IE Interfaces
Vol. 24, No. 2, pp. 97-104 , June 2011.

다변량 데이터의 분류 성능 향상을 위한 특질 추출 및 분류 기법을 통합한 신경망 알고리즘

윤현수 · 백준걸†

고려대학교 산업경영공학과

Feature Selecting and Classifying Integrated Neural Network Algorithm for Multi-variate Classification

Hyunsoo Yoon · Jun-Geol Baek†

School of Industrial Management Engineering, Korea University

Research for multi-variate classification has been studied through two kinds of procedures which are feature selection and classification. Feature Selection techniques have been applied to select important features and the other one has improved classification performances through classifier applications. In general, each technique has been independently studied, however consideration of the interaction between both procedures has not been widely explored which leads to a degraded performance. In this paper, through integrating these two procedures, classification performance can be improved. The proposed model takes advantage of KBANN (Knowledge-Based Artificial Neural Network) which uses prior knowledge to learn NN (Neural Network) as training information. Each NN learns characteristics of the Feature Selection and Classification techniques as training sets. The integrated NN can be learned again to modify features appropriately and enhance classification performance. This innovative technique is called ALBNN (Algorithm Learning-Based Neural Network). The experiments' results show improved performance in various classification problems.

Keyword: classification, feature selection, data mining, neural network, KBANN, multi-variate analysis

1. 서 론

분류 성능 향상은 다양한 분야에서 요구되는 중요한 이슈이다. 최근 들어 DNA Microarray 기술의 발달로 인하여 고차원의 특성을 갖는 방대한 양의 데이터를 통해 인체 기관이나 질병에 관련된 다양한 정보를 이용할 수 있다(Sarkar *et al.*, 2002). 따라서 환자의 상태에 영향을 미치는 특질들을 찾아내고 분류 성능을 향상시키기 위한 다변량 분석방법은 정확한 치료와 진단을 위해 많이 연구되고 있다. 또한 화학공정이나 반도체공정의 경우, 생산 과정의 여러 특성들 간의 상호작용에 의해 생산품의 품질이 결정된다. 따라서 제품의 품질에 영향을 미치는 여러 중요 특질들을 통계적인 기법으로 찾아내고 관리하는 방식은 불량 제품의 생산을 조기에 차단하여 경제적인 생산을 가능케 할 수 있다. 이와 같은 분석의 토대가 되는 다변량 데이터의 분류 문제에 관한 연구는 중요 변수를 찾아 차원을 축소하거나 특성의 변형을 통해 주요 특질들을 찾아내는 특질 추출(Feature

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0016510). 본 과제는 정부(지식경제부)의 지원을 받아 수행되었음(No. 10031812-2010-13). 이 논문은 2010년도 2단계 두뇌한국(BK)21 사업에 의하여 지원되었음.
† 연락저자 : 백준걸 교수, 136-701 서울시 성북구 안암동 5가 1번지 고려대학교 산업경영공학과
Fax : +82-2-929-5888, Tel : +82-2-3290-3396, E-mail : jungeol@korea.ac.kr
투고일(2010년 12월 15일), 심사일(1차 : 2011년 01월 06일), 게재확정일(2011년 01월 16일)

Selection) 절차와 선택된 특질들을 적합한 분류기(Classifier)를 통해 각 클래스별로 분류하는 절차, 두 단계로 이루어져 왔다. 본 연구에서는 신경망 알고리즘을 이용하여 두 가지 절차를 통합시킬 수 있는 새로운 방법을 제안하여 분류성능을 향상시키고자 한다.

기존의 특질 추출에 관한 연구에서는 세부적으로 클래스 정보를 학습에 사용하는지 여부에 따라 무감독 학습(Unsupervised Learning) 또는 감독 학습방식(Supervised Learning) 방식이 있다. 무감독 학습 방식에 주로 활용되고 있는 PCA(Principal Component Analysis) 또는 ICA(Independent Component Analysis)는 변수들의 특성을 고려하여 특질을 추출하고 원 정보의 큰 손실 없이 차원을 축소하는데 탁월하다. 하지만 클래스의 정보에 대한 고려가 없기 때문에 분류문제에 적용하기에 적합하지 않을 수 있다(Park *et al.*, 2009). 이에 반해서 감독방식은 클래스 정보와 변수간의 통계적인 유의성이나 변수들 간의 상관관계를 활용하여 변수를 선택하는 방식으로 무감독 방식의 문제점을 개선하였지만 변수들의 특성과 선택 방법에 따라 성능이 변동하고 분류기와의 상호작용을 고려하지 않는 단점이 있다(Saeys *et al.*, 2007). 감독 방식중의 하나인 Wrapper 방식은 분류기와의 상호작용까지도 고려하여 특질을 선택하는 방식이다. 이 방식은 주어진 특질 집합에서 전 방향(Forward) 또는 후 방향(Backward)으로 특질들을 순차적으로 추가하거나 제거해가면서 분류기에 적용시킨 성능을 통하여 주요 특질을 산출한다. 그러나 이 방식은 한번 선택되거나 제거된 변수는 계속 동일한 상태를 유지해야 하는 둥지 효과(Nesting Effect)가 발생하여 분류 성능의 저하를 유발할 수 있다(Pudil *et al.*, 1994; Kabir *et al.*, 2010).

