특징공간을 사선 분할하는 퍼지 결정트리 유도
(Fuaay Decision Tree Induction to Obliquely Partitioning a Feature Space)

이 우 향 † 이 건 명 ††
(Woo-Hang Lee) (Keon-Myung Lee)

요 약 결정트리 생성은 특징값들로 기술된 사례들로부터 분류 규칙을 추출하는 유용한 기계학습 방법중 하나이다. 결정트리는 특징공간을 분할하는 형태에 따라 단변수(univariate) 결정트리와 다변수(multivariate) 결정트리로 대별된다. 실제 현장에서 얻어지는 데이터는 관측오류, 불확실성, 주관적인 판단 등의 이유로 특징값 자체에 오류를 포함하는 경우가 많다. 이러한 오류에 대해 강건한 결정트리를 생성하기 위한 방법으로 퍼지 기법을 도입한 결정트리 생성 방법에 대한 연구가 진행되어 왔다. 현재까지 대부분의 퍼지 결정트리에 대한 연구는 단변수 결정트리에 퍼지 기법을 도입한 것들이며, 다변수 결정트리에 퍼지 기법을 적용한 것은 찾아보기 힘들다. 이 논문에서는 다변수 결정트리에 퍼지 기법을 적용하여 퍼지 사선형 결정트리라고 하는 퍼지 결정트리를 생성하는 방법을 제안한다. 또한 제안한 결정트리 생성 방법의 특성을 보이기 위한 실험 결과를 보인다.

키워드 : 분류, 결정트리, 퍼지 결정튈, 의사결정, 기계학습, 퍼지이론

Abstract Decision tree induction is a kind of useful machine learning approach for extracting classification rules from a set of feature-based examples. According to the partitioning style of the feature space, decision trees are categorized into univariate decision trees and multivariate decision trees. Due to observation error, uncertainty, subjective judgment, and so on, real-world data are prone to contain some errors in their feature values. For the purpose of making decision trees robust against such errors, there have been various trials to incorporate fuzzy techniques into decision tree construction. Several researches have been done on incorporating fuzzy techniques into univariate decision trees. However, for multivariate decision trees, few research has been done in the line of such study. This paper proposes a fuzzy decision tree induction method that builds fuzzy multivariate decision trees named fuzzy oblique decision trees. To show the effectiveness of the proposed method, it also presents some experimental results.

Key words : classification, decision tree, fuzzy decision tree, decision making, machine learning, fuzzy theory

1. 서 론

분류(classification)는 주어지는 데이타를 이미 정해진 부류(class)와 비교하여 가장 근접한 부류로 판별하는 것으로, 패턴 인식, 의사결정, 데이타 분석 등에서 핵심적인 작업이다. 분류 작업을 위해서는 각 부류에 대한 표준 패턴을 찾아내는 것이 전제되어야 한다. 이러한 표준 패턴을 부류 정보가 있는 데이타들로부터 자동으로 추출하기 위한 다양한 방법론들에 대한 연구가 진행되어 왔다[1-16]. 대표적인 것들로는 ID3, C4.5, CART 등과 같은 결정트리 생성(decision tree induction), 다층 퍼셉트론(multilayer perceptron), LVQ, IF-Then 규칙형태의 사전 지식을 이용하는 KBANN, KBCNN와 같은 규칙기반(rule-based) 신경회로망 등과 같은 신경회로망 모델, 베이지안 분류기(Bayesian classifier) 등과 같은 확률 기반 모델, 유전자 알고리즘(genetic algorithm), 유전자 프로그램(genetic programming) 등과 같은 진화연산 기법 등이 있다[1-16]. 이 논문에서

· 이 연구는 1998~2000년 정보통신관리단 대학기초연구지원사업의 지원을 받아 수행된 것임.
† 비 회 원 : 지식시스템(주) 연구원
artin@aicore.chungbuk.ac.kr
†† 정 회 원 : 충북대학교 컴퓨터과학과 교수
kmlee@cbucc.chungbuk.ac.kr
논문접수 : 2001년 2월 8일
심사완료 : 2001년 11월 9일

는 결정트리를 기반으로한 분류 규칙 학습방법의 한가지로 퍼지 경계영역(fuzzy boundary)를을 갖는 다변수 결정트리를 소개하고, 이에 대한 학습 방법을 제안한다.

결정트리 생성은 사례들로부터 이해하기 쉬운 트리 형태의 분류 규칙을 추출하는 학습방법으로, 데이타 분석, 분류, 일반화, 의사결정 등을 목적으로 폭넓게 사용되고 있다[1, 13]. 결정트리는 특징공간 분할 형태에 따라 단변수(univariate) 결정트리와 다변수(multivariate) 결정트리로 대별된다[1]. 단변수 결정트리는 트리의 각 노드에서 하나의 특징값만 비교하는 결정트리이고, 다변수 결정트리는 각 노드가 여러 특징값을 고려하여 비교 연산을 하는 결정트리이다. 동일 학습 데이타에 대해서 단변수 결정트리에 비하여, 다변수 결정트리가 일반적으로 더 단순한 형태를 가진다. 반면 다변수 결정트리는 노드가 일반적으로 여러 특징값에 대한 선형방정식의 값과 비교연산을 하기 때문에, 표현된 분류 규칙을 이해하는 것이 상대적으로 쉽지 않은 단점이 있다.

결정트리 생성을 위해 현장에서 수집되는 사례들, 즉 학습 데이타들은 관측 오류, 불확실성, 주관적 판단 등의 원인으로 참값이 아닌 근사값 형태로써 기술되는 경우가 많다. 이런 잠재적 오류를 내포하고 있는 데이타를 이용하여 명확한 기준으로 특징공간을 분할하는 결정트리를 생성하면, 특징 경계부근에서 분류 오류가 발생할 가능성이 커진다[2, 9]. 또한 명확한 경계를 기준으로 데이타를 분류하면, 특징값의 작은 변화에도 데이타가 갑자기 다른 클래스로 분류될 수도 있다. 이런 문제들을 해결하고자 작은 오류에 민감하지 않은 퍼지 결정트리 생성에 대한 많은 연구가 진행되어왔다[2, 7, 10]. 이들 연구에서는 특징공간 상에 분명한 분류 경계를 설정하는 대신에, 퍼지 경계 설정하는 방식을 이용한다.

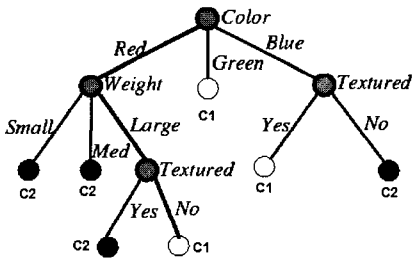
퍼지 결정트리에 대한 기존의 연구는 주로 단변수 결정트리에 퍼지 기법을 적용하는 방법을 중심으로 이루어져왔다[2, 7, 10]. 이 논문에서는 다변수 결정트리에 퍼지 기법을 도입하여 퍼지 다변수 결정트리를 만드는 방법을 제안한다. 이 논문에서 대상으로 하는 다변수 결정트리의 형태는 OC1 알고리즘[3]에 의해서 생성되는, 트리의 노드가 특징공간을 사선 분할하는 경계면에 대해서 비교 연산을 하는 결정트리이다. 클래스 경계면 부근에서의 작은 오류에 민감하지 않은 퍼지 결정트리로 만들기 위해서는, 특징공간을 분명한 경계면으로 분할하지 않고 퍼지 경계면(fuzzy boundary surface)으로 분할하도록 하는 것이 필요하다. 이 논문에서는 우선 특징공간을 사선 분할하는 경계면을 퍼지 경계면으로 정의하는 방법을 제안한 다음, 이를 이용하여 퍼지 다변수 결정트리를 생성하는 방법을 제안한다. 제안한 방법에 의해서 생성되는 퍼지 다변수 결정트리를 이 논문에서는 '퍼지 사선형 결정트리(fuzzy oblique decision tree)'라고 부른다.

