# 대용량 자료에서 핵심적인 소수의 변수들의 선별과 로지스틱 회귀 모형의 전개

임용빈[1] · 조재연 · 엄경아 · 이선아
이화여대 통계학과

# Screening vital few variables and development of logistic regression model on a large data set

Yong B. Lim · J. Cho · Kyung-A Um · Sun-Ah Lee
Department of Statistics, Ewha Womans University

Key words: Classification Tree, Screening vital few variables, Logistic regression model

Abstract

In the advance of computer technology, it is possible to keep all the related informations for monitoring equipments in control and huge amount of real time manufacturing data in a data base. Thus, the statistical analysis of large data sets with hundreds of thousands observations and hundred of independent variables whose some of values are missing at many observations is needed even though it is a formidable computational task. A tree structured approach to classification is capable of screening important independent variables and their interactions. In a Six Sigma project handling large amount of manufacturing data, one of the goals is to screen vital few variables among trivial many variables. In this paper we have reviewed and summarized CART, C4.5 and CHAID algorithms and proposed a simple method of screening vital few variables by selecting common variables screened by all the three algorithms. Also how to develop a logistics regression model on a large data set is discussed and illustrated through a large finance data set collected by a credit bureau for th purpose of predicting the bankruptcy of the company.

1) 교신저자 yblim@ewha.ac.kr

## 1. 서론

컴퓨터의 발전과 더불어서 관련된 많은 정보들의 데이터 베이스화가 가능하여져서 대용량의 자료를 다룰 기회가 많아지게 되고, 프로젝트 목적에 따라서 대용량의 자료를 처리하고 분석하기 위한 통계적인 방법의 필요성이 대두된다. LCD나 반도체 등의 첨단 산업에서는 오븐 온도, 배기압, 평균 암페어, 평균 볼테지, 통과 호기, 램프 수명, 막 두께, 결점수, 중간 검사 통과 여부 등 공정의 실처리 변수들과 검사 변수들이 자동 계측되어서 데이터 베이스에 자동 집적된다. 또한 신용 평가 회사에서는 기업체들을 정상인 기업과 부도를 낸 기업으로 분류하고, 각 기업체의 재무 상태를 설명하는데 도움이 되는 수 십 개의 재무 비율 변수들을 계산하고, 경영 상태에 관한 변수들을 정의하여, 대규모의 자료를 저장한다. 대용량의 자료는 관측치들의 수는 물론, 변수들의 수도 많을 뿐만 아니라, 각각의 자료점에서 결측된 변수들이 존재하는 것을 그 특징으로 한다. 우리는 이런 대용량의 자료를 이용하여, 최종검사의 불합격이나, 기업체의 부도 여부 등과 같은 특정 현상에 영향을 주는 공정의 실처리 변수들이나 재무 비율 변수들을 선별하고, 선별된 변수들을 활용하여, 앞으로의 특정 현상의 발생 가능성을 예측하고 싶을 때가 많다.

식스 시그마 프로젝트에서 주된 관심사항들 중의 하나는 품질 특성치에 영향을 주리라 기대되는 사소한 다수(trivial many)의 변수들 중에서 결정적으로 영향을 주는 핵심적 소수(vital few)의 변수들을 선별하는 것이다. 경험적으로 공업 통계에서 잘 알려진 원칙이 'factor sparsity'의 원칙으로 품질특성치에 영향을 주리라 기대되는 많은 변수(trivial many)들 중에서 실질적으로 영향을 주는 변수들(vital few)의 수는 많아야 3, 4 개라는 사실이다. 품질특성치 변동의 80-90%가 사소한 다수의 변수들 중에서 10-20%의 변수들 만에 의해서 설명이 된다는 '파레토의 원칙'이 또한 'factor sparsity'의 원칙을 뒷받침해준다. 이 논문에서는 사소한 다수의 변수들 중에서 핵심적 소수의 변수들을 선별하는 대용량 자료의 분석 방법을 소개하려 한다. 또한 제품의 불량률이나 기업의 부도율을 로지스틱 회귀 모형을 통해서 예측하려는 경우, 단계적 방법(stepwise method)에 의하여 중요 변수들을 선별하고 적절한 모형을 선택하는 것이 일반적이다. 그런데 단계적 방법을 적용할 시에 중요 변수 선별의 후보가 되는 목록은 모든 관측치에서 결측치가 존재하지 않는 변수들만으로 구성되기 때문에, 단계적 방법을 적용하기 위해서는 사소한 다수의 변수들 중에서 목록에 들어갈 후보 변수들을 선별하여, 분석용 자료에 목록에 있는 후보변수들의 결측이 없도록 해야 한다. 따라서 단계적인 방법에 의한 로지스틱 회귀 모형의 결정시에도 변수 선별이 선행되어야 한다. 이 논문에서는 결측치들이 많은 대용량의 자료에서 적절한 로지스틱 회귀 모형 선택을 위한 전개 방법에 관해서 논하고, 모 신용평가 회사에서 수집된 기업체의 재무 및 경영 상태와 부도 여부에 관한 자료를 가지고 제안된 방법들을 예시한다.

