유전 알고리즘을 이용한 특징 결합과 선택

Feature Combination and Selection Using Genetic Algorithm for Character Recognition

이진선
우석대학교 컴퓨터공학과

Jin-Seon Lee(jslee@woosuk.ac.kr)

요약

문자 패턴에서 추출한 서로 다른 특징 집합을 결합함으로써 문자 인식 시스템의 성능을 향상시킬 수 있다. 이때 결합된 특징 벡터의 차원을 줄이기 위해 특징 선택을 수행해야 한다. 이 논문은 문자 인식 문제에서 특징 결합과 선택을 위한 일반적인 틀을 제시한다. 또한 필기 숫자 인식을 위한 설계와 구현을 제시한다. 이 설계에서는 필기 숫자 패턴에서 DDD 특징 집합과 AGD 특징 집합을 추출하며 특징 선택을 위해 유전 알고리즘을 사용한다. 실험 결과 CENPARMI 필기 숫자 데이터베이스에 대해 0.7%의 정확률 향상을 얻었다

■ 중심어 : | 특징 결합 | 특징 선택 | 유전알고리즘 | 문자인식 |

Abstract

By using a combination of different feature sets extracted from input character patterns, we can improve the character recognition system performance. To reduce the dimensionality of the combined feature vector, we conduct the feature selection. This paper proposes a general framework for the feature combination and selection for character recognition problems. It also presents a specific design for the handwritten numeral recognition. In the design, DDD and AGD feature sets are extracted from handwritten numeral patterns, and a genetic algorithm is used for the feature selection. Experimental result showed a significant accuracy improvement by about 0.7% for the CENPARMI handwrittennumeral database.

■ keyword : | Feature Combination | Feature Selection | Genetic Algorithm | Character Recognition |

I. 서 론

좋은 특징의 설계는 높은 정확률을 갖는 문자 인식 시스템을 만드는데 매우 중요한 요소이다. 문자 인식에서 원래 입력은 스캔된 문서 영상에서 분할한 비트맵 패턴이다. 인식 시스템의 첫 번째 일은 이 비트맵에서 특징 집합을 추출하는 것이다. 지금까지 많은 특징 추출 방법이 제시되어 있다[1]. 이들 특징 집합은 인식 성능에서 매우 다양한 특성을 나타낸다는 것은 이미 잘 알려져 있다.

정보 결합 (information fusion)이 인식 성능 향상에 매우 효과적이라는 사실 때문에 이에 대한 연구는 갈수

* 본 연구는 한국과학재단 특정기초연구 (R01-2003-000-10879-0)의 지원으로 수행되었습니다.

접수번호 : #050920-002
접수일자 : 2005년 9월 20일
심사완료일 : 2005년 10월 05일
교신저자 : 이진선, e-mail : jslee@woosuk.ac.kr

록 폭이 넓어지고 그 깊이도 깊어지고 있다. 패턴 인식에서 정보 결합은 특징 수준과 분류기 수준의 두 가지 수준으로 나누어 볼 수 있다. 분류기 수준 결합은 다중 인식기 결합이라고 하며, 특징 수준에 비해 보다 오랫동안 연구되어 왔다 [2][3].

분류기 수준에 비해 특징 수준 결합에 대한 연구는 적은 편이다 [4-7]. 간단한 방법 중의 하나는 서로 다른 특징 집합을 단순히 붙여서 결합하는 것으로, 결합된 특징 집합을 인식기의 입력으로 한다. 이 방법을 구현하는 데에서 장애물은 특징 벡터 크기가 매우 커진다는 것이다. 따라서 특징 선택을 통해 적절한 크기의 특징 벡터로 만들어야 한다.

