

데이터 분할 평가 진화알고리즘을 이용한 효율적인 퍼지 분류규칙의 생성

Generation of Efficient Fuzzy Classification Rules Using Evolutionary Algorithm with Data Partition Evaluation

류정우¹ · 김성은² · 김명원³⁺

Joung Woo Ryu¹, Sungeun Kim² and Myung Won Kim³⁺

1 한국전자통신연구원 지능형로봇연구단

2 (주)퓨처시스템 정보통신연구소

3 숭실대학교 컴퓨터학부

요 약

데이터 속성 값이 연속적이고 애매할 때 퍼지 규칙으로 분류규칙을 표현하는 것은 매우 유용하면서도 효과적이다. 그러나 효과적인 퍼지 분류규칙을 생성하기 위한 소속함수를 결정하기는 어렵다.

본 논문에서는 진화알고리즘을 이용하여 효과적인 퍼지 분류규칙을 자동으로 생성하는 방법을 제안한다. 제안한 방법은 지도 군집화로 클래스 분포에 따라 초기 소속함수를 생성하고, 정확하고 간결한 규칙을 생성할 수 있도록 초기 소속함수를 진화시키는 방법이다. 또한 진화알고리즘의 시간에 대한 효율성을 높이기 위한 방법으로 데이터 분할 평가 진화 방법을 제안한다. 데이터 분할 평가 진화 방법은 전체 학습 데이터를 여러 개의 부분 학습 데이터들로 나누고 개체는 전체 학습 데이터 대신 부분 학습 데이터를 임의로 선택하여 평가하는 방법이다. UCI 벤치마크 데이터로 기존 방법과 비교 실험을 통해 평균적으로 제안한 방법이 효과적임을 보였다. 또한 KDD'99 Cup의 침입탐지 데이터에서 KDD'99 Cup 우승자에 비해 1.54% 향상된 인식률과 20.8% 절감된 탐지비용을 보였고 데이터 분할 평가 진화 방법으로 개체평가 시간을 약 70% 감소시켰다.

키워드 : 퍼지 분류규칙, 진화알고리즘, 데이터 분할 평가, 소속함수, 지도군집화

Abstract

Fuzzy rules are very useful and efficient to describe classification rules especially when the attribute values are continuous and fuzzy in nature. However, it is generally difficult to determine membership functions for generating efficient fuzzy classification rules.

In this paper, we propose a method of automatic generation of efficient fuzzy classification rules using evolutionary algorithm. In our method we generate a set of initial membership functions for evolutionary algorithm by supervised clustering the training data set and we evolve the set of initial membership functions in order to generate fuzzy classification rules taking into consideration both classification accuracy and rule comprehensibility. To reduce time to evaluate an individual we also propose an evolutionary algorithm with data partition evaluation in which the training data set is partitioned into a number of subsets and individuals are evaluated using a randomly selected subset of data at a time instead of the whole training data set. We experimented our algorithm with the UCI learning data sets, the experiment results showed that our method was more efficient at average compared with the existing algorithms. For the evolutionary algorithm with data partition evaluation, we experimented with our method over the intrusion detection data of KDD'99 Cup, and confirmed that evaluation time was reduced by about 70%. Compared with the KDD'99 Cup winner, the accuracy was increased by 1.54% while the cost was reduced by 20.8%.

Key Word : fuzzy classification rule, evolutionary algorithm, data partition evaluation, membership function, supervised clustering

1. 서 론

데이터 속성 값이 연속적이고 애매할 때 퍼지 규칙으로 분류규칙을 표현하는 것은 매우 유용하면서도 효과적이다. 그러나 효과적인 퍼지 분류규칙을 생성하기 위한 소속함수를 결정하기는 어렵다.

연속형(continuous) 속성을 갖는 데이터에서 분류규칙을 생성할 경우, 연속형 속성의 범위를 일정한 기준에 따라 여러 구간으로 나누고 각 구간을 기호화하여 범주형

접수일자 : 2007년 11월 20일

완료일자 : 2007년 12월 13일

본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
+ 교신저자

(categorical) 속성으로 바꾸어야 한다. 이와 같이 연속형 속성을 범주형 속성으로 변환하는 것을 이산화(discretization)라 한다. 이산화된 범주형 속성은 구간들 간의 경계가 명확하므로 경계 부근에서 작은 값의 변화에도 범주가 달라지는 예민한 경계 문제(sharp boundary problem)를 갖는다. 이러한 문제는 생성되는 분류규칙의 정확성(accuracy)에 큰 영향을 미친다.

퍼지 분류규칙은 퍼지화(fuzzification)를 통해 연속형 속성에서 구간을 소속함수로 정의하고 이웃한 소속함수들을 중첩시켜 구간들 간의 경계를 퍼지경계(fuzzy boundary)로 정의함으로써 예민한 경계 문제를 완화시킬 수 있다. 그러나 퍼지 분류규칙의 정확성은 소속함수의 모양에 따라 영향을 받는다. 또한 이산화뿐만 아니라 퍼지화에서도 연속형 속성에 대해서 소속함수를 정의할 구간의 개수를 결정해야 하는 알갱이(granularity) 문제가 존재하고, 이는 생성되는 분류규칙의 간결성(comprehensibility)에 영향을 미친다.

생성할 분류규칙의 정확성을 높이려면 규칙 개수와 규칙의 조건항 개수가 많아지므로 간결성이 떨어지고 반면 간결성을 높이려면 정확성이 떨어진다. 이와 같이 효율적인 분류규칙을 생성할 때 고려되는 두 가지 척도는 서로 상반관계(trade-off)에 있다.

본 논문에서는 효율적인 퍼지 분류규칙을 생성하기 위해 지도 근집화로 클래스 분포에 따라 초기 소속함수의 개수와 모양을 결정하고 정확성이 높은 간결한 퍼지 분류규칙을 생성할 수 있도록 소속함수의 개수와 모양을 진화알고리즘으로 진화시키는 방법을 제안한다.

제한한 방법에서 퍼지 분류규칙은 퍼지 의사결정트리로 생성하므로 학습 데이터 개수가 많아지면 퍼지 분류규칙을 생성하는 개체평가 단계에서 시간이 많이 소요되어 진화알고리즘의 수행시간이 증가된다. 따라서 본 논문에서는 모델 최적화를 위한 진화알고리즘의 수행시간 단축을 위해, 개체평가 단계에서 모델을 생성할 때 학습 데이터 전체를 이용하는 대신 샘플링된 부분 학습 데이터로 개체를 평가하는 데이터 분할 평가 진화알고리즘(EA-DPE: evolutionary algorithm with data partition evaluation)을 제안한다.