## 2. 변수 선별 방법

대용량 자료를 분석할 때 많이 활용되는 데이터 마이닝 기법중 하나로 중요한 설명변수들과 그들의 교호작용효과를 쉽게 판별할 수 있는 분석 방법인 의사결정나무(Decision Tree)가 있다. 의사결정나무(Decision Tree)의 분석 방법은 나무구조의 접근 방법(Tree-structured approach)인 축차분할(Recursive Partitioning)이다. 축차분할에서는 설명변수들의 공간(space of explanatory variables)이 축차적인 분리에 의해서 끝마디(terminal nodes)로 분할된다. 끝마디에 도달하는 길(path)이 그 마디에 배치된 자료점들의 구조 정보(structure information)를 제공하는데, 길을 이루고 있는 마디의 개수를 깊이(depth)라고 한다. 반응변수가 범주형 변수(categorical variable)인 분류나무(classification tree)인 경우에 끝마디에서의 예측치는 배치된 자료점들의 최빈값(mode)이다. 분류나무 예측치는 구하는 과정이 간단하고 각 자료점은 그 자료점이 속한 끝마디에 도달하는 길에 의해서 설명되어 해석이 쉽다는 장점이 있다.

축차분할의 잘 알려진 알고리즘에는 CART(Classification and Regression Tree), CHAID(Chi-squared Automatic Interaction Detection) 그리고 C4.5가 있다. 의사 결정 나무의 축차적인 분리는 각 마디의 불순도(node impurity)를 가장 크게 감소시키는 방향으로 분리가 일어나는데, 이때 불순도의 차이를 가장 크게 해주는 변수가 가장 좋은 분할자로 선택된다. 의사결정나무에서 나무가 계속 성장하도록 방치하는 것은 구조를 복잡하게 만들기 때문에 바람직하지 않으므로,

더 이상 분리가 일어나지 않도록 하는 정지규칙(stopping rule)이 있다. 정지규칙은 모든 자료가 한 그룹에 속할 때, 마디에 속하는 자료가 일정 수 이하일 때, 불순도의 감소량이 작을 때, 깊이(depth)가 일정 수 이상일 때 등의 규칙이 있으며 SAS E-Miner에서는 이를 사용자가 직접 설정해 줄 수 있다. Briemen 등(1984)이 제시한 CART는 축차분할의 잘 알려진 알고리즘이다. CART는 마디의 불순도의 측도로 $2p(1-p)$ 로 정의된 지니 지수(Gini index)를 사용한다. 여기서 $p$는 해당 마디의 불량률 또는 부도율이고, 지니 지수는 목표변수인 반응변수의 불량 여부 또는 부도 여부를 나타내는 베르누이 확률변수의 분산이 된다. CART는 주어진 마디에서 정지규칙을 만족할 때까지 분리가 일어나고, 최종 나무의 크기와 끝마디들을 결정하기 위해서 가지치기 기법(pruning technique)과 교차 타당성을 사용한다.

C4.5는 Quinlan(1993)에 의해 개발된 의사결정나무 알고리즘으로, 마디의 불순도의 측도로 $-p\log p-(1-p)\log(1-p)$로 정의된 엔트로피 지수(Entropy index)를 사용하는 것을 제외하고, CART와 유사하다. (C4.5에서도 가지치기를 해주는데, CART와의 차이점은 검증용 자료를 사용하지 않고 모델에 사용되는 자료만을 가지고 각 마디에서의 오류율에 근거하여 가지치기한다.)

CHAID는 Hadrian(1975)이 제시한 $X^2$검정을 이용한 알고리즘이다. CHAID는 각 마디에서의 축차적인 분리의 기준으로 $X^2$검정에서 통계적으로 가장 유의한 차이를 보여주는 입력변수를 분리 변수로 선택하고, 정지규칙은 유의확률(p-value)의 값이다.

사소한 다수의 변수들 중에서 핵심적 소수의 변수들을 선별하기 위한 일차적인 전략으로 각각의 알고리즘을 적용하여 구해진 분류나무의 축차적인 분리에 사용된 변수들을 선별하고, 선별된 변수들의 투표를 통하여 핵심적 소수의 변수들을 선별하기를 제안한다.

