Frequency Matrix 기법을 이용한 결측치 자료로부터의 개인신용예측

배 재 권* · 김 진 화** · 황 국 재***

Predicting Personal Credit Rating with Incomplete Data Sets Using Frequency Matrix technique

Jae Kwon Bae* · Jinhwa Kim** · Kook Jae Hwang***

Abstract

This study suggests a frequency matrix technique to predict personal credit rate more efficiently using incomplete data sets. At first this study test on multiple discriminant analysis and logistic regression analysis for predicting personal credit rate with incomplete data sets. Missing values are predicted with mean imputation method and regression imputation method here. An artificial neural network and frequency matrix technique are also tested on their performance in predicting personal credit rating.

A data set of 8,234 customers in 2004 on personal credit information of Bank A are collected for the test. The performance of frequency matrix technique is compared with that of other methods. The results from the experiments show that the performance of frequency matrix technique is superior to that of all other models such as MDA-mean, Logit-mean, MDA-regression, Logit-regression, and artificial neural networks.

Keywords : Prediction, Personal Credit Rating, Incomplete Data, Imputation Methods, Artificial Neural Networks, Frequency Matrix technique

논문접수일 : 2006년 05월 19일 논문게재확정일 : 2006년 12월 11일

* 교신저자, 서강대학교 경영학과 박사과정, 서울시 마포구 신수동 1번지, Tel: 02-705-8860, e-mail: baejaekwon@sogang.ac.kr
** 서강대학교 경영학과 부교수
*** 서강대학교 경영학과 부교수

1. 서 론

외환위기 이후 경기침체와 개인들의 과도한 신용카드 사용 등으로 가계신용이 타격을 입으면서 금융기관들의 개인고객에 대한 신용예측의 중요성이 부각되고 있다. 따라서 금융기관들은 과학적인 신용예측 모형 개발에 많은 노력을 기울이게 되었다. 학계에서도 효과적인 개인신용예측 모형 개발을 위하여 다양한 통계적 기법과 인공지능기법들을 사용하여 정확도를 향상하는 연구가 여러 각도에서 진행되고 있다. 이러한 연구를 위해서는 무엇보다도 충분하고 정확한 자료가 필수적이다. 그러나 현재의 신용예측 자료는 일반적으로 데이터 양이 충분하지 않고, 그나마 결측치가 많아서 보다 정확한 모형의 개발을 어렵게 하는 요인이 되고 있다.

기존 연구에서는 결측치가 많은 경우 관찰치나 변수를 데이터에서 제거하거나, 다른 의미있는 값(예를 들어 평균이나 중앙값)으로 대체하여 사용하였다. 그러나 관찰치나 변수를 제거하는 경우는 결측치의 비율이 높으면 데이터의 양이 감소하게 되며, 이로 인해 신뢰성이 떨어지는 모형이 개발될 가능성이 높아진다. 즉, 변수 중 한 개의 결측치가 있더라도 데이터에서 관찰치를 제거하거나, 결측치가 많은 변수를 임의적으로 제거하면 데이터가 가진 진정한 정보를 왜곡함으로써 효과적인 모형개발이 어렵다. 또한 결측치를 가진 자료를 모형에 적용하여 개인부도여부를 판단하는 것도 적절하지 않다. 따라서 단순히 자료를 제거하거나 평균값 등으로 대체하는 ad-hoc방법들은 비록 실행이 간단하다는 장점이 있으나, 아주 심각한 오류를 범할 수 있다[Little and Rubin, 1987]. 따라서 본 연구에서는 결측치 데이터를 포함한 모든 데이터의 특성을 이용할 수 있는 매트릭스로 시뮬레이션 구현한 Frequency Matrix(FM) 기법을 제안하고 기존 방법론과의 성능을 비교하고자 한다. 위 모형의 예측성과를 비교하기 위해 본 연구에서는 결측치 대체 방법 중에서 가장 많이 쓰이고 있는 평균값 대체 방법론을 이용한 다변량 판별분석과 로지스틱 회귀분석, 결측치 대체 방법 중에서 회귀식대체 방법론을 이용한 다변량 판별분석과 로지스틱 회귀분석, 인공지능적인 방법으로서 최근 널리 사용되고 있는 인공신경망 모형과 본 연구에서 제시하고자 하는 Frequency Matrix 기법의 6가지 개인신용 예측모형을 제시한다. 이러한 Frequency Matrix 모형의 성과를 증명하기 위해 A 은행이 보유하고 있는 8,234개의 우량, 불량 고객의 데이터를 기초로 분석하였고 그 결과를 통계적 방법과 인공지능 방법 등의 기존 모형과 성과비교를 통하여 그 유용성을 검증해보고자 한다.

본 연구는 다음과 같이 구성되어 있다. 제 2 장에서는 다양한 예측변수의 결측치 처리 알고리즘과 결측치 대체 방법론에 관련된 선행연구, 개인신용평가에 관한 일반적 고찰, 연결빈도행렬과 브레인 매핑에 대한 개념 등을 검토한다. 제 3 장에서는 본 연구에서 사용될 변수선정 작업과 자료분석, 분석절차를 설명한다. 제 4 장에서는 기존의 5가지 모형과 Frequency Matrix 모형을 구축하고 6가지 개인신용예측 모형의 특성을 상호비교 설명한다. 제 5 장에서는 검증데이터의 결과를 모아 기존의 5가지 모형과 Frequency Matrix 모형의 예측력 성과비교를 통하여 그 유용성을 검증해본다. 제 6 장에서는 결론과 향후 연구방향에 대해 논의한다.

2. 이론적 고찰

2.1 예측변수의 결측값 처리 알고리즘

대부분의 자료분석기법에서 결측치를 해결하

기 위하여 활용하고 있는 방식으로는 통계적 대체의 방식과 의사결정나무 모형을 이용한 대체 방식이 있다. 통계적 대체의 방식에는 전통적으로 사용하는 평균값 대체 방식, 변수간의 상관관계를 이용한 회귀분석 방식 등이 있다.

평균대치법(mean imputation)은 각 변수에서 결측치가 포함되어 있는 경우에 각 변수의 완전 자료에서 구한 표본평균으로 대치한 후에 자료가 완전히 관찰되었다는 가정 하에 자료를 분석하는 방법이다. 이 방법은 동일한 대체층 내에서 결측값이 모두 한 개의 값 즉, 평균으로 대체됨으로 인해 관심 변수의 경험적 분포가 상당히 왜곡된다는 단점이 있지만, 사용하기가 쉽고 평균이나 총합들과 같은 일변량 모수에 대한 추정에 있어 결측치 편차들을 감소시키는 데는 상당히 효과적이어서 간단한 점추정이나 예산상의 제약이 있을 때 주로 사용된다.

회귀대체법(regression imputation)은 결측값 대체가 필요한 변수를 종속변수로 하고 일련의 연관 요인들을 설명변수로 하는 회귀 모형을 적용하는 방법으로 결측값 대체에 활용하는 방법이다. 회귀대체법은 각 변수에서 결측치가 포함되어 있는 경우에 관찰된 다른 변수들을 독립변수로 사용하여 결측변수에 관한 회귀분석과 예측치를 대입한 후에 자료가 완전히 관찰되었다는 가정 하에 자료를 분석하는 방법이다. 이 방법은 똑같은 보조변수 값을 갖는 결측값이 다수인 경우 같은 값이 여러 번 사용될 수 있으며 적합된 모형이나 가정이 맞지 않는 경우에는 분포의 왜곡을 초래하는 문제점이 있다. 또한 회귀대체의 경우 표본오차가 과소 추정될 가능성이 있다는 점을 주의해야 한다[Feelders, 1999].

