*Journal of the Korean Data & Information Science Society*
*2009*, **20**(1), 179–190

한국데이터정보과학회지

# 생명보험사의 개인연금 보험예측 사례를 통해서 본 의사결정나무 분석의 설명변수 축소에 관한 비교 연구[†]

이용구[1] · 허준[2]

[1] 중앙대학교 수학통계학부 · [2]SPSS KOREA (주)데이타솔루션 컨설팅팀

접수 2008년 12월 26일, 수정 2009년 1월 16일, 게재확정 2009년 1월 23일

**요 약**

금융 산업에서, 의사결정나무 분석은 분류분석을 위해서 널리 사용되는 분석기법이다. 그러나 금융 산업에서 실제로 의사결정나무 분석을 적용할 때, 발생하는 문제점 중 하나는 설명변수의 수가 너무 많다는 점이다. 따라서 모형의 결과에 별 영향을 미치지 않으면서 설명변수의 수를 줄이는 효과적인 방법을 연구할 필요가 있다. 본 연구에서는 의사결정 나무 분석에서 모형의 정확성에 근거한 최선의 변수 선택 방법을 구하기 위하여 다양한 변수 선택방법들을 비교 분석 하였다. 이를 위하여 본 연구에서는 한 보험회사의 연금 보험 상품 자료에 다양한 설명변수 축소방법을 적용하여, 가장 적은 수의 설명변수를 가지고 가장 높은 정확도를 제공하여 주는 설명변수 축소방법을 구하는 실증적인 연구를 시행하였다. 이러한 실험결과, 신경망의 민감도 분석을 이용하여 변수를 축소하고, 그 축소된 변수를 이용하여 의사결정나무 분석 모델을 생성하는 경우가 가장 효율적인 설명변수 축소방법임을 알 수 있었다.

주요용어: 변수선택 (축소), 신경망, 연금보험, 요인분석, 의사결정나무 분석.

## 1. 서론

현재 생명보험 시장에서, 가장 화두가 되고 있는 것이 바로 개인연금보험 또는 개인은퇴보험 시장이라고 할 수 있다. 개인연금보험 또는 개인은퇴보험이란 직장에서의 은퇴 또는 일을 더 이상할 수 없는 연령에 처한 사람들에게 일종의 노후의 생계유지를 지원하는 상품으로 우리 사회의 고령화 문제 및 각종 퇴직 후 생활불안 심리와 맞물려서 큰 인기를 끌고 있다. (향후 본 논문에서는 개인연금보험과 개인은퇴보험을 같은 의미로 하여, 개인연금보험으로 지칭하고자 한다.) 이로 인해, 대형 생명보험사들뿐만 아니라, 중소규모의 여러 보험사들이 다양한 개인연금보험 상품들을 내놓고 있으며, 보험일보에 따르면, 현재 본 시장은 향후 더욱 더 급속하게 성장하여, 2010년에는 31조원 규모의 시장이 될 것으로 예측하고 있다. 또한 한국의 개인연금보험 시장의 잠재력을 간파한, 외국계 대형 보험사들도 시장에 눈을 돌리면서, 향후 이 시장에서의 경쟁이 매우 치열해짐을 예고하고 있다. 이렇게 모든 보험사가 개인연금보험의 판매에 총력을 기울이는 시점에서, 개인연금보험을 판매하는 대상은 2가지 종류로 나눌 수 있다. 첫 번째는 신규 고객을 대상으로, 보험설계사 및 홈쇼핑 등의 판매 채널을 통해서, 개인연금보험을 판매하는 것이고, 두 번째는 기존의 다른 보험 가입고객들에게, 추가 판매를 하는 것이다.

[†] 이 논문은 2007년도 중앙대학교 우수 연구자 연구비 지원에 의한 것임.
[1] 교신저자: (156-756) 서울시 동작구 흑석동 221, 중앙대학교 수학통계학부, 교수. E-mail: leeyg@cau.ac.kr
[2] (135-513) 서울시 강남구 역삼동 701-2 삼정개발빌딩, SPSS Korea (주)데이타솔루션 컨설팅팀, 수석연구원.

본 논문의 사례가 되는 생명보험사는 국내 중위권 회사로, 기존 고객들의 개인연금보험으로의 추가 판매 방법을 통해서 타사와 대비되는 CRM 경쟁력을 달성하고자 한다 (Kang, 2004). 기존 고객들에게 개인연금보험 추가 판매를 위해서 사례의 보험사가 선택한 방법으로, 데이터 마이닝의 지도학습 기법 중 의사결정 나무 (Decision Tree Induction)를 이용하여, 기존 다른 보험 상품가입자의 개인연금보험 추가 가입 가능성을 예측하고, 가능성과 불가능성의 규칙을 만드는 것이다. 의사결정나무 분석을 이용하는 것은, 예측 가능성 점수와 함께 동시에, 개인연금보험에 가입 가능성이 높은 규칙을 도출할 수 있어서, 최종적으로 이를 적용하기에 편리하기 때문이다. 그러나 실제로 의사결정나무 기법을 적용함에 있어, 많은 문제점이 나타나게 되는데, 그 중 하나가 많은 수의 설명변수 중 의미 있는 변수를 선택하는 문제이다. 변수 선택이란, 의사결정나무 분석과 같은 지도학습 기법에 사용되는 설명변수들 중에서, 모델에 활용성이 적거나 또는 활용할 때 문제가 되는 변수를 파악하여, 제거하는 것을 의미한다. 실제 금융권의 경우 유사한 성격의 데이터가 매우 많으며, 고객의 다양한 정보를 보유하고 있어, 많은 수의 설명변수를 가지게 된다. 설명변수가 많다는 것은, 고객의 상황을 파악하기에 매우 좋은 정보를 가지고 있다는 의미도 되지만, 다른 한 편으로 보면, 유사한 정보의 중복도 많을 수 있고, 정확도 및 예측력의 향상에는 기여하는 바 없이, 모델링의 시간만 낭비하는 설명변수도 있다는 것을 의미한다. 따라서 모델링 시간 및 분석 자원을 절약하면서, 모델의 정확성 및 예측력은 유지하는 설명변수 선택 방법은 특히 많은 수의 설명변수를 가진 모델에서는 우선적으로 고려해야 하는 필수 단계이기도 하다. 그러나 여기에는 많은 주의가 필요하다. 어떤 방법으로 변수 선택을 하는 것이 옳은지, 그리고 변수 선택이 자칫 잘못되어, 예측력이나 정확도의 문제를 저하시키지는 않는지 등이 대표적인 주의점이다. 그리고 기업의 입장에서는 수집된 정보는 모두 활용하고 싶어 하는 측면도 있어서, 변수 선택을 하는데 더욱 어려움이 따른다. 본 논문에서는 다양한 방법으로 설명변수 데이터를 축소하는 방법들 중 가장 성능이 우수한 방법을 찾기 위해, 비교연구를 수행하는 것이 목적이다. 이를 위해, 본 논문의 구성은 1장의 서론에 이어서, 2장에서는 논문과 관련된 문헌 연구를 수행하였고, 3장에서는 본 논문에서 사용할 데이터의, 설명과 실험을 위한 설계를 정리하였다. 그리고 4장에서는 데이터 마이닝 모델링을 수행한 결과를 수록하고, 마지막 5장에서 결론 및 향후 보완점을 언급하고자 한다.

