변수간 공분산구조와 가중치를 고려한 사례기반추론의 적용: 기업부도예측

홍효정
서강대학교 경영대학
(rina216@hanmail.net)
조성빈
서강대학교 경영대학
(sungbincho@sogang.ac.kr)

본 연구는 기업부도예측을 위한 사례기반추론모형을 제시하고 전통적 통계모형과 인공지능모형과의 성과를 비교하고자 하였다. 기존의 연구에서는 모형 입력변수의 가중치를 차별화하는 것만 고려하였는데, 본 연구에서는 입력변수의 공분산구조를 추가적으로 고려하였다. 2001년에서 2003년에 걸쳐 수집된 중소규모의 제조업분야의 데이터를 대상으로 실험을 실시하였다. 사례기반추론 모형에서는 두 가지 실험요인(변수간 공분산구조를 고려함/고려하지 않음 × 변수가중치를 차별화함/차별화하지 않음)을 고려하여 총 네 가지 모형을 적용하였다. 가장 근접한 이웃에 포함되는 이웃의 수는 탐색을 통하여 15개로 결정하였다. 시뮬레이션기법을 적용하여 훈련용 데이터에서 가장 우수한 결과를 주는 모형을 선별한 후, 검증용 데이터에서 모형의 성과를 비교하였다. 실험 결과, 변수간 공분산구조와 가중치를 고려한 사례기반추론모형이 기존 모형에 비하여 높은 예측률을 나타내었다.

주제어: 사례기반추론, 부도예측, 변수간 공분산구조, 변수가중치

Ⅰ. 서 론

기업의 재무적 건전성 진단과 신용위험의 측정은 지난 60여년 간 학계와 금융산업 실무자들에게 매우 중요하고 해결하기 어려운 과제 중의 하나로 여겨져왔다. 정부 관리, 투자자, 신용평가 전문가 등 많은 이해당사자들이 기업의 채무불이행 위험을 평가하고 관리하는데 노력을 기울여왔다. 한국은행(2006)의 자료에 따르면 2005년 매달 평균 285개의 기업이 파산을 신청하였다. 기업을 둘러싼 환경의 불확실성이 증가함에 따라 기업부도사태는 이제 사회경제적으로 잘 알려져있고 건실하다고 판단되어왔던 대기업에도 드물지 않게 일어나고 있는 실정이다. 기업부도 예측에 있어서 정확성의 추구는 신용평가비용의 감소, 보다 철저한 모니터링과 관리, 부채상환비율의 증가와 같은 여러 가지 이득을 줄 수 있다.

부도예측에 적용되어온 통계분석방법론과 신용평가모형의 발전을 살펴 보면 다음과 같다. 은행산업은 전통적으로 신용리스크를 평가하는 내부평가시스템을 운영하여왔다. 이 평가시스템은 재무비율과 같은 정량적 측면과 함께 기업의 평판과 같은 정성적 측면도 고려하였다(Treacy and Carey, 2000). 좀 더 발달된 체계적인 방법으로는 통계적 모형과 기계학습적 접근방법이 있다. 이러한 방법들은 기업의 계량적 측면이 많이 반영되는 특징이 있다. 선행연구에 따르면 다변량 판별분석이 부도예측에 쓰인

논문접수일: 2009. 1 게재확정일: 2009. 7

최초의 통계적 기법이라고 할 수 있을 것이다. 판별분석은 널리 알려진 것처럼 통계모형의 엄격한 가정을 충족시켜야 하는 한계점을 지니고 있었다. 그리하여 엄격한 통계적 가정에서 벗어난 로지스틱 회귀분석이 대안으로서 점차 인기를 얻어왔다. 기업경영문제의 해결에 인공지능적 접근방법이 1990년대부터 적용되기 시작하면서, 최근에는 이러한 기법이 매우 진보되고 세련되어지고 있는 추세이다. 기존연구에 대한 조사에 따르면, 이러한 인공지능기법들은 전통적 통계기법 보다 많은 경우에 높은 예측력을 나타내고 있다.

본 연구는 1990년대 말부터 널리 사용되고 있는 사례기반추론(case-based reasoning: CBR)을 적용하여 부도를 예측하는데 목적이 있다. CBR은 문제와 해답 쌍(pair)의 형식으로 저장된 사례를 이용하여 문제를 해결하는 추론방식으로, 사례 간의 유사성/비유사성을 측정하여 가장 가까운 이웃을 결정하고, 이들 이웃의 결과값을 투표하는 방식을 적용하여 목표집단에 대한 결과를 예측한다. 부도예측에 관련된 정확성을 향상시키기 위하여 본 연구는 입력변수 간의 공분산구조를 고려하여 기업 간의 유사성/비유사성을 계산할 수 있는 마할라노비스 거리(Mahalanobis distance)개념을 도입하여 기업 간의 거리 메트릭스(distance matrix)를 설계하였다. 유클리디안 거리(Euclidean distance)를 이용한 기존의 사례기반추론에서는 입력변수 간의 공분산 혹은 상관계수가 모두 0으로 취급되어 왔다. 즉, 변수간의 선형 상관관계가 모두 0인 경우에는 기존의 사례기반추론이 이상적인 모형이지만, 그러한 경우는 현실적으로 매우 드물다고 할 수 있다. 결과적으로 사례기반추론은 부도예측에 있어서 인공신경망이나 로지스틱 회귀분석 등의 기법에 비하여 예측력이 떨어졌고 덜 적용되어왔다. 또한 거리를 계산 함에 있어서 입력변수의 가중치를 차별화 함으로써 추가적으로 예측력을 향상시키고자 하였다. 실제기업의 재무데이터를 이용하여 본 연구가 제시하는 모형의 유용성을 평가하고자 하였다.

II. 이론적 배경

2.1 선행 연구

기업부도예측은 기본적으로 이분적 의사결정(dichotomous decision) - 부도 혹은 정상(비부도) - 라고 할 수 있다. 통계적 기법과 인공지능 기법은 부도가 될 확률을 계산하여 임계값(예를 들어 0.5) 보다 크면 "부도"라고 예측하는 것이다. 부도예측에 관련된 초기의 연구는 통계적 기법에서 비롯되었다.

Beaver(1966)는 수익성, 유동성, 지급성에 관련된 재무비율을 이용하여 만기에 기업이 채무를 상환하지 못할 리스크를 연구하였다. 그의 연구는 각 재무비율 별로 부도/정상을 구분하는 한계값(threshold value)을 개발하는데 중점을 두었다. 이러한 비율분석의 뒤를 이어 통계적 선형 모형이 개발되었다. Altman(1968)은 다변량 판별분석(multiple discriminant analysis)를 이용하여 개별기업의 판별점수를 산출하였다. 이 모형의 예측력은 도산이 발생하기 2년 전까지는 우수하였고 그 이후에는 예측력이 현저히 떨어졌다. Ohlson(1980)은 시그모이드 함수(sigmoid function)를 이용한 로지스틱 회귀(logistic regression)모형을 도입하여 기업도산문제를 예측하고자 하였다. 판별분석에 비하여 로지스틱 회귀분석은 0과 1사이의 값을 갖는 로지스틱 점수를 산출하여 확률적으로 해석할 수 있다는 장점

을 가지고 있었다.

1990년대에 들어서 빠른 속도로 발전하는 컴퓨터 기술의 발전에 힘입어, 데이터 마이닝 기법들이 경영문제의 해결에 적용되기 시작하였다. 데이터 마이닝 기법들은 개인용 컴퓨터의 메인 프로세서가 팬티엄급 이상으로 진보되는 시점에서 개인 분석자의 입장에서도 SAS Enterprise Miner나 SPSS Clemantine 등의 패키지들을 구동할 수 있게 됨에 따라 널리 적용되기 시작하였다. 데이터 마이닝 기법은 비선형 패턴을 갖는 대용량의 데이터에서 모형의 유효성이 검증되어 왔다. 인공지능 기법을 이용한 초기 연구의 대다수는 전통적 통계기법 보다 좀 더 정확한 예측을 할 수 있는 가에 초점을 두었다(Boritz and Kennedy, 1995; Coates and Fant, 1992; Klersey and Dugan, 1995; Pompe and Feelders, 1997). Fletcher and Goss(1993)와 Desai et al. (1996)의 연구에서는 판별분석이나 로지스틱 회귀분석에 비하여 인공신경망모형의 예측력이 우수함을 보였다.

