경영정보학연구
제6권 2호
1996년 12월

전문가시스템 실용화를 위한 지식오류분석방법론 연구[1)]

김 현 수[2)]

A Development of Knowledge Error Analysis Methodology for practical use of Expert Systems

The accuracy of knowledge is a major concern for expert system developers and users. Machine learning approaches have recently been found to be useful in knowledge acquisition for expert systems. However, the accuracy of concept acquired from machine learning could not be analyzed in most cases. In this paper we develop a comprehensive knowledge error analysis methodology for practical use of expert systems. Decision tree induction is an important type of machine learning method for business expert systems. Here we start to analyze with knowledge acquired from decision tree induction method, and extend the results to develop error analysis methodology for general machine learning methods. We give several examples and illustrations for these results. We also discuss the applicability of these results to multistrategy learning approaches.

1) 이 논문은 1995년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

2) 국민대학교 정보관리학과(Department of Management Information Systems, Kookmin University)

Ⅰ. 서 론

전문가시스템이 기업경영을 위한 도구로 고려되기 시작하면서부터, 시스템의 성능향상과 실용성 증진을 위한 연구가 꾸준하게 진행되어 왔다. 많은 연구자와 실무자들이 지식획득이 전문가시스템 실용화의 난제임을 인식하고, 자동학습(machine learning) 기법의 개발 및 개선에 노력하였다. 그러나, 전문가지식획득은 여전히 전문가시스템 실용화의 걸림돌이 되고 있으며, 자동학습방법은 지식획득의 유망한 대안임에도 불구하고 실제적인 활용이 부진한 상황이다. 실제적인 활용이 부진한 가장 큰 이유중의 하나가, 자동학습을 통하여 획득된 지식의 정확성과, 신뢰성을 검증할 수 있는 수단이 약하기 때문이다. 즉, 시험사례등을 통하여 부분적으로 검증된 지식을 실제 상황에서 사용하기에는 많은 위험을 감수해야하기 때문이다.

본 연구는 이와 같은 상황에서 자동학습 방법으로 획득되는 지식의 정확성과 신뢰성을 평가하고 검증할 수 있는 오류분석방법론을 개발하여, 시스템의 성능향상을 위한 이론적인 체계를 제시하고, 실용성을 증진하기 위한 수단을 제공하고자 한다.

경영분야에서 가장 많이 사용되는 자동학습 방법은 의사결정 사례(Examples)로부터 의사결정나무(Decision tree)를 추론하는 귀납추론방법(Induction Method)이다. 신경망(Neural Network)이나 유전자알고리즘(Genetic Algorithm)도 활용도가 높은 자동학습기법이다. 본 연구는 의사결정나무 추론 방법을 중심으로 오류분석방법론을 개발하고, 이를 신경망을 포함하는 일반적인 자동학습방법에 적용한다.

의사결정나무는 노드(node)와 가지(branch)로 구성되어 규칙들의 집합을 표시한다. 의사결정나무를 구축할때는 나무의 뿌리로부터 시작하여 잎으로 진행되는데 이러한 접근방법은 예측이나 설명을 하기위한 목적으로 직접 사용될 수도 있으며 [Braun and Chandler, 1987 ; Carter and Catlett, 1987 ; Messier and Hansen, 1988 ; Shaw and Gentry, 1988] 전문가시스템을 개발하기위한 지식획득의 방법으로 이용될 수도 있다.[Michalski and Chilausky, 1980 ; Quinlan, 1979, 1987b] 의사결정나무의 귀납적 추론방법은 대부분의 통계적 추론방법(discriminant analysis 등)이 근거하고 있는 파라메터나 구조적 가정(parametric or structural assumptions) 등과 무관하게 적용되는 방법이다.

많은 연구자들은 훈련에 이용된 사례의 집합인 훈련집합(training set)으로부터 구축된 대형 의사결정나무는 대개 전체 사례공간(instance space)에서 그 정확성을 유지하지 못한다는 사실을 발견하였다.[Breiman et al., 1984 ; Quinlan, 1983 ; Spangler et al., 1989] 이에 따라 많은 논문들이 훈련사례로부터 구축된 대형 의사결정나무를 단순화하는 기법(pruning)들에 대하여 연구하고 있다. [Fisher and Schlimmer, 1988 ; Mingers, 1986 ; Niblett, 1987 ; Quinlan, 1987a]

구축된 의사결정나무의 많은 가지들은 실제의 내포된 관계를 나타내기보다는 우연히 발생하는 특정 자료를 반영하는 경우가 빈번하다. 이렇게 우연히 발생한 자료들은 추후에 다시 발생할 가능성이 희박한 경우가 많다. 이러한 신뢰도가 낮은 가지들은 프루닝(pruning)에 의해 제거될 수 있기 때문에 프루닝된 의사결정나무는 비록 훈련집합(training set)에서는 오류 확률이 높지만 전체 사례공간(instance space)에서는 오류확률을 줄이고 보다 정확한 규칙으로 작용할 수 있다.

기존의 연구들이 추론된 개념의 정확성을 높인다는 측면에서 프루닝(pruning : 이하에서 "단순화기법"이라 칭한다)의 유용성을 경험적으로 증명하였지만, 이러한 연구결과들은 연구에서 선택한 특정한 훈련집합과 이들이 적용되는 분야(domain)에 깊이 의존하고 있다. 그러므로 이러한 연구결과들이 일반적으로 성립되는지 그리고 어느정도로 단순화기법이 개념(concepts)의 정확성을 높여주는지에 대해서는 알려져있지 않다.

본 논문에서는 학습되는 개념의 오류를 분석하는 제반 연구를 하나의 통합된 체계로 정립하고 부족한 이론을 개발하여, 이를 오류분석 방법론으로 구축한다. 우선, 우리는 일정수준 이하의 오류를 보장하는데 필요한 적정한 표본의 크기(sample size)에 대한 연구 결과를 요약하고, 그러한 충분한 양의 표본을 구하지 못하였을 경우 오류수준에 대한 사후 분석(posterior analysis)을 수행하는 여러가지 방법을 제공한다. 또한 사례로부터 지식을 추론하는 경우에 예상되는 오류수준에 대한 사전분석을 수행하는 과정과, 학습된 지식에 대한 오류의 사후분석을 수행하는 과정을 실제 경영문제를 통하여 적용함으로써 본 방법론이 통합된 체계로 활용될 수 있도록 하였다.

마지막으로 의사결정나무와 같은 단일 학습기법 뿐만 아니라, 여러가지 학습기법을 동시에 적용하여 학습된 지식의 오류를 분석할 수 있는 방법론으로 발전시키기 위한 탐색적인 연구를 수행하였다.

II 장에서는 일반적인 오류분석 이론에 대해 소개하고, 의사결정나무의 귀납적 추론 방법 및 단순화 기법에 이를 적용하여 원하는 오류수준을 보장하는데 필요한 사례(example)의 충분한 양(sufficient bound)을 제공한다. 이 결과는 오류분석을 위한 기초이론으로서 다음 장에서의 사후적 분석의 토대를 제공한다. III 장에서는 이미 학습이 완료된 지식이거나, 충분한 양의 사례를 구하기가 어렵거나 불가능한 경우에 추론된 지식의 정확성 및 신뢰성을 사후적으로 평가하는 방법을 제시한다. IV 장에서는 이 결과들을 예시하는 사례를 들어 실용성을 입증한다. V 장에서는 의사결정나무추론 방식이 신경망이나 유전자알고리즘과 결합되어 다단계 자동학습 전략으로서 보다 성능과 활용성을 강화할 수 있는지를 탐색적으로 분석하고, 이 경우에 앞에서 도출한 오류분석방법론을 적용할 수 있는 방안을 논의한다. VI 장에서는 요약 및 미래의 연구방향에 대하여 언급한다.

II. 자동학습의 오류분석 이론

본 단원에서는 자동학습(machine learning) 기법을 이용하여 학습된 지식이 전문가시스템의 지식기반으로 사용되는 경우를 위하여, 오류분석의 새로운 이론을 소개하고 대표적인 전문가시스템을 위한 지식획득 방법인 의사결정나무 추론 방법에 이를 적용하여 획득 지식에 대한 오류 분석을 수행한다.

3.1 자동학습 일반 이론

자동학습 방법을 분석하고, 학습 알고리즘의 성능을 평가하여 이를 알고리즘의 선택이나 개선에 반영하는 연구가 최근에 활성화 되고 있다. 일반적으로 학습알고리즘의 성능에 관계되는 요소는 학습된 개념의 정확성(concept accuracy), 속성의 증가에 따른 알고리즘의 효율성(scalability with input dimensionality), 관련없는 속성추가에 대한 알고리즘의 견고성(robustness), 메모리 사용 및 계산 자원 사용 효율성(complexity), 모형의 이해성(model comprehension) 등 다양한 요소가 있다. 본 연구에서는 이 중에서 학습된 개념의 정확성과 자원 사용의 효율성 관점에서 학습알고리즘을 분석하는 이론을 제시한다.

우리는 이를 표본 복잡도(sample complexity)와 계산 복잡도(computational complexity)로 나타낸다. 개념들의 상세정의는 다음과 같다.

정의 2.1 :

1) 표본 복잡도 : 표본복잡도(sample complexity)는 높은 신뢰도를 가지고 오류가 적은 개념(또는 가설, hypothesis)을 학습하는데 필요한 사례의 수이다. 이는 사례공간(instance space)의 모든 가능한 확률분포와 모든 가능한 목표개념(target concept) 상에서 최악의 경우에 필요로하게 되는 사례의 수를 취하는 것으로 정의된다. 여기서 사례공간은 관계되는 모든 가능한 사례를 생성할 수 있는 공간으로서, 주로 속성 벡터로서 표시된다.

2) 계산복잡도 : 계산복잡도(computational complexity)는 주어진 크기의 사례 표본으로부터 가설(hypothesis)을 생성하는데 소요되는 최대 계산 시간을 의미한다.

이 두가지의 복잡도는 O( )로 표시한다.

예를 들어 기업의 신용평가를 위해 이에 관련되는 요소를 조사한 결과 유동비율, 부채비율, 매출신장율, 자기자본수익율 등 4가지 요소만이 관련있다고 가정하자. 그러면 이 문제의 사례공간 X는 다음과 같이 표시된다.

X=(유동비율, 부채비율, 매출신장율,
자기자본수익율)

따라서 4차원상의 실수 공간이 사례공간이 되고, 발생가능한 사례의 종류는 무한히 많게 된다. 그러나 각 변수의 공간을 유한개의 구간으로 나누어 값을 부여하면 발생가능한 사례의 수는 유한개가 된다. 예를 들어, 위의 X에서 첫번째 변수인 유동비율을(3.0이상, 3.0과 1.5사이, 1.5이하)의 3구간으로 나누어 각각 1, 2, 3으로 값을 부여하고, 나머지 변수들도 모두 3

개씩의 구간으로 나누어 값을 표현하면, 전체 발생가능한 사례의 종류는 $3^4=81$개가 된다. 여기에 신용평가 등급이 k개의 등급으로 나누어 진다면 $k*81$이 분류를 고려한 전체 사례 종류 수가 된다.

가설공간(hypothesis space)은 발생가능한 사례의 집합을 나타내는데, 규칙공간(rule space)이라고도 한다. 대표적인 가설공간의 예로는 의사결정나무(decision trees), DNF (disjunctive normal form), CNF(conjunctive normal form) 등이 있다. 가설공간을 이와 같이 제한하는 이유는 학습의 효율을 높이기 위해서이다. 즉, 어떤 방식으로든 가설공간을 제한하지 않으면 학습은 평균적으로 임의추측(random guessing)과 같은 성능을 보이기 때문이다. 이렇게 가설공간을 확률이 높은 공간으로 제한하는 과정을 귀납적 편향(inductive bias)이라고 한다.

