전문가시스템을 위한 실제적인 상황에서의 지식획득 방법

A knowledge acquisition method for expert systems in a practical situation

김 현 수 (한국정보문화센터 정책연구부)

Induction methods have recently been found to be useful in a wide variety of business related problems, including in the construction of expert systems. Decision tree induction is an important type of inductive learning method. Empirical results have shown that pruning a decision tree usually improves its accuracy. In this paper we summarize theoretical results of pruning and illustrate these results with an example. We give a sample size sufficient for decision tree induction with pruning based on recently developed learning theory. For situations where it is difficult to obtain large enough sample, we provide several methods for a posterior evaluation of the accuracy of a pruned decision tree. Finally we summarize conditions under which pruning is necessary for better prediction accuracy.

I. 서론

전문가시스템은 현대의 기업경영에서 필요불가결한 도구가 되어가고 있다. 그러나, 전문가를 면담하는등의 재구성방법(reconstruction methods)으로는 전문가의 지식을 끌어내기 어렵기 때문에 [Musen, 1989] 전문가시스템개발에 있어서 지식획득이 가장 어려운 문제가 되고 있다. 귀납적 추론 방법 (Induction Methods)은 경영분야에의 응용을 위한 지식획득방법의 희망적인 대안이 되고 있다. [Braun and Chandler, 1987], [Messier and Hansen, 1988]

단순하지만 가장 성공적인 귀납적 추론방법은 사례(Examples)로부터 의사결정나무(Decision tree)를 추론하는 방법이다. 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] 최근에 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 나타내기

보다는 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 확율이 높지만 전체 사례공간(instance space)에서는 오류확율을 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.

이 논문에서 우리는 의사결정나무에 적용되는 단순화 기법에 대한 여러가지 이론적인 결과들을 요약하고 그 유용성을 증명하고자 한다.

본 논문의 주요목적은 적절한 오류분석 방법론의 결여를 이유로 귀납적 추론 방법의 사용에 대하여 회의적인 입장에 대해 반론을 제기하려는데 있다. 여기서 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 방법을 제공한다.

II 장에서는 의사결정나무의 귀납적 추론 방법 및 단순화 기법에 대하여 간단히 살펴보고, III 장에서는 원하는 오류수준을 보장하는데 필요한 사례(example)의 충분한 양(sufficient bound)을 제공한다. 또한 충분한 양의 사례를 구하기가 어렵거나 불가능한 경우에 추론된 의사결정나무의 정확성을 사후적으로 평가하는 방법을 제시한다. IV 장에서는 이러한 결과들을 예시하는 보기를 들어 설명하고 V 장에서는 단순화기법을 사용하면 예측정확성을 높일 수 있게 되는 이론적인 조건에 대한 연구를 요약한다. VI 장에서는 요약 및 미래의 연구방향에 대하여 언급한다.

II. 의사결정나무의 귀납적 추론

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 전문가시스템을 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. [Niblett,1987].

1. 의사결정나무 구축

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].

2. 의사결정나무의 단순화

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 없거나 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. [Niblett, 1987].

의사결정나무 구축시의 단순화 기법은 국지적인 정보에만 의존하여 나무확대에 대한 의사결정을 한다는데에 그 단점이 있다. 즉, 어느 한 노드에서 확대 중지 결정을 내리게 되었을때 그 노드의 하위 노드에서 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 있

다. 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.

3. 예제

여기에서 우리는 의사결정나무를 단순화하는 예제를 보여주고자 한다. Messier and Hansen [1988] 은 ID3를 이용하여 회사의 채무불이행(loan default) 행태를 예측하는 규칙을 제공하였다. 그들은 18개의 재무비율(financial ratio)과 추세를 이용하여 규칙을 추론하였다. 그림 1 은 추론된 의사결정나무를 수정하여 모든 변수를 이진 변수로 표현한 의사결정나무이다. 각 속성에 대한 "높음" 과 "낮음" 의 정확한 의미는 다음과 같다.

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

그림 1: 채무불이행을 예측하는 의사결정나무

```
                 유동비율
        낮음 /            \높음
            /              \
불이행                     장기부채/자기자본
(10)*                  낮음/            \높음
                          /              \
          저장기부채/자기자본        운전자본/매출액
              낮음/    \높음          낮음/        \높음
                 /      \                /          \
      순이익/총자산      이행      순이익/매출액      불이행
      낮음/      \높음    (11)     낮음 / \높음        (4)
         /        \                    /   \
불이행         이행             불이행    이행
 (1)            (1)              (1)       (4)
```

*: 각 최종 노드에 속하는 사례의 수

의사결정나무 단순화의 예를 들기 위하여 Breiman et al.[1984]이 제안한 오류-복잡도 방법과 Niblett and Bratko [1986] 의 최소오류법을 사용하기로 한다. 오류-복잡도의 척도인 α 를 각 단말 노드 t 에 대하여 다음과 같이 정의한다.

$$\alpha = (e_p - e_u)/(n_t - 1)$$

여기서 e_p = 노드 t 를 잘라냈을 경우의 오류율

e_u = 노드 t 를 잘라내지 않았을 경우의 오류율

n_t = 노드 t 의 아래에 있는 단말 노드의 수.

