효율적인 신용평가를 위한 데이터마이닝 모형의 비교·분석에 관한 연구

김 갑 식*

Study on the Comparison and Analysis of Data Mining Models for the Efficient Customer Credit Evaluation

Kap-Sik Kim*

Abstract

This study is intended to suggest the optimized data mining model for the efficient customer credit evaluation in the capital finance industry. To accomplish the research objective, various data mining models for the customer credit evaluation are compared and analyzed. Furthermore, existing models such as Multi-Layered Perceptrons, Multivariate Discrimination Analysis, Radial Basis Function, Decision Tree, and Logistic Regression are employed for analyzing the customer information in the capital finance market and the detailed data of capital financing transactions. Finally, the data from the integrated model utilizing a genetic algorithm is compared with those of each individual model mentioned above. The results reveals that the integrated model is superior to other existing models.

Keywords : Genetic Algorithm, Data Mining Models, Customer Credit Evaluation

논문접수일 : 2003년 10월 4일 논문게재확정일 : 2004년 3월 7일

※ 본 논문은 2003년도 대구산업정보대학 연구조성비(교비)에 의한 것임.

* 대구산업정보대학 컴퓨터정보계열 교수

1. 서 론

할부금융거래에서 고객에 대한 신용평가는 우·불량고객의 판별 및 고객 신용등급의 관리에 활용되어 오고 있다. 특히, 신용평가는 불량채권 발생률을 미연에 감소시켜 궁극적으로 할부금융회사의 수익을 증대시키는 역할을 수행하므로 그 중요성이 매우 크다. 또한 신용평가를 위해 수집된 자료들과 신용평가결과를 바탕으로 할부금융회사는 고객에 따라 차별화 된 금융상품과 여러 가지 혜택을 제공한다. 뿐만 아니라, 위험상황을 사전에 통지하는 등의 고객관계마케팅(CRM)을 실현할 수 있게 되므로 신용평가는 할부금융회사의 운영에 있어서 필수적인 부분이라 하겠다[김홍철, 2001].

신용평가는 신규고객이 융자를 처음 신청할 때, 그 고객이 제시하는 인구통계학적 자료만을 가지고 재정적인 위험을 판단하는 협의의 신용평가와 인구통계학적 자료 이외에 기존 고객의 거래 내역에 의해 그 고객의 현재 상태를 평가하는 행태평가(behavior scoring)로 대별된다. 이 두 가지 모두 같은 방식으로 측정할 수 있으나 입력되는 자료에 있어 후자의 경우에는 전자에서 사용된 자료 이외에 거래내역이 포함된다는 점에서 차이가 있다[Thomas, 2000].

협의의 신용평가와 행태평가 모두를 포함한 광의의 신용평가(이하 신용평가)는 데이터마이닝의 영역 중에 하나인 분류(classification) 문제이다. 분류 문제란 미리 몇 개의 부류(class)를 설정하여 놓고 어떤 부류에 속하는지를 판단하는 것으로서, 일반적으로 통계적 기법이나 인공지능기법을 손쉽게 사용할 수 있는 문제영역으로서 데이터마이닝에서 가장 중요한 영역들 중의 하나에 속한다[Berry and Linoff, 1997].

신용평가문제를 해결하기 위해서 학계에서는 꽤 오래 전부터 다양한 모형들이 신용평가에 사용되어 왔다. 대표적인 예가 선형회귀분석이나 다변량판별분석과 같은 전통적인 통계모형과 선형계획법과 같은 경영과학모형 이외에도 의사결정나무, 인공신경망과 같은 모형들을 들 수 있다. 그러나 실제 이와 같이 여러 가지 모형들을 다양하게 적용했음에도 불구하고 어떠한 모형도 뚜렷하게 우위를 보인 접근법은 없는 것으로 알려져 있다[Thomas, 2000]. 특히 1980년대 이후, 현저하게 많이 사용되고 있는 인공 신경망이나 의사결정나무 등과 같은 학습을 기반으로 하는 인공지능모형에서 뚜렷하게 우위를 보인 접근법이 없는 데는 여러 가지 이유가 있겠으나 가장 근본적인 이유는 학습방법이 항상 과학습(overfitting)이나 최적모형 설정의 어려움 또는 학습방식의 부적절성 등과 같은 문제 때문에 광역최적(global optimum)을 보장하지 못하기 때문이다. 이와 같은 현상은 학습을 기반으로 하고 있는 단일 모형에 의존하는 기법이 가지고 있는 한계점이라 할 수 있다[Hansen, 1999].

단일 모형의 기계학습 모형의 한계를 극복하기 위해 대두된 접근방식이 여러 가지 모형을 결합하는 것이다. 다양한 모형을 결합할 경우 각 단일 모형의 과학습의 문제 또는 학습방식의 부적절성을 해소시킬 가능성이 높기 때문에 여러 분류문제에서 널리 사용되어 왔다. 모형을 결합하는 방식으로 대표적인 방식은 각 모형의 값에 대한 단순평균과 각 모형별 중요도를 고려한 가중평균 그리고 다수결에 의한 투표(voting) 방식, Borda 카운팅에 의한 투표 방식 등을 들 수 있다[Cho, 1999 ; Kim, et al., 2002].

본 연구에서는 할부금융 서비스에 의해 물품을 구입한 일반 개인 고객들의 신용평가, 특히 거래내역을 기반으로 한 행태 평가를 위한 방법론으로 단일모형을 결합한 통합모형을 얻고자 한다. 이를 위하여 본 연구에서는 기존 신용평

가 모형에서 많이 사용되고 있는 기법인 인공신경망의 다계층 퍼셉트론(Multi-Layered Perceptrons : MLP)과 반경기반함수(Radial Basis Function : RBF), 대표적 선형모형인 다변량 판별분석(Multivariate Discrimination Analysis : MDA), 로지스틱 회귀분석(Logistic Regression : LR) 그리고 의사결정나무(Desicion Tree) 등을 이용하여 각각의 단일모형을 얻어 신용평가의 예측결과를 비교·분석한 후, 유전자알고리즘(genetic algorithm)방식에 의해 이들 각각의 단일 모형을 결합해서 이 통합모형과 각 단일모형을 비교·분석하여 할부금융 이용고객의 행태평가에 의한 신용평가 예측의 최적화를 제시하려는데 그 목적을 두고 있다.