이 논문은 다음과 같이 구성된다. 2절에서는 일반 결정트리와 퍼지 결정트리에서 대해서 간단히 소개하고, 3절에서는 퍼지 사선형 결정트리를 생성하는 방법과 퍼지 사선형 결정트리를 이용하여 데이타의 클래스를 결정하는 방법을 제안한다. 4절에서는 제안한 방법의 유용성을 보이기 위해 대표적인 벤치마크 데이타인 Iris 데이타에 대한 퍼지 사선형 결정트리와 다층 퍼셉트론의 실험 결과를 보이고, 5절에서는 결론을 맺는다.

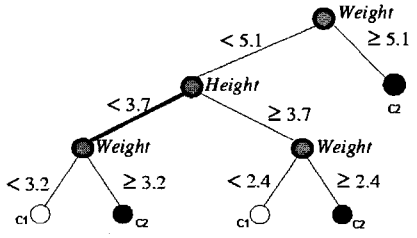
2. 결정트리와 퍼지 결정트리

2.1 결정트리

결정트리는 (그림 1)과 같이 분류규칙을 표현하는 트리로서, 비단말노드에는 분류를 위해 비교할 데이타의 특징이 명시되고, 링크에는 비교할 조건 또는 특징값이 부여되며, 단말노드에는 근(root) 노드에서 해당 노드까지의 경로상에 있는 모든 조건을 만족하는 데이타가 속하는 클래스 값이 부여된다. 이러한 결정트리로 표현된 분류규칙은 IF-THEN 형태의 분류규칙으로 쉽게 변환될 수 있다. 결정트리는 데이타를 정확히 분류할 수 있으면서, 크기(깊이, 노드 수 등)가 작을수록 바람직하다.



(a) 특징값이 기호인 결정트리



(b) 특징값이 수치인 결정트리

그림 1 결정트리의 예

주어진 학습 데이타로부터 결정트리를 생성하는 방법은 일반적으로 다음과 같은 과정을 따른다.

알고리즘 1 전형적인 결정트리 생성 알고리즘

procedure 결정트리 생성
begin
1. 모든 학습 데이타를 포함하는 근노드를 생성한다.
2. 모든 단말노드가 동일한 클래스의 데이타만을 포함한다든가 하는 종료조건을 만족하면, 각 단말노드에 대해 현재 포함하고 있는 데이타를 참고하여 클래스를 부여한 다음, 트리 생성 과정을 종료한다.
3. 생성된 노드에 대한 가능한 분할 방법들에 대하여, 이들이 해당 노드의 데이타를 얼마나 잘 분류할 수 있는지 평가한다.
4. 가장 높은 평가를 받은 분할 방법을 선택하여, 이를 기준으로 자식노드를 생성하고 분할 조건에 따라 데이타를 이들 자식노드에 전달한다.
5. 생성된 자식노드에 대하여 위의 단계 2-4를 반복적으로 수행한다.

end.

결정트리는 노드의 특징공간 분할 형태에 따라 단변수 결정트리와 다변수 결정트리로 대별된다. 단변수 결정트리란 트리 확장을 위한 특징공간 분할 기준으로서 단 하나의 특징만을 고려하는 것이다. 특징공간의 분할 형태를 보면 (그림 2)와 같이 분할 경계가 특징값의 축에 평행하다. 따라서 단변수 결정트리에 의해 기술되는 분류규칙은 비교적 쉽게 이해할 수 있다. 단변수 결정트리를 생성하는 대표적인 알고리즘에는 ID3[11], C4.5[4], CART[8] 등이 있다.

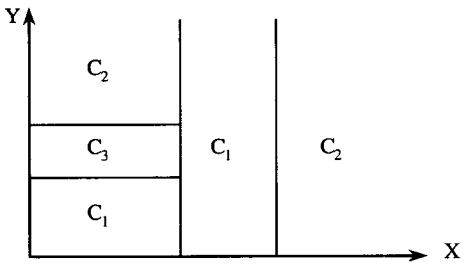


그림 2 단변수 결정트리의 특징공간 분할

다변수 결정트리는 노드들이 여러 개의 특징값을 동시에 고려하여 비교연산을 수행하는 결정트리로서, 이러한 결정트리를 생성하는 알고리즘에는 LMDT[1], OC1[3], CART-LC[8] 등이 있다. 다변수 결정트리에서는 (그림 3)에 보인 것과 같이 특징공간을 분할하는 분할 경계가 특징값의 축에 평행하지 않고 임의의 방향을 가질 수 있다.

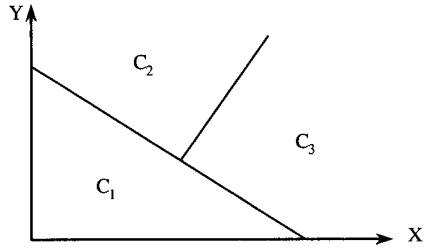


그림 3 다변수 결정트리의 특징공간 분할

다변수 결정트리는 특징공간 분할이 임의의 형태로 이루어지기 때문에 다차원 특징을 갖는 데이타에 대한 다변수 결정트리인 경우에는 표현된 분류규칙을 직관적으로 이해하는 것이 다소 어렵기는 하지만, 비교적 트리 구조가 간단하다는 장점이 있다. 이 논문에서는 다변수 결정트리를 생성하는 알고리즘의 하나인 OC1을 이용하는 퍼지 결정트리 생성 방법을 제안한다.

2.2 퍼지 결정트리

일반 결정트리는 명확한 값을 기준으로 특징공간을 분할하기 때문에, 미세한 차이를 갖는 두 데이타를 서로 다른 클래스로 분류할 수 있다. 이런 문제을 완화시키기 위해 특징공간을 소속함수(membership function)[5]를 이용하여 정의한 퍼지 경계면(fuzzy boundary surface)으로 분할하는 퍼지 결정트리 생성 방법에 대한 연구가 활발히 수행되고 있다[2,7,10].

퍼지 결정트리 생성을 위해 노드를 확장할 때는, 우선 일반 결정트리와 마찬가지로 분류 특성이 가장 좋은 특징을 선택한다. 이를 위해 각 특징에 대해서 미리 주어지거나 트리 생성중에 정의한 소속함수에 의해 표현된 퍼지 언어항(fuzzy linguistic term)을 이용하여 특징공간을 퍼지 분할(fuzzy partitioning)한 다음, 정보이득(information gain)[13] 등의 척도를 이용하여 특징별 분류 성능을 평가하며, 가장 성능이 좋은 특징을 선택한다. 노드를 선택된 특징영역에 정의된 퍼지 언어항 개수만큼의 자식노드로 확장하고, 자식노드에 연결되는 링크에는 선택된 특징영역을 퍼지 분할하는 퍼지 언어항을 부여한다. (그림 4)는 특정 특징영역(x)을 소속함수에

의해 정의되는 세 개의 퍼지 언어항(X_1, X_2, X_3)으로 퍼지 분할한 예를 보인 것이다.

