다중 특징의 반복적 분석에 의한 퍼지 분류기의 설계

Design of a Fuzzy Classifier by Repetitive Analyses of Multifeatures

신 대 정*, 나 승 유**
Dae-Jung Shin*, Seung-You Na**

요 약

유전자 알고리즘을 이용한 다양한 특징의 분석이 필요한 퍼지 분류기의 설계 방법을 제안한다. 본 논문에서 제안한 퍼지 분류기는 퍼지 논리를 이용한 분류 부분과 유전자 알고리즘을 이용한 규칙 생성 부분으로 구성된다. 유전자 알고리즘을 이용한 규칙 생성 부분에서는 최적의 퍼지 멤버쉽 함수를 결정하고, 각 특징이 규칙에 포함되는지 포함되지 않는지의 여부도 결정하게 된다. 또한, 특정 대상에 대한 인식률을 분석하여 큰 오인식률을 갖는 부분에 세부 특징을 추가하는 방법과 문자열과 population의 최소 크기, 인식률 개선을 위한 반복적 분석 방법을 사용한다. 제안된 퍼지 분류기의 적용 예로서, 아이리스 데이터와 갑상선 종양 세포, 그리고 필기된 숫자와 인쇄된 숫자의 인식을 든다. 필기된 숫자와 인쇄된 숫자의 인식을 위해서 각 숫자를 구조적인 정보가 동일한 그룹으로 분류한다. 본 논문에서 제안한 퍼지 분류기는 아이리스 데이터에 대해 98.67%의 인식률을, 갑상선 종양 세포에 대해서 98.25%의 인식률을, 필기된 숫자와 인쇄된 숫자에 대해서 96.3%의 인식률을 얻었다.

ABSTRACT

A fuzzy classifier which needs various analyses of features using genetic algorithms is proposed. The fuzzy classifier has a simple structure, which contains a *classification part* based on fuzzy logic theory and a *rule generation part* using genetic algorithms. The rule generation part determines optimal fuzzy membership functions and inclusion or exclusion of each feature in fuzzy classification rules. We analyzed recognition rate of a specific object, then added finer features repetitively, if necessary, to the object which has large misclassification rate. And we introduce repetitive analyses method for the minimum size of string and population, and for the improvement of recognition rates. This classifier is applied to three examples of the classification of iris data, the discrimination of thyroid gland cancer cells and the recognition of confusing handwritten and printed numerals. In the recognition of confusing handwritten and printed numerals, each sample numeral is classified into one of the groups which are divided according to the sample structure. The fuzzy classifier proposed in this paper has recognition rates of 98.67% for iris data, 98.25% for thyroid gland cancer cells and 96.3% for confusing handwritten and printed numerals.

*전남대학교 대학원 전자공학과
**전남대학교 전자공학과

Ⅰ. 서 론

Zadeh 교수에 의해 제창된 퍼지 논리 이론이 다양한 분야에서 그 적용이 시도되고 좋은 성과를 보인 것은 퍼지 논리가 갖는 추론 능력과 언어의 불분명함이나 개념이 잘 정의되지 않는 모호한 현상에 적용할 수 있는 강력한 일반화 능력 때문이다.[1] 전문가에 의하여 판단되는 결론은 다분히 주관적이고, 그 판단 근거 또한 모호한 경우가 많다. 퍼지 논리 이론은 이와 같은 모호한 정보를 효과적으로 다룰 수 있는 방법을 제시해 주었다.[2] 퍼지 논리 이론은 주로 "If ~ Then ~" 규칙을 사용하며 제어 문제, 패턴 인식 문제 등을 포함한 여러 분야에 널리 적용되고 있다. 대부분의 적용 분야에서, 퍼지 규칙은 전문가의 지식으로부터 얻어지고 있으며, 최근 들어 수치 데이터에 의해 퍼지 규칙을 자동으로 생성하는 많은 연구가 활발히 실행되어 가고 있으며, 패턴 분류 문제에서도 자동화된 퍼지 분류 규칙의 생성을 제안하고 있는 추세이다.[3][4]

유전자 알고리즘(Genetic Algorithms)은 적자 생존의 자연 선택과 자연계의 유전학에 근거한 탐색 알고리즘이다. 특히 유전자 알고리즘은 패턴 인식과 최적화 문제에 그 유용성이 뛰어나다고 알려져 있다.[5]

패턴 인식에 있어 특징 선택(Feature Selection)은 분류 대상이 가지는 초기의 많은 개수의 특징 중 다른 대상과 차별되는 소수 특징들의 선택이라 할 수 있다. 다른 대상과의 명확한 차별성을 갖는 소수 특징의 선택은 가장 이상적이고 효율적인 방법이다. 하지만, 패턴 인식에서 인식기에 대한 연구가 아주 활발한 반면 특징 선택에 대한 연구는 아주 미비하며 최적의 특징을 선택하는 일반적인 방법은 없는 것으로 알려져 있다.[3]

또한, 인식기에 있어 미리 선택된 특징은 인식기의 성능에 큰 영향을 미치며, 대부분의 경우 특징의 인식 성능에 대한 기여 정도에 따른 가중치의 변화 혹은 기여도의 변화로 열등한 특징 변수는 인식 과정에 작은 영향만을 미치게 된다. 하지만 이러한 경우에도 인식기의 입력으로 사용되는 모든 특징 변수들은 그 영향이 작든 크든 인식 과정에 반드시 포함되며 이는 인식기의 성능과도 직접 관련될 수 있다.[3]

제안된 퍼지 분류기는 퍼지 논리를 이용한 분류 부분과 유전자 알고리즘을 이용한 간단한 규칙 생성 부분으로 구성된다. 퍼지 논리를 이용한 분류 부분은 표준 패턴과 입력 패턴과의 유사도를 측정하고, 입력 패턴과 가장 큰 유사도를 갖는 클래스가 입력 패턴이 소속되어 있는 클래스라는 비교적 간단한 개념에서 출발한다. 유전자 알고리즘을 이용한 규칙 생성 부분에서는 최적의 퍼지 멤버쉽 함수를 결정하고, 각 특징이 규칙에 포함되는지 포함되지 않는지의 여부도 결정하게 된다.

