Executive Report(2009-10)

효율적인 비리적발을 위한
데이터마이닝 기법의 감사활용

요약

□ 사회복지, 정부조달, 조세분야 등 다수를 대상으로 소액을 지원 또는 징수하는 '소액다수사업'은 대형사업과 다른 특성을 가지고 있으므로 이에 적합한 감사방법론을 개발할 필요가 있음

□ 민간부문과 선진국에서는 이미 다양한 과학적 기법들을 적극 활용 중인데 특히 데이터마이닝이 주목을 받고 있으며, 이에 데이터마이닝의 개념 및 기법, 감사적용시 고려사항, 감사적용절차, 그리고 적용의 한계 등에 대해 체계적으로 소개하고 연구하는 것은 가치가 있음

□ 데이터마이닝은 통계분석이나 모델링 등을 통해 대용량의 데이터에서 감추어진 패턴이나 관계 등을 발견하는 절차로써, 통신, 신용카드, 보험, 소매, 대테러 등의 분야에서 성과개선이나 특이집단 식별에 주로 사용되고 있음

□ 데이터마이닝을 감사에 적용할 경우
- 첫째, 데이터마이닝은 많은 시간이 소요되고 기존자료가 없거나 부실한 경우 효과적이므로 이러한 실행가능성과 적합성을 고려하여 분석대상을 적절히 선정해야 하며,
- 둘째, 목표변수나 대리목표변수 유무에 따라 세 가지 분석모형 중 하나를 선택해야 하며,
- 셋째, 데이터마이닝의 감사적용은 엄밀한 이론이나 모델을 수립하는 것이 아니라 최종확인을 위한 사전 리스트를 도출하는 데 있다는 점을 명심해야 함

□ 데이터마이닝의 적용은 문제정의단계 → 데이터처리단계 → 모델링 및 평가 단계 → 적용단계 등을 통해 이루어지는데, 여기서는 각각의 단계에서 무슨 내용이 어떻게 진행되어야 하는지를 자세히 설명하였으며, 특히 실제 국민연금 관련 자료를 이용한 두 가지 사례분석을 제시함으로써 현실 설명력을 높이고자 하였음

□ 향후 데이터마이닝의 감사활용도를 높이기 위해서는 일차적으로 인력, 교육 등 기본 인프라 구축과 시범적용을 통한 유용성 제고 등이 필요하며 이에 기반한 점진적인 확대가 요망

Executive Report Series

문의 : **감사연구원 연구부**

김찬수(02-2011-3073)

차경엽(02-2011-3054)

2009년 8월

목차

Ⅰ 연구 배경

◎ 사회복지, 정부조달, 조세분야 등 다수를 대상으로 소액을 지원 또는 징수하는 '소액다수사업'에 적합한 감사방법론을 개발할 필요가 있음
◎ 민간부문과 선진국에서는 데이터마이닝 등 이에 적합한 방법론을 이미 사용하고 있는데, 우리 院도 내 · 외부 감사환경 변화를 고려할 때 이에 대한 소개와 감사적용 방법에 대한 연구가 필요함

□ 사회복지, 정부조달, 조세 등 다수를 대상으로 소액을 지원 · 징수하는 '少額多數 사업'에 적합한 감사방법론 개발 필요

○ 대형(국책)사업을 대상으로 개발된 비리적발 감사방법론(심층점검, 프로젝트 위험관리 등)을 소액다수사업에 적용하는 데에는 한계

사업 유형	예 시	감사방법론
대형국책사업	대형 SOC 투자 등	심층점검, 프로젝트 관리 등
소액다수사업	사회복지, 정부조달, 조세 등	-

- 수만 건에서 많게는 수천만 건이 넘는 대상을 일일이 심층점검하거나 프로젝트처럼 개별 관리하는 것은 시간과 비용 측면에서 비효율적

○ 소액다수사업의 특징은 축적된 대용량의 데이터이며, 이를 잘 활용할 수 있는 과학적인 기법을 개발 · 적용하는 것이 중요

□ 민간부문과 선진국에서는 이미 이러한 분야에 적합한 다양한 과학적 기법들을 적극적으로 활용 중

○ (민간부문) 통신, 신용카드, 보험 등 대규모 고객자료가 축적된 분야를 중심으로 마케팅, 고객성향분석, 부정탐지, 과다청구적발 등을 위해 데이터마이닝(data mining)이라는 통계기법을 활용 중

○ (선진국) 미국 연방정부는 성과개선, 테러리스트 탐지, 부정적발 등을 위해

데이터마이닝 기법을 활용 중 (미국감사원 조사에 따르면, 128개 연방기관 중 51개 기관에서 데이터마이닝 사용 중(GAO, 2004))

데이터마이닝과 데이터매칭의 차이 (비리집단 적발의 경우)

- 데이터매칭(data matching)은 둘 이상의 자료를 서로 비교하여 일치나 불일치를 확인하는 방법으로 이를 통해 비리를 곧바로 확인할 수 있는데, 이 방법론의 관건은 비교할 수 있는 자료의 존재(existence)와 이용가능성(availability)

- 반면 데이터마이닝은 비리를 곧바로 확인할 수 있는 자료가 없다는 데에서 출발하며, 모델링(modelling)을 통해 비리집단의 특성이나 패턴을 도출하고 이에 입각하여 비리가능성이 높은 집단을 추출하는 방법론

□ 데이터 분석기법의 발전단계에 비추어 보았을 때 우리 院도 데이터마이닝에 관심을 가져야 할 시기

○ 데이터분석기법은 데이터점검 ⇨ 데이터매칭 ⇨ 데이터마이닝 순으로 발전하는데 현재 우리 院은 데이터매칭 단계

분석방법	내 용	우리나라	미국
데이터 점검	입력오류나 이상치 점검, 중복확인 등	자체감사기구	
데이터 매칭	기존자료와 일치 또는 불일치 확인	감사원	자체감사기구 *
데이터마이닝	모델링을 통한 패턴이나 이상치 발견	민간	민간, GAO