2. 이론적 배경

전통적으로 신용평가를 위한 기법으로는 로지스틱 회귀분석이나 프로빗 분석법과 같은 통계학적 기법[Wington, 1980 ; Grablowsky and Talley, 1981]과 선형계획법과 같은 경영과학적 기법을 들 수 있다.

통계적기법 또는 경영과학적 기법이 고객에 대한 신용을 점수화하여 평가하는데 비해 의사결정트리의 경우는 고객을 성격에 따라 그룹화하는 방식을 취함으로써 우량고객과 불량고객에 대한 분류를 좀 더 알기 쉽게 하고 있기 때문에 신용평가에서 널리 사용되고 있는 기법이다[Carter and Cartlett, 1987].

한편 1980년대 중반부터 경영분야에서 널리 사용되기 시작한 인공신경망 모형은 통계적 가설이 필요 없으면서도 비선형적인 회귀모형을 설명하기에 적당하기 때문에 신용평가에서 널리 사용되어 뛰어난 성과를 보여 주고 있다[Altman, Marco and Varetto, 1994 ; Desai, Convay, Crook and Overstreet, 1997].

최근 들어 의사결정나무(decision tree), 인공신경망 등의 인공지능(artificial intelligence) 모형을 이용한 연구가 활발하게 진행되었고, 그 중에서도 특히 인공신경망 모형을 이용한 연구가 좋은 결과를 보여주고 있다[West, 2000 ; Jain and Nag, 1997]. 이 모형은 다양한 응용분야를 가지고 있는 많은 문제들에 대해 널리 적용될 수 있고, 통계적 가정이 필요 없으면서도 비선형적인 회귀모형을 설명하기에 적당하며, 신용평가에 매우 적합하다고 증명되었다[Cheng & Titterington, 1994]. 판별분석(discrimination Analysis)은 사회현상의 여러 특성들을 토대로 하여 주어진 상황에서 응답자들이 어떻게 행동할 것인지를 예측하는 하나의 통계모형이다[정충영, 최이규, 1998]. 구현이 간단하고 학습시간도 짧지만 독립변수들이 다변량 판별분석의 기본적인 통계학적 가정들을 만족해야 하므로 이에 대한 검증이 필요하다는 한계점을 가지고 있다[채서일, 1999]. 의사결정나무는 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것으로, 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다[최종후, 한상태, 2000]. 이 모형은 빠르고 간단할 뿐만 아니라 이해하기 쉬운 규칙으로 전환될 수 있으며[Imielinski & Mannila, 1996] 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다[Mehta, 1968].

이들 모형통합에 강력한 방법론인 유전자 알고리즘은 인공지능의 한 모형으로서 2차원 이상의 복잡한 탐색공간에서 전 범위의 최적해(global optimal solution)를 탐색하는데 아주 효율적이며, 유연하다고 증명되어져 왔다[Gupta, 1995]. 이러한 유전자 알고리즘은 새로운 집단

(new population)을 형성할 때에 과거의 집단(oldpopulation)에서 높은 적합도를 가지는 개체(string)가 높은 확률을 가지고 새로운 집단으로 유전한다는 것이 그 기본적인 원리이다[Hon & Chi, 1994 ; Cho, 1999]는 유전자 알고리즘을 통한 인공신경망 결합모형을 이용해서 결합방법(hybrid method)에 의한 통합모형의 예측 성능향상을 연구하였다.

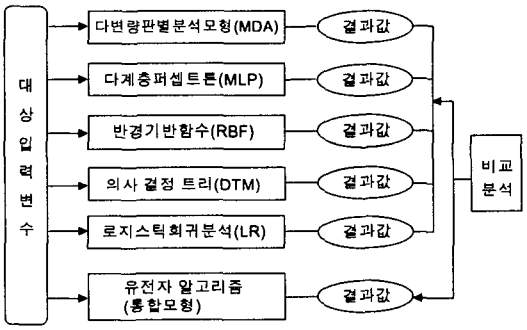
이상의 데이터마이닝 기법을 이용한 신용평가에 대한 기존연구를 살펴보면, Thomas[2000]는 신용 및 행동 스코어링 연구에서 기업이 신용 스코어링과 행동 스코어링 기법을 소비자의 신용을 승인할 것인지 혹은 거부할 것인가와 관련하여 기존의 신용평가 스코어링에 대한 연구를 종합적으로 검토하였다. West[2000]는 신경망을 이용한 신용평가 모델에서 MOE, RBF, MLP, LVQ, FAR 이 다섯 개의 인공신경망들이 얼마나 정확하게 신용정도를 측정할 수 있는지를 연구해서 MLP모형이 압도적으로 우수한 모형이 아니라 다른 모형도 나름대로의 좋은 장점이 있다는 것을 증명하였다. Feelders[1999]는 EM-알고리즘을 이용하여 혼합모형의 예측성능을 높여주는 실험연구를 시도하였는데, 이 연구는 혼합모형이라는 측면에서 신용평가에 대한 예측성능 향상을 위해 통합모형을 시도한 의의가 있다. Kim과 Street[2003]는 유전자 알고리즘과 인공신경망을 이용하여 개발된 모델이 우편광고의 수익성을 예측할 수 있음을 보여주었는데, 개발된 모델이 민감도 분석을 통하여 다른 모델보다 우수함을 입증하였다.