제안된 퍼지 분류기는 최소의 규칙 선택을 위한 방법으로써 1) 특징의 개수 및 문자열의 개수를 늘려 가는 반복적 분석 방법과 2) 최소의 문자열의 길이와 population의 크기를 위하여 규칙 선택 비트를 사용하는 방법, 그리고 그 성능에 대해 논하며, 세 가지 예제에 적용하여 제안한 방법의 효용성을 검증한다.

Ⅱ장에서 표준 패턴과 입력 패턴을 이용한 퍼지 분류기의 간단한 개념을, Ⅲ장에서는 유전자 알고리즘을 이용한 퍼지 멤버쉽 함수의 생성과 규칙에의 포함 여부에 대하여 설명한다. 그리고 Ⅳ장에서는 본 논문에서 제안한 방법을 아이리스 데이터, 갑상선의 종양 세포 데이터, 필기된 숫자와 인쇄된 숫자에 적용한 결과를, Ⅴ장에서는 반복적 분석 방법과 규칙 선택 비트 사용에 의한 성능 평가를 통해 제안한 방법의 타당성을 확인하고, 마지막으로 Ⅵ장에서 결론을 맺는다.

Ⅱ. 퍼지 분류기

표준 패턴 P와 입력 패턴 I와의 유사도가 크면 입력 패턴이 표준 패턴의 해당 클래스에 소속될 가능성이 크고 유사도가 작으면 소속될 가능성이 작다[3]는 아주 간단한 개념을 분류 알고리즘의 기본으로 한다. 입력 변수 x에 대한 퍼지 소속 정도를 나타내는 퍼지 멤버쉽 함수 $\mu(x)$를 식 (1)과 같이 정의한다. 그 형태는 Fig. 1과 같다.

$$\mu(x)=\begin{cases}\dfrac{x-p_M}{p_M-p_L}+1, & where\ p_L\le x\le p_M\\ \dfrac{x-p_M}{p_M-p_R}+1, & where\ p_M< x\le p_R\\ 0, & o/w\end{cases} \quad (1)$$

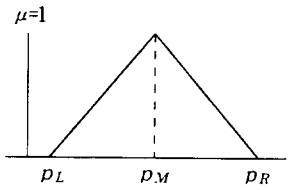

Fig. 1 Fuzzy membership function

p_L, p_M, p_R은 퍼지 멤버쉽 함수의 파라메터이며 각각 하한, 중앙, 상한을 의미한다. 표준 패턴 $P(i, j)$는 식 (2)와 같이 개개 클래스의 특징들의 평균으로 정의한다. 클래스에 소속되어 있는 패턴의 개수는 서로 다를 수 있으나, 특징 벡터는 항상 같은 차원을 갖는다. $I_k(i, j)$는 소속 클래스가 i이고, j번째 특징을 갖는 k번째 입력 패턴을 의미한다.

$$P(i,j) = \sum_{k=1}^{N_I} I_k(i,j)/N_I \tag{2}$$

$i = 1, 2, \cdots, N_C, \quad j = 1, 2, \cdots, NF.$

N_C: Number of class, N_F: Number of feature,

N_I: Number of input patterns in each class.

각 클래스에 대한 특징의 퍼지 패턴 매칭(fuzzy pattern matching)[1]을 구하기 위하여 식 (3)을 이용한다. 이는 표준 패턴과 입력 패턴의 유사도를 의미한다. FPM_{ij}는 퍼지 패턴 매칭 $\boldsymbol{FPM}$의 원소로서 클래스 i, 특징 j에 대한 퍼지 패턴 매칭 값이다.

$$\begin{aligned} \boldsymbol{FPM} &= \{ FPM_{ij} | 1 \le i \le N_c, 1 \le j \le N_F \} \\ &= Min_{k=1}^{N_I}(1 - |\mu_P(x) - \mu_{I_k}(x)|) \end{aligned} \tag{3}$$

$\mu_P(x)$는 표준 패턴의 퍼지 소속 정도이며, $\mu_{I_k}(x)$는 입력 패턴의 퍼지 소속 정도이다. 표준 패턴과 입력 패턴과의 유사도가 크면 클수록 퍼지 패턴 매칭 값은 커지게 되며 입력 패턴 I가 표준 패턴 P에 소속될 가능성은 커진다. 입력 패턴의 표준 패턴에 대한 소속 클래스를 구하기 위하여 입력 패턴의 매칭 정도(Degree of Matching) M_i를 식 (4)와 같이 정의한다. s_R은 유전자 알고리즘에서 각 특징들이 규칙에 포함되는지 포함되지 않는지의 여부를 나타내는 규칙 선택 비트이며, N은 규칙에 포함된 특징의 개수로서 "1"인 s_R의 개수와 같다.

$$M_i = \sum_{\substack{j=1 \\ \text{if } s_R=1}}^{N_F} \frac{FPM_{ij}}{N} \tag{4}$$

입력 패턴의 표준 패턴에 대한 소속 정도로부터 다음의 분류 규칙에 의하여 입력 패턴의 소속 클래스를 구할 수 있다.

$$If\ Max_{i=1}^{N_C} M_i\ is\ M_c,\ \ then\ \ Class(I) = c \tag{5}$$

입력 패턴이 소속되는 클래스의 번호를 나타내는 c는 i중의 하나이다. 또한, M_i는 입력 패턴이 i번째 클래스에 소속되는 정도 즉 i번째 클래스와의 매칭 정도이고, M_c는 클래스 c와의 매칭 정도이며, $Class(I)$는 입력 패턴이 속해있는 $Class$의 번호이다. 식 (5)에 의하여 최대의 유사도를 가지는 클래스를 입력 패턴의 소속 클래스로 판단한다. 본 논문의 인식 결과는 식 (5)를 만족하는 하나의 문자열을 이용하여 얻어진 것이다.

