감리지적기업의 분류적 특성에 관한 연구: 베이지안 망과 C5.0, 그리고 앙상블 방법간의 비교를 중심으로

이건창
성균관대학교 경영학부 교수
(kunchanglee@naver.com)
최 관
성균관대학교 경영학부 교수
(kwanchoi@skku.ac.kr)

본 연구는 감사보고서 감리지적기업과 감리비지적기업을 효율적으로 구분할 수 있는 분류방법에 대한 연구이다. 본 연구에서는 선행연구에서 널리 사용되어온 분류방법인 로짓회귀분석 방법이 종속변수와 설명변수간에 획일적인 선형함수만을 가정하는 데에서 나오는 문제점을 두 가지 측면에서 극복하고자 한다. 첫째는 감리지적 여부에 영향을 미치는 설명변수간에 존재하는 인과관계(causal relationship)를 도출할 필요가 있다. 이는 어떤 변수가 다른 어떤 변수와 직접 또는 간접적 인과관계를 통하여 감리지적 여부에 영향을 주는지를 의사결정자에게 알려줌으로써 보다 효과적인 감리작업을 할 수 있도록 지원할 수 있다. 이를 위하여 본 연구에서는 일반 베이지안 망(GBN: General Bayesian Network)을 제안하고 GBN에서 유도되는 마코프 블랭킷(Markov Blanket)을 제시한다. 둘째는 감리지적 예측을 보다 정확하게 하기 위하여 기존에 사용되던 분류방법인 GBN과 나이브 베이지안 망(NBN: Naive Bayesian Network) 및 C5.0을 결합한 앙상블 방법을 제시한다. 1990년에서 1999년까지의 감리지적 및 감리비지적기업의 자료를 기초로 실험한 결과 본 연구에서 제안하는 두 가지 방법이 모두 통계적으로 유의한 결과를 제공한다는 것이 실증적으로 검증되었다.

주제어: 감리, 베이지안 망, 마코프 블랭킷, C5.0, 앙상블 방법, 분류방법

Ⅰ. 서 론

본 논문은 감사보고서 감리지적기업과 감리비지적기업간의 분류방법에 관하여 연구하였다. 감사보고서 감리란 공인회계사가 수행한 외부감사에 대한 감리로서, 기업의 회계분식과 감사인의 부실감사에 대한 중요한 감시 및 감독제도의 하나이다. 감리지적기업이란 감리에서 회계부정이 있다고 밝혀진 기업을 의미하고, 감리비지적기업이란 감리대상은 되었으나 회계부정기업으로 지적되지 않은 기업을 의미한다. 우리나라는 「주식회사의 외부감사에 관한 법률」 제15조에 의거하여 증권선물위원회로 하여금 감리업무를 수행하도록 규정하고 있고 금융감독원에서 감리실무를 맡고 있다.[1)]

그런데 현행 감리제도에서는 감리대상기업을 선정하는 감리선정방법에 임의성이 많고 체계적인 선정기준이 미흡하다는 지적이 있다. 이와 같은 문제점을 보완하기 위하여 여러 선행연구가 수행되었으나 두 가지 한계점이 있다.

첫째, 지금까지의 연구는 어떤 기업특성변수들이 감리지적 여부와 밀접한 관계를 가지고 있는지에 대하여 분석하였으나(윤중옥과 김명환, 2001; 최관과 최국현, 2003), 이들 변수간에 존재하는 인

논문접수일: 2006. 11 게재확정일: 2007. 5

1) 동 법률 제15조 ⑤에 따라서 증권선물위원회는 비상장기업과 신규상장대상기업의 감리를 한국공인회계사회에 위탁하고 있고, 금융감독원은 상장기업에 대한 감리를 실시하고 있다.

과관계를 도출하여 의사결정자로 하여금 보다 풍부한 정보를 제공하는 연구는 없었다.

둘째, 감리지적 여부에 대한 예측율을 획기적으로 향상시킬 수 있는 연구가 없었다. 있다고 하더라도 기업특성변수와 감리지적 여부와 기업특성변수간의 단변량분석에 그치거나, 통계적으로 유의한 변수만을 가지고 판별함수인 로짓 회귀모형으로 분석하는 것에 그치고 있다(최 관과 최국현, 2003). 그러나 문제는 회계부정기업은 여러 가지 기업특성변수들이 서로 복잡한 영향을 줄 수 있으므로 개별 기업특성변수로는 감리지적을 설명하기에 한계가 있다. 따라서 기업특성변수들 중에서 어떤 변수들이 서로 영향을 미쳐 감리지적의 결과를 나타내는지에 대한 연구가 부족한 실정이다.

본 연구에서는 이와 같은 기존연구의 문제점을 극복하기 위하여 두 가지 연구목적을 제안한다. 첫째는 감리지적기업과 감리비지적기업을 통계적으로 의미 있게 구분시켜주는 설명변수간의 숨어있는 인과관계를 분명하게 보여주고자 한다. 이러한 목적을 위하여 본 연구에서는 일반 베이지안 망(GBN: General Bayesian Network)을 적용한다. 두 번째 연구목적은 감리지적기업을 정확하게 예측할 수 있는 예측방법을 찾기 위하여 선행연구에서 전혀 다루지 않았던 앙상블 방법을 제시한다. 앙상블 방법은 베이지안 망 방법과(일반 베이지안 망인 GBN과 나이브 베이지안 망(Naive Bayesian Network)인 NBN) 여타 인공지능방법인 C5.0 방법을 결합한(앙상블) 방법으로서 감리지적 여부를 기존방법과 비교하여 상대적으로 정확하게 예측할 수 있다.

첫 번째 연구목적에서 주로 다루는 베이지안 망(Bayesian Network)은 불확실성이 많은 영역 데이터로부터 분류에 영향을 미치는 속성들 간의 상호의존성을 잘 표현하고 이것을 바탕으로 비교적 클래스를 정확하게 예측할 수 있는 견고한 확률적 도구로 알려져 있다(Neapolitan, 1990). 이와 같이 베이지안 망에 대한 연구가 본격화 된 것은 베이지안 망 중에서도 단순한 형태의 나이브 베이지안 망(NBN)이 분류문제에서 상당히 높은 정확도를 보여주면서부터이다(Langley et al., 1992). 그러나 NBN은 클라스 노드(또는 종속변수, 결과변수)를 여타 다른 노드와는 다른 별도의 특별한 변수로 간주하는데 반하여, 일반 베이지안 망인 GBN은 클라스 노드도 다른 노드와 마찬가지인 평범한 노드 중의 하나로 본다. Friedman et al. (1997)은 이와 같은 GBN을 보다 효과적으로 학습하고 이를 다양한 문제에 적용한 바 있다.

두 번째 연구목적에서 다루는 앙상블 방법은 여러 분류방법을 결합하는 방법이다. 즉, 결합 이라는 의미를 갖는 앙상블(Ensemble)은 말 그대로 여러개의 분류방법을 결합할 때가 단일방법으로 분류를 할때보다 더 성과가 좋다는 가정을 가지고 출발한 방법이다. 특히 Ji and Ma(1997)는 앙상블 방법이 성능이 약한 분류방법을 상호보완적으로 결합하면 상당히 성능이 좋은 분류방법을 산출할 수 있음을 이론과 실증적으로 보여 주었고, Dzeroski and Zenko(2002)와 Chebrolu et al.(2004) 등에 의해서도 앙상블 방법이 기존의 단일 분류방법보다 실증적으로 더 우수하다는 연구가 지속적으로 이뤄지고 있다. 본 연구는 이와 같은 앙상블 방법에 관한 연구를 토대로 감리지적기업과 감리비지적기업을 구분시켜 주는 효과적인 분류방법을 제시한다. 본 연구에서 고려하는 분류방법으로는 C5.0, GBN, NBN 모두 3개의 분류방법이다. 이들 분류방법은 각각 개별적으로도 우수한 분류방법이지만 앙상블방법을 사용하면 감리지적 여부의 예측능력

을 크게 향상시킬 수 있다. 특히, 본 연구에서는 감리지적기업을 1, 감리비지적기업을 0으로 표시하고 이를 클라스 노드, 즉 결과변수로 처리하고 있는데, 0을 0으로 (즉, 0 → 0) 예측하는 것보다는 (다시 말하면, 감리비지적기업을 감리비지적기업으로 예측), 1을 1로 예측하는 것이 훨씬 더 중요하다(즉, 감리지적기업을 감리지적기업으로 예측). 이와 같은 맥락하에 본 연구에서는 특히 앙상블 방법이 1 → 1의 경우 예측능력이 얼마나 향상되는지를 실증분석 하고자 한다.

본 연구는 기업의 재무적 특성을 이용하여 감리지적기업을 분류할 수 있는 효율적인 방법을 모색하므로 실무적으로도 큰 공헌점이 있다고 생각한다. 선행연구에서는 감리지적기업의 특성을 주로 연구하였고, 로짓(logit) 회귀분석모형으로 감리지적기업의 구분을 시도하였으나 연구의 주된 목적은 감리지적기업의 중요 특성변수를 찾는 것에 머물렀다. 그러나 본 연구는 선행연구에서 한 단계 더 나아가 감리지적기업을 효율적으로 분류하는 방법을 제시함으로써 금융감독원의 감리업무에 실질적으로 사용될 수 있는 도구를 제공할 수 있을 것이다. 특히 증권관련 집단소송법의 도입으로 인하여 2005년부터는 감리대상기업 선정방법을 심사감리로 변경하였다. 심사감리는 무작위표본추출로 선정된 많은 상장기업들을 대상으로 회계부정 혐의기업을 일차적으로 선정하게 된다. 감독당국은 이 단계에서 선정기준이 있어야 하는데, 선행연구들은 주로 재무적으로 한정된 특성변수들만 제시한데 반하여, 본 연구는 좀더 활용가능성이 높고 또 효율적인 분류방법을 보여주고 있다.

그리고 본 연구는 회계학 분야에서 사용하지 않았던 여러 분류방법들을 사용하고 있다. 특히 앙상블 방법을 회계학 분야에서 최초로 적용하여 그 효율성을 검증하고 있어서 학술적인 가치도 크다고 생각한다. 또한 인접 경영학 분야의 연구방법을 회계학에 적용하고 있어서 학제간 연구로서의 기능도 수행하고 있다고 본다.

본 연구의 2장에서는 선행연구를 살펴보고, 본 연구가 갖는 학술적 의미와 실무적 의의를 재확인한다. 특히 본 연구에서 다루는 분류방법인 베이지안 망인 GBN, NBN과 인공지능 방법인 C5.0에 대한 설명과 관련 기존연구를 소개한다. 3장에서는 연구자료 및 변수선정과정을 설명하였고, 4장에서는 실험 및 그 결과를 해석한다. 그리고 마지막 5장에서는 결론 및 향후 연구주제를 제시한다.

II. 선행연구의 검토

2.1 회계부정기업과 감리지적기업의 특성

회계부정기업의 특성을 알 수 있으면 회계부정기업이나 회계부정의 가능성이 높은 기업을 파악할 수 있어서 이해관계자들이 기업에 대한 경제적 의사결정을 내리는데 큰 도움을 줄 수 있다. 하지만 선행연구의 필요성은 있으나 회계부정기업의 특성을 찾는 연구는 많지 않다. 그 이유는 연구대상이 되는 회계부정기업을 찾기가 쉽지 않기 때문이다. 어느 기업도 스스로 회계부정을 했다고 밝히지는 않을 것이다. 따라서 회계부정기업에 대한 연구는 권위 있는 기관이 회계부정기업이라고 밝힌 기업들을 대상으로 연구를 수행하고 있다. 이러한 기업들은 미국의 경우에는 SEC에서 회계부정을 했다고 공시한 기업들이고 한국의 경우에는 금융감독원의 감리에 지적을 받은 기업들이다.

Beneish(1994)는 미국 SEC에서 회계부정으로 지적받은 48개의 기업과 각종 미디어 서치(media search)에서 이익조작기업으로 확인된 26개 기업의 재무적 특성을 연구하였다. 연구방법으로는 이들 기업과 통제기업간의 재무적 특성을 서술적(descriptive)으로 비교하고, 프로빗(probit) 모형으로 양 집단간에 유의적인 차이를 나타내는 재무변수를 분석하였다. 연구결과에 의하면, 매출채권 증가율, 자산 질(asset quality)의 감소율, 매출액 성장률, 그리고 발생액(accruals)의 증가율 등이 회계부정기업의 특성으로 나타났다. Beneish(1997)도 Beneish(1994)와 유사한 연구방법으로 회계부정기업의 특성을 연구하였는데 신용매출 증가율, 총자산에서 발생액이 차지하는 비율, 주가수익률, 현금매출 증감률이 회계부정기업의 재무적 특성으로 밝혀졌다. 특히 Beneish(1997)는 회계부정기업의 통제기업을 선정할 때 재량적 발생액이 유사한 기업들을 선택하였다. 이는 대부분의 이익조작기업들은 재량적 발생액이 크기 때문에 이익조작기업에 재량적 발생액이 큰 기업을 대응시켜서 두 집단의 특성을 비교, 분석하면 재량적 발생액의 수준이 통제되어 이익조작기업의 특성을 좀더 보수적으로 파악할 수 있는 장점이 있다..

