GA를 이용한 특징 가중치 알고리즘과 Modified KNN규칙을 결합한 Classifier 설계

The Design of a Classifier Combining GA-based Feature Weighting Algorithm and Modified KNN Rule

이희성*, 김은태**, 박민용***
Hee-sung Lee, Euntai Kim, Mignon Park

Abstract - This paper proposes a new classification system combining the adaptive feature weighting algorithm using the genetic algorithm and the modified KNN rule. GA is employed to choose the middle value of weights and weights of features for high performance of the system. The modified KNN rule is proposed to estimate the class of test pattern using adaptive feature space. Experiments with the unconstrained handwritten digit database of Concordia University in Canada are conducted to show the performance of the proposed method.

Key Words : Pattern recognition, Genetic Algorithm, Feature Weighting, Knn rule

1. 서 론

패턴 인식 시스템의 성능은 보통 인식률에 의해 좌우된다. 패턴 인식 시스템의 인식률을 높이기 위해서는 다음과 같은 두 가지 요소를 고려하여야 한다. 첫째는, 패턴에서 쓸모없는 특징들은 최소화하면서, 중요한 특징들을 추출하는 특징 추출기의 설계이다. 다음으로는 패턴의 분류 에러를 일반적인 데이터에서도 최소화하는 분류기(classifier)의 설계가 중요하다. 본 논문에서는 다음과 같은 두 가지 접근을 고려하여 높은 성능을 갖는 분류기를 설계하려고 한다.

전통적인 패턴인식 시스템에서, 패턴들은 일반적으로 특징공간(feature space)에서의 벡터로 표현되기 때문에, 특징들의 추출과 또 추출된 특징의 선택은 패턴 분류 알고리즘의 결과에 중요한 영향을 미친다. 따라서 패턴을 정확하게 분류하기 위해서는, 적합한 특징의 추출은 매우 중요하다. 하지만 어느 특징들이 클래스들 간의 가장 좋은 차별성을 제공하는지 알 수 없다. 또한 패턴을 표현 가능하게 하는 선택된 하위 특징공간의 구성도 무수히 많다. 그 결과 특징의 추출과 추출된 특징의 선택들은 패턴 인식 시스템의 성능에 중요한 역할을 한다[1].

자연 선택과 자연 발생의 과정을 기초로 다수의 개체를 동시에 진화시켜 가면서 최적의 해를 찾는 유전자 알고리즘은 많은 최적화 문제에서 사용되고 있다. 패턴 인식 시스템의 정확도를 향상시키면서 특징의 숫자를 줄이기 위해 적절한 특징공간을 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, m-차원($m < d$)으로 변환하는 것이다. 일반적으로 변화된 패턴의 특징 벡터들은 그들의 차원, 클래스의 분리 또는 정확성에 의거하여 평가를 받기 때문에 분류기나 데이터의 분포상태에 맞는 최적의 특징 공간을 찾을 수 있다[2].

하지만 기존의 유전자 알고리즘을 기본으로 한 특징 추출 방식은 overfitting 문제로 분류기의 성능이 저하된다. 또한 모든 클래스에 같은 특징 공간을 사용하여 클래스에 따라 각각 최적의 특징 공간을 찾을 수 있는 점을 간과하였다.

본 논문에서는 유전자 알고리즘을 이용한 적응적 특징 가중치(Adaptive-3FW)방식과 클래스별로 적용된 KNN규칙을 결합한 새로운 패턴 인식 시스템을 제안한다. 우선, 학습 데이터의 특징 벡터에 적응적-3FW알고리즘을 각 클래스별로 적용하여, 각 클래스를 표현하는 최적의 특징 공간들을 구한다. 다음으로 미지패턴의 클래스를 결정하기 위하여, 선들로 이루어져 있는 숫자들의 특성을 이용하여 미지패턴의 특징을 추출한다. 클래스에 따라 특징 공간이 적응적(adaptive)으로 변하고, 각 클래스별 특징 공간에서 최소 거리를 갖는 학습 데이터의 클래스를 미지패턴의 클래스로 선택한다.