오류-복잡도 방법은 각 비단말 노드에 대해 α 를 계산하고 가장 작은 α 값을 가지는 노드를 잘라내기로 결정한다. 즉, 가장 약한 연결고리를 발견하여 끊어주는 방법이다. 이 과정을 반복하여 별도의 시험사례집합을 이용하여 가장 오류확율이 작은 의사결정나무를 최종적으로 선택한다. 각 노드의 α 값이 그림 2에 나타나 있다. 여기서 가장 α 값이 작은 노드인 "저장기부채/자기자본" 노드를 잘라내고 난 후의 의사결정나무가 그림

3에 나타나 있다.

그림 2: 각 노드의 α 값을 표시한 의사결정나무

```
              유동비율(1/12)**
        낮음 /             \높음
            /               \
     불이행                  장기부채/자기자본(3/80)
     (10)*                낮음/           \높음
                             /             \
   (1/64)저장기부채/자기자본           운전자본/매출액(1/16)
              낮음/   \높음            낮음/          \높음
                 /     \                  /            \
(1/32)순이익/총자산      이행     순이익/매출액(1/32)      불이행
     낮음/     \높음     (11)     낮음 / \높음             (4)
        /       \                    /   \
     불이행      이행             불이행    이행
      (1)        (1)               (1)     (4)
```

*: 각 최종노드에 속하는 사례의 수

**: 각 노드의 α 값

그림 3: 채무불이행을 예측하는 단순화된 의사결정나무

```
              유동비율(15/128)**
        낮음 /             \높음
            /               \
     불이행                  장기부채/자기자본(5/96)
     (10)*                낮음/           \높음
                             /             \
                         이행            운전자본/매출액(1/16)
                        (12:1)***         낮음/          \높음
                                             /            \
                              (1/32)순이익/매출액          불이행
                                    낮음 /  \높음          (4)
                                        /    \
                                     불이행    이행
                                      (1)     (4)
```

*: 각 최종노드에 속하는 사례의 수

**: 각 노드의 α 값

***: 12개의 사례가 "이행"에 속하고 1개의 사례가 "불이행"에 속한다.

그림 3 에서 "순이익/매출액" 노드가 가장작은 α 값을 가지고 있으므로 이 노드를 잘라내게 되는데 그림 4에서 그 결과를 볼 수 있다.

그림 4: 채무불이행을 예측하는 단순화된 의사결정나무

```
          유동비율(14/96)**
    낮음 /          \높음
        /            \
 불이행              장기부채/자기자본(1/16)
 (10)*             낮음/            \높음
                    /                \
                  이행               운전자본/매출액(3/32)
                (12:1)***            낮음/       \높음
                                      /           \
                                    이행         불이행
                                    (4:1)          (4)
```

*: 각 최종노드에 속하는 사례의 수

**: 각 노드의 α 값

***: 12개의 사례가 "이행"에 속하고 1개의 사례가 "불이행"에 속한다.

그림 5는 그림 4에서 가장 작은 α 값을 가지는 노드인 "장기부채/자기자본" 노드를 잘라낸 후의 의사결정나무를 보여주고 있다.

그림 5: 채무불이행을 예측하는 단순화된 의사결정나무

```
              유동비율
      낮음 /          \높음
          /            \
      불이행          이 행
      (10)*          (16:6)**
```

*: 10개의 사례가 "불이행" 그룹에 속한다

**: 16개의 사례가 "이행" 그룹에 속하고, 6개의 사례가 "불이행" 그룹에 속
한다.

각각의 가지쳐진 의사결정나무는 독립 시험사례 (independent test set)를 분류하는데 사용되는데 최소 오류에서 표준오차(one standard error) 범위내의 오류를 가지는 가장 크기가 작은 의사결정나무를 최종적으로 선택하게 된다. [Breiman et al., 1984], [Mingers, 1989b].

Messier and Hansen [1988] 이 사용했던것과 동일한 유보표본(hold-out sample)을 사용하여 다음과 같은 결과를 얻었다.

단순화된 의사결정나무의 노드수	오류 비율
6 (그림 1)	2/16
4 (그림 3)	2/16
3 (그림 4)	2/16
1 (그림 5)	2/16
0	0 또는 1*

*: 동일한 숫자의 "이행" 및 "불이행" 사례가 있으므로 오류비율은 레이블 부여 방법에 달려있다.

