Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.4 pp. 825~836

A Study of Data Mining Optimization Model for the Credit Evaluation

Kap-Sik Kim[1] · Chang-Soon Lee[2]

Abstract

Based on customer information and financing processes in capital market, we derived individual models by applying multi-layered perceptrons, MDA, and decision tree. Further, the results from the existing single models were compared with the results from the integrated model that was developed using genetic algorithm. This study contributes not only to verifying the existing individual models and but also to overcoming the limitations of the existing approaches. We have depended upon the approaches that compare individual models and search for the best-fit model. However, this study presents a methodology to build an integrated data mining model using genetic algorithm.

Keywords : data mining, genetic algorithm, multi-layered perceptrons

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 증대시켜 주기 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, 후

1) 대구광역시 수성구 만촌3동 산395번지 대구산업정보대학 컴퓨터정보계열 교수
E-mail : kskim@mail.tpic.ac.kr
2) 경북 경산시 유곡동 290번지 대구한의대학교 정보과학부 부교수
E-mail : cslee@dhu.ac.kr

자는 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(Thomas, 2000).

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 학술적 관심이 높아지게 되었다.

데이터마이닝 분석을 위한 도구에는 인공신경망(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, 한가지 모형만 연구문제에 맞게 최적화하는 과정은 많은 시간과 노력이 요구되기 때문에 각 단일모형의 장점들만을 취하여 최적의 통합신용평가모형으로 연구문제에 맞게 최적화하는 것이 보다 효율적일 것으로 생각된다(김갑식, 2003).

따라서 본 연구는 다계층 퍼셉트론(Multi-Layered Perceptrons:MLP) 구조를 갖는 인공신경망 모형과 다변량 판별분석(MDA)모형 그리고 의사결정나무모형 등을 이용하여 각각의 단일 모형을 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 판별분석이나 로지스틱 회귀분석, 프로빗 분석과 같은 통계학적 모형(Wiginton, 1980 ; Grablowsky and Talley, 1981)과, 선형계획법(linear programming : LP)(Mangasarian, 1965)과 같은 경영과학적 모형을 들 수 있다. 최근 들어 의사결정나무(decision tree), 인공신경망 등의 인공지능(artificial intelligence) 모형을 이용한 연구가 활발하게 진행되었다. 특히 기존 연구에서 인공신경망 모형을 이용한 연구가 좋은 결과를 보여주고 있다(West, 2000 ; Wong, et. al, 1997; Jain and Nag, 1997).

인공신경망 모형은 데이터 마이닝에 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(Cheng & Titterington, 1994).

판별분석(discrimination Analysis)은 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(정충영, 최이규, 1998). 데이터마이닝을 위한 다변량 판별분석모형의 경우 구현이 간단하고 학습시간도 짧지만 독립변수들이 다변량 판별분석의

기본적인 통계학적 가정들을 만족해야 하므로 이에 대한 검증이 필요하다는 한계점을 가지고 있다(채서일, 1999).

의사결정나무는 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, 인공신경망, 판별분석(discriminant analysis), 회귀분석(regression analysis) 등에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(최종후, 한상태; 2000). 이 모형은 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단할 뿐만 아니라 이해하기 쉬운 규칙으로 전환될 수 있다(Imielinski & Mannila, 1996). 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다.(Mehta, 1968)

유전자 알고리즘은 인공지능의 한 모형으로서 2차원 이상의 복잡한 탐색공간에서 전 범위의 최적해(global optimal solution)를 탐색하는데 아주 효율적이며, 유연하다(Gupta, 1995). 이러한 유전자 알고리즘은 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(Hon & Chi, 1994). 이러한 유전자 알고리즘은 Holland(1975)가 그 이론적인 근거를 마련했으며, Goldberg(1989)에 의해 공학분야에서 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(Kim, et. al, 2000).