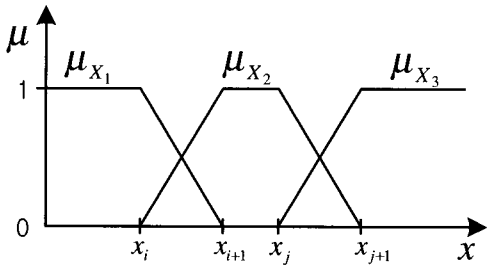


그림 4 특징영역의 퍼지 분할

일반 결정트리에서는 노드에서 특징값의 비교 후에 데이타를 하나의 하위 노드에만 전달하지만, 퍼지 결정트리에서는 데이타를 각기 다른 소속정도로 여러 자식 노드에 전달할 수도 있다. 퍼지 결정트리에서는 단말노드에 클래스를 부여할 때 하나의 클래스만을 지정하지 않고, 여러 클래스를 확신도와 함께 지정하기도 한다. (그림 5)는 전형적인 퍼지 결정트리를 보인 것으로 노드에는 비교할 특징이 부여되어 있고, 링크에는 특징영역을 퍼지 분할하는 소속함수에 의해 정의되는 퍼지 언어항이 부여되어 있으며, 단말노드에는 근노드에서 해당 단말노드까지의 경로상에 있는 조건을 모든 만족하는 데이타가 속하는 클래스 값이 부여되어 있다.

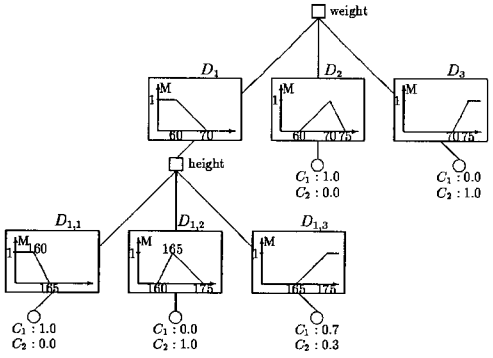
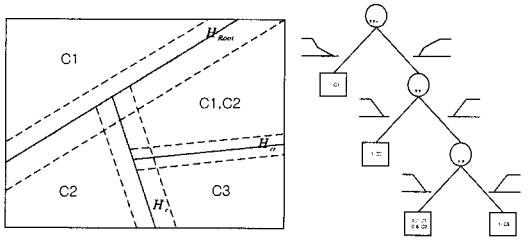


그림 5 퍼지 결정트리

3. 퍼지 사선형 결정트리

이 논문에서 제안하는 방법에 의해 생성되는 퍼지 사선형 결정트리(fuzzy oblique decision tree)는 (그림 6-a)와 같이 특징공간을 사선형으로 분할하면서, 분할 경계면이 소속함수에 의해 퍼지 경계면으로 정의되는 (그림 6-b)와 같은 이진(binary) 결정트리이다. 퍼지 사선형 결정트리에서 각 노드는 여러 특징값에 대한 선형방정식으로 표현되는 초평면(hyperplane)에 대한 정보를 가지고 있으며, 노드에서 분기하는 링크에는 이 초평면이 나타내는 경계면으로부터의 거리에 대해 정의된 소속함수가 부여된다. 단말노드에는 근노드로부터 해당 노드 사이에 있는 조건을 모두 만족하는 데이타가 속하는 클래스가 부여되는데, 이때 클래스는 소속정도가 있는 퍼지 집합으로 정의된다. 이 논문에서는 이러한 퍼지 사선형 결정트리를 생성하는 방법을 제안한다. 제안한 퍼지 사선형 결정트리 생성 방법은 (알고리즘 1)에 보인 전형적인 결정트리 생성과정을 따르는데, 이때 이 절에서는 설명할 특징공간 분할 방법, 결정트리 생성을 위한 노드 확장 방법, 일반화(generalization) 특성 향상을 위한 트리 가지치기 방법, 단말노드에 대한 클래스 부여 방법, 퍼지 사선형 결정트리를 이용한 데이타의 클래스 추론 방법을 결정트리 생성과정에서 이용한다.



(a) 특징 공간 분할 (b) 결정트리

그림 6 퍼지 사선형 결정트리

3.1 특징공간 분할

퍼지 사선형 결정트리를 생성하기 위해서는 특징공간을 사선형으로 퍼지 분할하는 특징공간 분할 방법이 필요하다. 제안한 퍼지 사선형 결정트리 생성 방법에서는 먼저 특징공간을 분명하게 구분하는 기본 경계면을 찾은 다음, 이 경계면으로부터의 거리에 대한 소속함수를 정의하여 퍼지 경계영역을 설정하는 방법을 사용한다. 기본 경계면은 OC1 알고리즘[3]에서 사용되는 초평면 탐색 알고리즘을 이용하여 구한다. 기본 경계면으로 사용되는 d개의 특징으로 구성된 특징공간을 분할하는 초평면 $H(x)$의 식은 $H(x)= a_1x_1+a_2x_2+ \ldots +a_dx_d+a_{d+1}$와 같이 특징값 변수($x_i$)에 대한 선형방정식으로 표현된다.

퍼지 사선형 결정트리의 퍼지 경계영역은, OC1의 초평면 탐색 알고리즘을 이용하여 얻어진 초평면으로부터의 거리에 대한 소속함수에 의해서 정의된다. 소속함수를 정의하기 위해서 특징공간상의 임의의 데이타 점과 초평면 사이의 거리를 다음과 같이 구한다. 임의의 데이타 $x=(x_1, x_2, \ldots, x_d)$와 초평면 $H(x)= a_1x_1+a_2x_2+\ldots+a_dx_d+a_{d+1}$간의 거리에 대응하는 스칼라 값 t는 식 (1)의 벡터의 연산을 통해 구할 수 있다.

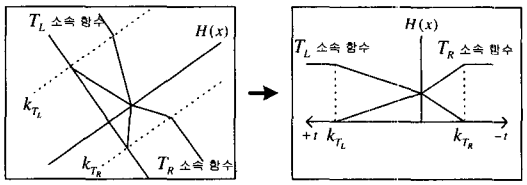
$$t(x) = \frac{\vec{a}\cdot(\vec{x}-\vec{o})}{|\vec{a}|} \quad (1)$$

여기에서 $\vec{a}=(a_1, a_2, \ldots, a_d)$는 초평면 $H(x)$의 법선 벡터이고, $\vec{o}$는 초평면상의 임의의 한 점에 대한 벡터로서, 가장 간단한 형태는 $(0, 0, \ldots, -a_{d+1}/a_d)$이고, $\vec{x}$는 데이타 x에 대응하는 벡터이다. 거리에 대응하는 스칼라 값 $t(x)$의 절대값 $|t(x)|$은 특징공간상에서 데이타에 해당하는 점과 초평면 $H(x)$ 사이의 거리가 된다. 한편, $t(x)$의 부호는 데이타 점의 초평면에 대한 상대적 위치를 나타내는데, $t(x)$가 양수이면 데이타 점이 초평면의 법선 벡터 방향으로 초평면 위쪽 부분에 위치하고, $t(x)$가 음수이면 데이타 점이 초평면의 아래쪽 부분에 위치한다는 것을 의미한다.