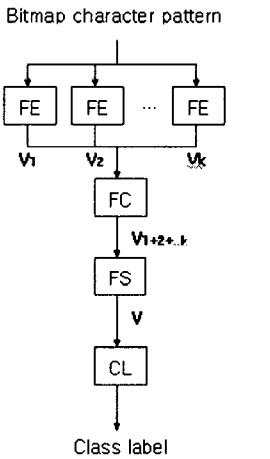
특징 결합과 선택의 직관적인 장점에도 불구하고 단지 몇 논문만이 이를 다루고 있다. 논문 [4]에서는 오목 모양 (concavity shape)이나 외곽 모양 (contour shape)과 같은 특징들을 결합하였다. 이 논문은 단지 적은 수의 특징만을 사용하였으며 특징 선택은 사용하지 않았다. 논문 [5]에서는 우리 논문과 비슷한 아이디어를 사용하였다. 하지만 특징 선택에서 개별 특징을 독립적으로 평가하여 정렬하는 방법을 사용하여 특징간의 상호 작용을 고려하지 않았다.

이 논문에서는 특징 결합과 선택을 위한 일반적인 틀을 제시한다. 제시한 방법의 효과를 측정하기 위해 필기 숫자 인식을 위한 특정한 설계와 실험을 기술한다. DDD와 AGD라는 두 개의 특징 집합을 추출하며 이들을 단순히 붙여 결합한다. 유전 알고리즘을 이용하여 특징 선택을 수행하며 분류기로는 모듈러 신경망을 사용하였다. 실험 결과 CENPARMI 숫자 데이터베이스에 대해 0.7%의 정확률 향상을 얻었다. 우리가 얻은 정확률은 현재 문헌에 보고된 state-of-the-art 성능과 비슷하거나 우수하다.

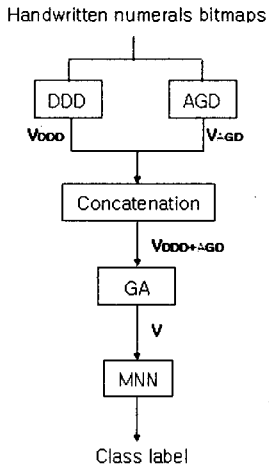
II. 일반 틀과 필기 숫자를 위한 설계

1. 일반 틀

우리가 제안하는 특징 결합과 선택을 위한 일반적인 틀은 [그림 1]에서 보듯이 간단하다. 문자 비트맵 패턴으로부터 여러 개의 특징 추출기 (FE: Feature Extractor)가 특징 벡터들 (Vi, $1 \le i \le k$) 을 추출한다. k는 결합에 참여하는 특징 벡터의 개수이다. 이 과정에서 특징 벡터들 간의 상호 보완성 (complementariness)이 고려될 수도 있다. 특징 결합기 (FC: Feature Combiner) 는 이들 특징 벡터를 결합하여 하나의 큰 특징 벡터를 만든다. 이후 처리 과정의 계산 부담을 덜기 위해 약한 특징들을 미리 제거하는 과정을 둘 수도 있다.



(a) 일반 틀



(b) 숫자 인식을 위한 특정 설계

그림 1. 문자 인식을 위한 특징 결합과 선택의 일반 틀과 특정 설계

특징 선택 (FS: Feature Selection) 과정은 결합된 큰 특징 벡터의 차원을 줄여 합리적인 크기로 바꾸어 준다. 이 과정을 위한 알고리즘은 전체 성능 향상에 중요한 요소로 작용한다.

2. 필기 숫자 인식을 위한 설계

[그림 1(b)]는 필기 숫자 인식을 위한 설계를 보여준다. 특징 추출을 위해서는 두 개의 알고리즘, DDD (Directional Distance Distribution)와 AGD(Area-based Gradient Distribution)를 사용하였다. 각각 256-차원을 갖는다. 보다 자세한 사항은 참고 논문 [8]과 [9]를 참고하기 바란다. 특징 결합기는 단순 붙이기 (concatenation)에 의해 512-차원의 특징 벡터를 만든다. 아래에서 두 가지 특징 벡터와 분류를 위한 모듈러 신경망을 기술한다. 특징 선택을 위해서는 유전 알고리즘을 사용하였다. 특징 선택 과정은 그 중요성 때문에 III장에서 따로 기술한다.

