韓國産業經濟學會
産業經濟硏究 第13卷 第5號(2000. 9), pp 1~17

租稅專門家시스템의 知識獲得을 위한 規則推論模型의 比較

The Comparison of Rule Reference Models for Knowledge Acquisiton In Tax-Expert System

金 鍾 元*

〈 目 次 〉

Ⅰ. 序 論

불확실성하에서 지식획득에 관한 연구는 많은 관심을 받고 있는 영역이다. 전통적으로 불확실성은 확률이론(특히 베이지안이론)과 효용이론과 같은 표준도구에 의해 다루어졌다. 최근 많은 연구자들은 대부분의 경우에 이러한 도구가 심각한 한계점을 가지고 있다는 사실을 인식하게 되었다. 본 연구에서는 전통적인 확률이론에 기초한 의사결정과정의 규칙추론모형이 가지고 있는 한계점을 극복할 수 있는 러프집합이론을 소개하며, 이 개념을 불확실성과 복잡성 및 모호성이 내재된 납세신고의사결정분야에서 적용하여 의사결정 규칙을 추론한다. 또한 이 결과를 실제로 전문가시스템구

* 서남대학교 경영학부 조교수

축 분야에서 성공적으로 응용되고 있는 기계학습알고리즘(machine learning)-ID3의 결과와 비교한다.

다음 2장에서는 전통적인 규칙추론의 한계점 및 이를 극복하기 위한 기계학습알고리즘과 러프집합이론의 개념 및 규칙추론방법을 소개한다. 3장에서는 본 연구의 설문결과를 분석하며, 납세의 사결정과정에 러프집합이론을 적용하여 납세신고의사결정과정의 지식을 추론하며, 이 결과를 기계학습알고리즘과 비교한다. 4장에서는 추론된 규칙들을 비교하고, 이 결과의 의미와 시사점을 제시한다.

Ⅱ. 專門家시스템을 위한 規則推論過程

1. 전통적인 확률이론에 의한 규칙추론의 한계점

납세신고분야는 전통적으로 정확한 측정치를 고집하는 경향이 있는데, 이러한 정확한 자료는 실제로 얻기가 곤란하다. 따라서 정확한 측정치라는 전제하에 현실적으로 적용가능한 모형을 도출할 때 이용가능한 정보를 통합하는 과정에서 심각한 문제가 발생하며, 통합된 정보가 의미를 잃을 수도 있다. 정확한 자료를 이용할 수 없기 때문에 발생하는 또 다른 문제는 '과잉자료(data enrichment)'문제이며, 이로 인해 모형을 도출하는 과정에서 신뢰성이 낮은 자료가 모형에 도입될 수 있다. Grzymala-Busse(1988)의 연구에 의하면 모형에 모호성개념을 도입했을 때 모형의 예측가치가 높아지는 것으로 나타났다.

확률모형에 의해 추론된 규칙들은 모호성을 기술하는데 적합하지 않으며 확률모형을 이용하여 모호성문제를 다룰 수는 없다. Ellsberg(1961)는 중간개념(middle term)이 포함되는 경우 확률의 전체 합계는 1이 되지 않는다는 사실을 입증하였으며, 몇몇 실증연구에서 Ellsberg의 이와 같은 주장이 지지되는 것으로 나타났다[Einhorn과 Hogarth, 1986; Slovic와 Tversky, 1974]. 효용이론도 기업관련 의사결정문제에 널리 응용되는 도구이다. 그러나 어떤 객체를 다차원속성에 의해 평가하는 경우, 의사결정자의 선호도에 대한 정보가 필요하다. 의사결정자에 의해 선호되는 정보가 이용가능했을 때, 전체선호도 관점에서 통합된 다차원점수를 부여하게 되는데, 이 때 정렬과 관련된 문제가 발생하게 된다. 선호되는 정보라는 개념에는 불일치성개념이 포함되고 선호속성을 기술하는 규칙은 확률개념이기 때문에 확률분포와 관련된 한계점에 직면하게 된다[Slowinski, 1993].

공정성, 효율성, 이해가능성 등의 각종 조세관련 제개념을 동시에 충족시키면서 다양한 이해관계자집단과 조세상황에서, 적용되는 세법에 내제된 수많은 수준과 차원 때문에 나타나는 복잡성과 애매모호성을 무작위성과 같은 개념에 의해 단순하게 처리하거나 무시하여서는 안된다. 따라서 납세신고 의사결정과정을 서술하거나 실제 납세신고 의사결정을 지원하기 위한 수단들은 이러한 세법

의 특성을 고려할 수 있는 모형 또는 접근방법에 기초해야 하며, 속성변수(조세모수)에 대한 측정치, 속성변수(조세모수)간의 관계, 확률값 등의 부정확성을 조작할 수 있는 개념 또는 이론을 응용하게 되면 보다 효율적인 납세신고행위에 대한 분석과 납세신고행위를 지원하기 위한 시스템의 설계가 가능할 것이다.

전통적인 방법하에서 의사결정과정에 관한 규칙을 추론하기 위한 지식베이스는 사례중심의 학습에 의해 획득된다. 학습자들은 각 속성치에 대해 자신이 판단한 중요도에 따라 가치를 부여하고 이러한 속성값과 의사결정과의 상관관계를 찾으려 한다. 전문가들의 의사결정은 전문가 전체적인 차원에서는 일관성을 유지하나 개별전문가의 차원에서는 의사결정속성변수와 의사결정속성치의 집합이 전문가들이 행하는 복잡한 의사결정을 예측하는데 충분한 정보를 담고 있지 못하기 때문에 비일관적인 판단을 하게 된다.

