한국정보전자통신기술학회논문지(jkiiect)'17-12, Vol.10 No.6
논문 17-12-06-594 http://dx.doi.org/10.17661/jkiiect.2017.10.6.594

주성분 분석법을 이용한 회귀다항식 기반 모델 및 패턴 분류기 설계

노 석 범, 이 동 윤*

Design of Regression Model and Pattern Classifier by Using Principal Component Analysis

Seok-Beom Roh, Dong-Yoon Lee*

요 약 본 논문에서는 매우 높은 차원을 가진 데이터에서 의미 있는 특징 벡터 추출하여 입력 공간의 차원을 줄이기 위하여 주성분 분석법을 사용하였다. 주성분 분석법을 이용하여 축소된 차원을 가진 입력 데이터를 이용하여 회귀 다항식의 입력벡터로 사용하는 모델과 패턴 분류기의 설계 방법을 제안하였다. 제안된 모델 및 패턴 분류기는 매우 단순한 구조를 가진 회귀다항식을 기반으로 설계하여 모델 및 패턴 분류기의 과적합 문제를 해결 하고자 하였다. 제안된 설계방법을 적용하여 설계된 모델과 패턴 분류기의 성능을 비교 및 평가하기 위하여, 다양한 기계 학습 데이터 집합을 사용하였다.

Abstract The new design methodology of prediction model and pattern classification, which is based on the dimension reduction algorithm called principal component analysis, is introduced in this paper. Principal component analysis is one of dimension reduction techniques which are used to reduce the dimension of the input space and extract some good features from the original input variables. The extracted input variables are applied to the prediction model and pattern classifier as the input variables. The introduced prediction model and pattern classifier are based on the very simple regression which is the key point of the paper. The structural simplicity of the prediction model and pattern classifier leads to reducing the over-fitting problem. In order to validate the proposed prediction model and pattern classifier, several machine learning data sets are used.

Key Words : Dimension Reduction, Feature Extraction, Pattern Classification, Prediction Model, Principal Component Analysis

1. 서론

현재 우리나라는 컴퓨터, 모바일 기기들의 보급으로 인하여 매우 큰 용량의 데이터들이 매일 개인 사용자들에 의하여 생산되어지고 있다. 이렇게 생산되어지는 데이터를 분석하여 정보를 추출하기 위하여 매우 다양한 시도들이 계속해서 이어지고 있는 상황이다. 이와 같이 개인 사용자들이 생산하는 데이터는 동영상, 이미지 및 텍스트 형태의 비정형 데이터들이 대부분이며, 이와 같은 비정형 데이터를 처리, 가공하여 정보를 추출할 수 있는 새로운 데이터 처리 및 정보추출 기술의 도입이 필요한 시점이다. 특히 획득된 데이터에 기반을 두고 미래를 예측하는 예측모델 [1]과 데이터들을 분류하는 패턴 분류기 [2, 3, 4]에 대한 요구가 늘어나고 있다. 비정형 데이터의 경우, 데이터의 차원이 매우 큰 경우가 대부분이며, 데이터의 차원이 매우 큰 경우, 이를 이용하여 설계되고 구축된 예측 모델 및 패턴 분류기의 성능이 우수하지 않은 단점을 가지고 있다[5]. 본 논문에서는 이와 같은 차원이 큰 데이터를 처리하기 위하여, 차원 축소 알고리즘인 주성분 분석법을 이용하여 데이터의 차원을 축소한다. 주성분 분석법을 적용하여 축소된 입력데이터를 이용하여 예측 모델 및 패턴

*Corresponding Author : Department of Electrical & Electronic Eng. Joongbu University(dylee@jbm.ac.kr)
Received December 01, 2017 Revised December 12, 2017 Accepted December 12, 2017

분류기를 설계하기 위하여, 예측 모델 및 패턴 분류기로 단순한 구조를 가진 회귀 다항식을 이용한다. 일반적으로 예측 모델 및 패턴 분류기의 구조가 복잡하면 복잡할수록 모델 및 분류기의 일반화 성(혹은 예측 성능)이 저하 된다는 특성을 보인다. 이와 같은 일반화 성능저하의 원인이 되는 과적합 (over-fitting) 문제를 해결하기 위하여 단순한 구조를 가진 예측 모델 및 패턴 분류기 설계 방법을 제안한다. 본 논문에서 제안된 주성분 분석기반 회귀 다항식 예측 모델 및 패턴 분류기의 예측 성능 및 일반화 패턴 분류성능을 평가하기 위하여 benchmark 데이터의 일종인 여러 개의 머신러닝 데이터 집합들을 이용한다.