2.1 DDD 특징 벡터

입력 패턴 P_{m*n}을 우선 16*16 메쉬 R_{16*16}으로 크기 정규화하고 이를 이진 영상으로 변환한다. 이진 맵 R의 각 화소 p는 8 방향으로 광선을 쏘고, 각 광선은 p와 반대가 되는 색을 갖는 화소까지의 거리를 계산한다. 이 방향 거리 정보를 이용하여 방향 거리 분포 맵을 계산한다. 이 맵은 256개의 실수 값을 가지며, 그들 각각이 특징이 되어 256-차원의 특징 벡터를 만든다.

2.2 AGD 특징 벡터

크기 정규화된 메쉬 R에 Sobel 에지 연산자를 적용하여, 에지 강도 맵, M_{16*16}과 에지 방향 맵, D_{16*16}을 얻는다[10]. D 맵의 방향 0~360 도를 16단계로 계수화한다. 이렇게 얻은 방향 맵을 4*4의 16개의 블록으로 나눈다. 각 블록에 대해 16개 구간을 갖는 (16개 방향 각각이 하나의 구간) 히스토그램을 구한다. 히스토그램 누적 과정에서 에지 강도 정보를 가중치로 사용한다. 이렇게 얻은 16개의 히스토그램의 각 구간이 하나의 특징이 되어 총 256개의 특징을 갖는 특징 벡터가 만들어 진다.

2.3 부류-모듈러 신경망 분류기

분류기를 위해서는 논문 [11]이 제안한 부류-모듈러 신경망 구조를 사용하였다. 이 구조는 10개의 부 신경망으로 구성되는데, 각 부 신경망은 10개의 숫자 부류 중의 하나를 책임진다. 부류 i를 위한 부 신경망은 부류 i와 나머지 9개 부류를 분류하는 역할을 한다. 각 부 신경망은 오류 역전파 알고리즘을 사용하여 독립적으로 훈련된다.

III. 유전 알고리즘을 사용한 특징 선택

이 장에서는 결합된 특징 벡터의 차원을 줄이기 위해 사용된 유전 알고리즘에 대해 기술한다. 유전 알고리즘이 할 일은 D개의 특징으로부터 최적의 성능을 갖는 d개를 선택하는 것이다. 우리가 사용한 알고리즘은 참고문헌 [12]에 있는 알고리즘에 기초를 두며, 아래에 있는 안정형(steady-state) 제어 구조를 사용한다.

```
steady_state_GA()
{
    initialize population P;
    repeat {
        select two parents p1 and p2 from P;
        offspring = crossover(p1,p2);
        mutation(offspring);
        replace(P, offspring);
    } until (stopping condition);
}
```

사용한 유전 알고리즘의 상세한 명세는 다음과 같다.

1. 염색체 표현

D개의 비트를 갖는 이진 스트링을 사용한다. 하나의 비트는 특징 하나를 나타내며, 값 1과 0은 각각 선택된 상태와 선택되지 않은 상태를 의미한다. 예를 들어 D=8일 때 염색체 00101000은 세 번째와 다섯 번째 특징이 선택되어 있는 해를 나타낸다. 이는 X와 Y가 각각 선택

된 상태와 선택되지 않은 상태를 나타낼 때 X={3,5}와 Y={1,2,4,6,7,8}과 동일하다. 우리 문제에서 D의 개수는 수백 개이다.

