논문 99-24-6T-13 한국통신학회논문지 '99-6 Vol.24 No.6T

변형된 유전자 알고리즘을 이용한 참조벡터의 최적화

정회원 김 대 극*, 이 재 곤*, 이 상 희**

Optimization of Reference Vectors using a Modified Genetic Algorithm

Dae-Keuk Kim*, Jae-Kon Lee*, Sang-Hee Lee** *Regular Members*

요 약

본 논문에서는 패턴인식에서의 오분류의 수를 최소화하기 위한 참조벡터 생성에 관한 것으로 신경회로망에서의 학습기능과 유전자 알고리즘의 최적화 방법들의 특성을 이용한 변형된 유전자 알고리즘을 제안한다. 유전자 알고리즘은 자연환경에서의 적자생존의 경쟁을 통해 개체 집단을 환경에 적응하도록 진화시켜 나가는 것에 기초하여 제안된 적응 탐색기법의 하나이다. 따라서 방대하고 복잡한 함수에 대하여 전역적 최적화의 특성을 가지고 있다. 그러나 왜곡도를 최소화하는 참조벡터의 설정은 가장 가까운 참조벡터로 분류가 이루어지는 NN에 근거한 패턴인식기로서는 적합하지 않다. 이에 신경회로망에서의 오인식 율를 최소화하는 최적화 방법의 하나인 LVQ 모델의 학습조건을 갖는 기능을 추가하면 각 모델들의 장점으로부터 패턴 분류문제에서 인식률 향상에 기여할 것이라는 것에 기초하여 학습기능을 갖는 변형된 유전자알고리즘을 제안하였다. 제안된 알고리즘의 성능 평가를 위해 모의 데이터들을 사용하여 기존방법들과의 비교실험을 통해 향상된 인식성능 결과를 얻을수 있었다.

ABSTRACT

In this paper, we present a modified genetic algorithm to minimize misclassification rate for determining the reference vectors. Genetic Algorithms(GAs) are adaptive methods which may be used to solve search and optimization problems based on the genetic processes of biological organisms. This algorithm is to minimize a distortion ,but didn't minimize misclassification rate for the classifier problem. LVQ is a supervised classifier, which approximates the Bayesian decision boundary. The modified genetic algorithm use that combines features taken from conventional GA and learning conditions of LVQ(Learning Vector Quantization). Experimental result showed to compare the performance of our method to conventional ones.

I. 서 론

정보화 시대에 따른 컴퓨터의 이용은 단순한 계산기능 이외에 인간이 보다 쓰기 편리하고 효율적인 방향으로 확산되고 있다. 이러한 추세는 인공지능(Arificial Inteligence), 신경회로망(Neural Network),퍼지이론(Fuzzy Theory) 그리고 최근에 유전자 알고리즘(Genetic Algorithm)에 대한 연구가 활발해 지면서 인간의 시각과 청각을 이용한 문자, 음성, 화상 등의 인식능력을 컴퓨터에 부여하고자 하는 패턴인식 분야의 연구가 폭넓게 이루어지고 있으며 부분적으로 실용화되고 있는 추세이다. 이러한 패턴에 대한 인식접근 방법으로 템플릿에 기초한 방법들이 많이 사용되어 왔다. 이것은 주어진 패턴에 대해 기준 템플릿을 만들어 인식하고자 하는 패턴과 비교를 통한 인식방법이다. 이러한 방법을 사용하기 위해서는 주어진 패턴으로부터 특징벡터를 추출하는 전처리 과정과, 전처리 과정을 통해서 구해진 특징벡터를 기준 참조벡터들과 비교 하게 되는 인식과정이 필요하게 되는데 이러한 특징벡터의

* 한림정보산업대학 전자통신과
** 강원대 전기전자공학부 교수

논문번호 : 99020-0430, 접수일자 : 1999년 4월 30일

선택이나 참조벡터의 설정, 참조벡터와의 비교방법들은 인식성능을 크게 좌우하게된다.

