부분 최소 자승법과 잔차 보상기를 이용한 비선형 데이터 분류

Non-linear Data Classification Using Partial Least Square and Residual Compensator

김 경 훈, 김 태 영, 최 원 호*
(Kyung-Hun Kim, Tae-Young Kim, and Won-Ho Choi)

Abstract : Partial least squares(PLS) is one of multivariate statistical process methods and has been developed in various algorithms with the characteristics of principal component analysis, dimensionality reduction, and analysis of the relationship between input variables and output variables. But it has been limited somewhat by their dependency on linear mathematics. The algorithm is proposed to classify for the non-linear data using PLS and the residual compensator(RC) based on radial basis function network (RBFN). It compensates for the error of the non-linear data using the RC based on RBFN. The experimental result is given to verify its efficiency compared with those of previous works.

Keywords : partial least squares, radial basis function network, residual compensator.

I. 서론

부분 최소 자승법(Partial Least Squares, PLS)은 다수의 입력 변수와 출력 변수의 관계를 정립하는데 효과적인 방법이다. 이것은 주 성분 분석(Principal Component Analysis)을 기반으로 개발된 알고리즘으로, 다변량 문제의 변수 상호 간 관련 구조분을 분석하기 위하여 사용하였다. 이 방법의 요지는 다수의 입력 변수와 출력 변수가 존재할 경우 변수 사이의 상호 관계를 파악하여, 기존 변수의 차원을 줄인다는 것이다[1]. 그러나 이러한 방법은 선형적 계산에 기초 하기 때문에 입력과 출력 사이의 함수 관계가 비선형이거나 데이터 집합 간의 상관관계가 존재하는 경우에 이들 관계를 분석하고 표현하기에는 적합하지 않다는 단점을 갖는다[2-4].

이러한 선형적 제약을 극복하고 비선형 데이터를 분석하기 위하여 PLS와 신경 회로망을 결합하는 방법[5] 또는 PLS의 내부 모델(Inner model)을 방사 기저 함수 신경망(Radial Basis Function Network, RBFN)으로 변환하는 방법[6,7] 등과 같은 다양한 알고리즘이 제안되고 있다.

본 논문에서는 RBFN을 기반으로 하는 잔차 보상기를 이용함으로써 기존 입력 변수의 차원을 축소하고 복잡도를 최소화하면서 PLS의 선형 결과값과 비선형 데이터의 오차를 보정하여 더욱 효율적으로 데이터를 분류하는 알고리즘을 제안한다.

또한, PLS, RBFN, PLS/RBF 모델을 설계하여 Irish 데이터와 Glass 데이터의 분류에 적용한 모의 실험을 행함으로써 RBFN을 기반으로 하는 잔차 보상기의 성능을 입증한다.

* 책임저자(Corresponding Author)
논문 접수 : 2003. 9. 2., 채택확정 : 2003. 12. 8.
김경훈, 최원호 : 울산대학교 전기전자정보시스템공학부
(kkh00111@hanmail.net/ whchoi@mail.ulsan.ac.kr)
김태영 : 알칸 대한 주식회사 (taeyoung.kim@alcan.com)
※ 이 논문은 2002년 울산대학교의 연구비와 한국과학재단 지정 울산대학교 네트워크 기반 자동화연구센터의 지원에 의한 것입니다.

II. PLS와 RBFN

1. PLS

PLS는 프로세스 변수로부터의 입력(predictor) 행렬과 출력(predicted) 행렬 사이의 최대의 공분산을 가지는 스코어 행렬(score matrix)과 로딩 행렬(loading matrix)을 결정하는 차원 축소 기술이다[2].

입력 행렬을 $X \in R^{n\times m}$ 로 표시하고 출력 행렬을 $Y \in R^{n\times p}$ 로 표시한다. 여기서 m 은 관찰 변수의 수, n 은 데이터의 총수이고 p 는 출력 행렬에서의 관찰 변수의 수이다.

입력 행렬, X 와 출력 행렬, Y 는 각각 스코어 행렬 $T \in R^{n\times a}$, $U \in R^{n\times a}$ 와 로딩 행렬 $P \in R^{m\times a}$, $Q \in R^{p\times a}$ 로 분해 된다.

$$X = TP^T + E = \sum_{j=1}^{a} t_j p_j^T + E \tag{1}$$

$$Y = UQ^T + F = \sum_{j=1}^{a} u_j q_j^T + F \tag{2}$$

여기서 a 는 PLS의 성분 차수이고 E 와 F 는 잔차 행렬이다.

(3)을 이용하여 출력 Y 의 스코어 $\hat{u}$ 를 추정한다.

$$\hat{u}_j = b_j t_j \tag{3}$$

$$b_j = \frac{u_j^T t_j}{t_j^T t_j} \tag{4}$$

(3)과 (4)에서 u_j 는 스코어 행렬 U 의 j-번째 열(column)이고 b_j 는 회귀 상수이다.