일단 영역 분할을 위한 기본 경계면에 대한 초평면이 정해지면, 초평면으로부터의 거리에 대한 소속함수를 정의하여 퍼지 경계면을 설정한다. 거리에 대한 소속함수는 (그림 7)과 같이 정의한다. 즉, 기본 경계면을 기준으로 경계면의 왼쪽과 오른쪽에 대해서 소속함수를 정의하는데, 기본 경계면에서 소속함수 값이 0.5가 되도록 하고, 기본 경계면 주변에서 왼쪽, 오른쪽 영역에 대한 소속함수들이 중첩되도록 한다. 이때 문제는 왼쪽 또는 오른쪽 영역에 대한 소속함수가 반대편 영역의 어느 부분까지 영향을 미치도록 할 것인가 하는 것이다. 즉, (그림 7)에서 볼 때 k_{TR}, k_{TL}을 어떻게 정의할 것인가 하는 것이다.

제안한 방법에서는 기준 초평면을 기준으로 주어진 데이타의 분포를 이용하는데, 기준 초평면으로부터의 거리에 따라 데이타를 정렬하여 해당 영역 전체 데이타 중 k%번째 데이타가 위치하는 거리를 상대쪽 소속함수가 영향을 미치는 최대 한계 거리로 한다. 이와 같이 소속함수 한계 영역을 결정하기 위해 사용하는, 데이타의 기본 초평면으로부터의 데이타의 분포 비율을 이 논문에서는 '퍼지 적용비율'이라고 부른다. (그림 7)은 특징공간상에서 정의된 퍼지 경계영역에 대한 소속함수를 거리 공간의 소속함수로 변환하여 표현한 것이다.

다음 식 (2), (3)의 $\mu_{TL}(x)$, $\mu_{TR}(x)$는 주어진 초평면



(a) 특징공간상에서 퍼지 경계영역 (b)거리 t기준의 퍼지 경계영역

그림 7 퍼지 경계영역

에 대해서 k%의 퍼지 적용비율을 적용할 때 정의되는 퍼지 경계면의 왼쪽 영역 T_L과 오른쪽 영역 T_R에 대한 소속함수를 나타낸 것이다. 여기에서는 초평면의 법선벡터 방향에 있는 영역을 왼쪽 영역으로 전제한다.

$$\mu_{T_L}(x) = \begin{cases} 1 & \text{if } k_{T_L} \le t(x) \\ 0.5t(x)/k_{T_L}+0.5 & \text{if } 0 \le t(x) \le k_{T_L} \\ 0.5(t(x)-k_{T_R})/k_{T_R} & \text{if } k_{T_R} \le t(x) \le 0 \\ 0 & \text{if } t(x) < k_{T_R} \end{cases} \quad (2)$$

$$\mu_{T_R}(x) = \begin{cases} 1 & \text{if } t(x) \le k_{T_R} \\ 0.5t(x)/k_{T_R}+0.5 & \text{if } k_{T_L} \le t(x) \le 0 \\ 0.5(k_{T_L}-t(x))/k_{T_L} & \text{if } 0 \le t(x) \le k_{T_L} \\ 0 & \text{if } k_{T_R} < t(x) \end{cases} \quad (3)$$

3.2 결정트리 생성을 위한 노드 확장

제안한 퍼지 사선형 결정트리 생성 알고리즘도 (알고리즘 1)에 보인 전형적인 결정트리 생성 절차를 따라, 전체 데이타를 포함하는 근노드에서 시작하여 현재 데이타들을 최적으로 분할하는 사선형 퍼지 경계면을 선택한 후, 이를 기준으로 데이타들을 나누어 두 개의 자식노드를 만들고, 새로 만들어진 노드에 대해서 동일한 과정을 반복하는 방법으로 노드를 확장해가면서 결정트리를 생성한다. 퍼지 사선형 결정트리에서는 초평면으로 부터의 거리에 대한 소속함수로 정의된 퍼지 경계면을 특징공간의 분할에 이용하기 때문에, 퍼지 경계면을 기준으로 데이타의 위치를 결정할 때 주의가 필요하다. 일반 결정트리에서는 데이타가 경계면을 중심으로 어느 쪽에 위치하는가를 분명히 판정할 수 있지만, 퍼지 경계면을 사용하는 퍼지 결정트리의 경우 경계면 부근의 데이타에 대해서는 경계면에 대한 상대적 위치를 분명하게 판정하기 곤란하다. 따라서 제안한 방법에서는 경계면을 정의하는 소속함수에 대한 데이타의 소속정도를 계산하여, 경계면을 기준으로 데이타의 양쪽 영역에 대한 소속정도로 사용한다. 이러한 특성으로 인해서 퍼지 사선형 결정트리에서는 하나의 데이타가 동시에 두 개의 자식노드에 전달될 수도 있다. 데이타가 트리 생성과정이나 클래스

추론 과정에서 자식노드에 전달될 때, 어떤 경우에는 소속정도가 매우 낮은 의미가 적은 데이타도 자식노드로 계속해서 전달될 수 있다. 따라서 제안한 방법에서는 '소속정도 임계값'을 지정하여, 이 임계값보다 낮은 데이타는 자식노드에 전달되지 못하도록 한다.

제안한 방법에서는 특징공간을 분할하는 퍼지 경계면의 분할 특성을 평가하기 위해 다음 식 (4)와 같이 이분 규칙(twoing rule)[3]을 데이타의 소속정도값을 반영하도록 다음과 같이 수정하여 정의한 퍼지 이분 규칙(fuzzy twoing rule)을 이용한다.

$$I_N = \frac{\sum_{x\in D_N}\mu_{T_L}(x)}{\sum_{x\in D_N}\mu_N(x)} \times \frac{\sum_{x\in D_N}\mu_{T_R}(x)}{\sum_{x\in D_N}\mu_N(x)} \times \left(\sum_{j=1}^{k} \left| \frac{\sum_{x\in D_N, x.class=j}\mu_{T_L}(x)}{\sum_{x\in D_N}\mu_{T_L}(x)} - \frac{\sum_{x\in D_N, x.class=j}\mu_{T_R}(x)}{\sum_{x\in D_N}\mu_{T_R}(x)} \right| \right)^2 \quad (4)$$

위 식에서 I_N은 노드 N의 데이터 D_N에 대한 임의의 주어진 퍼지 경계면의 분할 특성을 나타내는 척도로서, 값이 클수록 분할 특성이 좋은 것이다. $\mu_{TL}(x)$, $\mu_{TR}(x)$는 각각 학습 데이타 x가 주어진 퍼지 경계면의 왼쪽 편 T_L과 오른쪽 편 T_R에 속하는 정도를 나타내는 것이고, $x.class=j$는 학습 데이타 x가 속하는 클래스가 j인 것을 의미한다.