본 논문은 오분류의 수를 최소화하기 위한 참조벡터 설정에 관한 것으로, 1970년대 미국의 John Holland에 의해 제안된 유전자 알고리즘(Genetic Algorithms)[1]과 LVQ[2]의 학습조건들을 결합한 형태의 새로운 알고리즘을 제안하였다. 유전자 알고리즘은 적응 탐색 기법의 하나로 자연환경에서 종(species)들이 저자 생존의 경쟁을 통해 개체 집단(population)을 환경에 적응하도록 진화시켜 나가는 것에 기초하여 제안된 알고리즘이다. 유전자 알고리즘은 다른 종류의 탐색 알고리즘과는 달리 방대하고 복잡한 함수에 대하여 전역적인(global optimization) 최적화를 할 수 있는 장점을 가지고 있기 때문에 최근에 크게 주목을 받으며 기존의 알고리즘으로 해결하기 힘든 여러 분야에 적용되고 있다[5-7]. 그러나 이러한 유전자 알고리즘은 지역적 최소치(local minimum)에 빠지지 않고 전역적인 최적해를 발견할 가능성은 높지만, 신경회로망에서의 학습기능이 갖고 있는 미세 조정되는 지역적 탐색 메카니즘이 존재하지 않으므로 최적해 부근의 탐색에서는 수렴속도가 급격히 떨어진다는 단점이 있다[3]. 이에 미세 조정이 가능한 신경회로망의 학습조건기능을 추가하여 유전자 알고리즘이 갖는 미세조정 부분의 단점을 보완하고자 하는데 목적이 있다. 특히 이러한 패턴 분류 문제 있어서는 참조벡터의 왜곡도를 최소화하는 최적화보다는 오인식률을 최소화하는 최적화 방법이 전체 패턴 인식시스템의 성능을 좌우하게 된다. 이에 신경회로망에서 지도학습에 근거한 패턴인식기의 일종으로 오인식(misclassification)의 수를 최소화하기 위해 Bayes 결정경계면(decision boundary)를 근사화하여 높은 분류능력을 갖는 LVQ 알고리즘의 학습조건들을 기존 GA에 추가시킨 변형된 유전자 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 유전자 알고리즘과 변형된 유전자 알고리즘에 사용된 연산자의 정의 및 알고리즘을 소개하고, 3장에서 모의 실험 데이타를 이용하여 기존의 양자화방법과 성능실험 결과를 서술한다. 마지막으로 4장에서는 결론 및 향후 연구과제에 대하여 기술한다.

Ⅱ. 변형된 유전자 알고리즘을 이용한 패턴인식모델

2.1 Simple Genetic Algorithm

유전자 알고리즘은 자연선택과 자연계 진화과정의 적자생존을 원칙으로 하는 적응적인 탐색 기법으로서 최적해에 대한 효과적인 탐색 도구로 널리 사용되고 있다. 이것은 다른 탐색방법 및 최적화 방법과 다른 점은 파라미터를 코딩한 것을 직접 이용하며, 점이 아닌 다점 탐색 방법을 취하는 것이다. 또한 탐색에 적합도를 이용하고 다른 부가적인 지식을 요구하지 않으며 결정론적인 규칙이 없이 확률적 연산자를 사용하여 수행된다. 이와 같은 특징으로 인해 다른 탐색 또는 최적화 방법 중 하나인 계산에 의존한 방법(calculus-based method)에 비하여 전역적 해를 구할 가능성이 높으며 다른 여러 탐색방법에 비하여 효율적이다. 유전자 알고리즘을 실제 응용에서 사용하기 위해 먼저 문제를 유전자형에 대응하는 문자열로 변환한다. 그리고 이진수 문자열의 나열인 염색체를 표현하고, 이 염색체들의 모임인 개체집단을 생성한다. 문자열을 평가하여 평가치가 높은 집단을 선택하여 남도록 하는데, 이것은 자연계에 있어서 도태에 대응하는 것이다. 이렇게 선택된 집단에 대하여 연산자를 적용함으로써 새로운 문자열을 생성한다. 기본적인 연산자 재생은 문자열을 적합도에 따라 개체집단에서 두 개의 염색체를 선택하는 역할을 하는데 개체집단내의 모든 염색체의 적합도의 합과 각 염색체의 적합도의 상대적 값, 즉 선택될 확률을 구하게 된다. 교차 연산자는 재생에 의해 선택된 두 염색체의 인자값을 서로 맞바꾸어 새로운 염색체를 생성한다. 이때 무작위로 선택된 교차 위치와 개수에 따라 교차 알고리즘이 달라지게 된다.