본 논문의 구성은 다음과 같다. 2장에서는 우선 배경지식으로 FW알고리즘에 대하여 설명하고, 유전자 알고리즘과 KNN규칙을 이용한 새로운 분류 시스템을 제안한다. 3장에서는 제안한 시스템의 효용성을 보이기 위한 실험과 그의 고찰을 한 뒤 마지막으로, 4장에서는 결론과 추후 과제에 대한 설명을 한다.

2. 패턴인식 시스템의 설계

2.1 FW알고리즘

특징추출 단계에서 중복되거나, 부적절한 특징을 추출할 수 있다. 이런 특징들이 패턴 인식 시스템의 입력으로 선택되면 시스템의 성능은 떨어진다. 하지만 일반적으로 특징의 선택은 사람의 경험이나 지식에 의존하였다. 유전자 알고리즘을 이용한 특징 벡터의 선택 알고리즘(FS, Feature Selection)은 Siedlecki와 Sklansky에 의해 처음 소개되었다[3]. 그들의 연구에서, 유전자 알고리즘은 특징과 연관되어있는 최적의 이진 벡터를 찾기 위해 사용되었다. 만약, 이진 벡터에서 i번째 비트가 1이라면 i번째 특징은 분류기의 입력에 포함된다. 반대로 만약, 비트가 0이라면 대응하는 i번째 특징은 입력에 포함되지 않는다. 유전자 알고리즘의 염색체인 이

저자 소개
* 學生會員 : 延世大學 電氣電子工學科 碩士課程
** 正 會 員 : 延世大學 電氣電子工學科 助敎授 · 工博
***正 會 員 : 延世大學 電氣電子工學科 正敎授 · 工博

진 벡터에 의해 만들어지는 특징들의 부분집합은 분류기의 정확도에 의해 평가된다. 유전자 알고리즘은 이 정보를 이용하여 최적의 특징들의 부분집합을 찾게 된다.

FS알고리즘의 단점은 사용되는 특징들이 모두 같은 중요함을 갖는다는 것이다. 간단한 해결책은 특징에 가중치를 부여하는 것이다. Kelly와 Davis[4]는 각각의 특징들을 독립적으로 선형 스케일링(linear scaling)할 수 있도록 염색체의 각각의 특징에 대응하는 비트들을 실수로 확장하였다.(FW, Feature Weight)

특징 벡터가 식(1)과 같이 주어지면,

$$X = \{x_1, x_2, x_3, ..., x_d\} \quad (1)$$

유전자 알고리즘은 이 특징 벡터를 식(2)와 같이 바꾼다.

$$X' = \{w_1 x_1, w_2 x_2, w_3 x_3, ..., w_d x_d\} \quad (2)$$

여기서 w_i는 i번째 특징과 대응되는 비트에 들어가는 수이다. 즉, 각 특징에 대응하는 가중치(weight)가 된다. 이런 선형 스케일링은 FS알고리즘에 비해 자세하게 분류기가 미지패턴을 구분할 수 있도록 도와준다. 만약 가중치가 1, 0 으로만 구성되어 있다면, FS알고리즘과 같아지기 때문에 FS알고리즘은 FW알고리즘의 특별한 경우이다. 하지만 과잉 맞춤 문제로 각 염색체의 비트가 실수의 가중치로 구성된 일반적인 FW알고리즘 보다는, 가중치가 0, 0.5, 1만을 고려하는 3FW알고리즘이 일반적인 경우에 좋은 성능을 갖는다고 알려져 있다[5].

2.2 패턴 인식 시스템 설계

2.2.1 Feature Extraction

숫자들은 2차원 공간에서 1차원 선들의 집합이다. 그러므로 선들의 지역적 검출은 숫자의 적절한 특징 추출중의 하나이다. 본 논문에서는 빠른 계산과 숫자들을 구성하는 선들의 검출을 모두 만족시키는 일차 미분 에지 디텍터인 Kirsch edge detector를 사용한다. Kirsch는 비선형 에지 향상 알고리즘을 다음과 같이 정의 하였다.

$$G(i,j) = max\{max[1, |5S_k - 3T_k|]\} \quad (3)$$

여기서,

$$S_k = A_k + A_{k+1} + A_{k+2},$$

$$T_k = A_{k+3} + A_{k+4} + A_{k+5} + A_{k+6} + A_{k+7}.$$

식(3)에서 $G(i,j)$는 (i,j)픽셀의 기울기(gradient) 값이고, $A_k(k=0,1,...,7)$는 픽셀 (i,j)의 8-Neighbor들이다.