여기서 우리는 뿌리노드(root node) 하나만 가지고 있는 그림 5 의 의사결정나무를 단순화된 가장 좋은 나무로 선택하게 된다. 위의 유보표본은 모두 하나의 집단(class)에 속해있기 때문에 무작위의 표본이라고 생각되지 않을 수도 있다. 일반적으로 오류비율은 무작위의 표본에서는 컨벡스(convex)에 가까운 형태를 보여준다. 즉, 단순화의 정도가 작으면 오류비율은 감소하여 가고, 너무 잘라내는 양이 많아지면 오히려 오류비율이 증가하는것이 일반적인 현상이다.

이번에는 최소오류법을 적용하여 보기로 한다. 최소오류법은 각 집단(class)이 발생할 확율이 동일하다는 가정하에 그림 1의 단순화되지 않은 의사결정나무가 가장 작은 기대오류비율(expected error rate)을 가지며, 다른 어떠한 단순화된 의사결정나무도 기대오류비율이 이것보다 크다는것을 보여준다. 최소오류법은 단순화한 후에 기대오류비율이 작아지게 되는 경우에 한해서 그러한 노드를 잘라낸다. Niblett and Bratko [1986]는 이진나무(binary tree)에 있는 노드의 기대오류비율(E)을 다음과 같이

정의한다.

$$E = (N-n+1)/(N+2),$$

여기서 N = 총 사례의 수

n = 사례의 수가 많은 집단에 속한 사례의 수

각 노드의 기대오류비율은 각 분기(branch)에 속한 사례(example)의 비율에 따라 각 분기의 오류비율을 가중평균하여 계산한다. 예를들어, 그림 1에서 "순이익/매출액" 노드를 잘라내면 E 는 다음과 같이 계산된다.

$$E = (5-4+1)/(5+2) = 2/7$$

위의 노드를 잘라내지 않고 그대로 둘 경우 기대오류비율은 다음과 같이 계산된다.

$$E = (1/5)(1-1+1)/(1+2) + (4/5)(4-4+1)/(4+2) = 1/5$$

위에서 보는바와 같이 의사결정나무를 단순화 하는것이 기대오류비율을 높여주기 때문에 "순이익/매출액" 노드를 그대로 두게 된다. 이와 같은 절차와 계산을 모든 비단말 노드(nonterminal node)에 적용시키면 그림 1의 단순화하지 않은 의사결정나무가 가장 작은 기대오류비율을 갖게된다는 것을 알 수 있다.

III. 이론적인 결과 및 사후적 추정방법

이단원에서는 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 ε 보다 클 확율이 δ 보다 작게되는 의사결정나무를 발견할 가능성을 보장하기에 충분한 사례의 수를 결정하고자 한다. 사례의 추출은 반복을 허용하는 추출법을 사용한다. 이러한 유형의 학습방법을 팩(PAC: Probably

Approximately Correct) 학습이라고 한다. [Valiant, 1984], [Angluin and Laird, 1988].

1. 단순화 과정이 없는 경우의 분석

여기서는 사례간에 모순(inconsistency)이 없다는 제약하에서 의사결정나무를 구축한다. 즉, 속성의 값들이 동일한 두개의 사례는 반드시 동일한 분류에 속하여야 한다.

추론된 의사결정나무는 모든 훈련사례들(Training set)을 완벽하게 분류한다. Tsai and Koehler [forthcoming]는 주어진 오류수준과 신뢰수준에 대한 충분한 사례의 수를 제공하는 연구결과를 발표하였다.

사례수를 결정하기 위하여 컴비네토리알 파라메터(Combinatorial parameter)로서 가설공간(Hypothesis Space), H, 의 Vapnik-Chervonenkis 차원인 VCdim(H)을 정의할 필요가 있다. 사례의 속성들의 집합을 X 라 하면 모든 가능한 개념의 집합은 C = 2^X 가 되며 가설공간(H)은 C의 부분집합으로서 추론알고리즘에서 채택하고 있는 인덕티브 바이아스(Inductive Bias)에 의해 결정된다. (VCdim(H)의 상세 정의는 Vapnik[1982] 참조)

d 를 가설공간(H)이 의사결정나무인 경우의 VCdim(H)이라 하자. 다음 정리 1 은 ID3에 대한 충분한 사례수를 제공한다.

