論文98-35C-10-7

퍼지 결정 트리를 이용한 효율적인 퍼지 규칙 생성

(Efficient Fuzzy Rule Generation Using Fuzzy Decision Tree)

閔 昌 宇 *, 金 明 源 *, 金 修 光 **

(Changwoo Min, Myung Won Kim, and Soo Kwang Kim)

요 약

데이터 마이닝의 목적은 유용한 패턴을 찾음으로써 데이터를 이해하는데 있으므로, 찾아진 패턴은 정확할 뿐 아니라 이해하기 쉬워야한다. 따라서 정확하고 이해하기 쉬운 패턴을 추출하는 데이터 마이닝에 대한 연구가 필요하다. 본 논문에서는 퍼지 결정 트리를 이용한 효과적인 데이터 마이닝 알고리즘을 제안한다. 제안된 알고리즘은 ID3, C4.5와 같은 결정 트리 알고리즘의 이해하기 쉬운 장점과 퍼지의 표현력을 결합하여 간결하고 이해하기 쉬운 규칙을 생성한다. 제안된 알고리즘은 히스토그램에 기반하여 퍼지 소속함수를 생성하는 단계와 생성된 소속 함수를 이용하여 퍼지 결정 트리를 구성하는 두 단계로 이루어진다. 또한 제안된 방법의 타당성을 검증하기 위하여 표준적인 패턴 분류 벤치마크 데이터인 Iris 데이터와 Wisconsin Breast Cancer 데이터에 대한 실험 결과를 보인다.

Abstract

The goal of data mining is to develop the automatic and intelligent tools and technologies that can find useful knowledge from databases. To meet this goal, we propose an efficient data mining algorithm based on the fuzzy decision tree. The proposed method combines comprehensibility of decision tree such as ID3 and C4.5 and representation power of fuzzy set theory. So, it can generate simple and comprehensive rules describing data. The proposed algorithm consists of two stages: the first stage generates the fuzzy membership functions using histogram analysis, and the second stage constructs a fuzzy decision tree using the fuzzy membership functions. From the testing of the proposed algorithm on the IRIS data and the Wisconsin Breast Cancer data, we found that the proposed method can generate a set of fuzzy rules from data efficiently.

I. 서 론

컴퓨터의 사용이 일반화됨에 따라 데이터를 생성하고 수집하는 것이 용이해졌다. 인간 게놈 데이터베이스 프로젝트(human genome database project)에서는 수십 기가 바이트(gigabyte)에 달하는 인간의 유전정보에 관한 데이터베이스가 작성되고 있고, NASA의 EOS(Earth Observing System)에서는 한 시간에 50 기가 바이트 이상의 지구에 대한 데이터가 수집되고 있다. 또한 상품거래에 있어서 바코드를 사용함으로써 상품거래에 대한 많은 데이터가 쉽게 축적되고 있다. 또한 최근에는 인터넷을 기반으로 한 전자상거래의 활성화로 상품거래에 대한 데이터 축적이 보다 쉬워졌다. 그 외에도 경제와 산업의 많은 부분에서

* 正會員, 崇實大學校 컴퓨터學部
(School of computing, Soongsil University)
** 正會員, 浦港 綜合 製鐵 技術研究所 計測制御研究팀
(Instrumentation and control department, Pohang Iron & Steel Co.)

接受日字:1997年12月3日, 수정완료일:1998年9月4日

컴퓨터 사용의 보편화로 많은 양의 데이터가 손쉽게 축적되고 있다. 또한 DBMS(database management system), 데이터 웨어하우징(data warehousing)등의 기술의 발달로 대용량의 데이터를 빠르고 손쉽게 저장, 검색할 수 있게 되었다[1].

그러나 이러한 원시 데이터는 분석됨으로써 중요한 지식이 추출되어 인간의 의사결정에 도움이 될 때에만 의미가 있다. 그러나 위와 같이 방대한 양의 데이터를 수작업으로 분석하여 중요한 지식을 발견하는 데에는 한계가 있다. 따라서 방대한 양의 데이터를 지능적으로 자동 분석하여 중요한 지식을 추출할 수 있는 도구 및 방법이 절실히 요구된다. KDD(Knowledge Discovery in Databases)는 이러한 요구로부터 시작된 분야로서 데이터로부터 유용하고 궁극적으로 이해 가능한 패턴을 찾아내는 과정이다. 데이터로부터 유용하고 이해 가능한 패턴을 추출하기 위해서 KDD는 다음과 같은 다섯 단계로 나눌 수 있다.

(a)자료 선택(selection)은 적용하는 분야와 관련된 사전 지식에 관하여 이해하고 사용자의 관점에서 지식 발견의 최종 목표를 설정하고, 하나의 데이터 집합을 선택하거나, 데이터 집합 중 일부만 선택하거나 변수의 일부만 선택하여 목표 데이터 집합을 만든다.

(b)전처리(preprocessing)단계에서는 자료에 포함되어 있는 노이즈(noise)를 제거하고 데이터 값이 없는 경우 어떻게 처리할 지를 결정한다.

(c)자료 변환(transformation)은 목표에 따라 데이터를 잘 나타내는 변수를 찾고 고려할 변수의 실제 개수를 줄이거나 데이터를 나타내는데 무관한 변수를 찾아 제거하거나 변환시킨다.

(d)데이터 마이닝 (data mining)은 KDD에서 가장 중요한 단계로 지식 추출의 목적에 따라 적절한 데이터 마이닝 알고리즘을 적용하여 지식을 추출하는 단계이다.

(e)해석/평가(interpretation/evaluation)단계에서는 데이터 마이닝을 통하여 얻어진 패턴을 해석한다. 또한 추출된 지식이 기존에 가지고 있던 지식과 상충되는지 비교해보고 상충되는 것을 해결한다.

