

자동차 검출을 위한 GAVaPS를 이용한 최적 분류기 앙상블 설계

Optimal Classifier Ensemble Design for Vehicle Detection Using GAVaPS

이 희 성, 이 제 현, 김 은 태*
(Heesung Lee, Jaehun Lee, and Euntai Kim)

Abstract: This paper proposes novel genetic design of optimal classifier ensemble for vehicle detection using Genetic Algorithm with Varying Population Size (GAVaPS). Recently, many classifiers are used in classifier ensemble to deal with tremendous amounts of data. However the problem has a exponential large search space due to the increasing the number of classifier pool. To solve this problem, we employ the GAVaPS which outperforms comparison with simple genetic algorithm (SGA). Experiments are performed to demonstrate the efficiency of the proposed method.

Keywords: GA, GAVaPS, classifier ensemble, vehicle detection

I. 서론

패턴 인식 시스템은 응용분야가 크고 활용도가 방대하기 때문에, 여러 응용분야의 인식이론 및 기술들이 많은 연구자, 공학자들에 의해 개발되고 있다. 최근에는 패턴 인식 시스템의 성능을 높이기 위하여 여러 분류기들을 결합하는 연구가 활발히 진행 중이다. 이런 다중 분류기 시스템 설계는 결정 최적화(decision optimization)와 범위 최적화(coverage optimization) 문제로 구분된다[1]. 결정 최적화를 이용한 다중 분류기 설계 방법은 분류기 앙상블 연구의 초창기에 많이 연구되었다. 결정 최적화는 앙상블에 참가하는 단일 분류기들의 출력을 어떻게 결합하여야 최적의 성능을 얻을 수 있는지를 다룬다. 최근 들어 활발히 연구되고 있는 범위 최적화를 이용한 다중 분류기 설계는 앙상블에 참여할 단일 분류기들의 집합을 결정하는 문제이다. 아무리 결합 방법이 좋더라도 참여하는 분류기들의 상호 보완성이 약하다면 의미 있는 성능 향상을 기대하기 어렵다. 따라서 범위 최적화는 결정 최적화 못지않게 중요하다[2]. 하지만, 어떤 분류기들의 집합이 상호 보완성이 크고 클래스들 간의 가장 좋은 차별성(discrimination)을 제공하는지 알 수 없다. 또한 선택할 수 있는 하위 분류기의 집합은 분류기의 크기에 따라 무수히 늘어날 수 있기 때문에 분류기 앙상블 선택 문제는 해결하기 어려운 최적화 문제 중 하나이다.

자연 선택과 자연 발생의 과정을 기초로 다수의 개체를 동시에 진화시켜 가면서 최적의 해를 찾는 유전자 알고리즘은 많은 최적화 문제에서 사용되고 있다[3]. 패턴 인식 시스템의 정확도를 향상시키면서 분류기의 숫자를 줄이기 위해 적절한 분류기들의 부분 집합의 구성을 구하는 문제 역시 최적화 문제이다. D 개의 분류기를 갖는 분류기 풀(classifier pool)이 있을 때, 유전자 알고리즘의 역할은 최적

화의 제약 조건을 지키며 d 개의 분류기 부분 집합을 찾는 것이다. 일반적으로 앙상블에 참가하는 분류기들의 부분 집합은 그들의 차원, 클래스의 분리 또는 정확성에 의거하여 평가를 받기 때문에 유전자 알고리즘을 이용하여 분류기나 데이터의 분포상태에 맞는 최적의 분류기의 집합을 찾을 수 있다. 하지만, 탐색 공간의 크기가 D^{C_d} 이기 때문에 D 가 커질수록 시간비용이 지수적으로 커지는 문제를 안고 있다. 특히 최근에는 복잡하고 크기가 매우 큰 데이터들을 다루기 위해 매우 많은 수의 분류기들이 사용된다. 따라서 많은 수의 다양한 분류기들을 포함하기 위해서 충분히 큰 분류기 풀이 있어야 하고 탐색의 시간은 분류기 풀의 크기(D)에 지수적으로 커지므로 계산 효율이 충분히 고려되어야 한다. 우리는 이전 연구[4]에서 매우 탐색 영역이 크고 해결하기 어려운 문제인 분류기들이 사용하는 특징들을 선택하는 문제에 개체군 변환 유전자 알고리즘을 적용하였고, 이는 일반적인 유전자 알고리즘보다 좋은 계산 효율과 함께 뛰어난 성능을 갖고 있음을 볼 수 있었다. 따라서 본 논문에서는 앙상블에 사용되는 다수 분류기들을 선정하는 방법으로 일반적인 유전자 알고리즘보다 효율적인 개체군 변환 유전자 알고리즘을 이용하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. II 장에서는 유전자 알고리즘과 개체군 변환 유전자 알고리즘을 설명하고, III 장에서는 개체군 변환 유전자 알고리즘 이용한 새로운 분류기 앙상블 선택 알고리즘을 제안한다. IV 장에서는 제안된 시스템의 효용성을 보이기 위한 실험과 그의 고찰을 한 뒤 마지막으로, V 장에서는 결론과 추후 과제에 대한 설명을 한다.

