한국산학기술학회논문지
Vol. 11, No. 4, pp. 1419-1429, 2010

대출심사의 예측 정확도 향상을 위한 방법 제안

양유영[1], 박상성[1], 신영근[1], 장동식[1*]
[1]고려대학교 정보경영공학과

Proposing the Method for Improving the Forecast Accuracy of Loan Underwriting

Yu-Young Yang[1], Sang-Sung Park[1], Young-Geun Shin[1] and Dong-Sik Jang[1*]
[1]**Division of Information Management Engineering, Korea University**

요 약 외환위기 이후 본격적으로 시작된 외국계 대형 은행의 국내 진출 및 선진 금융상품의 수입은 국내 은행 산업 구조와 환경을 변화시키고 경쟁을 가속화시켰다. 앞으로 일어날 변화 및 추세에 대한 정확한 예측은 경쟁이 치열한 환경에서 국내의 은행이 생존하고 발전하기 위해 필수적인 요소이며 그 중에서도 대출 신청 고객에 대한 승인 여부에 대한 예측은 대출 상품이 은행 경영에 있어 가장 큰 비중을 차지하는 수익의 원천이자 신용 리스크 관리의 중심이 된다는 점에서 큰 의미가 있다. 따라서 본 논문에서는 대출 심사 결과의 예측 정확성을 높이기 위한 방법을 제시하고자 한다. 수행 단계로는 상관관계 분석과 특징선택 기법을 통해 대출승인 결과에 유의한 영향을 주는 예측변수들을 선별하고 선별된 변수로 2-Step 군집화 기법을 통해 고객을 군집화 하였다. 이후 각 군집에 LR, NN, SVM 기법을 활용하여 구축한 예측 모형을 적용하여 정확도가 가장 높은 모형을 찾아보았다. 최종적으로 기존 방식의 대출 심사 모형에 LR, NN, SVM 예측 모형을 적용했을 때 산출된 결과와 제안한 모형의 결과를 비교하여 예측의 정확도를 평가하였다.

Abstract Industry structure and environment of the domestic bank have been changed by an influx of large foreign-banks and advanced financial products when the currency crisis erupted in Korea. In a competitive environment, accurate forecasts of changes and tendencies are essential for the survival and development. Forecast of whether to approve loan applications for customer or not is an important matter because that is related to profit generation and risk management on the bank. Therefore, this paper proposes the method to improve forecast accuracy of loan underwriting. Processes in experiments are as follows. First, we select the predictor variables which affect significantly to the result of loan underwriting by correlation analysis and feature selection technique, and then cluster the customers by the 2-Step clustering technique based on selected variables. Second, we find the most accurate forecasting model for each clustering by applying LR, NN and SVM. Finally, we compare the forecasting accuracy of the proposed method with the forecasting accuracy of existing application way.

Key Words : Loan Underwriting, Correlation Analysis, Feature Selection, 2-Step Clustering, Logistic Regression, Neural Network, Support Vector Machine, Forecasting

1. 서론

외환위기 이후 본격적으로 시작된 외국계 대형 은행들의 국내 진출 및 선진 금융상품의 수입은 국내 은행 산업의 구조와 환경을 변화시키고 경쟁을 가속화시켰다. 경쟁이 치열한 환경에서 생존하고 발전하기 위해서는 다른 경쟁자들보다 먼저 변화를 감지하고 정확한 상황 분석을 통해 효과적인 대응을 해야 하는데 이는 앞으로 일어날

본 논문은 2010년도 두뇌한국 21사업에 의하여 지원되었음.
*교신저자 : 장동식(jang@korea.ac.kr)
접수일 10년 03월 09일 수정일 10년 04월 08일 게재확정일 10년 04월 09일

변화 추세에 대한 정확한 예측을 바탕으로 실현 가능하다. 예측이란 불확실성이 내재된 복잡한 미래의 문제를 과거 데이터를 토대로 근접한 해를 얻고자 하는 기법으로 은행 산업에서는 기업 및 개인의 도산, 환율, 금리, 신용평점, 대출상환, 연체 여부 등을 미리 알아보고자 할 때 사용된다. 이러한 분야들 중 대출을 신청한 고객에 대한 승인 여부에 대한 예측의 경우 은행의 수익 창출에 있어 다른 분야에 비해 매우 큰 중요성을 가진다고 할 수 있다. 대출은 은행 수익의 상당 부분을 차지하는 이자 수입의 원천[1]인 동시에 만약 대출자가 채무를 이행하지 않을 경우 은행에 막대한 손실을 줄 수도 있는 잠재된 위험 요소이기 때문이다. 따라서 대출 신청 고객의 신용 위험을 측정하여 그 고객의 채무 이행 여부를 정확하게 판단하는 것은 은행 경영에 있어 매우 중요한 부분을 차지한다. 일반적으로 고객이 대출 신청 등록을 하면 은행은 CB (Credit Bureau)에서 제공받은 신용 정보 및 고객 개인의 은행 거래내역 등 심사에 필요한 정보를 수집하여 자체평가시스템인 CSS(Credit Scoring System) 모형을 통해 신용 평점을 계산, 대출 가능 여부(가/부) 및 적용 금리를 결정하여 결과를 통보한다. 이 때 사용되는 CSS 모형에는 은행의 대출 전략, 운용 자산의 정도와 고객의 신용 평가를 위한 요소들이 반영되어 있으나 모든 대상에 대해 일괄적인 기준과 단일한 예측모형이 적용되고 있는 현실이다.[4] 은행의 경우 대출 심사에 있어 리스크를 줄이기 위해 보수적인 기준을 설정, 운용하고 있는데 기존에 설계, 제작된 대출 모형은 주로 과거의 우량한 고객의 데이터를 기반으로 모형이 구축되어 있기 때문에 차후 상환할 능력, 의지가 있음에도 불구하고 현재 구축되어 있는 모형에 적용될 경우 대출 승인 기준에 미달되어 승인이 거절되는 결과가 발생할 수 있다. 이는 은행의 입장에서 향후 수익을 창출할 수 있는 고객을 잃을 수도 있는 있기 때문에 현재의 모형을 보완할 수 있는 정확하고 세분화 된 예측 모형이 필요하다고 할 수 있다. 따라서 본 논문에서는 고객의 특성에 따라 유사한 고객들을 각각의 군집으로 분류한 뒤 분류된 각 군집에 다양한 예측 기법을 적용한 후 예측 정확도를 비교하여 군집별로 정확도가 가장 높은 기법을 찾아 이를 신용 대출에 대한 심사에 적용하는 방안을 제안하는 것을 목적으로 한다. 만약 새로운 고객이 대출을 신청하는 경우, 누적된 고객 정보를 기반으로 군집화를 수행하고 생성된 군집들의 특성을 기준으로 해당 고객을 가장 유사한 군집으로 분류한 뒤 그 군집에 가장 높은 정확도를 보인 예측 모형을 적용하여 심사함으로써 대출 심사 결과의 정확성을 높일 수 있을 것이다. 이와 같은 방법으로 대출 심사의 적용 범위와 정확도를 확장, 향상시켜 은행은 고객들의 채무 불이행으로 발생하는 손실 위험을 줄이는 동시에 은행의 주된 수익 원천이라 할 수 있는 대출을 통한 이자수익을 높여 경쟁력을 키울 수 있을 것으로 기대한다. 본 논문의 구성은 다음과 같다. 먼저 2장에서는 신용 대출과 예측 기법에 대한 선행 연구 및 이론적 배경에 대하여 알아본다. 그리고 제안한 대출 심사 프로세스의 구조 및 설계 모형을 3장에서 보여주며 4장에서는 국내 A은행에서 수집한 대출 데이터를 이용하여 제안한 대출 심사 모형의 성능을 알아보고 결과를 비교, 분석하였다. 마지막으로 5장에서는 결론 및 대출 심사모형의 성능을 향상시킬 수 있는 추후 연구 방향을 제시하였다.