이와 같이 데이터 마이닝 단계는 KDD에서 가장 중요한 단계로 KDD를 성공적으로 수행하는데 있어서 핵심적인 기술이다. KDD의 목적이 잘 알려지지 않은 시스템에 대한 지식을 추출하는 것이기 때문에 데이터 마이닝 알고리즘을 통하여 추출된 지식의 형태는 데이터를 잘 기술할 뿐 아니라 인간이 이해하기 쉬운 형태여야 한다.

이해의 용이성(comprehensibility)은 정량적으로 측정하기가 어렵다. 따라서 본 논문에서는 추출된 지식의 간결성(compactness)을 이해의 용의성의 측정 단위로 사용한다. 본 논문에서는 많은 양의 데이터로부터 효과적으로 이해하기 쉬운 형태의 지식을 추출하기 위한 데이터 마이닝 알고리즘으로서 퍼지 결정 트리를 이용한 방법을 제안한다. 본 논문에서 제안된 방법은 간결한 규칙을 생성하는 결정 트리 생성알고리즘인 ID3를 퍼지 결정 트리를 생성할 수 있도록 확장한 것으로 규칙의 형태, 생성되는 규칙의 수, 생성된 각 규칙의 조건부의 항 수가 간결한 것이 특징이다.

본 논문은 모두 7장으로 구성되어 있으며 그 내용은 다음과 같다. 2 장에서는 데이터 마이닝 알고리즘으로 사용되는 여러 기계 학습(machine learning) 알고리즘에 대하여 살펴보고, 3장에서는 본 논문에서 사용하는 퍼지 규칙의 형태 및 추론 방법에 대하여 살펴본다. 4장에서는 히스토그램에 의한 퍼지 소속 함수의 생성 방법에 대하여 기술하며, 5장에서는 퍼지 결정 트리 생성 알고리즘에 대하여 살펴본다. 6장에서는 제안된 알고리즘을 Iris 데이터와 Wisconsin Breast Cancer 데이터에 대하여 실험한 결과를 보이고 7장에서는 결론 및 향후 연구 방향에 대하여 검토한다.

Ⅱ. 관련 연구

많은 기계학습 알고리즘들이 데이터 마이닝 알고리즘으로 사용되고 있다. 기계 학습은 크게 연역적 학습(deductive learning)과 귀납적 학습(inductive learning)으로 분류할 수 있다. 연역적 학습은 주어진 영역 이론(domain theory)로부터 하나의 예제를 통하여 영역 이론보다 효율적인 규칙을 생성하는 것으로 EBL(Explanation-Based Learning)과 Soar에서 사용되는 청킹(chunking)이 있다. 귀납적 학습은 데이터로부터 데이터를 기술하는 모델을 생성하는 것으로서 대표적인 학습 방법으로는 ID3, C4.5와 같은 결정 트리 계열의 학습 방법이 있다[2, 3, 4]. 또한 오류 역전파 학습 알고리즘과 같은 신경망을 이용한 학습 알고리즘도 역시 귀납적 학습 알고리즘으로 볼 수 있다. 귀납적 기계학습은 데이터로부터 데이터를 기술하는 모델을 생성하는 것으로 데이터 마이닝 알고리즘으로

서 사용될 수 있다. 그러나 데이터 마이닝의 경우 그 목적이 데이터로부터 유용하고 이해 가능한 지식을 추출하는 것에 있으므로 데이터 마이닝 알고리즘으로 적합한 귀납적 기계학습 알고리즘은 데이터로부터 유용하고 이해 가능한 지식을 추출할 수 있어야 한다. 본 장에서는 대표적인 기계학습 방법인 결정 트리 생성 알고리즘과 퍼지 규칙 생성 방법에 대하여 살펴본다.

1. 결정 트리 생성에 의한 규칙 생성 방법

결정 트리는 널리 알려진 기호주의적 지식 습득 방법으로, 생성되는 지식이 규칙의 형태를 가지며 이해하기 쉽다는 장점이 있다. 결정 트리는 노드와 링크로 구성된 일반적인 트리로서, 각각의 노드는 데이터들이 표현된 특징(feature)중의 하나가 되며, 링크는 그 특징이 가질 수 있는 값이 된다. 결정 트리에 생성에 의한 규칙 생성 방법은 데이터로부터 그 데이터들이 암시적으로 포함하고 있는 개념을 추출하여 결정 트리의 형태로 일반화하고 이를 이용하여 새로운 데이터를 분류하게 된다. 결정 트리는 곧바로 규칙으로 변환될 수 있기 때문에 결정 트리가 복잡할수록 생성되는 규칙의 수가 많고 각 규칙의 조건부가 길어지게 된다. 따라서 Quillan이 제안한 ID3와 C4.5에서는 주어진 패턴을 올바르게 분류할 수 있으면서 간단한 결정 트리를 구성하기 위해 정보 이론(information theory)에 따른 무질서도(entropy) 개념을 이용한다. 무질서도는 다른 종류의 패턴들이 섞여있는 정도를 정량적으로 나타내는 것으로, 결정 트리에서 어떤 노드에 속하는 데이터들이 여러 다른 클래스의 데이터가 많이 섞여 있을수록 무질서도가 높고, 반대로 단일한 클래스로 되어 있으면 무질서도가 낮게된다. ID3와 C4.5 알고리즘은 루트 노드(root node)에서부터 무질서도를 가장 낮게 할 수 있는 특징을 선택하여 트리를 확장함으로써 간단한 결정 트리를 구성한다[3, 4].

결정 트리에 의하여 생성되는 지식은 그 형태가 규칙이므로 이해하기 쉽다는 장점이 있으나 생성되는 규칙은 항상 특징 축에 평행하게 패턴 공간을 분할함으로써 복잡한 패턴을 분류하는데 어려움이 있다[1, 4].

그림 1은 패턴의 분포가 특징 축에 평행하지 않게 판단 경계선(decision boundary)을 형성하는 경우 결정 트리가 어떻게 근사하는가를 보여준다. 그림에서 볼 수 있듯이 특징 축에 평행하지 않은 판단 경계선을 여러 개의 규칙으로 근사하고 있음을 볼 수 있다. 이와 같이 패턴이 복잡하게 분포되어 있는 경우 결정 트리는 매우 복잡해지며 분류율이 나빠지게 되어 효율적이지 못하다는 것을 알 수 있다.