분류기 관한 연구에서는 원 정보나 선택된 특질들을 이용하여 분류 성능을 향상시키기 위한 다양한 통계적인 기법들이 제안되어 왔다. 대표적인 SVM(Support Vector Machine) 분류기는 특질들의 Margin을 최대화하는 SV(Support Vector)를 찾아 집단을 구분하는 방식이다. 다차원 분류 문제에서 SVM 분류기는 차원이 축소된 선택된 특질공간에서 SV를 찾기 위한 기계학습(Machine Learning)을 한다. 따라서 특질 정보를 잘못 선택하게 되면 분류기의 성능이 저하된다. 그리고 인공신경망을 활용한 분류기에 관한 연구는 분류 모형에 적합한 복잡한 형태의 구조를 모델링하고 반복적인 학습을 통해 분류 성능을 향상시키는 방식이다. 하지만 인공신경망을 통한 분류기는 임의로 결정되는 초기 연결가중치로 인하여 신경망간의 복잡한 비선형적인 특성에 영향을 받아 지역 최소값(local minima)에 빠질 수 있다. 이런 한계점으로 인공신경망 분류기는 좋은 성능을 보장하지 못하고 결과 값의 변동성이 커지는 단점이 있다(Gori and Tesi, 1994).

따라서 본 연구에서는 앞서 지적한 기법들의 문제점을 해결하기 위해 KBANN(Knowledge Based Artificial Neural Network)방식이 갖는 장점을 활용한 ALBNN(Algorithm Learning Based Neural Network)을 제안한다. ALBNN은 기존 특질 추출 알고리즘과 분류기 자체의 특성을 신경망간 연결가중치를 통해 선행지식(Prior knowledge) 형태로 학습하고 이를 통합하여 재학습하는 방식이다. 통합과정에서 특질의 추가적인 변형을 통해 특질들과 분류기와의 상호작용, 변수들 간 상관관계 등을 분류문제에 적합하게 변형하여 분류 성능을 향상시킨다.

본 논문의 구성 내용은 다음과 같다. 제 2장에서는 본 논문에서 활용하는 신경망 학습 알고리즘인 Levenberg-Marquardt 기법과 본 연구에서 제안하는 ALBNN에 대해서 서술하였다. 제 3장에서는 실험계획 및 비교 알고리즘들의 특성과 실험결과 등을 기술하였으며, 마지막 제 4장에서는 결론을 기술하였다.

2. 본 론

2.1 인공신경망(Artificial Neural Network)

인공신경망(ANN)은 입력과 출력에 따라 행동을 결정하는 특성 때문에 패턴인식, 함수근사, 분류(Classification) 기법 등 다양한 분야에서 응용되어 왔다. 특히 컴퓨터의 성능 향상으로 빠른 학습의 구현이 용이해졌고, 복잡한 형태의 모델링이 가능하다는 강점을 지닌다. 인공신경망은 생물학적인 특성에서 영감을 얻어 만들어진 기법 중의 하나이며, 그 구조는 여러 개의 층(layers), 노드(nodes), 신경망간 연결 가중치로 구성된다.

본 연구에서 사용하는 신경망 층간 연결 방식은 전향(Feed-forward) 방식이다. 입력 패턴에 따라 각 노드에 대한 신경망 연결 가중치와 활성화 함수를 이용하여 출력 값을 산출하였다. 활성화 함수는 hyperbolic tangent sigmoid 함수를 식 (1)과 같이 사용하였다.

$$f(x) = \frac{2}{1+\exp^{-2x}} - 1 \tag{1}$$

신경망은 학습을 통해 복잡한 비선형의 특성을 지닌 함수의 추정에 장점을 지닌다. 또한 출력 노드(output node)의 특성에 따라 실수 값을 도출하는 회귀함수(regressor)의 기능을 할 수 있고, 정수 형태를 도출하는 분류기의 역할도 수행할 수 있다. 따라서 본 연구에서는 특질 추출 및 분류의 두 가지 단계가 가진 복잡한 특성을 추정하기 위해 신경망을 사용하였다.

다양한 학습 알고리즘이 신경망 학습에 사용되는데 일반적으로 사용되는 것이 역전파(Backpropagation) 알고리즘이다. 역전파 학습 방법은 출력오차를 활용하여 신경망간의 연결 가중치를 오차가 최소가 되도록 반복적으로 조정하는 방식이다. David Rumellhart(1995)에 의해 발전된 역전파 알고리즘은 우리가 통상적으로 어떤 것을 학습하는 것과 유사하다. 주어진 정보를 통하여 어떤 일을 처리했을 때 두뇌를 통해 추정한 값과 실제 결과 값이 상이한 경우, 산출된 값을 실제 결과 값과 비교하면서 오차를 줄이는 과정을 반복해서 찾아내는 방식이다.

2.2 LM(Levenberg-Marquardt) 알고리즘

본 연구에서 적용한 신경망 학습 방법은 다양한 역전파 알고리즘 중에서 비선형 최적화 문제를 해결하는데 뛰어난 성능을 보이는 LM 알고리즘을 사용하였다(Hagan and Menhaj, 1994). 기존의 역전파 학습 방법이 1차 도함수를 이용한 경사 하강만을 적용하여 오차를 최소화했다면, LM 알고리즘은 Gauss-Newton 방법을 통해 근사하는 방법을 적용한다. Gauss-Newton 방법은 오차함수를 테일러 전개(Taylor expansion)로 근사하고 2차 미분을 통하여 근사 최소점을 찾는 방식이다. LM 알고리즘에서 비중 변경방정식은 Gauss-Newton 방법의 변형을 통해 식 (2)와 같이 도출 된다. 1차 도함수(Gradient)와 헤시안(Hessian)행렬을 적용하여 신경망의 에러를 최소화하는 방법이다(Marquardt, 1963).