2. 이론적 배경

지금까지 신용예측 및 평가에 관한 연구에서 사용된 예측기법으로는 통계학적 기법, 데이터 마이닝 기법 등이 있다. 먼저 통계학적 기법으로는 초기 신용평가 문제에 사용되어 신용평점 시스템을 구현하는데 사용된 다변량 회귀분석[15], LR(Logistic Regression)분석[12], 판별분석[19,22] 모형 등이 사용되었으며 데이터 마이닝 기법으로는 통계적 가설이 필요 없으면서 비선형적 회귀모형의 설명에 적합한 인공신경망[11,16,17], 최근에 뛰어난 안정성으로 인해 신용예측 분야에서 활용되고 있는 SVM(Support Vector Machine) 기법[10], 군집 분석과 신경망의 혼합[9], 다변량 회귀분석과 인공 신경망의 혼합[24], NN(Neural Network)과 GA(Genetic Algorithm)의 혼합[18] 등 다양한 형태의 데이터 마이닝 기법들이 신용예측 및 평가 분야에서 사용되고 있다. 각각의 예측 모형들은 적용 한 데이터의 특성, 모형의 구현 방법 및 고유의 특징에 따라 성능이 다르게 나타나기 때문에 특정 모델이 우수하다고 단정할 수는 없으나 일반적으로 선형성을 가지지 않는 실제 데이터를 적용했을 경우, 데이터 마이닝 기법을 기반으로 한 모델의 예측 정확도가 비교적 우수하게 나타나는 경향을 보인다. 본 논문에서는 그 중 통계적 기법에서 분류문제를 해결하는데 있어 비교적 좋은 성능을 보여주는 LR 기법과 비선형 데이터를 다루는데 있어 성능이 탁월한 데이터 마이닝 기법의 NN기법과 SVM 기법을 선택하여 대출 심사 모델로 사용하였다. 이에 앞서 종속변수에 유의한 영향을 주는 변수를 선정하기 위해 상관분석 (Correlation Analysis) 및 특징 선택 (Feature Selection) 기법을 사용하였으며 선택한 변수들을 활용하여 2-step 군집화 기법을 통해 데이터들을 비슷한 속성을 가지는 군집으로 분류하였다.

2.1 상관분석 및 특징 선택(Feature Selecti on)

상관분석[5]은 두 개 이상의 변수 사이에 존재하는 밀접함을 측정하는 것으로 두 측정치 사이의 공통적 변이(Joint Variation)를 다루며, 어느 변수도 독립변수나 종속변수로 규정되지 않는다. 변수 관계를 하나의 수치로 나타내는 상관계수 $CC(\rho_{xy})$는 다음 식(1)과 같다.

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y}, \quad -1 \le \rho_{xy} \le 1 \qquad (1)$$

단순 상관관계에 관한 가장 보편적인 측정치는 피어슨 적률상관계수(Pearson Product Correlation Coefficient)로 이는 간격 척도 이상의 수준에서 측정된 데이터로부터 계산될 수 있으며 +1의 상관계수는 완전한 정의 상관관계, -1의 상관계수는 완전한 부의 상관관계를 나타내며 0은 상관관계가 없음을 나타낸다. 피어슨 상관계수 r_{xy}는 식(2)을 통해 도출한다.

$$r_{xy} = \frac{s_x \cdot s_y}{\sqrt{\sum s_x^2 \cdot \sum s_y^2}} \ , \ -1 \le r_{xy} \le 1 \qquad (2)$$

특징 선택 기법[13,25]은 데이터의 입력변수 수가 많을 때 학습을 통해 예측 모형을 구축하기 전 입력 변수의 수를 줄여줌으로써 시간과 메모리 효율성을 높일 수 있으며 noise 변수로 인한 오류를 줄이는데 사용된다. 이 과정은 중요도가 떨어지거나 결측값(missing value)을 많이 포함한 변수, 변동범위가 너무 작거나 큰 변수를 제거하는 Screening 단계, 중요도에 따라 남은 변수들의 순위를 매기는 Ranking 단계, 남은 변수들에 대한 확인 및 반영 여부를 결정하는 Selecting 단계로 이루어진다. 이 때 변수의 중요도는 연속형 변수인 경우 피어슨 카이 스퀘어 값을, 명목 변수인 경우 F 통계량 값을 기준으로 산출된다.

2.2 군집화 (Clustering)

데이터 마이닝 분야에서 대표적으로 사용되는 군집분석 방법[8] 으로는 군집을 순차적으로 병합·분할하는 계층적 군집화, K-means 군집분석으로 대표되는 비계층적 군집화, 신경망 기법의 하나인 SOM(Self Organizing Map), 2-Step 군집화(Two-Step Clustering) 기법으로 나눌 수 있다. K-means 의 경우, 군집의 수 K 값이 분석자의 주관에 의해 결정되며 거리 유사성을 기반으로 군집화하기 때문에 원칙적으로 변수들은 연속적 척도여야 하며 범주형 변수는 더미(Dummy)화 시키는 과정을 필요로 한다. SOM기법의 경우, 범주형 변수로도 군집분석이 가능하지만 기본적으로 승자 독식 전략의 경쟁학습을 사용하기 때문에 유연성은 좋으나 정방행렬의 구성에 따라 매번 결과가 상이하게 나타난다.[6] 이에 반해 2-Step 군집화[21] 방법은 모형 기반 거리 척도를 사용하기 때문에 범주형 변수와 연속형 변수를 모두 사용할 수 있어 범주형 변수를 더미화하는 가공 과정이 불필요하며 군집 개수의 범위를 설정할 수 있어 최적의 군집수를 빠르고 효과적으로 찾을 수 있다. 2-Step 군집화의 1단계는 예비 군집화 단계로 거리를 기준으로 기존의 군집에 병합시킬 것인지 새 군집을 생성할 지 판단하여 순차적으로 개체들을 묶는데 이 때 군집수가 사용자가 지정한 수 보다 커질 경우, 거리 기준을 상향조정하여 군집간 거리가 새 기준에 미달하는 군집들을 병합시켜 총 군집 수를 줄인다. 2단계는 계층적 군집화 시키는 과정으로 유사한 예비 군집들을 병합시켜 몇 개의 군집해를 만든 후 통계적 기준을 적용하여 가장 좋은 군집해를 찾아 출력한다.