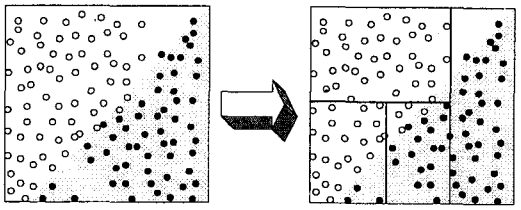


그림 1. 특징 축에 평행하지 않은 판단 경계선을 가지는 경우 결정 트리의 근사

Fig. 1. Approximation of nonaxis-parallel decision boundaries in the decision tree.

2. 퍼지 규칙 생성 방법

퍼지 이론은 이러한 한계를 극복할 수 있는 방법을 제시하고 있다. 퍼지 추론은 언어적 불확실성을 퍼지 집합의 개념을 이용하여 정량적으로 표현하고 추론할 수 있는 수단으로 제어 문제, 패턴 분류 문제 등 여러 분야에서 활용되고 있다. 또한 언어적 불확실성을 표현하는데 있어서 기호주의적 틀을 유지함으로서 이해가 쉽다는 장점이 있다. 다시 말하면 결정 트리에 의해서 생성된 규칙과 같이 언어적으로 쉽게 표현할 수 있으면서 규칙들이 이루는 판단 경계선은 특징 축에 평행하지 않을 수 있음으로서 패턴이 복잡하게 분포되어있는 경우에도 이해하기 쉬우면서 분류율이 높은 규칙의 생성이 가능하다.

퍼지 규칙을 생성하기 위한 방법은 패턴 공간의 분할을 통하여 퍼지 규칙을 자동으로 생성하고자 하는 퍼지 공간 분할 방법[5, 6], 신경망의 학습성을 퍼지와 결합하고자 하는 퍼지 신경망[7, 8]등이 있다. Ishibuchi[5]는 계층적 퍼지 공간 분할 방법으로 오류가 있는 부공간(subspace)들에 대하여 동일한 크기의 사분면(quadrant)으로 재귀적으로 분할한다. 이때 생성된 퍼지 규칙의 형태는 사진트리(quadtree)형태가 된다. Abe[6]는 퍼지 규칙의 개수와 분류율을 높이기 위해서 활성화 영역(activation region)과 억제성 영역(inhibition region)에 의한 규칙 생성을 방법을 제안하였다. 초기에 각 클래스 별로 클래스 별 데이터를 모두 포함하는 최소 사각형인 활성화 영역을 정의하고 활성화 영역이 서로 중복되는 영역인 억제성 영역에 대하여 재귀적으로 활성화 영역을 정의함으로써 퍼지 규칙을 생성한다. 퍼지 공간 분할 방법은 알고리즘이

간단하다는 장점이 있지만 생성되는 퍼지 규칙의 수가 매우 많고 규칙이 지나치게 세분화되기 때문에 생성된 규칙을 이해하기가 어렵다. Rhee[7]는 입력층의 활성 함수(activation function)을 퍼지 소속 함수로 하고 은닉 층의 활성 함수를 퍼지 연산자(fuzzy operator)로하는 전향 신경망(feed forward network) 형태의 퍼지 신경망을 제안하였다. 또한 여기에서는 제안한 퍼지 신경망을 학습 시키기 위하여 경사 하강법(gradient descent)에 기반한 학습 방법을 사용하였다[7]. Jang[8]은 퍼지 연상 기억 장치(FAM: Fuzzy Associative memory)에 기반한 학습 방법으로 히스토그램에 기반하여 퍼지 소속 함수를 생성하고 입력층의 퍼지 소속 함수와 출력층의 퍼지 소속 함수와의 관계를 퍼지 Hebb 학습 규칙을 이용하여 학습 함으로써 퍼지 규칙을 생성한다. Rhee[7]와 Jang[8]이 제안한 퍼지 신경망의 경우 대체로 퍼지 공간 분할 방법에 비하여 생성되는 규칙의 수가 적지만 퍼지 규칙의 형태가 조건부에 가중치가 결합된 형태로 역시 생성된 규칙을 이해하는 것이 어렵다.

따라서 결정 트리의 이해하기 쉬운 장점과 퍼지의 표현력을 결합하여 간결한 퍼지 규칙을 생성하기 위한 연구가 진행되고 있다[9]. 그러나 Umano[9]가 제안한 퍼지 결정 트리 생성알고리즘의 경우 이해하기 쉬운 퍼지 규칙을 생성하기 위하여 사용자가 퍼지 소속 함수를 정의해주어야하는 단점이 있다. 그러나 데이터의 양이 많고 데이터에 대한 사전 지식이 적은 경우에 소속 함수를 결정하는 것은 매우 어렵다. 따라서 본 논문에서는 히스토그램에 의하여 자동적으로 소속 함수를 생성하고 생성된 소속 함수를 이용하여 퍼지 결정 트리를 생성함으로서 분류율이 높으면서 이해하기 쉬운 규칙을 생성할 수 있는 데이터 마이닝 알고리즘을 제안한다.

