한국지능시스템학회 논문지 2008, Vol. 18, No. 3, pp. 360-366

퍼지 규칙기반 분류시스템에서 퍼지 분할의 선택방법

Selection Method of Fuzzy Partitions in Fuzzy Rule-Based Classification Systems

손창식[b] · 정환묵[a] · 권순학[b*]

Chang S. Son, Hwan M. Chung, Soon H. Kwon

[a] 대구가톨릭대학교 컴퓨터정보통신공학부

[b] 영남대학교 전기공학과

요 약

퍼지 규칙기반 분류 시스템에서 초기의 퍼지 분할은 주어진 데이터가 가진 속성들의 도메인을 고려함으로서 결정되어지고, 최적의 분류 경계면은 초기에 정의된 퍼지 분할의 파라미터들을 조정함으로서 찾을 수 있다. 본 논문에서는 학습과정들을 사용하지 않고 패턴분류의 성능을 최대화하기 위해 통계적 정보에 기반을 둔 퍼지 분할의 선택방법을 제안한다. 제안된 방법에서 통계적 정보는 주어진 수치적인 데이터로부터 각 입력 속성의 '불확실성 영역', 즉 패턴분류문제에서 분류 경계면이 결정되는 영역을 추출하기 위해 사용되었다. 또한 통계적인 정보에 의해서 생성된 퍼지 분할구간에 대응하는 후보 규칙들을 추출하기 위한 방법과 그 후보 규칙들 간의 커플링 문제를 최소화하기 위한 방법도 추가적으로 논의하였다. 실험에서는 제안된 방법의 효용성을 보이기 위해 IRIS와 New Thyroid Cancer 데이터를 사용한 기존 패턴분류 방법들과의 분류 정확성을 비교하였고, 그 결과들로부터 제안된 방법이 기존의 방법들보다 더 좋은 분류 정확성을 제공함을 확인할 수 있었다.

키워드 : 패턴분류, 퍼지 분할, 통계적 정보, 커플링 문제

Abstract

The initial fuzzy partitions in fuzzy rule-based classification systems are determined by considering the domain region of each attribute with the given data, and the optimal classification boundaries within the fuzzy partitions can be discovered by tuning their parameters using various learning processes such as neural network, genetic algorithm, and so on. In this paper, we propose a selection method for fuzzy partition based on statistical information to maximize the performance of pattern classification without learning processes where statistical information is used to extract the uncertainty regions (i.e., the regions which the classification boundaries in pattern classification problems are determined) in each input attribute from the numerical data. Moreover the methods for extracting the candidate rules which are associated with the partition intervals generated by statistical information and for minimizing the coupling problem between the candidate rules are additionally discussed. In order to show the effectiveness of the proposed method, we compared the classification accuracy of the proposed with those of conventional methods on the IRIS and New Thyroid Cancer data. From experimental results, we can confirm the fact that the proposed method only considering statistical information of the numerical patterns provides equal to or better classification accuracy than that of the conventional methods.

Key Words : pattern classification, fuzzy partition, statistical information, coupling problem

1. 서 론

퍼지 규칙을 기반으로 한 분류 시스템들은 제어, 의사결정 등과 같은 다양한 응용분야에서 성공적으로 적용되어져 왔고, 최근에는 패턴분류를 위한 문제들에 적용되어지고 있다. 일반적으로 퍼지 규칙기반 분류시스템의 분류 성능은 주어진 수치적인 입력패턴들이 대응하는 분할구간들과 규칙들의 수에 의해 결정된다. 그러나 이들을 결정하기란 쉽지가 않으며, 대부분 도메인 전문가들에 의해서 정의되거나 발견적인(heuristic) 방법으로 결정되어진다. 이러한 이유로 인해서 퍼지 분할과 퍼지 규칙을 주어진 수치적인 입력패턴들로부터 결정하기 위한 다양한 연구들이 수행되었고 [1-8,11-14,17,19], 그들의 특징을 다음과 같이 요약될 수 있다.

가. 퍼지 규칙생성 및 특징추출 방법

Ⅰ) 규칙 선택 (Rule selection) 방법

대부분의 퍼지 규칙기반 분류시스템에서 주어진 수치적인 데이터로부터 규칙을 생성하기 위해 사용되는 방법으로, 퍼지규칙들의 수를 최소화하기 위해서 신경망[6,11,17], 유전자 알고리즘[1-4,12,13], 그리고 엔트로피[14]와 같은 학습기법들을 통해 최적의 규칙들을 획득한다.

Ⅱ) 특징 선택(Feature selection)

주어진 데이터가 가진 특징들이 고차원 혹은 다차원 분

접수일자 : 2007년 10월 6일
완료일자 : 2008년 1월 15일
* Corresponding Author

류문제에서 모두 필요한 것은 아니기 때문에 Thawonmas와 Abe[19]는 퍼지 분류기에 의해서 생성된 클래스 구간들(class regions)의 분석을 근거로 부적절한 특징들을 제거할 수 있는 알고리즘을 제안하였다.

나. 퍼지 분할 결정방법

Ⅰ) 그리드 분할(Grid partition)

주어진 도메인 구간을 정사각형 또는 직사각형 형태의 일정한 그리드로 나누어 퍼지 분할을 고정된 형태의 공간으로부터 획득하기 때문에, 퍼지 규칙이 언어적으로 해석이 가능하다는 장점을 가진다[1,2,4,13].

Ⅱ) 분산 분할(Scatter partition)

그리드 분할과 비교해볼 때, 분할의 구간이 보다 세분화된 공간을 형성하기 때문에 다차원의 속성을 가진 패턴분류 문제에서 가장 효과적인 방법이며[1], 퍼지 영역(fuzzy region)의 형태에 따라 polyhedron[5]와 hyperbox 분할[6]로 세분화될 수 있다.