Ⅲ. 유전자 알고리즘을 이용한 퍼지 규칙 생성

유전자 알고리즘은 적자 생존의 자연 선택과 자연계의 유전학에 근거한 탐색 알고리즘이다.[5] 유전자 알고리즘은 미시간 대학의 John Holland와 그의 동료들에 의하여 소개되었으며, 최근 들어 다양한 응용분야에서 강인한 확률적 탐색 방법(Robust Stochastic Search Method)으로서 많은 주목을 받고 있다.[6][7]

1. 문자열의 표현(String Representation)

유전자 알고리즘을 이용하여 어떤 최적화 문제를 해결하기 위하여 문제에서 사용되는 변수들을 하나의 문자열로 코딩하여야 할 필요가 있다. 서론에서 언급했듯이 유전자 알고리즘을 이용하여 최적의 분류 규칙을 위한 퍼지 멤버쉽 함수의 형태, 개수, 그리고 각 특징이 분류 규칙에 포함되는지, 포함되지 않는지의 여부를 결정한다. 따라서 각 문자열의 구조는 각 특징들에 대한 멤버쉽 함수의 형태 및 개수를 결정하는 부분과 각 특징들이 분류 규칙에 포함되는지, 포함되지 않는지의 여부를 결정하는 부분으로 구성된다. 멤버쉽 함수의 형태 및 개수를 결정하는 부분을 멤버쉽

함수 결정 부분이라 하여 m +2 비트를 할당하고, 분류 규칙에 포함되는지의 여부를 결정하는 부분을 분류 규칙 결정 부분이라 하고 1 비트를 할당한다. m = NM − 1이며, NM은 문자열의 크기이다.

멤버쉽 함수 결정 부분의 양쪽 맨 끝 비트는 항상 "1"로 고정되며 나머지 부분은 "0" 혹은 "1"을 갖는다. 따라서 전체 문자열의 길이는 m +3 비트이나 실제 사용되는 문자열은 Fig. 2의 (b)와 같이 m +1 비트가 된다. 멤버쉽 함수 결정 부분에서 "1"은 퍼지 멤버쉽 함수의 파라메터가 됨을 의미하고, 규칙 선택 부분의 "1"은 특징이 규칙에 포함됨을 "0"은 규칙에 포함되지 않음을 의미한다. 본 논문에서 사용한 문자열과 한 예를 Fig. 2의 (c)에 보인다.

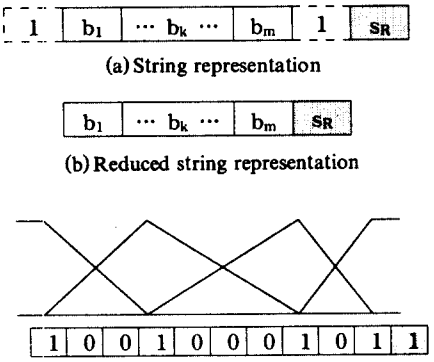

Fig. 2 String representation and its example

b_k(k = 1, 2, ..., m)는 각 특징 벡터의 한 원소에 대한 퍼지 멤버쉽 함수의 파라메터가 될 수 있는 값이며 0을 제외한 연속된 3개의 1은 각각 하한, 중앙치, 상한을 결정하며, 마지막 비트는 특징이 분류 규칙에 포함되는지 포함되지 않는지의 여부를 나타내는 규칙 선택 비트(Rule Selection bit)이다.

논문에서 제안한 문자열을 $S = b_1 b_2 \cdots b_k \cdots b_m s_R$, 규칙 선택 비트가 제외된 문자열을 $S' = b_1 b_2 \cdots b_k \cdots b_m$이라고 하자. 유전자 알고리즘을 사용하여 최적의 S'를 생성할 수 있다는 것은 널리 알려져 있다.[5][13][14] 본 논문에서 문자열의 정의에 의해 최적의 문자열을 생성할 수 있음은 곧 최적의 퍼지 멤버쉽 함수를 생성할 수 있음을 의미하며, 곧 전체적인 인식률이 증가할 수 있다는 의미한다. 그리고, S에 의하여 생성되는 퍼지 멤버쉽 함수들의 집합 F_S는 S'에 의해 생성될 수 있는 퍼지 멤버쉽 함수의 집합 $F_{S'}$를 포함한다. 따라서 문자열 S에 의하여 생성되는 퍼지 멤버쉽 함수에 의한 인식률은 S'에 의해 생성되는 퍼지 멤버쉽 함수에 의한 인식률 보다 최소한 같다는 것을 보일 수 있다.

인식률이라는 면에서 보면, 어떤 특징이 임의의 패턴 인식기의 입력으로 추가되었을 때 인식률은 증가, 감소, 혹은 변화를 보이지 않는 세가지 경우가 존재하며 이러한 경우 각각 그 특징을 우세 특징 변수, 열등 특징 변수, 잉여 특징 변수라고 하자. 제안한 문자열에서 규칙 선택 비트는 멤버쉽 함수 결정 부분과 그 역할이 독립적이다. 따라서 규칙 선택 비트는 멤버쉽 함수 결정 부분과 마찬가지로 최적화될 수 있으며, 이는 우세 특징 변수를 추출하고 열등 특징 변수와 잉여 특징 변수를 배제할 수 있음을 의미한다. 곧 인식률이 증가 될 수 있음을 의미한다. 따라서 논문에서 제안한 특징 선택 비트를 사용함으로써 인식률의 증가를 기대할 수 있다.