## 2. 관련연구

변수 선택은 가장 적은 노력을 통해서, 가장 정확한 분석을 할 수 있다는 경제적인 관점에서, 분석을 할 때 매우 중요한 사전 고려 사항이다. 이에 따라 과거에도. 데이터 마이닝 및 각종 통계분석에 있어서, 변수선택에 관련된 다양한 연구들이 있어왔다. 먼저 국내에서 권철신과 홍순욱 (2001)은 범주형 변수를 선택하는데 있어서, 요인분석의 단점을 보완한 유사상관계수라는 개념을 도입하여, 새로운 변수 축약방법을 제시하였으며, 허명회 등 (2008)은 다변량 분석 등에서 효율적으로 사용할 수 있는 주(主)변수 선택방법에 의한 방법론을 개발하여 제시하였다. 또한 허문열과 박영석 (2005)은 결합상호정보 (JMI: Joint Mutual Information)를 사용하여 변수들의 상호 작용이, 고려된 변수 부분 집합들로서 종속 변수에 대하여, 최대 정보량을 나타내는 순서로 변수를 선택하는 동적 모델링을 제안하고, 성능에 대한 연구를 수행하였다. 새로운 변수 선택 방법들에 대한 제안 이외에도, 박성민과 박영준 (2005)은 회귀분석의 변수선택 알고리즘을 이용하여, 인터넷 통신의 네트워크 품질 특성에 필요한 정보만을 추출하는 연구를 수행하였으며, 정석훈과 서용무 (2008)는 Rough Set 이론을 신용카드 연체자를 분류하는 업무에서, 불필요한 속성들을 제거하는 사전 단계로 사용하는 실증적 연구를 수행하기도 하는 등 변수 선택 방법을 실제 비즈니스에 적용하는 연구들이 있었다. 그리고 송문섭과 윤영주 (2001)는 범용 데이터 마이닝 패키지에서 의사결정나무 분석들의 변수 선택에 대한 편의에 대하여 시뮬레이션을 수행하여, 의사결정 나무 분석에서 변수 선택 편의가, 적은 알고리즘을 권장하는 연구를 수행하였다.

국내뿐 아니라 해외에서도, 변수 선택에 관한 많은 연구들이 있었다. 먼저 Battiti (1994)가 상호정보를 이용하여, 신경망 알고리즘을 적용할 때, 이산형 자료에서, 변수를 선택하는 알고리즘을 제안하였으며, Krzanowski (1987, 1996) 등이 후진제거법을 응용한 정지규칙을 이용하여, 주성분이 되는 변수를 선택하는 알고리즘을 제안하고, 이에 대한 실험을 수행한 연구가 있었다. 또한 Lu 등 (1996)은 본 논문과 유사한 방법으로, 신경망 분석을 적용하기 전에 의사결정나무 분석 기법을 적용하여, 의미 있는 변수를 선택한 다음 모델링을 하는 연구를 통해, 신경망 모델의 성능 향상에 대한 연구를 수행하기도 하였다. 이와 유사한 연구로, Anand 등 (1998)도 계층적 일반화 나무모형이라는 방법을 이용하여, 단계적으로 불필요한 정보를 계속적으로, 제거해 나가면서 Rule 기법을 적용한 연구도 있었으며, 그 외에 변수 선택에 관련된 많은 연구가 있었다.

## 3. 사례 데이터 설명과 실험의 설계

### 3.1. 사례 데이터 설명

실험에 사용된 데이터는 사례 보험사의, 서울 지역 고객 77,031명의 데이터이다. 고객들 중에서, 개인연금보험에 가입한 고객은 10.03% (약 10%)인 7,725명이고, 나머지 89.97% (약 90%)의 고객인 69,306명은 개인연금보험이 아닌 다른 보험 상품들에 가입을 한 고객들이다. 분석에 필요한 고객의 선정 조건 중 하나는, 최소한 2006년 12월 29일 이전에 가입한 고객들만을 선택하였다는 것이다. 이는 만약 바로 분석시점으로부터 전월이나 전일에 가입한 고객의 경우 보험료 납입 등에서 행동패턴을 정확하게 파악할 수 없기 때문이다. 고객들의 데이터에 대한 자세한 설명은 표 3.1과 같다.

표 3.1에서 보면, 주피보험자의 ID와 목표변수를 제외하고 총 42개의 설명변수 데이터로 구성이 되어져 있는 것을 알 수 있다. 42개의 설명변수 중 변수유형이 범주형인 것이 7개, 연속형인 것이 35개이다. 연속형인 변수 중에서, 미납차수 합계와 평균과 같이, 만약 1개의 보험 상품만 보유한 경우 거의 유사한 성격을 가지는, 데이터도 많이 있는 것을 알 수 있다. 참고적으로 77,031명의 고객들은 평균 1.6개의 보험 상품에 가입하고 있는 것으로 나타났다.

