유전자 알고리즘기반 복수 분류모형 통합에 의한 할부금융고객의 신용예측모형

A credit prediction model of a capital company's customers using genetic algorithm based integration of multiple classifiers

이웅규* · 김홍철

* 대구대학교 경영학과

Abstract

본 연구에서는 할부금융시장에서의 고객신용예측을 위한 모형으로 여러 가지 인공신경망(Neural Network) 모형들을 유전자 알고리즘(Genetic Algorithm)을 이용하여 통합한 신용예측모형을 제안한다. 10개의 학습된 인공신경망 모형들을 유전자알고리즘을 이용하여 종류별로 통합하여 MLP (Multi-Layered Perceptrons), Linear, RBF(Radial Basis Function) 세 가지의 대표모델을 얻고 이를 다시 하나의 인공신경망 모델로 통합하였다. 이를 통합되기 이전의 각각의 인공신경망 모형들과 성능을 비교, 분석하여 본 연구에서 제안한 통합모형의 유효성과 통합방법의 타당성을 제시하였다.

1. 서론

"할부금융"이란 재화 및 용역의 매매계약에 대하여 매도인(기업에 한하되, 주택매매에 있어서는 개인을 포함한다) 및 매수인과 각각 약정을 체결하여 매수인에게 융자한 재화 및 용역의 구매자금을 매도인에게 지급하고 매수인으로부터 그 원리금을 분할하여 상환 받는 방식의 금융을 말한다.(여신전문금융업법 2조 13항)

할부금융회사에서는 우·불량고객의 판별 및 고객 신용등급 관리를 통하여 불량채권 발생율을 미연에 감소시켜 궁극적으로 할부금융회사의 수익을 증대 시키는 효과를 거둘 수 있으므로 그 중요성이 매우 크다. 신용평가를 위해 수집된 자료들을 바탕으로 고객에 따라 차별화 된 금융상품과 여러 혜택들을 제공하고 고객에게 위험상황을 사전에 통지함으로써 고객관계마케팅을 실현할 수 있게 된다.

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, 반경기반함수(Radial Basis Function: RBF)에 의한 인공신경망모형 그리고 다변량판별분석에 의한 선형모형에서부터 각각 복수개의 분류모형을 얻은 다음 이를 유전자 알고리즘에 의해 세 가지 부류의 대표 모형을 도출하고 다시 같은 방식으로 이들 세 가지 대표모형을 통합하여 최종 모형을 구했다.

2절에서는 신용평가와 관련된 선행 연구들을 살펴보고, 3절에서는 신용예측을 위한 통계적 기법과 인공지능 기법들에 대해 살펴보며, 복수 분류모형의 통합기법과 그 과정을 알아본다. 4절에서는 본 연구의 실험설계 및 신용예측모형 통합 과정을 보이며, 5절에서는 제안한 통합예측모형과 기존의 개별모형의 성능을 비교 분석한다.

2. 문헌고찰

신용평가는 고객에 대하여 신용을 부여할지 여부를 결정하는 기법으로 크게 신규고객이 대출신청을 할 때 고객의 정보만을 가지고 신용을 평가하는 협의의 신용평가(Credit Scoring)와 기존고객의 신용을 평가하기 위해 고객의 거래 내역을 바탕으로 신용을 평가하는 행태평가(Behavioral Scoring)로 나뉠 수 있다 (Thomas, 2000; Johnson,1992).

전통적으로 신용평가를 위한 기법으로는 로지스틱 회귀분석이나 프로빗 분석법과 같은 통계학적 기법 (Wington, 1980,; Grablowsky and Talley, 1981)과 선형계획법(Mangasaria, 1965)과 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(Makowsky, 1985; Carter and Cartlett, 1987).

한편 1980년대 중반부터 경영분야에서 널리 사용되기 시작한 인공신경망 모형은 통계적 가설이 필요없으면서도 비선형적인 회귀모형을 설명하기에 적당하기 때문에 신용평가에서 널리 사용되어 뛰어난 성과를 보여 주고 있다 (Altman, Marco and Varetto, 1994; Desai, Crook and Overstreet, 1996; Desai, Convay, Crook and Overstreet, 1997).