* 자체감사기구 감사관을 대상으로 데이터마이닝에 대한 교육이 활성화

- 최근 일부 지방자치단체에서는 데이터매칭을 감사에 활용하기 시작했으며, 일부 공공기관에서는 이미 데이터마이닝 체계도 구축 · 활용 중

⇨ 데이터마이닝의 개념 및 기법, 감사 적용 방법론 및 절차, 적용 사례, 적용의 의미와 한계 등에 대한 체계적인 소개와 접근 필요

Ⅱ 데이터마이닝 개관

1. 개념 및 절차

□ 개념

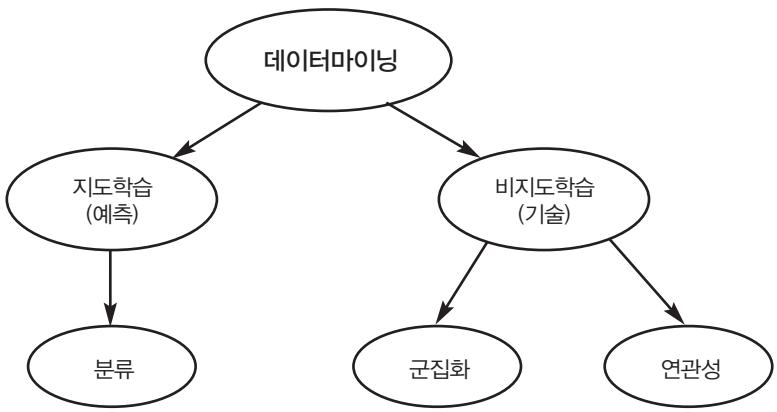
○ 통계분석과 모델링 등의 분석기법(method)을 이용하여 **대용량의 데이터에서 감추어진 패턴이나 미묘한 관계** 등 **유용한 지식**을 발견하는 절차 (GAO, 2004)

– 데이터마이닝은 기본적으로 대용량 데이터에 적합한 분석방법론

※ 기존의 분석방법론은 대용량 데이터 처리에서 한계를 드러냈음

○ 데이터마이닝에서의 지식 구분

기 준	내 용
학습방법	지도학습 vs. 비지도학습
사용목적	예측 vs. 기술
유형	분류, 군집화, 연관성



– 지식은 **학습방법**, 즉 어떻게 지식을 학습하는가에 따라 지도학습(supervised learning)과 非지도학습(unsupervised learning)으로 구분

※ 지도학습은 아이가 부모라는 지도자(supervisor)에 의해 지식을 학습하는 것처럼 데이터마이닝에서는 분석하고자 하는 특정한 목표변수(target variable)가 존재할 경우 이를 지도학습이라 하며, 그렇지 않을 경우 非지도학습이라 함.

– 지식은 **사용목적**에 따라 미래 '예측'(prediction)과 현재패턴 '기술'(description)로 구분하는데, 지도학습은 예측, 非지도학습은 기술과 관련

– 지식은 **유형**에 따라 분류(classification), 군집화(clustering), 연관성(association) 등 3가지로 구분

유형	내 용	예 시
분류	목표변수에 의해 미리 나뉘어진 집단들의 특성을 추출하고 이를 통해 미래를 예측	신용불량자와 우량자의 특성을 도출하고 특정인의 불량가능성 예측
군집화	미리 정의된 집단이 없는 상태에서 유사성을 가진 몇개의 클러스터를 만드는 것	고객을 소비액, 지역, 연령 등 몇 가지 속성으로 세분화
연관성	비슷한 상품을 찾아내는 기법	필기구와 공책 구입의 관계

○ 분석 기법

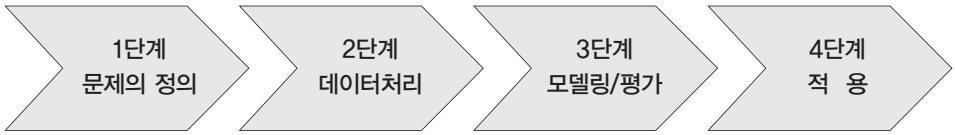
– 분류 : 회귀분석, 신경망, 의사결정나무 등

– 군집화 : K-means 클러스터링

– 연관성 : 연관성 규칙

□ 데이터마이닝 절차

○ 지식발견과정으로서의 데이터마이닝



– 문제정의 단계 : 분석목적 및 대상을 선정하고 전체계획을 수립

– 데이터처리 단계 : 데이터수집, 품질점검, 샘플링, 데이터변환, 데이터분할 등을 통해 분석에 맞도록 데이터를 준비

– 모델링/평가 단계 : 여러 가지 분석기법들을 모델링하여 실행하고, 평가를 통해 이 중 설명력이나 예측력이 제일 높은 모델을 선정

– 적용 단계 : 선정된 모델을 관심있는 새로운 데이터에 적용

○ 실제 데이터마이닝 수행은 다양한 통계패키지 프로그램에서 지원하므로 (SAS의 E*Miner, SPSS의 클레멘타인 등) 이를 이용하면 됨

2. 데이터마이닝의 활용

○ 데이터마이닝은 **성과개선**(performance improvement)과 **특이집단 식별**(anomaly detection)이라는 두 가지 목적으로 주로 활용

– **(성과개선)** ① 맞춤형 사업 · 정책 : 군집화(clustering)를 통해 대상(예를 들면, 고객)을 몇 개의 그룹으로 적절히 나누고 각각에 대해 차별적인 마케팅이나 상품조합을 제시할 경우 성과개선 가능
② 연관성(association)을 적절히 이용할 경우 동일한 효과 가능

예) 타겟 마케팅, Amazon.com의 유사도서 추천 시스템 등

– **(특이집단 식별)** ① 분류(classification)를 통해 정상집단과 특이집단의 특징을 도출하고 특정대상이 특이집단이 될 가능성 예측

② 군집화 등의 방법을 이용하여 정상적인 집단에서 멀리 떨어져 있는 특이한 개체나 집단을 식별

예) 신용카드 연체자, 금융사기거래, 부정수급자, 테러리스트 등의 식별과 예측

※ 여기서 특이집단은 정상집단에 비해 그 수가 상대적으로 적다는 의미일 뿐 그 속성이 반드시 부정적인 것(사기, 연체, 부정 등)은 아님. 예를 들면 벤치마킹을 위해 '크게 성공한 사람'의 특징을 도출하고 성공 가능성을 예측할 수도 있음