Ⅲ) 계층적인 분할(Hierarchical partition)

그리드 분할의 한 형태이며, 각 입력패턴에 대응하는 퍼지 IF-THEN 규칙의 확신도(the grade of certainty)와 정의된 임계치에 따라 각 그리드 형태의 분할 공간을 2^N개의 퍼지 부분공간으로 분해하는 방법[7]이다. 예를 들어, 만약 임의의 수치적 입력패턴에 대응된 퍼지 IF-THEN 규칙의 확신도가 정의된 임계치보다 큰 값을 가지는 경우 그 규칙과 연관된 그리드 분할 공간은 퍼지 부분공간으로 세분화되지 않고, 그렇지 않은 경우 그 공간은 퍼지 부분공간으로 세분화된다. 또한 이 방법은 퍼지 분할을 결정하는데 있어서 학습기법을 사용하지 않기 때문에 추가적인 학습 파라미터들을 고려하지 않아도 되는 장점을 가진다.

지금까지 살펴본 이들 방법들은 도메인 전문가들의 개입 없이 객관적인 관점에서 퍼지 분할과 퍼지규칙을 생성할 수 있다는 장점을 제공하지만, 다음과 같은 몇몇 제약점들을 가지고 있다: ⅰ) 주어진 데이터가 가진 특성에 관계없이 그 데이터가 가진 모든 도메인 영역을 퍼지 분할공간으로 고려하고, ⅱ) 그 공간에서 정의된 퍼지 분할은 여전히 발견적 방법으로 결정되고, ⅲ) 최적의 분류 경계면을 찾기 위해 초기에 정의된 퍼지 소속함수들의 파라미터들을 조정해야 한다.

따라서 본 논문에서는 이러한 제약점을 보다 단순한 방법으로 개선시키기 위해 주어진 데이터의 통계적인 정보만을 고려한 퍼지 분할의 선택방법을 제안한다. 제안된 방법에서 통계적인 정보는 주어진 데이터로부터 각 입력 속성의 '불확실성 영역', 즉 패턴분류문제에서 분류 경계면이 결정되는 구간을 추출하기 위해 사용된다. 또한 통계적인 정보에 의해서 생성된 분할구간에 대응하는 후보 규칙들을 추출하기 위한 방법과 그 후보 규칙들 간의 커플링(충돌) 문제를 최소화하기 위한 방법도 추가적으로 논의한다.

2. 퍼지 분할 결정구간

일반적으로 패턴분류 문제에서 고려해야할 분할구간으로는 다음과 같은 2부분 즉, '불확실성 영역(uncertainty region)'과 '확실성 영역(certainty region)'으로 나뉜다. 여기서 '불확실성 영역'은 패턴분류의 성능에 직접적인 영향을 미치는 영역 즉 서로 다른 클래스로 정의된 패턴들이 밀집된 영역을 의미하고, '확실성 영역'은 분류성능에 간접적인 영향을 제공하는 영역 즉 정의된 패턴들이 밀집되지 않은 영역을 의미한다. 그림 1은 전체 도메인영역에서 '불확실성 영역'과 '확실성 영역'을 나타내기 위한 IRIS 데이터의 2개의 속성 (Petal Length와 Petal Width)만을 고려한 예이다. 그림에서 점선으로 표시된 영역은 '불확실성 영역'을 의미하고, 나머지 영역은 '확실성 영역'을 나타낸다.

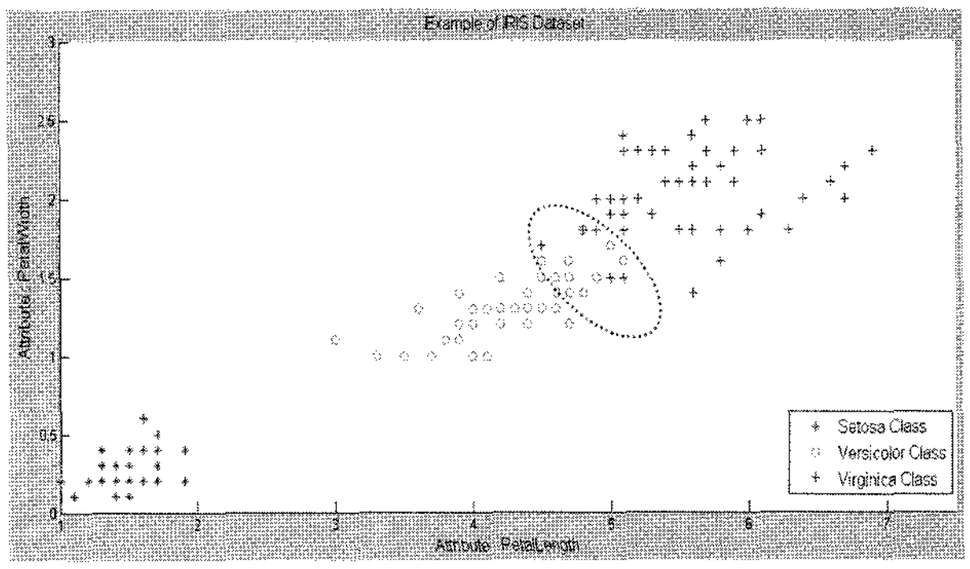

그림 1. 패턴분류를 위한 IRIS 데이터의 예
Fig. 1. Example of IRIS data for pattern classification

예를 들어, 입력패턴들이 그림 1과 같다면, 기존의 분류방법들은 주어진 패턴들을 분류하기 위해서 속성 Petal Length와 Petal Width가 가진 전체 도메인 영역(구간 [1, 7]과 구간 [0, 2.5])을 퍼지 분할구간으로 고려한다. 또한 초기의 퍼지 분할을 구성하기 위해 그 도메인 구간을 몇 개의 퍼지 부분공간으로 나누고, 분류 성능이 최대가 되는 분류 경계면을 찾기 위해서 그 정의된 퍼지 분할구간들을 지속적으로 조정하게 된다. 그러나 그림에서 볼 수 있듯이 3개의 클래스 (Setosa, Versicolor, Virginica)를 분류하기 위한 최적의 분류 경계면은 점선으로 표시된 '불확실성 영역' (구간 [4.5, 5.1]과 구간 [1.4, 1.8])에서 결정되므로, 나머지 영역은 퍼지 분할을 위해 직접적으로 고려되어질 필요는 없다. 따라서 본 논문에서는 이러한 '불확실성 영역'과 '확실성 영역'들을 추출하고, 이들 영역에서의 퍼지 분할의 수를 결정하기 위해서 출력 클래스와 연관된 각 입력속성의 통계적인 정보 (즉 최소값, 최대값, 그리고 표준편차)를 이용하였다.