이와 같이 다양한 기법들이 제안되고 학술적인 성과를 보이고 있음에도 불구하고 어떤 기법이 가장 뛰어난 기법인지 절대적인 판단을 내릴 수는 없다. 실제 신용평가에 적용된 기법들의 성능을 비교한 여러 보고에서는 상이한 결과를 제시하고 있다 (Thomas, 2000). 따라서 어떠한 기법도 절대적으로 우수하다고 단정할 수 없으며 연구문제에 대한 최적기법을 찾기 위한 기법들에 대한 통합의 필요성과 방법론들이 제안되고 있다.(Kim, et. al, 2000)

3. 접근 방법론

본 연구에서는 할부금융시장에서의 고객정보 및

할부진행과정에 대한 세부 내역을 바탕으로 여러 가지 분류모형(Classifier)들을 유전자 알고리즘(Genetic Algorithm)을 이용하여 통합한 신용예측 모형을 제안한다. 이를 위해 다층퍼셉트론(Multi-Layered Perceptrons: MLP)구조를 갖는 인공신경망모형, 반경기반함수(Radial Basis Function : RBF)에 의한 인공신경망모형 그리고 다변량판별분석에 의한 선형모형에서부터 각각 복수 개의 분류모형을 얻은 다음 이를 유전자 알고리즘에 의해 세 가지의 부류의 대표 모형을 도출하고 다시 같은 방식으로 이들 세 가지 대표모형을 통합하여 최종 모형을 구했다.

MLP 모형은 역전파(Backpropagation) 알고리즘에 의한 학습으로 우수한 예측성능을 검증 받은 바 있으나 최적 모형의 설계를 위해서는 상당한 학습시간을 필요로 하며, 은닉계층의 노드들이 Radial Unit들로 구성된 RBF의 경우에는 MLP에 비해 짧은 시간 내에 학습시킬 수 있으나 모형의 최적화하는데 있어 MLP에 비해 상대적으로 어려운 단점이 있다(Hwang, 1997). 한편 선영모형의 경우 구현이 간단하고 학습시간도 짧지만 독립변수들이 다변량 정규분포나 다중공선성과 같은 통계적 가정들을 만족해야 하는 한계점을 가지고 있다 (채서일, 1999).

이와 같이 신경망 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(Goldberg, 1989)을 이용하여 복수 분류모형 통합모듈의 가중치행렬을 최적화하는 방식으로 개별 모형들을 병렬식으로 가중통합을 하였다 (Kim, et. al., 2000).

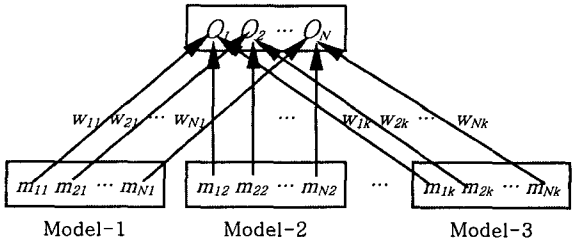
이를 위해서 가령 종속변수가 취할 수 있는 값이 N개 즉, 분류해야할 집단의 개수가 N개 이고, 통합해야 할 분류모형이 K개 있다고 가정할 때 모형의 결과 값 O_i는 <식 1>과 정의될 수 있다. 그리고 <식 2>에서 보는 바와 같이 한 패턴이 취할 수 있는 값 E(x)는 O_i 가운데 최대값을 골라 그 값이 일정값(α)을 넘어갈 경우에 그 값으로 하고 그렇지 않을 경우에는 값을 주지 않는다. 이 때 E(x)의 값이 원래의 값과 같을 경우에는 <식3>에서 보는 바와 같이 유전자 알고리즘의 Fitness Function에 1의 값을 주고 그렇지 않을 경우에는 0의 값을 주었다. 이와 같은 방식으로 그림-1에서 보는 바와 같이 복수개의 분류모형을 결합할 수 있는 가중벡터인 W를 구한다.

$$O_i = \sum_{k=1}^{K} w_{ik} m_{ik} \qquad \text{<식1>}$$

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

$$E(x) = \begin{cases} j, \ if \ o_j = \max_{i \in \Lambda}(o_i) \ and \ o_j \geq \alpha \\ reject, \quad otherwise \end{cases} \qquad \text{<식2>}$$

$$HF(WS_q) = \begin{cases} 1 \ , \ if \ correctly \ matched \\ 0 \ , \ otherwise \end{cases} \qquad \text{<식3>}$$



<그림1> **복수 분류모형 통합모듈의 구조**

4. 실험설계

1999년 1월을 기준으로 이전 12개월(1998년 1월부터 12월까지)의 할부진행과정을 관측하여 그 이후 6개월간의 행동양상(우 · 불량)을 예측한다.