$$\triangle W = [J^T(W)J(W) + \mu I]^{-1} J^T(W)e(W) \qquad (2)$$

단, $\mu > 0$ and I : 단위 행렬
$J(W)$: Jacobian 행렬
$e(W)$: 출력 오차
W : 비중 vector

2.3 ALBNN(Algorithm Learning Based Neural Network)

Towell과 Shavlik(1994)가 제안한 KBANN 방식은 기존 도메인 지식을 인공신경망의 트레이닝 정보로 활용하여 학습 성능을 향상시키는 기계학습 방법이다. <Figure 1>은 신경망간 연결 가중치의 초기 값에 따라 학습 후 수렴하여 해가 찾아지는 영역 및 범위를 보여준다. 일반적인 신경망은 신경망간 연결 가중치가 임의로 주어지기 때문에 학습 후에 수렴되는 해 영역의 범위가 상대적으로 넓게 분포되고 수렴속도가 느린 특성을 지닌다. 반면, 기존 지식을 활용해 적절한 시작점을 선정하는

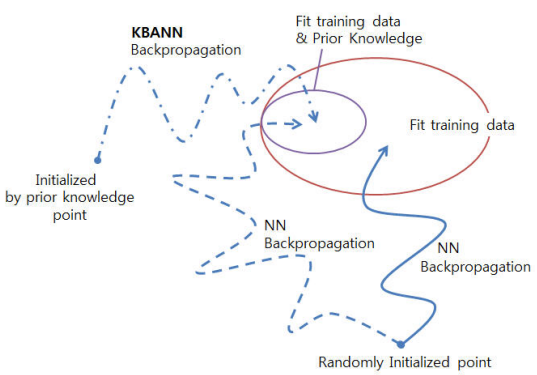


Figure 1. 초기 연결 가중치에 따른 수렴 영역

KBANN 방식은 해 공간의 범위가 훨씬 줄어들게 되고, 추정에 적합한 영역으로 수렴하게 하는 특징이 있다. 이러한 효과에 착안하여 본 논문에서 제안하는 ALBNN은 특질 추출 기법과 분류 알고리즘 자체의 특성을 도메인 지식으로 사용하고 학습을 통해 분류 성능을 향상 시키는 방식이다. 제안하는 알고리즘은 다음과 같은 세 단계 절차에 따라 모델을 구성한다.

(1단계) 특질 추출 알고리즘의 특성을 학습한 신경망 구축
(2단계) 분류기의 특성을 학습한 신경망 구축
(3단계) 두 개의 신경망을 결합한 통합 신경망 구축

각 단계별 자세한 설명은 다음과 같다.

(1) 1단계(특질 추출 학습 신경망 구축) : 주어진 전체 데이터를 특질 추출 알고리즘에 적용하고, 그 결과 값을 활용하여 <Figure 2>와 같이 신경망에 특질 추출 알고리즘 자체 특성을 학습을 통해 반영하는 단계이다. 먼저, 기존에 연구된 특질 추출 알고리즘 중 특질 추출 성능이 좋은 기법을 선택한 뒤, 이를 이용하여 원 자료의 정보(raw data)를 변환하고 중요한 몇 가지 특질들을 선택한다. 분류 문제를 해결하는 일반적인 방식은 이 선택된 특질들을 기존 분류기의 입력정보로 활용하고 상이한 집단을 구별하는 작업을 수행해 왔다. 하지만 제안 알고리즘에서는 원자료를 신경망 입력정보로 사용하고 선택된 특질들을 출력 정보로 하는 학습데이터 집합으로 활용한다. 신경망간의 연결 가중치 정보는 제 2.2절에서 소개한 LM 알고리즘을 통해 학습 오류가 목표 값 이하가 될 때까지 반복적으로 조정하여 학습한다. 이와 같은 방식으로 학습된 인공신경망에는 연결 가중치를 통해 특질 추출 알고리즘 자체가 가진 고유한 특징을 추정한 정보가 표현된다.

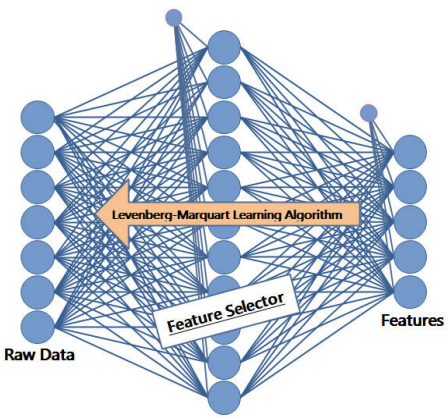


Figure 2. 특질 추출 학습 신경망 구축

(2) 2단계(분류기 학습 신경망 구축) : 선택된 특질과 분류기에 의해 분류된 클래스 정보를 활용하여 분류 기법의 특성 자체를 <Figure 3>과 같이 학습을 통해 추정하고 신경망에 반영하여 인공신경망이 분류기의 특성을 지니도록 하는 단계이다. 먼저, 분류 성능이 뛰어난 분류기법을 선택한다. 주어진 데이터의 특성(선택된 특질들과 클래스 정보)에 맞게 분류 알고리즘의 파라미터를 조정하여 분류 작업을 수행한다. 1단계에서 이용한 특질 추출 알고리즘에 의해 선택된 특질들을 입력정보로, 분류기에 의해 도출된 분류정보를 출력정보로 활용한 학

습 데이터 집합을 만든다. LM 학습 알고리즘을 통해 신경망의 학습 오류가 최소가 될 때까지 신경망간 연결 가중치를 조정한다. 이와 같은 절차를 통해 두 번째 신경망에는 기존 분류기 자체의 특성을 지닌 신경망을 구성할 수 있다.

