金融工學硏究, 제18권 제3호 (2019. 9.) 131~152

기업부도예측과 기계학습

김형준* · 류두진** · 조 훈***

<요 약>

기업부도에 대한 예측은 경제 전반의 각 분야에서 다양하게 활용된다. 기업은 예측 결과를 토대로 경영상태를 진단하고 경영전략을 수립하며, 투자자는 신용위험을 관리하고 투자전략을 조정한다. 정부가 거시건전성 정책을 만들고 금융제도를 설계·개선하는 데에도 이러한 기업부도예측 방법론을 근거로 활용한다. 현재 기업부도예측은 수리통계 모형뿐만 아니라, 기계학습 알고리즘(machine learning algorithm)을 활용한 첨단 금융공학의 일선에 있다. 본 연구는 지금까지 진행된 기업부도예측에 관한 연구를 살펴보고, 통계적 모형과 기계학습 알고리즘의 대표적 방법론을 소개함으로써 관련 분야를 조망한다. 기업부도예측을 위한 주요 통계적 모형은 3세대로 구분할 수 있으며, 각 세대는 대표적으로 판별분석(discriminant analysis), 이항반응모형(binary response model), 위험모형(hazard model)을 사용하여 연구되었다. 기계학습 알고리즘에는 주로 분류방법론이 사용되었으며, SVM(support vector machine), 의사결정나무(decision tree), 인공신경망(artificial neural networks) 알고리즘이 대표적이다. 기계학습 방법론의 발달은 금융 분야의 혁신을 더욱 가속화하며, 이에 따라 새로운 금융서비스의 출현과 데이터 유통에 기반한 데이터 경제의 부상이 뒤따를 것으로 전망된다.

주제어: 기계학습, 기업부도, 부도예측, 신용위험, 인공신경망

논문접수일 : 2019. 07. 29. 1차 수정일 : 2019. 08. 28. 게재확정일 : 2019. 09. 04.

본 논문은 “빅데이터 분석 방법론과 데이터 과학자(2017)”의 내용 중 김형준 교수가 작성한 부분을 수정·보완 및 확장하여 작성되었습니다. 본 연구와 관련하여 유익한 조언을 주신 송완영 박사님, 안세륭 박사님, 그리고 익명의 심사위원님들께 감사드립니다.

* 주저자, 영남대학교 경영대학 경영학과 교수, 053-810-2740, hkim@yu.ac.kr

** 교신저자, 성균관대학교 경제대학 경제학과 정교수, 02-760-0429, sharpjin@skku.ac.kr

*** 공동저자, 한국과학기술원 경영대학 금융전문대학원 교수, 02-958-3413, hooncho@kaist.ac.kr

Ⅰ. 서 론

기업부도(corporate default)에 대한 예측은 기업의 재무제표와 관련된 자료는 물론, 산업 및 거시경제에 영향을 미치는 변수를 활용하여 기업의 신용위험(credit risk)을 측정하는 것을 의미한다. 이렇게 측정된 기업의 부도위험은 경제 전반의 각 분야에서 매우 중요하게 활용된다. 기업은 이를 토대로 경영상태를 진단하고 경영전략을 수립하며, 투자자는 신용위험을 관리하고 투자전략을 조정한다. 정부는 기업의 신용위험을 관리할 수 있는 거시건전성 정책을 만들고, 금융 규제 및 제도를 설계·개선하는 데에도 이러한 기업부도예측 방법론을 활용한다. 현재 기업부도예측은 통계적 방법론뿐만 아니라, 기계학습 알고리즘(machine learning algorithm)을 활용한 첨단 금융공학의 일선에 있다. 본 연구는 지금까지 진행된 기업부도예측에 관한 연구를 살펴보고 대표적 방법론을 소개함으로써 관련 분야를 조망한다.

기업부도에 대한 예측은 학계와 업계의 많은 관심 속에서 오래전부터 다양하게 연구되었다. Beaver(1966)와 Altman(1968)은 판별분석(discriminant analysis)을 이용한 축약형 모형(reduced-form model)을 제시하고, 신용점수를 통해 부도위험 정도를 순서대로 나타내었다. Ohlson(1980)과 Zmijewski(1984)는 기업의 부도위험을 로짓(logit) 회귀분석과 프로빗(probit) 회귀분석을 이용하여 분석하였다. 이러한 2진 모형(binary model)을 활용하면 다음 1기간 동안 기업의 부도확률을 계산할 수 있다. Shumway(2001)는 위험모형(hazard model)을 이용한 기간 분석(duration analysis)을 통해 기업의 부도위험을 측정하였으며, 이러한 접근방법은 기존의 1기간 모형(single-period model)을 다기간 모형(multi-period model)으로 확장하고 시간 경과에 따른 기업의 부도위험 변화를 예측하였다. 그리고 Nam et al.(2008)은 Shumway(2001) 방법론의 기저위험(baseline hazard)에 거시경제변수를 포함한 시계열 변수를 활용할 수 있도록 확장하였다. Campbell et al.(2008)은 기업부도위험을 다중로짓모형(multiple logit model)으로 분석하고 다기간에 대한 기업의 부도확률을 계산하였다. Bonfim(2009)은 기업부도의 예측에 있어서 거시경제조건의 중요성을 강조하였다. Dakovic et al.(2010)은 기업부도위험 예측모형에 산업별 비관측 이질성(unobserved heterogeneity)의 개념을 도입하고, 이러한 방법이 기존 Altman방식의 변수들보다 예측력이 우수하다고 주장했다. Finglewski et al.(2012)은 reduced-form Cox intensity model을 사용하여 일반적인 경제적 상황이 신용위험에 미치는 영향을 시험하였다. Kukuk and Rönnberg(2013)는 혼합로짓모형(mixed logit model)을 사용하여 기업부도위험을 예측하였다. Tian et al.(2015)은 부도예측을 위하여 LASSO (least absolute shrinkage and selection operator) 변수선택방법

을 사용하였으며, 이러한 방법이 기존 모형보다 예측력이 우수함을 보였다. Jessen and Lando(2015)는 변동성조정 부도거리(volatility-adjusted distance-to-default) 방법론을 이용하여 기업부도위험을 진단하였다. Glover(2016)는 동적자본구조모형(dynamic capital structure model)을 사용하여 기업의 기대부도비용을 추정하였다. Pan et al.(2018)은 경영위험(management risk)이 기업의 부도위험에 미치는 영향을 시험하고, 경영진 교체(management turnover)가 있을 때 부도위험이 증가한다고 주장하였다. Brogaard et al.(2017)은 주식 유동성 증가가 부도위험을 감소시키는 것을 보였다. Traczynski(2017)는 베이지안 모형 평균(Bayesian model averaging)으로 기업부도위험을 평가하고, 모형 평균 예측(model-averaged forecast)이 개별 모형 예측보다 예측력이 우수하다고 주장하였다. Aretz et al.(2018)은 미국 외 기업들을 대상으로 Campbell et al.(2008)의 모형을 적용하였으며, 유의하게 양의 부도위험 프리미엄(premium)이 있음을 보였다.

최근에는 다양한 기계학습 방법론을 활용하여 기업부도에 대한 예측을 시도한다. 기계학습 방법론을 활용한 기업부도에 대한 예측은 새롭게 등장한 것이 아니며, 실제로 1990년대에는 인공신경망(artificial neural networks) 방법론을 이용한 많은 연구가 시도되었다. Yang et al.(1999)은 여러 종류의 신경망 모형을 시험하였으며, 이 가운데 확률신경망(probabilistic neural networks) 및 Fisher 판별분석 모형의 예측력이 가장 높은 것을 확인하였다. Hinton et al.(2006)의 독창적인 연구로 기계학습 방법론이 획기적으로 발전한 이후, 기업부도예측 분야에서도 기계학습을 활용한 연구가 진행되고 있다. Falavigna(2012)는 회계정보가 충분하지 않은 이탈리아 소기업을 대상으로 신경망 모형을 적용하여 신용위험을 예측하였다. Duan et al.(2012)은 기업부도예측을 위한 전방강도모형(forward intensity model)을 제시하고 기존 모형보다 예측력이 높다고 주장하였다.