2. 본론

2.1 주성분 분석

고차원 데이터 패턴들을 분석하여 저차원의 데이터로 만들고 우수한 특징벡터들을 추출하기 위하여 주성분 분석을 사용한다. 주성분 분석법은 머신러닝, 패턴인식 등 다양한 분야에서 대표적으로 쓰이는 특징 추출 알고리즘이다[6, 7]. 고차원의 데이터를 저차원으로 축소할 경우, 주성분 분석법을 사용하면 기존의 데이터가 가지고 있는 정보들의 손실을 최소로 하는 주성분으로 축소한다. 아래의 그림은 주성분 분석법의 차원축소 과정을 나타냈다.

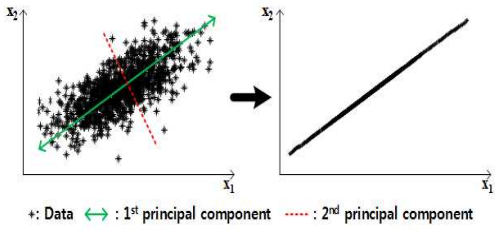

그림 1. 주성분 분석법을 사용한 차원축소
Fig. 1. Dimension reduction using PCA

그림 1을 예시로 간단하게 설명하면, 그림에서 제시된 데이터는 2차원의 데이터이다. 가령 2차원의 데이터를 1차원으로 축소한다고 할 때, 그림 1의 왼쪽처럼 제 1 차 주성분, 제 2 차 주성분으로 나타낼 수 있고, 이러한 두 개의 주성분들은 상호 수직으로 나타난다. 주성분 분석법을 사용하여 입력변수 축소하는 과정은 아래와 같이 진행된다.

[Step 1] 학습하고자 하는 데이터 집합(S)을 구성한다.

$$S = \{\Gamma_1, \Gamma_2, \Gamma_3, \cdots, \Gamma_K, \cdots, \Gamma_N\}$$
$$\Gamma_K = \{k_1, k_2, \cdots, k_D\} \quad (2.1)$$

식 (2.1)에서 N은 학습 데이터의 개수, D는 입력변수의 개수(차원수)를 나타낸다. 그리고 집합 S는 $N \times D$의 행렬로 표현할 수 있다.

[Step 2] 집합S 에서의 평균(Ψ)을 계산한다.

$$\Psi = \frac{1}{N}\sum_{i=1}^{N}\Gamma_i \quad (2.2)$$

[Step 3] 집합S 와 평균Ψ의 차이(Φ)를 계산한다.

$$\Phi_i = \Gamma_i - \Psi \quad (2.3)$$

[Step 4] Φ집합의 공분산행렬(C)을 계산한다.

$$C = \frac{1}{N}\sum_{i=1}^{N}\Phi_i \cdot \Phi_i^T \quad (2.4)$$

이 때, 공분산행렬(C)은 $D \times D$의 행렬의 결과로 나타나게 된다.

[Step 5] 고유값 분석을 통해 공분산행렬의 고유값(λ)과 고유벡터 행렬(U)을 구하고, 고유값이 가장 큰 순서부터 축소하고자하는 차원의 개수(d)만큼 선택하여 변환행렬(W)을 구한다.

$$C = U\lambda U^T$$
$$W = \{w_1, w_2, \cdots, w_K, \cdots, w_D\}, \quad (2.5)$$
$$w_K = \{a_1, a_2, \cdots, a_d\}$$

위 식(2.5)에서 구한 변환행렬(W)은 $d \times D$의 행렬로 표현할 수 있다.