2. 적합도 함수(Fitness Function)

모든 적용 대상에 대해 각 개체에 대한 입력 패턴의 인식률 ***RR***(*Recognition Rate*)을 적합도 값(*Fitness value*)으로 정의한다. 실험적으로 적합도 값(***Fitness*** = ***RR***)이 높은 상위 30%의 문자열을 우수 문자열로 하여 우수 문자열과 나머지 문자열과의 교배 연산, 변이 연산을 수행하며 최고의 인식률을 보이는 문자열은 다음 세대에도 살아남게 된다. 따라서 모든 문자열은 인식률이 높은 방향으로 진화한다.

3. 유전 연산자(Genetic Operator)

유전 연산자는 현재 세대의 탐색 노드에 근거해 새로운 탐색 노드를 발생시킨다. 새로운 탐색 노드는 이전의 탐색 노드의 조합 혹은 재배치에 의하여 구성된다.

가. 재생산(Reproduction)

높은 인식률을 보이는 문자열은 다음 세대에 재생산될 가능성이 높아야 한다. 적합도 값의 크기에 따라 아래의 연산자들을 위해 선택될 가능성은 높아진

다. 현재 세대의 우수 문자열과 나머지 문자열을 이용하여 아래의 방법으로 교배 연산을 실행한다. 따라서 재생산은 단순히 인식률이 높은 상위 30%의 문자열들을 선택하는 아주 간단한 과정이다. 모든 실험에서 population의 크기를 20, 우수 문자열의 개수를 6으로 하였다.

나. 교배 연산(Crossover Operation)

재생산 과정에서 얻어진 우수 문자열과 나머지 문자열과의 조합에 의하여 새로운 문자열을 생성한다. 재생산 과정에서 선택된 상위 30%의 인식률을 보이는 우수 문자열들과 나머지 문자열들을 선택한다. 선택된 문자열의 각 쌍에 대하여, 랜덤하게 선택된 문자열의 두 비트를 포함하여 두 비트를 경계로 세 영역 중 임의로 선택된 한 영역 내부의 문자열을 교환한다. 교배 연산을 위한 확률은 0.5를 사용하였다.

다. 돌연변이 연산(Mutation Operation)

재생산과 교배 연산에 의해 발생되는 새로운 문자열은 전체 탐색 공간을 효과적으로 탐색할 수 있다. 하지만 특별한 위치에서 과포화 상태에 빠지는 경우가 보고되고 있다.[7] 인공 유전 시스템에서 돌연변이 연산자는 이러한 과포화 상태를 예방해 줄 수 있으며 교배 연산자와 더불어 대표적인 연산자 중의 하나이다. 문자열의 모든 비트에 대하여 조건을 만족하면 선택된 비트의 값을 "0"에서 "1"로 혹은 "1"에서 "0"으로 변화시킨다. 변이 연산을 위한 확률은 0.1을 사용하였다.

Ⅳ. 실험 결과

본 논문에서 제안한 퍼지 분류기의 효용성을 검증하기 위하여 150개의 아이리스 데이터, 57개의 갑상선 종양 세포 데이터[8][9], 700개의 필기 숫자와 300개의 인쇄 숫자로 이루어진 1,000개의 숫자 데이터에 대한 인식 결과를 보인다.

1. 아이리스 데이터[10][11]

아이리스 데이터는 패턴 분류 문제의 표준 자료로서 4개의 특징과 3개의 클래스로, 각 클래스에 대해 50개의 패턴들로 이루어져 있다. 분류기의 성능 측정을 위한 150개의 아이리스 데이터에 대한 인식률을 Table 1의 (a)에 보인다. 규칙 선택 비트를 제외한 문자열의 크기를 1에서 7까지 확장시키면서 30회의 반복 연산과 10회의 실험을 통해 얻어진 최대 인식 결과이다. 하나의 분류 규칙으로 문자열의 크기가 4일 때 148개의 패턴을 분류할 수 있었으며 인식률은 98.67%이다.

Table 1. (a) Simulation results with different string size

String size	With RS bit		Without RS bit	
	No. of patterns	No. of rules	No. of patterns	No. of rules
1	144	1	140	1
2	145	1	144	1
3	147	1	146	1
4	148	1	146	1
5	147	1	146	1
6	146	1	145	1
7	146	1	146	1

(b) Simulation results with different crossover operations[10]

Trial number	Uniform crossover		One-point crossover	
	No. of patterns	No. of rules	No. of patterns	No. of rules
1	149	5	149	5
2	149	6	149	5
3	148	6	150	7
4	149	7	149	6
5	149	7	149	6
Average	148.8	6.2	149.2	5.8

RS bit : Rule Selection bit
No. of patterns : the number of correctly classified patterns.
No. of rules : the number of selected rules.

Table 1의 (b)에 여러 개의 분류 규칙을 사용한 Ishibuchi[10]의 결과를 보인다. Ishibuchi는 5개의 분류 규칙으로 149개, 7개의 분류 규칙으로 150개 전부의 패턴을 분류하였다. 또한 Young[11]은 1개의 분류 규칙으로 146개의 패턴을 분류하였다. Young의 결과는 규칙 선택 비트가 없을 경우의 최대 인식률과 같다. 이는 규칙의 개수에 대한 인식률의 측면에서 보면 제안된 방법이 우수한 인식 결과를 보임을 알 수 있다. 또한 Ishibuchi는 문자열의 크기로 441을 사용하고, population의 크기로 50을 사용하였지만, 본 논문에서는 각각 1~7과 20을 사용하였다.