국내에서도 기업의 부도예측을 위한 많은 연구가 시도되었다. 장욱(2008)은 구조모형과 축약모형을 결합함으로써 부도기업의 부도시손실율을 추정하는 방법을 제안하였다. 김성환(2013)은 금융 시장구조에 따른 기업 부도처리를 모형화하고, 금융기관의 장기적 성과를 위하여 과감한 부도처리가 필요하다고 주장했다. 박종원・안성만(2014)은 117개의 재무비율 가운데 17개 변수를 선정하여 기업부도에 대한 예측모형을 설계하였으며, 약 80% 이상의 분류 정확도를 나타낸다고 밝혔다. 오세경 외 2인(2015)은 옵션가격결정모형을 이용한 부도예측모형을 확장하여, 관측기간 내에 부도 발생이 가능하도록 완화한 모형을 제시하였다. 최정원 외 3인(2015)은 신문기사의 텍스트마이닝을 이용한 기업부도 예측모형을 제안하며, 분석 결과 기존 모형 수준의 높은 예측력을 보이지는 못하였으나 기업부도 예측모형의 정확도를 높이는 가능성을 확인하였다. 이인로・김동철

(2015)은 부도예측모형을 회계모형, 시장모형, 해저드모형(hazard model)으로 분류하고 각 모형의 예측력을 비교하였으며, 이 가운데 국내 환경을 반영한 수정 해저드모형의 예측력이 가장 높음을 보였다. 도영호 외 3인(2016)은 구조적 VAR모형(structural vector autoregressive model)을 사용하여 실질경제성장률로 측정되는 외부환경이 중소기업 부도에 미치는 영향을 산업별로 분석하고, 실질경제성장률 둔화가 중소기업 부도율 증가에 양의 영향을 미치는 것을 보였다. 이인로・김동철(2016)은 국내 주식시장에서 기업의 주식수익률과 부도위험의 관계를 분석하고, 닷컴 버블(dot-com bubble)이 일어난 기간을 제외하는 경우 2001년부터 2014년 중 매우 유의한 음(-)의 관계가 나타난다고 주장하였다. 기업의 신용위험과 관련하여, 안경희 외 2인(2018)은 기업의 신용등급 변경 가능성이 채권수익률에 영향을 미치므로 채권의 내재등급의 변화가 향후 기업 신용등급의 변화에 선행하는 것을 확인하였다.

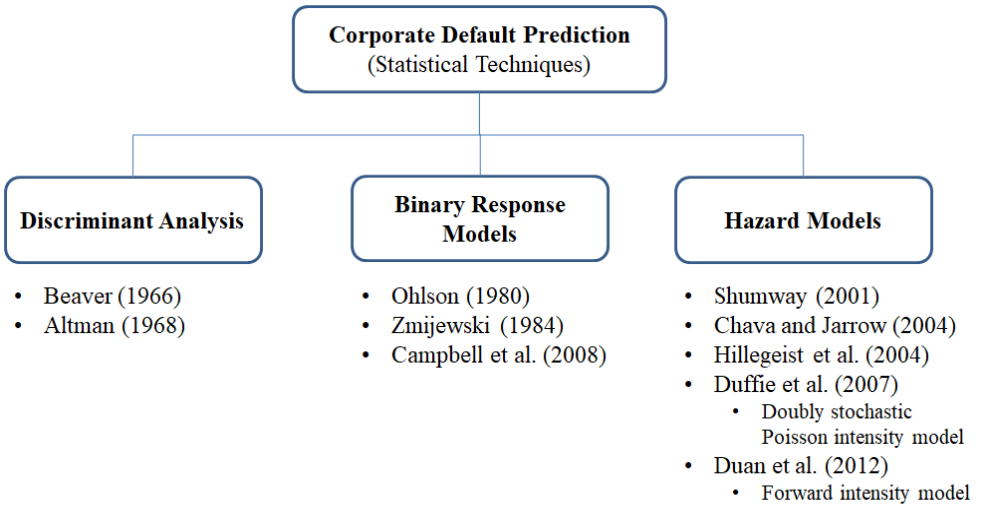
본 연구는 기업부도예측을 위해 사용된 분석 방법론을 통계적 모형을 활용한 방법과 기계학습 알고리즘을 활용한 방법으로 구분하고, 각각의 대표적인 분석 방법론을 살펴본다. 본 논문의 구성은 다음과 같다. 제2장에서는 통계적 모형을 활용한 기업부도예측의 주요 방법론을 3세대로 구분하여 살펴본다. 제3장에서는 기계학습을 활용한 대표적인 기업부도예측 방법론 세 가지를 살펴보며, 제4장에서는 기계학습 발달이 초래하는 금융 분야의 혁신을 전망한다. 마지막 제5장에서는 결론과 함께 본 연구의 시사점을 간략히 제시한다.

Ⅱ. 통계적 모형을 활용한 기업부도예측

1. 통계적 모형을 활용한 기업부도예측 개괄

통계적 모형을 활용한 기업부도예측은 크게 3세대로 구분할 수 있다([그림 1] 참조). 첫 세대는 판별분석(discriminant analysis)을 활용한 Altman Z-score가 대표적이다(Beaver, 1966; Altman, 1968). 판별분석을 활용하면 부도위험 정도에 따라 순서대로 점수화가 가능하다는 장점이 있다. 두 번째 세대는 기업의 상태를 정상(=0)과 부도(=1)로 구분하는 이항반응모형(binary response model)을 사용하였으며, 로짓 회귀분석이나 프로빗 회귀분석으로 기업부도를 분석하였다. Ohlson(1980)의 연구에서 소개된 Ohlson O-score가 대표적으로, 이와 같은 이진반응모형은 다음 기간의 기업부도확률을 계산할 수 있다는 장점이 있다. 다만 다기간의 기업부도확률을 계산하지는 못하는

데, Campbell et al.(2008)은 다중로짓모형을 사용하여 이를 확장하였다. 세 번째 세대는 Shumway(2001)로 대표되는 위험모형을 사용한 연구이다. 위험모형은 생존분석(survival analysis)이라고도 지칭되는데, 이 방법론을 사용하면 기간별 기업부도확률을 계산할 수 있다. 이 외에도 각 방법론의 확장을 통해 다양한 연구가 등장하였으며, 현재까지도 활발히 연구되고 있다.



본 그림은 기업부도예측 연구에서 사용된 통계적 모형을 크게 세 가지로 분류하여 나타냄. 판별분석(discriminant analysis), 이항반응모형(binary response models), 위험모형(hazard models)으로 나누었으며, 대표적인 연구 및 주요 확장 연구를 소개함.