[Step 6] [Step 1]의 학습 데이터 집합S와 [Step 5]의 변환행렬W 의 선형변환에 의해 축소된 특징벡터

(X)를 추출한다.

$$X = SW^T \tag{2.6}$$

식 (2.6)의 계산에 의해 [Step 1]에서 정했던 기존 D차원의 학습데이터 집합$S(N\times D)$가 변환행렬 W^T $(D\times d)$과의 행렬 곱으로 인해 d차원의 행렬인 X $(N\times d)$의 행렬로 차원축소가 일어나는 것을 확인할 수 있다.

2.2 회귀 다항식 기반 예측 모델 및 패턴 분류기

본 논문에서는 예측 모델 및 패턴 분류기의 구조를 단순화 하여 과적합 문제를 해결하기 위하여 단순한 구조를 가진 회귀 다항식 기반 예측 모델 및 패턴 분류기를 제안한다.

2.2.1 회귀 다항식 기반 예측 모델

회귀 다항식을 기반으로 한 예측 모델은 미리 획득된 데이터들로부터 의미 있는 다항식을 추출하고 추출된 다항식을 이용하여 새롭게 주어진 입력데이터와 연관성이 높은 출력데이터를 예측하는 모델이다. 이러한 모델을 설계하기 위하여, 주어진 데이터를 기반으로 다항식 회귀분석을 하고 회귀다항식의 계수를 추정한다. 다항식의 계수를 추정하기 위한 알고리즘으로 일반적으로 사용되는 최소 자승법 (least square estimation)을 적용한다. 표 1은 본 논문에서 사용되는 다항식의 형태를 나타낸다.

표 1. 회귀 다항식의 구조
Table 1. Structure of polynomial regression

Type of polynomial	Polynomial
Linear	$y = a_0 + a_1x_1 + \cdots + a_mx_m$
Quadratic	$y = a_0 + \sum_{i=1}^{m} a_ix_i + \sum_{j=1}^{m}\sum_{k=1}^{m} a_{jk}x_jx_k$
Modified Quadratic	$y = a_0 + \sum_{i=1}^{m} a_ix_i + \sum_{j=1}^{m}\sum_{k=i+1}^{m} a_{jk}x_jx_k$

본 논문에서는 3가지 종류의 다항식을 이용하여 회귀 다항식 기반 예측 모델을 설계한다. 회귀 다항식 예측 모델을 설계하기 위한 목적함수는 식 (2.7)과 같다.

$$J = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{2.7}$$

여기서, n은 데이터의 개수를 의미한다.

1) Linear Type Polynomial

목적함수 식 (2.7)을 행렬과 벡터형태로 표현하면 (2.8)과 같이 표현 할 수 있다.

$$J = (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}}(\mathbf{Y} - \hat{\mathbf{Y}}) \tag{2.8}$$

여기서, $\mathbf{Y} = [y_1\, y_2 \cdots y_n]^T$, $\hat{\mathbf{Y}} = [\hat{y_1}\, \hat{y_2} \cdots \hat{y_n}]^T$.

$$\hat{\mathbf{Y}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \mathbf{X}\,\mathbf{a},$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \mathbf{a} = [a_0\, a_1\, a_2 \cdots a_m]^T$$

여기서, m은 입력변수의 수를 나타낸다.

선형 회귀 다항식의 계수를 추정하기 위하여 최소자승법을 적용하면 아래와 같은 식을 이용하여 다항식의 계수를 추정할 수 있다.

$$\frac{\partial J}{\partial \mathbf{a}} = \frac{\partial (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}}(\mathbf{Y} - \hat{\mathbf{Y}})}{\partial \mathbf{a}} = 0 \tag{2.9}$$

식 (2.9)를 만족하는 다항식의 계수 벡터 $\mathbf{a}$는 식 (2.10)을 이용하여 구할 수 있다.

$$\mathbf{a} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) \tag{2.10}$$

2) Quadratic Type Polynomial

위에 설명한 추정 방법과 동일하게 계수 벡터 $\mathbf{a}$를 구할 수 있으나 행렬 $\mathbf{X}$는 아래와 같이 수정되어야 한다.