미국 연방기관의 데이터마이닝 활용 현황(GAO, 2004)

감사원(GAO) : 정부부처의 부정, 낭비, 남용에 대한 감사
국방부 : 서비스 및 성과개선, 인적자원 관리(데이터마이닝을 가장 많이 사용)
국토안전부 및 법무부 : 테러리스트 활동분석, 정보분석
연금관리부처 : 부정수급, 낭비 등 비리 탐지
국세청 : 탈세자 탐지 및 탈세액 추정
중소기업청 : 기업에 대한 자금대출 모니터링

Ⅲ 감사적용시 고려사항

□ 분석대상 선정

○ **실행가능성과 적합성**을 **고려**하여 분석대상 선정

– 데이터마이닝은 대용량 자료수집, 자료처리, 모델링 등에 시간이 많이 소요되기 때문에 우선순위를 고려하여 몇 가지로 분석대상을 한정

– 데이터마이닝은 모델링을 통해 패턴을 발견하는 것이기 때문에 객관적 자료와의 매칭을 통해 분석목적을 달성할 수 있다면 불필요

소득이 일정수준 이하일 경우에만 지원하는 정부사업에서
소득관련 부정수급자 적발 (예)

핵심은 신고소득의 정확성, 달리 말하면 **거짓보고**(false reporting) 여부인데,

① 소득을 비교 검증할 수 있는 객관적 자료가 존재하고 이용 가능한 경우 :
⇨ 객관적 자료와의 비교를 통한 부정수급자 적발 **(데이터매칭)**

② 객관적인 자료가 없거나 이용 불가능한 경우 :
⇨ 여러 요인(직업, 경력, 학력 등)을 이용해서 소득결정모형을 구축하고 소득을 추정한 이후, 이 '추정소득과 신고소득의 차이가 큰 집단'을 거짓보고 가능성이 높은 집단으로 인식 **(데이터마이닝)**

➞ 데이터마이닝은 기존 자료가 없거나 부실한 경우 또는 그러한 자료를 이용하는데 시간과 비용이 많이 소요되는 경우에 적합

➞ 감사에 데이터마이닝을 적용하는 목적은 '성과개선' 보다는 '특이집단 식별'인 경우가 더 많을 것임

□ 분석모형 설정

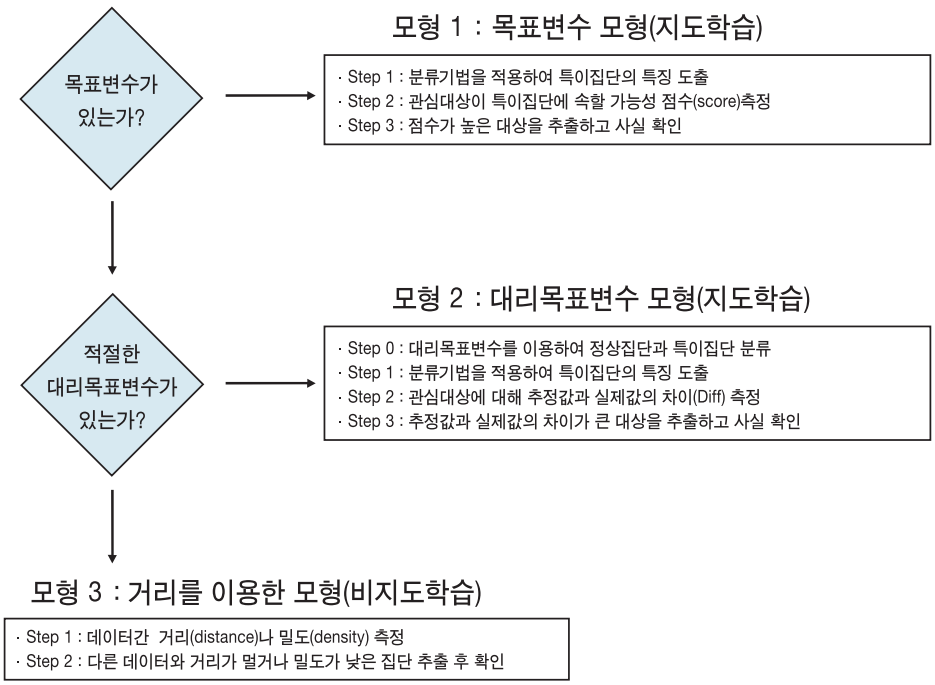
○ '특이집단 식별'을 위한 3가지 분석모형

– 특이집단 여부를 나타내는 목표변수가 있을 경우 : 〈모형 1〉 적용

– 목표변수가 없을 경우 : 적절한 대리목표변수가 있을 경우에는 〈모형 2〉를 적용하고, 그렇지 않을 경우 〈모형 3〉을 적용

※ 농업보조금 과다수급자를 적발하는 상황에서 목표변수인 과다수급자 여부 자료가 없을 경우, 만약 신고소득이 높을수록 수급액이 많아진다면 과다소득신고 여부가 과다수급 여부에 대한 대리목표변수가 될 수 있음

※ 적절한 대리목표변수를 찾는 것은 이론의 영역이라기보다는 경험과 직관의 영역



1) 목표변수는 특이집단 여부와 같은 이산적(discrete) 값뿐만 아니라 연속적인(continuous)인 값일 수도 있다. 예를 들면 목표변수가 소득대비 지출액 비중일 경우 이 값은 연속적이다.