3.3 결정트리 가지치기

결정트리가 학습 데이타에 대해서 너무 완벽하게 학습(overfitting)되면, 학습 데이타에 대해 약간의 오류를 갖는 결정트리보다, 학습되지 않은 데이타에 대해서는 오히려 분류 정확성이 떨어질 수 있다. 이러한 문제를 완화시키기 위해서 완성된 결정트리를 단말노드의 삭제를 통해 단순화시키는 가지치기(pruning) 작업을 수행하게 된다[12]. 제안한 퍼지 사선형 결정트리 생성 방법에서는 대표적인 가지치기 방법중 하나인 Breiman 등[8]이 개발한 최소비용 복잡도(Minimal Cost Complexity) 척도를 소속정도가 있는 데이타에 대해서도 적용할 수 있도록, 다음 식 (5)과 같이 수정하여 정의한 가지치기 척도 α_N을 이용하여 가지치기를 수행한다.

$$\alpha_N = \left(\frac{\sum_{x.class\neq N.class}\mu_N(x) - \sum_{N_T\in N_T, x.class\neq N.class}\mu_{N_T}(x)}{|D|} \right) / |N_T| \quad (5)$$

위 식에서, D는 전체 학습 데이타의 집합을 나타내고, $x.class$는 데이타 x가 속하는 클래스를 의미하고, $N.class$는 노드 N에 의해 표현되는 클래스이며, N_T는 노드 N을 조상 노드로 갖는 단말노드의 집합이다.

가지치기 평가 척도 α_N의 값은, 노드 N이 자식노드들을 이용하여 확장될 때 얻을 수 있는 분류 정확도 향상 정도에 비례하고, 노드 확장에 의해 추가되는 하위 노드의 개수에 반비례한다. 가지치기는 모든 노드 N에 대해서 가지치기 평가척도 α_N의 값을 구한 다음, 가장 낮은 값을 갖는 노드를 찾아 해당 노드의 하위 노드들을 모두 제거하고 해당 노드를 단말노드로 만드는 과정을 반복하는 방법으로 수행한다.

3.4 단말노드의 클래스(class) 부여

결정트리는 분류규칙을 트리 형태로 표현하는 것으로, 트리의 비단말노드와 링크는 분류규칙의 조건 부분에 해당하고, 단말노드는 분류규칙의 결과 부분에 해당한다. 따라서 결정트리의 단말노드에는, 근노드에서 해당 단말노드까지의 경로 상에 있는 모든 조건을 만족하는 데이타들이 속하는 클래스가 부여된다. 제안한 퍼지 사선형 결정트리에서는 특징공간의 분할이 퍼지 경계면에 의해서 이루어지기 때문에, 단말노드에 단 하나의 클래스만이 아니라, 만족정도 값이 서로 다른 여러 개의 클래스가 부여되는 것을 허용한다.

제안한 퍼지 사선형 결정트리 생성방법에서는 단말노드의 클래스별 만족정도를 다음의 식 (6)과 같이 결정한다. 우선, 각 단말노드에 도달한 모든 데이타에 대하여 클래스 별로 단말노드에 대한 데이타의 소속정도를 합산한다. 클래스 별로 합산된 소속정도를 이용하여 해당 단말노드에 대한 클래스 차지 비율을 계산하고, 이 값을 단말노드의 클래스별 만족정도로 한다.

$$\chi_{N_i}(k) = \frac{\sum_{x\in N_i, x.class=k}\mu_{N_i}(x)}{\sum_{x\in N_i}\mu_{N_i}(x)} \quad (6)$$

여기에서 $x_{Ni}(k)$는 단말노드 N_i의 클래스 k에 대한 만족정도를 나타내며, μ_{Ni}는 단말노드 N_i대한 데이타 x의 소속정도를 나타낸다. (그림 8)은 퍼지 사선형 결정트리의 단말노드에 부여되는 클래스 정보의 형태를 예시한 것이다. 그림에서 왼쪽 단말노드는, 해당 노드에 도달한 데이타가 0.2의 만족정도로 클래스 C_1에는 속하고, 0.8의 만족정도로 클래스 C_2에 속한다는 것을 나타낸다.

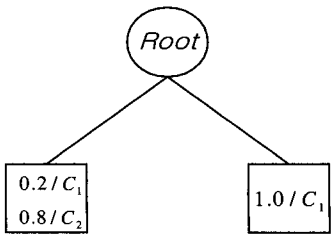


그림 8 단말노드에 대한 클래스 부여

3.5 퍼지 사선형 결정트리를 이용한 클래스 추론

새로운 데이타의 클래스 결정을 위해 생성된 퍼지 사선형 결정트리를 사용할 때는 다음 과정을 따른다. 입력 데이타는 초기 소속정도를 1로 하여 결정트리의 근노드에 입력된다. 데이타를 받은 각 비단말노드 N은 노드에 부여된 퍼지 경계면에 대한 정보를 이용하여 해당 데이타의 퍼지 경계면의 왼쪽 T_L 및 오른쪽 T_R영역에 대한 소속정도 $\mu_{TL}(x)$와 $\mu_{TR}(x)$를 계산한다. 계산된 소속정도를 이용하여 왼쪽 자식노드 N_L 및 오른쪽 자식노드 N_R에 대한 데이타 x의 소속정도 $\mu_{N_L}(x)$, $\mu_{N_R}(x)$를 다음과 같이 계산한다.

$$\begin{aligned}\mu_{N_L}(x) &= f(\mu_N(x), \mu_{T_L}(x)) \\ \mu_{N_R}(x) &= f(\mu_N(x), \mu_{T_R}(x))\end{aligned} \quad (7)$$

여기에서 $\mu_N(x)$는 노드 N에 대한 데이타 x의 소속정도이고, f는 퍼지 결정트리를 생성할 때 선택한 min 연산 등과 같은 T-norm 연산자[5]이다. 계산된 자식노드에 대한 소속정도가 주어진 소속정도 임계값보다 크면 데이타를 자식노드에 전달한다. 이 과정을 데이타가 단말노드에 도달하거나 더 이상 자식노드에 전달될 수 없을 때까지 반복한다.

퍼지 결정트리에서는 하나의 입력 데이타가 여러 개의 단말노드에 도달할 수 있으므로, 모든 단말노드에서의 데이타의 소속정도와 단말노드의 클래스별 만족정도를 다음과 같이 종합하여 입력 데이터 x에 대한 클래스 k의 확신도 $cf_k(x)$를 결정한다.

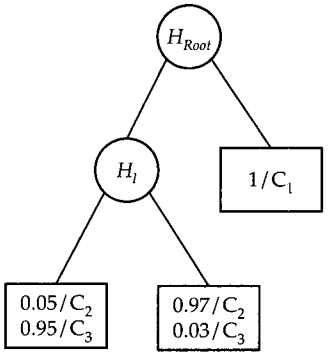
$$cf_k(x) = \frac{c_k(x)}{\max_i c_i(x)} \quad (8)$$

$$c_k(x) = \sum_{N_i \in TN} \chi_{N_i}(k) \times \mu_{N_i}(x) \quad (9)$$

위 식은 입력 데이터 x를 클래스 k로 분류하는 것에 대한 확신도 $cf_k(x)$를 구하는 것으로, TN은 주어진 퍼지 사선형 결정트리의 단말노드들의 집합을 나타내고, $\chi_{N_i}(k)$는 단말노드 N_i의 클래스 k에 대한 만족정도를 나타내고, $\mu_{N_i}(x)$는 데이타 x의 단말노드 N_i에 대한 소속정도를 나타낸다.

