Journal of the Korean Institute of Industrial Engineers
Vol. 36, No. 1, pp. 13-21, March 2010.

# 불균형 이분 데이터 분류분석을 위한 데이터마이닝 절차

정한나 · 이정화 · 전치혁†

포항공과대학교 산업경영공학과

# A Data Mining Procedure for Unbalanced Binary Classification

Han-Na Jung · Jeong-Hwa Lee · Chi-Hyuck Jun

Department of Industrial and Management Engineering, Pohang University of Science and Technology

The prediction of contract cancellation of customers is essential in insurance companies but it is a difficult problem because the customer database is large and the target or cancelled customers are a small proportion of the database. This paper proposes a new data mining approach to the binary classification by handling a large-scale unbalanced data. Over-sampling, clustering, regularized logistic regression and boosting are also incorporated in the proposed approach. The proposed approach was applied to a real data set in the area of insurance and the results were compared with some other classification techniques.

***Keywords:*** Clustering, Large-scale Data, Over-sampling, Regularized Logistic Regression, Unbalanced Data

## 1. 서 론

보험업계의 경우 고객의 보험계약 체결이 재정에 직접적인 영향을 주며 계약의 해지 또는 이탈을 예측하는 것은 중요한 역할을 한다. 이탈고객 예측은 주로 데이터마이닝 기법 중 분류분석(Classification)을 통하여 이루어지며 모델을 구축하기 위해서는 충분한 학습데이터를 확보하여야 한다. 그러나 관심이 되는 이탈고객(이를 목표클래스라 함)은 전체 고객 데이터에서 적은 부분만을 차지하기 때문에 기존의 방법으로는 효과적인 모델을 만들기 어려울 뿐 아니라 데이터의 사이즈가 크기 때문에 분석에 시간이 많이 소요된다. 분류분석에서 클래스를 지속고객과 이탈고객으로 구분할 때 두 클래스의 관측수가 현저하게 차이가 있는 경우를 불균형(Unbalanced; Imbalanced; Skewed) 데이터라 한다.

이와 같은 불균형 데이터를 사용하여 분류분석 할 때 관측수가 많은 클래스의 데이터가 분류기 생성에 지배적으로 작용하게 된다. 그러나 데이터가 적은 클래스의 정보 역시 중요하기 때문에 모델링에서 어려움을 내포한다. 고객 이탈예측이나 신용사기 예측 등에 대한 기존 연구에서 크게 세 가지를 고려하고 있다. 첫 번째로 어떤 분류분석을 사용할 것인지를 결정하여야 한다. 주로 이용되는 분류분석 방법은 크게 의사결정나무(Decision tree)를 이용한 방법, 신경망(Neural network)을 이용한 방법 또는 새로운 분석기법을 제시한 것으로 나누어 볼 수 있다. 두 번째로 불균형 데이터를 다루기 위해 필요한 절차를 결정해야 하며, 세 번째로 모델검증에 어떤 척도를 사용할지를 고려해야 한다.

위에서 언급한 분류분석 방법들은 종종 복합되어 사용된다. 이탈예측에 분류분석을 사용한 연구로는 다음과 같은 것들이 있다. Hung *et al.*(2006)은 K-means 방법을 이용해 고객을 군집하고 의사결정나무와 신경망을 사용하여 통신 고객이탈을 분석 하였으며, Datta *et al.*(2000)은 Forward stepwise selection을 사용하여 데이터의 차원을 줄이고 의사결정나무와 신경망을 사용하였다. 그 밖에도 Wei and Chiu(2002) 은 의사결정나무를 사용하여 분석하였으며, Coussement and Van den Poel(2008)는 Support vector machine, Random forest 및 로지스틱 회귀분석 등을 사용하고 비교하였으며, Mozer *et al.*(2000)은 로지스틱 회귀분석, 의사결정나무, 신경망, 부스팅을 사용하였다. Viaene *et al.*(2002)는 C4.5, Naïve Bayes, 로지스틱 회귀분석을 사용하였으며, Au *et al.*(2003)는 evolutionary learning을 통한 데이터마이닝 방안을 제안하고 C4.5, 신경망과 비교하여 이탈고객예측

이이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2009-0072598).
† 연락저자 : 전치혁 교수, 790-784 경북 포항시 남구 효자동 산 31 포항공과대학교 산업경영공학과, Tel : 054-279-2197, Fax : 054-279-2870, E-mail : chjun@postech.ac.kr
2009년 8월 5일 접수; 2010년 2월 2일 수정본 접수; 2010년 2월 9일 게재 확정.

에 이용하였다. 이탈고객 예측과 유사한 신용카드 사기 예측에 이용된 연구로 Phua *et al*.(2004)은 Back propagation 신경망, Naïve Bayes, C4.5가 효율성, 확장성, 속도에서 각기 장단점이 있는 것을 이용하여 이들을 결합한 복합 분류 시스템(Multiple classifier system)을 제안하였다. 이 논문에서 하나의 분류기를 선택하는 것보다는 몇 개의 분류기를 결합하는 Stacking-bagging이 더 낫다는 결과를 보였으며 이를 Meta-learning이라 하였다.

한편, 클래스별 불균형 데이터를 처리하기 위해서 샘플링을 이용하거나 가중치를 주는 방법이 이용되어 왔다. Kubat and Marwin(1997)은 One-sided sampling를 제안하였는데 이것은 데이터의 수가 적은 목표 클래스의 모든 데이터를 포함하고 데이터가 많은 클래스의 데이터 중에서는 클래스의 경계부분에 있는 데이터들만 샘플링하여 불균형 데이터에 이용하는 것이다. Stolfo *et al*.(1997) 또한 샘플링을 사용하여 불균형 문제에 적용하였는데 이 때에 목표 클래스의 비율을 달리하며 분류분석하였다. Pazzaniet *et al*.(1994)은 가중치를 사용하여 Training 데이터 각각에 다른 중요도를 부여하여 분석하였고 Gorden and Perlis(1989)는 클래스별로 다른 비용을 할당하는 방법으로 분석하였다. 그 외에도 Windowing와 Bootstrapping을 이용한 연구가 있다(Catlett, 1991; Sung and Poggio,1995).