초기의 인공신경망 모형이 역전파(backpropagation) 알고리즘에 의존한 반면, 최근의 연구들은 다양한 알고리즘을 소개하고 있다. Charalambous et al.(2000)은 learning vector quantization과 feedforward network를 도입하여 역전파에 기초한 인공신경망과의 예측력을 비교하였다. Anandarajan et al.(2001)은 유전자 알고리즘에 기초한 인공신경망 모형을 소개하였다. Pendharkar(2005)는 임계값이 변하는 인공신경망을 유전자 알고리즘과 결합하였으며, Lee et al.(2005)은 데이터 크기가 인공신경망의 예측력이 미치는 영향을 조사하였다. 인공신경망 외에도 의사결정나무 추론(decision tree induction)과 사례기반추론도 기업부도 예측에 적용되었다(Bryant, 1997; Curram and Mingers, 1994).

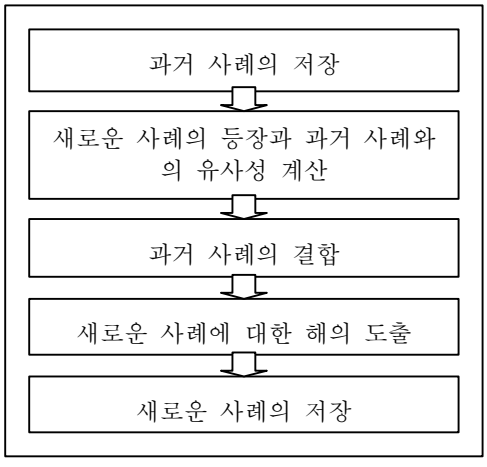
2.2 사례기반추론 모형

인간이 미래를 예측할 때 과거경험을 근거로 판단하듯이, CBR은 지식을 "if-then"의 규칙 형식으로 기술하는 것이 곤란한 경우에 유효한 추론 방식으로서, 데이터 베이스에서 사례를 검색하고 적용하여 해답을 얻는 방식이다. CBR을 구현하려면 규칙을 작성하는 대신에 사례를 수집하여 데이터 베이스에 저장하면 된다. CBR을 실행하는 데 문제가 되는 것은 유사도의 정의와 얻어진 유사한 사례의 해답을 목적하는 문제에 맞추기 위한 수정 방법인데, 범용적인 방법은 없기 때문에 개발하는 시스템에 따라 결정되는 것이 보통이다.

CBR 모형의 시스템은 일반적으로 〈그림 1〉이 나타내는 것처럼 다섯 단계의 순환 구조로 구성된다. CBR 시스템은 새로운 의사결정 문제가 발생하면, 일차적으로 사례를 모은 데이터 베이스로부터 현재의 사례와의 거리를 계산하여 거리가 적은 유사하다고 판단되는 과거의 사례들을 추출(retrieve)한다. 그리고 추출된 사례의 정보를 그대로 적용(reuse)하거나 수정(revise) 함으로써 새로운 사례에 대한 최적해를 도출하는 것이다. 새로운 사례는 다시 데이터 베이스에 저장되고 미래의 문제해결에 활용된다.

CBR을 이용한 기업부도예측은 1990년대 후반 이후에 활발히 연구되고 있다(Bryant, 1997; Elhadi and Vamos, 1999; Elhadi, 2000; Jo et al., 1997; Park and Han, 2002).

Bryant(1997)는 사례기반분석모형을 기업부도 예측에 적용하여 Logit 모형과의 성과를 비교하였다. 독립변수는 Baldwin and Glezen(1992)과 Ohlson(1980)의 연구에서 중요하다고 판단되었던



〈그림 1〉 사례기반추론에서의 추론과정

25개를 사용하였으며, 기간별 Hold out방식에 의하여 10%의 표본을 추출하여CBR의사결정나무 클러스터링으로 분석하는 방식과 전체 표본으로 분석하는 방식을 사용하였다. Type I error(부도 기업을 비부도기업으로 예측)와 Type II error(비부도기업을 부도기업으로 예측)값으로 모형의 유용성을 판단하였다. 결과적으로 Type I error는 CBR모형이 Logit모형보다 높은 반면, Type II error는 CBR모형이 Logit모형보다 낮게 나타났다. 비용적 측면으로 보았을 때 Type I error와 Type II error가 같은 비용을 부담한다면 본 모형을 어느정도 유용하게 사용할 수 있겠으나, Type I error가 비용손실 리스크가 높다는 점에서 판단한다면, 현실문제의 적용에 문제점이 있다고 판단된다.

Elhadi and Vamos(1999)의 연구와 Elhadi (2000)의 연구에서는 CBR과 Information Retrieval (IR)방식을 결합하여 기업부도와 관련된 법령을 보다 쉽고 적합한 목차로 검색하는 모형을 제시하였다. 법조인들이 어떻게 조사결과를 법조문에서 유용한 정보화하여 나타내는지 모형화 하였다. 본 연구에서 주제어는 제한된 개념과 내용, 구조, 정형화된 시나리오로 정해져 있다고 가정하였으며, 입력변수는 상황에 대한 자연스러운 서술로 하고 출력변수는 비슷한 사례의 검색결과로 나타내도록 하였다. 연구결과, 본 연구가 제시하는 Statutes-Seeded Automatic Indexing and Retrieval(SSAIR) 방법은 정확성이 95%로 나타나, 비교 모형인 식별가치기반분석모형(DV-Based)과 법령기반분석모형(Textbook-Cases)의 정확성인 각각 85%, 87%보다 높은 값을 나타내고 있다. 이와 같이 SSAIR모형은 IR과 CBR을 이용하여 인간의 의사결정 결과와 비슷한 사례를 찾는데 더 유용하였다.

Jo et al.(1997)은 다변량 판별분석(Discriminant Analysis: DA), 사례기반추론(Case-Based Forecasting System: CBFS), 인공신경망(Neural Network: NN)을 이용하여 한국기업의 부도기업과 비부도기업을 예측하였다. 기존 기업부도예측 연구에서 유의하다고 알려진 61개의 재무비율 독립변수를 정리하고 이를 다시 단계선택법과 t-test를 통해 변수를 추출하였다. 세 모형의 예측률은 81.5~83.8%로 나타났으며 NN 모형이 가장 우수하고 DA와 CBFS는 특이할 만한 다른 점을 발견하지 못하였다. 본 연구에서는 이분(binary)으로 나타나는 종속변수에 대한 예측과 독립변수와 종속변수의 상관도가 낮은 경우에 사례기반추론의 예측력이 떨어진다고 추정하였다.

Park and Han(2002)은 Analytic Hierarchy Process(AHP)를 이용하여 CBR의 중요한 변수들에 대한 가중치를 결정하였다. CBR 모형에서 사용되어지는 k-nearest neighbor(k-NN)방식은 거리측정 방식에 따라 그 값이 민감하게 달라질 수 있으므로 가중치를 적용하여 민감성을 줄이고 모형의 예측력을 높이는 데에 그 목적을 두었다. 본 연구에서

는 AHP모형을 통한 전문가 집단의 의견을 변수가중치에 사용하였고, 정량적 변수뿐 아니라 정성적 변수도 도입하였다. 본 연구의 정량적 변수(13개)는 통계적 방법으로 유의한 변수들을 추출하였고, 정성적 변수(15개)는 선행연구 혹은 신용평가모형에서 사용되는 변수를 선정하였다. AHP의 가중치를 적용한 CBR모형은 정량적 변수만 쓴 경우 K=10일 때 통계적 방법으로 가중치를 적용한CBR이나 가중치가 적용되지 않은 CBR과 비교하여 4개의 실험군에서 가장 높은 예측력을 나타내었으며 정량적 변수와 정성적 변수를 모두 쓴 경우 K=15일 때 가중치가 적용되지 않은 CBR과 비교하여 높은 예측력을 나타내었다. 유사성/비유사성의 측정에 새로운 함수를 도입하거나 새로운 변수가중치 선정방법을 선택하는 것이 향후 연구로 지적되었다.