4. 실험 및 분석

제안한 퍼지 사선형 결정트리 생성방법의 유용성을 보이기 위해 분류 문제 분야에서 대표적인 벤치마크 문제인 Iris 데이타에 대해서 실험을 하였다. Iris 데이타는 4개의 수치 특성값(꽃잎의 길이와 폭, 꽃받침의 길이와 폭)을 갖는 150개의 붓꽃(iris) 데이타를 3개의 클래스로 분류해 놓은 것이다. 생성된 결정트리의 특성을 알아보기 위해, 전체 학습 데이타들을 임의로 5등분하여 그중 4개는 결정트리에 생성을 위한 학습 데이타로 사용하고 나머지 1개는 평가를 위한 검증 데이타로 사용하는데, 5개의 데이타 집합이 모두 한번씩 검증 데이타 역할을 하도록 결정트리 생성과 검증을 총 5번 수행하는 5중 교차 검증(5-fold cross validation)을 10회 수행하였다. (그림 9)는 Iris 데이타에 대해서 제안한 학습 방법을 이용하여 구한 퍼지 사선형 결정트리의 형태 및 각 노드의 특징값을 보인 것이다. 여기에서 k_{TR}, k_{TL}은 (그림 7)에 보인 것과 같이 퍼지 경계영역을 정의하는 파라미터 값이고, x_1, x_2, x_3, x_4은 각각 붓꽃 꽃잎의 길이와 폭, 꽃받침의 길이와 폭에 대한 특징값이고, $H(\cdot)$는 해당 노드에 대응하는 초평면의 방정식이다.



H_{Root} 노드 :

$k_{T_R} = 1.827,\ k_{T_L} = -6.934$

$H_{\text{Root}}(x_1, x_2, x_3, x_4) = -6.291x_1 - 0.543x_2 + 1.735x_3 - 0.115x_4 + 1.254$

H_l 노드

$k_{T_R} = 1.423, k_{T_L} = -0.562$

$H_l(x_1, x_2, x_3, x_4) = 1.001x_1 - 0.0003x_2 - 0.0004x_3 - 0.0009x_4 - 16.501$

그림 9 Iris 데이타에 대한 퍼지 사선형 결정트리

실험에서는 먼저 퍼지 경계면을 도입하지 않은 OC1 알고리즘에 의해서 생성된 일반 사선형 결정트리와 퍼지 경계면을 도입한 제안한 방법에 의해 생성된 퍼지 사선형 결정트리와의 분류 정확도를 비교하였다. 실험에서 데이타의 자식노드로의 전달을 위한 소속정도 임계값으로 0.01을 사용하였고, 퍼지 적용비율을 1%, 2%, 3%, 3%, 5%, 10%, 15%에 대해서 실험하였다. 퍼지 적용비율이 0%인 것은 OC1에 의해 생성된 일반 사선형 결정트리이다. (그림 10)는 각 퍼지 적용비율에 대해서 5중 교차검증을 10회 수행하여, 분류 정확도의 평균값을 보인 것이다. 그림에서 보는 바와 같이 퍼지 적용비율

0%인 기존 OC1에 기반한 일반 사선형 결정트리에 비해, 퍼지 적용비율을 3~10% 정도 적용한 구간에서의 퍼지 사선형 결정트리의 분류 정확도가 0.2~0.6% 정도 향상될 수 있음을 확인할 수 있다. 퍼지 적용비율이 지나치게(15% 이상) 커지게 되면 특징공간 상에서 다른 클래스에 속하는 영역간에 겹치는 부분이 넓어지게 되어, 퍼지 사선형 결정트리의 분류 특성이 나빠지는 결과가 나왔다. 이와 같은 결과가 발생하는 것은, 퍼지 적용비율이 2개 이상의 부류를 허용하는 퍼지 경계영역의 크기를 결정하는데, 퍼지 적용비율이 커지면 이러한 퍼지 경계영역이 커지게 되고, 소속정도의 차이가 미미한 퍼지 경계영역 부분에서는 퍼지 사선형 결정트리의 분별력이 떨어지기 때문으로 판단된다.

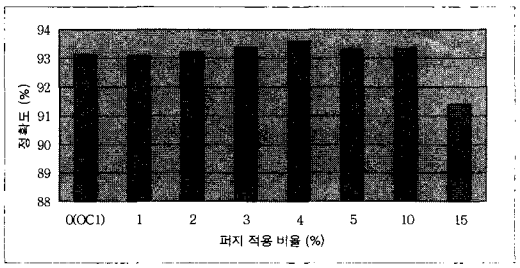


그림 10 Iris 데이타에 대한 결정트리의 성능 평가

퍼지 사선형 결정트리는 (그림 10)의 실험 결과에서 보는 바와 같이 퍼지 적용비율 조정에 따라 기존 사선형 결정트리보다 향상된 분류 정확도를 보일 수 있다. 한편, 특징공간에서 클래스 경계영역 근처의 데이타에 대해서 기존 사선형 결정트리는 데이타를 하나의 클래스로 단정적으로 분류하는 반면에, 제안된 퍼지 사선형 결정트리는 데이타가 위치한 영역 부근 여러 클래스에 대한 정보를 확신도와 함께 제공하기 때문에, 의사결정자가 최종 의사결정에 이들 정보를 유용하게 사용할 수 있다. 퍼지 사선형 결정트리에 의해서 분류된 데이타에 대한 가장 큰 확신도를 갖는 클래스와 두 번째로 큰 확신도를 갖는 클래스 간의 확신도 차이가 근소하면, 의사결정자가 의사결정을 할 때 가장 큰 확신도를 갖는 클래스만을 고려하는 것이 아니라 두번째로 큰 확신도를 갖는 클래스도 참고할 것이다. 따라서 다음 실험에서는 가장 큰 확신도를 갖는 클래스와 그 다음으로 큰 확신도를 갖는 클래스의 확신도의 크기비가 식 (10)과 같이 어느 정도 (a) 이상일 때, 데이타가 실제 속하는 클래스가, 의사결정자가 고려하게 될 두 클래스 중 어느 하나와 일치하는 경우가 얼마나 되는지 실험하였다.

$$\frac{C_{2ndClass}(x)}{C_{1stClass}(x)} \geq a \quad (0 < a \leq 1) \tag{10}$$

위 식에서 a값은 데이터 x에 대해 두 번째로 큰 확신도를 갖는 클래스의 확신도값 ($C_{2and\ classt}(x)$)이 가장 큰 확신도를 갖는 클래스의 확신도값 ($C_{1st\ class}(x)$)과 얼마나 근접한지를 나타낸다. (그림 11)은 Iris 데이타에 대해서 a값을 0.8로 사용하였을 때, 데이타의 실제 클래스가 퍼지 사선형 결정트리에 의해 판정된 가장 큰 확신도를 갖는 클래스와 일치하는 비율(1st에 의해 표현된 것)과, 실제 클래스가 확신도가 가장 큰 클래스와 그 다음으로 큰 클래스 중 하나와 일치하는 비율(1st & 2nd에 의해 표현된 것)을 나타낸 것이다. (그림 11)에서 유추할 수 있듯이 최종 의사결정시에 확신도 정보를 활용하는 경우에는 그렇지 않은 경우보다 정확한 분류를 할 가능성을 높일 수 있다. 퍼지 사선형 결정트리가 확신도를 갖는 클래스 정보를 제공하는 것은 의사결정 측면에서 볼 때 매우 유용한 특성이다.