2. 초기 해 집단

해 집단 (population)의 생성은 아래에 보여진 바와 같이 간단하다. 함수 random_uniform()은 [0,1] 사이의 임의의 실수를 생성한다. 임의의 초기 해에서 선택된 특징의 기대 개수는 d이다. P는 해 집단을, |P|는 해 집단 크기를 나타낸다.

```
초기 해 집단 :
for (i=1 to |P|)
  for (each gene g in i-th chromosome)
      if (random_uniform()<d/D) g=1;
       else g=0;
```

3. 적합도 계산, 선택, 대치, 그리고 중단

한 염색체의 적합도는 분류기의 인식 정확도이며 J(X)로 나타낸다. 이 논문에서는 J를 계산하기 위해 신경망을 사용한다. 특징 부분집합이 주어진 부분집합 크기 요구를 만족하도록 하기 위해, 부분 집합 크기가 d여야 한다는 사실을 제약 조건으로 설정하고, 이 조건을 만족하지 않는 염색체에는 벌점 (penalty)을 준다. 염색체 C의 적합도는 다음과 같이 정의한다. 여기서 XC는 C에 해당하는 부분 집합이고, w가 벌점 계수일 때 penalty(XC) = w*||XC|-d| 이다. 여기서 |XC|는 집합 XC의 크기이다.

```
fitness(C) = J(XC) - penalty(XC)
```

다음 세대를 위한 염색체 선택은 적합도에 기반하여 수행한다. 해 집단에 있는 염색체들의 적합도 차이가 작기 때문에 순위에 기반한 룰렛 휠 방법을 사용한다. 해 집단에 있는 염색체들을 적합도에 따라 정렬하고 i-번째 염색체에게 비선형 함수 $P(i)=q(1-q)^{i-1}$로 선택 확률을 부여한다.

안정형 GA에서, 위의 방법을 통해 두 개의 부모 염색체를 선택한다. 교배 연산은 두 부모로부터 새로운 자손 염색체를 생성하고, 돌연변이 연산은 이 염색체에 약간의 변형을 가한다. 이렇게 만들어진 염색체가 두 부모에 비해 모두 우세하면 두 부모 중에서 자신과 비슷한 부모를 대치한다. 만일 두 부모의 사이에 있다면 열세한 부모를 대치한다. 그렇지 않으면 해 집단에서 가장 열세한 염색체를 대치한다. 유전 알고리즘은 전체 세대 수가 미리 정한 최대 세대 수 T에 도달하면 중단한다.

4. 유전 연산자

일반적으로 사용하는 교배와 돌연변이 연산자를 약간 변형하여 사용한다. m개 자름 점을 임의로 선택한 후 두 부모의 부분 염색체를 서로 교차하여 자식 염색체를 만드는 m-점 교배 연산자를 사용한다. 돌연변이는 자식에게 적용한다. 돌연변이는 부분집합 크기 요구를 위반할 수 있으므로, 1-0 변환과 0-1 변환의 개수를 비슷하도록 조절할 필요가 있다. 아래 코드에서는 이 개수들을 유사하게 조절하며, 이때 매개변수 p_m은 돌연변이 확률을 나타낸다.

Controlled mutation:

1. Let n_0 and n_1 to be numbers of 0-bits and 1-bits in the chromosome.
2. $p_1=p_m$; $p_0=p_m \cdot n_1/n_0$;
3. for (each gene g in the chromosome)
4. Generate a random number r within [0,1].
5. if(g=1 and r<p_1) convert g to 0;
   else if(g=0 and r<p_0) convert g to 1;

IV. 실험과 토론

1. 실험

실험에서는 CENPARMI 필기 숫자 데이터베이스를 사용하였다. 이 데이터베이스는 훈련 집합이 4000개의 샘플, 테스트 집합이 2000개의 샘플을 가지고 있다. 실제 우편물에서 추출한 샘플이어서 객관적인 성능 실험에 많이 사용되고 있는 데이터베이스이다.

유전 알고리즘에서 매개변수 설정은 exploration과 exploitation 사이에 균형이 잘 이루어지도록 설정되어야만 좋은 성능을 얻을 수 있다. 우리 실험에서는 아래에 주어진 매개변수 설정을 사용하였다.

