한국지능시스템학회 논문지 2008, Vol. 18, No. 6, pp. 871-875

GAVaPS를 이용한 다수 K-Nearest Neighbor classifier들의 Feature 선택

Feature Selection for Multiple K-Nearest Neighbor classifiers using GAVaPS

이희성 · 이제헌 · 김은태*

Heesung Lee, Jaehun Lee, and Euntai Kim

* 연세대학교 전기전자공학부

요 약

본 논문은 개체군 변환 유전자 알고리즘을 (GAVaPS) 이용하여 k-nearest neighbor (k-NN) 분류기에서 사용되는 특징들을 선정하는 방법을 제시한다. 우리는 다수의 k-NN 분류기들을 사용하기 때문에 사용되는 특징들을 선정하는 문제는 매우 탐색 영역이 크고 해결하기 어려운 문제이다. 따라서 우리는 효과적인 특징들의 선정을 위해 일반적인 유전자 알고리즘 (GA) 보다 효율적이라고 알려진 개체군 변환 유전자 알고리즘을 사용한다. 또한 다수 k-NN 분류기를 개체군 변환 유전자 알고리즘으로 효과적으로 결합하는 방법을 제시한다. 제안하는 알고리즘의 우수성을 여러 실험을 통해 보여준다.

키워드 : 특징선정, 유전자 알고리즘, 개체군 변환 유전자 알고리즘, k-NN, 분류

Abstract

This paper deals with the feature selection for multiple k-nearest neighbor (k-NN) classifiers using Genetic Algorithm with Varying Population Size (GAVaPS). Because we use multiple k-NN classifiers, the feature selection problem for them is vary hard and has large search region. To solve this problem, we employ the GAVaPS which outperforms comparison with simple genetic algorithm (SGA). Further, we propose the efficient combining method for multiple k-NN classifiers using GAVaPS. Experiments are performed to demonstrate the efficiency of the proposed method.

Key Words : Feature selection, GA, GAVaPS, k-NN, classification

1. 서 론

패턴 인식 시스템은 응용분야가 크고 활용도가 방대하기 때문에, 여러 응용분야의 인식이론 및 기술들이 많은 연구자, 공학자들에 의해 개발되고 있다. 일반적으로 패턴은 특징 공간에서의 벡터로 표현되기 때문에, 특징들의 측정과 측정된 특징의 선택은 패턴 인식 알고리즘의 결과에 중요한 영향을 미친다[1]. 따라서 패턴인식 시스템의 정확도를 높이기 위해서는, 적합한 특징의 선택이 매우 중요하다. 특히 k-nearest neighbor (k-NN) 분류기 (classifier)[2]는 패턴들 간의 거리를 통해 클래스(class)를 결정하기 때문에 선정된 특징(feature)들은 분류기의 성능을 좌우하는 결정적인 요소이다. 하지만, 어느 특징들이 클래스들 간의 가장 좋은 차별성(discrimination)을 제공하는지 알 수 없다. 또한 패턴을 표현 가능하게 하는 선택된 하위 특징 공간 (sub-feature space)은 특징의 크기에 따라 무수히 늘어날 수 있기 때문에 특징선정 문제는 해결하기 어려운 최적화 문제 중 하나이다.

접수일자 : 2008년 2월 18일
완료일자 : 2008년 6월 5일
* 교신 저자
본 연구는 한국과학재단의 특정기초연구사업의 연구비 지원에 의하여 수행되었음.(R01-2006-000-11016-0)

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 특징의 숫자를 줄이기 위해 적절한 특징 공간을 이루는 특징들의 구성을 구하는 문제 역시 최적화 문제이다. D-차원의 입력 패턴의 집합이 있을 때, 유전자 알고리즘의 역할은 최적화의 제약 조건(ex. 정확도)을 지키며, d-차원($D \gg d$)으로 변환하는 것이다. 일반적으로 변화된 패턴의 특징 벡터들은 그들의 차원, 클래스의 분리 또는 정확성에 의거하여 평가를 받기 때문에 유전자 알고리즘을 이용하여 분류기나 데이터의 분포상태에 맞는 최적의 특징 공간을 찾을 수 있다. 또한 기존의 고전적인 특징 선택 방법 중에서 유전자 알고리즘을 이용한 특징 선택 방법은 가장 좋은 성능을 보여준다[3]. 하지만, 본 연구에서 우리는 다수의 분류기들을 사용한다. 일반적으로 한 개의 분류기보다 다수의 분류기들은 더욱 좋은 성능을 보여준다[4]. 다수의 분류기들이 사용하는 특징을 선정하는 문제는 한 개의 분류기에서 사용되는 특징의 선정보다 더욱 복잡한 문제이며 그 계산시간도 길어진다. 따라서 우리는 일반적인 유전자 알고리즘보다 효율적인 개체군 변환 유전자 알고리즘을 이용하여 특징들을 선정하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 유전자 알고리즘과 개체군 변환 유전자 알고리즘을 설명하고, 3장에서는 개체군 변환 유전자 알고리즘 이용한 새로운 특징 선정 방

