신용평가를 위한 데이터마이닝 분류모형의 통합모형에 관한 연구

김 갑 식†

요 약

본 연구는 금융기관에서의 고객신용평가를 위한 최적의 데이터마이닝 모형을 제안한다. 이를 위해 할부금융시장에서의 고객정보 및 할부진행 과정에 대한 세부 내역을 바탕으로 다계층 퍼셉트론(Multi-Layered Perceptrons:MLP)과 다변량 판별분석(Multivariate Discrimination Analysis : MDA), 그리고 의사결정나무(Decision Tree)를 적용하여 각각의 개별모형을 도출하고 이를 유전자 알고리즘을 이용하여 통합한 최종 모형을 구해 그 결과를 각 단일모형과 비교·분석하였다. 그 결과 유전자 알고리즘을 통해 결합한 통합모형의 성능이 가장 우수한 것으로 나타났다. 이에 본 연구는 기존에 진행되었던 개별모형에 대한 검증은 물론, 단순히 여러 개의 모형을 비교·분석하여 우월한 모형을 평가하는 기존 방법론상의 한계를 극복하기 위해 각각의 개별모형을 유전자알고리즘을 통해 통합모형으로 구축하는 하나의 방법론을 제시하였다는데 그 의의가 있다.

A Study of the Integration of Individual Classification Model in Data Mining for the Credit Evaluation

Kap Sik Kim†

ABSTRACT

This study presents an integrated data mining model for the credit evaluation of the customers of a capital company. Based on customer information and financing processes in capital market, we derived individual models from multi-layered perceptrons(MLP), multivariate discrimination analysis(MDA), and decision tree. Further, the results from the existing models were compared with the results from the integrated model using genetic algorithm. The integrated model presented by this study turned out to be superior to the existing models. This study contributes not only to verifying the existing individual models but also to overcoming the limitations of the existing approaches.

키워드 : 데이터마이닝(Data Mining), 신용평가(Credit Evaluation), 유전자 알고리즘(Genetic Algorithm), 다계층 퍼셉트론(Multi-layered Perceptrons), 통합모형(Integrated Model), 다변량 판별분석(Multivariate Discrimination Analysis)

1. 서 론

오늘날 금융기관에서는 우·불량고객의 판별 및 고객의 신용등급 관리를 위해 신용평가(credit scoring)를 매우 중요시한다. 신용평가는 불량채권 발생률을 미연에 감소시키고 고객에 따라 차별화된 금융상품과 혜택을 제공함으로써 고객관계관리(customer relationship management)를 실현시켜 궁극적으로 기업의 수익을 증대시켜주기 때문에 수많은 금융기관 및 금융관련 기업들이 신용평가 예측율 향상을 위해 다각적으로 대안을 모색하고 있다.

신용평가에 대한 기존 접근방법은 일반적으로 협의의 신용평가와 행태평가(behavior scoring)로 대별된다. 전자는 신규고객이 융자를 처음 신청할 때 그 고객이 제시하는 인구통계학적 자료만을 가지고 재정적인 위험을 판단하는 접근방법이며, 후자는 협의의 신용평가에서 활용하는 인구통계학적 자료 이외에 기존 고객의 거래 내역에 의해 그 고객의 현재 상태를 평가하는 접근방법이다. 이 두 가지 접근방법 모두 같은 방식으로 측정할 수 있으나 입력되는 자료에 있어 후자의 경우에는 전자에서 사용된 자료 이외에 거래내역이 포함된다는 점에서 차이가 있다[21].

그러나 두 가지 신용평가 기법에 대한 접근 방식이 크게 차이가 없음에도 불구하고 여러 가지 정보 기술상의 문제로 인해 대부분의 학술적인 연구는 협의의 신용평가에 집중되어 왔다. 그러나 최근 들어 컴퓨터의 강력한 가공·처리능력을 이용하여 다량의 데이터로 새로운 패턴을 찾아낼 수 있는 데이터마이닝 기법이 발달됨에 따라 행태평가에 대한

† 정 회 원 : 대구산업정보대학 인터넷비즈니스과 교수
논문접수 : 2004년 2월 28일, 심사완료 : 2004년 11월 29일

학술적 관심이 높아지게 되었다.

