논문 04-29-5C-13 한국통신학회논문지 '04-5 Vol.29 No.5C

AutoEncoder와 FCM을 이용한 불완전한 데이터의 군집화

정회원 박 동 철*, 장 병 근**

Clustering of Incomplete Data Using Autoencoder and Fuzzy c-Means Algorithm

Dong-Chul Park*, Bung-Geun Jang** *Regular Members*

요 약

Autoencoder와 Fuzzy c-Means 알고리즘을 이용하여, 불완전한 데이터의 군집화를 위한 알고리즘이 본 논문에서 제안되었다. 본 논문에서 제안된 Optimal Completion Autoencoder Fuzzy c-Means (OCAEFCM)은 손상되어 불완전한 데이터의 최적 복원과 데이터의 군집화를 위해 Autoencoder Neural Network (AENN) 과 Gradient-based FCM (GBFCM)을 이용하였다. OCAEFCM 의 성능평가를 위해 IRIS 데이터와 금융기관에서 취득한 실제 데이터를 사용하였다 기존의 Optimal Completion Strategy FCM (OCSFCM)과 비교했을 때, 제안된 OCAEFCM 이 OCSFCM 보다 18%-20%의 성능 향상을 보여준다.

Key Words : Neural Network, OCAEFCM, GBFCM

ABSTRACT

Clustering of incomplete data using the Autoencoder and the Fuzzy c-Means(FCM) is proposed in this paper. The proposed algorithm, called Optimal Completion Autoencoder Fuzzy c-Means(OCAEFCM), utilizes the Autoencoder Neural Network (AENN) and the Gradiant-based FCM (GBFCM) for optimal completion of missing data and clustering of the reconstructed data. The proposed OCAEFCM is applied to the IRIS data and a data set from a financial institution to evaluate the performance. When compared with the existing Optimal Completion Strategy FCM (OCSFCM), the OCAEFCM shows 18%-20% improvement of performance over OCSFCM.

Ⅰ. 서 론

일반적으로 다차원의 숫자로 주어지는 데이터들은 각 차원의 값이 대상의 특징값을 나타내는데, 때때로 이러한 특징값들 중의 일부는 종종 손실된 상태로 주어진다. 이러한 불완전한 데이터에서 손실된 데이터의 값을 추정하고, 군집화 시키는 것에 대한 연구가 1962년부터 시작되었으며, Jainn과 Dubes에 의해 정리되었다[1-3]. Hathaway와 Bezdek은 Fuzzy C-Means (FCM)의 변형된 형태인 Optimal Completion Strategy FCM (OCSFCM)을 제안하고 불완전한 데이터의 손실된 값들을 추정하는 방법에 대해 연구하였다[4]. 그러나, 이 OCSFCM은 완전한 원본 데이터에 대한 군집 중심값의 정보를 초기에 필요로 하기 때문에, 실제 원본 데이터를 알지 못한다면 불완전한 데이터를 군집화 시킬 수 없다. 따라서, 본 논문에서는 원본 데이터를 모르는 상황에서 불완전한 데이터를 군집화 시키는 방법에 대해 제안한다. 본 논문에서 제안되는 알고리즘은 Autoencoder Neural Network (AENN)을 사용하여

* 명지대학교 정보공학과
논문번호 : 030534-1202, 접수일자 : 2003년 12월 2일
※본 연구는 한국과학재단 지역대학우수과학자 지원연구(R05-2003-000-10992-0)의 지원으로 수행되었음.

불완전한 데이터의 손실부분을 추정하고, 그 결과를 Gradient-Based FCM (GBFCM)에 인가하여 군집화를 수행한다.

본 논문의 구성은 다음과 같다. 기존의 알고리즘인 FCM[5,6], GBFCM[7], OCSFCM[4]에 대한 설명을 제 2장에서, AENN[9]과 본 논문에서 제안하는 OCAEFCM에 대하여 각각 제 2장, 3장에서 기술한다. 제 4장은 IRIS 데이터[10]에 대한 각 알고리즘 별 불완전한 데이터의 군집화에 대한 실험 및 결과를 보이며, 실제 금융관련 데이터를 적용하여 손실된 신상 정보를 가지는 데이터를 이용해 개인의 신용 등급을 결정하는 응용에 대한 결과를 보여준다. 마지막으로 제 5장에서는 전체적인 요약과 결론을 내린다.