2. 기계학습에 의한 규칙추론 - ID3 알고리즘

기계학습에 대한 연구는 인공지능분야에서 핵심적인 영역을 차지하고 있다. 지식지향 전문가시스템에서 구현되는 알고리즘은 지식이 암묵적(implicit)인 형태가 아닌 명시적(explicit)인 형태로 제시되는 경우 강력한 성과를 얻게 된다(Quinlan, 1986). 성능이 뛰어난 전문가시스템은 특정분야에 대한 전문가와 지식엔지니어와의 지속적인 상호교류를 통하여 조직화되고 구축된다. 이러한 방법에 의해 지식을 획득하고 표현하면 지식엔지니어 한 사람이 생성하는 규칙의 수는 제한을 받게 된다. 그러나 복잡한 과업을 지원하기 위한 전문가시스템에서 제시해야 하는 규칙의 수는 수백 또는 수천개에 해당되기 때문에, 이러한 면담에 의해 지식을 획득하는 방법은 전문가시스템의 요구사항을 충족시켜주지 못한다. 이러한 문제점을 해소하기 위한 노력으로 지식을 조직화하기 위한 수단으로써 기계학습에 대한 연구가 활발하게 진행되었다.

Carbonell, Michalski 그리고 Mitchell(1983) 등은 (1) 적용된 학습전략, (2) 시스템에 의해 획득된 지식의 표현, (3) 시스템의 응용분야 등 3가지 차원에 의해 기계학습에 관한 연구를 구분하고 인식하였다. 이 분류기준에 의하면 Quinlan(1983)이 개발한 ID3 알고리즘은 첫째로 분류의사결정에 응용가능한 일반목적시스템(general-purpose system)이고, 둘째로 획득된 지식은 의사결정나무(decision tree)로 표현되기 때문에 복잡하지 않은 단순한 시스템의 구현이 가능하며, 셋째로 사례를 통하여 학습이 이루어지며 제시된 사례의 순서를 중시하는 자료중심접근방법이 아닌 사례에 포함된 정보의 빈도에 의해서 의사결정나무가 생성하기 때문에 특정사례집단에 의해서 학습이 향상되지 않는 특성을 갖는다. 또한 ID3는 Hunt의 CLS(concept learning system)의 개념적토대(Hunt, Marin 그리고 Stone, 1966)와 정보이론(information theory)을 도입한 알고리즘으로써 비용중심보다는 정보중심의 평가함수를 채택하고 있다. 정보이론적인 접근방법을 도입한

ID3에서는 제공되는 정보이득이 최대가 되는, 다시 말해서 속성변수에 의해 앤트로피[2]의 감소가 최대가 되는 속성변수를 찾는 과정을 통하여 의사결정나무의 각 수준이 결정된다.

3. 러프집합이론에 의한 규칙추론

러프집합 개념은 모호성(vagueness)과 불확실성(uncertainty : ambiguity)에 대한 새로운 수학적 접근방법이다. 러프집합의 사고는 어떤 정보(자료, 증거, 지식)와 전체집합을 구성하는 객체간에는 어떤 관계가 있다는 가정에 기초한다. 러프집합이론에서는 모호성개념을 도입하여 'IF..., THEN ...' 형태의 규칙을 추론한다. 러프집합이론에 의해 최종적으로 추론되는 규칙은 확실성규칙과 가능성규칙의 두 가지 규칙집합군으로 대별된다.

1) 러프집합이론의 개념

러프집합이론에서는 정보표 또는 정보시스템속성치표라는 테이블이 이용되며(이하 정보표라함), 이 정보표에서 종축은 속성변수(attribute), 횡축은 관심대상의 객체(object) 그리고 표의 몸체는 속성값(attribute value)을 나타낸다.

동일한 정보에 의해 특징지워지는 객체는 이들에 대한 이용가능한 정보를 통해서 서로 구분(또는 분리)할 수 없다(indiscernible). 즉 서로 유사한 것 또는 동일한 것으로 간주된다. 이와 같은 분리불능관계(indiscernibility relation)가 러프집합이론의 수학적 기초가 된다.

주어진 정보에 의해 분리가 곤란한 객체의 특정 집합을 원소집합(elementary set)이라고 부르며, 이 원소집합이 전체에 대한 지식의 기본요인(granule, atom)이다. 원소집합의 합집합에 의해 표현되는 집합을 크리습(crisp, precise)집합이라 하며, 그렇지 않은 집합을 러프(rough, imprecise, vague)집합이라 한다.

각 러프집합은 경계(boundary-line)가 있는 객체 -다시 말해서 특정 집합의 원소 또는 여집합 원소에 의해 확실하게 구분될 수 없는 객체를 갖게 된다. 크리습집합은 경계를 갖는 원소가 있을 수 없다. 경계가 있는 객체는 이용가능한 지식에 의해 적절하게 분류할 수 없음을 뜻한다.

이와 같이 객체는 오로지 이들에 대한 이용가능한 정보에 의해 투시된다는 가정을 통해서 '지식은 과립형태의 구조를 갖는다(Knowledge has granular structure)[3]'라는 사고를 이끌어 낼 수

---

2) Jackson(1988)에 의하면 정보이론이란 정보를 수량화, 기호화 그리고 전달하는 것과 관련된 이론이며, 어떤 사상이 전달하는 정보의 크기를 합리적으로 정의하기 위해서는 그 사상의 확률이 도입되어야 하며, 또한 확률개념에 의한 정보는 서로 합산과 승산이 가능하기 때문에 정보의 크기는 확률에 대해서 로그를 취하여 산출할 수 있다고 하였다.

3) '지식은 과립형태의 구조를 갖는다(Knowledge has granular structure)'는 의미는 사물(또는 물체)을 식별함에 있어 분자 또는 원자단위의 정보(자료, 증거, 지식)가 제공되는 경우 사물은 식별이 곤란하지만, 분자 또는 원자보다 더 광의의 개념인 과립단위의 정보가 이용가능한 경우, 특정한 사물을 정확하게 식별

있다. 지식의 '과립성(granularity)'때문에, 관심대상이 되는 어떤 객체는 구분할 수 없으며 유사한 것 또는 동일한 것으로 간주될 수 있다. 정확한 개념(precise concept)과는 달리 모호한 개념은 이들 객체에 대한 이용가능한 정보에 의해 특성화될 수 없다. 따라서 러프집합이론에서 모호한 개념은 한 쌍의 정확한 개념 -모호한 개념의 하한근사치(lower approximation)와 상한근사치(upper approximation)라고 부른다- 에 의해 대체된다. 하한근사치는 이 개념에 확실하게 포함이 되는 확실한(sure) 객체로 구성되며, 상한근사치는 이 개념에 포함이 가능한(possible) 객체에 의해 구성된다. 상한근사치와 하한근사치 사이의 차이에 의해 표현되는 모호한 개념에는 범위(boundary region)가 있게 된다. 이 하한근사치와 상한근사치는 러프집합이론의 기본이 되는 연산자이며, 이들을 통하여 확실성규칙과 가능성규칙이 추론된다.

