데이터 마이닝을 위한 퍼지 결정트리

이중근[1*], 민창우[1**], 김명원[2]
한국 IBM 소프트웨어 연구소[1]
숭실대학교 컴퓨터학부[2]

A Fuzzy Decision Tree for Data Mining

Joong Geun Lee, ChangWoo Min, Myung Won Kim
Korea Software Development Institute, IBM Korea[1*]
School of Computing, Soongsil Univ.[2]

요 약

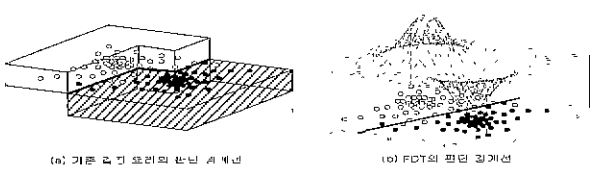
사회 전 분야에서 데이터가 폭발적으로 증가함에 따라 데이터를 이해하고 분석하는 새로운 자동적이고 지능적인 데이터 분석 도구의 기술이 필요하게 되었다. KDD(Knowledge Discovery in Databases)는 이러한 필요로부터 데이터에서 유용하고 이해 가능한 지식을 추출하는 연구이다. 데이터 마이닝(Data Mining)은 KDD에서 가장 중요한 단계로 데이터로부터 지식을 추출하는 단계이다. 데이터 마이닝에서 생성된 지식은 좋은 분류율을 가지고 이해하고 이해하기 쉬워야한다. 본 논문에서는 퍼지 결정트리(FDT : Fuzzy Decision Tree)에 기반한 효율적인 데이터 마이닝 알고리즘을 제안한다. FDT의 각 링크는 속성(attribute) 값을 갖는 퍼지 집합이며, FDT의 각 경로는 퍼지 규칙을 생성한다. 제안된 알고리즘은 ID3의 이해성과 퍼지이론의 추론과 표현력을 결합한 방법으로 히스토그램에 기반한 각 속성별 소속 함수 생성 단계와 ID3와 유사한 방법을 이용하여 결정트리를 생성하는 두 단계로 이루어진다. 마지막으로 제안된 방법의 타당성을 검증하기 위해 표준적인 패턴 분류 벤치마크 데이터에 대한 실험 결과를 보인다.

1 서론

데이터 마이닝을 통하여 생성된 지식을 활용하여 일반 기업에서는 소비자의 미래 소비 성향이나 제품에 대한 판매 예측, 금융업계에서는 고객의 신용 등급, 금리나 환율의 변동 예측, 공장에서는 제품 생산비용에 대한 예측, 의학 분야에서는 환자의 병원 방문 예측이나 효과적인 의학 치료방법 예측과 같은 중요한 지식을 이끌어 낼 수 있게 된다. 하지만 실세 응용 분야에서 수집된 데이터는 잡음이 섞인 데이터나 데이터의 특성상 모호한 값들을 가지는 데이터들도 많이 존재하게 된다. 또한 데이터 마이닝의 단계를 거쳐 생성된 지식이 인간이 이해하기에 어렵거나 정확도가 떨어지게 된다면 생성된 지식이 의미가 없게된다[1].

현재 많이 이용되는 데이터 마이닝 기술은 통계 자료 분석, 기호주의 학습(symbolic learning), 신경망, 시각화(visualization) 등이 있다. 본 논문에서는 보다 효과적인 방법을 이용하여 지식을 추출하는 것으로 군집화(clustering), 분류(classification), 패턴(pattern), 연관성 발견(association discovery) 등에 대한 규칙 생성을 목적으로 한다. 규칙 생성을 목적으로 하는 기존의 데이터 마이닝 기술은 결정트리 생성 방법, 퍼지 공간 분할 방법, 퍼지 신경망 등이 있다.

