(그림 (e) 참조). 따라서, 샘플 A, B, C를 I로, D와 E를 K로, 그리고 F와 G를 J로 표시하여 7개 샘플 벡터를 대표하는 3개의 프로토타입을 추출할 수 있다.

PRS는 크게 프로토타입을 생성시키는 PRS(Creative PRS)와 주어진 패턴들 중에서 선택하는 PRS(Selective PRS)로 나눌 수 있다. Selective PRS는 주어진 패턴구조를 대표할 수 있는 소수의 샘플패턴을 학습패턴중에서 선택하는 방법인 반면, Creative PRS는 소수의 샘플패턴을 새롭게 생성하는 방법이다.

Selective PRS와 Creative PRS의 추출성능은 적용하는 문제의 특성 (차원 및 샘플 수 등)과 알고리즘의 파라미터 (초기 코드북 및 학습 파라미터)의 설정 등에 따라 다르다. 문헌[9]에 따르면, 고차원 응용에서 PNN이 다른 PRS에 비해 우수한 반면, 저차원에서는 HYB가 우수하다. 따라서 본 논문에서는 저차원 및 고차원 학습패턴에 대한 RDA 성능을 고찰하기 위하여 PNN과 HYB를 이용하여 프로토타입을 추출한다. 또한, Creative PRS와의 비교를 위하여 Selective PRS인 CNN을 실험에 포함시킨다.

나. 마할라노비스 거리에 의한 비유사도 측정

비유사도 공간에서 학습한 RDA 식별기의 성능은 비

유사도 측정방법 및 프로토타입 추출방법에 따라 민감하게 변한다. 두 패턴 $\boldsymbol{x}$와 $\boldsymbol{y}$의 비유사도 $d(\boldsymbol{x},\boldsymbol{y})$를 측정하는 방법으로 유클리드 거리 ED, 평균 제곱오차 (mean squre error), 해밍 거리 (Hamming distance), 마할라노비스 거리 MD 등을 이용할 수 있다[1]. 두 패턴 $\boldsymbol{x}=(x_1,x_2,\cdots,x_m)^T$, $\boldsymbol{y}=(y_1,y_2,\cdots,y_m)^T$에 대한 유클리드 거리 (ED)는

$$d_{ED}(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{k=1}^{m}|x_k-y_k|^2} \tag{4}$$

이고, 패턴 $\boldsymbol{x}$와 분산이 Σ인 패턴 $\boldsymbol{y}$와의 마할라노비스 거리 (MD)는

$$d_{MD}(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}-\boldsymbol{y})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{y}) \tag{5}$$

이다. 여기서, Σ^{-1}는 공분산 행렬 Σ의 역행렬이다.

식 (4), (5)에서 d_{ED}는 두 패턴 벡터 사이의 거리를 나타내며, d_{MD}는 한 포인트에서 어떤 분포를 이루고 있는 군집 중심까지의 거리나 또는 두 분포의 중심간 거리를 나타낸다. 따라서 두 분포의 분산구조를 변경시킬 경우, 두 중심의 유클리드 거리는 동일한 값을 갖는 반면, 마할라노비스 거리는 다른 값을 갖게 된다. 예를 들어, 두개의 1차원 정규분포 $N(\mu_1,\sigma_1)$, $N(\mu_2,\sigma_2)$에서 분산이 서로 같으면 ($\sigma_1=\sigma_2$), 유클리드 거리와 마할라노비스 거리는 $d_{ED}(\mu_1,\mu_2)=d_{MD}(\mu_1,\mu_2)$가 되지만, 두 패턴이 속한 분산이 서로 다를 경우($\sigma_1\neq\sigma_2$)에는 $d_{ED}(\mu_1,\mu_2)\neq d_{MD}(\mu_1,\mu_2)$가 된다. 따라서, 두 패턴 $\boldsymbol{x}$와 $\boldsymbol{y}$의 상관성을 나타내는 비유사도 측정법으로 두 점사이의 기하거리만을 나타내는 d_{ED}보다는 분산구조를 고려하는 d_{MD}가 효과적이다.