CBR에서는 객체 간의 유사성의 정도를 정량적으로 나타내기 위하여 일종의 척도가 필요하다. 가장 보편적으로 많이 사용되는 것은 거리(distance)인데, 거리와 같이 클수록 유사성이 적어지는 척도 즉, 비유사성 척도(dissimilarity measure)를 사용한다.

n 개의 객체가 각각 p 개의 속성 또는 변수(attribute, variable, 또는 데이터 마이닝분야에서는 feature로 표현함)를 가진다고 하자. 객체 i의 p차원 공간에서의 좌표는 다음과 같이 열 벡터로 표현될 수 있다.

$$\mathbf{Y_{i,p}} = (Y_{i,1}, Y_{i,2}, ..., Y_{i,\mathbf{p}})$$

이 때, 서로 독립적인 입력변수들 간의 거리를 일반화시킨 거리를 민코우스키 거리(Minkowski distance)라 하는데 이는 다음과 같이 정의된다.

$$Minkowski\_d_{ij} = \left(\sum\nolimits_{p=1}^{\mathbf{p}} | \mathbf{Y_{i,p}} - \mathbf{Y_{j,p}} |^m\right)^{1/m}$$

여기에서 m = 2일 때 유클리디안 거리(Euclidean distance)를 나타내며, m = 1일 때는 맨하탄 거리(Manhattan or rectilinear distance)를 나타낸다. 군집분석에서는 통상적으로 유클리디안 거리가 가장 많이 사용되므로, 별도의 언급이 없는 한 "거리"라고 하면 유클리디안 거리를 지칭한다.

본 연구에서 객체로 표현되는 기업 i 와 기업 j 간의 유클리디안 거리를 구하는 공식은 다음과 같다 ($i \neq j$):

$$ed_{ij} = \sqrt{(\mathbf{Y_{i,p}} - \mathbf{Y_{j,p}})(\mathbf{Y_{i,p}} - \mathbf{Y_{j,p}})^T}$$

($\mathbf{Y_{i,p}}$ = 기업 i의 p속성수준 벡터, $(\ )^T$ = 벡터의 전치행렬).

유클리디안 거리의 경우, 변수들 간의 단위 등이 다를 때 큰 값을 갖는 변수가 거리의 값을 주도할 수 있다. 이러한 단점을 해소하기 위하여 종종 표준화된 거리(standardized distance)를 사용한다. 마할라노비스 거리(Mahalanobis distance)는 표준화된 거리를 보다 일반화한 것으로 두 지점의 단순한 거리 뿐만이 아니라, 변수 간의 특성을 나타내는 표준편차와 상관계수가 함께 고려된다는 특징을 가지고 있다. 그러므로 각 변수가 어떤 절대값을 가지던 간에, 변수의 분산-공분산 행렬(variance-covariance matrix)의 역함수를 이용하여 어느 한 변수의 크기에 영향 받지 않고 거리를 측정할 수 있다.

마할라노비스 거리를 CBR 방법에 적용시킨 기존의 연구는 없는 것으로 파악되었으며, 마할라노비스 거리를 이용한 기타 논문은 자료융합에 관한 연구가

있었다. 김성호와 조성빈(2005)의 연구에서는 자료융합에서 효과적인 데이터베이스의 축소와 설계에 기여할 수 있는 여러 가지 기증자선택전략을 탐구하고 그 성과를 비교하였다. 공통속성변수를 미리 정하여 놓은 계획된 자료융합을 설정하였으며, 자료융합의 영역에 최초로 마할라노비스 거리의 개념을 도입하고 응답자 간의 비유사성을 측정하였다. 다섯 가지의 기증자선택전략이 설계되었다: 상관계수를 적용하는 전략; 유클리디안 거리를 적용하는 전략; 마할라노비스 거리를 적용하는 전략; 대응일치분석을 거쳐 유클리디안 거리를 적용하는 전략; 대응일치분석을 거쳐 마할라노비스 거리를 적용하는 전략. 몬테카를로 시뮬레이션을 실시한 결과, 대응일치분석을 거쳐 마할라노비스 거리를 적용하는 전략의 성과가 가장 우수하게 도출되었다.

부도 여부를 알아야 하는 기업 i 와 기준이 되는 기업 j 간의 마할라노비스 거리를 구하는 공식은 다음과 같다($i \neq j$):

$$md_{ij} = \sqrt{(\mathbf{Y_{i,p}} - \mathbf{Y_{j,p}})\Gamma^{-1}(\mathbf{Y_{i,p}} - \mathbf{Y_{j,p}})^T}$$

($\mathbf{Y_{i,p}}$ =기업 i의 p속성수준 벡터, $(\ )^T$ = 벡터의 전치행렬, Γ^{-1} = 속성의 분산-공분산 역행렬).

만일 모든 변수가 표준화되어 있고 서로 통계학적으로 독립적인 관계를 가지고 있다면, 마할라노비스 거리는 유클리디안 거리와 일치하게 된다.

$$ed_{ij} = \sqrt{(\mathbf{Y_{i,p}} - \mathbf{Y_{j,p}})\mathbf{I}(\mathbf{Y_{i,p}} - \mathbf{Y_{j,p}})^T}$$

($\mathbf{I}$ = 단위행렬(identity matrix)).

추출된 유사사례로부터 부도여부를 측정하는 데에는 다양한 방법이 있다. 가장 쉬운 방식은 민주주의(democracy) 기준에 의하여 동일한 가중치로 투표하는 것이다. 예측하고자 하는 종속변수가 C 개의 분류에 속한다고 할 때, 적절한 비교대상의 수를 선택하여 최고 득표수를 이용하여 예측할 수 있다. 그러나 동일한 가중치를 가정 할 경우 투표 대상의 수를 짝수로 정하면 동수 득표 상황이 발생하여 정확한 결과가 나오지 않을 수도 있다. 이를 보완하기 위하여 투표에 가중치를 적용할 수도 있는데, 보통 가중치를 적용한 투표는 거리의 역수를 가중치에 가산하는 방식을 통해 실제로 측정된 거리를 투표에 반영할 수 있다 (Berry and Linoff, 2003).

2.3 사례기반추론 모형의 비교 대상인 통계 기법과 인공지능 기법

본 연구에서는 사례기반추론 모형의 유용성을 비교평가하기 위하여 두 가지의 대표적인 전통적 통계기법과 두 가지의 대표적 인공지능기법을 선택하였다. 각 모형의 특성은 다음과 같다.

첫째, 판별분석은 두 그룹이 명확하게 구분되어 나뉘어진 상황에서 연구대상이 어떠한 그룹에 속할 것인지를 측정변수를 이용하여 판단하는 통계적 기법이다. 판별분석을 실시하는 경우, 자료가 무작위 표본이어야 하므로 동일한 분포에서 서로 독립된 측정값을 형성하여야 한다. 또한 그룹을 나누는 명확한 기준이 있어야 한다. 그리고 판별분석은 가능한 많은 변수를 고려하고 적절한 변수를 선택하는 과정이 중요하며 다변량 정규분포를 가정하고 있으므로 로지스틱 회귀분석에 비하여 분포의 제한이 있다.

둘째, 로지스틱 회귀분석은 종속변수가 범주형 혹은 명목척도인 경우에 연구대상의 범주를 예측하는 회귀분석 방법이다. 분포에 대한 통계적 가정이 필요하지 않고, 각 재무변수가 어느 정도의 비중으로

모형구축에 공헌하는지 모형화 하기 쉬우므로 결과의 통계적 유의성을 계산할 수 있다. 그러나 변수들 간에 복잡한 비선형 관계가 존재할 경우에는 분석하기가 어렵고 추정한 분포가 실제 현상을 설명하지 못한다는 단점이 있다.

셋째, 인공신경망(artificial neural networks: ANN)모형은 복잡하고 비선형적인 자료에서 패턴을 추출할 수 있고, 특정한 통계적 가정을 전제로 하지 않으며, 상대적으로 적응력이 뛰어나고 견고한(robust) 모형으로 간주되고 있다. 특히 과거의 통계학적 방법에 비해 학습을 통해 분석하므로 분석시간이 짧고 계산비용(computational cost)이 적다는 점이 장점이 있다. 그러나 모형의 결과가 네트워크 상에서 내부가중치를 통해 구현되기 때문에 검증하기가 어려운 것이 문제로 남아있다. 이 가중치들이 왜 최적의 솔루션을 주는지 검증하기 어렵고, 결과값이 지나치게 모형의 구축(training)에 초점이 맞춰지면 모형의 과적합(over fitting)이 발생할 수 있다. 또한 잡음(noise)이 심한 데이터일 경우, 일관성 있고 예측 가능한 성과를 보이지 못하는 점도 문제점으로 지적되어왔다.