(3) 3단계(두 개의 신경망을 결합한 통합 신경망) : 두 개의 신경망을 통합하여 분류 성능 향상을 위한 최종 모형을 만드는 단계이다. 동일한 특질 추출 알고리즘에 의해 선택된 특질들의 집합을 <Figure 2>의 특질 추출 학습 신경망에서는 출력정보로 사용하고 <Figure 3>의 분류기 학습 신경망에서는 입력정보로 사용하기 때문에 첫 번째 신경망의 출력 노드와 두 번째 신경망의 입력 노드의 개수가 동일하다. 따라서 신경망 연결 가중치 정보는 그대로 보유한 상태에서 <Figure 4>와 같이 선택된 특질들을 하나의 중간 은닉 층(Hidden Layer)으로 하는 신경망으로 결합할 수 있다. 결합된 신경망의 연결 가중치의 전면부에는 특질 추출기법의 특성이 반영되고, 후면부에는 분류기의 특성이 선행지식 형태로 반영된다. 신경망의 구조(은닉 층의 개수, 은닉 층 내 노드 개수 및 연결 가중치 초기 값)가 결정 된 상태에서, 원 자료를 입력 정보로 클래스 정보를 출력값으로 하는 학습 데이터 집합을 만들고 LM 알고리즘을 통해 추가적인 학습을 수행하고 최종 모델을 구성한다.

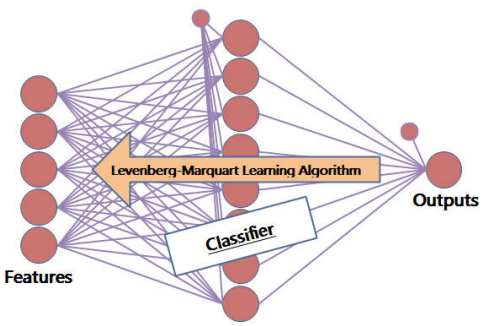


Figure 3. 분류기 학습 신경망 구축

본 논문에서 제안하는 최종 ALBNN 모델은 특질 추출 기법과 분류 기법 자체의 특성을 추정한 상태에서 통합 단계로 재학습을 한다. 그 과정에서 선택되는 특질의 개수는 고정되기 때문에 변수의 차원을 효율적으로 감소시키는 장점은 그대로 보유할 수 있다. 또한 제안 모델은 두 단계의 알고리즘의 특성 자체를 선행지식으로 학습하기 때문에 초기해의 공간을 분류문제에 적합하게 시작점을 이동시킬 수 있고 목표 값에 보다 정확하게 수렴시킬 수 있는 특징이 있다. 추가적으로 두 단계를 통합하여 학습하기 때문에 그동안 고려하지 못했던 특질과 분류기법과의 상호작용 등을 반영할 수 있고 최종 분류 모형을 만들면서 목적 값에 맞게 각 단계에서 발생되었던 오류를 다시 줄이는 과정을 거치게 되어 분류 성능 향상이 기대된다.

2.4 제안 알고리즘의 효과

본 연구에서 제시한 ALBNN의 효과를 설명하기 위해 이 절에서는 통계 패키지인 R을 활용하여 시뮬레이션으로 다변량 데이터를 생성하고 분류하는 실험을 실시하였다. 먼저, 두 개의 집단별로 각각 상이한 모수를 지닌 다변량 정규분포를 따르는 3차원의 데이터를 1000개씩 생성하였다. 두 개의 집단의 분포는 3차원 공간상에서 혼재된 영역이 상당히 존재하도록 각각의 분포에서 사용한 모수를 조정시켰다. 전체 데이터의 산포는 <Figure 5>와 같이 분포하게 된다. 두 집단의 분포가 3차원 공간상에서 겹쳐지는 영역이 상당히 존재하므로 성능이 좋은 분류기를 직접 적용하여 2차원의 초평면(Hyperplain)을 만들어도 집단을 오분류 하는 경우가 발생할 가능성이 크다. 두 개의 집단 분류문제에서 일반적으로 많이 사용되고 있는 통계적 기법인 SVM을 사용하여 분류한 결과 6.90%의 오분류가 있었다. 이와 같은 특징이 있는 다변량 데이터는 원 데이터를 이용하는 것보다 정보의 변형을 통해 적합한 특질을 추출한다면 분류 성능을 향상시킬 수 있다.