데이터마이닝에서 방대한 데이터를 이용하는 경우가 많은데 기존 소프트웨어와 하드웨어의 성능이 이와 같은 방대한 데이터의 분석에 적합하지 않아 속도를 향상시키는 것도 중요한 문제이다. 따라서, 불균형한 데이터이면서 방대한 데이터인 경우 정확도뿐 아니라 속도 측면도 고려를 하여야 하며 이를 해결하기 위해 데이터마이닝 기법들이 요구되고 있다. 본 논문에서는 대용량 데이터를 처리하는 데 효과적으로 알려져 있는 Trust region Newton method를 적용한 로지스틱 회귀분석 기법을 사용하며 불균형한 데이터에서의 예측정확도를 높이기 위해 샘플링, 군집분석, 부스팅을 이용하는 새로운 데이터마이닝 절차를 제안한다. 제안된 절차를 보험회사의 이탈고객 예측에 적용하였으며 제안된 방법을 의사결정나무 또는 선형판별분석과 비교하여 제안된 방법의 타당성을 보였다.

이 후 본 논문의 구성은 다음과 같다. 제 2장은 제안분석절차의 기술로 데이터 탐색과정, 샘플링, 군집분석, 로지스틱 회귀모형, 부스팅, 모델의 정확도를 산출하는 과정을 설명한다. 제 3장은 적용사례를 소개하며 제안된 분석절차를 거침에 따라 분류분석한 결과가 어떻게 달라지는지 알아본다. 마지막으로 제 4장에서는 본 논문의 의미를 요약하고 장단점을 정리한다.

## 2. 제안분석절차

분류분석 시 가장 먼저 이루어지는 것은 데이터의 정제 과정이며, 다음으로 분류기법을 적용하여 분류기를 학습하고, 마지막으로 새로운 데이터에 대하여 클래스를 예측하게 된다. 그러나, 보험이탈고객 예측과 같은 문제에는 이와 같은 단순한 절차가 적당하지 않은데 이것은 고객정보 데이터가 대용량이며 불균형이라는 특성 때문이다. 데이터의 양이 적은 목표 클래스를 효과적으로 분류하기 위하여 추가적인 데이터마이닝 기법을 사용하여야 한다. 본 연구에서는 전처리(Preprocessing), 샘플링, 군집분석, Regularized 로지스틱 회귀분석, 부스팅의 단계를 거쳐 분석하였으며 개요는 <그림 1>과 같다. 추가된 기법들은 불균형 문제를 처리하기 위해, 모델의 정확도를 높이기 위해 사용되었다. 아래에 각 단계를 자세히 기술하도록 하겠다.

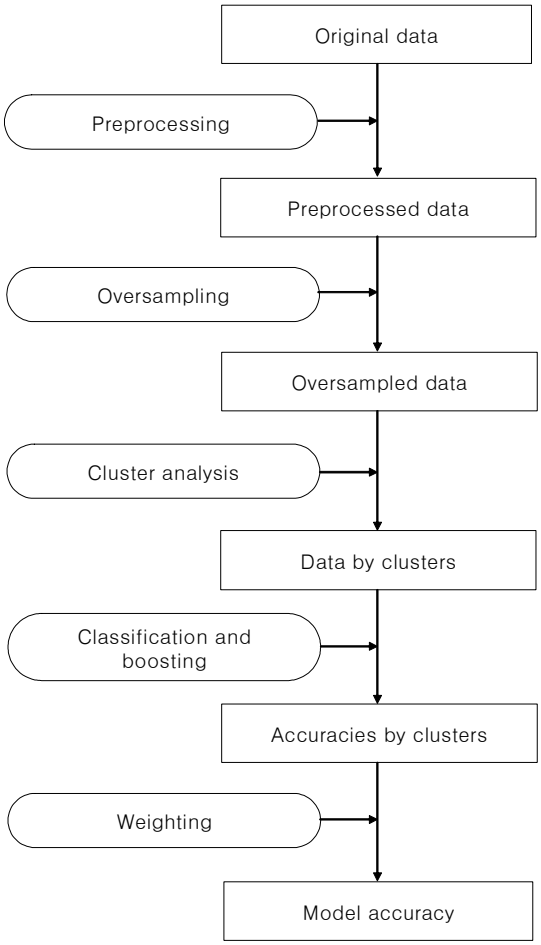


**Figure 1.** Proposed Classification Procedure

### 2.1 데이터 탐색 및 전처리

우선 데이터의 최소값, 최대값, 평균, 분산, 상관관계 등의 특징을 파악하고 이상치를 제거하는 것이 필수적인 과정이다. 데이터의 배경이나 목적에 따라 결측치 처리, 이상치 제거, 변수 재정의 등이 있을 수 있다. 데이터의 성질에 따라 단위를 바꾼다거나 명목, 서열, 구간 데이터의 형태를 적절하게 변경한다든지 일자를 기간으로 나타내는 등의 전처리를 실행하여 데이터가 목적을 좀 더 잘 나타내도록 조정한다.

### 2.2 샘플링

관측수가 작은 클래스 데이터는 모두 사용하고 큰 클래스는 샘플링으로 일부를 사용하는 것을 언더샘플링(Under-sampling)이라 하며, 관측수가 큰 클래스를 모두 사용하고 작은 클래스의 관측수를 증대시키는 것을 오버샘플링(Over-sampling)이라 한다. 기존에 이탈고객 예측을 위한 많은 연구가 존재하였지만 대용량 데이터를 모두 사용하기보다는 언더샘플링을 통해 사이즈를 줄이고 불균형 데이터를 보정하여 계산이 가능하게 한 경우가 많았다. 이런 경우 데이터가 많은 클래스의 정보를 어느 정도 손해본채 균형이 맞는 데이터를 만들게 된다. 언더샘플링을 하게 되면 데이터의 사이즈가 줄어들기 때문에 메모리나 처리속도 측면에서 유리할 수 있다. 그러나 대용량 데이터 분석이 기술과 메모리 등으로 충분히 가능하다면 굳이 샘플링을 통해 정보의 손실을 감수할 필요가 없다. 본 연구에서는 대용량 데이터의 문제를 추후에 설명할 새로운 최적화 알고리즘으로 처리하기 때문에 오버샘플링을 사용한다.

