

# 신용카드 연체자 분류모형의 성능평가 척도 비교 : 예측률과 유틸리티 중심으로

정 석 훈\* · 서 용 무\*\*

## Comparison of Performance Measures for Credit-Card Delinquents Classification Models : Measured by Hit Ratio vs. by Utility

Sukhoon Chung\* · Yongmoo Suh\*\*

### Abstract

As the great disturbance from abusing credit cards in Korea becomes stabilized, credit card companies need to interpret credit-card delinquents classification models from the viewpoint of profit. However, hit ratio which has been used as a measure of goodness of classification models just tells us how much correctly they classified rather than how much profits can be obtained as a result of using classification models. In this research, we tried to develop a new utility-based measure from the viewpoint of profit and then used this new measure to analyze two classification models(Neural Networks and Decision Tree models). We found that the hit ratio of neural model is higher than that of decision tree model, but the utility value of decision tree model is higher than that of neural model. This experiment shows the importance of utility based measure for credit-card delinquents classification models. We expect this new measure will contribute to increasing profits of credit card companies.

Keywords : Utility-Based Data Mining, Credit Classification Model, Neural Networks, Decision Tree

논문접수일 : 2008년 04월 23일      논문게재확정일 : 2008년 12월 19일

※ 고려대학교 특별연구비에 의하여 수행되었음.

\* 고려대학교 경영대학 박사과정, e-mail : sukhoon@korea.ac.kr

\*\* 고려대학교 경영대학 교수, e-mail : ymsuh@korea.ac.kr

## 1. 서 론

과거 1990년대 말에서 2000년대 초반에 이르는 한국의 신용대란은 신용카드사들의 무분별한 카드 발급과, 이용 고객들의 무절제한 카드 사용으로 인해 수많은 신용불량자들이 양산되는 결과를 가져왔다. 이들 신용 불량자들의 결제 부도는 개인의 신용불량은 물론 신용카드사들의 재무 상태에도 많은 부담을 주어 결국 기업의 존립에까지 위협을 가하는 상황에 이르게 하였다. 이러한 신용대란은 다른 형태의 금융기관들에게까지 영향을 미치게 되었는데 이들은 신용불량자들에 대한 엄격한 금융 제재를 가하기 시작하였고, 그 결과 신용불량자들은 스스로의 신용상태를 회복하기가 더욱 어려워지는 악순환이 발생하였다. 이러한 악순환은 국가 차원에서라도 건전하지 못한 개인 및 신용카드사들의 재무상태를 감내하지 않으면 안 되는 형국으로 발전되어 갔다.

이와 같은 혼란 속에서 정부는 2004년 말 '신용정보의 이용 및 보호에 관한 법률'을 개정함으로써 이러한 악순환의 고리를 끊고자 하는 의지를 보였다. 개정된 법률안의 요지는 신용불량자 스스로 신용을 회복할 수 있는 기회를 마련해 주면서 도덕적 해이는 줄이고자 한다는 것이다[전국 은행 연합회, 2005]. 신용카드사들도 과거의 무분별한 카드발급을 줄이고 좀 더 세분화된 고객관리를 수행하는 등 여러 가지 노력들을 통해 점차 과거의 위험에서부터 벗어나기 시작하였다. 이러한 경제 각 주체들의 적극적인 노력으로 2000년대 초반 발생했던 신용대란은 차츰 안정을 찾기 시작하였다.

이러한 신용대란은 신용카드사들의 입장에서 보았을 때, '수익'의 측면 보다는 엄청난 액수로 발생하고 있었던 '손실' 또는 '비용'을 줄이는 측면으로 기업활동에 반영되었었다. 이러한 '손실'

및 '비용'의 감소 측면에서는 분류 모형 성능의 평가에 있어서도 기존에 사용해 왔던 평가 척도인 예측률(hit ratio)과 ROC 곡선의 아래 면적값(AUC) 정도만을 이용해도 큰 불편함이 없었다. 그러나 점차 혼란이 안정화 되어감에 따라 이제는 발생하는 '손실'이나 '비용'을 무조건 감소시키려 하기 보다는 '수익'의 증대 관점에서 신용 분류 모형의 예측력을 평가해야 할 필요성이 증대되었으며, 이에 따라 좀 더 세밀한 분석을 통하여 신용카드사들의 수익 증대 전략 수립에 도움을 줄 수 있는 분류 모형이 필요하게 되었다.

본 연구의 목적은 신용카드 고객의 신용도 분류모형의 성능을 평가함에 있어, 유틸리티 기반의 함수를 개발하고 그 평가결과를 예측률로 평가하는 경우의 결과와 비교해 보는 것이다. 이를 위하여, 분류모형이 불량자를 불량자로 잘 예측한 경우, 불량자를 우량자로 잘못 예측한 경우, 우량자를 불량자로 잘못 예측한 경우, 그리고 우량자를 우량자로 잘 예측한 경우 등 각 상황에서 신용카드사의 수익에 미치는 영향을 고려한 유틸리티 함수를 정의하여 분류모형의 성능 평가에 사용하였다. 분류 모형은 가장 많이 쓰이고 있는 신경망 기법과 의사결정나무 기법(C5.0)을 이용하였고, 실제 신용카드 거래 요약 데이터를 사용하였다. 실험 결과, 예측률을 기반으로 작성된 분류모형과 수익을 기반으로 작성된 분류모형의 평가 결과가 상이할 수 있음을 관찰 하였으며 평가 결과에 대한 해석도 좀 더 체계적으로 이해할 수 있었다. 이와 같은 수익 기반 분류모형은 신용카드사들의 수익 개선에 기여할 것으로 예상된다.

본 논문의 구성은 다음과 같다. 이어지는 제 2장에서는 신용 분류 모형과 유틸리티 기반 데이터마이닝에 대한 기존의 연구들을 소개하였고, 제 3장에서는 실험에 사용된 데이터와 실험

모형에 대하여 자세하게 설명하였다. 제 4장에서 는 실험 결과에 대하여 그 의미를 설명하였으며 마지막으로 제 5장에서는 결론으로 본 연구 결과가 시사하는 바와 향후 연구할 바를 기술하였다.

## 2. 문헌연구

### 2.1 신용 분류 모형

분류란 여러 가지 속성들의 값과 클래스가 정해진 많은 수의 개체 데이터로부터 일반화하여 분류 모형을 만들고, 이를 기반으로 새로운 개체가 어떤 클래스에 속할지를 예측하는 것을 말한다[Han, Kamber, 2001]. 다양한 기법들이 이러한 분류 모형을 만들기 위해서 사용되었는데, 대표적인 기법으로는 신경망 모형, 의사결정나무 등의 인공지능 기법들과 판별분석, 로짓분석 등의 통계적 기법들이 있으며, 근래에는 SVM (Support Vector Machine), Rough Sets 기법, 그리고 여러 가지 기법들이 함께 사용된 하이브리드 기법 등이 자주 사용되고 있다.

이 중에서 신경망 모형은 가장 빈번하게 사용된 기법으로 비선형 문제에 특히 좋은 성능을 발휘하는 것으로 알려져 있다[Desai 등, 1997; Jesen, 1992; Piramuthu, 1999]. 신용도 평가에 있어서 신경망 모형은 다른 모형들과의 성능 비교 연구가 많이 이루어졌는데, West[2000]의 연구에서는 5개의 신경망 모형<sup>1)</sup> 사이의 성능을 비교하여 그 중에서 mixture-of-experts 모형과 Radial Bases Function 모형이 가장 좋은 성능을 가졌음을 보였다.