데이터마이닝 분석을 위한 도구에는 인공 신경망(artificial neural network) 모형, 의사결정 나무(decision tree) 모형, 통계학적 모형, 경영 과학적 모형, 유전자 알고리즘(genetic algorithm) 모형 등이 있다. 이러한 각각의 단일모형들은 구현기법에 따라 각기 다른 고유의 특성을 가지고 있으며, 연구 상황에 따라 그 성능이 다르게 나타나기 때문에 어떤 모형이 우수하다고 단정할 수는 없다. 또한, 한 가지 모형만 연구문제에 맞게 최적화하는 과정은 많은 시간과 노력이 요구되기 때문에 각 단일모형의 장점들만을 취하여 최적의 통합신용평가모형으로 연구문제에 맞게 최적화하는 것이 보다 효율적일 것으로 생각된다[1].

따라서 본 연구는 다계층 퍼셉트론(Multi-Layered Perceptrons : MLP) 구조를 갖는 인공신경망 모형과 다변량 판별분석(Multivariate Discrimination Analysis : MDA)모형 그리고 의사결정나무(Decision Tree) 모형 등을 이용하여 각각의 단일모형을 얻어 신용평가의 예측결과를 비교·분석한 후, 유전자 알고리즘 방식에 의해 이들 단일모형에 대한 통합모형을 구축함으로써 할부금융 이용고객의 행태평가에 의한 신용평가 예측의 최적화를 입증하려는데 그 목적을 두고 있다.

2. 데이터마이닝모형의 특성

전통적으로 신용평가를 위한 연구에 쓰여진 모형으로는 다변량 판별분석(multivariate discrimination analysis)이나 로지스틱 회귀분석(logistic regression analysis), 프로빗 분석과 같은 통계학적 모형[10, 23]과, 선형계획법(linear programming : LP)[18]과 같은 경영과학적 모형을 들 수 있다. 최근 들어 의사결정나무(decision tree), 인공 신경망 등의 인공지능(artificial intelligence) 모형을 이용한 연구가 활발하게 진행되었다. 특히 기존 연구에서 인공 신경망 모형을 이용한 연구가 좋은 결과를 보여주고 있다[16, 22, 24].

인공 신경망 모형은 데이터마이닝에 대한 관심이 높아지면서 최근 가장 많이 언급되고 있다. 이 모형은 다양한 응용분야를 가지고 있는 많은 문제들에 대해 널리 적용될 수 있다. 또한 통계적 가정이 필요 없으면서도 비선형적인 회귀모형을 설명하기에 적당하며, 신용평가에 매우 적합하다고 증명되었다[7].

판별분석 모형은 관찰된 자료의 어떤 특성을 바탕으로 관찰 값을 두 개 이상의 그룹에 각각 구분되도록 도와주는 통계적 모형이다. 사회현상의 여러 특성들을 토대로 하여 주어진 상황에서 응답자들이 어떻게 행동할 것인지를 예측하는 하나의 통계모형이다[3]. 데이터마이닝을 위한 다변량 판별분석모형의 경우 구현이 간단하고 학습시간도 짧지만 독립변수들이 다변량 판별분석의 기본적인 통계학적 가정들을 만족해야 하므로 이에 대한 검증이 필요하다는 한계점을 가지고 있다[4].

의사결정나무 모형은 데이터마이닝의 분류작업에 주로 사용되는 모형으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다. 분류를 목적으로 하는 다른 방법들 즉, 인공 신경망 모형, 판별분석 모형, 회귀분석 모형 등에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다[5]. 이모형은 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단할 뿐만 아니라 이해하기 쉬운 규칙으로 전환될 수 있다[15]. 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다[19].

유전자 알고리즘모형은 인공지능의 한 모형으로서 2차원 이상의 복잡한 탐색공간에서 전 범위의 최적해(global optimal solution)를 탐색하는데 아주 효율적이며, 유연하다[11]. 이러한 유전자 알고리즘모형은 생태계의 자연선택(natural selection)과 적자생존(survival of the fittest)에 근거를 두고 있다. 새로운 집단(new population)을 형성할 때에 과거의 집단(old population)에서 높은 적합도를 가지는 개체(string)가 높은 확률을 가지고 새로운 집단으로 유전한다는 것이 그 기본적인 원리이다[14]. 이러한 유전자 알고리즘은 Holland [13]가 그 이론적인 근거를 마련했으며, Goldberg[9]에 의해 공학 분야에서 가스 송수관문제에 대한 최적 설계가 최초로 시도된 이래 많은 발전이 되어오고 있다.