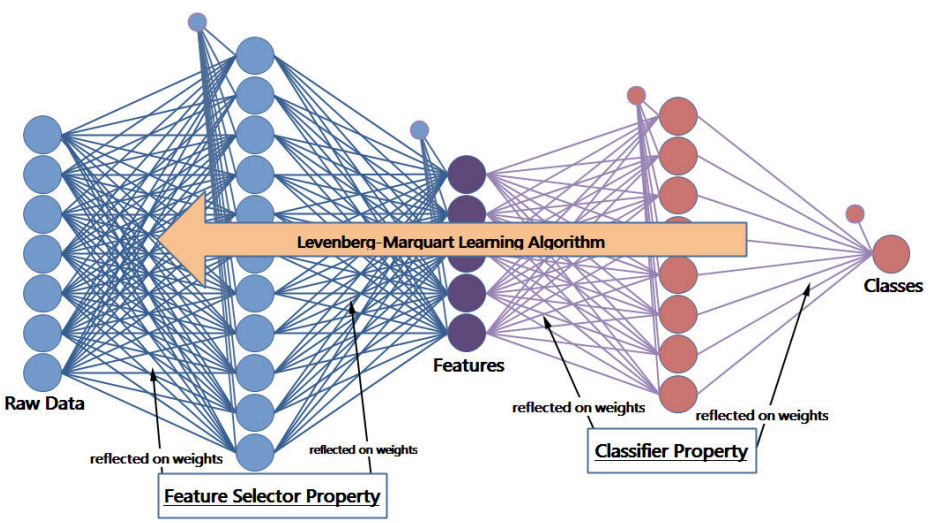


Figure 4. 두 개의 신경망을 결합한 통합 신경망

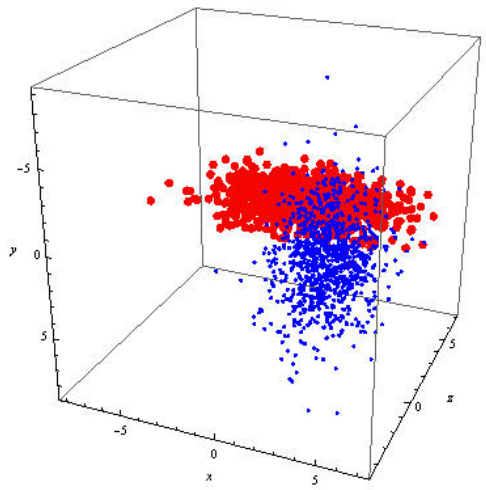


Figure 5. 상이한 두 집단 데이터의 3차원 산포도

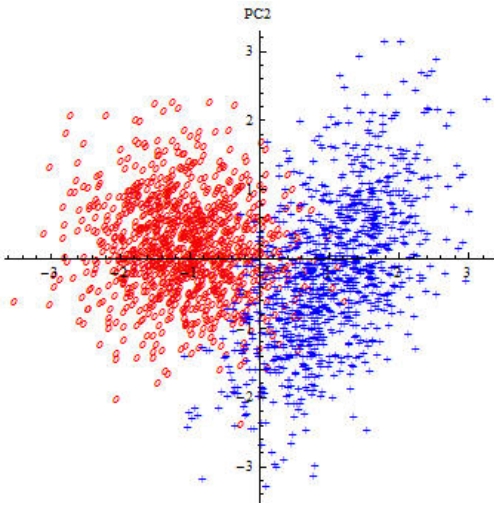
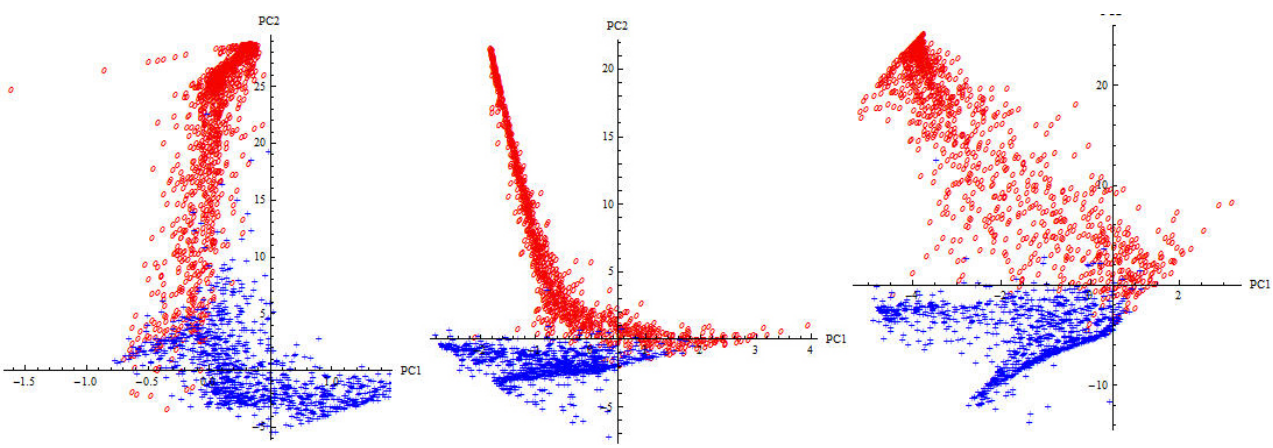


Figure 6. 두 개의 PC를 활용한 산포도

<Figure 6>은 PCA를 통해 분산이 큰 두 개의 PC(주성분)를 선택하여, 이를 기반으로 데이터를 표시한 그림이다. PCA는 무감독 기법 중 하나로 클래스 정보의 반영 없이 변수들만의 특성만을 활용하여 데이터를 변환하는 방식이다. 따라서 각각 다른 집단의 특성 정보가 반영된 것이 아니라 전체 데이터의 특성을 변형하여 특질을 선택하기 때문에 분류에 적합하게 특질의 변형이 이루어지지 않았다. 두 집단이 중첩된 영역이 아직도 꽤 남아 있음을 확인할 수 있다. PCA를 통해 산출된 특질들을 SVM에 적용한 결과 5.56%의 오분류가 발생하였다.