학습알고리즘의 표본복잡도에 대해서 Vapnik[1982]이 처음으로 유의한 결과를 도출하였다. 다음 정리 2.1은 그의 결과를 이용한 정리이다.

정리 2.1 : N을 가설공간 H에 있는 규칙의 집합이라하고, f를 찾아내고자 하는 목표개념(target concept)이라고 하자. h가 최소한 아래 m의 크기를 가지는 사례들과 일치하는 가설이라하면 $\text{Prob}\{d(h,f) \geq \varepsilon\} \leq \delta$이 성립한다.

$$m=(1/\varepsilon)\ln(N/\delta)$$

여기서, d(h,f)는 학습된 개념 h의 오류를 의미하고 공식적으로는

$$d(h,f)=\text{Prob}\{x \in X : h(x) \neq f(x)\}$$

로 표시한다.

또한 ε은 오류수준, δ는 신뢰도 수준을 의미하며 이 값들은 모두 0과 1사이의 값을 가진다.

그러나 위의 정리는 적용이 상당히 제한적이며 많은 개선의 여지를 안고 있다. 즉, 앞서 언급한 바와 같이 실수공간의 변수가 있는 경우 N 이 유한한 값이 아닐 수 있다. 따라서, 보다 개선된 결과를 얻기 위하여 가설공간의 규칙의 수가 아닌 다른 척도의 개발이 필요하다. 이를 위해 아래와 같은 두개의 조합(combinatorial) 파라메터를 활용할 수 있다.[Haussler, 1988 ; Vapnik, 1982].

정의 2.2 : 성장함수 및 VC차원(*Growth function, VC dimension*)

1. 성장함수(Growth function : $\pi_H(m)$) : 크기가 m 인 임의의 집합의 원소를 가설공간 H상의 어떤 개념에 의해 양의 사례와 음의 사례로 양분할 수 있는 최대 가짓수를 의미하고, 이를 $\pi_H(m)$으로 표현한다.

2. VC차원(Vapnik-Chervonenkis dimension of H : VCdim(H)) : 성장함수 $\pi_H(m)=2^m$이 성립되는 최대의 m이 H의 VC차원이 된다. 만약 임의의 집합에 대해 항상 등식이 성립하면, $\text{VCdim}(H)=\infty$이다.

다음 정리 2.2는 위의 척도를 이용하여 개선되는 결과를 보여준다.

정리 2.2 : [Haussler, 1988] H를 가설공간이라하고, 유한한 값 d를 H의 VC차원이라 하자. f를 목표개념이라 하면, $0<\varepsilon<1$에 대해, h가 최소한 다음 크기 m개의 사례와 일치하는 가설이면, $Prob\{d(h,f)\geq\varepsilon\}\leq\delta$이 성립한다.

$$m=\min[(1/\varepsilon)\{\ln(1/\delta)+\ln|H|\},$$
$$(1/\varepsilon)\{4\log(2/\delta)+8d\log(13/\varepsilon)\}]$$

이와 같이 어떠한 학습알고리즘이 목표개념 f를 다항식의 표본복잡도와 다항식의 계산복잡도를 가지는 절차로서 학습할수 있고, 학습된 개념 h가 $Prob\{d(h,f)\geq\varepsilon\}\leq\delta$의 구조를 가지면, 이러한 학습알고리즘을 PAC(Probably Approximately Correct)학습알고리즘이라 부른다.[Angluin과 Laird, 1988]

이 정리의 활용예를 가설공간의 규칙의 수가 유한한 경우와 무한한 경우 두가지로 나누어 아래에 제시한다.

예 1 : $|H| < \infty$ 인 경우

위에 예로든 사례공간 X=(유동비율, 부채비율, 매출신장율, 자기자본수익율) 에 대해 각 변수가 3개의 구간으로 나누어 표시되고, 신용평가를 A, B 두 등급으로만 한다고 가정하자. 그리고 가설공간 H를 결합개념(conjunctive concept)이라 하면 H의 규칙 수 $|H|=4^4=256$이 된다. 즉, N=256이 된다.

정리 2.1에 의해, 다음 크기 이상의 표본사례를 통하여 학습된 개념 h는 사례공간의 분포(distribution)에 상관없이 $1-\delta$ 이상의 확률로서, ε보다 작은 오류를 가지게 된다.

$$(1/\varepsilon)(\ln(1/\delta)+\ln 256)$$
$$=(1/\varepsilon)(\ln(1/\delta)+5.55)$$

위 식에서 우리는 가설공간의 크기가 증가하는 속도에 비해 사례크기가 증가하는 속도는 매우 느림을 알 수 있다.

$\varepsilon=0.1$이고 $\delta=0.05$인 경우, 정리 2.1을 통해 PAC학습을 위해 필요한 표본크기는 $m=(1/\varepsilon)(\ln(1/\delta)+\ln 256)=86$임을 알 수 있다. 다시말하면, 사례의 분포에 상관없이 86개 이상의 임의 표본(random sample)을 취하여 학습을 수행하면, 학습된 개념 h가 10% 이하의 오류를 가질 확률이 95% 이상이 된다.

Haussler [1988]의 정리에 의하면 순수 결합(conjunctive) 개념의 VC차원은

$n\leq VCdim(H)\leq 2n$(여기서 n은 속성의 수)이 된다.

따라서 이 문제의 VC차원은 최대 8이 된다. $\varepsilon=0.1$과 $\delta=0.05$인 경우, 정리 2.2를 이용하여 PAC 학습에 요구되는 사례수(표본복잡도)를 구하면 역시 86이 된다.

$$m=\min[(1/\varepsilon)(\ln(1/\delta)+\ln 256),$$
$$(1/\varepsilon)\{4\log(2/\delta)+8d\log(13/\varepsilon)\}]$$
$$=\min[(1/0.1)(\ln(1/0.05)+\ln 256),$$
$$(1/0.1)\{4\log(2/0.05)+8\times 8\log$$
$$(13/0.1)\}]=86$$

즉 이 경우에는 새로운 가설공간의 척도인 VC차원을 이용하여 개선 효과를 거두지 못하였다. 이는 가설공간이 유한 공간이기 때문이다.

예 2 : H=∞인 경우

성공적인 기업이 가지는 자산대비 부채의 비율(부채/자산)의 범위에 대하여 학습한다고 가정하자. 그러면, 사례공간 X는 [0,1] 구간이 된다. 가설공간 H를 [x, y] 구간과 공집합이라고 하면(여기서 $0 \le x \le y \le 1$), [0, 1] 구간에는 무한대의 부분 구간이 존재하므로 H=∞이다. H의 성장함수는 다음과 같이 계산된다.

하나의 사례, 예를 들어 사업을 성공적으로 영위하는 어느 회사의 자산대비 부채비율이 0.6이라고 하자. 이 사례 $0.6 \in X$은 [0.5, 1]구간에 의해 +(성공적인 회사)로 분류될 수도 있고 [0, 0.5] 구간에 의해 −(성공적이 아닌 회사)로 분류될 수 있다. 즉 가능한 모든 경우(여기서는 +, −의 2가지 경우)에 대해 가설공간의 개념으로서 분류가 가능하다. 즉 $\pi_H(1) = 2 = 2^1$이다.

이제 두개의 사례, 예를 들어, 비율이 0.3과 0.6인 2개의 회사가 있다고 가정하자. 아래 4가지 경우는 이 두개의 사례를 분류할 수 있는 모든 가능한 경우를 사례공간의 개념으로서 생성할 수 있음을 나타낸다.

[0.1, 0.4] 구간인 개념은 두개의 사례를 각각 (+,−);

[0.4, 0.7] 구간인 개념은 두개의 사례를 각각 (−,+);

[0.2, 0.7] 구간인 개념은 두개의 사례를 각각 (+,+);

[0.4, 0.5] 구간인 개념은 두개의 사례를 각각 (−,−) 로 분류한다.

따라서 $\pi_H(2) = 4 = 2^2$이다.

이제 a, b, c (여기서 a<b<c)의 자산대비 부채비율을 가지는 3개의 사례회사가 있다고 가정하자. 복수개의 구간을 OR로 연결할 수 있는 분리(disjunction)가 허용되지 않기 때문에, H내의 어떠한 개념도 (a, b, c)를 (+, −, +)로 분류할 수 없다. 나머지 모든 경우는 분류할 수 있기 때문에 $\pi_H(3) = 7 < 2^3$이 된다.

따라서 정의에 의해 VCdim(H)=2가 된다.

실수구간에는 무한개의 부분 구간이 존재하므로, H내의 규칙 갯수인 $|H| = \infty$ 이다. 즉, 규칙의 수 $N = \infty$ 이다.

그러므로 정리 2.1로는 무한개의 표본사례가 필요하다는 결론에 도달하게 된다. 그러나, 정리 2.2를 이용하면, $\varepsilon = 0.1$과 $\delta = 0.01$인 경우 PAC 학습에 요구되는 사례의 수 m은 다음과 같이 유한한 값이 된다.

$$m = \min[\infty, (1/0.1)\{4 \log(2/0.01) + 8 \times 2 \log(13/0.1)\}] = 1429.3$$

이와 같이 우리는 VC차원을 이용하여 학습에 필요한 사례의 수에 대한 결과를 개선할 수 있다.

2.2 의사결정나무 학습에 대한 분석

본 단원에서는 앞에서 논의한 학습이론을 이용하여 의사결정나무를 추론할때 정해진 신뢰도 수준의 한도내에서 오류수준을 보장하기에 충분한 사례의 수(sample size)에 대한 분석결과를 제시한다. 먼저 의사결정나무를 단순화하지 않은 경우에 대해서 분석하고, 다음에는

단순화하는 경우에 대하여 분석한다.

두가지 모두의 경우에 오류수준인 ε, $0 \leq \varepsilon \leq 1$, 과 신뢰수준인 δ, $0 \leq \delta \leq 1$, 가 주어진 상태에서 시작한다. 우리는 여기서 예측오류가 ε보다 클 확률이 δ보다 작게되는 의사결정나무를 발견할 가능성을 보장하기에 충분한 사례의 수를 결정하고자 한다. 사례의 추출은 반복을 허용하는 추출법을 사용한다.

2.2.1 의사결정나무 학습 방법

의사결정나무를 추론하는 많은 알고리즘이 있는데[Mingers, 1989a, 1989b ; Niblett, 1987 ; Quinlan, 1979, 1983, 1986, 1987a ; Utgoff, 1989], 대다수의 알고리즘은 다음 3가지 단계를 포함한다.

1) 모든 사례(example)를 정확하게 분류하는 완벽한 의사결정나무를 구축한다.
2) 위의 의사결정나무를 단순화하여 전체 사례공간(instance space)에서의 신뢰도와 예측가능성을 높인다.
3) 단순화된 의사결정나무를 처리하여 이해성을 높인다.

세번째 단계에서 많은 알고리즘들은 전문가 시스템을 위한 규칙(production rule)등의 이해하기 용이한 규칙을 생성한다.[Quinlan, 1987b]. 어떤 알고리즘들은 단계 2)와 3)을 결합하여 의사결정나무 구축시에 단순화 기법을 실행하기도 한다.

의사결정나무는 점증적으로(incrementally) 또는 비점증적으로(nonincrementally) 구축될 수 있다. 점증적인 귀납적 추론 알고리즘은 훈련사례(training instance)가 발생함에 따라 매번 현재의 의사결정나무를 수정 및 개선하는 방법이다. 이러한 방법은 훈련사례가 연속적으로 발생하는 경우에 적합하다. [Utgoff, 1989 ; Van de Velde, 1990].