Table 1의 (a)에서 알 수 있듯이 문자열의 크기와 최대 인식률과는 밀접한 관계를 갖는다. 문자열의 크기가 증가할수록 최대 인식률은 점차 증가하다가 감소하는 경향을 보인다. 이는 문자열 크기의 증가가 미세한 퍼지 멤버쉽 함수의 생성을 제공할 수 있지만 탐색 노드의 기하급수적인 증가로 인해 짧은 시간 동안 충분한 탐색을 보장할 수 없음을 의미한다.

또한 문자열이 구성하는 퍼지 멤버쉽 함수의 기하학적인 구조상 문자열의 길이가 7인 경우는 3인 경우를 모두 포함하므로, 문자열의 길이가 7인 경우의 인식률은 3인 경우보다 최소한 같거나 커야 한다. 문자열의 크기가 커질 경우 효과적인 탐색을 위해서는 population의 크기도 커져야 하며, 반복 회수도 커져야 하는 단점이 발생한다. 이에 대해서는 V장에서 논한다.

2. 갑상선의 종양 세포 식별[8][9]

갑상선의 종양 세포에는 여포성 종양 세포와 유두상 종양 세포의 두 가지 종양 세포가 존재한다. 전문가에 의하여 구분되는 정상 세포와 종양 세포의 구분은 다분히 주관적이며, 복합적인 판단 기준에 의한다. 실험에서 사용한 데이터의 각 세포는 16개의 특징으로 구성되어 있으며 16개의 정상 세포, 25개의 여포성 종양 세포, 16개의 유두상 종양 세포 등 3개의 클래스로 구성된 57개의 갑상선 세포 데이터에 대한 분류율을 Table 2~Table 8에 보인다. 종양 세포의 특징을 4개에서 16개까지 증가시키면서 규칙 선택 비트(Rule Selection Bit)를 사용했을 경우와 사용하지 않았을 경우를 비교하였다. 괄호 안의 숫자는 오인식된 패턴의 개수이다.

규칙 선택 비트를 포함한 실험에서 특징의 개수가 10개이고 규칙 선택 비트를 제외한 문자열의 크기가 3일 때와 특징의 개수가 16개이고 문자열의 크기가 3개일 때 최대의 인식률을 보였으며, 이때의 오인식된 개수는 1개이다. 또한 규칙 선택 비트를 포함하지 않은 경우는 특징의 개수가 12개이고 문자열의 크기가 3일 때, 특징의 개수가 16이고 문자열의 크기가 2, 3이었을 경우 92.98%의 인식률을 보였으며, 이때의 오인식된 데이터의 개수는 4개이다. 본 실험에서도 아이리스 데이터에서와 마찬가지로 문자열 크기의 증가에 따른 인식률의 저하가 나타났다. Table 8에서 알 수 있듯이 문자열의 크기가 2이었을 때의 최대 인식률은 96.49%이었으며 5일 때의 최대 인식률은 94.74%이었다.

참고문헌 [8]의 최대 인식률은 88.75%이고, [9]의 최대 인식률은 90.0%이며, 이는 규칙 선택 비트가 없는 경우의 인식률인 92.98% 보다도 더 낮은 인식률을 보였다.

Table 2. String size and recognition rates with/without RS bit, Feature size : 4

string size	with RS bit	without RS bit
1	75.44% (14)	73.68% (15)
2	75.44% (14)	73.68% (15)
3	78.95% (12)	73.68% (15)
4	77.19% (13)	77.19% (13)
5	74.44% (14)	75.44% (14)
6	77.19% (13)	77.19% (13)
7	75.44% (14)	75.44% (14)

Table 3. String size and recognition rates with/without RS bit, Feature size : 6

string size	with RS bit	without RS bit
1	84.21% (12)	78.95% (12)
2	87.72% (7)	78.95% (12)
3	85.96% (8)	78.95% (12)
4	84.21% (9)	80.70% (11)
5	85.96% (8)	78.95% (12)
6	87.72% (7)	78.95% (12)
7	85.96% (8)	80.70% (11)

Table 4. String size and recognition rates with/without RS bit, Feature size : 8

string size	with RS bit	without RS bit
1	87.72% (7)	80.71% (11)
2	94.74% (3)	85.96% (8)
3	92.98% (4)	85.96% (8)
4	91.23% (5)	84.21% (9)
5	92.98% (4)	84.21% (9)
6	91.23% (5)	85.96% (8)
7	87.72% (7)	82.46% (10)

Table 5. String size and recognition rates with/without RS bit, Feature size : 10

string size	with RS bit	without RS bit
1	89.47% (6)	85.96% (8)
2	92.98% (4)	89.47% (6)
3	98.25% (1)	89.47% (6)
4	94.74% (3)	85.96% (8)
5	91.23% (5)	84.21% (9)
6	91.23% (5)	87.72% (7)
7	91.23% (5)	84.21% (9)

Table 6. String size and recognition rates with/without RS bit, Feature size : 12

string size	with RS bit	without RS bit
1	91.23% (5)	87.72% (7)
2	94.74% (3)	91.23% (5)
3	94.74% (3)	92.98% (4)
4	94.74% (3)	89.47% (6)
5	92.98% (4)	89.47% (6)
6	91.23% (5)	89.47% (6)
7	91.23% (5)	87.72% (7)

Table 7. String size and recognition rates with/without RS bit, Feature size : 14

string size	with RS bit	without RS bit
1	91.23% (5)	89.47% (6)
2	94.74% (3)	91.23% (5)
3	96.49% (2)	91.23% (5)
4	94.74% (3)	91.23% (5)
5	92.98% (4)	87.72% (7)
6	92.98% (4)	89.47% (6)
7	92.98% (4)	87.72% (7)

