변형된 팩터 분석 모델을 이용한 생체데이타 분류 시스템

(Bio-data Classification using Modified Additive Factor Model)

조 민 국 † 박 혜 영 ††
(Minkook Cho) (Hyeyoung Park)

요 약 생체데이타 프로세싱이란 인간개체로부터 얻을 수 있는 고유의 생체 신호를 이용하여 다양한 목적으로 사용하는 것으로, 최근 이에 대한 요구가 높아지고 있다. 생체데이타는 도메인의 특성상, 클래스의 수는 많고 해당 클래스 내의 데이타는 상당히 제한적일 수 있어서 그만큼 데이타 내에 포함된 노이즈에 민감하게 된다. 따라서 기존의 패턴 인식과 분류 방법을 그대로 적용하여 개발된 시스템의 경우는 높은 일반화 성능을 기대하기 힘들다. 이를 해결하기 위해 본 논문에서는 생체데이타가 가지는 특성을 고려하여 각 클래스 고유의 특성에 영향을 미치는 클래스 요인과 노이즈와 같이 전체 데이타에 영향을 미치는 환경 요인으로 구성된 변형된 팩터 분석 모델로 생체데이타 생성 모델을 정의한다. 이를 바탕으로 분류에 필요한 데이타간 이격(inter-data discrepancy) 정보를 추출하고 새로운 유사도 함수를 정의하여 분류기에 적용한다. 제안하는 방법은 분류 대상이 되는 클래스의 정보 활용을 극대화 하여 적은 수의 데이타로부터 노이즈에 강인한 결과를 얻을 수 있다. 실제 생체데이타를 적용한 실험에서 제안하는 방법이 기존의 방법보다 우수한 분류 성능을 보임을 확인할 수 있었다.

키워드 : 생체데이타 프로세싱, 패턴 인식, 데이타 생성 모델, 팩터 분석 모델, 유사도 함수, 클래스 요인, 환경 요인

Abstract The bio-data processing is used for a suitable purpose with bio-signals, which are obtained from human individuals. Recently, there is increasing demand that the bio-data has been widely applied to various applications. However, it is often that the number of data within each class is limited and the number of classes is large due to the property of problem domain. Therefore, the conventional pattern recognition systems and classification methods are suffering form low generalization performance because the system using the lack of data is influenced by noises of that. To solve this problem, we propose a modified additive factor model for bio-data generation, with two factors; the class factor which affects properties of each individuals and the environment factor such as noises which affects all classes. We then develop a classification system through defining a new similarity function using the proposed model. The proposed method maximizes to use an information of the class classification. So, we can expect to obtain good generalization performances with robust noises from small number of datas for bio-data. Experimental results show that proposed method outperforms significantly conventional method with real bio-data.

Key words : bio-data processing, pattern recognition, data generation model, additive factor model, similarity function, class factor, environment factor

1. 서 론

생체데이타 프로세싱이란 인간개체로부터 얻을 수 있는 고유의 생체 신호를 이용하여 목적에 맞도록 사용하는 것이다. 최근 다양한 목적을 위해 이를 이용하고자 하는 요구가 높아지고 있다. 대표적인 예로는 생체인식, 생물 정보학, 그리고 의료정보시스템 등이 있다[1-3]. 이러한 요구에 부응하기 위해서는 생체데이타가 공통적으로 가지는 데이타 특성을 고려한 방법론의 개발이 중

· 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2006-311-D00807)
† 학생회원 : 경북대학교 전자전기컴퓨터학부
ucaresoft@paran.com
†† 종신회원 : 경북대학교 전자전기컴퓨터학부 교수
hypark@knu.ac.kr
(Corresponding author)
논문접수 : 2006년 10월 4일
심사완료 : 2007년 4월 17일

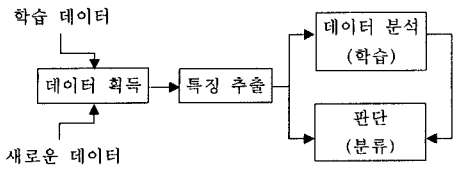


그림 1 생체데이타 프로세싱 시스템의 구조

요하다. 전형적인 생체데이타 프로세싱 시스템의 공통적인 처리의 흐름을 그림 1에 나타내었다. 이 시스템은 크게 데이타 획득, 특징 추출과 데이타 분석(학습), 그리고 판단(분류)의 네 가지 모듈로 나눌 수 있다.

데이타 획득 모듈은 인간개체로부터 얻을 수 있는 고유의 생체 신호를 이미지, 소리, 마이크로 어레이 등 목적에 맞는 방법으로 획득하는 부분이다. 특징 추출 모듈은 하나의 데이타를 입력받아 노이즈 제거 등의 전처리를 한 후, 판단에 중요한 역할을 하는 정보만을 추출해 내는 부분이다. 데이타 분석 모듈은 수집된 데이타 집합을 이용하여 그 분포 특성을 분석한다. 이를 바탕으로 새로운 데이타가 주어졌을 때 올바른 판단을 위한 기준을 설정하고 필요한 정보를 데이타베이스화하여 저장한다. 만약, 이 모듈에서 사용되는 데이타 특성을 고려하지 않는다면 올바른 판단 기준을 설정하지 못해 시스템의 성능이 저하될 것이다. 판단(분류)모델에서는 새로운 데이타가 주어진 경우 기존의 데이타 집합을 이용하여 설정된 기준과 정보를 바탕으로 적절한 판단을 내린다. 즉, 판단 기준이 몇 가지 카테고리로 나누어 진 경우, 그 중 어디에 속하는 지를 분류한다. 본 논문에서는 데이타 획득 및 특징추출 모듈은 데이타의 특성에 맞도록 획득되었다 가정하고 이후 데이타 분석과 분류 모듈에 대해 논의한다.

좋은 일반화 성능을 보이기 위해선 분류의 대상이 되는 생체 데이타의 특성을 먼저 분석해야 한다. 생체데이타가 가지는 중요한 특성으로는 크게 세 가지가 있다[4]. 첫째로, 생체데이타의 주된 사용 분야인 생체인식과 같은 문제에서는 분류 대상이 되는 클래스의 수가 많다. 많은 수의 클래스에 대해서 정확한 결정 경계를 찾기 위해선 높은 분류능력(capacity)을 가진 복잡한 분류기들이 사용되어야 한다. 하지만, 생체데이타의 두 번째 특성은 상반된 문제를 야기한다. 즉, 개개인으로부터 데이타를 획득하는 비용이 높아 클래스에 해당하는 데이타의 수가 현저히 적은 경우가 자주 있다. 이렇듯, 클래스별 데이타 수의 불균형은 해당 클래스의 분포 특성 추출을 어렵게 하고, 이는 곧 분류기의 성능을 저하시키는 요인이 된다. 마지막 특성으로 영상데이타, 마이크로 어레이 데이타와 같이 입력 데이타의 차원이 높은 경우가 많다. 높은 입력 차원은 입력 공간상에서 데이타의 희소성(sparsity)을 야기할 수 있고, 이는 통계적으로 안정된 결정 경계를 찾기 힘들게 한다.

기존의 연구에서는 이러한 데이타의 특성을 고려하지 않고 전통적인 패턴 분류 방법론을 그대로 적용해 왔다[5-8]. 즉, 데이타를 M개의 클래스로 분류하는 경우 데이타 분석(학습) 단계에서는 먼저 M개의 클래스 각각에 대한 분포 특성을 분석한다. 이때 통계적 추정방법이나 신경망과 같은 기계학습 방법을 이용하여 그 분포 특성을 결정한다. 이후 새로운 데이타가 주어지면, 분석된 결과를 이용하여 해당 데이타가 M개의 클래스 중 어디에 속하는지를 판단한다. 이렇게 개발된 시스템은 데이타의 분포를 얼마나 잘 분석해 내느냐에 따라 분류 성능이 좌우된다.

최근의 다중 클래스 분류를 위한 다른 방법으로 이진 분류기를 확장한 클래스 쌍(pairwise) 방법이 많이 연구되어 왔다[9,10]. 클래스 쌍 방법은 각각의 클래스가 쌍을 이루도록 분류기를 만들어 각 클래스를 분류하는 방법으로, 이진 분류기의 성능을 높이는데 초점이 맞추어지거나[9], 이진 분류기의 결과를 일반화 성능이 높아지도록 조합하는데 관심이 있다[10]. 하지만 이러한 방법은 각 클래스 내의 데이타 수가 제한적인 바이오 데이타의 특성을 반영하지는 못 한다. 즉, 각 클래스 쌍 분류기의 성능은 클래스 내의 데이타 수에 의존하므로, 분류해야할 클래스의 수는 많고 해당하는 클래스 내의 데이타 수가 적을 때, 분류기의 높은 일반화 성능을 기대하기 힘들다. 따라서 본 논문에서는 데이타 간의 유사성을 학습을 통해 추정하고, 이를 분류에 적용하는 방법으로 다중 클래스 분류를 시도한다.

다중 클래스 분류를 위해 데이타 간의 유사도 함수를 학습하는 접근 방법도 최근에 활발히 연구되고 있다[11-13]. 이러한 방법들은 기본적으로 같은 클래스 내의 데이타와는 유사하고, 서로 다른 클래스 내의 데이타와는 유사하지 않도록 하는 유사도 함수를 학습을 통해 찾아낸다. 대표적인 방법으로 데이타의 부분 공간으로의 사영과 유클리디안 거리를 이용하여 유사도 함수를 정의하고, 각 클래스 간의 마진을 최대화 하도록 유사도 함수를 학습하는 방법[11,12]이 있다. 다음으로 같은 클래스의 데이타 쌍에 대한 에너지는 줄이고, 다른 클래스의 데이타 쌍에 대한 에너지는 높이도록 손실 함수를 정의하여 컨볼루션 신경망으로 이를 학습하는 방법[13]이 있다.