2) 러프집합이론에 의한 규칙추론

러프집합이론에서는 리덕션개념을 이용하여 의사결정표로부터 불필요한 속성변수와 속성값을 제거함으로써 최종적으로 간략화된 규칙을 얻을 수 있다.

자료를 분석할 때 속성변수간의 종속성(dependency)은 매우 주의 있게 다루어져야 한다. 만일 어떤 속성변수집합 D가 다른 속성변수집합 Q에 의해서 완전(totally) 종속적이라면(기호로는 Q → D), 속성변수집합 D의 모든 속성값들은 C의 속성값에 의해 유일하게 결정될 것이다. 다시 말해서 D와 Q값 사이에 함수적인 관계가 성립되면 완전 종속이라고 할 수 있다. 그러나 변수간의 일반적인 관계는 완전종속성이 아닌 부분종속성(partial dependency)일 것이다. 조건속성변수인 불확실성은 의사결정속성변수인 몇몇 속성값에 대해서만 유일하게 결정된다. 부분종속성이란 D의 어떤 값만이 Q값에 의해 결정된다는 것을 말한다. 속성변수집합 Q와 속성변수집합 D의 종속의 정도는 일치성측정치라는 개념이 이용되며, 일치성측정치 $\gamma(Q, D)$는 다음과 같다.

$$\gamma(Q, D) = \frac{\mathrm{card}(POS_c(D))}{\mathrm{card}(U)}.$$

여기서 card(U)는 전체집합의 원소 숫자이며, $POS_c(D)$는 속성값집합의 하한근사치의 합집합을 나타낸다.

리덕션개념을 정의하면, $P' \subset P \subset Q$ 인 조건하에서, 만일 $\gamma(P',D) < \gamma(P, D) = \gamma(Q, D)$이면 , 이 때의 P를 Q의 리덕트라고 부른다.

간략화된 규칙을 얻기 위한 다음 단계는 조건속성값들에 대한 리덕션이다. 조건속성값들에 대한 리덕션의 정의는 조건속성변수에 대한 리덕션개념이 그대로 준용된다. 이와 같은 절차를 통하여 얻

할 수도 있고 그렇치 못할 수도 있다는 것을 뜻하며, 현재 주어진 정보에 의해서는 사물의 식별이 명확하지 못하지만 지식이 추가되면 그 사물은 정확하게 식별가능하다는 것을 의미한다.

어진 간략화된 의사결정알고리즘은 확실성 알고리즘과 가능성 알고리즘으로 나누어진다.

3) 규칙추론을 위한 리덕션과정

러프집합이론에서 간략화된 규칙을 얻기 위한 리덕션과정은 다음과 같다.

단계 1 : 먼저 각 규칙에 대하여 상대 리덕션개념을 적용하여 수정리덕션을 구한다. 이를 위해 리덕션후보와 다른 규칙과의 일관성여부를 확인한다. 일관성이 존재하는 경우, 속성변수를 하나씩 제거하면서 제거된 속성변수를 뺀 나머지 속성변수집합을 리덕션후보로 두고 리덕션여부를 확인한다. 이와 같은 방법으로 점점 작은 숫자의 속성변수로 구성된 집합이 일관성을 유지하는지를 확인한다.

단계 2 : 단계 1에서 구한 리덕션에 의해 규칙을 만든다. 여기서 규칙은 리덕션에 속하는 조건속성들이 해당 속성값을 가질 때 판단속성이 어떤 값을 갖는지를 나타낸다.

단계 3 : 다음으로 각 행에서 구해진 규칙들 중에서 각 규칙마다 하나의 규칙을 선택하되 선택된 총규칙의 수가 최소가 되게 한다.

단계 4 : 마지막으로 판단속성의 속성값이 같은 규칙에 대해서는 OR를 이용하여 규칙을 하나로 묶는다.

4. 기계학습알고리즘과 러프집합이론의 비교

실무적으로 러프집합이론과 ID3 알고리즘에 의해 최종적으로 도출된 규칙은 'IF..., THEN' 형태로 지식지향 시스템에 포함된다. 러프집합이론과 ID3의 공통점은 첫째, 사례(example 또는 instance)를 통해 규칙을 추론하는 귀납적 추론이라는 점이다. 연역적 추론과는 달리 이들 알고리즘에 의해 추론되는 규칙들은 현재의 경험(또는 사상)에 대한 추론이기 때문에 추가적인 또는 미래의 사건에 의해 변화될 수 있다. 둘째, 알고리즘의 전개는 자료지향적(data-driven) 접근방법이 아닌 정보지향적(information driven) 접근방법에 의하기 때문에 제시되는 속성변수의 순서에 가중치를 부여하지 않는다. 셋째, 이들은 분류(classification)의사결정과 관련된 일반목적시스템에 응용될 수 있으며, 마지막으로 이들 알고리즘에서는 질적변수의 조작이 가능하다.

러프집합이론과 ID3의 차이점은 첫째, 규칙을 추론하는 과정에서 ID3는 조건속성변수의 상대적 중요성을 고려하여 진입변수의 결정하는 순차적처리개념이 적용되나, 러프집합이론에서는 조건속성변수의 개별적이 차원이 아닌 집합차원에서 속성변수들을 고려하는 병렬처리개념이 적용되며, 규칙을 추론하는 과정에서 진입변수가 아닌 탈락변수가 먼저 결정된다. 둘째, ID3에서는 의사결정속성변수들이 주어진 정보에 의해 진실 또는 거짓으로 구분되는 성격을 가지나 러프집합이론에서 적용되는 의사결정속성변수들은 주어진 정보에 의해 객관적으로 정확하게 집단구분이 곤란한 경우도