Table 8. String size and recognition rates with/without RS bit, Feature size : 16

string size	with RS bit	without RS bit
1	92.98% (4)	91.23% (5)
2	96.49% (2)	92.98% (4)
3	98.25% (1)	92.98% (4)
4	94.74% (3)	91.23% (5)
5	94.74% (3)	91.23% (5)
6	94.74% (3)	91.23% (5)
7	92.98% (4)	89.47% (6)

3. 필기체 및 인쇄체 숫자 인식[12]

문자 인식은 패턴 인식의 한 분야로서 시각 정보를 통하여 문자를 인식하고 나아가 그 의미를 이해하는 사람의 능력을 컴퓨터로 실현하려는 시도로서 1970년대 이후 활발한 연구가 수행되어 왔다. 본 논문의 필기 및 인쇄된 숫자 인식은 구조적인 정보가 동일한 부류를 동일 그룹으로 분류하여 각 그룹에 대하여 최적의 특징 변수를 획득하고 최대의 인식률을 얻을 수 있도록 한다. 본 논문에서 사용한 구조적인 정보는 *SEP* (*Starting or Ending Point*)와 *Loop*이다. 본 실험에서는 제안된 퍼지 분류기를 이용하여 700개의 필기된 숫자와 300개의 인쇄된 숫자의 인식을 다룬다. 먼저 숫자 영상의 구조적인 정보를 얻어내기 위하여 숫자 영상의 골격선을 구한다. 본 실험에서 정확한 골격선을 구하기 위하여 윤곽선 검출 과정에서의 세선화 방안을 이용하였다. 정확한 골격선을 검출하기 위하여 2진 영상에 대한 윤곽선으로 검출된 픽셀에 대해 골격선의 여부를 판단하여 윤곽점들을 줄여 나가는 방법을 사용하였다.

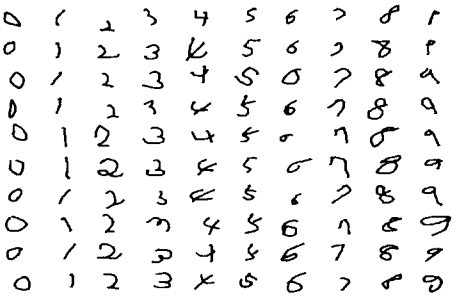

Fig. 3 Handwritten numerals

문자를 구성하는 기본 요소 중의 하나인 *Loop*의 개수를 결정하는 방법은 다음과 같다. 라벨링 과정에서 숫자를 이루는 영역의 개수 $N_{Numeral}$과 배경 영역의 개수 $N_{Background}$를 이용하여 Loop의 개수 N_{Loop}를 구한다.

$$N_{Loop}=N_{Background}-N_{Numeral} \tag{6}$$

$N_{Loop}=0$: 루프가 존재하지 않음
$N_{Loop}>0$: 루프가 존재
$N_{Loop}<0$: 문자 영역이 서로 분리되어 있음

1,000개의 숫자 영상에 대하여 문자의 구조적인 최소 정보인 *SEP*의 개수와 *Loop*의 개수를 이용하여 Table 9와 같이 소 그룹으로 분류하였다.[12] 문자의 구조적인 최소 정보인 *SEP*와 *Loop*의 존재 유무는 숫자 인식에 있어 중요한 역할을 한다. Group5와 Group8은 구조적인 정보가 곧 숫자를 인식하는 규칙이 된다. Group5는 루프의 개수가 1이고 시작점이나 끝점이 골격선을 통해 구해지지 않는 경우이며 10개의 숫자 데이터 중 0만이 속해 있으며, Group8은 루프의 개수가 2이며 시작점이나 끝점이 존재하지 않는 경우로서 숫자 8임을 의미한다. 본 실험에서 사용한 필기 숫자의 예를 Fig. 3에 보인다.

실험에서 사용한 소 그룹은 각 그룹이 갖는 숫자 집합의 개수에 따라, 소수의 숫자 집합이 소속되어 있는 그룹들(Group7, Group8, Group10, Group11)을

Table 9. Group

Group	SEP	Loop	Numerals	Number
Group1	4	-1	5,7	29
Group2	2	0	0,1,2,3,4,5,6,7,8,9	337
Group3	3	0	1,2,3,4,5,7,9	155
Group4	4	0	4,5	60
Group5	0	1	0	90
Group6	1	1	0,4,6,8,9	182
Group7	2	1	2,4,6,8,9	41
Group8	3	1	4	13
Group9	0	2	8	88
Group10	1	2	8	4
Group11	2	2	8	1

Table 10. Recognition result

Group	Recognition Rate	Number of Misclassification
Group1	100%	0
Group2	81.30%	63
Group3	80.65%	30
Group4	100%	0
Group5	100%	0
Group6	86.26%	25
Group7	100%	0
Group8	100%	0
Total	88.2%	118

Table 11. (a) R.R.(Recognition Rate): 90.8% (Feature size: 5/axis)

	Number	R.R.	Misclassification Number
Group21	145	90.35%	14
Group22	192	88.02%	23
Group3	155	80.65%	30
Group6	182	86.26%	25

(b) R.R.(Recognition Rate): 93.5% (Feature size: 6/axis)

	Number	R.R.	Misclassification Number
Group21	145	93.84%	9
Group22	192	91.15%	17
Group3	155	85.81%	22
Group6	182	90.71%	17

(c) R.R.(Recognition Rate): 96.3% (Feature size: 7/axis)

	Number	R.R.	Misclassification Number
Group21	145	95.17%	5
Group22	192	94.27%	11
Group3	155	92.90%	11
Group6	182	94.50%	10

(d) R.R.(Recognition Rate): 94.2% (Feature size: 8/axis)

	Number	R.R.	Misclassification Number
Group21	145	94.48%	8
Group22	192	90.62%	18
Group3	155	89.03%	17
Group6	182	91.76%	15