기존의 결정트리 생성 방법[2,3,4]의 경우 규칙의 형태가 생성 규칙(production rule)으로 이해하기 쉬운 형태이나 패턴이 복잡하게 분포되어 있는 경우 분류율이 나빠지는 단점이 있으며, 퍼지 공간 분할 방법[8]의 경우 규칙의 형태가 퍼지 규칙으로 패턴이 복잡하게 분포되어 있어도 비교적 분류율이 좋으나 생성되는 규칙의 수가 많고 각 규칙의 조건부의 항의 수가 데이터의 속성(attribute) 수와 같으므로 속성수가 많은 실제 데이터에 적용하여 생성된 규칙은 이해하기 어렵다. 퍼지 신경망[9]의 경우 신경망의 학습 기능을 퍼지와 결합함으로써 분류율은 좋으나, 규칙의 형태가 조건부에 가중치가 결합된 퍼지 규칙의 형태를 가지며, 각 규칙의 조건부 항의 수가 퍼지 공간 분할 방법과 같이 데이터의 속성 수와 같으므로 생성된 규칙을 이해하기 어렵다.



(a) 기존 결정 트리의 분할 경계선 (b) FDT의 퍼지 경계선

그림 1 기존의 결정트리와 FDT의 분할 경계선 비교

퍼지이론은 언어적 불확실성을 퍼지 집합의 개념을 이용하여 정량적으로 표현하며 제어, 패턴분류 문제 등 여러 분야에서 활용되고 있다. 또한 잡음이 있거나 불확실한 데이터에도 잘 적용되며 언어적 불확실성을 표현하는데 있어서 기호주의적 틀을 유지하므로 이해가 쉽고 규칙들이 이후는 퍼지경계선이 속성 축에 평행 할 필요가 없으며 추론을 할 때 규칙들의 합성 방법을 이용하여 신뢰성 있는 추론을 한다(그림 1).

따라서 본 논문에서는 실제 응용분야에서 수집된 데이터들로부터 인간이 이해하기 쉽고 데이터들의 특성을 잘 기술할 수 있는 퍼지 규칙을 생성하는 데이터 마이닝 알고리즘으로서 퍼지 결정트리(FDT : Fuzzy Decision Tree)를 제안한다. FDT는 간결한 규칙을 생성하는 결정트리 생성 알고리즘인 ID3에 퍼지 이론의 추론과 표현력을 결합한 방법이다.

본 논문의 구성은 다음과 같다. 2장에서는 퍼지 결정트리 알고리즘을 제안한다. 3장에서는 제안된 알고리즘을 이용하여 표준적인 패턴 분류 벤치마크 데이터들에 대한 실험 결과를 비교 분석한다. 마지막 장에서는 결론 및 향후 연구 방향을 제시한다.

2. 퍼지 결정트리

2.1 규칙의 형태 및 추론 방법

본 논문에서는 사람이 이해하기 쉽도록 단순한 형태의 퍼지 규칙 생성 방법을 사용한다. 각 퍼지 규칙은 식 (1)과 같이 다수 개의 퍼지 집합으로 구성되어 있는 조건절과 1개의 결론을 가지는 MISO(Multiple Input Single Output) 형태의 규칙을 사용하며 각 규칙에는 규칙에 대한 CF(Certainty Factor)가 있다.

$$Rule_i: \text{if } x_{i1} \text{ is } U_{i1} \text{ and } x_{i2} \text{ is } U_{i2} \ldots \text{ and } x_{in} \text{ is } U_{in} \text{ then } Class_j\ (CF_i) \quad (1)$$

식 (1)에서 x_{in}은 i번째 규칙의 n번째 속성을 나타내며 U_{in}은 i번째 규칙의 n번째 퍼지 소속 함수를 나타낸다. $Class_j$는 j번째 클래스를 나타낸다.

추론을 하기 위해서는 두 가지 연산자를 결정해야 한다. 각 조건절의 소속 정도(degree of membership) 합성 방법과 결론부의 합성 방법을 결정해야 한다. 본 논문에서는 추론 방법으로 많이 쓰이는 방법으로써 식 (2)와 같이 조건절의 소속 정도 합성 방법에 최소 연산자를 사용하며 결론부의 합성 방법으로 대수적 합 연산자를 사용한다.

$$Concl_i = \sum_i \min_A(\mu_{U_{in}}) * CF_i \quad (2)$$

식 (2)에서 $Concl_i$는 결론이 i번째 클래스를 나타내는 규칙들의 합을 나타내며 $Concl_i$가 가장 큰 i번째 클래스를 최종 결론으로 한다