의사결정나무 기법도 분류 작업에 자주 쓰이는 기법 중 하나로, Stark와 Pfeiffer[2000]는 여

러 가지 의사결정나무 기법들<sup>2)</sup>을 이용한 모형과 선형 회귀분석 모형의 성능을 비교하는 연구를 수행하였다. Colombet 등[2000]도 의료분야 분류 문제에 있어서 심혈관의 위험 정도를 분류하는데 CART 모형, 신경망 모형 그리고 로지스틱 모형의 성능을 비교하는 연구를 수행하였다. 이 연구에서는 신경망 모형과 로지스틱 모형의 경우 AUC 값이 모두 0.78로 비슷한 성능을 보였고 CART 모형의 경우 0.76으로 다소 예측력이 떨어지는 것으로 평가되었다.

이러한 분류모형들 중에서 특히 조직 또는 개인 고객의 신용을 분류 평가하는 모형을 신용분류모형이라고 하는데 신용 산업의 규모가 커지면서 그 사용 빈도 및 연구의 관심이 점점 증가하여 왔다[Thomas, 2000]. 과거의 데이터를 이용하여 새로운 조직 또는 고객의 신용도를 '우수' 또는 '불량' 등으로 분류하는 신용분류 모형을 사용하게 되면 신용분석에 대한 비용을 줄일 수 있을 뿐만 아니라 신용에 대한 의사결정을 신속하게 수행할 수 있는 등의 장점이 있다 [Brill, 1998]. 신용 분류모형들 역시 다양한 기법들을 이용하여 개발 되어 왔는데 초기에는 주로 통계적 기법들이 사용되었으나[Hand, Henley, 1997], 점차 기계학습, 인공지능 등 여러 분야의 기법들이 이용되기 시작하였다.

Desai[1996]는 통계기법을 이용한 모형들과 신경망 모형의 성능 비교를 위하여 전통적인 통계적 기법들인 판별분석, 로짓회귀분석의 성능과 신경망 모형의 성능을 비교하는 연구를 수행하였다. 이 실험에서는 전통적으로 쓰였던 통계적 기법보다 신경망 모형의 성능이 비교적 성능이 좋은 것으로 평가되었다. Huang 등[2007]은 하이브리드 형태의 SVM 기법을 이용하여 신용상품 구매 의뢰자의 신용을 평가하는 분류모형을

1) Multilayer Perceptron, Mixture-of-Experts, Radial Basis Function, Learning Vector Quantization 그리고 Fuzzy Adaptive Resonance

2) ID3, C4.5, CHAID 그리고 CART.

만들고 이를 신경망 모형, 유전자 모형 그리고 의사결정나무 모형과 비교하였다. 그 결과 하이브리드 형태의 SVM 모형이 비교적 적은 수의 속성으로도 만족할 만한 결과를 가져올 수 있음을 확인했다. 이 이외에도 Lee 등[2002]의 신용 분류 모형, Lee와 Chen[2005]의 신용 분류모형 그리고 Hsieh[2005]의 신용 분류모형에서도 하이브리드 형태의 모형을 발견할 수 있었다.

신용평가의 대상이 되는 조직의 규모에 의해서도 연구를 나누어 볼 수 있다. 대부분의 연구는 평가의 대상이 되는 조직의 규모를 밝히지 않았거나 또는 대기업에서 나온 자료를 이용하여 모형을 만드는 경우가 많은데, Cavalluzzo 등[1998, 1999]의 연구와 Tsaih 등[2004]의 연구에서는 주로 규모가 작은 기업들을 대상으로 신용평가 모형을 개발하였다. 그 이유는, 대기업에서 발생하는 데이터와 작은 기업에서 발생하는 데이터는 서로 그 성질이 다르기 때문에, 분류모형의 대상이 되는 조직의 특성을 무시하고는 정확한 분류모형을 만들 수 없었기 때문이었다. 이것은 모든 신용평가 분야에 적용할 수 있는 일반적인 모형의 개발 보다는 해당 영역에 맞는 특화된 모형의 개발이 더욱 의미 있을 수 있음을 말해준다.

Hoffmann 등[2007]은 모형의 예측력 및 설명력을 강화하기 위하여 퍼지 분류 규칙(fuzzy classification rules)을 이용하여 신용 분류모형을 만들었다. 특히 이들은 진화적 전략을 적용한 퍼지 알고리즘과 유전자 알고리즘을 이용하여 유연하고 설명력 높은 모형을 만들 수 있었다. 이 모형은 예측력에 있어서도 기타 Nef-class 모형, Neurofuzzy 모형과 비교했을 때 더욱 우수했다. Huang 등[2006]도 2SGP(Two-Stage Genetic Programming) 기법을 이용하여 If-Then 규칙과 판별 함수의 장점을 살리면서 예측력을 높인 모형을 만들었다. 이처럼 규칙(rules) 등을

이용하여 사용자가 그 결과에 대해 이해하기 쉽고 설명력이 높은 신용 분류모형을 만들려는 노력들은 개발된 분류 모형의 사용성을 높이기 위함이라고 볼 수 있다. Blöchlinger와 Leippold [2006]는 ROC 곡선과 가격 곡선의 연결을 이용하여 대출 상품의 최적 가격 산정에 이용할 수 있는 신용 분류 모형의 규칙들을 만들었다. 이 연구는 신용 분류모형에 경제적 수익 개념을 접목시킨 연구라 볼 수 있는데 이처럼 신용 분류 모형에 관련된 최근의 연구들에서는 단순한 분류모형의 예측력 향상 보다는 설명력이나 수익적 측면을 함께 고려하여 모형의 사용성을 높여려는 노력들이 조금씩 보이고 있다.

그러나 아직도 대부분의 신용분류 모형에서는 예측률을 이용한 모형 성능평가 방법을 사용하고 있는데 예측률을 이용한 평가 방법은 개발된 분류모형을 실제 문제에 적용하는데 많은 해석적 어려움을 갖고 있다. 따라서 신용분류 모형의 출력결과에 대한 이해도와 사용성을 높일 수 있는 방법의 연구가 매우 절실하며, 또 이를 위해서 범용적인 모형의 개발에 앞서 특정 분야 혹은 조직에 적용될 수 있는 특화된 모형의 개발이 우선적으로 이루어져야 할 것이다.

## 2.2 유틸리티 기반 데이터마이닝 (Utility-Based Data Mining)

앞 절에서 살펴본 바와 같이 지금까지의 분류 모형들은 대부분 예측률을 이용하여 모형의 성능을 평가하였다. <그림 1>은 목적변수가 이진형인 경우의 전형적인 혼동행렬(confusion matrix)을 보여주고 있는데, 이러한 상황에서 우리는 예측률(HR)을 다음 식 (1)과 같이 정의하여 사용해왔다. 즉 예측률(HR)은 전체 예측한 개체 수 중에서 분류모형이 정확하게 예측한 개체의 수를 의미한다.

$$HR = (TP + TN) / (TP + FP + FN + TN) \quad (1)$$

실제치 \ 예측치	긍정 클래스	부정 클래스
긍정 클래스	맞힌 긍정(TP)	틀린 긍정(FP)
부정 클래스	틀린 부정(FN)	맞힌 부정(TN)

(그림 1) 혼동행렬(confusion matrix)

그러나 이러한 예측물은 각각의 경우( $TP$ ,  $FP$ ,  $FN$  그리고  $TN$ )에 예측결과의 비용이 모두 동일하다는 가정하에서 사용할 수 있다. 예를 들어 실제치 ‘긍정 클래스’를 ‘긍정 클래스’로 잘 예측한 경우의 수익과 실제치 ‘부정 클래스’를 ‘부정 클래스’로 잘 예측한 경우의 수익이 동일하다라는 가정을 하고 있는 것이다. 그러나 실제 존재하고 있는 많은 실생활의 사례는 이러한 가정을 만족 시키지 못하는 경우가 대부분이다. 따라서 예측물을 모형의 성능 평가척도로 사용한 경우에는 그 결과의 해석이나 사용에 문제가 발생할 수 있다[Holte, Drummond, 2005].