오버샘플링에는 이탈고객 데이터에 노이즈를 발생시켜 새로운 데이터를 얻는 방법과 기존 데이터를 중복하여 사용하는 방법이 있다. 노이즈를 발생시킨 새로운 데이터는 보다 현실감이 있으나 노이즈 정도에 따른 결과해석상의 복잡도가 야기되므로 본 연구에서는 오버샘플링 방법 중 데이터를 중복하여 사용하였고 적은 클래스의 데이터를 반복하여 전체 데이터에 포함시킨다. 오버샘플링의 비율이 너무 커지면 계산량이 증대되고 본 데이터와 특성이 달라질 수 있으므로 분석에 필요한 최소한의 중복을 허용하도록 한다. 목표 클래스의 비율을 10%씩 늘려가면서 분류분석을 적용하여 학습데이터에 대한 목표정확도의 상승이 관측될 때까지 오버샘플링하는 것을 추천한다.

### 2.3 군집분석

고객의 특성이 다양할 수 있기 때문에 모든 데이터를 하나의 분류기로 처리하기 보다는 군집분석을 통하여 고객을 그룹핑하고 각 군집별로 분류기를 만드는 방안을 사용한다. 이 경우 특정 관측치에 대해 우선 속할 군집을 찾고 해당 군집에 따라 적합한 모델을 선택하여 클래스를 예측하게 된다. 이러한 방안은 Vapnik(1998)에 의해 제안된 Transduction이란 개념으로 전체 데이터에 맞는 한 가지 모델을 만드는 것이 아니라 특정 데이터 그룹에 맞는 다수의 모델을 생성하는 것이다. 군집별 분류기는 유사한 특징을 가진 고객군 중에서 이탈과 지속고객을 나누는 요소를 좀 더 반영시킴으로 써 각 군집별 정확도를 높이며 궁극적으로 전체 정확도를 높일 수 있다(Heo *et al*., 2008). 본 연구에서는 군집분석 방법 중 속도가 빠르며 이해와 구현이 쉽다는 장점을 가진(Tan *et al*., 2006) K-means 기법을 사용하여 고객군을 구분한다.

### 2.4 로지스틱 회귀모형에 의한 분류분석

군집분석을 통해 나뉘어진 각 군집에 대하여 다음에 설명할 조정 로지스틱 회귀모형(Regularized logistic regression model)을 이용하여 분류분석을 한다. 제안된 절차에서 반드시 로지스틱 회귀모형을 사용할 필요는 없으나 대용량 데이터 하에서 계산시간을 단축시킬수 있는 최적화방법이 이 모형에 대해 가용하기 때문이다. 두 개의 클래스(±1로 표기)를 갖는 통상의 이분 로지스틱 회귀모형은 식 (1)과 같은 구조를 가진다.

$$P\{y_i = \pm 1\} = \frac{1}{1+\exp(-y_i(w^T x_i))} \quad i = 1, \cdots, N \quad (1)$$

여기서 $y_i$는 i번째 고객의 클래스를 나타내는 종속변수이며 $x_i$는 i번째 고객의 속성들을 나타내는 독립변수 벡터, $w$는 관련 회귀계수 벡터이다. 그리고 $N$은 고객의 수이다.

$w$를 구하기 위해서 최우추정법을 사용하는데 $w$의 절대값이 크게 증가하여 특정 데이터에 대해 과적합 (overfit)되는 것을 방지하기 위해 종종 $w$의 절대값 또는 제곱합에 제약을 가하는 조정 로지스틱 회귀분석을 사용한다. 식 (1)의 역수를 고려하고 조정항을 추가하면 이 모형에 대한 최소화시켜야 할 우도 함수는 다음과 같다(Koh *et al*., 2007).

$$f(w) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \log\left[1+\exp(-y_i(w^T x_i))\right] \quad (2)$$

여기서, $C$는 양의 값을 가지는 상수로 조정항과의 균형을 맞추는 역할을 한다.

식 (2)의 우도함수식을 최소화시키는 해를 구하는 데에는 Iterative scaling(Darroch and Ratcliff, 1972; Della Pietra *et al*., 1997; Goodman, 2002; Jin *et al*., 2003), Nonlinear conjugate gradient, Quasi-Newton(Liu and Nocedal, 1989; Benson and Moré, 2002), Truncated Newton(Komarek and Moore, 2005) 등의 방법이 있을 수 있다. 이 중 Quasi-Newton의 방법인 Limited Broyden-Fletcher-Goldfarb-Shanno(L-BFGS)가 데이터의 규모가 클 때에 적합하다는 것이 알려져 있었는데(Malouf, 2002; Sutton and McCallum, 2006), 그 후 Truncated Newton방법의 발전적 형태인 Trust Region Newton Method(TRON)가 대용량 데이터 분석에서 L-BFGS와 같은 결과를 내면서 수렴속도가 훨씬 빠르다는 것이 입증되었다(Lin *et al*., 2008).

따라서 본 연구에서는 TRON을 통하여 식 (2)를 최소화시키는 회귀계수벡터인 $w$를 구한다. 이때 TRON software(Lin *et al*., 2008)에서 제공하는 기본 파라미터를 이용한다. 고객의 클래스 예측을 위해서 식 (1)을 사용하여 +1 및 -1 클래스에 대한 확률을 산출하고 확률이 큰 클래스를 부여한다.