Ⅲ. 퍼지 규칙의 형태 및 추론 방법

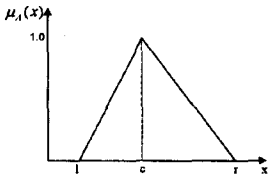
본 논문에서는 사람의 이해가 쉽도록 단순한 형태의 퍼지 규칙 및 추론 방법을 사용한다. 각 퍼지 규칙은 식 (1)과 같이 다수 개의 퍼지 집합으로 구성되어 있는 조건부와 1개의 결론부를 가지는 MISO(Multiple Input Single Output)형태의 규칙을 사용한다. 또한 각 규칙에는 규칙에 대한 CF(Certainty Factor)가 있다.

$$Rule_i: \text{ if } x_{i1} \text{ is } U_{i1} \text{ and } x_{i2} \text{ is } U_{i2} \ldots \text{ and } x_{in} \text{ is } U_{in} \text{ then class } j\,(CF_i) \tag{1}$$

클래스 i에 대한 추론 결과는 식 (2)와 같이 클래스 i를 결론부로 하는 규칙들 각각에 대하여 입력 조건에 대한 퍼지 소속값 중 가장 작은 값과 규칙의 CF를 곱한 값을 합한 것이다. 식 (2)에서 j는 결론이 클래스 i인 퍼지 규칙이다.

$$Concl_i = \sum_j \min_k(\mu_{U_{jk}}) * CF_j \tag{2}$$

최종 결론은 $Concl$ 이 가장 큰 클래스를 최종 결론으로 한다.



$$\mu_A(x) = \begin{cases} \frac{x-l}{c-l} & : \text{if } l \le x \le c \\ \frac{r-x}{r-c} & : \text{if } c \le x \le r \\ 0 & : \text{otherwise} \end{cases} \tag{3}$$

그림 2. 삼각형 형태의 퍼지 소속 함수
Fig. 2. Triangular fuzzy membership function.

Ⅳ. 히스토그램에 의한 퍼지 소속 함수의 생성

퍼지 소속 함수는 퍼지 규칙에서 가장 중요한 요소로 삼각형, 사다리꼴, 가우시안 함수 형태의 퍼지 소속 함수가 많이 사용된다. 본 논문에서는 각 클래스의 각 특징에 대한 히스토그램에서의 극대점(maxima)과 극소점(minima)를 이용하여 삼각형 형태의 퍼지 소속 함수를 생성한다. 히스토그램은 데이터 분포에 대한 통계적 특성을 나타내고 있음으로 히스토그램을 이용하여 소속 함수를 생성할 경우 효율적인 퍼지 결정 트리를 생성할 수 있다. 삼각형 형태의 소속 함수는 식 (3)과 같이 정의된다.

데이터의 수가 적은 경우 일반적으로 히스토그램에는 많은 수의 극대점과 극소점이 있게 된다. 따라서 퍼지 소속 함수를 생성하기 위해 먼저 각 클래스의 각 특징에 대한 히스토그램을 작성하고 N_s번 평활화(smoothing)함으로써 작은 극대점과 극소점을 없앤다. 히스토그램을 평활화하기 위해서는 이동 평균법(moving average)을 이용하였다. 또한 본 논문에서는 보다 구분력 있는 퍼지 소속 함수를 생성하기 위하

여 각 특징의 각 클래스별로 히스토그램을 생성하였다. 이 방법은 여러 개의 클래스가 하나의 군집을 이루고 있는 경우에도 구분력 있는 퍼지 소속 함수를 생성할 수 있게한다.

두 번째 단계로 평활화된 히스토그램에서 극대점과 극소점을 찾는다. 본 논문에서는 히스토그램에 대한 미분값을 이용하여 극대점과 극소점을 찾는다. x에 대한 히스토그램을 $h(x)$라고 할 때 $h(x)$의 일차 도함수는 식 (4)와 같이 정의된다.

$$\frac{dh(x)}{dx} = h(x) - h(x-1) \tag{4}$$

히스토그램에 대한 미분값이 양수에서 음수로 변하는 지점을 극대점으로 정의하며, 미분값이 양수, 영, 음수로 변할 때는 미분값이 영인 부분의 중간을 극대점으로 정의한다. 반대로 히스토그램에 대한 미분값이 음수에서 양수로 변하는 지점을 극소점으로 정의하며, 미분값이 음수, 영, 양수로 변하는 경우 미분값이 영인 구간의 중간을 극소점으로 정의한다.

식 (5)는 본 논문에서 사용한 극대점에 대한 정의이고 식(6)은 극소점에 대한 정의이다.

$$\begin{cases} \text{if } \dfrac{dh(t)}{dt} > 0 \text{ and } \dfrac{dh(t+1)}{dt} < 0 \\ \text{then t point is maxima} \end{cases}$$
$$\begin{cases} \text{if } \dfrac{dh(t)}{dt} > 0 \text{ and } \dfrac{dh(t+1)}{dt} = \ldots = \dfrac{dh(t+n)}{dt} = 0\ (n \geq 1) \text{ and } \dfrac{dh(t+n+1)}{dt} < 0 \\ \text{then } \dfrac{n-1}{2} + t \text{ point is maxima} \end{cases} \tag{5}$$

$$\begin{cases} \text{if } \dfrac{dh(t)}{dt} < 0 \text{ and } \dfrac{dh(t+1)}{dt} > 0 \\ \text{then t point is minima} \end{cases}$$
$$\begin{cases} \text{if } \dfrac{dh(t)}{dt} < 0 \text{ and } \dfrac{dh(t+1)}{dt} = \ldots = \dfrac{dh(t+n)}{dt} = 0\ (n \geq 1) \text{ and } \dfrac{dh(t+n+1)}{dt} > 0 \\ \text{then } \dfrac{n-1}{2} + t \text{ point is minima} \end{cases} \tag{6}$$

세 번째 단계는 히스토그램의 극대점의 높이를 현재 히스토그램에서 높이가 가장 높은 극대점의 높이로 바꾼다. 극대점 보정은 서로 중복이 있는 소속 함수의 경우 극소점이 두 소속 함수의 교차점이 되도록 유지시켜주는 역할을 한다. 극대점 보정을 통하여 원래 히스토그램에서 극대점, 극소점과 같이 패턴 분류에서 중요한 특징이 삼각형 형태의 소속 함수에 반영되도록 하였다.