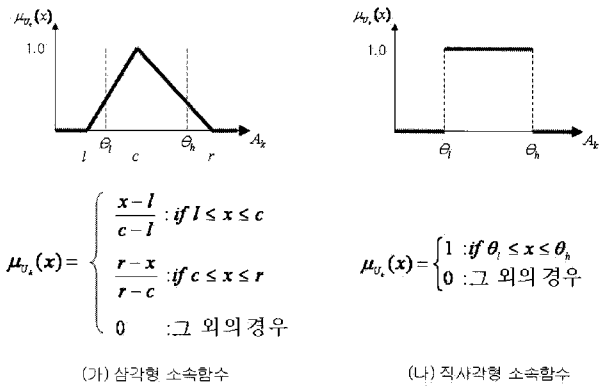
2절에서는 본 논문에서 사용하고 있는 퍼지 분류규칙의 형태와 소속함수의 표현법을 기술하면서 예를 통해 퍼지 분류규칙이 일반 분류규칙(crisp classification)보다 효과적임을 보여준다. 그리고 진화알고리즘에 대하여 간략히 설명한다. 3절에서는 퍼지 분류규칙을 생성하는 기존 방법들을 기술하고 4절에서는 제한한 방법을 기술한다. UCI 벤치마크 데이터와 KDD'99 Cup 침입탐지 데이터로 제한한 방법의 타당성을 실험한 결과를 5절에서 제시하며 마지막으로 6절에서는 결론 및 향후 연구를 논의한다.

2. 퍼지 분류규칙과 진화알고리즘

본 논문에서 사용하는 퍼지 분류규칙은 식 (1)과 같이 여러 개의 조건항이 논리곱으로 연결된 형태의 조건부와 클래스 분류를 나타내는 결론부, 그리고 해당 규칙에 대한 확신도(CF: Certainty Factor)로 구성되어 있다.

$$R_i : \text{IF } A_1 \text{ is } U_1 \text{ AND } A_2 \text{ is } U_2 \dots \text{ AND } A_m \text{ is } U_m \text{ THEN Class is } k \text{ (CF}_i\text{)} \quad (1)$$

식 (1)에서 A_k 는 k 번째 데이터 속성을 나타내며 U_k 는 속성 A_k 의 값을 나타내는 언어항으로서 퍼지집합을 의미한다.

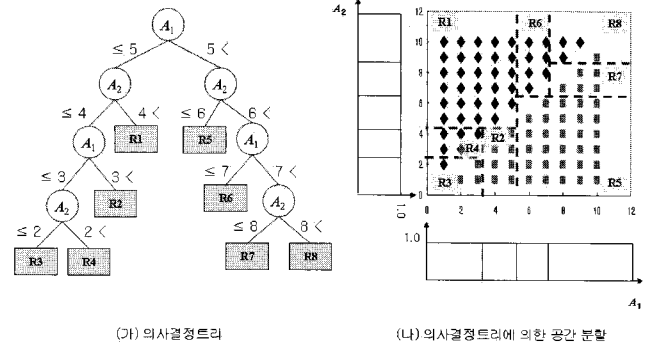


(가) 삼각형 소속함수 (나) 직사각형 소속함수

그림 1. 퍼지집합의 소속함수

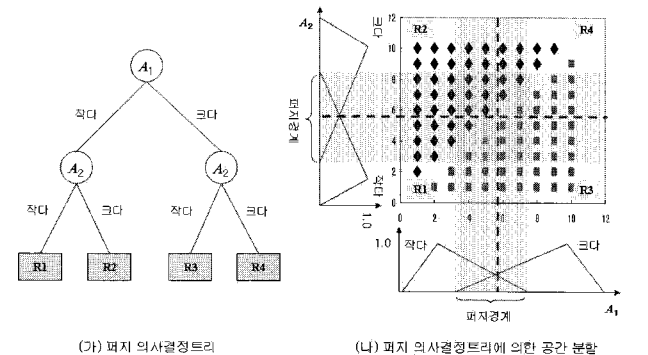
Fig. 1. Membership functions of fuzzy set

본 논문에서 퍼지집합, U_k 는 <그림 1>의 (가)와 같이 삼각형 소속함수로 표현한다. 만약 소속함수를 <그림 1>의 (나)와 같은 직사각형 소속함수로 표현하면 일반 분류규칙이 된다. 따라서 일반 분류규칙을 사용할 경우 직사각형 소속함수에 의해 0 또는 1의 소속정도($\mu_{U_k}(x)$: 퍼지집합 U_k 에 대한 x 의 소속정도)를 가지므로 경계점(θ_l, θ_h) 근처에서 작은 값의 변화에도 집합의 소속 여부가 달라지는 예민한 경계 문제를 갖는다.



(가) 의사결정트리 (나) 의사결정트리에 의한 공간 분할

그림 2. C4.5에 의해 생성된 의사결정트리와 공간 분할
Fig. 2. A decision tree and an input space partition generated by C4.5



(가) 퍼지 의사결정트리 (나) 퍼지 의사결정트리에 의한 공간 분할

그림 3. FDT에 의해 생성된 퍼지 의사결정트리와 공간 분할
Fig. 3. A fuzzy decision tree and an input space fuzzy partition generated by FDT

벡터공간에서 분류규칙은 클래스 분포에 따라 분할된 부

분공간을 의미하고 일반적으로 의사결정트리로 생성한다. <그림 2>의 (가)는 C4.5[1]로 생성한 의사결정트리이고 (나)는 생성된 일반 분류규칙에 의해 분할된 부분공간을 보여준다. <그림 3>는 FDT(fuzzy decision tree)[2]로 생성한 퍼지 의사결정트리와 생성된 퍼지 분류규칙에 의해 분할된 부분공간이다. 퍼지 분류규칙에 의해 분할된 부분공간은 이웃한 소속함수들을 중첩시켜 퍼지경계를 형성함으로써 경계부근에서 데이터의 애매성을 표현하고, 이는 퍼지 분류규칙이 예민한 경계 문제의 영향을 받지 않는 이유이다. 또한 그림과 같이 서로 다른 클래스가 축에 평행하지 않게 분포되어 있을 경우 퍼지 분류규칙이 일반 분류규칙보다 적은 개수로 데이터를 분류할 수 있다. 본 논문에서는 FDT를 이용하여 퍼지 분류규칙을 생성한다.

```

procedure EA
begin
    t=0
    초기화(P)
    while not 종료조건 do
        begin
            개체평가(P)
            t=t+1
            Pt=재생산(Pt-1)
            교배(P)
            돌연변이(P)
        end
    end
end
    
```

그림 4. 진화알고리즘
Fig. 4. Evolutionary algorithm

진화알고리즘은 자연계 생물의 진화 메커니즘을 모방한 확률적 탐색 알고리즘으로서 주로 최적화 문제에 많이 응용되고 있다. 주어진 문제의 잠재적 해를 개체 (individual)로 인코딩하여 초기 개체집단(세대)을 만들고 개체가 주어진 문제의 제약조건에 얼마나 부합한지 목적함수를 평가하여 각 개체의 적합도(fitness score)를 계산한다. 적합도가 상대적으로 우수한 개체(부모 개체)들을 선택(재생산)한 후 교배(crossover), 돌연변이(mutation) 등의 진화연산을 적용하여 새로운 개체(자식 개체)를 생성하고 이 새로운 개체들이 새로운 세대를 이룬다. 이와 같이 개체의 평가, 진화연산 적용 등의 과정을 종료조건이 만족될 때까지 반복적으로 수행함으로써 개체를 시켜 문제에 대한 최적의 해를 구한다. <그림 4>는 진화알고리즘을 나타내며 P_t 는 t 세대의 개체집단을 나타낸다.