Ⅳ. 실 험

가. 실험 데이터

본 논문에서 제안한 방법을 평가하기 위해 다양한 실험을 하였다. 실험 데이터로는, 시각적 관찰을 위해 2차원 평면에서 균일 분포로 랜덤하게 생성한 Random 인공 데이터와 8-차원 가우시안(Gaussian) 분포를 혼합하여 생성한 Non_normal2[9](이 데이터에 대한 자세한 설명은 문헌[1][9][12]를 참조할 수 있으며, 여기서는 간단히 Non_n2로 표기한다)를 이용하였다. 또한, 실-생활 벤취마크 데이터로 Iris2 (Iris 데이터에서 Versicolor와 Virginica 만을 취하여 구성한 데이터[9]), Ionosphere, Sonar, Arrhythmia 및 Adult4 등을 UCL machine learning database[12]에서 내려 받아 이용하였다. 이 때, 모든 데이터를 랜덤하게 두 부분집합으로 나누어 하나는 학습 데이터로 다른 하나는 테스트 데이터로 이용하였다. 또한, 모든 패턴 벡터는 −1부터 +1 사이의 실수로 정규화하였다. 본 실험에서 이용한 데이터별 패턴 벡터의 수, 특징 차원 및 클래스 수는 표 1과 같다.

표 1. 실험을 위해 사용한 벤취마크 데이터 집합
Table 1. The benchmark data sets for experiments.

Dataset Types	Dataset Names	Total Patterns (Training, Test)	# of Features	# of Classes
Artificial	Random	400(200,200)	2	2
	Non_n2	1000(500,500)	8	2
Real-life	Iris2	100(50,50)	4	2
	Ionosphere	351(176,175)	34	2
	Sonar	208(104,104)	60	2
	Arrhythmia	452(226,226)	279	16
	Adult4	8336(4168,4168)	14	2

기존의 RDA 식별기에서는 비유사도 상관관계를 나타내기 위하여 유클리드 거리 d_{ED}를 이용하고, 학습패턴을 대표하는 부분집합을 추출하기 위하여 프로토타입을 임의로 랜덤하게 선택하는 방법 (앞으로, 이 방법을 RAND로 표기한다) 등을 이용하였다. 본 논문의 제안방법에서는 d_{ED} 대신에 패턴집합의 분포특성을 고려할 수 있는 마할라노비스 거리 d_{MD}를 이용하였다. 또한, 학습 패턴을 대표하는 부분집합을 추출하기 위하여 Selective PRS인 CNN과 Creative PRS인 PNN, HYB 등으로 실험하여 기존의 방법과 식별성능을 비교하였다. RAND 방법으로 프로토타입을 선정할 때 프로토타입의 수 m은 CNN, PNN, HYB로 선정한 프로토타입의 수와 동일하게 하였다.

나. 실험 결과

인공 데이터와 실-생활 벤취마크 데이터를 대상으로 기존의 유클리드 거리와 마할라노비스 거리를 이용한 RDA의 식별 성능을 실험하였다.

먼저, CNN, PNN, HYB법으로 실험 데이터에서 프로토타입을 추출하였다. 그림 2(a)와 같은 Random 학습 데이터에서 CNN, PNN, HYB법으로 추출한 프로토타입은 각각 그림 2(b), (c), (d)와 같다. 여기서, CNN 및

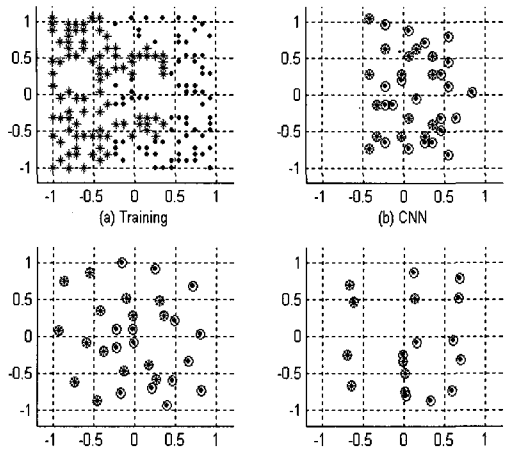

그림 2. 학습데이터 "Random"을 생성하여 CNN, PNN, HYB법으로 프로토타입을 선정한 결과

Fig. 2. Training dataset "Random" and its prototype vectors obtained with CNN, PNN, and HYB.