Beneish(1999)는 회계부정기업과 관련된 경영자의 행위를 살펴보았다. 이 연구에서도 연구방법은 프로빗(probit) 모형을 이용하였다. 그는 이익조작기업의 경영자들은 이익조작 기간 동안 자신의 주식을 매각하거나 주식매수청구권을 행사함을 발견하였다. 이러한 연구결과는 경영자가 자신의 이기적인 목적을 달성하기 위해서도 이익조작을 하고 있음을 나타내고, 동시에 경영자의 주식매매행위를 통해서 회계부정을 발견할 수도 있음을 의미한다. Summers and Sweeney(1998)도 재무제표를 분식한 기업을 연구하면서 Beneish(1999)와 유사한 결과를 보고하였다.

Beasley(1996)는 Beneish의 선행연구들과 다르게 재무제표 분식기업의 지배구조에 대하여 연구하였다. 연구방법으로는 로지스틱(logistic) 회귀분석모형을 사용하였는데, 회계부정기업들은 이사회에서 사외이사의 비중이 낮고, 사외이사 중에서 기업과 독립적인 구성원(independent board members)이 적으며, 감사위원회의 독립성도 상대적으로 낮게 나타났다. 이 결과는 기업지배구조가 회계부정을 방지할 수 있는 수단이 될 수 있음을 보여주는 것이다. Beasley et al.(1999)도 SEC에서 감리지적을 받은 200개 기업의 기업특성과 지배구조를 연구하면서, 회계부정기업들은 감사위원회가 제 구실을 하지 않으며, 이사회 구성원 중에서 사외이사가 차지하는 비중이 낮고, 창업자나 소유자가 이사회의 구성원인 경우가 많으며, 대부분의 경우(72%) 최고경영자가 회계분식에 연관되어 있음을 보고하였다.

국내에서는 최 관과 백원선(1998)이 금융감독원의 감리지적을 받은 기업들의 재무적 특성을 연구하였다. 연구결과에 따르면, 회계부정기업은 금융비용의 부담이 크고, 전기손익수정손익이 크며, 재량적 발생액도 높게 나타났다. 유사한 연구로서 박종성(1999)은 재무적 특성 이외에 감사인의 특성과 계속감사기간의 차이도 분석하였다. 그리고 최 관과 최국현(2003)은 금융감독원뿐만 아니라 한국공인회계사회의 감리에 지적을 받은 기업들까지 연구대상기업에 포함하여 회계부정기업의 특성을 연구하였다. 이 연구들도 연구방법으로는 로짓(logit) 회귀분석을 사용하여 회계부정기업의 특성을 파악하였다. 연구결과에 따르면, 매출액순이익율의 크기, 현금흐름 대비 유동부채 비율, 금융비용부담

률, 특수관계자와의 거래금액, 감사인 변경여부 등이 금융감독원의 감리지적 기업의 특성으로 나타났다. 한편, 한국공인회계사회의 감리지적기업은 현금흐름 대비 유동부채비율, 회계법인의 Big5 여부만이 통제기업과 유의한 차이를 보이고 있다고 보고하였다.

이상과 같이 회계부정기업의 특성은 여러 선행연구에서 체계적으로 연구되고 있다. 그러나 기존 연구들은 공통적으로 몇 가지 한계점을 갖고 있다. 첫째, 회계부정기업을 통제기업과 비교분석할 때 사용하는 연구모형이 프로빗(probit) 모형이나 로짓(logit) 회귀분석모형에 국한되어 있다. 이 연구모형들은 종속변수와 설명변수의 관계를 획일적인 선형으로 가정하고 있어서 설명변수간의 인과관계 등을 고려하지 못한다. 따라서 연구모형에 대하여 좀더 폭넓은 시각으로 다양한 분석방법을 적용해 보는 시도가 필요할 것으로 생각한다. 본 연구에서는 인접 학문분야에서 활발히 사용되는 연구모형을 적용시켜보고 있다. 둘째, 로지스틱(logistic) 모형을 사용한 기존연구에서는 회계부정기업의 예측확률이 그리 높게 산출되지 않았다. 만약 인접학문분야에서 사용되는 연구모형들이 회계부정기업의 예측확률을 높일 수 있다면 의미 있는 일이 될 것이고, 또한 이러한 모형을 회계감리 실무에서도 적용할 수 있는 가능성도 제시할 것이다.

2.2 베이지안 망(Bayesian Network)

베이지안 망은 경영과학의 문제 해결, 특히 재무와 마케팅 분야에서 유용하게 사용되어온 의사결정 지원 방법이다(Pearl, 1988; Sprites et al., 1993; Jensen, 1996; Neapoltan, 2004). Sarkar and Sriram(2001)은 파산과 관계된 베이지안 망의 조기경보(early warning) 능력을 보여주었고, Gemela(2001)는 재무비율간에 존재하는 상호의존성을 고려하여 대출평가의 질을 향상시켰다. Wong et al.(2004)과 Baesens et al.(2004)는 데이터베이스 마케팅 문제를 푸는데 베이지안 망이 얼마나 유효한지를 실증적으로 분석하였다. 베이지안 망이 경영의사결정문제에 다양하게 적용된 데에는 베이지안 망이 가지고 있는 다음과 같은 몇 가지 특성 때문이다.

첫째, 베이지안 망은 주어진 의사결정문제의 영역지식을 확률적으로 표현한다. 즉, 주어진 문제를 구성하는 변수들간에 존재하는 확률적 의존관계의 방향을 아크(arc)로 나타내고 각 변수들이 갖는 조건부 확률을 계산함으로써 문제에 포함된 변수들간의 인과관계를 고려할 수 있다(Heckerman, 1995; Jensen, 1996).

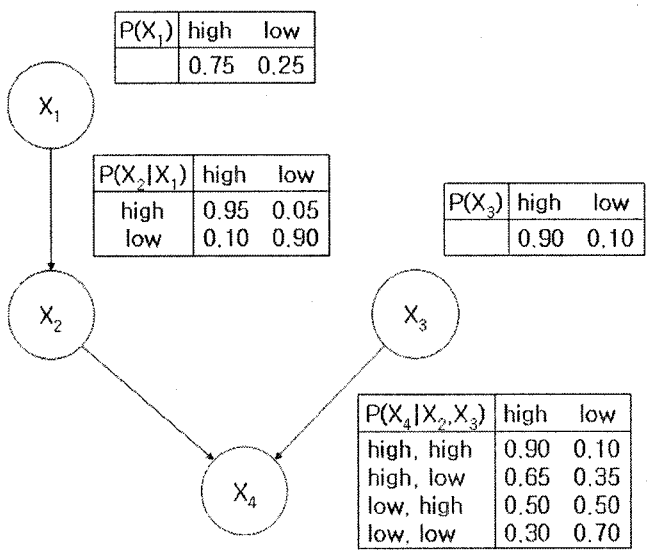
둘째, 베이지안 망은 주어진 클라스 노드, 즉 결과변수에 영향을 주는 최소한의 설명변수의 집합을 찾아냄으로써 이른바 특징추출(feature selection)[2] 이 가능하다. 일반적으로 이와 같은 특징추출을 통해 처리 대상 데이터의 양을 줄임으로써 계산의 효율성을 높일 수 있고, 보다 함축된 분류 지식이나 패턴을 얻을 수 있으며, 때로는 분류방법의 성능을 크게 향상시킬 수 있다. 이와 같이 베이지안 망에 의하여 선택된 최소한의 설명변수의 집합을 마코프

2) 일반적으로 한 의사결정문제를 표현하는 중요한 속성(attribute)들을 특징(feature)이라고 한다. 주어진 의사결정문제의 결과변수, 즉 클라스 노드(class node)를 판단하는데 큰 영향을 미치지 못하는 특징들은 삭제하고 반대로 중요도가 높은 특징들만을 골라내는 과정을 특징추출(feature selection)이라고 하며, 이는 일반적으로 주어진 의사결정문제를 구성하는 변수의 숫자를 줄여서 문제해결을 보다 쉽게 해준다.

블랭킷(Markov Blanket)이라고 하며(Pearl, 1988), 주로 일반 베이지안 망인 GBN에서 연구되고 있다(Cheng and Greiner, 1999; Margaritis and Thrun, 1999).

베이지안 망의 첫 번째 특징을 자세하게 살펴보자. 하나의 베이지안 망은 각 노드마다 하나의 조건부 확률표(conditional probability table)를 갖는 비순환 유향 그래프(direct acyclic graph)인 G=〈N, A〉이다. 따라서 베이지안 망을 B라고 표기하면, B=〈N, A, Θ〉으로 정의할 수 있다. 이때 각 노드 n∈N은 하나의 영역변수를, 각 아크 a∈A는 두 변수간의 확률적 의존성을 나타내며, Θ는 조건부 확률들의 집합을 나타낸다. 일반적으로, 하나의 베이지안 망은 다른 노드들에 배정된 값들을 기초로 특정 노드가 가질 값에 대한 조건부 확률을 계산하는데 이용할 수 있다. 따라서 하나의 베이지안 망은 한 개체의 다른 속성들의 값이 주어졌을 때 클라스 노드 (또는 결과변수)의 사후 확률분포(posterior probability distribution)를 구해줌으로써 주어진 의사결정문제의 클라스 노드에 대한 분류함수(classifier)로 이용될 수 있다(Kevin, 2001; Pearl, 1998). 하나의 데이터 집합으로부터 베이지안 망을 학습할 때 베이지안 망의 각 노드는 데이터 집합의 각 속성을, 각 아크는 속성들 간의 확률적 의존성을 표현하게 되며, 이렇게 학습된 베이지안 망을 기초로 주어진 클라스 노드를 확률적으로 예측할 수 있다. 예를 들어 변수 X_1, X_2, X_3, X_4가 있고, 각 변수들이 높은 값(high)과 낮은(low) 값을 갖는다고 할 때, 이를 통하여 얻어지는 베이지안 망은 〈그림 1〉과 같이 각 변수들에 대한 의존성을 그래프로 표현할 뿐 아니라 각 변수별로 조건부 확률도 함께 표현할 수 있다.

베이지안 망에는 여러 가지 유형이 있다. 즉, 나이브 베이지안 망인 NBN(Naïve Bayesian Network), NBN을 트리형태로 확장한 TAN(Tree Augmented

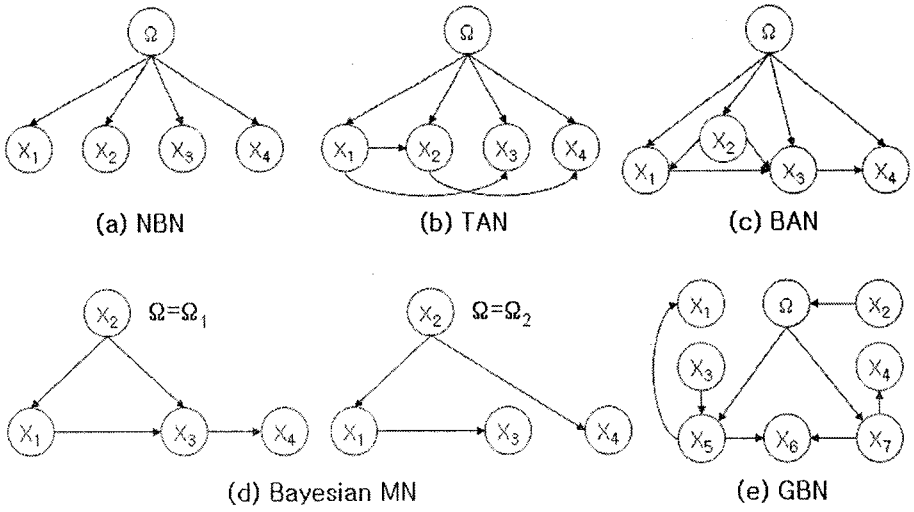


〈그림 1〉 조건부 확률표를 갖는 베이지안 망

Naive Bayesian Network), 역시 NBN을 확장한 BAN(Bayesian Augmented Naive bayesian network), 베이지안 멀티넷(Bayesian Multi-net: 베이지안 MN), 그리고 GBN(General Bayesian Network)이 있다. NBN은 〈그림 2(a)〉와 같이 클라스 노드 Ω를 제외한 다른 모든 설명변수, 즉 속성노드들이 클라스 노드에만 의존적이고, 그들 간에는 서로 독립적이라는 것을 가정한 베이지안 망이다. NBN은 가정은 단순하지만 많은 연구를 통해 비교적 높은 분류 성능을 보여주는 것으로 알려져 있다. 하지만 이 가정은 실세계 문제들에서는 만족하지 않은 경우도 있기 때문에 성능이 떨어질 수 있다(Domingos and Pazzani, 1996).

Friedman et al.(1997)에 의해 소개된 TAN은 NBN이 자식노드들 사이에 너무 독립적이라는 가정을 하는 것을 완화하기 위하여 〈그림 2 (b)〉와 같이 자식노드들 사이에 트리형태의 관계가 있음을 가정한 베이지안 망이다. TAN의 성능은 Friedman et al.(1997)과 Cheng and Greiner (1999)의 의해 연구되었다. BAN은 NBN과는 달리 설명변수인 속성노드들 간에도 상호의존성이 존재한다고 가정하고 이러한 속성노드간 상호의존성을 하나의 일반 베이지안 망 형태로 표현 가능하도록 NBN을 확장한 것이다. 즉, BAN은 〈그림 2(c)〉와 같이 클라스 노드 Ω를 제외한 다른 모든 속성노드들 간의 상호의존성을 또 하나의 베이지안 망으로 표현할 수 있다. BAN 분류기의 성능은 Friedman et al.(1997)과 Cheng and Greiner (1999)에 의해 연구되었다. 베이지안 MN은 Geiger and Heckerman(1996)이 처음 소개하였고, Friedman et al.(1997)에 의해 연구되었다. 이 베이지안 망은 〈그림 2(d)〉와 같이 클라스 노드 값에 따라 자식노드 간의 관계가 변화한다는 가정 하에 클라스 노드가 취하는 값에 따라 매번 다른 베이지안 망을 그려내고 그 베이지안 망 중에서 가장 설명력이 높은 망을 취하는 방식이다.