[그림 1] 기업부도예측을 위한 통계적 모형의 분류

2. 통계적 모형을 활용한 기업부도예측 방법론

2.1 판별분석(Discriminant Analysis)

판별분석은 Beaver(1966)와 Altman(1968) 이후 지금까지도 널리 사용되는 부도예측 모형이며, Altman Z-score 및 이를 개량한 후속 연구(Altman, 1993; Mare et al., 2017)가 대표적이다. 판별분석은 기업의 부도 여부를 가장 잘 분류할 수 있는 변수들을 선정하고, 이들의 선형결합을 이용하여 다음과 같이 판별함수(discriminant function)를 계산한다.

$$D = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{1}$$

이때 D는 판별함수로부터 계산한 판별득점(discriminant score)이며, β_0는 상수항, β_m은 추정계수, X_m은 설명변수를 의미한다. 분석 대상은 판별득점이 임계치 이하일 경우 정상으로, 임계치보다 높으면 부도 그룹으로 분류된다. 그러나 판별분석의 경우, 독립변수들이 다변수 정규분포(multivariate normal distribution)를 따른다는 가정이 있어야만 하고, 정상 기업과 부도 기업으로 정의되는 두 그룹 간의 공분산 행렬(covariance matrix)이 동일하다는 가정을 필요로 한다는 점에서 비판의 대상이 된다(Charitou et al., 2004).

2.2 이항반응모형(Binary Response Models)

이항반응모형은 기업의 상태를 정상(=0)과 부도(=1)로 나누고 부도사건이 발생할 확률을 설명변수를 사용하여 추정하는 모형으로, Ohlson(1980)에서 소개된 Ohlson O-score가 대표적이다. 일반적으로 로짓 함수나 프로빗 함수를 사용하며, 본 연구에서는 Foreman(2003)과 Charitou et al.(2004)의 연구를 참고하여 로짓 함수를 이용한 방법론을 설명한다. 로짓 회귀분석을 이용한 기업부도예측은 판별분석에 비해 몇 가지 우위점을 지니고 있다. 먼저, 로짓 모형은 실패의 사전확률(prior probability)이나 예측변수(predictor variable)의 분포에 대한 가정을 필요로 하지 않는다. 또한, 개별 독립변수의 유의성을 검정할 수 있다. 마지막으로, 로짓 모형을 사용하면 다음 기간의 부도확률을 계산할 수 있다. 기업 $n = 1, \cdots, N$에서 다음 기간 동안 부도가 발생할 확률을 P_n이라 하자. 로짓 모형에서 P_n은 다음과 같이 정의된다.

$$\begin{aligned} P_n(y_n = 1) &= \frac{1}{(1 + e^{-Z})} \\ &= \frac{1}{\{1 + \exp[-(\beta_0 + \beta_1 X_{1,n} + \beta_2 X_{2,n} + \cdots + \beta_m X_{m,n})]\}} \end{aligned} \tag{2}$$

이때 y_n은 n번째 기업의 부도가 발생하는 경우 1, 그렇지 않고 정상인 경우 0을 가지는 변수이다. $P_n(y_n = 1)$은 n번째 기업의 부도확률이며, $\beta_1, \beta_2, \cdots, \beta_m$는 추정계수를 의미한다. $X_{1,n}, X_{2,n}, \cdots, X_{m,n}$는 n번째 기업의 설명변수이다. 따라서 우도함수(likelihood function)는 다음과 같이 계산된다.

$$L = \prod_{n=1}^{N} F(\beta' X_n)^{y_n} (1 - F(\beta' X_n))^{1-y_n} \tag{3}$$

이때 $F(\beta' X_n) = \frac{1}{\{1 + \exp[-(\beta_0 + \beta_1 X_{1,n} + \beta_2 X_{2,n} + \cdots + \beta_m X_{m,n})]\}}$ 이다. 모형의 계수 추정을 위하여 최대우도(maximum likelihood) 방법을 사용한다.

2.3 위험모형(Hazard Model)

위험모형(hazard model), 혹은 생존분석(survival analysis) 방법은 Cox(1972)의 위험회귀분석 모형(hazard regression model)을 사용하는 부도예측 방법론이다. 위험모형은 기업의 상태를 정상(=0)과 부도(=1)로 구분하고, 부도 사건이 발생하면 기업의 상태에 대한 관측을 종료한다. 기업의 부도가 발생하는 시점을 T라 할 때, t 시점에서 기업의 생존함수(survival function)인 $S(t)$는 다음과 같이 나타낼 수 있다.

$$S(t) = \Pr(T \geq t) \tag{4}$$

이때 위험함수(hazard function) $\lambda(t)$는 t시점에 부도가 발생하는 순간의 실패율(instantaneous failure rate)을 의미하며 다음과 같이 정의된다.

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{5}$$

Cox 비례위험모형(Cox proportional hazard model)은 순간실패율 $\lambda(t)$를 다음과 같이 비정형 기저위험율(unspecified baseline hazard rate) $\lambda_0(t)$에 비례하는 모형으로 나타난다.

$$\lambda(t|X_n) = \lambda_0(t) \exp(\beta' X_n) \tag{6}$$

이때 β는 회귀계수로 구성된 열벡터(column vector)이고, X_n은 n번째 기업의 설명변수이다. Cox 모형은 준모수적(semi-parametric) 모형으로, 비모수적 요소인 기저위험율 $\lambda_0(t)$와 모수적 요소인 $\exp(\beta' X_n)$로 구성되어있다. 회귀계수 β에 대한 부분우도함수(partial likelihood function)는 다음과 같이 계산된다.

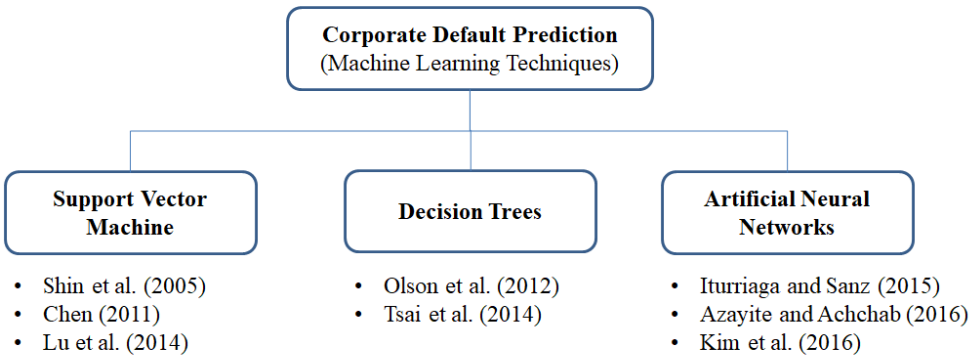
$$PL(\beta)=\prod_{i=1}^{N}\left[\frac{e^{\beta' X_i}}{\sum_{j=1}^{N} Y_{ij}e^{\beta' X_j}}\right]^{\delta_i} \tag{7}$$

이때, Y_{ij}은 $t_j \geq t_i$인 경우 1, 그렇지 않고 $t_j < t_i$인 경우 0을 가지는 변수이다. δ_i는 관측이 중도절단되지 않은 경우(not censored) 1, 중도절단된 경우(censored) 0을 가지는 변수이다. 이 모형의 모수는 부분우도함수를 최대화하는 값으로 추정한다.

Ⅲ. 기계학습을 활용한 기업부도예측

1. 기계학습을 활용한 기업부도예측 개괄

Samuel(1959)은 기계학습(machine learning)이라는 개념을 제시하고, 이를 “명확한 프로그램 없이 컴퓨터에게 학습하는 능력을 부여하는 학문 분야”로 정의하였다. Mitchell(1997)은 이러한 기계학습에 관한 연구를 더욱 발전시켜, “성과지표(P)로 측정한 컴퓨터 프로그램의 작업(T) 수행 결과가 수행경험(E)으로 향상되는 경우, 이 프로그램은 성과지표(P) 차원에서 작업(T) 수행능력을 수행경험(E)을 통해 학습하였다”고 정의하였다. 이러한 측면에서 볼 때, 기업부도예측을 위한 기계학습 알고리즘은 기업과 관련된 정보를 이용하여 기업의 신용위험을 예측하는 작업(T)을 실시하기 위하여 실제 기업의 신용정보를 활용(E)함으로써 실제로 정확하게 예측할 확률(P)을 향상시키는 일련의 과정으로 이해할 수 있다. 기계학습 알고리즘을 활용한 기업부도예측 연구 또한 다양하게 진행되었으며, 특히 컴퓨터 과학 분야를 중심으로 많이 발전하였다([그림 2] 참조). 기계학습을 활용한 연구는 대부분 기업의 경영상태를 정상(=0)과 부도(=1)라는 2가지, 혹은 그 이상의 상태로 정의하고 기업이 특정 상태에 포함될 확률을 계산하는 분류(classification) 문제로 접근하였다는 공통점이 있다. 따라서 분류 문제 해결을 위한 기계학습 알고리즘이 주로 사용되었으며, 대표적으로 SVM(support vector machine), 의사결정나무(decision tree), 인공신경망 알고리즘 등이 사용되었다.