행렬 형태에서, 이들 관계는 다음과 같이 표현된다.

$$\hat{U} = TB \tag{5}$$

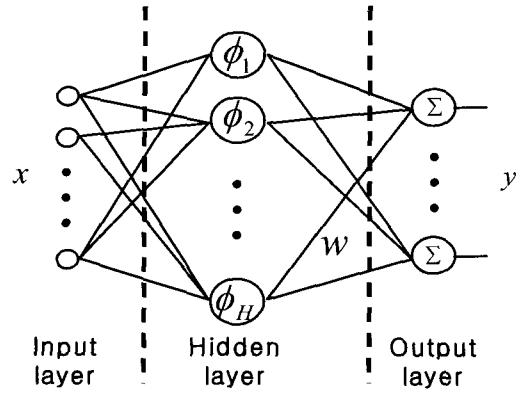


그림 1. RBFN의 기본 구조.

Fig. 1. The structure of RBFN.

현재 이들 스코어 행렬, 로딩 행렬, 회귀 행렬을 계산하기 위한 여러 알고리즘이 제안되고 있으나 일반적으로 NIPALS (Non-iterative Partial Least Squares) 알고리즘이 사용된다[2].

2. RBFN

RBFN은 중심점과 함수폭을 이용하여 입력 공간을 지역화된 영역들로 나누는 특징을 가지고 있으며 네트워크의 구조는 전방향 신경망의 구조와 비슷하다. RBFN은 다양한 기법에 의해 학습될 수 있으며 이중 가장 일반적인 학습방법은 K-평균 군집화(K-means clustering) 알고리즘, P-최근접 이웃(Pnearest neighbor) 알고리즘에 의해 은닉층 뉴런의 중심점과 분산을 계산하고 최소 자승법(least square) 알고리즘에 의해 은닉층과 출력층 사이의 연결 강도를 결정한다[8].

2.1 RBFN의 기본 구조

RBFN은 그림 1과 같이 입력층, 은닉층, 출력층으로 구성된다.

각각의 은닉 뉴런들은 고유의 중심점과 함수폭을 가지고 있고 은닉층의 출력은 이들 파라미터를 이용한 가우시안 함수를 통해 계산할 수 있다.

$$\tilde{\phi}_h(x) = \exp\left(\frac{\| x - c_h \|}{2\sigma_h^2}\right) \tag{6}$$

여기서 c_h 와 σ_h 는 각각 은닉 뉴런의 중심점과 함수폭을 나타내고 x 는 입력값을 나타낸다.

(7)은 은닉 뉴런의 출력값을 정규화한다[9,10].

$$\phi_h(x) = \frac{\tilde{\phi}_h(x)}{\sum_{h=1}^{H} \tilde{\phi}_h(x)} \tag{7}$$

출력층의 출력값은 아래 (8)과 같이 은닉층의 출력값과 연결강도의 곱으로 표현 되어진다.

$$y_i = \sum_{h=1}^{H} w_{ih}\phi_h(x) \tag{8}$$

여기서 w_{ih} 는 h-번째 은닉층 뉴런과 i-번째 출력 층 간의 연결 강도이고 H 는 은닉 뉴런의 수이다.

각 뉴런의 파라미터와 연결강도는 일반적으로 세 단계에 의해 결정된다.

2.2 은닉 뉴런의 중심점 결정

은닉 뉴런의 중심점을 결정하는 방법은 여러 가지가 있지만 가장 대중적으로 사용하는 방법은 K-평균 군집화이다[8]. 중심점의 수는 은닉 뉴런의 수와 동일하며 데이터들은 가장 근처에 있는 중심점에 할당된다. 이 알고리즘은 각각의 중심점에 할당되어 있는 훈련 데이터들과 중심점 사이의 유클리디안 거리의 총 합을 최소화 하는 중심점을 결정한다.

$$E_{K-mean} = \sum_{h=1}^{H}\sum_{k=1}^{K} \left\| c_h - x_k \right\|^2 \tag{9}$$

각 중심점은 학습 데이터에 의해 임의로 초기화 되고 각 학습 데이터들은 가장 가까운 초기화된 중심점에 할당된다. 할당된 데이터들을 이용하여 평균을 구하고 중심점은 그 평균값으로 이동한다. 모든 중심점이 갱신 된 후, 그 중심점이 수렴할 때까지 앞의 과정을 반복한다.

2.3 함수폭, σ 의 결정

은닉 뉴런들의 중심점을 결정한 후, 각각의 중심점에 대한 함수폭을 결정한다. 함수폭을 결정하는 목적은 원하는 출력이 유연성을 가지도록 하기 위해서이다. 함수폭, σ 를 결정하는 방법으로 (10)과 같은 P-최근접 이웃 추정법을 이용한다[8].

$$\sigma_h = \left[\frac{1}{P} \sum_{j=1}^{P} \| c_h - c_j \|^2 \right]^{1/2} \tag{10}$$

제안한 알고리즘에서는 $P = 2$ 로 설정하였다.

2.4 연결 강도의 결정

은닉층으로부터 출력층으로 전달되는 연결 강도를 구하는 목적은 목표 값과 RBFN의 출력 값의 차이를 최소화하기 위해서이다. O 를 학습 데이터의 목표값 이라고 가정하면 연결 강도는 최소 자승법(least squares)을 이용하여 행렬 연산으로 계산할 수 있다[8].

$$w = O\phi^T(\phi\phi^T)^{-1} \tag{11}$$

RBFN의 출력 변수의 수가 M 이고 훈련 데이터의 수가 K 라고 가정하면 O 는 $M \times K$ 행렬이 되고 ϕ 는 $H \times K$ 행렬이 되며 연결 강도 W 는 $M \times H$ 행렬이 된다.