본 논문에서 제안하는 ALBNN 방식은 두 번의 학습과정을 통해 특질 추출 기법 및 분류 기법의 특성이 선행 지식으로 반영된 상태에서, 통합 단계를 통해 학습 오차를 줄이도록 신경망 연결 가중치를 추가적으로 조정한다. 따라서 목표 값에 맞게 신경망 구조를 변형하는 과정에서 추출된 특질과 변수들간의 간의 상간관계 뿐만 아니라 분류기법 및 특질간의 상관관계 등을 반영하여 특질 들이 변형된다. 분류 오차를 줄이도록 신경망 연결 가중치를 변형시키기 때문에 통해 분류 문제에 적합한 새로운 특질을 도출할 수 있다. <Figure 7>은 ALBNN 알고리즘을 적용하는 세 번째 단계에서 통합하여 학습하는 과정에서 변형되는 특질들의 모습을 반영한 그래프이다. 신경망의 학습 과정에서 epoch(신경망의 학습 횟수)가 늘어나면서 점차 상이한 집단의 정보가 혼재되는 영역은 줄어드는 것을 확인할 수 있다. 최종 26번째 epoch에서 산출된 특질을 PCA를 통해 도출한 특질과 비교해 보면 상이한 집단 정보가 겹쳐진 개수가 상당히 줄어들었다. 이는 제안 알고리즘을 통해 새롭게 생성된 특질들이 분류에 더 적합한 형태로 변형되었다는 것을 알 수 있다. 최종학습이 끝난 뒤에 분류 성능을 평가한 결과 3.74%의 오분류가 있었다. 앞에서 기술한 두 가지 방식에 비해 상당히 분류 성능이 향상 된 것을 알 수 있다.



(a) 2 epochs 학습 (b) 14 epochs 학습 (c) 26 epochs 학습

Figure 7. 신경망의 학습 과정 중 각 epoch 별 변형된 특질의 산포도

3. 실험 및 결과 분석

3.1 실험 설계

본 논문에서 제안한 알고리즘의 성능 평가를 위해 UCI(the University of California, Irvine) repository에 있는 2개의 클래스(2 Classes) 분류 문제에 해당하는 7가지 데이터를 활용하여 실험을 진행하였고 사용한 데이터의 특징은 <Table 1>과 같다. 의료데이터는 결과에 영향을 미치는 변수가 많고 각 변수의 분포가 상이한 다변량 데이터로 중요 특질 추출이 의미가 있고 분류의 정확성이 다른 도메인보다 더 중요하고 치명적이기 때문에 본 실험에 이용하였다. 특히, 공정의 이상이나 정상상태를 판별하기 위한 목적을 위한 특성을 반영하기 위해 집단 간 빈도가 상이한 데이터를 활용하여 실험을 진행하였다.

학습 데이터 집합(dataset)에 종속적인 성능 평가 결과를 방지하기 위해 5-집단 교차 검증(cross validation) 방법을 활용하였다.

Table 1. 실험 데이터

Dataset	Instances	Variables	Proportions (normal)
Parkinsons	195	22	0.7538
Ionosphere	351	33	0.6410
WDBC	569	30	0.6274
Spectf	267	44	0.7940
Hepatitis	155	19	0.7935
Horse colic	368	20	0.6630
Pima Indians diabetes	768	8	0.6510

즉, 전체 자료를 임의로 5등분하여 5개 집단을 구성하고 하나의 집단의 자료는 제외하고 나머지 4개 집단의 자료를 활용하여 학습 데이터(Training Data)로 모델을 학습시키고 제외되었던 자료를 테스트 데이터(Test Data)로 사용하여 알고리즘의 성능을 평가하였다. 동일한 방법을 집단을 변경해가면서 5번 반복하고, 얻어진 분류결과의 정확성(Accuracy)의 평균값을 성능 평가의 기준으로 삼는다.

실험에 사용한 랜덤 시드의 변경을 통해 30개의 각기 다른 초기 비중 값을 생성하여 일반적인 인공신경망 분류 알고리즘이 임의로 정해지는 초기 신경망 연결 가중치 값에 영향을 받는 특성을 반영하였다. 또한 신경망 분류기의 성능에 상당한 영향을 미치는 것이 은닉 층에 있는 노드 개수이다. 본 연구에서는 신경망 단위로 은닉 층은 1개로 구성된 모델을 만들고 노드의 개수를 변경해가며 반복 실험을 통해 MSE(Mean Squared Error)를 가장 작게 만드는 노드의 개수를 찾아 해당 신경망의 은닉 층 노드 개수로 결정하였다.

특질 추출기법을 통해 얻어진 특질들의 개수에 따라 성능에 영향을 미치는 은닉 층 노드 개수가 결정된다. 또한 선정된 특질의 개수가 실험의 비교대상이 되는 PCA-SVM의 성능에도 큰 영향을 미친다. 따라서 본 연구에서 선택할 특질(PC)의 개수는 일반적으로 사용되는 고유벡터나 누적분산의 크기 기준이 아닌 반복 실험을 통하여 해당 데이터의 분류 성능을 가장 향상 시키는 개수를 해당 문제에 적합한 특질의 수로 결정하였다.