정리 1: <u>(ID3에 대한 사례수: Tsai and Koehler[forthcoming])</u>

1. 0 과 1 사이의 값을 가지는 어떠한 ε과 δ에 대해서도 사례수(sample size)가 $[\ln(1/\delta) + d \ln 2] / \varepsilon$ 보다 크게되면 ID3는 최소한 $1-\delta$ 의 확율로서 오류가 ε 보다 작은 개념 h 를 찾아낼 수 있다.
2. ε 이 0 과 1/2 사이의 값을 가질때 ID3는 최소한 다음 크기의 사례를이용해야 한다.
   $\max\{[(1-\varepsilon)/\varepsilon]\ln(1/\delta),\ d[1-2(\varepsilon(1-\delta)+\delta)]\}$

예를들어, 속성의 갯수가 5 이고, 각 속성은 5가지의 가능한 값을 가진

다고 가정하자. 그러면 VCdim(H)은 3,125 이다. 또 우리는 $\varepsilon = 0.1$ 과 $\delta = 0.01$ 을 필요로 한다고 가정하자. 그러면 정리 1에 의하여,

$m = \max\{ 41.25,\ 2443.75 \} = 2,444$ 개의 사례가 어떠한 경우에도 필요하게 된다. 그리고

$m = 21,707$ 개의 사례가 있으면 위의 정확도수준을 보장하기에 충분하게됨을 알 수 있다.

$\varepsilon = 0.5$ 이고 $\delta = 0.01$ 인 경우에는

$m = 4,342$ 개의 사례가 있으면 위의 정확성을 얻는데 충분한 수준이 된다.

실수값을 가지는 속성이 있는 경우에, VCdim(H) 은 무한대가 되기 때문에 정리 1은 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 1을 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 Tsai and Koehler[forthcoming]에 상세 절차가 나타나 있다. 이 방법은 베타 사전분포를 가정하고 일관성있는 의사결정나무(consistent decision tree)의 오류에 대한 상한선을 찾아낸다.

2. 단순화를 하는 경우의 분석

이 단원에서는 먼저 Kim and Koehler[1991,1992] 의 결과를 요약한다. 첫째 단순화의 과정을 거쳐서 추론된 의사결정나무가 진짜 개념(true concept)과 상치할 확율이 $1-\delta$ 보다 큰 가능성을 가지고 ε 보다 작게되기를 보장하는 충분한 크기의 훈련사례수를 제공하고자 한다. 또한 단순화의 과정을 거쳐서 추론된 의사결정나무의 정확성을 추정하는 3 가지 사후적 추정방법을 요약하고자 한다.

논문의 초점을 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 있다. 먼저 이진의사결정나무와 그 등급(rank)을 정의한다.

정의 1: $V_n = \{v_1, \ldots, v_n\}$ 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_n 상의 이진의사결정나무이다. (이하에서 우리는 이경우를 간편하게 "Q=0" 또는 "Q=1" 이라고 부르기로 한다).

(ii) Q(0)를 Q의 왼쪽나무(left subtree)라고 하고, Q(1)을 Q의 오른쪽 나무라고 하자. 만약 Q의 뿌리노드 v가 V_n에 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_0 와 r_1 이 각각 Q의 왼쪽나무와 오른쪽나무의 등급이라고 할때

$$r(Q) = \begin{cases} \max(r_0, r_1), & r_0 \neq r_1 \text{ 인 경우} \\ r_0+1 \ (=r_1+1), & r_0 = r_1 \text{ 인 경우} \end{cases}$$

Ehrenfeucht and Haussler[1988]의 의사결정나무 추론 알고리즘, Findmin(S) 와 Kim and Koehler[1991]의 의사결정나무 단순화 알고리즘, Prune(r, k, Q, S), 에 의해서 다음결과가 도출된다. 여기서 r, k, Q, S 는 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: (단순화 과정을 포함하는 의사결정나무 추론에 충분한 사례수:Kim and Koehler[1991]) 어떠한 $n \geq r > 0$, 어떤 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) $m \geq [2/\{\varepsilon^2(1-2\mu_{n,r})^2\}]\{(e*n/r)^r \ln(8n)+\ln(2/\delta)\}$, $0<\varepsilon, \delta<1$ 인 경우

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

인 경우 $1-\delta$ 보다 큰 확율로서 Findmin(S)와 Prune(r,k,Q,S)는 등급이 r 이하이며 오류가 ε 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 는 하나의 무작위 사례를 추출하는데 한 단위시간이 소요됨을 고려할때 n 개의 속성을 가지는 등급(rank)이 r 이하인 의사결정나무를 r 이 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.

현실적인 추론 상황에서는 이론적으로 도출된 어떤 정확성 수준을 보장하는데 충분한 수의 사례들을 얻기 어려운 경우가 많으므로, 그러한 경우에 대비하여 단순화된 의사결정나무의 예측정확성에 대한 사후적 평가 방

법을 고안할 필요가 있다.

여기서 Kim and Koehler[1992]의 3가지 사후적 오류 추정방법을 요약한다. 아래에서 C_k^m 은 m 개에서 k 개의 순서를 고려하지 않은 샘플을 얻는 방법의 수이다.