본 논문에서 제안하는 방법도 마찬가지로 데이타 간의 유사성을 찾는 접근 방법이지만, 데이타에만 의존하여 유사도 함수를 학습하기 보다는, 먼저 대상이 되는 바이오 데이타의 특성을 고려한 데이타 생성 모델을 가정하고, 이에 기반으로 하여 클래스 내 왜곡 정보(y)의

분포를 추정하여 유사도 함수를 얻어낸다. 데이타 생성 모델에 기반한 이러한 접근 방법은 대상이 되는 데이타의 특성을 사전지식으로 활용함으로써 적은 수의 데이타에 대해서도 비교적 안정적인 성능을 기대할 수 있다. 또한 클래스 내 왜곡 정보를 추정(학습)함에 있어서도 다양한 방법을 적용할 수 있는 확장성을 가지고 있다.

한편, 데이타의 분포를 더 잘 분석하기 위해 데이타 생성 모델(Generative Model)을 이용하는 연구들이 존재한다. 대표적으로 주성분분석(Principal Component Analysis)이나 독립성분분석(Independent Component Analysis) 등을 비롯한 팩터 분석 모델(Factor Analysis Model)[14-17], 계층적 요인 모델(Hierarchical Factorial Model)[18-20]과 이선형 모델(Bilinear Model)[21] 등이 있다. 이중에서 이선형 모델은 데이타를 스타일 요소와 컨텐츠 요소로 구분하여 하나의 요소가 고정되었을 때 선형 모델과 동일하게 동작한다. 이것에 대한 변형으로 조명 효과에 의해 발생되는 노이즈에 강인하도록 만들어진 쿼션트 이미지 모델이 있다[22]. 이러한 모델들의 대부분이 클래스 정보를 활용하지 않고 입력 값만으로 데이타를 분석하는 비교사 학습(Unsupervised Learning)을 취한다. 이선형 모델이나 쿼션트 모델은 클래스 정보를 활용하지만 이러한 모델도 각 클래스에 해당하는 데이타의 수가 많지 않은 경우, 데이타 내에 포함되어 있는 노이즈에 영향을 많이 받게 되고 그만큼 신뢰성 있는 정보를 찾아내기가 어려워진다. 이는 기존 방법에 의해 개발된 시스템의 성능 저하의 주된 원인으로 작용한다고 볼 수 있다. 이를 해결하기 위해서는 데이타 특성을 고려한 새로운 학습 및 분류 방법론의 개발이 필요하다.

이에 본 논문에는 기존의 팩터 분석 모델에 각 클래스 고유의 특성에 영향을 미치는 클래스 요인과 노이즈와 같이 전체 데이타에 영향을 미치는 환경 요인의 개념을 넣어 변형된 데이타 생성 모델을 만들었다. 이어 본 논문의 핵심으로, 분류 대상이 되는 클래스의 정보 활용을 극대화 하여 적은 수의 데이타로부터 노이즈에 강인하도록 데이타의 분포 특성을 추출할 수 있는 방법을 제안한다. 추출된 데이타의 분포 특성을 이용하여 새롭게 유사도 함수를 정의하여 분류기를 설계한다. 유사도 함수의 정의를 위해서는 간단한 가우시안 분포를 가정한 방법과 SVM(Support Vector Machine)[23]을 이용한 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 데이타 생성 모델을 설명하고 이를 이용해 데이타의 분포 특성 추출을 위해 정의된 새로운 랜덤 변수에 대해 설명한다. 3장에서는 제안하는 유사도 함수 분류 방법에 대해 설명하고, 4장에서는 간단한 인공 데이타와 실제 생체데이타를 이용한 실험과 그에 대한 결과를 보이고, 마지막으로 결론이 뒤따른다.

2. 데이타 생성 모델

2.1 변형된 팩터 분석 모델

앞에서 언급했듯이, 본 논문에서는 데이타가 각 클래스 고유의 특성에 영향을 미치는 클래스 요인과 노이즈와 같이 전체 데이타에 영향을 미치는 환경 요인으로 구성된 데이타 생성 모델을 생각한다. 우리가 고정된 사물을 여러 번 촬영 했다고 가정하자. 같은 사물을 찍은 이미지이지만 그 각각이 일치하기는 힘들다. 이는 그 사물에 대한 유일하고 이상적인 이미지가 있지만, 여러 환경적인 요인들로서 상이해진다 볼 수 있다. 만약, 우리가 데이타로부터 이러한 유일하고 이상적인 이미지와 환경적인 요인들을 분리할 수 있다면, 보다 좋은 분류기를 만들 수 있는 것은 자명하다. 하지만, 데이타에서 이러한 요인들은 서로 상관관계를 가지고 있으므로 정확히 구분지을 수 없다. 그러므로 본 논문에서 데이타는 각 클래스 고유의 특성에 영향을 미치는 클래스 요인과 노이즈와 같이 전체 데이타에 영향을 미치는 환경 요인으로 구성된다고 가정하고 다음과 같이 함수로 표현한다.

$$x = f(\xi, \delta) \tag{1}$$

여기서 ξ는 데이타의 각 클래스의 특성과 관련된 요인을 나타내는 변수로 본 논문에서는 클래스 요인이라 칭한다. 한편 δ는 모든 클래스에 공통으로 가해지는 환경적 변화 요인을 나타내는 변수로 본 논문에서는 환경 요인으로 부른다. 관찰된 데이타 x를 생성해 내는 숨어있는 환경 요인과 클래스 요인을 적절히 분리해 낼 수 있다면 이를 이용하여 효과적인 분류기를 설계할 수 있을 것이다. 이러한 데이타 생성 모델을 만들기 위해서 함수 f를 정의해야 하고 실제 앞에서 언급한 다른 모델과 같이 다양한 함수 f를 정의할 수 있다.

본 논문에서는 함수 f는 다음과 같이 팩터 분석 모델[24]에 각 클래스 고유의 특성에 영향을 미치는 클래스 요인과 노이즈와 같이 전체 데이타에 영향을 미치는 환경 요인의 개념을 넣은 변형된 모델로 정의한다.

$$x_i = \xi_i W + \delta \tag{2}$$

여기서 x_i는 클래스 i에 속한 데이타이다. 즉, 각 클래스 별로 가산 팩터 모델을 만들지만 클래스 요인은 각 클래스별로 다르고 환경 요인은 모든 클래스에 영향을 준다. 본 논문에서는 이후 분류를 위한 정보추출을 용이하게 하기 위해 클래스 요인은 랜덤 변수가 아닌 고정된 파라미터 값으로 고정한다. 여기서 획득된 데이타 x가 1 × n 차원이라면, ξ_i은 각 클래스 고유의 성질의 나타내는 클래스 요인(1 × p), 그리고 δ은 전체 데

이타에 노이즈 영향을 미치는 환경 요소(1 × n)이다. 또 W(p × n)는 클래스 요인의 선형 변환으로, 본 논문에서는 이것을 모든 클래스에 동일하다고 가정한다. 만약, $x \in C_i$이면 클래스 요인은 ξ_i이 되고, 같은 클래스의 데이타들은 환경 요인 δ에 의해 변동된다. 이를 그림으로 표현하면 그림 2와 같다.

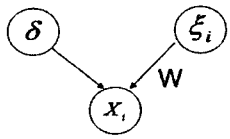


그림 2 클래스 i에 대한 데이타 생성 모델

본 논문에서 제안하는 모델은 기존의 혼합 팩터 분석 모델(Mixture Of Factor Analysis)과 다른 모델이다. 즉, 혼합 팩터 분석 모델에서는 공통의 팩터와 각 클래스별 특징을 나타내는 팩터 로딩 행렬을 가지는 것에 반해, 제안하는 모델은 클래스 유일한 특징을 나타내는 클래스 요인 ξ가 존재하고 이는 랜덤 변수가 아닌 고정된 파라미터 값으로 고정된다. 또한 W는 클래스 요인 ξ를 단순히 선형변환 시키는 것으로 모든 클래스에서 동일하다고 가정한다. 이는 앞서 설명한 바와 같이 분류의 대상으로 하고 있는 클래스들은 개개인으로부터 얻은 생체 신호로, 각 클래스들에서 그 원형(혹은 팩터들)으로부터 발생하는 변형은 개개인(클래스)의 특성에 의존하기보다 전체 클래스에 공통적인 신호 획득 환경(카메라 특성, 조명 등등)에 의존한다는 가정에 바탕을 둔다. 그러므로 생성된 같은 클래스 내의 데이타는 클래스 특징 팩터 ξ와 공통의 W에 의해 같은 데이타가 만들어지지만, 공통의 환경 변수 δ에 의해 변형되어 진다. 결국, 제안하는 모델은 기존의 혼합 팩터 분석 모델과 학습 방법이 다르고, 본 논문의 4장에서 기술한 것과 같이 실제 인공 데이타와 생체 데이타에 대한 실험에서 우수한 성능이 검증 되었다.