또 다른 변수선별 방법으로는 제품의 불량률이나 기업의 부도율을 예측하기 위해서 로지스틱 회귀분석의 단계적 방법(stepwise method)에 의하여 중요 변수들을 선별하고 적절한 모형을 선택하는 것이다. 그런데 단계적 방법을 적용할 시에 자료 점들의 목록은 모든 변수들에 대하여 결측치가 존재하지 않는 자료 점들만으로 구성되기 때문에, 단계적 방법을 적용하기 위해서는 사소한 다수의 변수들 중에서 목록에 들어갈 핵심적 소수 변수들의 후보 변수들을 일차적으로 선별하여, 후보변수들의 결측이 없는 자료점들 만으로 구성된 분석용 자료에 최대한 많은 자료점들이 포함되도록 해야 한다. 일차적으로 분석용 자료에 대체(replacement) 노드를 통해 결측치를 연속형 변수의 경우 각 변수의 평균값으로, 그리고 범주형 변수의 경우 최빈값으로 각각 그 값을 대체한 후에, 로지스틱 회귀분석의 단계적 방법(stepwise method)으로 일차적인 핵심적인 소수 변수들의 후보 변수들을 선별할 수도 있다.

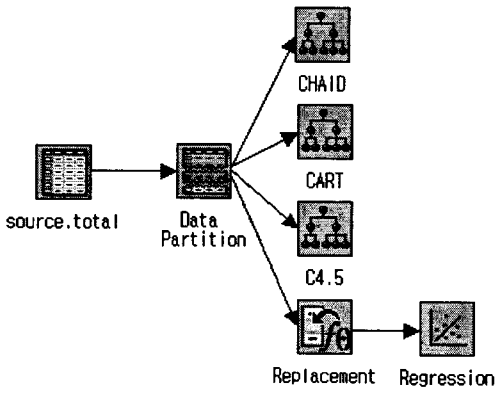
이차적인 전략은 일차적인 전략에서 선별된 핵심적 소수 변수의 후보 변수들만을 가지고, 로지스틱 회귀모형의 단계적 방법을 적용하여 핵심적 소수 변수들을 선별하는 것이다. 첫 번째 모형으로 CART, C4.5, CHAID 알고리즘을 적용하여 구해진 분류나무의 축차적인 분리에 사용된 변수들을 모두 모아서 로지스틱 회귀분석에 의해서 선별될 변수들의 후보 목록을 만든다. 분석시 목록에 있는 변수들 중에서 적어도 하나의 결측치를 가지고 있는 변수들의 자료점은 자동적으로 제외된다. 이때 나무구조를 반영하여 변수들 간의 교호작용효과를 선별적으로 후보 변수들의 목록에 추가하여, 교호작용 항을 갖는 모형을 만들 수 있다. 세 번째 모형은 일차적인 전략에서 분석용 자료의 결측치들을 대체 노드를 통해서 대체한 후에 로지스틱 회귀분석의 단계적 방법을 통해서 선별된 핵심적 소수 변수들의 후보 변수들만으로 구성되고, 첫 번째 모형과 동일한 방법에 의해서 핵심적 소수 변수들을 선별한다.

## 3. 기업 부도 여부 예제를 통한 핵심적 소수의 변수 선별

예제 자료는 기업 상태가 정상인 기업체와 부도가 난 기업체의 재무와 경영 상태에 대한 자료로서, 부도 여부를 베르누이 확률변수인 good으로 정의하고, 부도를 낸 경우에 good은 0이다. 설명변수들은 재무 비율 변수들과 경영 상태에 관한 변수들로 구성된다. 분석에 사용되는 자료의 개수는 32231개이며, 자료의 보안상 입력변수는 X2 - X146까지로 명명한다.

의사결정나무분석과 로지스틱 회귀분석을 수행하는 프로그램으로는 SAS E-miner, Clementine 등 여러 통계 프로그램이 있는데, 본 논문에서는

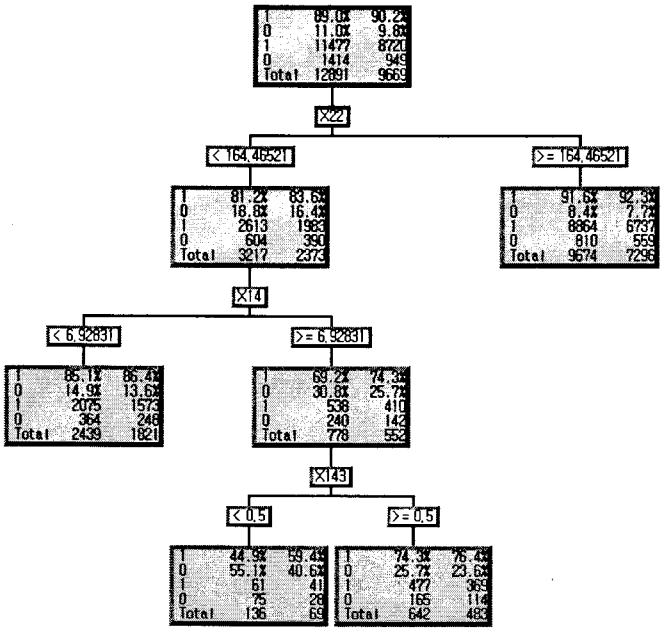
SAS E-miner를 이용하여 분석하고자 한다.