우선, 입력 패턴을 16×16으로 크기를 정규화 시킨 후, 수평(H), 수직(V), 우-대각(R), 좌-대각(L)의 방향 특징 벡터를 아래와 같은 식으로 계산한다.

$$\begin{aligned} G(i,j)_H &= max[|5S_0 - 3T_0|, |5S_4 - 3T_4|], \\ G(i,j)_V &= max[|5S_2 - 3T_2|, |5S_6 - 3T_6|], \\ G(i,j)_R &= max[|5S_1 - 3T_1|, |5S_5 - 3T_5|], \\ G(i,j)_L &= max[|5S_3 - 3T_3|, |5S_7 - 3T_7|]. \end{aligned} \quad (4)$$

각각의 4×4의 픽셀을 하나의 픽셀로 축적시키는 방법으로 식(4)로 계산된 4개의 16×16 특징 방향 벡터들을 4×4의 특징 벡터들로 압축시킨다.

위에서 계산된 $4\times4\times4$의 특징 벡터는 입력 패턴의 지역적인 특징만을 고려하였다. 이것을 보완하기 위하여, 16×16의 입력 영상을 4×4로 압축시키는 광역 특징 벡터를 고려한다. 그 결과, 최종 특징 벡터는 $4\times4\times4$의 지역 특징 벡터와 $1\times4\times4$의 광역 특징 벡터로 구성되어 있다. 그림 1은 특징을 추출하는 전체적인 과정을 보여준다[6].

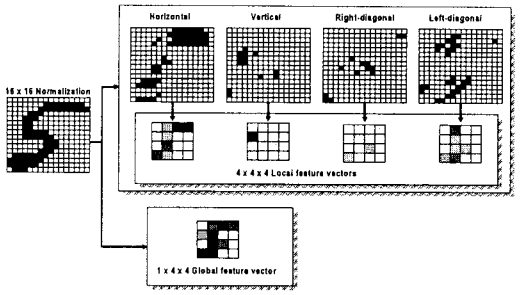

그림 1. 특징 추출의 전체 과정

2.2.2 Adaptive 3FW

본 논문에서는 기존의 3FW의 방법과는 달리 weight의 중간값을 유전자 알고리즘으로 결정하는 adaptive-3FW알고리즘을 제안한다. 또한 새로운 교차 연산자와 돌연변이 연산자를 제안한다.

제안하는 알고리즘에서 사용되는 유전자 알고리즘의 염색체는 특징 벡터의 수에서 1을 더한 만큼의 숫자로 구성된다. 염색체의 가장 마지막 자리에 들어오는 M은 0과 1사이의 실수로써 weight의 중간값이다. 일반적인 3FW알고리즘이 0, 0.5, 1의 웨이트를 갖는 반면 적응적 3FW알고리즘은 0, M, 1의 웨이트를 갖고 있어, 3FW의 장점인 과잉 맞춤(overfitting)을 피하면서, 데이터의 분포에 따라 적절한 특징의 가중치를 찾을 수 있는 장점을 갖는다.

일반적인 교차 연산을 제안하는 알고리즘에서는 적용하지 못한다. 이런 문제점을 해결하기 위하여 수정된 교차 연산을 제안한다. 수정된 교차 연산은 그림 2와 같이 정의된다.

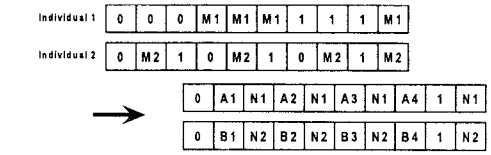

그림 2. 수정된 교차연산

$N1, N2$는 다음과 같다.

$$N1 = \frac{aM1 + (1-a)M2}{2}, \quad N2 = \frac{aM2 + (1-a)M1}{2} \quad (5)$$

여기서 a는 0과 1사이의 임의의 실수이다.