2.3 LR(Logistic Regression)

로지스틱 회귀분석(LR: Logistic Regression)[2]은 종속변수가 연속형 변수가 아닌 이항변수 일 때, 즉 입력변수에 대한 그 분포를 수식화하기 위해 일반적인 선형회귀나 다항회귀분석을 사용하기 어려운 경우 사용할 수 있다. 선형 회귀분석이 종속변수와 예측변수 사이의 선형관계를 전제로 하는데 반해 로지스틱 회귀분석(LR)은 비선형의 로지스틱 형태를 취하며 단지 2개의 값을 가지는 종속변수와 예측변수 사이의 인과관계를 밝히는 대표적인 통계기법 중 하나이다. LR은 분산과 공분산의 행렬이 동일해야 한다는 가정에서 자유롭고, 계수의 유의성 검증이 가능하며, 결과값이 0과 1사이의 확률값을 가지므로 수치적 해석이 가능한 장점이 있다. 종속변수 범주가 1일 확률을 p라 할 때, 이를 독립변수 X에 대한 다음의 식(3)의 로지스틱 함수에 적합 시키려면

$$P = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \qquad (3)$$

실제 모형 추정을 위해서 p 대신 식(4)와 같이 로짓(logit) 변환을 통해 선형 형태의 모형을 사용하며

$$\log\ddot{y}(p) = \log\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X \qquad (4)$$

모형의 적합도 평가는 식(5)의 피어슨 카이제곱 테스트를 통해 검정할 수 있다.

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{j}\frac{(n_{ij} - n_i\widehat{\pi}_{ij})^2}{n_i\widehat{\pi}_{ij}} \quad (5)$$

2.4 NN(Neural Network)

신경망 이론(NN:Neural Network)[20]은 인간 의 신경세포인 뉴런의 시스템을 컴퓨터를 이용해 구현한 인공지능 분야의 이론 중 하나로, 패턴인식, 식별 및 분류, 의사결정 등에 널리 응용되고 있는 기법이다. 기존의 계산 기법과는 달리 학습을 통해 추가되거나 변경된 데이터로부터 스스로 규칙을 생성하고 그 결과에 따라 가중치를 갱신하여 모델을 새롭게 구축· 적용시킬 수 있다. 그리고 병렬 연산 방식으로 인해 처리 시간이 빠르고 관련된 데이터들을 연관해서 기억할 수 있어 여러 가지 처리를 한꺼번에 할 수 있다. 신경망은 세 가지 노드 즉, 입력, 은닉, 출력 노드의 구조로 이루어지는데 입력 노드를 $x = (x_1, x_{2,\ldots,} x_i)$, 은닉 노드를 $z = (z_1, z_2, \ldots, z_j)$, 출력 노드를 $y = (y_1, y_2, \ldots, y_k)$라고 하면 신경망 모형은 다음식 (6), (7)과 같이 수식화 할 수 있다.

$$z_j = f_j(a_j + \sum_{n=1}^{i} w_{nj}x_n) \quad (6)$$

$$y_k = f_k(a_k + \sum_{n=1}^{j} w_{nk}z_n) \quad (7)$$

위의 식에서 w_{nj}, w_{nk}는 각각 입력노드와 은닉노드, 은닉노드와 출력노드 간의 연결강도이며 활성화 함수 $f(x)$는 Sigmoid 활성화 함수로 다음식(8)과 같이 나타낼 수 있다.

$$f(x) = \frac{1}{1+e^{-\beta x}} \quad (8)$$

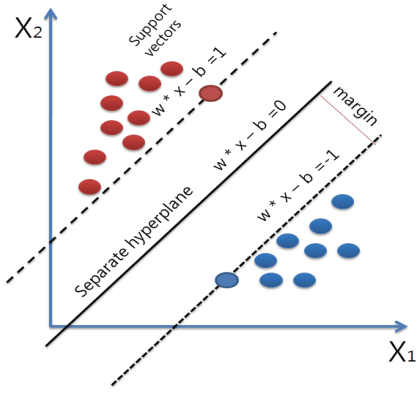
상기 식에서 β는 함수의 기울기 경사를 결정하는 상수이다. 이렇게 구축된 신경망 모형은 학습을 통해 초기 연결강도 값을 데이터에 적합한 값으로 변환하는 과정을 거치게 된다. 활성화 함수를 통과하여 나온 출력값과 목표값 간의 오차를 다음 식(9)와 같이 구할 수 있으며 이러한 오차 e를 최소화 하는 방향으로 반복 학습이 이루어져 최종적으로 오차가 가장 적은 모형을 구축하게 된다.

$$e = \frac{\sum_{n=1}^{k}(y_n - t_n)^2}{2} \quad (9)$$

신경망은 다수의 뉴런으로 연결되어있기 때문에 일부의 오류가 발생했을 때 결과에 치명적인 영향을 미치지 않는 결점 포용성(Fault Tolerance)과 지속적인 학습을 통해 가중치를 변화시키면서 적합한 모델을 만들어 가는 적응성(Adaptability)을 가장 큰 장점으로 갖는다.[7] 또한 통계적 가정을 기반으로 파라미터와 함수 추정을 필요로 했던 기존의 기법들과는 달리 이 과정을 신경망 구조가 대체하면서 입출력 과정 및 구조를 다양하게 설계할 수 있어 정보의 추가나 변경이 용이하다. 신경망은 이러한 장점들을 가지는 반면 데이터에 존재할 수 있는 잡음까지 학습함으로 인해 과적합(overfitting) 문제가 발생될 수 있으며 결과에 대한 추론을 설명하기 어렵다는 단점이 있다.

2.5 SVM(Support Vector Machine)

SVM[3]은 이분 분류를 위해 고안된 기법으로 신경망을 포함한 기존의 분류기법들이 오류율을 최소화하는 것을 목적으로 했던 것과 달리 여백(Margin)을 최대화함으로써 구조적 위험을 최소화하는 것을 목적으로 하며 새로운 데이터에 대한 일반화능력이 우수하다. 그림 1은 2차원 특징 공간(Feature Space)에서 이진 분류를 위한 초평면(Hyperplane)을 보여준다.



[그림 1] Hyperplane

SVM은 각 클래스를 분리하는 초평면 가운데 가장 가까운 데이터 간 거리를 최대화시킬 수 있는 최적의 초평면(Optimal Hyperplane)을 찾으며 초평면은 아래 식(10)과 같이 정의할 수 있으며 각 데이터 x_i에 대해 $y_i \in \{+1, -1\}$의 결과값을 갖는다.

$$y = w^T \bullet x + b = 0 \quad (10)$$

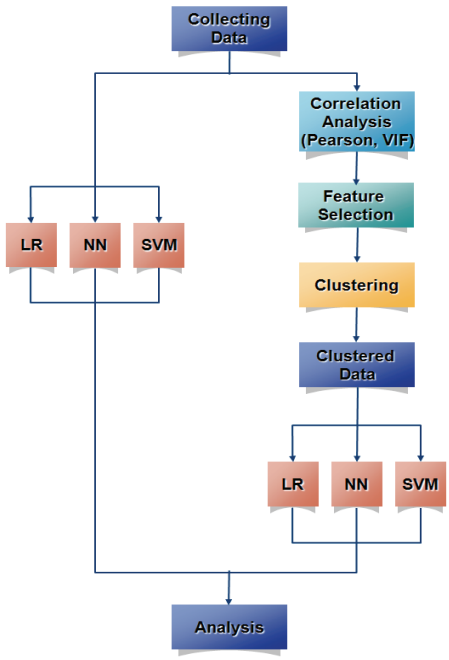
최적의 초평면을 찾기 위해 여백을 공식화하여 목적함수 $h = \frac{2}{\| w \|}$로 만든 뒤 최대화시켜야 하는데 $\frac{1}{\| w \|}$의 최대화는 $\| w \|^2$의 최소화와 같으므로 이를 $J(w) = \frac{1}{2} \| w \|^2$로 바꾸어 최소화 문제로 변형시킨다. 선형 분리가 불가능한 경우 목적함수에 슬랙변수(ζ)와, 패널티 함수(C)를 추가하여 아래 식(11)을 최소화하여 최적화시킨다.