또한 돌연변이 연산자는 염색체내의 인자를 무작위로 선택하고, 그 값을 임의대로 바꾸어 새로운 염색체를 만드는 기능을 담당한다. 이것은 개체 집단의 특성을 다양하게 변화시키며 국부적인 최소 상태를 벗어 날 수 있게 한다. 위에서 설명한 연산자의 사이클을 반복함으로써 환경에 대응하는 평가치가 높은 문자열을 만들어 내어 문자열의 집단 전체의 평가치를 향상 시켜 나간다. GA 연산자를 이용하여 새로운 세대를 생성하였을 때 부모 세대 보다 더 낮은 적합도를 갖는 염색체가 생성될 수도 있는데, 세대를 거듭하면서 적합도가 낮은 것들은 염색체끼리의 경쟁에서 도태되어 사라지게 되므로 낮은 적합도를 갖는 염색체가 다음 세대에 생기더라도 문제가 되지 않는다. 기본적인 단순 유전자 알고리즘(Simple Genetic Algorithm: SGA)의 흐름도는 다음과 같다[4].

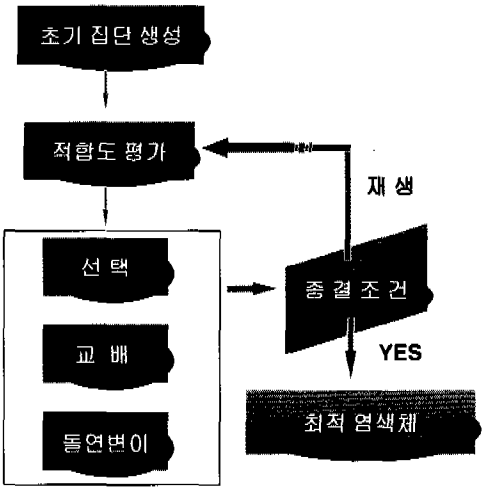


그림 1. 단순 유전 알고리즘 및 흐름도

2.2 변형된 유전자 알고리즘의 연산자

변형된 유전자 알고리즘은 방대하고 복잡한 함수에 대하여 전역적인 최적화를 할 수 있는 장점을 갖는 유전자 알고리즘과 신경회로망에서 Bayesian 결정경계면을 근사화하여 오인식의 수를 최소화하는 LVQ 알고리즘의 결합된 형태에 기초를 두고 있다. 인식기로서의 LVQ 모델은 통계적인 패턴인식 방법의 하나로서 각 집단사이의 최적 결정경계면을 갖는 참조벡터를 설정하기 위한 방법으로 각 집단의 왜곡도를 최소화 시키기 보다는 오인식의 수를 최소화하는 참조벡터를 생성시키기 위한 모델이다. 여기에 유전자 알고리즘은 전역적인 해를 구하거나 최적화 방법의 하나로서 자연 환경에서의 적자생존으로 생성과 도태의 반복으로 최적화 시키는 알고리즘이다. 제안된 유전자 알고리즘에서는 이러한 각 모델들의 특성을 이용하여 LVQ에서의 학습조건 즉, 오분류의 수를 최소화하는 참조벡터의 학습과 전체 오분류에 가장 크게 영향을 미치는 참조벡터의 도태와 생성을 통한 유전자 알고리즘을 이용한 오분류 최소화 알고리즘 구현에 기초한다.