2.2 랜덤 변수의 정의

앞장에서 정의한 모델에 의하면 하나의 클래스내의 데이타들은 대표 벡터 ξ_i을 중심으로 하여 그로부터 흩어진 양상은 랜덤 변수 δ의 분포에 의해 결정된다고 볼 수 있다. 그리고 이 랜덤 변수는 전체 클래스에 영향을 미치는 환경 요인이므로, δ의 분포는 모든 클래스에서 같다고 생각할 수 있다. 그러므로 ξ와 δ의 관계를 적절히 활용하면 분류 대상이 되는 클래스의 정보 활용을 극대화 하여 적은 수의 데이타로부터 노이즈에 강인하도록 데이타 분포 특성을 추출할 수 있다. 이를 위해 새로운 랜덤 변수 y를 다음과 같이 정의한다. 즉, 두 데이타 x_i, x_j을 빼준 값을 새로운 랜덤 변수 y_{ij}로 정의한다. 이를 식으로 표현하면 다음과 같다.

$$\begin{aligned} y_{ij} &= x_i - x_j \\ &= \xi_i W + \delta' - (\xi_j W + \delta'') \\ &= (\xi_i - \xi_j) W + \delta' - \delta'' \end{aligned} \tag{3}$$

여기서 만약, i, j가 같은 클래스이면 우변의 첫 항은 소거되고 두 번째 항은 공통의 환경 요인 분포에서 기인한 확률 변수이므로 $\delta' - \delta''$만 남는다. i, j가 다른 클래스이면 y_{ij}은 첫 항이 소거되지 않아 같은 클래스의 경우의 y_{ij}의 분포와는 다를 것이다. 그리고 클래스의 수가 K개이고 각 클래스 내에 데이타 수가 N개라면 우리는 총 KN(N-1)/2개의 y_{ij}에 대한 샘플을 획득할 수 있다. 이러한 y의 분포 특성을 이용하여 새로운 유사도 함수를 정의할 수 있고, 이를 이용한 분류 시스템을 개발할 수 있다. 즉, 본 논문에서는 제안하는 모델을 직접 학습하여 분류 시스템을 개발하는 것이 아니라, 새로운 랜덤 변수 y의 분포 학습을 통해 추정하여 두 데이타 x_i, x_j의 유사도 함수를 얻어내고, 이를 이용하여 분류 시스템을 개발하는 것이다. 제안하는 방법에서 y의 분포를 학습하기 위해서는, 첫 번째로 y가 가우시안 분포를 따른다고 가정한 경우와 두 번째로 y의 분포 함수에 대한 가정을 별도로 하지 않은 경우가 있는데, 이는 3장에서 기술하겠다.

3. 분류 시스템

3.1 제안하는 분류 시스템

그림 3은 개발하고자 하는 새로운 모델의 처리과정이다. 우선, 데이타 분석(학습) 단계에서는 주어진 데이타 집합에서 두 개씩 데이타를 쌍을 지운다. 즉, 각 클래스 별로 같은 클래스 내의 데이타끼리 쌍($(x, x') \in C_S$)을 지우고, 또 다른 클래스의 데이타끼리도 쌍($(x, x') \in C_D$)을 지워서 새로운 랜덤 데이타(y)를 생성한다. 그러므로 이때 만들어진 데이타 y는 같은 클래스에서 온 데이타 쌍으로 이루어진 그룹(C_S)과 서로 다른 클래스 데이타로 구성된 데이타 쌍으로 이루어진 그룹(C_D)으로 분류될 수 있다. 이때 클래스의 수가 K이고 각 클래스내의 데이타 수가 N이라면, 만들어지는 데이타 y의 수는 K × N × (N-1)/2개가 된다. 데이타 분석(학습) 단계에서는 같은 클래스에서 온 데이타 쌍으로 이루어진 그룹(C_S)의 분포와 서로 다른 클래스에서 온 데이타로 구성된 그룹(C_D)의 분포를 분석함으로써 두 데이타 쌍의 유사도를 측정하는 함수 $S(y) = S(x, x')$을 구한다. 이때 얻어지는 유사도 함수 $S(x, x')$는 x와 x'의 물리적 거리를 나타내기보다, x와 x'이 같은 클래스에 속하는 경우에 높은 값을 가지도록 학습한다.

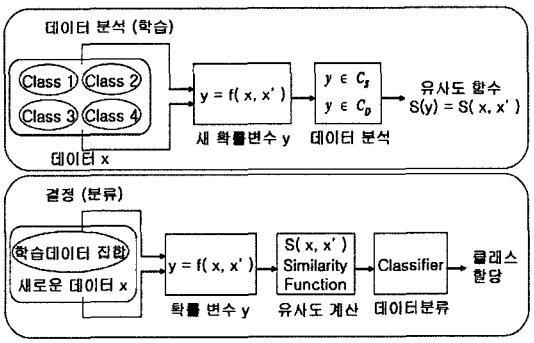


그림 3 제안하는 생체 데이타 처리 모델

다음으로 결정(분류) 단계이다. 새로운 데이타가 입력될 때, 이 데이타를 학습 단계에서 사용된 데이타와 쌍을 지워 유사도 계산을 위한 데이타 y를 생성한다. 학습에 사용된 데이타의 수가 K × N개 이므로 새로운 입력 데이타와 쌍을 지어 만들어지는 새로운 데이타 y는 K × N개가 된다. 이 데이타들에 대해, 학습 단계에서 만들어진 유사도를 계산한다. 얻어진 유사도 값을 이용하여 K-근접이웃 방법을 비롯한 적절한 분류 방법을 이용하여 새로운 입력 데이타가 어떤 클래스에 속하는지를 분류할 수 있다. 다음 장에서는 제안하는 분류 시스템을 사용하기 위해 유사도 함수를 어떻게 학습을 통해 얻어 내는지를 설명하고 데이타 분류 방법에 대해 기술하겠다.

3.2 유사도 함수의 정의

3.1에서 가장 중요한 특징은 주어진 입력(랜덤 데이타) x를 그대로 사용하지 않고, 두 데이타의 쌍 (x_i, x_j)으로부터 정의되는 새로운 랜덤 변수 y_{ij}를 사용하는 것이다. 이 랜덤 변수 y_{ij}의 분포 특성을 분석함으로써 두 입력 x_i와 x_j 사이의 유사도를 계산하는 유사도 함수를 찾는다. 이때 y_{ij}의 분포 특성은 x_i와 x_j가 같은 클래스로 부터 나온 경우 (즉, $y_{ij} \in C_S$)와 그렇지 않은 경우 (즉, $y_{ij} \in C_D$)로 구분될 수 있고, 유사도 함수는 y_{ij}가 두 부류 중 어디에 속하느냐에 따라 그 값이 크게 차이가 나도록 정의되어야 할 것이다. 본 논문에서는 유사도 함수를 y_{ij}가 C_S에 속할 확률 값에 비례하도록 정의한다.

우선, 유사도 함수를 제안하기에 앞서 y의 확률 분포 p(y)를 추정해야 한다. 이를 위해 매개변수에 의한 밀도 추정 방법(parametric density estimation method)를 이용한다[25]. 이때 새로운 랜덤 변수 y가 평균 $\mu = 0$이고 공분산 $\Sigma = I$인 표준정규분포를 따른다고 가정하면 이로부터 얻어지는 p(y)에 의한 유사도 함수는 대표적으로 사용되는 유클리디안 거리와 동일해 진다. 이를 본 논문에서는 S_E로 정하고 식으로 표현하면 다음과 같다.

$$S_E(x, x') = \|x - x'\|^2 \tag{4}$$

하지만 이 유사도 함수는 제약 사항이 강하기 때문에 일반화 성능이 좋은 분류 시스템을 만들기 힘들다. 이후 이 가정을 완화하여 가우시안 분포의 평균과 공분산을 추정하여 만드는 새로운 유사도 함수를 제안하겠다.

클래스내의 환경 요인은 여러 가지 요인이 통합되어 나타나는 것으로 볼 때, 중심 극한 정리에 기반하여 δ가 가우시안 분포를 따른다고 가정한다. 따라서 y 또한 가우시안 분포를 따르게 된다. 그러므로 p(y)를 추정하기 위해서는 파라미터인 평균과 분산을 추정하면 된다. 최대우도 추정(Maximum Likelihood Estimation)에 의해 우리는 쉽게 평균과 분산을 추정할 수 있다.

$$\hat{\mu}_y = \frac{1}{N(N-1)K} \sum_{i=1}^{N(N-1)K} y_i \,, \tag{5}$$

$$\hat{\Sigma}_y = \frac{1}{N(N-1)K} \sum_{i=1}^{N(N-1)K} (y - \hat{\mu}_y)(y - \hat{\mu}_y)^t \tag{6}$$

이제, x와 x'의 유사도는 아래 식과 같이 두 개의 데이타가 같은 클래스에 속할 확률로 정의한다.

$$S(x, x') = P\{x \in C_i \text{ and } x' \in C_j \,; i = j\} \tag{7}$$

그러므로 우리는 유사도 함수를 $y = x - x'$의 확률 함수로 재정의 할 수 있고 y가 가우시안 분포를 따르므로 우리는 새로운 유사도 함수 S_G를 아래와 같이 얻을 수 있다.

$$\begin{aligned} S_G(x, x') &= p(x - x') \\ &= \frac{1}{\sqrt{2\pi|\Sigma|}} exp\left\{-\frac{1}{2}(x - x' - \hat{\mu}_y)^t \hat{\Sigma}_y^{-1}(x - x' - \hat{\mu}_y)\right\} \\ &\propto 1/\left\{-\frac{1}{2}(x - x' - \hat{\mu}_y)^t \hat{\Sigma}_y^{-1}(x - x' - \hat{\mu}_y)\right\} \end{aligned} \tag{8}$$

여기서 y의 평균 공분산 행렬은 x의 것과 다르기 때문에 제안하는 유사도 함수는 마하라노비스(Mahalanobis) 거리[26]와 다르다.