본 그림은 기업부도예측 연구에서 사용된 기계학습 알고리즘 가운데 대표적인 세 가지를 나타냄. 본고에서는 SVM(support vector machine), 의사결정나무(decision tree), 인공신경망(artificial neural networks) 알고리즘을 선정하였으며, 대표적인 연구 및 주요 확장 연구를 소개함.

[그림 2] 기업부도예측을 위한 기계학습 알고리즘의 분류

2. 기계학습을 활용한 기업부도예측 방법론

2.1 SVM(Support Vector Machine)

SVM 알고리즘은 주로 분류 문제 해결을 위해 사용되는 기계학습 방법론으로, 독립변수가 많은 고차원 공간에서도 관측치를 적절하게 분류할 수 있도록 약간의 오류를 허용하면서 관측치 간의 거리를 의미하는 마진(margin)을 최대화하는 초평면(hyperplane)을 찾는 것을 목표로 한다. SVM 방법론은 이러한 초평면을 결정한 후, 이를 기준으로 관측치를 분류한다. p차원 공간에서의 초평면이란 $p-1$차원의 평면 아핀 부분공간(flat affine subspace)을 의미하며, 선형 커널(linear kernel)을 사용하는 SVM을 수식으로 나타내면 다음과 같다.

$$\max_{\beta_0,\beta_1,\cdots,\beta_p,\epsilon_0,\epsilon_1,\cdots,\epsilon_n} M \tag{8}$$

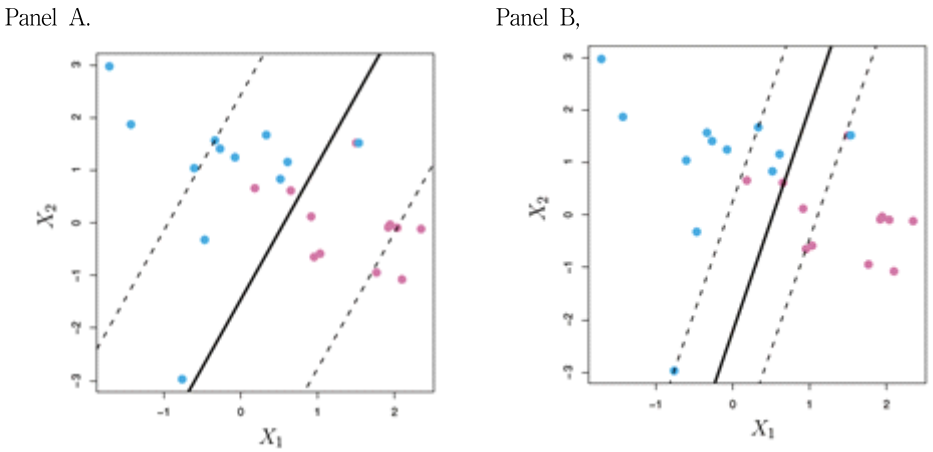
subject to

$$\sum_{j=1}^{p}\beta_j^2 = 1$$

$$y_i(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}) \geq M(1-\epsilon_i)$$

$$\epsilon_i \geq 0,\ \sum_{i=1}^{n}\epsilon_i \leq C$$

이때 C는 음이 아닌 조정 매개변수(nonnegative tuning parameter)를 의미한다. [그림 3]과 같이 조정 매개변수의 크기에 따라 기존 훈련 자료의 분류 정확도가 달라진다. 그림 3의 패널 B는 패널 A와 비교하여, 분류 오류의 한계치(tuning parameter)를 더욱 줄인 것으로, 오류를 허용하는 점선 안에 있는 서포트 벡터(support vector)에 의해서만 초평면이 결정되는 것을 확인할 수 있다. SVM 방법론은 커널 함수에 따라 비선형 초평면을 이용한 분류도 가능하여, 간결하면서도 뛰어난 성능을 보이는 집단 분류 방법으로 널리 사용된다. 따라서 다양한 사례에서 우수한 분류 정확도를 나타내는 것으로 보고되고 있으며, 부도예측 분야에서도 좋은 성과를 보인다(Devi and Radhika, 2018).



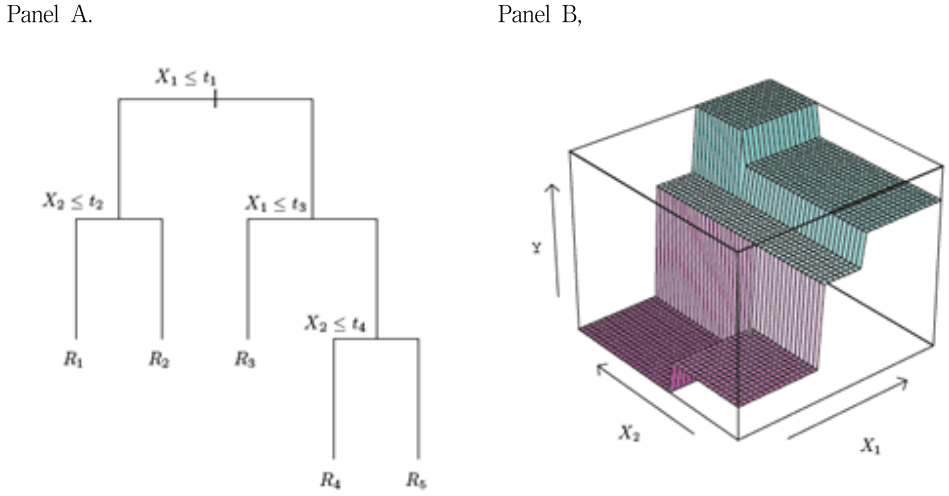
본 그림은 조정 매개변수(tuning parameter)의 변화에 따른 support vector classifier의 차이를 나타냄. Panel A는 조정 매개변수가 클 때, Panel B는 조정 매개변수가 작을 때 각 그룹을 분류하는 support vector classifier를 나타냄. (출처: James et al., 2013)

[그림 3] 조정 매개변수(tuning parameter)에 따른 support vector classifier

2.2 의사결정나무(Decision Tree)

의사결정나무, 혹은 DT 알고리즘은 [그림 4]의 패널 A와 같이 의사결정규칙(decision rule)을 나무(tree) 구조로 도표화하여 회귀분석이나 분류 문제를 해결하는 방법론이다. DT 알고리즘은 p개의 설명변수 $X_1, X_2, \cdots, X_p$의 조합으로 구성된 특징 공간(feature space)을 J개의 비중첩 영역(non-overlapping regions) $R_1, R_2, \cdots, R_J$으로 분할한다. 그리고 동일한 영역 R_j에 속하는 관측치에 대해서는 동일한 예측을 실시하는 방법론이다([그림 4] 참조). DT 알고리즘은 모형의 이해가 직관적이고 누구나 해석이 용이하다는 장점이 있지만, 특징 공간을 분할하거나 분기를 생산하는 과정에 과적합(over-fitting) 문

제가 발생하기 쉽고 예측의 정확도가 다소 떨어진다는 한계가 있다.