마지막으로, 의사결정나무(decision tree)모형은 선형 회귀분석으로 쉽게 나타나지 않는 군집 간의 특성을 미리 정해진 가지 수를 통해 알맞은 변수에 따라 분류하는 모형이다. 의사결정 규칙을 나무구조로 도표화하여 분류와 예측을 수행하기 때문에 다른 방법들(위에서 언급된 신경망, 판별분석, 회귀분석 등)에 비하여 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점이 있다. 분석방법이 간단하고 비모수적 모형이며 예측력이 높고 증명이 쉬워서 많이 사용되지만, 비연속성과 비안정성을 갖고 있고 선형성이나 주효과에 대한 개념이 없기 때문에 선형관계를 분석할 때는 오분류가 커질 위험이 있다.

III. 연구 모형

3.1 데이터 특성

본 연구에 사용한 표본의 선정은 다음과 같은 과정을 거쳐 이루어졌다. 2001년에서 2003년 사이의 국내 중소규모 제조기업의 재무비율을 재무기준일을 기준으로 수집하였다. 부도기업의 경우, 사후적 자료이용의 문제점으로 지적되어왔던 회계연도 직후에 도산하는 경우를 고려하기 위하여 전년도 회계연도 기준으로 3개월 이내에 부도 처리된 기업은 2년 전 재무비율을 사용하였다(전성빈과 김영일, 2001). 본 연구의 자료는 3년 사이에 부도 처리된 기업을 부도로 정의하고 그 기업의 재무비율을 전년도 회계기준으로 500개를 추출한 후, 같은 산업 군에 속하는 정상기업을 무작위로 500개를 추출하여 총 1,000개의 표본을 구성하였다.

학자마다 어떠한 재무비율을 사용할 것인가에 대한 논의도 활발하여 왔다. Altman(1968)은 유동성, 수익성, 레버리지, 지급능력, 활동성을 대표하는 다섯 가지 재무비율을 사용하여 기업부실 예측모형을 개발하였다. 한국은행(2002)은 수익성, 금융비용관련비율, 성장성, 생산성, 안정성으로 분류하여 사용하여왔고, 한국산업은행(2002)은 수익성, 성장성, 생산성, 안정성, 활동성 비율로 구분하였으며, 중소기업진흥공단(2002)은 성장성, 자산·자본의 관계비율, 회전율, 생산성비율로 구분하였다. 대부분의 연구결과는 공통적으로 수익성비율이 판별변수로 높은 분석력을 갖고 있었다. 본 연구에서는 총 132개의 재무변수를 정규화(normalization)하고 다음의 단계를 거쳐 입력변수를 결정하였다. 첫째 단계로서, 개별 변수별로 t 검정(유의수준 5%)을

실시하여 총 59개의 변수를 선택하였다. 둘째 단계로서, SAS 8.2프로그램의 stepwise logistic regression 방법(entry 유의수준 1%, stay 유의수준 5%, link function: logit, stepwise criterion: SBC)을 적용하여 26개의 변수를 선택하였다. 셋째 단계로서, SAS 8.2 Enterprise Miner 프로그램의 Decision Trees(CHAID, CART, entrophy algorithm)에서 한 번 이상 선택된 변수 15개를 추출하였다. 최종적으로 선행연구에서 자주 쓰인 일곱 개의 변수를 선택하였다. 〈표 1〉은 본 연구에서 입력변수로 선택된 변수와 정의를 기술하고 있다.

변수들을 분류하는 기준으로서 변수의 속성을 적용하였으며 그 의미는 다음과 같다.

먼저, 유동성 비율은 레버리지 비율과 함께 재무분석의 비율분석에서 안정성 분석을 위해 산출하는 안정성 비율의 하나로 단기 채무에 충당할 수 있는 유동성 자산이 얼마나 되는가를 나타낸다. 본 연구에서는 X2(매출채권 대 매입채무) 변수가 유동성 비율로 쓰였다.

유동성 비율이 단기채무 관리능력을 나타낸다면, 안정성 비율은 장기적인 관리능력을 나타낸다고 할 수 있으며, X4(순금융비용)과 X5(순운전자본 대 총자본) 변수가 안정성 비율에 해당된다.

수익성 비율은 대부분의 연구에서 기업의 부도 여부를 구분하기 위한 가장 중요한 변수로 간주되는 것으로서 손익계산서 상의 매출액 수익성과 대차대조표 상의 자본수익성으로 구분할 수 있다. 본 연구에서는 X1(매출액영업이익률)이 이 그룹에 속한다.

활동성 비율은 기업이 보유하고 있는 자원(자산, 자본) 등을 얼마나 효율적으로 운용하고 있는 지를 나타내는 비율로서 회전율(turnover ratio)로 표시된다. X3(고정자산회전율)이 활동성 비율에 속하였다.

성장성 비율은 재무비율의 전기에서 당기로의 증가비율을 측정하여 기업의 경영규모 및 기업활동의 성과가 어느 정도 변화 하였는가를 측정하는 지표로서 본 연구에서는 X6(손익분기점률)이 성장성 비율에 속하였다.

생산성 비율은 경영합리화의 척도라고 할 수 있는

〈표 1〉 선택된 입력변수의 산술식과 속성

변수	변수 이름	산술식	분 류
X1	매출액영업이익율	= 영업이익/매출액*100	수익성
X2	매출채권 대 매입채무비율	= 매출채권/매입채무*100	유동성
X3	고정자산회전율	= 매출액/고정자산	활동성
X4	순금융비용	= 이자비용-이자수익	안정성
X5	순운전자본/총자본	= (유동자산-유동부채)/총자본*100	안정성
X6	손익분기점률	= 손익분기점에서의 매출액/매출액*100 → 손익분기점에서의 매출액= (고정비-영업외수익)/ (1-변동비/매출액)	성장성
X7	인건비	= 급여+퇴직급여+급여(제조원가) +퇴직급여(제조원가)	생산성

생산성의 향상으로 얻은 성과에 대한 분배기준이 된다. 이런 생산성비율로는 크게 노동생산성과 자본생산성으로 구분되는데, 인건비 대 매출액 비율, 총자본투자효율 지표를 사용할 수 있다. X7(인건비)가 본 연구에서 생산성 비율에 속하였다.

입력변수 간 상관관계를 살펴보기 위하여 각 변수별로 1000개의 데이터에 대한 상관관계 분석표를 작성하였다. 〈표 2〉의 변수 중 높은 상관관계를 보이는 변수조합은 X3(고정자산회전률)과 X4(순 금융비용), X3(고정자산 회전률)과 X5(순운전자본 대 총자본), X4(순금융비용)과 X6(손익분기점률), X4(순금융비용)과 X7(인건비), X6(손익분기점률)과 X7(인건비)이다. 이 중 X3(고정자산 회전률)과 X4(순 금융비용)의 조합에서는 음의 방향으로 상관관계가 높고, 다른 네 개의 조합에서는 상관관계가 양의 방향으로 높게 나타나고 있다. 상관계수의 절대치를 살펴보면, X6과 X4의 경우 0.6 이상, X7과 X4의 경우 0.5 이상, X7과 X6의 경우 0.7 이상으로 매우 높은 값을 나타내고 있다. 통계적 유의성을 살펴보면, 총 21개의 상관관계 중 12개의 조합이 유의하게 나타나고 있다. 많은 변수 중에서 소수의 입력변수를 선택하는 과정에서 stepwise 방법을 적용하였음에도 최종변수 7개는 상당한 상관성을 가지고 있음을 보여주고 있다. 따라서 객체 간의 거리를 측정함에 있어서 변수 간의 상관성을 고려하는 것이 진정한 의미에서 객체 간 유사성/비유사성을 판단하는 방법이 될 수 있을 것으로 전망된다.

3.2 CBR 모형의 소개

본 연구가 제시하는 모형은 마할라노비스 거리에 기초한 사례기반추론 모형이며, 이 모형과의 비교를 위하여 추가로 유클리디안 거리에 기초한 사례기반추론 모형도 성과를 비교 측정하였다. 〈표 3〉에서 네 가지 사례기반추론 모형에 관한 거리의 계산방법을 설명하고 있다. 첫째 모형은 가장 기본적 모형으로 유클리디안 거리에 기초한 사례기반추론 모형이다. 둘째 모형은 유클리디안 거리에 기초하고 각 변수의 가중치를 차별화한 모형이다. 셋째 모형은 마할라노비스 거리에 기초한 사례기반추론 모형이다. 넷째 모형은 마할라노비스 거리에 기초하고 각 변수의 가중치를 차별화한 사례기반추론 모형이다.