네 번째 단계는 각각의 극대점과 극대점의 양쪽 방향으로 가장 가까운 극소점을 직선으로 연결한다. 이 때 각 직선의 x 절편을 삼각형의 소속 함수의 l, r로 설정하고 마루를 c로 설정함으로써 삼각형 형태의 퍼지 소속 함수를 생성한다. 그림 3는 히스토그램을 이용하여 퍼지 소속 함수를 생성하는 과정을 보여준다.

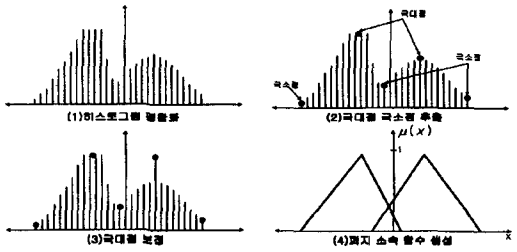


그림 3. 히스토그램에 의한 소속 함수의 생성
Fig. 3. Generating fuzzy membership functions based on histograms.

V. 퍼지 결정 트리 생성 알고리즘

퍼지 결정 트리는 그림 4과 같이 단말 노드(terminal node), 비단말 노드(nonterminal node), 호(arc)로 구성되어 있다. 비단말 노드는 분기를 위한 특징을 가지고 있으며, 단말 노드는 클래스 명과 CF를 가지고 있다. 노드와 노드 사이를 연결하는 호는 특징에 대한 퍼지 소속 함수이다.

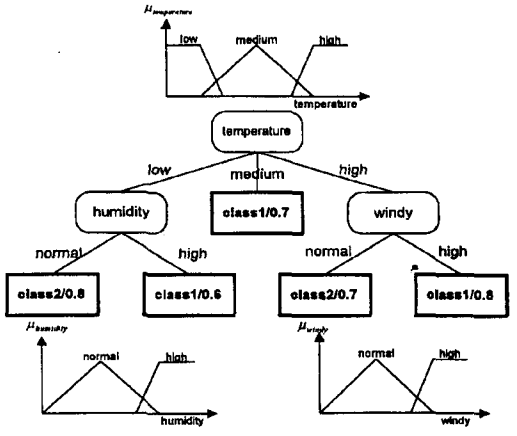


그림 4. 퍼지 결정 트리
Fig. 4. Fuzzy decision tree.

그림 4와 같은 형태의 퍼지 결정 트리를 생성하기 위한 알고리즘은 기존의 ID3에서 무질서도의 계산에

퍼지 개념을 결합함으로서 이루어진다. 결정 트리의 근 노드(root node)에서 i번째 노드까지의 퍼지 소속 함수가 규정하는 부공간에 대한 j번째 훈련 데이터의 소속 정도는 식 (7)과 같이 계산된다.

$$m_{ij} = \begin{cases} \min_{f \in FSet_i} (\mu_f(d_{jf})) & : FSet_i \neq \emptyset \\ 1 & : FSet_i = \emptyset \end{cases} \quad (7)$$

식 (7)에서 $FSet_i$는 결정 트리의 근 노드에서 i번째 노드까지의 퍼지 집합에 대한 집합이며 d_{jf}는 j번째 훈련 데이터에서 퍼지 소속 함수 f에 대한 특징값을 나타낸다.

ID3나 C4.5와 같은 기존의 결정 트리 생성 방법은 무질서도를 계산하는데 있어서 데이터가 각 부공간에 속하는 정도가 0과 1의 두 값만을 가지나, 퍼지 결정 트리에 있어서는 데이터의 부공간에 대한 소속 정도가 식 (7)에서 정의한 바와 같이 0에서 1사이의 연속적인 값을 가진다. 따라서 퍼지 결정 트리에서 무질서도는 데이터가 부공간에 속하는 정도를 위에서 정의한 m_{ij}를 이용함으로써 계산한다. i번째 노드에서 특징 F에 대한 무질서도 E_F^i는 다음과 같다.

$$E_F^i = \sum_j (p_{ij} I^j) \quad (8)$$

$$I^j = - \sum_{\text{for each class } k} (p_k^j \log_2 p_k^j) \quad (9)$$

$$p_k^j = \frac{\sum_{m \in \text{class } k} m_{jm}}{\sum_{\text{for all training data } l} m_{jl}} \quad (10)$$

$$p_{ij} = \frac{\sum_k m_{jk}}{\sum_k m_{ik}} \quad (11)$$

식 (8)에서 f는 특징 F의 퍼지 소속 함수를 가리키며, j는 i번째 노드에서 퍼지 소속 함수 f를 적용하여 생성된 자식 노드를 나타낸다.

다음은 퍼지 결정 트리 생성을 위한 알고리즘이다.

<단계 1> 아래의 조건이 만족되면 자식 노드를 더 이상 확장하지 않고 결론부를 생성한다.

- $\sum_j m_{ij}$가 θ_n보다 작거나
- $\frac{\sum_{c \in \text{major class}} m_{ic}}{\sum_j m_{ij}}$가 θ_m 보다 크거나
- 더 이상 사용할 수 있는 특징이 없을 때

위에서 major class는 클래스별로 부공간에 대한 소속 정도를 합했을 때 소속 정도의 합이 가장 큰 클래스를 나타낸다.

위의 조건이 만족될 때 단말 노드에 대한 결론은 major class로 하고 CF는 $\frac{\sum_{c \in \text{major class}} m_{ic}}{\sum_j m_{ij}}$로 한다.

<단계 2> 그렇지 않으면,

(a) 무질서도 E가 가장 낮은 특징을 F_{best}라고 정의한다.

(b) F_{best}의 모든 퍼지 소속 함수에 대하여 자식 노드를 생성하고 각각의 자식 노드에 대하여 <step 1>부터 알고리즘을 재귀적으로 적용한다.