<그림 1> 일차적인 변수 선택

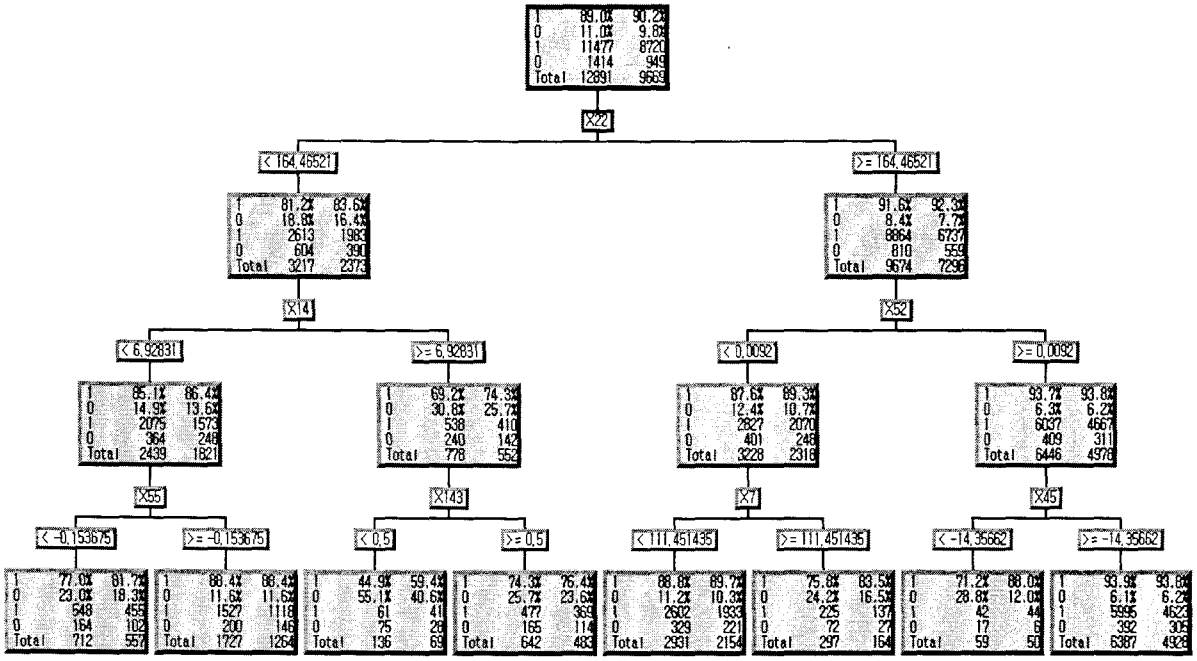
자료 분할(data partition) 노드에서 분석용 자료(training dataset), 타당성 자료 (validation dataset) 그리고, 검증용 자료(test dataset)를 각각 40%, 30%, 30%로 분리하였다. 그리고 로지스틱 회귀분석의 단계적 방법(stepwise method)으로 1차적인 변수선별을 하기 위해 대체(replacement) 노드를 통해 결측치를 연속형 입력변수의 경우 각 값의 평균값으로, 그리고 범주형 변수의 경우 최빈값으로 각각 그 값을 대치하였다. 단, 의사결정나무를 이용하여 변수선별을 하는 경우에는 결측치 대체 단계를 거치지 않는다.

먼저 CHAID 알고리즘은 Tree의 Basic Tab에서 Chi-squared test를 선택함으로서 실행할 수 있는데, 나무의 크기를 결정하는 역할을 하는 유의수준으로는 default값인 0.2를 주었다. CHAID 알고리즘 수행 결과 얻어진 분류나무는 <그림 2>에 주어지며, 변수 X22, X14, X143이 선별되었다.