2.2 퍼지 소속 함수의 생성

퍼지 소속 함수는 퍼지 규칙에서 가장 중요한 요소로 삼각형, 사다리꼴, 기우시안 함수 형태의 퍼지 소속 함수가 많이 사용된다. 논 논문에서는 각 클래스의 각 속성에 대한 히스토그램에서의 극대점과 극소점을 이용하여 삼각형 형태의 퍼지 소속 함수를 생성한다. 그러나 성별과 같이 이산적인(distributed) 속성은 삼각형의 체점이 일치하는 소속 함수를 사용하였다. 히스토그램은 데이터 분포에 대한 통계적 특성을 나타내고 있음으로 히스토그램을 이용하여 소속 함수를 생성할 경우 효율적인 FDT를 생성할 수 있다. 삼각형 형태의 소속 함수는 수식 (3)과 같이 정의된다

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

데이터의 수가 적은 경우 일반적으로 히스토그램에는 많은 수의 극대점과 극소점이 있게 된다. 따라서 퍼지 소속 함수를 생성하기 위해 먼저 각 클래스의 각 속성에 대한 히스토그램을 작성하고 N번 평활화(smoothing)함으로써 작은 극대점과 극소점을 없앤다. 보다 구분력 있는 퍼지 소속 함수를 생성하기 위하여 각 속성의 각 클래스별로 히스토그램을 생성하였다.

두 번째 단계로 평활화된 히스토그램에서 극대점과 극소점을 찾는다. 본 논문에서는 히스토그램에 대한 평균변화율을 이용하여 극대점과 극소점을 찾는다. x에 대한 히스토그램을 $h(x)$라고 할 때 $h(x)$의 평균변화율은 식 (4)과 같이 정의된다. 이때 극대값이나 극소점이 평지를 이루고 있으면 그 중심점을 극대점이나 극소점으로 한다

$$\Delta h(x)=h(x)-h(x-1) \tag{4}$$

세 번째 단계는 각각의 극대점과 극대점의 양쪽 방향으로 가장 가까운 극소값을 직선으로 연결한다. 이때 각 각선의 x 절편을 삼각형의 소속 함수의 l, r로 설정하고 극대점을 c로 설정함으로써 삼각형 형태의 퍼지 소속 함수를 생성한다. 그림 2는 히스토그램을 이용하여 퍼지 소속 함수를 생성하는 과정을 보여준다

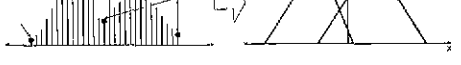


그림 2 히스토그램에 의한 소속 함수의 생성

2.3 퍼지 결정트리 생성 알고리즘

실세계에서 수집되는 데이터들은 모든 속성에 대한 값을 가지고 작성되지 않는다. 따라서 거래 되지 않은 값을 가진 훈련 데이터와 테스트 데이터에 대한 처리가 필요하다. 제안된 퍼지 결정트리에서는 히스토그램 특성상 반드시 이러한 데이터의 속성 값은 부지에서 제외되었으며 트리 생성시에 소속값이 없기 때문에 부정적도 제신에 반영하지 않았다. 테스트 데이터에서 알 수 없는 데이터의 속성 값은 부시하고 소속 값을 갖는 다른 속성들의 소속값만을 이용하여 추론하였다

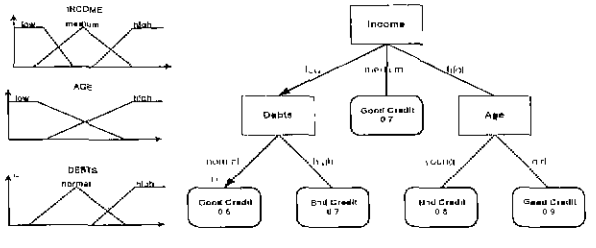


그림 3 고객 신용도 평가를 위한 FDT

FDT는 그림 3에 같이 비단말 노드(nonterminal node)인 Income, Debts, Age와 단말 노드(terminal node)인 Good Credit, Bad Credit 그리고 low, medium, high와 같은 링크(link)으로 구성되어 있다. 비단말 노드는 분기를 위한 속성을 가지고 있으며, 단말 노드는 클래스명과 CF를 가지고 있다. 노드와 노드 사이를 연결하는 링크는 속성에 대한 퍼지 소속 함수이다