4.1. 변수

<부록1>에 나타난 항목들은 원시데이터를 예측모형의 입력변수로 사용하기 위해 정규화 등의 과정을 거쳐 적절히 가공한 변수목록이다.

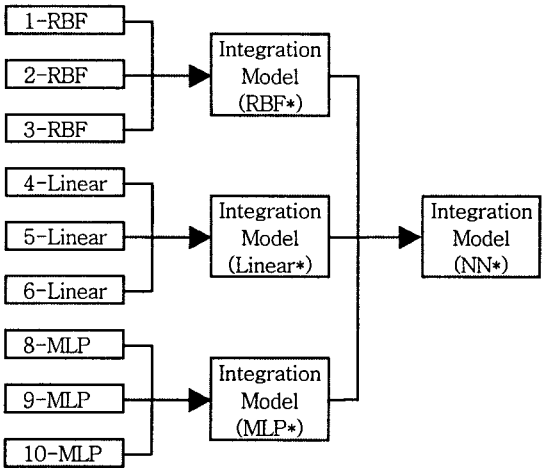
변수 g_b_u_ex는 채권의 우 · 불량을 판별하는 종속변수로서 1998년 1월 이후 6개월 동안의 연체개월 수가 4개월 이상이면 1(불량), 2개월 이하이면 2(우량), 3개월인 것은 3(미정)의 값을 갖는다. 나머지 변수들(채권번호 제외)은 대상입력변수들이며 금액과 관련된 변수들은 평균값으로 나누어 주는 방법을 통해 정규화(Normalization) 하였다.

4.2 예측 모형의 개발과 통합절차

인공신경망 개발도구(Statistica-Neural Network)를 이용하여 선형모형-3개, RBF모형-3개, MLP모형-4개 총 10개의 개별 예측모형을 개발한 다음 각 모형별로 예측성능을 평가해 보았다. <표1>은 10개의 예측모형에 대한 정보이며 대상입력변수 중에서 실제 입력변수로 채택된 변수의 개수, 은닉노드의 개수, 평가성능(상세기록포함)을 나타내고 있다.

<표1> **각 모형별 특성 및 성능**

No	Type	Inputs	Hidden	ScPerf (%)	ScPerf (불량)	ScPerf (우량)	ScPerf (미정)
1	RBF	18	19	79.8	76.29	93.95	0
2	RBF	18	17	80.3	77.59	94.25	0
3	RBF	18	18	80.2	78.45	93.8	0
4	Linear	25	-	80	85.34	91.07	0
5	Linear	26	-	80.4	86.21	91.23	0.93
6	Linear	27	-	81.6	90.95	91.23	1.87
7	MLP	1	2	81	87.07	91.98	0
8	MLP	5	3	80.7	90.52	89.71	3.74
9	MLP	6	5	79.6	89.22	87.9	7.48
10	MLP	5	3	81.3	90.95	90.02	6.54



<그림2> **복수 신용예측모형의 통합과정**

<표1>에서 나타나는 개별 모형 10개 중에서, 입력변수가 하나만 선정된 MLP모형(7번)을 제외한 9개의 분류모형을 유전자 알고리즘을 이용하여 통합하였다. 각 모형은 번호순으로 3절의 <식1>에 적용되었고 일차로는 같은 형태별로 통합하여 3개의 대표모형 Linear*, RBF*, MLP*를 얻고 이것을 같은 방법으로 다시 통합하여 최종모형 NN*를 얻었다. <그림2>는 모형의 통합과정을 나타낸 그림이다.

4.3 표본

본 연구에서 사용된 표본자료는 1997년 8월부터 2000년 5월까지의 국내의 모 할부금융회사의 고객정보 및 할부진행과정에 대한 세부 내역이다.

10,229개의 무결한 표본 중에서 우량·불량·미정데이터를 각각 무작위 추출하여 적정 비율로 구성(Balancing)한 3500개(우량:1500, 불량:1500, 미정:500)의 데이터를 개발도구에 의한 예측모형의 개발(학습:1750, 검증:875, 시험용:875)에 사용하였고, 예측성능평가(Scoring)에는 앞서 사용한 3500개를 제외한 다른 1000개의 데이터를 사용하였다. 유전자 알고리즘을 이용한 개별모형의 통합에는 아직까지 사용하지 않은 데이터 중에서 1000개를 적정비율(우량:450, 불량:450, 미정:100)로 추출하여 사용하였고 통합모형의 예측성능평가에는 또 다른 1000개의 데이터를 사용하여 중복 사용된 데이터가 없도록 하였다.