3. 연구의 설계

3.1 연구모형

본 연구에서는 할부금융시장에서의 고객정보 및 할부진행과정에 대한 세부 내역을 바탕으로 여러 가지 개별분류모형(classifier)들을 유전자 알고리즘을 이용하여 통합한 신용예측모형을 개발해서 이 통합모형과 각 단일모형을 비교·분석해서 최적의 신용평가모형을 제안한다. <그림 1>은 개별분류모형들이 각각의 결과 값들을 도출하고 이 결과 값들을 통합모듈(combining module)에서 가중치를 주어 통합하여 하나의 결과 값을 얻은 후 이 통합모형과 각 단일모형을 비교하는 것을 나타낸 것이다.

통합할 다섯 가지 단위 분류모형으로는 인공신경망의 MLP와 RBF모형, 다변량 판별분석모형(MDA), 의사결정모형, 로지스틱 회귀분석 모형 등이 사용되며, 이들 단위 모형의 학습이 마친 후에는 대상입력변수들 중 각 단위모형별로 선택된 입력변수의 값을 기준으로 하여 각각의 측정 결과 값을 도출한다. 이렇게 해서 도출된 각 단위모형의 결과 값들을 유전자알고리즘 의한 가중치 최적화를 통하여 최종결과 값으로 가중 통합하는 것이다.이렇게 통합된 결과 값을 다섯 개의 각 단위모형과 서로 비교해 본다.



<그림 1> 연구 모형

3.2 표본 및 모형 적용

본 연구에서 사용된 표본자료는 1997년 7월부터 2000년 5월까지의 국내 X할부금융회사의

고객정보 및 할부진행과정에 대한 데이터이다. 약 200,000개의 개인별 데이터를 대상으로 missing value가 없는 데이터 중 신용우량과 불량을 판단 기준으로 하여 총 6,500개의 데이터를 추출하였다. 이 중에서 <표 1>에서 보는 바와 같이, 개별 분류 모형 개발에 3,500개를 사용하였는데, 이것을 다시 학습(training) 1,750개, 검증(validation) 875개, 시험(test) 875개를 사용하였다.그리고 개별모형 예측성능평가에는 앞서 사용한 3,500개를 제외한 다른 1,000개의 데이터를 사용하였다. 유전자 알고리즘을 이용한 개별모형의 통합(학습용)에는 아직까지 사용하지 않은 데이터 중에서 1,000개를 우량 450개, 불량 450개, 미정 100개의 적정비율로 추출하여 사용하였고, 통합모형의 최종 예측성능평가(scoring)에 또 다른 1,000개의 데이터를 사용하였다.

3.3 변 수

<표 2>의 변수들은 원시데이터를 예측모형의 입력변수로 사용하기 위해 정규화 등의 과정을 거쳐 적절히 가공한 변수목록이다. 변수 B19는 채권의 우·불량을 판별하는 종속변수로서 X할부금융회사 자체기준에 의해 판단기간(1999년 2월~7월)동안의 연체 개월 수가 4개월 이상이면 1(불량), 3개월인 것은 2(미정), 2개월 이하이면 3(우량)의 값을 갖는다. 나머지 변수들(채권번호 제외)은 대상입력변수들이며 금액과 관련된 변수들은 평균값으로 나누어주는 방법을 통해 정규화 하였다.

〈표 2〉 변수 상세 설명

변수명	설 명
A1	나 이
A2	성 별
A3	보증인수
A4	**매입지역**
ACA5	차량원부
A6	차 종
A7	차량년식
A8	배기량
A9	신용조사방법
A10	구매자구분
B1	3개월의무납입액평균/3개월잔액평균
B2	6개월의무납입액평균/6개월잔액평균
B3	1998년 1월 의무납입액/잔액
B4	**1998년 12월 의무납입액/잔액**
B5	1998년 1월 실납입액/의무납입액
B6	**1998년 12월 실납입액/의무납입액**
B7	3개월 납입액평균/3개월 의무납입액평균
B8	6개월 납입액평균/6개월 의무납입액평균
B9	**12개월 납입액평균/12개월 의무납입액평균 (1998년 1월 납입액을 이전 6개월 평균으로)**
B10	**1998년 12개월간 최장연체횟수**
B11	1998년 12월 잔액 / 3개월 잔액평균
B12	1998년 12월 잔액 / 6개월 잔액평균
B13	1998년 12개월간 연체액평균
B14	1998년 12개월간 연체개월수/총할부개월수
B15	1998년 12개월간 연체개월수/12개월
B16	매월 납부액
B17	**총할부개월수**
B18	할부가격(할부원금+할부이자)
B19	우·불량판별(1 : 불량, 2 : 미정, 3 : 우량)

주) 변수 설명 중에서 3개월은 1997년 11월~1998년 1월 6개월은 1997년 8월~1998년 1월을 말한다.)

〈표 1〉 표본데이터의 사용내역

용 도	표본수	균형화(Balancing)
개별 분류모형의 개발 (학습, 검증, 시험)	3,500	(우량 ; 1500, 불량 ; 1500, 미정 ; 500) (학습 ; 1750, 검증 ; 875, 시험 ; 875)
개별모형 예측성 평가(scoring)	1,000	
유전자 알고리즘을 이용한 개별모형의 통합 (학습용)	1,000	(우량 ; 450, 불량 ; 450, 미정 ; 100)
최종 예측성 평가(scoring)	1,000	

<표 2>에 나타난 28개의 변수목록은 신용평가와 관련된 기존의 많은 변수 중 객관적인 신뢰성이 떨어지는 인구통계학적인 정보들을 제외하고 신용평가시 신뢰성이 높은 변수라고 판단할 수 있는 변수들을 정리한 것이다.

<표 2>의 변수들을 살펴보면 B1, B2, B7, B8, B9, B11, B12 등의 변수에 관측기간 이전의 할부 진행 기록들을 반영하기 위하여 관측기간 이전 3개월, 또는 6개월의 할부 내역을 반영시켰으며각 입력변수들의 값이 개월 수, 금액 등으로그 스케일이 현저하게 차이가 나기 때문에 이를 1에서 0사이의 실수 값으로 만들어 주기 위해 변수 값들을 해당하는 변수의 평균값으로 나누어주는 방법을 통해 정규화 하였다.