○ **〈모형 1 : 목표변수 모형(지도학습)〉**

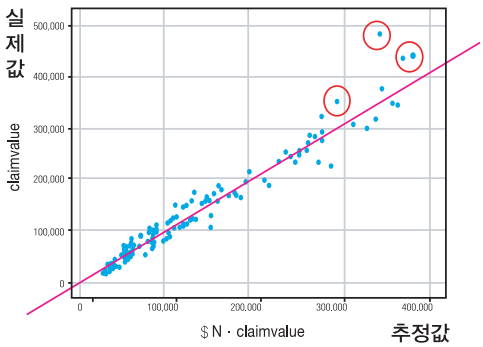
– 통상의 지도학습모형을 그대로 적용하면 되는데, ① 분류기법을 이용하여 특이집단의 특성을 도출하고 ② 새로운 데이터에 적용하여 **특이집단에 속할 가능성을 점수화(scoring)**한 이후 ③ 점수가 큰 집단을 추출

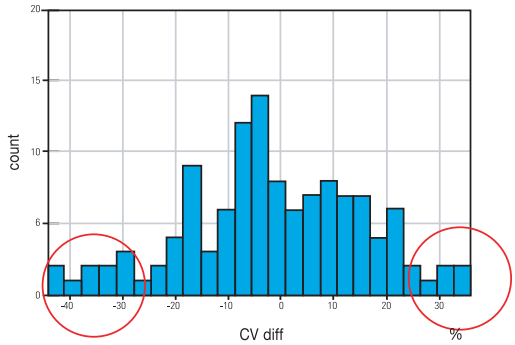
○ **〈모형 2 : 대리목표변수 모형(지도학습)〉**

– 대리목표변수를 이용한 지도학습모형인데, 통상 대리목표변수가 연속형 변수이기 때문에 ① 모델링을 통해 대리목표변수의 추정값을 구하고, ②**추정값과 실제값의 상대적 차이**를 계산하고 ③ 그 값이 큰 집단을 추출

※ 추정값과 실제값의 추세선에서 멀리 떨어진 값들이 특이집단 가능성 높음

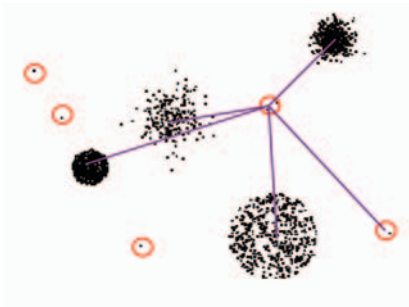
※ 상대적 차이(CV.diff) = (추정값–실제값)/추정값





○ **〈모형 3 : 거리를 이용한 모형(비지도학습)〉**

– ① 데이터 사이의 거리나 밀도를 측정한 이후 ② 서로 유사성이 높은(거리가 가깝거나 밀도가 높은) 데이터들은 클러스터로 묶고, ③ 여기서 **멀리 떨어진 데이터**를 특이집단으로 추출



□ 모델평가와 적용의 관계

○ 일반적인 데이터마이닝에서 모델평가는 그 자체로(즉 새로운 자료에 적용하는 것과 독립적으로) 상당히 중요한 의미를 가짐

– 여러 가지 모델들 중에서 설명력이나 예측력이 가장 좋은 모델을 찾아내는 것은 이론적 엄밀성 측면에서 의미가 큼

○ **그러나 데이터마이닝을 감사에 적용하는 경우, 모델평가는 그 의미가 제한적이며 적용에 종속되는 성격으로 전환**

– 감사에 데이터마이닝을 적용하는 목적은 이론적으로 엄밀한 모델을 도출하는데 있는 것이 아니라, **실제적용을 통해 문제발생 가능성이 높은 데이터(audit short list)**를 추출하는 데 있으며, 이는 전체 데이터를 모두 점검하는 것에 비해 감사의 효율성을 높이는 효과를 가짐

– 따라서 **모델평가에 과도한 노력을 기울일 필요없이,** 모든 모델, 또는 설명력이 아주 낮은 것을 제외한 나머지 모델들을 **모두 실제 데이터에 적용**해보고 문제발생가능성이 높은 집단을 추출한 이후 서로 교차점검(cross-check)하는 것이 보다 감사효율성을 높일 수 있을 것임

※ 또한 실제 모델평가를 해보면, 대부분 모델평가에 사용되는 값이 서로 큰 차이가 없는 경우가 다수

☞ 데이터마이닝의 목적은 어디까지나 최종확인을 위한 리스트를 도출하는데 있는 것이지 그 자체로 마무리되는 작업이 아님을 분명히 이해하여야 함

Ⅳ 데이터마이닝의 감사적용 절차

1) 1단계 : 문제정의 단계

□ 대상기관의 현황파악 및 대책이해

○ 기관의 비리현황 파악 및 비리취약분야 발굴

– 비리를 유형화하고 유형별 발생빈도 파악

※ 특히 자체감사기구의 감사결과를 적극 활용

※ 위험평가 등을 통해 비리에 쉽게 노출되고 파급효과가 큰 비리취약분야 발굴

○ 기관의 비리관리체계 이해

– 초점(예방, 적발, 억제)과 관리주체(해당기관, 감독기관)에 따라 분류

〈사 례 : 국민연금공단의 부당이득 환수〉

◈ 2008년 부당이득 환수현황을 보면 총 27,834건에 206억 원으로 확인

– 건수 기준으로는 부양가족 미해당, 자격징수내역변경, 소득정지 순으로 많았고, 금액기준으로는 소득정지, 자격징수내역변경, 수급권취소 등이 많았음

유형	건수	비중(%)	금액(백만 원)	비중(%)
부양가족 미해당	13,238	47.6	1,043	5.1
소멸	3,438	12.4	1,967	9.5
소득정지	4,341	15.6	8,806	42.7
수급권취소〈원천삭제〉	791	2.8	3,614	17.5
자격징수내역변경 등	5,727	20.6	4,683	22.7
손해배상금수령 등	299	1.1	524	2.5
합계	27,834		20,637	