3. 관련 연구

퍼지 분류규칙 생성에 가장 일반적으로 적용되는 방법은 퍼지 의사결정트리를 생성하여 추출하는 방법[3,4]이다. [3]에서는 사전에 정의된 소속함수로부터 전역적 이산화(global discretization)로 퍼지 의사결정트리를 생성하고, 보다 간결하고 정확한 트리를 생성하기 위해 분류에서 속성의 기여도를 이용한 휴리스틱 방법을 제안하였다. FID(fuzzy ID3)[4]는 C4.5처럼 퍼지 무질서도(fuzzy entropy)가 가장 낮은 분할점을 찾고 이를 기준으로 퍼지경계가 생길 수 있도록 두 개의 사다리꼴 소속함수를 생성하였다. FID는 <그림 2>와

같이 지역적 이산화(local discretization)로 트리를 생성함으로써 트리가 생성되면 각 속성마다 소속함수 개수가 결정된다. 트리 생성 과정에서 전역적 이산화는 한 속성을 한 번만 사용하여 공간을 분할하는 반면 지역적 이산화는 한 속성을 여러 번 사용하여 공간을 분할한다. [3,4]에서 사용된 퍼지 의사결정트리는 트리가 생성될 때까지 한번 결정된 소속함수의 형태는 고정되어 있으므로 생성된 규칙의 정확성과 간결성을 보장하기가 어렵다.

Hisao Ishibuchi는 사전에 균등하게 정의된 소속함수로 벡터공간을 격자(grid)로 분할하여 초기 퍼지 분류규칙을 생성한 후, 정확한 퍼지 분류규칙을 생성하기 위해 [5]에서는 오류보정학습 방법으로 데이터를 잘못 분류한 규칙의 소속함수를 학습시켰고, [6]에서는 학습 데이터의 분류 결과에 따라 퍼지 분류규칙의 확신도를 학습시켜 퍼지 분류규칙을 생성하였다. 그러나 이러한 방법들은 사전에 적절한 소속함수 개수를 결정하기 어렵기 때문에 연속형 속성을 두 개에서부터 다섯 개까지 균등하게 분할된 14개의 삼각형 소속함수와 속성의 전체 범위에서 정의된 직사각형 소속함수 1개를 포함한 총 15개의 소속함수에서 최대 조건항의 개수(벡터공간 차원 개수)만큼 중복 선택하면서 분류규칙의 조건부를 생성한 후에 결론부와 확신도를 계산하여 규칙을 생성하였고, 진화 다중목적 최적화(evolutionary multi-objective optimization)로 정확하고 간결한 퍼지 분류규칙을 선택하였다.[7,8,9].

[10],[11]에서는 사전에 소속함수를 결정하고 초기 퍼지 분류규칙을 생성하여 진화알고리즘으로 정확하고 간결한 퍼지 분류규칙을 생성하였다. 특히, 간결한 규칙을 위해 진화 중에 규칙기반 간소화(rule base simplification) 방법을 적용하여 규칙의 조건부를 간소화시켰다. [10]에서는 지도 퍼지 군집화로 소속함수를 결정하고 초기 퍼지 분류규칙을 생성하였다. 그러나 생성될 수 있는 퍼지 분류규칙의 개수를 클래스의 개수로 제한시키므로 동일한 클래스가 다수의 군집을 갖는 분류 문제에서는 정확한 규칙을 생성하기 어렵다. 이러한 문제를 보완하기 위해 [11]에서는 지도 퍼지 군집화 대신 C4.5로 의사결정트리를 생성한 후 초기 퍼지 분류규칙을 생성하였다. C4.5로 초기 퍼지 분류규칙을 생성하는 것은 <그림 2>와 같이 많은 규칙들이 생성되므로 간결한 퍼지 분류규칙 생성을 위한 진화알고리즘의 효율성이 떨어진다.

제안한 방법은 앞서 기술한 기존 방법과 다르게 소속함수를 진화시키면서 퍼지 의사결정트리를 생성하고 평가하므로 소속함수와 퍼지 분류규칙이 자동으로 생성된다. 따라서 퍼지화에서 발생하는 알갱이 문제를 해결할 수 있다.

또한 본 논문에서는 KDD'99 Cup의 침입탐지 데이터로 제안한 방법의 타당성을 확인하였다. 동일한 데이터에서 퍼지 분류규칙을 생성한 기존 연구로, 소속함수를 사전에 정하고 퍼지 분류규칙을 생성하는 EFRID(evolving fuzzy rules for intrusion detection) 방법[12]과 에이전트 기반 진화 프레임워크로 소속함수를 자동으로 생성하여 최적화하는 방법인 MOGFIDS(multi-objective genetic fuzzy intrusion detection system) 방법[13]이 제안되었다.

4. 데이터 분할 평가 진화알고리즘을 이용한 퍼지 분류규칙 생성

제안한 방법은 <그림 5>와 같이 클래스 분포에 따라 초기 소속함수를 생성하고 정확성이 높은 간결한 퍼지 분류규

칙을 생성할 수 있도록 소속함수를 진화알고리즘으로 최적화시키는 방법이다. 퍼지 분류규칙은 진화되는 소속함수들을 이용하여 FDT[2]로 퍼지 의사결정트리를 생성한 다음 추출한다. 또한 진화알고리즘의 개체평가 시간을 단축시키기 위해 전체 학습 데이터를 n 개의 부분 학습 데이터로 분할하여 개체를 평가하는 데이터 분할 평가 진화 방법을 제안한다.

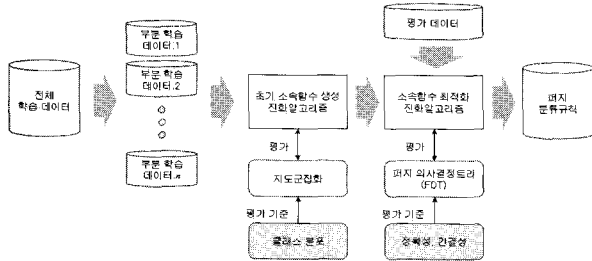


그림 5. 퍼지 분류규칙 생성 과정

Fig. 5. Process of generating fuzzy classification rules

4.1 초기 소속함수 생성

퍼지 분류규칙을 생성할 때 소속함수의 모양과 개수를 먼저 결정해야 한다. 소속함수를 적게 사용하면 생성되는 퍼지 분류규칙의 정확성이 떨어지고, 너무 많이 사용하면 생성되는 퍼지 분류규칙이 많아져 간결성이 떨어진다. 본 논문에서는 지도군집화(supervised clustering)로 클래스의 분포를 고려하여 적절한 소속함수의 개수를 결정한다.