$$J(w,\zeta) = \frac{1}{2} \| w \|^2 + C\sum_{i}^{n} \zeta_i \qquad (11)$$

만약 데이터의 경계가 비선형인 경우, 매핑함수 $\Phi(\bullet)$를 통해 데이터 공간의 입력 벡터의 차원보다 높은 차원으로 변환해줌으로써 선형분류를 가능하게 해준다. 차원 변환의 과정을 하는데 있어서는 커널(kernel) 함수가 필요하며 대표적으로 다항식커널, RBF커널, 하이퍼볼릭 탄젠트 커널이 사용되는데 커널함수의 선택은 SVM의 성능에 중요한 영향을 미친다.

3. 대출심사 모형 설계

본 논문에서는 대출 심사의 정확성을 향상시키기 위해 고객들의 특성에 따라 고객을 다수의 군집으로 분류하고 각각의 군집에 따라 최적의 정확도를 보여주는 예측 기법을 선택하고자 한다. 이를 위해 먼저 종속변수인 대출가능여부(가/부)에 유의한 영향을 미치는 예측변수 추출하는 방법으로는 특징 선택(Feature Selection) 기법을 사용하였다. 그런데 특징 선택 기법을 사용하려면 먼저 예측 변수들 간의 상관 여부를 판별하기 위한 상관 분석과 다중공선성 문제를 해결하는 것이 선행되어야 한다. 이와 같은 방법으로 종속변수에 유의한 영향을 미치는 예측변수를 선별한 뒤, 해당 변수들을 사용하여 2-Step 군집기법을 통해 데이터를 군집화한다. 마지막으로 각 군집에 LR, NN, SVM 예측 모형을 적용하여 군집에 따라 가장 정확도가 높은 모형을 찾아보았다. 또한 기존 대출 심사에 시행되는 방법인 군집분석을 하지 않은 데이터에 대해 각 예측 모형들을 적용한 결과와 본 연구에서 제안한 방법을 통해 도출한 결과를 비교해 보았다. 연구절차는 그림 2와 같다.



[그림 2] 연구 절차

예측 모형을 설계하는데 있어 예측 변수 간 상관관계가 높을 경우, 추정량의 분산이 확대되어 결과값에 좋지 않은 영향을 미칠 수 있다. 따라서 변수들 간의 상관관계 분석을 통해 높은 상관계수를 갖는 변수를 제거함으로써 입력변수의 상관관계에 의한 종속변수에 대한 효과를 줄이고자 하였다. 피어슨 상관계수 등에 의해 계산되는 상관성의 경우 독립변수와 종속변수를 구분 없이 두 변수 간의 상관정도를 측정하게 되는데 이와 더불어 예측변수와 종속변수간의 관계를 모델링 할 때 발생할 수 있는 예측변수 간의 상관관계인 다중공선성(multicollinearity) [14] 문제도 해결되어야 한다. 다중공선성이 존재하는 입력변수와 종속변수의 모델이 구축된 경우 회귀 계수가 불안정해져 예측 변수가 종속변수에 미치는 영향력에 대한 잘못된 결론, 즉 변수 중요성에 대한 판단 오류를 일으킬 수 있다. 따라서 다중공선성 문제를 해결하기 위해 예측변수들의 분산팽창계수(VIF: Variance Inflation Factor) 값을 도출하여 일반적으로 다중공선성이 있다고 여겨지는 기준값인 10 이상이 나오는 예측변수를 제거하고자 하였다. 이와 같은 방법으로 입력 변수간의 상관관계를 제거한 후 특징 선택 기법을 통해 다수의 예측변수들 중 종속변수에 가장 큰 영향을 끼치는 변수를 독립적으로 선별하였다. 이를 통해 데이터를 분류 모형에 적용하여 학습시키기 전 잠음 특징(Noise Feature)들로 인한 부적합한 분류자(Classifier) 생성을 방지할 수 있고, 가장

필수적인 변수만을 예측 모형에 사용함으로써 모형 구축 시 메모리 사용 및 시간 측면에서 효율성을 높일 수 있을 뿐 아니라 분류군의 해석도 용이하게 해준다[3]. 이렇게 선택된 예측 변수들을 기반으로 비슷한 특징을 가지는 군집으로 데이터를 나누기 위해 2-Step 군집화 기법을 사용하였다. 본 논문에서 사용되는 데이터는 예측변수의 경우 범주형 변수와 연속형 변수가 혼합된 형태이고 종속변수는 범주형 변수이다. 앞서 설명한 것처럼 2-Step 군집화 기법은 모형 기반 거리 척도를 사용하기 때문에 범주형 변수와 연속형 변수를 모두 사용할 수 있어 범주형 변수를 더미화하는 가공과정이 불필요하며 군집화를 위한 개체간 거리를 측정하는데 있어 데이터가 범주형 변수를 포함하기 때문에 로그-우도(Log- Likelyhood) 거리를 사용하였다. 이때 연속형 변수의 경우 정규분포, 범주형 변수의 경우 다항분포를 따르는 것으로 가정하였다. 거리 측정을 위해서는 변수간의 상호 독립이 가정되어야 하는 데 본 논문에서는 앞서 상관관계와 다중공선성 분석을 통해 추출된 변수들이므로 이들의 상호 독립성이 만족된 것으로 판단하였다. 최적의 군집수를 결정하기 위해 군집을 실행하기에 앞서 미리 설정한 군집수 범위에서 각각의 군집수마다 산출한 Schwarz's Bayesian Criterion(BIC) 통계량[23]을 바탕으로 최적의 군집수를 결정하게 된다. 이와 같은 방법으로 데이터를 군집화 한 후 각각의 군집에 LR, NN, SVM 예측 모델을 적용하여 산출된 결과값들의 비교를 통해 군집별로 최적의 예측 모델을 제시하였다. 그리고 각각의 군집이 가지는 데이터의 특성을 분석함으로써 차후 새로운 대출 신청 데이터가 들어올 때 가장 바람직한 예측 모델을 제시할 수 있게 하였다. 본 논문에서 사용한 예측 모델 중 인공 신경망(NN)은 각각 한 개의 입력층, 은닉층, 출력층을 가지는 다층 신경망으로 은닉층의 노드 수는 기존의 심사 방법에서 가장 좋은 결과를 보인 4개로 설정하였으며 학습 방법으로는 Back Propagation(BP) 알고리즘을 사용하였다. SVM 모델의 경우 기존의 심사 방법에 적용하여 반복 수행한 결과 커널이 RBF이고 C값이 8인 경우 가장 높은 정확도를 보여 이 값을 제안한 예측 모델에 적용하였다.