◈ 연금공단은 외부자료 활용과 수급자 관리를 통해 부당이득을 환수

– 외부자료는 타 기관으로부터 정기적으로 받아 확인하고 있으나, 수급자 관리는 지점/관리자별로 이루어지고 있어 체계적 관리가 미흡한 실정

□ 분석대상 선정

○ 실행가능성과 적합성을 고려하여 분석대상 선정

– 데이터점검이나 데이터매칭으로 해결하기 어려운 과제 중 선정

□ 분석모형 설정

○ 목표변수나 대리목표변수 등을 고려하여 〈목표변수모형〉, 〈대리목표변수모형〉, 〈거리를 이용한 모형〉 중 어떤 것을 선택할 것인지 결정

– 특히 목표변수가 없을 경우 무엇을 대리목표변수로 설정할 것인가에 대해 여러 가지 각도에서 다양한 시나리오 작성 필요

□ 설명요인 도출

○ 목표변수나 대리목표변수를 설명해 줄 수 있는 설명변수 나열

– 기존문헌, 논리적 추론, 대상기관 인터뷰 등을 최대한 활용

〈사 례 1 : 손해배상금 불성실 신고를 통한 연금 부당청구〉

◈ 분석대상

– 부당이득 환수는 대부분 데이터점검이나 데이터매칭을 통해 확인 가능

– 손해배상금 불성실 신고란 장애 · 유족연금 수급자가 손해배상금을 수령했음에도 불구하고 신고하지 않거나 감액신고하는 것이며 적발될 경우 부당이득 환수조치가 취해짐

– 문제는 손해배상금 수령여부를 위해서는 민간보험사의 자료가 필요하나 개인정보 보호 등 때문에 확보가 어렵다는 데 있음

⇨ **데이터마이닝을 통해 부당이득 환수조치를 받은 집단의 특성을 도출하고, 나머지 집단의 불성실 신고가능성 점수를 측정하고 점수가 큰 집단 추출**

◈ 분석모형

– 손해배상금 불성실 신고로 인한 부당이득 환수조치를 받은 집단과 그렇지 않은 집단을 구별하는 변수가 있으므로 〈모형1〉 목표변수 모형을 적용

– 목표변수 : 손해배상금 불성실 신고로 인한 부당이득 환수조치 여부

◈ 설명요인

구분	연속형 변수	범주형 변수
인적속성	· 수급 전년도 평균소득 월액 · 평균 가계소득 대비 개인소득 등	· 부양가족 수
연금가입	· 가입기간 연금지급율 · 평균보험료 등급, 가입기간	· 가입형태, 지역가입 경력 · 사업장가입 경력 등
급여		· 산재등 타급여 가입형태 · 타급여 대상 여부
사고내역	· 사고비용, 일실이익	· 장애등급, 손해배상 수령 여부

〈사 례 2 : 국민연금 지역가입자의 신고소득 적정성 점검〉

◈ 분석대상

– 연금공단은 매해 신고소득과 국세청 소득자료를 비교하여 신고소득이 과세 소득보다 낮은 사람들에 대해 직권조정을 통해 보험료를 재산정하고 있음

– 그런데 지역가입 대상자 중 절반가량이 국세청 과세소득자료가 없어 신고소득의 적정성을 확인하지 못하고 있는 실정

⇨ **데이터마이닝을 통해 소득자료가 없는 지역가입자의 신고소득 적정성 확인**

⇨ **즉, 국세청 자료가 있는 데이터를 이용하여 과세소득결정모형을 수립하고, 이에 기반하여 소득자료가 없는 가입자의 소득을 추정하고, 신고소득이 추정소득에 비해 과도하게 작은 집단을 과소소득 신고집단으로 추출**

◈ 분석모형

– 소득 과소신고 여부를 통해 보험료 과소납부 여부를 점검하는 것이기 때문에 〈**모형2**〉 **대리목표변수 모형**을 적용

– 대리목표변수 : 국세청 과세소득

◈ 설명요인

– 사업자등록증이 있는 지역가입자의 소득에 영향을 미치는 요인들

구분	연속형 변수	범주형 변수
인적속성	· 연령	· 성, 학력, 경력 등
기업특성	· 규모, 설립기간 등	
산업특성		· 산업분류코드
연관변수	· 공시지가, 재산 등	

2) 2단계 : 데이터처리 단계

□ 데이터 접근(access)

○ 필요 데이터 목록 및 항목 등 현황 파악

– 정보시스템관리자와 협의하여 내 · 외부적으로 수집가능한 DB 목록 파악

○ 데이터 가용성, 정확성, 적시성, 수집비용 등을 고려하여 자료수집

– 데이터 수집은 **자료수집계획서**를 작성하여 체계적으로 진행

변수명	데이터 관리 현황		데이터 유형	기타
	출처	담당자		
과세소득	국세청	김△△	연속형	
산업분류코드	국세청	김△△	범주형	
공시지가	국토해양부	이△△	연속형	
지역가입자 여부	국민연금공단	박△△	범주형	

○ 최종적으로 수집된 데이터 및 변수 정리

– 특히 데이터들을 서로 연결 및 통합시킬 수 있는 변수 확인

※ 아래 〈사례〉에서는 내부관리번호, 증서번호, 주민번호 등이 매개변수 역할

구분	연속형 변수	범주형 변수
자격	가입자 기본	내부관리번호, 주민번호, 지역 · 임의 · 사업자 가입경력 등
	사업장 가입자	내부관리번호, 사업장 기호, 취득사유, 등급, 소득월액 등
	지역 가입자	내부관리번호, 취득일자, 가입종별, 등급, 소득월액 등
징수	연금보험료 납입	내부관리번호, 주민번호, 고지월수, 수납월수, 미납월수 등
급여	수급권기본사항	주민번호, 급여종류, 급여세부종류, 장애등급 등
	장애연금	증서번호, 상병구분, 타급여 가입여부, 제3자 가해여부 등
	유족연금	증서번호, 타급여 가입여부, 타급여 대상여부 등
	제3자 가해이력	증서번호, 합의금, 손해배상금수령여부, 일실이익 등
	연금최종산정내역	증서번호, 가입기간 연금지급율, 최종5년간 평균소득액 등
부당이득	환수결정내역	증서번호, 처리연도, 환수결정사유, 발생사유 등

□ **데이터 정제**(cleaning)

○ 결측값(missing value)이나 이상치(outlier)를 파악하고 처리

□ **표본추출**(sampling)

○ 모집단의 규모가 너무 클 경우 분석의 편의성을 위해 표본추출

– 주로 임의추출(random sampling)이나 층화추출(stratified sampling) 적용

※ 층화추출이란 모집단을 특성에 따라 몇 개의 그룹으로 나눈 후 그 그룹내에서는 임의추출하는 방법으로 각 그룹의 구성비율 만큼 샘플을 추출하는 방법