비점증적인 알고리즘은 현재 이용가능한 모든 훈련사례를 이용하여 의사결정나무를 한번에 추론하는 방법이다. Quinlan[1979, 1986]의 ID3는 의사결정나무를 추론하는 대표적인 비점증적인 알고리즘이다. 이 알고리즘에서는 분기할 가지를 선택하기 위해 각 단계마다 하나의 속성(attribute)을 선택하게 되는데, 이 속성을 선택하는 많은 방법이 발표되어 있다. [Breiman et al., 1984 ; Quinlan, 1986 ; Marshall, 1986 ; Mingers, 1986, 1989a].

의사결정나무 추론 알고리즘이 확정적 자료가 아닌 불확실한 자료를 사용하는 경우에 통계적 신뢰도가 낮은 가지를 제거하는 단순화 단계는 매우 중요하게 적용된다. "불확실하다" 함은 진짜 개념(true concept)을 나타내는데 있어서 오류를 수반함을 의미한다. 자료의 불확실성은 측정의 오류에 기인되거나, 측정될 수 없는 요소 또는 숨겨진 요소가 존재함에 기인되기도 한다. 의사결정나무를 단순화하는 여러개의 경험적인 방법이 제안되었는데 크게 두가지 종류로 구분할 수 있다. 그중의 하나는 의사결정나무 구축시에 수행하는 단순화(construction-time pruning)이고 다른 하나는 의사결정나무를 완전히 구축한 다음에 단순화를 적용하는 (post-pruning) 방법이다.

의사결정나무 구축시에 수행하는 단순화 기법은 의사결정나무의 확대를 언제 중지해야 될지를 결정하는데 이용된다. 기존의 중지 기준(termination criteria)은 현재의 훈련집합속에 있는 모든사례가 동일집단(same class)에 속할때 나무의 확대를 중지하는 것이었는데 반해 새로운 중지기준은 의사결정나무를 구축할 때 사용되는 속성선택방법(selection measure)과 관계가 있다. 대표적인 의사결정나무 구축시의 단순화 기법은 문지방 방법(threshold method)과 카이-제곱 검사법(chi-square test method)등이 있다. 의사결정나무 구축시의 단순화 기법은 국지적인 정보에만 의존하여 나무확대에 대한 의사결정을 한다는데에 그 단점이 있다. 즉, 어느 한 노드에서 확대중지 결정을 내리게 되었을때 그 노드의 하위노드에서 의사결정나무 전체의 판별력이 높게될 가능성을 배제할 수 없다는 것이다. [Breiman et al., 1984 ; Niblett, 1987]. 전체 이용가능한 정보를 모두 이용하여 완전히 확대된 의사결정나무를 단순화하는 여러가지 기법이 있다. 그중에 대표적인 것으로는 오류-복잡도 방법(error-complexity method) [Breiman et al., 1984], 임계치 방법(critical value method) [Mingers,1989b], 최소오류법(minimum-error method) [Niblett and Bratko, 1986], 축소오류법(reduced error method) [Quinlan, 1987a] 등이 있다.

2.2.2 단순화 과정이 없는 경우의 분석

여기서는 사례간에 모순(inconsistency)이 없다는 제약하에서 의사결정나무를 구축한다. 즉, 속성의 값이 동일한 두개의 사례는 반드시 동일한 분류에 속하여야 한다. 추론된 의사결정나무는 모든 훈련사례들(Training set)을 완벽하게 분류한다.

표본복잡도를 산출하기 위하여 가설공간(H)이 의사결정나무인 경우의 VC차원(VCdim(H))을 먼저 도출해야 한다.

C_i 를 속성 i($1 \le i \le n$, n은 속성의 수)가 취할 수 있는 가능한 값의 갯수라고 하자. 그러면 사례공간 X에서 가능한 사례종류의 수는 다음과 같이 나타낼 수 있다.

$$d = C_1 \times C_2 \times \cdots \times C_i \times \cdots \times C_n$$

각각의 사례는 학습목표 개념에 대한 양의 사례 또는 음의 사례로 나타날 수 있으므로, $2^d = |H|$ 가 된다. 이 경우 의사결정나무는 2^d 개의 개념을 생성할 수 있으므로, VCdim(H) =d가 된다.

따라서 d를 가설공간(H)이 의사결정나무인 경우의 VCdim(H)이라 할때, 다음 정리 2.3은 의사결정나무 추론알고리즘인 ID3에 대한 충분한 사례수를 제공한다.

정리 2.3 : [Tsai and Koehler, 1993]

1. 0과 1 사이의 값을 가지는 어떠한 ϵ과 δ에 대해서도 사례수(sample size)가
   $[\ln(1/\delta) + d \ln 2]/\epsilon$
   보다 크게되면 ID3는 최소한 $1-\delta$의 확률로서 오류가 ϵ보다 작은 개념 h를 찾아낼 수 있다.

2. ε이 0과 1/2사이의 값을 가질때 ID3는 최소한 다음 크기의 사례를 이용해야 한다.

$$\max\{[(1-\varepsilon)/\varepsilon]\ln(1/\delta),\ d[1-2(\varepsilon(1-\delta)+\delta)]\}$$

예를 들어, 속성의 갯수가 5이고, 각 속성은 5가지의 가능한 값을 가진다고 가정하자. 그러면 VCdim(H)은 3,125이다. 또 우리는 $\varepsilon=0.1$과 $\delta=0.01$을 필요로 한다고 가정하자. 그러면 정리 2.3에 의하여, $m=\max\{41.25, 2443.75\}=2,444$개의 사례가 어떠한 경우에도 필요하게 된다. 그리고, m=21,707개의 사례가 있으면 위의 정확도 수준을 보장하기에 충분하게 됨을 알 수 있다. $\varepsilon=0.5$이고 $\delta=0.01$인 경우에는 m=4,342개의 사례가 있으면 위의 정확성을 얻는데 충분한 수준이 된다.

실수값을 가지는 속성이 있는 경우에, VCdim(H)은 무한대가 되기 때문에 정리 2.5는 성립하지 않는다. 그러나 많은 의사결정나무 추론알고리즘들은 이러한 경우에 대처하기 위하여 알고리즘을 수정하고 있다. 예를 들어, 실수값을 갖는 속성의 구간을 반으로 나누거나, 구간을 훈련사례에 근거하여 유한개의 의미있는 구간으로 나누는 방법들이 이용되고 있으며, 이렇게 수정된 알고리즘에는 정리 2.3을 적용할 수 있다.

만약 필요한 만큼의 사례를 얻을 수 없다면 독립시험사례(independent test set)를 이용하는 사후적 추정 방법에 의하여 의사결정나무의 정확성을 추정할 수 있는데 Tsai와 Koehler [1993]에 상세 절차가 나타나 있다. 이 방법은 베타사전분포를 가정하고 일관성있는 의사결정나무(consistent decision tree)의 오류에 대한 상한값을 찾아낸다.

2.2.3 단순화를 하는 경우의 분석

여기에서는 Kim과 Koehler[1996]의 결과를 중심으로 분석방법을 제시한다. 단순화의 과정을 거쳐서 추론된 의사결정나무가 진짜 개념(true concept)과 상치할 확률이 $1-\delta$보다 큰 가능성을 가지고 보다 작게되기를 보장하는 충분한 크기의 훈련사례수를 제공하고자 한다. 논문의 촛점을 이진의사결정나무(binary decision tree)에 맞추어 서술하고자 한다. 유한한 수, k, 의 값을 가지는 어떠한 속성도 $\log_2(k)$개의 이진 변수로 다시 나타낼 수 있기 때문이다. 먼저 이진의사결정나무와 그 등급(rank)을 정의한다.

정의 2.3 : $V_n=\{v_1, ..., v_n\}$은 n개의 Boolean 변수이고, $X_n=\{0,1\}^n$ 이라고 가정하자.

1. (이진의사결정나무) : 이진의사결정나무는 다음과 같이 정의된다.
   (i) Q가 0 또는 1의 레이블을 가지는 뿌리노드만의 나무라고 하면, Q는 V_n상의 이진의사결정나무이다.(이하에서 우리는 이경우를 간편하게 "Q=0" 또는 "Q=1" 이라고 부른다).
   (ii) Q(0)를 Q의 왼쪽나무(left subtree)라고 하고, Q(1)을 Q의 오른쪽 나무라고 하자. 만약 Q의 뿌리노드 v가 V_n

에 있고 Q(0)와 Q(1)이 이진의사결정나무이면 Q도 이진의사결정나무이다.

2. (의사결정나무의 등급) : 이진의사결정나무의 등급은 r(Q)로 표현되며 그 정의는 다음과 같다.

(i) Q=0 또는 Q=1 이면 r(Q)=0 이다.

(ii) r_0와 r_1이 각각 Q의 왼쪽나무와 오른쪽나무의 등급이라고 할때

$r(Q)=\max(r_0, r_1)$, $r_0 \neq r_1$ 인 경우

$r_0+1(=r_1+1)$, $r_0=r_1$ 인 경우

Ehrenfeucht와 Haussler[1988]의 의사결정나무 추론 알고리즘, Findmin(S)와 Kim과 Koehler[1996]의 의사결정나무 단순화 알고리즘, Prune(r, k, Q, S)에 의해서 다음결 과가 도출된다. 여기서 r, k, Q, S는 각각 다음과 같다.

r = 단순화의 결과로 얻어진 의사결정나무의 등급, 즉 우리가 원하는 의사결정나무의 단순화 수준.

k = 단순화 대상인 현재의 의사결정나무의 등급.

Q = 현재의 의사결정나무.

S = 추론에 사용되는 훈련사례.

정리 2.4 : (단순화 과정을 포함하는 의사결정나무 추론에 충분한 사례수 : Kim and Koehler[1996]) 어떠한 $n \geq r > 0$, 여떤 목표개념(target concept) f, X_n에 대한 어떤 균등분포 D, 그리고 D로부터 독립적으로 추출된 다음 크기, m, 의 f에 대한 무작위 사례인 표본 S가 있으면

a) $m \geq [2/\{\varepsilon^2(1-2\mu_{n,r})^2\}]\{(e \times n/r)^r \ln(8n)+\ln(2/\delta)\}$, $0<\varepsilon, \delta<1$인 경우

b) $m \geq [1/\{\varepsilon(1-\exp(-(0.5)(1-2\mu_{n,r})^2))\}]\{(e \times n/r)^r \ln(8n)+\ln(1/\delta)\}$, $0<\varepsilon, \delta<1/2$인 경우

$1-\delta$보다 큰 확률로서 Findmin(S)와 Prune(r, k, Q, S)는 등급이 r이하이며 오류가 보다 작거나 같은 의사결정나무 h를 추론할 수 있다. 여기서 e는 자연로 그의 밑수이며, $\mu_{n,r}$은 다음과 같다.

$\mu_{n,r}=0.5-(0.5)^{n-r+1}$, $n \geq r=1$인 경우

$\mu_{n,r}=0.5-\{1+(n-r)(0.5)^2\}(0.5)^{n-r+1}$, $n \geq r>1$인 경우.

위의 정리 2.4 는 하나의 무작위 사례를 추출하는데 한 단위시간이 소요됨을 고려할때 n개의 속성을 가지는 등급(rank)이 r이하인 의사결정나무를 r이 주어졌을때 $1/\varepsilon$, $1/\delta$, $1/(1-2\mu_{n,r})$과 n의 다항(polynomial)시간내에 $1-\varepsilon$의 정확성과 $1-\delta$의 신뢰도를 가지고 추론할 수 있음을 보여준다. 여기서 우리는 단순화 과정이 추가됨으로 하여 획득해야하는 사례의 수가 증가되기는 하였지만 여전히 다항 특성을 유지함을 볼 수 있다.

III. 사후적 오류분석방법

현실적인 추론 상황에서는 이론적으로 도출된 어떤 정확성 수준을 보장하는데 충분한 수

의 사례들을 얻기 어려운 경우가 많으므로, 그러한 경우에 대비하여 단순화된 의사결정나무의 예측정확성에 대한 사후적 평가 방법을 고안할 필요가 있다. 학습으로 부터 획득된 정보를 이용하는 방법과 학습정보가 없어도 이용할 수 있는 방법을 제시한다.