4. 실험분석

본 논문에서 제안한 예측 모델의 평가를 위해 국내 A 은행의 실제 대출 심사 데이터를 수집하였는데 2009년 9월부터 10월까지, 총 두 달간 A 은행에 접수된 가계 신용대출에 대한 데이터를 수집하였다. 수집한 데이터는 총 360개이며 이 중 결손 데이터가 있는 46개의 데이터는 제거한 후 실험을 진행하였다. 데이터 정리 및 예측 모델 적용을 위한 전처리 과정은 SPSS 17.0을 사용하였으며 클레멘타인 12(Clementine 12)를 이용하여 앞서 제시한

[표 1] 데이터의 요인, 세부항목 및 비율

요인	세부 항목	비율
성별	1:남 2:녀	1:88% 2:12%
결혼여부	1:Y 2:N	1:90% 2:10%
소득방식	1:급여소득 2:사업소득 3:기타	1:97% 2:1% 3:2%
재직구분	1:정규직 2:계약직 3:용역직 4: 기타	1:96% 2:2% 3:2%
직업분류	1:대기업 2:증기업 3:소기업 4:공기업 및 비영리기관 5:공무원 6:전문직 7:자영업 8:기타	1:49% 2:7% 3:9% 4:9% 5:20% 6:4% 7:0% 8:2%
직무구분	1:관리/사무 2:연구/개발 3:생산/노무 4:영업/판매 5:모집/외판 6:전문직 7:기타	1:55% 2:2% 3:34% 4:7% 5:0% 6:0% 7:2%
주택소유구분	1:자가 2:전세 3:월세	1:67% 2:9% 3:24%
주거상황	1:아파트 2:고급빌라(165m2이상) 3:연립주택(빌라포함) 4:다세대주택 5:단독 6:오피스텔 7:기타	1:78% 2:0% 3:1% 4:4% 5:8% 6:1% 7:8%
취급구분	1:신규 2:채무인수 3:한도증액 4:기대출상환 5:대환 6:기타	1:13% 2:0% 3:1% 4:0% 5:1% 6:85%
자금용도	1:주택자금 2:결혼 3:학자금 4:대출상환 5:부업자금 6:내구소비재구입7:공과금및세금납부 8:개업자금 9:기타	1:4% 2:0% 3:0% 4:2% 5:0% 6:0% 7:0% 8:0% 9:94%
대출상품	1:가계신용 2:공무원가계자금 3:공무원플러스 4:닥터클럽	1:79% 2:13% 3:3% 4:5%
심사	1:가 2:부	1:82% 2:18%
나이(/세)	[24, 58]	Mean:40.93 Std:8.244
근무기간(/개월)	[0, 404]	Mean:163.24 Std:105.970
연소득(/백만원)	[0,146]	Mean:47.93 Std:22.224
신용등급(/등급)	[1, 10]	Mean:5.59 Std:2.543
대출신청횟수(/회)	[1, 7]	Mean:1.41 Std:0.922
신청금액(/십만원)	[17, 880]	Mean:184.29 Std:125.57

예측 기법들을 활용한 모델을 구축하였다. 수집한 데이터의 요인, 세부 항목과 비율에 대한 상세한 설명은 표 1과 같다.

4.1 기존 심사 모형의 실험 결과

일반적으로 은행에서 대출 심사를 위해 수집하는 데이터는 상기 표 1과 같이 다양한 정보를 포함하고 있다. 기존의 예측 기법들에서 통계학적 방법 중 주로 활용되는 LR 모형과 데이터마이닝 기법 중 NN, SVM 모형을 활용하여 대출 심사를 위해 받은 데이터를 기반으로 예측 성능을 알아보고자 하였다. 이를 위해 본 논문에서는 나이부터 대출상품 종류까지의 17개 변수를 예측변수로, 대출 가능 여부(가/부)를 종속변수로 하여 각각의 기법들의 예측 성능을 알아보았으며 총 314개의 데이터 중 학습을 위한 데이터로 188개(60%), 학습 된 모형의 성능 평가를 위한 테스트 데이터로 126개(40%)를 분할하여 사용하였다. 그 결과는 다음 표 2와 같다.

[표 2] 기존 방법의 예측 성능

	Correct	Wrong	Total
LR	92	34	126
	73.01%	26.99%	100%
NN	98	28	126
	77.78%	22.22%	100%
SVM	99	27	126
	78.57%	21.43	100%

LR, NN, SVM 모형에 11개의 명목형 예측변수와 6개의 연속형 예측 변수를 사용하여 대출 가능 여부(가/부)를 실험해 본 결과 LR 모형의 정확도는 73.01%로 77.78%와 78.57%의 정확도를 보인 NN, SVM에 비해 다소 낮게 나타났다.

4.2 제안한 심사 모형의 실험 결과

4.2.1 예측변수 선정

먼저 종속변수인 대출 가능여부에 유의한 영향을 미치는 변수를 선정하기 위해 예측변수간의 상관관계를 분석하였다. 예측 변수간의 상관 관계 분석 결과는 표 3과 같다.