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11}\cdots x_{1m} & x_{11}^2\cdots x_{1m}^2 & x_{11}x_{12}\cdots x_{1m-1}x_{1m} \\ 1 & x_{21}\cdots x_{2m} & x_{21}^2\cdots x_{2m}^2 & x_{21}x_{22}\cdots x_{2m-1}x_{2m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1}\cdots x_{nm} & x_{n1}^2\cdots x_{nm}^2 & x_{n1}x_{n2}\cdots x_{nm-1}x_{nm} \end{bmatrix},$$

$$\mathbf{a}=[a_0\, a_1\, a_2\cdots a_m\, a_{11}\, a_{12}\cdots a_{mm}]^T \qquad (2.11)$$

3) Modified Quadratic Type Polynomial

위에 설명한 추정 방법과 동일하게 계수 벡터 $\mathbf{a}$를 구할 수 있으나 행렬 $\mathbf{X}$는 아래와 같이 수정되어야 한다.

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11}\cdots x_{1m} & x_{11}x_{12}\cdots x_{1m-1}x_{1m} \\ 1 & x_{21}\cdots x_{2m} & x_{21}x_{22}\cdots x_{2m-1}x_{2m} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1}\cdots x_{nm} & x_{n1}x_{n2}\cdots x_{nm-1}x_{nm} \end{bmatrix},$$

$$\mathbf{a}=[a_0\, a_1\, a_2\cdots a_m\, a_{11}\, a_{12}\cdots a_{mm}]^T \qquad (2.12)$$

2.2.2 회귀 다항식 기반 패턴 분류기

회귀 다항식기 패턴 분류기를 설계하기 위해서 앞장에서 설명한 3가지 구조를 가진 다항식을 사용하였다. 본 논문에서 사용된 패턴 분류 문제는 가장 기본적인 패턴 분류 문제인 2진 분류 문제이다. 2진 분류 문제에서 사용되어진 회귀 다항식 패턴분류기는 확률 기반 모델로서 아래와 같은 출력 값을 가진다.

1) Linear Type

$$f_i(\mathbf{x})=a_{i0}+\sum_{p=1}^{n}a_{ip}x_p \qquad (2.13)$$

2) Quadratic Type

$$f_i(\mathbf{x})=a_0^i+\sum_{p=1}^{m}a_p^i x_p+\sum_{j=1}^{m}\sum_{k=1}^{m}a_{jk}^i x_j x_k \qquad (2.14)$$

3) Linear Type

$$f_i(\mathbf{x})=a_0^i+\sum_{p=1}^{m}a_p^i x_p+\sum_{j=1}^{m}\sum_{k=j+1}^{m}a_{jk}^i x_j x_k \qquad (2.15)$$

위와 같이 정의된 3가지 형태의 회귀 다항식 기반 확률 모델인 패턴 분류기의 출력은 아래 식과 같다.

$$y_j=\frac{e^{f_j}}{\sum_{q=1}^{c}e^{f_q}} \qquad (2.16)$$

여기서, c는 클래스의 개수를 의미한다.

제안된 회귀 다항식 기반 패턴 분류기의 계수를 추정하기 위하여 사용된 목적함수는 일반화된 크로스 엔트로피 함수이다.

$$CE=-\sum_{k=1}^{n}\sum_{j=1}^{c}y_{jk}\ln\widehat{y_{jk}} \qquad (2.17)$$

여기서, y_{jk}(j=1, 2, ..., c; k=1, 2, ..., n)은 클래스 레이블을 의미한다. $\widehat{y_{jk}}$은 패턴분류기를 추정된 해당 입력 데이터의 클래스 레이블을 나타낸다. 크로스 엔트로피 오차 함수를 목적함수로 사용할 경우, 주어진 학습데이터의 클래스 레이블 정보를 이용하여 새로운 행렬형태의 출력으로 변경하여 사용하여야 한다. 다시 말하면, 일반적으로 사용되는 클래스의 레이블을 나열한 벡터 ($n\times 1$ 열벡터, 여기서 n은 데이터의 개수) 형태의 출력벡터가 아니라 $n\times c$ (c: 클래스 의 종류) 의 행렬로 변환하여 사용하여야 한다. $n\times 1$ 출력 벡터를 $n\times c$ 출력 행렬로 바꾸기 위하여 데이터 패턴의 클래스에 해당하는 열의 요소 값을 1로 설정하고 나머지 열의 요소 값은 0으로 설정한다. 이와 같이 만들어진 행렬의 행 벡터의 요소들의 합은 1이 되어 마치 이산적인 확률 값으로 취급 할 수 있다. 비용함수로 크로스 엔트로피 함수를 사용할 경우 일반적으로 사용되는 오차 제곱합을 사용 할 경우보다 아래와 같은 장점을 가질 수 있다. 임의 학습 데이터의 실제 출력 클래스 벡터 ($\mathbf{y}_k$)와 패턴 분류기를 통해 얻어진 클래스 레이블 추정치($\hat{\mathbf{y}}_k$)를 아래와 같다고 가정하자.