□ **데이터 통합 및 변환**

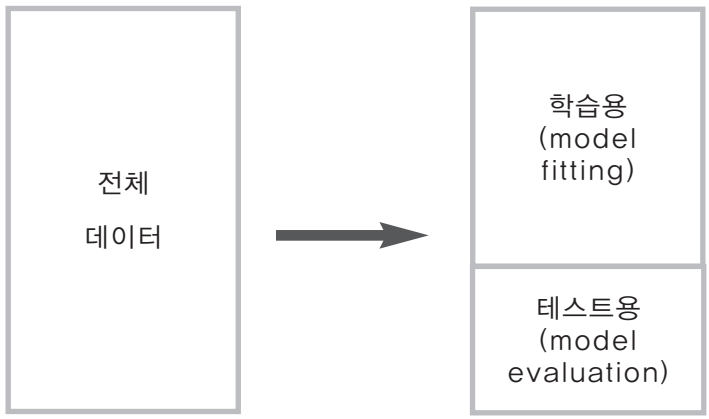
○ 매개변수를 활용하여 흩어져 있는 여러 가지 데이터를 통합

○ 변수를 분석모형에 적합하도록 변환 : 집계, 일반화, 정규화 등

□ **데이터 분할**(partitioning)

○ 전체 데이터를 모형적합을 위한 학습용 데이터(training data)와 모형을 테스트할 테스트용 데이터(test data)로 분할

– 이 경우 한 데이터가 두 가지 용도로 동시에 사용되지 않도록 유의



3) 3단계 : 모델링 및 평가 단계

□ 모델 실행

○ 여러가지 모델을 학습용 데이터에 실행하고 각각 모델최적화

- 모델최적화란 변수의 유의성을 고려하여 모델의 설명력을 극대화하기 위해 어떤 변수를 포함하고 제거할 것인지 결정하는 지속적인 과정

- 이는 변수의 형태를 다양하게 변환시키는 것도 포함하는데, 예를 들면 변수를 제곱하거나 다른 변수와 교차하거나 로그변환 등이 대표적

- 또한 필요에 따라서는 새로운 변수를 생성하기도 함

➡ 모델최적화는 상당한 시행착오와 노력을 수반하는 과정

○ 실제 모델최적화는 변수가 많은 경우 변수별 검정과 다변량 검정을 순차적으로 거치고, 그렇지 않을 경우 주로 다변량 검정만 수행

구분	변수결정 방법
인적속성	변수별 검정 → 다변량분석방법 검정
기업특성	다변량분석방법 검정

- 변수별 검정이란 특정변수의 유의성을 검정할 때, 다른 변수들을 고려하지 않고 그 변수와 목표변수의 관계를 통해 검정하는 방법 : x^2검정, T검정 등

	유형	유의성 검정방법
지도학습	범주형	x^2검정, 로그선형모형 등
	연속형	T검정, 로짓모형의 검정(우도비, Wald, 일치도) 등
비지도 학습	-	F검정, 상관계수 등

- 다변량분석방법 검정은 전체 모형의 설명력이 어떻게 변화하는 가에 따라 변수를 선택하는 방법 : 전진선택법(추가), 후진소거법(제거), 단계적 방법 등

※ 데이터마이닝 통계프로그램에서 이와 관련된 다양한 옵션을 지원

〈 사례 : 손해배상금 불성실 신고 탐지 모형에서 검정 〉

◈ 설명변수를 범주형과 연속형 변수로 구분하여 검정

- 범주형 변수는 검정, 연속형 변수는 Wald 이용
- 부당이득 환수는 대부분 데이터점검이나 데이터매칭을 통해 확인 가능

◈ 범주형 변수의 유의성 검정 결과 (1% 유의수준)

- 장애연금에서는 거의 모든 변수(임의가입경력 제외)가 유의하였으며, 유족연금에서는 가입형태, 타급여가입경력, 손해배상금수령여부가 유의

범주형 변수의 유의성 검정 결과

구분	내용	PROB〉| x^2 |	
		장애연금	유족연금
인적속성	부양가족 수	0.0001	0.0571
연금가입	가입형태	0.0001	0.0001
	최종가입형태	0.0005	0.1386
	지역가입경력	0.0053	0.0327
	사업장가입경력	0.0001	0.7920
	임의가입경력	0.6539	0.1799
급여	산재등타급여가입	0.0001	0.0001
	타급여대상여부	0.0001	0.8260
사고내역	장애등급	0.0001	-
	손해배상수령여부	0.0001	0.0001

◈ 연속형 변수의 유의성 검정 결과 (1% 유의수준)

- 장애연금에서는 최종 5년간 평균소득액, 가입기간 연금지급율, 사고비용 등이 유의했으며 장애연금에서는 소득변동액, 사고비용, 일실이익 등이 유의

연속형 변수의 유의성 검정 결과

구분	내용	PROB〉| Wald x^2 |	
		장애연금	유족연금
소득	수급전년도평균소득월액	0.5693	0.0005
	최종5년간 평균소득액	0.0001	0.0002
	소득변동액	0.0004	0.0001
	평균가계소득대비개인소득	0.0001	0.0001
연금가입	가입기간 연금지급율	0.0001	0.0001
	가입기간	0.0001	0.0213
	평균보험료 등급	0.0884	0.5499
사고내역	사고비용	0.0001	0.0001
	일실이익	0.0001	0.0001

□ 모델 평가

○ 각기 최적화된 모델의 성능을 상호 비교 평가

– 학습용 데이터에서 적합된 모델들을 테스트용 데이터에 적용하여 모델들의 성능을 비교하고 가장 우수한 모델 선택

○ 지도학습 모형에서는 **연속형 목표변수에서는 Root ASE, 범주형 목표변수에서는 Lift 도표, ROC 그래프, 오분류율** 등을 이용

지도학습 모형에서 목표변수 유형에 따른 모델평가 방법

목표변수 유형	모델 평가 방법	
	통계량	그래프
연속형	Root ASE*, RMSE*, SBC*, R^2 **등**	
범주형	오분류율*, SBC* 등	Lift, ROC

주 : * 표시가 된 값은 작을수록 모델이 우수하며 R^2는 클수록 모델이 우수.