PNN법의 경우에는 각각 문헌 [6]과 [7]의 알고리즘을 이용하였다. 그리고 HYB에서는 SVM^{light}[13]를 이용하여 초기 벡터를 선택한 다음, LVQ3 타입의 알고리즘으로 최적의 프로토타입을 학습하였다.

기타, 인공 데이터와 실-생활 벤취마크 데이터를 대상으로 프로토타입을 추출한 실험 결과는 표 2와 같다. 앞에서 설명한 바와 같이, 학습할 벡터 차원을 축소시키기 위해서는 적은 수의 프로토타입이 바람직하다. CNN, PNN, HYB중에서 Random, Non_n2, Iris2, Adult4의 데이터에 대해서는 HYB가 가장 적은 수의 프로토타입을 추출하였고, Ionosphere, Arrhythmia에서는 PNN이 가장 효율적으로 프로토타입을 추출하였다.

추출한 프로토타입을 대상으로 유클리드 거리와 마할라노비스 거리를 이용하여 비유사도를 측정한 다음, 비유사도 공간에서 RDA의 식별 성능을 실험하였다. 이 때 식별기는 NN 식별기 (k-NN에서 k=1로 한 식별기)로 실험하였다. 일곱가지 인공데이터 및 실-생활 벤취마크 데이터를 대상으로 식별한 실험 결과는 표 3, 표 4 및 표 5와 같다. 여기서, 표 3은 일곱 개의 실험 데이터를 대상으로 특징공간에서 CNN, PNN, HYB 방법으로 프로토타입을 추출한 다음, NN 식별기로 식별한 평균 인식률(%)이다. 즉, 표 3은 전통적인 방법으로 수행한, 특징공간 식별기의 인식률에 해당한다. 반면에, 표 4와 표 5는 비유사도 식별기로 실험 데이터를 식별한 평균 인식률(%)이다. 단, 표 4는 기존의 방법대로 프로토타입을 랜덤하게 선택하였을 경우의 인식률이고, 표 5는 CNN, PNN, HYB 방법으로 프로토타입을 추출한 경우의 인식률이다. 이 때, 표 4의 RAND_1, RAND_2, RAND_3에서 선택한 프로토타입의 수는 각각 CNN, PNN, HYB에서 추출한 수와 같도록하였다.

먼저, 표 4와 표 5의 평균 인식률을 비교해보면, (Non_n2 데이터를 제외한 모든 종류의 데이터에 대해) 랜덤하게 프로토타입을 추출하는 기존의 방법에 비해

표 2. CNN, PNN, HYB법을 이용하여 실험 데이터 집합으로부터 추출한 프로토타입 벡터의 수

Table 2. The number of prototype vectors extracted from experimental datasets using CNN, PNN, and HYB.

Dataset Types	Dataset Names	CNN	PNN	HYB
Artificial	Random	36, 30	30, 25	18, 15
	Non_n2	64, 66	56, 380	63, 57
Real-life	Iris2	15, 12	10, 7	6, 8
	Ionosphere	51, 42	37, 33	44, 46
	Sonar	52, 53	34, 33	53, 59
	Arrhythmia	32, 28	8, 7	65, 69
	Adult4	755, 752	659, 658	430, 448

표 3. 실험데이터에 대한 CNN, PNN, HYB를 이용한 특징-기반 식별기의 인식률(%)

Table 3. The classification accuracy rates (%) of the feature-based classifiers with the CNN, PNN, and HYB on the experimental datasets.

Dataset Types	Dataset Names	ORG	CNN	PNN	HYB
Artificial	Random	96.50	96.20	95.75	89.50
	Non_n2	92.50	91.90	92.10	94.00
Real-life	Iris2	92.00	89.00	94.00	94.00
	Ionosphere	78.69	81.82	82.67	83.24
	Sonar	82.21	79.81	82.69	80.77
	Arrhythmia	97.56	96.47	99.12	99.12
	Adult4	93.40	91.58	89.36	92.78

표 4. 실험 데이터에 대한 랜덤선택 방법을 이용한 기존의 비유사도-기반 식별기의 인식률(%)

Table 4. The classification accuracy rates(%) of the conventional dissimilarity-based classifiers with the random selections on the experimental datasets.