3. 연구의 설계

3.1 접근 방법론

본 연구에서는 할부금융시장에서의 고객정보 및 할부진행과정에 대한 세부 내역을 바탕으로 여러 가지 분류모형(Classifier)들을 유전자 알고리즘(Genetic Algorithm)을 이용하여 통합한 신용예측모형을 제안한다. 이를 위해 다층퍼셉트론 (Multi-Layered Perceptrons: MLP)구조를 갖는 인공신경망모형, 다변량판별분석에 의한 선형모형에서부터 각각 복수 개의 분류모형을 얻는다. 그리고 의사결정나무 모형에서 단수개의 모형을 얻는다. 그 다음 복수개의 분류모형을 갖는 MLP모형과 다변량 판별분석 모형을 유전자 알고리즘에 의해 세 가지의 부류의 대표 모형을 도출단다. 그 다음으로 이 두 가지 모형의 대표모형과 의사결정나무 모형을 다시 같은 방식으로 통합하여 최종 모형을 구했다.

신경망 모형 및 기타 모형들은 각기 다른 특성을 가지며 연구상황에 따라 서로 다

른 성능을 보이므로 어떤 모형이 우수하다고 단정할 수 없으며 연구문제에 대한 최적 모형을 얻기 위한 통합의 필요성이 제기된다.

본 연구에서는 유전자 알고리즘(Goldberg, 1989)을 이용하여 복수 분류모형 통합모듈의 가중치행렬을 최적화하는 방식으로 개별 모형들을 병렬식으로 가중통합을 하였다 (Kim, et. al., 2000).

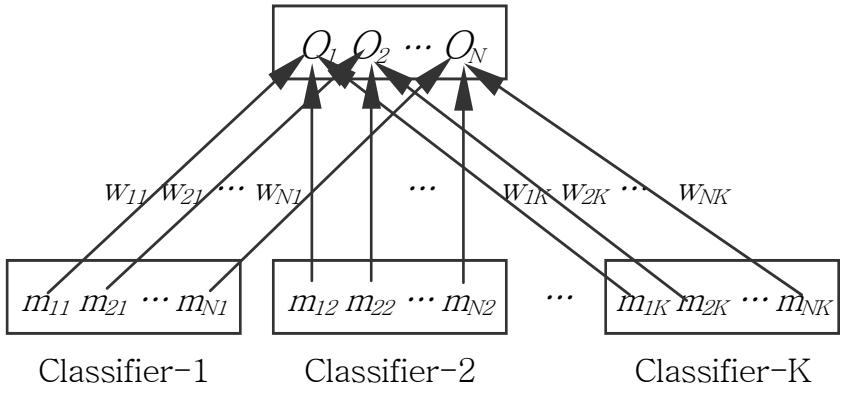
이를 위해서 가령 종속변수가 취할 수 있는 값이 N개 즉, 분류해야할 집단의 개수가 N개 이고, 통합해야 할 분류모형이 K개 있다고 가정할 때 모형의 결과 값 O_i는 식 (1)과 정의될 수 있다. 그리고 식 (2)에서 보는 바와 같이 한 패턴이 취할 수 있는 값 E(x)는 O_i 가운데 최대값을 골라 그 값이 일정값(α)을 넘어갈 경우에 그 값으로 하고 그렇지 않을 경우에는 값을 주지 않는다. 이 때 E(x)의 값이 원래의 값과 같을 경우에는 식 (3)에서 보는 바와 같이 유전자 알고리즘의 Fitness Function에 1의 값을 주고 그렇지 않을 경우에는 0의 값을 주었다. 이와 같은 방식으로 <그림 1>에서 보는 바와 같이 복수개의 분류모형을 결합할 수 있는 가중벡터인 W를 구한다(김홍철, 2001).

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

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

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

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



<그림 1> 복수 분류모형 통합모듈의 구조

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 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 0사이의 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 스케일변

수를 개월수 또는 평균금액으로 나누어주는 과정이 필요하다. 이를 통해서 금액과 개월 수 등의 스케일이 다른 변수 값을 0에서 1사이의 값으로 정규화 하였다.