하나의 그룹(Group7)으로 통합하였다. 소 그룹으로 분류한 후의 인식률을 Table 9에 보인다. 본 실험에서 각 숫자 영상을 구성하는 픽셀들의 x, y축으로의 히스토그램을 분류기의 특징으로 사용하였다. Table 10을 통해 알 수 있듯이 Group2, 3, 6을 제외한 그룹은 모두 분류할 수 있었으며 분포 비율이 비교적 크고 인식률이 낮은 Group2를 SEP의 x좌표의 상대 위치에 따라 Group21과 Group22로 다시 분류하였다. 최적의 특징의 개수를 구하기 위하여 특징의 개수를 늘려 가면서 인식률의 변화를 알아보았다. 각 축을 5개에서 8개까지 등간격으로 분할하여 각 축에 대한 히스토그램을 각 패턴의 특징으로 사용하였다. 각 축을 7개의 등간격으로 영역 분할한 히스토그램을 특징으로 사용하였을 경우 최대의 인식률을 얻을 수 있었다. 최대 인식률은 96.3%이다.

V. 반복적 분석 방법의 성능 고찰

문자열의 크기는 퍼지 멤버쉽 함수의 형태를 결정하는 중요한 요인이 되며, 이는 분류기의 인식 성능과도 직접적인 관련을 갖는다. 전체 탐색 공간을 충분히 탐색할 수만 있다면, 일반적으로 문자열의 크기가 커지면 그 인식률 또한 증가할 수 있다. 문자열의 크기가 큰 상태에서 최적의 노드를 탐색하는 것보다는 최적화 된 작은 문자열을 바탕으로 문자열을 증가시켜 나가면서 최적의 노드를 탐색해 가는 방법을 사용한다.

Fig. 4에 문자열의 길이가 1, 3, 7인 경우의 대표적인 퍼지 멤버쉽 함수의 예를 든다. 문자열의 길이가 3인 경우는 1인 경우의 가능한 모든 멤버쉽 함수의 형태를 포함하며, 마찬가지로 문자열의 길이가 7인 경우는 1과 3인 모든 경우를 포함하게 된다. 이러한 이유로 문자열의 길이가 3인 경우는 1인 경우보다 그 인식률이 최소한 같거나 커야 하며, 7인 경우는 1과 3

인 경우보다 최소한 같거나 커야 한다. 하지만 Ⅳ장의 대부분의 결과와 같이 실제의 경우에는 반드시 이러한 조건을 만족하지는 않는다. 그 이유는 문자열의 길이가 증가함으로써 탐색하여야 할 공간이 증가하게 되어 충분한 탐색이 보장되지 않을 수도 있기 때문이다. 본 논문에서는 규칙 선택 비트를 사용하고 또, 반복적인 분석 방법을 사용함으로써 population의 크기를 20으로, 문자열의 크기는 1에서 7로 충분히 작게 할 수 있었다. 이는 최고의 인식률을 보이는 최소의 문자열의 크기를 유지함으로써 최적의 퍼지 멤버쉽 함수를 구성하는 효율적인 분류기를 설계하기 위함이다.

문자열의 증가는 Fig. 4와 같이 사선으로 분리되어 있는 각 구간을 이등분하는 방법을 사용한다. 일정 반복 횟수만큼 인식률의 개선이 없으면 각 구간은 정확히 절반으로 쪼개지고, 따라서 초기 문자열의 길이가 1, 확장된 문자열의 길이가 3, 7, 15, ...가 된다. 초기 문자열의 크기에 따라 다음 문자열의 크기가 결정된다. 여기에서 일정한 횟수란 현재 문자열의 크기에 따라 다르며 실험에서는 현재 문자열의 크기 $L_{String}(n)$의 10배를 사용하였다. 주어진 조건을 만족하면 다음 문자열의 크기 $L_{String}(n+1)$은 현재 문자열의 크기 $L_{String}(n)$의 2배에 1을 더한 값을 갖는다.

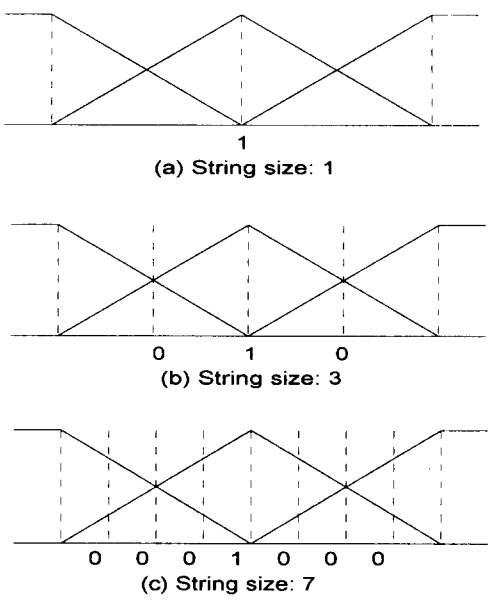

Fig. 4 Fuzzy membership functions for each string size

$$L_{String}(n+1) = 2L_{String}(n) + 1 \quad (7)$$

초기 문자열의 크기 L_{String}은 1, 2, 4, 6, 8, ...의 값을 갖는다. 제안된 방법을 사용함으로써 Ⅳ장의 결과와 같은 문자열의 증가(Table 12의 문자열의 크기가 각각 "1, 3, 7", "2, 5", "4, 9"일 때)로 인한 인식률의 저하는 발생하지 않으며, 최소의 문자열의 길이를 보장할 수 있게 된다. 제안된 방법에 의한 아이리스 데이터와 갑상선 종양세포 인식의 개선 결과를 Table 12, Table 13에 보인다.