지금까지 우리는 δ가 가우시안 분포를 따른다는 가정에서 y를 추정하고 클래스 분류를 하였다. 하지만 이러한 가정은 많은 문제에 대해서 적당하지 못 할 수도 있다. 이 점을 해결하기 위해 우리는 SVM을 이용해 유사도 함수를 직접적으로 추정하는 방법을 제안한다. 즉, 주어진 두 개의 x_i, x_j 데이타 쌍에 대해 우리는 SVM 입력 데이타를 $x_i - x_j$와 같이 정의하고 타겟 z값을 다음과 같이 정의한다.

z = 1 만약, x_i, x_j이 같은 클래스일 때

z = -1 그 외의 경우 (9)

이를 이용하여 학습 데이타 집합으로부터 모든 가능한 쌍이 학습 되어진다. 학습한 후, SVM의 타겟 값 O_{SVM}이 같은 클래스의 데이타 쌍을 입력으로 할 땐 양

의 값을, 그 외의 경우에는 음의 값을 가질 것이다. 그러므로 O_{SVM}은 다음의 새로운 유사도 함수로 고려 할 수 있다.

$$S_V(\boldsymbol{x}_i, \boldsymbol{x}_j) = O_{SVM}(\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (10)$$

이러한 유사도 함수와 여러 분류 방법을 이용하여 우리는 새로운 데이타 $\boldsymbol{x}_{new}$를 분류할 수 있다. 또한 이밖에도 다양한 형태의 $\boldsymbol{y}$를 생성하는 함수 f와 유사도 함수 S의 정의가 가능하므로 여러 가지 형태로 확장이 가능할 것이다.

3.3 분류 방법

분류 단계에서는 새롭게 주어진 데이타에 대해, 학습시에 사용된 데이타들과의 유사도를 각각 계산하고, 이를 바탕으로 새로운 데이타가 어떤 클래스에 속하는 지를 결정한다. 본 논문에서는 앞의 유사도 함수 값을 이용하여 기존의 잘 알려져 있는 근접 이웃방법(Nearest Neighbor Method)을 적용한 분류 방법을 사용한다.

근접 이웃방법에서는 $\boldsymbol{x}_{new}$와 모든 데이타 $\boldsymbol{x}_i$사이에서 유사도를 계산하여 $\boldsymbol{x}_{new}$와 제일 근접한 데이타 $\boldsymbol{x}_{nn}$을 찾는 것이고 이를 식으로 표현하면 다음과 같다.

$$\boldsymbol{x}_{nn} = \arg max\{S(\boldsymbol{x}_{new}, \boldsymbol{x}_i) \mid \boldsymbol{x}_i \in X\} \quad (11)$$

K-근접 이웃방법에서는 모든 학습 데이타에 대해 유사도를 계산하고 그 중 가장 높은 유사도를 가진 K개의 이웃들을 찾고 찾아진 K개의 후보들의 소속 클래스에 보팅을 한다. 이때 가장 많은 보팅 값을 획득한 클래스에 $\boldsymbol{x}_{new}$를 할당한다. 표 1은 이를 알고리즘으로 나타내고 있다.

표 1 K-근접 이웃을 이용한 분류 방법

1. 학습 데이타 $\boldsymbol{x}_i (1 \le i \le n)$에 대해 $S(\boldsymbol{x}_{new}, \boldsymbol{x}_i)$를 계산.
2. 학습 데이타 $\boldsymbol{x}_i (i = 1,2,...n)$ 중 가장 높은 유사도를 가진 K개의 데이타
   $\boldsymbol{X}_{\max} = \{\boldsymbol{x}_{\max_1}, \cdots, \boldsymbol{x}_{\max_K}\}$를 찾기.
3. $i = 1, \ldots, K$ 에 대해 다음 n_i를 계산.
   $n_i =$ 집합 $\boldsymbol{X}_{\max}$에 속하는 C_i의 원소의 수.
4. n_i의 값이 가장 큰 클래스에 $\boldsymbol{x}_{new}$를 할당.
   $\boldsymbol{x}_{new} \in C_i$, 만약 $n_i = \max\{n_1, \ldots, n_k\}$

본 논문에서 사용한 근접 이웃 방법대신에 신경망 등 보다 일반화 성능이 우수한 분류 방법을 사용할 수 있다. 즉, 이전의 학습 데이타들의 유사도 함수 값들을 신경망의 입력 값으로 하여 원하는 분류 결과를 얻을 수 있다.

4. 실 험

제안하는 방법의 유효성을 검증하기 위해 분류 시스템을 적용하여 여러 가지 데이타들에 대해 실험하였다. 제안하는 방법과 함께 기존의 SVM에 의한 다중 클래스 분류방법인 OVA(One Versus All)[27] 방법도 함께 실험하여 인식 성능을 비교하였다. SVM을 학습 시킬 때는 RBF 커널 함수를 사용하였고 각각의 파라미터는 실험을 통하여 최적화한 값을 사용하였다. 4.1절과 4.2절에서 실험한 Toy I, II의 데이타는 인공적으로 생성한 2차원 데이타이고, 4.3절과 4.4절에서는 각각 홍채와 사람 얼굴 등 실제 생체 데이타에 대한 실험을 하였다.

4.1 Toy I 데이타

Toy I 데이타는 2차원 데이타로 세 개의 클래스로 구성되어 있고, 각각의 클래스는 가우시안 분포를 따른다. 각 클래스별 평균은 다르나 분산은 동일하게 하여 학습 데이타로는 각 클래스별로 5개씩 생성하고 각 클래스에서 1000개의 데이타를 이용해 테스트 데이타로 사용하였다. 그림 4는 Toy I 데이타의 테스트 데이타를 나타내었다. Toy I 데이타는 δ의 분포가 가우시안이라는 제안하는 방법의 초기 가정에 정확하게 부합된다.

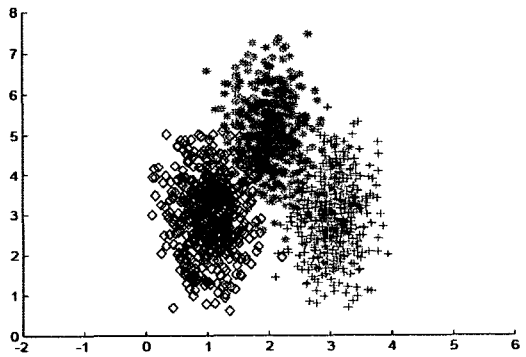
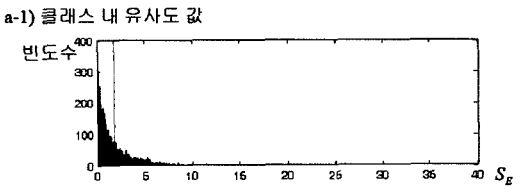
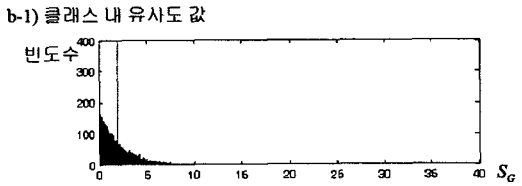


그림 4 실험 데이타 I : 각 클래스 당 하나의 클러스트

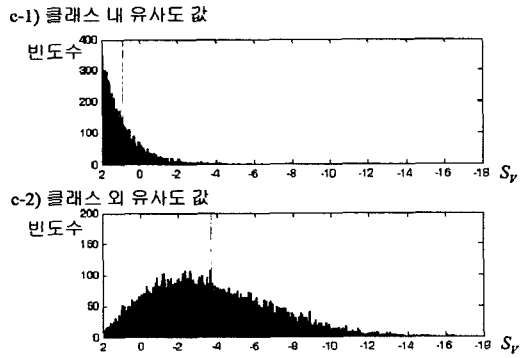
먼저 학습 데이타를 사용하여 앞장에서 정의한 각 유사도 함수(S_E, S_G, S_V)를 이용하여 각 데이타 쌍의 유사도 값을 계산할 수 있다. 만약, 두 데이타 쌍이 같은 클래스로 부터 선택된 데이타이면, 이 유사도 값이 높을 것이고 그렇지 않으면 유사도 값이 낮을 것이다. 좋은 유사도 함수란 같은 클래스의 데이타 쌍으로부터 얻어지는 유사도 값의 분포와 서로 다른 클래스의 데이타 쌍으로부터 얻어지는 유사도 값의 분포의 차를 크게 하는 것으로 볼 수 있다. 여기서 한 가지 유의할 점은 유사도 값은 3.2절의 유사도 함수의 정의에 따라 S_E, S_G은 값이 낮을수록 유사도가 높고 S_V은 값이 높을수록 유사도가 높다. 유사도 함수의 성능을 비교하기 위해 각 유사도 함수를 이용하여, 같은 클래스에 속하는 학습 데이타와 테스트 데이타와의 유사도 값의 분포를 히스토그램으로 표현하고(그림 5의 a-1, b-1, c-1) 마찬가지로



(a) S_E



(b) S_G



(c) S_V

그림 5 Toy I 데이타에 대한 클래스 내, 외 각 방법별 유사도 함수

서로 다른 클래스에 속하는 학습 데이타와 테스트 데이타 쌍의 유사도 값의 분포를 히스토그램으로 표현하였다(그림 5의 a-2, b-2, c-2).