이상에서 논한 모형들을 분석해 볼 때 어떤 방법이 최선의 방법인지를 결정하는 것은 무척 어려운 일이다. 이러한 이유는 신용평가기관의 보고와 같은 가장 의미 있는 자료들의 대부분이 너무 민감하거나 비싼 이유로 학자들의 비교연구는 종종 한계를 가지기도 하지만 그들의 연구결과는 나름대로 성격을 갖기 때문이다. 그러한 이유로 연구문제에 대한 최적모형을 찾기 위하여 여러 가지 신용평가 모형들에 대한 통합의 필요성과 방법론들이 제안되고 있다[17].

3. 연구의 설계

3.1 연구방법 및 모형

신경망 모형 및 기타 모형들은 각기 다른 특성을 가지며 연구상황에 따라 서로 다른 성능을 보이므로 어떤 모형이 우수하다고 단정할 수 없으며 연구문제에 대한 최적모형을 얻기 위한 통합의 필요성이 제기된다.

본 연구에서는 유전자 알고리즘[9]을 이용하여 복수 분류모형 통합모듈의 가중치행렬을 최적화하는 방식으로 개별모형들을 병렬식으로 가중통합을 하였다[17].

이를 위해서 가령 종속변수가 취할 수 있는 값이 N개 즉, 분류해야할 집단의 개수가 N개 이고, 통합해야 할 분류모형이 K개 있다고 가정할 때 모형의 결과 값 O_i는 식 (1)과 같이 정의될 수 있다. 그리고 식 (2)에서 보는 바와 같이 한 패턴이 취할 수 있는 값 E(x)는 O_i 가운데 최대값을 골라 그 값이 일정값(a)을 넘어갈 경우에 그 값으로 하고 그렇지 않을 경우에는 값을 주지 않는다. 이 때 E(x)의 값이 원래

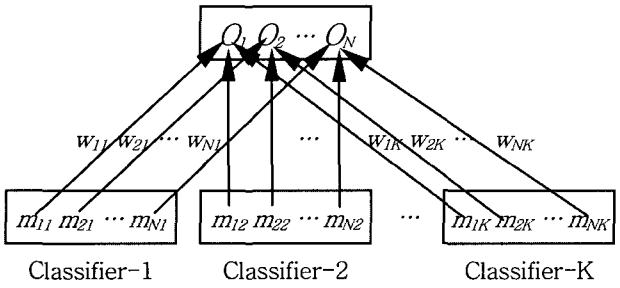
의 값과 같을 경우에는 식 (3)에서 보는 바와 같이 유전자 알고리즘의 Fitness Function에 1의 값을 주고 그렇지 않을 경우에는 0의 값을 주었다. 이와 같은 방식으로 <그림 1>에서 보는 바와 같이 복수개의 분류모형을 결합할 수 있는 가중벡터인 W를 구한다[2].

$$O_i = \sum_{j=1}^{K} w_{ij} m_{ij}$$

$$\begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_N \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1K} \\ m_{21} & m_{22} & \cdots & m_{2K} \\ \vdots & \vdots & & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{NK} \end{bmatrix} \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{N1} \\ w_{12} & w_{22} & \cdots & w_{N2} \\ \vdots & \vdots & & \vdots \\ w_{1K} & w_{2K} & \cdots & w_{NK} \end{bmatrix} \quad (1)$$

$$E(x) = \begin{cases} S, \ if \ o_s = \max_{i \in \Lambda}(o_i) \ and \ o_s \geq \alpha \\ reject, \quad otherwise \end{cases} \quad (2)$$