3. PLS/RBF 모델

PLS/RBF 모델은 부분 최소 자승법의 선형적인 내부 모델을 비선형의 관계로 표현하기 위하여 외부 모델은 변화 시키지 않고 내부 모델인 선형 회귀 행렬을 RBFN으로 대체한 모델이다[6,7].

방사 기저 함수 신경망의 입력은 입력 행렬, X 로 부터 계산된 t 스코어 행렬이 되고 출력은 출력 행렬, Y 로 부터 계산된 u 스코어 행렬이 된다. 따라서 PLS/RBF 모델에서의 RBFN 출력값은 (12)가 된다.

$$\hat{u} = \sum_{i=1}^{H} w_i f(\|t - c_i\|) \tag{12}$$

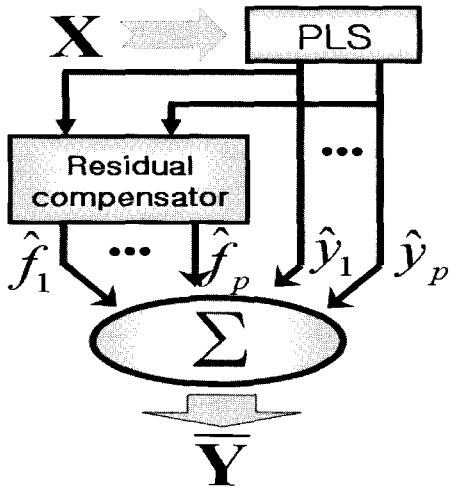


그림 2. 잔차 보상기의 구성.
Fig. 2. The organization of the residual compensator.

여기서, w는 방사 기저 함수 신경망의 연결 강도이고 c는 중심값, $f(\cdot)$는 방사 기저 함수 신경망의 활성함수이다. 활성 함수로 가우시안 함수를 사용한다.

$$f(s) = \exp(-s^2 / 2\sigma^2) \tag{13}$$

방사기저 함수 신경망의 훈련을 위해서는 hybrid BFGS 알고리즘을 사용한다[11].

III. 제안한 알고리즘

1. 잔차 보상기

PLS는 선형적 계산에 의한 알고리즘이므로 비선형 데이터의 분류나 추정에는 적합하지 않다. 이러한 단점을 보완하기 위하여 PLS에 의해 추정된 값과 원하는 목표값 사이의 잔차를 RBFN을 기반으로 하는 잔차 보상기로 추정하여 보정하였다.

그림 2는 잔차 보상기의 구성을 보여준다. 입력 변수, X를 NIPALS 알고리즘을 이용하여 PLS 변수로 변환시키고 선형 PLS를 이용하여 각 차수에 대한 추정값, $\hat{Y}$을 계산한다.

(14)는 PLS에 의한 출력값을 나타낸다.

$$\hat{Y} = TBQ^T \tag{14}$$

PLS 출력값, $\hat{Y}$을 RBFN을 기반으로 하는 잔차 보상기의 입력으로 하고 PLS의 출력값과 희망하는 출력값, Y의 차이를 RBFN의 목표값, O으로 하여 학습시킴으로써 RBFN의 중심값 $[c_1 \cdots c_H]$, 함수폭 $[\sigma_1 \cdots \sigma_H]$, 연결강도 w를 결정한다.

(4)에서 보는 바와 같이 PLS의 차수가 변하더라도 그 출력값의 차수는 변하지 않기 때문에 잔차 보상기의 입력 차수는 PLS의 출력 차수와 동일하다.

(5)은 잔차 보상기의 은닉 뉴런의 정규화된 출력값이고 (16)은 출력층의 출력값을 나타낸다[9].

$$\phi_h(\hat{y}) = \frac{\exp \dfrac{\|\hat{y} - c_h\|}{2\sigma_h^2}}{\sum_{j=1}^{H} \exp \dfrac{\|\hat{y} - c_j\|}{2\sigma_j^2}} \tag{15}$$

$$\hat{f}_i = \sum_{h=1}^{H} w_{ih} \varphi_h(\hat{y}) \tag{16}$$

잔차 보상기의 출력은 PLS에 의해 추정된 $\hat{Y}$을 보정하기 때문에 출력층의 뉴런 수는 $\hat{Y}$의 차수와 동일하다. 은닉층과 출력층 사이의 연결 강도는 (11)의 최소 자승법을 이용하여 계산한다.