한편, 베이지안 망 중에서 가장 일반화된 형태가 GBN이다. GBN의 학습방법과 성능에 대해서는

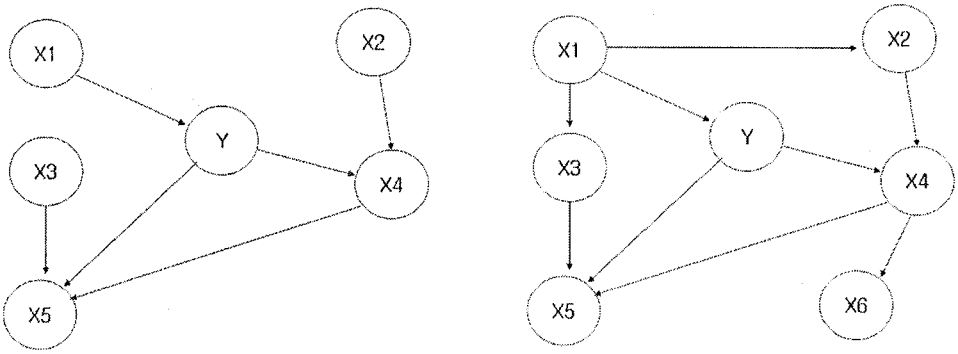


〈그림 2〉 주요 베이지안 망

Friedman et al.(1997)과 Cheng and Greiner(1999)에 의해 연구된 바가 있다. GBN에서는 기존의 다른 베이지안 망들과는 달리 클라스 노드조차 일반 속성노드와 차이를 두지 않고 모든 노드들 간의 상호의존성을 하나의 베이지안 망으로 표현한다(Bouckaert, 1995). 따라서 GBN에서는 클라스 노드도 부모노드들을 가질 수 있기 때문에 (〈그림 2(e)〉 참조), GBN은 주어진 의사결정문제에 속하는 여러 변수 (즉, 속성)간에 존재하는 확률적 인과관계 (또는 상호의존성)를 가장 자연스럽게 표시할 수 있다는 장점이 있다. 본 연구에서는 이러한 베이지안 망 중에서 문헌에서 분류방법으로서 가장 널리 활용된 NBN과 GBN을 사용한다.

베이지안 망의 두 번째 특징인 마코프 블랭킷은 주어진 의사결정 문제를 구성하는 변수들 중에서 의미있는 최소한의 변수만을 추출하는 특징선택의 기법으로 널리 사용된다 (Koller and Sahami, 1996; Tsamardinos et al., 2003). 즉, 마코프 블랭킷에 속한 변수들에 대한 지식만 가지고 있으면 결과변수인 클라스 노드의 확률분포를 결정하기에 충분하다는 의미이다. 따라서 마코프 블랭킷을 정확하게 구할 수만 있으면 해당 마코프 블랭킷 내에 포함된 변수들간의 인과관계를 통하여 주어진 의사결정문제의 클라스 노드에 대한 분류작업을 확률적으로 정확하게 수행할 수가 있다. 마코프 블랭킷을 구조적으로 설명하면, 마코프 블랭킷은 클라스 노드의 부모노드들과 자식 노드들, 그리고 자식 노드들의 또 다른 부모 노드들을 포함하는 모든 노드들의 부분 집합이다. 예를 들면, 〈그림 2(e)〉에서 클래스 노드 Ω의 마코프 블랭킷은 클라스 노드의 부모 노드{X_2}, 클라스 노드의 자식 노드{X_5, X_7}, 그리고 자식 노드의 부모 노드인{X_3}로 구성된다.

본 연구에서는 이러한 마코프 블랭킷의 특징을 이용하여 주어진 감리지적과 감리비지적간의 차이를 분석하고자 하므로 마코프 블랭킷의 의미를 더욱 자세하게 부연 설명해보자. 즉, X를 변수의 집합이라고 하고, Y를 타겟변수라고 하자. 이때 Y의 마코프 블랭킷은 X에서 가장 작은 부분집합인 Q로 구성되는데, 이때 Y는 Q를 제외한 나머지 X에 대해서는 독립이다. 이를 노드의 타입으로 설명하면, Y의 마코프 블랭킷은 세가지 타입의 노드로 구성되어 있다. 즉, Y의 부모노드, Y의 자식노드, 그리



(a) 베이지안 네트워크 (b) 마코프 블랭킷

〈그림 3〉 마코프 블랭킷의 예

고 Y의 자식노드의 부모노드 가 그것이다. 이같은 마코프 블랭킷의 개념은 다음과 같이 그림 3(a)와 3(b)를 비교하면 더욱 명확해 진다. 그림 3(b)에 나타나 있는 Y의 마코프 블랭킷은 3(a)의 노드중에서 Y의 부모노드인 X1, Y의 자식노드인 X4와 X5, 그리고 Y의 자식노드의 부모노드인 X3와 X2로 구성되어 있음을 알 수 있다.

결국, GBN은 기존의 NBN, TAN등의 베이지안 망 방법과 비교하여 볼때, 변수간에 존재하는 인과관계를 보다 자유롭게 찾을 수 있다는 측면에서 본 연구주제에 더 부합된다고 볼 수 있다. 본 연구에서는 GBN이 가지고 있는 변수간의 자유로운 인과관계 설정 기능을 주요 장점으로 채택하여 감리지적기업이 갖는 관련 변수간의 인과관계를 마코프 블랭킷으로 찾고자 하기 때문이다.

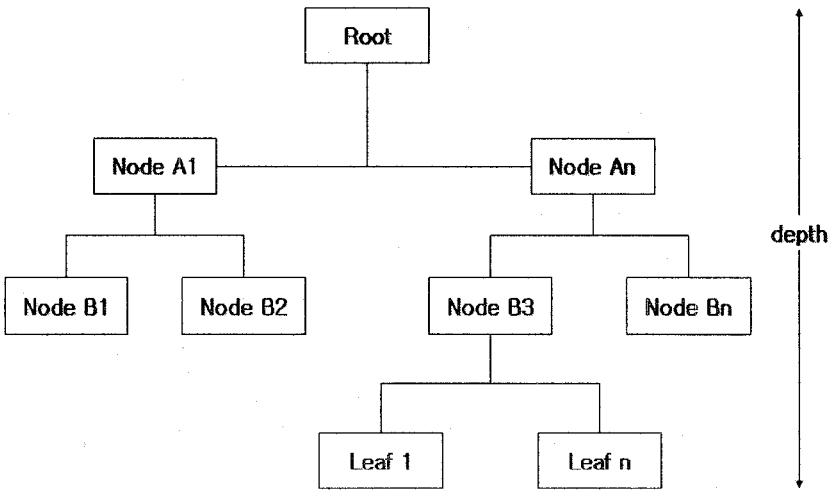
한편, GBN과 유사한 것으로 보이는 기존의 방법론인 요인분석과 구조방정식모형(SEM: Structural Equation Model), 그리고 인공신경망 방법을 비교하여 보자. 우선 요인분석은 사용자들에게 설문문항을 물어서 설문문항끼리 어떤 특정 개념, 즉 컨스트럭트 (또는 변수)로 묶이는지를 분석하는 다변량 분석기법이다. 그리고 요인분석에서는 해당 컨스트럭트간에 인과관계를 존재하는지 여부를 분석하지는 않는다. 반면에 GBN은 주어진 변수간(따로 변수를 찾기 위하여 요인분석을 돌리지 않는다는 의미임)에 존재하는 인과관계를 찾고자 하는데 사용된다는 측면에서 차이가 있다. 또한 GBN과 SEM에도 많은 차이점이 존재한다. 첫째, SEM에서는 주어진 설문자료를 토대로 우선 컨스트럭트가 제대로 묶이는지를 먼저 찾아야 한다. 반면 GBN은 미리 정해져 있는 변수간의 인과관계를 찾기만 하지 설문문항에서 컨스트럭트를 찾지는 못한다. 둘째, SEM은 컨스트럭트 사이에 미리 설정한 연구모형에서 기술되어 있는 인과관계, 즉 경로가 통계적으로 유의하게 존재하는지를 회귀분석을 기초로 분석한다. GBN 역시 변수간의 인과관계를 찾기는 하지만, 이는 어디까지나 베이지안 정리에 의한 조건부 확률에 기초하여 분석한다는 측면에서 SEM과는 확연한 차이가 있다. 인공신경망 방법은 기본적으로 입력층과 은닉층, 그리고 출력층의 다계층간에 연결가중선이 있다고 가정하여 분석하는 방법이다. 이같은 구조적 특징으로 인하여 인공신경망은 학습의 효과성이 매우 뛰어나며 또한 추론의 효과 또한 크다고 알려져 있다. 그러나, 이같은 장점에도 불구하고 인공신경망은 주어진 입력변수간에 존재하는 인과관계에 대해서는 아무런 정보를 제공하지 못한다. 오로지 입력변수와 출력변수간에 존재하는 비선형함수만을 숫자로 추정하여 그 추론결과만 알려줄 뿐이다. 따라서, 인공신경망은 높은 추론능력과 학습능력에도 불구하고 실제 의사결정 현장에서는 "어떤 입력변수간의 인과관계를 통하여 이러한 출력값이 나왔는가?"라는 핵심질문에 답을 줄 수 없다는 약점을 가지고 있다. 반면 GBN은 입력변수간의 인과관계는 물론이려니와 입력변수와 타겟변수간의 인과관계도 찾아낼 수 있고, 아울러 학습능력을 동시에 갖추고 있기 때문에 인공신경망과는 차별화된 의사결정을 내릴 수 있다.

2.3 C5.0

C5.0은 주어진 의사결정 문제를 구성하는 결과변수와 설명변수와의 관계를 〈그림 4〉에서와 같이 의사결정트리 형태로 표현한다. 특히 C5.0은 의사결정트리 형태에 입각한 귀납적 추론방법으로 다양한 분류의사결정문제에 성공적으로 적용되어 왔다. 〈그림 4〉에서 보는 바와 같이, 의사결정나무는 맨

상단에 위치하는 루트노드에서 맨 아래에 위치하는 리프노드까지가 가지로 연결되어 있는 구조이다. 그리고 루트노드로부터 리프노드까지의 단계를 깊이(depth)라고 한다.

의사결정나무는 인공지능, 기계학습 방법에서 대용량의 자료를 처리할 때에 변수의 성격이나 변수간의 관계가 불분명할 때에 종종 사용되는 분류방법이다. 특히 의사결정나무가 갖는 장점은 정보를 쉽게 이해할 수가 있다는 점이다. 즉, 루트노드에 존재하는 변수로부터 출발하여 나머지 하위 노드들 간에는 이른바 If-Then의 관계가 설정되어 있어서 최종 단말 노드인 리프노드(leaf node)까지 If-Then의 관계로 쉽게 해석이 된다. 따라서 의사결정나무는 규칙베이스로서도 널리 활용되어 왔다(Quinlan, 1986). 현재 사용되어지는 대표적인 의사결정나무 알고리즘으로는 카이제곱검정(이산형 변수인 경우), 또는 F검정(연속형 변수인 경우)을 이용하는 CHAID알고리즘, 지니지수(Gini Index)(이산형 변수의 경우) 또는 분산의 감소량 (연속형 변수)을 이용하는 CART알고리즘, 엔트로피 지수를 분리기준으로 사용하는 C5.0, 예측변수의 측도에 따라서 서로 다른 분리규칙을 사용하는 QUEST 알고리즘이 있다. 본 연구에서는 기존의 회계 및 경영분야에서 널리 활용되어온 C5.0방법을 이용하여 의사결정나무 기법을 감리지적 예측에 적용한다. C5.0은 Quinlan(1986, 1993)의 방법을 확장한 것으로서 SPSS의 클레멘타인에서 사용할 수 있다. 특히 C5.0은 기존 연구에서 많이 적용하고 있는 로짓분석과 비교하여 볼때, 다음과 같은 장점이 있다. 첫째, 의사결정나무로 결과를 보여주고 있기 때문에 변수간의 인과관계를 루트노드에서 리프노드까지 내려가면서 설명을 할 수가 있다. 둘째, 로짓분석과 달리 C5.0에서는 상위노드의 변수들이 하위노드의 변수보다 상대적으로 더 중요하다는 추가적인 정보를 알 수 있다.