분류모형의 성능 평가나 데이터마이닝 각 단계에서의 의사결정시, 평가의 척도로 수익이나 비용 개념을 이용한 연구들은 2005년과 2006년에 열린 두 번의 국제 학술대회에서 찾아볼 수 있다. Arnt와 Zilberstein[2005]의 연구, Ciraco 등[2005]의 연구 그리고 Zadrozny[2005]의 연구는 ‘비용’을 중심으로 한 연구이다. Arnt와 Zilberstein[2005]은 모형의 비용을 오분류 비용, 속성 측정 비용 그리고 응답 시간 비용 등 세 가지로 나누고 이들을 묶어서 시간과 비용에 반응하는 모형을 개발하였다. Ciraco 등[2005]은 잘못된 예측의 경우에만 비용이 발생한다는 가정하에서 ‘틀린 긍정’과 ‘틀린 부정’의 비용 비율을 1 : 10, 2 : 10, 3 : 10, ~, 10 : 10, 10 : 9, 10 : 8 ~, 10 : 1로 점차적으로 변화시켜 주면서 총 비용의 변화 양상을 분석하였다. Zadrozny[2005]

는 비용에 대한 정보에 제한이 있다라는 가정하에서 비용-민감도 학습(Cost-Sensitive Learning) 방법을 제안하였다.

Chawla와 Li[2006]의 연구 및 Ling 등[2006]은 ‘수익’ 측면에서 모형을 개발하였다. Chawla와 Li[2006]는 수익 함수를 이용하여 모형의 수익 변화를 관찰하면서 가장 적절한 가격을 결정할 수 있는 프레임워크를 고안하였다. 또 Ling 등[2006]은 소프트웨어를 개발할 때 발생하는 오류(bugs)의 발생을 예측하면서 모형이 가져다 주는 수익을 고려하여 모형의 성능을 평가하였다.

Melville 등[2005a, 2005b]은 데이터 전처리 단계 중 결측값 예측에 대한 의사결정에서 유틸리티를 고려한 연구를 수행하였다. 첫 번째 연구에서는 결측값을 예측할 때 경제적으로 비용을 절약할 수 있는 방법을 제안하였고, 두 번째 연구에서는 결측값을 예측 할 때의 비용을 고려하여 비용이 낮은 속성들만을 선별하여 예측하는 방법을 제안했다. Chawla 등[2005]과 McCarthy 등[2005]은 목적변수 값의 분포가 한 쪽 클래스에 치중되어 있는 경우에 취할 수 있는 조치에 대한 연구를 수행하였다. Chawla 등[2005]은 자신들이 고안한 Wrapper 방식의 무작위 추출방법을 소개하였고, McCarthy 등[2005]은 이렇게 한 쪽 클래스에 목적 변수의 값이 치중되어 있는 경우 비용-민감도 학습 방식과 무작위 추출을 통해서 두 개의 클래스의 균형을 맞추는 방식을 비교하였는데 그 결과 비용-민감도 학습 방식이 더 우수한 결과를 보여주는 것을 확인하였다.

이 이외에도 Tseng 등[2006]은 과거에 발생 빈도가 높은 아이템 세트를 선정하던 방식에서 유틸리티가 높은 아이템 세트를 선정하는 방안을 제시했으며, Weiss와 Tian[2006]은 학습(Learning)에 들어가는 비용이 매우 높을 때, 가장 적절한 학습 데이터의 양을 결정하는 방식을

제안하면서 여러 가지 무작위 추출 방법들의 성능을 비교하였다. Yan과 Baldasare[2006]는 기존의 비용-민감도 학습 방법에서 수행하지 못했던 ROI의 최대화 방안을 제안하였으며 Yao 등[2006]은 기존의 유틸리티 기반의 측정방법들을 정리하여 통합된 프레임워크를 제시하기도 하였다.

이처럼 유틸리티 기반의 데이터마이닝에 대한 연구들이 서서히 관심을 받기 시작했으며 비용과 수익을 고려한다는 점에서 데이터마이닝 연구자들뿐만 아니라 직접 모형을 사용하는 실무자들에게도 구체적이고 금전적인 해석을 내릴 수 있는 모형의 평가들을 제시해 줄 수 있을 것으로 기대된다. <표 1>에 지금까지 소개된 연구들을 ‘연구 중점’ 별로 분류하여 정리하였다.

### 3. 데이터 및 실험 모형

#### 3.1 실험 데이터

실험에 사용된 원천 데이터는 2002년 2월부터 2003년 8월까지 총 1년 7개월 간 수집한 데이터이며, 연체가 한 번 이상 발생한 신용카드 고객들을 수집 대상으로 하였다. 5개 카드사 연체정보군, 복수카드 정보군, 그리고 신용정보군 등 총 세 종류의 정보군에서 데이터를 수집한 후 개인별로 데이터를 통합하여 사용하였다. 5

개 카드사 연체정보군은 5일 이상 10만 원 이상 연체한 고객에 대한 정보를 공유하기로 협약한 5개 카드사간에 공유되고 있는 정보를 말하며 그 내용은 연체일, 연체해제일, 그리고 연체금액 등으로 이루어져 있다. 복수카드 정보군은 신용카드업을 하고 있는 업체들이 가입한 여신전문협회에서 취합한 정보로 신용카드를 4개 이상 보유하고 있으면서 2개 이상의 신용카드를 사용한 고객들에 관련된 정보를 말한다. 이 정보군은 신용판매(일시불 및 할부 포함) 사용금액, 현금 서비스 사용금액, 그리고 연체금액 등으로 구성되어 있다. 마지막으로 신용정보군은 신용조회정보, 대출정보, 그리고 신용개설 정보 등으로 구성된 정보군이다. 이 중에서 연체에 관련된 정보는 목적변수 생성에 사용하였으며 나머지 정보를 이용하여 입력변수를 생성하였다.