있다. 세째, ID3에서는 속성변수의 측정상의 오류와 규칙의 오분류에 따른 노이즈(noise)를 고려하나, 러프집합이론에서는 규칙의 오분류에 따른 노이즈는 고려하지 않는다. 즉, 조건속성변수의 값이 동일함에도 불구하고 의사결정결과가 다른 경우, 이 결과를 그대로 규칙에 이용한다. 넷째, ID3에서는 근노드(또는 진입변수)를 결정하는 기준으로 확률이론에 의한 엔트로피개념이 이용되나, 러프집합이론에서는 조건속성변수 중 탈락변수를 결정함에 있어서 집합이론에 의한 종속성개념이 적용된다. 다섯째, ID3에서 최종적으로 추론된 규칙은 의사결정나무형태로 제시가 가능하나, 러프집합이론에서 최종적으로 추론된 규칙은 각 속성변수에 대한 상대적 중요성이 고려되지 않기 때문에 의사결정나무형태가 아닌 의사결정표에 의해 제시된다. 따라서 ID3에 의한 결과는 의사결정나무형태로 제시할 수 있으나, 러프집합이론에 의한 결과는 의사결정나무형태가 아닌 다른 형태로 제시되어야 한다.

Ⅲ. 規則推論模型의 比較 - 納稅申告意思決定過程

본 연구에서는 가상의 조세상황을 반영한 사례를 이용하여, 즉 소득원천, 세율체계, 벌금 등의 3가지 독립변수와 직접적으로 관계가 있는 소득원천의 불확실성, 세무조사에 대한 인식, 공정성에 대한 인식, 제재위협에 대한 인식 등에 대한 실험 및 설문응답결과를 이용하여 피실험자의 납세신고의사결정과정에 관한 규칙을 추론한다.

1. 설문지의 구성

납세신고의사결정자들의 지식획득을 위한 설문지는 총 5문항으로 구성되었으며, 1번 문항은 과세표준을 신고하는 과정에서 가장 중요하게 고려한 요인은 무엇인가?에 대한 질문이며 소득원천, 벌금, 세율구조, 세무조사확률 요인 중 하나에 대하여 응답을 하면 된다. 2번 문항부터 5번 문항까지는 각 조세변수에 대한 피실험자들의 주관적인 평가와 관련된 질문으로 구성되었는데, 2번 문항은 소득원천의 불확실성, 3번 문항은 인식 세무조사율, 4번 문항은 세율의 공정성, 5번 문항은 제재에 대한 위협 등에 관한 질문이다. 2번부터 5번까지의 응답은 0.0부터 1.0까지의 11점 척도로 구성되는데, 각각의 문항에서 0.0은 (불확실성 : 매우 낮음), (인식 세무조사율 : 매우 낮음), (공정성 : 불공정함), (제재에 대한 위협 : 매우 낮음)을 나타내며, 0.5의 수치는 (불확실성 : 보통), (인식 세무조사율 : 보통), (공정성 : 보통), (제재에 대한 위협 : 보통)을 나타낸다. 그리고 1.0의 수치는 (불확실성 : 매우 높음), (인식 세무조사율 : 매우 높음), (공정성 : 공정함), (제재에 대한 위협 : 매우 높음)을 나타낸다.

2. 실험 및 설문결과

먼저 과세표준을 신고하는 과정에서 중요하게 고려한 요인에 대한 응답을 분석한 결과(〈표 1〉 참조), 전체 피실험자들이 인식하는 중요요인은 소득원천(30.7)〉세무조사확률(26.3)〉벌금(24.6)〉세율구조(18.4) 순으로 나타났으며, 소득원천에 따른 집단간 신고금액의 차이는 통계적으로 유의적이었다(p〉0.035). 그러나 세금체계와 벌금에 따른 차이는 유의적이지 않았다(각각 p〉0.455, (p〉0.550).

표 1. 과세표준신고시 중요하게 고려한 요인

(단위 : 명)

		소득원천	벌금	세율구조	세무조사확 률	합계	통계량
소득원천	근로〉자영	33(43.4)	14(18.4)	11(14.5)	18(23.7)	76 (100)	$x^2=13.53$ (p〉0.035)
	근로=자영	13(17.1)	21(27.6)	17(22.4)	25(32.9)		
	근로〈자영	24(31.6)	21(27.6)	14(18.4)	17(22.4)		
세금체계	누진세	38(33.3)	31(27.2)	18(15.8)	27(23.7)	76 (100)	$x^2=2.614$ (p〉0.455)
	비례세	32(28.1)	25(21.9)	24(21.0)	33(29.0)		
벌금	벌금 高	32(28.1)	27(23.7)	25(21.9)	30(26.3)	76 (100)	$x^2=2.110$ (p〉0.550)
	벌금 底	38(33.3)	29(25.4)	17(14.9)	30(26.3)		
전체		70(30.7)	56(24.6)	42(18.4)	60(26.3)	288 (100)	

( )안은 %임.

조세모수에 대한 설문응답을 분석한 결과는 〈표 2〉와 같다. 이러한 결과를 통하여, 첫째 총수입 중에서 자영소득이 차지하는 비율이 높을수록 소득원천의 불확실정도가 높다고 응답하였으며, 통계적으로 유의적이었다(p〉0.0001). 둘째, 총수입 중에서 자영소득이 차지하는 비율이 높을수록 세무조사율이 높다고 피실험자에게 알렸으나 실제로 과세표준을 신고하는 과정에서 피실험자들은 자영소득이 차지하는 비율이 높을수록 세무조사율이 높을 것이라고 인식하지 않고 있는 것으로 나타났다(p〉0.5832). 피실험자의 세무조사에 대한 인식에 미치는 영향요인은 추가적으로 분석이 요구된다. 셋째, 전체 피실험자들의 세율공정성에 대한 응답을 세율체계가 누진세로 구성된 사례집단과 비례세로 구성된 사례집단에 대하여 평균한 결과 각각 0.5000과 0.5500으로 나타났으며 이 결과는

통계적으로 유의적이다(p〉 0.0376). 그러나 세율체계에 따른 과세표준신고에 미치는 영향은 없었다. 넷째, 제재에 대한 인식은 벌금이 높을수록 위협이 더 크다고 응답하고 있으며, 통계적으로 매우 유의적이었다(p〉0.001).