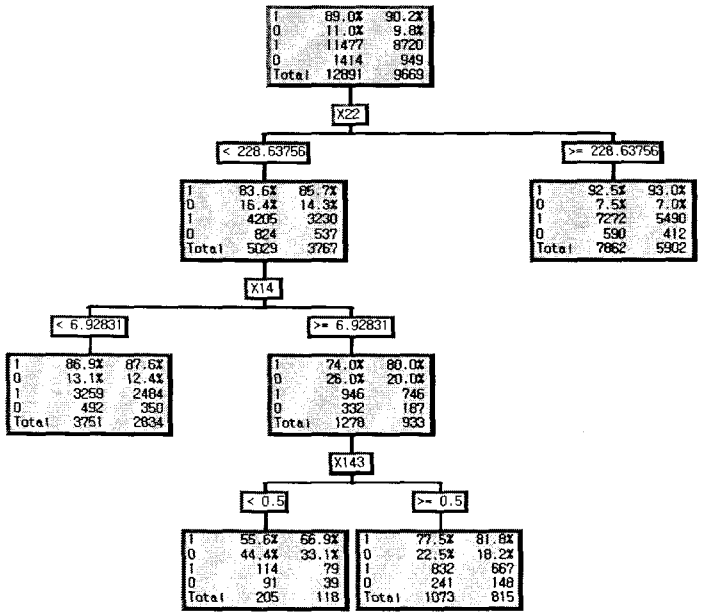

<그림 2> CHAID 알고리즘 수행으로 얻어진 분류나무

다음으로 Tree의 Basic Tab에서 Gini Reduction을 선택하고 Advanced Tab에서 Model Assessment Measure로 Total Leaf Impurity(Gini index)를 선택하여 CART 알고리즘을 실행한다. 이 때, Total Leaf Impurity 옵션은 생성되는 마디에 대한 Gini 계수의 합이 최소가 되도록 나무 구조를 결정한다. CART 알고리즘을 실행시킨 결과 얻어진 분류나무는 <그림 3>에 주어지며, 변수 X22, X14, X52, X55, X143, X7, X45가 선별되었다.




<그림 3> CART 알고리즘 수행으로 얻어진 분류나무(depth 3까지 이용)

마지막으로 Tree의 Basic Tab에서 Entropy Reduction을 선택하여 C4.5 알고리즘을 실행한다. C4.5 알고리즘을 실행시켜 얻어진 분류나무는 <그림 4>에 주어지며, 변수 X22, X14, X143이 선별되었다. 우연히도 CHAID 알고리즘 실행에 의해서 선별된 변수는 같으나, 변수 X22의 임계값이 164.46과 228.64로 서로 다름을 알 수 있다.




<그림 4> C4.5 알고리즘 수행으로 얻어진 분류나무

우선 2절에서 제시된 일차적인 전략에 의해서 핵심적 소수 변수들을 선별해 보자. CART, C4.5, CHAID 알고리즘을 각각 실행하여 선별된 변수를 정리하면 <표 1>과 같다. 분리기준으로 선택되지는 않았지만, 설명력이 뛰어난 후보 변수들을 추가하기 위해서 첫 번째로 분리되는 노드의 각각의 분류나무에서 상위 2개의 변수들을 선택하기를 제안한다. 9개의 변수 중 X13과 X21 두 개의 변수는 가장 처음 분리되는 노드에서 View competing split을 클릭하여 나오는 상위 2위와 3위 변수에 해당된다. 1위 변수는 분류나무를 처음 분리하는데 이미 사용되었다. 각 변수가 나타난 빈도수에 의해서 변수들을 선별하면, 세 가지 알고리즘에 의해서 모두 선별된 X13, X14, X21, X22, X143이다.

결측치를 대체시킨 분석용 자료를 가지고, 유의수준 0.05에서 로지스틱 회귀분석의 단계적 방법에 의해 중요변수를 선별하면 X2, X4, X6, X12, X14, X25, X37, X59, X68, X101, X104, X116, X120, X126, X143의 변수가 선택된다.

<표 1> CHAID, CART, C4.5 알고리즘을 적용하여 선별된 변수

| | X7 | X13 | X14 | X21 | X22 | X45 | X52 | X55 | X143 |
|---|---|---|---|---|---|---|---|---|---|
| CHAID | | o | o | o | o | | | | o |
| CART | o | o | o | o | o | o | o | o | o |
| C4.5 | | o | o | o | o | | | | o |

이차적인 전략으로, 일차적인 전략에서 선별된 핵심적 변수들의 후보 변수들의 결측치를 포함하지 않는 자료만으로 구성된 분석용 자료를 가지고 로지스틱 회귀분석의 단계적 방법(stepwise method)을 적용하여 적절한 모형을 선택하여 보자. 우선, <표 1>에 나타난 의사결정나무에 의하여 선별된 9개 변수들로 로지스틱 회귀 모형의 후보 변수들의 목록을 구성한다. 9개 변수들 중에서 변수 X13, X14, X21, X45, X52, X143이 선별되었고, 분석결과가 <표 2>에 주어진다.

<표 2>로지스틱 회귀 모형의 단계적 방법에 의한 분석 결과 (주효과모형)

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | Standardized Estimate | Exp(Est) |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -2.0743 | 0.0467 | 1975.82 | <.0001 | | 0.126 |
| X13 | 1 | 0.00739 | 0.00199 | 13.87 | 0.0002 | 0.062 | 1.007 |
| X14 | 1 | 0.0252 | 0.00522 | 23.2 | <.0001 | 0.0821 | 1.025 |
| X143 | 1 | -0.00572 | 0.00133 | 18.47 | <.0001 | -0.1638 | 0.994 |
| X21 | 1 | -0.0109 | 0.00228 | 22.93 | <.0001 | -0.0881 | 0.989 |
| X45 | 1 | -0.00113 | 0.00047 | 5.82 | 0.0158 | -0.0527 | 0.999 |
| X52 | 1 | -0.2811 | 0.0549 | 26.24 | <.0001 | -0.1728 | 0.755 |