3. 퍼지 분할구간의 선택과 규칙 추출

3.1 퍼지 분할의 수와 구간의 선택

만약 n개의 입력속성을 가진 데이터가 k개의 클래스로 분류된다면,

$$\begin{aligned} &A=\{a_{i1}, a_{i2}, ..., a_{in}\},\ (i=1,2,...,s)\\ &C=\{c_1, c_2, ..., c_k\},\\ &a_{ij} \in c_l,\ (j=1,2,...,n;\ l=1,2,...,k) \end{aligned} \tag{1}$$

여기서 $a_{i1}, a_{i2}, ..., a_{in}$은 주어진 데이터가 가진 n개의 입력속성을 의미하고, i는 각 입력속성에 대해서 s개의 데이터를 나타낸다. 또한 C는 출력 속성 즉 클래스의 집합을 나타내고, $a_{ij} \in c_l$은 k개의 클래스들 중에서 l번째 클래스로 분류됨을 의미한다. 이때 각 입력속성의 도메인 영역에서 퍼지 분할의 수와 그들의 구간은 다음과 같은 단계들에 의해서 결정된다.

단계 1: 각 입력속성에 대해서 s개의 데이터의 통계적 정보를 계산한다.

$$
\begin{aligned}
a_{ij} &= (a_{ij}^{\min}, a_{ij}^{\max}, a_{ij}^{\sigma}) \\
a_{ij}^{\min} &= \min(a_{1j}, a_{2j}, ..., a_{sj}) \\
a_{ij}^{\max} &= \max(a_{1j}, a_{2j}, ..., a_{sj}) \\
a_{ij}^{\sigma} &= \sqrt{\left\{\sum_{i=1}^{s}(a_{ij} - M)^2\right\}/s} \\
m &= \left\{\sum_{i=1}^{s} a_{ij}\right\}/s
\end{aligned}
\tag{2}
$$

여기서 j번째 입력속성 a_{ij}는 3개의 요소로 구성되어 있으며, $a_{ij}^{\min}$, $a_{ij}^{\max}$, a_{ij}^{σ}는 각각 j번째 속성의 최소값, 최대값, 표준편차를 의미하고, m은 그 속성의 평균값을 나타낸다.

단계 2: 주어진 데이터를 클래스별로 오름차순 또는 내림차순으로 정렬한다.

단계 3: *단계 2*에서 정렬된 데이터를 근거로 각 클래스에 대응하는 입력속성들의 데이터들을 추출한다.

$$
\begin{aligned}
&(a_{ij,m} | c_{l=1,...,k}^{m}) = \begin{cases} 1, & \text{if } a_{ij} \in c_{l=1,2,...,k}^{m} \\ 0, & else \end{cases} \\
&i = 1,2,..,s;\ j = 1,2,...,n;
\end{aligned}
\tag{3}
$$

여기서 $(a_{ij,m} | c_{l=1,...,k}^{m})$은 k개의 클래스들 중에서 m번째 클래스에 속하는 j번째 입력속성의 데이터 $a_{ij,m}$을 추출함을 나타낸다. 또한 $a_{ij} \in c_{l=1,2,...,k}^{m}$은 j번째 입력속성에서 m번째의 클래스를 포함하는 데이터가 존재하면 1, 그렇지 않으면 0임을 나타낸다.

단계 4: 식 (3)으로부터 추출된 각 입력속성의 데이터의 통계적인 정보를 계산한다.

$$
\begin{aligned}
a_{ij,m} &= [a_{ij,m}^{\min}, a_{ij,m}^{\max}] \\
a_{ij,m}^{\min} &= \min(a_{1j,m}, a_{2j,m}, ..., a_{sj,m}) \\
a_{ij,m}^{\max} &= \max(a_{1j,m}, a_{2j,m}, ..., a_{sj,m}) \\
\\
a_{ij,n} &= [a_{ij,n}^{\min}, a_{ij,n}^{\max}] \\
a_{ij,n}^{\min} &= \min(a_{1j,n}, a_{2j,n}, ..., a_{sj,n}) \\
a_{ij,n}^{\max} &= \max(a_{1j,n}, a_{2j,n}, ..., a_{sj,n})
\end{aligned}
\tag{4}
$$

여기서 $a_{ij,m} = [a_{ij,m}^{\min}, a_{ij,m}^{\max}]$와 $a_{ij,n} = [a_{ij,n}^{\min}, a_{ij,n}^{\max}]$은 각각 j번째 입력속성에 대해서 m번째와 n번째 클래스에 대응하는 데이터들의 최소값과 최대값을 의미하고, 이들 값들은 그 입력속성 내에서 정의된 클래스들의 구간을 나타낸다.

단계 5: *단계 4*로부터 획득한 각 입력속성의 클래스 구간을 근거로, 입력속성들이 가지는 클래스들 간의 중첩영역(overlapping region)을 추출한다.

$$
\begin{aligned}
UI &= [a_{ij,n}^{\min}, a_{ij,m}^{\max}] \\
CI &= [a_{ij,m}^{\min}, a_{ij,n}^{\min}], [a_{ij,m}^{\max}, a_{ij,n}^{\max}]
\end{aligned}
\quad \text{if } a_{ij,m}^{\min} < a_{ij,n}^{\min} < a_{ij,m}^{\max} < a_{ij,n}^{\max}
$$

여기서 UI는 m번째 클래스 구간 $[a_{ij,m}^{\min}, a_{ij,m}^{\max}]$과 n번째 클래스 구간 $[a_{ij,n}^{\min}, a_{ij,n}^{\max}]$이 서로 중첩될 때의 분류 경계 영역이고, CI는 그렇지 않은 경우의 영역을 나타낸다.