Dataset Names	RAND_1		RAND_2		RAND_3	
	ED	MD	ED	MD	ED	MD
Random	82.90	84.58	81.33	84.30	82.75	85.40
Non_n2	95.14	94.50	95.08	94.92	95.07	94.99
Iris2	75.80	89.40	76.20	90.20	76.10	90.40
Ionosphere	69.80	86.14	69.46	86.05	70.09	86.14
Sonar	56.11	69.81	55.34	69.33	55.72	69.47
Arrhythmia	85.80	93.01	77.52	93.58	82.74	95.11
Adult4	73.16	18.29	73.14	18.27	72.69	18.42

표 5. 실험데이터에 대한 CNN, PNN, HYB를 이용한 비유사도-기반 식별기의 인식률(%)

Table 5. The classification accuracy rates(%) of the dissimilarity-based classifiers with the CNN, PNN, and HYB on the experimental datasets.

Dataset Names	CNN		PNN		HYB	
	ED	MD	ED	MD	ED	MD
Random	82.25	84.00	84.50	84.50	83.25	84.25
Non_n2	91.90	89.50	92.70	91.00	91.70	90.10
Iris2	77.00	90.00	73.00	88.00	78.00	88.00
Ionosphere	70.74	86.08	70.45	86.08	69.03	86.08
Sonar	59.62	70.19	57.69	69.23	59.13	69.71
Arrhythmia	89.82	92.92	85.84	92.92	78.76	95.13
Adult4	71.48	18.30	71.96	18.32	77.18	19.25

PRS를 이용하는 방법이 더 우수함을 알 수 있다. 여기서 랜덤하게 추출한 프로토타입의 수와 PRS로추출한 수는 같도록 하였다.

또한, 표 4와 표 5의 ED 및 MD란의 평균 인식률 (비유사도 측정방법에 따른 인식률)을 비교해보면, 마할라노비스 거리를 이용할 경우 식별 성능이 큰 폭으로 증가하는 것을 볼 수 있다. 예를 들어, 표 4에서 Ionosphere 데이터의 경우, 유클리드 거리를 이용한 경우의 인식률 (RAND_1, RAND_2, RAND_3의 ED 란의 평균 값)이 각각 69.80, 69.46, 70.09 (%)이었으나, 마할라노비스 거리를 이용하면 각각 86.14, 86.05, 86.14 (%)가 됨을 보이고 있다. 또한, 표 5에서도 인식률 (CNN, PNN, HYB의 ED 란의 값) 70.74, 70.45, 69.03 (%)가 각각 86.08, 86.08, 86.08 (%)로 향상되었음을 보인다.

그러나, Adult4 데이터의 경우 표 4와 표 5에서, 마할라노비스 거리를 이용한 경우 인식률이 큰 폭으로 떨어진 것을 보이고 있다. 이는 클래스 별 샘플 데이터의 불균형에 기인한 것으로 보인다. 즉, Arrhythmia 데이터의 경우 클래스 별 샘플데이터 수는 122와 104이고, CNN으로 추출한 프로토타입은 각각 16이다. 그러나 Adult4의 4167샘플 데이터는 클래스 별로 3961과 206이며, CNN으로 추출한 프로토타입은 각각 583과 172이다. 따라서 클래스 별로 추정한 공분산 행렬이 서로 다르고, 또 이를 이용하는 비유사도 벡터가 특징공간의 정보를 제대로 표현하지 못하기 때문인 것으로 보여진다.

끝으로, 표 3과 표 4, 표 5의 결과를 비교해 보면 전통적인 특징기반의 식별 법에 비하여 비유사도 기반의 RDA 식별이 다소 저조함을 알 수 있다. 이는 프로토타입을 적절하게 선정하고 비유사도 식별에 적합한 식별기를 설계하여 극복할 수 있을 것으로 보여지며, 앞으로의 과제이다. 그러나, Arrhythmia 데이터의 경우, 주어진 특징공간의 차원이 279인데 비하여 비유사도공간의 차원은 32 (=16+16)가 되어 효율적으로 차원을 축소할 수 있음을 알 수 있다.