알고리즘의 <단계 1>에서 θ_n과 θ_m은 분할 종료 조건을 나타내는 파라미터이다. 퍼지 집합이 규정하는 부공간에 대한 데이터의 소속 정도의 합이 θ_n보다 작을 때 분할을 종료한다. 이것은 퍼지 규칙이 과적응되는 것을 줄여준다. 또한 퍼지 집합이 규정하는 부공간에 대한 소속 정도의 합이 가장 큰 클래스와 모든 클래스의 소속 정도의 합의 비율이 θ_m보다 클 경우 분할을 종료한다. 이것도 역시 퍼지 규칙이 과적응되는 것을 줄여준다. 이와 같은 방법으로 퍼지 결정 트리를 생성한 후에는 퍼지 결정 트리에서 사용된 퍼지 소속 함수에 대하여 사용자에게 질의를 통하여 퍼지 소속 함수에 대한 어휘적 표현을 결정한다.

VI. 실험 결과

본 장에서는 본 논문에서 제안하는 퍼지 결정 트리 생성 알고리즘의 타당성을 검증하기 위해 패턴 분류 문제에 표준적으로 사용되는 벤치마크 데이터인 Iris 데이터와 Wisconsin Breast Cancer 데이터에 대하여 실험하였다. 실험에 사용된 Iris 데이터와 Wisconsin Breast Cancer 데이터는 UCI 기계 학습 데이터 베이스로부터 얻을 수 있다[10]. 구현을 위하여 사용된 컴퓨터는 Pentium-166이며 구현을 위해 사용된 언어는 Borland Delphi 2.0이다.

1. Iris 데이터에 대한 실험

Iris 데이터는 setosa(c1), versicolor(c2), virginica(c3)의 3개의 클래스로 구성되어 있는 데이터로 꽃받침(sepal)의 길이(f1)와 폭(f2), 꽃잎(petal)의 길이(f3)와 폭(f4)의 4개의 특징으로 기술된다. 각 클래

스는 50개씩의 데이터가 있어 총 150개의 데이터로 구성되어있다[10]. 퍼지 결정 트리를 구성하기 위한 데이터는 각 클래스 별로 같은 수의 데이터를 무작위로 추출하여 사용하였으며 시험 데이터는 150개의 데이터를 모두 사용하였다. 학습 자료의 수에 따른 성능을 평가하기 위해 훈련 데이터의 수를 21, 30, 60, 90으로 하여 실험하였다. 실험에서 N_s는 10, θ_n은 2, θ_m은 0.7로 하였다. 표 I은 훈련데이터의 수에 따라 생성되는 규칙의 수와 분류율을 Ishibuchi[5]와 Jang[8]과 비교한 것이다.

퍼지 규칙의 수/분류율

표 1. 각 방법에 의해 생성되는 퍼지 규칙의 수와 분류율

Table 1. Classification rate and the number of generated fuzzy rules for each algorithm.

학습 데이터 수	단순 퍼지 격자[5]	분산된 퍼지 규칙[5]	CF criterion[5]	NM criterion[5]	RM criterion[5]	Jang[8]	제안된 방법
21	455 /91.3%	8328 /92.4%	253 /91.1%	71 /89.8%	72 /89.6%	32 /88.3%	3 /91.3%
30	1727 /92.7%	20512 /93.8%	307 /91.8%	83 /93.0%	87 /93.3%	36 /91.6%	3 /95.3%
60	2452 /93.5%	63069 /95.4%	449 /93.8%	105 /93.9%	107 /94.1%	46 /93.3%	3 /96.0%
90	3440 /94.5%	140498 /95.8%	528 /95.1%	150 /94.8%	150 /94.6%	46 /95.0%	3 /96.0%

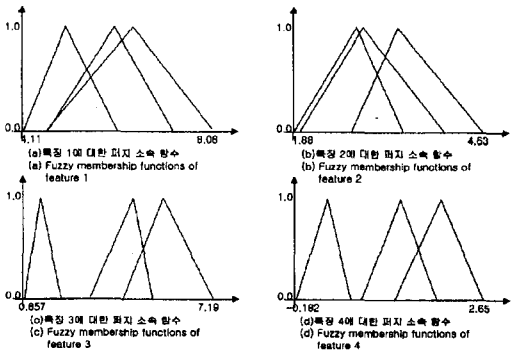


그림 5. 각 특징에 대한 퍼지 소속 함수

Fig. 5. The fuzzy membership functions for each feature.

표 1에서 볼 수 있듯이 본 논문에서 제안하는 방법이 기존의 방법에 비하여 규칙의 수가 작음은 물론 분류율도 높음을 알 수 있다. 그림 5는 훈련 데이터의 수가 90개일 때 생성된 각 특징에 대한 퍼지 소속 함수를 나타낸다.

또한 다음은 역시 훈련 데이터의 수가 90개일 때 생성된 퍼지 규칙이다.

- if f4 is small(-0.182,0.298,0.653) then c1(1)
- if f4 is medium(0.796,1.4,1.95) then c2(0.839)
- if f4 is large(1.3,2.01,2.65) then c3(0.809)

Ishibuchi[5]는 퍼지 공간 분할 방법으로써 퍼지 소속 함수를 생성하는데 있어서 패턴의 통계적 분포를 이용하지 않기 때문에 표 I에서와 같이 많은 수의 퍼지 규칙이 생성된다. 또한 Ishibuchi[5]의 경우 생성되는 퍼지 규칙의 조건부에는 항상 모든 특징이 모두 포함되므로 특징 수가 많은 경우 생성된 퍼지 규칙을 이해하기 어렵다. Jang[8]의 경우 히스토그램을 이용하여 소속 함수를 생성하지만 각 특징에 대한 히스토그램을 생성한다. 이러한 방법은 여러 개의 클래스가 하나의 군집을 이루고 있는 경우, 본 논문에서 사용한 방법인 각 특징의 각 클래스에 대하여 히스토그램을 생성하여 소속 함수를 생성하는 방법에 비하여 정확한 소속 함수를 생성하기 어렵다는 단점이 있다. 또한 생성되는 퍼지 규칙은 각 조건부에 각 조건부의 중요도를 표시하는 가중치가 있으므로 생성된 퍼지 규칙을 이해하기 어렵다. 그러나 본 논문에서 제안하는 퍼지 결정 트리에 기반한 퍼지 결정 규칙 생성 방법에 의하여 생성된 규칙은 각 규칙의 조건부가 짧고 규칙의 형태가 간단하여 Ishibuchi[5]와 Jang[8]에서 생성된 퍼지 규칙보다 이해가 쉬움을 알 수 있다.