데이터마이닝 관련 알고리즘 중 많은 사용자층을 갖고 있는 의사결정나무에서 결측치를 포함하는 자료를 활용하는 대표적인 방법 중 하나는 CART에서 사용하는 대리분리(Surrogate splits)이다. Breiman et al.[1984]이 제안한 대리분리의 알고리즘은 결측이 있는 변수가 그것과 가장 높은 상관을 가지는 비결측 관찰치에 의한 선형모델로 대체되는 것과 유사한데 선형회귀 등에서 나타날 수 있는 다중 공선성 문제 등을 피할 수 있는 점에서 보다 강건(robust)하다고 볼 수 있다. 또한 의사결정나무 기법의 가장 큰 장점은 모델생성 후 해석이 용이하다는 것이다. 따라서 현장의 자료 분석자들이 선호하는 마이닝 기법 중의 하나로 CART의 대리분리가 널리 사용되고 있다.

Breiman et al.[1984]이 CART에서 대리분리를 주요 기법으로 사용한 반면 Quinlan[1993]의 C4.5에서는 해당변수에 결측치가 있는 개체를 자신의 값을 갖고 있는 다른 개체들이 분류되는 분포를 따라 부분적으로 할당하는 방식을 이용하였다.

통계적기법 또는 경영과학적 기법이 고객에 대한 신용을 점수화하여 평가하는데 비해 의사결정트리의 경우는 고객을 성격에 따라 그룹화하는 방식을 취함으로써 우량고객과 불량고객에 대한 분류를 좀더 알기 쉽게 하고 있기 때문에 신용평가에서 널리 사용되고 있는 기법이다[Carter and Catlett, 1987; Makowski, 1985].

최근에는 결측치 대체에 대한 방법으로 MCMC (Markov Chain Monte Carlo)[Schafer, 1997]나, EM(Expectation and Maximization) [Little and Rubin, 1987] 알고리즘이 널리 이용되고 있으며 결측값이 랜덤하게 발생(MAR : Missing At Random)하고 어떤 분포를 따른다는 가정에 부합되면 결측치 대체를 위한 모수추정에 있어서 정확도가 높은 알고리즘으로 알려져 있다[김상용, 박재인, 2001].

2.2 결측값 대체 방법론 관련 연구

이훈영과 이시환[1999]은 기업부실예측 모형

의 개발에 있어 과학적 결측 처리 방법인 Little and Rubin[1987]의 MI(Multiple Imputation) 방법과 Markov Chain Monte Carlo를 이용한 Schafer[1997]의 MI 방법의 우수성을 기존의 결측치 처리 방법과 비교하여 분석하였다. 연구자들은 결측치를 가진 관찰치를 데이터에서 제거하는 방법, 전체 평균으로 대체하는 방법, 전체 중앙값으로 대체하는 방법, EM 알고리즘을 사용한 MI, DA 알고리즘을 사용한 MI 방법으로 결측치를 처리한 후 모형을 개발하여 모형의 예측정확도를 비교하였다. 정확도 비교 결과 DA 알고리즘을 사용한 MI 방법, EM 알고리즘을 사용한 MI 방법이 관찰치를 제거하거나 다른 의미 있는 값으로 대체하는 방법보다 예측정확도가 높게 나타났다. 그러나 여러 방법들 간의 차이에 대한 T-검정을 실시하였으나 통계적으로 유의한 결과는 얻지 못하였다.

신형원과 손소영[2001]은 하나의 관측치에서 하나의 결측이 발생하는 경우 최빈 범주법(modal category method), 로지스틱 회귀분석(logistic regression), 연관규칙(association rule), 투표융합(voting scheme based fusion algorithm), 신경망 융합(neural network based fusion algorithm) 등 다섯 가지의 방법을 이용하여 범주형 결측 데이터의 다양한 특성에 따른 결측치 추정 성능을 비교하였다. 그들은 범주형 자료의 특성을 입출력 변수간 연결함수, 데이터의 크기, 노이즈의 크기, 결측치의 비율, 결측발생 형태로 나누고 각 시나리오별로 위의 다섯 가지 방법을 이용하여 결측치를 추정하고 결측치를 추정한 상태에서의 검증용 데이터에 대한 분류정확성의 차이를 오분류율을 바탕으로 분석하였다. 연구결과, 입출력 변수간의 함수가 선형일 경우 신경망 융합이 다른 네 가지 방법에 비하여 높은 성능을 나타냈으며 나머지 네 가지 방법간에는 유의한 차이가 없었다. 그리고 데이터의 크기가 작고 결측치의 비율이 높으면 로지스틱 회귀분석은 다른 네 가지 방법에 비하여 예측성능이 떨어짐을 보였다. 또한 데이터의 크기가 작고 Y 변수의 결측 발생 확률이 X 변수들과 강한 함수관계를 가질 때는 신경망 융합과 로지스틱 회귀분석이 높은 성능을 보였으며 데이터의 크기가 크고 노이즈가 크면서 결측치 비율은 작고 결측 발생 상관관계가 강하면 신경망 융합이 가장 우수한 성능을 보였다.

Parthasarathy and Aggarwal[2003]은 결측치의 문제를 해결하기 위하여 데이터마이닝 알고리즘을 사용하는 개념적 복원(conceptual reconstruction)의 방법을 소개하였다. 이 방법의 장점은 원래의 자료구조 차원에서 결측치를 복원하는 대신 결측치 자료의 개념을 자료들 사이의 상관관계 구조를 사용하여 복원한다는 점이다. 따라서 복원절차는 결측치를 갖고 있는 데이터군에서 무리하게 결측치 자체를 추측하기보다는 결측자료의 개념적인 측면만을 예측한다. 연구자들은 이러한 방법이 효과적이라는 것을 여러 가지 경우의 실제 자료를 사용하여 보여주었다.

Strike et al.[2001]는 결측치가 있는 자료를 사용하여 소프트웨어 원가를 추측하는 경우 결측치를 다루는 세 가지 방법을 비교하였다. 그들이 비교한 방법은 listwise deletion, 평균값 대체, 그리고 여덟 가지의 핫 덱 대체(hot-deck imputation)이다. 연구결과, 모든 방법이 비교적 편이가 적고 높은 정확도를 보여 주었다. 이는 세 가지 방법 중에서 가장 단순한 listwise deletion 방법이 최선의 선택이라는 것을 제시한다. 그러나 지속적인 최상의 성과는 유클리디안 거리와 Z-score 표준화를 사용하는 핫 덱 대체 방법이라고 결론을 맺고 있다.

Schoenberg and Arminger[1988]는 결측치를 갖는 데이터를 가우스 언어를 사용한 MISS 프

로그램을 사용하여 분석한 결과를 다음과 같이 보고하였다. 그들에 의하면 결측치 분석을 위하여 MISS는 EM 알고리즘을 활용하여 공분산 행렬의 최대우도 추정치와 평균 벡터를 구하여 준다. 또한 결측치를 대신한 추정치를 포함하는 새로운 데이터를 생산한다. 연구결과, 결국 MISS는 다중대체방법을 사용하는 것과 동일한 결과를 나타내었다.

Raghunathan[2004]는 결측치 예측의 중요성을 강조하였으며 결측치를 무시한 연구는 편이로 인하여 무의미한 결과를 도출한다고 하였다. 그는 결측치를 다루는 다양한 세 가지의 방법을 그 복잡성 순서에 따라 소개하였다. 첫 번째 방법은 결측치를 이유로 표본에서 제외된 자료를 고려하기 위하여 표본에 포함된 데이터에 가중치를 부여하는 방법이다. 두 번째 방법은 multiple imputation 방법을 사용하여 결측치를 둘 혹은 그 이상의 값으로 대체하는 것이다. 마지막으로 불완전한 데이터에 기초하여 최빈값을 구하는 방법이다.

Wang[2003]은 데이터마이닝에서 유용하게 사용되는 self-organizing maps(SOM)이 결측치를 처리할 수 없다는 문제점을 개선하기 위해서 새로운 방법인 SOM에 기반한 퍼지맵 모형을 제안하였다. 이 방법에 의하면 결측치가 퍼지 데이터로 전환되어 퍼지 관찰치를 구하는데 활용된다. 이러한 퍼지 관찰치는 결측치가 없는 완전한 관찰치와 함께 SOM이 퍼지 맵을 구하도록 만든다. 이 방법은 기존의 SOM 접근법에 비해 더 많은 정보를 제공한다.