법을 제안한다. 4장에서는 제안된 시스템의 효용성을 보이기 위한 실험과 그의 고찰을 한 뒤 마지막으로, 5장에서는 결론과 추후 과제에 대한 설명을 한다.

2. 배경지식

2.1 Genetic Algorithm (GA)

1970년대 초 John Holland는 유전학과 진화원리를 컴퓨터 알고리즘과 결합시키는 가능성을 연구하게 되었다. 오랜 노력 끝에 그는 이진 스트링의 개체집단 위에서 모의진화를 일으켜 효율적으로 최적의 해를 탐색하는 유전자 알고리즘(GA, Genetic Algorithm)을 제안하였다. 이 알고리즘은 생명체의 유전 및 진화과정을 전산학적으로 모델링(modeling)한 기계학습의 방법으로, 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 유전연산자에 의하여 진화과정을 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. 일반적인 유전자 알고리즘의 절차는 다음과 같다[5].

```
procedure GA
begin
  t=0;
  initialize P(t);
  evaluate P(t);
  while termination condition not satisfied do
  begin
    t=t+1;
    select P(t) from P(t-1);
    recombine and mutate P(t);
    evaluate P(t);
  end
end
```

초기 염색체들의 집단인 P(0)를 생성한다. 초기 집단은 해 공간 내에서 무작위로 분포되도록 선택되거나 경험적인 방법으로 선택된다. 그리고 각 염색체들의 적합도를 평가한 다음 평가된 적합도에 비례하여 선택된 염색체들을 P(t)에 복사한다. P(t)안의 염색체들에게 유전자 연산을 적용시킨 후 P(t)를 재생성한다. 그리고 그 결과를 바탕으로 자식 세대 P(t+1)을 생성한다.

2.2 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, Gentic Algorithm with Varying Population Size)이 제안되었다[7]. 이 알고리즘의 수행과정은 다음과 같다.

```
procedure GAVaPS
begin
  t=0;
  initialize P(t);
  evaluate P(t);
  while termination condition not satisfied do
  begin
    t=t+1;
    age(P(t))=age(P(t))+1;
    recombine and mutate P(t);
    evaluate P(t);
    remove chromosomes with age greater than
      their lifetime;
  end
end
```

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:

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

2. linear allocation:

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

3. bi-linear allocation:

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

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 의미

한다. 또한 일반적인 유전자 알고리즘과는 달리 개체군 변환 유전자 알고리즘에서는 개체군내의 모든 염색체는 염색체의 적합도와 무관하게 똑같은 확률로 교차와 돌연변이 등의 유전 연산자를 적용하여 얻는다.

3. 개체군 변화 유전자 알고리즘을 이용한 Feature Selection

제안하는 시스템의 염색체의 구성은 그림 1과 같다.

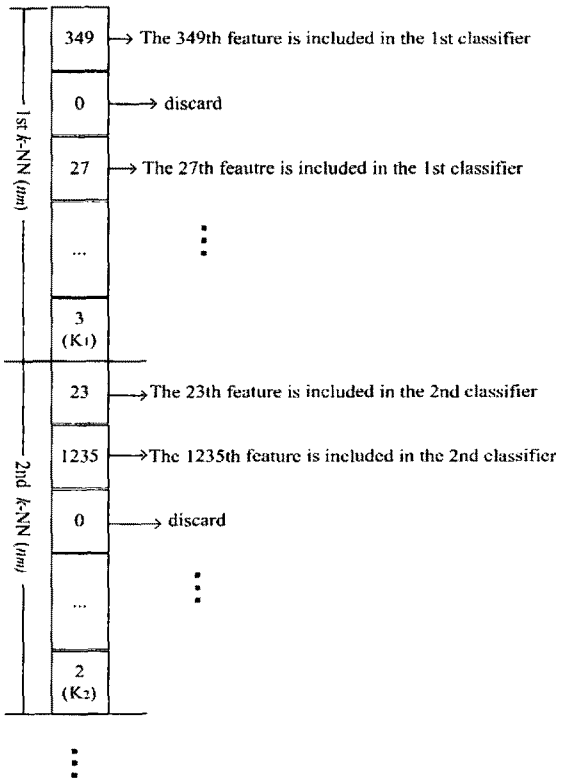

그림 1. 염색체의 구성.
Fig. 1. Structure of the chromosome.