II. 배경지식

1. Genetic Algorithm (GA)

유전자 알고리즘은 진화 알고리즘의 한 부류이며, 진화 알고리즘은 컴퓨터를 기반으로 한 문제 풀이 체계의 설계나 실험에 이미 알려진 진화 메커니즘 중 일부를 중요 요소로 사용한다. 유전자 알고리즘은 생명체의 유전 및 진화 과정을 전산학적으로 모델링(modeling)한 기계학습의 방법

* 책임저자(Corresponding Author)

논문접수: 2009. 9. 11., 수정: 2009. 11. 24., 채택확정: 2009. 11. 26.
이희성, 이제현, 김은태: 연세대학교 전기전자공학부
(4u2u@yonsei.ac.kr/aznable17@yonsei.ac.kr/etkim@yonsei.ac.kr)

※ This research was supported by grant number R01-2006-000-11016-0 from Basic Research Program of the Korea Science & Engineering Foundation and by LG Yonam Foundation under the overseas research professor program.

으로, 탐색해야 할 공간이 매우 넓은 경우 유용하게 사용되는 탐색 및 최적화 기법이다. 유전자 알고리즘의 가장 큰 특징은 잠재적 해인 염색체(chromosome)들이 집단을 이루어 만들어진 해의 집단(population)을 운용한다는 것이다. 각 염색체는 적자생존의 법칙에 의하여 상대적으로 우수한 것이 살아남을 확률이 크며, 또한 유전 연산자에 의하여 진화 과정을 거치게 된다. 적합도가 높은 개체의 집합이 선택(selection)되어 다음 세대의 자손을 생성하는 부모가 되며, 자손은 교차(crossover), 돌연변이(mutation)의 유전 연산자를 통해 생성된다[5].

2. Genetic Algorithm with Varying Population Size (GAVaPS)

유전자 알고리즘은 전역 탐색 능력은 우수하나 지역 탐색 능력이 떨어지고 세대가 지남에 따라 불완전 수렴을 하거나 근사 최적의 해에 수렴하는 단점이 있다[6]. 이러한 단점을 극복하기 위해 각 세대마다 개체군의 크기를 유동적으로 적용하여 유전적 다양성을 향상시키고자 하는 개체군 변환 유전자 알고리즘 (GAVaPS: Genetic Algorithm with Varying Population Size)이 제안되었다[7]. 개체군 변환 유전자 알고리즘은 일반적인 유전자 알고리즘의 선택 단계 대신 나이(age) 개념을 도입한다. 나이는 개체가 생존한 세대수로 정의된다. 또한 수명(lifetime)을 정의하는데 이것은 평가 단계에서 각각의 염색체에 한 번 할당된다. 만약 염색체의 나이가 수명을 초과하게 되면 그 염색체를 삭제한다. 따라서 다음 개체군의 크기($PopSize(t+1)$)는 다음과 같이 정의된다.

$$PopSize(t+1) = PopSize(t) + AuxPopSize(t) - D(t) \quad (1)$$

여기에서 $AuxPopSize(t)$ 와 $PopSize(t)$ 는 현 세대와 추가 개체군의 크기, $D(t)$ 는 수명이 다한 개체의 수를 나타낸다.

수명은 개체가 개체군 내에 존재할 세대수를 결정하기 때문에 각 자손들의 기대치는 수명값에 비례한다. 따라서 높은 적합도를 갖는 염색체에는 긴 수명을 할당하고, 낮은 적합도를 갖는 염색체에는 짧은 수명을 할당한다. 이러한 점들을 고려한 수명 할당 방식에는 비례 할당(proportional allocation), 선형 할당(linear allocation), 쌍선형 할당(bilinear allocation)이 있고 다음과 같이 각각 계산된다.