지도군집화는 학습데이터의 클래스를 사용하여 군집을 생성한다. 본 논문에서 사용하고 있는 지도군집화는 k-means와 수행과정은 같고 단지, 군집 중심에 클래스를 정의하고 데이터의 클래스와 군집 중심의 클래스가 같은 경우에만 유사도를 계산하여 데이터를 가장 가까운 군집에 할당한다. 군집 개수는 진화알고리즘을 이용하여 결정한다.

군집 개수를 결정하기 위한 진화알고리즘에서 개체의 유전인자를 군집의 중심좌표로 인코딩하여 유전인자 개수가 군집의 개수를 나타낼 수 있도록 실수형 가변 길이로 개체를 설계한다. 초기 개체집단에서 개체의 길이는 $[[CLASS], 2 \times [CLASS]]$ 범위에서 임의로 결정하고 유전인자는 클래스마다 임의로 선택된 데이터로 인코딩한다. CLASS는 클래스 집합을 나타낸다.

개체평가 단계에서는 개체에 표현된 군집의 중심과 개수가 클래스 분포를 잘 반영할 수 있는지 지도군집화를 통해 평가한다. 이때 군집의 개수가 초기 소속함수의 개수를 결정하므로 가능한 한 적은 군집의 개수로 클래스 분포를 잘 반영할 수 있도록 식 (2)와 같이 적합도 함수를 정의한다.

$$CFit(C) = vHom(C) - (1-v)CcpX(C), 0 \leq v \leq 1 \quad (2)$$

적합도 함수는 군집의 동질성(homogeneity)과 군집화 복잡도(clustering complexity) 항목으로 나타낸다. 동질성(Hom(C))은 식 (3)과 같이 군집의 평균 순도로 계산하고 군집의 순도는 식 (4)와 같이 군집 내에서 군집 중심의 클래스와 동일한 클래스의 데이터 비율로 나타낸다. 군집화 복잡도(CcpX(C))는 군집 개수를 학습데이터 개수로 나눈 것(군집이 얼마나 많이 생성되는가)을 나타내는 군집화율(clustering rate)로 식 (2)에서 페널티로 작용한다.

$$Hom(C) = \frac{\sum_{i=1}^{|C|} Pur(c_i)}{|C|}, c_i \in C \quad (3)$$

$$Pur(c_i) = \frac{\sum_{m \in N(k_i)} \mu_{im}}{\sum_{l \in N} \mu_{il}} \quad (4)$$

$$CcpX(C) = \frac{|C|}{|M|} \quad (5)$$

N, C 는 각각 전체 학습 데이터 집합, 생성된 군집의 집합을 나타낸다. k_i 는 군집 i 의 중심 c_i 의 클래스이고, $N(k_i)$ 는 k_i 와 동일한 클래스를 갖는 데이터 집합이다. μ_{il} 는 데이터 l 이 군집 i 에 속하는 소속정도를 의미한다. v 는 적합도 함수에서 동질성과 군집화율의 비중을 제어하는 가중치이다.

부모 개체에 진화연산인 클래스별 한 점 교배 연산과 가우시안 돌연변이[10] 연산을 적용하여 자식 개체를 생성한다. 클래스별 한 점 교배연산은 각각 클래스별로 임의의 한 점을 선택하고 클래스별로 선택된 점을 기준으로 교배한다. 이 교배연산은 클래스 별로 군집의 개수를 다양화하는 효과를 준다.

진화된 군집의 중심좌표를 각 차원에 투영하여 삼각형 소속함수의 가운데 점(소속정도 1.0)으로 설정하고, 삼각형 소속함수의 좌우 끝점은 임의로 선택하여 설정하되 이웃하는 소속함수의 경우 중간 부분이 서로 겹치게 끝점을 선택하여 퍼지경계를 형성하도록 한다. 따라서 각 차원(속성)에서 생성될 수 있는 소속함수의 최대 개수는 생성된 군집의 개수와 같다.

4.2 소속함수 최적화

클래스 분포에 따라 생성된 초기 소속함수로부터 퍼지의 사결정트리를 생성하고, 생성된 퍼지 의사결정트리로부터 퍼지 분류규칙을 추출한다. 이 때 정확성과 간결성이 높은 퍼지 분류규칙을 생성하기 위해 진화알고리즘을 사용하여 소속함수를 최적화시킨다.

소속함수의 최적화를 위한 진화알고리즘에서 개체는 <그림 6>과 같이 그 길이가 데이터 속성 개수와 같다. 각각의 유전인자는 속성을 퍼지화한 소속함수의 집합을 나타내고, 트리플(triple)로 표현된 삼각형 소속함수의 좌표를 집합의 원소로 정의한다. 이와 같이 삼각형 소속함수의 좌표값을 개체로 인코딩함으로써 보다 다양한 모양의 소속함수를 표현할 수 있다.

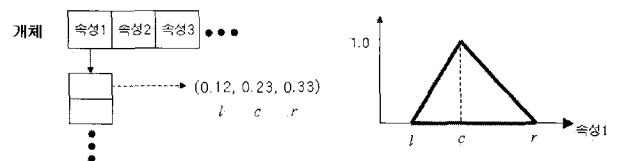


그림 6. 개체 인코딩

Fig. 6. Individual encoding

초기 개체집단에서 모든 개체의 길이는 클래스 분포에 따라 생성된 각 속성의 소속함수 개수의 합과 같고, 모든 개체의 삼각형 소속함수에서 소속정도가 1.0인 중앙값(c)도 클래스 분포에 따라 생성된 좌표값과 같다. 단지, 삼각형 소속함수의

양 끝점인 l, r 의 좌표값만 다르다.

개체집단에 표현된 소속함수로부터 FDT를 이용하여 퍼지 의사결정트리를 생성하고 생성된 트리로부터 퍼지 분류규칙을 추출한다. 추출된 규칙이 간결하면서 정확성이 높은 규칙인지 개체평가 단계에서 적합도 함수를 통해 평가한다. 적합도 함수는 식 (6)과 같이 정의하고 간결성과 정확성의 비중을 나타내기 위해 가중치 w 를 고려한다. 정확성보다 간결성에 비중을 높게 주었을 때 (예를 들어 가중치 w 가 0.1일 때), 많은 규칙이 생성되면서 정확성이 낮은 경우가 발생되면 적합도에 음수가 발생할 수 있으므로 식 (6)에 1을 더한다. 제안한 방법은 엘리트(elitist) 방법과 룰렛휠(roulette wheel) 방법으로 다음 세대의 개체들을 선택하기 때문에, 적합도는 반드시 0보다 같거나 커야한다.

$$TFIt_i = 1 + wAcc(\tau_i) - (1 - w)Tcpx(\tau_i), 0 \leq w \leq 1 \quad (6)$$

$$Tcpx(\tau_i) = \frac{R(\tau_i)}{|M|} \quad (7)$$

τ_i 는 개체 i 에 대응하는 퍼지 의사결정트리를 의미하고, $Acc(\tau_i)$ 는 개체 i 에 의해 생성된 퍼지 의사결정트리의 정확성을 나타낸다. $Tcpx(\tau_i)$ 는 퍼지 의사결정트리 τ_i 의 복잡도(complexity)를 나타내고 식 (7)을 사용하여 계산한다. $R(\tau_i)$ 는 퍼지 의사결정트리 τ_i 로부터 생성된 규칙 개수를 나타내고 N 은 학습데이터 집합을 나타낸다.