그림 3.1은 모델링을 하는 과정을 단계적으로 설명하고 있다. 먼저 1단계의 목표변수 균형화는 목표변수의 2개 범주 중 수가 작은 쪽인 연금보험 가입자 수에 맞추어서, 비 가입자를 표본추출 (Sampling)하는 방법인 과소적합 표본추출 (Under Sampling)을 수행하였다 (허준과 김종우, 2007). 따라서, 전체 데이터 수는 표본추출을 할 때 마다, 약간씩 차이가 나겠지만, 약 15,000개 이상의 값을 사용하도록 구성하였다. 2단계의 데이터 분할 (partition) 과정 역시 일반적으로 데이터 마이닝 패키지에서 기본설정으로 정의된 훈련 (training) 데이터와 검증용 (test)데이터의 비율이 80:20이 되도록 설정을 하였다 (SPSS Inc, 2007). 3단계는 본 논문의 핵심인 비교 연구를 위한 과정이므로 다음 절에 상세하게 설명하고자 한다. 다음 4단계에서 의사결정나무 분석 기법 4가지는 과거 관련 연구 (Chung 등, 2005) 등에서 활용된 대표적인 의사결정나무 분석 기법들인 CART (Brieman, 1996), C5.0 (Quinlan, 1993), CHAID (Kass, 1980) 그리고 QUEST (Loh와 Shih, 1997)로 선정을 하고, 각각 실험을 수행하였다. 이렇게 다양한 의사결정나무 분석 기법을 이용한 것은 의사결정나무분석 기법의 알고리즘에 따른 결과의 편의를 최대한 줄이기 위함이다. 마지막 5단계에서 지도학습을 통해서 나온 결과를 검증용 데이터에 적용을 하여, 최종 정확도를 산출하였다.

### 3.2. 설명변수 축소 방법

그림 3.1에서 3단계는, 앞의 1, 2단계에서 기본적으로 정리가 된 데이터 중 설명변수를 선택하는 단계이다. 본 논문의 목적은 이 단계의 축소 변화 (요인 분석을 통한 변수 축약이나 또는 별도의 변수 선

**표 3.1** 사례 데이터 설명

| 번호 | 변수명 | 변수유형 | 비고 |
|---|---|---|---|
| 1 | 주피보험자 ID | ID | 구분자 분석변수 아님 |
| 2 | 주피보험자 연령 | 연속형 | 2008년을 기준으로 현재 연령 |
| 3 | 주피보험자 성별 | 범주형 (이분형) | 남/여 |
| 4 | 주피보험자 직업코드 | 범주형 | 직업코드 (12개 범주) |
| 5 | 주피보험자 서울거주여부 | 범주형 (이분형) | 서울/비서울 |
| 7 | 계약자 주피보험자 동일여부 | 범주형 (이분형) | 예/아니오 |
| 8 | 계약자 연령 | 연속형 | 2008년을 기준으로 현재 연령 |
| 9 | 계약자 성별 | 범주형 (이분형) | 남/여 |
| 10 | 계약자 직업코드 | 범주형 | 직업코드 (12개 범주) |
| 11 | 계약자 서울거주여부 | 범주형 (이분형) | 서울/비서울 |
| 12 | 가입상품수 | 연속형 | |
| 13 | 최근계약일로부터일수 | 연속형 | 최근계약일 - 2008년 1월 1일 |
| 14 | 최장계약일로부터일수 | 연속형 | 최장계약일 - 2008년 1월 1일 기준 |
| 15 | 평균가입주기 (월 단위) | 연속형 | |
| 16 | 유지건수 합계 | 연속형 | 가입상품 중 현재 유지건수 |
| 17 | 제13회이상유지건수합계 | 연속형 | 가입 후 13개월 이상 유지한 상품수 |
| 18 | 제18회이상유지건수합계 | 연속형 | 가입 후 18개월 이상 유지한 상품수 |
| 19 | 제24회이상유지건수합계 | 연속형 | 가입 후 24개월 이상 유지한 상품수 |
| 20 | 실효건수합계 | 연속형 | 보험금을 일정기간 납입치 않아 보험 효력을 상실한 건수의 합 |
| 계 21 | 부활건수합계 | 연속형 | 실효한 상품을 다시 부활한 건수 합계 |
| 22 | 보험금지급건수합계 | 연속형 | 실제 보험금을 지급한 건수의 합계 |
| 23 | 보험금지급금액합계 | 연속형 | |
| 24 | 보험금지급금액평균 | 연속형 | 보험금지급금액합계/가입상품수 |
| 25 | 납입완료거래개월수합계 | 연속형 | |
| 26 | 납입완료거래개월수평균 | 연속형 | 납입완료거래개월수합계/가입상품수 |
| 27 | 월납환산보험료합계 | 연속형 | |
| 28 | 월납환산보험료평균 | 연속형 | 월납환산보험료합계/가입상품수 |
| 29 | 기납입보험료합계 | 연속형 | |
| 30 | 기납입보험료평균 | 연속형 | 기납입보험료합계/가입상품수 |
| 31 | 미납상품수합계 | 연속형 | 미납이 1번이라도 있었던 상품의 수 합계 |
| 32 | 미납차수합계 | 연속형 | |
| 33 | 미납차수평균 | 연속형 | 미납차수합계/가입상품수 |
| 34 | 미납보험료합계 | 연속형 | |
| 35 | 미납보험료평균 | 연속형 | |
| 36 | 보험계약대출보유건수합계 | 연속형 | 대출건수 |
| 37 | 보험계약대출납입금액합계 | 연속형 | |
| 38 | 보험계약대출납입금액평균 | 연속형 | 보험계약대출납입금액합계/대출건수 |
| 39 | 보험계약대출잔액합계 | 연속형 | 보험계약대출금합계-상환금합계 |
| 40 | 보험계약대출잔액평균 | 연속형 | 보험계약대출금잔액합계/대출건수 |
| 41 | 보험계약대출연체일수합계 | 연속형 | |
| 42 | 보험계약대출연체일수평균 | 연속형 | 보험계약대출연체일수합계/대출건수 |
| 43 | 신용대출건수합계 | 연속형 | |
| 44 | 목표변수 | 범주형 (이분형) | 개인연금보험가입여부 (예/아니오) |

택 방법)를 통해서, 전체 데이터를 이용한 것과 유사하게 모델 정확도가 나오는지 비교하여, 현재 사례의 모델에서 가장 효과적인 데이터 축소 방안을 찾고자 하는 것이다. 표 3.2는 본 논문에서 실험에 사용할, 설명변수 축소방법에 대한 정리이다.