Ⅱ. 기존의 알고리즘

2.1 FCM 알고리즘

Bezdek 에 의해 제안되어 널리 사용되고 있는 FCM에서 목적함수 J_m은 다음의 식으로 정의된다[5,6].

$$J_m(U,v)=\sum_{k=1}^{N}\sum_{i=1}^{C}(\mu_i(\overrightarrow{x_k}))^m(d_i(\overrightarrow{x_k}))^2 \quad (1)$$

여기서 $d_i(\overrightarrow{x_k})^2$은 입력 $\overrightarrow{x_k}$패턴 와 군집 중심값인 $\overrightarrow{v_i}$사이의 거리로서 Euclidean 거리를 사용하고, $\mu_i(\overrightarrow{x_k})$는 군집 i와 입력 패턴 $\overrightarrow{x_k}$사이의 멤버십 등급, m은 가중지수, C는 군집의 수이고, N은 입력데이터의 수이다. Bezdek 은 이 목적 함수를 최소화시키기 위한 조건을 다음과 같이 제안하였다[5,6].

$$\mu_i(\overrightarrow{x_k})=\frac{1}{\sum_{j=1}^{C}\left(\frac{d_i(\overrightarrow{x_k})}{d_j(\overrightarrow{x_k})}\right)^{\frac{2}{m-1}}} \quad (2)$$

$$\overrightarrow{v_i}=\frac{\sum_{k=1}^{N}(\mu_i(\overrightarrow{x_k}))^m\overrightarrow{x_k}}{\sum_{k=1}^{N}(\mu_i(\overrightarrow{x_k}))^m} \quad (3)$$

2.2 GBFCM 알고리즘

Bezdek 에 의해 제안되어 널리 사용되고 위의 식 (2) 와 (3)에서 보듯이, FCM은 중심값을 갱신하기 위해 모든 데이터를 사용한다. 그러나, GBFCM은 오직 하나의 데이터를 가지고 중심값을 갱신할 수 있도록 제안되었다[7] 즉, 하나의 데이터 $\overrightarrow{x_i}$와 군집 중심값 $\overrightarrow{v_j}$ $(j=1,2,\cdots,K)$를 갖는 K개의 군집이 주어질 때, 목적함수를 다음과 같이 정의되며, 멤버십 등급의 합은 항상 1로 제한된다.

$$J_i=\mu_{1i}^2(\overrightarrow{v_1}-\overrightarrow{x_i})^2+\cdots+\mu_{ci}^2(\overrightarrow{v_c}-\overrightarrow{x_i})^2 \quad (4)$$

$$\mu_{1i}+\mu_{2i}+\cdots+\mu_{ci}=1 \quad (5)$$

위의 조건에 최적의 군집중심값을 찾기 위한 조건으로 기울기 감소 방법을 다음과 같이 적용할 때,

$$\overrightarrow{v}_{k+1}=\overrightarrow{v}_k-\eta\mu_{ki}^2(\overrightarrow{v_k}-\overrightarrow{x_i}) \quad (6)$$

최적의 중심값을 위한 필요조건은 다음과 같다. 즉,

$$\frac{\partial J_i}{\partial \mu_i}=0 \quad (7)$$

이며, 이를 적용한 경우의 멤버십 등급은 다음의 경우로 요약된다.

$$\mu_i(\overrightarrow{x_k})=\frac{1}{\sum_{j=1}^{c}\left(\frac{d_i(\overrightarrow{x_k})}{d_j(\overrightarrow{x_k})}\right)^2} \quad (8)$$

FCM 과 GBFCM 은 둘 다 목적함수를 사용한다는 것이 공통점이지만, FCM은 중심값을 갱신하기 위해, 모든 데이터가 사용되는 반면, GBFCM은 주어진 각각의 데이터에 대해 식 (6) 과 (8)을 교대로 적용하여, 최적의 중심값과 멤버십 등급을 구하게 되므로, 학습시간과 정확성에서 더욱 효율적인 결과를 얻을 수 있다.[7,8]