그림 5는 각 유사도 함수별로 유사도 값의 두 그룹 - 같은 클래스의 데이타 쌍으로부터 얻어진 그룹과 다른 클래스의 데이타 쌍으로부터 얻어진 그룹 - 의 분포를 히스토그램으로 나타낸 것이다. a-1, b-1, c-1은 같은 클래스 내의 유사도 함수 값을 나타낸 것이고, a-2, b-2, c-2은 서로 다른 클래스 간의 유사도 함수 값을 나타낸 것이다. 각 히스토그램에서 분포의 평균값에 선을 그어 표시하였다. 여기서 각 유사도 함수별로 두 유사도 값의 그룹의 분포가 상당부분 겹치고 있다. 이는 그림 4의 데이타 분포에서 알 수 있듯이 같은 클래스의 데이타 쌍이지만 다른 클래스의 인접한 데이타 쌍보다 유사도 값이 낮을 수 있기 때문이다. 하지만, 실제 클래스를 분류할 때에는 가장 높은 유사도 값을 가지는 쌍만을 이용하므로, 두 유사도 값의 그룹의 분포가 상당부분 겹친다고 해서 그만큼 인식률이 나빠지는 것은 아니다.

각 유사도 함수별 성능 차이를 알아보기 위해, 두 그룹의 유사도 값의 평균 차에 두 그룹의 유사도 값의 표준편차 합을 나누어준 값을 다음 식과 같이 계산하였다.

$$\text{그룹간 유사도 분포 차} = \frac{\text{클래스 내 유사도 값 평균} - \text{클래스 외 유사도 값 평균}}{\text{클래스 내 유사도 값 표준편차} + \text{클래스 외 유사도 값 표준편차}} \quad (11)$$

이 값은 두 그룹의 유사도 값의 표준편차를 고려한 유사도 값의 평균차로 두 그룹간의 유사도 분포 차를 설명한다. 그림 6의 (a)는 이 값을 히스토그램으로 나타낸 것이다. (b)는 각 유사도 함수를 이용한 분류율을 히스토그램으로 나타낸 것이다. 이때 분류는 K-근접이웃 방법으로 클래스를 분류하였고, 여기서 K의 값은 1이며 최적화된 값이다.

그림 6의 (a)에서 알 수 있듯이, S_G의 클래스 내, 외의 유사도 분포의 차가 S_E의 것보다 더 큰 차이를 보이고 있다. 이는 환경 요인을 나타내는 확률변수 δ가 가우시안 분포를 가지고 있어 S_G의 기본 가정에 부합되기 때문으로 볼 수 있다. 그러므로 제안하는 방법 S_G가 두 그룹간 유사도 분포의 차를 더 크게 함을 확인할 수 있다. 또한 S_V의 방법도 S_G과 같이 좋은 성능을 보이는데 이는 SVM 자체가 좋은 유사도 함수를 학습을 통해 찾아내기 때문으로 해석할 수 있다. 이상과 같은 유사도 함수의 성능이 인식률에 그대로 반영되어 제안하는 두 유사도 함수 S_G와 S_V를 사용한 경우가 높은 인식률을 보임을 (b)에서 확인할 수 있다. 마지막으로 기존의 SVM에 의한 다중 클래스 분류방법인 OVA 방법과 비교했을 때도 제안하는 방법이 더 좋은 성능을 보임을 알 수 있다. 이는 OVA 방법이 적은 수의 데이타에 효과적이지 못한 반면, 제안하는 방법은 클래스의 정보 활용을 극대화 하여 적은 수의 데이타로부터 노이즈에 강인하도록 데이타 분포 특성을 추출할 수 있었기 때문으로 해석할 수 있다.

4.2 Toy II 데이타

실험 데이타 I과 마찬가지로 실험 데이타 II도 세 개

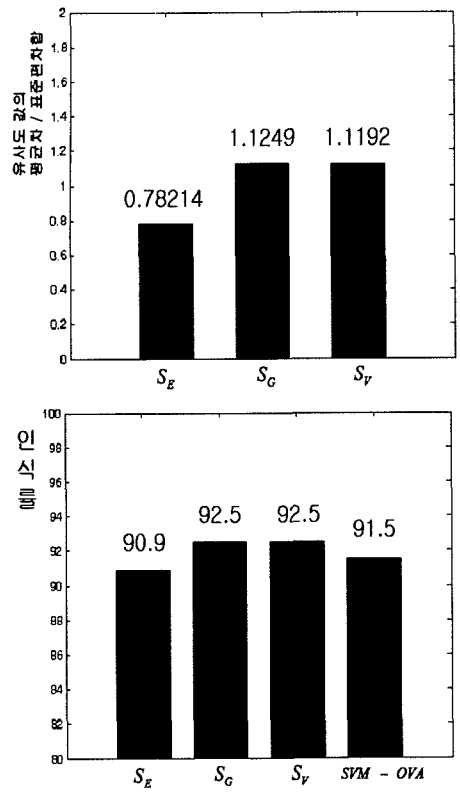


그림 6 (a) 각 방법별 유사도 분포 차, (b) 그에 따른 인식률

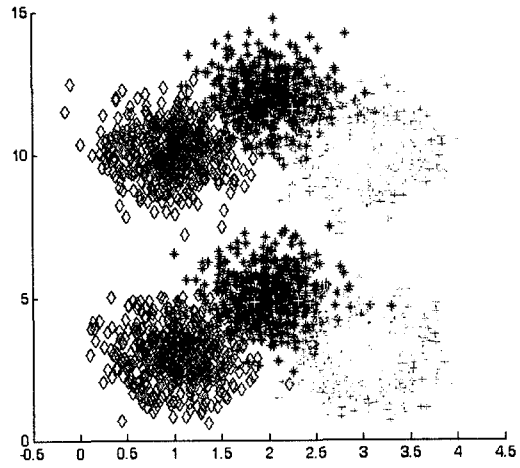
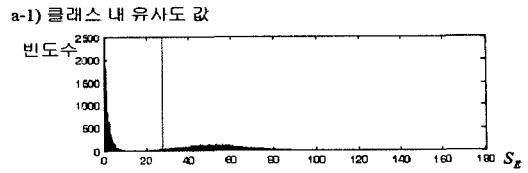
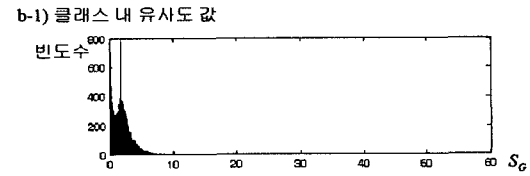


그림 7 실험 데이타 II : 각 클래스 당 두개의 클러스트

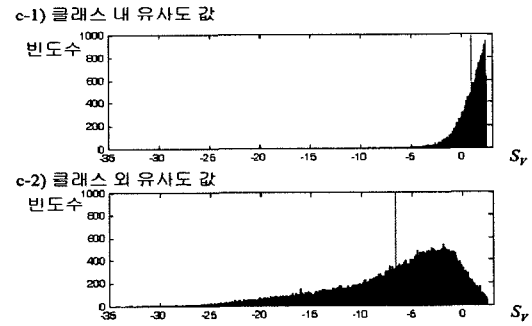
의 클래스로 구성되어 있다. 하지만 각 클래스의 데이타 분포가 두 개의 가우시안이 혼합 되어 클러스터를 이루고 있다. 이는 제안하는 방법에서 유사도 함수 S_G를 얻어낸 기본 가정, 즉 δ가 가우시안 분포를 이룬다는 가정에 위반되는 보다 일반적인 경우에 해당한다. 실험



(a) S_E



(b) S_G



(c) S_V

그림 8 Toy II 데이타에 대한 클래스 내, 외 각 방법별 유사도 함수

데이타 II는 각 클러스터에서 5개의 데이타, 즉 각 클래스당 10개의 데이타를 이용해 유사도 함수를 찾고, 각 클러스터당 500개의 데이타, 즉 각 클래스당 1000개를 이용해 테스트하였다. 그림 7는 테스트에 사용된 각 클래스당 1000개의 데이타를 나타내었다. 그림 7에서 같은 클래스는 가로 축에서 같은 위치에서 위 아래로 클러스터를 이루고 있다.

4.1장에서와 같이 각 유사도 함수 별 두 그룹의 유사도 값 분포를 그림 8과 같이 히스토그램으로 나타내었다. 그림 8의 a-1, b-1, c-1은 같은 클래스 내의 유사도

함수 값을 나타낸 것이고, a-2, b-2, c-2은 서로 다른 클래스 간의 유사도 함수 값을 나타낸 것이다. 각 히스토그램에서 분포의 평균값에 선을 그어 표시하였다. 그림 8에서 S_E, S_G의 클래스 내의 유사도 분포는 4.1에서와 다르게 두 개의 봉우리를 가지고 있다. 이는 그림 7의 데이타 분포에서 알 수 있듯이, 하나의 클래스가 두 개의 클러스터로 이루어져 있어 하나의 클러스터 내에서의 데이타 쌍의 유사도 값이 다른 클러스터의 데이타 쌍의 유사도 값보다 크기 때문이다. 또한 같은 클래스의 데이타 쌍이지만 다른 클래스의 인접한 데이타 쌍보다 유사도 값이 낮을 수 있기 때문에 두 그룹간 유사도 분포는 상당부분 겹쳐져 있다. 하지만, 4.1에서 언급했듯이 실제 클래스를 분류할 때에는 같은 클래스의 인접한 데이타 쌍과 다른 클래스의 인접한 데이타 쌍 중에서 보다 높은 유사도 값을 이용하므로, 두 유사도 값의 그룹의 분포가 상당 부분 겹친다고 해서 그만큼 인식률이 나빠지는 것은 아니다.