2.2.1 타입(Type)의 정의

유전자 알고리즘은 생성과 교차, 도태를 반복적으로 변화하면서 최적의 값을 갖도록 반복 진행한다. 변형된 유전자 알고리즘에서는 오인식에 영향을 주는 참조벡터의 교차 및 도태 생성하기 위하여 오인식되는 경우의 상태를 정의하여 선별적인 학습을 하도록 정의된다. 이것은 학습과정에서 주어진 패턴에 대한 가장 가까운 참조벡터(Ref1)와 두번째로 가까운 참조벡터(Ref2)를 계산하여 정의하였다. NN에 근거한 분류시 Type Ⅰ,Ⅱ는 정인식을 의미하며, 나머지 Type들은 오인식 Type을 나타낸다. 변형된 유전자 알고리즘에서는 오인식 타입에서 대부분의 유전자 조작들이 이루어지며, 이와 같은 타입이 나타나는 패턴들을 최소화하는데 목적이 있다.

- Type Ⅰ : 학습패턴과 Ref1이 같은 클래스이고 정인식된 경우
- Type Ⅱ : 학습패턴과 Ref1은 같은 클래스, Ref2는 다른 클래스
- Type Ⅲ : 학습패턴과 Ref1은 다른 클래스, Ref2는 같은 클래스
- Type Ⅳ : 학습패턴과 Ref1, Ref2이 다른 클래스

2.2.2 재생과 교차

기존의 유전자 알고리즘에서 재생은 적합도에 따라 개체 집단에서의 두 개의 염색체를 선택하는 역할을 하는데 개체 집단내의 모든 염색체의 적합도 합과 각 염색체의 상대적 값, 즉 선택될 확률을 구하게 된다. 또한 교차는 재생에 의해 선택된 두 염색체의 인자를 맞바꾸어 새로운 염색체를 생성한다. 본 논문에서 적합도 함수는 정인식과 오인식으로 정의하며 교차에서 필요한 두 염색체는 오인식된 경우에 학습패턴과 가까운 자기 참조벡터를 선택하여 교차하게 하는데, 식(1)과 같이 윈도우를 정의하여 윈도우 안에 놓여 있을때 교차 연산을 수행하게 한다.

$$\min\left(\frac{d_i}{d_j}, \frac{d_j}{d_i}\right) > s \tag{1}$$

$$d_i = \| m_i - x \| = \sqrt{\sum_{j=1}^{n} (m_{ij} - x_j)^2} \tag{2}$$

여기서, s 는 윈도우 폭을 결정하는 파라미터이며, d_i, d_j는 주어진 패턴 x와 가장 가까운 참조벡터 mi, 두 번째 가까운 참조벡터 mj 와의 유클리디언 거리(euclidean distance)이다.

이것은 신경회로망에서의 LVQ 학습조건과 유사하다. 교차점은 이진수의 LSB(최하위 bit)에서 최대로 base*0.7 까지의 범위에서 랜덤한 position의 위치를 갖는다. 즉, 이진수의 최대 자릿수(base)의 70%의 MSB에 해당한다. 이것은 base의 최대 MSB의 보호차원에서 생성된 숫자라고 할 수 있다.

2.2.3 생 성

돌연변이 및 생성은 기존의 유전자 알고리즘에서는 무작위로 선택하는 것에 대하여 Type Ⅲ 조건에 있는 참조벡터 중에서 오분류에 영향을 많이 준, 즉, 에러 카운트가 일정 개수이상 발생하면, 그 때의 학습패턴과 Ref2의 두 점을 이은 선분의 내분점 중 계수 a 에 따르는 점을 구해 GA 알고리즘의 생성점으로 삼는다. 물론 생성점은 Ref2와의 거리가 0.2이고, 학습패턴과의 거리가 0.8인 내분점이 된다. 여기서 파라미터값은 문제에 따라 다른 값을 주어줄 수 있다.