3.2 비교 알고리즘

본 연구에서 제시한 ALBNN은 특질 추출 기법으로 PCA를 사용하였고 분류 기법을 적용하기 위해 SVM을 사용하였다. 먼저, 특질 추출 기법으로 널리 이용되고 있는 PCA는 변수들의 분포에 따라 변수들 간의 분산이 최대가 되게 축을 변환 하는 방법이다(Turk *et al.*, 1991). 적절한 PC를 특질로 선택하면 소수의 PC만으로도 데이터 전체의 특성을 반영할 수 있고 잘 식별할 수 있는 장점이 있다. 정보의 손실을 최소화하면서 변수들 간의 분산을 고려하기 때문에 데이터의 차원 축소에 적용할 수 있다. 따라서 차원이 큰 데이터의 분류 문제에서 데이터의 전 처리 과정으로 널리 활용되고 있다. 따라서 본 논문에서는 특질 추출 알고리즘으로 PCA를 활용하였다.

SVM 분류기는(Cortes and Vapnik, 1995)에 의해 제안된 학습 알고리즘으로 다양한 분야에 활용되고 있는데 전형적으로 2개 그룹의 분류 문제에 주로 적용되어 왔다. SVM은 선형으로 분리되지 않는 데이터를 다른 차원 공간으로 변환하여 선형으로 분리되도록 만든 후에, 2차 수리계획법을 통해 데이터를 분류하는 기법이다. 마진을 최대화하는 형태로 초평면을 계산하는 통계적 학습 알고리즘으로 분류 성능이 뛰어난 특징이 있다(Burges, 1998). 따라서 본 연구에서는 분류기로 SVM을 활용하여 ALBNN에 특성을 학습시켰고 성능을 비교하였다.

차원이 크지 않은 데이터의 경우 전 처리과정으로 사용되는 특질 변환 단계 없이 직접 분류기를 적용할 수도 있다. 따라서 SVM 분류기를 특질 추출 단계 없이 직접 적용하는 방식도 분류 성능을 평가하는 비교 대상으로 한다. 그리고 신경망 분류기의 특성을 비교하기 위해 제안방식이 아닌 일반 신경망 분류기도 비교 알고리즘으로 선정하였다. 신경망 연결 가중치의 초기 값의 영향을 반영하여 성능을 평가하기 위함이다.

3.3 실험 결과

<Table 2>는 실험 데이터를 활용하여 네 가지 분류기의 분류 성능을 비교한 결과이다. 실험 결과를 보면 6개의 데이터에서 PCA+SVM이 SVM보다 성능이 좋았다. 이 데이터들은 직접 분류 기법을 적용한 것보다 특질 추출 과정을 거친 후 성능이 향상되었기 때문에 특질 추출 알고리즘이 효과가 있었다. 이 데이터들 총 6개 중 4개에서 본 연구에서 제안하는 알고리즘이

Table 2. 실험 결과(분류 정확도)

Dataset	SVM		PCA+SVM		ANN		ALBNN	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Parkinsons	86.94	0.57	87.69	0.70	83.76	7.5	**88.07***	1.15
Ionosphere	93.87	0.50	95.40	0.47	90.99	2.94	**96.22***	0.83
WDBC	97.40	0.33	96.07	0.44	**97.87***	0.92	95.82	0.60
Spectf	79.30	0.70	80.35	1.50	71.76	3.82	**81.30***	2.31
Hepatitis	84.06	1.00	**85.98***	1.44	82.05	4.77	84.73	2.82
Horse colic	67.33	1.60	**69.33***	1.61	66.68	3.61	69.09	2.36
Pima Indians diabetes	76.23	0.80	76.41	0.72	72.12	1.44	**78.57***	1.22

PCA+SVM보다 우수하였다. 기존 알고리즘으로 분류하지 못했던 정보를 통합신경망을 통해 특질들의 추가적인 변형을 하게 되면 분류 성능을 향상 시키는 효과가 있었다. 나머지 2개의 데이터에서는 SVM 분류기를 단독으로 사용했을 때 보다 성능의 향상을 보였지만 PCA+SVM 분류기에 비해 향상된 성능을 보이지 못했다. 이는 두 개의 알고리즘의 특성을 학습과정에서 발생한 문제라고 판단된다. 기존 알고리즘이 가진 특성을 제대로 반영한 상태에서 특질이 변환되어야 하는데 독립된 신경망 두 개를 학습시키는 과정에서 알고리즘의 특성이 충분히 반영되지 않아 분류성능이 다소 차이가 난 것으로 보인다. 기존 알고리즘을 학습 과정에서 원래의 특성을 잘 추정할 수 있도록 신경망 학습 방식에 대한 추후연구를 통해 분류성능을 개선 시켜나갈 수 있을 것이다.

WDBC 데이터의 경우는 PCA+SVM보다 SVM의 성능이 좋기 때문에 특질 추출 알고리즘을 적용하는 것 자체가 오히려 성능을 저하시킨 경우이다. 이와 같이 추출된 특질 자체가 원 정보 보다 분류에 적합하지 않은 경우, 이를 학습하는 ALBNN 은 성능이 더 떨어지는 특징을 보였다. 학습에 사용된 특질 자체가 원래 데이터보다 오분류 가능성이 높은 특성을 갖고 있기 때문에, 특징을 추정하는 과정에서 분류문제에 적절하지 않은 방향으로 학습하여 성능이 낮아진 것으로 볼 수 있다. 따라서 본 연구에서 제안하는 방식은 선택한 특질 추출 알고리즘 자체를 학습하기 때문에 특질 추출의 효과가 있는 데이터에 적용하는 것이 바람직하다.