$$\mathbf{y}_k=[0\,1\,0\,0\,0],\hat{\mathbf{y}}_k=\left[\hat{y}_{k1}\,\hat{y}_{k2}\,\hat{y}_{k3}\,\hat{y}_{k4}\,\hat{y}_{k5}\right].$$

비용함수로 일반적으로 사용되는 오차 제곱값을 사용할 경우의 비용함수 값은 아래와 같다.

$$E=\sum_{j=1}^{c}(y_{kj}-\hat{y}_{kj})^2$$
$$=\hat{y}_{k1}^2+(1-\hat{y}_{k2})^2+\hat{y}_{k3}^2+\hat{y}_{k4}^2+\hat{y}_{k5}^2$$

위 식에서는 모든 클래스의 오차들의 제곱들이 더하여져서 최종 오차를 결정한다. 그러나 비용함수로 크로스 엔트로피 함수를 적용하였을 경우는 위 식과는 다른 형태의 오차 값을 갖게 된다. 크로스 엔트로피 함수를 적용한 오차 값은 아래식과 같다.

$$CE=\sum_{j=1}^{c}y_{kj}\ln\hat{y}_{kj}$$
$$=0\cdot\hat{y}_{k1}+1\cdot\hat{y}_{k2}+0\cdot\hat{y}_{k3}+0\cdot\hat{y}_{k4}+0\cdot\hat{y}_{k5}$$
$$=\hat{y}_{k2}$$

크로스 엔트로피 함수의 경우 실제 클래스 레이블이 1인 열의 값만 존재하고 나머지는 모두 사라지게 된다. 다시 정리하여 말하자면, 크로스 엔트로피 함수를 비용함수로 적용할 경우, 해당 입력데이터의 클래스 레이블과 연관 있는 추정 값 $\hat{y}_{kj}$만 강조하여 최적화 할 수 있는 장점을 가지고 있다.

$$A_j(r+1)=(X^TR_jX)^{-1}\{(X^TR_jX)A_j(r)-X^TY_j\} \quad (2.18)$$
$$=(X^TR_jX)^{-1}X^TR_jZ_j$$
$$(\because Z_j=XA_j(r)-R_j^{-1}Y_j)$$

여기서, $X^T=\begin{bmatrix}1 & 1 & \cdots & 1\\ x_{11} & x_{12} & \cdots & x_{1N}\\ x_{21} & x_{22} & \cdots & x_{2N}\\ \vdots & \vdots & \vdots & \vdots\\ x_{m1} & x_{m2} & \cdots & x_{mN}\end{bmatrix}$,

$$R_j=\begin{bmatrix}y_{j1}(1-y_{j1}) & 0 & \cdots & 0\\ 0 & y_{j2}(1-y_{j2}) & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & y_{jN}(1-y_{jN})\end{bmatrix}$$

$$Y_j=[y_{j1}\,y_{j2}\cdots y_{jN}]^T, A_j(r)=[a_{10}^j(r),\ a_{11}^j(r),\ \ldots,\ a_{1n}^j(r)]^T$$

본 연구에서는 비용함수로 크로스 엔트로피 함수를 사용함으로써 연결가중치의 학습을 위해 뉴턴법(Newton's method)을 이용한 비선형 최소자승법을 적용하게 되었다[8]. 비선형 최소 자승법을 이용하여 추정되는 계수는 (2.18)과 같은 정의된 방법을 반복적으로 수행하여 계수 벡터가 임의의 값으로 수렴하게 하여 구할 수 있다.