3.1 학습사례의 정보를 이용하는 방법

학습정보와 일치하는 파라메터를 얻기 위하여 베타사전분포(Beta prior)를 가정한다. 오류에 대한 사후적 추정치를 얻기 위하여 다음과 같은 호프딩의 부등식(Hoeffding's Inequalities : [Hoeffding, 1963])을 이용한다

보조정리 3.1 :

$x_1, x_2, \ldots, x_n$을 $0 \le x_i \le 1$인 독립 확률변수라 하고, $i=1, 2, \ldots, n$에 대해 기대치 $E[x_i]=\mu$라 하자.

그러면,

$Prob\{\bar{x}-\mu \ge c\} \le \exp(-2nc^2)$이고,

또한

$Prob\{\mu-\bar{x} \ge c\} \le \exp(-2nc^2)$이 성립한다.

m개의 학습 사례에서 b개의 오류가 발생하였다면, 위의 보조정리에 의하여 다음 결과가 성립한다.

보조정리 3.2 :

모든 $b/m \le \varepsilon$에 대해

$Prob\{\theta \ge \varepsilon\} \le \exp(-2(\varepsilon-b/m)^2m)$.

증명 : 학습 알고리즘이 i번째 사례를 잘못 분류하였다면 $x_i=1$이라 하고 그렇지 않다면 $x_i=0$라 하자. 잘못 분류할 확률 θ는 $E[x_i]$가 된다.

표본사례에서 b개의 오류가 있었으므로, 보조정리 3.1에 의해 다음이 성립한다. $Prob\{2\theta \ge \varepsilon\} = Prob\{\theta - b/m \ge \varepsilon - b/m\} \le \exp(-2(\varepsilon-b/m)^2m)$.□

Z^+를 양의 정수 집합이라 하고, S(b,m)을 위의 보조정리 3.2와 모순이 없는 불완전베타분포(Incomplete Beta distribution)의 파라메터 집합$\{(p,q) : p,q \in Z^+\}$이라 하자. S(b,m)은 다음과 같이 일련의 보조정리에 의해 값이 결정된다.

보조정리 3.3 : $p,q \in Z^+$이고 $b/m \le \varepsilon < 1 (1 \le b \le m-1)$일때,

ε이 $-2(1-\varepsilon)\ln(1-\varepsilon) = \varepsilon - b/m$의 근이고, $p+q \ge 1+4m(1-\varepsilon(\varepsilon-b/m))$인 경우, $(p,q) \in S(b,m)$이다.

증명 : 보조정리 3.2에 의해서, 모든 $b/m \le \varepsilon < 1$에 대해서, 모순이 없는 베타사전분포는 다음을 만족한다.

$Prob\{\theta \ge \varepsilon\} = 1 - I_\varepsilon(p,q) \le \exp(-2(\varepsilon-b/m)^2m)$

여기서 $I_\varepsilon(p,q)$는 베타사전분포가 이하의 값을 가질 확률이다. p와 q가 정수이므로, 위의 부등식을 이항분포로 나타낸후, 항을 재정리하여 $k=0$을 대입하면 다음식이 성립한다.

$(p+q-1)\ln(1-\varepsilon)+2(\varepsilon-b/m)^2m\le 0.$

즉, $p+q-1\ge -2(\varepsilon-b/m)^2m/\ln(1-\varepsilon)$이 성립한다.

$f(\varepsilon)=-2(\varepsilon-b/m)^2m/\ln(1-\varepsilon)$라 하면 $f(\varepsilon)$는 최고치가 나타나는 구간에서 연속 오목(concave) 함수이다. 따라서 $f'(\varepsilon)=0$ 식은 다음과 같이 표현된다.

$$-2(1-\varepsilon)\ln(1-\varepsilon)=\varepsilon-b/m.$$

그러므로 $-2(1-\varepsilon)\ln(1-\varepsilon)=\varepsilon-b/m$를 만족하는 ε에 대해 다음이 성립한다.,

$$p+q\ge 1-2(\varepsilon-b/m)^2m/\ln(1-\varepsilon)$$
$$=1+4m(1-\varepsilon)(\varepsilon-b/m).\square$$

다음 보조정리 3.4는 S(b,m)에서 모순 없는 베타사전분포의 관계를 나타낸다..

보조정리 3.4 : [Tsai and Koehler, 1993]

$p, q\in Z^+$이고 $(p,q)\in S(b,m)$이면,

$(t,q)\in S(b,m)$ $1\le t\le p$ 및

$(p,t)\in S(b,m)$ $t\ge q$이 성립한다.

다음 보조정리 3.5는 S(b,m)에서 q의 필요조건을 나타낸다.

보조정리 3.5 : $p, q\in Z^+$이고 $b/m\le\varepsilon<1(1\le b\le m-1)$이면,

$-2(1-\varepsilon)\ln(1-\varepsilon)=\varepsilon-b/m$를 만족하는 ε에 대해 $q\ge 4m(1-\varepsilon)(\varepsilon-b/m)$를 만족할 경우에만 $(p,q)\in S(b,m)$이 된다.

q^*를 위의 조건을 만족하는 최소 q라 할때, $(1,q^*)$는 모순이 없는 베타사전분포(consistent Beta prior) 이다.

증명 : $p=1$ 이면 $\text{Prob}\{\theta\ge\varepsilon\}=1-I_\varepsilon(1,q)=(1-\varepsilon)^q\le\exp(-2(\varepsilon-b/m)^2m)$이 된다.

로그함수를 취하여 $q\ln(1-\varepsilon)\ge -2(\varepsilon-b/m)^2m$ 을 얻을 수 있다.

보조정리 3.3의 증명과 유사한 방법을 사용하여 다음결과를 얻을 수 있다.

$-2(1-\varepsilon)\ln(1-\varepsilon)=\varepsilon-b/m$를 만족하는 ε에 대해 $q\ge 4m(1-\varepsilon)(\varepsilon-b/m)$이 된다.

q^*를 위의 조건을 만족하는 최소 q 값이라 하면, $(1,q^*)$는 보조정리 3.3에 의해 모순이 없는 베타사전분포가 되며, 이 식은 어떤 $\varepsilon(b/m\le\varepsilon<1)$에 대해서도 성립한다.

반대로 $(1+k, q^*-k)$, $k\in Z^+$, 를 모순없는 베타사전분포라 가정하자. 그러면 보조정리 3.4에 의해서 $(1,q^*-k)$ 도 모순이 없는 베타사전분포이다. 따라서, $p=1$인 경우, q^*는 q의 최소값이 될수 없으며 모순이 발생한다. 그러므로 $(1+k,q^*-k)$는 어떤 $k\in Z^+$에 대해서도 모순없는 베타사전분포가 될 수 없다.

어떠한 $p\in Z^+$에 대해서도 q는 q^*보다 크거나 같아야 하며, $-2(1-\varepsilon)\ln(1-\varepsilon)=\varepsilon-b/m$를 만족하는 ε에 대해 $q\ge 4m(1-\varepsilon)(\varepsilon-b/m)$를 만족할 경우에만 $(p,q)\in S(b,m)$이 된다.$\square$

(p,q)를 모순이 없는 베타사전분포라 하고, m_2개의 시험사례중에 b_2개의 오류가 나타났다고 가정하면, 사후분포의 파라메터는 $(p+b_2, q+m_2-b_2)$가 된다. 최악의 신뢰도 수준인 δ는 다음과 같이 정식화할 수 있다.

$$\delta=\text{Sup } 1-I_\varepsilon(p+b_2,q+m_2-b_2)$$

s. t.

$(p,q)\in S(b,m)$

$p, q\in Z^+$

보조정리 3.4에 의해 $I_\epsilon(p+b_2,q+m_2-b_2)$는 p에 대해서 증가함수이고, q에 대해서 감소함수임을 알수 있다. 따라서 supremum 값은 모든 가능한 p중에서 최대치 p와, 모든 가능한 q 중에서 최소치 q의 집합 (p, q)에서 발생한다. 그러나, p가 증가함에 따라, 모순이없는 q 값은 비감소함수이다. 따라서 대부분의 경우에, 우리는 이상적인 (최대 p, 최소 q) 집합을 얻을 수 없다. 즉, 파라메터 집합에 대한 보다 정밀한 분석이 요구된다.

아래 보조정리 3.6 은 $p'>p$와 $q'>q$를 만족하는 파라메타 집합 (p,q)와 (p′,q′)의 관계를 보여준다. 아래에서 우리는 높은 신뢰도 δ를 얻는 파라메터 결정은 ϵ값에 관계가 있음을 알수 있다.

보조정리 3.6 : $k\in Z^+$ 이면 다음이 성립한다.

a) (p,q)와 (p+k,q+1)이 모순없는 베타 사전분포의 파라메터 집합이라 하자.

$q/(p+q)$ b/m이면, (p+k,q+1)은 $b/m\le\epsilon<\alpha$일때 신뢰도 값이 높고, (p,q)는 $\alpha<\epsilon<1$ 구간에서 신뢰도값이 높다.(여기서 $b/m\le\alpha<1$) k가 증가함에 따라, α값도 증가한다. $q/(p+q)<b/m$ 이고 k=1이면, (p,q)는 (p+1,q+1)보다 모든 $\epsilon(b/m\le\epsilon<1)$ 값에 대해 높은 신뢰도 값을 준다.

b) (p,q)와 (p+1,q+k)가 모순없는 베타 사전분포의 파라메터 집합이라 하자.

$q/(p+q)\ge b/m$이면, (p+1,q+k)는 $b/m\le\epsilon<\beta$ 일때 신뢰도 값이 높고, (p,q)는 $\beta<\epsilon<1$ 구간에서 신뢰도값이 높다.(여기서 $b/m\le\beta<1$) k가 증가함에 따라, β값은 감소한다.

증명 : a) k=1일때 Abramowitz와 Segun [1968]의 공식 26.5.15와 26.5.16을 결합하여 $g(k)=g(1)=I_\epsilon(p,q)-I_\epsilon(p+1,q+1)=K^*(1-(p+q)/q\ \epsilon)$을 얻는다. 여기서 $K=\epsilon^p(1-\epsilon)^q\Gamma(p+q)/\Gamma(p+1)\Gamma(q)$이다.

K가 양수라면, $b/m\le\epsilon<q/(p+q)$에 대해서 $g(1)=I_\epsilon(p,q)-I_\epsilon(p+1,q+1)>0$ 이고, $q/(p+q)<\epsilon<1$에 대해서 $g(1)<0$ 이다. 따라서 $q/(p+q)>b/m$ 이면, (p+1,q+1) 이 $b/m\le\epsilon<q/(p+q)$ 구간에서 높은 신뢰도 값을 주고, (p,q)는 $q/(p+q)<\epsilon<1$ 구간에서 높은 신뢰도 값을 준다. $q/(p+q)<b/m$인 경우에는 모든 $\epsilon(b/m\le\epsilon<1)$ 구간에서 (p,q)가 높은 신뢰도 값을 주게 된다.

g(k)의 일반식을 도출하기 위하여 Abramowitz와 Segun의 공식 26.5.16을 이용하여 $I_\epsilon(p+(k-1),q+1)$를 $I_\epsilon(p+k,q+1)$로 대치하여 다음을 얻는다.

$$g(k)=I_\epsilon(p,q)-I_\epsilon(p+k,q+1)$$
$$=K\times(1-(p+q)/q\ \epsilon+(1-\epsilon)R(k,\epsilon)).$$

여기서 $R(k,\epsilon)$ is 0 보다 큰 k−1개 항의 합이다. k가 증가함에 따라 $R(k,\epsilon)$의 값이 커지므로 ϵ은 $g(k)>0$ 으로 하기 위하여 값이 커져야 한다. 즉, k가 증가함에 따라 값이 증가한

다.