그림 2는 j번째 입력속성에서 m번째 클래스 $a_{ij,m}$와 n번째 클래스 $a_{ij,n}$ 간의 중첩영역과 퍼지 분할의 예를 보여준다.

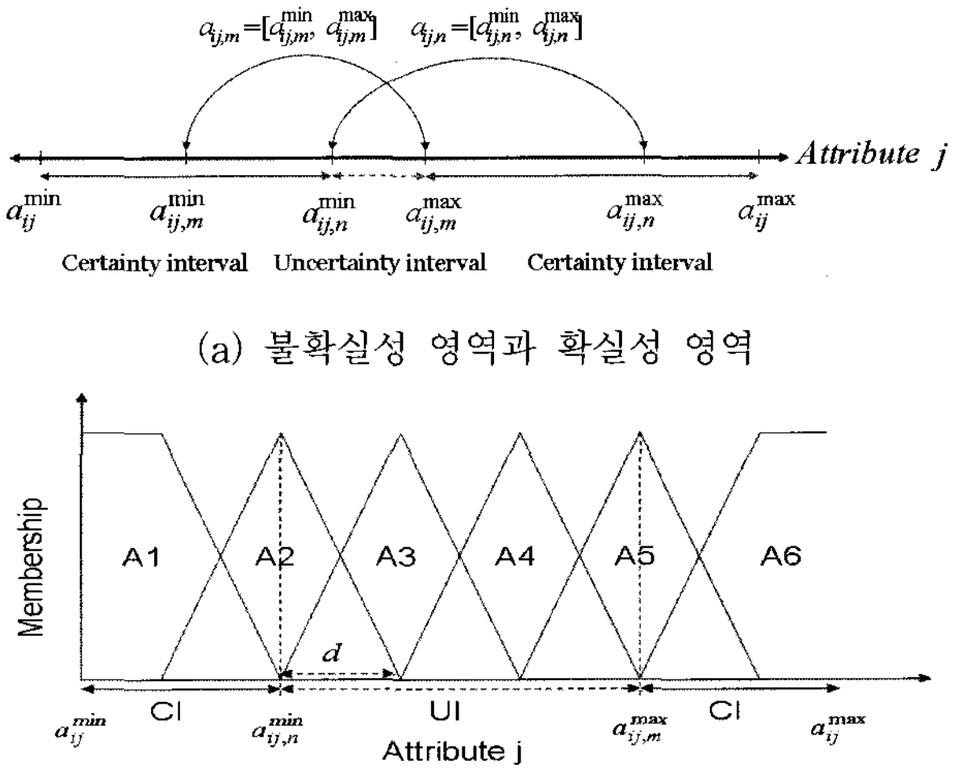

그림 2. 퍼지 분할의 예

Fig. 2. Example of fuzzy partitions (a) Uncertainty and certainty intervals (b) Fuzzy partitions for uncertainty and certain intervals

그림 2(a)는 이들 클래스 구간들의 '불확실성 영역'과 '확실성 영역'을 나타낸 것이고, 그림 2(b)는 그 영역들로부터 퍼지 분할한 예를 보여준다. 또한 그림에서 '불확실성 영역'(구간$[a_{ij,n}^{\min}, a_{ij,m}^{\max}]$)은 주어진 패턴들을 분류하는데 반드시 고려되어야 하는 퍼지 분할영역을 나타내고, 이 영역 내에서 퍼지 분할의 수가 어떻게 결정되느냐에 의해서 분류의 성능이 결정된다. 또한 '확실성 영역' (구간$[a_{ij}^{\min}, a_{ij,n}^{\min}]$과 구간$[a_{ij,m}^{\max}, a_{ij}^{\max}]$)은 그림 1에서 보여준 예에서처럼, 분류의 성능에 그다지 영향을 미치지 않는 영역을 나타낸다. 따라서 제안된 방법에서는 '불확실성 영역'에서 퍼지 분할구간의 수를 결정하기 위해서 출력 클래스와 연관된 각 속성 데이터의 통계적인 정보를 사용하였다 (단계 6 참조).

단계 6: *단계 5*에서 추출된 '불확실성 영역'과 '확실성 영역'에서의 퍼지 분할의 수와 폭을 계산한다.

$$
P = \{p_1, p_2, ..., p_n\} \tag{5}
$$

여기서 $P = \{p_1, p_2, ..., p_n\}$은 각 입력속성에 대한 퍼지 분할의 수를 나타내고, p_i는 i번째 입력속성에 대한 전체 분할의 수를 의미한다.

$$
\begin{aligned}
&p_i = (p_i^{UI} + p_i^{CI}),\ p_i \in [\alpha_i^{\min}, \alpha_i^{\max}] \\
&(\alpha_i^{\min} = a_{ij}^{\min} - a_{ij}^{\sigma};\ \alpha_i^{\max} = a_{ij}^{\max} + a_{ij}^{\sigma}) \\
&p_i^{UI} = \left(\frac{a_{ij}^{\max} - a_{ij}^{\min}}{a_{ij}^{\sigma}}\right),\ p_i^{UI} \in [p_{UI}^{\min}, p_{UI}^{\max}] \\
&p_i^{CI} = 2, \\
&d = \frac{(p_{UI}^{\max} - p_{UI}^{\min})}{p_i^{UI} - 1}
\end{aligned}
\tag{6}
$$

여기서 p_i, i번째 속성의 도메인영역은 구간 $[\alpha_i^{\min}, \alpha_i^{\max}]$에서 정의되고, 퍼지 분할의 수는 p_i^{UI}와 p_i^{CI}의 분할 수들의 합과 같다. 또한 p_i^{UI}의 퍼지 분할의 수는 그림 2에서 나타낸 중첩영역 내에서의 수를 나타내고, p_i^{CI}의 퍼지 분할의 수는 p_i^{UI}을 제외한 나머지 구간에서 각각 1개씩의 퍼지 분할의 수를 가진다. d는 각 퍼지 분할들의 간격 (하나의 퍼지 소속함수의 half-width는 $2d$의 폭을 가짐)이고, p_i는 반올림된 정수 값을 가진 i번째 속성의 퍼지 분할 수를 나타낸다.