Orre et al.[2005]은 수많은 결측치를 갖고 있는 세계보건기구의 약물 부작용자료를 데이터로 사용하여 인공신경망의 우수성을 보여 주었다. 인공신경망의 우수성은 기존의 알려진 방법인 AutoClass 방법과의 비교를 통하여 보여주었다. AutoClass는 가우시안의 혼합을 이용한 데이터를 모델링하는 Bayesian군 기법으로 신경망과는 다른 무감독 분류 기법이다. 두 방법의 성능은 시뮬레이션된 자료를 사용하여 분석하였고 분석결과 인공신경망은 AutoClass 방법과 대등하였으며 특히 실제 데이터의 경우 더욱 우수하였다. 인공신경망은 우수한 스케일링 성능으로 결측치를 갖고 있는 대용량의 데이터 분석에 매우 유용한 도구로서 사용될 수 있다고 결론을 맺고 있다.

2.3 개인신용평가의 일반적 고찰

신용이란 경제활동의 주체에 따라 여러 가지로 정의될 수 있으나 일반적으로 장래의 어느 시점에서 그 대가를 지급할 것을 약속하고 현재의 경제적 가치를 획득할 수 있는 능력이라고 정의할 수 있다. 금융기관은 필연적으로 이러한 능력을 갖춘 자에게 자금을 안정적으로 운용하고 신용공여를 함으로써 국민경제의 건전한 발전은 물론 금융기관 자신의 건전한 경영과 수익성을 도모하고자 하는 것이 필연적인 사실이다. 이런 관점에서 금융기관은 개인신용대출 대상의 건전성, 자금용도의 타당성 및 상환능력에 관한 사항들을 정확하게 파악한 후 여신의 가부 판단과 신용대출 가부를 결정해야 하는데 이를 신용평가라 한다[Robert, 1980]. 이러한 신용평가는 크게 기업신용평가와 개인신용평가로 나누어질 수 있는데 본 논문에서 다루고자 하는 것은 개인신용평가에 대한 부분이다. 개인신용평가는 소비자 금융거래에 있어서 개인의 신용도를 종합적으로 평가하는 것으로서 거래자의 상환의사(willing to pay)와 상환능력(ability to pay)을 판단할 수 있는 신용정보에 대한 조사분석 및 평가를 의미한다. 개인신용평가는 일반적으로 개인의 신용정도를 신용평점 제도(Credit scoring system) 등을 이용하여 계량화해서

신용도에 따라 금리 기간 한도 등 신용공여 조건을 차등화하고 자금을 합리적으로 배분하는 공정한 평가 기준과 전문능력이 있는 평가자의 객관적이고 정확한 판단이 부가되는 새로운 가치 창조 활동이라고 말할 수 있다[이명식, 1992].

전통적으로 개인신용평가를 위한 기법으로는 로지스틱 회귀분석이나 프로빗 분석법과 같은 통계학적 기법[Ohlson, 1980; Wiginton, 1980]과 선형계획법[Mangasarian, 1965]과 같은 경영과학적 기법을 들 수 있다.

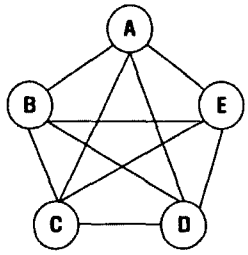
한편 1980년대 중반부터 경영분야에서 널리 사용되기 시작한 인공신경망 모형은 통계적 가설이 필요 없으면서도 비선형적인 회귀모형을 설명하기에 적당하기 때문에 개인신용평가에 널리 사용되어 뛰어난 성과를 보여주고 있다[Altman et al. 1994; Desai et al. 1997].

이와 같이 다양한 기법들이 제안되고 학술적인 성과를 보이고 있음에도 불구하고 어떤 기법이 가장 뛰어난 기법인지 절대적인 판단을 내릴 수는 없다. 또한 결측치가 있는 개인신용 데이터를 활용하여 보다 효과적인 개인신용예측을 위한 방법론들이 제안되어야 한다.

2.4 연결빈도행렬과 브레인 매핑

연결빈도행렬(Connection Frequency Matrix)은 인접 행렬의 개념에서 출발한다. 인접 행렬(Adjacency Matrix)은 데이터의 인접성을 이용하여 의사결정공간에서 유용하게 쓰일 수 있는 개념[김진화 외 2인, 2004]으로 품목 A와 B가 존재할 때 품목 A와 B가 동시에 구매되었는지, 또는 품목 A가 B의 구매에 영향을 주었는지 그 여부를 확인할 수 있어 데이터마이닝의 기법 중 연관규칙 분석[Berry and Linoff, 1997]이나, 또는 추천 시스템[Schafer, 1997] 및 데이터 시각화[Condon et al. 2002] 등에서 이용되고 있다.

이와 같은 데이터의 인접성은 유한개의 점과 선으로 구성된 도형인 연결그래프(Connected Graph)와 사상(Mapping)의 개념이나 도형의 위상적 성질을 이용하여 알 수 있다. 연결그래프에서 단위 정보를 표현하는 점을 "Vertex", 각 점을 잇는 선을 "Edge" 라고 하고 Edge에 방향성이 있는가에 따라서 유향 그래프(Directed Graph)와 무향 그래프(Undirected Graph)로 구분한다. 그 밖에 그래프 내에서 여러 Vertex 들의 연결과정을 경로(Path)라고 하며, 시작점과 끝점이 연결된 경로는 특별히 순환(Cycle)이라고 한다. 연결그래프에서 선(Edge)이 점(Vertex)을 공유하고 있다면 1, 공유하고 있지 않으면 0으로 나타낸 행렬을 인접행렬이라고 한다. 인접행렬은 n개의 Vertex를 가지고 있는 $n \times n$ 정방행렬이다. 인접행렬의 어떤 원소 $A_{ij} = 1$이면 두 Vertex가 인접해 있다는 것이며, $A_{ij} = 0$이면 두 Vertex는 인접해 있지 않은 것이다.



	A	B	C	D	E
A	0	1	1	1	1
B	1	0	1	1	1
C	1	1	0	1	1
D	1	1	1	0	1
E	1	1	1	1	0

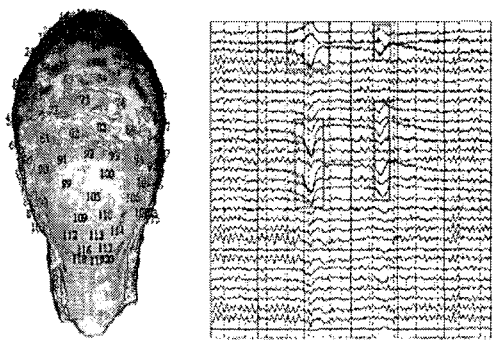
〈그림 1〉 연결그래프와 인접행렬의 예

<그림 1>은 연결그래프와 이에 대한 인접행렬의 한 예 이다. 연결빈도행렬은 이러한 인접행렬의 특성에 방향과 누적빈도를 추가한 개념이다.

연결빈도행렬의 알고리즘은 브레인 매핑(BM : Brain Mapping)에 기원하고 있다. 한번에 하나의 명령을 정보로 변환하고 이 정보에 기초하여 다음의 과정을 결정하고 한번에 하나의 정보를 처리하는 컴퓨터의 직렬 정보처리 방식과는

다른 인간의 정보처리 과정을 인공지능에 많은 응용을 하고 있다. 인간의 기능적 브레인 매핑(HFBM : Human Functional Brain Mapping) 개념은 1928년 Hans Berger가 최초로 뇌의 활동을 시각화하기 위해 두피의 다른 지점 사이의 전기 활동성을 측정하는 데에서 시작하였으며 측정결과는 뇌파도, 뇌전도라 한다[Law et al. 1991]. 의학에서의 기능적 브레인 매핑에 대한 연구는 관련 학문들과의 학제적 연구를 통해 다양하게 연구되고 있다.