2.2.4 도 태

도태는 생성과 밀접한 관계를 갖는 것으로 오인식률에 가장 크게 영향을 미치는 참조벡터를 선택하여 도태시킨다. 도태조건으로는 Type 정의에서 Ⅲ,Ⅳ의 요인으로 동작되는 참조벡터에 인덱스 카운터를 증가시킨다. 그리고 반복 수행되는 동안 학습이 더 이상 수행되지 않고 평가가 수렴되었을 경우 인덱스된 참조벡터를 도태시킨다.

2.3 변형된 유전자 알고리즘

유전자 알고리즘은 여러 종류의 최적화문제에 뛰어난 성능을 보이는 반면 양질의 해를 얻는데는 많은 시간이 걸린다는 단점이 있다. 또한 분류기로서 왜곡도를 최소화하는 것보다는 오인식률을 최소화하는 것이 인식성능에서 더 효율적이라는 것이 증명되고 있다. 이에 본 논문에서는 초기 양질의 학습조건을 갖도록 양자화 방법을 이용하여 참조벡터를 설정하고 유전자 알고리즘이 갖는 뛰어난 최적화 기술에 선택적인 학습을 통한 알고리즘을 제안한다. 학습기능을 갖는 변형된 유전자 알고리즘은 다음과 같다.

단계 1. 학습패턴으로부터 k-means방법에 의해 초기 참조벡터 및 파라미터 초기화한다.

단계 2. GA의 반복횟수 만큼 단계 3-5를 반복한다.

단계 3. 주어진 학습패턴과 참조벡터들과의 유클리디언 거리를 구한후 가장 가까운 참조벡터(Ref1)와 두 번째 가까운 참조벡터(Ref2)를 구한다.

단계 4. 타입의 정의에서 Type을 판정한다.

1) Type Ⅲ의 경우 :

① Ref2의 참조벡터의 에러 인덱스에 1를 카운트한다.

② 윈도우 안에 떨어진 경우 학습패턴과 Ref2와 일점교차를 수행한다.

2) Type Ⅳ의 경우 : 주어진 패턴은 학습 및 교차시 제외시킨다.

단계 5. 참조벡터의 에러 인덱스가 임의의 값 이상이고, 학습이 더 이상 수행되지 않을시 Ref2와 주어진 패턴 사이에 가중치를 둔 내분점 구한후 새로운 참조벡터를 생성한다

Ⅲ. 컴퓨터 모의실험

변형된 유전자 알고리즘에 대한 모의 실험은 그림2에서 볼 수 있듯이 3개의 클래스에 대하여 가우시안 분포를 갖고 [0,1] 사이의 양의 실수를 갖는 각각 100개의 모의 개체를 갖는 학습패턴과 시험패턴을 가지고 팬티엄II-200MHz 프로세서 이상의 환경 하에서 MATLAB 5.2 버전을 사용하여 시뮬레이션을 하였다. 각 패턴들을 GA에 염색체로 적용시키기 위하여 실수를 12자리의 2진수로 변환하였다. 정수부는 2자리, 소수부의 표현에 10자리를 사용하였으며 각 염색체는 클래스 인덱스, 그리고 에러가 중인덱스를 갖는다. 본 논문에서의 목적은 전술한 바와 같이 오분류의 수를 최소화하는 참조벡터 설정에 관한 것으로 기존의 유전자 알고리즘에서는 초기 랜덤한 값을 가지고 최적화 과정을 진행하였지만 본 논문에서는 초기에 유용한 참조벡터를 이용한다. 이에 학습기능을 갖는 유전자 알고리즘을 수행하기 전에 기존의 벡터 양자화 방법을 이용하여 초기 참조벡터를 설정하게 된다. 본 논문에서는 초기 참조벡터 설정 방법으로 k-means 클러스트링 알고리즘을 사용하였으며 각 집단에 대해 각각 3개의 참조벡터를 설정하였으며 학습패턴 및 참조벡터는 GA의 염색체로 사용한다. 변형된 유전자 알고리즘의 성능 평가를 위하여 초기 학습패턴과 K-means 방법을 이용하였을때의 초기 참조벡터 위치를 그림2에 나타내었으며, 신경회로망에서의 참조벡터 최적화 방법의 하나인 LVQ 알고리즘을 수행하였을 때 k-means에 의해 생성된 초기 참조벡터의 변화된 위치를 그림 3에 나타내었다. K-means에 의해 생성된 초기 참조벡터는 왜곡도를 최소화 하기 때문에 각 집단의 평균적으로 중앙에 위치함을 알 수 있었으며, LVQ 학습 과정을 거친 참조벡터는 보다 중앙에 밀집되어짐을 알 수 있다.