### 2.5 부스팅을 추가한 분류분석

부스팅은 데이터의 중요도를 반복적으로 조절하여 분류하

는 방법으로 분류분석의 정확도를 높이기 위하여 부가적으로 쓰인다(Tan *et al*., 2006). 부스팅의 대표적 알고리즘인 Freud and Schapire(1996)에 의해 제안된 Adaboost는 초기에 각 데이터에 동일한 중요도를 부여하고 클래스 예측 후 잘못 분류된 정도가 큰 데이터에 큰 중요도를 부여하는 과정을 반복한다.

본 연구에서도 Adaboost를 사용하였으며 반복은 20회 시행되었다. 조정 로지스틱 회귀분석을 통해 얻어진 계수를 학습데이터에 적용하여 그 결과 잘못 분류된 데이터에 가중치를 주고 다시 분류하게 된다. 보통 반복할 때마다 중요도를 재계산하는데 오분류율이 비교적 높은 경우 이전 중요도를 그대로 적용하고 새롭게 오분류된 데이터에 대하여 가중치를 추가적으로 산출 적용하는 것을 추천한다.

### 2.6 정확도

데이터마이닝 기법의 성능을 비교할 때에 정오분류표(Confusion matrix)와 연결된 전체정확도(Overall accuracy), 목표정확도(Target accuracy), ROC 곡선, Lift 등의 척도 등이 사용된다. 이분 클래스는 양(Positive)과 음(Negative)으로 구분할 수 있는데 양은 예측의 대상이 되는 보험이나 통신업의 이탈고객, 신용카드 사기거래 등을 의미하며 분류분석에서의 목표 클래스가 된다. 본 연구에서는 목표정확도를 주로 사용코자 한다.

<표 1>은 정오분류표의 일반적인 구조를 나타낸다. True positive(TP)는 실제 양인 클래스를 양으로 분류한 것이고 False positive(FP)는 실제 음인데 양으로 잘못 분류한 것이다. 이때, 전체정확도는(TN+TP)/(TN+TP+FN+FP)로 계산하며 목표 정확도는 True positive rate(TPR)와 동일한 TP/(TP+FP)로 나타낸다. 그밖에, FP/(FP+TN)로 계산되는 False positive rate(FPR) 등이 있으며 TPR과 FPR의 관계를 그래프로 그린 것을 ROC 곡선이라 한다.

기존의 연구에서 Stolfo *et al*.(1997)는 True positive와 False positive를 척도로 사용하였으며 이것이 전체 정확도보다 불량고객을 탐지하는 데에 더 효과적임을 보인 바 있다. Viaene *et al*.(2002)은 ROC 곡선을 사용하여 분류분석 모델을 비교하였고 Lift는 Datta *et al*.(2000)의 논문과 Hung *et al*.(2006), Mozer *et al*.(2000)의 논문에서 사용된 바 있다. Wei and Chiu(2002)는 Detection error tradeoff 곡선을 사용하여 FPR과 FNR간의 트레이드 오프를 나타내었다.

**Table 1.** Confusion Matrix of Binary Classification

| | | Classified | |
|---|---|---|---|
| | | Positive | Negative |
| True: | Positive | True positive(TP) | False negative(FN) |
| | Negative | False positive(FP) | True negative(TN) |

앞에서 기술한 바와 같이 본 연구에서는 군집별로 다른 모델을 구축하기 때문에 군집 정확도가 우선 산출되며 이로부터 전체 정확도 또는 목표 정확도를 구하여야 한다. 후자를 모델 정확도라 칭할 때, 이는 군집 정확도 (전체 또는 목표)의 가중평균으로 도출된다. 즉, K를 군집수, $A_k$를 k 군집의 정확도라 할 때(k = 1, …, K) 모델 정확도는 다음 식으로 산출된다.

$$A_m = \frac{\sum_{k=1}^{K} N_k A_k}{\sum_{k=1}^{K} N_k} \tag{3}$$

여기서 $N_k$는 k 군집의 관측수이다. 테스트 데이터에 대한 정확도 역시 윗 식을 사용할 수 있다. 테스트 데이터인 경우 우선 각 관측치가 기존 군집 중에 어떤 군집에 가장 가까운지를 유클리디언 거리로 결정하고 해당 군집의 학습된 분류모델을 사용하여 클래스를 예측한다.

## 3. 사례연구

### 3.1 데이터 탐색

본 데이터는 특정 보험회사에서 일 년 사이에 얻어진 것으로 총 40,000명의 고객의 정보를 담고 있다. 고객의 특징을 나타내는 변수 35개는 <표 2>와 같다.

실험의 편의를 위해 변수명을 코드화하였으며 증권번호, 고객번호, 계약자 가입연령 등이 실제 변수명이다. 변수의 형태에는 구간(Interval), 명목(Nominal), 이분(Binary), ID 등이 포함되었다.

### 3.2 데이터 전처리

고객의 속성 외에 고객이 관찰된 시점에서 계약을 이탈하였는지 지속하였는지 나타나있는데, 이탈고객은 전체 40,000명 중 1,557명으로 3.89%를 차지하여 불균형데이터에 속하는 것을 알 수 있다.