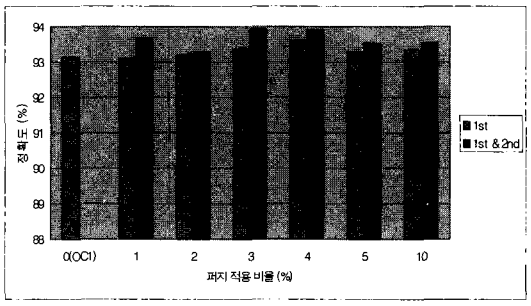


그림 11 확신도가 근사할 경우 두 클래스를 고려한 경우

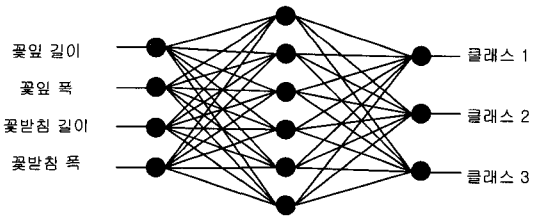


그림 12 Iris 데이타 분류를 위해 사용한 다층 퍼셉트론

또한 제안한 퍼지 사선형 결정트리의 유용성을 보이기 위해 대표적인 분류 방법중의 하나인 다층 퍼셉트론(multi-layer perceptron)[17]을 이용하여 Iris 데이타를 분류하는 실험을 하였다. 실험에서는 (그림

12)와 같은 입력층에 Iris 데이타의 4개의 특징값을 받아들이기 위해 4개의 노드를 갖고, 출력층에는 3개의 클래스 각각에 대응하는 3개의 노드가 있고, 하나의 은닉층(hidden layer)에 6개의 노드가 있는 다층 퍼셉트론을 이용하였다.

네 개의 특징값과 하나의 클래스 값을 갖은 원래 학습데이타(예, (*a, b, c, d*, 클래스 1), 여기에서 *a, b, c, d*는 특징값)를, 목표 클래스에 대응하는 출력노드에는 0.8, 나머지 출력노드에는 0.2값을 갖도록(즉, (*a, b, c, d*, 0.8, 0.2, 0.2)) 변환하여, 다층 퍼셉트론 학습에 이용하였다. 출력값으로 1과 0값 대신 0.8, 0.2를 사용한 것은 0값을 사용한 것은, 오차역전파(Error Backpropagation) 알고리즘[17]을 적용할 때 출력값이 0이거나 0에 가깝게 되는 경우, 역전파되는 연결강도 수정항의 값이 없거나 너무 작아져서 연결강도의 학습이 더디게 일어나는 경향이 있기 때문이다. 실험에서는 여러번의 시행착오를 통해 결정한 모멘트(moment) 값 0.9, 학습율 0.7, 학습횟수 1,000,000를 사용하여, 5중 교차검증을 10회 수행하였다. 다층 퍼셉트론의 분류 성능은 무작위로 초기화되는 노드간 연결강도에 크게 영향을 받는 경향을 보인다. 따라서 실험에서는, 다층 퍼셉트론을 실제 현장에 적용할 때 학습 데이타에 대한 분류 정확도가 만족할 수준인 것만을 적용한다는 전제하에, 주어진 학습 데이타와 검증 데이타 각 쌍에 대해서 30번의 학습과 검증 실험을 하여 학습 데이타에 대한 분류 성능이 85%미만인 것들의 실험결과는 무시하고, 나머지 것들에 대해서 평균을 구하였다. 실험에서 사용한 다층 퍼셉트론은 Iris 데이타에 대해서 5중 교차검증을 하였을 때 평균적으로 93.0%의 분류 정확도를 보였다.

제안한 퍼지 사선형 결정트리는 (그림 11)에 보인바와 같은 2개의 분류 경계면(즉, 2개의 내부 노드)를 가지고 93.1 ~ 93.6% 정도의 분류 정확도를 보인 반면, 실험에서 사용한 6개의 분류 경계면(즉, 6개의 은닉층 노드)을 갖는 다층 퍼셉트론은 93.0%의 분류 정확도를 보였다. 따라서, 퍼지 사선형 결정트리는 Iris 데이타에 대해서는 다층 퍼셉트론 이상의 분류 성능을 보임을 확인할 수 있었다. 한편, 다층 퍼셉트론은 초기 연결강도값에 분류 성능이 크게 영향을 받는 반면, 퍼지 사선형 결정트리 생성 방법은 초기 설정값에 비교적 둔감한 특성을 보였다. 또한 퍼지 사선형 결정트리는 트리 구조를 갖기 때문에 다층 퍼셉트론에 비해 구조화된 형태의 분류 규칙 정보를 추출하고 분석하는 것이 상대적으로 용이하다.

5 결 론

이 논문에서는 수치 특징값으로 기술되는 데이타들로부터 다변수 결정트리 형태의 퍼지 사선형 결정트리를 생성하는 방법을 제안하였다. 제안한 퍼지 사선형 결정트리는 소속함수에 의해 정의된 퍼지 경계면을 사용하여 특징공간을 분할하기 때문에, 분명한 경계면을 사용하는 사선형 결정트리보다는 데이타 특징값의 작은 오류에 대한 민감도가 완화되는 장점이 있다. 따라서 실제 현장에서 얻어진 학습 데이타로부터 분류규칙을 추출하기 위해 결정트리를 생성할 때는, 이 논문에서 제안한 퍼지 사선형 결정트리 생성방법을 유용하게 사용될 수 있다.

퍼지 사선형 결정트리는 일반 사선형 결정트리와는 달리 분류되는 데이타에 대해 확신도와 함께 복수개의 클래스 정보를 제공한다. 따라서 특징공간에서 클래스 경계영역 근처의 데이타에 대해서 기존 사선형 결정트리는 데이타를 하나의 클래스로 단정적으로 분류하는 반면에, 제안된 퍼지 사선형 결정트리는 데이타가 위치한 영역 부근의 여러 클래스 정보를 확신도와 함께 제공하기 때문에 의사결정자가 최종 의사결정에 이들 정보를 유용하게 사용할 수 있다.

퍼지 사선형 결정트리가 내부적으로 표현된 분류규칙을 다소 이해하기 힘들고, 결정트리 생성 소요 시간이 다소 많이 걸리는 단점은 있으나, 분류 정확도 향상 및 부가적인 분류 정보 제공등과 같은 점은 기존 사선형 결정트리에 대한 퍼지 사선형 결정트리의 장점이다. 퍼지 적용비율에 따라 정의되는 퍼지 경계면을 이용하는 퍼지 사선형 결정트리가 실험을 통해서 확인한 바와 같이 일반 사선형 결정트리보다 분류 정확도를 향상시킬 수 있지만, 아직 어느 정도의 퍼지 적용비율을 적용하는 것이 바람직한가에 대해서는 충분한 연구가 이루어지지 않았으므로, 이에 대해서는 향후 연구가 필요하다. 향후 연구로서 우선 고려하는 것은 제안한 방법에 유전자 알고리즘을 도입하여 퍼지 적용비율을 변경하면서 퍼지 사선형 결정트리를 생성하는 방법과, 모든 퍼지 경계면이 동일한 퍼지 적용비율을 갖는 것이 아니라 퍼지 경계면별로 임의의 퍼지 적용비율을 가질 수 있도록 하는 퍼지 사선형 결정트리를 생성하는 방법에 대한 연구이다.