이미 선택된 최대 특징의 수(N_m)에서 1을 더한 만큼의 숫자에 k-NN 분류기의 개수(N) 만큼 곱한 수 $((N_m+1)\times N\times bits)$의 유전인자가 유전자 알고리즘에서 사용할 염색체를 이룬다. 그림에서 볼 수 있는 염색체 "349, 0, 27, ..., 3, 24, 1245, ..."는 349번째 특징과 27번째 feature가 첫 번째 k-NN 분류기에서 사용되고 첫 번째 k-NN 분류기의 k값으로 3을 사용하는 것을 의미한다. 같은 방식으로 24번째와 1245번째 특징들은 두 번째 k-NN 분류기에서 사용된다. 만약 유전인자가 영 값을 갖는다면 그 특징은 포함되지 않기 때문에 각 분류기들은 각기 다른 수의 특징들을 가질 수 있다. 또한 여러 k-NN 분류기들은 다음과 같은 방법으로 결합하게 된다.

$$\omega_T = \arg max_\alpha C[\tau_\alpha(T)] \qquad (5)$$

여기에서 $C[.]$은 계수 함수 (counting function)이고 T는 테스트 데이터, ω_T는 T의 클래스이다. 또한 $\tau_\alpha(T)$는 α클래스에 포함되어 있는 T와 nearest neighbor (NN)의 관계를 갖고 있는 학습 데이터들의 집합이다. $\tau_\alpha(T)$는 n번째 k-NN이 사용하는 M개의 학습 데이터 $x_{n1},x_{n2},...,x_{nM}$중 T와 가까운 k_n개의 데이터들의 합집합으로 계산된다.

$$\tau_\alpha(T) = \cup_{n=1,...,N} NN_{nt}(T,x_{nj}) \quad \text{for } t=1,...,k_n,\ j=1,...,M \qquad (6)$$

여기에서 NN_{nt}는 n번째 k-NN 분류기의 t번째 nearest neighbor이다. 우리는 염색체에 k_n을 구하는 유전인자도 포함시키기 때문에, 분류기나 데이터의 분포상태에 맞는 최적의 특징들을 찾을 수 있을 뿐만 아니라 사용하는 여러 k-NN 분류기의 k값들도 자동으로 결정할 수 있다.

4. 실 험

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] (PCA, principal component analysis)을 적용하여 2000개의 특징을 추출하였다. 다음 그림은 두 개의 클래스의 샘플들을 보여준다.

그림 2. car 이미지.
Fig. 2. car images.
(Fig. 2, 3의 크기를 조절하였습니다.)

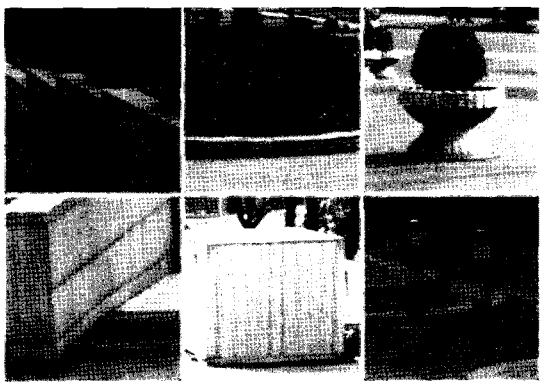

그림 3. non-car 이미지.
Fig. 3. non-car images.

우리는 데이터베이스를 각각 학습, 튜닝, 테스트 데이터로 나누었다. k-NN 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 다음과 같다.

표 1. 사용된 Database들.
Table 1. Used Databases.

종 류	개 수
Training data	60
Tuning data	30
Test data	30
Total	120

본 실험에 사용된 유전자 알고리즘과 개체군 변환 유전자 알고리즘의 파라미터 값을 표 2에 나타내었다.