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

<표 2> 변수 상세 설명

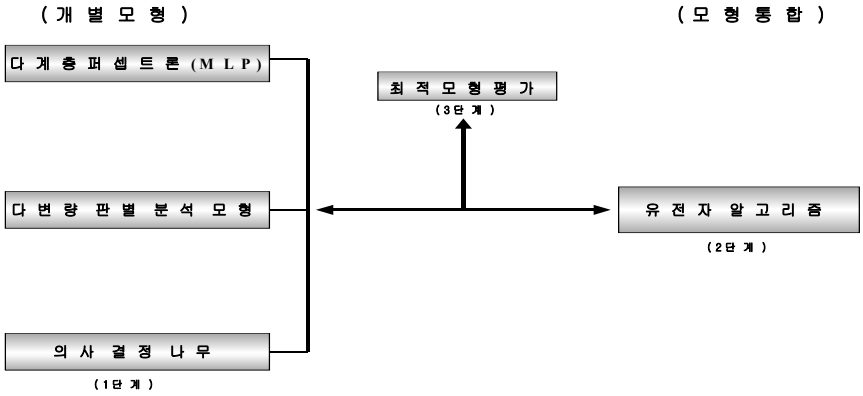
변수명	설 명	상 세 설 명	
A1	나이		※ 아래의 변수 설명 중에서 3개월은 1997. 11 ~ 1998. 1을 6개월은 1997. 8 ~ 1998. 1을 말한다.
A2	성별		
A3	보증인수		
A4	매입지역		
A5	차량원부		
A6	차종		
A7	차량년식		
A8	배기량		
A9	신용조사방법		
A10	구매자구분		
B1	3개월의무납입액평균/ 3개월잔액평균	1997. 11 ~ 1998. 1까지 납부해야할금액의 월평균/1997. 11 ~ 1998. 1까지 각 월의 총할부잔액 평균 (관측기간이전의 내역을 반영하기 위한 변수)	
B2	6개월의무납입액평균/ 6개월잔액평균	1997. 8 ~ 1998. 1까지 납부해야할금액의 월평균/1997. 8 ~ 1998. 1까지 각 월의 총할부잔액 평균 (관측기간이전의 내역을 반영하기위한 변수)	
B3	98년1월 의무납입액/잔액	98년 1월의 의무납입액 / 총할부잔액 - 관측기간의 처음 상태를 보기 위한 변수	
B4	98년12월 의무납입액/잔액	98년 12월의 의무납입액 / 총할부잔액 - 관측기간의 마지막 상태를 보기 위한 변수	
B5	98년1월 실납입액/의무납입액	98년1월의 실납입액 / 98년1월의 의무납입액 - 관측기간의 처음 납부실적 상태를 보기 위한 변수	
B6	98년12월 실납입액/의무납입액	98년12월의 실납입액 / 98년12월의 의무납입액 - 관측기간의 마지막 납부실적 상태를 보기 위한 변수	
B7	3개월납입액평균/ 3개월의무납입액평균	1997. 11 ~ 1998. 1까지 3개월동안 실납입액평균 / 1997. 11 ~ 1998. 1까지 3개월 의무납입액평균 (관측기간이전의 내역을 반영하기위한 변수)	
B8	6개월납입액평균/ 6개월의무납입액평균	1997. 8 ~ 1998. 1까지 6개월동안 실납입액평균 / 1997. 8 ~ 1998. 1까지 6개월 의무납입액평균 (관측기간이전의 내역을 반영하기위한 변수)	
B9	12개월납입액평균/ 12개월의무납입액평균 (98년1월 납입액을 이전6개월 평균으로)	98년 12개월동안 납입액평균 / 98년 12개월동안의 의무납입액평균 (98년 1월의 납입액을 이전 6개월 간의 납입액 평균으로 하였다.- 역시 관측기간 이전의 내역을 반영하기 위함이다.	
B10	98년 12개월간 최장연체횟수	98년 12개월간의 최장연체 개월수 (98년 1년 동안의 연속연체개월수)	