[표 3] 입력변수의 상관관계 분석

변수	나이	성별	결혼여부	소득방식	재직구분	직업분류	직무구분	근무년수	연소득	주택소유	주거상황	신용등급	대출신청횟수	취급구분	자금용도	신청금액	대출상품
나이	1	-.181 (.072)	-.534 (.000)	-.229 (.022)	-.245 (.014)	.053 (.599)	-.200 (.046)	.785 (.000)	.528 (.000)	-.153 (.129)	.070 (.491)	-.151 (.133)	.195 (.052)	.138 (.170)	-.006 (.949)	.147 (.144)	.138 (.172)
성별	-.181 (.072)	1	.185 (.066)	-062 (.537)	-.067 (.507)	.162 (.107)	-.080 (.429)	-.005 (.959)	-.157 (.118)	.007 (.946)	-.045 (.655)	-.074 (.465)	.103 (.306)	-.122 (.227)	.091 (.366)	.033 (.744)	-.043 (.668)
결혼여부	-.534 (.000)	.185 (.066)	1	.394 (.000)	.394 (.000)	.164 (.104)	.250 (.012)	-.385 (.000)	-.392 (.000)	.288 (.004)	.257 (.010)	.107 (.291)	-.113 (.265)	-.159 (.113)	-.075 (.461)	-.012 (.908)	-.148 (.141)
소득방식	-.229 (.022)	-062 (.537)	.394 (.000)	1	.891 (.000)	.436 (.000)	.493 (.000)	-250 (.012)	-.143 (.155)	.293 (.003)	.122 (.227)	.001 (.995)	-.076 (.455)	-.131 (.195)	.042 (.680)	.124 (.219)	.057 (.570).
재직구분	-.245 (.014)	-.067 (.507)	.394 (.000)	.891 (.000)	1	.365 (.000)	.524 (.000)	-.234 (.019)	-.304 (.002)	.261 (.009)	.013 (.901)	.110 (.275)	-.056 (.577)	-.127 (.208)	.045 (.657)	-.114 (.261)	-.081 (.424)
직업분류	053 (.599)	.162 (.107)	.164 (.104)	.436 (.000)	.365 (.000)	1	.077 (.447)	-.037 (.714)	-.053 (.599)	.180 (.073)	-.004 (.972)	-.023 (.823)	.335 (.001)	-.041 (.682)	.128 (.203)	.163 (.104)	.605 (.000)
직무구분	-.200 (.046)	-.080 (.429)	.250 (.012)	.493 (.000)	.524 (.000)	.077 (.447)	1	-.100 (.325)	-.337 (.001)	.089 (.377)	-.079 (.435)	.056 (.582)	-.002 (.985)	-091 (.368)	-.017 (.866)	-.281 (.005)	-.081 (.422)
근무년수	.785 (.000)	-.005 (.959)	-.385 (.000)	-250 (.012)	-.234 (.019)	-.037 (.714)	-.100 (.325)	1	.503 (.000)	-.259 (.009)	-.009 (.930)	-.187 (.063)	.151 (.134)	.074 (.462)	-.063 (.531)	.082 (.418).	.015 (.884)
연소득	.528 (.000)	-.157 (.118)	-.392 (.000)	-.143 (.155)	-.304(.002)	-.053 (.599)	-.337 (.001)	.503 (.000)	1	-.069 (.498)	.137 (.173)	-.234 (.019)	.018 (.848)	.116 (.250)	.008 (.935)	.541 (.000)	.281 (.005)
주택소율	-.153 (.129)	.007 (.946)	.288 (.004)	.293 (.003)	.261 (.009)	.180 (.073)	.089 (.377)	-.259 (.009)	-.069 (.498)	1	.376 (.000)	.063 (.532)	-.088 (.383)	.060 (.553)	.121 (.229)	.121 (.229)	.025 (.801)
주거상황	.070 (.491)	-.045 (.655)	.257 (.010)	.122 (.227)	.013 (.901)	-.004 (.972)	-.079 (.435)	-.009 (.930)	.137 (.173)	.376 (.000)	1	-.003 (.978)	-.159 (.114)	-.041 (.684)	-.001 (.990)	.204 (.041).	-.024 (.809)
신용등급	-.151 (.133)	-.074 (.465)	.107 (.291)	.001 (.995)	.110 (.275)	-.023 (.823)	.056 (.582)	-.187 (.063)	-.234 (.019)	.063 (.532)	-.003 (.978)	1	-.048 (.634)	.258 (.010)	.116 (.251)	-.316 (.001)	-.084 (.411)
대출신청횟수	.195 (.052)	.103 (.306)	-.113 (.265)	-.076 (.455)	-.056 (.577)	.335 (.001)	-.002 (.985)	.151(.134)	.018 (.848)	-.088 (.383)	-.159 (.114)	-.048 (.634)	1	-.191 (.057)	-.287 (.004)	-.090 (.374)	.400 (.000)
취급구분	.138 (.170)	-.122 (.227)	-.159 (.113)	-.131 (.195)	-.127 (.208)	-.041 (.682)	-091 (.368)	.074 (.462)	.116 (.250)	.060 (.553)	-.041 (.684)	.258 (.010)	-.191 (.057)	1	.420 (.000)	.100 (.320)	.104 (.304)
자금용도	-.006 (.949)	.091 (.366)	-.075 (.461)	042 (.680)	.045 (.657)	.128 (.203)	-.017 (.866)	-.063 (.531)	.008 (.935)	.121 (.229)	-.001 (.990)	.116 (.251)	-.287 (.004)	.420 (.000)	1	-130 (.197)	.033 (.745)
신청금액	.147 (.144)	.033 (.744)	-.012 (.908)	.124 (.219)	-.114 (.261)	.163 (.104)	-.281 (.005)	.082 (.418).	.541 (.000)	.121 (.229)	.204 (.041).	-316 (.001)	-.090 (.374)	.100 (.320)	-130 (.197)	1	.339 (.001)
대출상품	138 (.172)	-.043 (.668)	-.148 (.141)	.057 (.570).	-.081 (.424)	.605 (.000)	-.081 (.422)	.015 (.884)	.281 (.005)	.025 (.801)	-.024 (.809)	-.084 (.411)	.400 (.000)	.104 (.304)	.033 (.745)	.339 (.001)	1

일반적으로 피어슨상관계수(Pearson Correlation Coefficient)가 0.70 이상 이면 강한 양의 상관관계가 있다고 판단할 수 있다. 이에 상관계수가 0.07 이상이고 유의확률이 0.05 이하로 결과가 나온 변수들을 제거하였다. 이 과정에서 예측 변수 중 나이, 소득방식 변수가 제거되었다. 다음으로 상관분석을 통해 도출한 예측변수와 종속변수간의 관계를 모델링 할 때 발생 가능한 예측 변수들간의 상관관계인 다중공선성 발생 여부를 검증하기 위해 예측변수들의 분산팽창계수(VIF: Variance Inflation Factor) 값을 도출하였다. 도출된 예측변수별 VIF 값은 표 4에 나타내었다.

[표 4] 다중공선성 진단

입력변수	공선성 통계량	
	공차	VIF
성별	.769	1.301
결혼여부	.571	1.752
재직구분	.451	2.215
직업분류	.342	2.924
직무구분	.612	1.635
근무기간(/개월)	.530	1.888
연소득(/백만원)	.360	2.774
주택소유구분	.713	1.403
주거상황	.725	1.380
신용등급(/등급)	.753	1.328
대출신청횟수	.583	1.716
취급구분	.627	1.596
자금용도	.585	1.709
신청금액	.425	2.354
대출상품	.359	2.784
▪ 종속변수: 심사		

VIF 값이 10 이상이 나오는 예측변수는 다중공선성을 가지기 때문에 이러한 값이 나오는 변수는 제거하고자 하였는데 표 4에서 볼 수 있는 바와 같이 입력한 예측변수들의 경우 VIF 값이 모두 10보다 작게 나와 예측변수간의 상관관계가 적은 것으로 판단하였다. 종속변수에 유의한 영향을 미치는 변수를 선택하는 마지막 단계로 특징 선택 기법을 사용하였다. 특징선택 기법을 통해 유의한 영향을 미치는 변수를 추출한 결과는 표 5와 같다.

[표 5] 특징선택을 통한 변수 추출

Variable	Value	Rank	Selecting
신용등급	1	1	TRUE
대출신청금액	1	2	TRUE
직무구분	1	3	TRUE
대출신청횟수	0.998	4	TRUE
직업분류	0.998	5	TRUE
주택소유구분	0.983	6	TRUE
대출상품	0.954	7	TRUE
연소득	0.953	8	TRUE
근무기간	0.913	9	TRUE
주거상황	0.879	10	FALSE
성별	0.625	11	FALSE
결혼여부	0.182	12	FALSE
재직구분		13	FALSE
취급구분		14	FALSE
자금용도		15	FALSE

특징선택을 통한 변수 추출 결과 신용등급, 대출 신청금액, 직무구분, 대출 신청횟수, 직업분류, 주택소유구분, 대출상품, 연소득 변수는 종속변수에 미치는 영향이 큰 것으로 나타났고, 근무기간 변수는 보통의 영향을 미치는 것으로 나타났다. 반면 하나의 범주에만 개체들이 집중되어 있다는 이유로 취급구분, 자금용도, 재직구분 변수는 제거되었고, 주거상황, 성별, 결혼여부 변수는 종속변수에 미치는 영향력이 떨어져 변수에서 제거되었다. 최종적으로 군집분석에 사용될 수 있는 변수에는 신용등급, 대출 신청금액, 직무구분, 직업분류, 대출 신청횟수, 주택소유구분, 대출상품, 연소득으로 9개의 변수가 선택되었다.