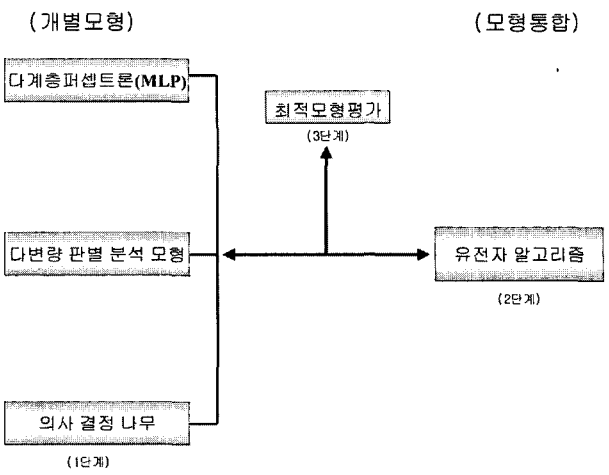
$$HF(WS_q) = \begin{cases} 1 \ , \ if \ correctly \ matched \\ 0 \ , \ otherwise \end{cases} \quad (3)$$



(그림 1) 복수 분류모형 통합모듈의 구조

본 연구에서는 할부금융시장에서의 고객정보 및 할부진행과정에 대한 세부 내역을 바탕으로 여러 가지 개별분류모형들을 도출하고 이를 유전자 알고리즘(genetic algorithm)을 이용하여 통합하여 최종모형을 구했다. 그리고 이 통합모형과 각 단일모형을 비교·분석해서 최적의 신용평가모형을 제안한다. <그림 2>는 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 세 가지 단위 분류모형으로는 인공 신경망의 다계층 퍼셉트론(MLP)모형, 다변량 판별분석(MDA)모형, 의사결정나무 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 유전자 알고리즘에 의한 가중치 최적화를 통하여 최종결과 값으로 가중 통합하는 것이다. 이렇게 통합된 결과 값을 세 개의 각 단위모형과 서로 비교해 본다.



(그림 2) 단위 분류모형의 통합방법

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 개별분류모형개발에 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.

〈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	

3.3 변 수

<표 2>에 나타난 항목들은 원시데이터를 예측모형의 입력변수로 사용하기 위해 정규화 등의 과정을 거쳐 적절히

〈표 2〉 변수 상세 설명

변수명	설 명
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	98년1월 의무납입액/잔액
B4	98년12월 의무납입액/잔액
B5	98년1월 실납입액/의무납입액
B6	98년12월 실납입액/의무납입액
B7	3개월납입액평균/3개월의무납입액평균
B8	6개월납입액평균/6개월의무납입액평균
B9	12개월납입액평균/12개월의무납입액평균 (98년1월 납입액을 이전6개월 평균으로)
B10	98년 12개월간 최장연체횟수
B11	98년 12월잔액/3개월잔액평균
B12	98년 12월잔액/6개월잔액평균
B13	98년 12개월간 연체액평균
B14	98년 12개월간 연체개월수/총할부개월수
B15	98년 12개월간 연체개월수/12개월
B16	매월 납부액
B17	총할부개월수
B18	할부가격(할부원금+할부이자)
B19	우·불량판별 (1:불량, 2:미정, 3:우량)

(아래의 변수 설명 중에서 3개월은 1997. 11~1998. 1을 6개월은 1997. 8 ~1998. 1을 말한다.)

가공한 변수목록이다.

변수 B19는 채권의 우·불량을 판별하는 종속변수로서 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>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 반영시켰으며 각 입력변수들의 값이 개월 수, 금액 등으로 그 스케일이 현저하게 차이가 나기 때문에 이를 1에서 0 사이의 실수 값으로 만들어 주기 위해 변수 값들을 해당하는 변수의 평균값으로 나누어주는 방법을 통해 정규화 하였다. 즉 모델에서 스케일 변수에 대한 조건으로 데이터의 범위가 0.0~1.0 또는 -1.0~1.0이 되어야 한다. 이러한 조건에 부합시키기 위해서 모형에 사용된 입력변수 중 스케일변수를 개월수 또는 평균금액으로 나누어주는 과정이 필요하다. 이를 통해서 금액과 개월 수 등의 스케일이 다른 변수 값을 0에서 1사이의 값으로 정규화 하였다.

대부분의 변수들에서 개월 수 또는 평균금액으로 나누어주는 이유는 금액과 개월 수 등의 변수 값이 스케일이 다르므로 0에서 1사이의 값으로 정규화 시켜주기 위함이다.