이번에는 CART, C4.5, CHAID 알고리즘을 각각 실행하여 선별된 7개 변수와 각각의 분류나무의 구조를 반영하여 선별적으로 선택된 변수들의 교호작용항들을 추가하여, 후보 변수 항들의 목록을 구성하여 로지스틱 회귀 모형의 단계적 방법(stepwise method)에 의하여 적절한 모형을 선택해 보자. 예를 들어서 CHAID 분류나무로부터, X22*X14, X22*X14*X143 항이 추가된다. 단계적 방법에 의해서 모형 선택을 한 결과가 <표 3>에 주어진다.

<표 3> 로지스틱 회귀 모형의 단계적 방법에 의한 분석 결과 I
(나무 구조에 의해서 선택된 교호작용항들을 포함)

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | Standardized Estimate | Exp(Est) |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -2.2484 | 0.0561 | 1605.99 | <.0001 | | 0.106 |
| X14 | 1 | 0.0804 | 0.00806 | 99.48 | <.0001 | 0.2628 | 1.084 |
| X143 | 1 | -0.00666 | 0.00143 | 21.64 | <.0001 | -0.1906 | 0.993 |
| X22_X14 | 1 | -0.00011 | 9.81E-06 | 126.41 | <.0001 | -0.3566 | 1 |
| X22_X14_X143 | 1 | -1.22E-06 | 3.69E-07 | 10.92 | 0.0010 | -0.0897 | 1 |
| X22_X52 | 1 | -0.00016 | 0.000038 | 17.08 | <.0001 | -0.2354 | 1 |
| X52 | 1 | -0.2705 | 0.053 | 26.06 | <.0001 | -0.1663 | 0.763 |

회귀분석에서는 일반적으로 교호작용항에 포함되는 변수에 대해서는 개별적인 변수도 모형에 포함시킨다. 따라서 위의 모형에 X22를 포함하여 다시 로지스틱 회귀분석을 수행하면 <표 4>가 얻어진다.

<표 4> 로지스틱 회귀 모형의 단계적 방법에 의한 분석 결과 II
(나무 구조에 의해서 선택된 교호작용항들을 포함)

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | Standardized Estimate | Exp(Est) |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -2.2672 | 0.0586 | 1497.03 | <.0001 | | 0.104 |
| X14 | 1 | 0.0826 | 0.00851 | 94.31 | <.0001 | 0.2697 | 1.086 |
| X143 | 1 | -0.00663 | 0.00144 | 21.27 | <.0001 | -0.1897 | 0.993 |
| X22 | 1 | 0.000037 | 0.000028 | 1.75 | 0.1858 | 0.0308 | 1 |
| X22_X14 | 1 | -0.00011 | 0.00001 | 115.05 | <.0001 | -0.3651 | 1 |
| X22_X14_X143 | 1 | -1.38E-06 | 3.82E-07 | 12.97 | 0.0003 | -0.1013 | 1 |
| X22_X52 | 1 | -0.00016 | 0.000038 | 16.48 | <.0001 | -0.2326 | 1 |
| X52 | 1 | -0.273 | 0.0557 | 24.03 | <.0001 | -0.1678 | 0.761 |

X14, X143, X52가 중요한 변수로 선별되었고, X22는 p-value가 매우 큰 값을 가지지만, 이 두 변수들을 포함하는 교호작용항들이 반응변수인 부도 여부를 설명하는데 중요한 역할을 하기에 추가되었다. 따라서 선별된 핵심적 소수 변수로 X14, X22, X52, X143이 선별된다.

마지막으로 일차적인 전략에서 로지스틱 회귀분석의 단계적 방법에 의해 선별된 변수들의 목록을 가지고, 다시 한 번 단계적 방법으로 선별하면 X2, X12, X14, X101, X104, X116, X120, X126, X143가 중요변수라는 결론을 내릴 수 있다. 그 결과는 아래의 <표 5>와 같다.

<표 5> 로지스틱 모형의 단계적 방법에 의한 분석결과 (로지스틱 모형)

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | Standardized Estimate | Exp(Est) |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -1.7545 | 0.0748 | 549.91 | <.0001 | | 0.173 |
| X101 | 1 | 0.000166 | 0.000067 | 6.16 | 0.0131 | 0.0276 | 1 |
| X104 | 1 | -0.0274 | 0.00838 | 10.71 | 0.0011 | -0.072 | 0.973 |
| X116 | 1 | 0.1032 | 0.00955 | 116.92 | <.0001 | 0.1723 | 1.109 |
| X12 | 1 | -0.00842 | 0.00231 | 13.28 | 0.0003 | -0.0666 | 0.992 |
| X120 | 1 | -0.00213 | 0.000635 | 11.26 | 0.0008 | -0.0785 | 0.998 |
| X126 | 1 | 0.000751 | 0.000152 | 24.35 | <.0001 | 0.0938 | 1.001 |
| X14 | 1 | 0.00849 | 0.00388 | 4.77 | 0.0289 | 0.027 | 1.009 |
| X143 | 1 | -0.00686 | 0.00157 | 19.04 | <.0001 | -0.1803 | 0.993 |
| X2 | 1 | -0.0396 | 0.00607 | 42.66 | <.0001 | -0.1246 | 0.961 |