고객의 속성 중 계약자 학력, 계약자 결혼유무, 계약자 결혼일자, 계약자 자녀수, 계약자 주거상태, 계약자 주거형태의 경우 결측정도가 적게는 65%, 많게는 98%에 달해 의미있는 정보를 얻기가 힘들기 때문에 사용하지 않았다. 납입기간이나 보험금 지급 만기일자의 경우 만기가 없는 보험이 존재하여 기록이 불가하였으나 최대 납입기간이나 최대 만기년도에 1을 더하여 임의로 만기를 산출하였다. 보험료의 금액은 소수의 고액 납입자가 존재하여 변수의 표준화가 어려웠으므로 백만원 이상을 납입한 경우는 백만 원에 최소액을 더한 금액으로 변경하였다. 약관 대출일자나 부활일자의 경우 존재하지 않을 수 있으므로 논리에 맞게 처리하였다. 또한 분석에 있어서 좀 더 의미있는 정보를 포함하게 하기 위해 몇 개의 변수들은 재정의하였는데 주로 주기나 기간, 액수, 나이 등의 시간의 흐름

**Table 2.** Variables Representing Customers Attributes

| 변수코드 | 변수명 | 형태 | 변수코드 | 변수명 | 형태 |
|---|---|---|---|---|---|
| C1_01 | 증권변호 | ID | C1_19 | 계약자결혼일자 | Interval |
| C1_02 | 고객번호 | ID | C1_20 | 계약자자녀수 | Interval |
| C1_03 | 계약자가입연령 | Interval | C1_21 | 계약자주거상태 | Nominal |
| C1_04 | 납입방법 | Ordinal | C1_22 | 계약자주거형태 | Nominal |
| C1_05 | 납입기간 | Interval | C1_23 | 계약자우편번호 | Nominal |
| C1_06 | 수금방법 | Nominal | C1_24 | 계약자출생년도 | Interval |
| C1_07 | 보험료 | Interval | C1_25 | 계약자성별 | Binary |
| C1_08 | 특약유무 | Binary | C1_26 | 계약자직업군 | Nominal |
| C1_09 | 약관대출유무 | Nominal | C1_27 | 모집설계사번호 | ID |
| C1_10 | 약관대출일자 | Interval | C1_28 | 수금설계사번호 | ID |
| C1_11 | 약관대출잔고 | interval | C1_29 | 상품중분류 | Nominal |
| C1_12 | 부활유무 | Binary | C1_30 | 상품소분류 | Nominal |
| C1_13 | 부활일자 | Interval | C1_31 | 피보험자가입연령 | Interval |
| C1_14 | 계약일자 | Interval | C1_32 | 피보험자출생년도 | Interval |
| C1_15 | 보험금지급만기일자 | Interval | C1_33 | 피보험자성별 | Nominal |
| C1_16 | 최종납입횟수 | Interval | C1_34 | 피보험자직업군 | Nominal |
| C1_17 | 계약자학력 | Nominal | C1_35 | 가입경로 | Binary |
| C1_18 | 계약자결혼유무 | Binary | | | |

**Table 3.** Modification of Some Variables

| 변수명 | 기존 | 변경 | 내용 | 단위 |
|---|---|---|---|---|
| C1_04_N | 납입방법 | 납입주기 | 기존의 Ordinal로 되어있는 변수를 Interval로 변경 (1→1, 2→3, 3→6, 4→12) | 월 |
| C1_10_N | 약관 대출일자 | 약관 대출기간 | (기준일-대출일) 약관대출이 없는 경우 '0' | 년 |
| C1_13_N | 부활일자 | 부활후 경과기간 | (기준일-계약일) 부활이 없는 경우 계약일로부터의 경과기간 | 년 |
| C1_14_N | 계약일자 | 계약후 경과기간 | (기준일-계약일) 계약일로부터의 경과기간 | 년 |
| C1_15_N | 보험금지급 만기일자 | 만기 잔여기간 | (만기일-기준일) 만기가 없는 상품의 경우 최대 만기년도+1 | 년 |
| C1_16_N | 최종납입횟수 | 총 납입액 | (1회 보험료×최종 납입횟수) | 원 |
| C1_24_N | 계약자출생년도 | 계약자 현재나이 | (기준일-계약자 출생일) | 나이 |
| C1_32_N | 피보험자 출생년도 | 피보험자 현재나이 | (기준일-피보험자 출생일) | 나이 |

에 관계 없고 정량적인 데이터로 재정의하였다. 변수 재정의 내용을 <표 3>에 나타내었다.

### 3.3 오버 샘플링

제 2장에서 언급한 바와 같이 본 사례에서는 목표 클래스의 비율을 10%씩 늘려가며 중복에 의한 오버 샘플링하여 분류분석을 실행하였고 목표정확도가 높아졌을 때 중복을 멈추었다. 본 데이터의 경우 목표 클래스의 비율이 30% 이상이 되었을 때에 결과가 이전에 비해 좋아졌으며 최종적으로 34.5%의 이탈고객을 생성하였다. 원래 데이터와 오버샘플링을 거친 후의 데이터에서 각 클래스의 수와 비율을 비교한 결과를 <표 4>에 나타내었다.

**Table 4.** Sample Sizes before and after Oversampling

| | Before Oversampling | After Oversampling |
|---|---|---|
| Cancelled customers | 1557(3.9%) | 20241(34.5%) |
| Continuing customers | 38443(96.1%) | 38443(65.5%) |
| Total | 40000(100%) | 58684(100%) |

### 3.4 군집분석

고객군을 K-means 방법으로 군집화한 결과 8개 군집이 적절함을 알았다. 이탈고객은 8개의 군집에 고루 분포되었으며 각 군집별 이탈고객의 비율은 <표 5>와 같다. 이탈고객의 비율이 가장 낮은 군집은 21.3%, 가장 높은 군집은 41.9%로 나타났으며 이탈고객이 특정군집에 치우치는 결과가 도출되지 않았다. 이것으로 고객특징을 고려해 비슷한 특징을 보이는 세분화된 군집내에서 분류분석을 하였을 때에 이탈 여부를 좀 더 자세히 예측할 수 있는 모델이 만들어질 수 있음을 알 수 있다.

