박 사 학 위 논 문

기업 재무 정보를 활용한 머신 러닝 기반 경영 예측 시스템

2016년

한성대학교 대학원

정보컴퓨터공학과

컴퓨터공학 전공

양 진 용

박 사 학 위 논 문
지도교수 허준영

기업 재무 정보를 활용한 머신 러닝 기반 경영 예측 시스템

Machine-learning-based Management Prediction Systems using the Corporate Financial Information

2016년 12월 일

한성대학교 대학원
정보컴퓨터공학과
컴퓨터공학 전공
양 진 용

박 사 학 위 논 문
지도교수 허준영

기업 재무 정보를 활용한 머신 러닝 기반 경영 예측 시스템

Machine-learning-based Management Prediction Systems using the Corporate Financial Information

위 논문을 공학 박사학위 논문으로 제출함

2016년 12월 일

한성대학교 대학원
정보컴퓨터공학과
컴퓨터공학 전공
양 진 용

양진용의 공학 박사학위논문을 인준함

2016년 12월 일

심사위원장 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

국 문 초 록

기업 재무 정보를 활용한 머신 러닝 기반 경영 예측 시스템

한성대학교 대학원
정보컴퓨터공학과
컴퓨터공학전공
양 진 용

갈수록 복잡해지고 고도화되는 기업경영 환경에서 올바른 경영 예측이 기업의 생존과 발전에 있어서 무엇보다 중요해지고 있다. 아울러 쏟아지는 데이터를 적절하게 분류하고 분석하여 가치 있는 정보를 도출하고 예측할 수 있는 기계 학습 기법이 나날이 발전하고 있다. 이에 따라 본 논문은 기업의 재무제표 등 재무 정보와 주가데이터를 활용하여 기계 학습 기법으로 기업파산 예측 및 주가 예측과 같은 경영 예측 응용을 제안한다.

최근 기계 학습을 바탕으로 한 기업 파산 예측 연구가 활발하다. 기계 학습의 대표적 응용 분야인 패턴 인식을 기업의 파산 예측에 응용한 것이다. 기업의 재무 정보를 바탕으로 패턴을 작성하고 이 패턴이 파산 위험 군에 속하는지 안전한 군에 속하는지 판단하는 것이다. 전통적인 Z-Score와 기계 학습을 이용한 파산 예측과 같은 기존 연구들은 특정 산업 분야가 아닌 일반적인 기업을 대상으로 하기 때문에 산업별 특성을 잘 고려하고 있지 못하다. 건설업은 오랜 사업 기간과 대규모 투자, 그리고 투자 회수가 오래 걸리는 특징을 갖는 자본 집약 산업이다. 이로 인해 다른 산업과는 상이한 자본 구조를 갖기 마련이고, 다른 산업의 기업 재무 위험도를 판단하는 기준과 동일한 적용이 곤란할 수 있다. 본 논문은 건설 기업을 규모에 따라 세 등급으로 분류하고 각 기법들의 예측 능력을 비교하였다. 실험 결과 적응형 부스팅이 다른 기법

에 비해 예측 결과가 좋았고, 특히 자본금 규모가 500억 이상인 기업의 경우 아주 우수한 결과를 보였다.

주식 시장 예측은 금융 시계열 예측의 흥미롭고 도전적인 연구 주제로 학계와 비즈니스에서 많은 주목을 받아 왔다. 주식 시장을 효율적으로 예측하는 것은 투자자들에게 매우 중요한 문제이다. 정확한 예측 알고리즘은 투자자들에게 수익 창출과 손실 회피를 가져다 줄 수 있다. 전통적인 시계열 분석으로 주식 시장을 예측하는 것은 어렵다는 것이 입증되었다. 그 대안으로 기계 학습이 부상하고 있다. 기계 학습은 컴퓨터를 학습시켜 분류나 예측에 사용하는 기술이다. 그 중 SVM은 빠르고 신뢰할만한 기계 학습 방법으로 분류나 예측에 널리 사용되고 있다. 본 논문에서는 재무 정보를 기반으로 SVM을 이용하여 주식 가격의 예측력을 검증하였다. 이를 통해 회사의 내재 가치를 나타내는 재무정보가 주식 가격 예측에 얼마나 효과인지를 평가할 수 있다. 회사 재무 정보를 SVM의 입력으로 하여 주가의 상승이나 하락 여부를 예측하고, 다른 기법과의 비교를 위해 전문가 점수와 기계 학습방법인 인공신경망, 결정나무, 적응형부스팅을 통한 예측 결과와 비교하였다. 비교 결과 SVM의 성능이 실행 시간이나 예측력 면에서 모두 우수하였다.

【주요어】 머신 러닝, 예측 시스템, 기업 경영, 파산 예측, 주가 예측

목 차

표 목 차

그 림 목 차

이 논문을 사랑하는 어머니 김초연 선생님의 영전에 바칩니다.

Veritas Liberabit Vos!
진리가 너희를 자유롭게 하리라 요 8:32

I. 서 론

1.1 연구의 목적

갈수록 복잡해지고 고도화되는 기업경영 환경에서 기업의 생존과 발전에 있어 합리적인 경영 예측이 무엇보다 중요해지고 있다. 올바른 경영 예측은 기업의 지속가능한 성장을 위해 필수적인 요소라 할 수 있다. 잘못된 경영 예측은 기업을 생존의 위기로 내몰 수 있을 정도로 그 영향이 매우 크다. 기업 환경의 불확실성과 변동성이 커질수록 경영 예측의 중요성은 더욱 커진다. 이에 따라 기업들은 장, 단기 경영 예측을 위해 많은 연구와 투자를 하고 있으며 예측 경영을 위한 여러 기법과 다양한 예측 시스템들이 개발되고 있다.

아울러 쏟아지는 데이터를 적절하게 분류하고 분석하여 가치 있는 정보를 도출하고 예측할 수 있는 데이터 마이닝과 머신 러닝 기법이 나날이 발전하고 있다. 머신 러닝(기계 학습)은 경험적 데이터를 기반으로 학습 및 예측을 수행하는 기법인데, 머신 러닝 알고리즘은 입력 데이터로부터 예측이나 결정을 이끌어내기 위한 모형을 구축하는 방식을 따른다. 이러한 머신 러닝은 데이터로부터 판단 기준이나 규칙, 지식 표현 같은 것을 추출한다는 점에서 데이터마이닝이나 통계 및 수학적 최적화 문제와 닿아 있다. 우리는 이러한 기법이 우리 생활 깊숙이 들어와 있으며, 여러 산업 분야에 응용되어 탁월한 성과를 내고 있음을 알고 있다.

이에 따라 본 논문은 기업의 재무제표 등 재무 정보와 주가 데이터를 활용하여 머신 러닝 기법으로 1)기업파산 예측 및 2)주가 예측과 같은 경영 예측 응용을 제안하고자 한다. 이는 개별 기업뿐만 아니라 산업적으로도 매우 의미 있는 작업이 될 것이라 생각한다.

1.1.1 기업 파산 예측

근래 들어 부동산 경기침체 여파로 건설기업이 심각한 위기를 겪으며 파산이 급증하였다. 2008년 이후 건설공사의 이윤율이 지속적으로 하락함에 따라 건설기업의 유동성에 부정적인 영향을 주었으며, 2013년 건설 경기 전망 보고서에 따르면 주택건설경기 침체 상황의 지속으로 건설 기업의 유동성 위기가 지속될 것으로 전망된다(CERIK, 2013). 건설업은 파산으로 인한 사회적 파급효과가 다른 산업에 비해 큰 편이지만, 업종의 특성상 다른 산업과는 상이한 자본구조와 부채비율, 현금흐름을 가지고 있어서 기업의 파산 예측이 더 어려운 측면이 있다(Heo & Yang, 2014).

우리나라 전체 GDP에서 건설업이 차지하는 비중은 OECD 국가평균 5.1%에 비해 5.9%로 높은 편이다. 건설 투자는 대규모 자금이 투입되어 경제성장률을 끌어올리는 효과가 있는데 미국의 경우 최근 건설투자의 증가에 힘입어 뚜렷한 경기회복세를 보이고 있는 반면, 우리나라의 경우 부동산침체 여파로 건설투자가 큰 폭으로 감소하고 있어 경기회복의 발목을 잡고 있다. 건설업은 레버리지가 큰 산업으로 부채비율이 매우 높은 업종이며 현금흐름이 프로젝트 후반부에 집중되는 특성이 있다. 그리고 경기사이클에 따른 부침이 매우 심하여 경기하강국면에선 파산이 급증하는 양상을 보인다. 건설업이 레버리지 산업인 이상 건설업체의 파산율 증가는 여신을 공여한 은행에 큰 부담으로 작용한다. 실제 부동산 프로젝트 파이낸싱(project financing, PF)[1]에 집중했던 저축은행들이 부동산 경기침체로 인한 건설업체의 파산으로 동반 몰락한 것이 최근의 일이다. 저축은행뿐만 아니라 시중은행도 상당한 충당금을 쌓아야 하는 등 그 여파를 피해갈 수 없었다. 아울러 복잡한 하도급 관계로 얽혀 있어 건설업체의 파산은 다수 회사의 연쇄 파산으로 이어지기 쉽다. 또한 건

1) 담보나 신용 대신 프로젝트(사업계획)의 수익성을 평가하여 여신을 제공하는 금융기법

설업은 후방산업이라 할 수 있는 시멘트, 철강 등의 업종과 산업연관효과가 커서 건설업의 침체는 산업전반에 영향을 미친다. 한편 건설업은 고용유발효과가 매우 큰 산업으로 건설업체의 파산은 일반서민경제에 미치는 영향이 매우 크다. 그럼에도 그간의 파산예측모형이 주로 금융기관에 집중되어 왔고 건설업종에 특화된 연구는 드물었다.

이러한 상황에서 건설업종에 특화된 파산예측모형을 구축하는 것은 학술적인 의미뿐만 아니라 여러 경제 이해 당사자들에 대한 실무적인 의미가 매우 크다 할 수 있다.

1.1.2 주가 예측

주식 투자자들은 미래주가의 예측을 위해 기본적 분석과 기술적 분석을 사용하여 왔다. 기본적 분석은 주가에 영향을 주는 기업 내외부의 여러 요인들을 분석하여 주식의 내재가치를 추정하고자 하는 방법이다. 장기적으로 주가는 주식의 내재가치에 수렴한다는 믿음을 기반으로 향후 주가흐름을 예측한다. 따라서 내재가치에 영향을 주는 재무요인과 경제요인을 찾아 그 관계를 추정하는 것이 기본적 분석의 주된 방법이다.[2)]

반면 기술적 분석은 주가나 거래량의 움직임에서 미래주가를 예측할 수 있는 패턴을 찾아내고자 하는 방법이다(Heo & Yang, 2015). 기술적 분석에서는 기본적 분석과 달리 주식의 내재가치 요인은 고려하지 않고 주가나 거래량의 움직임 같은 과거자료를 이용하여 미래주가를 예측하고자 한다.[3)] 이러한 기술적 분석은

2) 기본적 분석은 질적 분석(qualitative analysis) 과 양적 분석(quantitative analysis)으로 나눌 수 있는데, 질적 분석은 산업동향, 경제정책, 경영능력이나 기업의 성장성 등과 같은 계량화가 불가능한 것에 대한 분석이고, 양적 분석은 경제지표, 산업지표, 재무제표 등과 같은 계량화가 가능한 것에 대한 분석이다.