2.3 OCSFCM 알고리즘

불완전한 데이터를 군집화 시키는데 효과적인 OCSFCM은 기존의 FCM 목적 함수 J_m의 값을 최소화시키기 위한 조건에 손실된 특징값들을 변수로 하는 새로운 조건을 더하여 목적 함수 J_m을 최소화시킴으로서 최종적으로 손실된 부분을 원본의 값으로 추정/복원시킨다. 즉, OCSFCM은 먼저 전체 데이터가 손실된 데이터가 없다는 것을 가정하고, 전체 데이터에 FCM을 사용하여 각 군집의 중심값을 구한다. 구해진 군집의 중심값을 이용하여, 불완전한 데이터의 손실된 특징값을 초기화시키고, 반복적으로 FCM을 이용해 손실된 데이터 x_k의 j번째 특징값 x_{kj}을 다음의 식으로 구하는 과정을 추가하여 손실된 데이터를 복원시키고 최적화된 데이터의 군집 중심값을 구하는 것이다[4].

$$x_{kj} = \frac{\sum_{i=1}^{C} (\mu_i(\overrightarrow{x_k}))^m v_{ij}}{\sum_{i=1}^{C} (\mu_i(\overrightarrow{x_k}))^m} \qquad (9)$$

Ⅲ. 제안된 알고리즘

OCSFCM은 불완전한 데이터를 군집화하는 과정에 사용되는 초기 군집 중심값 $\overrightarrow{v_i}$를 정하고, 손실된 데이터의 특징값 x_{kj}를 초기화하는데 있어 완전한 원본 데이터의 군집 중심값을 사용하였다. 이는 실제로 완전한 원본 데이터에 대한 군집 중심값의 정보를 알지 못하는 상황에서는 불가능하기 때문에, 본 논문에서는 AENN의 학습을 통해 손실된 데이터를 원본 데이터로 추정/복원시키는 방법을 제안한다[9].

3.1 AENN

그림 1에서 보는 것과 같이 AENN은 중요한 두 가지 특징을 가진다.

먼저, "자동 연상"의 성질을 가지는 것으로, AENN의 입력값은 학습과정에 필요한 목적값과 동일한 값을 가진다는 것이다. 즉, 입력 패턴과 동일

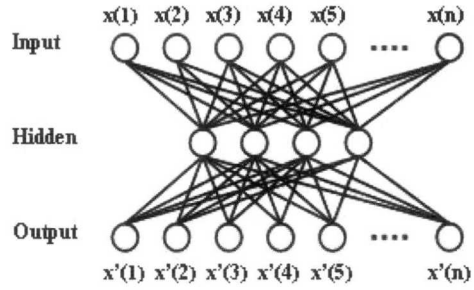

그림 1. Autoencoder 신경망의 기본 구조

한 형태의 목적 값으로 학습을 한다. 다른 하나는, AENN의 구조로 나비 모양을 이루고 있어 "병목"으로 불리는 은닉층이 존재하는데, 은닉층의 뉴런의 수는 입·출력층의 뉴런 수보다 항상 작다. 특히, AENN 학습 후에 입력 패턴의 정보는 "병목" 은닉층의 형태로 압축되어 표현된다[9]. AENN의 기본적인 학습의 개념은 오류 역전파 학습 알고리즘으로 주어진 입력 데이터에 대해 신경망의 출력이 목표 값에 가까이 접근하도록 가중치의 최적값을 구하는 것이다[11]. 본 논문에서는 이와 같은 AENN의 성질을 이용해 불완전한 데이터를 추정/복원시키는데 그림 2와 같이 적용한다. 즉, 전체 데이터에서 완전한 데이터 패턴들만을 이용해 GBFCM 군집화 과정을 통해 완전한 데이터 패턴들의 군집 중심값을 구하고, AENN을 먼저 학습시킨다. 손실된 특징값을 가지는 불완전한 데이터 패턴들에 대한 복원을 위해서 GBFCM 군집화 과정에서 생성된 군집 중심값을 이용하여 손실된 특징값을 초기화 시킨 후, 학습된 AENN을 이용해 손실된 특징값을 순차적으로 반복 추정한다.