간단히 언급하면, 방법(1)은 단순화된 의사결정나무에 의한 오류비율의 분포가 해당구간에서 균등분포(uniform distribution)를 갖는다고 가정하는것이 무난할때 사용되는 방법이다. 그러나 이 방법은 오류비율이 어떤 특정구간에서 다른 구간보다 높은 확율을 가지는 경우에는 부적당하다.

방법(2)는 훈련사례(training example)를 분류하는데 오류가 있게되는 어떠한 추론된 의사결정나무에도 사용할 수 있다. 이 방법은 정수 파라메타를 가지는 베타 사전분포(Beta Prior)를 가정하지만 어느 특정한 베타분포에 가정을 두지는 않는다.

방법(3)은 오류비율의 사전분포(prior distribution)에 무관하게 적용될 수 있다.

방법(1): <u>사전정보가 전혀없을때 적용</u>

m개의 시험사례를 단순화된 의사결정나무를 이용하여 분류할때 b 개의 오류가 발생하였다면, 균등사전분포(uniform prior)를 사용한 2항 파라메타 θ 의 사후적 추정치는 다음과 같다.

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

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

방법(2): <u>훈련사례로부터의 정보를 이용</u>

베타사전분포를 가정하고 훈련의 정보와 일치하는(즉, 모순이 없는) 파라메타의 집합을 얻는다. 여기서 특정한 파라메타는 필요로 하지 않는다. 단순화의 결과로 m개의 훈련사례에서 b개의 분류오류가 발생하였다고

가정하자. 여기서 S(b,m)는 훈련정보와 일치하는 파라메타의 집합이다.

양의 정수 p, q 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 ε 은 $-2(1-\varepsilon)\ln(1-\varepsilon) = \varepsilon - b/m$의 근이다.

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

m_2 개의 사례로 시험을 하여 b_2 개의 사례를 잘못 분류하였다면,

$\mathrm{Prob}\{\ \theta \ge \varepsilon\ \} \le \delta$ 이며 여기서 δ 의 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.

방법(3):<u>일반적인 오류 추정치</u>

단순화된 의사결정나무를 사용하여 m개의 독립사례중에서 b개의 분류오류가 발생하였다고 가정하자.

어떠한 $b/m \le \varepsilon$ 에서도 다음 부등식이 성립한다.

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

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

또 어떤 $b/m \le \varepsilon$ 과 $0 \le \varepsilon' \le b/m$에 대해서도 다음 부등식이 성립한다.

c) $\mathrm{Prob}\{\ \varepsilon' \le \theta \le \varepsilon\ \} \ge 1 - \exp(-2(\varepsilon - b/m)^2 m) - \exp(-2(b/m - \varepsilon')^2 m)$.

다음 단원에서는 Messier and Hansen[1988]의 채무불이행(loan default)자료를 사용하여 위의 방법들을 설명한다.

IV. 실제 문제 적용 사례

이 단원에서는 Messier and Hansen[1988]의 채무불이행(loan default) 문제를 사용하여 단원 III 에서 요약된 결과를 실제문제에 적용한다. 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 Messier and Hansen[1988]으로 부터 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

우리가 채택한 Ehrenfeucht and Haussler[1988]의 의사결정나무구축 알고리즘 Findmin(S)를 이용하면 그림 1 (II장) 의 의사결정나무가 최소 등급(minimal rank)를 가진 나무로서 추론되어진다. (데이터는 Messier and Hansen[1988]이 사용한 32 개의 훈련사례를 사용하였다). 이 의사결정나무의 등급은 2 이고, 이 나무는 훈련사례(S)를 완벽하게 분류(오류없이 추정)한다.

이제 우리는 등급이 1 인 보다 단순화된 의사결정나무를 원한다고 가정하여 보자. 그러면 r=1, k=2 가 된다. Kim and Koehler[1991]의 알고리즘 Prune(r,k,Q,S)을 사용하여 의사결정나무를 단순화시키면 그림 3 (II장) 의 등급이 1 인 의사결정나무가 얻어진다.

1. 단순화과정을 포함하는 의사결정나무 추론에 요구되는 충분한 양의 사례수

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

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

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

정리 2 의 b) 를 사용하고 $\mu_{n,r}$ 에 대한 보다 엄격한 경계(tight bound)를 사용하면 (상세한 내용은 Kim and Koehler[1991] 참고) m 이 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 $m = 4,168$ 이다.

즉, 위의 채무불이행 문제에서 단순화의 과정을 거쳐 의사결정나무의 형태로 추론된 규칙 h 가 4,168 개 이상의 무작위의 독립된 훈련사례를 사용하여 추론되었다면 h 의 추정 오류가 10 % 이하일 확율이 99 % 이상이 된다는 것이다.

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.