최종의 추정값, $\overline{Y}$은 (17)와 같이 PLS의 출력값, $\hat{Y}$과 잔차 보상기에 의한 출력값, $\hat{F}$을 더하여 계산할 수 있다.

$$\overline{Y} = \hat{Y} + \hat{F} \tag{17}$$

2. PLS의 차수와 은닉 뉴런의 수 결정

최적 구조의 잔차 보상기와 PLS 모델을 결정 하기 위한 제한 조건으로는 다음과 같다.

i) 추정된 출력 값의 오차, prediction residual sum of squares (PRESS)

$$PRESS = \frac{1}{n} \| Y - \overline{Y} \|_F^2 \tag{18}$$

ii) 분류에 관한 문제의 경우 오분류율 ($m(a,h)$)

iii) PLS의 성분 차수 (a)

iv) RBFN 내 은닉 노드의 수 (H)

본 논문에서는 비용 함수(Cost Function)를 (19)과 같이 결정하고 $J_{a,h}$을 최소화 하는 성분차수(a)와 은닉 뉴런의 수 (H)를 결정하였다.

$$J_{a,h} = \frac{1}{n} \| Y - \overline{Y} \|_F^2 + \frac{a}{n} + \left(\frac{h}{n}\right)^2 + m(a,h) \tag{19}$$

여기서 n은 테스트 데이터의 총수이고, $\| \bullet \|_F$는 Frobenius norm을 나타낸다. $m(a,h)$는 성분 차수가 a이고 은닉 뉴런의 수 h가 일 때의 오분류율 이다.

IV. 실험 및 결과

1. Irish data

제안한 알고리즘의 검증을 위하여 피셔의 아이리쉬 분류 데이터 집합을 이용하였다[12,13].

표 1의 Fisher의 Irish 분류 데이터 집합은 세 종류의 꽃에 대하여 꽃받침의 폭과 길이, 꽃잎의 폭과 길이에 대한 평균과 표준편차를 나타낸다.

그림 3은 표 1의 평균과 표준 편차를 기준으로 생성된 각 클래스의 데이터 분포를 나타낸다. 클래스 1의 경우를 보면 Petal Length가 다른 클래스의 그것과 확연한 분포의 차이를 보이기 때문에 PLS 알고리즘 이나 다른 분류 알고리즘을 사

표 1. 피셔의 아이리쉬 데이터 집합.
Table 1. Fisher's Irish Classic Data Set.

Class 1: Iris Setosa	Mean	Std. Deviation
Sepal Length	5.01	0.35
Sepal Width	3.43	0.38
Petal Length	1.46	0.17
Petal Width	0.30	0.13
Class 2: Iris Versicolor	**Mean**	**Std. Deviation**
Sepal Length	5.94	0.52
Sepal Width	2.77	0.31
Petal Length	4.29	0.47
Petal Width	1.33	0.20
Class 3: Iris Virginica	**Mean**	**Std. Deviation**
Sepal Length	6.59	0.64
Sepal Width	2.98	0.32
Petal Length	5.55	0.55
Petal Width	2.03	0.27

표 2. 강건성 테스트를 위한 데이터 집합.
Table 2. Data set for robustness.

Class 1: Iris Setosa	Mean	Std. Deviation
Sepal Length	4.9933	0.3465
Sepal Width	3.4369	0.3804
Petal Length	1.4574	0.1714
Petal Width	0.3232	0.1411
Class 2: Iris Versicolor	**Mean**	**Std. Deviation**
Sepal Length	5.9049	0.5247
Sepal Width	2.7632	0.3081
Petal Length	4.2862	0.4670
Petal Width	1.3392	0.2042
Class 3: Iris Virginica	**Mean**	**Std. Deviation**
Sepal Length	6.6059	0.6507
Sepal Width	2.9658	0.3058
Petal Length	5.5778	0.5460
Petal Width	2.0048	0.2758

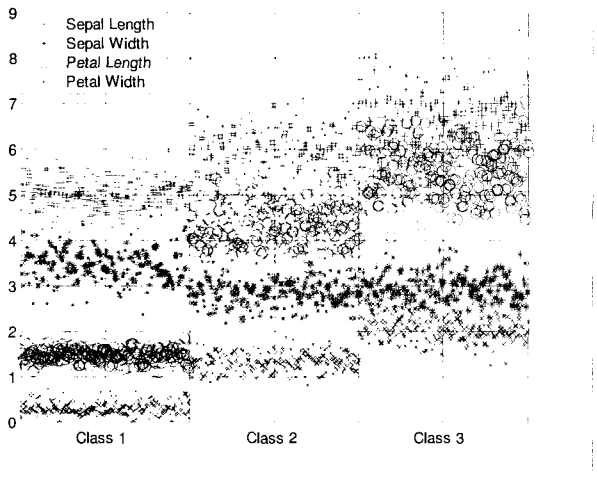


그림 3. 피셔의 아이리쉬 데이터 집합.
Fig. 3. Fisher's Irish Classic Data Set.

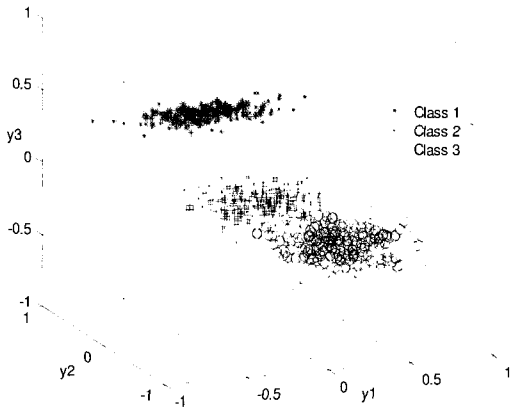


그림 4. PLS 알고리즘의 결과 (차수=3).
Fig. 4. The Result of PLS (order=3).