3) 기술적 분석가들은 주가나 거래량의 움직임이 주가에 영향을 미칠 만한 여러 요인들을 모두 반

1900년대 초 다우이론(Dow theory)(Hamilton, 1922)[4]을 시작으로 수많은 방법들이 개발되었다. 대부분의 기술적 분석은 도표를 그려 분석하기 때문에 도표분석 방법이라고도 한다.

하지만 효율적 시장가설(efficient market hypothesis: EMH)) 하에서는 주가에 영향을 줄 수 있는 정보가 정확하고 신속하게 주가에 반영되므로 기본적 분석이나 기술적 분석을 통해 미래 주가를 예측하는 것이 불가능해진다(Malkiel, 1987). 효율적 시장가설이 성립하는 시장에서는 주식의 내재가치를 결정짓는 기본요인에 대한 정보가 이미 현재의 주가에 반영되어 있으므로 기본적 분석을 통해 초과수익을 얻을 수 없으며, 주가의 움직임은 임의 보행에 가깝기 때문에 어떤 특정 패턴을 찾아내어 초과수익을 얻는 것도 불가능하다(Conrad & Kaul, 1989). 이러한 효율적 시장가설의 실증분석을 위해 사건 연구, 수익률의 예측가능성에 관한 연구들이 이루어져 왔다(Timmermann & Granger, 2004). 전자는 새로운 정보가, 후자는 과거정보가 시장가격에 잘 반영되고 있는지를 검증하는 것이다.

새로운 정보에 대한 시장가격의 반응을 통한 시장효율성검증, 즉 사건연구(event study)는 Fama와 Fisher, Jensen, Roll에 의해 시작되었다(Fama, 1991; Lim & Brooks, 2011; Sewell, 2011). 이들의 연구결과는 효율적 시장가설을 지지하지만 이후 다른 연구자들의 연구결과는 엇갈리고 있다. Ball, Brown의 연구는 지지하는 결과를, Rendleman, Ibbotson 등의 연구는 부정하는 결과를 보여주고 있다(Rendleman, Jones & Latane, 1982).

영한 것으로 판단하며 이에 근거하여 미래 주가의 움직임을 예측한다.

4) Charles H. Dow로부터 비롯된 이론으로 기술적 분석의 시초가 된 이론이다. William Peter Hamilton이 이론적으로 체계화하였으며, Robert Rhea가 더욱 발전시켰다. 다우이론은 주가가 임의보행(random walk)을 하는 것이 아니라 주기적 추세에 의해 영향을 받는다는 것이 핵심이다. 즉, 주가가 어떤 방향성을 갖게 되면 그 추세가 전환되는 신호가 나타날 때까지 그 방향을 관성적으로 유지한다는 가설이다.

그러나 점점 더 많은 연구가 효율적 시장가설(EMH)를 비판적으로 검토하기 시작하였는데, 특히 행동 경제학(behavioral ecnomics) 및 행동 재무학(behavioral finance), 사회경제적 금융이론 (Socionomic Theory of Finance, STF)[5]의 관점을 따르는 비판적 연구들은 효율적 시장가설에 의문을 제기하였다. 더욱이 많은 연구 결과들이 주식 시장 가격은 랜덤 워크(random walk)를 따르지 않아 어느 정도 예측이 가능하다는 것을 보여주면서 효율적 시장가설의 기본 가정은 큰 도전에 직면하게 되었다.

이처럼 효율적 시장가설에 대한 여러 반대증거가 나오고 있는 상황에서 통계에 기반 한 주가예측 연구와 더불어 컴퓨터 기술의 비약적인 발전에 힘입은 기계학습 기반 주가예측연구들이 활발히 이루어지고 있다. 인공 신경망(Neural Network), 유전알고리즘(Genetic Algorithm), 은닉 마르코프 모형(Hidden Markov Model, HMM), 파형요소(wavelet), 서포트 벡터 머신(Support Vector Machine, SVM), 퍼지 논리(fuzzy logic), 결정나무(Decision Tree), 적응형 부스팅(AdaBoost) 등과 이것을 조합한 무수한 방법들이 시도되고 있다.

이는 알고리즘 트레이딩(Algorithm Trading)[6]을 넘어 새롭게 대두되는 로보어드바이저(Robo-Advisor)[7]의 등장에 이르기까지 금융 산업에도 큰 영향을 미치고 있다. 기계학습 기반 헤지펀드들의 등장은 이러한 흐름을 잘 말해주고 있다.

---

5) 이 이론들은 경제주체들이 온전히 합리적이라는 주장을 부정하고 때로는 감정적으로 선택하는 경향이 있다고 주장한다. 이 이론들에 따르면, 재무적 가치를 결정하는 것은 객관적, 의식적, 합리적인 결정이 아니라 주관적, 무의식적, 비합리적인 충동이 재무적 가치 결정에 큰 영향을 미친다. 보유효과, 손실회피성향, 심리적 회계, 무리 효과, 프레이밍 효과 등이 그 예이다.

6) 데이터를 조합해 구성된 일정한 논리구조에 따라 컴퓨터 시스템이 주식, 파생, 외환 등의 거래를 자동으로 수행하는 것을 말한다.

7) 로보어드바이저(Robo-Advisor)는 로봇(robot)과 어드바이저(advisor)'의 합성어로 알고리즘을 활용해 자산 운용의 자문 및 관리를 해주는 것을 말한다. 로보어드바이저는 과거 데이터로부터 미래를 예측하는 것을 넘어서 스스로 데이터 조합을 익히고 학습하는 기계학습 기술을 적용하여 다양한 미래 변수를 고려하는 예측을 수행한다.

본 논문에서는 재무정보에 기초한 주가 예측에 있어서 여러 기계학습 방법의 성과를 비교 분석하였는데 학술적 의미뿐만 아니라 산업적 응용도 가능하리라 생각한다. 이러한 주가 예측은 경영자, 투자자 및 관련 업계 종사자 등 주가 관련 이해당사자들에게 유의미한 정보를 제공해 줄 수 있기 때문이다.

1.2 연구 내용

1.2.1 파산 예측

전통적인 Z-Score와 기계 학습을 이용한 파산 예측과 같은 기존 연구들은 특정 산업 분야가 아닌 일반적인 기업을 대상으로 하기 때문에 기업들의 특성을 전혀 고려하고 있지 못하다. 본 논문에서는 건설 기업을 규모에 따라 각 기법들의 예측 능력을 비교하여 적응형 부스팅(AdaBoost)이 가장 우수함을 확인하였다.

본 논문은 건설 기업을 자본금 규모에 따라 세 등급으로 분류하고 각각에 대해 적응형 부스팅(AdaBoost)의 예측력을 분석하였다. 적응형 부스팅(AdaBoost)의 상대적 성능 측정을 위해 다른 기계 학습 방법인 인공 신경망(ANN), SVM(Support Vector Machine), 결정 나무(decision tree)와 비교를 수행하였다. 실제로 2008년부터 2012년까지 파산한 기업과 2012년 정상 운영 중인 기업의 재무 자료를 사용하여 각 기법들의 예측력을 측정하였다. 실험 결과 적응형 부스팅(AdaBoost)이 다른 기계학습 기법에 비해 예측 결과가 좋았고, 특히 자본금 규모가 500억 이상인 기업의 경우 아주 우수한 결과를 보였다.(Heo & Yang, 2014)

1.2.2 주가 예측

본 논문에서는 Han이 연구한 방법(Han & Chen, 2007)과 유사하게 SVM을 사용하여 기본적 분석을 통한 기업 주식 가격 등락을 예측한다. 즉, 기업의 재무정보를 SVM의 입력으로 사용하여 기본적 분석, 즉 기업의 내재 가치를 나타내는 재무 정보에 따른 분석을 하고 이 결과로 주식의 향후 등락을 예측하는 것이다. Han의 방법(Han & Chen, 2007)과 차이는, 효율적 시장가설의 실증 분석을 위한 사건 연구로서 기업의 재무 정보 발표를 사건으로 보고 그로 인한 주식 가격 등락을 예측한다는 것이다. 이를 위해 재무 정보 발표 후 1개월과 2개월 후의 가격을 예측하였다.

기본적 분석을 위한 재무 정보로 자산과 이익에 대한 정보를 활용하였다. 재무 정보에서 자산과 이익에 대한 정보는 대표적으로 해당 기업의 재무 상태를 설명해줄 수 있는 지표이다(Chung & Kim, 2010). 본 논문은 이 지표에 따라 주가의 등락을 예측할 수 있는지, 그리고 그 예측이 어느 시점까지 가능한지 기계학습 기법인 SVM을 사용하여 평가하였다. 본 논문에서 사용한 재무정보는 상대가치지표로 주당순이익(EPS), 주당순자산(BPS)과 성장성을 고려해 볼 수 있는 순이익증가율(NPGR)이다(Han & Chen, 2007; Song, 2011).

실제로 주가는 학술적 연구뿐 아니라 증권 애널리스트들에 의해 분석되고 예측되고 있다. 애널리스트들의 예측은 주식 시장에서 그나마 신뢰받는 정보로, 이들의 분석은 어느 정도 검증된 분석, 예측 기술에 바탕을 둔 것으로 볼 수 있다. 따라서 전문가의 예측과 기술적 예측 방법의 비교가 필요하다. 본 논문에서는 이런 전문가의 예측과 본 논문에서 제안하는 방법의 예측을 비교하여 전문가의 예측보다 더 우수함을 보였다(Heo & Yang, 2015). 아울러 다른 기계학습 방법인 인공 신경망

(Artificial Neural Network, ANN), 결정 나무(Decision Tree, DT), 적응형 부스팅(Adaptive Boosting, AdaBoost) 등과 비교하여 SVM이 더 우수한 결과를 보여주었다.

II. 관련 연구

2.1 기업 파산 예측 관련 연구

기업의 재무 자료를 바탕으로 한 파산 예측 모델에 대한 연구는 오래 전부터 다양하게 진행되었다. 하지만, 일반적인 기업 전체를 대상으로 하는 모델이기 때문에, 건설 기업과 같이 유동성이 큰 기업의 예측에는 적절하지 못할 수 있다. 건설 산업은 오랜 사업 기간과 대규모 투자, 그리고 투자 회수가 오래 걸리는 특징을 갖는 자본 집약 산업이다. 이로 인해 다른 산업과는 상이한 자본 구조를 갖기 마련이고, 다른 산업의 기업 재무 위험도를 판단하는 기준과 동일한 적용이 곤란할 수 있다(Sun et al., 2013).