참 고 문 헌

[1] C. E. Brodley, P. E. Utgoff, Multivariate versus univariate decision trees, Technical Report 92-8, Dep. of Computer Science, Univ. of Massachusetts, Amherst, MA, 1992.

[2] Y. Yuan, M. J. Shaw, Induction of fuzzy decision trees, *International Journal for Fuzzy Sets and*

Systems, Vol. 69, pp.125-139, 1995.
[3] S. K. Murthy, S. Kasif, S. Salzberg, A System for Induction of Oblique Decision Trees, *Journal of Artificial Intelligence Research* Vol. 2, pp.1-33, 1994.
[4] J.R. Quinlan, *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA. 1993
[5] G. J. Klir, T. A. Folger, *Fuzzy Sets, Uncertainty, and Information*, Prentice-Hall, 1992.
[6] C. E. Brodley, P. E. Utgoff, Multivariate decision trees, *Machine Learning*, Vol.19, pp.45-77, 1995.
[7] C. Z. Janikow, Fuzzy Decision Trees: Issues and Methods, *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*, Vol.28, No. 1, pp. 1-14, 1998.
[8] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees*, Wadsworth International Group, 1984.
[9] J.R. Quinlan, Induction trees at probabilistic classifiers, *Proc. 4th Int. Workshop on Machine Learning*, pp.31-37, 1987
[10] J. Zeidler, M. Schlosser, Continuous-Valued Attributes in Fuzzy Decision Trees, *Proc. of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96)*, Granada, Spain, Volume I, pp. 395-400. 1996.
[11] J. R. Quinlan, Induction on Decision Trees, *Machine Learning*, Vol.1, 1986, pp.81-106.
[12] L. A. Breslow, D. W. Aha, Simplifying decision trees: A Survey, *NCARAI Technical Report No. AIC-96-014*, Naval Research Laboratory, 1996.
[13] T. M. Mitchell, *Machine Learning*, The McGraw-Hill Co.,414p, 1997.
[14] X. Wang, J. Hong, Learning optimization in simplifying fuzzy rules, *Fuzzy Sets and Systems*, Vol.106, pp.349-356, 1999.
[15] K.-M. Lee, K.M. Lee, J.-H. Lee, H. Lee-Kwang, A Fuzzy Decision Tree Induction Method for Fuzzy Data, *Proc. of Int. Conf. on FUZZ-IEEE*, Seoul, Korea, pp.16-21, 1999.
[16] H. Kim, L. Fu, Generalization and Fault Tolerance in Rule-based Neural Networks, Proc. of 1994 IEEE Conf. on Neural Networks, Vol.3, pp.1550-1555, 1994.
[17] y. -H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley, 1989.

부 록

다음은 제안한 퍼지 사선형 결정트리 생성 방법에서 기본 경계면으로 사용할 초평면을 구하기 위해 사용하는 OC1의 초평면 탐색 알고리즘[3]을 기술한 것이다.

procedure OC1의 초평면 탐색 알고리즘

입력 : 초평면을 이용해 분할할 데이타의 집합 T

출력 : 데이타의 집합 T를 양분하는 최적의 초평면 H

begin

1. 데이타의 집합 T를 양분하는 축에 평행한(axis-parallel) 초평면 H_0를 찾고, 이 초평면에 의해 분할될 때의 데이타 집합 T의 혼잡도(impurity) I를 계산한다.
2. 다음 과정을 미리 지정한 R번 반복한다.
   2.1 임의의 초평면 H를 선택한다(첫 수행시에는 H를 과정 1에서 구한 축에 평행한 초평면 H_0로 초기화한다).
   2.2 혼잡도 척도 값이 더 이상 개선되지 않을 때까지 H의 각 계수 a_i를 순차적으로 *Perturbation* 한다.
   2.3 다음 과정을 미리 지정한 J번 반복한다.

   임의의 방향을 선택하여, 해당 방향으로 H의 각 계수를 무작위로 약간씩 수정하여, 이때 H에 의한 혼잡도가 줄어들면 단계 2.2로 간다.

   2.4 H에 대한 혼잡도 I_1를 구하고, $I_1 < I$이면 I에 I_1값을 저장한다.
3. I값을 갖는 초평면 H을 반환한다.

end.

위에 기술한 OC1의 초평면 탐색 알고리즘에서 사용하는 *Perturbation*은 다음과 같은 방법으로 초평면 방정식 $H(x) = a_1x_1 + a_2x_2 + \ldots + a_dx_d + a_{d+1}$의 계수값을 조정한다.

procedure *Perturbation*

입력 : 초평면 H, 초평면 H의 조정할 계수의 첨자 m

출력 : 조정된 계수값 a_m

begin

1. 모든 데이타 x^i $(i=1,\ldots,n)$에 대해서 다음과 같이 U_i를 계산하여, 오름차순으로 정렬한다.

$$U_i = \frac{a_m x^i_m - H(x^i)}{x^i_m}$$

2. 정렬한 U_i값의 배열에서 인접한 두 값 사이의 중간값중에서 가장 좋은 분할 특성을 보이는 값을 새로운 계수의 후보 a'_m로 한다.
3. H의 a_m을 a'_m로 대치해서 만들어지는 초평면을 H_1으로 한다.

if (H_1에 의한 혼잡도 < H에 의한 혼잡도)

H의 a_m을 a'_m으로 대치하고, P_{move}를 P_{stag}로 대치한다.

else if (H_1에 의한 혼잡도=H에 의한 혼잡도)

H의 a_m을 P_{move}의 확률로 a'_m으로 대치한다.

P_{move}를 $P_{move}-0.1P_{stag}$로 대치한다.

endif

end.

위 알고리즘에서 P_{move}는 계수에 대한 변경을 확률적으로 수행하기 위한, P_{stag}에 의해 결정되는 확률값이다. P_{stag}는 혼잡도의 개선이 없을 때 계수 값을 변경시키지 않을 확률값으로, 초기값으로 1을 가지며 초평면 탐색 알고리즘이 진행되면서 값이 점점 줄어들도록 설정된다. 혼잡도 척도는 초평면의 분할 특성을 평가하는 척도로서, 대표적인 혼잡도 척도로는 이분 규칙(Towing rule), 지니 계수(gini index), 엔트로피 척도 등이 있다[3, 4, 13].

이 우 향

1998년 2월 충북대학교 수학과 졸업. 2000년 2월 충북대학교 전자계산학과 석사. 2000년 2월 ~ 현재 지식시스템(주) 연구원. 관심분야는 Rule-based System, Fuzzy Theory, Data Mining, Knowledge Management Sytem, CRM, Billing Automation

이 건 명

1990년 2월 한국과학기술원 전산학과 학사. 1992년 2월 한국과학기술원 전산학과 석사. 1995년 2월 한국과학기술원 전산학과 박사. 1995년 5월 ~ 1996년 4월 프랑스 INSA de Lyon 연구원. 1996년 5월 ~ 1996년 8월 미국 PSI 연구원. 1996년 9월 ~ 현재 충북대학교 컴퓨터과학과 조교수. 관심분야는 에니전트 시스템, 전자상거래, 무선 인터넷 응용, 데이타 마이닝, 소프트 컴퓨팅