Fox 등[2005]은 적절한 실험디자인을 통한 브레인 맵을 이용해 데이터 필터링에 응용할 수 있음을 보였으며 또한 브레인 맵 분류방법을 통해서 데이터베이스 구축을 위한 메타데이터 스키마를 만드는 데에도 유용함을 보였다. 또한 Law et al.[1991]은 전자 두뇌 그림(EEG : electroencephalogram)의 사용을 통해서 두뇌가 무엇을 하고 있는지를 시각화하는 방법을 통해서 확장 발전시켰는데 EEG의 결과를 보면 인간의 뇌는 끊임없는 전기적 활동을 하고 있으며 이러한 활동의 결과는 뇌에 존재하는 수천 개 뉴런의 활동의 결과로 기록할수 있다. 뉴런의 활동 패턴은 인간의 심리 상태에 따라 빠르고 느리게 나타나며 이러한 패턴을 단순한 시각화로 <그림 2>와 같은 패턴 특징으로 나타낼 수 있다.



〈그림 2〉 뉴런의 활동 패턴과 뇌에서의 기억장소

3. 연구 방법

3.1 자료수집과 사전처리

개인신용예측모형을 구축하기 위해 본 연구에서는 A 은행이 보유하고 있는 2004년 개인신용 관련데이터를 이용하였다. 데이터는 총 28개 변수와 8,234개의 우량, 불량 고객으로 분류된 데이터이다.

분석에 사용할 변수들은 모두 [0, 1] 사이의 값을 가지도록 단위를 조정하였다. 이와 같이 자료를 정규화(normalize)하게 되면 분석에 사용되는 모든 변수의 분산이 동일한 범위 내에 있게 되므로 측정 단위에 따른 예측오차를 줄일 수 있게 된다[Peel et al. 1986]. 분석에 사용될 변수의 선정으로는 처음에 고려된 28개의 변수 중에서 개별 독립 t검정을 거쳐 1차로 변수 12개(alpha = 0.05)를 선정하였고 단계별 로지스틱 회귀분석을 거쳐 2차로 선정된 변수 4개를 최종분석에 사용할 변수로 선정하였다. 선정된 변수 4개 항목과 산출식을 <표 1>로 나타내었다.

〈표 1〉 선정된 변수 목록

변 수 명	설 명
av_outcome6	은행측 지급액 6개월평잔액
sumdangcheilsu	당해연체 일수합
CASH_AMT	현금서비스 이용액
TA_TOT_AMT	타사총이용액

모든 분석은 훈련용과 검증용의 두 가지 데이터 셋으로 구성되었으며 전체 데이터의 70%(5,764/8,234)는 훈련용 데이터 셋으로 사용하고, 나머지 30%(2,470/8,234)는 검증용 데이터 셋으로 사용하였다. 보다 일반화된 연구결과를 얻기 위하여 본 연구에서는 상호검증방법(cross-val-

idation method)을 사용하였다[Weiss and Kulikowski, 1991]. 따라서 본 연구에서는 총 10회에 걸친 상호검증방법을 실시하였다.

3.2 자료 분석

개인 결측치 자료는 두 가지의 경우로 나누어 볼 수 있는데 하나는 예측변수(독립변수)에 결측치가 포함된 자료이고, 다른 하나는 목표변수(종속변수)에 결측치가 포함된 자료이다. 본 논문에서는 예측변수(독립변수)에 결측치가 포함된 자료에 해당된다.

개인신용예측 데이터에서 결측값의 비율은 93.1%(7,666/8,234)이다. 본 연구의 결측치 데이터는 한 개의 예측변수가 결측치를 가지고 있는 경우와 두 예측변수가 동시에 결측치를 갖는 경우의 두 가지로 구성되어 있다.

〈표 2〉 결측변수별 구성비율

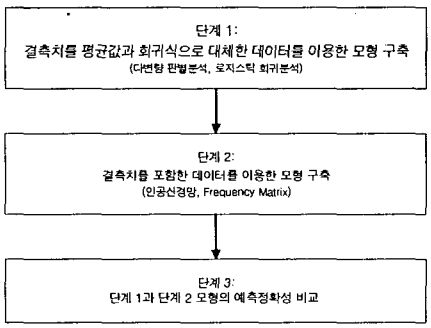
결 측 치 변 수	결측개수	결측비율 (%)
X1 (은행측 지급액 6개월평잔액)	411/8,234	4.99%
X2 (당해연체 일수합)	2,724/8,234	33.08%
X3 (현금서비스 이용액)	1,596/8,234	19.38%
X4 (타사총이용액)	2,935/8,234	35.64%
X1 & X2	336/8,234	4.08%
X1 & X3	3/8,234	0.04%
X1 & X4	28/8,234	0.34%
X2 & X3	1,567/8,234	19.03%
X2 & X4	2,366/8,234	28.73%
X3 & X4	7/8,234	0.09%
총 결측비율	7,666/8,234	93.10%

<표 2>에서 보는 바와 같이 한 개의 예측변수가 결측치를 갖는 경우를 살펴보면 X1이 결측치를 갖는 경우는 전체 8,234개 데이터 중에서 4.99%, X2가 결측치를 갖는 경우는 33.08%, X3가 결측치를 갖는 경우는 19.38%, X4가 결측치를 갖는 경우는 35.64%이다. 두 예측변수가 동시에 결측치를 갖는 경우를 살펴보면 X1과 X2가 동시에 결측치를 갖는 경우는 4.08%, X1과 X3가 동시에 결측치를 갖는 경우는 0.04%, X1과 X4가 동시에 결측치를 갖는 경우는 0.34%, X2와 X3가 동시에 결측치를 갖는 경우는 19.03%, X2와 X4가 동시에 결측치를 갖는 경우는 28.73%, X3와 X4가 동시에 결측치를 갖는 경우는 0.09%이다. 따라서 총 결측비율은 전체 8,234개 데이터 중에서 7,666개인 93.1%에 해당한다.

3.3 분석절차

본 연구의 진행절차는 <그림 3>과 같은 분석절차에 따라 진행된다.

단계 1에서는 결측치를 평균값과 회귀식으로 대체한 데이터를 이용하여 MDA-평균값대체, Logit-평균값대체, MDA-회귀식대체, Logit-회귀식대체의 4가지 개인신용예측모형을 구축한다. 단계 2에서는 결측치를 대체하지 않은 상태에서의 데이터를 이용하여 인공신경망 모형과 Frequency Matrix 모형의 2가지 개인신용예측모형을 구축한다. 단계 3에서는 단계 1과 단계 2에서 구축한 6가지 개인신용예측모형을 가지고 각 모형의 예측정확성을 비교한다.



〈그림 3〉 분석절차

4. 연구 모형

4.1 결측치 대체한 데이터를 이용한 모형 구축

(1) MDA - 평균값대체

MDA-평균값대체는 결측치를 평균대치법을 이용하여 채운 후에 이것을 다변량 판별분석 모형을 이용하여 개인신용예측모형의 예측력을 검증하는 방법론이다. 평균대치법은 각 변수에서 결측치가 포함되어 있는 경우에는 각 변수의 완전자료에서 구한 표본평균으로 대치한 후에 자료가 완전히 관찰되었다는 가정 하에 자료를 분석하는 방법이다.