본 연구에서 제시하는 네 가지 모형이 거리측정에 어떠한 차이를 보이는지 다음의 예시를 통하여 살펴

〈표 2〉 입력변수 간 상관관계 분석표: Pearson's product moment correlation coefficient test

	X1	X2	X3	X4	X5	X6	X7
X1	1.00						
X2	0.081*	1.00					
X3	0.013	0.074*	1.00				
X4	0.044	0.110*	0.208*	1.00			
X5	0.011	0.038	0.248**	0.037	1.00		
X6	0.068*	0.175**	0.049	0.6462**	0.005	1.00	
X7	0.020	0.132**	0.094*	0.5306**	0.060	0.776**	1.00

(*: p-value < 0.05; **: p-value < 0.0001)

〈표 3〉 네 가지 CBR 모형의 거리계산방법

구 분	계 산 방 법
Eucli-w/o w	가중치를 고려하지 않은 유클리디안 거리 모형 $w/o\_ed_{ij} = \sqrt{\sum_i (Y_{i,p} - Y_{j,p})(Y_{i,p} - Y_{j,p})^T}$
Eucli-with w	가중치를 고려한 유클리디안 거리 모형 $with\_ed_{ij} = \sqrt{\sum_i w_i (Y_{i,p} - Y_{j,p})(Y_{i,p} - Y_{j,p})^T}$
Mahala-w/o w	가중치를 고려하지 않은 마할라노비스 거리 모형 $w/o\_md_{ij} = \sqrt{\sum_i (Y_{i,p} - Y_{j,p})\Gamma^{-1}(Y_{i,p} - Y_{j,p})^T}$
Mahala-with w	가중치를 고려한 마할라노비스 거리 모형 $with\_md_{ij} = \sqrt{\sum_i w_i (Y_{i,p} - Y_{j,p})\Gamma^{-1}(Y_{i,p} - Y_{j,p})^T}$

보고자 한다. 〈표 4〉는 다섯 기업에 대한 세 가지 변수의 값을 보여주고 요약하고 있다.

〈표 4〉 다섯 기업에 대한 세 가지 변수값

기업 No.	X1	X2	X3
1	0.72	0.18	1.21
2	0.55	0.45	1.54
3	0.31	0.97	0.98
4	0.24	0.88	0.75
5	1.25	0.45	0.87

예를 들어, 기업 1과 기업 2의 거리를 계산함에 있어 네 가지 모형은 다음과 같이 차이를 보이고 있다.

첫째, Eucli -w/o w 모형에 의한 기업 1과 기업 2의 거리는 다음과 같이 계산된다.

$$\begin{bmatrix}0.72-0.55 & 0.18-0.45 & 1.21-1.54\end{bmatrix} \times \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}0.72-0.55\\ 0.18-0.45\\ 1.21-1.54\end{bmatrix}$$

둘째, 변수 X1, X2, X3에 대한 가중치를 각각 .30, .45, .25라고 가정했을 때, Eucli -with w 모형에 의한 기업 1과 기업 2의 거리는 다음과 같이 계산된다.

$$\begin{bmatrix}.30(0.72-0.55) & .45(0.18-0.45) & .25(1.21-1.54)\end{bmatrix} \times \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}0.72-0.55\\ 0.18-0.45\\ 1.21-1.54\end{bmatrix}$$

셋째, Mahala -w/o w 모형에 의한 기업 1과 기업 2의 거리는 다음과 같이 계산된다.

$$\begin{bmatrix}0.72-0.55 & 0.18-0.45 & 1.21-1.54\end{bmatrix} \times \begin{bmatrix}0.1304 & -0.0695 & 0.0009\\ -0.0695 & 0.0871 & -0.0444\\ 0.0009 & -0.0444 & 0.0782\end{bmatrix}^{-1} \times \begin{bmatrix}0.72-0.55\\ 0.18-0.45\\ 1.21-1.54\end{bmatrix}$$

넷째, 변수 X1, X2, X3에 대한 가중치를 각각 .30, .45, .25라고 가정했을 때, Mahala-with w

모형에 의한 기업 1과 기업 2의 거리는 다음과 같이 계산된다.

$$\left[.30(0.72-0.55)\quad .45(0.18-0.45)\quad .25(1.21-1.54)\right] \times \begin{bmatrix} 0.1304 & -0.0695 & 0.0009 \\ -0.0695 & 0.0871 & -0.0444 \\ 0.0009 & -0.0444 & 0.0782 \end{bmatrix}^{-1} \times \begin{bmatrix} 0.72-0.55 \\ 0.18-0.45 \\ 1.21-1.54 \end{bmatrix}$$

위에서 예시한 바와 같이, 기업 1과 기업 2의 거리는 네 가지 모형에 의하여 다르게 계산된다. 〈표 5〉는 네 가지 모형을 적용하여 기업 1과 나머지 기업 간의 거리를 계산한 결과를 보여주고 있다. 괄호 안에는 각 기업과 기업 1의 거리가 가까운 순서를 보여주고 있다. 각 모형에서 기업 1과 가까운 순서가 다르게 나타남을 확인할 수 있다. Eucli -w/o w 모형의 경우 기업 1과 가까운 이웃의 순서는 기업 2, 5, 3, 4이고, Eucli -with w 모형의 경우 기업 2, 5, 4, 3의 순이며, Mahala -w/o w 모형의 경우 기업 4, 2, 5, 3의 순이며, Mahala-with w 모형의 경우 기업 2, 4, 5, 3의 순으로 나타나고 있다. 따라서 training data set에서 가장 가까운 이웃을 선정함에 있어서 네 모형이 서로 다른 이웃을 선정하게 되고, testing data set에 대한 예측을 하는 경우에도 서로 다른 이웃의 조합에 의하여 목표변수를 예측함으로 정분류율에 차이를 발생시킨다.

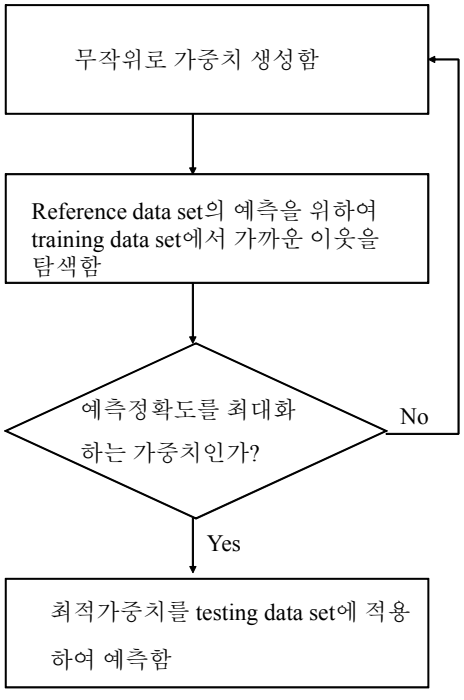
위의 예에서는, Eucli -with w 모형과 Mahala-with w 모형에서 세 가지 변수 X1, X2, X3에 대하여 동일한 가중치(각각 .30, .45, .25)를 부여한 경우에도 객체 간 거리가 상이하게 결정됨을 보여주고 있다. 예측 정확도를 최대화 하는 가중치를 결정함에 있어서, training data set을 training data set과 reference data set으로 분할하고, reference data set의 예측을 위한 이웃을 training data set에서 탐색하므로, 변수에 대한 가중치는 두 모형에서 당연히 다르게 결정될 것이고, testing data set에 대한 예측에도 서로 다른 이웃의 조합이 사용되게 된다.