– 비지도학습모형은 클러스터 중앙(center)과 개별 값들의 거리의 제곱합이 작을수록 모델이 우수한 것으로 평가

○ 종합적인 판단을 통한 최종 모델 선정

– 비교결과 모델들 사이에 성능에 큰 차이가 없거나, 또는 평가방법에 따라 비교결과가 서로 달라질 경우 **반드시 하나의 모델을 최종모델로 선택할 필요는 없으며 여러 가지 모델을 선택**하여 실제로 적용한 이후 공통점을 추출하는 것도 좋은 대안

〈 모델 평가 사례 〉

◈ 연속형 목표변수 (지역가입자 신고소득 적정성에 대한 모델평가)

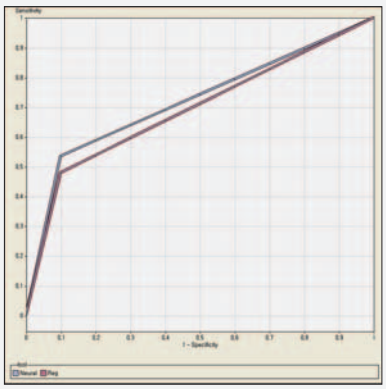
– 비교 결과 신경망 〉 의사결정나무 〉 회귀분석 순. 그러나 모델간 차이는 미미

목표변수 유형	Root ASE	
	학습용 데이터	테스트용 데이터
회귀분석 모델	1.1118	1.1186
의사결정나무 모델	1.0945	1.1149
신경망 모델	1.0986	1.1066

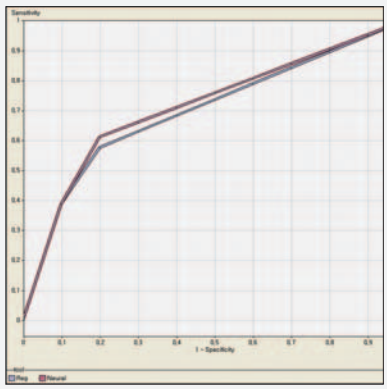
◈ 범주형 변수 (손해배상금 불성실 신고 탐지에 대한 모델평가)

– ROC 곡선이나 Lift 도표를 볼 때 장애연금과 유족연금 모두 신경망 모델이 로지스틱 모델에 비해 더 우수

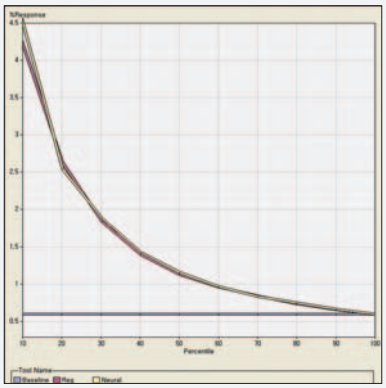
ROC Curve : 장애연금



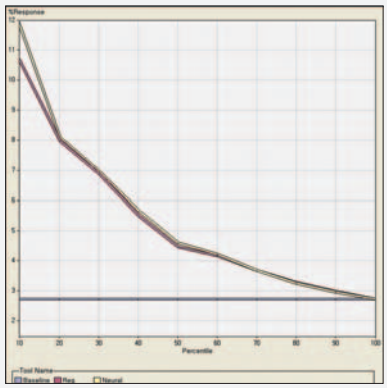
ROC Curve : 유족연금



Lift Chart : 장애연금



Lift Chart : 유족연금



주: Roc 곡선과 Lift 도표 모두 처음에 Y값이 클수록 모델 성능이 좋음

4) 4단계 : 적용 단계

○ **최종 선정된 모델을 관심있는 데이터에 적용**

– 모델 평가를 통해 최종 선정된 모델을 새로운 데이터에 적용

○ 점수화(Scoring)나 '추정값과 실제값의 차이'등을 통해 **특이집단에 속할 가능성을 측정**

– 범주형 목표변수는 점수화, 연속형 변수는 추정값과 실제값의 차이 이용

목표변수 유형	특이집단 소속 가능성	예시
연속형	추정값과 실제값의 차이	$100 \times \frac{(\text{실제값} - \text{추정값})}{\text{추정값}}$
범주형	점수화	$\frac{\exp(\hat{y})}{1+\exp(\hat{y})}$ (로지스틱모형)

○ 특정 기준을 적용하여 **특이집단 가능성이 높은 리스트 추출**

– 검증기간이나 비용 등을 고려하여 적절한 수의 리스트 추출

※ 예를 들면 특이집단 가능성 점수가 높은 상위 1%, 5%, 10% 추출

– 둘 이상의 모델을 적용하였을 경우 각각 특이집단 가능성이 높은 리스트를 추출하되 공통으로 도출되는 집단에 대해서는 특별한 주의가 필요

〈모델 적용 사례〉

◈ 범주형 목표변수 〈불성실 신고 가능성이 높은 상위 10% 추출(일부)〉

INCOME_AMT	BIZ_OPEN_CHK	BIZ2	LAND_AMT	REAL_TAX_AMT	L_INCOME_AMT	PRE_1	RESID_1
300000	1	741	6470000	375991500	12,611537754	14,754841281	-2.143303528
370000	2	505	201000	359600446	12,821258285	14,623751856	-1.802493571
220000	1	940	1240000	84150000	12,301382825	14,042802824	-1.741419999
440000	2	505	59692	1630285913	12,994530006	14,607095379	-1.612565373
730000	3	505	232000	149593600	13,500799813	15,059834157	-1.559034344
250000	1	292	618000	172136510	12,429216197	13,975837296	-1.546621099
670000	1	505	270000	702157140	13,415032991	14,911568085	-1.496535094
220000	1	519	688000	14337659	12,301382825	13,794428789	-1.493045963
570000	2	505	226000	276060200	13,25339164	14,710415488	-1.457023848