3.4 유전자알고리즘의 개요

유전자 알고리즘은 생물의 진화를 모방한 탐색기법으로, 최적화문제, 확률적 탐색 등에 많이 적용되고 있다. 유전자 알고리즘은 특히 다른 인공지능 기법들과 통합하여 많이 적용되어 왔는데, 이는 인공지능 모형구축에 있어서 최적화 필요를 만족시키는데 효과적이라는 점에서다. 유전자 알고리즘은 초기화(initialization), 선택(selection), 교배(crossover), 그리고 변이(mutation)와 같은 절차를 통해 탐색을 수행하게 된다[이건창, 1999]. 유전자 알고리즘의 구조는 <그림 2>와 같다.

초기화 단계에서는 염색체라고 불리는 개체를 미리 결정된 개체 수만큼 임의 생성하게 된다. 이렇게 생성된 개체들은 적합도 함수(fitness function)를 사용해서 그 적응도를 평가한다. 그러므로 적합도 함수의 선정은 적용 분야의 특성과 목적을 반영하는 만큼 중요한 절차라고 할 수 있다.

각각 개체의 적응도가 평가되면 교차를 수행할 개체를 선정하며, 이 개체들은 적응도가 높은 개체들로, 교차조작에 의해 자손들을 생성한다. 적응도로 표현되는 우수한 유전적 특성을 가진 개체들을 통하여 생성된 자손들의 적응도가 높을 것이라는 시각이 내포되어 있다.

```
 t ; 세대
P(t) ; t세대의 모집단

Genetic Algorithm
{
    t = 0 ;
    initializeP(t) ; // 모집단 생성 및 초기화
    evaluateP(t) ; // 평가

    while (종료조건이 만족되지 않으면)
    {
        t = t + 1
        select P(t) from P(t-1) ; // 새로운
        세대 선별
        alter P(t) ; // 유전연산(교차, 돌연변이)
        evaluate P(t) ; // 적응도 평가
    }
}
```

<그림 2> 유전자 알고리즘의 구조

교차는 두개의 염색체를 조합하는 방법이며, 정해진 crossover 위치에 따라 게놈을 잘라서 서로 바꿔 재결합한다. 교차는 확률에 의해 이루어지는데 다양한 조작 방법론이 제시되고 있다.

변이는 염색체의 일정하지 않은 위치에 가끔씩 무작위적으로 변화를 주는 것을 말한다. 생물학에서의 돌연변이에 해당하며 이를 통해 원래의 개체군이 갖지 못했던 특성을 가질 수 있게 된다.

이 때 게놈의 적합도를 구하는 함수를 적합도함수(Fitness Function)라고 하며 이 연구문제에 있어서는 통합된 결과 값의 참 값에 대한 예측적중함수(HF : Hit function)를 적합도함수로 볼 수 있다. 식 (1)은 예측적중함수 식 (2)는 적합도함수를 나타낸 것이다[김홍철, 2001].

$$HF(WS_q) = \begin{cases} 1, if\ correctly\ matched \\ 0, otherwise \end{cases} \quad (1)$$

$$Fitness(WS_q) = \frac{\sum_{i=1}^{N} HF(WS_q)}{N}$$

$$N: Totalnumberof\ trainingdata \qquad (2)$$

3.5 표본특성

실험에 사용된 표본의 주요 특성을 연령, 성별, 보증인수, 지역, 차종 등으로 학습에 사용된 3,500개 데이터를 중심으로 살펴보면 다음과 같다.

<표 3>에서 <표 7>까지의 표본에서 연령은 20대~30대가 74.7%로 가장 많은 비중을 차지하는 것을 보여주고 있다. 특히 30대는 전체의 40.6%를 차지하고 있다. 이를 볼 때 우리나라 할부금융은 주로 20~30대에 의해 가장 많이 이용되고 있음을 알 수 있다. 남성이 거의 90%가까이 차지하고 있어 할부금융 이용고객 층이 그동안 남성에 치중되어 있음을 보여준다. 할부금융을 위한 보증인수는 법인 또는 무보증 형태가 전체의 94%로 대부분 할부금융이 이루어지고 있음을 보여준다. 매입지역은 서울지역이 가장 많은 비중을 차지하였다. 이외에 대구, 대전, 원주 지역에 대한 비중이 높은 반면에 상대적으로 부산과 인천은 인구수에 대비하여 차지 비중이 낮음을 보여주고 있다. 차종은 승용차가 거의 대부분을 차지하고 있다.

〈표 3〉 연령 분포

연 령	빈 도	%	유효 %	누적비중
20대	1193	34.1	34.1	34.1
30대	1421	40.6	40.6	74.7
40대	635	18.1	18.1	92.8
50대	211	6.0	6.0	98.9
60대	40	1.1	1.1	100.0
계	3500	100.0	100.0	

〈표 4〉 성별 분포

성 별	빈 도	%	유효 %	누적비중
남성	3043	86.9	86.9	86.9
여성	457	13.1	13.1	100.0
계	3500	100.0	100.0	

〈표 5〉 보증인 분포

보 증 인 수	빈 도	%	유효 %	누적비중
법인 또는 0명	3291	94.0	94.0	94.0
1명	195	5.6	5.6	99.6
2명	11	.3	.3	99.9
3명	3	.1	.1	100.0
계	3500	100.0	100.0	

〈표 6〉 지역 분포

매입지역	빈 도	%	유효 %	누적비중
서울지점	860	24.6	24.6	24.6
부산지점	89	2.5	2.5	27.1
인천지점	160	4.6	4.6	31.7
대구지점	491	14.0	14.0	45.7
광주지점	355	10.1	10.1	55.9
수원지점	254	7.3	7.3	63.1
대전지점	365	10.4	10.4	73.5
전주지점	290	8.3	8.3	81.8
원주지점	376	10.7	10.7	92.6
마산지점	260	7.4	7.4	100.0
계	3500	100.0	100.0	