가변길이 실수 표현으로 인코딩한 개체에 대해 진화연산은 가우시안 돌연변이[10]와 전체산술 교배(whole arithmetic crossover)[10], 휴리스틱 교배(heuristic crossover)[10] 연산을 사용하고 속성교체 교배(attribute change crossover) 연산을 정의한다. 속성교체 교배 연산은 부모 개체에서 동일한 속성을 표현하는 소속함수의 개수가 다를 경우 적용함으로써 자식 개체에서 부모 개체와 다른 소속함수 개수를 가질 수 있고 그 이외의 진화연산은 소속함수의 좌표값을 조정하는 역할을 한다.

4.3 데이터 분할 평가 진화

제안한 방법과 같이 진화알고리즘을 사용하여 모델을 최적화할 경우 학습 데이터 양은 수행시간과 밀접한 관련이 있다. 모델을 최적화하기 위한 진화알고리즘에서 개체평가 과정의 수행시간은 모델 생성 시간과 모델 평가 시간으로 구분된다.

데이터 분할 평가 진화 방법은 개체평가 단계에서 학습 데이터 전체를 사용하는 대신 n 개로 나누어진 부분 학습 데이터를 이용하여 개체를 평가함으로써 수행시간을 단축시키는 방법이다.

만약 크기가 M 인 전체 학습 데이터 집합(S)로부터 시간 복잡도가 $O(M)$ 인 모델 생성 방법으로 모델을 최적화한다고 가정하면, 전체 학습 데이터 대신 크기가 m 인 부분 학습 데이터 집합(s)을 이용할 경우 시간 복잡도는 $\frac{m}{M}$ 만큼 단축된

$O(m)$ 이 된다. 그러나 부분 학습 데이터 집합의 크기를 작게 할수록 수행시간은 단축되지만 전체 학습 데이터의 클래스 분포를 나타내기 어렵기 때문에 왜곡된 모델로 진화될 수 있다.

<그림 7>은 모델 최적화를 위한 진화알고리즘에서 데이터 분할 평가 진화 방법이 적용된 알고리즘이다. S 는 모델을 생성하기 위한 전체 학습 데이터이고 S_{test} 는 모델을 평가

하기 위한 평가 데이터이다. 먼저 진화하기 전에 전체 학습 데이터 S 를 n 개의 부분 학습 데이터(s_1, s_2, \dots, s_n)로 분할한다. 개체평가 단계에서는 개체마다 부분 학습 데이터 중에 한 개를 임의로 선택하여 개체에 인코딩된 정보로 모델을 생성한 후, 생성된 모델(π)을 평가 데이터(S_{test})로 평가하여 적합도를 계산한다. pop_size 는 개체집단의 크기를 나타낸다.

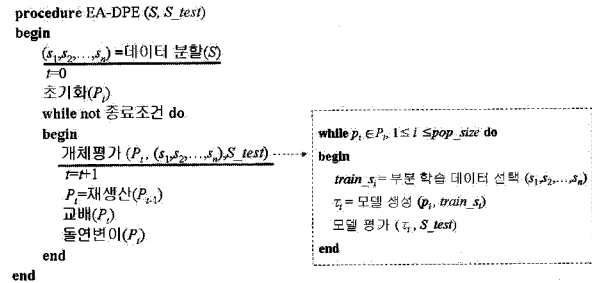


그림 7. 데이터 분할 평가 진화알고리즘

Fig. 7. Evolutionary algorithm with data partition evaluation

5. 실험

제안한 퍼지 분류규칙 생성 방법의 타당성을 검증하기 위해 UCI[14]에서 제공되는 벤치마크 데이터와 1999년도 KDD(knowledge discovery and data-mining) Cup[15]에서 제공되었던 침입탐지 데이터를 이용하여 기존 방법과 비교 평가하였다. 비교 평가 척도로 생성된 퍼지 분류규칙의 정확성과 간결성을 고려하였다. 정확성은 평가 데이터 중에 정확하게 분류한 백분율인 인식률로 계산하였고, 간결성은 생성된 규칙 개수와 조건항 개수로 나타내었다.

5.1 벤치마크 데이터를 이용한 기존 방법과 비교 실험

표 1. 벤치마크 데이터

Table 1. Benchmark data sets

데이터	데이터 개수	속성 개수	클래스 개수	클래스 별 데이터 개수
iris	150	4	3	(50,50,50)
pima	768	8	2	(500,268)
thyroid	215	5	3	(150,35,30)
ionosphere	351	34	2	(126,225)
glass	214	9	6	(70,76,17,13,9,29)
BCW1)	683	9	2	(444,239)
wine	178	13	3	(59,71,48)

본 실험에서는 [3,11]에서 사용되었던 UCI 벤치마크 데이터들을 이용하여 기존 방법과 비교하였다. 사용된 데이터에 대한 세부 사항은 <표 1>과 같다.

본 실험에서 진화알고리즘의 기본 매개변수인 개체집단의

1) Wisconsin breast cancer

표 2, 기존 방법과 제안한 방법의 성능 비교
Table 2. Performance comparison of our method and exiting methods

데이터	규칙의 정확성(%)				규칙의 간결성(규칙 개수 / 조건항 개수)				비교 (군집 개수/ θ_a/θ_e)
	[3]	[11]	제안한 방법		[3]	[11]	제안한 방법		
			w=0.1	w=0.9			w=0.1	w=0.9	
iris	97.0	96.1	98.0	98.0	9.7/16.0	3.0/4.0	3.0/4.0	3.0/4.0	3.0/1.0/0.02
pima	80.0	73.1	75.5	78.5	34.7/35.2	11.2/40.0	2.0/3.0	10.0/36.0	3.0/0.8/0.02
thyroid	86.0	-	93.5	96.7	7.8/15.2	-	3.0/3.0	7.0/15.0	3.0/0.9/0.02
ionosphere	-	86.4	89.7	97.2	-	3.4/10.2	3.0/6.0	26.0/80.0	3.0/0.9/0.02
glass	-	66.0	74.3	88.3	-	19.2/90.8	20.0/36.0	57.0/159.0	7.0/0.7/0.02
BCW	-	96.8	97.7	97.8	-	2.0/3.0	3.0/9.0	6.0/19.0	2.0/1.0/0.02
wine	-	91.2	91.6	100.0	-	3.6/8.8	4.0/7.0	13.0/38.0	3.0/0.9/0.02

크기는 30, 교배 확률은 0.9, 돌연변이 확률은 0.9로 설정하였고, 5000세대까지 진화시켰다. 그 결과 <표 2>와 같다.