표 2. 유전자 알고리즘의 파라미터들.
Table 2. Parameters for Genetic Algorithm.

Parameter	Value
Max. generation number	500
Crossover rate	0.7
Mutation rate	0.2
Max. feature number	300
Classifier number	3
MaxLT	20
MinLT	5

일반적인 유전자 알고리즘과 개체군 변환 유전자 알고리즘을 사용하여 다수 k-NN 분류기에서 사용되는 특징들을 선정하고 테스트 한 결과를 표 3에 도시하였다. 적은 trial을 갖고도 높은 정확도를 보여주기 때문에 일반적인 유전자 알고리즘(GA)에 비해 개체군 변환 유전자 알고리즘(GAVaPS)이 feature 선정 문제에 적합함을 표 3을 통해 확인할 수 있다.

표 3. 제안된 방법의 인식 결과.
Table 3. Results of the proposed method.

Methods	GA	GAVaPS
Tuning accuracy	100.0	100.0
Test accuracy	86.7	90.0
Trial	50,000	10,956

5. 결 론

일반적으로 패턴은 특징 공간에서 벡터로 표현되기 때문에, 특징들의 측정과 측정된 특징의 선택은 패턴 인식 시스템의 결과에 중요한 영향을 미친다. 따라서 패턴인식 시스템의 정확도를 높이기 위해서는, 적합한 특징의 선택이 매우 중요하다. 다수의 분류기들에서 사용되는 특징들을 선정하는 문제는 한 개의 분류기에서 사용되는 특징 선정보다 더욱 복잡하다. 이 문제를 해결하기 위하여 본 논문에서는 개체군 변환 유전자 알고리즘을 이용하여 특징들을 선정하는 방법을 제안하였다. 또한 다수의 k-NN을 유전자 알고리즘으로 결합하는 방법도 제시하였다. 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.

참 고 문 헌

[1] 이희성, "KNN규칙과 새로운 특징 가중치 알고리즘을 결합한 패턴 인식 시스템 설계," *연세대학교 석사학위논문*, 2005.

[2] 최병인, 이정훈, "영상 분할을 위한 퍼지 커널 K-nearest neighbor 알고리즘," *퍼지 및 지능 시스템학회 논문지*, vol. 15, no. 7, pp. 828-833, 2005.

[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] L. Hansen, L. Liisberg, and P. Salamon, "Ensemble methods for handwritten digit recognition," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 333-342, 1992.

[5] Z. Michalewicz, *Genetic algorithm + data structures = evolution programs*, Springer-Verlag, New York 1999.

[6] 권기호, "개체군 변환 유전자 알고리즘의 새로운 수명 할당 방식에 관한 연구," *전자공학회논문지*, 제36권 C편, 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, P. 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.

저 자 소 개

이희성(Heesung Lee)
2003년 : 연세대학교 전기전자공학부 졸업 (공학사)
2005년 : 연세대학교 전기전자공학부 석사과정 졸업(공학석사)
2005년~현재 : 동 대학원 전기전자공학과 박사과정

관심분야 : Computational intelligence, 로봇 비전, 패턴 인식
E-mail : 4u2u@yonsei.ac.kr

이제헌(Jaehun Lee)
2005년 : 연세대학교 전기전자공학부 졸업 (공학사)
2007년 : 연세대학교 전기전자공학부 석사과정 졸업(공학석사)
2007년~현재 : 동 대학원 전기전자공학과 박사과정

관심분야 : 지능형 Home network, 유전자 알고리즘
E-mail : aznable17@yonsei.ac.kr

김은태(Euntai Kim)
1992년 : 연세대학교 전자공학과 졸업 (공학사, 전체수석)
1994년 : 연세대학교 전자공학과 석사과정 졸업(공학석사)
1999년 : 연세대학교 전자공학과 박사과정 졸업(공학박사)
1999년 3월~2002년 2월 : 국립한경대학교 제어계측공학과 조교수
2002년 3월~현재 : 연세대학교 전기전자공학부 부교수
2003년 : University of Alberta, visiting researcher
1998년~현재 : IEEE TFS, IEEE SMC, IEEE CAS, FSS 등에서 심의위원 활동 중
2003년 : 대한 전자공학회 해동상 수상

관심분야 : Computational intelligence, 지능형 로봇
Phone : +82-2-2123-2863
E-mail : etkim@yonsei.ac.kr