◈ 연속형 목표변수 〈허위 소득신고 가능성 높은 지역가입자〉

– 회귀분석모델을 적용한 허위신고가능성이 높은 상위 1%(1,300개)를 추출

허위소득신고 가능성 높은 집단(회귀분석 모델 적용 결과)

LOG_LOST_PROFIT	D22	D31	D51	D52	D61	D62	Predicted: budang=1
16.811242832	0	0	1	0	1	0	0.35942822
18.875659539	0	1	1	0	1	0	0.31620495
15.538277156	0	0	1	0	1	0	0.31339005
18.66811985	0	1	1	0	1	0	0.30897195
18.198537193	0	1	1	0	1	0	0.292917
18.153781968	0	1	1	0	1	0	0.29144963
18.002336196	0	1	1	0	1	0	0.28640309
20.668873521	1	1	1	0	1	0	0.29407115

– 나머지 두 가지 모델에 대해서도 동일하게 상위 1% 추출하였는데, 276개는 3가지 모델에서 모두 추출되었으며, 701개는 2가지 모델에서 공통적으로 추출

허위신고가능성이 높은 집단으로 중복 추출된 횟수

중복 횟수	빈도	비율(%)	누적빈도
3	276	10.43	276
2	701	26.48	977
1	1,670	63.09	2,647
계	2,647		

– 허위신고가능성이 높다고 3번 중복추출된 집단에 대해서는 특별히 주의

허위신고가능성이 높은 집단으로 3회 중복 추출 리스트(일부)

INCOME_AMT	BIZ_OPEN_CHK	BIZ2	AGE	SEX	LAND_AMT	REAL_TAX_AMT	L_LAND_AMT	L_REAL_TAX_AMT	L_INCOME_AMT	NUM
400000	0	701	59	1	970000	3442174320	13.78505135	21.959369178	12.899219826	3
220000	0	552	59	0	7430000	412815690	15.821036417	19.83851189	12.301382825	3
350000	0	701	59	0	979000	1451569948	13.794286912	21.09591153	12.765988433	3
270000	1	701	58	1	299000	283154088	12.608198852	19.461501788	12.506177238	3
300000	1	701	58	0	1770000	311194892	14.385490105	19.555629336	12.611537754	3
340000	1	659	57	1	4040000	132550000	15.211755215	18.702470491	12.736700897	3
370000	1	701	57	1	704000	354956593	13.464533615	19.687506067	12.821256285	3
220000	0	672	57	0	2730000	598815394	14.819812167	20.210466388	12.301382825	3
440000	1	701	57	1	272000	499002502	12.513557345	20.028121668	12.994530006	3
370000	1	701	57	1	48887	1362506387	10.797266791	21.032592139	12.821256285	3

Ⅴ 데이터마이닝의 감사활용 과제

1. 데이터마이닝기법의 용도에 대한 정확한 인식

□ **데이터마이닝은 대용량 데이터 처리를 용이하게 하지만, 이는 어디까지나 감사를 사전적으로 지원하는 기법**

○ 데이터마이닝을 감사에 적용할 경우 데이터마이닝의 최종목적은 '**가장 그럴듯한 리스트를 추출**'하는 것임

○ 그것이 **사실인지 여부**는 데이터마이닝을 통해 해결될 수 없으며, **실제 감사를 통해서만 확인 가능**

2. 데이터마이닝기법 적용을 위한 기반 구축

□ **전문인력 확보**

○ 실제 데이터마이닝을 수행하기 위해서는 상당한 수준의 데이터 분석능력과 경험을 갖춘 전문인력 필요

– 데이터마이닝뿐만 아니라 **최근 합법성감사를 지원하는 기법들**(Undercover Test[1] **등)이 빠르게 발전**하고 있는데, 이를 본격적으로 연구하고 실무에 적용할 수 있는 전문인력 확보계획 필요

1) 가짜(bogus)를 만들어 여권발급이나 정부지원금 등을 신청해보고 이것이 어떻게 처리되는지를 통해 정부행정시스템을 테스트하는 기법으로 GAO가 주로 사용

※ 감사연구원은 하반기에 통계전문가를 충원할 예정

□ 교육을 통한 역량 강화

○ **민간과 선진국 사례에 대한 적극적인 교육**과 벤치마킹 필요

– 신용카드 사기적발, 부정금융거래 적발 등 민간분야와 선진국, 특히 미국에서는 주 · 지방정부 감사관을 대상으로 한 데이터마이닝 교육이 활성화

○ **데이터마이닝 매뉴얼 개발**을 통한 자체교육 시행

– 우리나라의 감사환경에 맞는 데이터마이닝 매뉴얼 개발이 필요

– (가칭) 「**데이터분석기법의 감사적용 방법**」이라는 교육 프로그램을 개발하여 데이터점검, 데이터매칭, 데이터마이닝에 대한 교육실시

3. 데이터마이닝기법의 단계적 적용 확대

□ 시범적용을 통한 유용성 점검

○ 데이터마이닝에 적합한 분야를 선정하여 시범적용을 하고 이를 통해 유용성이 확인될 경우 점진적으로 확대

– 우선적용 가능분야(예시) : 건강 · 실업보험 등 사회복지, 조세징수 분야 등

– 감사원과 감사연구원의 효율적인 역할 분담이 특히 중요

※ 감사원은 경험과 직관을 통한 문제정의와 추출된 리스트에 대한 확인 분야에서, 감사연구원은 데이터 처리와 모델링 및 평가 분야에서 각각 강점 보유

※ 특히 관세, 국세 등 공통 자료의 경우 院 지식관리담당관실과 협조

4. 감사연구원 내 실무지원팀 운영

□ Ad hoc 팀(3–4인) 운영

○ 院에서 데이터마이닝 요청이 있을 경우 이를 실무적으로 지원

※ 실무적으로 볼 때, 본감사 초기에 '리스트 추출'이 어느 정도 마무리 되어 이를 본감사에서 확인할 수 있어야 이상적이기 때문에 이에 맞추어 선행계획을 설계

작 성 자
감사연구원 연구2팀 연구관 김 찬 수
감사연구원 연구3팀 연구관 차 경 엽