전처리 과정을 수행하기 전 변수의 수는 총 74개이다. <표 2>에서 이들 중 몇 가지의 예를 볼 수 있는데, 시간적 흐름에 따라 누적된 값을 갖는 변수들을 발견할 수 있다. 예를 들어 변수 ‘조회 총 건수 90’, ‘조회 총 건수 183’, 그리고 ‘조회 총 건수 365’는 은행, 신용카드사, 개인 대출 업체 등 여러 종류의 대출기관들이 대출 심사를 위하여 각 기간별로 해당인의 신용을 조회한 횟수를 뜻한다. 즉, ‘조회 총 건수 90’은 과거

<표 1> 유틸리티 기반 데이터마이닝의 기존 연구들

연구 중점	해당 연구
비용	Arnt와 Zilberstein(2005), Ciraco 등(2005), Zadrozny(2005)
수익	Chawla와 Li(2006), Ling 등(2006)
결측값 예측 시 비용	Melville 등(2005a, 2005b)
편향된 목표값 분포 시 비용-민감도 학습측면	Chawla 등(2005), McCarthy 등(2005)
아이템 세트 선정 방식	Tseng 등(2006)
적정 학습 데이터량 선정	Weiss와 Tian(2006)
ROI 최대화 방안	Yan과 Baldasre(2006)
유틸리티 기반 측정방법 정리	Yao 등(2006)

90일 전부터 데이터 수집일(현재)까지의 조회 총 건수를 말하며 '조회 총 건수 183'은 마찬가지로 과거 183일 전부터 데이터 수집일(현재)까지의 조회 총 건수를 뜻한다. 따라서 변수 '조회 총 건수 183'에는 변수 '조회 총 건수 90'의 값이 누적되어 있다. 이렇게 누적된 값을 지니는 변수들은 서로 많은 상관성이 발생할 것으로 예상되어 원래의 변수형태를 그대로 사용하지 않고 전처리 과정에서 파생변수를 생성하여 사용하였다.

전술한 바와 같이 연구에 사용된 데이터는 최소한 한 번이라도 연체가 발생한 고객들을 대상으로 하여 수집하였다. 같은 연체자 중에서도 실제 고질적인 재정적 어려움에 처해서 장기적인 연체자가 될 고객인지 아니면 단순한 실수(망각에 의한 입금 지연, 현금 흐름 상 짧은 기간 동안 상환 기일을 못 맞춘 경우 등)로 인하여 연체가 된 고객인지를 분류하는 것은 큰 의미가 있다. 실제 신용카드 업계에서도 연체 후 10일 이내에 모든 대출금이 상환되는 경우에는

단순 실수에 의한 연체자로 분류한다. 그리고 30일 이후에도 또 다시 연체가 발생한 고객의 경우에서부터 실제 금전적 어려움에 의한 연체자로 구분하고 있다. 따라서 우리의 연구에서는 연체자 중에서 연체 후 10일 이내에 모든 대출 금액을 상환한 고객은 우량자로 분류하였으며, 2차 연체(최초 연체 발생 후 30일 이후에도 또 다시 연체가 발생한 경우)가 발생한 경우에만 불량자로 분류하였다. 따라서 본 연구의 목적변수 값은 '우량자' 그리고 '불량자'로 두 개의 값을 갖는다.

### 3.2 실험 모형

#### (1) 데이터 전처리

결측 값을 포함하고 있거나 극단치를 갖고 있는 인스턴스는 제거하였으며 모든 숫자형 변수는 0에서 1사이의 값으로 정규화 시켰다. 파생 변수는 앞 기간에 대하여 뒤 기간 동안 발생한

〈표 2〉 원천 데이터 변수의 예

번호	변수 이름	설 명
1	불량 총 건수	불량 발생 총 건수(신용카드 신청 이전 불량 포함)
2	신용카드 총 건수	보유 신용카드 수
3	최초신용개설기간	최초 신용상품 개설 후 경과일
4	최초신용카드개설기간	최초 신용카드 개설 후 경과일
5	1년 이내 신용카드 개설 건수	1년 내 신용카드 개설 건수
6	조회 총 건수 90	과거 90일 전부터 현재까지 대출기관에서 해당인의 신용정보를 조회한 건수
7	조회 총 건수 183	과거 183일 전부터 현재까지 대출기관에서 해당인의 신용정보를 조회한 건수
8	조회 총 건수 365	과거 365일 전부터 현재까지 대출기관에서 해당인의 신용정보를 조회한 건수
9	대출 총 건수	여러 금융기관으로부터 받은 대출의 현재 총 건수
...	.....	.....
72	신용판매 사용금액 90	과거 90일 전부터 현재까지 사용한 신용판매 금액
73	신용판매 사용금액 180	과거 180일 전부터 현재까지 사용한 신용판매 금액
74	신용판매 사용금액 365	과거 365일 전부터 현재까지 사용한 신용판매 금액

값의 증감율을 나타내는 값을 가지도록 만들어 사용하였다. 이렇게 증감율을 파생변수로 사용한 이유는 고객의 신용도 변화는 보다 최근의 카드 사용 패턴의 변화와 밀접하게 관계가 있다고 예상했기 때문이다. 파생된 변수는 총 12개이며 파생 변수를 만드는데 사용된 변수들을 제외하고 남은 원 변수는 14개이다. 따라서 전처리 후 26개의 입력 후보 변수와 한 개의 출력변수로 데이터 세트를 구성하였다. 이외에 실험 후 모델의 유틸리티 값을 계산하기 위한 변수 두 개가 추가되었는데, 이 두 개의 변수는 '신용판매 사용금액'과 '현금 서비스 사용금액'이다. 총 인스턴스의 수는 80,518개이며 이 중 70%인 56,313개는 학습용 데이터로 사용하였고 나머지 30%인 24,155개는 테스트용 데이터로 사용하였다.

예측 모형을 만드는데는 가장 널리 쓰이고 있는 데이터마이닝 기법인 인공신경망과 의사결정나무(C5.0)를 사용하였다. 예측률이 가장 좋은 모형을 찾기 위하여 신경망 모형의 민감도 분석과 Information Gain 값을 이용하여 입력변수들의 중요도 순위를 산출한 후, 가장 중요도가 높은 변수에서부터 가장 중요도가 낮은 변수까지 하나씩 순서대로 모형에 변수를 추가시키며 예측률을 관찰하였다. 최종 예측 모형은 모형의 예측률이 가장 높게 나타나는 입력변수의 구성으로 결정하였다. 이렇게 찾아진 예측 모형의 유틸리티를 계산하여 실제 선정된 예측모형이 가져다 줄 수 있는 수익을 추정해 보았다. 신경망 모형의 민감도 분석과 예측 모형의 생성은 Clementine v10.1을 이용하였고 Information Gain의 계산은 Weka3-4-10<sup>3)</sup>을 사용하였다.

인공신경망 모형의 경우 Clementine v10.1에서 제공하고 있는 '동적 모형'을 이용하여 최적의 위상을 찾았는데, 하나의 입력층과, 하나의

출력층 그리고 각각 두 개씩의 노드를 갖는 두 개의 은닉층으로 구성된 위상으로 결정되었다. 인공신경망 모형의 경우 가지치기 강도는 72로 주었으며, 자식마디의 최소 레코드 수는 2개로 하여 나무를 형성하였다.

## (2) 유틸리티 계산

본 실험에서 사용된 예측 모형의 유틸리티는 모형의 사용자가 영업행위를 수행하면서 예측 모형을 사용했을 때 기대할 수 있는 수익으로 정의하였다. 본 실험에서 사용된 데이터처럼 이진형의 목적 변수를 갖는 경우 아래 <그림 2>와 같이 모형의 예측결과를 분할해서 고려할 수 있다. 실제치가 '우량자'이고 모형의 예측치 역시 '우량자'라면 '가' 영역에 해당한다. 이 경우에는 실제 우량 고객에게 대출 상품이 판매 되었을 것이기 때문에 모델을 사용하는 신용카드사의 입장에서는 고객에게 대출된 금액에 대한 수수료 수익 등이 발생하게 된다. 그러나 '다' 영역의 경우 실제 우량 고객임에도 불구하고 불량 고객으로 예측 되었기 때문에, 만약 대출이 발생했었다면 취할 수 있었던 수수료 수익 등을 얻지 못한 경우가 된다. 즉, 모델의 사용자에게는 모델의 잘못된 예측 결과로 인하여 사실상 손실이 발생한 것과 마찬가지로 볼 수 있다.