표 2. 조세모수의 조작과 관련된 통계자료[4)]

	평균응답정도			F값	Pr 〉 F
소득원천의 불확실정도	근로〉자영	근로＝자영	근로〈자영	14.44	0.0001
	0.4407	0.5697	0.6158		
세무조사에 대한 인식	근로〉자영	근로＝자영	근로〈자영	0.54	0.5832
	0.5197	0.5461	0.5211		
공정성에 대한 인식	누진세	비례세		4.37	0.0376
	0.5000	0.5500			
제재에 대한 인식	벌금 : 고	벌금 : 저		54.78	0.0001
	0.5780	0.3964			

3. 러프집합이론에 의한 규칙추론

실험에 참가한 전체 피실험자의 평균응답자료와 평균신고금액을 이용하여 의사결정표를 작성한 결과는 〈표 3〉과 같으며, 이 표는 전체 피실험자의 과세표준신고정보표 또는 과세표준의사결정표가 되며 이를 전체의사결정표라 하기로 한다. 의사결정표에서 q_1은 소득원천에 대한 불확실성, q_2는 인식 세무조사율, q_3은 세금체계에 대한 공정성, 그리고 q_4는 제재에 대한 위협이며, d는 과세표준의 사결정이다. 또한 개별사례에 대해 붙여진 C1,..., C12는 의사결정표에서 그 자체가 규칙의 이름이 된다.

4) 표에서 제시한 수치는 해당 설문항목에 대하여 사례에 따라 평균한 값이다. 예를 들어 표에서 근로〉자영 부분은 불확실성과 인식세무조사율에 대한 피실험자가 응답한 결과를 사례 1, 사례 2, 사례 3, 사례 4에 대하여 종합한 다음 평균한 값이 제시되었으며, 비례세 부분은 공정성과 관련된 문항에 대한 응답결과를 사례 3, 사례 4, 사례 7, 사례 8, 사례 11, 사례 12에 대하여 종합한 다음 평균한 값이다.

표 3. 전체 피실험자집단의 과세표준의사결정표

	q_1	q_2	q_3	q_4	d
C1	1	1	1	2	1
C2	0	2	1	0	0
C3	0	1	2	2	1
C4	1	1	1	0	0
C5	2	1	1	2	0
C6	1	1	1	0	0
C7	2	1	1	2	1
C8	2	2	2	0	0
C9	2	1	2	2	2
C10	2	2	1	0	2
C11	2	1	2	2	2
C12	2	1	2	0	1

위 〈표 3〉에서 전체집합 U는 {C1, C2, C3, C4, C5, C6, C7, C8, C9, C10, C11, C12}이며, 전체 조건속성변수집합 Q는 $\{q_1, q_2, q_3, q_4\}$이고, 전체 의사결정속성변수 D는 {d}이다. 러프집합이론을 이용하여 의사결정규칙을 추론하기 위해서 먼저 12개의 사례를 3가지 의사결정집합으로 분할하면,

$E_1(d_0)$ = {C2, C4, C5, C6, C8}

$E_2(d_1)$ = {C1, C3, C7, C12}

$E_3(d_2)$ = {C9, C10, C11}이다.

전체조건변수집합 $Q=\{q_1, q_2, q_3, q_4\}$의 요소집합[5]을 구하면,

Q에 대해서 : $E_1(Q)$ = {C1}, $E_2(Q)$ = {C2}, $E_3(Q)$ = {C3},

$E_4(Q)$ = {C4, C6}, $E_5(Q)$ = {C5, C7}, $E_6(Q)$ = {C8},

$E_7(Q)$ = {C9, C11}, $E_8(Q)$ = {C10}, $E_9(Q)$ = {C12}이다.

따라서 d_0, d_1, d_2의 하한근사치(A_*)와 상한근사치(A^*)를 구하면

$A_*(d_0)$ = {$E_2(Q)$, $E_4(Q)$, $E_8(Q)$} = {C2, C4, C6, C8},

$A_*(d_1)$ = {$E_1(Q)$, $E_3(Q)$, $E_{12}(Q)$} = {C1, C3, C12},

5) 여기서 $E^*(Q)$는 전체집합 Q = {q1,q2,q3,q4}의 각 원소에 대응되는 각 조세모수의 속성값의 묶음을 나타낸다. 따라서 E1(Q) = (1,1,1,2), E2(Q) = (0,2,1,0), E3(Q) = (0,1,2,2), E4(Q) = (1,1,1,0), E5(Q) = (2,1,1,2), E6(Q) = (2,2,2,0), E7(Q) = (2,1,2,2), E8(Q) = (2,2,1,0), E9(Q) = (2,1,2,0)이다.

$A_*(d_2) = \{E_7(Q), E_8(Q)\} = \{C9, C10, C11\}$.

$A^*(d_0) = \{E_2(Q), E_4(Q), E_5(Q), E_6(Q)\} = \{C2, C4, C5, C6, C7, C8\}$.

$A^*(d_1) = \{E_1(Q), E_3(Q), E_5(Q), E_{12}(Q)\} = \{C1, C3, C5, C7, C12\}$.

$A^*(d_2) = \{E_7(Q), E_8(Q)\} = \{C9, C10, C11\}$.

따라서 개별의사결정속성값 d_0, d_1, d_2의 경계영역은 다음과 같다.

$bdry(d_0) = A^*(d_0) - A_*(d_0) = \{C5, C7\}$

$bdry(d_1) = A^*(d_1) - A_*(d_1) = \{C5, C7\}$

$bdry(d_2) = A^*(d_2) - A_*(d_2) = \phi$

그러나 의사결정속성값 d_0, d_1, d_2는 3가지이므로 특정 의사결정속성집합간의 경계영역은 다음과 같이 설정되어야 한다.

$bdry(d_0, d_1) = bdry(d_0) \wedge bdry(d_1) \wedge - bdry(d_2)$.

$bdry(d_0, d_2) = bdry(d_0) \wedge bdry(d_2) \wedge - bdry(d_1)$.

$bdry(d_1, d_2) = bdry(d_1) \wedge bdry(d_2) \wedge - bdry(d_0)$.

$bdry(d_0, d_1, d_2) = bdry(d_0) \wedge bdry(d_1) \wedge bdry(d_2)$.