Messier and Hansen[1988]의 채무불이행(loan default) 자료를 사용하여 이 경우를 설명한다. 훈련사례의 수는 $m = 32$, 그림 3 의 등급 1 인 의사결정나무에 의해 2 개의 분류오류가 발견된다. 시험사례의 수는 16 이고 그중 2 개의 사례가 잘못 분류된다.

방법(1): 시험사례의 수가 16 이고 2 개의 분류오류가 발생하였으므로, 균등분포(uniform prior)의 가정하에서, 방법(1)은 다음과 같이 개념의 오류를 추정한다.

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

여기서 ε 을 변화시키면서 여러가지 다른 수준의 오류 추정치를 얻을 수 있다.

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

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

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

방법(2): 조건을 만족하는 $q^* = 22$ 이고 따라서 ($p^* = 1,\ q^* = 22$)가 얻어진다.

$\varepsilon = 0.1$ 인 경우에

$\mathrm{Prob}\{\ \theta \geq 0.1\ \} \leq \delta$, 이고

$1 - I_{\varepsilon}(p^*+b_2,\ q^*+m_2-b_2) = 1 - I_{0.1}\ (1+2,\ 22+16-2) = \underline{\delta} \leq \delta$ 이 된다.

왼쪽항을 계산하면 다음과 같이 된다.

$\mathrm{Prob}\{\ \theta \geq 0.1\ \} \leq \delta$, 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 확율이 δ 보다 작은데 δ 의 하한 경계는 0.253670 이다.

여기서 여러 다른 수준의 ε 값에 대해 다음 결과를 얻게된다.

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

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

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

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

작은 ε 값에 대해서는 p 값을 크게 하여 일치되는 베타 사전분포

(consistent Beta prior)를 찾아냄으로써 하한경계(lower bound)를 개선할 수 있다. 예를 들면, p 가 2 일때 일치되는 q 값중에 가장 작은 값은 30 이다. $\varepsilon = 0.1$ 인 경우에

$\underline{\delta} = 1 - I_{\varepsilon}(p^{*}+b_2,\ q^{*}+m_2-b_2) = 1 - I_{0.1}\ (2+2,\ 30+16-2) = 0.295587.$

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

방법(3): 방법 3 의 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 $\text{Prob}\{\ \theta \geq 0.3\ \} \leq 0.37531.$

방법 3 의 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.

이번에는 단순화된 의사결정나무가 단순화과정을 거치지 않은 원래의 의사결정나무보다 시험사례에서 더 나은 추정력(better classification power)을 가지는 일반적인 경우를 예로들어 보기로 한다. 훈련사례가 40 개이고 단순화된 의사결정나무에 의해서 8 개의 사례가 잘못 분류된다고 하자. 또한 20 개의 시험사례를 취하여 단순화된 의사결정나무를 시험한 결과 2 개의 분류오류가 발견되었다고 하자. 이 경우 방법(1)은 다음과 같이 오류를 추정한다.

방법 (1):

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

여기서 ε 을 변화시키면서 여러가지 다른 수준의 오류 추정치를 얻을

수 있다.

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

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

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

방법(2): 조건을 만족하는 $q^* = 18$ 이고 따라서 ($p^* = 1$, $q^* = 18$)이 얻어진다.

$\varepsilon = 0.2$ 인 경우에

Prob{ $\theta \ge 0.2$ } $\le \delta$, 이고

$1 - I_\varepsilon(p^*+b_2,\ q^*+m_2-b_2) = 1 - I_{0.2}\ (1+2,\ 18+16-2) = \underline{\delta} \le \delta$ 이 된다.

왼쪽항을 계산하면 다음과 같이 된다.

Prob{ $\theta \ge 0.20$ } $\le \delta$, 이고

$1 - I_{0.2}\ (3, 32) = 0.0113 = \underline{\delta} \le \delta$ 이 된다.

즉, 단순화과정을 거친 의사결정나무의 추정상의 오류가 20 % 보다 크거나 같을 확율이 δ 보다 작은데 δ 의 하한 경계는 0.0113 이다.

여기서 여러 다른 수준의 ε 값에 대해 다음 결과를 얻게된다.

Prob{ $\theta \ge 0.25$ } $\le \delta$, $1 - I_{0.25}\ (3,\ 32) = 0.001641 = \underline{\delta} \le \delta$.

Prob{ $\theta \ge 0.30$ } $\le \delta$, $1 - I_{0.30}\ (3,\ 32) = 0.000190 = \underline{\delta} \le \delta$.

Prob{ $\theta \ge 0.35$ } $\le \delta$, $1 - I_{0.35}\ (3,\ 32) = 0.000018 = \underline{\delta} \le \delta$.

$\varepsilon = 0.2$ 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 ε 값에 대하여는 오류를 정확하게 추정하는 것이 가능하여 진다. 그러나 큰 ε 값에 대하여는 이 방법이 엄격한 하한 경계값을 제공한다. 예를 들어 ε 이 0.75 보다 크면, (p, q) = (2, 20) 이 경계값을 개선하지 못한다.