〈그림 4〉 의사결정나무의 구조

III. 연구자료 및 변수선정

3.1 연구자료

본 연구에 사용한 감리지적자료는 1990년부터 1999년까지 금융감독원이 실시한 감리선정기업과 감리지적기업을 대상으로 하였다. 논문작성 시점인 2006년까지 연구자료를 갱신(update)하지 못한 이유는 금융감독원에서 감리선정기업의 자료를 대외비로 하고 있기 때문이다. 이는 감리선정이 되었으나 비지적된 기업이 추후에 회계부정기업으로 밝혀질 경우 감리기관의 신뢰성과 책임문제가 있기 때문으로 보인다. 본 논문의 연구자료는 2000년에 금융감독원으로부터 연구자료로만 사용할 목적으로 입수하였다.

〈표 1〉은 본 연구에서 사용한 감리선정기업과 감리지적기업의 연도별 자료와 지적비율을 보여주고 있다. 1990년부터 1999년까지 감리대상에 선정된 기업의 수는 1,442개이며 이중에서 감리지적기업의 수는 392개로서 전체 지적비율은 27.2%이다. 전체 감리(1,442개)에서 일반감리는 60.0%(866개), 수시감리는 33.0%(476개), 그리고 특별감리는 6.0%(100개)를 차지하고 있다.[3] 감리지적비율은 일반감리 21.8% (189/866), 수시감리 23.9% (114/476), 특별감리 89.0% (89/100)이다. 따라서 일반감리나 수시감리에 비하여 특별감리의 지적 비율이 매우 높음을 알 수 있다. 특히 특별감리

〈표 1〉 연도별 감리형태별 감리대상기업과 감리지적기업의 수[4]

감리연도	감리형태별 감리대상기업				감리지적기업			
	일반	수시	특별	합계	일반	수시	특별	합계
1990	189	121	2	312	31	34	2	67
1991	117	61	13	191	39	22	7	68
1992	81	14	22	117	17	1	22	40
1993	101	20	4	125	17	9	1	27
1994	74	51	9	134	11	12	7	30
1995	76	67	6	149	17	14	6	37
1996	43	95	2	140	6	14	2	22
1997	94	38	8	140	24	6	8	38
1998	50	8	17	75	17	2	17	36
1999	41	1	17	59	10	0	17	27
합계	866	476	100	1,442	189	114	89	392

3) 일반감리란 금융감독원이 정한 일반감리대상 선정기준에 부합하는 상장기업들과 표본추출방법에 의해 선정된 상장기업들에 대한 감리이다. 감사보고서 감리 중에서 연도에 따라서 차이는 있으나 약 60%가 일반감리이다. 특별감리는 중대한 분식회계 및 부실감사의 구체적인 사실이 인지된 회사나 부실금융기관에 대하여 실시하는 감리이다. 수시감리는 기업공개예정회사의 감사보고서를 대상으로 실시한다(금융감독원 「외부감사 및 회계등에 관한 규정」 참조. 개정 2000년 9월 1일).

4) 〈표 1〉의 연도별 일반감리 대상기업의 수와 감리지적기업의 수는 최 관과 백원선(1998)과 상당한 차이를 보이고 있다. 최관과 백원선(1998)에서는 감리지적기업의 목록과 감리지적 사항을 증권감독원에서 입수하여 연구자들이 수작업으로 자료를 정리한 연구결과인데 비하여, 본 연구는 금융감독원에서 직접 요약, 정리하여 전달해준 자료에 의존한 연구결과이다.

는 10개 연도 중에서 7개 연도에서 감리지적비율이 100%이다.

상장기업의 재무제표 자료는 한국신용평가주식회사의 데이터베이스인 KIS-FAS에서 추출하였다. 금융감독원의 감리에 선정된 1,442건 중에서 KIS-FAS에서 찾을 수 있는 자료는 1,192건이었다. 결측치가 많은 이유는 1990년과 1991년에는 금융감독원이 비상장기업에 대한 감리도 실시하였는데 KIS-FAS에는 비상장기업의 자료가 수록되어 있지 않고, 도산기업 중에서 일부기업에 대한 자료가 없기 때문이다. 1,192건의 재무제표 자료 중에서는 본 연구에서 사용되는 재무제표 자료가 없는 기업들과 금융업에 속하는 기업들을 제외하면 총 건수는 1,070건으로 감소한다. 금융업에 속하는 기업들은 일반기업과 재무구조에 큰 차이가 있어서 분석에서 제외하였다. 본 연구에 필요한 감사인과 감사의견에 관한 자료도 한국신용평가주식회사로부터 입수하였다.

3.2 특성변수

본 연구에서 사용된 변수는 선행연구에서 회계부정기업이나 이익조작기업의 분석에서 언급된 변수들과 그동안 금융감독원이 감리대상 선정기준으로 사용한 변수들을 중심으로 한국신용평가주식회사의 KIS-FAS와 동 회사에서 입수한 감사인 자료에서 활용 가능한 모든 변수들을 조사하였다. 본 연구에서 사용한 변수들은 최 관과 최국현(2003)의 연구에서 사용한 변수들과 동일하다. 〈표 2〉에 변수들의 정의를 기술하였다.

3.2.1 순이익 관련 변수

만약 기업의 경영성과가 양호하다면 경영자가 회계분식을 시도할 가능성이 줄어들고, 이에 따라서 감사인도 의도적인 부실감사를 할 가능성이 감소할 것이다. 그러나 경영성과가 예측보다 크게 향상될 경우에는 법인세를 줄이거나 기타의 이유로 순이익을 감소시키는 회계분식을 시도할 수도 있을 것이다. 경영성과 관련 변수들로는 기업의 손실발생 여부, 매출액순이익율의 크기, 총자산이익률, 매출액순이익률, 그리고 순이익증가율을 이용하였다. 손실발생 여부는 손실이 발생하면 1, 아니면 0을 부여하였다. 매출액순이익율의 크기는 당기순이익의 크기가 매출액의 -1%보다 크고 1%보다 작으면 1을, 그렇지 않으면 0을 부여한 더미변수이다. 이 변수는 금융감독원의 감리선정기준에 포함되어 있으며, Burgstahler and Dichev(1997)의 연구에서도 순이익이 0에 근접한 경우 기업의 이익조정이 매우 심한 것으로 나타났다.

3.2.2 현금흐름 관련 변수

경영성과를 나타내는 영업현금흐름도 기대보다 양호하다면 경영자가 회계부정을 시도할 가능성이 작을 것이고, 감사인의 부실감사 가능성도 낮을 것이다. 현금흐름 관련 변수들은 현금흐름/매출액, 현금흐름증가율, 그리고 현금흐름/유동부채를 사용하였다. 현금흐름 관련 변수 중에서 (현금흐름/유동부채)는 경영성과 변수보다는 기업의 유동성을 나타낼 수 있다. 즉 이 변수는 단기에 지급하여야 할 부채에 대한 현금창출능력을 나타낼 수 있어서 기업의 현금지급능력이나 도산가능성을 추정할 수 있는 변수도 될 수 있다.

3.2.3 재무구조 관련 변수

재무구조가 건실하면 이에 대한 회계부정의 가능성도 낮을 것이다. 재무구조가 악화되었거나 위약하면 재무구조를 잘 보이려는 회계조작의 가능성이 높아질 것이다. 따라서 부채비율이 높을수록, 금융비용이 매출액에서 차지하는 비중이 클수록, 부실공시와 부실감사의 가능성이 커져서 감리지적 가능성도 높아진다. 재무구조 관련 변수로는 유동비율, 부채비율(총부채/총자산), 총부채/매출액, 그리고 금융비용/매출액을 사용하였다.

3.2.4 발생액 관련 변수

발생액은 순이익에서 영업활동으로 인한 현금흐름을 차감한 금액이다. 많은 선행연구들은 경영자의 재량적 회계선택과 이익조정 및 회계추정 등이 발생액에 포함되어 있음을 보여주었다. 그리고 Jones (1991)와 Dechow et al.(1995)은 발생액 중에서 재량적 발생액(DA: discretionary accruals)을 추정하여 이익조정의 여부와 정도를 분석할 수 있는 측정치를 개발하였다. 본 연구에서는 선행연구에서 이익조정의 측정치로 사용한 총발생액의 차이, Jones 모형으로 산출한 DA, 수정 Jones 모형으로 산출한 DA를 이용하여 감리지적기업의 특성을 연구한다. 이때 재량적 발생액은 연도별-산업별(산업중분류) 횡단면 분석으로 추정하였다. 추가적으로 유동발생액과 비유동발생액의 변화율과 총자산에서 차지하는 비율도 고려하였다.

3.2.5 매출채권, 재고자산 관련 변수

자산 중에서 매출채권과 재고자산은 가장 회계부정의 대상이 되기 쉬운 자산이다. 왜냐하면, 이 자산들은 거래 빈도가 많고, 대손의 추정이나 자산의 평가 또는 원가배분 등 회계처리에 경영자의 재량권이 행사되기 쉽기 때문이다. 실제로 회계분식으로 밝혀진 기업 중에는 가공의 매출이나 재고자산, 수익의 조기인식 등이 회계부정의 도구로 빈번하게 사용되고 있다. 매출채권, 재고자산 관련변수로는 매출채권/총자산, 재고자산/총자산, △(매출채권/총자산), △(재고자산/총자산), 그리고 매출채권/재고자산을 이용하였다.

3.2.6 성장률 관련 변수

성장률 관련 변수로는 매출액성장률과 매출원가성장률을 사용하였다. Stice(1991)는 성장률이 높은 경우 내부통제제도가 이에 따르지 못하기 때문에 내부통제제도의 유효성이 낮아진다고 주장하였다. 따라서 성장률이 높을수록 부실공시의 위험과 부실감사의 가능성이 높아질 것으로 예측할 수 있다.

3.2.7 지배구조 관련 변수

지배구조 관련 변수로는 지분율 자료를 이용하였다. KIS-FAS에는 정부, 금융기관, 외국인, 개인의 지분율과 소액주주 지분율 및 대주주 1인 지분율을 자료가 포함되어 있다. 본 연구에서는 우선 기관지분율을 정부+증권회사+정부업체+금융기관+보험회사의 지분율을 기관지분율 1로 정의하고, 기관지분율 1에 기타법인과 외국인지분율을 합한 지분율을 기관지분율 2로 정의하여 사용한다. 기관지분율이 높을수록 기업을 감시하고 감독하는 기관의 기능이 높을 수 있기 때문에 회계부정의 가능성

이 감소할 수 있다. 또한 외국인지분율도 추가적 변수로 고려하였다. 외국인 지분율이 높을수록 기업의 감시, 감독기능이 강화될 수 있다. 그리고 우리나라에서는 최근까지도 소액주주의 권리보호가 제대로 되지 않았고 소액주주들의 권리주장도 크게 없었기 때문에 소액주주지분율이 높을수록 부실공시와 부실감사의 가능성이 높을 것으로 예측한다.

3.2.8 특수관계자와의 거래 관련 변수

특수관계자와의 거래에 관련되는 변수로는 (특수관계자와의 채권/총자산)과 (특수관계자와의 채권과 채무/총자산)을 사용하였다. 금융감독원의 감리선정기준 중에는 (최대주주 현금대여금/자기자본)과 (특수관계인 현금대여금/자기자본)이 있다. 이는 우리나라 기업들 중에서 많은 기업들이 소유와 경영이 명확히 분리되지 않고, 소유주가 기업의 자금을 유용하는 경우가 많아서 이 변수들이 사용되었다. 특수관계자와의 채권에는 현금대여금 이외의 채권도 포함되며, 특수관계자와의 거래비중은 총자산에 대한 특수관계자와의 채권과 채무를 모두 고려하여 측정치를 만들었다.

3.2.9 감사 관련 변수

이상의 변수들은 주로 기업의 재무관련 변수와 지분율 변수이었다. 본 연구는 이에 추가하여 감사인 관련 변수를 포함하였다. 이 변수들은 감사인의 부실감사 가능성을 측정한다. 선행연구에서는 감사인이 Big6(현재 Big 4)와의 업무제휴법인인가의 여부, 감사인 변경 여부, 그리고 감사의견 등이 이익조정이나 회계부정연구에서 사용되었다.[5] Big6 업무제휴법인들은 국내법인에 비하여 좀더 선진화된 감사기술을 가지고 있고 감사업무도 체계적 행하고 있으며, 소속 공인회계사의 훈련에도 보다 많은 투자를 하고 있고, 감사 후 공시 재무제표를 이용하여 계산한 재량적 발생액도 국내법인에 비하여 적다(나종길과 최 관, 2002). 따라서 Big6제휴법인이 부실감사로 감리에 지적될 가능성은 국내법인보다 낮을 것으로 예측할 수 있다. 그리고 만일 기업이 감사인을 변경한 경우에는 감사인과의 의견충돌이 있었거나 적정의견을 받을 수 있는 감사인을 선임하기 위하여 감사인 변경을 꾀할 수 있다. 따라서 감사인 변경이 있는 기업들이 부실감사의 가능성이 높고 또 감리에 지적될 가능성도 높다.

3.2.10 기타 변수

지금까지 분류한 기업특성변수 이외에 기업규모, 전기오류수정손익, 그리고 총자산회전율을 감리지적기업을 구분하기 위한 변수로 사용하였다. 김문철과 황인태(1998)는 전기손익수정손익(현재 전기오류수정손익)이 큰 경우 부실공시의 징후로 볼 수 있다고 주장하였다. 이는 고의적으로 당기 손익을 조작한 후 차기에 이를 수정할 가능성이 있기 때문이다. 총자산회전율은 기업의 활동성을 나타내는 대표적인 비율이다. 총자산회전율이 매우 낮은 경우 기업이 원활히 운영되고 있지 않음을 나타낼 수 있기 때문에 이 비율이 낮을 경우 부실공시 가능성이 높을 수 있다.