용하더라도 구분이 용이하지만 클래스 2와 클래스 3의 경우는 Sepal Length와 Sepal width가 서로 비슷한 분포를 가지며 나머지 두 변수는 중복되는 영역을 가진다.

우선 PLS와 RBFN을 기반으로 하는 잔차 보상기의 훈련을 위하여 표 1을 기준으로 각 클래스 당 120개의 표본 데이터를 생성하였고, 테스트를 위하여 각 클래스 당 400개의 데이터를 생성하였다.

학습 데이터의 출력 패턴은 '1 of n' 기법을 사용하여 다음과 같이 정의 하였다.

1) Class 1 (Iris Setosa) =>[1 0 0]
2) Class 2 (Iris Versicolor) =>[0 1 0]
3) Class 3 (Iris Virginica) =>[0 0 1]

표 2는 잔차 보상기의 강건성을 테스트하기 위하여 각 클래스 대해 변수의 평균값과 표준 편차를 각각 ±0.03 , ±0.02 의 범위 내에서 임의로 변화 시킨 데이터 집합의 평균과 표준 편차를 나타낸다.

강건성의 검증을 위하여 표 2의 평균과 표준편차를 기준으로 각 클래스 당 500개의 데이터를 생성하였다.

제안한 알고리즘의 검증을 위해 기존의 PLS, RBFN, PLS/RBF모델 등의 알고리즘과도 비교 하였다.

1.1 PLS

그림 4는 선형 PLS를 이용하여 세 개의 클래스를 분류한 결과를 보여준다. 학습 데이터와 독립성을 가지는 테스트 데이터에 대해 각 차수에서의 분류 결과를 보면 각 클래스는 각 클래스의 출력 패턴 방향으로 이동되어야 하지만 PLS 알고리즘의 선형적 특성 때문에 미리 정의되어 있는 출력 패턴 방향으로 이동이 가능하지 않다. 따라서, PLS의 결과에 대한 분류는 테스트 데이터의 결과 성분을 미리 정의되어 있는 출력 패턴 중에서 가장 유사한 출력 패턴을 결정하고 그 출력 패턴에 할당된 클래스로 분류하였다.

표 3은 PLS의 차수에 따른 학습 데이터와 테스트 데이터의 분류율을 보여준다. PLS의 차수가 3인 경우와 4인 경우, 테스트 데이터의 분류율은 각각 86.33%와 86.5%로 큰 차이를

표 3. PLS 차수에 따른 분류율.

Table 3. Classification rate of PLS.

a	Training data (%)	Test data (%)	PRESS
1	66.67	66.67	0.024
2	83.5	85.41	0.0228
3	85	86.33	0.0227
4	86	86.5	0.0224

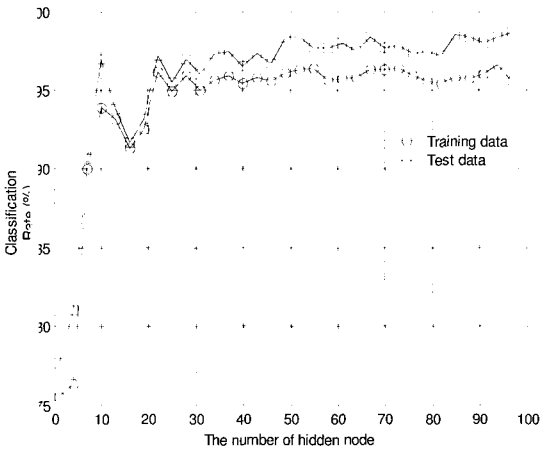


그림 5. RBFN의 은닉 뉴런 수에 따른 분류율.

Fig 5. Classification rate of RBF network.

보이지 않는다. 따라서, 모델의 복잡성과 분류율을 고려하였을 때의 PLS의 차수는 3이다.

1.2 RBFN

그림 5는 각 은닉 뉴런의 수에 따른 분류율을 보여 준다. 은닉 뉴런의 수가 25개 미만일 경우는 뉴런 수의 증가에 따라 분류율도 급격하게 증가하지만 20개의 이상일 경우는 포화 상태를 이룬다.

네트워크의 복잡성과 분류율, 그리고 PRESS를 고려하여 결정된 최적의 은닉 뉴런의 수는 54개이다.

이때의 분류율은 96.33%이고 PRESS는 0.0073이다.

1.3 PLS/RBF 모델

PLS의 선형적인 특성을 보완하기 위해 제안된 PLS/RBF 모델은 PLS의 내부 모델인 회귀 행렬을 RBFN으로 대체하였다.

그림 6은 PLS의 각 차수와 은닉 뉴런의 수의 변화에 따른 PLS/RBFN모델의 분류 결과를 보여준다.

PLS/RBF모델의 최적의 구조를 결정하기 위해서 RBFN 기반의 잔차 보상기에 적용한 비용함수를 이용하였다.