나머지 $A1, A2, A3, A4$는 각각 다음과 같이 정의된다.

$$A1 = \begin{cases} M1 & M2 \geq 0.5 \\ 0 & M2 < 0.5 \end{cases}, \quad A2 = \begin{cases} N1 & M1 \geq 0.5 \\ 0 & M1 < 0.5 \end{cases}$$
$$A3 = \begin{cases} 1 & M1 \geq 0.5 \\ M1 & M1 < 0.5 \end{cases}, \quad A4 = \begin{cases} 1 & M2 \geq 0.5 \\ M1 & M2 < 0.5 \end{cases} \quad (6)$$

$B1, B2, B3, B4$도 같은 방식으로 다음과 같이 정의된다.

$$B1 = \begin{cases} N2 & M2 \geq 0.5 \\ 0 & M2 < 0.5 \end{cases}, \quad B2 = \begin{cases} N2 & M1 \geq 0.5 \\ 0 & M1 < 0.5 \end{cases}$$
$$B3 = \begin{cases} 1 & M1 \geq 0.5 \\ N2 & M1 < 0.5 \end{cases}, \quad B4 = \begin{cases} 1 & M2 \geq 0.5 \\ N2 & M2 < 0.5 \end{cases} \quad (7)$$

교차 연산과 마찬가지로 일반적인 돌연변이 연산을 제안된 알고리즘에 직접적으로 적용할 수 없기 때문에 수정된 돌연변이 연산을 다음과 같이 정의한다.

Procedure Modified mutation operator
begin
generate random value r1 from the range [0,1];
if r1 < mutation_probability
Then generate random value N from the range [0,1];
last bit(M) = N;
t=0;

```
while t ≤ (size of chromosome)
begin
  t=t+1;
  generate random value r2 from the range [0,1];
  if r2 < mutation_probability
  Then mutate bit within {0, M, 1};
end
end
```

염색체(특징의 가중치 벡터)를 평가하기 위하여, 식(8)과 같은 평가 함수를 사용한다.

$$Maximize \ : \ fitness(w) = -(C_{pred}N_1 + C_{mask}N_2) \quad (8)$$

여기서 C_{pred}, C_{mask}는 평가 함수의 계수이고, N_1은 조율 데이터(tuning data)를 잘못 분류한 개수, N_2는 가중치 벡터에서 0이 아닌 가중치의 개수로 정의된다.

유전자 알고리즘의 염색체인 특징의 가중치와 학습 데이터를 곱하여 수정된 학습 데이터를 얻는다. 그리고 수정된 학습 데이터의 패턴과 조율 데이터의 패턴과의 거리를 계산하여 가장 최소 거리를 갖는 학습 데이터 패턴의 클래스를 조율 데이터 패턴의 클래스로 결정한다. 이렇게 계산된 조율 데이터 패턴의 클래스를 예측한 결과와 실제 조율 데이터 패턴의 클래스가 다른 개수의 합이 N_1이 된다.

평가 함수에서 N_1항은 가중치 벡터가 높은 정확도를 갖는 방향으로, N_2항은 가중치 벡터가 0을 많이 갖도록 하여, 측정해야할 특징의 숫자를 줄여가는 방향으로 유전자 알고리즘의 해가 수렴하도록 한다.

2.2 Classifier 설계

지금까지의 유전자 알고리즘을 기반으로 한 FW알고리즘은 학습 데이터의 모든 클래스를 가장 잘 표현할 수 있는 특징 공간을 찾는데 주력하였다. 이것과는 달리 본 논문에서는 모든 클래스가 아닌 각각의 클래스를 가장 잘 표현하는 특징 공간들을 개별적으로 찾는다. 우선, 적응적 3FW알고리즘을 학습 데이터의 각 클래스별로 적용하여, 각 클래스를 나타내는 가중치 벡터들을 구한다.