4.2.2 군집화(Clustering)

앞서 종속변수인 대출 가능 여부에 유의한 영향을 미치는 것으로 나온 예측 변수들을 이용해 수집한 데이터들의 군집화를 시행하였다. 군집화는 본 논문에서 사용된 변수가 범주형 변수와 연속형 변수를 모두 포함하고 있음을 감안하여 2-Step 군집화 기법을 사용하였으며 최적의 군집 개수 선정을 위해서는 Schwarz's Bayesian Criterion(BIC) 통계량을 기준값으로 사용하였다. 2-Step 군집화 기법을 사용한 군집결과는 표 6과 같다.

[표 6] 군집결과

	데이터 개수(비율)	데이터 특징
군집 1	63(20.1%)	직업: 공무원(80%) 근무기간: 평균 219개월 연봉: 평균 5천7백20만원
군집 2	66(21.0%)	대출신청 횟수: 평균 1.05회 근무기간: 평균 101개월 주택 소유: 소유자 없음 직업: 불명확
군집 3	97(30.9%)	직업: 공기업 및 비영리기관 재직(67%) 주택소유: 자가 주택 소유(46.3%)
군집 4	88(28.0%)	대출신청금액: 평균 1320만원 직무: 생산/노무직 종사(82.4%) 주택 소유 :전세(79.2%)

총 4개의 군집이 최적의 군집 수로 도출되었으며 각각의 군집에 속한 데이터의 비율은 20.1%, 21.0%, 30.9.%, 28.0% 로 나왔다. 군집 1의 경우 군집의 80%가 공무원이며 근무기간 평균 219개월, 연소득 평균 57000만원으로 전체 데이터의 평균과 비교했을 때 근무기간이 가장 길고 소득 또한 가장 높은 특징을 가진다. 군집 2는 근무 기간이 평균 101 개월로 가장 짧았으며, 직업이 불명확하고 주택 소유자가 없었으나 대출 신청 횟수는 평균 1.05회로 가장 낮은 것이 특징이다. 군집 3은공기업 및 비영리기관에 재직하는 사람이 군집의 67%를 차지하며 전체 평균과 비교했을 때 비교적 고 연봉 소득자에 장기간 근무한 집단인 것으로 파악되며 군집 내 자가 주택 소유율은 46.3% 가장 높게 나타났다. 마지막으로 군집 4는 대출 신청 금액이 평균 1320만원으로 군집들 가운데 가장 소액이었으며 군집의 82.4%가 생산 및 노무직 관련 직종에 종사 중인 것으로 나타났다. 또한 군집의 79.2%는 전세 형태의 주택을 소유한 것으로 파악되었다.

4.2.3 대출 심사 예측 기법별 실험

상기의 과정을 통해 도출된 군집화 결과를 토대로 각각의 군집에 예측 기법을 적용하여 군집별로 정확도의 차이에 대해 알아보았다. 앞서 기존 심사 방법에 사용되었던 기법인 LR, NN, SVM을 활용하여 예측 성능을 알아보았으며 군집별로 도출된 예측 성능은 다음 표 7과 같다.

[표 7] 군집별 예측 성능

Cluster	Method	Correct	Wrong
1	LR	84.00%	16.00%
	NN	88.00%	12.00%
	SVM	**92.00%**	8.00%
2	LR	88.46%	11.53%
	NN	**96.15%**	3.84%
	SVM	92.30%	7.69%
3	LR	74.35%	25.64%
	NN	89.74%	10.25%
	SVM	**92.30%**	7.69%
4	LR	74.28%	25.71%
	NN	**82.85%**	17.14%
	SVM	77.14%	22.85%

상관분석과 특징선택 기법을 통해 선별한 변수를 사용하여 2-step 군집기법으로 산출된 4개의 군집에 각각 LR, NN, SVM 예측 모형을 실행시킨 결과 전반적으로 기존에 사용하던 방법에 비해 예측의 정확도가 향상된 것으로 나타난다. 먼저 군집 1에서는 SVM이 92.00%의 정확도를 보이며 기법들 가운데 가장 좋은 성능을 보였으며 LR과 NN기법 역시 정확도가 많이 향상 되었다. 군집2에서는 NN기법과 SVM 기법의 정확도가 각각 96.15%, 92.30%의 매우 높게 나왔고 LR기법 역시 가장 좋은 성능을 보였다. 군집3에서는 SVM이 92.30% 정확도로 가장 좋은 성능을 보였고 NN기법 역시 정확도가 많이 높아졌으나 LR의 경우 74.35%의 정확도로 기존 방법과 비교하여 별다른 성능 향상을 보이지 않았다. 마지막으로 군집 4의 경우 LR과 NN기법 모두 정확도의 향상이 있긴 했지만 다른 군집들과 비교했을 때 현저하게 떨어졌으며 SVM기법의 경우 기존의 방법으로 예측했을 때보다 오히려 정확도가 감소한 것으로 나타났다.

4.3 기존 심사 방법과 제안한 심사 방법의 예측 성능 비교

지금까지 보다 정확한 예측을 하기 위해 상관분석과 특징선택 기법으로 유의한 변수를 선정하였고 이를 토대로 4개의 최적화된 군집을 생성하였다. 각 군집에 LR, NN, SVM 기법을 적용하였을 때 가장 좋은 성능을 발휘하는 예측 기법과 그 성능은 다음 표 8과 같다.

[표 8] 성능 비교

	제안한 심사 방법				기존 방법
	군집1	군집2	군집3	군집4	
Method	SVM	NN	SVM	NN	SVM
Accuracy	**92.00 %**	**96.15 %**	**92.30 %**	**82.85 %**	78.57 %

LR 모형의 경우, 기존의 심사 방법에서는 73.1% 정확도를 보였던 데 반해 본 논문에서 제안한 방법을 적용했을 시 군집 2의 경우 정확도가 88.46%까지 향상 되었으며 가장 정확도가 낮게 나온 군집4에서도 74.28%로 기존의 방법보다 정확도가 높은 것으로 나타났다. NN 모형의 경우, 기존의 심사방법에서는 77.78%의 정확도를 보였지만 제안 방법의 적용 시 군집 2의 경우 정확도가 96.15%까지 향상된 것을 알 수 있으며 전반적으로 모든 군집에서 성능이 향상된 것으로 나타난다. 기존의 방법에서 SVM은 78.57%로 세 가지 예측 모형 중 가장 높은 정확도를 보였다. 제안한 방법을 적용했을 때 나온 결과 역시 군집 4를 제외하면 모든 군집에서 90% 초반대의 정확률로 우수한 성능을 보였다. 따라서 군집에 따라 예측 모형의 정확도가 일부 감소한 것도 있고 향상 정도가 작은 군집도 있긴 하지만 각 군집에 적용된 모형 중 가장 높은 정확도를 가진 예측 기법을 선택하여 적용 한다면 매우 높은 성능 향상을 보이는 것으로 나타났다.