가중치를 고려한 유클리디안 거리 모형과 가중치를 고려한 마할라노비스 거리 모형에서 입력변수 별 최적 가중치를 찾는 과정은 〈그림 2〉에서 표현되고 있다. 1,000개의 관측치를 600: 400의 training data set과 testing data set으로 분할하였으며, 600개의 training data set은 다시 350: 250의 training data set과 reference data set으로 분할하였다. 각 변수 별 가중치를 무작위로 생성하여 reference data set의 예측을 위한 nearest neighbors를 training data set에서 탐색하였다. 이 과정은 reference data set에서의 예측정확도를

〈표 5〉 네 가지 모형에 의한 기업 1과 기업 2, 3, 4, 5의 거리 계산 결과 (가까운 순위)

	Eucli-w/o w	Eucli-with w	Mahala-w/o w	Mahala-with w
기업 No.	1	1	1	1
2	0.2107 (1^{st})	0.0687 (1^{st})	6.8852 (2^{nd})	2.4450 (1^{st})
3	0.8451 (3^{rd})	0.3445 (4^{th})	9.7041 (4^{th})	5.0613 (4^{th})
4	0.9320 (4^{th})	0.3425 (3^{rd})	5.8178 (1^{st})	2.4817 (2^{nd})
5	0.4694 (2^{nd})	0.1459 (2^{nd})	8.7452 (3^{rd})	3.2567 (3^{rd})

최대화 하는 변수의 가중치를 찾을 때가지 반복하였다. 이 최적 가중치를 testing data set에 적용하여 최종적으로 모형의 예측력을 평가하였다. 반면에 가중치를 고려하지 않은 유클리디안 거리 모형과 가중치를 고려하지 않은 마할라노비스 거리 모형에서는 600개의 training data set으로부터 400개의 testing data set에 속하는 개별 관측치의 거리를 구하여 가장 가까운 이웃(nearest neighbor)을 선정하였다.



〈그림 2〉 가중치를 고려한 사례기반추론 모형

Testing data set의 개별 관측치에 대한 부도여부를 예측할 때 고려된 nearest neighbors의 수는 k = 5, 15, 25 등을 이용하여 예비조사를 하여 본 결과, k = 5는 예측력이 다소 떨어졌고, k = 15와 k = 25의 경우에는 커다란 차이를 보이지 않아, 최종적으로 k = 15를 사용하였다. 15개 nearest neighbors를 합성함에 있어서는 1위의 이웃이 15위의 이웃에 비하여 월등하게 좋은 예측력을 준다고 확신하기 힘들므로 democracy 원칙을 적용하였다. 즉, 투표 알고리즘은 다음과 같다.

$$\hat{Y}_i = \begin{cases} 1, & \text{if } N_{for\,\hat{Y}=1} > N_{for\,\hat{Y}=0} \\ 0, & \text{otherwise} \end{cases}$$

where 1 = 부도 and 0 = 정상 (N = 부도여부를 나타내는 $\hat{Y}$의 개수).

다음 단계로서, 투표에 근거하여 결정한 예측 값($\hat{Y}_i$)과 실제 부도 여부를 비교하여 예측정확도를 산출한다.

$$I_i = \begin{cases} 1, & \text{if } Y_i = \hat{Y}_i \\ 0, & \text{otherwise} \end{cases}$$

where I_i = indicator variable for subject I.

모형의 목적함수는 비교대상이 되는 관측치들의 indicator variables의 합, 즉 $\sum I_i$를 최대화하는 것이며, 이 때의 각 변수 별 가중치가 의사결정변수의 역할을 한다.

최적가중치를 찾아내는 작업은 기술적으로 용이하지 않았다. 매 번 일곱 개의 가중치를 무작위로 생성하고 가중치의 합이 1.0이 되도록 조정한다. 이 가중치를 적용하여 training data set에 해당하는 350개 관측치와 예측의 대상이 되는 reference data set의 250개 관측치 간 거리 메트릭스(350 × 250)를 만든다. Reference data set의 각 관측치

별로 training data set의 350개 관측치를 가까운 순서로 정렬시키고, 가장 가까운 15개 이웃의 부도여부를 참조하여 투표를 실시한다. 투표에 의한 예측결과를 validation data set의 실제 부도여부와 비교하여 1 또는 0의 값을 갖는 indicator variable을 생성하고, 이를 합산하여 예측정확도(즉, 정분류율)를 산출하는 과정이 1번의 loop를 형성하는 것이다. 이처럼 1회의 loop 마다 거리 메트릭스를 산출하고 (특히, 마할라노비스 거리의 경우에는 공분산의 역행렬까지 고려한) 범주형 변수의 이분적 판단의 결합을 요하는 복잡한 목적함수 최적화작업은 Excel의 Solver나 Premium Solver 또는 유전자 알고리즘을 이용한 방법으로는 적절하게 구동이 되지 않았다. 전문 공학용 해찾기인 GAMS Solver를 쓴다면 향후 연구에서는 도전해볼수도 있으리라고 사료된다. 그리하여, 본 연구에서는 현실적인 대안으로 5,000개의 가중치 조합을 랜덤하게 산출하고, 5,000번의 시뮬레이션을 통하여 최적가중치에 가깝다고 추정될 수 있는 near optimal weights의 조합을 탐색하였다.

부도예측 문제를 위에서 설명된 네 가지의 사례기반추론 모형과 함께 전통적 통계 기법인 로지스틱 회귀분석과 다변량 판별분석, 그리고 인공지능 기법인 인공신경망과 의사결정나무를 이용하여 실험분석하였다.

Ⅳ. 실험 결과

앞 장에서 언급한 총 여덟 개의 예측모형에 관하여 training data set과 testing data set의 비율을 6: 4로 분할하여 training data set에서 얻어진 학습모형을 testing data set에 적용하여 정분류율(correct classification ratio)을 산출하였다.

〈표 6〉은 training data set에 대한 입력변수의 가중치를 보여주고 있다. 가중치를 고려하지 않은 두 모형은 모든 변수에 결과적으로 동일한 가중치를 부여한 것과 같다. 가중치를 차별화 한 나머지 두 모형의 경우, 변수 간 공분산구조가 독립적이냐 종속적이냐는 매우 상이한 접근법에서 출발하였으므로 당연히 가중치 간 공통점이 발견되지 않았다. CBR-Eucli -with w 모형에서 변수들의 중요도 순위는 {3, 4, 5, 1, 2, 7, 6} 인데 반하여, CBR-Mahala -with 모형에서의 중요도 순위는 {3, 7, 6, 5, 2, 1, 4}로 나타났다. 변수 간의 공분산구조를 제대로 반영하느냐 혹은 이를 무시하고 단위행렬로 대체하

〈표 6 〉 CBR 모형의 변수별 최적가중치

	CBR-Eucli-w/o	CBR-Eucli-with w	CBR-Mahala-w/o	CBR-Mahala-with
X1	n/a	0.1880	n/a	0.1476
X2	n/a	0.1569	n/a	0.0183
X3	n/a	0.1486	n/a	0.0513
X4	n/a	0.2447	n/a	0.1273
X5	n/a	0.2193	n/a	0.2352
X6	n/a	0.0066	n/a	0.2853
X7	n/a	0.0359	n/a	0.1350

는냐는 최적의 가중치조합을 결정하는데 중대한 영향을 미치고 있음을 확인할 수 있었다.

〈표 7〉은 각 모형의 정분류율을 보여주고 있다. 실험결과를 통하여 발견한 점은 다음과 같이 요약할 수 있다.

첫째, 가장 주목할만한 점은 가중치를 고려한 마할라노비스 모형(Mahala-with w)이 training data set과 testing data set 모두의 경우에 (0.7520, 0.7475로서) 다른 비교 모형 보다 높은 정분류율을 보여주고 있는 것이다. 두 데이터 간의 정분류율의 차이(0.0045)도 로지스틱 회귀모형(0.0050)과 함께 가장 적어 비교적 안정된 예측력을 보여주고 있음을 알 수 있다.

둘째, 가중치를 고려하지 않은 마할라노비스 모형은 training data set에서는 (0.7400으로서) 높은 적합도를 보여주고 있으나, 새로운 데이터에 대하여 평가할 시에는 그 예측도가 (0.6975로서) 상당히 떨어짐을 보여주고 있다.

셋째, 유클리디안 거리에 기초를 하던 마할라노비스 거리에 기초를 하던 두 경우 모두 가중치를 차별화 한 모형이 변수의 가중치를 일률적으로 한 모형보다 예측력이 좋게 (0.7475 vs. 0.6975, 0.6650 vs. 0.5900으로서) 나타나고 있었다.

넷째, 기존의 연구에서 부도예측에 널리 사용되어 온 인공신경망 모형, 로지스틱 회귀 모형, 판별분석 모형은 가중치를 고려한 마할라노비스 모형 보다는 예측력이 약간 떨어지지만 두번째로 좋은 예측력을 보여주고 있으며, training data set과 testing data set간의 예측력 차이도 적어 상당히 안정된 결과를 보여주고 있었다.