3.2 후보규칙 추출 및 선택

3.1절에서 추출된 각 속성의 퍼지 분할구간과 분할 수를 기반으로 다음과 같은 퍼지 IF-THEN 규칙이 생성되었다고 가정하자.

$$\begin{aligned} &Rule\ R^i: IF\ x_{i1}\ is\ A_{i1}\ \text{and}\ \ldots\ \text{and}\ x_{in}\ is\ A_{in} \\ &\qquad THEN\ C_j \end{aligned}$$

여기서 $x_{i1},\ldots,x_{in}$는 n개의 퍼지 입력변수를 의미하고, $A_{i1},\ldots,A_{in}$과 $C_j(j=1,2,\ldots,k)$는 각각 입력부 소속함수와 출력부 클래스를 나타낸다. 이때 n개의 수치적인 입력패턴이 주어졌을 때 후보규칙(candidate rules)의 추출은 다음과 같다.

$$\begin{aligned} &CR_o = \max(\mu_{A_{ij}}(X)),\ (o=1,2,\ldots,t) \\ &\mu_{A_{ij}}(X) = \min(\mu_{A_{i1}}(x_1),\ldots,\mu_{A_{in}}(x_n)) \end{aligned} \tag{7}$$

여기서 CR_o는 max-min 연산으로 획득한 t개의 후보규칙을 의미하고, $\mu_{A_{i1}}(x_1), \mu_{A_{i2}}(x_2),\ldots,\mu_{A_{in}}(x_n)$는 각 수치적인 입력패턴이 대응된 소속함수들의 적합도를 나타낸다.

식 (8)-(11)은 식 (7)에 의해서 추출된 후보규칙들 간의 커플링 문제를 최소화하기 위한 기준(criterion)이다[8].

Ⅰ) 커플링 규칙이 발생하지 않는 경우

각 클래스에 대응하는 후보규칙은 해당 클래스의 규칙으로 선택한다.

Ⅱ) 커플링 규칙이 발생하는 경우

$$f_{C_m} = \sum_{i=1,..,n;o=1}^{t} CR_o \tag{8}$$

$$f_{C_n} = \sum_{i=1,..,n;o=1}^{t} CR_o \tag{9}$$

여기서 f_{C_m}, f_{C_n}는 각각 전체 n개의 규칙 중에서 m와 n번째 클래스에 대한 후보규칙의 출현 빈도수를 나타낸다.

$$CR_o = \begin{cases} C_m, & \text{if } f_{C_m} > f_{C_n} \\ C_n, & else \end{cases} \tag{10}$$

여기서 $f_{C_m} > f_{C_n}$는 m번째 후보규칙의 출현 빈도수가 n번째 후보규칙의 빈도수보다 큰 경우 m번째 클래스의 규칙으로 선택하고, 그렇지 않으면 n번째 클래스의 규칙으로 선택함을 의미한다.

$$CR_o = \begin{cases} C_m, & \text{if } f_{C_m} = f_{C_n},\ \max_o(\mu_{C_m}) > \max_o(\mu_{C_m}) \\ C_n, & else \end{cases} \tag{11}$$

식 (11)은 m번째 후보규칙의 출현 빈도수와 n번째 후보규칙의 출현 빈도수가 같다면, 최종 규칙은 m번째 후보규칙과 대응하는 클래스의 최대-적합도와 n번째 후보규칙과 대응하는 클래스의 최대 적합도를 비교하여 더 큰 적합도를 가지는 후보규칙을 클래스의 규칙으로 선택함을 나타낸다.

3.3 규칙감축

본 절에서는 3.2절에서 획득한 규칙의 수를 최소화하기 위해서 Skowron[9]에 의해서 제안된 러프집합의 식별가능 행렬을 이용하였다.

$$(c_{ij}) = \{a \in A: a(x_i) \neq a(x_j)\},\ \text{for } i,j=1,2,\ldots,n \tag{12}$$

여기서 c_{ij}는 속성 x_i, x_j을 구별하게 하는 모든 속성들의 집합 (즉, $n \times n$ 식별가능행렬)을 의미한다. 따라서 식별가능 행렬에서 코어는 주어진 속성들 중에서 단일 원소로 이루어진 속성들의 집합이므로 다음과 같이 정의될 수 있다.

$$core(A) = \{a \in A: c_{ij} = (a),\ \text{for } some\ i,j\} \tag{13}$$

식 (13)으로부터 리덕트를 계산하기 위하여 다음과 같이 식별가능 함수 $f(A)$을 계산할 수 있다.

$$f(A) = \prod_{(x,y) \in U^2} \left\{ \sum \delta(x,y) : (x,y) \in U^2 \text{and}\ \delta(x,y) \neq 0 \right\} \tag{14}$$

식 (14)에 의해서 계산된 리덕트로부터 각 속성 x에 대한 최종 식별가능 함수 $f^x(A)$을 다시 정의할 수 있다.

$$f^x(A) = \prod_{y \in U} \{\delta(x,y): y \in U \text{and}\ \delta(x,y) \neq 0\} \tag{15}$$

위의 과정으로부터 추출된 규칙은 여러 개의 리덕트를 포함하기 때문에 패턴분류에서 사용되어지는 규칙들의 수가 증가할 수 있다. 그러므로 제안된 방법에서는 보다 적은 수의 규칙을 획득하기 위해서 한 개의 리덕트만을 포함하는 규칙을 선택함으로써 최종 규칙을 구성하였다.

4. 실험결과 및 검토

실험에서는 제안된 방법의 효과성을 보이기 위해서 패턴분류의 벤치마킹 데이터로 주로 활용되는 UCI Machine Learning Repository의 IRIS와 New Thyroid Cancer 데이터를 사용하였고[10], 다음은 이들 데이터들의 특징정보를 나타낸다.