**Table 5.** Numbers of Cancelled Customers by Clusters

| Cluster no. | # Customers | # Cancelled customers(%) |
|---|---|---|
| Cluster1 | 15275 | 6396(41.9%) |
| Cluster2 | 8544 | 2145(25.1%) |
| Cluster3 | 6296 | 2496(39.6%) |
| Cluster4 | 3918 | 1495(38.2%) |
| Cluster5 | 4955 | 1053(21.3%) |
| Cluster6 | 4031 | 1092(27.1%) |
| Cluster7 | 7337 | 2275(31.0%) |
| Cluster8 | 8328 | 3289(39.5%) |
| Total | 58684 | 20241(34.5%) |

### 3.5 분류분석 결과

본 연구에서 제안한 절차의 검증을 위하여 절차 중 일부 과정을 제외한 방법의 결과를 비교 분석하였다. 이를 위해 아래와 같이 4 단계에 따라 분류결과를 알아보고자 한다. 마지막 단계인 단계 4가 본 연구의 결과이다.

단계 1 : 로지스틱 회귀분석을 수행
단계 2 : 오버샘플링한 후 로지스틱 회귀분석을 수행
단계 3 : 오버샘플링을 거친 데이터를 군집분석으로 고객군을 분류하고 각 고객군 별로 로지스틱 회귀분석을 수행
단계 4: 오버샘플링, 군집분석을 모두 적용한 데이터를 고객군 별로 분류분석 할 때에 부스팅을 적용하여 반복적으로 로지스틱 회귀분석을 수행

**Table 6.** Classification Result at Step 1

| | | Classified | | |
|---|---|---|---|---|
| | | cancelled | continuing | total |
| True: | cancelled | **2** | 1555 | 1557 |
| | continuing | 9 | 38434 | 38443 |
| | total | 11 | 39989 | 40000 |

분석 결과는 정오분류표로 나타내었다. 단계 1의 분류분석을 적용한 결과를 <표 6>에 나타내었는데 실제로 이탈한 고객인 1557명 중 단 2명만을 이탈고객으로 구분해 내어 0.13%(=2/1557)의 목표 정확도를 나타낸다. 여기서 전체 정확도는 96.09%에 달하는데 이것은 모든 데이터를 같은 클래스에 분류한 결과이기 때문에 큰 의미가 없다. 따라서 불균형 데이터의 특성상 전체 정확도보다는 목표 정확도에 초점을 맞춰야 한다는 것을 알 수 있다.

다음으로 불균형 데이터 처리를 위하여 오버샘플링으로 이탈고객의 비율을 34.5%로 증대시킨 후 분류분석 하였을 때의 결과를 <표 7>에 나타내었다. 오버샘플링 후의 이탈고객 수가 1557에서 20241로 증가하였으며 이는 정오분류표를 통해 원래 데이터의 결과와 직접적인 비교를 힘들게 한다. 이에 따라 이탈고객 수를 원래 관측수인 1557개로 보정하였다. 오버샘플링의 방법이 같은 데이터를 중복하는 방식으로 일어났기 때문에 중복된 데이터가 모두 한 클래스에 속하므로 비율을 이용한 숫자환산이 가능하다. 샘플링 후 목표정확도는 19.27%(=300/1557)로 이전에 비해 150배의 목표정확도 증가를 보였다.

**Table 7.** Classification Result at Step 2

| | | Classified | | |
|---|---|---|---|---|
| | | cancelled | continuing | total |
| True: | cancelled | **300** | 1256 | 1557 |
| | continuing | 3259 | 35184 | 38443 |
| | total | 3559 | 36440 | 40000 |

오버샘플링을 거친 데이터를 K-means 방법으로 군집화한 후 각 군집별로 다른 로지스틱 회귀분석을 실행하여 정확도를 구한 결과는 <표 8>과 같다. 군집분석으로 나뉘어진 데이터를 각각 분류모델을 적용하였을 때의 목표정확도는 32.7%로 증가하였다.

**Table 8.** Classification Result at Step 3

| | | Classified | | |
|---|---|---|---|---|
| | | cancelled | continuing | total |
| True: | cancelled | **509** | 1048 | 1557 |
| | continuing | 5044 | 33399 | 38443 |
| | total | 5553 | 34447 | 40000 |

로지스틱 회귀분석을 사용하여 이탈고객 예측시 오버샘플링과 군집분석을 사용하여 목표정확도의 유의할만한 성과를 발견하였다. 여기에 목표정확도의 보다 높은 향상을 위해 분류분석시 부스팅을 적용하여 반복적으로 에러를 최소화한 결과는 <표 9>와 같으며 목표정확도는 39.4%로 이전단계에 비해 증가함을 볼 수 있다.

상기 결과를 종합하여 그래프로 나타낸 결과는 <그림 2>와 같으며 단계를 거침에 따라 목표정확도가 꾸준히 상승함을 볼 수 있다. 결국 제안된 절차에 따르면 목표정확도를 39.4%까지

달성할 수 있음을 알 수 있으며 이는 본 데이터가 실제의 고객 정보 데이터임을 고려할 때에 괄목할만한 결과임을 알 수 있다.

**Table 9.** Classification Result at Step 4(Proposed method)

| | | Classified | | |
|---|---|---|---|---|
| | | cancelled | continuing | total |
| True: | cancelled | **613** | 944 | 1557 |
| | continuing | 4524 | 33919 | 38443 |
| | total | 5137 | 34863 | 40000 |

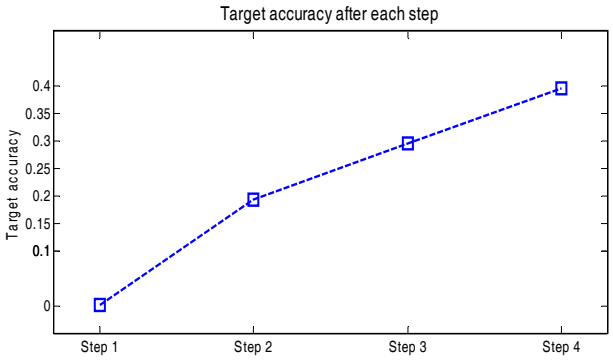


**Figure 2.** Target Accuracies by Steps

본 연구에서 데이터 탐색과 전처리는 SAS를 사용하였으며 분류분석은 Matlab과 C++를 연동하여 수행되었다. 조정 로지스틱 회귀분석 과정에서 TRON을 사용하였기 때문에 처리속도에 있어 상당한 향상을 보인다. 대용량 분석에 효과적인 L-BFGS와 각 단계를 비교한 결과는 <표 10>과 같다. 대부분의 경우 TRON이 L-BFGS에 비해 10배 이상 빠른 결과를 보였으며 이는 Lin *et al.*(2008)의 결과와 일치한다. 정확도는 모든 경우에 TRON과 동일하였기 때문에 L-BFGS의 정확도는 생략하였다.