다변량 판별분석은 사전에 정해진 집단(본 연구에서는 우량고객과 불량고객)을 가장 잘 판별해내는 선형판별함수를 도출하기 위한 통계적 기법이다. 선형판별함수는 집단내 분산대비 집단간 분산비율을 최대로 하는 통계적 의사결정 규칙을 생성하게 되는데, 이를 식으로 나타내면 식 (1)과 같은 형태이다[박정민외 2 인, 2005].

$$Z = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n \qquad (1)$$

여기서 얻은 Z값을 판별점수라 하고, $\omega_i (i=1, 2, \cdots, n)$는 판별함수의 계수로서, 두 그룹을 가장 잘 구분할 수 있도록 판별분석 과정에서 추정되었다. 이러한 계수들의 크기와 부호는 두 그룹으로 분류하는 과정에서 측정변수들이 기여하고 있는 정도와 어떠한 방향을 가지고 있는지를 파악할 수 있게 만들어 준다. 즉, 판별점수인 Z값이 일정한 판별점(Cut-off point)을 넘으면 우량고객으로, 판별점 이하인 경우에는 불량고객으로 판단한다.

다변량 판별분석은 독립변수가 다변량 정규분포를 따르고, 각 집단의 공분산행렬이 동일할 때는 유용하지만 불량고객의 경우 정규성에 대한 가정이 위배되는 경우가 많고, 집단별 공분산이 동일하다는 가정도 위배되는 경우가 많다. 특히, 독립변수간에 다중공선성이 존재할 경우 단계별 분석을 적용하게 되면 심각한 오류가 발생할 가능성이 높다[Hair et al. 1995].

본 연구에서는 다변량 판별분석을 위해 SPSS 12.0 프로그램을 사용하여 수행하였다.

(2) Logit - 평균값대체

Logit-평균값대체는 결측치를 평균대치법을 이용하여 채운 후에 이것을 로지스틱 회귀분석 모형을 이용하여 개인신용예측모형의 예측력을 검증하는 방법론이다.

로지스틱 회귀분석은 비선형의 로지스틱 형태를 취하며 단지 2개의 값을 가지는 종속변수(우량고객, 불량고객)와 독립변수 사이의 인과관계를 밝히는 통계기법이다[Ohlson, 1980]. 로지스틱 회귀분석을 불량고객예측에 사용할 경우 개인고객 설명변수의 관찰치벡터를 X_i로 하고, 그 계수 β_i를 추정한다면 불량고객 확률은 로지스틱 함수에 의해 식 (2)와 같이 유도된다.

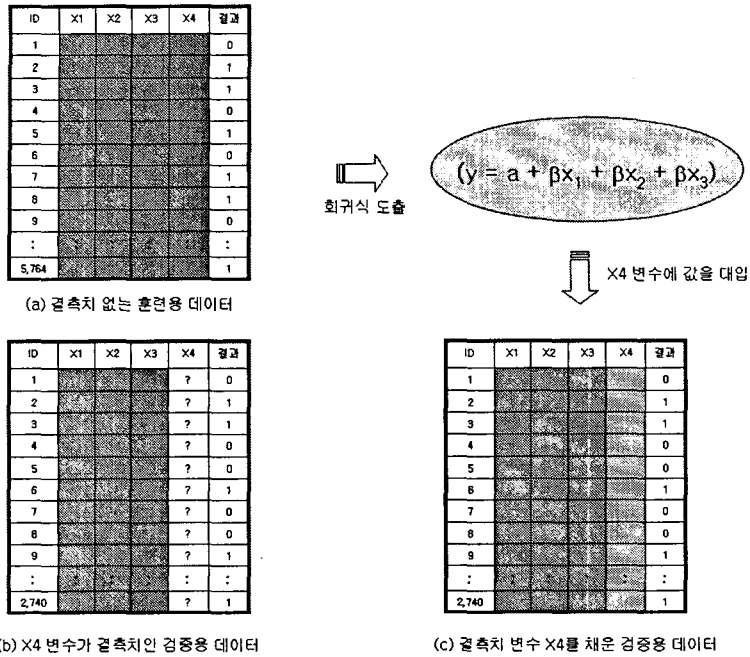
$$Y_i = 1/1 + exp(-p) \qquad (2)$$

여기서, $P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_i X_i$ 이다.

본 연구에서는 로지스틱 회귀분석 모형 구축을 위해 SPSS 12.0 프로그램을 사용하였고, 판별점은 0.5를 기준으로 설정하였다.

(3) MDA - 회귀식대체

MDA-회귀식대체는 훈련용 데이터로부터 회귀식을 도출한 후 검증용 데이터에 회귀식을 대입하여 결측치를 채운 후에 다변량 판별분석 모형을 이용하여 개인신용예측모형의 예측력을 검증하는 방법론이다.



(a) 결측치 없는 훈련용 데이터

(b) X4 변수가 결측치인 검증용 데이터

(c) 결측치 변수 X4를 채운 검증용 데이터

〈그림 4〉 하나의 변수가 결측된 데이터를 회귀식을 도출하여 대체하는 방법

<그림 4>에서 보는 바와 같이 (b)의 검증용 데이터 X4 변수가 결측치를 가지고 있는 경우에 회귀식대체 방법론을 이용하여 결측치를 대체하고자 할 때에는 (a)의 결측치 없는 훈련용 데이터에서 회귀식을 도출하여 검증용 데이터에 회귀식을 대입한다. 검증용 데이터의 X4 변수를 대체한 후, 다른 결측형태의 변수들도 이와 같은 방식으로 모두 대체한다. 검증용 데이터의 모든 결측치를 대체한 상태에서 다변량 판별분석을 이용하여 개인신용예측모형을 구축한다.

(4) Logit - 회귀식대체

Logit-회귀식대체는 MDA-회귀식대체와 마찬가지로 결측치를 회귀식으로 대체한 상태에서 로지스틱 회귀분석 모형을 이용하여 개인신용예측모형을 구축하는 방법론이다. 결측치를 회귀식으로 대체하는 방법은 <그림 4>와 같은 방식이다.

본 연구에서는 로지스틱 회귀분석 모형 구축을 위해 SPSS 12.0 프로그램을 사용하였고, 판별점은 0.5를 기준으로 설정하였다.

4.2 결측치 포함한 데이터를 이용한 모형구축

(1) 인공신경망 모형

인공지능의 한 분야인 인공신경망은 인간두뇌의 휴리스틱한 문제해결방법을 모형화한 것으로서 그 학습능력과 추론능력이 매우 뛰어난 것으로 알려져 있다. 인공신경망의 구조는 여러 가지 형태가 있으나, 가장 일반적으로 많이 쓰이는 형태는 관리학습(supervised learning)에 알맞은 다층 전향구조(multi-layered feedforward) 인공신경망이다. 이는 입력층, 은닉층, 출력층의 삼층구조를 이루며 각 층마다 다수의 뉴런 또는 노드, 즉 처리단위를 소유하고 있다. 서로 다른 층에 존재하는 처리단위는 서로 연결되어 있으며 그러한 연결강도는 연결가중치(int-

erconnection weight)로 계산된다.

일반적으로 신경망의 성능에 영향을 미치는 요인으로는 은닉층 수, 은닉노드 수, 학습 회수 등이 있는데, 어떤 값이 최적인지에 대한 일반적인 규칙은 없다. 다만, 분류 문제를 포함한 대부분의 문제에서 한 개의 은닉층으로도 만족할 만한 결과를 얻을 수 있다는 선행 연구[Hornik, 1991]를 토대로 본 연구에서도 은닉층이 하나인 3층 퍼셉트론을 사용하였다.

개인신용예측에 대한 인공신경망 구축과 평가를 위해 Clementine 8.1 프로그램을 사용하였으며 데이터 셋은 훈련용과 시험용, 그리고 검증용으로 구분하였다. 전체 데이터(8,234개)의 40%는 훈련용, 30%는 시험용, 30%는 검증용의 용도로 사용하여 분류 예측 정확도를 분석하였다. 인공신경망의 경우, 학습과정에서 과대적합이 발생할 가능성이 크기 때문에 시험용 데이터를 사용하여 학습과정이 적절히 이루어졌는가를 확인하게 된다.