기업 파산 예측은 Fitzpartrick(1932)의 연구로부터 시작되었다고 할 수 있으며, Beaver(1966)는 일반화 선형모형(generalized linear model) 같은 통계적 방법을 적용하여 30개의 재무 비율 분석에 기초한 모형을 만들었다. 이후 다변량 판별분석(multivariate discriminant analysis)과 로지스틱 회귀분석(logistic regression analysis)이 도입되었다. MDA는 예측 인자의 분포 특성에 따른 특정 통계 요건이 있는 점과 모형의 출력이 거의 직관적인 해석을 갖지 않는 점수라는 점, “matching” procedure와 관련된 문제가 있는 점 등 몇 가지 한계로 인하여 Ohlson (1980)은 조건부 로짓 분석(conditional logit analysis)을 사용하여 샘플의 파산확률분포를 분석하였다.

Kumar와 Ravi (2007)는 1968-2005년 동안의 파산 예측 연구를 포괄적으로 리뷰하면서 i)통계적 기법 (statistical techniques), ii)신경망(statistical techniques), iii)사례기반추론(case-based reasoning, CBR)[8], iv)결정 나무(decision trees), v)오퍼레

8) Case-based reasoning (CBR)은 비슷한 과거의 문제의 해결책을 바탕으로 새로운 문제를 해결하는 과정이라 할 수 있다.

이션 리서치(operational research)[9], vi)진화적 접근법(evolutionary approaches)[10], vii)러프 집합 기반 기법(rough set based techniques), viii)퍼지 논리(fuzzy logics), 서포트 벡터 머신(support vector machine), 아이소토닉 분리(isotonic separation)[11] 등을 포함하는 기타 기법, ix)위 기법들을 통합하는 하이브리드 소프트 컴퓨팅 등의 9가지 범주로 분류하였다.

가장 대표적인 기업 파산 예측 연구로는 알트만의 Z-score 모형을 들 수 있다(Altman, 1968). Z-score는 1968년 처음 발표되었으며, 간단한 식을 통해 기업의 파산 가능성을 예측한다. Z-score는 계산된 값을 세 구간으로 나누어 위험, 중간, 안전으로 판단한다.[12] 위험 구간에 속하면 향후 2년 내에 파산 가능성이 높으며 안전 구간에 속하면 파산 가능성이 적다고 보는 것이다. 하지만 중간 구간에 속하는 경우 판단을 하지 못하며, 본 논문에서 사용한 건설 기업의 경우 많은 수가 이 구간에 속하는 단점이 있다. Z-score 모형과 더불어 실무적으로 널리 쓰이는 것으로 ZETA® 신용 위험 모형이 있는데, ZETA Services, Inc.에서 가입자에게 배타적으로 제공하는 모형이다. 이는 1976년 Altman이 공동 개발한 2세대 모형이다.[13] Grice와 Ingram (2001)은 Altman 모형이

9) operational research는 보다 나은 결정을 내리는데 도움이 되는 고급 분석 방법의 적용을 다루는 분야이다.

10) 진화 및 유전의 원리를 문제 해결에 응용하려는 연구의 한 방법이다. evolutionary strategies, evolutionary programming, genetic algorithm, genetic programming 등이 있다.

11) Isotonic separation은 지도 기계학습 기법(supervised machine learning technique) 인데, 여기에서 분류(classification)는 misclassification의 수를 최소화하는 것을 목적으로 하는 선형 프로그래밍 문제(linear programming problem, LPP)로 표현된다.

12) Z = 0.012X1 + 0.014X2 + 0.033X3 + 0.006X4 + 0.999X5
(X1~working capital to total assets ratio, X2~retained earnings to total assets ratio, X3~earnings before interest and taxes to total assets ratio, X4~market value equity to book value of total debt ratio, X5~sales to total assets ratio)

13) 파산 시점 1년 이전에서는 Z-Score와 ZETA 모형이 비슷한 정확도를 보이지만 2~5년 이전에서는 ZETA모형의 정확도가 더 높은 결과를 보인다.

개발된 당시처럼 현재에도 유용한지, 비제조업에 대해서도 제조업만큼 유용한지에 대해서 다소 부정적인 평가 결과를 내놓았다.

이러한 모형들은 기업파산의 효과적인 지시자(indicators), 예측자(predictors) 역할을 할 수 있는 변수들을 특정하기 위해 기업 실패의 독특한 특성을 찾고자 한다. 특히, 복합적인 판별 통계 방법론(multiple discriminant statistical methodology)을 사용하여 재무 및 경제 비율 집합(a set of financial and economic ratio)을 분석하고 잠재적 파산의 수량화 가능한 특성들을 찾는다.[14] 이러한 모형들은 1960년대 후반에서 1970년대 중반에 개발된 이래 제조업, 비제조업[15] 등 다양한 부문에 적용이 확대되었으며, 채권평가모형(bond-rating equivalent model)으로도 발전하였다.[16] (Altman, 2000) 이러한 비율분석에 기초한 분석기술은 학술연구자들 뿐만 아니라 시장이론가와 실무가들에게 매우 다양한 형태로 변형 응용되어 여러 수정모형(revised model)들이 등장하였다.

Laitinen (2001)은 파산예측에 있어서 로지스틱 회귀 모형(logistic regression model)의 정확도를 증가시키기 위하여 테일러 전개(Taylor expansion)[17]를 적용하였다. 변수의 정규성(normality)과 관련된 문제

---

14) 모형에 사용되는 Traditional Ratio Analysis와 Discriminant Analysis, 그리고 표본 선정(Sample Selection)과 변수 선정(Variable Selection) 방법에 대해서 Altman(2000)을 참조하라.

15) Altman은 비제조업에 대해서 새로운 Z3 Score 모형을 구성하였다.
Z3 = 6.56X1 + 3.26X2 + 6.72X3 + 1.05X4
(Z3>2.6 ~ a good financial condition, 1.8<Z3<2.6 ~ an insecure financial condition, Z3<1.8 ~ a very high possibility of financial failure)

16) 예를 들면, Altman, Hartzell, and Peck(1995)는 original Altman Z-Score 모형을 변형하여 emerging market scoring (EMS) 모형을 만들었다. public bond markets에 대한 연구로 Altman & Eberhart (1994), Standard & Poor's(1995), Carty & Lieberman(1996) 등도 참고할 만하다.

17) 테일러 전개(Taylor expansion)는 비선형미분방정식체계를 분석을 할 수 있는 선형 근사(linear approximation))기법으로 균형점을 중심으로 하여 주어진 체계의 값을 구한다. 선형 근사(linear approximation)는 어떤 함수의 테일러급수로부터 2계 이상의 항을 제거함으로 구할 수 있다.

를 피하기 위해 로지스틱 모형(logistic model)을 적용한 다음 로지스틱 함수의 지수(exponent of the logistic function, or logit)를 근사하기 위해 테일러 전개(Taylor expansion)를 이용하였다. 총자산대비 현금(cash to total assets), 총자산대비 현금 흐름(cash flow to total assets), 총자산대비 자기 자본(shareholder's equity to total assets) 비율 등을 파산 위험에 영향을 미치는 요인으로 하는 분석으로 파산 1~2년 전 예측에서 높은 정확도를 보여 주었다.

로지스틱 회귀분석(logistic regression analysis)과 생존율(survival rate)의 개념을 혼합한 형태의 생존분석 모형인 해저드 모형을 사용한 연구도 있다. Chava와 Jarrow (2004)는 1962-1999년 기간 동안 미국 기업에 대하여 bankruptcy hazard rate model의 예측 정확도를 연구하였다. 확장된 파산 데이터베이스를 이용하여 해저드 모형 (Shumway, 2001)의 Altman (1968)과 Zmijewski (1984)모형 대비 우수한 예측성과를 검증하였고, hazard rate estimation 에 산업 효과(industry effects)를 포함시키는 것이 중요하다는 것을 증명하였다. Nam 등(2008)은 Shumway (2001)의 연구를 확장하여 시변 공변량(time-varying covariates)과 기저 위험 함수(baseline hazard function)를 가진 듀레이션 모형(duration model)을 제안하였다. 제안된 모형은 시간 및 거시경제 의존성(dependency)을 허용함으로써 얻어진 개선 효과와 예측 성능을 입증하였다. Bharath (2008)은 해저드 모형 (hazard model)이 Merton DD 모형 (distance to default model)[18] 과 축약형 모형(reduced-form model)[19] 보다 우수한 성과를 나타냄을 검증하였다. Beaver 등(2005)은 해저드 모형을 사용하여 1962-2002년 동안 파산을 예측할 수 있는 재무제표 데이터의 능력이 변화하는지를 조사하여 파산 예측 능력에 영향을 미칠 수 있는 세 가지 추세를 확인하였다. 여기에는 FASB 기준, 재량적 재무보고 행동의 증가, 인

18) Merton(1974)의 bond pricing model을 기초로 하는 신용 위험 모형(credit risk model)이다.

19) 머튼 모형과 축약형 모형은 신용 위험 관리 모형 (credit risk management model) 중 널리 사용되는 모형이다.

식되지 않은 자산 및 부채의 증가 등이 포함되었다. parsimonious three-variable model은 기간 전체에 걸쳐 중요한 설명력을 제공하였고 40년 기간 내내 예측 모형은 견고성(robustness)이 유지되었다.

Sung 등(1999)은 정상 및 위기 경제 상황에 적합한 예측 모형 개발을 위해 데이터 마이닝 접근 방식을 사용하는데 정상 상태에서 위기 상황으로 모형 변화의 역동성을 관찰하면서 파산을 예측하는 주요 변수로 정상 상태에서는 "총자산 대비 현금 흐름(cash flow to total assets)", "자본 생산성(productivity of capital)"을 위기 상황에서는 "부채 대비 현금 흐름(cash flow to liabilities)", "자본 생산성(productivity of capital)", "자기자본과 장기부채 대비 고정 자산(fixed assets to stockholders equity and long-term liabilities)"를 지적하였다.

파산 예측 모형을 크게 두 가지로 분류한다면, 재무비율 등을 이용한 회계 기반의 모형(accounting-based model)과 옵션(option) 등을 이용한 시장 기반의 모형(market-based model)로 나눌 수 있다. 회계 기반 모형의 일반적인 가정은 근본적인 경제적 요인과 기업의 특성이 재무제표에 반영된다는 것이고, 시장 기반 모형의 일반적인 가정은 시장에서 거래되는 가격에 모든 정보가 담겨있다는 것이다.

Berg (2006)는 파산예측을 위한 여러 가지 회계기반 모형을 비교한 결과 일반화 가법 모형(Generalized Additive Model, GAM)[20]이 모든 위험 수준에서 선형 판별 분석(linear discriminant analysis), 일반화 선형 모형(generalized linear model), 신경망 등과 같은 모형보다 좋은 성능을 나타냄을 보여주었다. 회계기반 모형의 전통적인 지표인 회계비율을 이른바 multinorm analysis의 결과로 대체하여 분석하려는 시도도 있다. Andrés 등(2013)은 비모수 분위 회귀분석(nonparametric quantile

20) 일반화 가법 모형(Generalized Additive Model, GAM)은 1990년 Trevor Hastie와 Rob Tibshirani에 의해 개발되었으며, 일반화 선형 모형(generalized linear model)과 가법 모형(additive models)의 속성을 혼합한 통계 모형(statistical model)이다. 일반 회귀식 형태에 비모수 형태의 식이 가산되어 있다. GAM은 비모수 회귀분석(nonparametric regressiom)과 smoothing technique을 통해 선형이 아닌 경우에도 정확한 추정을 할 수 있도록 해 준다.

regression)에 의해 계산된 일련의 industry norms로부터의 각 회사 편차를 분류기(classifier)의 입력변수로 사용한 결과, 선형 및 비선형 분류기(classifier) 모두에서 예측 정확도를 크게 향상시킬 수 있었다.