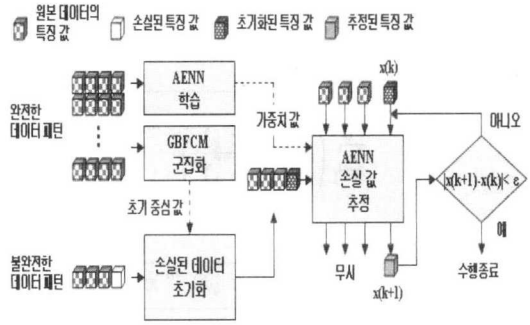

그림 2. AENN을 이용한 손실값 추정

3.2 OCAEFCM

본 논문에서는 GBFCM의 군집화 과정과 학습된 AENN를 이용한 손실된 데이터 추정/복원 과정을 조합하여 불완전한 데이터를 군집화 시키는 알고리즘을 제안한다. AENN의 학습된 가중치를 이용하여 손실된 데이터를 추정/복원한 후에 GBFCM의 군집화 과정을 통해 군집 중심값을 구하는데, 이것은 완전히 AENN의 추정/복원 과정이 종료된 후에만 데이터의 군집화를 수행할 수 있다. 이때, AENN의 출력에 의한 추정/복원값이 군집화에 많은 영향을 미치기 때문에, 제안된 알고리즘에서는 손실된 특징값을 순차적으로 반복/추정하는 것이 아니라 그림 3에서와 같이 한번 수행시마다 AENN을 통해 생성된 추정/복원값을 그대로 군집 중심값을 계산하는데 반영한다. 여기서, 수행과정 시 중심값의 변화가 선정값 보다 작으면 수행은 종료된다.

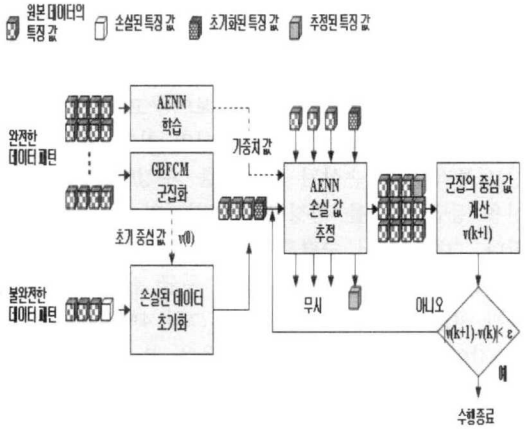

그림 3. OCAEFCM의 수행 과정 블록도

표 1. OCAEFCM의 수행 코드

```
Procedure main()
 Read c, ε, m   [c is initialize cluster , ε is small value , m is weighting exponent (m∈1,⋯∞)]
  While (error > ε)
    e ← 0
     While (input file is not empty)
        Read one datum  x_i
        μ_i(x_n) = 1 / Σ_{j=1}^{c} ( d_i(x_n) / d_j(x_n) )^2
        [Update OCAEFCM membership grade]
        If (missing input file)
          AENN Estimation()
        end if
        v_{n+1} = v_n − η μ^2_∋ (v_n − x_i)
            [Update OCAEFCM center value]
        e ← e + v_i(n+1) − v_i(n)
     end while
        error ← e
  end while
 Output  μ_i , v_{n+1}
 [Final membership grade and center value]
end main()
```

Ⅳ. 실험 및 결과

4.1 IRIS 데이터 문제

본 논문에서는 150개의 패턴을 가지는 4차원(총 600개의 특징값)의 IRIS 데이터[11]에 대한 실험을 하였다. IRIS 데이터는 50개의 패턴 씩 총 3개의 군집으로 구성되어 있다. 인위적으로 IRIS 데이터의 특징값들을 무작위로 손실을 시켰으며, 이 때, 손실된 데이터 패턴의 비율과 완전한 데이터 패턴의 비율을 10%에서 60%까지로 하여 각각의 조건에 대하여 실험을 하였다. 여기서 손실된 데이터의 패턴은 적어도 한개 이상의 손실된 특징값을 가지는 것으로, 이러한 실험방법은 Bezdek의 OCSFCM에 이용된 것이다[4]. 본 논문에서는 IRIS 데이터를 무작위로 불완전한 데이터로 만든 후, 군집화 시키는 실험을 각 알고리즘 별로 수행하여 그 성능을 비교하였다. 실험은 총 50번에 걸쳐 수행을 하였고, 한 번

표 1. 각 알고리즘 별 평균 원본 변형오차

	데이터 손실률					
	10%	20%	30%	40%	50%	60%
OCSFCM	0.067	0.074	0.085	0.086	0.091	0.109
AENN	0.055	0.061	0.073	0.071	0.079	0.085
OCAEFCM	0.043	0.052	0.063	0.067	0.075	0.081

수행 시 같은 형태의 불완전한 데이터를 이용해 각 알고리즘 별로 수행하여, 최종적으로 50번의 수행 결과에 대한 성능의 평균을 구했다. 표 1은 각 알고리즘 별 평균 원본변형오차(Mean Prototype Error : MPE)를 보여준다[12]. 표 1에서 보는 것과 같이 성능 비교기준의 하나인 MPE는 OCAEFCM이 OCSFCM에 비해 약 20%정도의 감소를 보여준다. OCSFCM에 비해 OCAEFCM이 불완전한 데이터를 군집화시키는데 있어서 보다 원본 데이터의 군집 중심값에 근접하고 있다는 결과를 보여주고 있는 것이다.