B11	98년 12월잔액/ 3개월잔액평균	98년 12월의 시점에서의 할부총잔액 / 1997. 11~1998. 1까지 3개월 할부 총잔액평균
B12	98년 12월잔액/ 6개월잔액평균	98년 12월의 시점에서의 할부총잔액 / 1997. 8~1998. 1까지 6개월 할부 총잔액평균
B13	98년 12개월간 연체액평균	98년 12개월간의 연체액 월평균
B14	98년 12개월간 연체개월수/ 총할부개월수	98년 12개월간 총연체개월수 / 총할부 계약개월수 (총할부개월수 대비 관측기간동안의 연체개월 수)
B15	98년 12개월간 연체개월수/ 12개월	98년 12개월간 총연체개월수 / 12 (년평균 연체개월수를 말한다.)
B16	매월 납부액	매월 의무납부금액
B17	총할부개월수	총할부 계약 개월수
B18	할부가격(할부원금+할부이자)	총할부원금액 + 총할부이자
B19	우:불량판별 (1:불량, 2:미정, 3:우량)	종속변수 (판단기간 동안의 연체개월 수에 따라 2개월 이하는 우량, 4개월 이상은 불량, 3개월은 미정으로 한다.)

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

3.4 연구방법 및 모형

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.



<그림 2> 단위 분류모형의 통합방법

4. 연구모형의 실험결과

유전자 알고리즘의 통합과정에서 도출된 가중치행렬은 <표 3>과 같다. 여기에 나타난 가중치행렬에서 각 행(row)은 통합되어지는 개별모형에 대한 가중치이며 열(column)은 분류 결과 값에 대한 가중치를 의미한다. MDA*와 MLP* 모두 1번째 모형의 2(미정)값에 가중치를 많이 두고 있음을 보여준다. <표 3>의 가중치 행렬은 Palisade Evolver 4.0 for Excel을 통해서 도출되었다.

<표 3> 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 4> 최종통합모형의 가중치행렬 (W)

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

최종 통합모형의 가중치행렬은 <표 4>와 같다. 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. 최종통합모형에서는 모형의 예측율을 높여주기 위해서 개별 모형 중에서 예측부분이 가장 미진한 부분에 해당하는 값에 가중치를 부여해주고 있다. 우량값과 미정에 가중치가 부여된 최종통합모형을 구체적으로 살펴보면 우량값의 경우 예측값이 가장 낮은 MDA모형에 0.87이 부여되었다. 미정값의 경우에도 MDA에 0.65의 가중치를 부여하였다. <표 5>의 최종통합 예측율은 Palisade Evolver 4.0 for Excel을 통해서 도출되었다.

<표 5> 모형별 예측성능 비교 - 최종통합(최적 가중치 부여후)

모형번호	모형형태	개별모형 예측율(%)	1차 통합	최종 통합	개별모형			1차통합			최종통합		
					(불량)	(미정)	(우량)	(불량)	(미정)	(우량)	(불량)	(미정)	(우량)
1	MDA	81.92	83.80	84.40	81.02	2.02	97.37	81.48	2.02	96.35	81.02	2.02	97.37
		82.27			82.87	3.03	94.01						
		81.81			81.48	6.06	95.04						
2	MLP	83.75	82.40		78.24	13.13	93.72	78.24	23.23	89.64			
		84.44			78.70	22.22	89.34						
		84.53			77.31	28.28	88.76						
3	DTM	82.20	82.20		80.56	53.54	86.86	80.56	53.54	86.86			

신용평가 예측을 위한 여러 가지 데이터마이닝 모형을 비교·분석한 결과 유전자 알고리즘을 통해 개별모형을 결합한 통합모형의 성능이 가장 우수한 결과를 보여주었다.

<표 5>를 보면 개별모형에 대한 최초 실험에 의한 신용평가결과는 81.81%~84.53%의 상당한 수준의 예측율이 나타났다. 유전자 알고리즘을 통한 MDA와 MLP의 1차 통합에서는 미정값에 가중치가 부여되었다. 1차 통합에 의한 개별모형의 예측비교에서는 MDA가 83.80%로 가장 우수한 성능 결과를 보였다. 그러나 MDA의 미정에 대한 예측율은 2.02%로 매우 낮은 결과를 보였다. DTM과 MLP는 전반적으로 불량, 미정, 우량 집단에 대하여 비교적 고른 예측성능을 나타내었다.

최종 통합과정에서는 MDA의 미정값과 DTM의 우량값에 가중치가 부여되었다. 이러한 가중부여를 통한 최종통합모형의 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 예측율 중 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 한 계점을 보였다.