두 개의 신경망 분류기의 성능을 비교해보면, 전체적으로 제안 알고리즘의 분류 성능이 일반적인 신경망 분류기보다 뛰어난 특징을 보이고 있다. 또한 분류 정확성의 표준편차가 상대적으로 상당히 줄어들었음을 확인할 수 있다. 이는 기존 신경망 알고리즘의 시작점이 임의로 결정되는 것에 반해 제안 알고리즘은 선행지식을 통해 초기해의 공간을 더 나은 시작점으로 이동시키고 분류 문제를 해결하기 때문에 성능의 변동이 적은 장점이 있다.

4. 결 론

본 연구에서는 다변량 데이터의 분류 성능을 향상시키기 위해 기존의 특질 추출 알고리즘과 분류 알고리즘을 통합할 수 있는 새로운 방식을 제안하였다. 신경망의 학습과정에서 기존 알고리즘의 특성을 추정하여 반영하고 최종 모형을 생성하면서 분류 문제에 적합한 새로운 특질들을 도출 할 수 있었다. 특질들을 변형하는 과정에서 변수들 간의 상관관계, 클래스 정보, 특질 추출 방식과 분류기법의 상관관계 등 다양한 정보들을 반영하기 때문에 다변량 데이터의 분류 성능을 개선시킬 수 있었다. 또한 본 연구는 특질 추출과 분류 단계를 통합할 수 있는 새로운 기계학습 방식이기 때문에 다양한 알고리즘을 통하여 확장시켜 나갈 수 있다. 즉, 본 연구에서는 PCA와 SVM에 한정된 비교 실험을 진행하였는데 데이터 집합의 특징 별로 어떤 특질 추출 및 분류 알고리즘의 조합을 사용하는 것이 분류 정확도를 향상시키는데 적합한지 추후 연구를 통해 찾고, 다양한 분류모형으로 제안 알고리즘의 장점을 확장시켜 나갈 수 있을 것이다.

본 연구에서는 특질 추출 기법 및 분류 기법을 활용하는 두 가지 단계 자체를 각각 특정한 역할을 하는 복잡한 형태의 함수라고 고려하고, 신경망의 학습을 통해 그 특성을 추정하였다. 기존 알고리즘이 갖고 있는 함수의 기능을 얼마나 잘 추정하느냐가 통합신경망을 구축하기 전에 선행지식의 형태로 초기 신경망의 연결가중치에 적절하게 반영하는 척도가 된다. 추후 연구를 통해 각 단계의 특성을 보다 정확하고 빠르게 추정할 수 있는 신경망 학습 알고리즘을 개선시킨다면 분류성능을 향상시킬 것으로 기대된다.

마지막으로 신경망의 구조는 학습 및 성능에 중요한 영향을 미치게 된다. 따라서 신경망의 은닉 층 개수 및 은닉 층 내 노드 개수를 성능에 적합하도록 결정하는 기준이 필요하다. 유전 알고리즘이나 다양한 최적화기법을 활용하여 제안 알고리즘에 적용할 수 있는 기준을 마련해야 할 것이다.

참고문헌

Burges, C. J. C. (1998), A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, **2**, 121-167.

Cortes, C. and Vapnik V. (1995), Support vector network, *Machine Learning*, **20**, 273-297.

Gori, M. and Tesi, A. (1994), On the Problem of Local Minima in Backpropagation, *IEEE Transactions on Pattern analysis and Machine Intelligence*, **14**(1), 76-86.

Hagan, M. T. and Menhaj, M. B. (1994), Training feedforward networks with the marquardt algorithm, *IEEE Transactions on Neural Networks*, **5**(6), 989-993.

Kabir, M. M. and Islam, M. M., Murase, K. (2010), A new wrapper feature selection approach using neural network, *Neruocomputing*(Article in Process).

Marquardt, D. W. (1963), An algorithm for least squares estimation of nonlinear parameters, *Journal of Society for Industrial and Applied Mathematics*, **11**(2), 431-441.

Park, M. S. and Choi, J. Y. (2009), Theoretical analysis on feature extraction capability of class-augmented PCA, *Journal of Pattern Recognition*, **42**, 2353-2362.

Pudil, P., Novovicova, J. and Kittler, J. (1994), Floating search methods in feature selection, *Pattern Recognition Letters*, **15**(11), 1119-1125.

Saeys, Y., Inza, I., and Larranaga, Y, P. (2007), *A review of feature selection techinques in bioinformatics Bioinformatics*, **23**(19), 2507-2517.

Sarkar, I., I. Sarkara, N., Planetb, P. J., Baelc, T. E., Stanleyd, S. E., Siddalle, M., and DeSalle, R. (2002), Characteristic attributes in cancer microarrays, *Journal of Biomedical Informatics*, **35**(2), 111-122.

Towell, G. G. and Shavlik, J. W. (1994), Knowledge based artificial neural networks, *Artificial Intelligence*, **70**(1), 119-165.

Turk, M. and Pentland, A. (1991), Eigenfaces for recognitions, *Journal of Cognitive Neuroscience*, **3**, 71-86.

Yves, C., David E. Rumelhart, A. (1995), Back Propagation : theory, architecture, and applications, *Lawrence Erlbaum Associates*, New Jersey, USA.

윤 현 수

고려대학교 산업공학과 학사

현재 : 고려대학교 산업공학과 석사과정

관심분야 : Multivariate Statistical Analysis for Data Mining, Machine Learning, FDC, Time Series Analysis

백 준 걸

고려대학교 산업공학과 학사

고려대학교 산업공학과 석사

고려대학교 산업공학과 박사

현재 : 고려대학교 산업경영공학과 부교수

관심분야 : 첨단공정제어, 지능형 이상 진단, 데이터마이닝 응용