이제 <표 2>의 주효과 모형과 <표 4>에 주어진 교호작용효과 포함 모형, <표 5>의 로지스틱모형에 대해서 사후확률 0.5를 기준으로 계산한 분석결과를 비교해보기로 하자.

<표 6> 세 모형의 ASE 비교

| | Root ASE(training) | Root ASE(validation) | Root ASE(test) |
|---|---|---|---|
| simple | 0.310885 | 0.29641 | 0.296658 |
| interaction | 0.309136 | 0.295694 | 0.295436 |
| logistic2 | ***0.308146*** | ***0.295603*** | ***0.292532*** |

먼저 ASE(Average Squared Error)값을 비교해 보면, 훈련용, 타당성, 검증용 자료에서 모두 <표 5>에 주어진 로지스틱모형이 근소하게 작은 값을 가져 약간 우월하다. 다음으로 오분류율(Misclassification rate)을 비교해본 결과는 <표 7>에 주어지는데, 타당성 자료를 제외한 훈련용, 검증용 자료에서 모두 <표 5>에 주어진 로지스틱 모형이 우월하다.

<표 7> 두 모형의 오분류율 비교

| | Misclassification Rate(training) | Misclassification Rate(validation) | Misclassification Rate(test) |
|---|---|---|---|
| simple | 0.110697 | ***0.098562*** | 0.098656 |
| interaction | 0.111008 | 0.099803 | 0.098966 |
| logistic2 | ***0.110465*** | 0.099597 | ***0.098552*** |

그러나 이와 같은 수치는 대부분 소수점 셋째자리 이하의 매우 적은 차이이고, 전체 자료에서 good=0인 즉, 부도기업의 비율이 10%정도인데 반해서 부도기업으로 판정한 자료수가 0.4% 이하인 점을 감안했을 때 사후확률 0.5를 기준으로 계산한 ASE와 오분류율에 큰 의미를 두는 것은 올바른 판단이 아니라 할 수 있다. 이제 자료 분석의 결과에서 부도인 기업을 부도가 아니라고 판단하는 경우

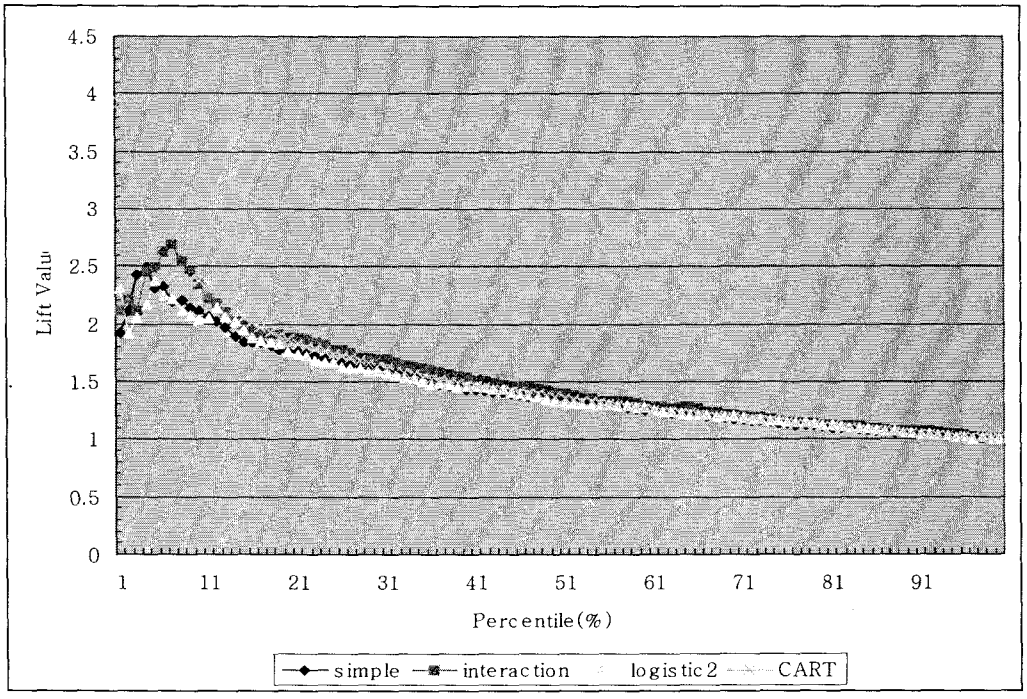
와 부도가 아닌 기업을 부도라고 판단하는 오류 중에서 어느 것이 손실이 클 지에 대해서 생각해 보자. 전자의 경우 기업의 재무상태와 경영자료를 보고 부도가 아니라고 판단하여 기업에 대출을 하는 사례가 발생할 수 있다. 이는 후자의 경우보다 매우 위험도가 큰 오류이므로 전자의 오류에 대한 손실 점수를 10배인 10점을 부과하고, 후자의 경우 1점을 부과하기로 한다. 각각의 모형에 대하여 기대손실을 가장 작게 하는 사후 확률과 기대 손실 값을 비교해 보자.