4.2 신용 등급 결정을 위한 금융 데이터 문제

금융 관련 회사에서 개인의 신용 등급의 결정은 심사자의 경험에 의해 이루어진다. 특히, 의뢰자의 최초 거래 시에는 의뢰자의 실적이 없으므로 신용 등급을 판정하는데 기준을 정하기가 어렵다. 본 논문에서는 실제로 개인의 신상 정보 중에서 누락되는 정보나, 의심되는 정보가 있는 상황에서도 개인의 신용 등급을 적절히 판정할 수 있도록 AENN을 이용해 손실된 값을 추정/복원한다. 실험에서 사용된 표본 데이터들은 국내 한 금융 회사의 근래 가입자들 중 최근 3년 동안 한 번이라도 이용한 실적이 있는 표본 중에서 무작위로 3,000명의 데이터 표본을 추출한 것이다. 개인의 신상 정보 중에서 신용 등급을 판정하는데 사용된 정보들은 생일, 성별, 주거현황, 주거 년 수, 세대주관계, 직장코드, 직위, 근속 년 수, 연소득, 배우자, 결제방법, 자가용유무, 자녀수(남), 자녀수(여)의 정보들로서 이들 정보들을 인위로 손실시킨 후, 각 알고리즘 별로 손실된 데이터를 추정, 복원시켜 최종적으로 신용 등급을 판정하도록 하였다. 신용등급은 1 등급부터 20 등급으로 주어진다. 주어진 3,000명의 데이터 표본 중에 무작위로 2,700명의 데이터 패턴을 추출하여 완전한 원본 데이터 패턴으로 사용하였으며, 나머지 300명의 데이터 패턴에 대해서는 손실 대상에 해당되는 부분에 대해 무작위로 손실시켜 데이터로 사용하였다. 무작위로 추출한 2,700명의 완전한 원본 데이터 패턴은 오류 역전파 알고리즘 학습 과정을 수행하여 나머지 300명의 손실된 데이터 패턴의 신용 등급을 판정하는데 사용되었다. 그리고, 각 알고리즘들은 최종적으로 손실된 데이터 값을 추정/복원시켜 신용 등급을 판정하도록 하였다. 즉, 300명의 신상 정보 데이터가 손실이 되지 않았을 때, 가지고 있던 신용 등급 정보(목적값)와 손실된 후 신용 등급을 추정하여 어느 정도 목적 값에 근접하는지를 결과로 나타내었다.

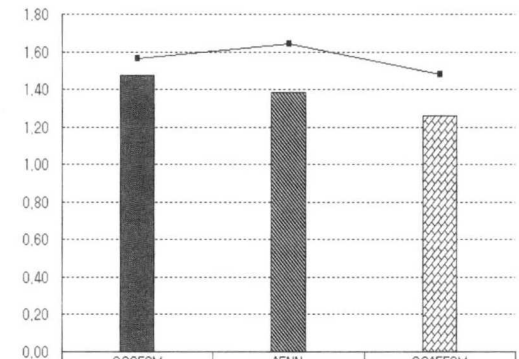

그림 4. 각 알고리즘 별 신용 등급 판정 오차율

실험은 무작위로 추출되는 50 쌍의 학습데이터/검증데이터에 대해 수행되었다. 그림 4는 각 알고리즘 별 신용 등급 판정 오차율을 보여주고 있다. 신용 등급 판정 오차율은 원본 데이터의 최초 신용 등급 정보(목적값)와 손실된 데이터를 추정/복원시킨 후 신용 등급 정보를 판정한 결과값과의 차이라 정의한다. 그림 4에서 수평축은 사용한 알고리즘의 종류를 나타내며, 아래 값들은 신용 등급 판정 평균오차율과 표준편차 값을 나타낸다. 그림 4에서 보는 것과 같이 신용 등급 판정 오차율은 본 논문에서 제안하는 OCAEFCM의 결과가 1.2등급 정도로 OCSFCM의 1.47등급에 비해 약 18.37%의 성능 향상을 보인다.