해 집단 크기 (|P|) = 20
돌연변이 확률 (P_m) = 0.1
q (순위 기반 선택) = 0.25
w (벌점 계수) = 0.5

[표 1]은 DDD와 AGD, 그리고 이들을 결합한 특징 벡터의 정확도를 보여준다. 신경망의 성능은 불안정성(unstability)을 가지므로 서로 다른 다섯 번의 훈련과 테스트를 수행하여 그들의 평균을 제시하였다. DDD가 AGD보다 좋은 성능을 보였다. DDD와 AGD를 결합한 특징 벡터가 DDD만 사용했을 때보다 열등한 성능을 보였다. 다섯 번의 평균 성능은 DDD, AGD, DDD+AGD가 각각 97.46%, 95.94%, 97.09%이다. 이 성능은 단순 결합이 항상 성능 향상을 가져오지는 않는다는 것을 보여주었다. 따라서 특징 선택이 필요함을 알 수 있다.

표 1. DDD와 AGD, 그리고 이들을 결합한 특징 벡터의 정확도

실행	DDD (256-D)	AGD (256-D)	DDD+AGD (512-D)
1	97.60%	95.90%	97.05%
2	97.35%	95.95%	97.05%
3	97.40%	95.80%	97.00%
4	97.50%	96.05%	97.25%
5	97.45%	96.00%	97.10%
평균	97.46%	95.94%	97.09%

특징 선택 과정에서 선택된 특징 부 집합의 크기는 200에서 300 사이로 지정하였다. 염색체의 적합도를 측정하기 위해 4-fold jackknife 기법을 사용하였다. 따라서 validation 정확도는 4-fold jackknife에서 얻은 네 개 정확도의 평균을 취하였다. 테스트 정확도는 테스트 집합으로 측정하였다. 선택된 특징의 개수도 [표 2]에 제시하였다.

표 2. 다섯 번의 독립된 특징 선택에 의한 정확도 (검증은 4-fold 잭 나이프 기법에 의한 것임)

GA 실행	특징 개수	정확도	
		검증	테스트
1	288	98.35%	98.35%
2	262	98.15%	98.25%
3	288	98.50%	98.40%
4	273	98.30%	97.85%
5	249	98.05%	97.90%
평균	272	98.27%	98.15%

2. 토론

[표 2]에서 가장 높은 정확도는 98.50%이었다. 이 값은 [표 1]의 DDD, AGD, DDD+AGD의 가장 좋은 정확도에 비해 0.9% 향상된 것이다. [표 1]과 [표 2]에 있는 평균을 비교하면, 특징 선택으로 얻은 특징 벡터는 98.15%이고, DDD는 97.46%이다. 이 사실로부터 우리가 제안한 특징 결합과 선택 방법으로 0.69%의 성능 향상을 얻었음을 알 수 있다. 선택된 특징 벡터는 272개의 특징을 가지므로 DDD 특징 벡터에 비해 단지 16개 특징을 더 갖는다.

우리가 사용한 유전 알고리즘의 단점은 느리다는 것이다. 512-차원의 특징 벡터에서 200~300개의 특징을 선택하기 위해 약 1주일의 시간이 소요되었다. 이러한 느린 속도 문제를 해결하기 위해 Beuwulf로 구현한 Linux 클러스터 컴퓨터를 사용하였다. 이 클러스터 컴퓨터는 1GHz dual Pentium 프로세서를 갖는 노드 여러 개로 구성되어 있다.

우리의 결론은 아래와 같다.

- 여러 개의 특징 벡터를 단순 결합하는 방법은 인식 성능을 저하시킬 수 있다. 따라서 적절한 특징 선택 방법이 매우 중요하다.
- 특징 결합과 선택 방법은 문자 인식 시스템의 성능 향상에 매우 효과적이다.
- 매개 변수가 잘 설정된 유전 알고리즘은 특징 결합과 선택 방법을 구현하는데 매우 효과적이다.
- 우리가 얻은 인식 성능은 필기 숫자 인식의 state-of-the-art 성능에 비해 비슷하거나 우수하다. 논문 [13]에서 제시된 Liu의 CENPARMI 데이터베이스

에 대한 벤치마킹 자료에 의하면, 논문 [14]의 Liu 방법이 98.45%로서 가장 우수한 성능으로 보고하였다.