항목	IRIS	New Thyroid Cancer
데이터 수	150개	215개
속성 수	4개	5개
클래스 수	3개	3개

4.1 IRIS 데이터의 퍼지 분할구간과 분할의 수

다음은 입력부 퍼지집합의 퍼지 분할을 결정하기 위한 IRIS 데이터의 통계적인 정보를 나타낸다.

표 1. IRIS 데이터의 통계적 정보
Table 1. Statistical information on IRIS data

클래스 / 속성	Setosa		Versicolor		Virginica		SD
	Min	Max	Min	Max	Min	Max	
SL	4.3	5.8	4.9	7.0	4.9	7.9	0.828
SW	2.3	4.4	2.0	3.4	2.2	3.8	0.436
PL	1.0	1.9	3.0	5.1	4.5	6.9	1.762
PW	0.1	0.6	1.0	1.8	1.4	2.5	0.765

SL : Sepal Length, SW : Sepal Width, PL : Petal Length, PW : Petal Width

표 1에서 Min과 Max는 주어진 150개의 데이터에 대해서 *단계 1*에서 *단계 4*를 수행한 후 생성된 각 속성의 최소값과 최대값을 나타내고, SD는 각 속성의 표준편차를 나타낸다.

표 2. IRIS 데이터의 퍼지 분할의 수와 분할구간
Table 2. Number of partitions and their intervals for IRIS data

속성	전체영역	불확실성 영역		확실성 영역	
		구간	분할 수	구간	분할 수
SL	[3.472 8.728]	[4.9,5.8]	4	[3.472,4.900] [5.800,8.728]	2
SW	[1.564 4.836]	[2.3,3.4]	6	[1.564,2.300] [3.400,4.836]	2
PL	[-0.762 8.665]	[4.5,5.1]	3	[-0.762,4.500] [5.100,8.665]	2
PW	[-0.665 3.262]	[1.4,1.8]	3	[-0.665,1.400] [1.800,3.262]	2

표 2는 *단계 5*와 *단계 6*으로부터 생성된 4개의 입력속성의 '불확실성 영역','확실성 영역', 그리고 퍼지 분할의 수를 나타낸다. 또한 전체 퍼지 분할의 수는 식 (6)에서처럼 '불확실성 영역'과 '확실성 영역'의 분할 수들의 합과 같다.

4.2 IRIS 데이터의 입력부 퍼지집합

표 2의 퍼지 분할의 수와 분할구간을 근거로 IRIS 데이터를 분류하기 위한 입력부 퍼지집합을 그림 3과 같이 구성하였다. 그림 3에서 속성 Sepal length에서 'SL1, SL6', 속성 Sepal width에서 'SW1, SW8', 속성 Petal length에서 'PL1, PL5', 그리고 속성 Petal width에서 'PW1, PW5'는 각각 그림 2(b)와 표 2의 '확실성 영역'에서 퍼지 분할구간을 의미하고, 그들을 제외한 나머지 구간들 (4개, 6개, 3개, 3개)은 '불확실성 영역'내에서 나누어진 퍼지 분할구간들을 나타낸다. 또한 이들 구간들의 폭(width)은 식 (6)에서 d에 의해 모두 동일한 크기로 분할되었다.

4.3 규칙 추출 및 선택

그림 3의 퍼지 분할구간을 이용하여 식 (7)의 후보규칙을 추출한 결과, Setosa, Versicolor, 그리고 Virginica 클래스의 규칙은 각각 14개, 43개, 27개 (총 84개)가 생성되었으며, 규칙감축 전 다음과 같은 규칙 패턴에서 커플링 규칙이 발생하였다.

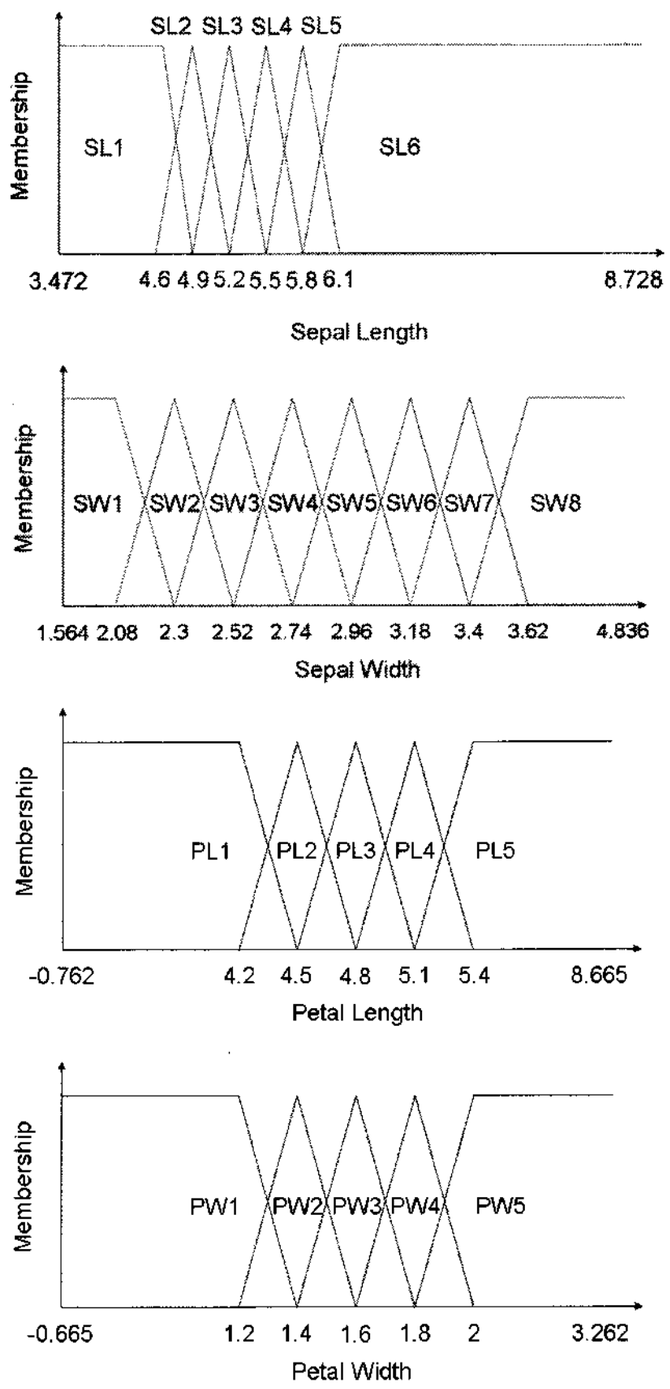

그림 3. IRIS의 입력부 퍼지집합
Fig. 3. Antecedent fuzzy sets for IRIS

IF Sepal length is SL6, Sepal width is SW4,
Petal length is PL4, Petal width is PW3
THEN Class is Versicolor (0.6667) or *Virginica* (0.5).