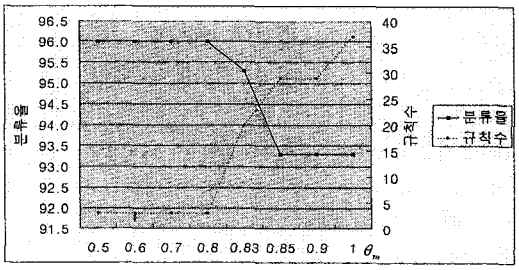


그림 6. N_s = 10, θ_n = 2일때, θ_m의 변화에 따른 규칙 수와 분류율의 변화

Fig. 6. Classification rate and the number of fuzzy rules as varing θ_m when N_s = 10 and θ_n = 2.

또한 그림 6과 그림 7은 θ_m과 θ_n의 변화에 따른

규칙의 수와 분류율의 변화를 보여준다. 그림을 통해 θ_m과 θ_n을 조절함으로써 규칙의 수와 분류율을 조절할 수 있음을 볼 수 있다. 이는 규칙을 생성하는 목적에 따라서 규칙의 정확성과 간결성을 조절할 수 있음을 보여준다.

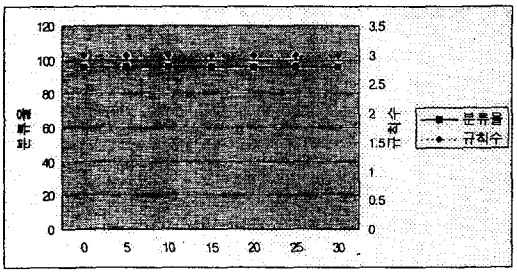


그림 7. N_s = 10, θ_m = 0.7일때, θ_n의 변화에 따른 규칙 수와 분류율의 변화

Fig. 7. Classification rate and the number of fuzzy rules as varing θ_n when N_s = 10 and θ_m = 0.7.

2. Wisconsin Breast Cancer 데이터에 대한 실험

Wisconsin Breast Cancer 데이터는 위스콘신 대학 병원의 William H. Wolberg가 수집한 데이터로 10개의 특징과 2개의 클래스로 이루어져 있으며 총 367개의 데이터로 이루어져 있다[10]. 훈련 데이터 중 빠진 값이 있는 데이터가 14개 있는데, 실험에서는 이러한 데이터를 제외하여 353개의 데이터만을 사용하였다. 퍼지 결정 트리를 생성하기 위해서 클래스 1이 106개, 클래스 2가 94개의 데이터가 사용되었으며 시험을 위해서는 353개 데이터 모두를 사용하였다.

실험에 사용된 파라미터는 N_s가 5, θ_n이 70, θ_m이 0.7로 하였다. 제안된 방법에 의해서 생성되는 퍼지 규칙의 수는 6개이고 시험 데이터에 대한 분류율은 91.8%이다. 그림 8은 퍼지 결정 트리에 사용된 퍼지 소속 함수로 10개의 특징 중 3개의 특징만이 퍼지 결정 트리에 사용되었다.

다음은 제안된 방법에 의해서 생성된 퍼지 규칙이다.

- if f3 is small(-2.09,2,5.68) then c1(0.738)
- if f3 is large(-0.997,4,13) and
  f4 is small(-1.79,2.01,8.02) and
  f7 is small(-1.8,2,12.5)
  then c1(0.656)
- if f3 is large(-0.997,4,13) and
  f4 is large(-0.637,5,12.3) and
  f7 is small(-1.8,2,12.5)
  then c2(0.502)
- if f3 is large(-0.997,4,13) and
  f4 is small(-1.79,2.01,8.02) and
  f7 is large(-0.635,8,12.2)
  then c2(0.526)
- if f3 is large(-0.997,4,13) and
  f4 is large(-0.637,5,12.3) and
  f7 is large(-0.635,8,12.2)
  then c2(0.652)

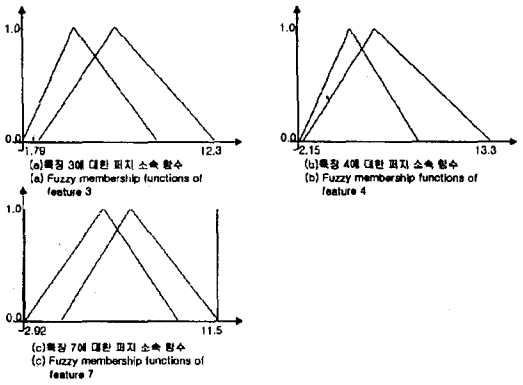


그림 8. 퍼지 결정 트리에 사용된 퍼지 소속 함수

Fig. 8. The fuzzy membership functions used in the fuzzy decision tree.

위의 규칙 역시 규칙의 수가 작고 규칙의 조건부가 짧아서 이해하기가 쉬움을 알 수 있다.

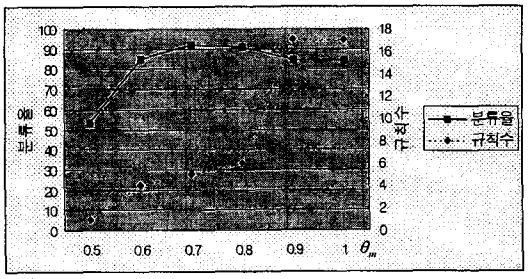


그림 9. N_s = 5, θ_n = 70일때, θ_m의 변화에 따른 규칙 수와 분류율의 변화

Fig. 9. Classification rate and the number of fuzzy rules as varing θ_m when N_s = 5 and θ_n = 70.