5. 실험결과

<표2>은 유전자 알고리즘에 의해 최적화된 가중치행렬 $W_k = \{w_{1k}, w_{2k}, \cdots, w_{Nk}\}$의 결과이다.

<표2> **통합모형별 가중치행렬 (W_k)**

RBF*	0.1 0.1 0.1 0.1 0.1 0.42 0.1 0.99 0.1	Linear*	0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
MLP*	0.1 0.69 0.1 0.1 0.1 0.1 0.1 0.1 0.1	NN*	0.1 0.1 0.1 0.1 0.1 0.1 0.39 0.47 0.1

<표2>에 나타난 가중치행렬의 행은 통합되어지는 모형에 대한 가중치이며 열은 종속변수의 값에 대한 가중치이므로 RBF*의 경우 2번째 모형의 3(미정)값과 3번째 모형의 2(우량)값에 가중치를 많이 주고 있음을 알 수 있다. 실제로 <표1>의 개별모형의 예측성능평가를 보면 RBF의 경우 2, 3번째 모형이 첫째모형보다 우수한 성능을 보였다. Linear*를 이루는 가중치는 일정한데 Linear 개별모형들의 예측성능또한 비슷하게 나타났음을 볼수 있다. 개별모델의 성능이 높은 쪽에 가중치를 많이 둔다는 것은 확률적으로 예측성능이 우수해질 가능성이 많음을 의미한다.

<표3>는 통합모형별 예측적중률이며 <표4>는 통합예측모형과 기존의 개별모형들의 성능을 비교한 것이다.

<표3> **통합모형별 예측성능**

	HIT	Hit(불량)	Hit(우량)	Hit(미정)
RBF*	80.5	78.45	94.25	0
Linear*	79.7	90.95	88.5	0.93
MLP*	82	88.79	90.62	14.02
NN*	82	88.79	90.62	14.02

<표4> **모형별 예측성능 비교**

No	Type	ScPerf (%)	ScPerf (불량)	ScPerf (우량)	ScPerf (미정)	1차 통합모형	최종 통합모형
1	RBF	79.8	76.29	93.95	0	(RBF*) 80.5	(NN*) 82.0
2	RBF	80.3	77.59	94.25	0		
3	RBF	80.2	78.45	93.8	0		
4	Linear	80.0	85.34	91.07	0	(Linear*) 79.7	
5	Linear	80.4	86.21	91.23	0.93		
6	Linear	81.6	90.95	91.23	1.87		
8	MLP	80.7	90.52	89.71	3.74	(MLP*) 82.0	
9	MLP	79.6	89.22	87.9	7.48		
10	MLP	81.3	90.95	90.02	6.54		

<표4>을 보면 RBF*의 예측성능은 RBF개별모형에 비하여 향상되었고 Linear*는 나빠졌으며 MLP*는 향상되었다. 최종통합모형인 NN*는 <표2>의 가중치 행렬에서 가장 성능이 우수하게 나타난 MLP* 모형에 높은 가중치를 두었으며 예측율 또한 가장 우수한 모형과 같은 같게 나타났다.

결과적으로 전체적으로 2차에 걸친 통합에 의해서 9개의 개별모델보다 우수한 최적모델이 선택되었음을 알 수 있다.

References

Altman, E. I., Marco, G., & Varetto, F. "Corporate distress diagnosis; Comparisons using linear discriminant analysis and neural networks (the Italian experience)" *Journal of Banking and Finance* 18, 505, 1994.

Berry, Michael J. A. and Gordon Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley and Sons, 1997

Billings, Steve A. and Guang L. Zheng, "Radial Basis Function Network Configuration

Using Genetic Algorithms", *Neural Networks*, vol. 8, no. 6, pp. 877-890, 1995

Capon, N., "Credit scoring systems: a critical analysis", *Journal of Marketing*, vol. 46, pp. 82-91, 1982

Carter, C., & Catlett, J. "Assessing credit card applications using machine learning", *IEEE Expert* 2, 71-79, 1987.

Desai, V. S., Convay, D. G., Crook, J. N., & 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* 8, 323-346, 1997.

Desai, V. S., Crook, J. N., & Overstreet, G. "A. A comparison of neural networks and linear scoring models in the credit environment". *European Journal of Operational Research* 95, 24-37, 1996..