	실제치	우량자	불량자
예측치			
우량자		가	나
불량자		다	라

<그림 2> 예측모델의 예측결과 유형

실제치가 '불량자'인 경우에도 동일하게 생각할 수 있다. 실제치가 '불량자'임에도 불구하고 예측 모형이 '우량자'로 예측한 경우('나' 영역)에는 실제 대출이 발생되기 때문에 대출된 금액

3) <http://www.cs.waikato.ac.nz/~ml/index.html>.



과 그 금액으로 얻을 수 있었던 수수료 수익만큼의 손실이 발생한다고 볼 수 있다. 그러나 실제치가 '불량자'인 경우 예측 모형이 다행히도 '불량자'라고 예측을 해 준다면('라' 영역) 대출되었을 뻔했던 만큼의 금액과 그 금액으로 얻을 수 있었던 수수료 수익을 손실에서 보전해 주었다고 볼 수 있다. 즉, 대출 되어서 발생했을 손실만큼의 수익이 예측 모델을 사용함으로써 발생했다고 볼 수 있다. 그러나 '라' 영역의 경우 실제 대출은 발생되지 않기 때문에 발생한 손실 및 수익을 0원으로도 볼 수 있을 것이다. 그러나 여기서 우리가 사용한 유틸리티는 예측 모형의 가치를 가늠해 보고자 한 것이기 때문에 보전 시켜 준 액수만큼을 모델의 수익으로 생각하였다. '라' 영역의 경우가 수익적 측면에서 가장 중요한 부분이기 때문에 '라' 영역에 대한 모형의 예측 가치를 0으로 볼 수는 없기 때문이다. 예측 모형을 사용했을 때, 수익이 발생할 수 있는 영역은 '가' 영역과 '라' 영역이다. 그리고 '나' 영역과 '다' 영역은 예측 모형을 사용했을 때 모델의 예측결과에 의해 손실이 발생하는 영역이라고 정리할 수 있다.

신용카드사의 수익구조는 사실 매우 복잡하게 구성되어 있다. 카드사마다 서로 다른 수익 모형으로 운영되고 있다. 신용판매의 경우에도 여러 가지 할부제도 및 협력사와의 관계 속에서 다양한 형태로 개발되어 있으며 현금 서비스 상품의 경우에도 개인마다 상이한 수수료를 및 이자율이 적용될 뿐만 아니라 초단기 카드 대출상품 이면서도 대출 기간에 있어서도 단기에서 중장기까지 사실상 다양한 형태의 상품들이 변형되어 개발돼 있다. 따라서 본 연구에서는 신용카드사의 수익구조를 이루는 여러 가지 형태의 상품들을 단순화 시켜 크게 대표적인 두 가지의 상품으로 일반화 하였다.

첫 번째는 신용판매 상품이다. 이것은 고객들

이 신용카드를 이용하여 물품이나 서비스를 구매한 경우를 말한다. 이 경우에는 고객이 사용한 신용판매 사용금액에 대하여 일정 부분의 수수료 수익이 발생한다. 이 수수료의 경우 업종마다 매우 상이하게 적용되는데 실험에 사용된 데이터에는 각 거래가 발생한 업종이 포함되어 있지 않았기 때문에 우리의 모형에서는 업종간의 평균인 3.25%를 사용하였다. 두 번째는 현금 서비스 상품이다. 현금 서비스의 경우 대출에서 상환까지의 기간이 보통 30일인 초단기 상품이다. 현금 서비스의 경우 수익이 두 가지로 분리된다. 첫 번째는 현금 서비스 사용금액에 대한 수수료이며 두 번째는 현금 서비스에 대한 취급 수수료이다. 이 중 수수료는 보통 이자라고 일컫는 부분인데 고객의 신용 등급 및 대출 기간에 따라 다양하게 적용이 되지만 역시 관련 데이터의 부재로 인해 우리의 연구에서는 평균 수치인 월 1.74%를 적용하였다. 취급수수료는 일률적으로 현금 서비스 사용금액의 0.5%를 부과하는 것이다. 따라서 사용금액에 따라 적용되는 금액이 달라진다. 지금까지의 가정을 기반으로 해서 유틸리티 함수를 각 영역별로 만들어 보면 다음과 같다. <그림 3>은 유틸리티 함수들을 <그림 2>의 각 영역에 표시한 것이다.

$$U(가) = p(가) * \{(Q * b) + (C * cb + C * ct)\} \quad (2)$$

$$U(나) = p(나) * \{-((Q + C) + (Q * b) + (C * cb + C * ct))\} \quad (3)$$

$$U(다) = p(다) * \{-((Q * b) + (C * cb + C * ct))\} \quad (4)$$

$$U(라) = p(라) * \{(Q + C) + (Q * b) + (C * cb + C * ct)\} \quad (5)$$

위 식 (2)~식 (5)에서, 각 변수의 의미는 다음과 같다.

$U(X)$  : 'X' 영역에 대한 유틸리티

실제치 예측치	우량자	불량자
우량자	$p(가) * \{(Q * b) + (C * cb + C * ct)\}$	$p(나) * \{-((Q + C) + (Q * b) + (C * cb + C * ct))\}$
불량자	$p(다) * \{-((Q * b) + (C * cb + C * ct))\}$	$p(라) * \{(Q + C) + (Q * b) + (C * cb + C * ct)\}$

〈그림 3〉 유틸리티 함수

$p(X)$ : 모형이 전체 예측 개수 중에서 'X' 영역으로 예측한 개수의 비율

Q: 신용판매 사용금액

C: 현금서비스 사용금액

b: 신용판매 평균 수수료율(3.25%)

cb: 현금서비스 30일 이용 평균 수수료율(1.74%)

ct: 현금서비스 취급 수수료율(0.5%)

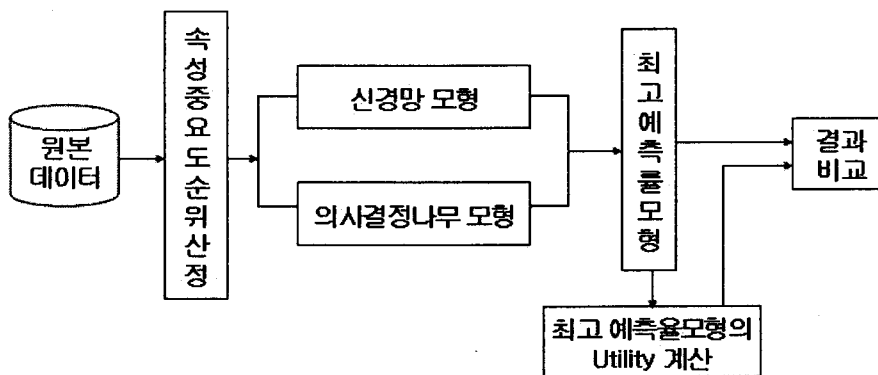
위에서 도출한 유틸리티 함수를 이용하여 모형의 예측 결과에 따라 각 고객의 유틸리티를 구한 후, 이들의 평균값을 해당 예측 모형의 유틸리티로 사용하였다. <그림 3>에서 볼 수 있듯이, 위와 같은 유틸리티 함수를 사용하게 되면 실제치 '우량자'에 대한 예측 보다 실제치 '불량자'에 대한 예측 결과가 모형의 성능 판정에 더 많은 영향을 미치게 된다. 즉, 실제치 '우량자'에 대한 예측 결과 얻게 되는 수익이나 손실은 실제치 '불량자'에 대한 예측 결과 얻게 되

는 수익이나 손실에 비해서 훨씬 작은 수치이다. 따라서 실제치 '불량자'를 잘 예측하는 모형에서 더 높은 유틸리티 값이 나올 확률이 크다고 할 수 있다.