본 그림은 2차원 특징 공간(feature space)에서 의사결정나무 알고리즘을 실시한 결과를 나타냄. Panel A는 반복이진분할(recursive binary splitting)을 통한 나무 모형을 나타내며, Panel B는 Panel A에서 구한 의사결정나무의 예측을 표현한 겨냥도(perspective plot)를 나타냄. (출처: James et al., 2013)

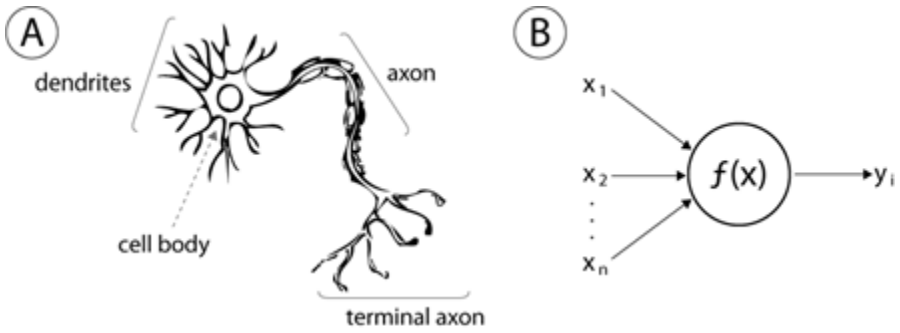
[그림 4] 2차원 특징 공간(feature space)의 의사결정나무

따라서 의사결정나무의 예측력을 높이기 위하여 bagging, random forests, boosting 등의 방법을 사용한다. bagging은 표본의 수를 높이면 예측의 분산이 줄어든다는 점에서 착안한 방법이다. bootstrap 방식으로 무작위로 추출한 표본으로 총 N번의 의사결정나무를 실행하고, N번의 결과 중 다수결에 의해 가장 많이 나온 집단으로 분류한다. random forests는 표본뿐만 아니라 독립변수들도 무작위로 추출하는 방식으로, 실행 과정은 bagging과 동일하다. bagging과 random forests 방법이 무작위로 추출한 여러 표본을 독립적으로 모델링하는 것과는 달리, boosting은 한 번 시행한 모형의 결과를 참조하여 순차적으로 N번 개선해나가는 방법이다.

2.3 인공신경망(Artificial Neural Networks)

인공신경망은 실제 두뇌가 작동하는 과정에서 착안된 기계학습 알고리즘으로, 두뇌의 구조를 모방하여 단순한 구조의 인공뉴런(artificial neuron)을 연결함으로써 복잡한 문제를 해결하고자 하는 방법론이다. 인간의 두뇌와 척수의 기본 단위를 구성하는 뉴런(neuron)은 수신한 신호를 연결된 다른 뉴런으로 전달하는 역할을 수행한다([그림

5.A] 참조). 뉴런은 전달받은 신호의 강도가 임계치(threshold) 이상일 때에만 다른 뉴런들로 신호를 전달하는 특징을 가지고 있다(김형준 외 2인, 2017).



본 그림은 실제 뉴런(neuron)과 인공뉴런(artificial neuron)을 비교함. 패널 A는 실제 뉴런의 구조를 나타내며, 패널 B는 인공뉴런의 구조를 나타냄. (출처: Maltarollo et al., 2013)

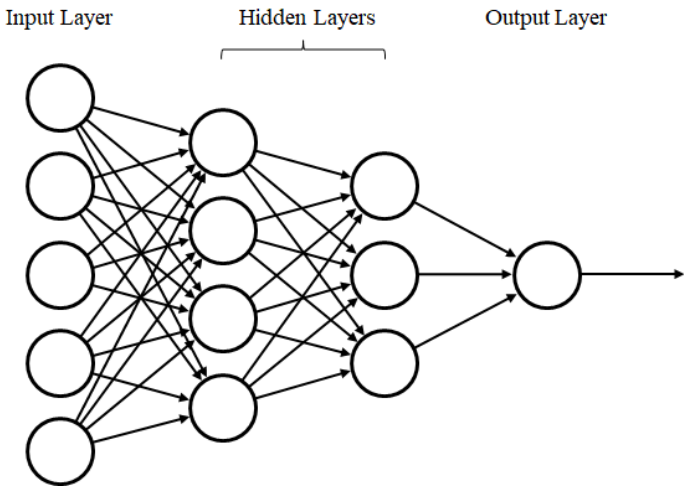
[그림 5] 실제 뉴런과 인공뉴런의 비교

인공뉴런은 이러한 실제 뉴런의 역할을 수학적 모형을 통해 모사한 것이다. 각 인공뉴런은 0과 1로 구성된 여러 신호 $x_1, x_2, \cdots, x_j$를 수신하며, 이들의 가중치 $w_1, w_2, \cdots, w_j$에 따라 전달받은 신호의 가중합을 계산한다([그림 5.B] 참조).[1] 그리고 입력된 신호의 가중합계가 일정 강도, 혹은 임계치 이상일 때에만 다음 인공뉴런으로 신호를 전달한다. 각 뉴런의 가중치와 임계치는 과거 경험이나 데이터에 따라 최적의 결과를 낼 수 있는 조합으로 결정된다. 인공뉴런을 수식으로 나타내면 다음과 같다.

$$y_i = output = \begin{cases} 0, \text{ if } \sum_j w_j x_j \leq threshold \\ 1, \text{ if } \sum_j w_j x_j > threshold \end{cases} \tag{8}$$

인공신경망은 단순한 인공뉴런의 조합으로 복잡한 문제를 해결할 수 있는 기계학습 방법론이다. 인공신경망은 [그림 6]과 같이 여러 계층의 인공뉴런의 연결로 구성되며, 각 계층은 입력계층(input layer), 은닉계층(hidden layer), 출력계층(output layer)으로 구분된다. 인공신경망의 훈련, 혹은 최적화 과정은 최선의 결과를 얻을 수 있는 각 인공뉴런의 가중치($w_1, w_2, \cdots, w_j$)와 임계치(threshold)를 찾는 것을 의미하며, 따라서 강력한 연산능력이 필요하다.

1) 종류에 따라 (-∞, ∞), [0, ∞)에 속하는 신호를 수신하기도 한다.



본 그림은 1개의 입력계층(input layer), 2개의 은닉계층(hidden layers), 1개의 출력계층(output layer)으로 구성된 4개 계층의 인공신경망의 구조를 도식화함. (출처: Nielsen(2015) 재구성)

[그림 6] 은닉계층이 2개인 인공신경망

인공신경망의 구조 가운데 은닉계층을 여러 개 중첩한 모형을 심층신경망(deep neural network)이라고 한다. 그리고 딥러닝(deep learning)은 이러한 심층신경망을 기계학습에 사용하는 것을 지칭한다. 이러한 접근방법은 기존의 인공신경망 방법론에서 나타나는 국소 최소(local minima)치 문제를 해결하기 위한 것으로, 데이터를 분석·학습하는 모든 과정을 프로그래밍하는 것이 아니라 지속적으로 데이터로부터 학습한 결과를 주어진 문제 해결능력 향상에 사용하는 것이 특징이다.