여기서 $\wedge$는 교집합을 -는 보집합을 의미한다. $bdry(d_0, d_1)$에 대한 해석은 집합 $bdry(d_0)$와 집합 $bdry(d_1)$에는 포함되지만 집합 $bdry(d_2)$에는 포함되지 않는다는 것을 의미한다. 이와 유사하게 $bdry(d_0, d_1, d_2)$는 집합 d_0 또는 d_1 또는 d_2를 의미하지만, 이들 3집합 중 어느 곳에 속한다고 말할 수 없다. 이상의 경계영역정의에 의해 본 연구에서 얻어진 자료를 가지고 특정의사결정속성집합간의 경계영역을 구하면,

$bdry(d_0, d_1) = \{C5, C7\}$, $bdry(d_0, d_2) = \phi$

$bdry(d_1, d_2) = \phi$ $bdry(d_0, d_1, d_3) = \phi$이다.

〈조건속성변수와 조건속성치에 대한 리덕션〉

조건속성변수에 대한 리덕션을 행하기 위해서 먼저 전체 조건속성변수집합의 하위속성변수집합을 정의해야 한다. 하위속성변수집합 Q'를 4개의 조건속성변수 중 3개의 조건속성변수로 구성된 하위속성변수집합은 $Q_1' = \{q_1, q_2, q_3\}$, $Q_2' = \{q_1, q_2, q_4\}$, $Q_3' = \{q_1, q_3, q_4\}$, $Q_4' = \{q_2, q_3, q_4\}$로 정의할 수 있다. 조건속성변수에 대한 리덕션을 행하기 위해서 먼저, 전체속성변수집합과 각 하위속성변수집합과의 종속 정도를 나타내는 일치성측정치를 구해야 된다. 러프집합이론의 정의에 의해 일치성측정치 $\gamma(Q, D) = card(POS_c(D)) / card(U)$이므로, 이를 구하기 위해서는 각 속성변수집합의 상한근사치를 구해야 한다. 여기서 $card(U)$는 전체집합의 원소 숫자이며, $POS_c(D)$는 속성값집합의 하한근사치의 합집합을 나타낸다. 각 속성변수집합의 요소집합과 하한근사치는 다음 〈표 4〉와 같다.

표 4. 하위속성변수집합의 요소집합과 하한근사치

요소집합		Q_1' $\{q_1, q_2, q_3\}$	Q_2' $\{q_1, q_2, q_4\}$	Q_3' $\{q_1, q_3, q_4\}$	Q_4' $\{q_2, q_3, q_4\}$
요소집합	E_1	{C1,C4,C6}	{C1}	{C1}	{C1,C5,C7,C9}
	E_2	{C2}	{C2}	{C2}	{C2,C10}
	E_3	{C3}	{C3}	{C3}	{C3,C11}
	E_4	{C5}	{C4,C6}	{C4,C6}	{C4}
	E_5	{C7}	{C5,C7,C9,C11}	{C5,C7,C9}	{C6}
	E_6	{C8}	{C8,C10}	{C8,C10,C12}	{C8}
	E_7	{C9}	{C12}	{C11}	{C12}
	E_8	{C10}			
	E_9	{C11,C12}			
하한근사치	$A_*(d_0)$	{C2,C5,C8}	{C2,C4,C6}	{C2,C4,C6}	{C4,C6,C8}
	$A_*(d_1)$	{C3,C7}	{C1,C3,C12}	{C1,C3}	{C12}
	$A_*(d_2)$	{C9,C10}		{C11}	

따라서 각 하위조건속성변수집합의 일치성정도를 구하면

$\gamma(Q_1', D) = 7 / 12$,

$\gamma(Q_2', D) = 6 / 12 = 1 / 2$,

$\gamma(Q_3', D) = 6 / 12 = 1 / 2$,

$\gamma(Q_4', D) = 4 / 12 = 1 / 3$이다.

전체조건속성변수집합의 일치성정도는 $\gamma(Q, D) = 10 / 12 = 5 / 6$이므로, $Q' \subset Q$이면서 $\gamma(Q', D) = \gamma(Q, D)$를 만족하는 하위조건속성변수집합이 존재하지 않는다. 따라서 이 전체의사결정표의 경우 조건속성변수에 대한 리덕션을 할 수 없으며, 추론된 의사결정규칙을 간략화할 수 없다.

다음으로 조건속성변수의 속성치에 대한 리덕션을 행하기 위하여 각 개별 조건속성변수의 요소집합을 구하면 다음과 같다.

q_1에 대하여 : $E_1(q_1) = \{C2, C3\}$, $E_2(q_1) = \{C1, C4, C6\}$,

$E_3(q_1) = \{C5, C7, C8, C9, C10, C11, C12\}$.

q_2에 대하여 : $E_1(q_2) = \{C1, C3, C4, C5, C6, C7, C9, C11, C12\}$,

$E_2(q_2) = \{C2, C8, C10\}$.

q_3에 대하여 : $E_1(q_3)$ = {C1, C2, C4, C5, C6, C7, C9, C10},
$E_2(q_3)$ = {C3, C8, C11, C12}.
q_4에 대하여 : $E_1(q_4)$ = {C2, C4, C6, C8, C10, C12},
$E_2(q_4)$ = {C1, C3, C5, C7, C9, C11}.

또한 앞에서 도출한 바와 같이, 의사결정속성집합의 요소집합은 $E_1(d)$ = {C2, C4, C5, C6, C8}, $E_2(d)$ = {C1, C3, C7, C12}, $E_3(d)$ = {C9, C10, C11}이다. 의사결정표에서 조건속성변수의 속성값을 리덕션하기 위해서는 특정의 요소집합에 대해서 $E_\cdot(q_\cdot) \subset E_\cdot(d_\cdot)$의 조건이 만족되어야 한다. 본 연구의 실험결과 각 조건속성변수의 요소집합이 이 조건을 만족하는 경우가 없기 때문에 조건속성변수의 속성값에 대한 리덕션을 할 수 없다.