1) proportional allocation:

$$\min\left(\text{MinLT} + \eta \frac{\text{fitness}[i]}{\text{AvgFit}}, \text{MaxLT}\right) \quad (2)$$

2) linear allocation:

$$\text{MinLT} + 2\eta \frac{\text{fitness}[i] - \text{AbsFitMin}}{\text{AbsFitMax} - \text{AbsFitMin}} \quad (3)$$

3) bi-linear allocation:

$$\begin{cases} \text{MinLT} + \eta \frac{\text{fitness}[i] - \text{MinFit}}{\text{AvgFit} - \text{MinFit}} & \text{AvgFit} > \text{fitness}[i] \\ 0.5(\text{MinLT} + \text{MaxLT}) + \eta \frac{\text{fitness}[i] - \text{AvgFit}}{\text{MaxFit} - \text{AvgFit}} & \text{AvgFit} < \text{fitness}[i] \end{cases} \quad (4)$$

여기에서 $MaxLT$ 와 $MinLT$ 는 허용되는 최대, 최소 수명이고, $\eta = 0.5(MaxLT - MinLT)$ 이다. $AvgFit$, $MaxFit$, 그리고 $MinFit$ 은 각각 평균, 최소, 최대 적합도 값을 의미한다. 또한 일반적인 유전자 알고리즘과는 달리 개체군 변환 유전자 알고리즘에서는 개체군내의 모든 염색체는 염색체의 적합도와 무관하게 똑같은 확률로 교차와 돌연변이 등의 유전 연산자를 적용하여 얻는다[4].

III. 개체군 변화 유전자 알고리즘을 이용한 분류기 앙상블 선택

범위 최적화를 이용한 다중 분류기 설계는 앙상블에 참여한 단일 분류기들의 집합을 결정하는 문제이다. 상호 보완성이 강한 분류기들을 이용하는 다중 분류기의 성능은 단일 분류기들에 비해 크게 향상된다. 최근에는 복잡하고 크기가 매우 큰 데이터들을 다루기 위해 매우 많은 수의 분류기들이 사용된다. 따라서 많은 수의 다양한 분류기들을 포함하기 위해서 충분히 큰 분류기 풀이 있어야 한다. 하지만, 분류기 풀의 탐색 시간은 분류기 풀의 크기에 지수적으로 커지므로 계산 효율이 충분히 고려되어야 한다. 일반적인 유전자 알고리즘을 이용한 분류기 선택 방법은 기존의 분류기 선택 방법에 비해 뛰어난 성능을 보여주지만, 매우 큰 분류기 풀을 다루는 문제에는 너무 많은 계산량을 필요로 한다. 따라서 우리는 탐색 영역이 클 때 일반적인 유전자 알고리즘에 비해 효율적으로 최적해를 찾아주는 개체군 변환 유전자 알고리즘을 사용하여 이 문제를 해결한다.

1. 염색체

본 논문에서 사용한 염색체의 구성은 그림 1과 같다. 유전자 알고리즘은 그림 1과 같이 각각의 비트가 각각의 분류기들과 연관되어 있는 최적의 이진 벡터를 찾기 위해 사용되었다. 만약 i 번째 비트가 1이라면 i 번째 분류기는 앙상블에 포함된다. 반대로 0이라면 앙상블에 포함되지 않는다. 예를 들어 염색체가 “1001...01”이라면 첫 번째, 네 번째 그리고 마지막 분류기들이 앙상블에 포함된다. 유전자 알고리즘의 염색체인 이진 벡터에 의해 만들어지는 분류기들의 부분집합은 정확도와 선택된 분류기의 숫자에 의해 평가된다. 유전자 알고리즘은 이 정보를 이용하여 최적의 분류기들의 조합을 찾게 된다.