**Table 10.** Computation Times of Two Optimization Methods

| Step No. | TRON(sec) | L-BFGS(sec) |
|---|---|---|
| Step 1 | 5.0 | 42.6 |
| Step 2 | 5.6 | 63.0 |
| Step 3 | 26.0 | 358.8 |
| Step 4 | 159.7 | 1793.8 |

### 3.6 다른 분류분석과의 비교

본 절에서는 제안된 절차를 동일하게 적용하되 조정로지스틱 회귀분석 대신 의사결정나무 (Decision Tree) 또는 선형판별분석 (Linear Discriminant Analysis; LDA)을 이용한 경우의 각 군집별 정확도 및 모델 정확도 (Total에 해당)를 <표 11>에 비교하였다.

**Table 11.** Accuracies by Clusters of Other Classification Methods

| Cluster No. | Proposed Method | Decision Tree | LDA |
|---|---|---|---|
| Cluster1 | 0.3463 | 0.3600 | 0.4046 |
| Cluster2 | 0.3394 | 0.4770 | 0.3366 |
| Cluster3 | 0.4104 | 0.3589 | 0.3869 |
| Cluster4 | 0.4609 | 0.4713 | 0.3466 |
| Cluster5 | 0.5494 | 0.4790 | 0.3589 |
| Cluster6 | 0.4219 | 0.4077 | 0.3509 |
| Cluster7 | 0.3829 | 0.4786 | 0.3514 |
| Cluster8 | 0.4257 | 0.4697 | 0.3352 |
| Total | 0.3936 | 0.4204 | 0.3684 |

<표 11>에서 보듯이 각 군집의 정확도를 TRON을 사용한 조정 로지스틱 회귀분석과 의사결정나무, 선형판별분석을 비교한 결과 의사결정나무의 경우 대체로 가장 높았다. 그러나 의사결정나무의 경우 고객 속성이 많으면 가지가 많아지고 모델이 복잡해 지며(현 데이터의 경우 총 131개의 노드 생성) 계산시간이 2배 이상 소요된다. 즉, 의사결정나무의 경우 정확도 측면에서 좋은 결과를 보이지만 속도가 많이 걸린다는 단점이 있으며 선형판별분석은 속도는 의사결정나무보다 빠르지만 로지스틱 회귀모형보다는 조금 느리며 정확도에 있어서도 로지스틱 회귀모형보다 조금 모자라는 것으로 나타났다. <표 10>에는 제안절차의 단계별 계산시간을 나타내고 있으나 선형판별분석 및 의사결정나무의 경우 이 보다 많은 계산시간을 요한다고 하겠다.

### 3.7 테스트 데이터의 정확도

보다 객관적인 성능분석을 위하여 총 4만개 중 20%인 8000개 데이터를 랜덤으로 추출하여 테스트 데이터로 활용하여 다시 분석하였다. 테스트 데이터 중 이탈고객은 311명으로 약 3.9%를 차지하고 있다. 이를 제외한 데이터에 모델을 학습시킨 후 테스트 데이터에 적용하여 클래스를 예측하고 분류 정확도를 산출하였다.

**Table 12.** Target Accuracies by Steps for Test Dataset

| Step No. | Proposed Method(%) | Decision Tree |
|---|---|---|
| Step 1 | 0.64 | - |
| Step 2 | 14.47 | - |
| Step 3 | 35.05 | 31.51 |

<표 12>에는 테스트 데이터에 대한 목표 정확도를 단계별로 나타내고 있다. 본 테스트 데이터의 경우 클래스 정보가 없는 것으로 간주하기 때문에 오분류된 데이터를 알아야 하는 단계 4의 부스팅은 실시하지 않았다. 참고로 이 표에 로지스틱 회귀

분석 대신 의사결정나무에 의한 단계 3의 결과를 표시하였는데 로지스틱 회귀분석 보다 다소 낮은 정확도를 보이고 있다.

## 4. 결 론

본 연구의 특징은 첫째로 오버샘플링, 군집분석, 부스팅을 거치며 불균형데이터의 정확도를 높이는 데이터마이닝 방법을 제안하였다는 데에 있다. 사례연구를 통하여 오버샘플링에 의해 목표고객의 비율이 늘어남에 따라 정확도가 상승하는 것을 관찰하였으며 군집분석이나 부스팅을 통하여 정확도가 더욱 증가하는 것을 보였다. 특히 군집분석으로 데이터를 그룹화한 후 조정 로지스틱 회귀모형을 적용하여 이탈고객의 정확도 향상에 괄목할만한 성과를 보였다. 둘째로 여러 종류의 분류분석방법을 데이터에 적용하여 정확도와 계산속도를 비교하였다. Trust region Newton method를 사용한 Regularized 로지스틱 회귀모형을 제안하고 있으며 이의 결과를 의사결정나무 및 선형판별분석과 비교하였는데, 의사결정나무의 경우 불균형 데이터에서 Training 데이터의 정확도가 높았으나 속도가 오래 걸렸으며 선형판별분석은 로지스틱 회귀분석에 비해 속도나 정확도 면에서 부족한 것을 확인하였다. 테스트 데이터에 대한 예측정확도의 경우는 조정항을 통해 과적합을 방지하기 때문에 다른 기법보다 나은 결과를 보였다. 결론적으로 정확도와 속도를 모두 고려할 때에 조정 로지스틱 회귀분석이 다른 기법에 비해 우수하다고 판단된다.