위의 표 3.2의 첫 번째 설명변수 축소방법인 요인분석을 통한 설명변수의 축소방법을 더 구체적으로 정리하면, 전체 데이터 집합 (data set)을 $L$이라 하고, 다음과 같이 정의한다. $L = \{(Y, X_p), p =$

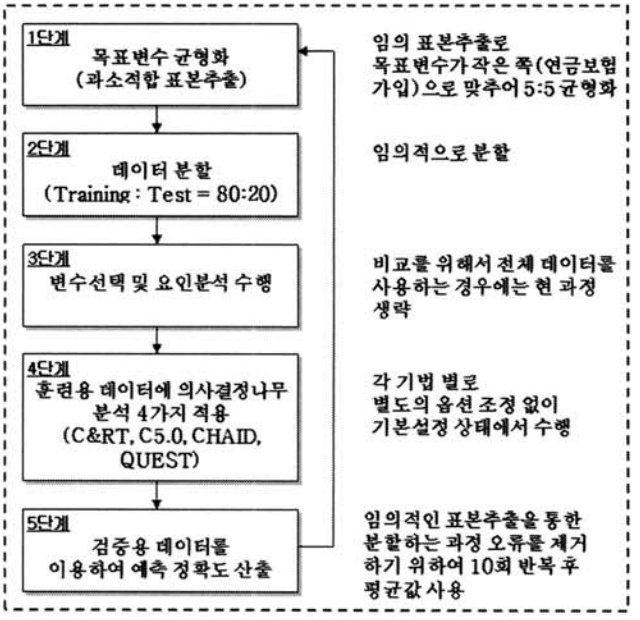


그림 **3.1** 실험의 설계

표 **3.2** 설명변수 축소 방법

| 번호 | 설명변수 축소 방법 | | 비고 |
|---|---|---|---|
| 1 | 요인분석을 통한 요약 | 연속형 변수를 요인분석을 수행하여,<br>변수 축소 후 요인적재값 활용 | 통계적 방법 |
| 2 | 통계검정 방법 (t-검정과 - $\chi^2$검정)<br>에 의한 변수선택 | 연속형 변수 (t-검정) / 범주형 변수 (카이스퀘어<br>검정)을 통해서 유의수준 99% 이상 유의한<br>변수만 선택하여 설명변수 집단으로 구성 | 통계적 방법 |
| 3 | 혼합법 1)<br>로지스틱 회귀분석의 단계선택법 (Stepwise)를<br>통해서 변수선택 | 로지스틱 회귀분석을 사전 수행하여,<br>로지스틱 회귀분석에서<br>사용되는 변수만을 선택 | Hybrid 방법 |
| 4 | 혼합법 2)<br>신경망 (MLP)의 민감도 분석을<br>통해서 변수선택 | 신경망 분석을 사전 수행하여,<br>민감도 분석 결과 상위 10개의 변수만을 선택 | Hybrid 방법 |

$1, 2, \cdots, p\}$ 여기서, $Y$는 1개의 목표변수를 의미하고, $X_p$는 p개의 설명변수 집합이라고 정의한다. 그리고 $X_m$를 설명변수 집합 중 범주형 (이분형 포함)설명변수 집합이라고 하고, $X_n$를 $X_p$라는 설명변수 전체 집합 중 연속형 설명변수 집합이라고, 정의하면 $X_p = (X_m, X_n)$ (단, p = m +n) 이라고 할 수 있다. 여기에서, 요인분석의 특성상 연속형 설명변수 집합 $X_n(n = 1, 2, \cdots, n)$만을 이용하여, 요인분석을 수행한다. 요인분석을 수행할 때 요인 추출 조건은 고유값이 1이상이고, 회전 방법은 VARIMAX 회전 방법을 이용한다. 연속형 설명변수 데이터 집합 $X_n$을 이용하여, 요인분석을 수행한 다음 나온 요인적재값 변수를 $X_l(l = 1, 2, \cdots, l)$이라고 정의하면, 최종적으로 활용하는 데이터 집합 $L_r$은 $L_r = (Y, M_m, X_l)$이 된다. 즉, 범주형 설명변수는 그대로 사용을 하고, 추가로 연속형 설명변수들은 요인분석을 이용하여, 데이터 축소를 한 요인적재값을 사용하여, 의사결정 나무 분석에 적용하는 것이다.

2번째 통계적 검정방법을 통한 변수선택과 3번째 로지스틱 회귀분석의 단계선택법은 많이 알려진 방법이므로 설명을 생략한다. 4번째 신경망 분석의 민감도 분석 (sensitivity analysis)을 이용한 변수 선택 방법은, 전체 42개 설명변수로 이루어진 전체 뉴런 (neuron)들의 네트워크를 가장 많이 변화시키는 변수를 가장 중요한 변수로 선택하는 방법으로, 민감도 분석결과 값이 클수록 높은 중요도를 가진다고 할 수 있다. (Engelbrecht와 Cloete, 1996) 민감도 분석에 대한 알고리즘을 자세히 알아보면, 먼저 신경망 알고리즘 자체적으로, 모든 입력변수들에 대하여 [0,1] 사이의 범위를 가지도록, 설명변수들을 조정한 다음 설명변수 $x_i$의 민감도를 계산하게 되는데, 설명변수 $x_i$의 민감도 $S_i$는 다음과 같이 계산을 하게 된다.

$$S_i = \frac{1}{m}\frac{1}{q}\sum_{L}\sum_{k=l}^{q}\left(\frac{|O_k^0 - O_k^i|}{O_k^0}\right) \tag{3.1}$$