〈표 7〉 차종 분포

차 종	빈 도	%	유효 %	누적비중
승용차	3125	89.3	89.3	89.3
상용차	153	4.4	4.4	93.7
승합차	219	6.3	6.3	99.9
기타	3	.1	.1	100.0
계	3500	100.0	100.0	

연구모형에 사용된 표본의 주요 고객행동변수에 대한 기술통계량은 <표 8>과 같다. 간략

〈표 8〉 고객행동변수(flow data)에 대한 기술통계량

변 수 명	범 위	평 균	표준편차
3개월의무납입액평균/3개월잔액평균(B1)	2.14	0.12	0.14
6개월의무납입액평균/6개월잔액평균(B2)	1.29	0.12	0.13
1998년 1월 의무납입액/잔액(B3)	11.40	0.13	0.25
1998년 12월 의무납입액/잔액(B4)	315.55	0.55	5.81
1998년 1월 실납입액/의무납입액(B5)	5.72	0.49	0.48
1998년 12월 실납입액/의무납입액(B6)	9.41	0.38	0.48
3개월납입액평균/3개월의무납입액평균(B7)	4.71	0.56	0.34
6개월납입액평균/6개월의무납입액평균(B8)	3.74	0.52	0.31
12개월납입액평균/12개월의무납입액평균(B9)	1.69	0.44	0.35
1998년 12개월간 최장연체횟수(B10)	12.00	3.66	3.80
1998년 12월잔액/3개월잔액평균(B11)	1.24	0.95	0.08
1998년 12월잔액/6개월잔액평균(B12)	1.36	0.88	0.14
1998년 12개월간 연체액평균(B13)	22476376.67	759741.81	1611025.46
1998년 12개월간 연체개월수/총할부개월수(B14)	3.83	0.47	0.31
1998년 12개월간 연체개월수/12개월(B15)	2.75	1.11	0.60
매월 납부액(B16)	1224200.00	212403.15	99497.72
총할부개월수(B17)	30.00	31.02	6.57
할부가격(할부원금+할부이자)(B18)	44768280.00	6675946.43	3674530.52

히 살펴보면, 3개월의무납입액평균/3개월잔액평균과 6개월의무납입액평균/6개월잔액평균은 평균 12%에 표준편차가 각 14%, 13%를 차지하고 있다.98년 1월 의무납입액/잔액은 평균 13%, 표준편차 25%인 반면에 98년 12월 의무납입액/잔액은 평균 55%, 표준편차 581%로 상당히 편차가 심한 것을 보여주고 있다.의무납입액은 1월 49%, 12월 38%로 감소하고 있으며 편차는 동일한 모습을 보이고 있다. 연체회수는 평균 3.6회, 표준편차 3.8회로 편차가 심한 것을 보여주고 있다. 이외에 총할부개월수는 평균 31개월 정도로 나타났다.

4. 실험결과

4.1 분류모형 개발 및 통합절차

본 연구의 분석 도구로 이용된 프로그램은 Statistica-Neural Networks V.4, C5 of Clementime V. 5.0 package, Evolve V.4 그리고 SPSS-WIN V.11.0 등이다. 본 연구에서 사용된 분석모형과 실험설계에 이용된 소프트웨어를 정리하면 <표 9>와 같다.

〈표 9〉 분석모형 및 도구

분 석 모 형	분 석 도 구
다계층퍼셉트론(MLP) 반경기반함수(RBF) 다변량판별분석(MDA)	Statistica-Neural Networks V. 4
의사결정트리(DTM)	C5 of Clementime V. 5.0 package
로지스틱회귀분석(LR)	SPSSWIN V.11.0
유전자알고리즘 통합모형(NN)	Evolve V.4

<표 10>은 본 연구에서 사용된 개발도구 중 유전자알고리즘 통합도구인 Evolve V.4를 제외

하고 나머지 도구들을 사용하여 추출한 10개의 분류모형에 대한 특성과 성능을 보여주고 있는데, 대상입력변수 중에서 실제 입력변수로 채택된 변수의 개수, 은닉노드의 개수, 예측율 등을 나타내고 있다. <표 10>에서 은닉노드의 개수는 일반적으로 SAS E-Miner나 SPSS Clementime의 경우에는 수동적으로 은닉노드에 대해 제어하는 과정들이 필요하지만 본 연구에 사용된 Statistica-Neural Networks V. 4는 이 도구자체에서 제공해주는 최적 은닉노드 도출 과정을 그대로 활용하여 적용하였기 때문에 별도의 선정기준 없이 나타난 것이다. 개별분류모형들로는 반경기반함수(RBF) 모형이 2개, 다변량 판별분석(MDA) 모형이 3개, 다계층 퍼셉트론(MLP) 모형 3개, 의사결정나무(DTM) 모형 1개, 로지스틱 회귀분석(LR) 모형 1개 등 총 10개가 개발되었으며 각각의 모형에 관한 예측성능을 평가하였다.

<표 10>에 나타나는 예측율을 보면 전체적으로는 81.58%~84.53%로 비슷한 성능을 보이고 있다. 그러나 상세 예측율에 대한 결과를 살펴보면 다변량 판별분석(MDA)모형들이 불량채권과 우량채권에 대한 예측율이 가장 높다. 반면에 MLP모형과 DTM은 미정에 대한 예측율이 매우 높으며, RBF모형은 우량에 대한 예측율이 높다. 특히 DTM은 미정채권에 대한 대단히 높은 예측율을 보여주고 있다. 이러한 평가결과는 각각의 개별모형들이 각기 다른 특성을 가지고 있음을 알려주며 통합의 필요성을 말해주는 지표들이라고 할 수 있다.