모델의 복잡성 (은닉 뉴런의 수, PLS의 차수), PRESS, 분류율(%)을 모두 고려하였을 때 PLS의 차수와 은닉 뉴런의 수는 각각 3와 60이다. 이때의 분류율은 97.83%이고 PRESS는 0.0062이다.

1.4 RBFN 기반의 잔차 보상기

그림 7은 RBFN을 기반으로 한 잔차 보상기를 이용하여 PLS의 선형적인 제약을 보완한 알고리즘의 결과를 보여준다.

모델의 복잡성 (은닉 뉴런의 수, PLS의 차수), PRESS, 분류율(%)를 모두 고려하기 위하여 (19)의 비용 함수를 이용하였고

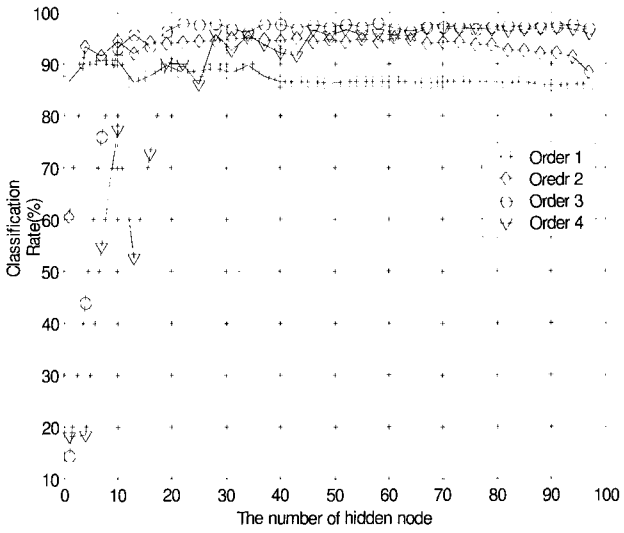


그림 6. PLS/RBFN의 분류율.

Fig. 6. Classification rate of PLS/RBFN.

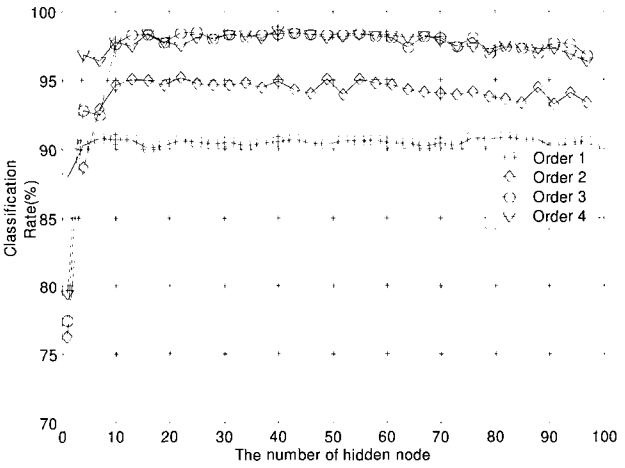


그림 7. 잔차 보상기의 분류율.

Fig. 7. Classification rate of residual compensator.

하였고, 비용 함수의 값이 최소가 될 때의 PLS의 차수와 은닉 뉴런의 수는 각각 3와 27이다. 이때의 분류율은 98.5%이고 PRESS는 0.0051이다.

1.5 각 모델간의 비교

표 4는 PLS , RBFN, PLS/RBF 모델, 잔차 보상기의 PRESS, 분류율, PLS의 차수, 은닉 뉴런의 수를 비교한 표이다.

성능면에서 잔차보상기는 PRESS와 분류율이 각각 0.0051, 98.5%로 RBFN 의0.0073, 96.33%나 PLS/RBF 모델의 0.0062, 97.83%

표 4. 각 분류기의 성능과 복잡성.

Table 4. Performance and complexity of classifiers.

	PRESS	분류율(%)	a	H
PLS	0.0224	86.33	3	
RBFN	0.0073	96.33	4	54
PLS/RBF 모델	0.0062	97.83	3	60
잔차 보상기	0.0051	98.50	3	27

표 5. 각 클래스의 오분류율.
Table 5. Classification error of each class.

	오분류율(%)			
	Total	Class 1	Class 2	Class 3
PLS	13.67	0.25	32.5	8
RBFN	3.67	0	2	9
PLS/RBF 모델	2.17	0	1.5	5
잔차 보상기	1.5	0	1.25	3.25

표 6. 강건성 검증 데이터에 대한 분류기의 성능.
Table 6. Performance of classifiers for robustness.

	PRESS	분류율(%)	a	H
PLS	0.0449	86.93	3	
RBFN	0.0150	95.80	4	54
PLS/RBFN	0.0122	96.93	3	60
잔차 보상기	0.0118	98.33	3	27

표 7. PROBEN 1의 "Glass"테스트 분류율.
Table 7. "Glass" test classification rate from PROBEN 1.

	linear (%)	best (%)	pivot (%)	no sc (%)
Glass	53.96	67.92	60.97	67.30

표 8. PLS 차수에 따른 "Glass"의 분류율.
Table 8. Classification rate of "Glass" by PLS order.