Reisz (2007)는 보통주를 자산에 대한 down-and-out barrier option으로 간주하는 모형으로 1988-2002년 기간 동안 5,784개의 기업에 대한 파산 가능성을 추정하였다. Agarwal와 Taffler (2008)은 contingent-claims valuation approach 같은 market-based approach와 전통적인 accounting-ratio-based approach를 비교 평가하여 영국 기업에서 두 모형의 예측력 사이에 별 차이가 없음을 밝혔다. Charitou와 Trigeorgis (2000)은 옵션 가격 이론(option-pricing theory)을 토대로 하여 1983-1994년 기간 동안 미국 기업의 파산을 설명하였는데, 변동성(volatility)같은 일차적인 option-motivated 변수가 1~3년 전 파산 예측에 중요한 역할을 한다는 것을 밝혔다. Hillegeist 등(2004)은 옵션가격이론에 따르는 market-based measure인 BSM-PB(Black and Sholes, Merton, probability of bankruptcy)가 accounting-based measure인 Altman's Z-score (1968)와 Ohlson의 O-score (1980) 보다 더 나은 설명력을 갖는다는 것을 보였다. 하지만, BSM-PB가 파산확률과 연관된 모든 가능한 시장기반 정보(market-based information)를 반영하지 않으며, 초과 수익(excess returns)과 상대 시장 크기(relative market size)가 부가적인 정보를 생산한다고 하였다. Bharath와 Shumway (2004)는 KMV-Merton 모형의 파산 확률이 CDS(Credit Default Swap)과 회사채 수익률 스프레드(corporate bond yield spreads)의 내재된 파산 확률과 단지 약한 상관관계를 가지고 있으며, 모형에서 요구되는 연립 비선형 방정식을 풀지 않고도 충분한 통계를 구성하는 것이 가능하다고 하는 다소 비판적인 입장을 보였다.

회계 기반 모형(accounting-based model)과 시장 기반 모형(market-based model)을 통합한 모형도 연구되었다. Li와 Miu (2009)는

BQR (binary quantile regression) 모형을 사용하여 회계 비율 기반 (accounting-ratio-based) 및 시장 기반 (market-based) 정보 모두에 동적 부하(dynamic loadings)를 지닌 하이브리드 파산 예측 모형(hybrid bankruptcy prediction model)을 확립하여 기존의 로짓 모형(logit model)보다 우수한 성능을 입증하였다. Das 등(2009)도 회계 기반 모형과 시장 기반 모형 둘 중 하나만을 사용하는 것 보다 둘 다 사용하는 것이 좋은 성과를 나타냄을 보여주어 두 종류의 정보가 상호보완적임을 입증하였다.

1990년대에 들어 컴퓨터를 이용한 기계 학습이 발전함에 따라 최근에는 기계 학습(e.g., neural networks, Support Vector Machines, decision trees)을 바탕으로 한 기업 파산 예측 연구가 활발하다. 기계 학습의 대표적 응용 분야인 패턴 인식을 기업의 파산 예측에 응용한 것이다. 기업의 재무 정보를 바탕으로 패턴을 작성하고 이 패턴이 파산 위험 군에 속하는지 안전한 군에 속하는지 판단하는 것이다. 기계학습은 일반적으로 선험적 가정(a priori assumption)이 필요하지 않고 비선형적인 관계를 추론할 수 있기 때문에 통계적 방법(statistical method)에 비해 복합적인 유추가 가능한 특성이 있다. 다만 이러한 모형을 해석하고 이해하는데 다소 어려움이 따를 수 있으며, 해석 가능성이 주요 이슈인 일부 문제에 있어서는 사용이 제한될 수 있다. 파산예측은 모형의 이해가능성을 요구하는 작업이며, 예측모형을 생성하기 위해 입력변수를 분석하면 파산을 유발할 수 있는 조건에 대한 더 많은 지식을 얻을 수 있다.

파산 예측에 사용되는 대표적인 기계 학습으로는 인공 신경망(Artificial Neural Networks) (Tae & Shin, 2010; Wilson & Sharda, 1994) 과 적응형 부스팅(AdaBoost)(Sun et al., 2013; Altman, 2000)이 있다. 이 외에도 SVM(Support Vector Machine)(Min & Lee, 2005; Shin et al., 2005)을 사용한 연구도 있다. 이 기법들을 결합한 하이브리드 연구도 다양하다(Kim, 2009; Shin & Hong 2011; Verikas et al., 2010).

Tam (1991)은 파산예측모형에 인공신경망(Artificial Neural Network, ANN)을 도입하여 예측 능력을 향상 시켰다. 이것은 입력(input), 숨김(hidden), 출력(output) 층(layers)의 시뮬레이션에 기초한다. Wilson (1994)과 Odom (1990) 역시 파산예측 신경망 모형을 개발하여 전통적인 파산예측방법인 다변량 판별분석 (multivariate discriminant analysis)과 비교분석한 결과, 신경망 모형이 더 우수한 예측력을 나타내어 파산예측문제에 매우 적합한 모형임을 보여주었다. Leshno와 Spector (1996)은 다양한 신경망 모형의 예측능력을 평가하였다. 이 연구에서 조사된 모형은 데이터 기간(data span), 학습기술(learning technique), 반복 횟수(number of iterations)와 같은 매개 변수에 따라 다르다. 모형의 예측 능력은 향상된 학습기술의 사용에 의해 개선되지만 향상 학습기법이 너무 강하면 모형이 훈련데이터 집합에 너무 specific하게 되므로 예측 능력이 떨어질 수 있다. Zhang 등(1997)은 파산 예측에서 인공 신경 신경망 (ANN)의 역할을 이해하기위한 일반적인 틀을 제시하고 신경망과 전통적인 베이지안 분류 이론(Bayesian classification theory)과 연결을 설명하였다. 아울러 신경망이 예측 및 분류 율 추정에서 로지스틱 회귀분석보다 월등함을 보여주었다. 또한 신경망은 전체 분류성능에서 샘플링 변화에 강건하였다. Yang 등(1999)은 신경 판별 모형(neural discriminant model)의 실제적인 유용성을 제한하는 단점을 지적하고 대안으로 확률 신경망(probabilistic neural network)[21]을 제안하였다. 미국 석유 및 가스 산업의 데이터를 사용하여 비교분석한 결과 패턴정규화(pattern normalization)가 없는 확률 신경망(probabilistic neural network)과 피셔 판별분석(Fisher discriminant analysis)이 전반적으로 최상의 추정결과를 얻었다. Atiya (2001)는 전통적인 재무비율과 Merton의 신용 위험 모형에서 차용한 indicators를 신경망 모형에 적용하여 예측력을 크게 향상시켰다. 일반적으로 단일 분류자 (single classifier)에 비해 여러 분류자 (multiple classifiers or classifier

21) 확률신경망 (probabilistic neural network, PNN)은 베이지안 네트워크(Bayesian network)와 커널 피셔 판별 분석 (Kernel Fisher discriminant analysis)이라 불리는 통계 알고리즘으로 부터 파생된 피드포워드 신경망(feedforward neural network이다. D.F. Specht가 1990 년대 초반 도입하였으며, 작동(operation)은 4 개의 층(layer)을 가진 다층 피드포워드 네트워크로부터 조직된다.

ensemble)를 조합하는 것이 더 낫다고 여겨지는데, Tsai와 Wu(2008)은 파산예측문제에 있어서 3개의 데이터 집합에 기초한 신경망 앙상블(neural net ensembles)을 이용하여 단일 분류자(single classifier)와 다중 분류자(multiple classifiers) 및 다양화된 다중 분류자(diversified multiple classifiers)의 성능을 조사한 결과 정확한 승자를 찾지 못하고, 최적의 결정을 내리기 위해서 이러한 세 가지 분류기 아키텍처(clsssifier architecture)를 고려해야 한다고 하였다. 아울러 Tsai(2009)는 다층 퍼셉트론 신경망(Multi-layer perceptron neural networks)을 파산 예측 모형으로 한 연구에서 5개의 잘 알려진 feature selection methods (t-test, correlation matrix, factor analysis, principle component analysis, stepwise regression)를 비교한 결과, t-test feature selection method가 가장 우수한 성과를 보였다. Ravi와 Pramodh (2008)는 파산예측문제에 주성분 신경망(principal component neural network, PCNN)을 적용하는 것과 새로운 feature subset selection (FSS) 알고리즘을 제안하였다. 여기서 hidden layer는 principal component layer로 대체되며 hidden node의 기능을 수행하는 몇 가지 선택된 principal component들로 구성된다. 아울러 PCNN을 훈련시키기 위하여 threshold accepting(TA) meta-heuristic 기반 알고리즘을 도입하였다. 실험 결과, PCNN의 높은 일반화 능력과 선택된 feature subset의 높은 판별능력(discriminating power)을 보여주었으며, PCA-TANN, PCA-BPNN 등 다른 분류기보다 성능이 우수하다는 것을 입증하였다.

사례 기반 추론 (CBR)은 복잡하고 변화하는 비즈니스 환경에서 문제를 해결하고 의사 결정을 내리는 방법이다. 많은 CBR 알고리즘은 k-nearest neighbor (k-NN) 방법에서 파생된 것으로, 저장된 사례로부터 분류를 생성하는 similarity function을 갖고 있는데 k-NN의 성능은 similarity function의 정의(definition)에 대해 매우 민감하다. Park과 Han (2002)는 AHP(analytic hierarchy process)[22]가중(weighted) k-NN 알고리즘이라는

22) AHP 모형은 수많은 전문가로부터 특정 분야의 지식을 얻고 지식 기반 색인을 작성하는 데 효과적인 방법론이다. 사례 색인 및 검색 (case indexing and retrieving)에서 상대적 중요도를 할당하는데

새로운 프레임워크(framework)를 이용한 유비 추론 구조(analogical reasoning structure)를 제안하였다. AHP 가중 k-NN 알고리즘은 순수한 k-NN 알고리즘보다 높은 분류 정확도를 달성하는 것으로 나타났다. 양적(재무 비율), 질적 (비재무 변수)기준을 관련시킨 파산 예측에 이러한 접근법을 적용시킨 결과 매우 좋은 성능을 보여 주었다. Bryant (1997)도 파산 예측모형에 사례 기반 추론 접근법(case-based reasoning approach)을 적용하였다.