미지 패턴이 들어오면 앞 절에서 설명한 특징 추출 방법을 이용하여 $5\times4\times4$의 특징 벡터를 계산한다. 미지 패턴의 특징 벡터는 클래스에 따라 적응적(adaptive)으로 변하고, 각 클래스별 특징 공간에서 최소 거리를 갖는 학습 데이터의 클래스를 미지 패턴의 클래스로 선택한다.

3. 실 험

본 논문에서 제안된 분류 시스템을 검증하기 위하여 그림2와 같은 Concordia 대학의 handwritten numeral database를 사용 하였다. 6000개의 데이터 중 3500개는 학습데이터, 500개는 조율데이터, 2000개는 테스트데이터로 사용되었다.

Concordia numeral database를 이용하여 제안된 알고리즘의 효용성을 알아보기 위한 실험을 한다. 제안된 방법의 인식 결과는 표 1과 같다.

Methods	Accuracy (Tuning)	Accuracy (Test)	Zero features
1NN	91.80	89.70	0
3FW	91.80	90.30	24
Adaptive-3FW	92.00	91.25	22
Proposed Method	93.80	93.75	23

표 1. 제안된 알고리즘의 인식 결과

기존의 방법인 3FW알고리즘은 1NN알고리즘에 비해 인식률이 높다. 하지만 제안된 방법들은 3FW 알고리즘보다도 정확도가 높음을 표 1에서 확인할 수 있다.

제안된 방법의 정확도와 다른 방법들의 정확도 및 신뢰도의 비교는 표 2와 같다.

Methods	Accuracy	Substituted	Reliability
1NN	89.70	10.30	89.70
3FW	90.30	9.70	90.30
Adaptive-3FW	91.25	8.75	91.25
Proposed Method	93.75	6.25	93.75

표 2. 다른 방법과의 정확도 비교

사용된 유전자 알고리즘의 파라미터의 값은 다음과 같다.

$C_{pred} = 20, C_{mask} = 1, Pop\ size = 40, P_m = 0.6, P_c = 0.05$

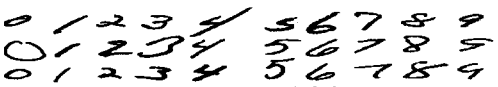

그림 2. 샘플 데이터

4. 결 론

본 논문에서는 유전자 알고리즘을 이용한 적응적 특징 가중치(Adaptive-3FW)방식과 클래스별로 적용된 KNN규칙을 결합한 새로운 패턴 인식 시스템을 제안하였다.

적응적 특징 가중치(Adaptive-3FW) 알고리즘은 기존의 3FW 방법과는 달리 가중치의 중간값을 유전자 알고리즘으로 결정함으로써 과잉 맞춤(overfitting)을 피하면서, 데이터의 분포에 따라 적절한 특징의 가중치를 찾을 수 있는 장점을 갖는다. 또한, 본 논문에서는 모든 클래스가 아닌 각각의 클래스를 가장 잘 표현하는 특징 공간들을 개별적으로 찾았다. 그리고 계산된 특징 공간들을 이용해 KNN분류기는 클래스에 따라 특징공간을 변화시켜 미지 패턴의 클래스를 예측하였다.

참 고 문 헌

[1] A. K. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 153-158, Feb. 1997.

[2] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn and A. K. Jain, "Dimensionality Reduction Using Genetic Algorithms," *IEEE Tran. Evolutionary Computation*, vol. 4, pp. 164-171, July 2000.

[3] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letter*, vol. 10, pp. 335-347, 1989.

[4] J. D. Kelly and L. Davis, "Hybridizing the genetic algorithm and the k nearest neighbors classification algorithm," in *Proc. 4th Int. Conf. Genetic Algorithms Appl.*, 1991, pp. 377-383

[5] Kohavi, R. Langlry, and Y. Yun, "The utility of feature weighting in nearest neighbor algorithms," *European Conf. Machine Learning*, ECML'97. 1997.

[6]S. Lee, "Off-Line Recognition of Totally Unconstrained Handwritten Numerals Using Multilayer Cluster Neural Network," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, pp. 648-652, June 1996.