다섯째, 의사결정나무 모형은 연속형 입력변수를 이용하여 범주형 목표변수를 예측하는 과제에 있어서는 분리의 경계부근에서 예측력이 떨어진다는 기존의 단점을 재확인할 수 있었다. 가중치를 고려하지 않은 사례기반추론 모형과 함께 가장 낮은 예측력을 보여주고 있었다.

마지막으로, 〈표 2〉의 변수 간 상관관계 분석에서 확인할 수 있는 것처럼, 입력변수 간에 선형 연관성이 확연히 존재함에도 불구하고 이를 모두 무시하고 유클리디안 거리를 이용하여 사례기반추론 모형을 적용한 경우에는 예측력이 기존에 많이 사용되고 있는 인공신경망 모형, 로지스틱 회귀 모형, 판별분석 모형 보다 떨어짐을 확인할 수 있었다.

Testing data set에 대한 예측정확도를 대상으로

〈표 7〉 각 모형의 정분류율 결과

	Training data set	Testing data set	difference
CBR-Eucli-w/o w	0.6200	0.5900	0.0300
CBR-Eucli-with w	0.7000	0.6650	0.0350
CBR-Mahala-w/o w	0.7400	0.6975	0.0425
CBR-Mahala-with w	0.7520	0.7475	0.0045
Neural (인공신경망)	0.7350	0.7125	0.0225
Logit (로지스틱 회귀분석)	0.7200	0.7150	0.0050
MDA (판별분석)	0.7217	0.7100	0.0117
DT (의사결정나무)	0.6883	0.6200	0.0683

McNemar test를 실시하였고 그 결과는 〈표 8〉에 요약되어 있다. 각 모형의 예측력 패턴에 대한 통계적 유의성 검증 결과, 가중치를 고려한 마할라노비스 모형은 인공신경망 모형과 로지스틱 회귀분석 모형을 제외한 나머지 다섯 가지 모형과는 유의한 차이를 보이고 있었다. 사례기반추론 모형 네 가지 간에는 분류 프로세스가 상당히 다르게 나타나고 있음을 알 수 있었다. 인공신경망 모형, 로지스틱 회귀분석 모형, 판별분석 모형 간에는 분류 패턴이 상이함을 확인할 수 없었다.

Ⅴ. 결론 및 시사점

기존의 문헌연구를 살펴보면 기업의 부도예측에 관한 대다수의 연구들은 현재 인공지능 기법에 의존하고 있으며 이 중에서도 인공신경망 기법이 주류를 이루고 있음을 알 수 있다. 모형의 예측정확도를 비교평가 함에 있어서는 전통적인 통계기법인 로지스틱 회귀분석과 판별분석이 가장 많이 사용되어 왔다. 본 연구는 상대적으로 개발이 덜 되어온 사례기반추론 모형을 진화시켜 예측력을 향상시켜 보고자 시도하였다. 사례기반추론 모형의 핵심은 예측 대상인 관측치와 진정으로 유사한 이웃 관측치들을 피예측 집단으로부터 선별하여 내는 데에 있다. 기존의 연구들을 조사하던 중 우리는 모든 사례기반추론 모형이 단순히 유클리디안 거리에 기초하고 있음을 파악하였다. 따라서, 본 연구는 변수 간의 공분산구조를 감안하여 거리를 계산하는 마할라노비스 거리를 적용할 만한 가치가 있을 것이라고 판단하였다. 그리하여, 마할라노비스 거리를 산출하고, 최근 예측

〈표 8〉 McNemar test 결과

	Maha-no	Eucli-w	Eucli-no	Neural	Logit	MDA	DT
Maha-w	9.025 (0.003)**	9.570 (0.002)**	24.800 (0.000)**	2.284[1] (0.131)[2]	1.694 (0.193)	2.925 (0.087)*	20.661 (0.000)**
Maha-w/o		1.321 (0.250)	12.000 (0.001)**	0.338 (0.561)	0.444 (0.505)	0.213 (0.644)	7.087 (0.008)**
Eucli-w			7.645 (0.006)**	2.919 (0.088)*	3.112 (0.078)*	2.535 (0.111)	2.094 (0.148)
Eucli-w/o				15.258 (0.000)**	16.223 (0.000)**	14.926 (0.000)**	0.688 (0.407)
Neural					0.000 (1.000)	0.000 (1.000)	10.537 (0.001)**
Logit						0.023 (0.880)	11.802 (0.001)**
MDA							9.879 (0.002)**

1. Chi-square value / 2. (p-value)
*significant at 0.10 / **significant at 0.05

문제에서 종종 쓰이고 있는 변수의 가중치를 차별화하는 방법도 함께 접목하였다.

본 연구는 시계열 표본의 한계로 인하여 2001년에서 2003년까지의 국내 중소규모 제조업체 1,000개 기업을 분석 대상으로 국한하였다. Training data set과 testing data set을 6: 4로 분할하여 training data set에서 모형을 학습시키고, 이 학습된 모형을 testing data set에 적용하여 모형의 예측력을 평가하였다. 본 연구는 가중치를 고려하지 않은 마할라노비스 거리 기초의 사례기반추론 모형과 가중치를 고려한 마할라노비스 거리 기초의 사례기반추론 모형을 소개하였다. 비교의 대상으로는 가중치를 고려하지 않은 유클리디안 거리 기초의 사례기반추론 모형, 가중치를 고려한 유클리디안 거리 기초의 사례기반추론 모형, 인공신경망 모형, 의사결정나무 모형, 로지스틱 회귀분석 모형, 그리고 판별분석 모형을 적용시켰다. 실험 결과, 본 연구가 제시한 가중치를 고려한 마할라노비스 거리에 기초한 사례기반추론 모형의 예측력이 가장 우수하게 나타나고 있었으며, McNemar test를 통하여 대부분의 비교 대상 모형과 분류 패턴에 차별성이 있음을 확인할 수 있었다.

한 가지 주목할 만한 점은 판별분석에서도 관측치 간의 거리를 계산함에 있어서 마할라노비스 거리를 사용한다는 점이다. 하지만, 판별분석에서는 모든 관측치의 선형결합함수로 예측대상이 되는 관측치의 목표변수를 결정한다. 이와는 대조적으로 사례기반추론에서는 가장 가까운 이웃이라고 판단되는 관측치 소수집단의 결정에만 마할라노비스 거리를 이용함으로써 모든 관측치에 근거하여 예측값을 결정하는 방식으로부터 차별화를 시도하였고, 그 결과는 상당히 고무적으로 나타났다.

본 연구의 결과를 일반화 하는 데는 아직 미흡한 점이 여러 가지 있다. 본 연구는 사례기반추론에 마할라노비스 거리를 적용한 탐색적 연구로서 그 의의를 찾을 수 있을 것이다. 본 연구의 한계점은 동시에 향후 연구의 과제가 될 수도 있을 것이다.

첫째, 최적의 변수 가중치를 결정함에 있어서, 경영학 연구분야에서 통상적으로 쓰이는 Excel Solver나 유전자 알고리즘을 이용한 해찾기로는 계산의 복잡성 때문에 적용시킬 수 없었다. 따라서 이에 대한 기술적 대안으로 5,000개의 가중치 조합을 무작위로 생성시키고, 5,000번의 시뮬레이션을 통하여 최적해에 근사하다고 판단되는 가중치를 구하여 모형에 적용시켰다. 만약 향후 연구에서 GAMS Solver와 같은 전문 공학용 해찾기 프로그램을 구입하여 적용할 수 있다면 최적의 가중치를 구하고 연구모형의 예측력을 한 차원 더 향상 시킬 수 있을 것이다.

둘째의 한계점은 위에서 언급한 컴퓨터 해찾기 프로그램의 용량과 상당한 관계가 있다. 사례기반추론 모형에서 가장 가까운 이웃의 수를 결정함에 있어서, 본 연구는 k = 5, 15, 25의 세 경우만 예비적으로 실시하여 보고, k = 15를 모형에 적용하였다. 전문 공학용 해찾기 프로그램을 이용하여 모형을 설계한다면 목적함수인 정분류율을 최대화 하는 이웃의 수도 최적으로 찾아낼 수 있을 것이다.