<표 11>은 예측모형별로 선택된 입력변수의 내용을 자세하게 보여주고 있다. 여기서 MDA, MLP, RBF모형의 선택된 입력변수는 본 연구에서 사용된 Statistica-Neural Networks V. 4에서 자동적으로 선택되어진 것이고, 로지스틱 회귀분석(LR)모형은 독립변수와 종속변수가 인공신경망 모델과 같기 때문에 MLP, RBF모형에 선택되어진 변수를 전부 선택하였다. 그리고 DTM 모형은 초기에 입력변수를 전부 투입한 후 목표경로를 찾아가면서 우선 추출된 부분을 선택하였다.

4.2 분류모형의 1차 통합

1차 통합과정에 개별모형들을 3개의 대표모형 MDA*, MLP*, RBF*로 통합하기 위해 가중치 행렬의 최적화를 시도하였다. 이를 위하여 유전자 알고리즘 구현도구인 EVOLVE 4.0 for Excel-Industrial을 사용하였다.

〈표 10〉 모형별 특성 및 성능

모형번호	모 형	입력변수개수	은닉노드 개수	예측율 (performance)	상세예측율(%)		
					불 량	미 정	우 량
1	MDA-1	21	-	81.92%	81.02%	2.02%	97.37%
2	MDA-2	25	-	82.27%	82.87%	3.03%	94.01%
3	MDA-3	24	-	81.81%	81.48%	6.06%	95.04%
4	MLP-1	4	6	83.75%	78.24%	13.13%	93.72%
5	MLP-2	5	4	84.44%	78.70%	22.22%	89.34%
6	MLP-3	8	7	84.53%	77.31%	28.28%	88.76%
7	RBF-1	9	19	81.58%	77.77%	2.02%	97.66%
8	RBF-2	9	20	81.92%	77.77%	5.05%	98.24%
9	DTM	21	-	82.20%	80.56%	53.54%	86.86%
10	LR	8	-	81.58%	78.57%	2.31%	92.54%

〈표 11〉 모형별 선택입력 변수

변수명	설 명	모형번호	1	2	3	4	5	6	7	8	9	10
		모형형태	MDA1	MDA2	MDA3	MLP1	MLP2	MLP3	RBF1	RBF2	**DTM**	LR
A1	나이										○	
A2	성별		○	○	○						○	
A3	보증인수										○	
A4	매입지역						○	○	○	○	○	○
A5	차량원부			○	○							
A6	차종			○			○	○			○	○
A7	차량년식		○	○	○		○	○			○	○
A8	배기량		○	○	○						○	
A9	신용조사방법		○	○	○							
A10	구매자구분											
B1	3개월의무납입액평균/3개월잔액평균		○	○	○							
B2	6개월의무납입액평균/6개월잔액평균		○	○	○						○	
B3	1998년 1월 의무납입액/잔액			○	○						○	
B4	1998년 12월 의무납입액/잔액		○	○	○						○	
B5	1998년 1월 실납입액/의무납입액		○	○	○						○	
B6	1998년 12월 실납입액/의무납입액		○	○	○	○	○	○	○	○		○
B7	3개월납입액평균/3개월의무납입액평균		○	○	○						○	
B8	6개월납입액평균/6개월의무납입액평균		○	○	○						○	
B9	12개월납입액평균/12개월의무납입액평균 (1998년1월 납입액을 이전6개월 평균으로)		○	○		○	○	○	○	○	○	○
B10	1998년 12개월간 최장연체횟수		○	○	○	○	○	○	○	○	○	○
B11	1998년 12월잔액/3개월잔액평균		○	○	○				○	○	○	○
B12	1998년 12월잔액/6개월잔액평균		○	○	○			○	○	○		○
B13	1998년 12개월간 연체액평균		○	○	○						○	
B14	1998년 12개월간 연체개월수/총할부개월수		○	○	○				○	○	○	○
B15	1998년 12개월간 연체개월수/12개월			○	○	○			○	○	○	○
B16	매월 납부액		○	○	○						○	
B17	총할부개월수		○	○	○			○	○	○		○
B18	할부가격(할부원금+할부이자)		○	○	○						○	
B19	우·불량판별 (0 : 불량, 1 : 우량, 2 : 미정)											

〈표 12〉 1차 통합모형별 최적가중치행렬(W)

1차 통합모형	가중치행렬(W)	
MDA*	W =	$\begin{bmatrix} 0.1 & 0.69 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
MLP*	W =	$\begin{bmatrix} 0.1 & 0.45 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
RBF*	W =	$\begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$

유전자 알고리즘의 통합과정에서 도출된 가중치 행렬은 <표 12>와 같다. 여기에 나타난 가중치행렬에서 각 행(row)은 통합되어지는 개별모형에 대한 가중치이며 열(column)은 불량, 미정, 우량의 분류 결과 값에 대한 가중치를 의미한다. MDA*와 MLP* 모두 1번째 모형의 2(미정)값에 가중치를 많이 두고 있고, RBF모형은 가중치가 일정한데 RBF개별 모형들의 예측성능 또한 비슷하게 나타났음을 볼 수 있다. 개별모델의 성능이 높은 쪽에 가중치를 많이 둔다는 것은 확률적으로 예측성능이 우수해질 가능성이 많음을 의미한다. 그렇지만 <표 13>의 모형별 예측율을 보면 MDA와 MLP의 첫 번째 모형의 2(미정)값이 다른 모형에 비해 예측율이 낮기 때문에 가중치가 집중되는 것으로 판단된다. 즉 일반적으로 성능이 높은 개별모형 쪽에 가중치가 집중되기도 하지만, 예측율 성능이 비슷할 경우에는 상세 예측율에서 가장 예측율이 떨어지는 부분에 가중치가 집중된다는 것을 알 수 있다.

4.3 각 단일모형의 최종 통합

1차 통합에 의해 얻어진 MDA*, MLP*, RBF*와 기존 단일 모형인 DTM과 LR 다섯 가지 모형들을 1차 통합 때와 같은 방법으로 최종통합을 실시하였다. 이때에도 유전자알고리즘 구현 도구는 EVOLVE 4.0 for Excel - Industrial 을 사용하였다.