이상의 결과를 토대로 추론된 의사결정규칙은 3개의 확실성규칙과 1개의 가능성규칙으로 대별된다. 본 연구에 참가한 피실험자의 자료를 토대로 과세표준신고 의사결정에 관해 추론된 의사결정표를 'IF..., THEN' 형태로 제시하면 다음과 같다

*** 확실성규칙 ***

〈규칙 1〉 IF {(소득원천의 불확실성 : 낮음)
and (세무조사율 : 높음)
and (공정성 : 보통)
and (제재에 대한 위협 : 낮음)}
or {(소득원천의 불확실성 and 세무조사율 and 공정성 : 보통)
and (제재에 대한 위협 : 낮음)}
or {(소득원천의 불확실성 and 세무조사율 and 공정성 : 높음)
and (제재에 대한 위협 : 낮음)},
THEN (과세표준신고 : 고위험신고).

〈규칙 2〉 IF {(소득원천의 불확실성 : 낮음)
and (세무조사율 : 보통)
and (공정성 and 제재에 대한 위협 : 높음)}
or {(소득원천의 불확실성 and 세무조사율 and 공정성 : 보통)
and (제재에 대한 위협 : 높음)},
THEN (과세표준신고 : 저위험신고).

〈규칙 3〉 IF {(소득원천의 불확실성 : 높음)
and (세무조사율 : 보통)
and (공정성 : 높음)
and (제재에 대한 위협 : 높음)}
or {(소득원천의 불확실성 and 세무조사율 : 높음)
and (공정성 : 보통) and (제재에 대한 위협 : 낮음)},
THEN (과세표준신고 : 무위험신고).

*** 가능성규칙 ***

〈규칙〉 IF(소득원천의 불확실성 : 높음)
and (인식된 세무조사율 and 공정성 : 보통)
and (제재에 대한 위협, 높음),
THEN (과세표준신고 : 고위험신고 또는 저위험신고).

4. 기계학습알고리즘 - ID3과의 비교

현재 실무에서 적용되고 있는 대부분의 지식지향 시스템에서 지식획득을 위한 접근방법은 하향식(top-down)이다. 이 방식에서는 먼저 근노드에 해당하는 중요변수를 결정하고, 그 다음 수준에 해당하는 변수를 결정하는 순차적 처리방식을 채택하고 있다. 현재 실무에서 성공적으로 응용되고 있는 귀납적 개념에 의한 기계학습 알고리즘의 하나인 ID3도 마찬가지로 순차적 처리방식을 채택한 하향식 알고리즘이다. 반면 러프집합이론은 귀납적 개념을 채택하고 있지만, 의사결정에 도입되는 변수들에 대한 상대적 우선순위를 고려하지 않는 병행적 처리방식이 도입된다. 따라서 조건속성변수간에 교호작용이 존재하고 단일속성변수 차원이 아닌 속성변수 집합차원에서 문제를 해결해야 하는 의사결정분야에서 지식을 추론하는 경우 순차적인 처리방식보다는 병행적인 처리방식을 채택하는 알고리즘의 의한 결과가 훨씬 효과적일 것이다.

본 연구의 실험에서 납세신고 의사결정과업과 관련하여 개발된 실험사례에서는 과세소득과 관련된 불확실성과 납세의사결정자들의 주관적인 판단이 포함되었으며, 조세모수들 사이의 독립성보다는 종속성에 초점을 맞추고 있으며 특정 조세모수에 대한 우선순위가 고려되지 않는다. 이러한 상황에서 납세신고 의사결정자들의 의사결정행태를 분석하기 위해서는 러프집합이론에 의한 알고리즘을 도입하는 것이 타당할 것이다. ID3와 러프집합이론에 의한 알고리즘은 각기 장·단점이 있기 때

문에 어떤 방법이 더 우수하다고 말할 수 없다.

이 기계학습알고리즘에 의해 최종적으로 결정되는 의사결정나무와 러프집합이론에 의한 추론결과 간의 가장 큰 차이는 기계학습알고리즘의 결과는 속성변수들의 우선순위에 따라 순차적으로 표현이 가능하나, 러프집합이론에 의한 결과에서는 속성변수들이 동시적으로 고려되고 있다는 점이다. 두 알고리즘에 의한 차이를 살펴보면, 첫째 러프집합에 의한 경우 인식세무조사율이 탈락되었으며, 규칙에 포함된 소득원천의 불확실성, 공정성, 제재에 대한 위협 등은 서로 우선순위가 존재하지 않으나, ID3에서는 근노드에 해당하는 가장 우선순위가 높은 조건속성변수는 소득원천의 불확실성이며 그 다음으로 제재에 대한 위협과 공정성이 동일한 2수준이며 인식세무조사율은 3수준으로 구성되고 있다. 둘째, 러프집합에 의한 결과를 보면 (공정성 : 낮거나 보통)이면 다른 조건속성변수의 값에 관계없이 (과세표준신고 : 저위험신고)가 된다. 그러나 ID3 알고리즘에 의한 결과에 의하면 조건속성변수인 공정성은 소득원천의 불확실성의 하위수준으로 구성되며, (소득원천의 불확실성 : 높음)인 경우에 고려되는 조건속성변수가 된다. 셋째, 러프집합이론에서 (제재에 대한 위협 : 보통)이면 규칙에서 고려가 되지 않으나, ID3에서는 이 경우 (소득원천의 불확실성 : 높음)을 제외하고 (과세표준신고 : 저위험신고)의 규칙이 추론된다.

Ⅳ. 結 論

본 연구에서 불확실성과 애매성이 수반되는 반구조적인 납세신고 의사결정상황에서 의사결정자들의 의사결정규칙을 추론하는데 적합한 이론으로 러프집합이론을 소개하고 있다. 러프집합이론은 사례를 통한 지식구축이 가능하며, 언어적 자료의 조작이 가능하고, 독립변수들을 동시적으로 고려할 수 있는 이론이다. 본 연구는 부정확한 언어적 표현과 불확실성과 애매성이 존재하는 의사결정에 적합한 이론으로써 러프집합이론을 소개하는 차원으로 해석될 수 있다. 실제 연구영역과 업무영역에서 이 이론이 광범위하게 적용되기 위해서는 의사결정자들의 의사결정과정에 대한 행태와 주관성이 포함된 언어적 자료에 대한 추가적인 연구가 행해져야 할 것이다.

러프집합이론을 적용하여 과세표준신고 의사결정과정에 관한 규칙을 추론하였으며,최종적으로 3개의 확실성규칙과 1개의 가능성규칙이 규칙이 추론되었다.