4. 연구모형의 실험결과 및 평가

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 등이다. 본 연구에서 사용된 분석모형과 실험설계에 이용된 소프트웨어를 정리하면 <표 3>과 같다.

〈표 3〉 분석모형 및 도구

분석모형	분석도구
다계층 퍼셉트론(MLP) 다변량 판별분석(MDA)	Statistica-Neural Networks V. 4
의사결정나무(DTM)	C5 of Clementime V. 5.0 package
유전자 알고리즘 통합모형(NN)	Evolve V.4

<표 4>는 본 연구에서 사용된 개발도구 중 유전자 알고리즘 통합도구인 Evolve V.4를 제외하고 나머지 도구들을 사용하여 추출한 7개의 분류모형에 대한 특성과 성능을 보여주고 있는데, 대상입력변수 중에서 실제 입력변수로 채택된 변수의 개수, 은닉노드의 개수, 예측율 등을 나타내고 있다. <표 4>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 제공

〈표 4〉 모형별 특성 및 성능

모형번호	모형	입력변수개수	은닉노드 개수	예측율 (performance)	상세예측율(%)		
					불량	미정	우량
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	DTM	21	-	82.20%	80.56%	53.54%	86.86%

해주는 최적 은닉노드 도출과정을 그대로 활용하여 적용하였기 때문에 별도의 선정기준 없이 나타난 것이다. 개별분류모형들로는 다변량 판별분석(MDA) 모형이 3개, 다계층 퍼셉트론(MLP) 모형 3개, 의사결정나무(DTM) 모형 1개 등 총 7개가 개발되었으며 각각의 모형에 관한 예측성능을 평가하였다.

<표 4>에 나타나는 예측율을 보면 전체적으로는 81.81%~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 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.

4.2 분류모형의 1차 통합

1차 통합과정에 개별모형들을 2개의 대표모형 MDA*, MLP*로 통합하기 위해 가중치 행렬의 최적화를 시도하였다. 이를 위하여 유전자 알고리즘 구현도구인 EVOLVE 4.0 for Excel-Industrial 을 사용하였다.

유전자 알고리즘의 통합과정에서 도출된 가중치 행렬은 <표 5>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 있음을 볼 수 있다. 개별모형의 성능이 높은 쪽에 가중치를 많이 둔다는 것은 확률적으로 예측성능이 우수해질 가능성이 많음을 의미한다. 그렇지만 <표 6>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.

〈표 5〉 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 =	$\begin{bmatrix} 0.1 & 0.45 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$

〈표 6〉 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	DTM	82.20	80.56	53.54	86.86

4.3 각 단일모형의 최종통합

1차 통합에 의해 얻어진 MDA*, MLP*와 기존 단일 모형인 DTM 등 세 가지 모형들을 1차 통합 때와 같은 방법으로 최종통합을 실시하였다. 이때에도 유전자 알고리즘 구현도구는 EVOLVE 4.0 for Excel - Industrial 을 사용하였다.

<표 7>은 최종통합 모형의 가중치 행렬을 보여주고 있는데, 여기에서 가중치 행렬의 원소값을 상세히 살펴보면 첫 번째 모형(MDA)과 세 번째 모형(DTM)의 1(불량)값에 많은 가중치를 두고 있음을 볼 수 있다. 최종통합 모형에서는

모형의 예측율을 높여주기 위해서 개별 모형 중에서 예측부분이 가장 미진한 부분에 해당하는 값에 가중치를 부여해주고 있다. 우량값과 미정에 가중치가 부여된 최종통합 모형을 구체적으로 살펴보면 우량값의 경우 예측값이 가장 낮은 MDA모형에 0.87이 부여되었다. 미정값의 경우에도 MDA에 0.65의 가중치를 부여하였다.

〈표 7〉 최종통합 모형의 가중치행렬 (W)

통합모형	가중치행렬(W)	
NN*	W=	$\begin{bmatrix} 0.57 & 0.65 & 0.87 \\ 0.1 & 0.1 & 0.1 \\ 0.09 & 0.1 & 0.1 \end{bmatrix}$

<표 7>의 가중치를 사용하여 최종통합모형의 예측성능을 비교하면 <표 8>과 같다. <표 8>의 전체최종통합모형인 NN*의 예측율이 84.40%로 1차 통합모형 중에서 예측성능이 가장 우수하게 나타난 MDA*의 예측율보다 높게 나타났다.