5) Big6와 업무를 제휴한 회계법인은 삼일, 안건, 산동, 세동 영화, 안진회계법인이다. 이 후에는 Big6가 합병 등의 이유로 Big4가 되었다. 현재 Big4와 업무를 제휴한 회계법인은 삼일, 삼정, Deloitte안진, 한영회계법인이다.

Ⅳ. 실험결과 및 평가

4.1 자료의 전처리

우선, 본 연구를 위하여 실시한 자료 전처리는 다음과 같다.

첫째, 1990년에서 1999년도까지의 자료 중에서 결측치는 모두 제거하였다.

둘째, 본 연구에서 사용한 변수는 3.2.에서 언급한 변수를 망라하되 이중 감사인 변경 여부 변수는 결측치가 많아서 실험에서 제외하였다. 따라서 첫째와 둘째의 전처리를 통하여 결측치를 제외한 결과 분석대상인 전체 자료수는 모두 823개이다.

셋째, 본 연구에서 사용하는 종속변수 (결과변수, 또는 클라스 노드)는 감리지적여부를 나타내는 변수(gamri)로서 (〈표 2〉를 참조) 감리지적기업이면 1, 감리비지적기업이면 0으로 표시된다. 감리비지적기업이란 감리대상기업으로 선정은 되었으나 감리에 지적되지 않은 기업이다.

넷째, 학습용 자료와 검증용 자료를 나누는 기준은 회계연도를 기준으로 했으며, 1990년도와 1995년도까지를 첫 번째 학습자료로 하고(이하 TRD1이라 칭함), 나머지 자료인 1996년도에서 1999년도까지의 자료를 첫 번째 검증자료로 하였다(이하 TED1이라 표기). TRD1의 개수는 667개,[6] TED1의 개수는 156개[7]이다.

다섯째, 1990년도에서 1996년도까지의 자료를 두 번째 학습자료로 하고 (이하 TRD2로 표기함), 나머지 자료인 1997년도에서 1999년도까지의 자료를 두 번째 검증자료로 하였다 (이하 TED2로 표기한다). TRD2의 개수는 742개[8]이고, TED2의 자료개수는 81개[9]이다.

여섯째, 본 연구에서 사용한 변수는 〈표 2〉에서 보듯이 전체 44개 변수이다. 이와 같이 고려한 변수가 많은 이유는 본 연구의 첫 번째 목적이 감리지적기업 여부에 영향을 주는 설명변수간의 인과관계를 마코프 블랭킷(Markov Blanket)으로 표시하여 의사결정을 지원하고자 함에 있기 때문이다. 따라서 GBN을 통하여 변수간의 인과관계가 자동으로 고려되어 최적의 설명변수를 추출할 수가 있기 때문에 선행연구인 최 관과 최국현(2003)과 달리 본 연구에서는 t검정과 윌콕슨(Wilcoxon)검정 등을 수행할 필요가 없었다.

4.2 실험 및 결과해석

서론에서 언급한 바와 같이 본 연구는 다음과 같이 두 개의 연구목적을 가지고 있다. 첫째, 본 연구에서는 회계학의 기존연구에서 전혀 다루지 않았던 감리지적기업과 감리비지적기업을 통계적으로 의미있게 구분시켜주는 설명변수간의 숨어있는 인과관계를 분명하게 보여주고자 한다. 본 연구에서는 이러한 목적을 위하여 일반 베이지안 망(GBN: General Bayesian Network)을 적용한다. 둘째, 본 연구에서는 감리지적기업을 정확하게 예측할 수 있는 예측방법을 찾기 위하여 기존연구에서 전혀 다루지 않았던 앙상블 방법을 제시한다. 앙상블 방

6) 이 중에서 감리비지적기업, 즉 결과변수가 0인 자료의 개수는 503개, 감리지적기업, 즉 결과변수가 1인 자료의 개수는 164개이다.
7) 이 중에서 결과변수가 0인 감리비지적기업의 개수는 96개, 결과변수가 1인 감리지적기업의 개수는 60개이다.
8) 이 중에서 감리비지적기업의 개수는 553개 (결과변수가 0), 감리지적기업의 개수는 189개 (결과변수가 1)이다.
9) 이 중에서 결과변수가 0인 감리비지적기업의 개수는 46개이고, 결과변수가 1인 감리지적기업의 개수는 35개이다.

〈표 2〉 사용변수와 정의

구분	변수명	정의와 설 명	역할
종속변수	gamri	1 이면 감리지적기업, 0이면 감리비지적기업	출력값
순이익 관련변수	loss	1이면 손실(순이익이 -), 0이면 이익이 난 기업	입력
	s_incom	매출총이익율이 매출액의 ±1% 안에 있으면 0, 아니면 1	입력
	roa	순이익/총자산	입력
	sincome	순이익/매출액	입력
	d_income	(순이익(t)-순이익(t-1))/총자산(t-1)	입력
현금흐름 관련변수	scfo	현금흐름/매출액	입력
	d_cfo	(영업현금흐름(t)-영업현금흐름(t-1))/총자산(t-1)	입력
	cfoliab	현금흐름/유동자산	입력
재무구조 관련변수	liq ratio	유동자산/유동부채	입력
	lev	총부채/총자산	입력
	slev	총부채/매출액	입력
	finexp	금융비용/매출액	입력
발생액 관련변수	d_tac	(총발생액(t)-총발생액(t-1))/총자산(t-1)	입력
	dcsca	재량적유동발생액 (Jones 모형)	입력
	dcstac	재량적총발생액 (Jones 모형)	입력
	mdcsca	재량적유동발생액 (수정 Jones 모형)	입력
	mdcstac	재량적총발생액 (수정 Jones 모형)	입력
	d_ca	(유동발생액(t)-유동발생액(t-1))/총자산(t-1)	입력
	d_nca	(비유동발생액(t)-비유동발생액(t-1))/총자산(t-1)	입력
	dca	유동발생액(t)/총자산(t-1)	입력
	dnca	비유동발생액(t)/총자산(t-1)	입력
	dtac	총발생액(t)/총자산(t-1)	입력
매출채권, 재고자산 관련변수	ar_ast	매출채권/총자산	입력
	inv_ast	재고자산/총자산	입력
	d_ar_ast	(ar_ast)-ar_ast(t-1))/총자산(t-1)	입력
	d_invast	(inv_ast(t)-inv_ast(t-1))/총자산(t-1)	입력
	ar_inv	매출채권/재고자산	입력
성장률 관련변수	sgrowth	매출액 성장률, (매출(t)-매출(t-1))/매출(t-1)	입력
	assetovr	매출액/(총자산-토지-건설중인자산)	입력
	cgschge	매출원가 변화율, (매출원가(t)-매출원가(t-1))/매출원가(t-1)	입력
지배구조 관련변수	instshr	기관지분율 1	입력
	intfshr	기관지분율 2	입력
	owner1	대주주 1인 지분율	입력
	mnrtysum	소액주주지분율	입력
	forgnr	외국인지분율	입력
특수관계자와의 거래 관련 변수	spcloan	특수관계인과의 채권/총자산	입력
	spctrade	특수관계자와의 채권과 채무/총자산	입력
감사 관련변수	b6	Big6 재휴법인이면 1, 아니면 0	입력
	gamsa	적정의견이면 1, 아니면 0	입력
기타변수	size	기업규모(1000원 단위)의 자연로그	입력
	prgainrt	전기손익수정이익/순이익	입력
	prlossrt	전기손익수정손실/순이익	입력
	sassets	매출액/총자산	입력

법은 베이지안 망 방법과 (일반 베이지안 망인 GBN과 나이브 베이지안 망인 NBN) 여타 인공지능방법인 C5.0 방법을 결합한 앙상블방법으로서 상대적으로 정확하게 감리지적기업 여부를 예측할 수 있다.

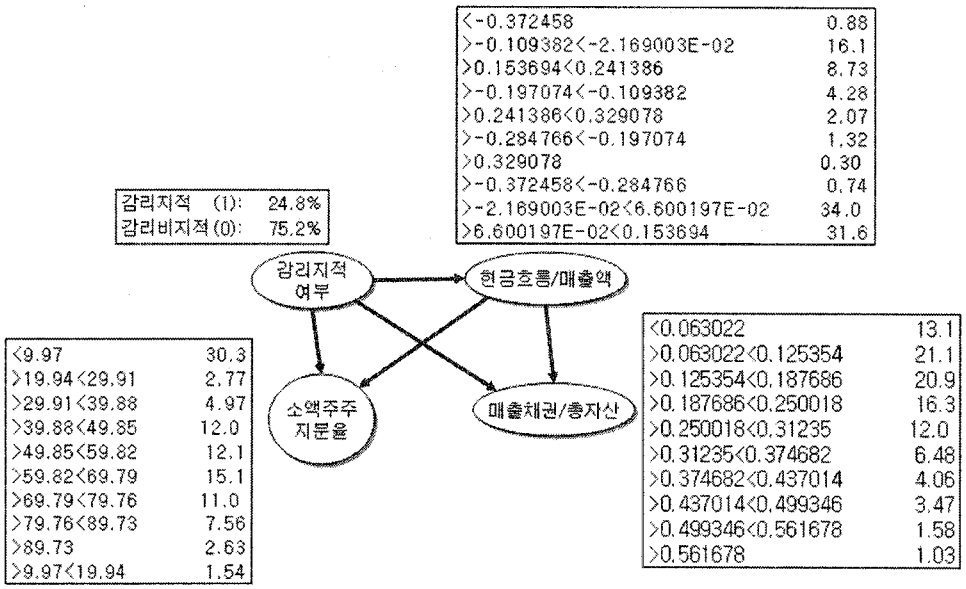
4.2.1 첫 번째 연구목적을 위한 실험

첫 번째 연구목적을 확인하기 위하여 본 연구에서는 GBN을 적용하여 감리지적여부에 영향을 주는 설명변수를 골라내고 이들 설명변수 간에 존재하는 인과관계를 분석한다. GBN을 통하여 추출되는 마코프 블랭킷(Cheng et al., 2002; Tsamardinos et al., 2003)은 감리지적기업의 특성에 영향을 주는 최소한의 설명변수의 집합을 의미한다. 여기서 중요한 것은 마코프 블랭킷에 포함되는 설명변수의 수가 최소한으로 유지된다는 점이다. 즉, 주어진 종속변수에 해당되는 감리지적기업과 감리비지적기업을 구분하는데에 도움이 되지 않는 불필요한 설명변수는 포함되지 않는다는 점이다. 이러한 마코프 블랭킷의 특성은 감리지적기업에 영향을 주는 의미있는 설명변수를 추출하고자 하는 회계학 선행연구들과 그 궤를 같이 한다. 만약, 마코프 블랭킷에 포함되는 설명변수를 도출할 수 있으면 이들 설명변수는 나머지 기타 설명변수와 달리 주어진 감리지적기업 특성을 가장 잘 설명해주는 변수로서의 의미가 있다.

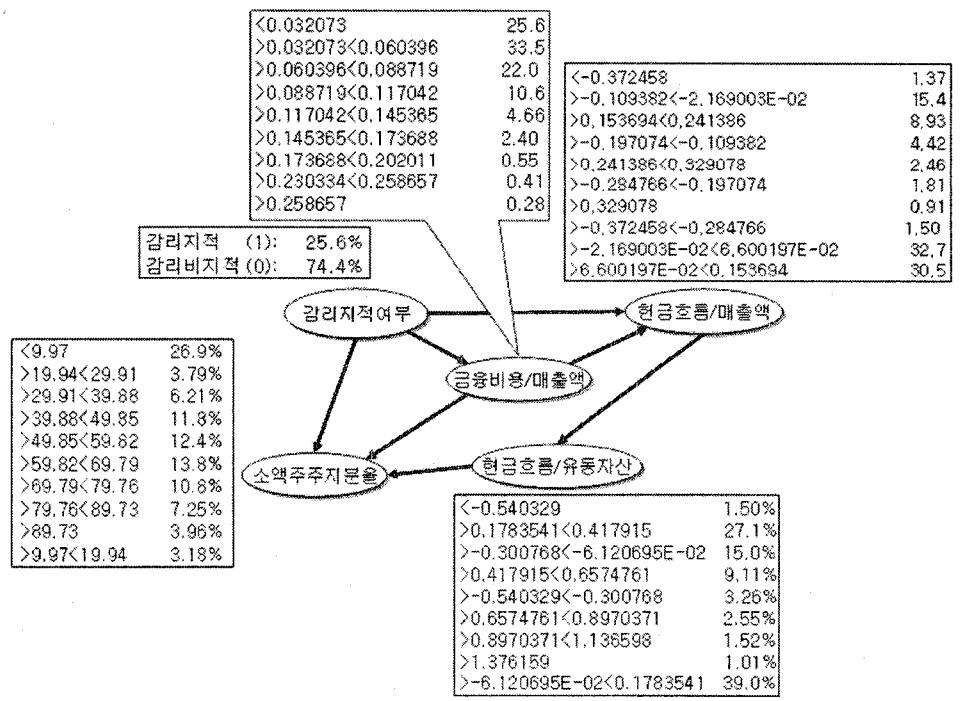
2장의 선행연구 검토에서 살펴본 바와 같이 GBN은 주어진 자료로부터 마코프 블랭킷을 추출하여 이른바 자료안에 숨겨진 변수간의 인과관계를 추출하는데 매우 유용한 방법이다(Cheng et al., 2002). 이같은 GBN의 특성을 이용하여 첫 번째 연구목적을 확인하기 위한 실험을 하였다. GBN실행을 위해서는 모든 연속형 변수를 이산화(discretization) 할 필요가 있다. 본 연구에서는 기본적으로 10개의 구간으로 나누어 이산화하는 것을 원칙으로 하였다. TRD1과 TRD2를 대상으로 GBN을 적용한 결과 〈그림 5〉과 〈그림 6〉과 같이 마코프 블랭킷이 도출되었다. 〈그림 5〉에서의 TRD1에서 도출된 마코프 블랭킷에는 '소액주주 지분율'과 '현금흐름/매출액' 및 '매출채권/총자산'이 포함되었고, 〈그림 6〉의 TRD2에서 도출된 마코프 블랭킷에는 '현금흐름/유동자산', '소액주주 지분율', '현금흐름/매출액', 그리고 '금융비용/매출액'이 포함되었다.