<표 2>에서 w는 식 (6)에서 간결성과 정확성의 비중을 나타내는 가중치이다. 가중치 값이 작을수록 간결성에 비중을 두어 생성되는 퍼지 분류규칙의 개수가 적어진다. θ_d 와 θ_e 는 FDT[2]로 퍼지 의사결정트리를 생성할 때 단말노드 조건인 임계값이다. θ_d (density)는 노드에 포함된 데이터의 소속정도 합에 대한 주 클래스의 소속정도 합의 비율을 의미한다. 주 클래스는 노드에서 소속정도 합이 가장 큰 클래스를 말한다. 본 실험에서는 기존 방법의 정확성을 고려해서 θ_d 를 설정하였다. θ_e (exception)는 전체 학습 데이터에 대한 노드에 포함된 데이터의 소속정도 합의 비율을 나타낸다. 군집 개수는 진화알고리즘을 이용한 지도 군집화로 클래스 분포에 따라 결정된 초기 소속함수의 최대 개수이다. 적절한 초기 소속함수의 개수를 결정하기 위한 초기 소속함수 생성 진화알고리즘에서 간결한 퍼지 분류규칙을 생성하기 위해 식 (2)의 가중치 v 을 0.1로 설정하였다.

[3]보다 약 3개적은 규칙으로 약 7% 향상된 정확성을 보였다.

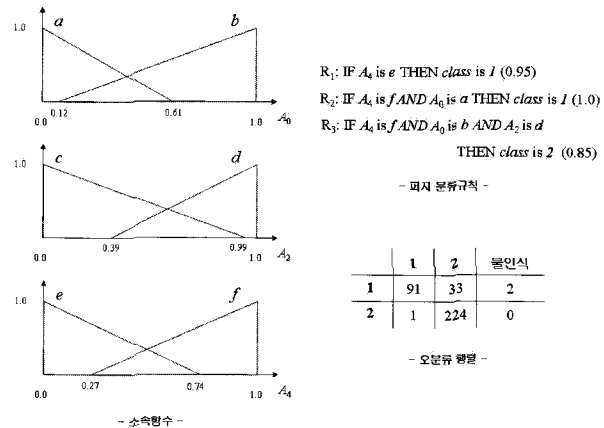


그림 9. Ionosphere 벤치마크 데이터 실험 결과

Fig. 9. Experimental result of the ionosphere data set

<그림 9>는 전리층에서 반사되는 17개의 레이더(radar) 펄스를 이용하여 레이더가 좋은지, 나쁜지를 분류하는 전리층(ionosphere) 데이터에 대한 실험 결과이다. 각 레이더 펄스는 두 개의 특성을 갖는다. 생성된 퍼지 분류규칙은 34개의 속성 중 세 개의 속성(A_0, A_2, A_4)만으로 구성되었으며 [11]보다 약 3.0% 높은 정확성을 보였다. 오분류 행렬(confusion matrix)에서 불인식은 생성된 퍼지 분류규칙으로 인식할 수 없는 경우를 말한다.

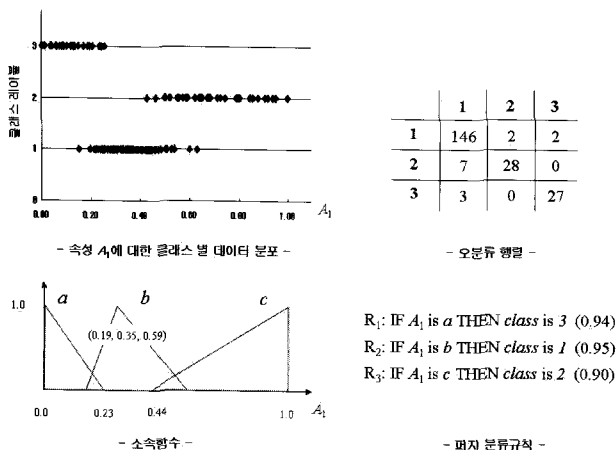


그림 8. Thyroid 벤치마크 데이터 실험 결과

Fig. 8. Experimental result of the thyroid data set

<그림 8>은 갑상선 질환 데이터(thyroid disease database)에 대한 실험 결과이다. 생성된 퍼지 분류규칙은 첫 번째 속성(A_1 : total serum thyroxin as measured by the isotopic displacement method)만 고려하여 세 개의 클래스(1:정상, 2: 갑상선 기능 항진증, 3: 갑상선 기능 저하증)를

5.2 침입탐지 데이터를 이용한 기존 방법과 비교 실험

본 실험은 1999년도 KDD Cup에서 제공되었던 침입탐지 데이터를 이용하여 기존 방법과 비교 평가하였고 대용량 데이터에서 제안한 데이터 분할 평가 진화 방법의 타당성을 확인하였다.

KDD'99 침입탐지 데이터는 침입탐지 연구의 평가를 목적으로 1998년 DARPA 침입탐지 평가 프로그램(intrusion detection evaluation program)에 의해 9주 동안 미국 군사 네트워크에서 시뮬레이션을 통해 제작되었다.

접속기록은 41개의 독립적인 속성과 클래스로 구성되어 있다. 접속기록의 클래스는 정상 또는 공격 유형으로 구분된다. 학습 데이터 집합과 평가 데이터 집합의 구성은 <표 3>

과 같다.

표 3. KDD'99 침입탐지 데이터의 구성

Table 3. Composition of KDD'99 intrusion detection data set

구분	학습 데이터	평가 데이터
속성 개수	41	41
클래스 개수	24	38
데이터 개수	494,021	311,029

<표 3>을 보면 학습 데이터와 평가 데이터의 클래스 개수가 다른 이유는 새로운 공격 유형으로 간주되는 14개의 클래스가 평가 데이터에 존재하기 때문이다. 본 실험에서는 KDD'99 Cup에서 제공한 클래스 범주화 스크립트를 사용하여 총 38개의 클래스를 대분류 5개('normal', 'probe', 'DOS', 'U2R', 'R2L')로 범주화 하였고, 41개 속성 중 범주형 속성을 제외한 38개의 연속형 속성만을 사용하였다.

인식률과 탐지비용, 오탐지율, 미탐지율의 항목으로 제안한 방법에 의해 생성된 침입탐지를 위한 퍼지 분류규칙의 정확성을 평가한다. 인식률은 전체 평가 데이터 311,029건 중에 정확히 분류한 백분율로 나타낸다.

표 4. KDD'99에서 사용한 침입 등급에 따른 탐지비용 행렬
Table 4. The cost matrix according to level of intrusion using in KDD'99

		예측 클래스				
		normal	probe	DOS	U2R	R2L
실 체 클 래 스	normal	0	1	2	2	2
	probe	1	0	2	2	2
	DOS	2	1	0	2	2
	U2R	3	2	2	0	2
	R2L	4	2	2	2	0

탐지비용은 <표 4>과 같이 행렬로 정의된 공격 유형에 대한 위험 등급에 따라 계산된 값이다. 이와 같이 탐지비용을 계산하기 위해 사용되는 행렬을 탐지비용 행렬(cost matrix)이라 하고 KDD'99 Cup에서 탐지비용을 계산하기 위해 정의된 것이다.