<표 8> 모형별 분류기준값에 따른 기대손실

| | 분류기준값 | 기대손실(training) | 기대손실(validation) | 기대손실(test) |
|---|---|---|---|---|
| simple | 0.111 | 0.739 | 0.777 | 0.765 |
| interaction | 0.104 | 0.718 | 0.744 | 0.745 |
| logistic2 | 0.098 | 0.771 | 0.764 | 0.745 |

기대손실을 보고 판단할 경우 나무 구조에 의해서 선택된 교호작용항들을 포함하는 모형으로부터 구해진 모형인 interaction 이 훈련용 자료, 타당성 자료와 검증용 자료에서 모두 우월하게 나타났다.

은행에서는 부도 가능성이 높은 기업을 우선적으로 선별하여 조사한다고 하자. 이 경우에 부도를 낼 확률인 사후 확률의 크기 순서로 정렬하여, 하나씩 조사하는 경우에 부도인 기업을 찾을 가능성을 랜덤하게 선택하는 경우에 부도 기업을 찾을 가능성과 비교하는 Lift Chart를 각각의 모형에 대해서 비교해 볼 수 있다. 검증용 자료(test dataset)에 대해서 3개 모형의 Lift Chart와 일차적인 전략에서 구한 분류나무들 중에서 가장 Lift Chart가 좋게 나타나는 CART 분류나무의 Lift Chart 를 비교해 보자. 3개 모형들의 Lift Chart 중에서는 교호작용 효과를 포함하는 모형인 interaction이 전반적으로 우월하게 나타났지만, 상위 4% 까지는 CART에 의한 선별이 효과적임을 확인할 수 있다.




<그림 5> 각각의 모형에 대한 Lift Chart

## 4. 요약

실처리 자료인 대용량의 자료를 이용하여, 최종 검사의 불합격이나, 기업체의 부도 여부 등과 같은 특정 현상에 영향을 주는 공정의 실처리 변수들이나 재무 비율 변수들을 선별하고, 선별된 변수들을 활용하여, 앞으로의 특정 현상의 발생 가능성을 예측하고 싶을 때가 많다. 즉, 관심사는 품질 특성치에 영향을 주리라 기대되는 사소한 다수(trivial many)의 변수들 중에서 결정적으로 영향을 주는 핵심적 소수(vital few)의 변수들을 선별하는 것이다. 핵심적 소수 변수를 선별하는 간결한 방법의 하나는(일차적인 전략은) 분류나무의 잘 알려진 알고리즘인 CART, CHAID와 C4.5을 적용하여 구해진 분류나무의 축차적인 분리에 사용된 변수들을 선별하고, 3개의 분류나무에 공통적으로 관련된 변수들을 결정하기 위해서 변수들의 투표를 통하여 핵심적 소수의 변수들을 선별하는 것이다. 기업체의 부도 여부와 재무 비율, 경영 상태에 관한 금융 자료를 가지고, 바람직한 로지스틱 회귀모형을 구축하기 위한 3가지 방법을 비교하였고, 각각의 분류나무에서 선별된 변수와 나무구조를 반영하여 변수들 간의 교호작용효과를 선별적으로 후보 변수들의 목록에 추가하여, 로지스틱 회귀분석의 단계적 방법에 의해서 결정된 interaction 모형이 기대손실과 Lift Chart의 관점에서 제안되었다.

## 참고문헌


[1] 강현철 등(1999), 「데이터마이닝, 방법론 및 활용」, 자유아카데미.

[2] 임용빈, 오만숙(2002), "분류와 회귀나무 분석에 관한 소고", 「품질경영학회지, 30권 1호, pp. 152-161.

[3] 허명회, 이용구(2003) 「데이터마이닝 모델링과 사례」, SPSS 아카데미.

[4] Abt, M., Lim, Y.B., Sacks, J., Xie, M. and Young, S. (2001), A sequential approach for identifying lead compounds in large chemical databases, vol. 16, No. 2, 154-168

[5] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees, Chapman and Hall, Belmont, CA, Wadsworth.

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

[7] Quinlan, J.R. (1993). C4.5 Programs for machine learning. San Mateo: Morgan Kaufmann.