FDT를 생성하기 위한 알고리즘은 기존의 ID3, C4.5에서 무질서도의 개선에 퍼지 개념을 결합함으로서 이루어진다. 결정트리의 근 노드(root node)에서 i번째 노드까지의 퍼지 소속 함수가 규정하는 부공간(subspace)에 대한 l번째 훈련 데이터의 소속 정도는 수식 (5)와 같이 계산된다

$$m_{il}=\begin{cases}\min_{f\in Fset_i}(\mu_f(d_{lf})) & Fset_i\neq\emptyset\\ 1 & Fset_i=\emptyset\end{cases} \tag{5}$$

수식 (5)에서 $Fset_i=\{F_1, F_2, \cdots, F_i\}$이며 결정트리의 근 노드에서 i번째 노드까지의 퍼지 집합을 나타내며 d_{lf}는 l번째 훈련 데이터에 대한 퍼지 소속 함수 f에 대응하는 속성값을 나타낸다

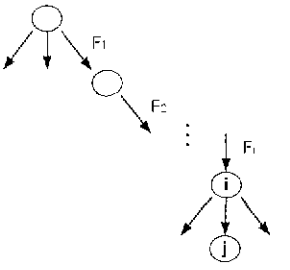


그림 4 퍼지 결정트리 생성 과정

클래스가 m개이며 데이터의 수가 n개의 데이터 집합에 대한 무질서도를 계산하는데 있어서 데이터가 부공간에 속하는 정도를 위에서 정의한 m_{il}을 이용하여 계산한다. i번째 노드에서 속성 F에 대한 무질서도 E_F는 다음과 같다(그림 4).

$$E'_F=\sum_j(p_{ij}I^j) \tag{6}$$

$$I^j=-\sum_{k=1}^{m}(p'_k\log_2 p'_k) \tag{7}$$

$$p'_k=\frac{\sum_{c\in k}m_{jc}}{\sum_{l=1}^{n}m_{jl}} \tag{8}$$

$$p_{ij}=\frac{\sum_{l=1}^{n}m_{jl}}{\sum_{l=1}^{n}m_{il}} \tag{9}$$

수식 (6)에서 f는 속성 F의 퍼지 소속 함수를 가리키며, j는 i번째 노드에서 퍼지 소속 함수 f를 적용하여 생성된 자식 노드를 나타낸다. I^j는 속성 F에 대한 정보 이득(information gain)을 나타내며 p_{ij}는 훈련 데이터가 j번째 노드에 속할 가능성(possibility)을 나타낸다. 수식 (7)에서 m은 총 클래스 수를 나타내며 p'_k는 훈련 데이터가 j번째 노드에서 클래스 k에 속할 가능성을 나타낸다. 수식 (8)에서 c는 특정 클래스 k에 속한 데이터들을 나타내며 수식 (8)과 수식 (9)에서 n은 총 데이터 수를 나타낸다

다음은 FDT 생성 알고리즘이다.

단계 1> 모든 훈련 데이터에 대하여 부공간에 대한 소속정도 m_{il}을 1로 초기화한다

단계 2> 이래의 조건중 일부가 만족되면 자식 노드를 더 이상 확장하지 않고 결론부를 생성한다

1) $\frac{1}{n}\sum_j m_{ij}\le\theta_m$

2) $\frac{\sum_{c\in major\ class}m_{ic}}{\sum_j m_{ij}}\ge\theta_m$, 단 major class는 클래스별로 해당 부공간에 대한 소속 정도를 합했을 때 소속 정도의 합이 가장 큰 클래스를 나타낸다.

3) 더 이상 사용할 수 있는 속성이 없을 때, 위의 조건이 만족될 때 단말 노드에 대한 결론은 major class로 하며 $CF=\frac{\sum_{c\in major\ class}m_{ic}}{\sum_j m_{ij}}$ 이다

뒤게 3> 그렇지 않으면,
무질서도 E가 가장 낮은 속성을 선택하여 그 속성의 모든 퍼지 소속 함수에 대하여 지식 노드를 생성하고 각각의 자식 노드에 대하여 <단계 2>부터 알고리즘을 재귀적으로 적용한다