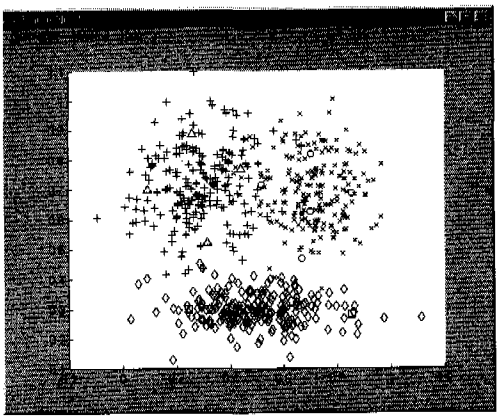


그림 2. K-means를 이용한 참조벡터 분포

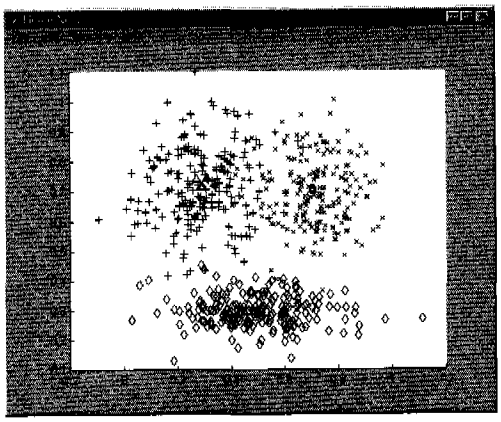


그림 3. LVQ을 수행한후의 참조벡터 분포

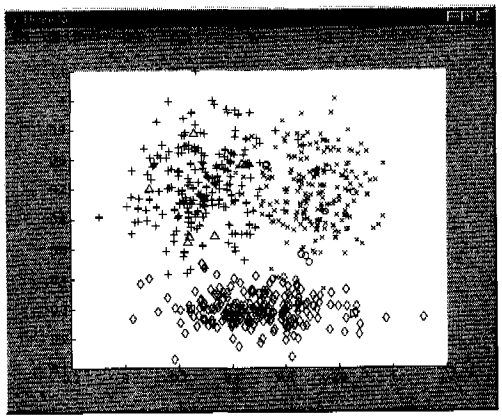


그림 4. 변형된 유전자 알고리즘을 수행한 후 참조벡터 분포

본 논문에서 제안된 변형된 유전자 알고리즘은 K-means와 LVQ에서와는 다르게 참조벡터의 개수는 점해지지 않는다. 이것은 생성과 도태를 반복적으로 수행하면서 인식률의 변화에 따라 참조벡터가 변화하기 때문이다. 그림 4.에서 볼 수 있듯이 참조벡터들의 위 변화는 초기 참조벡터에 비해 크게 변화 되지 않는 반면 결정경계면 부근에 위치한 참조벡터나 오인식에 영향을 받는 부근에서 참조벡터가 변화하거나 새로운 참조벡터가 생성됨을 알 수 있었다. 각 방법에 대한 학습패턴과 시험패턴에 대한 인식률를 표 1.에 나타내었다.