추론된 납세신고 의사결정규칙의 특성은 첫째, 인식세무조사율에 대한 변수는 다른 조세모수(소득원천의 불확실성, 공정성, 제재에 대한 위협)들에 의해 동시적으로 설명될 수 있기 때문에 규칙추론과정에서 제기되었다. 납세신고와 관련된 선행연구에서도 인식세무조사율은 매우 중요한 설명변수이지만 이 변수는 다른 조세모수에 의해 조작되거나 직접적인 영향을 받는 요인으로 해석되고 있다. 둘째, 소득원천의 불확실성과 제재에 대한 위협변수는 각각 소득원천구성비와 벌금율과 직접적으로 대응된 것으로 나타나 피실험자들의 반응과 실험조작이 차이가 없었다. 셋째, 개별변수에 대한

분석과정에서 유의적이지 않은 것으로 판단된 공정성변수의 경우, 공정성에 대한 인식이 보통이 아닌 경우(즉, 낮거나 높은 경우) 규칙추론과정에서 다른 조세모수의 특성에 관계없이 저위험보고를 권고하는 규칙이 추론되었다. 넷째, 공정성이 보통이며 소득원천의 불확실성이 높고 제재에 대한 위협이 낮은 경우 신고과세표준은 일관성(고위험신고 또는 저위험신고)을 유지하지 못하고 있는 것으로 나타났다.

본 연구에서 도입한 러프집합이론은 속성변수들을 동시적으로 고려하는 병렬처리방식을 채택하고 있다. 현재 실무에서 성공적으로 이용되고 있는 기계학습알고리즘-ID3은 러프집합이론과 마찬가지로 귀납적개념을 채택하고 있으며, 규칙을 추론하는 과정은 속성변수에 대한 우선순위를 결정하는 순차적처리방식을 채택하고 있다. 이들 두 알고리즘에 의해 추론된 규칙을 비교한 결과 상당한 차이가 있는 것으로 나타났다. ID3 알고리즘의 결과는 의사결정나무형태로 제시되며 러프집합이론에 의한 결과는 의사결정표형태로 제시되기 때문에 두 알고리즘간의 우수성에 대해서 본 연구에서는 직접적으로 평가하지 못하고 있다.

또한 본 연구는 전문지식이 요구되는 조세상황을 사례를 통하여 조작하였는데, 사례자체가 세밀하고 구체적이기보다는 일반적인 상황을 가정하고 있기 때문에 러프집합이론을 통한 실제 의사결정자들의 행위를 추론하기 위해서는 구체적이고 실제적인 사례를 통한 연구가 행해져야 할 것이다. 실험과 관련하여 피실험자들의 설문응답과 관련하여 성실하고 체계적인 응답을 유도하기 위한 엄격한 통제절차가 요구된다.

참 고 문 헌

Alm, J., "Uncertain Tax Policies, Individual Behavior, and Welfare," *American Economic Review*, 1988, pp. 323-338.

Berg, J. E., L. A. Daley, J. W. Dickhaut and J. R. O' Brien, "Controlling Preferences for Lotteries on Units of Experimental Exchange," *Quarterly Journal of Economics*, 1986, pp. 281-306.

Carnes, G., A. Korvin, and J. M. Hagan, "Dealing with Ambiguity in the Tax Law : An Application of Rough Set Theory to the Determination of Debt Worthlessness," *Applications of Fuzzy Sets and The Theory of Evidence to Accounting*, (London, England : JAI Press INC), 1995, pp. 105-119.

Einhorn, H. J. and R. M. Hogarth, "Decision Making under Ambiguity," *Journal of Business*, 1986, pp. 225-250.

_____ and _____ , "Order Effects in Belief Updating : The

Belief-Adjustment Model," working paper, University of Chicago, 1990.

Ellsberg, D., "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, 1961, pp. 643-669.

Grzymala-Busse, J. W., "Knowledge Acquisition under Uncertainty-A Rough Set Approach," *Journal of Intelligent & Robotic Systems*, 1988, pp. 3-16.

Hagan, J. M., A. Korvin, and P. H. Seigel, "Ambiguity and Vagueness in Determining Reasonable Compensation for Closely Held Corporations : The Use of Rough Set Theory and Fuzzy Set Theory to Develop Decision Rules," *Applications of Fuzzy Sets and The Theory of Evidence to Accounting*, (London, England : JAI Press INC), 1995, pp. 121-134.

Hideo Tanaka, "Rough Sets and Knowledge Acquisition", *KFIS*(한국 퍼지 및 지능시스템학회), 1997, pp. 3-17.

Jackson, A. H., "Machine Learning," *Expert Systems*, 1988, pp.132-150.

Pawlak, Z., "Rough Set Fundamentals," *KFIS*(한국 퍼지 및 지능시스템학회), 1996, pp. 1-32.

Quinlan, J. R., "Induction of Decision Trees," *Machine Learning*, 1986, pp. 81-106

Sansing, R. C.,"Information Acquisition in a Tax Compliance Game," *The Accounting Review*, 1993, pp. 874-884.

Scotchmer, S. and J. Slemrod, "Randomness in Tax Enforcement," *Journal of Public Economics*, 1989, pp. 17-32.

Schepanski, A., "Tests of Theories of Information Processing Behavior in Credit Judgment," *The Accounting Review*, 1983, pp. 581-599.

Slovic, P., and Tversky, A. " Who Accept Savage's Axiom?" *Behavioral Science*, 1974, pp. 368-372.

Slowinski and Sharif, E. S., " Rough Sets Approach to Analysis of Data of Diagnostic Peritoneal Lavage Applied for Multiple Injuries Patients," *Rough Sets, Fuzzy Seta and Knowledge Discovery. Proceedings of the International Workshop on Rough Sets and Knowledge Discovery*, 1993, pp. 420-425.

Steinbart, R. J.,"The Construction of a Rule-Based Expert System as a Method for Studying Materiality Judgments," *The Accounting Review*, 1987, pp. 97-116.

Zebda, A., "Fuzzy Set Theory and Accounting," *Journal of Accounting Literature*," 1989, pp. 76-105.