2.3 실험 및 결과 고찰

제안된 회귀 다항식 예측 모델 및 패턴 분류기의 예측 성능과 패턴 분류기의 일반화 성능을 테스트하기 위하여 3개의 regression 문제에 관련된 머신러닝 데이터 와 5개의 classification 문제 관련 머신러닝 데이터를 사용하였다. 제안된 모델의 예측 성능 및 회귀 다항식 기반 패턴 분류기의 성능 검증을 위해 사용된 데이터는 UCI machine learning repository에서 다운로드하여 사용하였고 제안된 패턴 분류기의 성능 평가를 위하여 10-fold 교차 확인 방법을 통해 검증하였다[8]. 표 2는 제안된 모델과 패턴 분류기의 설계에 필요한 파라미터를 나타낸다.

표 2. 제안된 모델 및 패턴 분류기의 설계 파라미터
Table 2. Design parameters of proposed model

Parameter	Value
Polynomial Type	1 (Linear function) 2 (Quadratic polynomial) 3 (Modified Quadratic polynomial)
Percentage of extracted features	20% 40% 60% 80%

실험에 적용된 기계학습 데이터 집합들의 특징 파라미터는 표 3과 같다.

표 3. 실험에 사용된 기계 학습 데이터 집합
Table 3. Machine learning data sets used in the experiments

Regression Problem		
Data set	Number of features	Number of patterns
Autompg	7	392
Boston Housing	13	506
Red Wine	11	1599

Classification problem			
	Number of features	Number of patterns	Number of classes
Australian	14	690	2
German	24	625	2
Heart	13	270	2
PIMA	8	768	2
Bupa	6	132	2

제안된 모델의 성능을 평가하기 위하여 Root Mean Square Error (RMSE)를 평가지수로 사용하였

다. 사용된 RMSE는 (2-19)와 같다.

$$SE = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \qquad (2\text{-}19)$$

제안된 회귀다항식 기반 패턴 분류기의 성능을 평가하기 위하여 분류율을 사용하였다.

$$\text{분류율} = \frac{1}{n} \sum_{k=1}^{n} f(t_k, \hat{y_k}) \cdot 100 \qquad (2\text{-}20)$$

여기서, $f(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases}$, n은 데이터의 크기를 나타내고, $\hat{y_k}$는 k번째 입력 데이터에 대한 회귀 다항식 기반 패턴 분류기의 출력 값을 의미한다.

표 4는 Regression 문제와 classification 문제에 회귀 다항식 기반 예측 모델과 패턴 분류기의 성능을 평가하기 위하여, 학습 데이터 집합과 테스트 데이터 집합을 이용하여 실험한 결과를 보인다.

표 4. 제안된 모델의 성능비교
Table 4. Results of Comparative analysis

Regression Problem				
Data set	No. of Features	Order	Training Data	Test Data
Autompg	6	2	2.54(0.04)	2.77(0.35)
Boston Housing	13	3	2.51(0.06)	3.73(1.21)
Red Wine	9	3	0.62(0.004)	0.64(0.04)
Classification problem				
Data set	No. of Features	Order	Training Data	Test Data
Australian	12	1	86.50(0.36)	85.51(3.98)
German	9	3	79.43(0.61)	77.3(4.37)
Heart	11	1	85.88(1.33)	84.81(7.50)
PIMA	7	1	77.40(0.68)	77.20(5.33)
Bupa	5	3	73.94(0.92)	70.14(10.1)

표 4의 실험 결과로부터 Regression 문제의 경우 Boston Housing 데이터를 제외하고 나머지 2개의 데이터는 주성분 분석법을 적용한 모델이 우수한 성능을 보임을 알 수 있다. Boston Housing 데이터의 경우 선택된 feature의 수가 11개이고 후반부 구조의 차수가 2일 때 테스트 데이터에 대한 예측 성능이 3.93(1.26) 이다. 이와 같은 예측성능은 정확도 측면에서는 전체 입력변수를 사용한 모델에 비해서 좋지 않다. Classification 문제의 경우, German, Heart 데이터 집합의 경우 주성분 분석을 이용하여 입력공간의 차원을 줄인 데이터를 적용한 패턴 분류기가 우수한 패턴 분류 성능을 보임을 알 수 있다.