정상을 공격으로 예측하는 경우를 오탐지라 하고 공격을 정상으로 예측하는 경우를 미탐지라 한다. 각각을 식 (8), 식 (9)과 같이 백분율로 나타낸 것을 오탐지율(FPR: false positive ratio), 미탐지율(FNR: false negative ratio)이라 한다.

$$FPR = \frac{\text{정상 데이터 중 침입으로 예측한 개수}}{\text{정상 데이터 개수}} \times 100 \quad (8)$$

$$FNR = \frac{\text{침입 데이터 중 정상으로 예측한 개수}}{\text{침입 데이터 개수}} \times 100 \quad (9)$$

본 실험에서 진화알고리즘의 기본 매개변수인 개체집단 크기를 20으로 설정하였고, 교배확률은 0.4, 돌연변이 확률은 0.1로 설정하였다. 그리고 개체평가 단계에서 FDT의 단말노드 조건인 θ_d 와 θ_e 을 각각 0.85와 0.01로 설정하였다. 진화 종료 조건은 엘리트 개체의 인식률이 95%이상 이거나 엘리트 개체의 적합도가 30세대 동안 변하지 않을 경우에 종료하

고 최대 100세대까지 진화하도록 설정하였다.

실험은 학습 데이터와 평가 데이터에서 각각 3,000(normal(900), probe(411), DOS(1231), U2R(52), R2L(406))건을 클래스별 계통추출 방법으로 샘플링하여 진화 과정에서 사용하고, 진화 종료 후 311,029건의 전체 평가 데이터로 평가하였다. 실험은 랜덤 초기값을 0에서 9까지 변경하면서 10회 반복 수행하여 평균을 구했다.

<표 5>은 데이터 분할 평가 진화방법에서 분할 개수와 적합도 함수에서 가중치(w) 변화에 따른 침입탐지 성능 변화를 보여준다. 가중치가 증가함에 따라 인식률은 증가하고 규칙 개수와 조건항 개수가 증가됨을 확인할 수 있다. 또한 인식률이 높아지면 탐지비용과 미탐지율(FNR)이 감소되는 것은 가중치가 증가함에 따라 적합도 함수에서 정확도 항목의 비중을 크게 하여 진화하게 되고, 그 결과 평가 시 사용된 데이터에서 U2R과 같은 낮은 비율의 클래스를 분류할 수 있는 규칙이 생성되기 때문이다. 이는 UCI 벤치마크 데이터 실험에서 데이터에 포함된 클래스 비율의 차이가 큰 glass 데이터의 결과와 유사하다.

표 5. 데이터 분할 개수와 가중치 변화에 따른 성능 비교
Fig. 5. Performance comparison for varying the number of data subsets and weights

데이터 분할	가중치 (w)	정확성				규칙 개수 / 조건항 개수	개체 평가시간 (초)
		인식률 (%)	탐지 비용	FPR (%)	FNR (%)		
전체	0.1	92.04	0.2500	4.9	7.8	69.2/224.0	2.47
	0.5	92.96	0.2281	5.4	6.7	71.6/184.2	2.78
	0.9	94.25	0.1847	5.4	4.9	178.4/513.3	3.28
10 분할	0.1	92.12	0.2440	5.7	7.3	63.5/166.8	0.75
	0.5	93.09	0.2221	5.8	6.1	65.6/158.9	0.76
	0.9	93.54	0.1960	8.4	4.8	144.7/399.5	0.95

표에서 개체 평가시간 항목은 진화알고리즘 수행 시 개체평가 단계에서 하나의 개체를 평가하기 위해 소요되는 시간을 나타낸다. 가중치가 증가함에 따라 생성되는 규칙 개수가 많아지고 개체 평가시간이 증가하게 된다. 규칙 개수가 많아지는 것은 FDT로 퍼지 분류규칙을 생성하는데 소요되는 시간이 증가되는 것을 의미한다. 본 실험에서는 제안한 데이터 분할 평가 진화 방법을 적용하여 개체평가 시간이 약 70%감소하고 탐지 성능도 전체 학습 데이터를 사용한 것과 유사한 효과를 얻었다.

<그림 10>은 전체 학습 데이터를 사용하여 가중치 0.9, 랜덤 초기값 0으로 설정하여 진화된 소속함수이다. 이때 165개의 퍼지 분류규칙이 생성되었다.

각 속성의 명칭은 KDD'99 데이터의 고유 명칭이다. count는 2초 동안 동일한 호스트에서의 연결 횟수, src_bytes는 소스 호스트에서 대상 호스트로 보낸 패킷의 누적 용량, dst_host_srv_count는 대상 호스트에서 2초 동안 동일한 서비스의 연결 횟수, dst_host_same_srv_rate는 대상 호스트에서 동일한 서비스의 비율, srv_diff_host_rate는 소스 호스트에서 서로 다른 호스트에 연결한 비율을 나타낸다.

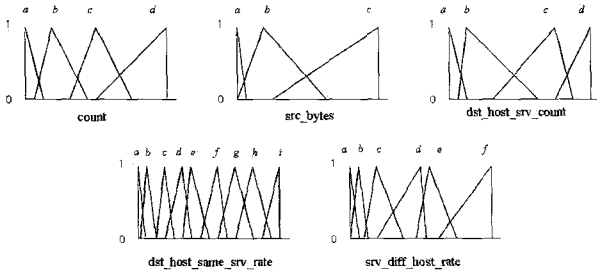


그림 10. 제안한 방법으로 진화된 소속함수

Fig.10. Membership functions evolved by our method

침입탐지 데이터에서 제안한 방법과 기존 방법과의 실험 결과는 <표 6>와 같다. 본 논문에서 클래스 분포를 고려하여 초기 소속함수를 생성하고 소속함수를 최적화 시키면서 퍼지 분류규칙을 생성한 제안한 방법이 임의로 초기 소속함수를 생성하여 규칙을 생성한 MOGFID보다 인식률이 1.48% 높아지고 탐지비용은 20.3% 절감되었다. 또한 KDD'99 Cup 우승자의 경우보다 인식률이 1.54% 높아지고 탐지비용은 20.8% 절감되었다.

표 6. 기존 방법과 제안한 방법의 성능 비교

Table 6. Performance comparison of our method with existing methods

구분	가중치(w)	정확성		간결성
		인식률 (%)	탐지비용	규칙 개수
MOGFIDS[13]	-	92.77	0.2317	148.0
KDD'99 Cup 우승자[16]	-	92.71	0.2331	-
제안한 방법 (전제)	0.1	92.04	0.2500	69.2
	0.5	92.96	0.2281	71.6
	0.9	94.25	0.1847	178.4
제안한 방법 (10 분할)	0.1	92.12	0.2440	63.5
	0.5	93.09	0.2221	65.6
	0.9	93.54	0.1960	5144.7

6. 결론 및 향후 연구

본 논문에서는 퍼지화에서 알갱이 문제를 해결하기 위해, 클래스 분포에 따라 초기 소속함수를 생성하고 정확하고 간결한 퍼지 분류규칙을 생성할 수 있도록 소속함수를 최적화 시키면서 규칙을 생성하는 방법을 제안하였다. 제안한 방법은 지도 군집화로 초기 소속함수를 생성하고 정확성이 높은 간결한 규칙을 생성할 수 있도록 소속함수를 진화알고리즘으로 진화시키면서 퍼지 분류규칙을 생성한다. 퍼지 분류규칙은 FDT를 이용하여 퍼지 의사결정트리를 생성하고 생성된 트리로부터 규칙들을 추출한다. 또한 진화알고리즘의 수행시간을 단축시키기 위해 개체평가 단계에서 전체 학습 데이터를 이용하는 대신 부분 학습 데이터로 개체를 진화시키는 데이터 분할 평가 진화 방법을 제안하였다.