방법(3): 방법 3 의 a)는 다음과 같이 오류를 분석한다.

$\varepsilon = 0.2$ 인 경우에

Prob{ $\theta \geq 0.2$ } $\leq$ exp($-2\ (0.2 - 2/20)^2\ 20$) = 0.6703.

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

방법 3 의 b)에 의하면, $\varepsilon = 0.2$ 인 경우에

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 Prob{ $\theta \geq 0.3$ } < 0.1935, ε = 0.4 인 경우에는 Prob{ $\theta \geq 0.4$ } < 0.0220 이 된다.

최저신뢰도 수준(the worst possible confidence factor) δ 에 대하여 방법 (2)는 하한경계를 제공하고 방법(3)은 일반적인 상한경계(upper bound)를 제공한다. 우리는 이 결과를 결합하여 δ 의 범위를 얻을 수 있다.

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

방법(2)에 의해 δ 의 하한 경계를 계산한다.

$b/m = 6\ /\ 20 = 0.3$ 이므로

$\varepsilon = 0.3$ 에 대하여 ($p^* = 1,\ q^* = 7$) 이다. 따라서

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.

방법(3)의 b)에 의해 δ 의 상한 경계를 계산한다.

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

= 0.6685

따라서 δ 의 범위가 얻어진다.

$0.0157 \leq \delta \leq 0.3678.$

즉, 단순화의 과정을 거쳐 추론된 개념의 오류가 30 % 보다 크게될 확

률이 최악의 경우에 0.3678 까지 될 수 있으나 결코 0.0157 이하로는 되지 않음을 보여준다.

V. 단순화기법의 이론적인 사용조건

단순화기법을 사용하면 항상 추론된 개념은 단순화된다. 그러나 어떠한 상황에서 단순화기법이 예측정확성을 높이게 되는지에 대해서는 잘 알려져 있지 않다. Kim and Koehler[forthcoming]는 더높은 예측정확성을 얻기 위해 단순화기법의 사용이 필요한 상황에 대한 연구를 하였다. 이들은 사례공간(instance space)의 균등분포를 가정하고 다음의 3 가지 방향으로 이 문제를 분석하였다.

1) 비대칭성(skewness)의 증가. (비대칭성은 사례중에서 분류군(class)의 규모가 큰 사례집합의 전체 사례중에서의 비율이다)

2) 사례수(sample size)의 증가.

3) 노이즈(noise)의 증가. (노이즈는 분류나 속성에서의 오류이다.)

이들의 연구는 위의 세 방향 모두에 대해서 값이 증가함에 따라 단순화기법이 예측정확성을 높이는데 상대적으로 유용한 기법이 되는것으로 밝히고 있다. 특히, 비대칭성이 커지거나 훈련사례(training set)의 수가 많아지면 노이즈가 전혀없다고 가정하더라도 단순화된 의사결정나무는 단순화되지 않은 의사결정나무의 예측정확성과 거의 같은 수준의 예측정확성을 가지게 된다. 속성에 노이즈가 있을 경우에는 단순화 기법의 유용성이 더욱 증가된다. 일반적으로 노이즈가 많아지면 단순화과정을 거친 의사결정나무의 예측정확성이 단순화되지 않은 의사결정나무보다 더 높게된다.

이상적으로는 단순화기법을 사용하여 추론된 의사결정나무가 무시할 수 있을 정도의 아주 작은 예측정확성을 잃거나 또는 더 높은 예측정확성을 가지는것이 바람직하다. Kim and Koehler[forthcoming]는 단순화과정을 거친 의사결정나무의 예측정확성이 어느정도로 다르게 되는지 그 정도를 분석하였다. 이러한 품질의 차이는 각 사례집단의 실제 발생확율분포에

의존하게 되는데, 균등분포를 가정하면 단순화기법을 사용한 의사결정나무가 사례수 10 개 이상인 경우에 잃게되는 예측정확성이 4 % 이하가 됨을 보여준다. 이러한 작은 양의 예측정확성 상실은 보다 단순하고 알기쉬운 개념이나 규칙을 얻기위하여 때에 따라서는 감수 할 수 있는 수준이다.

VI. 요약 및 결론

이 논문에서 우리는 의사결정나무 추론과 개념을 단순화하고 개념의 예측정확성을 높여주는 단순화기법에 대하여 최근의 결과를 요약하고 예를 들어 그 이용방법과 의미를 설명하였다. 이 논문의 주요 의도는 전문가시스템 개발자들에게 위의 여러가지 이론적인 결과들을 이용할 수 있도록 가능하게 해주는데 있다. 여기서는 세가지 종류의 결과를 제시하였다.