여기서 ( )는 주어진 수치적인 입력패턴에 대응된 Versicolor와 Virginica 클래스의 최대 적합도를 나타내고, 식 (11)에 의해서 위 규칙은 Versicolor 클래스로 선택하였다. 또한 식 (12)에서 (15)까지 수행 후 리덕트가 1개인 규칙만을 추출하여 최종 규칙을 구성하였고, 다음은 그 결과를 보여준다.

규칙 감축 후 클래스별 규칙의 수			
Setosa	Versicolor	Virginica	총 규칙 수
10	29	8	47

4.4 기존의 방법과의 분류결과 비교

기존의 분류방법과의 분류 정확성을 비교하기 위해서, 제안된 방법은 각 수치적인 입력패턴에 대해서 그림 3의 퍼지 분할구간에 대응된 퍼지 IF-THEN 규칙들 가운데 최대 적합도를 가지는 규칙의 출력부를 클래스로 분류하였다. 표 3은 기존의 패턴 분류방법들과 제안된 방법과의 분류 정확도를 비교한 결과들을 나타낸 것이다.

표 3에서, Nauck[11], Nozaki[12], Ishibuchi[13], 그리고 Guo[15]의 방법들은 제안된 방법에 비해 규칙생성의 측면에서 좋은 성능을 제공한다는 것을 볼 수 있으나, 분류 정확도 측면에서는 그렇지 않음을 알 수 있다. 또한

Ishibuchi[3], Tsai[14] 및 Chen[16]의 방법들 역시 제안된 방법보다 규칙생성의 측면에서는 좋은 성능을 제공하지만, 학습과정에 기반을 둔 이들 방법들은 패턴분류의 좋은 성능(분류 정확도와 규칙의 수)을 보장하기 위해 고려되어야 할 학습 파라미터들이 너무 많고, 그들을 최적화하는데 다소 많이 시간이 필요하다. 그러므로 제안된 방법은 주어진 데이터가 가진 통계적인 정보만을 고려하기 때문에 이들 방법들에 비해 단순하고 효과적인 방법이라 할 수 있다.

표 3. IRIS 데이터의 비교결과
Table 3. Comparison results of IRIS data

패턴분류 방법		규칙 수 및 분류 정확도	
Neuro-fuzzy approach	Nauck[11]	7개	145개(96.6%)
Adjustment of rule weight	Nozaki[12]	6개	142개(94.8%)
Genetic-algorithm approach	Ishibuchi[3]	6개	149개(99.3%)
	Ishibuchi[13]	5개	147개(98.0%)
Entropy based fuzzy-rough	Tsai[14]	16개	149개(99.3%)
Hierarchical fuzzy-based approach	Guo[15]	분할 수 : 6 (서브시스템 1 : 27개, 서브시스템 2 : 30개)	146개(97.33%)
	Chen[16]	16개	149개(99.6%)
Fuzzy-rough classification	Son[8]	41개 가중치 고려	147개(98.0%) 148개(98.7%)
The proposed method		47개	149개(99.3%)

표 4는 IRIS 실험과 같은 방법으로 New Thyroid Cancer 데이터의 규칙을 생성한 결과들을 나타낸다.

표 4. 규칙 감축 전과 후의 규칙들의 수
Table 4. Number of rules before and after rule reduction

클래스 / 규칙	Normal	Hyper	Hypo	규칙의 수
규칙감축 전	127	25	22	174
규칙감축 후	45	13	9	67

표 4에서 '규칙감축 전'은 주어진 데이터의 통계적인 정보만을 사용하여 '불확실성 영역'과 '확실성 영역'에서 퍼지 분할구간들을 결정하고, 그 분할구간들에 대응하는 수치적인 입력패턴들의 클래스-규칙들의 수를 나타낸다. '규칙감축 후'는 러프집합의 식별가능 행렬을 이용하여 규칙을 감축한 후 리덕트가 1개인 규칙들만 추출한 클래스-규칙들의 수를 나타낸다. 표 5는 Jang[17]의 ANFIS와 Chinneck[18]의 방법으로 New Thyroid Cancer 데이터의 분류 정확도를 비교한 결과들을 나타낸 것이다. 표 5에서 Jang의 분류 정확도는 5개의 입력속성에 대해서 소속함수의 수를 각각 2개와 3개로 실험하였을 때의 결과들을 나타낸다. 이들 실험결과들로부터 알 수 있는 사실은 주어진 데이터의 통계적인 정보만을 고려하여 패턴분류의 분류 경계면이 결정되는 '불확실성 영역'을 추출하고 그들의 퍼지 분할구간을 결정하더라도 기존의 분류방법들과 유사하거나 보다 좋은 성능을 제공할 수 있고, 패턴분류를 위해 전체 도메인영역을 퍼지 분할영역으로 고려하지 않아도 된다는 점이다.