Goldberg, David E., Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, 1989

Grablowsky, B. J., & Talley, W. K. 'Probit and discriminant functions for classifying credit applicants; a comparison", *Journal of Economics and Business* 33, 154. 254-261, 1981.

Hopper, M. A., & Lewis, E. M., Behaviour Scoring and adaptive control systems. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 257-276, 1992

Hwang, Young-Sup and Sung-Yang Bang, "An Efficient Method to Construct a Radial Basis Function Neural Network Classifier", Neural Networks, vol. 10, no. 8, pp. 1495-1503, 1997

Kim, Eunju, Wooju Kim, and Yillbyung Lee, "Purchase Propensity Prediction of EC Customer by Combining Multiple Classifiers base on GA", *International Conference on Electronic Commerce 2000*, Proceedings, pp. 274-280, 2000

Makowski, P. "Credit scoring branches out" *Credit World* 75, 30-37, 1985

Mangasarian, O. L, "Linear and nonlinear separation of patterns by linear programming", *Operations Research* 13, 444-452, 1965.

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

Wang, C. H., T.P.Hong, and S.S.Tseng, "Integrating Fuzzy Knowledge by Genetic Algorithms", *IEEE Trans. On Evolutionary Computation*, vol. 2, no. 4, pp. 138-149, 1998.

West, David, "Neural Network Credit Scoring Models", *Computers & Operations Research*, vol. 27, pp. 1131-1152, 2000.

Wiginton, J. C. "A note on the comparison of logit and discriminant models of consumer credit behaviour", *Journal of Financial and Quantitative Analysis* 15, 757-770,1980.

손동우, 이웅규, "약체연결뉴런 제거법에 의한 부도 예측용 인공신경망 모형에 관한 연구", 한국정보시스템학회, 2000년 춘계학술대회 논문집, 2000.5., pp.115-121.

이웅규, 손동우, "부도 예측용 인공신경망 모형의 최적 입력노드 설계: 연결강도판별분석 접근", 한국지능정보시스템학회, 2000년 춘계정기학술대회 논문집, 2000.6., pp.251-258.

이웅규, 손동우, "연결강도판별분석에 의한 부도 예측용 신경망 모형의 입력노드 설계: 강체연결뉴런 선정 및 약체연결뉴런 제거 접근법", 한국지능정보시스템학회, 2000년 추계정기학술대회 논문집, 2000.11, pp. 469-477

이재규, 송용욱, 권순범, 김우주, 김민용, 'UNIK를 이용한 전문가시스템의 개발', 법영사, 1996.

이재규, 최형림, 김현수, 서민수, 주석진, 지원철, '전문가시스템 원리와 개발', 법영사, 1996.

채서일, '사회과학 조사방법론', 학현사, 1999. 2, 2판

삼성할부금융, '할부금융제도', 고려인쇄사, 1996.

<부록1> 대상 입력변수 목록

변수명	설 명	
cheno	채권번호	※ 아래의 변수 설명 중에서 3개월은 1997. 11~1998. 1 을 6개월은 1997. 8 ~ 1998. 1 을 말한다.
old_div	나이	
Sex	성별	
bocnt	보증인수	
maeji_re	매입지역	
wonbu	차량원부	
carjung	차종	
caryy_re	차량년식	
baegi_re	배기량	
singu_re	신용조사방.법	
bubgu_re	구매자구분	
adu_aba3	3개월의무납입액평균/3개월잔액평균	
adu_aba6	6개월의무납입액평균/6개월잔액평균	
due_bal1	98년1월 의무납입액/잔액	
du_bal12	98년12월 의무납입액/잔액	
pay_due1	98년1월 실납입액/의무납입액	
pa_due12	98년12월 실납입액/의무납입액	
apa_adu3	3개월납입액평균/3개월의무납입액평균	
apa_adu6	6개월납입액평균/6개월의무납입액평균	
apd_N6	12개월납입액평균/12개월의무납입액평균 (98년1월 납입액을 이전6개월의 평균으로)	
mcinc_12	98년 12개월간 최장연체횟수	
bal_1_a3	98년 12월잔액/3개월잔액평균	
bal_1_a6	98년 12월잔액/6개월잔액평균	
avamt12	98년 12개월간 연체액평균	
yn_hal12	98년 12개월간 연체개월수/총할부개월수	
ynn#r_12	98년 12개월간 연체개월수/12개월	
eveamt	매월 납부액	
halbun	총할부개월수	
halamt	할부가격(할부원금+할부이자)	
g_b_u_ex	우·불량판별 (1:불량, 2:우량, 3:미정)	