둘째, 앞에서 설명한 바와 같이 마코프 블랭킷 내에 포함된 설명변수만이 감리지적기업의 특성과 확률적으로 연결이 되는 변수이고 이들 변수간에 존재하는 조건부 확률값을 계산하여 인과관계를 표시할 수가 있다(Dzeroski and Zenko, 2002). 이러한 인과관계 그래프는 의사결정자로 하여금 어떤 설명변수가 다른 여타 설명변수와의 직접적 또는 간접적 인과관계를 통하여 최종 결과변수인 감리지적기업 노드와 연결이 되는지를 파악할 수가 있게 하기 때문에 다양한 민감도 분석을 수행할 수가 있다.

본 연구에서 감리지적기업과 감리비지적기업을 구분하는 분류함수를 도출하기 위하여 사용되는 C5.0, 인공지능, 그리고 로짓분석방법과 달리 GBN은 마코프 블랭킷 내에 포함된 설명변수와 종속변수간의 인과관계 분석을 통하여 What-If 분석을 수행할 수 있다. 즉, 베이지안 망의 기본 원칙인 베이지안 정리에서 알 수 있듯이 사전확률과 사후확률의 변화를 통하여 특정 설명변수가 복잡한 인과관계를 통하여 종속변수에 영향을 얼마나 주는지 민감도 분석이 가능하다. 베이지안 망을 위한



〈그림 5〉 TRD1에서 도출된 마코프 블랭킷



〈그림 6〉 TRD2에서 도출된 마코프 블랭킷

소프트웨어인 네티카(NETICA)에 통해서 이미 위에서 구한 GBN의 결과값인 각 변수별 사전확률과 사후확률 값을 입력하여 What-If 분석과 같은 민감도 분석을 수행하면 다음과 같다.

우선, TRD2에서 구한 마코프 블랭킷은 감리비지적의 사전확률이 74.4%로 감리지적일 확률보다 약 3배 정도 높다. 현금흐름/매출액(scfo)이 0.153694보다 크고 0.241386보다 작을 경우에는 감리비지적의 사후 확률이 87.8%로 올라간다. 그러나 현금흐름/매출액이 -0.284766 보다 크고 -0.197074보다 작을 경우에는 감리비지적의 사후확률이 39.3%로 변하며, 반대로 감리지적의 사후 확률은 25.6%에서 60.7%로 크게 오른다.

금융비용/매출액(finexp)의 경우에는 예외는 존재하지만 값이 커질수록 감리지적일 확률이 높아지는 경향을 보인다. 마코프 블랭킷을 이용하여 what-if분석을 해보면, 현금흐름/매출액(scfo)이 -0.284766 이상 -0.197074 이하이고, 금융비용/매출액(finexp)이 0.117042 이상 0.145365 이하, 소액주주지분율(mnrtysum)이 59.82% 이상 69.79% 이하, 현금흐름/유동자산(cfoliab)이 -0.300768 이상 -6.120695E-02 이하인 경우에는 감리지적확률이 95.5%로 상승한다.

이와 같이 GBN은 감리지적 및 감리비지적 분야에서 두 가지 의의를 갖는다. 첫째는 마코프 블랭킷을 통하여 감리지적과 감리비지적에 영향을 미치는 최소한의 설명변수를 도출할 수 있다는 점이고, 둘째는 마코프 블랭킷내에 포함된 이들 설명변수간의 인과관계 분석을 통하여 다양한 민감도 분석을 할 수가 있어서 의사결정자들에게 매우 강력한 의사결정지원 기능을 제공한다.

4.2.2 두 번째 연구목적 확인을 위한 실험

감리지적과 비지적에 관한 선행연구를 살펴보면 연구주제가 감리지적기업과 감리비지적기업간의 특성을 구분할 수 있는 변수의 파악(최 관과 최국현, 2003)에 집중되어 있다.

그러나 이와 같은 연구주제는 사실 감리지적여부를 결정하여야 하는 실무계의 입장에서 본다면 의사결정지원의 수준이 낮을 수밖에 없다. 실제로 감리지적여부를 결정하여야 하는 실무적인 입장에서 본다면 본 연구에서 제시한 연구목적인 감리지적기업을 정확하게 예측할 수 있는 예측방법을 강구할 필요가 있다. 따라서 이러한 연구목적을 달성하기 위해서는 기존 회계학 분야에서 주로 사용되어온 로짓분석에만 의존할 수는 없다. 왜냐하면 이러한 로짓분석은 그 자체로서 충분히 통계적으로 의미있는 예측함수로서의 역할을 하지만, 본 연구의 대상인 감리지적 문제에서 기존 방법과 비교하여 얼마나 정확하게 예측하는지에 대한 면밀한 검토가 없었기 때문이다. 따라서 본 연구에서는 이와 같은 연구목적 달성을 위하여 인공지능방법인 C5.0, GBN, NBN방법을 시로 연결한 앙상블 방법을 제안한다.

이와 같은 앙상블 방법은 이미 여러 학자들에 의하여 그 우수성이 증명된 방법이다(Dzeroski and Zenko, 2002; Ji and Ma, 1997). 앙상블 방법이 갖는 특징은 다음과 같다.

첫째, 서로 다른 분류방법론을 결합한다. 분류방법론이란 주어진 분류형 의사결정문제에 적용할 수 있는 방법론을 총칭하는데 회계학분야에서는 주로 사용해온 방법론은 로짓회귀분석 방법이 있다.

둘째, 서로 다른 분류방법론을 결합하였을 때에 기대할 수 있는 가장 큰 장점은 각 방법론이 갖는 분류상의 한계점이 이러한 결합을 통하여 상쇄되어

예측력이 크게 향상될 수 있다는 점이다.

우선, 앙상블 방법을 강구하기 위하여 본 연구에서는 다음과 같이 두 단계를 거쳐서 앙상블 방법을 완성하고자 한다.

1단계: 분류기법간 가중치 결정

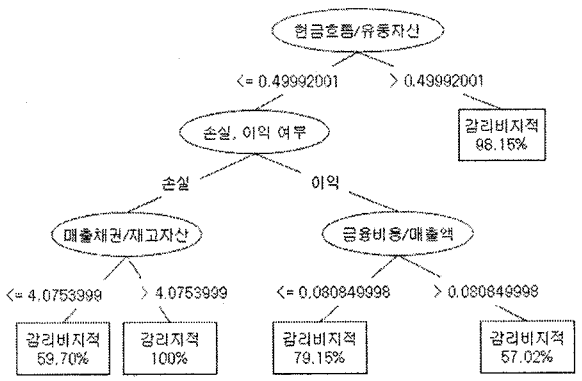
앙상블 방법을 위하여 사용된 기존의 분류기법은 모두 3개로서, 기존의 분류의사결정문제에서 널리 활용되어온 인공지능방법인 C5.0[10]과 NBN 방법, 그리고 첫 번째 연구목적에서 사용된 GBN을 적용한다. 이들 기법간의 가중치를 결정하기 위해서는 우선 각 방법을 TRD1과 TRD2를 통하여 학습을 시킨 후에 TED1과 TED2에 적용하여 예측율을 분석하여야 한다. 〈표 3〉에서 〈표 5〉까지 각 방법에 대한 검증결과가 요약 정리되어 있다.

이와 같은 3가지 분류방법론간 예측율을 보다 알기 쉽게 비교하기 위하여 〈표 6〉와 같이 정리하였다. 이 표를 보면, 전체적인 비율의 관점에서 예측율이 가장 좋은 방법은 TED1의 경우 GBN이고 TED2의 경우는 GBN과 C5.0이다. 그러나 감리지적기업을 감리지적기업으로 정확하게 예측하는 것이[11] 본 연구의 취지에 맞기 때문에 '1 → 1'의 경우만 따로 본다면 TED1의 경우 C5.0방법이 61.67%로서 가장 좋은 예측율을 보이고 있고, TED2의 경우에도 마찬가지로 C5.0방법이 34.29%로서 가장 좋은 예측율을 보여주고 있다.

한편, 세 가지 분류방법별 예측율간의 차이가 통계적으로 유의한지 여부를 확인하기 위하여 〈표 7〉에서와 같이 통계검증을 하였는바, '1 → 1' 의 경우 TED1에서는 C5.0 방법이 NBN과 GBN에 비해 95% 이상의 신뢰도하에서 통계적으로 유의한 차이를 보여주고 있으나, TED2에서는 각 방법론

10) C5.0실험은 SPSS사의 클레멘타인 8.1버전을 사용하였다. C5.0은 의사결정트리 형태로 주어진 학습자료를 학습한다. 따라서, 의사결정트리의 맨 꼭대기에 있는 이른바 루트노드(root node)에 위치한 변수가 하위노드에 위치한 다른 여타 변수에 비하여 상대적으로 중요한 분류파워가 있다고 해석한다. TRD1의 소액주주지분율(mnrtysum), 현금흐름/유동자산(cfoliab), 매출액증가율(d_sales), 손실 이익 여부(loss), 재고자산/총자산(inv_ast) 순으로 나타났다. 한편, TRD2의 C5.0분석 결과를 보면 상위 3개 층의 분류 노드에 현금흐름/유동자산(cfoliab), 손실 이익 여부(loss), 매출채권/재고자산(ar_inv), 금융비용/매출액(finexp) 순으로 나타났다. 결과는 다음 그림과 같다.

TRD2를 이용한 C5.0 분석 결과



11) 실험과정에서 감리지적의 경우 1, 감리비지적의 경우 0으로 표시하였기 때문에, 감리지적을 감리지적으로 정확하게 예측하는 경우를 〈표 7〉에서는 '1 → 1'로 표기하였다.

〈표 3〉 GBN 검증 결과

TRD1				
predict>	0	1	개별 정확도	전체 정확도
0	487	16	96.82%	
1	121	43	26.22%	79.46%
TRD2				
predict>	0	1		
0	466	12	97.49%	
1	122	67	35.45%	79.91%
TED1				
predict>	0	1		
0	90	6	93.75%	
1	51	9	15.00%	63.46%
TED2				
predict>	0	1		
0	36	10	78.26%	
1	28	7	20.00%	53.09%

〈표 4〉 C5.0의 검증 결과

TRD1				
predict>	0	1	개별 정확도	전체 정확도
0	495	8	98.41%	
1	21	143	87.20%	95.65%
TRD2				
predict>	0	1		
0	379	99	79.29%	
1	139	50	26.46%	64.32%
TED1				
predict>	0	1		
0	49	47	51.04%	
1	23	37	61.67%	55.13%
TED2				
predict>	0	1		
0	31	15	67.39%	
1	23	12	34.29%	53.09%

〈표 5〉 NBN 검증 결과

TRD1				
predict>	0	1	개별 정확도	전체 정확도
0	364	139	72.37%	
1	76	88	53.66%	67.77%
TRD2				
predict>	0	1		
0	326	152	68.20%	
1	122	67	35.45%	58.92%
TED1				
predict>	0	1		
0	59	37	61.46%	
1	36	24	40.00%	53.21%
TED2				
predict>	0	1		
0	29	17	63.04%	
1	24	11	31.43%	49.38%

간의 예측율 차이가 통계적으로 유의하지 못하다.

2단계: 가중치를 이용한 앙상블 방법도출

본 연구에서 제안하는 앙상블 분류기법의 2단계는 1단계에서 행한 세 가지 분류방법에서 나타난 '1 → 1'에 해당되는 예측율을 이용하여 이를 각 방법론별 가중치로 사용한다. 이같이 감리지적을 감리지적이라고 정확히 예측하는 비율을 중심으로

〈표 6〉 방법론간 비교(단위: %)

검증 자료	예측율 대상	GBN	C5.0	NBN	예측율 비교
TED1	전체	63.46%	55.13%	53.21%	GBN>C5.0>NBN
	0 → 0	93.75%	51.04%	61.46%	GBN>NBN>C5.0
	1 → 1	15.00%	61.67%	40.00%	C5.0>NBN>GBN
TED2	전체	53.09%	53.09%	49.38%	GBN=C5.0>NBN
	0 → 0	78.26%	67.39%	63.04%	GBN>C5.0>NBN
	1 → 1	20.00%	34.29%	31.43%	C5.0>NBN>GBN

* '전체'는 전체 예측율을 의미하고, '0 → 0'은 감리비지적기업을 감리비지적기업으로 예측한 정확도를 의미하고, '1 → 1'은 감리지적기업을 감리지적기업으로 예측한 정확도를 의미한다.