사례기반 추론(CBR)은 적용하기 쉽고 과적합(overfitting) 가능성이 없으며 출력(output)에 대한 좋은 설명(explanation)을 제공하는 반면, 예측 성능이 다소 낮은 편이다. Ahn과 Kim (2009)은 파산 예측을 위한 CBR의 예측성과를 높이기 위해 유전 알고리즘(genetic algorithm)을 사용하여 CBR에 대한 feature weighting과 instance selection을 동시에 최적화하였다. 이러한 방법은 연관성 높은 사례를 참조하고 노이즈(noise)를 제거하여 기존 CBR의 예측 정확도를 크게 향상시켰다. Jo와 Han (1997)은 사례기반 추론(case-based reasoning), NN, MDA 및 이들의 조합을 비교하여 조합 접근법(combination approach)이 예측 능력 측면에서 어떤 단일 방법보다 우수하다는 것을 발견했다. Cho 등(2010)은 마하라노비스 거리(Mahalanobis distance)[23]를 이용한 사례기반 추론과 결정 나무(decision tree)를 이용한 변수 선택을 결합한 하이브리드 파산예측 방법을 제안하였다. 결정 나무에 의해 선택된 변수는 회귀(regression)에 의해 생성된 변수와 비교하여 상호 작용을 하는 경향이 있으며, 변수끼리 서로 상관관계가 있을 때는 마하라노비스 거리가 유클리드 거리보다 근접성을 더 정확하게 측정한다. 실험 결과 제안된 방법은 현재 사용 중인 다른 기법보다 우수한

AHP 방법론을 사용할 수 있다.

23) 마하라노비스 거리(Mahalanobis distance)는 1936년 P. C. Mahalanobis에 의해 소개된 point P와 distribution D 사이의 거리를 측정한 값이다. 이것은 P가 D의 평균으로부터 몇 표준편차만큼 떨어져 있는지를 측정한다는 개념을 다차원적으로 일반화한 것이다. 각 축의 단위 분산을 조정한 경우, Mahalanobis 거리는 변형된 공간의 표준 유클리드 거리에 해당한다.

성능을 나타냈다.

인공 신경망은 파산예측 문제에 오랫동안 성공적으로 적용되어 왔는데, 역전파 신경망(back-propagation neural network, BPNN[24])이 예측 정확도 면에서 가장 우수한 것으로 알려져 있다.[25](Chen & Du, 2009; Liu & Marukawa, 2004). Lee와 Choi (2013)도 역전파 신경망(BPNN)을 이용해 한국 기업 파산에 대한 산업별 조사를 하였는데, 산업별 표본 예측이 전체 표본 예측보다 우수하였고, BPNN을 사용한 예측 정확도가 다변량 판별 분석(MDA)의 예측 정확도 보다 훨씬 우수하였다. 그런데 역전파 신경망은 적용 이전에 네트워크 토폴로지(network topology), 학습 매개변수(learning parameters), 입출력 벡터의 정규화 방법(normalization methods for the input and output vectors) 등과 같은 몇 가지 주요 이슈들을 먼저 숙고해야 한다. 예측 성능을 향상시키기 위하여 네트워크 토폴로지와 학습 매개변수를 최적화하는 방법에 대해 많은 연구자들이 관심을 가져왔는데, Tae와 Shin (2010)은 유전자 알고리즘 기반 정규화 변환(GA-based normalization transform)을 제안 한다[26]. 그밖에도 다양한 데이터 정규화 방법[27]이 기계학습에 적용되어 왔는데, Z-score(Jolai & Ghanbari, 2010), mean(Wang & Zhang, 2009), min-max(Shalabi & Shaaban, 2006), median(Jain et al., 2005),

---

24) VPN은 오랫동안 다층 망(multi-lauer networks)에서 가장 효율적인 학습 절차 중 하나이다. BPN 학습 방법은 감독 학습(supervised learning)이며, BPN의 activation function은 combination function과 trasfer function으로 구성되어 있다.

25) 로지스틱 회귀분석, 선형 판별분석, 다중 판별분석, 결정 트리 등과 비교해서도 좋은 예측 능력을 보여준다.

26) 여기서 GA(Genetic Algorithm)는 일반화를 위한 최적 가중치(the optimal weight for the generalization)를 추출하는데 사용된다. GA는 1975년 J. Holland에 의해 개발된 전역 최적화 기법으로 자연세계의 진화과정에 기초한 계산 모형이다. 일반적으로 GA는 초기화(initialization), 선택(selection), 교차(crossover), 돌연변이(mutation)의 4단계 프로세스를 수행하는데, 문제 영역에 특정 조건이 주어지면 크고 복잡한 공간을 탐색할 수 있다.

27) 정규화(normalization)란 “scaling down” 변환이라 할 수 있다.

range(Mazzatorta & Benfenati, 2002) 등이 있다.[28] Chauhan 등(2009)은 wavelet neural network(WNN)을 훈련하는데 differential evolution algorithm(DE)를 제안하였다. 이러한 differential evolution trained wavelet neural network (DEWNN) 모형은 파산예측의 정확도(accuracy)와 민감도(sensitivity)에서 original WNN과 threshold accepting trained wavelet neural network(TAWNN)을 능가하는 성능을 보여주었다. Lee 등(1996)은 파산예측을 위한 하이브리드 신경망 모형을 개발하였다. 제안된 하이브리드 신경망은 MDA 보조 신경망(MDA-assisted neural network), ID3[29] 보조 신경망(ID3-assisted neural network), SOFM 보조 신경망(self organizing feature map[30]-assisted neural network)인데, 실험 결과 하이브리드 신경망 모형이 에측 정확도와 적응성 측면에서 파산 예측을 위한 매우 유망한 신경망 모형임을 보여주었다.

퍼지 집합은 데이터의 불확실성과 부정확성을 처리할 수 있으므로, 기계학습과 퍼지 집합의 장점을 이용하여 전체 모형의 예측 정확도를 향상시킬 수 있다. Chen 등(2009)은 파산 예측을 위해 퍼지 논리의 기능성(functionality)과 신경망의 학습 능력(learning ability)을 결합한 신경 퍼지(neuro fuzzy 하이브리드 접근법을 제시하였는데, 신경 퍼지(neuro fuzzy)는 로짓 회귀(logit regression)보다 정확도와 탐지 능력이 높고 오분류 비용(misclassification cost)이 낮아 신경 퍼지(neuro fuzzy)가 임박한 파산에 대한 경고를 제공하는 데 유용함을 입증하였다. 특히 변수 간의 매핑 기능에 대한 포괄적인 설명을 통해 멤버십 함수 형태(membership function

28) 정규화 방법에 따라 서로 다른 결과를 가져올 수 있다. (Crone et al., 2006)

29) 결정 나무(Decision Tree)의 일종으로, ID3 알고리즘은 Information Gain의 크기가 가장 큰 것을 부모 노드로 하여 Tree를 구성한다. 1993년 C4.5를 거쳐 1998년 C5.0으로 완성된 알고리즘으로 명목형 목표변수만을 지원하는 단점이 있지만 분류의 정확도가 높은 장점이 있다.

30) self-organizing feature map은 신경망(neural network) 중 하나의 방법으로, randomized small valued initial weight vector를 사용하고, weight vector가 입력과 가장 잘 일치하는 neuron을 winning neuron으로 선택하며, activity bubble내의 neuron이 입력 벡터 쪽으로 이동되도록 weight vector를 훈련시킨다.

shape), 전달 함수(transfer function) 등에 대한 추가 정보를 제공하였다.

Shin과 Lee (2002)는 파산 예측의 양적 판별 규칙을 추출하기 위한 재무 비율의 컷오프를 찾기 위해 유전 알고리즘을 사용하여 좋은 성과를 냈다. Tsakonas 등(2006)도 예측 정확도를 향상시키기 위한 신경망 모형의 최적화를 위해 유전자 알고리즘을 사용하였다. Varetto (1998)는 유전 알고리즘에 의해 선형 함수와 판별 규칙을 추출하고 통계적 가정에 의해 제약받지 않는 최적의 선형 방정식을 얻었다. Back 등(1996)은 파산 예측 신경망 모형을 위한 예측자(predictor)를 선택하는데 이용될 수 있는 선형 판별 분석(linear discriminant analysis), 로짓 분석(logit analysis), 유전 알고리즘(genetic algorithm)을 비교 분석한 결과 유전 알고리즘을 사용할 때 최상의 예측결과가 얻어졌다. Etemadi 등(2009)은 파산 예측 모델링에 유전 프로그래밍(genetic programming, GP)을 적용하였다. McNemar 검정[31] 결과, 유전 프로그래밍 접근법이 파산 예측 문제에 대해 MDA를 능가하는 것으로 나타났다.

McKee (2000)는 파산 예측을 위해 러프 집합 이론 (rough set theory)[32]을 사용하였는데, 이 방법이 데이터 집합에 숨겨진 중요한 정보를 발견하고 이를 결정 규칙의 집합 (a set of decision rules)[33]으로 표현할 수 있음을 보였다. Park과 Han (2002)은 정량적 지표와 양적 지표의 장점을 통합하여 파산 예측 정확도를 향상시키기 위해 k-최근접 이웃 알고리즘[34]의 계량에 대한 분석적 계층 구조 프로세

31) McNemar test는 대응 명목 자료(paired nominal data)에 사용되는 통계 검정(test)이다.

32) 러프 집합은 Zdzisław I. Pawlak라는 폴란드의 컴퓨터 과학자가 처음으로 제창했다. 러프 집합은 전통적 집합이라고도 하는 crisp set의 정식 근사로 원래 집합의 하한, 상한 근사 두 가지 집합의 한 쌍이다. 러프 집합 이론은 하한과 상한 근사 집합이 crisp set이지만 다른 변수 하에서는 fuzzy set이 될 수도 있다.

33) Decisoin Rule이란 대안들에 대한 정보과 의사결정자의 선호도를 통합하여 총체적으로 평가하는 과정이라 할 수 있다. 각각의 규칙이 분류(classification)와 연관을 맺으며 결정 규칙의 집합(a set of decision ru;es)이 특정된다.

34) k-최근접 이웃 알고리즘(k-NN)은 분류나 회귀에 사용되는 가장 간단한 기계학습 알고리즘이다. 입력은 특징 공간 내 k개의 가장 가까운 훈련 데이터로 구성되며, 이웃은 항목(분류)이나 객체 특성값(회귀)이 알려진 객체의 집합으로부터 구해진다. k-NN은 함수가 지역적으로 근사하고 모든 계산

스(analytic hierarchy process)[35]를 사용했다. Mckee (2003)는 러프 집합 파산 예측 모형(rough sets bankruptcy prediction model)과 회계감사 신호 율(auditor signalling rate)을 비교 연구하였는데, 예측 정확도와 관련하여 러프 집합 모형과 실제 회계 감사인의 방법론 간에 큰 비교 우위를 발견하지 못하였다.

Lindsay와 Campbell (1996)은 파산기업의 pair-matched sample을 이용하여 파산 예측에 카오스 이론(chaos theory)을 적용하였다. 건강한 시스템이 그렇지 않은 시스템보다 더 많은 chaos를 나타내는 것을 감안할 때 파산에 근접한 회사의 수익률은 Lyapunov exponents로 측정된 chaos가 초기보다 유의적으로 적다는 가설을 세운 다음, 일변량 및 다변량 파산 예측 모형을 구해진 Lyapunov exponets를 이용하여 성공적으로 구성하였다. Fujiwara (2004)는 일본 파산 기업의 목록을 사용하여 부채가 많은 파산기업의 총 부채 분포에 대한 Zipf law를 발견했으며, 파산기업의 수명은 새로운 기업의 진입율과 상관관계가 있는 지수 분포를 보임을 밝혔다. Hong 등(2007)은 한국 파신 기업 수의 변동을 계측하여 파산기업 수의 분포에 대한 멱법칙(power law)을 관찰하였다. 파레토 지수(Pareto exponent)는 1에 가까웠다. 또한 파산기업 수의 일일 증분 확률 분포가 중앙 부분에서 가우시안 분포(Gaussian distribution)를 따르며, fat tail 부분이 멱 법칙(power law)을 따른다는 것을 관찰하였다.