차수	Training data (%)	Test data (%)	PRESS
1	33	33.3	0.0171
2	54	53.5	0.0162
3	57	59.8	0.0156
4	61.5	63.2	0.0156
5	64.67	53.8	0.0155
6	63.3	63.8	0.0155
7	66.67	64.3	0.0155
8	66.5	64.1	0.0155
9	66.7	64.5	0.0155

보다 나은 성능을 보였고 복잡성 면에서도 동일한 PLS 차수를 가지고 있지만 은닉 뉴런의 수는 21개로 RBFN의 54개나 PLS/RBF모델의 60개 보다 적은 수를 가진다.

표 5는 각 클래스의 오분류율을 보여준다. 잔차 보상기의 경우, 클래스 3은 변수의 분포가 다른 두 클래스의 분포와 상의하기 때문에 오분류율은 0%이고 다른 두 클래스의 경우도 다른 알고리즘보다 적은 오분류율을 보인다.

강건성의 검증을 위해 생성된 표 2의 데이터에 대한 실험한 결과는 표 6과 같다.

2. Glass data

PROBEN1의 "Glass" 데이터 집합은 9개의 변수, 6개의 클래스로 구성 되어 있다[13,14].

"Glass" 데이터 집합은 범죄 연구에서 법의학의 필요성에 의해 동기화 되었으며, 유리조각의 굴절률을 비롯한 8종류의 화학 분석을 통하여 float processed building window (CLASS 1), non float processed building window (CLASS 2), vehicle window (CLASS 3), container (CLASS 4), tableware (CLASS 5), head lamp (CLASS 6)을 분류한다.

UCI repository of machine learning database[13]의 "glass" problem의 데이터 집합을 기초로 각 클래스 당 100개의 Training 데이터 집합을 임의로 생성하였고 알고리즘의 검증을 위해 각 클래스 당 500개의 Test 데이터 집합을 생성하였다.

표 7은 "Glass"문제에 대한 PROBEN 1에서 제안한 여러 신경망의 분류오차율에 대한 결과를 보여준다.

'linear', 'best', 'pivot', 'no sc'는 PROBEN 1 에서 구성한 여러 신경망 구조를 나타낸다[14].

9개의 입력 변수와 6개의 출력 변수를 가진 "Glass"에 대한 분류 결과는 표 8과 같다.

2.1 PLS

PLS는 선형 회귀 방법을 이용하기 때문에 클래스1, 클래스 2, 클래스 3의 결과값이 매우 유사하여 분리가 용이 하지 않다. PLS의 차수가 7인 경우 각 클래스의 분류율은 표 9와 같다

2.2 RBFN

은닉 뉴런의 수가 증가함에 따라 분류율이 서서히 증가하여 뉴런수가 70 이상 일 때부터 포화 상태에 이르렀다. 은닉 뉴런의 수가 75일 때 71.70 %의 분류율을 보이며 이때의 PRESS는 0.0104이며 각 클래스의 분류율은 표 10과 같다.

2.3 PLS/RBF 모델

PLS/RBF 모델에서는 PLS의 차수가 5이고 은닉 뉴런의 수가 75일 때 가장 우수한 성능을 나타내었으며, 그때의 분류율은 69.53 %이고 이때의 PRESS는 0.0123이었다. 그림 17에서 보듯이 PLS의 차수가 2인 경우를 제외한 나머지 차수의

표 9. 각 클래스의 분류율(차수 = 7).
Table 9. Classification rate of Each class (order = 7).

	C1	C2	C3	C4	C5	C6	Tatal
분류율(%)	58	16.8	59.2	76.6	89.2	85.8	64.3

표 10. 각 클래스의 분류율.
Table 10. Classification rate of Each class.

	C1	C2	C3	C4	C5	C6	Tatal
분류율(%)	55.4	68.6	52.6	75.2	90.4	88	71.70

표 11. 각 클래스의 분류율(차수 =5, 은닉 뉴런 = 75).
Table 11. Classification rate and PRESS of Each class.

	C1	C2	C3	C4	C5	C6	Tatal
분류율(%)	58.4	65.2	56.6	70.6	81	85.4	69.53

표 12. 잔차 보상기의 분류율과 PRESS.
Table 12. Classification rate and PRESS of RC.

PLS차수	은닉 뉴런 수	분류율 (%)	PRESS
5	57	70.83	0.012
6	78	71.53	0.012
7	75	70.63	0.0119
8	84	72.17	0.0119
9	60	72.03	0.012

PRESS는 거의 동일하였다.

표 11은 PLS의 차수와 은닉뉴런의 수가 각각 5와 75인 경우, 각 클래스의 분류율을 나타낸다.

4.4 RBFN기반의 잔차 보상기

실험을 통해PLS의 차수가 5이상이고 은닉 뉴런의 수가 40이상의 잔차 보상기 구조에서는 대부분 분류율이 70%이상이었고, PRESS 도 0.012에 수렴하였다.

표 12는 PLS의 차수가 5이상 일 때 각 차수에서 가장 우수한 성능을 보이는 은닉 뉴런의 수와 그 성능을 나타낸다. PLS의 차수가 5인 경우, PLS/RBF 모델보다 적은 은닉 뉴런 수를 가지며 더 나은 성능을 보였으며, 9인 경우 RBFN 보다 우수한 성능을 보였다.