(2) Frequency Matrix

Frequency Matrix 기법은 결측치 자료를 포함한 모든 자료의 빈도 특성을 이용한 매트릭스 구현 방법이다. Frequency Matrix 기법은 선행연구에서 언급한 연결빈도행렬의 인접행렬특성과 브레인 매핑의 데이터 필터링 알고리즘을 따르고 있다. Frequency Matrix 방법론을 그림으로 구체화하면 <그림 5>와 같은 n×n 매트릭스로 나타낼 수 있다.

<그림 5>에서 보는 바와 같이 모든 데이터의 값을 0과 10사이의 11구간으로 나눈 다음 우량고객과 불량고객 매트릭스로 나누어 빈도를 계산한다. (b)와 (c)에서 보는 바와 같이 훈련용 데이터의 전체 데이터를 우량고객 매트릭스와 불량고객 매트릭스로 나눈 다음에 검증용 데이터에서 해당 변수의 구간에 대한 전체 우량, 불량 매트릭스의 빈도수를 각각 대입하여 각 레코드의 우량값과 불량값을 산출한다.

ID	X1	X2	X3	X4	결과
1	8	1	2	6	0
2	2	8	9	6	1
:	:	:	:	:	:
5,764	1	9	7	6	1

(a) Data-Set 1 훈련용 데이터

	X1	X2	X3	X4	결과
0					0
1					0
2					0
3					0
4					0
5					0
6					0
7					0
8					0
9					0
10					0

1번 레코드(우량고객)

	X1	X2	X3	X4	결과
0					1
1					1
2	1				1
3					1
4					1
5					1
6				1	1
7					1
8		1			1
9			1		1
10					1

2번 레코드(불량고객)

	X1	X2	X3	X4	결과
0	0	0	0	0	0
1	0	0	1,229	0	0
2	11	0	379	0	0
3	15	0	69	0	0
4	22	123	39	1910	0
5	980	42	35	367	0
6	1,727	9	20	107	0
7	59	2	24	45	0
8	33	0	27	22	0
9	5	0	45	10	0
10	8	0	60	4	0

(b) 전체 레코드(우량고객)

	X1	X2	X3	X4	결과
0	0	0	0	0	1
1	0	0	28	0	1
2	2,707	0	68	0	1
3	79	0	24	0	1
4	19	2	170	60	1
5	6	307	250	508	1
6	3	244	1,110	801	1
7	5	111	165	225	1
8	0	32	161	146	1
9	0	12	292	60	1
10	0	4	646	24	1

(c) 전체 레코드(불량고객)

ID	X1	X2	X3	X4	결과
1	7	2	2	5	0
2	3	9	8	7	1
:	:	:	:	:	:
2,470	2	8	8	6	1

(d) Data-Set 1 검증용 데이터

ID	X1	X2	X3	X4	우량값	결과
1	59	0	379	367	805	0
2	15	0	27	45	87	1
:	:	:	:	:	:	:
2,470	11	0	27	107	145	1

(e) 각 레코드의 우량값 산출

ID	X1	X2	X3	X4	불량값	결과
1	5	0	66	508	581	0
2	79	12	161	225	477	1
:	:	:	:	:	:	:
2,470	2707	32	161	801	3701	1

(f) 각 레코드의 불량값 산출

ID	X1	X2	X3	X4	우량값 (e)	불량값 (f)	예측 결과	실제 결과	정확도
1	7	2	2	5	805	581	0	0	1
2	3	6	7	2	87	477	1	1	1
:	:	:	:	:	:	:	:	:	:
2,470	2	8	8	6	145	3701	1	1	1

(g) Frequency Matrix 예측결과

〈그림 5〉 Frequency Matrix 예측기법

예를 들면 (d)번의 검증용 데이터 레코드 1번은 X1=7, X2=2, X3=2, X4=5 값을 가지고 있다. 이에 대한 (b)의 전체 레코드 우량고객 매트릭스 값은 X1=59, X2=0, X3=379, X4=367 값에 해당한다. 이 값을 모두 합한 (e)의 우량값은 805가 된다. 또한 (c)의 전체 레코드 불량고객 매트릭스도 같은 방식으로 대입하면 (f)의 불량값은 581이 산출된다.

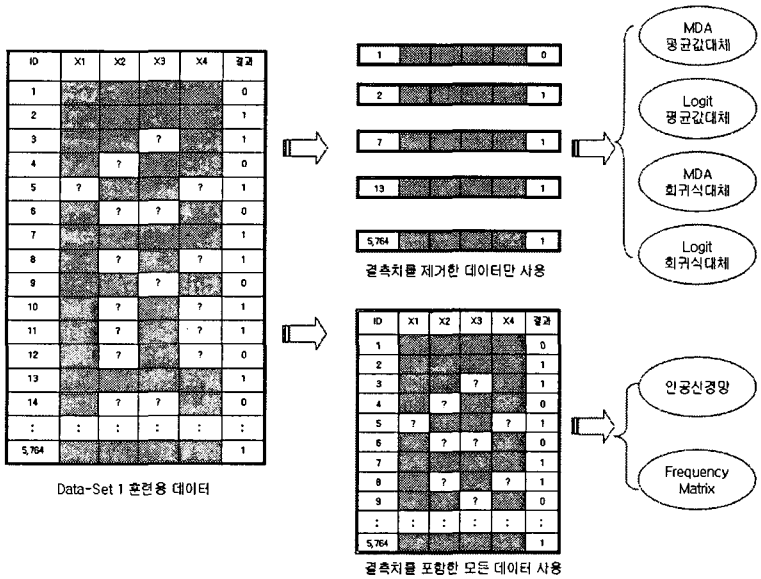
따라서 (g)의 Frequency Matrix 예측결과는 우량값과 불량값 중에서 큰 값이 예측결과로 채택되며 예제 1번 레코드에서는 우량고객으로 판별된다.

본 연구에서는 데이터 값의 구간 설정시 0과 10사이의 11구간, 0과 20사이의 21구간, 0과 30사이의 31구간의 3가지 방식으로 구분하여 실험을 수행하였으며 이 중에서 가장 높은 예측정확도를 가진 구간을 예측결과로 삼았다. Frequency Matrix 기법은 Microsoft의 Visual Basic .NET을 이용하여 구체화하였다.

4.3 개인신용예측모형 비교분석

앞에서 제시한 6가지 개인신용예측 모형의 특성을 <그림 6>으로 나타내어 비교해 본다.

<그림 6>에서 보는 바와 같이 MDA-평균값대체, Logit-평균값대체, MDA-회귀식대체, Logit-회귀식대체는 훈련용 데이터에서 결측치를 제거한 데이터만 사용하여 평균값과 회귀식을 도출한 다음 검증용 데이터에 대입하여 결측치를 대체하는 방법론이며 인공신경망과 Frequency Matrix 방법론은 결측치를 추정하지 않고 모든 데이터를 그대로 이용하여 개인신용을 예측하는 방법론이다.



〈그림 6〉 개인신용예측모형 비교분석

5. 연구 결과

연구결과 분석을 위하여 검증데이터의 결과만을 모아 <표 3>을 구성하였고 <표 3>의 결과를 바탕으로 모형별 평균 예측력을 보여주는 <그림 7>을 구성하였다.