ε값을 1에 매우 가깝게 취함으로서, g(k)는 어떠한 $k\in Z^+$에 대해서도 음수가 될 수 있다. 그러한 $\varepsilon<1$에 대해서, (p,q)는 (p+k,q+1)보다 높은 신뢰도 값을 주게 된다. 따라서, 어떠한 $k\in Z^+$에 대해서도 (p,q)가 더 높은 신뢰도 값을 주게 되는 의 범위는 공집합이 아니게 된다.

보조정리 3.4에 의해서, $1-I_\varepsilon(p+k,q+1)$는 k가 증가함에 따라 같이 증가한다. 따라서, 어떠한 $k\in Z^+$에 대해서도 $q/(p+q)$ b/m이면, (p+k,q+1)가 더 높은 신뢰도 값을 주게 되는 ε의 범위는 공집합이 아니게 된다.

b) 앞의 a)에서의 증명방법과 유사한 방법으로 증명됨.□

$q/(p+q)<b/m$가 아닌한, 모든 $\varepsilon(b/m\le\varepsilon<1)$ 값에 대해 최악의 신뢰도 값을 주는 (p,q) 값은 존재하지 않음을 보조정리 3.6을 통해 알수 있다. 모든 가능한 값에 대해 최대의 δ값을 주는 (p,q)값을 찾기 위해 모든 모순이 없는 베타사전분포 파라메터 집합을 찾아 비교하는 것은 너무 많은 노력이 소요되는 작업이므로, 다음과 같은 차선책을 통해 최적에 가까운 값을 찾아내는 방법을 생각할 수 있다. 예를 들어, p를 최대값에 고정하고, 모순이 없는 최소의 q를 찾아내는 방법이 있고, 또한 q를 최소값에 고정하고, 모순이 없는 최대의 p를 찾아내는 방법도 있다. 그런데, p의 가능한 최대값은 매우 큰 불확정적인 값이므로, q의 최소값을 먼저 찾아내고, 이에 대응하는 최대의 p값을 찾아나가는 것이 보다 경제적인 전략이 된다. 다음 정리는 이 전략을 통한 결과를 보여준다.

정리 3.7 : $p, q\in Z^+$와 $b/m\le\varepsilon<1$ $(1\le b\le m-1)$에 대하여

$q\ge 4m(1-\varepsilon)(\varepsilon-b/m)$인 경우에만

$(p,q)\in S(b,m)$ 이다.

여기서 ε은 $-2(1-\varepsilon)\ln(1-\varepsilon)=-b/m$의 근이다.

q^*를 위의 조건을 만족하는 가장 작은 q라고 하고, p^*를 q^*가 주어졌을때 일치되는 최대의 p라고 하자. 그러면 (p^*, q^*)는 최소한 어떤 유효한 ε의 구간에서 가장 낮은 신뢰도(confidence factor)를 제공하는 사전정보와 모순이 없는(consistent) 베타사전분포이다.

m_2개의 사례로 시험을 하여 b_2개의 사례를 잘못 분류하였다면,

$Prob\{\theta\ge\varepsilon\}\le\delta$이며 여기서 δ의 하한(lower bound)은 다음과 같다.

$1-I_\varepsilon(p^*+b_2, q^*+m_2-b_2)=\underline{\delta}\le\delta$.

즉, (p^*, q^*)는 δ의 하한을 제공한다.

여기서 $I_\varepsilon(a,b)$는 불완전베타분포(Incomplete Beta Distribution)이다.

증명 : 앞서 증명된 보조정리 3.3, 3.4, 3.5, 3.6에 의하여 증명됨.□

위의 결과는 δ의 하한에 대한 유용한 정보를 제공하므로, δ의 상한에 대한 결과를 도출해야 한다. 제 2절은 이에 대한 논의를 한다.

3.2 학습정보가 없는 경우에도 이용 가능한 방법

여기에서는 사전 분포에 대한 가정이 없는 Prob{$\theta \geq \varepsilon$}에 대한 일반 경계치를 도출한다. 이 방법은 또한 학습의 결과를 이용하지 않는 방법이다. 따라서 학습도메인과 시험 도메인이 다른 경우나 학습방법이 변화를 가지는 경우 특히 적합하다.

m 개의 시험사례로 시험하여 b개의 사례에서 오류가 발생했다고 가정하자.

정리 3.8 : 어떠한 $b/m \leq \varepsilon$ 에서도 다음 부등식이 성립한다.

a) $\text{Prob}\{\theta \geq \varepsilon\} \leq \exp(-2(\varepsilon - b/m)^2 m)$.

b) $\text{Prob}\{\theta \geq \varepsilon\} < \exp(-2(\varepsilon - b/m)^2 m - (4/3)(\varepsilon - b/m)^4 m)$.

증명 : a)는 보조정리 3.2에 의해 손쉽게 증명되며, b)는 Hoeffding의 부등식을 개선한 Johnson과 Kotz[1969]의 부등식을 이용하여 보조적으로 도출됨.□

따라서 앞서 도출된 신뢰도 값의 하한 경계에 대한 결과와 여기서의 결과를 결합하여 신뢰도 값의 구간을 결정할 수 있다.

사전정보가 전혀없을때 적용할 수 있는 또 하나의 방법이 있다. 이 방법은 앞의 방법과 달리 오류비율의 분포가 균등사전분포(Uniform prior)를 가정하고 도출되는 방법으로서, 이 가정이 무난할 때 사용되는 방법이다. 그러나 이 방법은 오류비율이 어떤 특정구간에서 다른 구간보다 높은 확률을 가지는 경우에는 부적당하다. 아래에서 C_k^m은 m개에서 k개의 순서를 고려하지 않은 표본을 얻는 방법의 수이다.

정리 3.9 : [Kim and Koehler, 1994a] m 개의 시험사례를 단순화된 의사결정나무를 이용하여 분류할때 b개의 오류가 발생하였다면, 균등사전분포(uniform prior)를 사용한 2항 파라메타 θ의 사후적 추정치는 다음과 같다.

$$\text{Prob}\{\theta \geq \varepsilon \mid b, m\} = \sum_{k=0}^{b} C_k^{m+1} \varepsilon^k (1-\varepsilon)^{m+1}, \quad \varepsilon \leq 0.5 \text{인 경우}$$

$$\text{Prob}\{\theta \geq \varepsilon \mid b, m\} = \sum_{k=m+1-b}^{m+1} C_k^{m+1} \varepsilon^{m+1-k} (1-\varepsilon)^k, \quad \varepsilon \geq 0.5 \text{인 경우}$$

본 단원에서 언급한 3가지 방법은 상호 보완적으로 학습 개념의 오류를 사후적으로 추정하는 목적으로 사용될 수 있다.

Ⅳ. 실제 문제 적용 사례

본 단원에서는 Messier와 Hansen[1988]의 채무불이행(loan default) 문제를 사용하여 단원 Ⅲ에서의 결과를 실제문제에 적용한다. 본 단원에서의 사례는 앞서의 오류분석 방법론을 일반 독자가 보다 쉽게 적용할 수 있도록 가

능하게 하는데 목적이 있다. 즉 수리적인 도출 과정보다 결과 적용에 주된 관심이 있는 독자들을 위하여 결과의 의미와 적용과정을 제시한다.

전개를 편리하게 하기위하여 6개의 재무비율로 사례공간(instance space)을 압축하여 문제를 표현한다. 다음과 같은 재무비율들을 사용하여 회사의 채무불이행을 예측하는 규칙을 추론하는 문제를 생각하여 보자. 여기서 우리는 '높음'과 '낮음'이라는 두개의 값을 가진 이진변수로 모든 변수를 표현하며 이러한 구분의 경계치는 Messier와 Hansen[1988]으로부터 도출되었다.

속 성	낮 음	높 음
유동비율	<1.912	≥1.912
장기부채/자기자본	< .486	≥.486
저장기부채/자기자본	< .046	≥.046 및 ≤.486
운전자본/매출액	< .222	≥.222
순이익/총자산	< .100	≥.100
순이익/매출액	< .010	≥.010

Ehrenfeucht와 Haussler[1988]의 의사결정나무구축 알고리즘 Findmin(S)를 이용하면 의사결정나무가 최소 등급(minimal rank)을 가진 나무로서 추론되어진다.(데이터는 Messier와 Hansen[1988]이 사용한 32개의 훈련사례를 사용하였다). 이 의사결정나무의 등급은 2이고, 이 나무는 훈련사례(S)를 완벽하게 분류(오류없이 추정)한다.

이제 우리는 등급이 1인 단순화된 의사결정나무를 원한다고 가정하여 보자. 그러면 r=1, k=2가 된다. Kim과 Koehler[1996]의 알고리즘 Prune(r,k,Q,S)을 사용하여 의사결정나무를 단순화시키면 등급이 1인 의사결정나무가 얻어진다.

4.1 단순화과정을 포함하는 의사결정나무 추론에 요구되는 충분한 양의 사례수

위의 단순화 과정을 거쳐 추론된 의사결정나무에 대하여 정리 2.4를 적용하여 보자. 여기서 속성의 수는 n=6, 현재의 의사결정나무 등급 k=2, 그리고 원하는 단순화 수준 r=1 이다. 정리 2.4의 a)에 의해 요구되는 충분량의 사례수 m은

$\varepsilon=0.5$이고 $\delta=0.1$일때 m=560,609이고

$\varepsilon=0.1$이고 $\delta=0.01$일때 m=14,015,240

이다.

정리 2.4의 b)를 사용하고 $\mu_{n,r}$에 대한 보다 엄격한 경계(tight bound)를 사용하면(상세한 내용은 Kim과 Koehler[1996] 참조) m이 다음과 같이 줄어들게 된다.

$\varepsilon=0.499$이고 $\delta=0.01$일때 m=833,

$\varepsilon=0.2$이고 $\delta=0.01$일때 m=2,084, 그리고

$\varepsilon=0.1$이고 $\delta=0.01$일때 m=4,168이다.

즉, 위의 채무불이행 문제에서 단순화의 과정을 거쳐 의사결정나무의 형태로 추론된 규칙 h가 4,168개 이상의 무작위의 독립된 훈련사례를 사용하여 추론되었다면 h의 추정 오류가 10% 이하일 확률이 99% 이상이 됨을 의미한다.

3.2 단순화된 의사결정나무의 추정오류에 대한 사후적 추정치

실제의 경영 환경에서는 위의 이론에 의하여 요구되는 충분한 양의 사례를 얻기 어려운 경우가 많다. 이러한 경우에 우리는 작은수의 사례를 사용하여 의사결정나무를 추론하고 사후적 추정방법에 의하여 추론된 개념의 정확성을 평가하는 방법을 사용할 수 있다.

Messier와 Hansen[1988]의 채무불이행(loan default) 자료를 사용하여 이 경우를 설명한다. 훈련사례의 수는 m=32이고, 등급 r=1 인 의사결정나무에 의해 2개의 분류오류가 발견된다. 시험사례의 수는 16이고 그중 2개의 사례가 잘못 분류된다.

학습사례의 정보를 이용하는 정리 3.7의 방법에 의한 오류추정치는 다음과 같다.

조건을 만족하는 $q^*=22$이고 따라서 $(p^*=1, q^*=22)$가 얻어진다.

$\varepsilon=0.1$인 경우에 $\text{Prob}\{\theta\geq 0.1\}\leq\delta$이고

$1-I_\varepsilon(p^*+b2, q^*+m2-b2)=1-I_{0.1}(1+2, 22+16-2)=\underline{\delta}\leq\delta$이 된다.

왼쪽항을 계산하면 다음과 같이 된다.