〈표 7〉 방법론간의 예측 정확도에 대한 t-검정 결과

방법론간 비교	TED1				TED2			
	0 → 0		1 → 1		0 → 0		1 → 1	
	t-값	유의확률	t-값	유의확률	t-값	유의확률	t-값	유의확률
GBN vs C5.0	7.495	0.000***	-5.942	0.000***	1.168	0.246	-1.342	0.184
GBN vs NBN	5.790	0.000***	-3.168	0.002***	1.608	0.111	-1.087	0.281
C5.0 vs NBN	-1.455	0.147	2.411	0.017**	0.433	0.666	0.251	0.803

*** $p < 0.01$
** $p < 0.05$

각 분류방법간의 상대적 가중치를 구하는 이유는 본 연구의 중요한 목적 중의 하나가 감리지적기업을 정확하게 예측하는 분류함수를 도출하는 것이기 때문이다. 분류방법별 상대적 가중치를 구하기 위하여 TRD1과 TRD2를 대상으로 각 분류방법별로 1 → 1에 대한 예측비율을 구하면 〈표 8〉과 같다. 따라서 〈표 8〉에서 구한 분류방법별 상대적 가중치를 적용하여 최종적인 앙상블 분류방법을 만들 수가 있다.

이때, TRD1에서 도출되는 앙상블 분류방법을 '앙상블_TRD1'이라고 하고, TRD2에서 도출되는 앙상블 분류방법을 '앙상블_TRD2'라고 하자. 각 앙상블 방법별 최종 결과값 산출은 다음과 같다. 이때, 앙상블 최종결과 값에서 사용되는 각 분류방법별 결과는 1 또는 0이다. 즉, 감리지적으로 추정된 경우는 1, 감리비지적으로 추정된 경우는 0이다. 따라서 0과 1사이의 값을 갖는 분류방법간 상대적 가중치를 곱한 앙상블_TRD1과 앙상블_TRD2의 최종결과 값은 항상 0과 1사이에 존재한다.

(식 1)
앙상블_TRD1의 최종결과값 = 0.1569*GBN결과 + 0.5219*C5.0결과 + 0.3212*NBN결과

(식 2)
앙상블_TRD2의 최종결과값 = 0.3641*GBN결과 + 0.2718*C5.0결과 + 0.3641*NBN결과

앙상블_TRD1에 적용될 임계치와 앙상블_TRD2에 적용될 임계치는 각각 0.15와 0.27로 결정되

〈표 8〉 TRD1에서의 분류방법별 1 → 1에 대한 상대적 가중치

분류방법	TRD1		TRD2	
	'1 → 1' 예측수 /전체수	분류방법간 상대적 가중치	'1 → 1' 예측수 /전체수	분류방법간 상대적 가중치
GBN	43/164	15.69%	67/189	36.41%
C5.0	143/164	52.19%	50/189	27.18%
NBN	88/164	32.12%	67/189	36.41%

었다.[12] 여기서 임계치란, 최종결과값이 해당 값보다 크면 감리지적 (즉, 1), 작으면 감리비지적이 (즉, 0) 되는 경계값이다. 앙상블_TRD1과 앙상블_TRD2 방법에 의한 각각의 검증결과는 〈표 9〉와 같다. 아울러, 앙상블 방법과 나머지 다른 분류방법간의 통계적 검증결과는 〈표 10〉에 정리되어 있

〈표 9〉 앙상블 분류방법에 의한 검증 결과

TRD1				
predict>	0	1	개별 정확도	전체 정확도
0	354	149	70.38%	
1	14	150	91.46%	75.56%
TRD2				
predict>	0	1		
0	339	214	61.30%	
1	76	113	59.79%	60.92%
TED1				
predict>	0	1		
0	41	55	42.71%	
1	16	44	73.33%	54.49%
TED2				
predict>	0	1		
0	20	26	43.48%	
1	16	19	54.29%	48.15%

〈표 10〉 앙상블 방법과 다른 분류방법과의 통계적 검증결과

	TED1				TED2			
	0→0		1→1		0→0		1→1	
	t값	유의확률	t값	유의확률	t값	유의확률	t값	유의확률
앙상블 vs GBN	-9.034	0.000***	7.883	0.000***	-3.618	0.000***	3.129	0.003***
앙상블 vs C5.0	-1.155	0.124	1.364	0.080*	-2.351	0.010**	1.695	0.047**
앙상블 vs NBN	-2.634	0.004***	3.880	0.000***	-1.897	0.031**	1.957	0.027**

*** p<0.01
** p<0.05
* p<0.1

12) 감리지적기업에 대해 감리비지적이라고 예측하는 경우에 오류에 따르는 비용이 크므로 보수적인 관점에서 각 앙상블 분류방법에 적용될 임계치는 TRD1과 TRD2 각각의 학습자료에서 감리지적기업을 비지적이라고 하는 오류를 최소화하는 구간에서 가장 큰 값을 임계치로 구하였다. 구간 중에서 가장 큰 값을 임계치로 결정한 이유는 앙상블 방법론의 일반화를 위해서이다. 그 결과 TRD1의 경우 임계치를 0.15로 하였을 때에 틀린 예측치가 14개로서 가장 작았으며, TRD2의 경우 임계치를 0.27로 하였을 때에 틀린 예측치가 76개로서 가장 작았다.

다. 〈표 10〉에서 알 수 있듯이 앙상블 방법은 TED2의 경우 95% 이상의 신뢰수준에서 다른 여타 방법과 통계적으로 유의한 차이를 보여주고 있다.

여기에서 우리가 유념하여야 할 사항은 앙상블 방법이 갖는 분류결과의 특이성이다.

첫째, 앙상블_TRD1은 TED1에 대해서 전체 예측율에서는 54.49%로서 여타 다른 분류방법과 비교하여 비슷하나 가장 중요한 예측율인 '1 → 1'에 대한 예측율에서는 73.33%로 가장 높은 정확도를 보여주고 있다.

둘째, 앙상블_TRD2는 TED2에 대해서 전체 예측율에서는 48.15%로서 여타 다른 분류방법과 비교하여 비슷하나 가장 중요한 예측율인 '1 → 1'에 대한 예측율에서는 54.29%로 가장 높은 정확도를 보여주고 있다.

셋째, 기존연구에서 많이 사용한 로짓회귀분석과의 비교를 위해서 Stepwise 로짓분석을 한 결과 기존의 로짓 회귀분석 방법은 본 연구에서 제안하는 앙상블 방법에 비해 '1->1'에 대한 예측율에서 큰 차이를 보였다.[13)]

13) SPSS를 이용하여 Stepwise 로짓분석을 하였다. Forward와 Backward 로짓분석을 하기 위하여, 변수진입(entry)과 변수잔류(stay)시 유의수준을 0.1로 하여 각각 분석하였다. 그 결과 TRD1에서 도출된 로짓함수는 Forward와 Backward가 약간 상이한 결과가 나왔으며(즉, 최종 설명변수 중에서 Backward에서는 Forward보다 d_ca(유동발생액증분/총자산)와 ar_ast(매출채권/총자산)가 더 선택됨) 이 중에서 카이제곱 값이 72.34로 더 큰 Backward 결과를 선택하였다.(Foward는 카이제곱 값이 65.82임)

TRD1	B	S.E.	Wald	유의확률	Exp(B)
MNRTYSUM	0.008967	0.003461	6.713747	0.009567	1.009007
OWNER1	-0.01319	0.006289	4.40042	0.03593	0.986895
cfoliab	-1.79297	0.519535	11.91007	0.000558	0.166466
d_ca	-1.91434	0.999968	3.664945	0.055568	0.147439
ar_ast	-1.42273	0.823988	2.981295	0.084232	0.241054
size	-0.25712	0.081366	9.985733	0.001578	0.773278
finexp	9.323298	2.928373	10.13646	0.001454	11195.84
prlossrt	0.115934	0.07671	2.284124	0.130704	1.122922
상수	3.285851	1.482384	4.913312	0.02665	26.73172

한편, TRD2에서도 Forward와 Backward가 약간 상이한 결과가 나왔으며 (즉, 최종 설명변수 중에서 Backward에서는 Forward보다 income(순이익), mnsrtysum(소액주주지분율), owner1(대주주 1인 지분율), intfshr(기관지분율2), prlossrt(전기손익수정손실/순이익)가 더 선택되고 d_cfo(영업현금흐름증분/총자산)가 빠짐) 이중에서 카이제곱 값이 78.02로 조금 더 큰 Backward 결과를 선택하였다(Forward는 카이제곱 값이 63.32임).

TRD2	B	S.E.	Wald	유의확률	Exp(B)
INCOME	-1.9E-08	1.15E-08	2.683893	0.101368	1
MNRTYSUM	0.011414	0.003872	8.688016	0.003203	1.011479
OWNER1	-0.01003	0.005931	2.857321	0.090959	0.990024
cfoliab	-1.25864	0.420492	8.959637	0.00276	0.284039
intfshr	-0.00901	0.004789	3.536397	0.060036	0.991035
finexp	8.102614	2.478782	10.68498	0.00108	3303.09
prlossrt	0.09349	0.065507	2.036815	0.153531	1.098
spcloan	4.445838	2.042758	4.736672	0.029526	85.27127
상수	-1.60004	0.237144	45.52366	1.51E-11	0.201889

마지막으로, 앙상블 방법의 경우는 (식 1)과 (식 2)와 같이 의사결정자가 알기 쉬운 구조인 가중합의 산술구조로 되어 있어서 이른바 화이트박스로서의 장점이 있다.

V. 결론 및 향후 연구과제

본 연구는 감리지적에 대한 예측율을 향상시키고 이를 토대로 보다 효과적인 의사결정지원이 되도록 베이지안 망을 이용한 방법을 제안하고 있다. 본 연구의 이같은 연구목적은 두가지로 세분할 수 있다. 즉, 첫 번째 목적은 어떤 기업특성변수들이 감리지적 여부와 밀접한 관계를 가지고 있는지에 대한 인과관계를 도출하여 의사결정지원을 보다 풍부하게 하는 것이고, 두 번째 목적은 감리지적에 대한 예측율을 기존연구와 비교하여 획기적으로 향상시키는 것이다. 이같은 연구목적을 달성하기 위하여 본 연구에서는 금융감독원이 실시한 감리에 선정된 기업과 이들 기업 중에서 감리지적기업들을 대상으로 하여, 감리지적기업을 설명하는 특성변수들을 추출하고 이 변수를 대상으로 C5.0, GBN, NBN, 그리고 앙상블 방법을 적용하였다. 그 결과 본 연구의 목적은 실증분석 과정을 통하여 통계적으로 유의하게 달성되었다. 선행연구와 비교하였을 때에 본연구의 대표적인 공헌점은 다음과 같다.

첫째, GBN의 마코프 블랭킷을 적용하여 감리지적 여부를 결정짓는 최소한의 설명변수를 추출하였고, 이들 설명변수간에 존재하는 인과관계를 그래프로 표시함으로써 의사결정자로 하여금 보다 알기 쉽게 감리지적 여부를 결정하도록 하였다. 감리지적 여부를 설명해 주는 이같은 인과관계 그래프는 기존의 연구에서는 볼 수 없었던 장점으로서 사용자 편의 측면에서 향후 감리지적 여부에 대한 설명력을 올리는데에 크게 기여하리라 판단된다. 또한,

이렇게 선택된 로짓 회귀분석식을 이용하여 자료를 분석해본 결과 다음과 같은 결과가 나왔다.

TRD1				
predict>	0	1	개별 정확도	전체 정확도
0	497	6	98.81%	
1	154	10	6.10%	76.01%
TRD2				
predict>	0	1		
0	465	13	97.28%	
1	187	2	1.06%	70.01%
TED1				
predict>	0	1		
0	87	9	90.63%	
1	57	3	5.00%	57.69%
TED2				
predict>	0	1		
0	20	26	43.48%	
1	16	19	54.29%	48.15%

마코프 블랭킷에 속한 설명변수와 결과변수간의 인과관계를 이용하여 민감도 분석을 할 수 있었다. 이러한 민감도 분석은 감리지적 비율을 정책적으로 조정하거나 또는 특정 설명변수의 사전확률값이 변할 때에 결과적으로 감리지적 비율이 어떻게 변하는지를 보다 체계적으로 분석할 수 있다. 이같은 what-if분석기능 역시 감리지적에 관한 기존연구에서는 볼 수 없었던 공헌점이다.

둘째, 기존문헌에서 분류방법으로서 이미 널리 사용되어온 C5.0, GBN, NBN 의 방법론을 결합한 새로운 개념의 앙상블 분류방법을 제시하여 그 성과를 실증적으로 검증하였다. 그 결과 기존의 분류방법에 비하여 매우 고무적인 결과가 나왔으며 특히 감리지적을 예측함에 있어서 탁월한 성과를 보여주었다. 감리지적 예측율을 획기적으로 올린다는 측면에서 이같은 앙상블 방법은 향후 감리지적 예측분야에서 많은 기여를 하리라 판단된다.