이상에서 고찰한 실험 결과로부터, 비유사도 기반의 RDA 식별은 프로토타입의 수를 적절히 선택하는 방법으로 패턴의 차원을 축소할 수 있음을 알 수 있다. 이러한 사실로부터, RDA는 패턴 샘플의 수가 차원에 비하여 적을 경우에 발생하는 희소성 문제에 대처할 수 있는 식별법임을 알 수 있다. 특히, 비유사도를 측정하는 방법으로 마할라노비스 거리를 이용할 경우 식별 성능을 크게 향상시킬 수 있으며, 랜덤하게 프로토타입을 추출하기보다는 PRS를 이용하는 방법이 효율적임을 알 수 있다.

V. 결 론

본 논문에서는 PRS로 프로토타입을 추출한 다음, 마할라노비스 거리를 이용하여 비유사도 공간을 구축하는 RDA의 식별성능을 인공 데이터 및 실-생활 벤취마크 데이터를 대상으로 실험하였다. 실험결과, 학습패턴으로부터 임의로 프로토타입을 추출하는 기존의 방법에 비해 PRS를 이용하는 방법이 우수함을 확인하였다. 특히, 마할라노비스 거리를 이용하여 비유사도를 측정하는 방법으로 식별 성능을 크게 향상시킬 수 있었다. 이러한 인식률의 향상은 비유사도를 측정할 때 패턴의 분포 구조를 고려하였기 때문으로 사료되며, 이에 대한 이론적인 고찰은 앞으로의 연구과제이다. 또한, 프로토타입의 수를 적절히 선택하여 비유사도 공간의 차원을 축소하는 방법으로, 패턴 샘플의 수가 차원에 비하여 적을 경우에 발생하는 희소성 문제의 해결 가능성을 확인하였다.

끝으로, 본 논문의 연구에 조언해 준 캐나다 Carleton 대학 John Oommen교수와 네덜란드 Delft공대 Bob Duin교수, David Tax박사께 감사드린다.

참 고 문 헌

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition, Second Edition*, Academic Press, San Diego, 1990.

[2] J. Ye, R. Janardan, C. H. Park and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems", *IEEE Trans. Pattern Anal. and Machine Intell.*, vol.

PAMI-26, no. 8, pp. 982 - 994, Aug. 2004.
[3] R. P. W. Duin, E. Pekalska and D. de Ridder, "Relational discriminant analysis", *Pattern Recognition Letters*, vol. 20, pp. 1175 - 1181, 1999.
[4] E. Pekalska and R. P. W. Duin, "Dissimilarity representations allow for buiilding good classifiers, *Pattern Recognition Letters*, vol. 23, pp. 943 - 956, 2002.
[5] Y. Horikawa, "On properties of nearest neighbor classifiers for high-dimensional patterns in dissimilarity-based classification", *IEICE Trans. Information & Systems*, vol. J88-D-II, no. 4, pp. 813 - 817, Apr. 2005, in Japanese.
[6] P. E. Hart, "The condensed nearest neighbor rule", *EEE Trans. Inform. Theory*, vol. IT-14, pp. 515 - 516, May 1968.
[7] C. L. Chang, "Finding prototypes for nearest neighbor classifiers", *IEEE Trans. Computers*, vol. C-23, no. 11, pp. 1179 - 1184, Nov. 1974.
[8] S. -W. Kim and B. J. Oommen, "Enhancing prototype reduction schemes with LVQ3-type algorithms", *Pattern Recognition*, vol. 36, no. 5, pp. 1083 - 1093, 2003.
[9] S. -W. Kim and B. J. Oommen, "A taxonomy and ranking of creative prototype reduction schemes," *Pattern Analysis and Applications*, Springer-Verlag, vol. 6, no. 3, pp. 232 - 244, December 2003.
[10] B. V. Dasarathy, *Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press, Los Alamitos, 1991.
[11] J. C. Bezdek and L. I. Kuncheva, "Nearest prototype classifier designs: An experimental study", *International Journal of Intelligent Systems*, vol. 16, no. 12, pp. 1445 - 1473, 2001.
[12] http://www.ics.uci.edu/~mlearn/ MLRepository.html
[13] http://www-ai.cs.uni-dortmund.de/ SOFTWARE /SVM_LIGHT/svm_light.eng.html

저 자 소 개

김 상 운 (정회원)
1978년 한국항공대학 학사
1980년 연세대학교 대학원 석사
1988년 동 대학원 공학박사
1989년~현재 명지대학교 컴퓨터공학과 교수
<주관심분야 : 패턴인식, 미디어처리>