그림 9의 (a)은 그림 6의 (b)와 마찬가지로 유사도 함수별 두 그룹간의 유사도 분포 차를 알아보기 위해, 두 그룹의 유사도 값의 평균차에 두 그룹의 유사도 값의 표준편차 합을 나누어준 값을 히스토그램으로 나타낸 것이고, (b)는 그에 따른 분류율을 히스토그램으로 나타낸 것이다. 이때 분류는 K-근접이웃 방법으로 클래스를 분류하였고, 여기서 K의 값은 1이며 최적화된 값이다. 이 실험에서 각 클래스의 데이타 분포가 두 개의 가우시안이 혼합 되어 클러스터를 이루고 있으므로 δ이 가우시안 분포를 이룬다는 가정은 위반된다. 하지만 그림 9의 b)에서 알 수 있듯이, 제안하는 S_G의 방법이 S_E의 방법 보다는 다소 좋은 결과를 보이고 있다. 이는 제안하는 S_G의 방법이 초기 가정에 위배됨에도 불구하고 그림 9의 (a)에서 알 수 있듯이, 클러스터 된 데이타 분포의 분산을 줄여 S_E의 두 그룹간의 유사도 분포 차보다 크게 하기 때문으로 해석할 수 있다. 또한 학습을 통해 분포 특성을 찾아낸 S_V의 경우는 S_G보다는 향상된 결과를 보였다. 마지막으로 기존의 SVM에 의한 다중 클래스 분류방법인 OVA 방법은 S_G 보다는 좋은 성능을 보이지만, S_V보다는 성능이 좋지 못하여 역시 적은 수의 데이타 셋의 경우에 제안하는 방법이 기존의 SVM 보다 우수함을 확인할 수 있다.

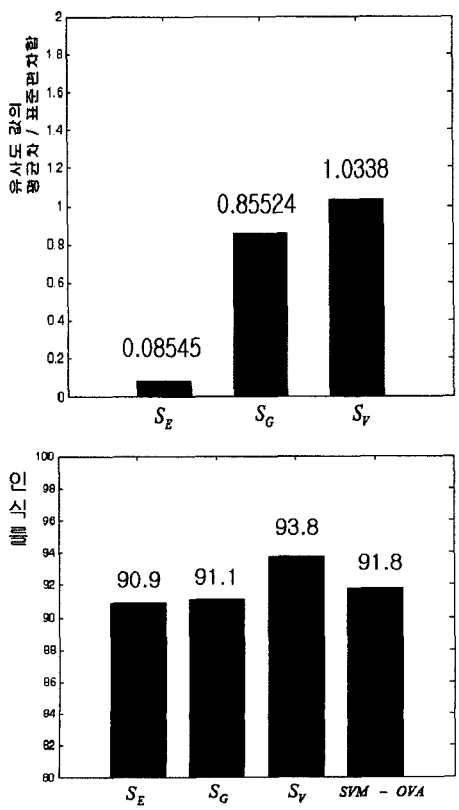


그림 9 (a) 각 방법별 유사도 분포 차, (b) 그에 따른 인식률

4.3 홍채 데이타

이어서 실제 생체 데이타들을 이용한 비교 실험을 수행하였다. 첫 번째로 홍채 이미지를 사용하였다. 그림 10에 이미지 샘플이 나타나 있다. 14명의 서로 다른 사람으로부터 얻어진 260개의 홍채 영상에 대해, 각 사람의 데이타로부터 무작위로 5개씩을 뽑아 70개를 학습 데이타로 사용하고 나머지 190개는 테스트 데이타로 사용하였다. 그림 10에 나타난 이미지는, 먼저 얼굴과 카메라 사이의 거리 변화에 따른 크기 변화를 보상하기 위해 국소화된 홍채 영역을 정규화하고, 홍채 영역이 표현된 극좌표를 직교 좌표로 변환함으로써 얻어진 것이다[28]. 이렇게 얻어진 7200화소(225×32)크기의 홍채 영상에 대해서 차원을 줄이기 위해 높은 차원의 데이타에 주로 사용되는 방법인 주성분분석 방법을 적용했다. 주성분 분석(PCA) 방법을 통해 최종적으로 각 이미지에 대해 70 차원의 특징 벡터를 획득하였다.

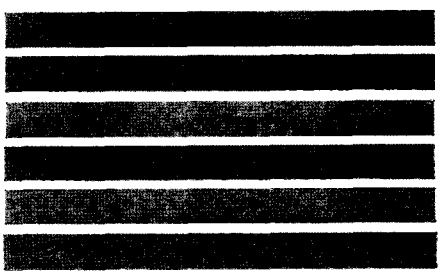
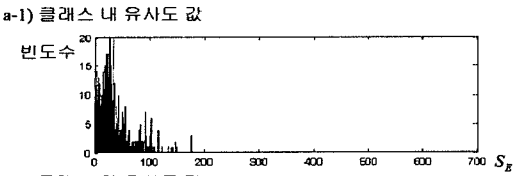
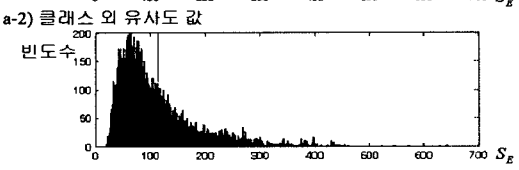
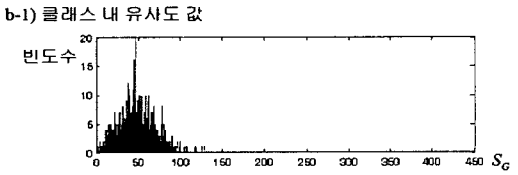


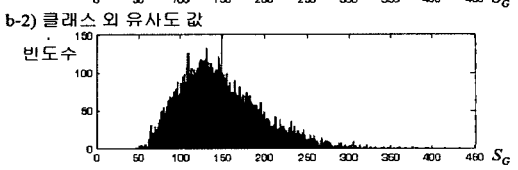
그림 10 사람 홍채 이미지



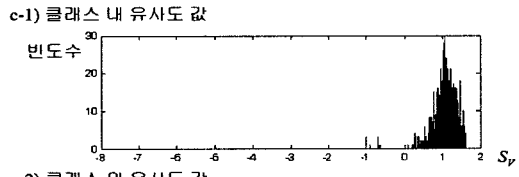


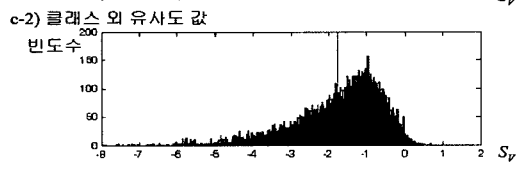
(a) S_E





(b) S_G



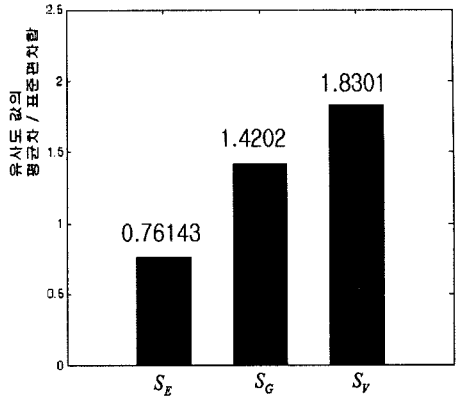


(c) S_V

그림 11 홍채 데이타에 대한 클래스 내, 외 각 방법별 유사도 함수

4.1에서와 같이 각 유사도 함수 별 두 그룹의 유사도 값 분포를 그림 11과 같이 히스토그램으로 나타내었다. 그림 11의 a-1, b-1, c-1은 같은 클래스 내의 유사도 함수 값을 나타낸 것이고, a-2, b-2, c-2은 서로 다른 클래스 간의 유사도 함수 값을 나타낸 것이다. 각 히스토그램에서 분포의 평균값에 선을 그어 표시하였다. 그림 11에서 S_E의 클래스 내, 외의 유사도 분포가 겹치는 것보다 S_G, S_V의 유사도 분포가 겹치는 것이 확연히 작은 것을 확인할 수 있다.

그림 12의 (a)는 그림 11를 이용하여 각 유사도 함수



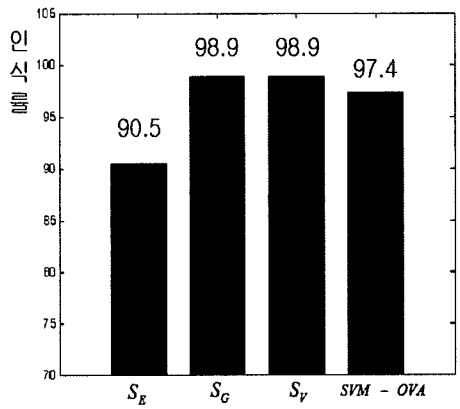


그림 12 (a) 각 방법별 유사도 분포 차, (b) 그에 따른 인식률

별 두 그룹간의 유사도 분포 차를 알아보기 위해, 두 그룹의 유사도 값의 평균차에 두 그룹의 유사도 값의 표준편차 합을 나누어준 값을 히스토그램으로 나타낸 것이고, (b)는 그에 따른 분류율을 히스토그램으로 나타낸 것이다. 그림 12의 (a)에서 알 수 있듯이, S_G, S_V의 클래스 내, 외의 유사도 분포의 차가 S_E의 것보다 확연히 큰 차이를 보이고 있다. 이로부터 홍채 데이타가 제안하는 데이타 생성 모델에 따르고, 뿐만 아니라 δ가 S_G를 얻는데 사용한 가우시안 분포라는 가정에도 어느 정도 부합됨을 추측할 수 있다. 이러한 유사도 분포의 차이를 이용하여 K-근접이웃 분류 방법으로 그림 12의 (b)와 같은 인식률을 얻었다. 여기서 K의 값은 1이며, 최적화된 값이다. 그림 12의 (b)에서 알 수 있듯이 S_E의 인식률 보다 S_G, S_V의 인식률이 훨씬 우수한 성능을 보이고 있다. 마지막으로 기존의 SVM에 의한 다중 클래스 분류방법인 OVA 방법 보다 제안하는 방법이 더 좋은 성능을 보이고 있어, 제안하는 방법이 실제 생체 데이타에도 성공적으로 적용됨을 확인할 수 있다.