위의 식 (3.1)에서 는 $O_k^0$각 학습 데이터의 신경망 출력 값을 나타낸다. 그리고 $O_k^i$는 설명변수 $x_i$가 제거된 경우의 신경망 출력값을 나타낸다. 여기서, 설명 변수 $x_i$가 제거되었다는 것은 $x_i$를 0이라고 가정하고 계산하는 것을 의미한다. 그리고 $L$은 전체 훈련용 학습 데이터를 의미하고, $m$은 훈련용 학습 데이터의 수를 $q$는 신경망 네트워크 내의 출력 노드 수를 의미한다. 위의 식 (3.1)의 의미는 훈련용 데이터 집단에서, 설명변수 $x_i$가 원래 데이터 값을 가졌을 때와 그렇지 않을 때의 값의 차이를 이용하여, 해당 변수가 전체 신경망에 영향을 미치는 정도를 계산한 것이고, 이 민감도 $S_i$값이 클수록 신경망에 미치는 영향도가 크다는 것을 의미한다 (강부식과 박상찬, 2001). 민감도 분석 점수가 얼마 이상이어야 활용도가 높은지는 각 네트워크 모델마다 전부 차이가 있을 수 있으므로, 본 논문에서는 Clementine의 기본설정 중 하나인 민감도 분석 결과 값이 높은 상위 10개 변수만을 선택하도록 한다. (SPSS Inc., 2007)

본 논문에서는 사례 보험사의 자료를 이용하여, 표 3.2에서 제시한 4개의 설명변수 축소방법과 전체 데이터를 전부 사용하는 경우에 대하여, 검증용 데이터를 이용한 정확도의 평가를 통해서, 설명변수를 어떻게 축소시키는 것이 가장 좋은 예측 정확도를 나타내는지, 그리고 가장 적은 수의 데이터를 이용하여, 가장 효율적인 모델을 만들 수 있는 방법은 무엇인지를 실험을 통해서 확인하고자 한다.

### 3.3. 실험의 가정

실험을 수행하기 전에, 2가지의 가정이 있다. 첫 번째 가정은 본 논문에서 활용되는 여러 분석방법들은 SPSS Clementine 12.0.2를 기준으로, 사용하는 기법에 따라 별도의 옵션을 설정하지 않고, 기본설정으로 하여 분석을 수행하였다. 두 번째 가정은 실험결과에 대한 평가방법이다. 설명변수를 축소하여, 정보가 유실이 되었는데도 불구하고, 정확도가 거의 유사하거나 혹은 더 좋아진다는 것은, 축소 방법이 매우 효과적이었다는 것을 증명하기 때문에, 이를 평가하기 위한 검증용 데이터의 정확도 계산법은 다음의 표 3.3과 같이 정의한다.

**표 3.3** 사례 데이터 설명

| 구분 | | 예측 | |
|---|---|---|---|
| | | T | F |
| 실제 | T | 참 (TP) | 거짓 (FN) |
| | F | 거짓 (FP) | 참 (TN) |

표 3.3과 같은 구조에서 검증용 데이터의 정확도를 산출하는 공식은 $(TP+TN)/(TP+FN+FP+TN) \times 100$ (%) 로 정의한다. 다음 고려할 사항은 설명변수 감소율이 될 것이다. 예를 들어 설명변수를 축소시키는 A와 B방법 중에서, 검증용 데이터에 대한 모델 정확도가 동일하다면, 다음에 생각해 볼

수 있는 것은, A와 B 방법 간의 설명변수 감소율이 될 것이다. 즉, 전체 42개 중 10개로 감소 후 모델링한 방법과 20개로 감소 후 모델링한 방법의 정확도가 동일하다면, 당연히 10개로 감소한 방법이 더 성능이 좋다고 얘기할 수 있을 것이다. 따라서 설명변수 감소율은 전체 42개의 설명변수를 기준으로 $(42 - \text{방법별축약사용설명변수수})/42 \times 100$ (%) 로 정의한다.

## 4. 실험 결과

표 4.1은 본 논문에서 수행하고자 하는 변수선택 방법 4가지와 전체 설명변수를 이용한, 5종류의 방법에 대하여, 4개의 의사결정나무 분석 기법을 적용하여, 산출된 정확도 값을 정리한 것이다. 표 4.1에서 시도 수 부분의 '합'이라는 것은 과소적합 표본추출을 통해서 나온 데이터 수를 의미하고, 그 아래 '훈' 이라는 것은 훈련용 데이터로 '합' 데이터의 약 80%를 차지하고 있다. 그 아래의 '검'이라는 것은 검증용 데이터로 나머지 20%를 데이터를 의미하며, 정확도는 훈련용 데이터를 기반으로 나온 모델 결과에 검증용 데이터를 적용하여, 산출된 것이다.

표 4.1에서 요인분석의 설명변수 수는 17개~19개가 나오는데, 이는 범주형 변수 7개를 제외하면, 요인수가 10개~12개가 나오는 것이라고 해석할 수 있다. 그리고 표 4.1의 가장 오른쪽의 평균은 4개 의사결정나무 분석기법을 통해서 나온 정확도 값의 산술 평균값을 의미한다.

표 4.1의 결과에서, 전체를 포함한 5개의 설명변수 선택 방법들에 관한, 정확도 평균의 차이에 대하여, 일원배치 분산분석 검정을 수행하였다. 그 결과가 표 4.2와 같다.

표 4.2에서 p-값이 0.000으로 5개의 기법들 간에는 차이가 있는 것으로 나타났다. 구체적으로 살펴보면, t/카이제곱 검정을 통해서 변수를 선택한 방법이, 가장 평균 정확도가 높은 것으로 되어져 있으나, Duncan 사후 검정 결과를 보면 전체 변수를 사용하여 나온 평균 정확도와 거의 차이가 없는 것을 알 수 있다. Duncan의 사후 검정에서는 로지스틱과 신경망 등도 유의한 차이가 나타나는 것으로 결과가 나왔다. 그러나 직관적 수치의 정확도차이에서 보면 요인분석으로 축약한 것을 제외하면, 전부 정확도가 약 86% 대인 것을 알 수 있어, 일반 분석 결과 적용 시, 큰 차이가 나타나지 않는다고 할 수 있다.

다음은 설명변수 감소율과의 관계에 대하여 알아본다. 그림 4.1은 위의 정확도 평균값과 각 기법별로 설명변수 평균 감소율에 대한 산점도이다.