2. 유전 연산자와 적합도 함수

유전 연산자는 현재 세대의 훌륭한 성질을 가지고 있는 염색체들을 엮어서 다음 세대에 전달하는 기능을 갖는다. 일반적으로 유전 연산자로는 교차와 돌연변이가 사용된다. 교차는 부모 세대의 두 개체간의 염색체를 부분적으로 서로

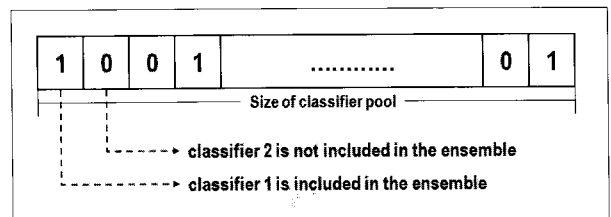


그림 1. 염색체의 구성.

Fig. 1. Structure of chromosome.

바꿈으로써 지식 세대의 새로운 개체를 생성하는 것이다. 이때 부모의 특성이 자손에게 적절히 계승되어야 한다. 유전 연산자를 위하여 one-point 교차와 bit-flip 돌연변이를 사용하였다. 또한 수명 할당 방식으로 비례 할당을 사용하였고, 허용되는 최대, 최소 수명은 20과 5를 각각 사용하였다.

염색체의 적합도를 측정하기 위하여 다음과 같은 함수를 사용하였다.

$$fitness = \sum_{j=1}^J P(v_j) - \alpha d \quad (5)$$

여기에서 $V = \{v_1, v_2, \dots, v_J\}$ 는 튜닝 데이터이고 J 는 튜닝 데이터의 개수이다. 만약 튜닝 데이터 v_j 가 올바르게 분류된다면 $P(v_j) = 1$ 이고 그렇지 않다면 $P(v_j) = 0$ 이다. α 는 적합도 함수의 coefficient으로써 분류기 앙상블의 정확성과 간결성을 조절한다. d 는 선택된 분류기들의 숫자로서 염색체의 각 비트의 1의 개수의 합이다. 따라서 염색체의 적합도는 튜닝 데이터의 인식률에 비례하고 앙상블에 참여하는 분류기들의 숫자에 반비례한다.

VI. 실험

카메라를 사용하는 지능형 자동차 시스템에서 영상으로 입력받은 물체가 차인지 아닌지를 구별하는 것은 중요한 작업이다. Intelligent Vehicle Database[8]는 이러한 시스템을 위해 만들어진 데이터베이스로써 60개의 자동차 이미지와 60개의 자동차가 아닌 이미지를 포함하고 있다. 모든 이미지의 사이즈는 50x50개이고 우리는 각 이미지 픽셀의 gray scale값에 주성분 분석[9]을 적용하여 2000개의 특징을 추출하였다. 다음 그림은 두 개의 클래스의 샘플들을 보여준다.

다양한 분류기들을 포함하는 분류기 풀을 만들기 위해 특징의 숫자를 100개에서 2000개까지 임의의 선택하여 총 100개의 분류기를 만들었다. 분류기는 K-nearest neighbor (KNN) [7]를 사용하였고 그 성능을 표 1에 나타내었다. 우리는 데이터베이스를 각각 학습, 튜닝, 테스트 데이터로 나누었다. KNN 분류기의 학습을 위하여 60개의 학습 데이터를 사용하였고, 유전자 알고리즘의 평가를 위하여 30개의 튜닝 데이터, 그리고 알고리즘의 평가를 위해 30개의 테스트 데이터를 각각 사용하였다. 실험을 위해 사용된 데이터베이스의 구성은 표 2와 같고, 본 실험에 사용된 파라미터 값은 표 3에 나타내었다.

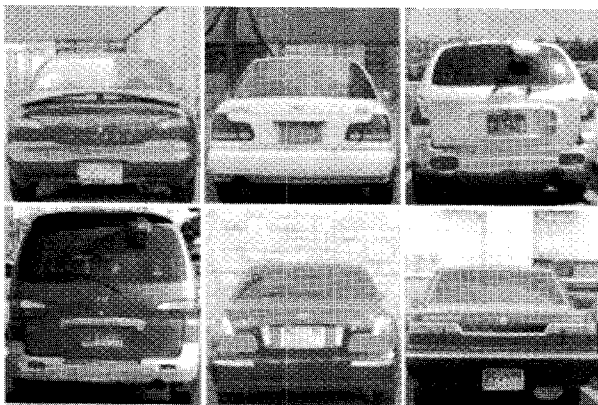


그림 2. Car 이미지.
Fig. 2. Car image.

트 데이터를 각각 사용하였다. 실험을 위해 사용된 데이터베이스의 구성은 표 2와 같고, 본 실험에 사용된 파라미터 값은 표 3에 나타내었다.