알고리즘의 <단계 2>에서 θ_n과 θ_m은 분할 종료 조건을 나타내는 메개변수이다. 퍼지 집합이 규정하는 부공간에 대한 데이터의 소속 정도의 합을 신체 훈련 데이터의 수로 나눈 값이 θ_n보다 작을 때 분할을 종료한다. 이는 퍼지 규칙이 과적응되는 것을 억제하며 퍼지 집합이 규정하는 부공간에 대한 소속 정도의 합이 가장 큰 클래스와 모든 클래스의 소속 정도의 합의 비율이 θ_m보다 클 경우 분할을 종료한다. 이것 역시 퍼지 규칙이 과적응되는 것을 억제한다. 이와 같은 방법으로 FDT를 생성한 후에는 FDT에서 사용된 퍼지 소속 함수에 대하여 사용자에게 질의를 통하여 퍼지 소속 함수에 대한 의미적 표현을 결정한다

3 실험 및 평가

이 장에서는 FDT 알고리즘의 타당성을 검증하기 위해 패턴 분류 문제에 표준적으로 사용되는 벤치마크 데이터들에 대하여 결정트리 생성의 대표적 알고리즘인 C4.5와 비교 실험한다

Iris 데이터는 Setosa, Versicolor, Virgina의 3개의 클래스로 구성되어 있는 데이터로 꽃받침(sepal)의 길이와 폭, 꽃잎(petal)의 길이와 폭의 4개 속성으로 기술된다. Breast Cancer Wisconsin 데이터는 위스콘신 대학 병원의 Wiliam H Wolberg가 수집한 데이터로 유방암 진단을 위한 환자들의 진료 상태에 대한 데이터이며 Credit Screening 데이터는 신용 카드 승인에 대한 데이터이다. Heart Disease 데이터는 환자들의 심장병 상태를 진단하기 위한 데이터이며 Sonar 데이터는 광물 탐사에서 사용하는 수중 음파 탐지기에서 얻어진 데이터로 광물의 여부를 판단한다. 실험에 사용된 데이터들은 UCI[11] 기계 학습 데이터 베이스로부터 얻을 수 있다. 각 데이터들에 대한 특성은 표 1과 같으며 실험에 사용된 데이터들은 무작위로 반씩 나누어 훈련 데이터와 테스트 데이터로 실험하였다

이 름	데이터 수	클래스 수	속 성	
			연속적	이산적
breast-w	699	2	9	-
credit-a	690	2	6	9
heart-c	303	2	8	5
Iris	150	3	4	-
sonar	208	2	60	-

표 1 실험 데이터 구성요소

3.1 퍼지 결정트리의 타당성 실험

C4.5[5]는 ID3 계열의 결정트리 알고리즘으로 속성이 연속적인 값을 보다 효율적으로 처리하기 위해서 Quinlan이 제안한 알고리즘이며 C4.5rl6]는 Breiman의 bagging과 Freund와 Schapire의 boosting 방법을 이용하여 C4.5 알고리즘을 개선한 것이다. C4.5[7]은 MDL (Minimum Description Length) 원리를 C4.5에 적용한 것이다. 표 2는 C4.5[5] 알고리즘의 실험 결과에 대하여 에러와 트리 크기를 FDT 방법과 비교한 것이며 표 3은 C4.5[6, 7] 알고리즘의 실험 결과에 대하여 에러와 규칙 수를 FDT 알고리즘과 비교한 것이다. 표 4는 각 실험 데이터에 사용된 메개변수를 나타낸다

	Rel 7 C4.5		Rel 8 C4.5		FDT	
	err(%)	tree size	err(%)	tree size	err(%)	tree size
breast-w	5.28 ± 0.9	20.3 ± 0.5	5.26 ± 1.9	25.0 ± 0.5	3.70	7
credit-a	15.80 ± 3.0	57.5 ± 1.2	14.70 ± 2.0	33.2 ± 1.1	13.60	14
heart-c	24.00 ± 1.0	45.3 ± 0.1	23.00 ± 5.0	39.9 ± 0.4	17.10	12
iris	4.87 ± 2.0	9.3 ± 0.1	4.80 ± 1.7	8.5 ± 0.0	4.00	4
sonar	28.40 ± 6.0	33.1 ± 0.5	25.60 ± 7.0	28.4 ± 0.2	18.30	8