또한 그림 9와 그림 10은 θ_m과 θ_n의 변화에 따른

규칙의 수와 분류율의 변화를 보여준다. 그림을 통해 θ_m과 θ_n을 조절함으로써 규칙의 수와 분류율을 조절할 수 있음을 볼 수 있다. 이는 규칙을 생성하는 목적에 따라서 규칙의 정확성과 간결성을 조절 할 수 있음을 보여준다.

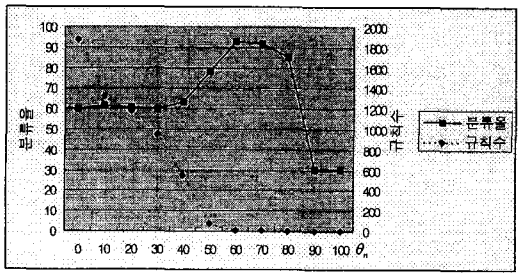


그림 10. N_s = 5, θ_m = 0.7일때, θ_n의 변화에 따른 규칙 수와 분류율의 변화

Fig. 10. Classification rate and the number of fuzzy rules as varing θ_n when N_s = 5 and θ_m = 0.7

Ⅶ. 결론 및 향후 연구 과제

본 논문은 컴퓨터 사용의 보편화에 따라 데이터의 수집과 저장이 용이해지면서, 데이터로부터 유용하고 이해가능한 정보를 추출하기 위한 방법인 KDD와 데이터 마이닝에 대한 연구이다. 기존에 대표적인 데이터 마이닝 알고리즘으로 사용된 ID3, C4.5와 같은 결정 트리 생성 알고리즘의 경우 생성되는 규칙의 수가 작고 이해하기가 쉽다는 장점이 있지만, 복잡한 패턴의 분류가 어렵다는 단점이 있었다. 그리고 퍼지 규칙 생성 방법의 경우 복잡한 패턴을 효과적으로 분류할 수 있다. 그러나 대부분 생성되는 규칙의 수가 너무 많거나 퍼지 규칙의 조건부에 가중치가 결합된 형태이므로 생성된 퍼지 규칙을 이해하는 것이 어려웠다.

따라서 본 논문에서는 간결한 규칙을 생성하는 결정 트리 생성 알고리즘과 퍼지를 결합함으로써 분류율이 높고 이해가 쉬운 퍼지 규칙을 생성할 수 있는 데이터 마이닝 알고리즘을 제안하였다. 제안한 알고리즘은 결정 트리 생성 알고리즘과 퍼지를 결합함으로서 분류율이 높으면서 이해하기 쉬운 규칙을 생성할 수 있으며, 데이터의 통계적 특성을 나타내고 있는 히스토그램을 이용하여 퍼지 소속 함수를 생성함으로써 더 효율적으로 퍼지 결정 트리를 생성할 수 있다. 또한 표준적인 벤치마크 데이터인 Iris와 Wisconsin Breast Cancer 데이터에 대하여 실험함으로써 제안한 방법이 효과적임을 보였다.

향후 연구 과제로는 보다 효율적인 소속 함수를 생성함으로써 생성된 퍼지 규칙의 정확성을 높이기 위하여 소속 함수를 조율하기 위한 연구와 생성된 퍼지 결정 트리를 가지 치기(pruning)함으로써 퍼지 규칙을 보다 간단히 하기 위한 연구가 필요하다. 또한 생성된 규칙의 정확성과 간결성을 결정하는데 있어서 중요한 변수인 θ_n과 θ_m을 결정하기 위한 알고리즘적인 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., "From Data mining to Knowledge Discovery: An Overview", in *Advances in Knowledge Discovery and Data Mining*, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., pp. 1-34, MIT Press, 1996.

[2] 박영택, 이강로, ID3 계열의 귀납적 기계학습, *정보과학회지*, 제 13권, 제 5호, pp. 6-19, 1995

[3] Quinlan, J. R., Induction of Decision Trees, Machine Learning, 1, pp. 81-106, 1986.

[4] Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[5] Ishibuchi, H., Nozaki, K., Tanaka, H., Effective fuzzy partition of pattern space for classification problems, *Fuzzy Sets and Systems*, vol. 59, pp. 295-304, 1993.

[6] Abe, S., Lan, M.-S., A Classifier Uning Fuzzy Rules Extracted Directly from Numerical Data, Proceedings of 2nd IEEE Conference on fuzzy Systems, pp. 1191-1198, 1993.

[7] Rhee, F. C., Krishnapuram, R., Fuzzy rule generation methods for high-level computer vision, Fuzzy Sets and Systems, vol. 60, pp. 245-258, North-Holland, 1993.

[8] Jang, D-S., Choi, H-I., Automatic

Generation of Fuzzy Rules with Fuzzy Associative Memory, In Proc. of the ISCA 5th International Conference, pp. 182-186, 1996.

[9] Umanno, M., Okamoto, H., Hatono, I. et al, Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis Systems, In *Proc. of 3th IEEE International Conference on Fuzzy Systems*, pp. 2113-2118, 1994.

[10] http://www.ics.uci.edu/~mlearn/MLRepository.html.

저 자 소 개

金 明 源(正會員) 第 35卷 C編 第 3號 參照
현재 숭실대 컴퓨터학부 교수

閔 昌 宇(正會員)
1995년 숭실대학교 컴퓨터 학부 졸업(학사). 1998년 숭실대학원 전자계산학과 졸업(석사). 1998년 한국 IBM 소프트웨어 연구소 재직중. 관심분야는 유연추론, 신경회로망, 전화 알고리즘, 퍼지 시스템, 인공생명 등

金 修 光(正會員)
1983년 부산대학교 기계공학과 졸업. 1985년 동 대학원(석사) 졸업. 1995년 ~ 1997년 포항제철 기술 연구소 계측제어 연구팀 책임 연구원. 1998년 ~ 현재 부산 정보 대학 기계산업계열 전임강사, 주관심분야는 열기관 최적자동 제어 및 CAM