<표 14>는 최종 통합모형의 가중치 행렬을 보여주고 있는데 가중치 행렬의 원소값을 상세히 살펴보면 첫 번째 모형(MDA)의 1(불량)값과 세 번째 모형(RBF) 2(미정)값에 많은 가중치를 두고 있음을 볼 수 있다.

〈표 14〉 최종통합모형의 가중치행렬(W)

통합모형	가중치행렬(W)	
NN*	W=	$\begin{bmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$

불량값과 미정에 가중치가 부여된 최종통합모형을 구체적으로 살펴보면 불량값의 경우 예측값이 높은 MDA모형에 0.5 가 부여되었다. 그리고 미정값의 경우 RBF모형에 0.3이 부여되었다.

<표 14>의 가중치를 사용하여 최종 통합모형의 예측성능을 비교하면 <표 15>와 같다. <표 15>의 전체최종통합모형인 NN*의 예측율이

〈표 13〉 모형별 예측성능 비교 - 1차통합(최초 가중치 부여후)

모형 번호	모형 형태	1차 통합	1차통합 상세 예측율		
			(불량)	(미정)	(우량)
1	MDA	83.80	81.48	2.02	96.35
2	MLP	82.40	78.24	23.23	89.64
3	RBF	84.20	77.31	6.06	97.66
3	DTM	82.20	80.56	53.54	86.86
4	LR	81.58	78.57	2.31	92.54

〈표 15〉 모형별 예측성능 비교 - 최종통합(최적 가중치 부여후)

모형 번호	모형 형태	전체 최종 통합	상세최종통합		
			(불량)	(미정)	(우량)
1	MDA	84.70	80.56	6.06	97.37
2	MLP				
3	RBF				
3	DTM				
4	LR				

84.70%로 1차 통합모형 중에서 예측성능이 가장 우수하게 나타난 RBF*의 예측율보다 높게 나타났다.

상세 예측의 경우에도 불량에 대한 예측은 80.56%로 가장 높은 예측율을 보이고 있는 MDA와 비슷한 값을 취하고 있으며 DTM과는 같다. 우량에 대한 예측은 97.37%로 매우 높은 예측율을 나타내고 있다. 한편 미정에 대한 예측율은 어느 정도 향상은 가져왔으나 여전히 6.06%에 머물고 있다. 전반적으로 최종통합모형을 요약 평가하면, 전체 예측율에 있어서 통합모형은 다른 개별모형과 비교하여 가장 우수한 예측율을 보이고 있다. 특히 상세최종통합모형에서는 우량에 대한 예측율은 매우 월등함을 보이고 있는 반면에, 미정에 대한 예측은 낮게 나타났다.

최종통합모형의 예측율이 개별모형과 비교하여 대폭 향상되지는 않았지만 <표 15>에 나타난 상세 예측율의 수치와 개별모형들의 상세 예측율을 비교해서 통합의 효과를 단계적으로 살펴보면 첫 번째 통합에서는 평균적으로 개별분류모형들의 각기 다른 특성들이 서로 절충되고 보완된 1차 통합모형이 얻어졌으며, 최종통합에서는 성능이 가장 우수한 MDA*와 RBF*모형을 선택하는 효과를 거둔 것으로 판단된다.

모형통합에 관한 연구결과를 전체적으로 종합해보면 초기에 얻어진 10개의 개별모형을 같은 형태별로 1차 통합한 모형 3개는 각 종류별로 통합 대상인 개별모형을 적절히 혼합한 특성을 나타냈으며 전반적인 성능 또한 향상되었다. 3개의 1차 통합모형을 다시 통합하여 얻은 최종통합 모형은 1차통합모형 중에서도 가장 성능이 우수한 모형을 선택한 것으로 보여지며 본 연구에서 실시한 2차에 걸친 단계적 모형통합에 의해서 각 개별모형들의 특성이 결합된, 보다 우수한 통합모형이 발견되었음을 알 수 있다.

5. 결 론

본 연구에서는 할부금융시장에서의 고객정보 및 할부진행과정에 대한 세부 내역을 바탕으로 각기 다른 기법들로 구현된 복수개의 분류모형(classifier)들을 유전자 알고리즘을 이용하여 하나의 모형으로 통합하는 방법을 통해 얻어진 신용평가모형을 제안하였다.

실험결과를 통해 여러 가지 분류모형들을 개별모형에 비하여 우수한 성능의 최종통합모형을 얻을 수 있었다. 예측성능의 수치를 볼 때 통합에 의해 성능이 대폭 향상된 최적모형을 구하려는 애초의 기대에는 못 미치는 듯 하지만 모형의 개발에 있어서 최적화의 어려움을 감안한다면 개별모형 이상의 성능을 가지며, 개별모형의 서로 다른 특성이 결합되어진 통합모형을 얻을 수 있었다는 점에서 연구의 의의를 찾을 수 있다.

또한 학습기반 모형의 개발에 있어서 데이터의 수가 매우 중요한 영향을 미친다는 것을 감안할 때 신용평가와 관련된 기존의 연구들이 수십 개 또는 수백 개 단위의 실험용 데이터를 사

용했던 것에 반하여 실세계에서 구해진 수십만 개의 데이터로부터 순화과정을 거친 데이터를 한 단계의 실험당 수천개 단위로 사용하였다는 점은본 연구의 결과에 대하여 보다 의미 있는 통찰과, 실무적으로 적용할 수 있는 가능성을 제공한다.

본 연구의 한계점을 들자면 첫째, 입력변수의 선정에 있어서 고객의 인구통계학적인 정보들이 대부분 탈락되어버린 이유로 본 연구에서 개발된 모형을 거래내역이 없는 신규고객의 신용평가에 적용하기 어렵다는 점이다. 둘째, 연구에서 신용평가의 단위가 하나의 채권(할부계약)별로 이루어졌으므로 복수의 채권과 관련된 고객의 신용평가가 채권별로 상이한 결과로 나타났을 때 이에 대한 해답을 제시할 수 없다는 점 또한 연구의 한계라고 보여진다.