최근 딥러닝은 convolutional neural network(CNN) 알고리즘과 recurrent neural network(RNN) 알고리즘 등을 사용하여 학습 과정에서 발생하는 과적합 문제를 해결하고 작업성과를 높이고 있다. CNN은 학습에 사용되는 정보를 여러 영역으로 분할한 뒤, 제한된 정보를 사용하여 관계성이 높은 영역끼리 분석하는 방법이다. 따라서 이미지 인식(image recognition) 분야에서 좋은 성과를 나타내며 많이 활용되고 있다. RNN은 학습에 사용되는 데이터를 각각 독립적으로 분석하는 것이 아니라 순차적으로 처리·분석하는 방법으로, 시계열 자료의 예측이나 텍스트(text) 자료의 문맥 인식 등에 사용된다. 현재 딥러닝은 입력된 정보의 맥락(context)을 이해하고, 이러한 분석을 토대로 패턴 인식, 자연어 처리, 자율주행 등 다양한 작업을 수행하고 있다.

Ⅳ. 기계학습의 발달과 금융 혁신

기계학습 알고리즘의 발달은 McCarthy et al.(2006)이 1956년 다트머스 학회(Dartmouth workshop)에서 주창한 인공지능(artificial intelligence, AI)의 구현을 가능하게 하며, 이는 금융 분야의 4차 산업 혁신 초래에 이바지한다. 현재의 기술 수준은 특정 작업에 특화된 좁은 인공지능(narrow AI)으로, IBM의 왓슨과 구글의 알파고 등이 대표적이다. 이러한 인공지능은 정보의 맥락(context)에 대한 이해를 바탕으로 기존 기계학습에서는 불가능했던 인간의 창의성이나 직관력을 흉내 내는 수준에 도달했다는 평가를 받고 있다. 특히 구글의 바둑 인공지능 알파고는 2개의 심층신경망을 탑재한 몬테카를로 트리 검색(Monte Carlo tree search)[2]을 실시하여, 정책망(policy network)으로 다음 돌을 착수할 위치를 선택하고 가치망(value network)으로 해당 지점에 착수하였을 때의 승리 확률을 평가하여 가장 승률이 높은 곳에 착수하는 방법으로 인공지능의 새로운 가능성을 선보였다. 이러한 인공지능 기술 발달은 부도예측 및 신용위험군 분류에서도 좋은 성과를 거둘 것으로 기대된다.

기계학습 방법론의 발달에 따라 데이터 분석의 유용성이 증가하면서 관련 경제 규모가 함께 확대될 전망이다. 딥러닝 등을 이용한 빅데이터(big data) 분석 기술 발달에 따라 다양한 분야에서 데이터 분석을 활용한 미래 예측을 실시하고 있다. 지금도 빅데이터 분석은 금융(신용위험 관리) 뿐만 아니라, 유통(상품수요 예측), 제조(불량원인 분석), 공공(가계소득 예측) 등 각 분야에서 유용성을 인정받고 있으며, 분석 방법론의 개선도 꾸준히 진행 중이다. 향후 양자 컴퓨터(quantum computer) 등과 같은 신기술과의 접목을 통해 더욱 정교한 분석이 가능하다. 사물인터넷(Internet of Things, IoT)의 발달로 새로운 데이터가 누적되는 속도 및 크기가 폭증하며, 이를 인공지능 알고리즘과 진일보한 컴퓨터로 처리할 필요성도 함께 증가하고 있다. 빅데이터 활용의 증가로 데이터가 새로운 형태의 자산으로 주목받게 되면서, 데이터 유통에 기반한 데이터 경제의 중요성도 부상하고 있다. 미국・영국・EU 등 선진국들은 데이터 기반 사회로의 이행을 위해 교육 과정 개선과 기술 표준 제정, 법적 책임 명시 등 대응 전략을 마련하고 있다. 하지만 우리나라는 엄격한 규제로 인하여 빅데이터의 구축과 활용에 어려움을 겪고 있으므로, 국가 경쟁력 강화를 위하여 기술 발전에 걸맞은 제도개선과 지원정책 수립 등이 필요하다(주강진 외 3인, 2017). 데이터 주도 경제를 실현하기 위해서는, 우선 신뢰 가능하고 상호 호환되는 데이터셋(data set)을 이용할 수 있는 인프라의 구축이 필요하고, 데이터셋에

2) 몬테카를로 트리 검색(Monte Carlo tree search, MCTS)은 무작위 대입 방법을 사용하여 사례별 예상 성공 확률을 평가한 다음, 가장 확률이 높은 대안을 선택하는 방법론을 의미한다.

서 가치를 창출할 수 있는 기술인력의 활용과 협력이 요구되며, 창출된 가치를 서비스로 연결할 수 있는 시스템과 이를 전파할 수 있는 얼리어답터가 필요하다. 무엇보다 이러한 환경을 뒷받침할 수 있는 정책적 지원이 요구된다.

또한, 기계학습의 발달로 촉발되는 인공지능이 사물인터넷 및 빅데이터 등과 결합하여 금융산업의 패러다임 변화를 견인하며 핀테크(FinTech)로 대표되는 새로운 금융서비스가 출현할 것으로 예상된다(주강진 외 3인, 2016). 자동차에 부착된 센서로 운전습관을 파악한 후 보험료를 산정하거나, 주변 지인들의 SNS 평가를 바탕으로 신용도 평가 및 투자성향 조정을 시행하는 금융상품도 만들어질 수 있다. 빅데이터 분석을 접목하여 개인신용평가를 정교하게 실시하는 시스템이 등장하고, 인공지능 알고리즘을 이용하여 개개인에게 맞춤형 위험관리와 생애주기별 자산관리 솔루션을 제공하는 서비스도 개발되고 있다. 인공지능에 기반한 로보어드바이저는 저렴한 수수료와 높은 접근성으로 이미 자산관리서비스 시장에 확산되고 있다(임혜진 외 2인, 2018; 최원우・류두진, 2018).

기계학습을 이용한 빅데이터 분석 기술의 발달은 데이터 자본이 주도하는 새로운 금융・경제 패러다임(paradigm)을 촉발할 수 있다. 美 메사추세츠 공과대학(Massachusetts Institute of Technology)의 최근 연구보고는 새로운 디지털 상품과 서비스를 만들어 가치를 창출하기 위한 데이터 자본(data capital)의 중요성을 강조한다(MIT, 2016). Newman(2011)은 소프트웨어나 하드웨어의 경제가 아니라, 빅데이터・오픈데이터(open data)・연결데이터(linked data) 등으로 구성된 데이터 경제(data economy)의 도래를 예견했다. 빅데이터 활용성 증대로 기업과 기관이 보유한 데이터의 가치가 높아지고, 데이터 자본의 가치를 극대화하기 위하여 데이터의 생성・관리・활용의 중요성이 증가한다. ICT(information and communications technology) 기술을 선도적으로 개발하는 기업이 아니더라도, 데이터 경제 시대에 적응하기 위해서는 보유한 데이터의 지속적인 품질관리를 포함한 데이터 거버넌스(data governance)[3]가 필요하다. 또한, 국가 행정에서 생산되는 공공 데이터와 기업에서 생산하는 민간 데이터의 융합 활용을 위한 국가 차원의 데이터 거버넌스 마련이 필요하다(정용찬, 2018).

한편, 금융과 기술의 융합에 따른 핀테크의 발달은 소비자보호 문제를 포함한 새로운 위험 가능성을 내포하며, 이에 대한 지속적인 모니터링이 필요하다(손진빈 외 2인, 2019). 또한, 데이터의 활용과 관리의 과정에서 개인정보보호의 중요도가 더욱 강조되고 있다. 동형암호(同型暗號, homomorphic encryption)는 빅데이터 활용에서 개인정보의 유출 위험을 원천적으로 차단하는 기술로, 금융 분야에서도 데이터 분석 기술의 발

3) Laudon and Laudon (2012)는 데이터 거버넌스를 기업이 사용하는 데이터의 가용성, 유용성, 통합성, 보안성의 관리를 위해 필요한 정책과 프로세스를 의미하며, 프라이버시, 보안성, 데이터품질, 관리규정의 준수를 강조한다고 설명한다.