<표 8>의 전체최종통합 모형인 NN*의 예측율이 84.40%로 1차 통합모형 중에서 예측성능이 가장 우수하게 나타난 MDA*의 예측율보다 높게 나타났다.

최종 통합과정에서는 MDA의 미정값과 DTM의 우량값에 가중치가 부여되었다. 이러한 가중부여를 통한 최종통합 모형의 신용평가 예측성능은 84.40%로 1차 통합에서 나타난 개별모형들의 예측성능보다 높게 나타났다. 이 통합모형 예측결과를 각 집단별로 살펴보면 다음과 같다. 우선, 우량집단의 예측율은 97.37%로서 1차통합의 예측율중 가장 높았던 MDA의 예측율 96.35%에 비해 약 1%의 예측율 증가를 보여주었다. 미정에 대한 예측은 2.02%로 상대적으로 낮게 나타났으나, 실무적인 할부금융회사의 고객신용관리측면에서는 이 값이 낮아지는 방향으로 전개되어야 하므로 큰 문제는 없는 것 같다. 다만, 불량에 대한 예측이 81.02%로 개별모형의 최고 예측치인 MDA의 불량예측 81.48%보다 낮게 나타나 불량집단에 대한 예측의 개선을 가져오지 못한 한계점을 보였다.

〈표 8〉 모형별 예측성능 비교 - 최종통합(최적 가중치 부여후)

모형번호	모형형태	전체최종통합	상세최종통합		
			(불량)	(미정)	(우량)
1	MDA	84.40	81.02	2.02	97.37
2	MLP				
3	DTM				

모형통합에 관한 연구결과를 전체적으로 종합해보면 초기에 얻어진 7개의 개별모형을 같은 형태별로 1차 통합한 모형 2개는 각 종류별로 통합 대상인 개별모형을 적절히 혼합한 특성을 나타냈으며 전반적인 성능 또한 향상되었다. 2개의 1차 통합모형을 다시 통합하여 얻은 최종통합 모형은 1차통합 모형 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.

4.4 본 연구의 시사점

본 연구의 시사점은 다음 몇 가지로 요약할 수 있다.

첫째, 본 연구는 기존에 진행되었던 개별모형에 대한 검증은 물론, 단순히 여러 개의 개별모형을 비교·분석하여 우월한 모형을 평가하는 기존방법론상의 한계[6, 8, 12, 20, 25]를 극복하기 위해 개별모형을 유전자 알고리즘을 통해 통합모형을 구축하는 하나의 방법론을 제시하였다는데 의의가 있다.

둘째, 본 연구의 결과는 실무적으로 할부금융시장에 있어서 고객신용평가모형 구축 및 실행에 보다 나은 예측모형을 제공해주고 있다. 이러한 모형을 토대로 하는 신용평가에 활용하여 적용할 때 우량고객에 대한 높은 예측율을 기할 수 있다. 특히 본 연구에서 실험과정에 사용된 데이터는 기존의 대부분 연구에서 수십 또는 수백 단위에 비해 200,000개의 실 데이터의 정제과정과 완전 정보를 취하고 있는 데이터를 무작위로 6,500개 추출하여 학습, 검증, 시험 등에 할당하여 활용하였기 때문에 자료의 신뢰성과 타당성은 매우 높다. 그러므로 본 연구결과에 나타난 모형 예측율은 상당히 현실적인 결과를 나타낼 가능성이 높기 때문에 실무적이며 직접적으로 활용할 수 있다는 의미가 있을 것이다.

그러나 본 연구의 진행에는 몇 가지 고려해야 할 한계점이 있다.

첫째, 본 연구에서는 MDA, MLP, DTM등 세 가지 개별모형만을 고려하였다는 것이다. 기존의 일부 연구에서는 이러한 모형 이외에 로지스틱 회귀분석과 사례기반 추론모형, 그리고 선형계획모형을 함께 진행하여 비교하고 있다[6, 20]. 그러나 본 연구 결과 도출된 통합모형을 좀 더 개선하기 위한 노력의 일환으로 비록 낮은 예측결과를 보여주지만, 선형계획모형 결과와 사례기반 추론 및 퍼지집합 모형을 모두 적용할 수 있는 유전자 알고리즘 방법론을 개발하여 이를 포함한 통합모형의 구축 노력이 필요하다.