향후의 연구과제로는 앞에서 지적한 연구의 한계점을 극복하기 위하여 고객의 인구통계학적인 정보들이 충실하게 포함된 자료를 통하여 복수개의 채권과 관련된 고객의 신용평가모형에 대한 연구가 더해져야 한다고 보여진다.

참 고 문 헌

[1] 김홍철, "유전자 알고리즘기반 복수 분류모형 통합에 의한 할부금융고객의 신용예측모형", *대구대학교 대학원 석사학위 논문*, 2001년.

[2] 이건창, "사례기반추론과 유전자 알고리즘을 결합한 지식경영 방법론에 관한 연구 : 신용평가를 중심으로", *정보기술응용연구 창간호*, 1999년 2월.

[3] 정충영, 최이규, *SPSSWIN을 이용한 통계분석*, 서울, 무역경영사, 1998년.

[4] 채서일, *사회과학 조사방법론*, 2판, 서울, 학현사, 1999년.

[5] 최종후, 한상태, *AnsweerTree를 이용한 데이터마이닝 의사결정나무분석*, 서울, SPSS 아카데미, 2000년.

[6] Altman, E.I., G. Marco, and F. Varetto, "Corporate Distress Diagnosis : Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience)", *Journal of Banking and Finance*, 18, 1994, pp. 505-520.

[7] Berry, J. and G. Linoff, *Data Mining Techniques : For Marketing, Sales, and Customer Support*, John Wiley and Sons, NY, 1997.

[8] Carter, C. and J. Catlett, "Assessing Credit Card Applications Using Machine Learning", *IEEE Expert*, 2, 1987, pp. 71-79.

[9] Cheng, B. and D.M. Titterington, "Neural Networks : A Review from a Statistical Perspective", *Statistical Science*, 9, 1994, pp. 2-30.

[10] Cho, Sung-Bae, "Pattern Recognition with Neural Networks Combined by Genetic Algorithm", *Fuzzy Sets and Systems*, 103, 1999, pp. 339-347.

[11] Desai, V.S., D.G. Convay, J.N. Crook, and G.A. Overstreet, "Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms", *IMA Journal of Mathematics Applied in Business and Industry*, 8, 1997, pp. 323-346.

[12] Feelders, A.J., "Credit Scoring and Reject Inference with Mixture Models", *International Journal of Intelligent Systems in Accounting Finance & Management*, 8, 1999, pp. 271-279.

[13] Grablowsky, B.J. and W.K. Talley, "Probit and Discriminant Functions for Clas-

sifying Credit Applicants : A Comparison", *Journal of Economics and Business*, 33, 1981, pp. 254-261.

[14] Gupta, Y.P., M.C. Gupta, A.K. Kumar, and C. Sundram, "Minimizing Total Intercell and Intracell Moves in Cellular Manufacturing : A Genetic Algorithm Approach", *INT. J. of Computer Integrated Manufacturing*, 8(2), 1995, pp. 92-101.

[15] Hansen, J.V., "Combining Predictors : Comparison of Five Meta Machine Learning Methods", *Information Sciences*, 119, 1999, pp. 91-105.

[16] Hon, K.K.B. and H. Chi, "A New Approach of Group Technology Part Families Optimization", *Annals of the CIRP*, 43(1), 1994.

[17] Imielinski, T. and H. Mannila, "A Database Perspective on Knowledge Discovery", *Communications of the ACM*, 39(11), 1996, pp. 214-225.

[18] Jain, Bharat A., and N. Nag, Barin, "Performance Evaluation of Neural Network Decision Models", *Journal of Management Information Systems*, 14(2), Fall, 1997, pp. 201-216.

[19] Kim, E., W. Kim, and Y. Lee, "Combining of Multiple Classifiers for the Customer's Purchase Behavior Prediction", Decision Support System, 2002, pp. 167-175.

[20] Kim, Y., and W.N. Street, "An Intelligent System for Customer Targeting : A Data Mining Approach", *Decision Support Systems, forthcoming*, 2003, pp. 1-14.

[21] Mehta, D., "The Formulation of Credit Policy Models", *Management Science*, 15, 1968, pp. 30-50.

[22] Thomas, L.C., "A Survey of Credit and Behavioral Scoring : Forecasting Financial Risk of Lending to Consumers", *International Journal of Forecasting*, 16, 2000, pp. 149-172.

[23] West, D., "Neural Network Credit Scoring Models", *Computers & Operations Research*, 27, 2000, pp. 1131-1152.

[24] Wiginton, J.C., "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behaviour", *Journal of Financial and Quantitative Analysis*, 15, 1980, pp. 757-770.

■ 저자소개

김 갑 식

계명대학교 일본학과와 경일대학교 전자계산학과를 졸업하고, 계명대학교 대학원 경영정보학과에서 경영학석사, 대구가톨릭대학교 대학원 경영학과에서 생산 및 경영정보학전공으로 경영학박사를 취득하고 1993년부터 현재까지 대구산업정보대학 컴퓨터정보계열에 재직중이다. 경상북도 21세기 발전위원회 과학·정보 분과위원과 중소기업청 및 중소기업정보화 경영원의 생산공정 정보화 평가위원 등 다수의 공공기관의 자문위원으로 활동중이며, 한국정보시스템학회와 한국산업정보학회, 한국산업경제학회에서 상임이사로도 활동중이다. 주요 관심분야는 데이터 마이닝, 데이터웨어 하우징, CRM, 중소기업 정보화(ERP, POP, SPC, SCM 등), e-비즈니스, 전자상거래 등이다.

◈ 이 논문은 2003년 10월 4일 접수하여 2차 수정을 거쳐 2004년 3월 7일 게재확정되었습니다.