달과 함께 주목할 필요가 있다.[4)]

Ⅴ. 요약 및 결론

본 논문은 기업부도예측을 위한 관련 연구의 진행을 조사하고, 주요 연구 방법론을 통계적 모형과 기계학습 알고리즘으로 구분하여 검토하였다. 또한, 기계학습 발전이 초래하는 금융 분야 혁신을 함께 전망하였다. 이를 통해 향후 재무·금융 분야와 컴퓨터 과학 분야의 융합 연구를 위한 단서를 제시하고, 기업부도에 대한 정밀한 예측을 위한 금융공학 방법론 확대의 토대를 마련하고자 하였다.

4차 산업혁명으로 명명된 일련의 기술발전으로 인해 기업부도예측을 포함한 금융공학 분야에서도 새로운 방법론 적용 필요성이 대두되고 있다. 특히 본 연구에서 제시된 빅데이터 분석 방법론은 기업의 경영상태 진단과 투자자의 정확한 투자의사결정을 위한 전사적 데이터 거버넌스의 필요성을 시사한다. 대기업뿐만 아니라 중소기업 역시 보유한 경영 데이터의 종류, 크기, 빈도 등을 정확하게 파악하고 지속적인 품질관리를 시행함으로써 새로운 기술을 보다 용이하고 신속하게 도입할 수 있는 토대 마련이 요구된다.

기계학습 방법론 발달에 따른 빅데이터 활용가치 증대는 동시에 개인정보보호 강화의 필요성을 수반한다. 빅데이터 활용 과정에서 개인정보보호를 위한 비식별화 조치에도 불구하고, 인터넷 등에 공개된 다른 정보와의 결합을 이용한 재식별 등으로 개인정보가 유출되는 사례가 발생하고 있다. 동형암호와 같은 최신 기술을 적극적으로 활용하여, 정보유출 위험을 낮추고 개인정보 침해 없는 빅데이터 분석 및 처리를 달성할 필요가 있다.

기계학습을 이용한 기업부도예측은 알고리즘에 따라 예측결과의 계산과정이 블랙박스(black box)로 남는 경우가 있어 주의가 필요하다. 이러한 방법론의 경우 기업부도위험을 계산할 수 있어도, 부도위험을 낮추기 위한 경영상태 개선전략을 제시하지는 못하는 한계가 있다. 따라서 기업부도예측을 실시할 때에는 예측 목적에 맞는 정보를 제공할 수 있는 적절한 방법론 선정이 요구되며, 이를 위해서는 각 방법론에 대한 상세한 이해와 활용능력이 필요하다.

---

4) 동형암호는 원자료와 암호화된 자료에서 같은 성질이 유지된다는 의미로, 원자료의 연산 결과와 암호문의 연산 결과가 같은 값을 지니는 특징을 가지고 있어 개인정보의 유출 위험 없이 빅데이터 활용을 가능하게 하는 차세대 암호 기술이다. (과학기술정보통신부 보도자료, “개인정보 활용기술로 주목받는 동형암호”, 2018. 5. 8.)

참고문헌

김성환 (2013), “금융 시장구조와 평판이 부도와 금융기관 성과에 미치는 영향,” 금융공학연구, 제12권 제2호, 139-160.

김형준・송완영・안세륭 (2017), “빅데이터 분석 방법론과 데이터 과학자,” 한국주택금융공사 주택금융연구원.

도영호・장영민・김경숙・김석진 (2016), “실질경제성장률이 중소기업 산업별 부도율에 미치는 영향: 구조적 VAR 모형을 이용하여,” 중소기업연구, 제38권 제3호, 25-48.

박종원・안성만 (2014), “재무비율을 이용한 부도예측에 대한 연구: 한국의 외부감사대상기업을 대상으로,” 경영학연구, 제43권 제3호, 639-669.

손진빈・류두진・박채진 (2019), “국내 핀테크 산업의 현황과 규제 및 지속가능성에 대한 논고,” 금융공학연구, 제18권 제2호, 119-150.

안경희・박래수・박종원 (2018), “신용등급변경가능성이 자본조달에 미치는 영향: 채권내재등급(BIR)과 신용등급(AR)의 차이를 중심으로,” 금융공학연구, 제17권 제2호, 23-52.

오세경・최시열・박기남 (2015), “임의의 부도발생 시점을 고려한 부도예측모형에 관한 연구,” 재무관리연구, 제32권 제4호, 23-51.

이인로・김동철 (2015), “회계정보와 시장정보를 이용한 부도예측모형의 평가 연구,” 재무연구, 제28권 제4호, 626-666.

이인로・김동철 (2016), “국내 주식시장의 부도위험 이례현상에 관한 연구,” 한국증권학회지, 제45권 제5호, 1097-1129.

임혜진・류두진・양희진 (2018), “금융시장 로보어드바이저 산업에 대한 고찰: 현황과 개선방안,” 경영학연구, 제47권 제3호, 725-749.

장욱 (2008), “구조모형과 축약모형을 결합한 내재 부도시손실율의 추정과 그 결정요인의 분석,” 금융공학연구, 제7권 제2호, 49-73.

정용찬 (2018), “4차 산업혁명 시대의 데이터 거버넌스 개선 방향”, Premium report, 18-05, 정보통신정책연구원.

주강진・이민화・양희진・류두진 (2016), “핀테크 산업의 발전방향에 관한 연구”, 한국증권학회지, 제45권 제1호, 145-170.

주강진・이민화・양희진・류두진 (2017), “4차 산업혁명과 인공지능: 현황, 사례, 규제에 대한 개괄적 고찰”, 한국경영과학회지, 제42권 제4호, 1-14.

최원우・류두진 (2018), “하이브리드 로보어드바이저 활용의 사례와 제언,” Korea Business Review, 제22권 제3호, 33-52.

최정원・한호선・이미영・안준모 (2015), “텍스트마이닝 방법론을 활용한 기업 부도 예측 연구,” 생산성논집, 제29권 제1호, 201-228.

Altman, E. I. (1968), “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy,” *Journal of Finance*, 23(4), 589-609.

Altman, E. I. (1993), *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting and Avoiding Distress and Profiting from Bankruptcy*, Second Edition, New York: John Wiley and Sons Inc.

Aretz, K., C. Florackis, and A. Kostakis (2018), “Do Stock Returns Really Decrease with Default Risk? New International Evidence,” *Management Science*, 64(8), 3469-3970.

Azayite, F. Z. and S. Achchab (2016), “Hybrid Discriminant Neural Networks for Bankruptcy Prediction and Risk Scoring,” *Procedia Computer Science*, 83, 670-674.

Beaver, W. H. (1966), “Financial Ratios as Predictors of Failure,” *Journal of Accounting Research*, 4, 71-111.

Bonfim, D. (2009), “Credit Risk Drivers: Evaluating the Contribution of Firm Level Information and of Macroeconomic Dynamics,” *Journal of Banking & Finance*, 33(2), 281-299.

Brogaard, J., D. Li, Y. Xia (2017), “Stock Liquidity and Default Risk,” *Journal of Financial Economics*, 124(3), 486-502.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008), “In Search of Distress Risk,” *Journal of Finance*, 63(6), 2899-2939.

Charitou, A., E. Neophytou, and C. Charalambous (2004), “Predicting Corporate Failure: Empirical Evidence for the UK,” *European Accounting Review*, 13(3), 465-497.

Chava, S. and R. A. Jarrow (2004), “Bankruptcy Prediction with Industry Effects,” *Review of Finance*, 8, 537-569.