V. 결 론

손실된 특징값들이 존재하는 불완전한 데이터에 대한 효율적인 군집화를 위해 Hathaway와 Bezdek에 의해 제안된 FCM을 이용하는 방법인 OCSFCM은 완전한 원본 데이터에 관한 초기 정보를 필요로 하는데, 실제의 상황에서는 완전한 원본 데이터에 대한 군집 중심값의 정보가 보통 주어지지 않는다. 이러한 OCSFCM의 문제점을 해결하기 위하여, AENN 과 FCM 알고리즘을 이용하여, 불완전한 데이터의 군집화를 위한 알고리즘이 본

논문에서 제안되었다. 본 논문에서 제안된 OCAEFCM은 AENN의 불완전한 데이터 복원 능력과 GBFCM의 군집화 능력을 적절히 이용하였다. OCAEFCM 의 성능평가를 위해 통계학과 패턴인식 분야에서 알고리즘의 성능검증을 위해 널리 사용되는 IRIS 데이터와 금융기관에서 취득한 실제 데이터를 사용하였다. 제안된 OCAEFCM은 기존의 OCSFCM에 비해 IRIS 데이터의 경우 약 20%의 성능향상이, 실제 금융 데이터의 복원과 신용등급판정에 응용한 결과에서는 15% 이상의 성능향상이 보여졌다. 따라서 본 논문에서 제안되는 OCAEFCM은 데이터 복원과 군집화에 효율적으로 적용될 수 있음을 알 수 있었다.

참 고 문 헌

[1] J. Dixon, "Pattern Recognition with partly Missing Data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, pp. 617-621, 1979.

[2] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewodd Cliffs, NJ: Prentice-Hall, 1988.

[3] G. Sebestyen, *Decision-making Processes in Pattern Recognition*. New York: Macmillan, 1962.

[4] R. Hathaway, and J. Bezdek, "Fuzzy c-Means Clustering of Incomplete Data," *IEEE Tr. Syst., Man, Cybern.*, vol. 31, pp. 735-744, 2001.

[5] J. Bezdek, "A convergence theroem for the fuzzy ISODATA clustering algorithms," *IEEE Tr. Pattern Anal. Mach. Int.*, vol. 2, pp. 1-8, 1980.

[6] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[7] Dong C. Park and Issam Dagher , "Gradient Based Fuzzy c_means (GBFCM) Algorithm," *Proc. of IEEE ICNN-94,* vol.3, pp1626-1631,1994

[8] K. Fukunage, *Introduction to Statistical Pattern Recognition,* Academic Press, Inc, 2nd edition, 1990.

[9] B. Thompson et al., *"Implicit Learning in Autoencoder Novelty Assessment," Proc. of IEEE IJCNN-2002*, pp. 2878-2883, 2002.

[10] J. Bezdek et al., "Will the real IRIS data please stand up?," *IEEE Tr. Fuzzy Syst.*, vol. 7, pp. 368-369, 1999.

[11] R. Reed and R. Marks II, *Neural Smithing : Supervised Learning in Feedforward Artificial Neural Network*. Cambridge, MA: MIT Press, 1999.

[12] M. Manning, and S. Banda, "Approxi- mate structured singular value compu- tation via Frobenius norms," *Proc. of IEEE Int. Conf on Systems Eng.*, pp. 103-106, 1989.

박 동 철(Dong-Chul Park) 정회원

1980년 2월 : 서강대학교 전자공학과(공학사)
1982년 2월; 한국과학기술원 전기 및 전자공학과 (공학석사)
1990년 6월: Ph.D. in Electrical Engineering Univ. of Washington (Seattle)
1990년 8월 - 1994년 2월: 조교수, Florida Int'l Univ. Dept. of Eelct. and Comp. Eng.
1994년 3월 - 현재: 명지대학교 정보공학과 교수
1997년-2000년: IEEE Tr. on Neural Networks, Associate Editor
1999년 - 현재: IEEE Senior Member
<주관심분야> 신경망 알고리즘 개발, 음성인식, 신경망의 금융공학에의 응용

장 병 근(Bung-Geun Jang) 준회원

2002년 2월: 명지대학교 제어계측공학과(학사)
2004년 2월: 명지대학교 정보제어공학과(석사)
2004년 3월~ 현재: (주) KDE 컴 근무

<주관심분야> 신경망 알고리즘개발, 음성인식, 패턴인식