5. 결 론

본 연구에서는 데이터마이닝에 사용되는 분석모형을 할부금융에 있어서 신용평가 예측에 적용하였다. 신용평가 예측을 위한 여러 가지 데이터마이닝 모형을 비교·분석한 결과 유전자 알고리즘을 통해 개별모형을 결합한 통합모형의 성능이 가장 우수

한 결과를 보여주었다.

이에 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(Henley,1995 ; Boyle et al.,1992 ; Srinivasan and Kim, 1987 ; Yobas et al., 1997 : Desai et al., 1997)를 극복하기 위해 개별모형을 유전자알고리즘을 통해 통합모형을 구축하는 하나의 방법론을 제시하였다는데 의의가 있다.

둘째, 본 연구를 통해 도출된 통합모형은 기존의 연구에서 제시된 60%에서 70%정도의 결과보다 높은 84.40%의 예측율을 보이고 있다. 따라서 데이터마이닝을 통한 신용평가모형 구축방법론으로써 본 연구의 유전자 알고리즘 통합방법론은 그 유용성이 높다. 그리고 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(Boyle et al., 1992; Srinivasan & Kim ,1987). 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 유전자알고리즘 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 기존신용거래고객의 행위 중심으로 적용될 수밖에 없는 한계점이 있다.

셋째, 본 연구를 위한 분석도구로 인공신경망 모형과 다변량 판별분석 모형에는 Statistica-Neural Networks Version 4, 의사결정트리 모형에는 C5 of Clementime V. 5.0 package, 그리고 통합모형인 유전자 알고리즘 모형에는 Evolve V.4를 이용하였다. 이러한 도구를 많이 사용한 것은 연구자의 기술적인 한계가 원인이 되었는데 향후의 연구에서는 데이터마이닝 실험을 위한 분석도구를 최대한 줄여서 1-2개의 도구를 사용해서 실험의 불편함을 최대한 줄이는 것이 연구의 효율성 및 편의성에 상당히 도움이 된다고 생각한다.

넷째, 본 연구에 사용된 데이터가 X할부금융회사의 고객데이터를 활용하였기 때문에, 이 회사에서 진행된 고객의 신용행동과 평가내용이 타 회사에서도 동일하게 진행되었는지를 파악하기 힘들다. 즉 신용의 우·불량 기준이 회사마다 다르고, 고객 또한 두 개 이상의 할부금융회사를 이용할 때 동일한 신용행위를 진행할 것인지에 대한

파악이 불가능하다.

향후의 연구에서는 이상의 한계점을 극복할 수 있는 통합모형도출상의 보완과 인구통계적 특성변수의 확장 방법, 통합모형의 선택입력 변수의 도출 및 실험도구 수의 축소, 그리고 동일한 고객에 대한 다양한 원천의 데이터와 기준에 의한 연구가 요청되는 바이다.

참고문헌

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. 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.
11. Henley, W. E. (1995). "Statistical Aspects of Credit Scoring", PhD Thesis, Open University.
12. Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, MI.
13. Hon, K. K. B., & Chi, H. (1994). "A New Approach of Group Technology Part Families Optimization", *Annals of the CIRP*, 43(1).
14. Imielinski, T., & Mannila, H. (1996). "A Database Perspective on Knowledge Discovery", *Communications of the ACM*, 39(11), pp 214-225.
15. 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.
16. 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.
17. Mangasarian, O. L. (1965). "Linear and Nonlinear Separation of Patterns by Linear Programming", *Operations Research*, 13, pp. 444-452.
18. Mehta, D. (1968). "The Formulation of Credit Policy Models", *Management Science*, 15, pp. 30-50.
19. Srinivasan, V., & Kim, Y. H. (1987). "The Bierman-Hausman Credit Granting Model: A Note", *Management Science*, 33, pp. 1361-1362.
20. 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.
21. West, D. (2000). "Neural Network Credit Scoring Models", *Computers & Operations Research*, 27, pp. 1131-1152.
22. 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.
23. 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.
24. Yobas, M. B., Crook, J. N., & Ross, P. (1997). "Credit Scoring Using Neural and Evolutionary Techniques", Credit Research Centre, University of Edinburgh, Working Paper.

[2003년 8월 접수, 2003년 10월 채택]