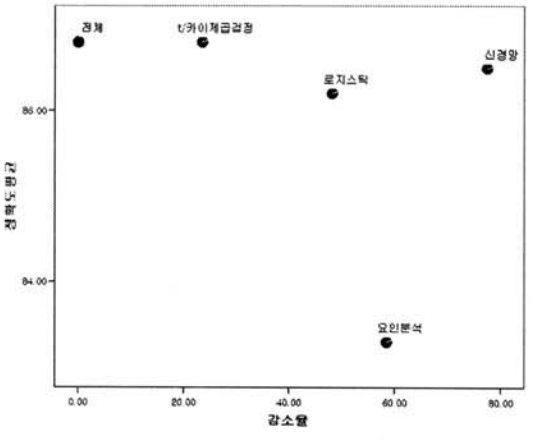


**그림 4.1** 정확도 평균값과 설명변수 감소율 간 산점도

그림 4.1에서 X축의 설명변수 감소율도 크면 클수록, 더욱 효과적인 모델이 될 것이고, 당연히 Y축

**표 4.1** 사례 데이터 설명

| 시도 수 | 축약방법 | 설명 변수 수 | C5.0 정확도 | C&RT 정확도 | CHAID 정확도 | QUEST 정확도 | 평균 |
|---|---|---|---|---|---|---|---|
| 1차 | 요인분석 | 19개 | 86.51% | 82.44% | 82.99% | 81.40% | 83.34% |
| 합:15,532건 | t/카이제곱 검정 | 31개 | 90.19% | 87.29% | 85.40% | 83.77% | 86.66% |
| 훈:12,426건 | 로지스틱 | 21개 | 89.61% | 87.29% | 85.37% | 83.28% | 86.39% |
| 검:3,106건 | 신경망 | 10개 | 89.61% | 86.10% | 85.14% | 82.77% | 85.91% |
| | 전체 | 42개 | 90.32% | 87.29% | 85.40% | 83.28% | 86.57% |
| 2차 | 요인분석 | 19개 | 86.49% | 84.52% | 81.88% | 79.88% | 83.19% |
| 합:15,374건 | t/카이제곱 검정 | 32개 | 89.25% | 87.57% | 85.89% | 83.53% | 86.56% |
| 훈:12,299건 | 로지스틱 | 22개 | 88.79% | 85.83% | 85.17% | 82.64% | 85.61% |
| 검:3,075건 | 신경망 | 10개 | 88.92% | 86.26% | 85.83% | 83.99% | 86.25% |
| | 전체 | 42개 | 89.11% | 87.57% | 85.89% | 83.53% | 86.53% |
| 3차 | 요인분석 | 18개 | 86.51% | 83.23% | 82.37% | 79.56% | 82.92% |
| 합:15,294건 | t/카이제곱 검정 | 33개 | 89.12% | 87.40% | 85.48% | 83.17% | 86.29% |
| 훈:12,235건 | 로지스틱 | 22개 | 89.10% | 87.40% | 85.35% | 83.17% | 86.26% |
| 검:3,059건 | 신경망 | 10개 | 89.99% | 86.94% | 85.32% | 84.49% | 86.69% |
| | 전체 | 42개 | 89.25% | 87.40% | 85.48% | 83.17% | 86.33% |
| 4차 | 요인분석 | 18개 | 86.49% | 83.20% | 81.84% | 80.34% | 82.97% |
| 합:15,546건 | t/카이제곱 검정 | 32개 | 89.49% | 87.44% | 85.71% | 82.65% | 86.32% |
| 훈:12,437건 | 로지스틱 | 21개 | 89.25% | 85.71% | 84.93% | 82.65% | 85.64% |
| 검:3,109건 | 신경망 | 10개 | 89.49% | 86.75% | 85.06% | 82.45% | 85.94% |
| | 전체 | 42개 | 89.52% | 87.44% | 85.71% | 82.65% | 86.33% |
| 5차 | 요인분석 | 18개 | 86.56% | 84.13% | 82.62% | 81.28% | 83.65% |
| 합:15,408건 | t/카이제곱 검정 | 34개 | 89.08% | 87.90% | 87.47% | 84.59% | 87.26% |
| 훈:12,326건 | 로지스틱 | 21개 | 89.27% | 87.80% | 85.77% | 82.72% | 86.39% |
| 검:3,082건 | 신경망 | 10개 | 90.16% | 87.57% | 87.08% | 85.41% | 87.56% |
| | 전체 | 42개 | 88.95% | 87.90% | 87.48% | 84.59% | 87.23% |
| 6차 | 요인분석 | 18개 | 86.68% | 82.59% | 81.38% | 80.03% | 82.67% |
| 합:15,459건 | t/카이제곱 검정 | 31개 | 89.33% | 86.71% | 85.63% | 84.19% | 86.47% |
| 훈:12,367건 | 로지스틱 | 24개 | 88.81% | 85.79% | 83.86% | 82.03% | 85.12% |
| 검:3,092건 | 신경망 | 10개 | 88.87% | 86.61% | 85.11% | 83.93% | 86.13% |
| | 전체 | 42개 | 89.49% | 86.81% | 85.63% | 84.19% | 86.53% |
| 7차 | 요인분석 | 17개 | 87.69% | 84.97% | 82.38% | 81.00% | 84.01% |
| 합:15,403건 | t/카이제곱 검정 | 32개 | 90.91% | 87.82% | 85.56% | 84.28% | 87.14% |
| 훈:12,322건 | 로지스틱 | 21개 | 89.14% | 86.87% | 85.20% | 82.97% | 86.05% |
| 검:3,081건 | 신경망 | 10개 | 90.35% | 87.53% | 85.69% | 83.72% | 86.82% |
| | 전체 | 42개 | 90.58% | 87.82% | 85.56% | 84.28% | 87.06% |
| 8차 | 요인분석 | 18개 | 87.43% | 82.97% | 82.78% | 81.70% | 83.72% |
| 합:15,442건 | t/카이제곱 검정 | 32개 | 90.83% | 88.15% | 85.40% | 85.13% | 87.38% |
| 훈:12,354건 | 로지스틱 | 22개 | 89.85% | 88.18% | 86.05% | 83.89% | 86.99% |
| 검:3,088건 | 신경망 | 10개 | 89.85% | 87.56% | 86.31% | 84.74% | 87.12% |
| | 전체 | 42개 | 90.06% | 88.15% | 85.40% | 85.13% | 87.19% |
| 9차 | 요인분석 | 18개 | 87.40% | 84.03% | 83.27% | 82.10% | 84.20% |
| 합:15,458건 | t/카이제곱 검정 | 32개 | 89.33% | 87.66% | 85.37% | 83.57% | 86.48% |
| 훈:12,366건 | 로지스틱 | 23개 | 89.53% | 87.53% | 85.76% | 83.57% | 86.60% |
| 검:3,092건 | 신경망 | 10개 | 88.64% | 86.97% | 86.06% | 83.18% | 86.21% |
| | 전체 | 42개 | 89.17% | 87.66% | 85.37% | 83.57% | 86.44% |
| 10차 | 요인분석 | 18개 | 87.00% | 82.79% | 81.40% | 80.91% | 83.03% |
| 합:15,356건 | t/카이제곱 검정 | 33개 | 89.53% | 87.29% | 86.97% | 85.32% | 87.28% |
| 훈:12,285건 | 로지스틱 | 22개 | 89.27% | 86.87% | 85.29% | 83.54% | 86.24% |
| 검:3,071건 | 신경망 | 10개 | 89.17% | 86.54% | 86.50% | 83.87% | 86.52% |
| | 전체 | 42개 | 89.70% | 87.29% | 86.97% | 85.32% | 87.32% |