표 2 C4.5와 FDT의 트리 크기와 에러율 비교

본 실험 결과 표 2와 표 3에서 볼 수 있듯이 본 논문에서 제안한 FDT가 C4.5 알고리즘 방법에 비하여 에러율이 적고 트리의 크기와 규칙 수가 적기 때문에 규칙이 간결하다. 따라서 본 논문에서 제안한 FDT에 의해 생성되는 규칙은 다른 방법과 비교하여 분류율이 높을 뿐만 아니라 규칙의 수가 적고 간결하여 이해가 쉬움을 알 수 있다

	C4.5	Bagged C4.5	Booting C4.5	RULES C4.5	FDT
	err(%)	err(%)	err(%)	err(%) 규칙수	err(%) 규칙수
breast-w	5.28	4.23	4.09	4.5 8.5	**3.70** **5**
credit-a	14.70	14.13	15.64	15.9 15.0	**13.60** **10**
heart-c	22.94	21.52	21.39	23.1 11.1	**17.10** **8**
iris	4.80	5.13	6.53	4.7 4.1	**4.00** **3**
sonar	25.62	23.80	19.62	31.1 7.0	**18.30** **7**

표 3 C4.5와 FDT의 규칙수와 에러율 비교

	N_c	θ_n	θ_m
breast-w	2	0.25	0.8
credit-a	3	0.09	0.8
heart-c	7	0.07	0.7
iris	10	0.07	0.7
sonar	3	0.47	0.7

표 4. 실험에 사용된 메개변수 값

5. 결론 및 향후 연구 과제

본 논문에서는 컴퓨터 사용의 보편화에 따라 데이터의 수집과 저장이 용이해지면서 데이터로부터 유용하고 이해 가능한 정보를 추출하기 위한 방법인 데이터 마이닝에 관한 연구이다. 기존의 데이터 마이닝 알고리즘들은 관단 경계선이 속성 축에 평행하지 않는 경우 분류가 어렵다는 단점과 생성되는 규칙의 수가 너무 많거나 퍼지 규칙의 조건부에 가중치가 결합된 형태이기 때문에 생성된 퍼지 규칙을 이해하는 것이 어려웠다

따라서 간결한 규칙을 생성하는 결정트리 생성 알고리즘과 퍼지를 결합함으로써 분류율이 높고 이해가 쉬운 퍼지 규칙을 생성할 수 있는 데이터 마이닝 알고리즘을 제안하였다. 제안한 알고리즘은 통계적 특성을 나타내고 있는 히스토그램을 이용하여 퍼지 소속 함수를 생성함으로써 더 효율적으로 FDT를 생성 할 수 있다. 실험을 통해 볼 수 있듯이 FDT 방법에 의해 생성되는 퍼지 규칙은 다른 방법과 비교하여 퍼지 규칙의 수가 적음은 물론 조건부가 짧고 규칙의 형태가 단순하여 이해하기 쉽다

향후 연구과제로는 추론의 정확성을 높이기 위해 새로운 퍼지 연산자의 개발과 생성된 규칙의 정확성을 높이기 위해 소속 함수를 조율(tuning)하는 연구가 필요하며 보다 간결한 규칙을 생성하기 위해 퍼지 규칙의 조건부를 절단(pruning)하는 방법에 대한 연구가 필요하다

참고 문헌

[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] 박명대, 이경근, ID3 계열의 귀납적 기계학습, 정보과학회지, 13권, 제 5호, pp 81-106, 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] Quinlan, J R, Improved use of continuous attributes in *C4.5*, *Journal of* Artificial Intelligence Research, vol 4, pp 77-90, 1996
[6] Quinlan, J R, Bagging, boosting, and C4.5, Proceedings 13th American Association for Artificial Intelligence National Conference on Artificial Intelligence, pp 725-730, AAAI Press, Menlo Park, CA, 1996
[7] Quinlan, J R, MDL and categorical theories(continued), Proceedings Twelfth International Conference on Machine Learning, pp 464-470, Morgan Kaufmann Montreal, 1995
[8] Ishibuchi, H, Nozaki, K, Tanaka, H, Effective fuzzy partition of pattern space for classification problems, *Fuzzy Sets and Systems*, Vol 59, pp 295-301, 1993
[9] 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
[10] Umano, 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
[11] http://www.ics.uci.edu/~mlearn/MLRepository.html