결정 나무(decision tree)는 신경망(neural network)이나 SVM 같은 알고리즘에 비해 사람이 이해하기 쉽고 직관적인 장점이 있다. 결정 나무(decision tree) 알고리즘에서 얻은 규칙(rule)의 수는 서로 다른 최소 지원 수준(minimum support level)을 설정하여 어느 정도 제어할 수 있다. Aoki와 Hosonuma (2004)는 결정 나무 모형(decision tree model)을 사용하여 일본 기업의 파산을 예측하였다. 73개 파산 기업과 73개 비 파산기업을 사용하여 91.2 %의 예측 정확도를 얻었으며, 이자보상배

이 분류될 때까지 연기되는 인스턴스 기반 학습의 일종이라 할 수 있는데, 데이터의 지역 구조에 민감하다는 단점이 있다.

35) AHP (Analytic Hierarchy Process)는 Thomas L. Saaty에 의해 개발된 것으로 수학 및 심리학을 기반으로 복잡한 의사 결정을 구성하고 분석하는 구조화된 기법을 말한다.

율(interest coverage ratio, CF ratio)[36]이 파산기업 예측의 가장 중요한 척도라는 사실을 밝혔다. Olson 등(2012)은 파산 데이터에 다양한 데이터 마이닝 도구를 적용한 결과, 결정 나무(decision tree)가 상대적으로 더 정확하였다. 너무 많은 규칙 노드(rule node)의 문제는 최소 지원 조정(adjustment of minimum support)으로 보다 다루기 쉬운 규칙 집합(rule set)을 얻어 해결할 수 있음을 보였다.

Garcia-Almanza 등(2013)은 파산 예측을 위해 진화적 기법인 MP-EDR(Multi-Population Evolving Decision Rules)을 사용하였고, MP-EDR에 의해 생성된 규칙과 조건을 사용하여 구축된 네트워크를 통해 피쳐(feature)들의 관련성(relevance)을 표현하였다.

Support Vector Machine(SVM)은 구조적 위험 최소화(Structural Risk Minimization)에 기초한 강력한 분류 기법이다. Shin 등(2005)은 SVM(Support Vector Machine)을 사용하여 한국 기업의 파산을 예측하였는데, 그 결과 다중판별분석(Multiple Discriminant Analysis, MDA)[37], 로짓(logit)[38] 및 신경망(NN) 보다 좋은 예측 능력을 보여주었다.[39] Min과 Lee (2005)도 역시 SVM의 커널 함수의 최적 파라미터를 구하기 위해 5-fold cross-validation을 이용한 grid-search technique를 사용하여, SVM이 MDA, Logit, BPNs 보다 성능이 좋음을 입증하였다. Min 등(2006)은 유전자 알고리즘(GA)을 이용하여 SVM의 feature subset과 파라미터를 최적화함으로써 SVM의 예측 성능을 향상시키는 하이브리드 파산예측모형을 제안하였다. Chaudhuri와 De (2011)은 SVM과 퍼지 논리를 결합한 Fuzzy Support Vector Machine (FSVM)을 사용하여 파산예측문제를 해결하고자 하였다. 미국 기업 데이터를 대상으로 한 실험 결과, FSVM은 optimal feature set과 파라미터를 찾

36) 기업의 영업이익을 지급이자로 나눈 값으로 기업의 채무상환능력을 나타내는 지표이다.

37) 분류 집단 간의 차이를 설명해 줄 수 있는 독립변수들의 선형 결합으로 판별식을 만들어 분류 대상들이 속하는 집단을 찾아내는 방법을 말한다.

38) 로짓 분석은 선택모델의 하나로 누적적 로지스틱 확률함수(cumulative logistic probability fuction)을 이용한다.

39) Min과 Lee(2005)의 결과도 유사하다.

는데 효과적이었으며 우수한 클러스터링 성능(clustering power)과 상당한 예측능력을 보여주었다.

Tseng과 Hu (2010)는 영국 기업들을 대상으로 로짓(logit), 이차 간격 로짓 모형(quadratic interval logit model)[40], 역전파 다층 퍼셉트론(backpropagation multi-layer perceptron, MLP)[41], 방사 기저함수 네트워크(radial basis function network, RBFN)[42]의 성능을 비교하였는데, RBFN이 다른 모형에 비해 우수한 성능을 보이는 것을 발견했다. 그밖에 Li와 Sun (2012)은 nearest-neighbour support vector model(N-N SVM)[43]을 이용하였고, Tserng 등(2011)은 enforced support vector machine model(ESVM)을 제창하였으며, Horta 등(2012)은 자료포락분석법(data envelopment analysis)[44]에 기초한 방법을 제안하였다. Cielen 등(2004)은 선형 계획 모형(linear programming model)[45], 데이터 포락 모형(data envelopment model, DEA), 규칙 귀납 모형 (rule induction model, C5.0)[46] 의 분류 성능을 비교하여, 정확성과 사용성면에서 DEA 모형이 다른 모형을 능가함을 밝혔다. Sun과

---

40) 이것은 logit (또는 logistic regression)과 Tanaka의 quadratic interval regression model의 장점을 결합한 모형이다.

41) 역전파(backpropagation) 방법으로 다층 퍼셉트론을 효율적으로 학습시키는 알고리즘

42) 방사형 기저 함수 네트워크(Radial Basis Function Network)는 방사형 기저 함수를 활성화 함수(activation function)로 사용하는 인공 신경망(Artificial Neural Netowork)이다. 네트워크의 출력은 입력과 뉴런 파라미터의 방사형 기저 함수 선형 조합(linear combination)이다.

43) 가장 가까운 이웃 (nearest neighbour)까지 임의의 비율 거리에 따라 새로운 소수 표본(minority sample)을 생성하고 서포트 벡터 머신 (SVM)을 어셈블 한다.

44) 1978년 차니스, 쿠퍼, 로즈 등이 개발. 의사결정단위(Decision Making Unit, DMU) 사이의 상대적 효율성을 비교하기 위해 선형계획법을 사용하며 이후 다양한 기법과 연계되어 발전되어 왔다.

45) 선형계획법은 제약 조건이 연립일차부등식 또는 연립일차방정식이고, 목적함수(objective function)도 일차식인 경우 이 일차식의 최댓값 또는 최솟값을 구하는 방법에 관한 이론으로, 최적화 이론의 한 분야이다.

46) 규칙 귀납(rule induction)은 일련의 관측으로 부터 공식 규칙을 추출하는 기계 학습의 한 영역인데, 추출된 규칙은 데이터의 완전한 과학적 모형을 나타낼 수도 있고 단순히 데이터의 로컬 패턴을 나타낼 수도 있다.

Shenoy (2009)는 파산예측을 위한 naïve Bayes Bayesian network (BN) 모형을 구축하고 가장 우수한 성능을 얻기 위한 운영 지침을 제공하였다. Tsakonas 등(2006)은 파산 예측을 위하여 neural logic networks와 grammar-guided genetic programming을 결합한 hybrid intelligent system을 제안하였다. 이는 전문가 규칙 집합(set of expert rules)을 통해 망구조(network structure)의 해석을 용이하게 하는 접근 방식인데 예측 정확도와 해석 가능성에서 매우 우수한 성능을 보여 주었다.

최근 연구에 따르면, 여러 분류자(multiple classifiers)를 결합하면 성능이 향상될 수 있다고 한다. 앙상블(ensemble)은 분류 및 예측 모델의 성능을 향상시키는데 널리 사용되는 방법 중 하나이다. 하지만, 앙상블(ensemble)은 대개 상이한 분류자(classifier)로부터 장점을 상속받을 뿐만 아니라 이들 (분류자classifier)의 단점을 겪을 수 있다. Hung과 Chen(2009)은 결정 나무(decision tree), 역 전파 신경망(back propagation neural network) 및 support vector machine 등 세 가지 분류자(classifier)의 선택적 앙상블(selective ensemble)을 제안한다. 파산과 비 파산의 예상 확률을 기반으로, 이 앙상블은 상이한 분류 기법의 장점을 상속하고 단점을 피하는 접근법을 제공한다. 이러한 선택적 앙상블은 파산 예측을 위한 다른 가중치(weighting) 또는 투표(voting) 앙상블보다 더 나은 수행 결과를 보여주었다.

Bagging과 Boosting이라는 널리 알려진 앙상블 방법이 주로 결정 나무(decision tree)를 기본 분류자(base classifier)로 사용하여 다양한 기계 학습 문제에 큰 성공을 거두었다. Kim과 Kang (2010)은 파산 예측 신경망의 성능을 향상시키기 위해 신경망 앙상블(ensemble with neural network)을 제안하였는데, bagged neural network과 boosted neural network은 모두 전통적인 신경망보다 향상된 성능을 보여주었다. Alfaro 등(2008)은 BP-NN(Back Propagation-Neural Network)과 AdaBoost[47] 앙상블 학습(ensemble learning)[48]을 이용하여 예측모형을 구축하였고,

47) Schapire (1990)가 처음으로 부스팅 아이디어를 제창한 이래 Drucker 등(1993)이 광특성 인식문제를 풀기 위해 부스팅을 이용하였으며 Freund(1995)가 보다 효율적인 부스팅 알고리즘을 제안하였다. Schapire (1995)는 Boosting 알고리즘의 개선된 버전을 가져 와서 AdaBoost라고 명명했다. AdaBoost는 훈련 표본의 분포를 자체적으로 적절하게 변경하고 이전 분류자에 의해 잘못 분류된 표본에 후속 분류자가 더 많은 주의를 기울일 수 있는 반복 프로세스이다.

1180여 개의 기업으로 구성된 데이터 세트를 기반으로 분석을 한 결과 AdaBoost 앙상블 학습이 1 년 전 예측에서 매우 좋은 성과를 나타냄을 보여주었다. Nanni & Lumini (2009)는 Bagging(Bootstrap Aggregating)[49], Random Subspace(Attribute Bagging or Feature Bagging)[50], 클래스 전환(Class Switching)[51], Rotation Forest[52] 포함 여러 앙상블의 성능을 비교하였는데, Random Subspace가 다른 앙상블 방법보다 성능이 우수함을 발견했다.[53] Finlay (2011)는 소비자 신용 위험 분류를 위한 여러 다중 분류 시스템의 성능을 평가하면서 Bagging과 Boosting이 다른 다중 분류 시스템보다 뛰어남을 발견했다. Sun 등(2012)은 BPNN[54] – AdaBoost 모형과 BPNN– Bagging 모형을 제안하여 BPNN – AdaBoost 모형은 단기, 중기 예측에, BPNN – Bagging 모형은 장기 예측에 적합함을 보여 주었다. Zięba 등(2016)은 파산 예측을 위한 새로운 접근법으로 결정 나무(decision trees)의 앙상블을 학습하기 위해 Extreme Gradient Boosting을 이용하는 방법을 제안하였다. 아울러, 데이터의 고차원 통계(higher-order statistics)를 반영하고 데이터 표현에 대한 사전 지

48) 기계 학습에서 앙상블 학습은 단일 학습 알고리즘 보다 예측 성능을 향상시키기 위하여 다수의 학습 알고리즘을 사용하는 방법을 말한다.