일반적인 유전자 알고리즘과 개체군 변환 유전자 알고리즘을 사용하여 최적의 분류기를 독립적으로 각각 10회씩 설계하여 그 성능을 비교하였다. 각각 설계된 분류기 앙상블들의 결과는 표 4와 그림 4에 도시하였다. 표 4에서 validation 항목은 각 유전자 알고리즘들의 최종 적합도를 나타내며 trial 항목은 계산된 염색체의 총 개수를 나타낸다. 또한 test 항목은 각 유전자 알고리즘으로 설계된 분류기 앙

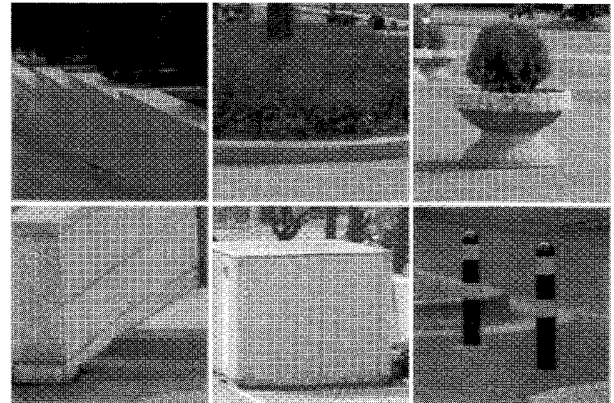


그림 3. Non-car 이미지.
Fig. 3. Non-car image.

표 1. 단일 분류기들의 성능.

Table 1. Performance of single classifiers.

Parameter	Value
Original feature number	2000
Used feature number	100~2000
Average recognition rate	0.82
Maximum recognition rate	0.87
Minimum recognition rate	0.80

표 2. 사용된 database들.

Table 2. Used databases.

Data set	Number of data
Training data	60
Tuning data	30
Test data	30
Total	120

표 3. 유전자 알고리즘의 파라미터들.

Table 3. Parameters for genetic algorithm.

Parameter	Value
Max. generation number	500
Crossover rate	0.65
Mutation rate	0.05
Classifier number	100

표 4. 제안된 방법의 인식 결과.

Table 4. Results of the proposed method.

Methods		SGA	GAVaPS
Validation	Average fitness value	28.56	28.75
	Maximum fitness value	28.80	29.40
	Minimum fitness value	28.30	28.20
Trial	Average trial	50090.10	4819.70
	Maximum trial	50096	5438
	Minimum trial	50086	4055
Test	Average recognition rate	0.90	0.91
	Maximum recognition rate	0.93	0.93
	Minimum recognition rate	0.83	0.87

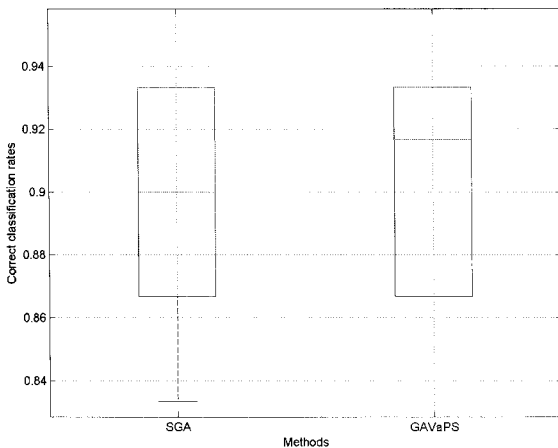


그림 4. 제안하는 알고리즘의 성능.

Fig. 4. Performance of the proposed method.

표 5. 학습시간의 비교.

Table 5. Comparison of training time.

Methods	Training time (min/100 generations)
SGA	69112.60
GAVaPS	7199.99

상블들의 인식률을 나타낸다. 제안하는 알고리즘은 적은 염색체의 계산을 갖고도 높은 인식률을 보여주기 때문에 일반적인 유전자 알고리즘(SGA)에 비해 개체군 변환 유전자 알고리즘(GAVaPS)이 앙상블 분류기 선택에 적합함을 알 수 있다. 탐색 시간의 효율성을 비교하기 위하여, 표 5에서는 제안하는 알고리즘과 기존의 방법의 학습 시간을 제시하였다. 제안하는 알고리즘은 기존의 방법보다 약 10분의 1 정도의 학습 시간을 필요로 했으며, 이는 계산된 염색체의 개수의 비율과 비슷함을 알 수 있다.