V. 결론

본 논문에서는 RBFN에 기반한 잔차 보상기를 이용하여 PLS의 선형적 계산에서 오는 오차를 보정하는 알고리즘을 제안하였다.

4개의 입력변수와 3개의 클래스를 가지는 피셔의 "Irish" 데이터와 9개의 입력변수와 6개의 클래스를 가지는 PROBEN 1의 'Glass" 데이터를 이용한 여러 알고리즘의 모의 실험 결과에서 알 수 있듯이 PLS와 RBFN을 기반으로 하는 잔차 보상기를 이용함으로써 동일한 분류 데이터를 적용한 RBFN이나 PLS/RBF 모델보다 낮은 차원과 은닉 뉴런을 사용함으로써 모델의 복잡성를 줄일 수 있었고 더욱 효율적으로 분류하여 성능을 향상 시킬 수 있었다.

참고문헌

[1] P. Geladi and B. R. Kowalski. "Partial least-squares re-gression: a tutorial", *Anal. Chim. Acta.*, vol. 195, pp. 1-17, 1986.

[2] W. Xun, U. Kruger, B. Lennox, and P. Goulding, "A novel multiblock method using latent vciable partial least squares", *Proc. Conf. on American Control,* vol. 4, pp. 3136-3141, 2001.

[3] U. Kruger, X. Wang, Q. Chen, and S. J. Qin, "An alternative PLS algorithm for the monitoring of industrial process", *Proc. Conf. on American Control,* vol. 6, pp. 4455-4459, 2001.

[4] M. J. Huang, H. Ye, and G. Z. Wang "A new PLS approach with hybrid internal models", *Proc. Int. Conf. on Machine Learning and Cybernetics,* vol. 1, pp. 161-164, 2002.

[5] Y. S. Kim, B. J. Yum, and M. Kim "A hybrid model of partial least squares and artificial neural network for analyzing process monitoring data", *Proc. Int. Joint Conf. on Neural Networks*, vol. 3, pp. 2292-2297, 2001.

[6] D. J. H. Wilson and G. W Irwin, "PLS modeling and fault detection on the Tennessee Eastman benchmark", *Proc.Conf. on American Control*, vol. 6, pp. 3975-3979, 1999.

[7] D. J. H. Wilson, G. W Irwin, and G. Lightbody, "Nonlinear PLS modelling using radial basis functions", *Proc. Conf. on American Control*, vol. 5, pp. 3275-3276, 1997.

[8] J. A. Leonard and M.A. Kramer, "Radial basis function networks for classifying process faults", *IEEE Control Systems Magazine*, vol. 11, pp. 31-38. 1991.

[9] Y. Li and J.-M. Deng, "WAV-a weight adaptation algorithm for normalized radial basis function networks", *Proc. IEEE Int. Conf. on Electronics, Circuits and Systems*, vol. 2, pp. 177-120, 1998.

[10] F. Heimes and B. Heuveln, "The normalized radial basis function neural network", *Proc. IEEE Int. Conf. On Systems, Man, and Cybernetics*, vol. 2, pp. 1609-1614, 1998.

[11] M. D. Brown, G.W. Irwin, and G. Lightbody, "Local model networks for nonlinear internal model control", *EURACO workshop on Robust and Adaptive Control of Integrated systems,* Munich, Germany, 1996.

[12] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[13] http://www.ics.uci.edu/~mlearn/MLRepository.html.

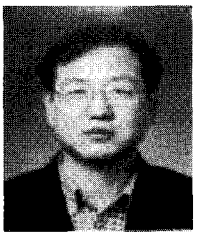
[14] P. Lutzt, "Proben1-a set of neural network benchmark problems and benchmarking rules", *Tech. Rep. 21/94,* Fakultat fur Informatik, Universitat Karlsruhe, 1994. http://ftp.ira.uka.de/pub/papers/techreports/1994/1994-21.ps.Z.

김 경 훈

1975년 4월 24일생. 2002년 울산 대학교 전자공학과 졸업. 2002~현재 울산대학교 전기 전자 정보 시스템 공학부 석사과정. 관심분야는 신경 망 응용, 시스템 고장 검출 및 진단 등.

최 원 호

1956년 2월 9일생. 1978년 연세 대학교 전자공학과 졸업. 동 대학원 석사(1980). 동 대학원 박사 (1990). 1979~1985 제일정밀 공업(주) 연구개발실 과장 대리. 1985~1986 삼성 휴렛 패커드 R & D Project Manager. 1986~현재 울산대학교 전기 전자 정보 시스템 공학부 교수. 관심분야는 신경망, 퍼지 이론, Image Processing, 고장 진단 등.



김 태 영

1962년 7월 14일생. 1986년 울산 대학교 전자공학과 졸업. 동 대학원 석사 (1991.8). 동 대학원 박사 과정 수료 (1994.8). 1987~현재 알칸 대한 (주) Automation Manager. 관심분야는 퍼지시스템, 공정자동화. 고장 검출 및 진단 등.