<그림 4>는 본 연구에서 수행한 실험의 흐름을 보여주고 있다. 최초 원본데이터에서 속성선택 기법을 통해 입력될 속성의 중요도 순위를 산정한 후, 중요도 순으로 입력속성의 수를 하나씩 증가시키면서 신경망 모형과 의사결정나무 무 모형에 입력시킨 다음 각 기법에서 최고 예측률이 나온 모형을 찾고, 개발된 유틸리티 함수를 이용하여 최고 예측률 모형의 유틸리티를 계산하여 그 결과를 비교하였다.

#### 4. 실험 결과 및 해석

입력 후보 변수들의 중요도를 산정한 결과 '6개월 내 현금서비스 최고액 사용 카드 수 증감율'이 가장 높은 중요도를 보였으며 뒤이어 '6개



〈그림 4〉 실험의 흐름

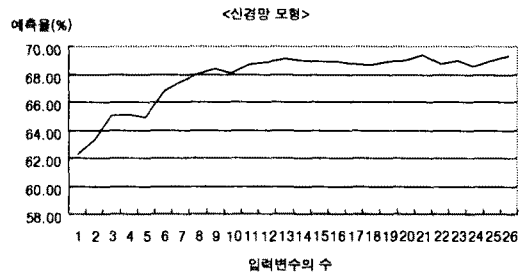
월 내 연체 최고 보유금액 증감율', '조회 총건수(저축은행업계)' 등이 높은 순위를 보였다. <표 3>에 중요도 순위에 따른 입력변수의 내용을 상위 10위까지 각 모형 별로 정리하였는데, 상위 5위권 안에 총 4개의 파생변수들이 위치하고 있는 것을 관찰할 수 있다. 이것은 전술한 바와 같이 가장 최근의 카드 사용 행위의 변화가 그 사람의 신용도와 밀접한 관계가 있을 수 있음을 말해준다.

<표 3> 중요도 순위에 따른 각 모형 별 입력변수의 내용

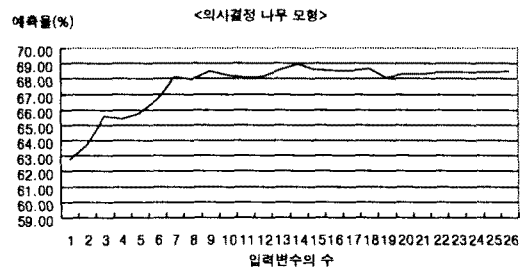
순위	변수 내용
1	(파생) 6개월 내 현금서비스 최고액 신용카드 수 증감율
2	(파생) 6개월 내 연체 최고보유금액 증감율
3	조회 총 건수(저축은행업계)
4	(파생) 6개월 내 신용판매 최고 사용액 증감율
5	(파생) 6개월 내 현금 서비스 최고사용금액 증감율
6	최초 신용카드 개설 후 경과일
7	1년 내 대출건수(신용카드업계)
8	최초 신용상품 개설 후 경과일
9	불량 발생 총 건수(신청이전 불량은 삭제)
10	(파생) 6개월 내 신용판매 최고액 신용카드 수 증감율

<그림 5>와 <그림 6>은 신경망 모형과 의사결정나무 모형의 입력 변수 증가에 따른 예측률의 변화를 보여주고 있다. 신경망 모형은 입력 변수가 21개일 때 예측률이 69.38%로 가장 높은 것을 확인할 수 있으며, 의사결정 모형은 입력 변수의 수가 14개일 때 예측률이 68.94%로 가장 높은 것을 관찰할 수 있다. 예측률의 결과로 보면 신경망 모형의 성능이 더 좋은 것으로 판단할 수 있다. 그러나 <표 4>에서 볼 수 있듯이, 이들 두 모형의 유틸리티를 계산해 보면

오히려 의사결정나무 모형의 성능이 더 좋은 것으로 판단할 수 있다. 입력변수가 21개 포함된 신경망 모형의 유틸리티는 328.20(천원)으로 14개의 입력변수로 만들어진 의사결정나무 모형의 354.46(천원)보다 1인당 26(천원)이 더 낮은 것을 확인할 수 있다.



<그림 5> 신경망 모형의 예측률 변화(테스트 데이터)



<그림 6> 의사결정나무 모형의 예측률 변화(테스트 데이터)

<표 4> 각 모형의 유틸리티(테스트 데이터)

모형	입력 변수의 수	예측률	유틸리티
신경망 모형	21	69.38%	328.20(천원)
의사결정나무 모형	14	68.94%	354.46(천원)

이러한 차이는 실제 모형들이 예측한 결과를 자세히 살펴보면 그 이유를 발견할 수 있다. <표 5>와 <표 6>은 실제 모형들이 예측한 결과를 각 경우 별 건수로 보여주고 있다. 실제치 '불량자'를 예측한 경우를 살펴보면, 신경망 모

형의 경우 실제 '불량자'를 '우량자'로 잘못 예측한 경우의 개수가 3,918건으로 의사결정나무 모형의 그것보다 340건 정도 더 손실로 작용한 것을 관찰할 수 있다. 실제치 '우량자'에 대한 예측의 경우에는 수익이 발생하는 상황인 실제 '우량자'를 '우량자'로 잘 예측한 경우에 신경망 모형이 7,537건으로 의사결정나무의 예측보다 446건을 더 잘 예측하였다. 즉, 실제 '불량자'를 잘 예측한 모형은 의사결정나무 모형이고 실제 '우량자'를 잘 예측한 모형은 신경망 모형이라고 할 수 있다.

하지만 실제의 상황에 바탕을 둔 우리의 유틸리티 함수 정의상 '불량자'에 대한 예측 결과가 '우량자'를 예측한 결과보다 유틸리티의 증감에 훨씬 더 많은 영향을 미치기 때문에 '불량자'를 잘 예측한 의사결정나무 모형이 모형의 유틸리티 측면에서는 더 우수한 평가를 받은 것이다. 즉, 신경망 모형이 전체적으로 정확히 예측한 건수는 16,758(=7,537+9,221)건으로 의사결정나무의 16,652(=7,091+9,561)건 보다 많지만 모형의 유틸리티 측면에서는 의사결정나무의 예측결과가 더 우수할 수 있다는 것이다. 이처럼 예측유틸리티를 모형의 성능 평가에 사용할 경우 좀 더 현실적인 모형 평가를 수행할 수 있을 뿐만 아니라 구체적으로 얼마의 수익 또는 손실이 예상되는지 가능해 볼 수 있게 된다.