4.4 얼굴 데이타

마지막으로 얼굴 데이타를 이용해 실제 생체 데이타

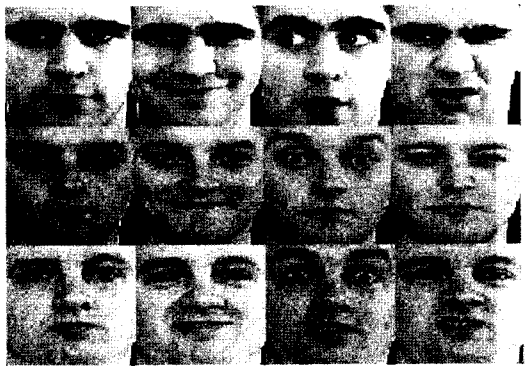
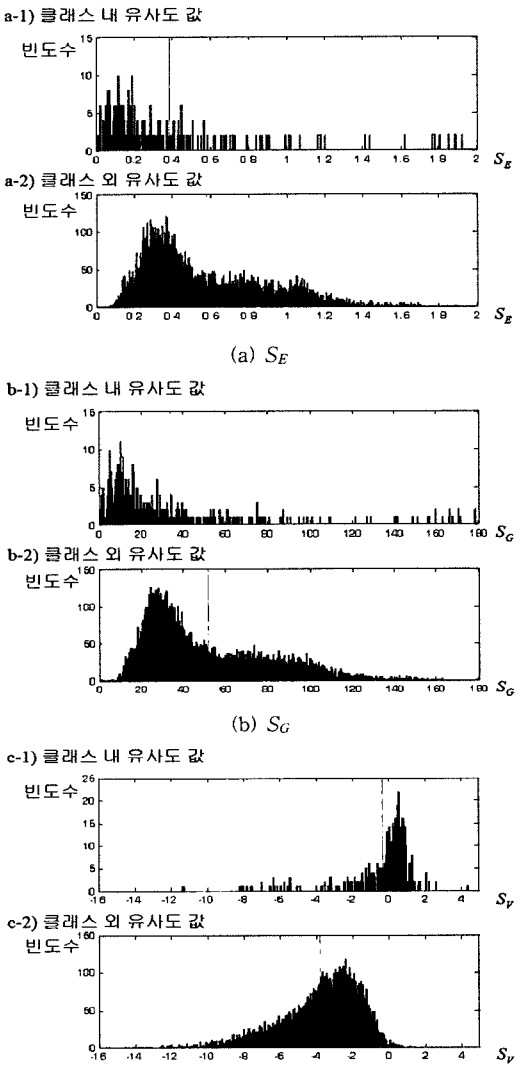


그림 13 얼굴 이미지

실험을 하였다. 얼굴 데이타는 PICS(Psychological Image Collection at Stirling) 홈페이지에서 획득하였고(http://pics.psych.stir.ac.uk/) 그림 13에 얼굴 이미지 샘플이 나타나 있다. 우리는 모두 30명의 사람으로부터 얼굴 이미지를 사용하는데, 각각의 사람은 4개의 다른 표정의 이미지로 구성되어 있다. 이 데이타는 전형적인 생체 데이타의 특성, 즉 클래스의 수는 많고 각 클래스에 속한 데이타의 양은 적다는 특성을 가지고 있다. 데이타의 수가 많지 않으므로 평균 성능을 얻기 위해, 4-폴드 크로스 밸리데이션을 수행하였다. 즉, 각각의 사람으로부터 3개의 이미지를 학습 시에 사용하였고, 나머지 하나를 테스트에 사용하는 것을 한 번의 실험으로 할 때, 테스트 데이타에 사용하는 데이타를 바꾸어 가며 모두 4번의 실험을 수행하여 평균 분류율을 얻었다. 이미지의 크기가 80 x 90이므로, 입력 차원이 7200이다. 이러한 각각의 이미지로부터 주성분 분석 방법을 통해 원래 데이타에 대해 50 차원의 특징 벡터를 획득하여 사용하였다.

4.1에서와 같이 각 유사도 함수 별 두 그룹의 유사도 값 분포를 그림 14와 같이 히스토그램으로 나타내었다. 그림 14의 a-1, b-1, c-1은 같은 클래스 내의 유사도 함수 값을 나타낸 것이고, a-2, b-2, c-2은 서로 다른 클래스 간의 유사도 함수 값을 나타낸 것이다. 각 히스토그램에서 분포의 평균값에 선을 그어 표시하였다. 여기서 각 유사도 함수별로 두 유사도 값의 그룹의 분포가 4.1절, 4.2절, 4.3절에 비해 상당부분 겹치고 있다. 이는 얼굴 표정의 변화에 의해 한 클래스 내의 변화가 결정되므로 δ가 가우시안 분포와는 많이 다른 형태를 가지기 때문으로 해석할 수 있다.

그림 15의 (a)는 두 그룹의 유사도 값의 평균차에 두 그룹의 유사도 값의 표준편차 합을 나누어준 값을 히스토그램으로 나타낸 것이고, (b)는 그에 따른 분류율을 히스토그램으로 나타낸 것이다. 이때 분류는 K-근접이웃 방법으로 클래스를 분류하였고, 여기서 K의 값은 1



(a) S_E

(b) S_G

(c) S_V

그림 14 얼굴 데이타에 대한 클래스 내, 외 각 방법별 유사도 함수

이며 최적화된 값이다. 그림 15의 (a)에서 알 수 있듯이, S_E와 S_G의 클래스 내, 외의 유사도 분포의 차보다 S_V의 유사도 분포의 차가 월등히 많이 나는 것을 확인할 수 있다. 이는 앞에서 언급한대로 세밀한 얼굴 표정에 의해 한 클래스 내의 변화량이 커지므로 δ이 가우시안 분포와는 많이 다른 분포를 가지는 경우에도 S_V는 SVM에 의해 그 분포 특성이 잘 학습되었기 때문으로 해석할 수 있다. 유사도 분포의 차이에 따라 그림 15의 (b)와 같은 인식률을 얻었다. 이때 각 밸리데이션의 K 값은 최적화된 값을 이용하였고, 각 벨리데이션 셋의 히

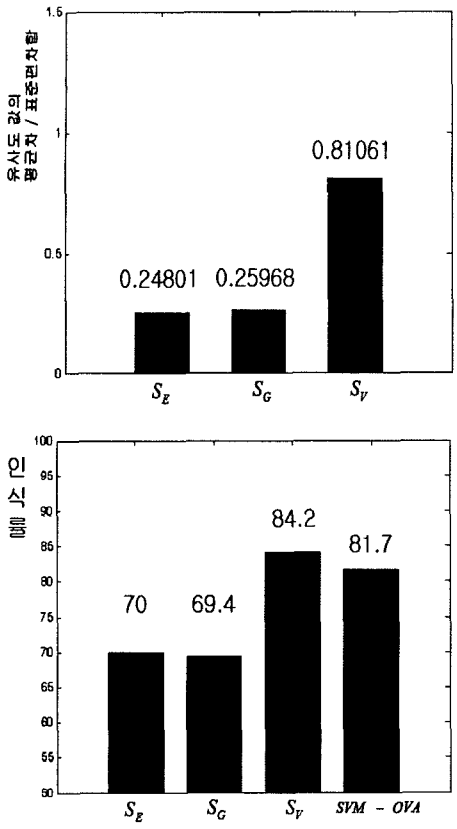


그림 15 (a) 각 방법별 유사도 분포 차, (b) 그에 따른 인식률

스토그램은 평균값을 나타내었다. 그림 15의 (b)에서 알 수 있듯이, S_E, S_G의 성능은 좋지 못 하다. 하지만 여전히 S_V가 전통적인 SVM의 OVA 방법보다도 현저히 좋은 성능을 보이고 있다. 이는 OVA 방법은 학습하는 데이타의 수가 너무 적어 학습을 제대로 하지 못하는데 반해, S_V은 제안하는 방법을 통해 보다 안정적인 통계 정보를 얻고, SVM이 좋은 유사도 함수를 찾아 주기 때문으로 생각할 수 있다.

4.5 실험 결과

실험 결과를 종합해 보면, 모두 네 가지 방법으로 각 데이타들에 대해 실험하였다. 우선, 단순히 유클리디안 거리를 유사도 함수 S_E로 하는 근접 이웃 방법, 가우시안 분포를 유사도 함수 S_G로 하는 방법, SVM을 유사도 함수 S_V로 하는 방법, 전통적으로 SVM을 OVA로 하는 방법이다. SVM을 학습 시킬 때는 RBF 커널 함수를 사용하였고 각각의 파라미터는 실험을 통하여 최적화한 값을 사용하였다.