그러나 본 연구도 몇가지 한계점이 있다. 첫째, 본 연구에서 제안한 앙상블 방법은 선행연구에서 사용하던 로짓회귀분석 방법의 예측성과(주석 11 참조) 보다 우수하고 선행연구(최 관과 최국현, 2003; 최 관과 백원선, 1998)와 비교해서도 손색이 없는 결과이지만 실무적인 정책에 바로 적용하기에는 예측율을 더욱 높힐 필요가 있다. 둘째, 2005년부터 실시되고 있는 심사감리 절차에는 회계부정 가능성이 있는 기업의 탐색절차가 필요한 만큼, 이 논문에서 제시하는 앙상블 방법과 더불어 다른 정책적 요소나 경제사회적 환경, 기업의 지배구조 및 최고경영자의 특성들도 함께 고려된다면 더욱 효율적인 심사감리를 실시할 수 있을 것이다.

본 연구와 관련된 대표적인 향후 연구주제로서는 마코프 블랭킷에서 선택된 설명변수를 기존의 분류방법에 입력자료로 사용함으로써 기존의 분류방법의 분류성과를 향상시킬 수가 있다. 이같은 시도는 이미 베이지안 망 방법을 사용하는 기존의 연구에서 마코프 블랭킷의 유용성을 분류 예측율을 향상시키기 위한 방법으로서 널리 사용되고 있기 때문에, 감리지적 예측율 향상에도 큰 기여를 하리라 예상된다.

참고문헌

김문철 · 황인태 (1998), "감사의 품질차이가 전기손익수정에 미치는 영향," **회계학연구**, 23, 2, 1-26.

나종길 · 최 관 (2003), "회계발생액과 차별적 감사수요," **회계학연구**, 28, 1, 1-31

박종성 (1999), "피감사회사 특성과 감사인 특성을 이용한 감리지적 예측," **회계학연구**, 24, 1, 1-32.

윤중옥 · 김명환 (2001), "감리지적기업의 회계특성에 관한 연구: 감리비지적기업과의 비교를 중심으로," 회계연구, 6, 2, 45-60.

박한순 (1996), "감사인 지정이 경영자의 이익조절 행위에 미치는 영향," **회계학연구**, 21, 3, 33-61.

최 관 · 백원선 (1998), "감리지적기업의 이익조작에 관한 실증적 연구," **회계학연구**, 23, 2, 133-161.

최 관 · 최국현 (2003), "회계부정기업의 특성에 관한 연구: 감리지적기업을 중심으로", **회계학연구**, 28, 2, 211-243.

Baesens, B., G. Verstraeten, D. Van den Poel, E.-P. Michael, P. V. Kenhove, and J. Vanthienen. (2004). Bayesian Network Classifiers for Identifying the Slope of the Customer Life Cycle of Longlife Customers. *European Journal of Operational Research* 156: 508-523.

Beasley, M. S. (1996). An Empirical Analysis of the Relation between the Board of Director

Composition and Financial Statement Fraud. *The Accounting Review* 71: 443-465.

Beasley, M. S., J. V. Carcello, and D. R. Hermanson. (1999). *Fraudulent Financial Reporting: 1987-1997 An Analysis of U.S. Public Companies*, SEC Committee of Sponsoring Organizations of the Treadway Commission.

Beneish, M. D. (1994). The Detection of Earnings Manipulation. Working Paper. *Duke University*, Durham, NC.

Beneish. M. D. (1997). Detecting GAAP Violation: Implication for Assessing Earnings Management among Firms with Extreme Financial Performance. *Journal of Accounting and Public Policy* 16,: 271-309.

Beneish, M. D. (1999). Incentives and Penalties Related to Earnings Overstatements that Violate GAAP. *The Accounting Review* 74: 425-457.

Bouckaert, R. (1995). Bayesian Belief Networks: From Construction to Inference. *Doctorial Dissertation, University of Utrecht*, The Netherlands.

Burgstahler, D. and I. Dichev. (1997). Earnings Management to Avoid Earnings Decreases and Losses. *Journal of Accounting and Economics* 24: 99-126.

Chebrolu, S., A. Ajith, and T. Johnson. (2004). Hybrid Feature Selection for Modeling Intrusion Detection Systems. in .R. Pal, et al. (Eds). *11th International Conference on Neural Information Processing (ICONIP)*. Lecture Notes in Computer Science. Vol. 3316., Germany: Springer-Verlag: 1020-1025.

Cheng, J. and R. Greiner (1999). Comparing Bayesian Network Classifiers. *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers: 101-107.

DeAngelo, L. (1986). Accounting Number as Market Valuation Substitutes: A Study of Management Buyouts of Public Shareholders. *The Accounting Review* 61: 400-420.

Dechow, P. M., R.G. Sloan, and A .P. Sweeney. (1995). Detecting Earnings Management. *The Accounting Review* 70: 193-225.

Dechow, P. M., R. G. Sloan, and A. P. Sweeny. (1996). Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC. *Contemporary Accounting Research* 13: 1-36.

Domingos, P. and M. Pazzani. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Proc. of the 13th International Conference on Machine Learning:* 105-112.

Dzeroski, S. and B. Zenko. (2002). Is Combining Classifiers Better than Selecting the Best One? *International Conference on Machine Learning (ICML):* 123-130.

Francis, J. and J. Krishnan. (1999). Accounting Accruals and Auditor Reporting Conservatism. *Contemporary Accounting Research* 16: 135-165.

Friedman, N., M. Geiger and M. Goldszmidt. (1997). Bayesian Network Classifiers. *Machine Learning* 29: 131-163.

Geiger, D and D. Heckerman. (1996). Knowledge Representation and Inference in Similarity

Networks and Bayesian Multinets. *Artificial Intelligence* 82: 45-74.

Gemela, J. (2001). Financial Analysis Using Bayesian Networks. *Applied Stochastic Models in Business and Industry* 17: 57-67.

Healy, P. (1985). The Effect of Bonus Schemes on Accounting Decisions. *Journal of Accounting and Economics* 7: 85-107.

Heckermann, D. (1995). A Tutorial on Learning Bayesian Networks. *Technical Report* MSR-TR-95-06. Microsoft Research.

Jensen, F.V. (1996). *An Introduction to Bayesian Networks*. UCL Press, London.

Ji, C. and S. Ma. (1997). Combinations of Weak Classifiers. *IEEE Transactions on Neural Networks*, 8(1): 32-42.

Jones, J. (1991). Earnings Management during Import Relief Investigation. *Journal of Accounting Research* 29: 193-228.

Kevin, P. M. (2001). A Brief Introduction to Graphical Models and Bayesian Networks. *Technical Report*, Department of Computer Science, UC Berkley.

Koller, D. and M. Sahami. (1996). Toward Optimal Feature Selection. *Proc. 13th Inter-national Conf. Machine Learning*: 284-292.

Lipmann, R. P. (1988). An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine* 3, 4: 4-22.

Margaritis, D. and S. Thrun. (1999). Bayesian Network Induction via Local Neighborhoods. *Advances in Neural Information Processing Systems* 12: 505-511.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*, Pearson Prentice Hall, Upper Saddle River, NJ.

Nelson, M. and W. T. Illingworth. (1991). *A Practical Guide to Neural Nets*. Addison-Wesley Publishing Company, Inc.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco.

Quinlan, J. R. (1986), *Induction of Decision Trees*. Machine Learning, 1(1)

Quinlan, J. R.. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Sarkar, S. and R.S. Sriram. (2001). Bayesian Models for Early Warning of Bank Failures. *Management Science* 47, 11: 1, 457-1, 475.

Spirtes, P., C. Glymour and R. Scheines. (1993). *Causation, Prediction, and Search*, New York: Springer-Verlag.

Stice, J. (1991). Using Financial and Market Information to Identify Pre-Engagement Factors Associated with Lawsuits against Auditors. *The Accounting Review* 66: 516-533.

Summers, S. L., and J. T. Sweeney. (1998). Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis. *The Accounting Review* 73: 131-146.

Tsamardinos, I., C. F. Aliferis, and A. Statnikov. (2003). Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations. *The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC: 24-27.

Wong, M. L., S. Y. Lee, and K. S. Leung. (2004). Data Mining of Bayesian Networks Using

Cooperative Coevolution. *Decision Support Systems* 38, 3: 451-472.

Yun, J. O. and Kim, M. H. (2001). A Study of Characteristics of Accounting Information about Substandard Companies Indicated by SSB (Securities Supervisory Board). *Korea Academic Society of Accounting* 6: 45-60.

A Study on the Classification Properties of Firms to be Subject to Accounting Disclosure Reviews and Investigations: Comparison of Bayesian Network, C5.0, and Ensemble Prediction Methods

Kun Chang Lee* · Kwan Choi**

Abstract

As many people is increasingly investing in those firms listed in the stock market, it becomes important than ever to review and evaluate the reliability of audit reports. To secure the reliability, the Korean Financial Supervisory Service has conducted the ADRI (accounting disclosure reviews and investigations) of the audit reports selectively. However, since there is a lack of systematic as well as highly predictive method that can classify the firms subject to ADRI precisely, the Korean Financial Supervisory Service needs more refined classification methods.

This study proposes using an ensemble method which is based on combining three predictive methods such as general Bayesian network (GBN), naive Bayesian network (NBN) and C5.0. Especially, in the process of proposing the ensemble method, we revealed that the Markov Blanket induced from GBN can show the underlying structure hidden in the data in terms of cause-effect relationships between minimal set of relevant variables explaining the target variables. Rigorous experiments with the ADRI sample data ranging from 1990 to 1999 showed that the proposed ensemble method surpasses other methods in terms of prediction accuracy, and it can be used significantly for the purpose of performing ADRI activities very systematically. In addition, we found the usefulness of the Markov Blanket obtained from GBN in explaining how a certain variable affects the target variables

* Professor, School of Business Administration, Sungkyunkwan University
** Professor, School of Business Administration, Sungkyunkwan University

through the causal relationships with other related variables.

Basically, this study is based on using Bayesian network as a main vehicle of developing an ensemble method to deal with the ADRI problem. Therefore, let us explain the meaning of Bayesian Network. BN or Bayesian Network is denoted as $B=\langle G, P\rangle$, is a directed acyclic graph (DAG) G with a set of conditional probability distributions P, that satisfies the *Markov condition*. Given its parents, each node is conditionally independent of the set of all of its non-descendants. Bayesian network classifiers theoretically seek an optimal error rate based on posteriori probabilities. If all child nodes were relevant to the root node (target concept) and conditionally independent of each other given the root, the naïve Bayesian network (NBN) would be the optimal classifier. However, in real-world applications its performance can be degraded since variables are often related to each other. The tree augmented naïve Bayesian network (TAN) introduced by Friedman et al. (1997) relaxes the unrealistic independence assumption of NBN by allowing arcs between the child nodes. The GBN can be viewed as a generalized form of TAN, where the nodes are allowed to form an arbitrary graph (rather than just a tree) and each child node need not be connected to the class node. Hence, the GBN is clearly more suitable to finding an underlying structure, since it can reveal true interdependencies among all variables. In this paper, we use the GBN structure as the basis from which to reveal the underlying structure of churn behavior. Its two byproducts, the Markov blanket and evidential beliefs, are then utilized for assisting complex models.

The proposed method for solving ADRI problem is based on the selection of salient features by Markov Blanket (MB) of GBN. The Markov blanket introduced here has four major advantages over traditional feature selection methods. First, it provides an explicit criterion for selecting salient features. Second, it can identify redundant features. Third, it can identify nonlinear interdependencies among features of interest, allowing the underlying structure of a problem to be found. Fourth, it produces a minimal set of variables required for learning a target concept, ultimately reducing the costs for collecting and maintaining variables. The concept of the Markov blanket was first introduced by Pearl in 1988, but has recently received renewed attention in the areas of Bayesian learning (Cheng and Greiner 1999, Margaritis and Thrun 1999) and features selection (Koller and Sahami 1996, Tsamardinos et al. 2003). The probabilistic nature of the Markov blanket can be explained using the concept of *d-separation* (*direction dependent separation*) (Pearl, 1988), a graphical

criterion related to the blocking of information flow among variables. In a faithful A DAG G is *faithful* to a joint probability distribution P over set of variables if and only if every conditional independency entailed by G is also present in P (Sprites et al. 1993). Bayesian network, *d-separation* captures all of the conditional independence relationships encoded in the network. If all variables in the Markov blanket of a node are instantiated, then we can say that the node is *d-separated* from the rest of the network.

When the Bayesian network is faithful, MB(T) probabilistically shields T from the rest of the variables. Given this property, knowledge about the variables belonging to MB(T) is enough to determine the probability distribution of T, thus rendering all other variables superfluous. In other words, MB(T) is a minimal feature subset required to predict T, which graphically corresponds to a set neighborhood of T: its parents, its children and the other parents of its children. It is straightforward that this set d-separates T from the set of all other nodes. See Neapolitan (2004) for the proof.

In conclusion, we used the Markov blanket here to show the underlying structure of ADRI problem under consideration. On the other hand, the ensemble prediction method was proposed to predict the ADRI better than the existing methods like GBN, NBN, C5.0, and even logistic regression. Rigorous experiment with real ADRI data showed that the proposed MB administered by GBN can reveal the underlying structure of the ADRI problem, and the proposed ensemble prediction method can yield more accurate ADRI prediction results.

Key words: Accounting disclosure reviews and investigations, Bayesian network, Markov Blanket, C5.0, Ensemble method