신경망 모형과 의사결정나무 모형의 예측 결과 비교 시, 맥니마 검정(McNemar's test)을 통하여 두 모형이 통계적으로 서로 같지 않음을 검증하였다. 그 결과  $\chi^2$  통계량이 178.03으로 99

<표 5> 신경망 모형(21개 입력변수)의 예측 결과(테스트 데이터)

실제치 \ 예측치	우량자	불량자
우량자	7,537건	3,918건
불량자	3,479건	9,221건

<표 6> 의사결정나무 모형(14개 입력변수)의 예측 결과  
(테스트 데이터)

실제치 \ 예측치	우량자	불량자
우량자	7,091건	3,578건
불량자	3,925건	9,561건

% 신뢰 수준에서 두 모형의 예측 결과는 동일하지 않음이 통계적으로 유의한 것을 확인할 수 있었다.

## 5. 결 론

지금까지 많은 연구자들은 예측 모형이나 분류 모형을 만들 때 예측률을 이용하여 모형의 성능을 평가하였다. 그러나 예측률과 같은 평가 척도는 모형의 우열을 가려주는 데서 그 역할이 끝나고 만다. 즉, 개발된 모형의 결과를 해석하거나, 실무에서 해당 모형을 사용할 때, 그 해석에 제약 사항이 많고 또 그 값이 우리의 수익에 얼마나 영향을 주고 있는지는 가늠하기 쉽지 않았다. 그러나 모형이 가져다 주는 비용과 수익을 기반으로 한 유틸리티 값을 평가척도로 이용하게 되면 구체적인 금전적 단위로 예측 모형의 가치를 표현해 줄 수 있기 때문에 개발된 모형을 해석하는데 매우 용이할 뿐만 아니라 구체적인 수익에 대한 영향을 체감적으로 평가할 수도 있다.

따라서 본 연구에서는 모형의 유틸리티를 평가척도로 이용하기 위하여 유틸리티 함수를 개발하고 이를 실제 데이터에 적용시켜 보았으며 그 결과 다음과 같은 결론들을 얻어낼 수 있었다. 첫째, 앞장의 실험결과에서 볼 수 있듯이 예측 모형의 성능을 예측률로 측정한 결과와 유틸리티로 측정한 결과는 서로 다르게 나타날 수 있다. 따라서 예측 모형을 통해 얻을 수 있는

수익의 정도를 표현해 주는 유틸리티를 모형의 평가척도로 고려하는 것이 수익 측면에서 잘못된 모형 선정을 하게 되는 위험을 줄여줄 수 있다. 둘째, 유틸리티는 구체적인 금전적 가치를 표현해 주기 때문에 개발된 모형의 이용도를 높여줄 수 있으며 데이터마이닝 각 단계에서 수행하게 되는 여러 가지 작업에서 의사결정에 이용될 수 있다. 예를 들어 변수의 중요도를 이용하여 속성선택 단계를 수행할 때 총 몇 개까지의 입력변수를 모형에 포함시켜야 할 것인가를 결정하는데 있어서, 변수 하나를 포함 시키는 데 필요한 비용과 해당 변수 하나를 포함시킴으로써 증가하는 유틸리티의 양을 비교함으로써 가장 최적의 입력 변수의 수를 결정할 수도 있을 것이다.

본 연구는 모형의 평가척도로서 유틸리티가 매우 유용하게 사용될 수 있는 가능성을 보여주었다. 기존의 평가척도로 측정하여 최고의 분류 모형이라고 평가되었던 모형이 반드시 최고의 수익을 가져다 주는 모형이 아닐 수 있다는 사실을 실제 데이터를 사용한 실험을 통하여 발견했으며, 유틸리티 기반의 평가 척도를 사용함으로써 모형간 성능차이를 체계적인 구체적 양으로 표현할 수 있었다.

그러나 본 연구에서 사용된 유틸리티 기반의 평가척도는 모형의 개발이 모두 끝난 후, 개발된 모형에 기대할 수 있는 유틸리티를 가늠하는 것이다. 향후에는 모형의 개발 단계에서부터 유틸리티를 높일 수 있는 학습방법이나 기법에 대한 연구가 필요하다. 그리고 유틸리티를 사용하기 위해서는 데이터 수집 단계에서부터 유틸리티 값을 계산해 낼 수 있는 데이터를 준비해야 한다. 유틸리티를 계산하기 위한 데이터가 다양하게 존재할 때 좀 더 정확한 유틸리티 값을 계산할 수 있기 때문이다. 예를 들어, 실제 신용카드사들의 고객 분류 유형에 있어서 '우량군'

과 '불량군' 이외에 '우량 가능군'<sup>4)</sup>이란 집단이 추가되어 고객을 삼분하여 관리하는 경우가 있는데 분류모형 개발 시 이들을 고려하기 위해서는 이들을 예측하면서 발생하는 비용과 수익에 대한 데이터를 확보해야만 한다. 그러나 이러한 데이터는 수집단계에서부터 고려하지 않으면 얻을 수 없는 값들이 대부분이다. 본 연구에서도 유틸리티를 계산할 때 많은 가정을 할 수 밖에 없었다. 업종마다 매우 상이한 카드 사용 수수료율도 각 거래가 어느 업종에서 발생했는지에 대한 데이터가 없었기 때문에 평균 수수료율을 이용할 수 밖에 없었으며, 현금서비스에 대한 수수료도 각 개인별, 기간별로 매우 상이하지만 단순화 시켜서 평균값을 사용할 수 밖에 없었다. 이러한 값들을 좀 더 상세하게 반영하기 위해서 데이터 수집 단계에서부터 유틸리티 함수를 고려해야 할 것이다. 또한 본 연구에서 제시한 유틸리티 함수는 신용카드사의 수익 모형에만 적용될 수 있는 것이다. 따라서 다른 수익모형을 사용하는 회사에서는 그에 적합한 유틸리티 함수를 새로 정의해서 사용해야 할 것이다.

## 참고 문헌

- [1] 전국 은행 연합회, 신용정보 관리규약, 대한민국, 2005.
- [2] Arnt, A. and Zilberstein, S., "Learning Policies for Sequential Time and Cost Sensitive Classification", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2005, pp. 39-45.

4) 현재 연체가 발생한 고객이지만 사실상우량으로 이동할 재정적 능력과 의지를 갖고 있는 고객들을 말한다. 이들은우량이 회사에 발생시켜 주는 수수료 수익 이외에도 연체료라고 하는 부가적인 수익을 발생시켜 줌으로써 사실상 가장 수익률이 좋은 집단이 될 수도 있다