아래 표 2는 위의 각 실험들을 정리한 것이다. 이 표에서 알 수 있듯이, 모든 데이타에 대해서 S_V의 방법이 제일 좋은 성능을 나타낸다. 이는 SVM 자체가 두 입력 x_i와 x_j사이의 유사도를 계산하는 유사도 함수를 제일 잘 찾는다고 생각할 수 있다. 그 외에 Toy I 데이타에 대해서 S_G가 S_V와 유사하게 좋은 성능을 보이는데 이는 데이타 δ가 가우시안 분포라는 가정에 부합되기 때문이다. 하지만, Toy II 데이타에 대해서는 이 가정이 맞지 않는다. 그러므로 S_G의 성능이 별로 좋지 못 하다. 하지만 여전히 S_E의 성능 보다는 좋은 성능을 보이고 있고 S_V 또한 좋은 성능을 보이고 있다. 다음으로 홍채 데이타의 경우, 제안하는 방법들이 S_E의 성능에 비해 현저하게 좋은 것을 알 수 있다. 이 경우 제안하는 두 가지 방법들이 모두 같은 성능을 보이는데, 이는 실제 데이타인 홍채 데이타에 대해서도 S_G에서 사용한 가우시안 가정이 적용될 수 있음을 보이고 있다. 마지막으로 얼굴 표정 데이타에 대해서는, S_E, S_G의 성능이 좋지 못 하다. 이는 얼굴 표정의 변화에 의해 만들어지는 클래스 내의 변화가 가우시안 분포와는 많이 다른 분포를 가지기 때문이라 생각할 수 있다. 하지만 이 경우에도 S_V는 좋은 유사도 함수를 찾아낼 수 있고, 전통적인 OVA 방법보다도 현저히 좋은 성능을 보이고 있다. 이는 OVA 방법은 학습하는 데이타의 수가 너무 적어 학습을 제대로 하지 못하는데 반해, S_V은 제안하는 방법을 통해 보다 안정적인 통계 정보를 얻어 낼 수 있음을 보인다.

표 2 각 실험 데이타에 대한 S_E, S_G, SVM-OVA, S_V 방법의 분류율

방법 \ 데이타	Toy I	Toy II	IRIS	FACE
S_E	90.9	90.9	90.5	70
S_G	**92.5**	91.1	**98.9**	69.4
S_V	**92.5**	**93.8**	**98.9**	**84.2**
SVM-OVA	91.5	91.8	97.4	81.7

5. 결 론

본 논문에서 생체인식을 위한 데이타 생성 모델을 제안하고, 이에 기반하여 새롭게 정의된 유사도 함수를 이용한 분류 방법을 제안하였다. 가장 기본적인 데이타 생성 모델인 팩터 분석 모델을 변형하여 데이타가 각 클래스 고유의 특성에 영향을 미치는 클래스 요인 ξ와 노이즈와 같이 전체 데이타에 영향을 미치는 환경 요인 δ으로 구성된 클래스 정보를 포함한 모델을 정의 하였다. 이 모델을 바탕으로 가우시안 분포를 유사도 함수 S_G로 하는 방법, SVM을 이용하여 유사도 함수 S_V를 학습하

는 방법에 대해 제안하였고, 성능 비교 검증을 위해 단순히 유클리디안 거리를 유사도 함수 S_E로 하는 근접 이웃 방법과 기존의 SVM을 OVA 방법에 대해 비교 실험하였다. 그 결과 S_E방법은 물론 기존의 SVM-OVA 방법보다 우수한 성능을 보였는데, 이는 제안하는 방법이 분류에 기준이 되는 정보를 안정적으로 추출하는 특징을 가지고 있기 때문이다. 즉, 제안하는 방법은 기존의 생성 모델에 클래스 정보를 넣어 학습하여 비교사 학습(Unsupervised Learning)보다 분류 시스템에 더 효과적인 정보를 이용하고, 다음으로 데이타 쌍으로부터 적은 수의 데이타에 대해서도 통계적으로 안정적인 정보를 추출하는 유사도 함수를 정의하여 사용함으로써 효율적인 분류가 가능하다.

본 논문에서 제안하는 방법은 데이타 생성 모델로 가장 기본적인 팩터 모델을 기반으로 하였다. 그러나 이는 2장에서 설명한 바와 같이 다양한 형태로 일반화가 가능하며, 이를 통해 각 클래스별로 하나의 대표 벡터를 가진다는 본 논문의 제약점을 완화하여 보다 복잡한 분류 문제에서도 좋은 성능을 내는 특징 추출 및 분류 모델로 발전시킬 수 있다. 추후 이에 관한 연구가 지속적으로 이루어질 것이다.

참 고 문 헌

[1] http://www.biometrics.org

[2] http://www.amia.org

[3] http://bioinformatics.org

[4] R. Rifkin and et al. "An Analytical Method for Multiclass Molecular Cancer Classification," SIAM Review, vol.45, issue 4, pp. 706-723, 2003.

[5] M. Bartlett, and T. Sejnowsky, "Viewpoint Invariant Face Recognition using Independent Component Analysis and Attractor Networks," Neural Information Proc. Systems - Natural and Synthetic, vol.9, pp. 817-823, 1997.

[6] T. S. Furey et al., "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, vol.16, pp. 906-914, 2000.

[7] R.P. Wildes, "Iris Recognition: An Emerging Biometric Technology," Proc. of the IEEE, vol.85, no.9, pp. 1348-1363, 1997.

[8] John D. Woodward, Jr., Nicholas M. Orlans, Peter T. Higgins, "BIOMETRICS," OSBORNE Press. 2003.

[9] Thomas Hofmann, Joachim M. Buhmann, "Pairwise Data Clustering by Deterministic Annealing," IEEE Trans on PAMI, vol.19, no.1, pp. 1-14, 1997.

[10] Johannes Fürnkranz, "Pairwise Classification as an Ensemble Technique," LNCS, vol.2430, pp. 97-110, 2002.

[11] Jacob Goldberger, Sam Roweis, Geoff Hinton, Ruslan Salakhutdinov, "Neighbourhood Components Analysis," Advances in Neural Information Processing Systems, vol.17, pp. 513-520, 2004.

[12] Kilian Q. Weinberger, John Blitzer, Lawrence K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," Advances in Neural Information Processing Systems, vol.18, pp. 1473-1480, 2005.

[13] Sumit Chopra, Raia Hadsell, Yann LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," Proc. of International Conference on Computer Vision on Pattern Recognition, pp. 539-546, 2005.

[14] Mardia, K., Kent, J., & Bibby, J., "Multivariate analysis. London," Academic Press. 1979.

[15] Bell, A., & Sejnowski, T., "An information-maximization approach to blind separation and blind deconvolution", Neural Computation, vol.7(6), pp. 1129-1159, 1995.

[16] Hinton, G.E., & Zemel, R., "Autoencoders, minimuum description length and Helmholtz free energy," In J. Cowan, G. Tesauro, and J. Alspector (Eds.), Advances in neural information processing systems, vol.6, pp. 3-10, San Mateo, CA: Morgan Kauffman, 1994.

[17] Ghahramani, Z., "Factorial learning and the EM algorithm," In G. Tesauro, D. Touretzky, and T. Leen (Eds), Advances in neural information processing systems Vol.7, pp. 617-624. Cambridge, MA: MIT Press, 1995.

[18] Hinton, G., Dayan, P., Frey, B., & Neal, R., "The wake-sleep algorithm for unsupervised neural networks," Science, vol.268, pp. 1158-1161, 1995.

[19] Dayan, P., Hinton, G., Neal, R., & Zemel, R. "The Helmholtz machine," Neural Computation, vol.7(5), pp. 889-904, 1995.

[20] Hinton, G., & Ghahramani, Z., "Generative models for discovering sparse distributed representations," Phil. Trans. Royal Soc. B, vol.352, pp. 1177-1190, 1997.

[21] Joshua B. Tenenbaum, William T. Freeman. "Separating Style and Content with Bilinear Models," Neural Computation, vol.12, pp. 1247-1283, 2000.

[22] Tammy Riklin-Raviv and Amnon Shashua, "The quotient image: class-based rerendering and recognition with varying illuminations," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.23, issue 2, pp. 129-139, 2001.

[23] J.G. Daugman, "High Confidence Visual Recognition of Persons by a Test of Statistical Independence," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.15(11), pp. 1148-1161, 1993.

[24] Gorsuch, Richard L., "Factor Analysis," Erlbaum, 1983.

[25] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2ed, Academic Press, 1990.

[26] Ethem Alpaydin, "Introduction to Machine Learning," MIT Press, 2004.
[27] Bernhard Schölkopf, Alexander J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)," MIT Press, 2001.
[28] Gyundo Kee, Kwanyong Lee, Hyeyoung Park, Yillbyung Lee, "A New Approach to Human Iris Recognition based on Statistical Information Theory," International Conference on Neural Information Processing, vol.1, pp. 134-139, 2000.

조 민 국

2005년 경북대학교 전자전기컴퓨터과학과를 졸업하고, 2007년 경북대학교 컴퓨터과학과에서 이학석사를 수여 받았다. 2007년 3월부터 현재까지 경북대학교 전자전기컴퓨터학과에서 박사학위 과정에 있다. 주요 관심 분야는 기계학습, 패턴인식 등이다.

박 혜 영

1994년 연세대학교 전산과학과를 졸업하고, 1996년 연세대 컴퓨터과학과에서 이학석사, 2000년 연세대 컴퓨터·산업시스템공학과에서 공학박사학위를 각각 수여하였다. 이후 2004년 2월까지 일본 이화학연구소 뇌과학연구센터(RIKEN BSI)에서 Research Scientist로 재직하였으며, 2004년 3월부터 현재까지 경북대학교 전자전기컴퓨터학부 조교수로 재직하고 있다. 주요 관심 분야는 기계학습, 패턴인식, 뇌과학 등이다.