5. 결론

본 논문은 316명의 대출 신청 고객에 대한 심사결과를 수집하여 상관관계 분석과 특성 선택기법을 통해 심사결과에 유의한 영향을 주는 변수를 선정, 2-Step 군집화 기법을 사용하여 고객의 특성에 따라 4개의 군집으로 분류하였다. 상기 과정을 통해 분류된 군집에 LR, NN, SVM 예측 기법을 적용한 후 정확도를 비교하여 군집별로 정확도가 가장 높은 기법을 찾아 군집 특성에 따른 최적의 모형을 제시하였다. 그 결과 본 논문에서 제안한 방법 심사 결과를 예측하였을 때 예측의 정확도가 상당히 향상됨을 알 수 있다. 이는 은행의 대출 신청 고객에 대한 판단 정확도를 높이는데 기여할 수 있으며 은행은 정확한 판단을 통해 채무 불이행으로 발생하는 손실 위험의 감소, 상환 능력이 있는 고객 유치를 통한 수익 향상을 기대할 수 있다. 본 연구를 수행하는데 있어 데이터 수집의 어려움이 제안한 모델의 성능을 평가하는데 한계점으로 작용하였다. 충분한 데이터 확보가 가능했다면 군집화 과정에서 보다 특징을 확연하게 보여줄 수 있는 군집을 생성하는 것이 가능했을 것이라 판단되며 실험에서 사용한 LR, NN, SVM과 같은 학습기반 모형은 데이터 수가 중요한 영향을 미친다는 점을 감안할 때 데이터가 충분했다면 더 좋은 결과를 도출할 수 있었을 것으로 생각한다.

참고문헌

[1] 금융감독원, 은행경영통계(09년판), Ⅱ-03. 은행별 이자수익·이자비용·비이자손익(은행계정).

[2] 박희정, 강호정, “로지스틱회귀분석을 이용한 코스닥 기업의 부실예측모형 연구“, 한국콘텐츠학회, vol 9(3), pages 305-311, 2009.

[3] 오일석, “패턴인식”, 교보문고, 2008.

[4] 윤성철, 윤명희, 서현석, “국내 소매은행의 개인 신용평가시스템에 대한 연구”, 한국 경영 과학회 학술대회논문집, pages 123-126, 2002.

[5] 윤영선, “상관분석(연구방법9)”, 교육과학사, 2000.

[6] 조용준, 김영화, “요인분석과 군집분석을 통한 세분화 및 전략방향 제시: 특수법인 사례를 중심으로, 응용통계연구”, vol 20(1), pages 23-38, 2007.

[7] 조홍규, “인공지능 방법을 이용한 신용평가 모형에 대한 개관”, 나이스채권평가 금융공학연구소, 2003.

[8] 허명회, 이용구, “데이터 마이닝 모델링과 사례 제2판”, 한나래 출판사, 2008.

[9] Nan-Chen Hsieh, “Hybrid mining approach in the design of credit scoring models, Expert systems with application”, vol 28, pages 655-665, 2005.

[10] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, “Credit scroring with a data mining approach based on support vector machines”, Expert systems with application, vol 33, pages 847-856, 2007.

[11] David West, “Neural network credit scoring models”, Computers & Operations Research, vol 27, pages 1131-1152 2000.

[12] Erkki K. Laitinen, “Predicting a corporate credit analyst’s risk estimate by logistic and linear models”, International Review of Financial Analysis, vol 8(2), pages 97–121, 1999.

[13] Fangming Zhu, Steven Guan, “Feature selection for modular GA-based classification”, Applied Soft Computing, Volume 4(4), pages 381-393 ,2004.

[14] H. Evangelaras, C. Koukouvinos, S. Stylianou, “Evaluation of some non-orthogonal saturated designs with two levels”, Statistics & Probability Letters, vol 74(4), pages 322-329 2005.

[15] James H.M, W.F. Edward, “The development of numerical credit evaluation system, journal of the American statistical association”, vol 58, pages 799-806, 1963.

[16] Mahlhotra R.,& Malhotra, D.K. , “Evaluating consumer loans using neural networks”, OMEGA: The International Journal of Management Science, vol 31(2), pages 83–96, 2003.

[17] MD Odom, R Sharda, “A neural network model for bankruptcy prediction”, Proceedings of IJCNN, vol 11, pages 163-168.

[18] Mu-Chen Chen, Shih-Hsien Huang, “Credit scoring and rejected instances reassigning through evolutionary computation techniques”, Expert systems with application, vol 24, pages 433-441,2003.

[19] P Falbo, “Credit-Scoring by Enlarged Discriminant Models”, Omega, vol 19(4), pages 275-289, 1991.

[20] Philip T. Quinlan, “Structural change and development in real and artificial neural networks”, Neural Networks, vol 11(4), pages 577-599, 1998.

[21] Radha Chitta, M. Narasimha Murty, “Two-level k-means clustering algorithm for k–τ relationship establishment and linear-time classification”, Pattern Recognition, volume 43(3), pages 796-804, 2010.

[22] R. Charles Moyer, “Forecasting financial failure: A Reexamination”, Financial management, vol 4, pages 11-19 , 1977.

[23] Schwarz G, “Estimating dimension of a model”,

annual statistics, vol 6, pages 461-464, 1978.
[24] Tian-Shyug Lee, I-Fei Chen, “A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines”, Expert systems with application, vol 28, pages 743-752, 2005.
[25] Tony Bellotti, Jonathan Crook, “Support vector machines for credit scoring and discovery of significant features”, Expert systems with application, vol 36, pages 3320-3308, 2009.

양 유 영(Yu-Young Yang) **[준회원]**

- 2008년 2월 : 경희대학교 사학과 (문학사)
- 2008년 9월 ~ 현재 : 고려대학교 정보경영공학전문대학원 정보경영공학과 석사과정

<관심분야>
신용예측, 데이터마이닝, 패턴인식

박 상 성(Sang-Sung Park) **[정회원]**

- 2006년 2월 : 고려대학교 산업시스템정보공학과 (공학박사)
- 2006년 5월 ~ 현재 : BK21 사업단 연구교수

<관심분야>
컴퓨터 비전, 패턴인식, 전문가시스템음용, 지식관리

신 영 근(Young-Geun Shin) **[정회원]**

- 2005년 2월 : 고려대학교 산업시스템정보공학과 (공학사)
- 2005년 9월 ~ 현재 : 고려대학교 정보경영공학부 석·박사 통합과정

<관심분야>
패턴인식, 스케쥴링, 인공지능

장 동 식(Dong-Sik Jang) **[정회원]**

- 1979년 2월 : 고려대학교 산업공학과 (공학사)
- 1985년 6월 : 텍사스 주립대학 산업공학과 (공학석사)
- 1988년 12월 : 텍사스 A&M 산업공학과 (공학박사)
- 1989년 3월 ~ 현재 : 고려대학교 정보경영공학부 교수

<관심분야>
Computer Vision, 최적화이론, 컴퓨터 알고리즘