49) 기계 학습에 Bootstrap을 사용하여 학습의 정확도와 안정성을 높이는 방법. Breiman (1996)은 기본 분류기가 Bagging 앙상블 시스템의 성능을 확실하게 하도록 다양하고 정확해야 하고, 불안정한 학습 알고리즘 (예 : NN 및 decision tree)이 기본 학습자(base learner)로 사용하기에 더 적합하다는 것을 지적했다.

50) 전체 feature set 대신 무작위 표본으로 훈련하여 앙상블 추정자(estimator) 간의 상관관계를 줄이려는 앙상블 학습 방법.

51) 클래스 전환 앙상블(class-switching ensemble)에서 각 학습자(learner)는 수정된 버전의 훈련 데이터를 사용하여 구성된다.

52) PCA(Principal Component Analysis, 주성분 분석)를 이용한 앙상블 기반 알고리즘. 기본 분류자에 대한 훈련 데이터를 생성하기 위해 feature set는 K subsets(K는 알고리즘의 파라미터)로 무작위 분할되고 주성분 분석(PCA)이 각 subsets에 적용된다.

53) 다른 앙상블 기법으로 Bagging(예제 선택의 임의성)과 Random space method(변수 선택의 임의성)를 사용하여 Decision Tree의 결점을 극복한 Random Forest가 있다. 이는 임의성에 의해 조금씩 다른 특성을 갖는 tree들로 구성되기 때문에 각 tree의 예측이 비상관화되어 일반화 성능을 향상시킨다.

54) back-propagation neural network

식을 부여하기 위해 synthetic features라고 하는 새로운 개념을 도입하였다.

2.2 주가 예측 관련 연구

주식 시장의 움직임을 예측하는 것은 많은 연구자들의 흥미를 끌어왔다. 하지만, 주식 시장 예측은 불확실성이 항상 시장의 움직임에 관련되어 있기 때문에 매우 어려운 문제이다. 수많은 과학적 시도가 있었지만 주가 움직임을 정확하게 예측하는 유일한 방법은 아직 발견되지 않았다. 예측의 어려움은 매개 변수가 끊임없이 움직이고 완전히 정의되지 않은 시장 역학과 관련된 복잡성에서 비롯된다. 또한 과거 시장 데이터 및 복잡한 수학적 모델을 사용하는 방법은 기존 정보의 범위 내에서 평가를 수행하도록 제한되기 때문에, 과거의 규칙을 벗어나는 예기치 않은 사건에 대해 반응 할 수 없는 단점이 있다(Schumaker & Chen, 2010). 일부 연구자는 이른바 효율적 시장가설(Efficient Market Hypothesis, EMH)에 따라 주식 시장 매개 변수의 시간 상관관계가 경제적으로나 통계적으로 유의하지 않다고 주장한다(Fama, 1965). 즉, 기본적으로 주식 시장 가격은 거래자가 이용할 수 있는 모든 정보를 즉각적으로 반영하여 결정되므로 주식 시장 예측이 불가능하다는 주장이다(Hawawini & Keim, 1995). 하지만 일군의 연구자들이 기술적 분석 전략으로 초과수익을 얻을 수 있다는 연구 결과를 내놓기 시작한 이래, 많은 연구자들이 효율적 시장가설을 반박하는 연구 결과를 내어 놓으면서 주가 추세가 통계적으로 예측될 수 있다는 주장이 힘을 얻고 있다(e.g. Brook, 1992). 물론 여전히 주식시장은 비교적 효율적이며 예측가능성은 떨어진다는 주장을 고수하는 목소리도 만만치 않다(Malkiel, 2003).

문헌에서 주가 예측을 위한 두 가지 전통적인 분석 방법이 있다. 하나는 회사의 재무 제표에 있는 정보를 사용하는 기본적 분석이며, 또 하나는 주식 시장의 추세를 연구하면 주가 변동의 규칙을 알 수 있을 것으로

믿는 기술적 분석인데, 둘 다 오랜 동안 주식 시장을 분석하는데 사용되어 왔다(Bettman 2009). 그러나 투자자의 불합리한 행동, 선택된 데이터 등 여러 요소가 기본적 분석에 편향을 일으킬 수 있고, 기술적 분석은 주식 시장의 과거 추세를 기반으로 주식 가격을 예측하기 때문에 주식 시장이 정기적인 규칙을 갖는다는 것을 보여주는 증거는 없다는 주장도 있다.

주가 예측과 관련하여 통계에 기반을 둔 연구들의 결과와 함께 컴퓨터 기술의 눈부신 발전에 힘입어 기계학습을 기반으로 한 주가 예측에 대한 연구가 많이 이루어지고 있다. ARCH-GARCH 모형, ANN 및 진화적 계산 방법(evolutionary computation methods)을 포함하여 전통적인 시계열 접근에서 인공 지능 기법에 이르기까지 금융 시장을 예측하기 위한 다양한 시도가 이루어져 왔다.

기계학습 기법은 다차원 공간에 복잡한 비선형 분류를 할 수 있다는 특징이 있어 다양한 실험적 시도가 가능하다는 장점이 있다. 대표적으로 인공신경망, 유전알고리즘, 퍼지이론(fuzzy theory)[55], SVM, 결정트리(decision tree)[56], 적응형 부스팅 등 다양한 기계학습 방법이 알려져 있다(Hadavandi, Shavandi & Ghanbari 2010; Kim & Cho, 2010; Pai & Lin, 2005; Wu, Lin & Lin 2006; Han & Chen, 2007;

55) 1965년 Loft A. Zadeh에 의해 도입된 퍼지집합의 사고방식을 기초로 하여 애매하고 불확실한 상황을 정량적으로 표현하여 수학적으로 접근하려는 이론을 퍼지 이론(fuzzy theory)이라고 한다. 퍼지집합(fuzzy set)이란 퍼지 논리 개념을 사용해 기존의 집합을 확장한 것으로, 각 원소가 그 집합에 속하는 정도(소속도)를 소속 함수로 나타냄으로써 수학적으로 표현하는데, 소속도는 0과 1 사이의 실수 값을 갖는다. 가전제품, 자동제어 분야에서 퍼지이론을 적용한 제품들이 있다.

56) 결정 트리 (decision tree, DT)는 많은 예측 문제에 대해 잘 알려진 데이터 마이닝 기술 중 하나인데, 합리적인 분류 및 예측 성능을 보인다. DT는 다양한 상황에 따라 여러 노드(node)와 분기(branch)를 만드는 트리 구조(tree structure)를 가지고 있다. 모든 노드(node)는 출력 / 대상 클래스(output/target class)를 나타내고 모든 분기(branch)는 분류(classification)를 위한 프로세스 / 결정(process/decision)을 보이며, 끝 노드(end node)는 예측 결과를 제공한다. 결정 트리 모형이 설정된 후 결정 트리를 pruning(전정, 가지치기)하기 위해 오류율을 계산할 수 있다. pruning(가지치기)은 결정 트리의 예측 능력과 분류 능력을 향상시키고 결정(decision)을 보다 효율적으로 수행하기 위한 것이다. ANN과는 달리 DT는 추가 분석을 위한 여러 결정 규칙(decision rule)을 생성한다.

Kazem, Ahmad, et al., 2013; Wen, Qinghua, et al., 2010; Zhiqiang, Huaiqing & Quan, 2013).

주가 예측 관련 연구는 예측 모형(forecasting model), 사용된 지시자(fundamental indexes, technique indexes, macroeconomic indexes, etc.), 발견된 결과(findings)등으로 요약될 수 있다. Phua 등(2003)은 예측 모형으로는 ANN, 지시자(indicator)로는 기술적 지표(technique indexes)를 사용하여 ANN이 평균 성공률 60%이상이며 최상의 예측결과는 74%에 이른다고 하였다. Chen et. al(2003)은 확률신경망(probabilistic neural network, PNN)[57] 과 GMM(generalized method of moments)[58] with Kalman filter[59]를 사용하여 PNN이 buy-and-hold 전략이나 parametric GMM 모형 예측에 따른 전략보다 더 높은 수익을 얻는다고 밝혔다. Kunhuang과 Tiffany(2006)은 Back-propagation ANN과 첸의 시계열모형(Chen's time series model)을 비교하여 ANN이 시계열 모형보다 더 좋은 예측력을 갖는다고 하였다. 한편 O'Connor와 Madden(2006)은 ANN을 기본적 지표(fundamental indexes)에 적용하여 ANN이 주식 시장에서 시장대비 초과수익률을 얻기 때문에 예측력이 있다고 주장하였다. Wang (2004)도 역전파(Back Propagation, BP) 알고리즘을 사용하여 주식 시장을 예측하는 시스템을 제시하였는데, 역전파 신경망은 주가와 금융 시장

57) 확률신경망 (probabilistic neural network, PNN)은 베이지안 네트워크(Bayesian network)와 커널 피셔 판별 분석 (Kernel Fisher discriminant analysis)이라 불리는 통계 알고리즘으로 부터 파생된 피드포워드 신경망(feedforward neural network이다. D.F. Specht가 1990 년대 초반 도입하였으며, 작동(operation)은 4 개의 층(layer)을 가진 다층 피드포워드 네트워크로부터 조직된다.

58) GMM은 계량경제학에서 통계모형의 파라미터를 추정하기 위한 일반적인 방법이다. 일반적으로 파라미터가 유한 차원인 semiparametric model의 컨텍스트에서 적용되는데, 데이터 분포함수의 전체 모양을 알 수 없으므로 최대 우두추정을 적용할 수 없다. L. P. Hansend이 1982년 도입하였으며, Peason의 method of moments를 일반화한 것이다. Hansen은 이러한 공로로 2013년 노벨 경제학상을 수상하였다.

59) 칼만 필터(Kalman filter)는 R. Kalman에 의해 개발된 것으로 잡음이 포함된 입력 데이터를 재귀적으로 처리하여 현재 상태에 대한 최적의 통계적 예측을 진행하는 재귀 필터이다. 알고리즘 전체는 예측과 업데이트로 나눌 수 있다.