- [3] Blöchliger, A. and Leippold, M., "Economic Benefit of Powerful Credit Scoring", *Journal of Banking and Finance*, Vol. 30, 2006, pp. 851-873.
- [4] Brill, J., "The Importance of Credit Scoring Models in Improving Cash Flow and Collection", *Business Credit*, Vol. 100, No. 1, 1998, pp. 16-17.
- [5] Cavalluzzo, K., "Market structure and discrimination : the case of small businesses", *Journal of Money, Credit, and Banking*, Vol. 30, No. 4, 1998, pp. 771-793.
- [6] Cavalluzzo, K., Cavalluzzo, L., Wolken, J., "Competition, Small Business Financing and Discrimination : Evidence from a New Survey", *Business Access to Capital and Credit*, *Federal Reserve System Research Conference Proceeding*, 1999, pp. 180-266.
- [7] Chawla, N. V., Hall, L. O. and Joshi, A., "Wrapper-based Computation and Evaluation of Sampling Methods for Imbalanced Datasets", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2005, pp. 24-33.
- [8] Chawla, N. and Li, X., "Pricing Based Framework for Benefit Scoring", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2006, pp. 65-69.
- [9] Ciraco, M., Rogalewski, M. and Weiss, G., "Improving Classifier Utility by Altering the Misclassification Cost Ratio", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2005, pp. 46-52.
- [10] Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P. and Jaulent, M.-C., "Models to Predict Cardiovascular Risk : Comparison of CART, Multilayer Perceptron and Logistic Regression", *Proceedings AMIA Symposium*, 2000, pp. 156-160.
- [11] Desai, V. S., Crook, J. N. and Overstreet, G. A., "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment", *European Journal of Operational Research*, Vol. 95, No. 1, 1996, pp. 24-37.
- [12] Desai, V. S., Conway, D. G., Crook, J. N. and Overstreet, G. A., "Credit-Scoring Models in the Credit-Union Environment Using Neural Networks and Genetic Algorithms", *IMA Journal of Mathematics Applied in Business and Industry*, Vol. 8, No. 4, 1997, pp. 323-346.
- [13] Han, J. and Kamber, M., *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
- [14] Hand, D. J. and Henley, W. E., "Statistical Classification Methods in Consumer Credit Scoring : a Review", *Journal of Royal Statistical Society, Series A(Statistics in Society)*, Vol. 160, No. 3, 1997, pp. 523-541.
- [15] Hoffmann, F., Baesens, B., Mues, C., Gestel, T. V. and Vanthienen, J., "Inferring Descriptive and Approximate fuzzy rules for Credit Scoring using Evolutionary Algorithms", *European Journal of Operational Research*, Vol. 177, 2007, pp. 540-555.
- [16] Holte, R. C. and Drummond, C., "Cost-Sensitive Classifier Evaluation", *Proceed-*

- ings of *KDD Workshop on Utility-Based Data Mining*, 2005, pp. 3-9.
- [17] Hsieh, N.-C., "Hybrid Mining Approach in the Design of Credit Scoring Models", *Expert Systems with Applications*, Vol. 28, 2005, pp. 655-665.
- [18] Huang, C.-L., Chen, M.-C., Wang, C.-J., "Credit Scoring with a Data Mining Approach based on Support Vector Machines", *Expert Systems with Applications*, Vol. 33, 2007, pp. 847-856.
- [19] Huang, J.-J., Tzeng, G.-H. and Ong, C.-S., "Two-stage Genetic Programming for the Credit Scoring Model", *Applied Mathematics and Computation*, Vol. 174, 2006, pp. 1039-1053.
- [20] Jesen, L. L., "Using Neural Networks for Credit Scoring", *Management Finance*, Vol. 18, No. 1, 1992, pp. 15-26.
- [21] Lee, T.-S., Chiu, C.-C., Lu, C.-J. and Chen, I.-F., "Credit Scoring using the Hybrid Neural Discriminant Technique", *Expert Systems with Applications*, Vol. 23, 2002, pp. 245-254.
- [22] Lee, T.-S., Chen, I.-F., "A Two-Stage Hybrid Credit Scoring Model using Artificial Neural Networks and Multivariate Adaptive Regression Splines", *Expert Systems with Applications*, Vol. 28, 2005, pp. 743-752.
- [23] Ling, C. X., Sheng, V. S., Bruckhaus, T. and Madhavji, N. H., "Maximum Profit Mining and Its Application in Software Development", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2006, pp. 70-75.
- [24] McCarthy, K., Zabar, B. and Weiss, G., "Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2005, pp. 69-77.
- [25] Melville P., Saar-Tsechansky, M., Provost, F. and Mooney, R., "Economical Active Feature-Value Acquisition through Expected Utility Estimation", *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2005a, pp. 10-16.
- [26] Melville P., Saar-Tsechansky, M., Provost, F. and Mooney, R., "An Expected Utility Approach to Active Feature-Value Acquisition", *Proceedings of the 15<sup>th</sup> IEEE International Conference on Data Mining*, 2005b.
- [27] Piramuthu, S., "Financial Credit-Risk Evaluation with Neural and Neurofuzzy Systems", *European Journal of Operational Research*, Vol. 112, No. 2, 1999, pp. 310-321.
- [28] Stark, K. D. C. and Pfeiffer, D. U., "The Application of Non-Parametric Techniques to Solve Classification Problems in Complex Data Set in Veterinary Epidemiology-An Example", *Intelligent Data Analysis*, Vol. 3, No. 1, 1999, pp. 23-35.
- [29] Thomas, L. C., "A Survey of Credit and Behavioural Scoring : Forecasting Financial Risk of Lending to Consumers", *International Journal of Forecasting*, Vol. 16, 2000, pp. 149-172.
- [30] Tsaih, R., Liu, Y.-J., Liu, W. and Lien, Y.-L., "Credit Scoring System for Small

Business Loans”, *Decision Support Systems*, Vol. 38, 2004, pp. 91-99.

- [31] Tseng, V. S., Chu, C.-J. and Liang, T., “Efficient Mining of Temporal High Utility Itemsets from Data Streams”, *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2006, pp. 18-26.
- [32] West, D., “Neural Network Credit Scoring Models”, *Computers and Operations Research*, Vol. 27, 2000, pp. 1131-1152.
- [33] Weiss G. M. and Tian, Y., “Maximizing Classifier Utility when Training Data is Costly”, *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2006, pp. 3-11.
- [34] Yan, L. and Baldasare, P., “Beyond Classification and Ranking : Constrained Optimization of the ROI”, *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2006, pp. 58-64.
- [35] Yao, H., Hamilton, H. J. and Geng, L., “A Unified framework for Utility Based Measures for Mining Itemsets”, *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2006, pp. 28-37.
- [36] Zadrozny, B., “One-Benefit Learning : Cost-Sensitive Learning with Restricted Cost Information”, *Proceedings of KDD Workshop on Utility-Based Data Mining*, 2005,

pp. 53-58.

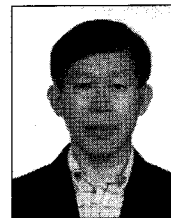
#### ■ 저자소개



##### 정 석 훈

아주대학교에서 정치학과와 경영학사를 취득하고 동 대학원에서 경영정보학 석사학위를 취득하였다. (주)EC Miner Data Mining 팀과 (주)대우증권 IT

Center에 근무하였으며, 현재 고려대학교 대학원 경영학과 박사과정에 재학 중이다. *Journal of Information Technology Applications and Management*, *경영정보학연구*, *Expert Systems with Applications* 그리고 *Entrue Journal of Information Technology* 등에 논문을 게재하였으며 주요 관심분야는 Business Intelligence, Utility based Data Mining 등이다.



##### 서 용 무

서울대학교 사범대학 수학과, 한국과학기술원 전산학과를 졸업하고, 한국과학기술 연구소 전산센터에서 연구원으로 재직 시

도미하여, University of Texas (at Austin)에서 전산학 석사, 경영정보학 박사를 취득한 후, 세종대학교, 건국대학교를 거쳐, 현재 고려대학교 경영대학에 재직하고 있다. 주요 관심분야는 use of ontology in business environment, business process management, e-collaboration, data mining 등이다.