UCI 벤치마크 데이터와 KDD'99 침입탐지 데이터를 이용하여 기존 방법과 비교한 결과, UCI 벤치마크 데이터에서는 적합도 함수에서 간결성 비중을 0.9로 고려하였을 경우 평균

적으로 규칙 개수는 약 5개가 적으며 정확성은 약 2% 향상되었고, KDD'99 침입탐지 데이터에서는 적합도 함수에서 정확성의 비중을 0.5 이상 주고 진화시킨 소속함수로 퍼지 분류규칙을 생성한 경우 기존 방법보다 정확성은 최대 1.54% 향상되었고 탐지비용은 최대 20.8% 감소되었다. 특히 데이터 분할 평가 진화 방법으로 평균 개체 평가시간이 전체 학습 데이터에서 부분 학습 데이터 비율에 비례해 단축되면서 정확성과 간결성은 전체 학습 데이터를 이용하였을 때와 유사한 결과를 보였다.

향후 연구로 진화알고리즘의 효율성을 보다 높이기 위해 진화 수렴 세대를 효과적으로 단축시킬 수 있는 휴리스틱 진화연산에 대한 연구가 요구된다. 또한 학습 데이터에서 상대적으로 빈도가 낮은 클래스에 대해서 효과적으로 분류할 수 있는 방법에 대한 연구가 필요하다. 마지막으로 제안한 방법을 로봇 분야 또는 센서 네트워크 분야에서 다양한 센서들로부터 상황인식을 위한 규칙 생성에 적용함으로써 보다 일반화된 타당성을 검증한다.

참 고 문 헌

- [1] Quinlan, J.R., "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, 4, pp.77-90, 1996
- [2] M. W. Kim, J. W. Ryu, "Optimized Fuzzy Classification for Data Mining," *Lecture Notes in Computer Science* Vol. 2973, pp.582-593, 2004
- [3] X.-Z. Wang, D.S. Yeung, E.C.C. Tsang, "A Comparative Study on Heuristic Algorithms for Generating Fuzzy Decision Trees," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 31, NO. 2, pp. 215-226, 2001
- [4] Janikow, C.Z., Fajfer, M., "Fuzzy partitioning with FID3.1," *18th International Conference of the North American, NAFIPS*, pp.467-471, 1999
- [5] Tomoharu Nakashima, Gaku Nakai, Hisao Ishibuchi, "Improving the Performance of Fuzzy Classification Systems by Membership Function Learning and Feature Selection," *FUZZ-IEEE '02 Proceedings of the 2002 International Conference on* Vol.1, pp.488-493, 2002
- [6] Tomoharu Nakashima, Gerald Schaefer, Yasuyuki Yokota, Hisao Ishibuchi, "A weighted fuzzy classifier and its application to image processing tasks," *Fuzzy Sets and Systems* 158, pp.284-294, 2007
- [7] Hisao Ishibuchi, Takashi Yamamoto, "Effects of Three-Objective Genetic Rule Selection on the Generalization Ability of Fuzzy Rule-Based Systems," *Lecture Notes in Computer Science*, Vol. 2632, pp.608-622, 2003
- [8] Hisao Ishibuchi, "Evolutionary Multiobjective Design of Fuzzy Rule-Based Systems," *Proceeding of the 2007 IEEE Symposium on Foundations of Computational Intelligence*, pp.9-16, 2007
- [9] Hisao Ishibuchi, Yusuke Nojima, "Analysis of interpretability-accuracy tradeoff of fuzzy systems

by multiobjective fuzzy genetics-based machine learning," *International Journal of Approximate Reasoning* 44, pp.4-31, 2007

- [10] J. Roubos, M. Setnes, J. Abonyi, "Learning Fuzzy Classification Rules from Labeled Data," *International Journal of Information Sciences*, 150(1-2), pp.77-93, 2003
- [11] J. Abonyi, J. Roubos, F. Szeifert, "Data-driven generation of compact, accurate, and linguistically sound fuzzy classifiers based on a decision-tree initialization," *International Journal of Approximate Reasoning*, 31(1), pp.1-21, 2003
- [12] J. Gomez, D. Dasgupta, "Evolving fuzzy classifiers for intrusion detection," *International Proceedings of the IEEE Workshop on Information Assurance*, 2002
- [13] Chi-Ho Tsang, S. Kwong and H. Wang, "Anomaly intrusion detection using multi-objective genetic fuzzy system and agent-based evolutionary computation framework," *International Conference on Data Mining*, pp.789-792, 2005
- [14] <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [15] KDD Cup 1999 data set, <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>, The UCI KDD Archive, University of California
- [16] C. Elkan, "Results of the KDD'99 classifier learning," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.63-64, 2000

저 자 소 개



류정우(Joung Woo Ryu)
 1998년 숭실대학교 컴퓨터학부 졸업.
 2002년 숭실대학교 컴퓨터학과 졸업 (석사)
 2005년 숭실대학교 컴퓨터학과 졸업 (박사)
 2005년~현재 한국전자통신연구원 지능형 로봇연구단 연구원

관심분야 : 기계학습, 데이터마이닝, soft computing
 Phone : 042-860-1143
 Fax : 042-860-1566
 E-mail : ryu0914@etri.re.kr



김성은(Sungeun Kim)
 2003년 동양대학교 소프트웨어학과 졸업
 2006년 숭실대학교 일반대학원 컴퓨터학과 졸업 (석사)
 2006년~현재 (주)퓨터시스템 정보통신연구소 주임연구원

관심분야 : soft computing, 데이터마이닝, 침입탐지
 Phone : 02-6220-7827
 E-mail : babystep@naver.com



김명원(Myung Won Kim)
 1972년 서울대학교 응용수학과 졸업
 1981년 University of Massachusetts (Amherst) Computer Science (석사)
 1986년 University of Texas (Austin) Computer Science (박사)

1975년~1978년 한국과학기술연구소 연구원
 1982년~1985년 Institute for Computing Science & Computer Application (Univ. of Texas) 연구원
 1985년~1987년 AT&T Bell Labs. (Naperville) 연구원
 1987년~1994년 한국전자통신연구소 책임연구원
 2000년~2001년 IBM T.J Watson 연구소 방문과학자
 1994년~현재 숭실대학교 컴퓨터학부 교수.

관심분야 : soft computing, 패턴인식, 자동추론, 기계학습, 데이터마이닝, creativity engineering 등
 Phone : 02-820-0923
 Fax : 02-813-8755
 E-mail : mkim@ssu.ac.kr