첫째, 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.

둘째, 위에서 도출된 충분한 양의 사례를 얻기가 어려울 경우에 사후적인 방법으로 오류의 수준을 측정하는 방법을 제공하였다. 세가지 방법을 제공하였는데 각각의 방법은 단계적으로 가정(assumptions)을 줄여나가면서 오류를 추정하고 있다. 여기서도 단순화기법을 사용한 의사결정나무와 사용하지 않은 의사결정나무에 대해 각각 언급하였다.

마지막으로, 의사결정나무의 예측정확성을 높이기 위하여 단순화기법이 필요하게되는 상황에 대하여 논의하였다. 이 결과는 단순화기법이 예측정확성(predictive accuracy)을 잃게되는 경우에 그 정도가 어떠한지를 추정하는데에도 이용할 수 있다.

앞으로의 연구방향은 단순화기법을 사용하여 의사결정나무를 추론하는 과정에서 각각의 오류수준을 보장하는데 요구되는 충분한 크기의 사례수(sample size)를 줄여가는 것이 중요한 연구방향의 하나이다. 또한 단순

화된 의사결정나무의 사후적 오류추정방법을 개선하여 보다 정확한 오류추정치를 얻는것이 또다른 중요한 연구방향이다. 그리고 보다 복잡한 상황에서 단순화기법이 유용하게 되는 환경을 찾아내는 것도 바람직한 연구방향이다.

參考文獻

Angluin, D. and Laird, P., "Learning From Noisy Examples", Machine Learning, 2, 343-370, 1988.

Boose, J. H. and Gaines, B.R., "Knowledge Acquisition for Knowledge-Based Systems: Notes on the State-of-the Art", Machine Learning, 4, 377-394, 1989.

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

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, 71-79, 1987.

Cheng, J., Fayyad, U.M., Irani, K.B. and Qian, Z., "Improved

Decision Trees: A Generalized Version of ID3", Proceedings of the 5th International Conference on Machine Learning, 100-106, Morgan Kaufmann, San Mateo, CA, 1988.

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

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

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

Johnson, D.E., "What kind of Expert should a System be?", Journal of Medicine and Philosophy, 1983, 8, 77-97.

Kim, H. and Koehler, G. J., "An Investigation on the Conditions of Pruning an Induced Decision Tree", The European Journal of Operational Research, forthcoming.

Kim, H. and Koehler, G. J., "PAC Learning a Decision Tree with Pruning", Department of Decision and Information Sciences, University of Florida, 1991.

Kim, H. and Koehler, G. J., "The Accuracy of Decision Trees with Pruning", Department of Decision and Information Sciences, University of Florida, 1992.

Laird, P., Learning from good data and bad. Doctoral Dissertation,Department of Computer Science, Yale University, New Haven, CT, 1987.

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

Messier, W.F. and Hansen, J.V., "Inducing rules for Expert Systems Development", Management Science, 34, 12, 1403-1415, 1988.

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, 125-161, 1980.

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

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

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

Musen, M. A., "Automated Support for Building and Extending Expert Models", Machine Learning, 4, 347-375, 1989.

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

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

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, 1, 86-106, 1986.

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

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

Schaffer, C., "When does Overfitting decrease Prediction Accuracy in Induced Decision Trees and Rule Sets?", Lecture Notes in Artificial Intelligence: Machine Learning-EWSL-91, Porto, Portugal, 1991.

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

Shaw, M. J. and Gentry, J.A., "Inductive Learning for Risk Classification", IEEE Expert, 5, 1, 47-53, 1990.

Simon, H., "Why Should Machines Learn?", In R.S. Michalski, J. Carbonell, T. Mitchell(Eds.), Machine Learning: An Artificial Intelligence Approach. Vol. I, Tioga, Palo Alto, CA, 1983.

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

Tsai, L. and Koehler, G.J., "The Accuracy of Concepts Learned from Induction", Decision Support System, forthcoming.

Utgoff, P., "Incremental Induction of Decision Trees", Machine Learning, 4, 161-186, 1989.

Valiant, L.G., "A Theory of the Learnable", Communications of the ACM., 27, 11, 1134-1142, 1984.

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

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

저자소개

저자 김현수는 서울대 공대를 졸업하고, 한국과학기술원 경영과학과에서 석사학위를 받았다. 미국 플로리다대학교에서 경영정보시스템(MIS)을 전공하여 경영학 박사학위를 취득하였으며 동 대학에서 객원조교수로 재직하였다. (주)데이콤의 시스템본부와 정보통신연구소에서 주임연구원으로 재직하였으며 현재 한국정보문화센터에서 정보통신 정책연구를 담당하고 있다. 주요 관심분야는 시스템 분석 및 설계, 전문가 시스템, 데이터 통신, 정보화 추진전략에 대한 연구등이다.