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.

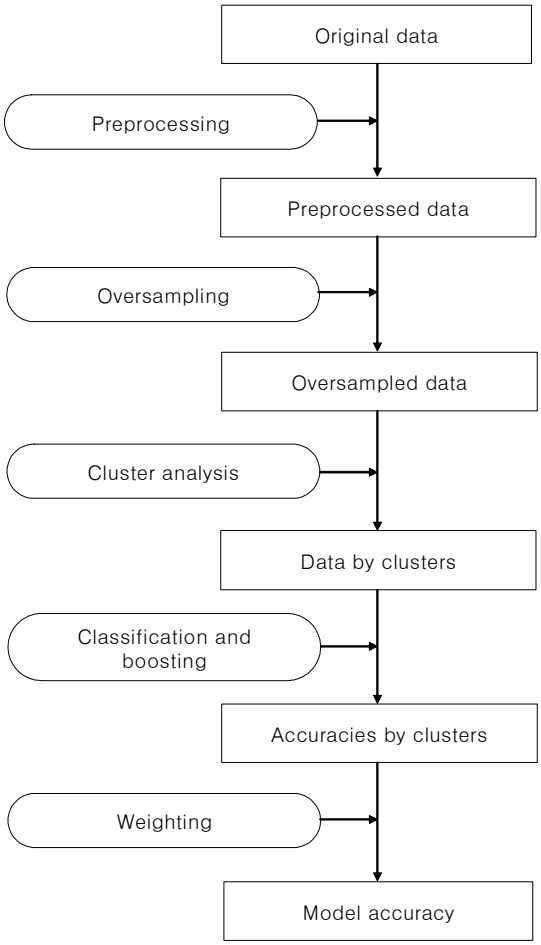

Figure 1. Proposed Classification Procedure

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.