**표 4.2** 사례 데이터 설명

| 축약/축소방법 | 설명변수 감소율 (평균) | 정확도 평균값 | F | p-값 |
|---|---|---|---|---|
| 요인분석 | 56.90% | 83.37% | | |
| 로지스틱 | 47.86% | 86.12% | | |
| 신경망 | 76.19% | 86.51% | 2178.85 | 0.000 |
| 전체변수사용 | 0% | 86.75% | | |
| t/카이제곱 검정 | 23.33% | 86.78% | | |
| Duncan 검정: t/카이제곱 검정 = 전체변수 사용 > 신경망 > 로지스틱 > 요인분석 (유의한 차이 순) | | | | |

의 정확도도 높으면, 높을수록 좋은 모델이라고 할 수 있어, 이를 기준으로 살펴보면 신경망의 민감도 분석을 이용하여, 42개를 10개로 축약한 방법이, 정확도 평균은 86.51%를 나타내고, 설명변수의 감소율도 76.19%가 되어서, 가장 효과적인 방법임을 알 수 있으며, 그림 4.1에서 보면 가장 우측 상단에 위치하고 있음을 알 수 있다. 따라서 본 사례의 생명보험회사에서, 은퇴보험 가입 예측가능성을 파악하기 위한, 의사결정나무 분석을 적용하기 위해서는 신경망 분석의 민감도 분석을 통해서, 상위 10개의 주요한 변수들만을 선택하고, 그 다음 의사결정나무 분석을 수행하는 것이, 가장 적은 수의 설명변수로, 정확도가 높은 모델을 만들어 낼 수 있다는 결과를 도출할 수 있다.

표 4.3은 10번의 반복 시행을 통해서, 신경망 기법의 민감도 분석결과 도출된 상위 10개의 주요 설명변수들에 대한 선정 빈도표이다.

**표 4.3** 사례 데이터 설명

| 변수명 | 선정 빈도수 | 변수명 | 선정 빈도수 |
|---|---|---|---|
| 월납환산보험료합계 | 10번 | 최근계약일로부터의일수 | 5번 |
| 월납환산보험료평균 | 10번 | 보험계약대출잔액평균 | 4번 |
| 납입완료기납입보험료합계 | 10번 | 가입상품수 | 3번 |
| 납입완료기납입보험료평균 | 10번 | 주피보험자 직업코드 | 3번 |
| 전체유지건수합계 | 9번 | 계약자 직업코드 | 3번 |
| 최장계약일로부터일수 | 9번 | 납입완료거래개월수평균 | 2번 |
| 보험계약대출잔액합계 | 7번 | 평균가입주기 (월단위) | 2번 |
| 납입완료거래개월수합계 | 6번 | 보험계약대출보유건수합계 | 1번 |
| 제24회이상유지건수합계 | 5번 | 신용대출건수합계 | 1번 |

표 4.3에서 월납환산보험료 합계 외 4개의 변수가 모두 10번씩 선정되었으며, 42개의 변수 중에서, 18개 변수가 10번의 신경망의 민감도 분석 수행 중 단 1번이라도 상위 10위 안에 포함되는 것으로 나타났다.

표 4.4를 사용된 4가지 방법들에 대한 정확도 값의 평균비교 검정을 일원배치 분산분석으로 수행한 것이다.

**표 4.4** 사례 데이터 설명

| 축약/축소방법 | 정확도 평균 | F | p-값 |
|---|---|---|---|
| QUEST | 83.14% | | |
| CHAID | 85.03% | 140.269 | 0.000 |
| CART | 86.47% | | |
| C5.0 | 88.99% | | |
| Duncan 검정: C5.0 > CART > CHAID > QUEST (유의한 차이 순) | | | |

표 4.4를 보면 C5.0으로 만든 의사결정나무 분석 모델이, 가장 높은 평균 정확도를 보이고, QUEST로 만든 모델이 가장 정확도가 낮은 것으로 나타났다. 따라서, 본 사례 보험사에서는 C5.0 알고리즘을 이용한 의사결정나무 분석이, 가장 정확도가 높을 것으로 판단되어진다.