셋째로, 본 연구에서는 가중치 결정에 정량적인 접근방법만을 적용하였다. 비록 하나의 기존 연구(Park and Han, 2002)에서만 정성적인 가중치 결정이 좋은 결과를 나타낸다고 보고하였지만, 기존의 방법을 보강하여, 가중치 결정에 있어 정성적 접근방법을 결합하고 거리 측정에 있어 마할라노비스 거리를 적용하여, 객체 간의 유사성/비유사성을 판단한다면 의미있는 연구가 될 것으로 사료된다.

마지막으로, 본 연구에서는 부도기업을 정상기업

으로 분류하였을 경우의 손실비용과 정상기업을 부도기업으로 분류하였을 경우의 손실비용을 동일하게 취급하였다. 현실적으로는 부도가 날 기업을 정상기업으로 예측하여 대출하고 손실이 발생했을 때의 비용이, 정상기업을 부도가 날 기업으로 오분류하여 대출거절을 하였을 경우의 비용 보다 훨씬 더 크기 때문에 이 비용간의 현실적인 비율을 산업 별로 조사하여 감안한다면, 실제문제에의 적용성을 높여줄 수 있을 것이다.

참고문헌

김성호, 조성빈, "마할라노비스 거리를 이용한 자료융합전략의 성과측정," **경영학연구**, 제34권, 제6호(2005), pp.1853-1867.

중소기업진흥공단, "중소기업경제·경영지표," **서울: 중소기업진흥공단**. (2002).

전성빈, 김영일, "도산예측모형의 예측력 검증," **회계저널**, 제10권, 제1호(2001), pp.151-182.

한국산업은행, "기업재무분석," **서울: 한국산업은행**. (2002).

한국은행, "기업경영분석," **서울: 한국은행**. (2002).

한국은행, "통계 데이터베이스," **서울: 한국은행**. (2006).

Altman, E.I., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance* Vol.23, No. 4(1968), pp589-609.

Anandarajan, M., P. Lee, and A. Anandarajan, "Bankruptcy Prediction of Financially Stressed Firms: An Examination of the Predictive Accuracy of Artificially Neural Networks," *International Journal of Intelligent Systems in Accounting, Finance and Management* Vol.10(2001), pp.69-81.

Baldwin, J. and Glezen, G.W., "Bankruptcy prediction using quarterly financial statement data," *Journal of Accounting, Auditing, & Finance* (1992), pp.269-282.

Beaver, W.H., "Financial Ratios as Predictors of Failure. Empirical Research in Accounting: Selected Studies," *Journal of Accounting Research* Vol.4, No.1(1966), pp.71-111.

Berry, M. A. and C.S. Linoff, Data Mining Techniques, Wiley (2003).

Boritz, J. E. and D.B. Kennedy, "Predicting Corporate Failure Using a Neural Network Approach," *Intelligent Systems in Accounting, Finance, and Management* Vol.4 (1995), pp.95-111.

Bryant, S.M., "A Case-Based Reasoning Approach to Bankruptcy Prediction Modeling," *Intelligent Systems in Accounting, Finance and Management* Vol.6(1997), pp.195-214.

Charlambous, C., A. Charitou, and F. Kaourou, "Comparative Analysis of Artificial Neural Network Models: Application in Bankruptcy Prediction," *Annals of Operations Research* Vol.99(2000), pp.403-425.

Coates, P.K. and L.F. Fant, "A Neural Network Approach to Forecasting Financial Distress," *Journal of Business Forecasting* Vol.3, No.4(1992), pp.8-12.

Curram, S.P. and J. Mingers, "Neural Networks, Decision Tree Induction and Discriminant Analysis: An Empirical Comparison," *Journal of the Operational Research Society* Vol. 45, No.4(1994), pp.440-450.

Desai, V.S., J.N. Crook, and G.A. Overstreet, "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union

Environment," *European Journal of Operational Research* Vol.95(1996), pp.24-37.

Elhadi, M.T. and T. Vamos, "An IR-CBR Approach to Legal Indexing and Retrieval in Bankruptcy Law," *Tenth proceedings in Database and Expert Systems Applications* (1999), pp.769-774.

Elhadi, M.T., "Bankruptcy Support System: Taking Advantage of Information Retrieval and Case-Based Reasoning," *Expert Systems With Applications* Vol.18(2000), pp.215-219.

Fletcher, D. and E. Goss, "Forecasting With Neural Networks," *Information & Management* Vol.24(1993), pp.159-167.

Jo, H., I. Han, and H. Lee, "Bankruptcy Prediction Using Case-Based Reasoning, Neural Networks, and Discriminant Analysis," *Expert Systems With Applications* Vol.13, No.2 (1997), pp.97-108.

Klersey, G.F. and M.T. Dugan, M.T., "Substantial Doubt: Using Artificial Neural Networks to Evaluate Going Concern," *In Advances in Accounting Information Systems* Vol.9 (1995), JAI Press: Greenwich: CT, pp. 267-273.

Lee, K., D. Booth, and P. Alam, "A Comparison of Supervised and Unsupervised Neural Networks in Predicting Bankruptcy of Korean Firms," *Expert Systems with Applications* Vol.29(2005), pp.1-16.

Ohlson, J., "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research* Vol.18, No.1(1980), pp.109-131.

Park, C. and I. Han, "A Case-Based Reasoning with the Feature Weights Derived by Analytic Hierarchy Process for Bankruptcy Prediction," *Expert Systems With Applications* Vol.23, No.3(2002), pp.255-264.

Pendharkar, P.C., "A Threshold-Varying Artificial Neural Network Approach for Classification and its Application to Bankruptcy Prediction Problem," *Computers & Operations Research* Vol.32(2005), pp.2561-2582.

Pompe, P.P.M. and A.J. Feelders, "Using Machine Learning, Neural Networks, and Statistics to Predict Corporate Bankruptcy," *Microcomputers in Civil Engineering* Vol.12 (1997), pp.267-276.

Treacy, W. and M. Carey, "Credit Risk Rating at Large US Banks," *Journal of Banking and Finance* Vol.24(2000), pp.167-201.

Case-Based Reasoning Approaches by Considering Variable Covariance Structure and Variable Weight: Corporate Bankruptcy Prediction

Hyojung Hong* · Sungbin Cho**

Abstract

Case-Based Reasoning (CBR) Approach has been used to predict future events by comparing past similar events. This paper proposes case-based reasoning models for corporate bankruptcy prediction problem and compares its performance with traditional statistical and artificial intelligence techniques. The Statistical techniques include multiple discriminant analysis and logistic regression, whereas artificial intelligence techniques include neural networks and decision trees. The purpose of this study is to develop a case-based reasoning model that reflects the covariance structure of variables and considers variable weights.

The literature review reveals that existing approaches in the area of CBR have mainly used the Euclidean distance for measuring dissimilarity between subjects and further differentiated variable weights for improving prediction accuracy. The distances between subjects are fairly different depending on whether the covariance structure of variables as well as variable weights are considered or not. Thus, we introduce four CBR models: Euclidean distance-based CBR without variable weight; Euclidean distance-based CBR with variable weight; Mahalanobis distance-based CBR without variable weight; Mahalanobis distance-based CBR with variable weight. We incorporate the covariance structure of input variables because the input variables of the model are commonly correlated with each other.

The data collected for analysis are the financial variables from the small-and medium-sized manufacturing firms in Korea during the fiscal year of 2001 to 2003. We selected 500 bankrupt firms and 500 non-bankrupt firms. After normalizing 132 variables, we obtained 15

* Research Fellow, School of Business, Sogang University
** Associate Professor, School of Business, Sogang University

variables by applying t-test, stepwise logistic regression, and decision trees induction. Then, by consulting former research, we finally selected seven variables, which include the various aspects of corporate activities such as profitability, liquidity, activity, stability, growth, and productivity.

The data are divided into the training data and the testing data set. The training data set is then divided into the training data set and the reference data set. Using 5,000 Monte Carlo simulations, the optimal weights of variables are determined while the nearest neighbors are searched from the training data set for predicting the bankruptcy of the reference data set. Then, these weights are used to predict the bankruptcy of the testing data set.

The experiment results indicate that the CBR based on the variable covariance structure and variable weight produces a higher correct classification ratio than other CBR models and currently-in-use approaches. The future study might improve practicality by applying different costs for the false positive prediction (predicting bankrupt for non-bankrupt firms) and the false negative prediction (predicting non-bankrupt for bankrupt firms).

Key words: Case-based Reasoning, Bankruptcy Prediction, Variable Covariance Structure, Variable Weight