본 연구 결과로 불균형 이분 데이터를 분류하기 위한 빠르고 정확도를 향상시킨 방법을 제안하였으나 몇 가지 논의될만한 사항이 있다. 첫째로 샘플링을 선택하는 대신 목표 데이터에 가중치를 부여하여 총 데이터 수의 변화없이 처리하는 방안의 개발이 요구된다. 두 번째로 목표정확도 이외에 ROC 곡선이나 Lift의 개념을 도입하여 성능을 평가하는 것이다. 이 때에 FP, FN, TP 등의 가중치를 결정해야 하는 문제가 있다. 마지막으로는 새로운 분류분석방법의 개발이다. 로지스틱 회귀분석은 이탈고객의 정확도와 더불어 전체 정확도를 향상시키는 데는 한계가 있을 수 있는데 이것은 선형이라는 특징 때문이다. 처리속도와 정확도의 향상이라는 두 가지를 동시에 달성할 수 있는 방안은 계속적인 연구가 필요하겠다.

## 참고문헌


Au, W-H., Chan, K. C. C. and Yao, X. (2003), A novel evolutionary data mining algorithm with applications to churn prediction, *IEEE Transactions on Evolutionary Computation*, **7**(6), 532-545.

Benson, S. and Moré, J. J. (2002), A limited memory variable metric method for bound constrained minimization, Tech. Rep. ANL-95/11-Revision 2.1.3, Argonne National Laboratory.

Catlett, J. (1991), Megainduction: a test flight, *Proceedings of the Eighth International Workshop on Machine Learning*, Morgan Kaufmann, 596-599.

Coussement, K. and Van den Poel, D. (2008), Churn prediction in subscription services : an application of support vector machines while comparing two parameter-selection techniques, *Expert Systems with Applications*, **34**(1), 313-327.

Darroch, J. N. and Ratcliff, D. (1972), Generalized iterative scaling for log-linear models, *The Annals of Mathematical Statistics*, **43**(5), 1470-1480.

Datta, P., Masand, B., Mani, D. R., and Li, B. (2000), Automated cellular modeling and prediction on a large scale, *Artificial Intelligence Review*, **14**(6), 485-502.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997), Inducing features of random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(4), 380-393.

Freund, Y. and Schapire, R. (1996), Experiments with a new boosting algorithm, *Machine Learning : Proceedings of the Thirteenth International Conference*, San Francisco, USA, 148-156.

Goodman, J. (2002), Sequential conditional generalized iterative scaling, *Proceedings of the 40th Meeting of the ACL*, Philadelphia, PA, 9-16.

Gordon, D. and Perlis, D. (1989), Explicitly biased generalization computational intelligence, *Computational Intelligence*, **5**(2), 67-81.

Heo, H., Park, H., Kim, N., and Lee J. (2008), Prediction of credit delinquents using locally transductive multi-layer perceptron, *Fifth International Symposium on Neural Networks*, Beijing, China, paper, 136.

Hung, S-Y., Yen, D. C. and Wang, H-Y. (2006), Applying data mining to telecom churn management, *Expert Systems with Applications* **31**(3), 515-524.

Jin, R., Yan, R., and Zhang, J. (2003), A faster iterative scaling algorithm for conditional exponential model, *Proceedings of the 20th International Conference on Machine Learning*, Washington DC.

Koh, K., Kim, S-J. and Boyd, S. (2007), An interior-point method for large-scale i1-regularized logistic regression, *Journal of Machine Learning Research*, **8**, 1519-1555.

Komarek, P., Moore, A. W. (2005), Making logistic regression a core data mining tool : A practical investigation of accuracy, speed, and simplicity. Technical Report. CMU-RI-TR-05-27, Carnegie Mellon University, USA.

Kubat, M. and Matwin S. (1997), Addressing the curse of imbalanced training sets : one-sided selection, *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, 179-186.

Lin, C-J., Weng, R. C., and Keerthi, S. S. (2008), Trust region Newton method for logistic regression, *The Journal of Machine Learning Research*, **9**, 627-650.

Liu, D. and Nocedal, J. (1989), On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, **45**(1), 503-528.

Malouf, R. (2002), A comparison of algorithms for maximum entropy parameter estimation, *Proceedings of the 6th Conference on Natural Language Learning*, **20**, 1-7.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., and Kaushansky, H. (2000), Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry, *IEEE Transactions on Neural Networks*, **11**(3), 690-696.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., and Brunk, C. (1994), Reducing misclassification costs, *Machine Learning : Proceedings of the Eleventh International Conference*, Morgan Kaufmann.

Phua, C., Alahakoon, D., and Lee, V. (2004), Minority report in fraud

detection: classification of skewed data, *ACM SIGKDD Explorations Newsletter*, **6**(1), 50-59.

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

Stolfo, S., Fan, D., Lee, W., Prodromidis, A., and Chan, P. (1997), Credit card fraud detection using meta-learning: issues and initial results, *Proceedings of the AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management* (AAAI Technical Report WS-97-07), Menlo Park : CA : AAAI Press, 83-90.

Sutton, C. and McCallum, A. (2006), An introduction to conditional random fields for relational learning, In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*, MIT Press.

Sung, K-K. and Poggio, T. (1995), Learning human face detection in cluttered scenes, *Lecture Notes in Computer Science*, **970**, 432-439.

Tan, P. N., Steinbach, M., and Kumar, V. (2006), *Introduction to Data Mining*. Addison-Wesley, Reading.

Vapnik, V. N. (1998), *Statistical Learning Theory*. Wiley, New York.

Viaene, S., Derrig, R. A., Baesens, B., and Dedene, G. (2002), A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, *The Journal of Risk and Insurance*, 69(3), 373-421.

Wei, C-P. and Chiu, I-T. (2002), Turning telecommunications call details to churn prediction : a data mining approach, *Expert Systems with Applications*, **23**(2), 103-112.