$\text{Prob}\{\theta\geq 0.1\}\leq\delta$이고 $1-I_{0.1}(3, 36)=0.253670=\underline{\delta}\leq\delta$이 된다.

즉, 단순화과정을 거친 의사결정나무의 추정상의 오류가 10% 보다 크거나 같을 확률이 δ보다 작은데 δ의 하한 경계는 0.253670이다.

여기서 여러 다른 수준의 ε값에 대해 다음 결과를 얻게된다.

$\text{Prob}\{\theta\geq 0.07\}\leq\delta$, $1-I_{0.07}(3, 36)$
$=0.497546=\underline{\delta}\leq\delta$.

$\text{Prob}\{\theta\geq 0.15\}\leq\delta$, $1-I_{0.15}(3, 36)$
$=0.061545=\underline{\delta}\leq\delta$.

$\text{Prob}\{\theta\geq 0.20\}\leq\delta$, $1-I_{0.20}(3, 36)$
$=0.011306=\underline{\delta}\leq\delta$.

$\text{Prob}\{\theta\geq 0.25\}\leq\delta$, $1-I_{0.25}(3, 36)$
$=0.001641=\underline{\delta}\leq\delta$.

작은 ε값에 대해서는 p값을 크게 하여 일치되는 베타 사전분포(consistent Beta prior)를 찾아냄으로써 하한경계(lower bound)를 개선할 수 있다. 예를 들면, p가 2일때 일치되는 q값 중에 가장 작은 값은 30이다. $\varepsilon=0.1$인 경우에

$\underline{\delta}=1-I_\varepsilon(p^*+b2, q^*+m2-b2)$
$=1-I_{0.1}(2+2, 30+16-2)=0.295587.$

위와 같은 방법으로 탐색을 계속하여 작은 ε값에 대하여는 오류를 정확하게 추정하는 것이 가능하여 진다. 그러나 큰 ε값에 대하여는 이 방법이 엄격한 하한 경계값을 제공한다. 예를 들어 ε이 0.20보다 크면, (p, q)=(2, 30) 이 경계값을 개선하지 못한다.

분포에 대한 가정이 전혀없는 정리 3.8의 방법 a)는 다음과 같이 오류를 분석한다.

$\varepsilon=0.2$인 경우에 $\text{Prob}\{\theta\geq 0.2\}\leq\exp(-2(0.2-2/16)^2\,16)=0.83527$,

$\varepsilon=0.3$인 경우에는

$\text{Prob}\{\theta\geq 0.3\}\leq 0.37531$이 된다.

정리 3.8의 b)에 의하면 다음과 같이 오류가 분석된다. $\varepsilon=0.2$인 경우에

$\text{Prob}\{\theta \geq 0.2\} < \exp(-2(0.2-2/16)^2 16 - (4/3)(0.2-2/16)^4 16) = 0.8347$이 되며, $\varepsilon = 0.3$인 경우에는 $\text{Prob}\{\theta \geq 0.3\} < 0.3678$이 된다.

시험사례의 수가 16이고 2개의 분류오류가 발생하였으므로, 균등분포를 가정하는 경우는 정리 3.9에 의해 다음과 같이 개념의 오류를 추정한다.

$$\text{Prob}\{\theta \geq \varepsilon \mid 2, 16\} = \sum_{k=0}^{2} C_k^{17} \varepsilon^k (1-\varepsilon)^{17-k},$$

여기서 ε을 변화시키면서 여러가지 다른 수준의 오류 추정치를 얻을 수 있다.

$\varepsilon = 0.2$이면 $\text{Prob}\{\theta \geq \varepsilon \mid 2, 16\} = 0.3096$이고,

$\varepsilon = 0.3$이면 $\text{Prob}\{\theta \geq \varepsilon \mid 2, 16\} = 0.07739$가 된다.

즉 단순화의 과정을 거쳐 추론된 개념의 추정오류가 20% 또는 30% 보다 클 확률이 각각 0.3096과 0.07739가 된다.

이번에는 단순화된 의사결정나무가 단순화 과정을 거치지 않은 원래의 의사결정나무보다 시험사례에서 더 나은 추정력(better classification power)을 가지는 일반적인 경우를 예로들어 보기로 한다. 훈련사례가 40개이고 단순화된 의사결정나무에 의해서 8개의 사례가 잘못 분류된다고 하자. 또한 20개의 시험사례를 취하여 단순화된 의사결정나무를 시험한 결과 2개의 분류오류가 발견되었다고 하자.

정리 3.7에 의한 방법 : 조건을 만족하는 $q^* = 18$이고 따라서$(p^* = 1, q^* = 18)$이 얻어진다.

$\varepsilon = 0.2$인 경우에 $\text{Prob}\{\theta \geq 0.2\} \leq \delta$, 이고 $1 - I_\varepsilon(p^* + b2, q^* + m2 - b2) = 1 - I_{0.2}(1 + 2, 18 + 16 - 2) = \underline{\delta} \leq \delta$이 된다.

왼쪽항을 계산하면 다음과 같이 된다.

$\text{Prob}\{\theta \geq 0.20\} \leq \delta$이고 $1 - I_{0.2}(3,32) = 0.0113 = \underline{\delta} \leq \delta$이 된다.

즉, 단순화과정을 거친 의사결정나무의 추정상의 오류가 20% 보다 크거나 같을 확률이 δ보다 작은데 δ의 하한 경계는 0.0113이다.

여러 다른 수준의 ε값에 대해 다음 결과를 얻을 수 있다.

$\text{Prob}\{\theta \geq 0.25\} \leq \delta$, $1 - I_{0.25}(3, 32) = 0.001641 = \underline{\delta} \leq \delta$

$\text{Prob}\{\theta \geq 0.30\} \leq \delta$, $1 - I_{0.30}(3, 32) = 0.000190 = \underline{\delta} \leq \delta$

$\text{Prob}\{\theta \geq 0.35\} \leq \delta$, $1 - I_{0.35}(3, 32) = 0.000018 = \underline{\delta} \leq \delta$

$\varepsilon = 0.2$인 경우에 일치되는 베타사전분포(consistent Beta prior)인 $(p, q) = (2, 20)$에 의하여 다음과 같이 하한 경계가 개선된다.

$\underline{\delta} = 1 - I_\varepsilon(p^* + b_2, q^* + m_2 - b_2)$

$= 1 - I_{0.1}(2+2, 20+16-2) = 0.0244.$

위와 같은 방법으로 탐색을 계속하여 작은 ε값에 대하여는 오류를 정확하게 추정하는 것이 가능하여 진다. 그러나 큰 ε값에 대하여는 이 방법이 엄격한 하한 경계값을 제공한다. 예를 들어 ε이 0.75보다 크면, $(p, q) = (2, 20)$이 경계값을 개선하지 못한다.

정리 3.8의 a)는 다음과 같이 오류를 분석한

다.

$\varepsilon=0.2$인 경우에 $\text{Prob}\{\theta\geq 0.2\}\leq \exp(-2(0.2-2/20)^2 20)=0.6703$,

$\varepsilon=0.3$인 경우에는 $\text{Prob}\{\theta\geq 0.3\}\leq 0.2019$이 된다.

정리 3.8의 b)에 의하면, $\varepsilon=0.2$인 경우에

$\text{Prob}\{\theta\geq 0.2\}<\exp(-2(0.2-2/20)^2 20-(4/3)(0.2-2/20)^4 20)=0.6685$이 되며, $\varepsilon=0.3$인 경우에는 $\text{Prob}\{\theta\geq 0.3\}<0.1935$, $\varepsilon=0.4$인 경우에는 $\text{Prob}\{\theta\geq 0.4\}<0.0220$이 된다.

정리 3.9에 의해 다음과 같이 추정된다.

$$\text{Prob}\{\theta\geq\varepsilon \mid 2, 20\}=\sum_{k=0}^{2} C_k^{21}\varepsilon^k(1-\varepsilon)^{21-k}.$$

여기서 ε을 변화시키면서 여러가지 다른 수준의 오류 추정치를 얻을 수 있다.

$\varepsilon=0.1$이면 $\text{Prob}\{\theta\geq\varepsilon \mid 2, 20\}=0.6484$이고,

$\varepsilon=0.2$이면 $\text{Prob}\{\theta\geq\varepsilon \mid 2, 20\}=0.1787$이고,

$\varepsilon=0.3$이면 $\text{Prob}\{\theta\geq\varepsilon \mid 2, 20\}=0.0271$이 된다.

이와 같이, 최저신뢰도 수준(the worst possible confidence factor) δ에 대하여, 학습정보를 이용하는 방법은 하한경계를 제공하고, 학습정보를 이용하지 않는 방법은 일반적인 상한경계(upper bound)를 제공한다. 우리는 이 결과를 결합하여 δ의 범위를 얻을 수 있다.

결과의 사용방법을 예시하기 위하여 20개의 사례를 사용하여 의사결정나무를 추론하였다고 하고, 단순화된 의사결정나무(pruned decision tree)에서는 6개의 분류 오류가 발생하였다고 가정하자. 또한 16개의 독립된 사례를 사용하여 시험한 결과 2개의 사례를 잘못 분류하였다고 가정하자.

정리 3.7에 의해 δ의 하한 경계를 계산한다.

$b/m=6/20=0.3$ 이므로, $\varepsilon=0.3$에 대하여 $(p^*=1, q^*=7)$이다.

따라서, $\text{Prob}\{\theta\geq 0.3\}\leq\delta$이고,

$1-I_\varepsilon(p^*+b_2, q^*+m_2-b_2)=1-I_{0.3}(1+2, 7+16-2)=0.0157=\underline{\delta}\leq\delta$이 된다.

정리 3.8의 b)에 의해 δ의 상한 경계를 계산한다.

$\text{Prob}\{\theta\geq 0.3\}<\exp(-2(0.3-2/16)^2 16-(4/3)(0.3-2/16)^4 16)=0.3678$

따라서 δ의 범위가 얻어진다.

$0.0157\leq\delta\leq 0.3678$.

즉, 단순화의 과정을 거쳐 추론된 개념의 오류가 30% 보다 크게될 확률이 최악의 경우에 0.3678까지 될 수 있으나 결코 0.0157 이하로는 되지 않음을 보여준다.

이와 같이 오류분석 방법론은 자동학습을 적용할 때, 학습의 결과로 도출되는 지식의 정확성에 대해 사전 예측을 가능하게 해주며, 오류의 수준을 조정할 수 있는 수단을 제공한다. 또한 이미 학습된 지식이나 사전 오류수준이 만족스럽지 못한 지식에 대하여 사후 오류분석을 수행하는 방법을 제공한다. 전문가시스템에서는 이와 같은 방법론을 포괄적으로 이용하여

시스템에서 사용하는 지식의 정확성에 대한 보다 정밀한 추정을 하고, 그 결과를 시스템 활용에 이용할 수 있다.

V. 다단계 자동학습 전략에의 적용

5.1 다단계 학습전략

본 절에서는 단일 학습 전략의 성능을 개선하기 위한 다단계 학습전략에 대해 분석한다. 다단계 학습전략의 필요성을 보이기 위하여 먼저 학습의 기본 구조를 살펴본다. 학습은 실 세계에 존재하는 많은 속성들로 부터 시작한다. 속성(feature 또는 attribute)중에는 표면적으로 잘 드러나는 속성도 있지만 분간하기 어렵거나, 숨어 있는 속성도 상당수 있을 수 있다. 대부분의 자동학습은 겉으로 드러난 일부의 속성만을 이용하여 학습을 수행하게 된다. 따라서 학습의 전체적인 성능을 높이기 위해서는 존재하는 모든 속성중에서 의사결정과 관련이 많은 소수의 속성을 추출해내는 작업과 추출된 속성을 이용하여 학습을 진행하는 과정 모두의 성능을 높여야 한다. 전체적인 학습과정을 도시하면 다음과 같다.