〈표 3〉 전체결과

구 분	MDA 평균값대체	Logit 평균값대체	MDA 회귀식대체	Logit 회귀식대체	인공신경망	Frequency Matrix
Set 1	61.74	63.81	63.84	66.52	62.75	69.35
Set 2	67.44	63.56	69.47	70.53	68.22	68.87
Set 3	63.36	61.86	65.71	70.20	63.48	69.27
Set 4	67.93	66.56	69.11	69.88	71.38	69.15
Set 5	67.57	64.70	67.61	68.14	66.28	69.31
Set 6	68.58	65.43	69.76	70.89	70.49	72.95
Set 7	63.36	63.16	64.45	68.18	66.19	69.15
Set 8	67.97	63.72	64.94	68.34	66.17	69.51
Set 9	60.40	62.39	64.09	67.53	63.00	69.31
Set 10	66.35	62.87	68.22	69.07	66.90	70.08
평 균	**65.47**	**63.81**	**66.72**	**68.93**	**66.49**	**69.70**
표준편차	2.97	1.42	2.36	1.42	2.95	1.19

<표 3>과 <그림 7>의 결과에서 나타난 것과 같이 평균예측력은 Frequency Matrix 기법이 전체 방법론과 비교하여 가장 예측력이 높았다. 데이터 셋 별로 살펴보면 Data Set 1, Data Set 5, Data Set 6, Data Set 7, Data Set 8, Data Set 9, Data Set 10에서는 Frequency Matrix 기법, Data Set 2와 Data Set 3에서는 Logit-회귀식대체, Data Set 4에서는 인공신경망 모형 등이 해당 각 데이터 셋에서 예측력이 가장 높다는 것을 알 수 있다.

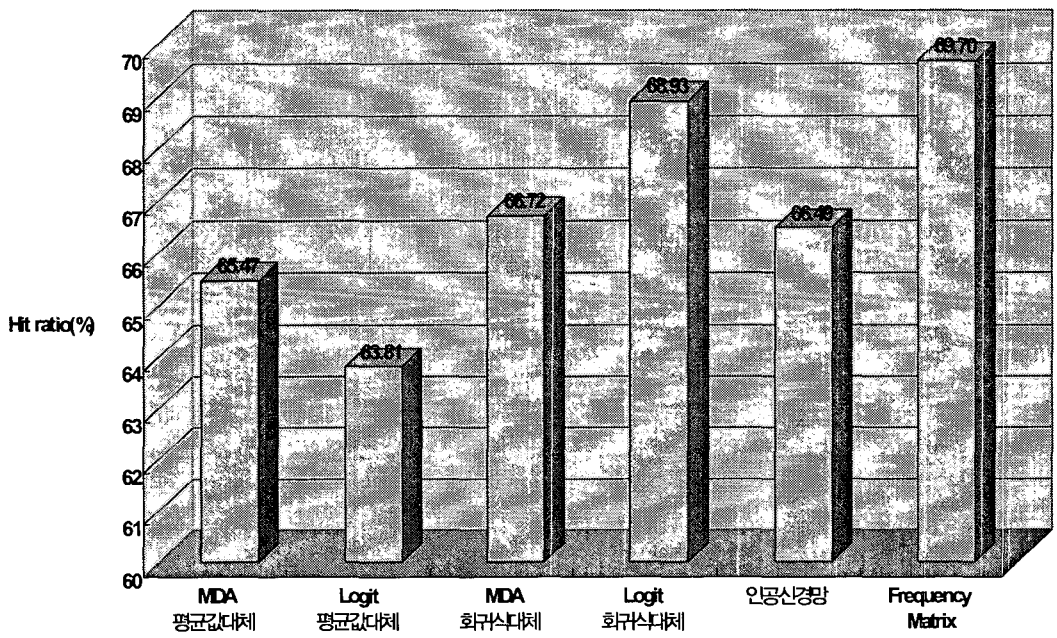
결측치를 대체한 방법론 중에서는 회귀식대체 집단(67.82%)이 평균값대체 집단(64.64%)보다 우수한 예측력을 보였다. 또한 평균값대체 집단에서는 다변량 판별분석이 로지스틱 회귀분석보다 예측력이 높았으며 회귀식대체 집단에서는 로지스틱 회귀분석이 다변량 판별분석보다 예측력이 높다는 것을 알 수 있다.

결측치를 대체하지 않은 방법론 중에서 인공신경망은 평균값대체 집단보다는 높은 예측력을 보여주고 있지만 회귀식대체 집단보다는 낮은 예측률을 보여주고 있다.

평균예측력을 모든 방법론과 비교하여 순서대로 나열하면 Frequency Matrix, Logit-회귀식대체, MDA-회귀식대체, 인공신경망, MDA-평균값대체, Logit-평균값대체의 순이다.

상기 기법들 간의 예측정확도 차이가 통계적으로 유의한지를 검증하기 위해 검증용 데이터를 대상으로 McNemar 검정을 실시하였다.

McNemar 검정은 비모수 통계분석기법으로 이진값을 가지는 명목형 변수에 대해 관련이 있는 두 집단간의 차이를 검정할 때 사용된다. 특히, McNemar 검정은 동일한 대상에 대한 처리 전·후의 측정치 비교에 매우 유용한 것으로 알려져 있다.



〈그림 7〉 모형별 평균예측력 비교

〈표 4〉 McNemar 검정결과

	MDA 평균값대체	Logit 평균값대체	MDA 회귀식대체	Logit 회귀식대체	인공신경망
Frequency Matrix	9.372 (0.002)**	17.738 (0.000)**	4.804 (0.028)**	0.728[1] (0.394)[2]	5.789 (0.016)**
MDA 평균값대체		1.356 (0.244)	0.664 (0.415)	4.501 (0.034)**	0.356 (0.551)
Logit 평균값대체			4.008 (0.045)**	11.080 (0.001)**	3.312 (0.069)*
MDA 회귀식대체				1.666 (0.197)	0.033 (0.855)
Logit 회귀식대체					2.207 (0.137)

1) McNemar 통계량 값 2)p-값
* : 유의수준 10%에서 통계적으로 유의함.
** : 유의수준 5%에서 통계적으로 유의함.

<표 4>에서 보는 바와 같이 Frequency Matrix 기법은 MDA-평균값대체, Logit-평균값대체, MDA-회귀식대체, 인공신경망과는 5% 수준에서 유의한 차이를 나타내어 예측성과가 뛰어남을 확인할 수 있었으나 Logit-회귀식대체와는 통계적으로 유의한 차이를 나타내지 않았다. 한편, MDA-평균값대체는 Logit-회귀식대체와 5% 수준에서 유의하였으나, Logit-평균값대체, MDA-회귀식대체, 인공신경망과는 통계적으로 유의한 차이를 나타내지 않았다. Logit-평균값대체는 MDA-회귀식대체, Logit-회귀식대체와 5% 수준에서 유의하였고 인공신경망과는 10% 수준에서 유의하였다. MDA-회귀식대체는 Logit-회귀식대체, 인공신경망과 통계적으로 유의한 차이를 나타내지 않았으며 Logit-회귀식대체도 인공신경망과 통계적으로 유의한 차이를 나타내지 않았다.

6. 결 론

본 연구에서는 결측치가 있는 개인신용 데이터를 활용하여 보다 효과적인 개인신용예측을 위해 Frequency Matrix(FM) 방법론을 제시하였다. 이를 위하여 평균값대체 방법론, 회귀식대체 방법론을 통하여 결측치를 추정하고 각각의 방법론으로 결측치를 추정한 상태에서 개인신용분류 예측을 위한 다변량 판별분석과 로지스틱 회귀분석 모형을 제시하였다. 또한 결측치를 추정하지 않고 개인신용을 예측하는 인공지능적인 방법인 인공신경망 모형, 결측치를 추정하지 않고 모든 데이터를 그대로 이용하여 개인신용을 예측하는 Frequency Matrix 기법의 6가지 개인신용예측모형을 제시하였다. 실험결과, 본 연구에서 제안한 개인신용예측 모형인 Frequency Matrix 기법이 평균값으로 결측치를 추정한 MDA-평균값대체, Logit-평균값대체, 회귀식으로 결측치를 추정한 MDA-회귀식대체, Logit-회귀식대체, 결측치를 추정하지 않은 인공신경망의 모형과 비교한 결과 가장 예측정확성이 우수한 것으로 나타났다.

연결빈도행렬의 인접행렬특성과 브레인 매핑의 데이터 필터링 알고리즘을 따르고 있는

Frequency Matrix 기법은 결측치를 대체하지 않은 상태에서 모든 데이터를 그대로 이용하여 개인신용을 예측할 수 있다는 장점이 있으나 데이터의 구간을 11구간, 21구간, 31구간으로 나누어 각각의 매트릭스를 구현해야 한다는 단점이 있다.