Chen, M. Y. (2011), “Bankruptcy Prediction in Firms with Statistical and Intelligent Techniques and a Comparison of Evolutionary Computation Approaches,” *Computers & Mathematics with Applications*, 62(12), 4514-4524.

Cox, D. R. (1972), “Regression Models and Life Tables (with Discussion),” *Journal*

of the Royal Statistical Society, 34(2), 187-220.

Dakovic, R., C. Czado, and D. Berg (2010), "Bankruptcy Prediction in Norway: a Comparison Study," *Applied Economics Letters*, 17(17), 1739-1746.

Devi, S. S. and Y. Radhika (2018), "A Survey on Machine Learning and Statistical Techniques in Bankruptcy Prediction," *International Journal of Machine Learning and Computing*, 8(2), 133-139.

Duan, J., J. Sun, and T. Wang (2012), "Multiperiod Corporate Default Prediction - A Forward Intensity Approach," *Journal of Econometrics*, 170(1), 191-209.

Duffie, D., L. Saita, K. Wang (2007), "Multi-period Corporate Default Prediction with Stochastic Covariates," *Journal of Financial Economics*, 83(3), 635-665.

Falavigna, G. (2012), "Financial Ratings with Scarce Information: A Neural Network Approach," *Expert Systems with Applications*, 39(2), 1784-1792.

Figlewski, S., H. Frydman, and W. Liang (2012), "Modeling the Effect of Macroeconomic Factors on Corporate Default and Credit Rating Transitions," *International Review of Economics and Finance*, 21(1), 87-105.

Foreman, R. D. (2003), "A Logistic Analysis of Bankruptcy within the US Local Telecommunications Industry," *Journal of Economics and Business*, 55(2), 135-166.

Glover, B. (2016), "The Expected Cost of Default," *Journal of Financial Economics*, 119(2), 284-299.

Hillegeist, S. A., E. K. Keating, D. P. Cram, and K. G. Lundstedt (2004), "Assessing the Probability of Bankruptcy," *Review of Accounting Studies*, 9, 5-34.

Hinton, G. E., S. Osindero, and Y.-W. Teh (2006), "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, 18(7), 1527-1554.

Iturriaga, F. J. L. and I. P. Sanz (2015), "Bankruptcy Visualization and Prediction using Neural Networks: A Study of US Commercial Banks," *Expert Systems with Applications*, 42(6), 2857-2869.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning*, Vol. 112. New York, Springer.

Jessen, C. and D. Lando (2015), "Robustness of Distance-to-Default," *Journal of Banking & Finance*, 50, 493-505.

Kim, H., N. O. Jo, and K. S. Shin (2016), "Optimization of Cluster-based Evolutionary Undersampling for the Artificial Neural Networks in Corporate

Bankruptcy Prediction," *Expert Systems with Applications*, 59, 226-234.

Kukuk, M., and M. Rönnberg (2013), "Corporate Credit Default Models: a Mixed Logit Approach," *Review of Quantitative Finance and Accounting*, 40(3), 467-483.

Laudon, K., and J. P. Laudon (2012), *Management Information Systems*, 12th Edition, Pearson.

Lu, Y., J. Zhu, N. Zhang, and Q. Shao (2014), "A Hybrid Switching PSO Algorithm and Support Vector Machines for Bankruptcy Prediction," *2014 International Conference on Mechatronics and Control*, IEEE, 1329-1333.

Maltarollo, V., K. Honorio, and A. da Silva (2013), "Applications of Artificial Neural Networks in Chemical Problems," *Artificial Neural Networks: Architectures and Applications*, 203-223.

Mare, D. S., F. Moreira, and R. Rossi (2017), "Nonstationary Z-Score Measures," *European Journal of Operational Research*, 260(1), 348-358.

McCarthy, J., M. L. Minsky, N. Rochester, and C. E. Shannon (2006), "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI magazine*, 27(4), 12-12.

MIT (2016), "The Rise of Data Capital", *MIT Technology Review*, 1-11.

Mitchell, T. (1997), *Machine Learning*, McGraw Hill.

Nam, C., T. Kim, N. Park, and H. Lee (2008), "Bankruptcy Prediction Using a Discrete Time Duration Model Incorporating Temporal and Macroeconomic Dependencies," *Journal of Forecasting*, 27(6), 493-506.

Newman, D. (2011), "How to Plan, Participate and Prosper in the Data Economy", *Gartner*.

Nielsen, M. (2015), *Neural Networks and Deep Learning*, Determination Press.

Ohlson, J. A. (1980), "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research*, 18(1), 109-131.

Olson, D. L., D. Delen, and Y. Meng (2012), "Comparative Analysis of Data Mining Methods for Bankruptcy Prediction," *Decision Support Systems*, 52(2), 464-473.

Pan, Y., T. Y. Wang, and M. S. Weisbach (2018), "How Management Risk Affects Corporate Debt," *Review of Financial Studies*, 31(9), 3491-3531.

Samuel, A. L. (1959), "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, 3(3), 210-229.

Shin, K.-S., T. S. Lee, and H. Kim (2005), "An Application of Support Vector Machines in Bankruptcy Prediction Model," *Expert Systems with Applications*, 28(1), 127-135.

Shumway, T. (2001), "Forecasting Bankruptcy More Accurately: a Simple Hazard Model," *Journal of Business*, 74(1), 101 - 124.

Tian, S., Y. Yu, and H. Guo (2015), "Variable Selection and Corporate Bankruptcy Forecasts," *Journal of Banking & Finance*, 52, 89-100.

Tsai, C.-F., Y.-F. Hsu, and D. C. Yen (2014), "A Comparative Study of Classifier Ensembles for Bankruptcy Prediction," *Applied Soft Computing*, 24, 977 - 984.

Traczynski, J. (2017), "Firm Default Prediction: a Bayesian Model-Averaging Approach," *Journal of Financial and Quantitative Analysis*, 52(3), 1211-1245.

Yang, Z. R., M. B. Platt, H. D. Platt (1999), "Probabilistic Neural Networks in Bankruptcy Prediction," *Journal of Business Research*, 44(2), 67-74.

Zmijewski, M. E. (1984), "Methodological Issues Related to the Estimation of Financial Distress Prediction Models," *Journal of Accounting Research*, 22, 59-82.

Abstract

Corporate Default Predictions and Machine Learning

Hyeongjun Kim, Doojin Ryu**, and Hoon Cho****

Corporate default predictions are essential in every economic sector. Based on these predictions, companies can diagnose their management statuses and establish management strategies. Investors can manage their credit risk and adjust their investment strategies. Governments can also use them to design macroeconomic policies and improve financial regulations. Currently, corporate default predictions are at the forefront of advanced financial engineering that utilizes not only statistical models but also machine learning algorithms. This study reviews the literature on corporate default predictions and the related field by introducing statistical and mathematical models and machine learning algorithms and their representative methodologies. Statistical models can be classified into three generations and can be studied using discriminant analyses, binary response models, and hazard models. For machine learning algorithms, classification methodologies such as support vector machines, decision trees, and artificial neural network algorithms are mainly used. The development of machine learning methodologies is expected to further accelerate innovation in the financial sector, which, in turn, will be followed by the emergence of new financial services and the rise of the data economy.

Key Words: Artificial Neural Networks, Credit Risk, Default, Default Prediction, Machine Learning

* First author, Professor, Department of Business Administration, Yeungnam University, Tel: +82-53-810-2740, E-mail: hkim@yu.ac.kr

** Corresponding author, Full Professor, College of Economics, Sungkyunkwan University, Tel: +82-2-760-0429, E-mail: sharpjin@skku.ac.kr

*** Professor, Graduate School of Finance, Korea Advanced Institute of Science and Technology, Tel: +82-2-958-3413, E-mail: hooncho@kaist.ac.kr