실세계의 속성 집합→속성 추출

→분류(classification)등의 학습 수행

의사결정나무추론과 같은 대부분의 자동학습방법은 문제에 적합한 속성이 이미 추출되어 정리되어 있다고 가정하고, 분류학습을 수행한다. 따라서, 학습의 성능이 학습에 사용된 속성 집합의 품질에 의해 많은 영향을 받게 된다. 즉 자체적으로 훌륭한 학습알고리즘일지라도, 속성의 대표성이 적어 좋은 평가를 받지 못하는 경우도 발생하고, 반대로 평범한 학습알고리즘도 속성의 대표성이 좋아 상당한 효과를 거두는 것으로 보고되기도 한다.

이러한 문제점을 인식하고 최근에 들어 속성추출에 대한 관심이 고조되고 있으나 아직 괄목할만한 가시적인 성과는 없는 실정이다. 앞서 언급한 바와 같이 의사결정나무 학습에서는 숨겨진 속성으로 인한 학습효율의 저하 방지를 위하여 단순화기법(pruning)기법을 활용하여 많은 성과를 거두고는 있으나, 아직 개선해야 할 부분이 많은 상황이다. 따라서, 자동학습 알고리즘의 전체적인 성능향상을 위해서는 속성추출 부분을 중심으로 많은 개선이 필요한 상황이다. 본절에서는 유전자 알고리즘과 의사결정나무, 신경망 기법 등을 활용하여 다단계 학습 전략을 구축하는 대안을 제시하여 전역적 성능향상 방안을 제시한다.

5.1.1 유전자알고리즘과 귀납학습

의사결정나무추론등의 귀납학습에서 의사결정에 가장 영향이 큰 속성을 차례로 선택하게 되는데, 정보이론(Information theory)에 근거한 방법등을 현재 사용하고 있다.[Breiman et al.,1984 ; Quinlan, 1986 ; Marshall, 1986 ; Mingers, 1986, 1989a]. 그러나 이들 방법은 모두 발견적인 해법으로 개선의 여지가 많은 방법들로서 추가적인 연구가 필요한 상황

이다. 또한 학습 대상 문제가 매우 큰 경우, 즉 속성의 수가 매우 많은 경우에는 변별력이 큰 속성을 추출하는 일이 매우 중요하게 된다.

유전자 알고리즘은 유전법칙을 이용하여 최적해에 빠르게 접근하는 능력을 보이고 있으므로, 이 능력을 이용하여 귀납학습에서 속성 선택의 불완전함을 보완할 수 있다. 즉, 전체 속성의 수가 매우 많은 경우에, 가장 정확도가 높은 개념(지식)을 학습하는 수단으로 유전자 알고리즘을 이용할 수 있다.

n개의 속성이 존재하는 학습상황에서는 2n개의 속성집합으로 모두 학습이 가능하다. 이 중 어느 하나의 속성 집합에 가장 정확성이 높은 지식이 존재한다. 즉, 우리가 학습하는 공간은 숨겨진 속성이 존재하고, 분류나 속성의 오류도 존재하는 사례 공간이기 때문에 모든 속성집합에 대해 최적해로서의 가능성을 부여해야 한다. 또한 시험사례(test set)에서의 오류 정확성이 선형적인 단순구조를 보이지 않기 때문에 속성집합의 각 부분집합에 대하여 학습을 시도하는 것이 의미가 있다. 그런데, n이 큰 경우에는(수십 이상) 가능한 모든 경우의 속성집합에 대하여 학습하는 것이 사실상 불가능하다. 따라서 유전자 알고리즘을 도입하여 정확성이 높은 속성 집합을 신속하게 찾아내어 학습능력을 높이는 것이 본 전략의 기본 개념이다.

전체적인 학습의 골격은 다음과 같이 구성할 수 있다.

1) 속성의 학습 활용여부를 각각 비트 스트링으로 표현한다. 즉 (x_1, x_2, ... ,x_i,..... , x_n)과 같이 n개의 속성이 있을 경우, n개 속성 전부가 학습에 활용되는 상황은 n개의 1(111111..... 11)로 표시한다. 또 x_1과 x_n만 학습에 활용될 경우는(100.... 0001)과 같이 표현한다.

2) 초기 집단(population)을 무작위로 생성한다. 문제의 크기에 따라 초기 집단을 수십 내지 수백개의 크기로 생성한다. 생성된 비트스트링은 일상적인 유전자알고리즘의 입력 초기집단으로 인식되어 학습을 개시한다.

3) 각각의 속성집합에 대하여 의사결정나무 학습방법이 사용하고 있는 속성선택 척도(selection measure)를 사용하여 일상적인 의사결정나무 학습을 진행한다. 채택된 학습알고리즘에 따라 단순화기법을 사용할 수 있다.

4) 학습된 지식을 시험사례로 시험하여 지식의 정확도를 계산한다. 계산된 정확도는 비트스트링(속성의 부분집합)의 적합도(fitness)로서의 의미를 가진다.

5) 재생산(reproduction), 교배(cross-over), 돌연변이(mutation)등 유전자알고리즘의 진화 연산자를 사용하여 다음 세대의 속성집합을 생성한다. 유전자 알고리즘 학습의 종료 조건이 만족되었으면, 학습을 중지하고 가장 정확도가 높은 속성집합 및 이 집합을 이용하여 학습된 지식(의사결정나무)을 학습의 결과로서 출력한다. 종료조건이 만족되지 않았을

경우 단계 3으로 돌아가 학습을 계속하고 매단계마다 종료조건을 검사한다.

이 방법으로 학습을 진행하면 속성의 모든 조합 가능성에 대해 학습을 진행함으로써, 속성 추출 기법의 불완전성을 많이 보완할 수 있고, 각각의 단위 알고리즘을 설계하는 방법에 따라, 숨겨진 속성 및 잡음의 문제도 동시에 해결할 수 있게 될 것으로 기대된다.

5.1.2 귀납학습과 신경망 기법

신경망 기법은 예측 능력면에 있어서는 매우 뛰어나지만 학습의 속도가 느리고, 학습된 내용의 설명력이 부족하다는 이유 때문에 경영의 사결정에 활용이 지연되고 있다. 우선 학습의 속도는 처리노드 수에 많은 영향을 받게 되는데, 처리노드수는 또한 입력데이터 수에 영향을 받게된다. 학습 대상 문제의 속성 수가 매우 많은 경우 일부의 신경망학습에서는 통계적인 기법을 활용하여 변별력이 높은 속성(변수)를 선택하는 전처리 작업을 수행하고 있다. 주로 많이 사용되는 통계적 방법은 요인분석(factor analysis)과 계단식회귀분석(stepwise regression)방법이다. 요인분석은 전체 속성을 몇개의 의미있는 그룹으로 나누면서 요인을 추출하는 방법으로서, 속성의 수를 줄이는데 기여하게 된다. 그러나 이 방법은 종속변수를 고려하지 않고 각 독립변수의 특성만을 고려하기 때문에 한계가 있는 방법으로 인식되고 있다. 계단식회귀분석 방법은 정방향으로 변수를 추가하고(forward selection), 역방향으로 변수를 제거하는(backward elimination) 회귀분석 방법이다. 미리 정해진 변수 선택 기준에 의해 가장 적합한 변수를 골라 모형에 포함한다. 첫번째 변수가 선택된 후에, 남은 변수중에서 가장 적합한 변수를 골라 다시 모형에 추가하고, 두번째 변수가 추가된 이후, 기존의 변수(여기서는 첫번째 변수)를 다시 심사하여 삭제조건에 해당되는지 판단하고, 변수가 삭제 조건에 해당하면 삭제한다. 이렇게 변수 선택과, 변수삭제를 반복하여 더 이상 진입조건이나, 삭제조건을 만족하는 변수가 없을때까지 계속한다. 이 방법은 종속변수를 중심으로 각 속성의 영향력을 판단하는 방법이므로 요인 분석보다 유의한 방법이라 할 수 있다. Kwon[1995]은 기업신용평가(Bond Rating)문제를 위해 회귀분석 방법을 적용하였는데, 126개의 재무비율(financial ratio) 변수 중에서 24개의 신용평가에 변별력있는 변수를 추출한 바 있다.

그러나 변수의 수를 간소화하는 목적과 설명력을 부여하는 목적을 위해서는 귀납학습의 원리를 이용하는 것이 보다 바람직한 대안으로 대두된다. 즉 의사결정나무 추론 기법의 변수 선택 기준(selection measure)을 이용하여 의사결정에 영향력 있는 변수를 선택하는 것이 신속성과 효과면에서 보다 희망적인 대안이 될 수 있다. 따라서 본 전략은 다음과 같이 구성된다.

1) 의사결정나무 추론 알고리즘을 이용하여 사례를 통한 학습을 수행한다. 이때 사례의 수가 매우 많은 경우 표본을 추출하여

수행할 수 있다.

2) 학습의 결과물인 의사결정나무에 사용된 변수(속성)들을 신경망의 입력 노드로 채택하여 신경망 학습을 수행한다.
3) 신경망 학습의 결과를 획득된 지식으로 채택하며, 최초에 구축된 의사결정나무는 설명을 위한 참고자료로 활용한다. 필요시 다양한 의사결정나무 구축 알고리즘을 사용하여 변수를 추출하고, 신경망 학습을 수행하여 학습된 개념의 정확도를 높일 수 있다.

5.1.3 신경망과 유전자알고리즘

또한 유전자 알고리즘과 신경망기법을 결합할 수 있다. 이 방법은 신경망의 가중치를 학습하는 방법을 유전자알고리즘을 이용하여 개선하는 구조로 시도될 수 있다.[de Garis,1994] 이 경우 학습과정은 다음과 같다.

1) 가중치를 코딩하고 초기집단을 구성한다.

   신경망이 n개의 가중치를 학습하는 구조로 설계되었다면, n*(p+1)개의 비트(bit)로 하나의 스터럭쳐를 나타낸다. 여기서 p는 각각의 가중치 하나를 나타내는 비트 수이며, 1을 더하는 것은 가중치의 부호를 표시하기 위함이다. 가중치의 수는 네트워크의 구조에 따라 노드수에 근거하여 계산된다. 초기집단은 무작위로 생성한다.
2) 각 스트링의 적합도(fitness)를 결정한다.

   적합도는 유전자 알고리즘의 학습 성과에 많은 영향을 미친다. 가장 손쉬운 적합도의 척도는 아래와 같이 요구되는 출력값(di)과 실제출력값(ai)의 오차합을 이용하는 것이다.

   적합도(fitness) $= 1/\sum(d_i - a_i)^2$

   위 식의 오차항을 가중합으로 변형하여 다음과 같은 적합도를 사용할 수도 있다.

   적합도(fitness) $= 1/\sum i*(d_i - a_i)^2$
3) 유전자알고리즘에 의해 학습을 수행한다.

   학습수행결과로 수렴된 스트링집합에서 적합도가 높은 스트링의 가중치들을 신경망의 학습결과로 추출하여 사용한다.

이와 같은 학습 전략은 인공신경시스템(artificial nervous systems)을 구축하는 도구로 활용될 수 있으며, 마케팅 등의 복잡한 경영분야 문제해결에도 적용될 수 있을 것이다.

5.2 다단계 전략의 성능 분석

의사결정나무 추론의 경우, 앞서 분석한 바와 같이 가능한 사례종류의 총 수를 d라 할때, $|H| = 2^d$가 된다. 여기에서 모든 속성이 범주변수(categorical variable)이거나 유한한 이산변수(discrete variable)인 경우는 $|H|$가 유한하지만, 속성중의 일부가 무한한 이산변수이거나 연속변수(continuous variable)이면 $|H| = \infty$가 되어 반드시 VC차원을 이

용해야 유의한 결과를 도출할 수 있다.