V. 결론

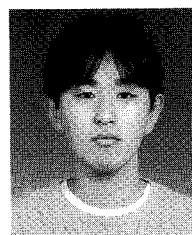
D 개의 분류기를 갖는 분류기 풀(classifier pool)이 있을 때, 유전자 알고리즘의 역할은 최적화의 제약 조건을 지키며 d 개의 분류기 부분 집합을 찾는 것이다. 탐색 공간의 크기가 $_D C_d$ 이기 때문에 D 가 커질수록 시간비용이 지수적

으로 커진다. 따라서 효율적으로 분류기의 부분집합을 찾기 위해서 본 논문에서는 개체군 변환 유전자 알고리즘을 이용하여 앙상블 분류기들을 선정하는 방법을 제안하였다. 제시한 방법의 유용성을 확인하기 위해 지능형 자동차 시스템에서 쓰이는 영상을 이용한 데이터베이스에 적용시킨 결과, 기존의 유전자 알고리즘보다 개체군 변환 유전자 알고리즘이 뛰어난 성능을 보임을 알 수 있었다.

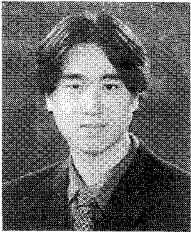
참고문헌

- [1] T. K. Ho, "Multiple classifier combination: lessons and next steps," *Hybrid Methods in Pattern Recognition*, pp. 171-198, 2002.
- [2] 김영원, 오일석, "군집화와 유전 알고리즘을 이용한 거친-섬세한 분류기 앙상블 선택," *정보과학회논문지*, vol. 34, no. 9, pp. 857-868, 2007.
- [3] H. Lee, E. Kim, and M. Park, "A genetic feature weighting scheme for pattern recognition," *Integrated Computer-Aided Engineering*, vol. 14, no. 2, pp. 161-171, 2007.
- [4] 이희성, 이제현, 김은태 "GAVaPS를 이용한 다수 K-Nearest Neighbor classifier들의 feature 선택," *한국지능시스템학회 논문지*, vol. 18, no. 6, pp. 871-875, 2008.
- [5] J. H. Holland, *Genetic Algorithms in Search, Optimization and Machine Learning* Addison-Wesley, Reading, MA, 1989.
- [6] 권기호, "개체군 변환 유전자 알고리즘의 새로운 수명 할당 방식에 관한 연구," *전자공학회논문지*, 제36권 C 편 제1호, pp. 66-72, 1999.
- [7] J. Arabas, Z. Michalewicz, and J. Mulawka "GAVaPS - a genetic algorithm with varying population size," in *Proc. Evolutionary Computation conf. part of the IEEE World Congress on Computational Intelligence*, 1994.
- [8] J. Hwang, K. Rou, S. Park, E. Kim, and H. Kang, "PCA based Vehicle Detection for ACC," in *Proc. 8th Int. Conf. on Elect., Inform., and Comm.*, pp 98-101, Jun. 2006.
- [9] P. Gelhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp 711-720, 1997.

이희성

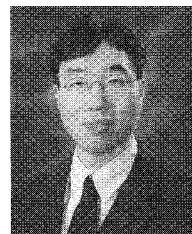


2003년 연세대학교 기계전자공학부(공학사). 2005년 동 대학원 전기전자공학부(공학석사). 2005년~현재 동 대학원 전기전자공학부 박사과정 재학중. 관심분야는 Computational Intelligence, Computer Vision, Pattern Recognition.



이 제 현

2005년 연세대학교 전기전자공학부(공학사). 2007년 동 대학원 전기전자공학부(공학석사). 2007년~현재 동 대학원 전기전자공학부 박사과정 재학중. 관심분야는 Computational Intelligence, machine learning.



김 은 태

1992년 연세대학교 전자공학과(공학사). 1994년 연세대학교 전자공학과(공학석사). 1999년 연세대학교 전자공학과(공학박사). 1999년~2002년 국립한경대학교 제어측공학과 조교수. 2002년~현재 연세대학교 전기전자공학부 교수. 2003년 University of Alberta, visiting researcher. 2008년 UC Berkeley visiting researcher. 2003년 대한 전자공학회 해동상 수상. 관심분야는 Computational Intelligence, machine learning.