표 1. 학습패턴 및 시험패턴에 대한 인식률 비교

방법 \ 인식률	학습패턴(%)	시험패턴(%)
K-means	95.33	94.55
LVQ	95.66	94.68
변형된 유전자 알고리즘	96.36	95.26

Ⅳ. 결 론

본 논문에서는 기존의 유전자 알고리즘이 교차, 생성, 도태등의 연산을 이용한 뛰어난 최적화 방법에 패턴 분류문제에 있어서 오분류의 수를 최소화할 수 있도록 학습 기능을 갖는 변형된 유전자 알고리즘에 관한 것이다. 이에 신경회로망의 LVQ 알고리즘이 갖는 미세 조정되는 지역적 탐색 메카니즘의 특성을 기존의 유전자 알고리즘에 첨가하여 최소의 오분류를 갖는 참조벡터 설정 하도록 하였다. 모의 실험결과에서도 볼 수 있듯이 변형된 유전자 알고리즘은 결정경계면에서의 참조벡터 변화 또는 생성과 도태가 발생되었음을 알 수 있었으며 인식성능면에서도 하나의 효율적인 방법으로 추정된다. 그러나 참조벡터 수의 제한 및 적합도 함수의 새로운 방향이 설정되어야 더욱 효율적인 성능을 보일 것으로 예상된다.

참 고 문 헌

[1] J.H. Holland, adaptation in Natural and Artificial Systems, MIT Press, 1975.

[2] T. Kohonen, "The self-organizing map," Proc. IEEE, Vol.78, No.9 pp.1464-1480, Sept. 1990

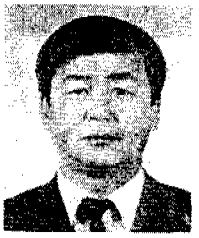
[3] H. Kitano, "Emprical studies on the speed of convergence of neural network training using genetic algorithm," In Proc. 8th JMIT National Conf. in Aritificial Intelligence,

[4] Lawrence Davis, Handbook of genetic algorithm, Van Nostrand Reinhold, New York, 1991
[5] Aha, D.W, "machine Learning," A tutorial presented at the 5th International Workshop on Aritifitial Intelligence & Statistics, pp.1-67, 1995
[6] Goldberg, D.E., "Genetic Algorithm in Search, Optimizations & Machine Learning," Addison Wesley, 1989
[7] Bir Bhanu, Sungkee Lee, and Jhon Ming, "Adaptive image segmentation using genetic algorithm," Image Understanding Workshop, pp.1043-1055, 1989

김 대 극(Dae-Keuk Kim) 정회원

1966년 7월 25일생
1992년 : 강원대학교 전자공학과 졸업(공학사)
1994년 : 강원대학교 대학원 전자공학과 졸업 (공학석사)
1997년 : 강원대학교 대학원 전자공학과(박사과정 수료)
1997년~현재 : 한림정보산업대학 전자통신과 조교수
<주관심 분야> 신경회로망, 음성신호처리, 패턴인식

이 재 곤(Jae-Kon Lee) 정회원



1951년 9월 19일생
1985년 : 숭실대학교 대학원 전자공학과 졸업 (공학석사)
1996년 : 강원대학교 대학원 전자공학과 (박사과정 수료)
1989년~현재 : 한림정보산업대학 전자통신과 부교수
<주관심 분야> 음성신호처리, 신경회로망

이 상 희(Sang-Hee Lee) 정회원

1974년 : 서울대학교 전기공학과 졸업
1978년 : 서울대학교 전기공학과 석사학위 취득
1984년 : NANCY 대학(불란서) 전자공학과 박사학위 취득
1994년 3월~1994년 12월 : 강원대학교 정보통신연구소 소장
1994년 12월~1995년 12월 : 버지니아 텍 교환교수
1985년~현재 : 강원대학교 전기전자공학부 교수
<주관심 분야> 음성신호처리, 신경회로망