Table 12. Recognition rates with/without RS bit using repetitive analyses method for Iris data.

string size	with RS bit	without RS bit
1	96.00% (6)	93.33% (10)
2	96.67% (5)	96.00% (6)
3	98.00% (3)	97.33% (4)
4	98.67% (2)	97.33% (4)
5	98.00% (3)	97.33% (4)
6	97.33% (4)	96.67% (5)
7	98.00% (3)	97.33% (4)
8	98.00% (3)	97.33% (4)
9	98.67% (2)	97.33% (4)

Table 13. Recognition rates with/without RS bit using repetitive analyses method for Gland Cancer cells data (Feature size : 10).

string size	with RS bit	without RS bit
1	89.47% (6)	85.96% (8)
2	92.98% (4)	89.47% (6)
3	98.25% (1)	89.47% (6)
4	94.74% (3)	85.94% (8)
5	94.74% (4)	89.47% (6)
6	91.23% (5)	87.72% (7)
7	98.25% (1)	89.47% (6)
8	94.74% (4)	87.72% (7)
9	94.74% (3)	85.94% (8)

또한 규칙 선택 비트를 사용함으로써 아이리스 데이터에 대해서는 평균 1.06%, 갑상선 종양세포 데이터에 대해서는 6.44%의 인식률 증가를 얻을 수 있었다. 규칙 선택 비트의 사용은 아이리스 데이터와 같이 특징의 개수가 작은 경우보다는 특징의 개수가 큰

종양 세포의 경우에 더욱 효과적임을 알 수 있다.

Ⅵ. 결 론

본 논문에서는 유전자 알고리즘을 이용하여 다양한 특징의 분석과 선택적인 적용이 필요한 대상에 대한 퍼지 분류기를 제안하였다. 제안된 퍼지 분류기는 퍼지 논리를 이용한 분류 부분과 유전자 알고리즘을 이용한 간단한 규칙 생성 부분으로 구성된다. 퍼지 논리를 이용한 분류 부분은 표준 패턴과 입력 패턴과의 유사도를 측정하고, 입력 패턴과 가장 큰 유사도를 갖는 클래스가 입력 패턴이 소속되어 있는 클래스라는 비교적 간단한 개념에서 출발하였다. 유전자 알고리즘을 이용한 규칙 생성 부분에서는 최적의 퍼지 멤버쉽 함수를 결정하고, 각 특징이 규칙에 포함되는지 포함되지 않는지의 여부도 결정하며, 인식률의 개선을 위한 반복적 규칙 생성의 방법도 제안하였다. 실험 결과를 통하여 특징 선택 비트의 사용과 반복적인 분석 방법이 많은 개수의 특징 변수에 대한 최적의 퍼지 멤버쉽 함수를 생성하는 데 매우 효과적임을 보일 수 있었다. 본 논문에서 제안한 방법은 뚜렷한 특징을 갖지 못하는 대상의 분류에 일반적으로 적용되리라 예상한다.

참 고 문 헌

1. R. J. Marks II, *Fuzzy Logic Technology and Applications*, Technical Activities Boards, 1994.
2. G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty, And Information*, Prentice-Hall, 1992.
3. J. C. Bezdek and S. K. Pal, *Fuzzy Models for Pattern Recognition*, IEEE Press, 1992.
4. M. Garbisch and M. Sugeno, "Multi-Attribute Classification Using Fuzzy Integral", FUZZ-IEEE '92, pp.47~54, 1992.
5. D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison Wesley, 1989.
6. A. A. Hopgood, *Knowledge-Based Systems for Engineers and Scientists*, CRC Press, 1993.
7. B. *Souček* and the IRIS Group, *Dynamic, Genetic, and Chaotic Programming*, Wiley Inter. Science, 1992.
8. 신대정, 나승유, 나철훈, "퍼지 논리와 유전 알고리즘을 이용한 종양 세포 인식에 관한 연구", 1995년도 대한전자공학회 하계종합학술대회 논문집 제18권 제 1호, pp.615~619, 1995.
9. 나철훈, 김창원, 김현재, "종양 세포 식별을 위한 공간 주파수 영역에서의 화상 해석", 한국통신학회 논문지, vol. 18, no. 3, pp.385~396, 1993.
10. Ishibuchi, K. Nozaki, N. Yamamoto and H. Tanaka, "Construction of fuzzy classification systems with rectangular fuzzy rules using genetic algorithms", Fuzzy Sets and Systems, vol. 65, pp.237~253, 1994.
11. Y. S. Young and S. Mitra, "An adaptive integrated fuzzy clustering model for pattern recognition", Fuzzy Sets and Systems, vol. 65, pp.297~310, 1994.
12. D. J. Shin, S. Y. Na and S. H. Kim, "A Fuzzy Genetic Classifier for Recognition of Confusing Handwritten Numerals 4, 6, and 9", Proc. of KFIS, pp.11~14, 1995.
13. M. Yamamura, H. Satoh, and S. Kobayashi, "An Analysis of Crossover's Effect in Genetic Algorithms", Proceedings of The 1st IEEE Conference on Evolutionary Computation, pp.613~618, 1994.
14. J. Stender, Parallel Genetic Algorithms: Theory and Applications, IOS Press, 1993.

신 대 정(Dae Jung Shin) 준회원
1972년 11월 23일생
1994년 : 전남대학교 전자공학과 졸업(공학사)
1996년 : 전남대학교 대학원 전자공학과 졸업(공학석사)
1996년~현재 : 전남대학교 대학원 박사과정
※주관심분야 : 제어시스템 설계, 지능시스템

나 승 유(Seung You Na) 종신회원

1954년 5월 1일생

1977년 : 서울대 전자공학과 졸업 (공학사)

1986년 : 미국 University of Iowa 전기 및 컴퓨터공학과 졸업(공학석사, 박사)

1987년~현재 : 전남대 전자공학과 부교수

※주관심분야 : 제어시스템 설계, 지능제어, 신호처리