V. 결 론

문자 인식을 위한 특징 결합과 선택 기법에 대한 일반적인 틀과 특정 설계를 기술하였다. 제안한 방법은 단순하지만 효과적인 성능 향상을 기대할 수 있다. 필기 숫자 인식을 이용한 실험으로 이러한 성능 향상을 입증하였다. 우리가 얻은 인식 성능은 state-of-the-art 기술의 성능과 비슷하거나 우수하였다. 이 논문의 또 다른 공헌은 유전 알고리즘이 특징 결합과 선택 방법을 구현하는데 좋은 도구임을 보인 것이다.

특징 수준의 결합과 분류기 수준의 결합을 동시에 사용하는 접근 방법을 중요한 향후 연구로 고려하고 있다. 이 접근 방법은 또 다른 큰 성능 향상을 제공할 것으로 기대된다. 또 다른 향후 연구로는 유전 알고리즘의 속도 향상이다. 약한 특징을 미리 제거하는 방법도 고려 대상이다. 보다 많은 특징 벡터를 사용하고 그들 간의 상호 보완성을 고려하는 것도 향후 연구 중의 하나이다.

참 고 문 헌

[1] Q. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition-a survey," Pattern Recognition, Vol.29, No.4, pp.641-662, 1996.

[2] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," IEEE Tr. PAMI, Vol.16, pp.66-75, 1994.

[3] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," IEEE Tr. PAMI, Vol.27, No.6, pp.942-956, 2005.

[4] J. T. Favata, G. Srikantan, and S. N. Srihari, "Handprinted character/digit recognition using a multiple feature/resolution philosophy," Proceedings of IWFHR'94, Taiwan, pp.57-66, 1994.

[5] I. S. Oh, J. S. Lee, and C. Y. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition," IEEE Tr. PAMI, Vol.21, No.10, pp.1089-1094, Oct. 1999.

[6] J. Yang, J. Y. Yang, D. Zhang, and J. F. Lu, "Feature fusion: parallel vs. serial strategy," Pattern Recognition, Vol.36, pp.1369-1381, 2003.

[7] Q. S. Sun, S. G. Zeng, Y. Liu, P. A. Heng, and D. S. Xia, "A new method of feature fusion and its application in image recognition," Pattern Recognition,(in printing).

[8] I. S. Oh and C. Y. Suen, "Distance features for neural network-based recognition of handwritten characters," International Journal on Document Analysis and Recognition, Vol.1, No.2, pp.73-88, 1998.

[9] G. Srikantan, S. W. Lam, and S. N. Srihari, "Gradient-based contour encoding for character recognition," Pattern Recognition, Vol.29, No.7, pp.1147-1160, 1996.

[10] L. G. Shapiro and G. C. Stockman, Computer Vision, Prentice Hall, 2001.

[11] I. S. Oh and C. Y. Suen, "A class-modular feedforward neural network for handwriting recognition," Pattern Recognition, Vol.35, pp.229-244, 2002.

[12] I. S. Oh, J. S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," IEEE Tr. PAMI, Vol.26, No.11, pp.1424-1437, Oct. 1999.

[13] C. L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition:

benchmarking of state-of-the-art techniques," Pattern Recognition, Vol.36, pp.2271-2285, 2003.

[14] C. L. Liu and M. Nakagawa, "Handwritten numeral recognition using neural networks: improving the accuracy by discriminative training," 5th International Conference on Document Analysis and Recognition, pp.257-260, 1999.

저 자 소 개

이 진 선(Jin-Seon Lee) 정회원

- 1985년 2월 : 전북대학교 전산통계학과(이학사)
- 1988년 2월 : 전북대학교 전산통계학과(이학석사)
- 1995년 8월 : 전북대학교 컴퓨터공학과(공학박사)
- 1988년 2월~1992년 : 한국전자통신연구원 연구원
- 1995년 3월~현재 : 우석대학교 컴퓨터공학과 교수