향후 연구에서는 임의적으로 결측치 비율을 조절(10%, 20%, 30%, 40%, 50%)하거나 결측치 생성 방법을 MCAR, MAR, Non-ignorable 등의 시나리오별로 결측치를 추정하는 방법, 의사결정나무를 이용하는 방법, MCMC와 EM 알고리즘을 이용하는 방법 등으로 결측치를 추정한 상태에서 개인신용예측모형을 개발한다면 보다 의미 있는 정보를 제공할 수 있을 것으로 기대한다.

참 고 문 헌

[1] 김상용, 박재인, "시계열 모형에서 결측치 보정에 관한 연구", *수학 · 통계논문집*, 제8권, 2001, pp. 1-18.

[2] 김진화, 남기찬, 변현수, "웹 방문 패턴 시각화 및 상품추천 방법에 관한 연구", *한국경영정보학회 추계학술대회 발표논문집*, 2004, pp. 47-55.

[3] 박정민, 김경재, 한인구, "Support Vector Machine을 이용한 기업부도예측", *경영정보학연구*, 제15권, 제2호, 2005, pp. 52-62.

[4] 신형원, 손소영, "범주형 자료의 결측치 추정방법 성능 비교", *한국경영과학회/대한산업공학회* 춘계공동학술대회, 2001, pp. 813-816.

[5] 이명식, *소비자 신용평가 제도, 정립 필요하다*, 국민경제리뷰, 1992, pp. 28.

[6] 이훈영, 이시환, "다양한 결측치 처리방법에 따른 부실예측모형의 정확도 비교 연구", *경희대학교 경영연구*, 제4호, 1999, pp. 182-195.

[7] Altman, E. I., Marco, G., and Varetto, F., "Corporate distress diagnosis : Comparisons using linear discriminant analysis and neural networks (the Italian experience)", *Journal of Banking and Finance*, Vol. 18, No. 3, 1994, pp. 505-529.

[8] Berry, M. and Linoff, G., *Data Mining Techniques*, Wiley, New York, 1997.

[9] Breiman, L., Friedman, J., Olshen, R., and Stone, C., *Classification and Regression Trees,* Champman and Hall, New York, 1984.

[10] Carter, C. and Catlett, J. "Assessing credit card applications using machine learning", *IEEE Expert*, Vol. 2, 1987, pp. 71-79.

[11] Condon, E., Golden, B., Lele, S., Raghavan, S., and Wasil, E., "A visualization model based on adjacency data", *Decision Support Systems*, Vol. 33, No. 4, 2002, pp. 349-362.

[12] Desai, V. S., Convay, D. G., Crook, J. N., and 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*, Vol. 8, 1997, pp. 323-346.

[13] Feelders, A. J., "Handling missing data in trees: surrogate splits or statistical imputation", *Proceedings of third European conference on principles and practice of knowledge discovery in databases (PKD D99)*, Springer, 1999, pp. 329-334.

[14] Fox, P. T., Laird, A. R., Fox, S. P., Fox,

P. M., Uecker, A. M., Crank, M., Koenig, S. F., and Lancaster, J. L., "BrainMap Taxonomy of Experimental Design Description and Evaluation", *Human Brain Mapping*, Vol. 25, 2005, pp. 185-198.

[15] Hair, J.F., R.E. Anderson, R.E. Tatham and W.C. Black, *Multivariate Data Analysis with Readings*, Prentice Hall, 1995.

[16] Hornik, K., "Approximation capabilities of multilayer feedforward networks", *Neural Networks*, Vol. 4, 1991, pp. 251-257.

[17] Law, S. K., Nunez, P. L., Westdorp, A. F., Nelson, A. V., and Pilgreen, K. L., "Topographical mapping of brain electrical activity", *Proceedings of the IEEE Conference*, 1991, pp. 194-201.

[18] Little, R. J. A., and Rubin, D. B., *Statistical Analysis with Missing Data*, J. Wiley & Sons, New York, 1987.

[19] Makowski, P. "Credit scoring branches out", *Credit World*, Vol. 75, 1985, pp. 30-37.

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

[21] Ohlson, J. A., "Financial ratios and probabilistic prediction of bankruptcy", *Journal of Accounting Research*, Vol. 18, No. 1, 1980, pp. 109-131.

[22] Orre, R., Bate, A., Noren, G.N., Swahn, E., Arnborg, S., Edwards, I.R., "A bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets", *International Journal of Neural Systems*, Vol. 15, No. 3, 2005, pp. 207-222.

[23] Parthasarathy, S., and Aggarwal, C., "On the Use of Conceptual Reconstruction for Mining Massively Incomplete Data Sets", *IEEE Transactions on Knowledge & Data Engineering*, Vol. 15, No. 6, 2003, pp. 1512-1521.

[24] Peel, M. J., D. A. Peel, and P. F. Pope, "Predicting corporate failure-some results for the UK corporate sector", *Omega*, Vol. 14, No. 1, 1986, pp. 5-12.

[25] J. Quinlan, C4.5 : *Programs for Machine Learning*, Morgan Kaufmann, 1993.

[26] Raghunathan, T. E., "What do we do with missing data? Some options for analysis of incomplete data", *Annual Review of Public Health*, Vol. 25, 2004, pp. 99-117.

[27] Robert, H. Cole, *Consumer and Commercial Credit Management*, Boston Irwin Home Wood IL, 1992, pp. 13.

[28] Schafer, J., Konstan, J., and Riedl, J., "E-commerce recommendation applications", *Data Mining and Knowledge Discovery*, Vol. 5, No. 1&2, 2001, pp. 115-153.

[29] Schafer, J. L., *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London, 1997.

[30] Schoenberg, R., and Arminger, G., "MISS: Analysis of Incomplete Data", *Multivariate Behavioral Research*, Vol. 23, 1988, pp. 275-276.

[31] Strike, K., El Emam, K., and Madhavji, N., "Software Cost Estimation with Incomplete Data", *IEEE Transactions on Software Engineering*, Vol. 27, No. 10, 2001, pp. 890-908.

[32] Wang, S. "Application of self-organizing maps for data mining with incomplete data sets", *Neural Computing and Applications*, Vol. 12, 2003, pp. 42-48.

[33] Weiss, S. and Kulikowski, C., *Computer Systems That Learn, Morgan Kaufmann Publishers*, Inc., 1991.

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

□ 저자소개

배 재 권

현재 서강대학교 경영학과 MIS 전공 박사과정에 재학 중이다. 한남대학교 MIS전공 학사, 서강대학교 재무관리 전공으로 경영학석사 학위를 취득하였다. 주요 관심분야는 기업/개인 신용평가, 재무 정보시스템, Data Mining, 인공지능, EC, ERP 등이다.

김 진 화

University of Wisconsin-Madison에서 전산학 석사 그리고 경영정보학 석·박사를 취득하였다. Oklahoma State University에서 MIS분야 조교수로 재직하였으며, 현재 서강대학교 경영학과에 경영정보학 분야 부교수로 재직중이다. 주요 연구관심분야는 Data Mining, Customer Relations Management, Simulation of Human Learning, Heuristic Optimization 등이다.

황 국 재

현재 서강대학교 경영학과 회계학 분야 부교수로 재직 중이다. 서강대학교 경영학사, Michigan State University에서 MBA를 취득하였고, Syracuse University에서 관리회계 분야로 Ph. D.를 취득하였다. 주요 연구관심분야는 복식부기 적용시 국가재정정책 변화 및 전략의 변화양상, 대리인비용 이론의 기업관리 및 전략의사결정 응용, 예산슬랙의 현실적용 등이다.

◈ 이 논문은 2006년 05월 19일 접수하여 수정없이 2006년 12월 11일 게재확정되었습니다.