3. 결론

본 논문에서는 주성분 분석법을 이용하여 입력공간의 차원을 축소시키고, 우수한 특징 벡터를 추출하여 입력변수로 사용하는 회귀다항식 기반 예측 모델과 패턴 분류기를 설계하였다. 회귀 다항식 기반 모델과 패턴 분류기의 파라미터를 추정하기 위하여, 최소자승법과 비선형 최소자승법을 각각 회귀 다항식 기반 모델과 패턴 분류기 계수 추정에 적용하였다. 제안된 모델 및 패턴 분류기의 성능을 평가하기 위하여 여러 개의 기계학습 데이터 집합을 사용하여 실험을 진행하였다. Regression 문제의 경우 Autompg데이터의 경우 주성분 분석을 통해 얻은 6개의 특징을 적용하여 모델의 출력을 얻을 경우 가장 우수한 성능을 보였다. Boston Housing 데이터의 경우는 13개의 특징들을 모두 사용할 경우 가장 우수한 모델링 성능을 보였으나. Red wine 데이터의 경우에는 총 11개의 특징들 중에서 9개의 특징을 사용 할 경우가 가장 우수 하였다. 데이터 분류 문제의 경우, German과 Heart 데이터의 경우, 전체 입력 데이터를 사용하여 데이터를 분류 한 결과보다 주성분 분석으로 얻은 일부 특징을 이용하는 것이 우수한 결과를 보임을 알 수 있었다. 위에 열거한 실험결과를 통해 주성분 분석법을 이용하여 획득한 특징벡터를 입력으로 사용할 경우 우수한 모델링 성능 및 패턴 분류 성능을 확인 할 수 있었다.

REFERENCES

[1] J. T. Seong, “Analysis of Signal Recovery for Compressed Sensing using Deep Learning Technique,” The Korea Institute of Information & Electronic Communication Technology, Vol. 10 , no. 4, pp. 257-267, 2017.

[2] I.-H. Lee, T.-S. Choi, "Shape from focus algorithm with optimization of focus measure for cell image," The Korea Institute of Information & Electronic Communication Technology, Vol. 3, pp. 8-13, 2010.

[3] E. H. Jeong, and B. K. Lee, "A Design of Customized Market Analysis Scheme Using SVM and Collaboration Filtering Scheme," The Korea Institute of Information & Electronic Communication Technology, vol. 9, no. 6, pp. 609-616, 2016.

[4] J. H. Son, S. Y. Kim, "Texture Classification Based on Gabor-like Feature," The Korea Institute of Information & Electronic Communication Technology, vol. 10, no. 2, 147-153, 2017.

[5] G. B. Hwang, Q. U. Zhu, C. K. Siew, "Extreme learning machine: theory and applications," Neurocomputing, vol. 70, pp. 489-501, 2006.

[6] I. T. Jolliffe, "Principal Component Analysis", second edition, Springer-Verlog, 2002.

[7] M. Imaizumi, and K. Kato, "PCA-based estimation for functional linear regression with functional responses," Journal of multivariate analysis, vol. 163, pp. 15-36, 2018.

[8] M. Lichman, "UCI Machine Learning Repository", 2013, http://archive.ics.uci.edu/ml.

저자약력

이 동 윤(Dong-Yoon Lee) **[정회원]**

- 1990년 2월 : 연세대학교 전기공학과 (공학석사)
- 2001년 2월 : 연세대학교 전기전자공학과 (공학박사)
- 2001년 3월 ~ 2002년 2월 : 원광대학교 BK21교수
- 2002년 3월 ~ 현재 : 중부대학교 전기전자공학과 교수

<관심분야> IT 융합

노 석 범(Seok-Beom Roh) **[정회원]**

- 1997년 2월 : 원광대학교 컴퓨터공학과 (공학석사)
- 2006년 8월 : 원광대학교 제어공학과 (공학박사)
- 2016년 4월 ~ 현재 : 수원대학교 연구교수

<관심분야> 계산 지능, 기계 학습