표 5. New Thyroid Cancer 데이터의 비교 결과
Table 5. Comparison results of New Thyroid Cancer data

패턴분류 방법	규칙 수 및 분류 정확도	
Jang[17]	32개 243개	212개(98.6%) 214개(99.5%)
Chinneck[18]	204개	204개(94.9%)
The proposed method	67개	215개(100%)

5. 결 론

기존의 패턴분류 방법들은 주어진 데이터로부터 초기에 퍼지 분할을 생성하기 위해 그 데이터가 가진 각 속성의 전체 도메인 영역을 퍼지 분할구간으로 고려하고, 그 구간 내에서 최적의 분류 경계면을 찾기 위해 다양한 학습과정들을 고려한다. 그러나 패턴분류에서 사용되는 모든 데이터들은 그 데이터들이 표현되어지는 영역들 내에서 몇몇의 특정 영역('불확실성 영역')에서 분류 경계면이 결정된다.

본 논문에서는 주어진 데이터의 통계적인 정보만을 활용하여 이러한 '불확실성 영역'을 가진 구간들을 추출하고, 그 추출된 구간에서 퍼지 분할을 결정할 수 있는 방법을 제안하였다. 또한 제안된 방법에서는 그 퍼지 분할구간에 대응하는 후보규칙을 추출하고 후보규칙간의 커플링 문제를 최소화할 수 있는 방법에 대해서도 추가적으로 논의하였다. 제안된 방법의 효용성을 보이기 위해 IRIS와 New Thyroid Cancer 데이터를 사용하여 기존의 패턴분류 방법과의 분류 정확성을 비교하였다. 그 결과, 패턴분류를 위한 분할구간의 결정에 있어서 학습과정(신경망, 유전자 알고리즘, 엔트로피 등)들을 고려하지 않고 주어진 데이터의 통계적인 정보만을 고려하더라도 보다 좋은 성능을 얻을 수 있음을 확인하였다. 향후 연구과제로는 퍼지 규칙의 수를 최소화하는 방법과 라벨링되지 않은 데이터를 분류할 수 있는 방법으로의 확장이 필요하다고 생각된다.

참 고 문 헌

[1] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," *Information sciences*, vol. 136, no. 1-4, pp. 109-133, 2001.

[2] H. Ishibuchi and T. Yamamoto, "Fuzzy rule selection by data mining criteria and genetic algorithms," *Genetic evolutionary computation conf. (GECCO-2002)*, pp. 399-406, 2002.

[3] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, "Construction of fuzzy classification systems with rectangular fuzzy rules using ge-

netic algorithm," *Fuzzy sets and systems*, vol. 65, no. 2-3, pp. 237-253, 1994.
[4] J. Casillas, O.Cordon, M.J.Del Jesus, and F. Herrera, "Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems," *Information sciences*, vol. 136, pp. 135-157, 2001.
[5] V. Uebele, S. Abe, and M.S. Lan, "A neural network based fuzzy classifier," *IEEE Transactions on SMC*, vol. 25, pp. 353-361, 1995.
[6] S. Abe and M.S. Lan, "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification," *IEEE Transactions on fuzzy systems*, vol. 3, pp. 18-28, 1995.
[7] M.A. Kbir, H. Benkirane, K. Maalmi, and R. Benslimane, "Hierarchical fuzzy partition for pattern classification with fuzzy if-then rules," *Pattern recognition letters*, vol. 21, no. 6-7, pp. 503-509, 2000.
[8] 손창식, 정환묵, 서석태, 권순학, "규칙의 커플링 문제를 최소화하기 위한 퍼지-러프 분류 방법," *한국 퍼지 및 지능시스템 학회 논문지*, vol. 17, no. 4, pp. 460-465, 2007.
[9] A. Skowron and C.M. Rauszer, "The Discernibility matrices and functions in information systems," *Institute of computer sciences report 1/91, Technical University of Warsaw*, pp. 1-41, 1991.
[10] UCI Repository of Machine Learning Databases, *Depatment of Information and Computer Science, University of California, Irvine, CA*, Available: http://mlearn.ics.uci.edu/MLRepository.html.
[11] D. Nauck, U. Nauck, and R. Kruse, "Generating classification rules with the neuro-fuzzy system NEFCLASS," *In proceedings of the biennial conference of NAFIPS, Berkeley*, pp. 19-22, 1996.
[12] S.M. Nozaki, H. Ishibuchi, and H. Tanaka, "Adaptive fuzzy rule based classification systems," *IEEE Transactions on fuzzy systems*, vol. 4, no. 3, pp. 238-250, 1996.
[13] H. Ishibuchi, T. Murata, and I.B. Turksen, "Single objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems," *Fuzzy sets and systems*, vol. 89, no. 2, pp. 135-150, 1997.
[14] Y.C. Tsai, C.H. Cheng, and J.R. Chang, "Entropy-based fuzzy rough classification approach for extracting classification rules," *Experts systems with applications*, vol. 31, no. 2, pp. 436-443, 2006.
[15] N. R. Guo, T.H.S. Li, and C.L. Kuo, "Hierarchical fuzzy model for classification problem," *IECON 2002*, vol. 3, pp. 2096-2101, 2002.
[16] Y. Chen, B. Yang, A. Abraham, and L. Peng, "Automatic design of hierarchical takagi-sugeno type fuzzy systems using evolutionary algorithms," *IEEE transactions on fuzzy systems*, vol. 15, no. 3, pp. 385-397, 2007.
[17] J-S. R. Jang, "ANFIS: Adaptive network based fuzzy inference systems," *IEEE Transactions on SMC*, vol. 23, no. 3, pp. 665-695, 1993.
[18] J.W. Chinneck, "Fast heuristics for the maximum feasible subsystem problem," *INFORMS Journal on computing*, vol. 13, no. 3, pp. 210-223, 2001.
[19] R. Thawonmas and S. Abe, "A novel approach to feature selection based on analysis of class regions," *IEEE Transactions on SMC*, vol. 27, pp. 196-207, 1997.

저 자 소 개

손창식(Chang S. Son)
제17권 제6호 참조

fuzzyrisk@paran.com

정환묵(Hwan M. Chung)
제17권 제6호 참조

hmchung@cu.ac.kr

권순학(Soon H. Kwon)
제17권 제6호 참조

shkwon@yu.ac.kr