둘째, 본 연구모형에 선택된 입력변수 중 인구 통계적 특성 변수들이 상당수 탈락하고 있기 때문에 실무적으로 적용함에 있어서 기존에 거래내역이 없는 신규고객의 신용을 평가하기가 곤란하다. 따라서 본 연구의 결과를 실무적으로 활용함에 있어서 기존 신용거래고객의 행위 중심으로 적용될 수밖에 없는 한계점이 있다.

5. 결 론

본 연구에서는 할부금융시장에서의 고객정보 및 할부진행

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

앞으로 향후의 연구에서는 본 연구에서 한계점을 극복할 수 있는 통합모형 도출상의 보완과 인구 통계적 특성변수의 확장 방법, 통합모형의 선택입력 변수의 도출 및 동일한 고객에 대한 다양한 원천의 데이터와 기준에 의한 연구가 요청되는 바이다.

참 고 문 헌

[1] 김갑식. (2003). "할부금융고객의 신용평가를 위한 데이터마이닝 통합모형구축", 대구가톨릭대학교 대학원 박사학위 논문.

[2] 김홍철. (2001). "유전자 알고리즘기반 복수 분류모형 통합에 의한 할부금융고객의 신용예측모형", 대구대학교 대학원 석사학위 논문.

[3] 정충영, 최이규. (1998). SPSSWIN을 이용한 통계분석, 서울, 무역경영사.

[4] 채서일. (1999). 사회과학 조사방법론, 2판, 서울, 학현사.

[5] 최종후, 한상태. (2000). AnsweerTree를 이용한 데이터마이닝 의사결정나무분석, 서울, SPSS 아카데미.

[6] Boyle, M., Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). Methods for Credit Scoring Applied to Slow Payers, In Thomas, L. C., Crook, J. N., & Edelman, D. B.(eds.), *Credit Scoring and Credit Control, Oxford University Press*, Oxford, pp.75-90.

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

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

[9] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA.

[10] Grablowsky, B. J., & Talley, W. K. (1981). "Probit and Discriminant Functions for Classifying Credit Applicants: A Comparison", *Journal of Economics and Business*, 33, pp. 254-261.

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

[12] Henley, W. E. (1995). "Statistical Aspects of Credit Scoring", PhD Thesis, Open University.

[13] Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, MI.

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

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

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

[17] Kim, E., Kim, W., & Lee, Y., (2000). "Purchase Propensity Prediction of EC Customer by Combining Multiple Classifiers Base on GA", *Proceedings of International Conference on Electronic Commerce*, pp.274-280.

[18] Mangasarian, O. L. (1965). "Linear and Nonlinear Separation of Patterns by Linear Programming", *Operations Research*, 13, pp.444-452.

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

[20] Srinivasan, V., & Kim, Y. H. (1987). "The Bierman-Hausman Credit Granting Model: A Note", *Management Science*, 33, pp.1361-1362.

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

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

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

[24] Wong, B. K., Bodnovich, T. A., & Selvi, Y. (1997). "Neural Network Applications in Business: A Review and Analysis of the Literature(1988-95)", *Decision Support Systems*, 19, pp.301-320.

[25] Yobas, M. B., Crook, J. N., & Ross, P. (1997). "Credit Scoring Using Neural and Evolutionary Techniques", Credit Research Centre, University of Edinburgh, Working Paper.

김 갑 식

e-mail : kskim@mail.tpic.ac.kr

1989년 2월 계명대학교 일본학과(문학사)

1991년 2월 경일대학교 전자계산학과 (공학사)

1991년 8월 계명대학교 경영대학원 경영정보학과(경영학석사)

2003년 8월 대구가톨릭대학교 대학원 경영학과 경영정보학 전공 (경영학박사)

1993년 3월~현재 대구산업정보대학 인터넷비즈니스과 교수

관심분야 : 데이터마이닝, 데이터웨어하우징, 중소기업정보화, 생산공정정보화, ERP, CRM, 전자상거래, e-비즈니스