실수 변수의 경우, 대개 가설공간이 하나 또는 2~3개의 구간으로 정의되는데, 구간의 수에 따라 VC차원의 크기가 달라지게 된다. 예를 들어, 하나의 구간으로 개념을 나타내는 경우 2개 까지의 사례는 하나의 구간으로 모든 경우에 대해 개념을 나타낼 수 있지만, 3개의 사례는 나타낼 수 없는 경우가 존재한다.(즉, 가운데의 사례만이 음(-)의 사례인 경우는 단일 구간 개념으로는 나타낼 수 없다). 이 경우 VC차원은 2가 된다.

유전자 알고리즘의 경우는 각각의 속성이 아닌, 개별적인 개념이 하나의 단위가 되어 학습되는 형태이므로 위의 구조로 분석하기는 어렵다. 신경망 기법의 경우 학습대상은 가중치(weights)이므로, 모두 실수값으로서 무한대의 S 값을 가지게 된다. 따라서 이러한 체계로는 분석이 불가능하고, 대안으로서 의사결정이론을 이용한 분석이 시도되고 있다.[Haussler, 1990]

그러나 앞 단원에서 제시한 사후적 추정방법은 어떠한 학습방법에도 적용할 수 있는 오류분석 방법이므로 다단계 학습전략에도 적용할 수 있는 개념 정확성 분석 방법이 된다. 이 방법은 3가지 기본 구조를 가지는데, 사전정보가 없는 경우, 학습으로부터의 정보를 이용하는 경우, 일반적인 경우 등 3가지로 상황을 나누어 오류분석 방법을 제시한다.

사전정보가 전혀 없는 경우, 균등사전분포(uniform proir)를 가정하고 사후적 추정치를 도출하여 사용한다. 균등사전분포는 사전정보가 없는 경우에 가장 흔히 사용되는 가정방법이다. 학습으로부터의 정보를 이용할 수 있는 경우는, 베타사전분포(beta prior)를 가정하고 학습정보와 일치하는 사후분포를 도출하여 보다 정확한 오류 추정이 가능하다. 또한 어떠한 가정도 하지 않은 상태에서 일반적인 오류경계를 도출하여 사용함으로서 사전분포 가정으로 인한 위험을 최소화하고, 오류범위에 대한 포괄적인 정보를 획득할 수 있다.

Ⅵ. 요약 및 결론

본 연구에서는 자동학습을 통하여 획득된 지식의 오류를 분석하는 통합된 체계의 오류분석 방법론을 정립하였다.

우선 최근에 개발된 자동학습이론을 의사결정 도메인에 도입하여 재정립하고, 이를 의사결정나무 추론에 적용하여 응용 결과를 도출하였다. 이는 두단계를 거쳐 적용되는데 먼저 지정된 신뢰도 수준(specified confidence level)을 가지고 지정된 오류보다 작은 수준의 오류를 가지는 의사결정나무를 추론할 수 있도록 보장하는데 충분한 사례의 양(sample size)을 도출하였다. 단순화기법을 사용한 의사결정나무와 사용하지 않은 의사결정나무에 대해 각각의 결과를 언급하였다.

다음으로는 위에서 도출된 충분한 양의 사례를 얻기가 어려울 경우에 사후적인 방법으로 오류의 수준을 측정하는 방법을 제공하였다. 학습의 정보를 이용하는 경우와 학습정보를 이용하지 않는 경우로 구분하여 분석하였고, 학

습정보를 이용하지 않는 경우는 오류수준의 분포에 대한 가정을 하는 경우와 가정이 전혀 없는 일반적인 경우에 적용할 수 있는 방법을 각각 제시하였다.

세번째로 오류분석 방법론의 적용범위를 확대하기 위하여 다단계 학습전략을 제시하고, 이러한 전략을 통하여 학습된 개념의 정확성을 측정하는데 앞서 개발된 방법들이 어떻게 적용될 수 있는지를 분석하였다.

본 논문은 전문가시스템 실무자들 및 자동학습의 활용에 관심이 있는 연구자들을 위하여 오류분석 방법론을 개발하고, 이의 활용방법을 경영문제 사례를 통하여 보여주는데 의의가 있다. 향후 의사결정나무 학습이외의 단일 학습전략에 대해서도 실용적인 결과도출 노력이 필요하며, 다단계 전략에 대한 보다 심층적인 연구 및 이론적인 체계 정립이 중요한 연구방향으로 대두된다.

참 고 문 헌

Han, In-goo, Young-sik Kwon, and Hong-kyu Jo, A Review of Artificial Intelligence Models in Business Classification, 한국전문가시스템학회지, 창간호, 1995, pp.23-41

Kwon, Young-sik, Ordinal Pairwise Partitioning (OPP) Approach to Neural Networks Training : Bond Rating Case, unpublished manuscript, 동국대학교, 1995

Abramowitz, M. and Segun, I., *Handbook of Mathematical Functions*, Dover Publications, New York, 1968.

Angluin, D. and Laird, P., "Learning from Noisy Examples." *Machine Learning*, Vol. 2, 1988, pp. 343-373.

Braun, H. and Chandler, J.S., "Predicting Stock Market Behavior through Rule Induction : An Application of the Learning-from-Example Approach", *Decision Sciences*, 18, 1987, pp.415-429.

Breiman, L., Freidman, J., Olshen, R. and Stone, C., *Classification and Regression Trees*, Wadsworth International, California, 1984.

Carter, C. and Catlett, J., "Assessing Credit Card Applications using Machine Learning", *IEEE Expert*, Fall, 1987, pp.71-79.

de Garis, Hugo, "Genetic Programming : Evolutionary Approaches to Multistrategy Learning", in *Machine Learning Vol.IV-Multistrategy Approach*, Ryszard Michalski and Gheorghe Tecuci eds., 1994, pp. 549-577

Ehrenfeucht, A. and Haussler, D., "Learning Decision Trees From Random Examples", *Proceedings of the 1988 Workshop on Computational Learning Theory*, Morgan Kaufmann, San Mateo, CA, 1988, pp.182-194.

Fisher, D. H. and Schlimmer, J. C., "Concept Simplification and Prediction Accuracy", *Proceedings of the 5th International Conference on Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1988, pp.22-28.

Haussler, D., "Quantifying Inductive Bias : AI Learning Algorithms and Valiant's Learning Framework," *Artificial Intelligence*, 36, 1988, pp. 177-221.

Haussler, D., "Decision Theoretic Generalization of the PAC Model for Neural Net and Other Learning Applications," Technical Report, UCSC -CRL-91-02, University of California, Santa Cruz, 1990.

Hoeffding, W., "Probability Inequalities for Sums of Bounded Random Variables", J. *American Statistical Association*, Vol. 58, 1963, pp.13-30.

Johnson, N. and Kotz, S., Discrete Distributions, Houghton Mifflin Co., Boston, 1969.

Kim, H. and Koehler, G. J., "PAC Learning a Decision Tree with Pruning", *The European Journal of Operational Research*, Volume 94, 1996, pp.405 -418.

Kim, H. and Koehler, G. J., "The Accuracy of Decision Tree Induction in a Noisy Domain for Expert Systems Construction", *Intelligent Systems in Accounting, Finance & Management*, Volume 3, Number 2, 1994a, pp. 89-98.

Kim, H. and Koehler, G. J., "An Investigation on the Conditions of Pruning an Induced Decision Tree", *The European Journal of Operational Research*, Volume 77, Number 1, 1994b, pp. 89-95.

Kim, H. and Koehler, G. J., "Theory and Practice of Decision Tree Induction", *Omega*, Vol.23, No.6, 1995, pp. 637-652.

Marshall, R., "Partitioning Methods for Classification and Decision making in medicine", *Statistics in Medicine*, Vol. 5, 1986, pp.517-526.

Messier, W.F. and Hansen, J.V., "Inducing rules for Expert Systems Development", *Management Science*, Vol. 34, No.12, 1988, pp.1403-1415.

Michalski, R.S. and Chilausky, C., "Learning by being told and learning from examples : An Experimental comparision of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis", *International Journal of Policy Analysis and Information Systems*, 4, 1980, pp.125-161.

Mingers, J., "Expert Systems"Experiments with rule induction", *Journal of the Operational Research Society*, Vol.37, 1986, pp.1031-1037.

Mingers, J., "An Empirical Comparison of Selection Measures for Decision Tree Induction", *Machine Learning*, Vol. 3, 1989a, pp.319-342.

Mingers, J., "An Empirical Comparison of Pruning Methods for Decision Tree Induction", *Machine Learning*, Vol. 4, 1989b, pp.227-243.

Niblett, T. and Bratko, I., "Learning Decision Rules in Noisy Domains", In M.A. Bramer (Ed.), *Research and Development in Expert Systems* III, Cambridge University Press, Cambridge, 1986, pp.25-34.

Niblett, T., "Constructing Decision Trees in Noisy Domains", *Proceedings of the Second European Working Session on Learning*, Bled., Yugoslavia : Sigma Press, 1987, pp.67-78.

Quinlan, R., "Discovering Rules from large collection of examples : A case study" In D. Michie (Ed.), *Expert systems in the microelectronic age.* Edinburgh : Edinburgh University Press, 1979.

Quinlan, R., "The effect of Noise in Concept Learning", In R.S. Michalski, J. Carbonell, T. Mitchell(Eds.), *Machine Learning : An Artificial Intelligence Approach. Vol.* II, Morgan Kaufmann, Los Altos, CA, 1983.

Quinlan, R., "Induction of Decision Trees," *Machine Learning*, Vol. 1, 1986, pp.86-106.

Quinlan, R., "Simplifying Decision Trees", *International Journal of Man-Machine Studies*, 27, 1987a, pp.221-234.

Quinlan, R., "Generating Production Rules from DecisionTrees", *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Los Altos, CA, 1987b, pp. 304-307.

Shaw, M. J. and Gentry, J.A., "Using an Expert System with Inductive Learning to Evaluate Business Loans", *Financial Management*, Autumn 1988, pp. 45-56.

Smyth, P., Machine Learning : Theory and Application, *Tutorial in 3rd World Congress on Expert Systems*, Seoul, Korea, 1996.

Spangler, S., Fayyad, U.M. and Uthurusamy, R., "Inductionof Decision Trees from Inconclusive Data", *Proceedings of the 6th International Conference on Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1989, pp.146-150.

Tsai, L. and Koehler, G.J., "The Accuracy of Concepts Learned from Induction", *Decision Support System*, Vol.10, 1993, pp.161-172.

Utgoff, P., "Incremental Induction of Decision

Trees", *Machine Learning*, Vol.4, 1989, pp.161-186.

Vafaie, Heleh and Kenneth De Jong, "Improving a Rule Induction System using Genetic Algorithms", in *Machine Learning Vol.IV-Multistrategy Approach*, Ryszard Michalski and Gheorghe Tecuci eds., 1994, pp. 453-470

Van de Velde, W., "Incremental Induction of Topologically Minimal Decision Trees", *Proceedings of the 7th International Conference on Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1990, pp.66-74.

Vapnik, V.N., Estimation of dependencies based on empirical data, Springer-Verlag, New York, 1982.

◇ 저자소개 ◇

저자 김현수는 서울대 공대에서 학사, 한국과학기술원에서 경영과학으로 석사, 미국 University of Florida에서 경영정보학 박사를 취득한후, 현재 국민대학교 경상대학 정보관리학과 교수로 재직하고 있다. (주)데이콤의 주임연구원, 한국정보문화센터의 정책연구부장으로 재직한바 있으며, 주요관심분야는 정보시스템계획, 시스템분석, 전문가시스템, 정보통신정책등이며, Omega, European Journal of Operational Research, Intelligent Systems in Accounting, Finance and Management 등의 학술지에 논문을 발표한 바 있다.