## 5. 결론 및 시사점

앞에서, 서술한 본 논문의 목적을 다시 정리하면, 사례의 생명보험회사에서 기존 고객들 중 개인연금 보험을 추천할 가능성이, 높은 고객을 선정하는 의사결정나무 분석 모델을 개발할 때, 많은 수의 설명변수를 최소화하면서, 정확도는 가장 높은 그러한 의사결정나무 분석 모델을 만드는 것이다. 이를 위한, 실험 결과 변수 선택 방법으로, 신경망의 민감도 분석을 수행하여, 상위 10개의 변수를 선택하는 방법이 가장 효과적인 것으로 나타났으며, 의사결정나무 분석 기법 중에서는 C5.0을 활용하는 것이 가장 높은 정확도를 나타내었다. 본 실험 결과를 통해서, 사례 보험사는 3가지 성과를 얻을 것으로 예상된다. 첫째는 모든 데이터를 사용하지 않으므로, 모델링을 수행하기 전에 데이터를 저장하는, 데이터 마트 자원을 절약할 수 있다. 둘째는 적은 수의 데이터를 이용하여 모델을 생성하므로, 더욱 빨리 분석 결과를 도출하여 다음 업무에 적용을 할 수 있으며, 마지막으로, 의사결정나무 분석의 결과를 해석할 때, 설명변수 수가 적어서, 특성을 해석할 때 편리하고 쉽게 해석이 될 것이다. 본 연구는 국내의 한 중위권 생명보험사 사례이기에, 모든 생명보험사에 이 결과를 적용하기에는 무리가 있을 수 있다. 그러나 보험사들에게는 공통적으로, 보험에 가입한 고객의 전반적인 특성이나 또는 고객들로부터 수집되는 정보의 차이가 유사한 부분도 많아서, 본 사례를 유사하게 적용을 하는 경우, 시행착오 등을 줄일 수 있을 것으로 기대된다. 그리고 본 연구를 더욱 확장하기 위한, 향후 연구에서는 본 논문에서 사용하지 않은, 설명변수의 변수 선택 방법을 적용하여, 결과를 비교 실험해보는 연구와 함께, 더욱 더 다양한 산업 사례와 시뮬레이션을 통한 추가적인 연구가 이루어질 필요가 있다.

## 참고문헌


강부식, 박상찬 (2001). 신경망의 민감도 분석을 이용한 귀납적 학습 기법의 변수 부분 집합 선정. <한국지능정보시스템학회논문지>, **7**, 51-63.

권철신, 홍순욱 (2001). 유사상관계수의 개념을 도입한 범주형 변수의 축약에 관한 연구. <산업공학>, **14**, 79-83.

박성민, 박영준 (2005). 회귀분석변수선택 절차를 이용한 인터넷 네트워크 품질 특성과 고객 만족도와의 관계 실증분석. <2005 한국경영과학회/대한산업공학회 춘계공동학술대회 논문집>, 822-828.

송문섭, 윤영주 (2001). 데이터 마이닝 패키지에서 변수 선택 편의에 관한 연구. <응용통계연구>, **14**, 475-486.

정석훈, 서용무 (2008). Rough Set 기법을 이용한 신용카드 연체자 분류. *Entrue Journal of Information Technology,* **7**, 141-150.

허명회, 임용빈, 이용구 (2008). 다목적 다변량 자료분석을 위한 변수선택. <응용통계연구>, **21**, 141-149.

허문열, 박영석 (2005). 상호정보를 사용한 변수선택의 동적 모델링. <통계연구>, **13**, 57-74.

허준, 김종우 (2007). 불균형 데이터 집합에서의 의사결정나무 추론: 종합병원의 건강 보험료 청구 심사 사례. *Information Systems Review,* **9**, 45-65.

Anand, S. S., Patrick, A. R., Hughes, J. G., and Bell, D. A. (1998). A data mining methodology for cross-sales. *Knowledge-Based Systems*, **10**, 449-461.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks,* **5**, 537-550.

Brieman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Chung, S. S., Lee, K. H. and Lee, S. S. (2005). A study on split variable selection using transformation of variables in decision trees. *Journal of Korean Data & Information Science Society,* **16**, 195-205.

Engelbrecht, A.P. and Cloete, I. (1996). A sensitivity analysis algorithm for pruning feedforward neural networks, neural networks. *1996,IEEE International Conference,* **2**, 1274-1278.

Kang, J. (2004). A study on factors associated with the success of CRM in the insurance company. *Journal of Korean Data & Information Science Society,* **15**, 141-172.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* **29**, 119-127.

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal component. *Applied Statistics,* **36**, 22-33.

Krzanowski, W. J. (1996). A stopping rule for structure-preserving variable selection. *Statistics and Computing,* **6**, 51-56.

Loh, W. and Shih, Y. (1997). Split selection methods for classification trees. *Statistica Sinica,* **7**, 815-840.

Lu, H., Setiono, R. and Liu, H. (1996). Effective data mining using neural networks. *IEEE Transactions on Knowledge and Data Engineering,* **8**, 957-961.

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

SPSS Inc., (2007). *Clementine 12.0 User's Guide,* SPSS Inc.

*Journal of the Korean Data & Information Science Society*
*2009*, **20**(1), 179–190

한국데이터정보과학회지

# A study on the comparison of descriptive variables reduction methods in decision tree induction: A case of prediction models of pension insurance in life insurance company[†]

Yong Goo Lee[1] · Joon Hur[2]

[1] Department of Mathematics & Statistics, Chung-Ang University
[2]Consulting Team, SPSS Korea Data Solution Inc.

Received 26 December 2008, revised 16 January 2009, accepted 23 January 2009

**Abstract**

In the financial industry, the decision tree algorithm has been widely used for classification analysis. In this case one of the major difficulties is that there are so many explanatory variables to be considered for modeling. So we do need to find effective method for reducing the number of explanatory variables under condition that the modeling results are not affected seriously. In this research, we try to compare the various variable reducing methods and to find the best method based on the modeling accuracy for the tree algorithm. We applied the methods on the pension insurance of a insurance company for getting empirical results. As a result, we found that selecting variables by using the sensitivity analysis of neural network method is the most effective method for reducing the number of variables while keeping the accuracy.

*Keywords*: Decision tree induction, factor analysis, neural networks, pension insurance, variables selection (Reduction).

[†] This research was supported by 2007 Chung-Ang University research fund for excellent researcher.
[1] Corresponding Author: Professor, Department of Mathematics & Statistics, Chung-Ang University, Seoul 156-756, Korea. E-mail: leeyg@cau.ac.kr
[2] Principal consultant, Consulting Team, SPSS Korea Data Solution Inc., Seoul 135-513, Korea.