뒤에 숨겨진 영향 요인 간의 상관관계를 찾는 데 도움을 준다. 지도 알고리즘(supervised algorithm)은 Feed-forward, Cascade-forward 및 Elman BP 인데, 기울기 하강(Gradient Descent)[60], Gradient Descent With Momentum, Gradient Descent With Adaptive Learning Rate, Gradient Descent With Momentum & Adaptive Learning Rate, Levenberg-Marquardt[61], Broyden-Fletcher-Goldfarb-Shanno (BFGS)[62], Resilient Propagation (RPROP)[63]와 같은 7 가지 BP 기법으로 각각 훈련되었다. Lam (2003)도 기본적 분석과 기술적 분석을 통합하여 BP-NN(Back Propagation Neural Networks) 모형의 예측 능력을 검증하였다. Majhi (2009)는 장단기 주가예측을 위해 trigonometric functional link artificial neural network(FLANN)모형을 개발하였는데, 특히 RLS(recursive least square)기반 FLANN모형이 온라인 예측(online prediction)에 더욱 적합함을 밝혔다. Lin (2009)은 주식시장에서 다음날 종가예측을 위해 ESN (Echo State Network)[64]을 이용하였다. Hurst

60) 기울기 하강(Gradient descent)은 1 차 반복 최적화 알고리즘(first-order iterative optimization algorithm)이다. 기울기 하강(Gradient descent)을 사용하여 함수의 로컬 최솟값을 찾으려면 현재 점에서 함수의 Gradient의 음수에 비례하는 단계를 취한다. Gradient의 양수에 비례하는 단계를 취하면 그 함수의 로컬 최댓값에 접근하며 이는 기울기 상승(Gradient ascent)이다.

61) Levenberg-Marquardt (LMA) 알고리즘은 비선형 최소 제곱 문제(non-linear least squares problem)를 푸는 데 사용한다. 이러한 최소화 문제는 특히 최소 제곱 커브 피팅(least squares curve fitting)에서 발생한다. LMA는 global minimum이 아닌 local minimum을 찾으며, Gauss-Newton 알고리즘 (GNA)과 gradient descent 사이에 위치한다. 또한 LMA는 DLS (damped least-squares)로도 알려져 있으며, 신뢰영역 접근법을 이용한 Gauss-Newton으로 볼 수도 있다

62) Gradient decent보다 빠르며 learning rate를 직접 선택하지 않아도 되는 효율적인 알고리즘이다. Quasi-Newton method의 일종으로 advanced optimization algorithm이라 할 수 있다.

63) Rprop (resilient backpropagation)은 1992년 Martin Riedmiller와 Heinrich Braun이 만든 1차 최적화 알고리즘(first-order optimization algorithm)이며, feedforward신경망에서 지도 학습(supervised learning)을 위한 learning heuristic이다.

64) ESN (Echo State Network)는 띄엄띄엄 연결된 숨겨진 레이어(sparsely connected hidden layer, 일반적으로 1% 연결성) 가 있는 recurrent neural network 이다. 숨겨진 뉴런의 연결성과 가중치는 고정되고 무작위로 할당된다. 출력 뉴런의 가중치는 망(network)이 specific temporal patterns를 생성하도록 학습될 수 있다.

exponent는 초기 과도현상을 순응적으로 결정하고 훈련 중 예측 가능성이 가장 큰 sub-series를 선택하는데 적용된다. 실험 결과, 대부분의 경우 ESN (Echo State Network)이 기존의 다른 신경망보다 우위에 있음을 보여 주었다. 실험은 또한 데이터 전처리에서 노이즈를 필터링하고 적절한 파라미터를 선택하기 위해 주성분 분석(principle component analysis, PCA)을 포함하면 조악한 예측성과를 방지하고 예측 정확도를 약간 향상시킬 수 있음을 나타내었다.

Kim & Lee(2004)는 주식시장 예측을 위한 ANN 모형의 feature transformation method로 유전자 알고리즘(genetic algorithm, GA)을 이용한다. GA는 주식시장 예측을 위한 ANN의 학습(learning) 및 일반화가능성(generalizability)을 향상시켜주는 역할을 한다. 이러한 접근법은 특징공간의 차원수(dimensionality of the feature space)를 축소하고 주식시장 예측과 무관한 요소를 감소시킨다. Armano 등(2002)도 ANN과 GA를 혼합한 모형을 사용하여 buy-and-hold 전략과 RNN(Recurrent Neural Networks)[65]모형을 초과하는 성과를 보여준다. Arau´jo(2010)는 주식시장 예측을 위해 QIEHI 모형(quantum-inspired evolutionary hybrid intelligent model)을 사용한다. 이것은 ANN with MQIEA(modified quantum-inspired evolutionary algorithm)[66], ANN training algorithm[67], suitable time lags[68]로 구성된다. Asadi (2012) 도 주식시장 예측을 위해 hybrid intelligent model을 제안하는데, 이것은 데이터 전처리 방법, 유전자 알고리즘, FNN(Feed-forward Neural Network)[69]학습을 위한

65) RNN은 단위 간의 연결이 지시된 순환(directed cycle)을 형성하는 인공 신경망의 부류이다. 이것은 동적 시간 동작(dynamic temporal behavior)을 나타낼 수 있도록 네트워크의 내부 상태를 생성한다. FNN(Feedforward Neural Networks)과 달리 RNN은 내부 메모리를 사용하여 임의의 입력 시퀀스를 처리 할 수 있다.

66) 이것은 완전한 ANN 아키텍처 및 파라미터를 진화시킬 수 있다. (pruning process)

67) MQIEA에 의해 제공된 ANN 파라미터를 더욱 향상시키기 위해 사용된다.

68) 시계열 현상을 더 잘 설명하기 위한 것이다.

69) FNN (Feed-forward Neural Network)은 유닛 간의 연결이 사이클을 형성하지 않는 신경망이다.

LM(Levenberg-Marquardt) 알고리즘[70]의 조합이다. Chang 등(2004)은 Back-propagation ANN과 ARIMA를 기본적 지표(fundamental indexes)와 기술적 지표(technique indexes)에 적용하여 하이브리드 모형이 단일 모형보다 예측력이 더 뛰어남을 보였으며, Wang(2007)역시 하이브리드 ANN을 사용하여 다른 기법과 결합된 ANN이 주식시장에서 좋은 예측력을 갖는다고 밝혔고, Roh(2006)도 ANN, NN-EWMA[71], NN-GARCH[72], NN-EGARCH[73] 등을 거시경제 지표(macroeconomic indexes)에 적용하여 역시 하이브리드 ANN이 단일모형보다 더 나은 예측력을 갖는다고 하였다. Tsai와 Wang(2009)는 ANN과 DT(Decision Tree)모형을 사용하여 DT + ANN 모형이 ANN, DT 단일 모형 보다 훨씬 높은 77%의 정확도를 보인다고 하였다. Kim과 Shin(2007)도 ANN, ATNN (Adaptive Time delay Neural Network), TDNN (Time Delay Neural Network), GA를 사용하여 역시 같은 결론에 도달하였다. 이들 연구의 공통점은 기계학습 모형이 일반 통계모형에 비해 우수한 예측력을 보인다는 점과 단일 모형보다 여러 모형을 결합한 하이브리드 모형이 더 나은 예측력을 나타낸다는 점이다.

주식시장 예측모형의 흥미로운 특성은 이 문제의 시간 의존성(time dependence), 변동성(volatility), 기타 복합적인 의존성(complex dependence)이다. 이러한 점을 포괄하여 주식시장 예측에 적용한 것이 은

---

70) Levenberg-Marquardt Algorithm (LMA)은 DLS (damped least-squares) method로도 알려져 있으며 비선형 최소 제곱 문제를 해결하는데 사용된다.

71) 여기서, NN은 Neural Network을 EWMA는 Exponentially Weighted Moving Average(지수가중 이동평균)을 뜻한다. MA(이동 평균)은 전체 데이터 집합의 하위 집합 평균을 만들어서 데이터 요소를 분석하는 방법이다.

72) GARCH (Generalized Autoregressive Conditional Heteroscedasticity) 모형은 오차분산에 대해 자기회귀 이동평균(ARMA) 모형을 가정한 것으로 시계열 데이터 분석에 널리 사용된다.

73) EGARCH (Exponential Generalized Autoregressive Conditional Heteroskedastic (EGARCH) 모형은 Nelson이 1991년 개발한 것으로 GARCH 모형의 한 형태이다.

닉 마르코프 모형(Hidden Markov Model, HMM)이다. HMM은 동적 시스템 모델링(dynamic system modelling)에 대한 입증된 적합성 때문에 패턴 인식 및 분류 문제에 광범위하게 사용되어 왔다. Gupta (2012)는 주가 예측을 위해 Maximum a Posteriori[74] HMM 접근법을 사용하여 ANN보다 좋은 성과를 보여 주었다. Hassan (2005)도 상호 연관된 시장에 대한 주가 예측을 위해 은닉 마르코프 모형(Hidden Markov Model, HMM)을 제시한다. HMM을 사용하여 얻은 결과는 고무적이며 주식시장 예측을 위한 새로운 패러다임을 제공한다. HMM과 시계열(time series)을 적합 시키기 위해서는 다른 파라미터를 학습하기 전에 숨겨진 상태(hidden state)의 수를 결정해야한다. 이것은 HMM의 복잡성과 정밀도에 큰 영향을 주기 때문이다. 하지만, 관찰된 계열(observed series)에 대한 사전 지식이 충분치 않으면 이것은 너무 어려워지고 예측과정에서 평균오차가 증가하게 된다. 이런 단점을 극복하기 위해 Duan (2007)은 adaptive model selection 기반의 prediction algorithm (PAAMS)을 제안한다. PAAMS에서 예측 평균오차가 증가하면 모형을 동적으로 업데이트할 수 있는데, 업데이트 프로세스 중에 최선의 hidden state number를 얻기 위해 AMSA(automatic model selection method)가 적용된다. Wu(2009)는 비균질 은닉 마르코프 모형(Non-Homogeneous Hidden Markov Model, NHMM)을 사용하여 주식 시장 프로세스를 모델링하는 접근법을 제안하는데, 주식 가격과 뉴스 기사를 둘 다 고려하는 이벤트 중심 접근법(event driven approach)이라는 특징을 갖는다. NHMM을 구축할 때 주가 변동에 중요한 영향을 주는 일련의 버스티[75] 피쳐(bursty features, keywords)를 이용하여 특정 주식에 대한 관련 이벤트를 식별한다. 이러한 접근법은 매우 실질적이고 효과적인 성과를 보여주었다. Hassan 등(2007)은 ANN,

74) 베이지안 통계(Bayesian statistics)에서 Maximum a Posteriori probability (MAP) 추정치는 posterior distribution의 모드와 동일한 알려지지 않은 수량(unknown quantity)의 추정치이다. MAP는 경험적 데이터에 기초하여 관측되지 않은 수량(unobserved quantity)의 점 추정치(point estimate)를 얻는 데 사용될 수 있다.

75) 데이터가 갑자기 집중적으로 한 번씩 소규모로 발송되는 것을 가리킨다.

은닉 마르코프 모형(Hidden Markov Model, HMM)[76], 유전 알고리즘(Genetic Algorithm, GA)을 사용한 결합 모형을 제안하였다. ANN을 사용하여, 일일 주가를 HMM에 입력되는 값의 독립적인 집합으로 변환하고, HMM의 초기 매개변수(initial parameter)를 최적화하기 위해 GA를 사용한다. 훈련된 HMM은 과거 데이터에서 유사한 패턴을 식별하고 찾아내는데 사용된다.

퍼지 논리(fuzzy logic)와 규칙 귀납(rule induction)의 영역을 합치는 것은 높은 이해력을 가진 일반화 기술의 출현을 위한 길을 열었다. Romahi와 Shen(2000)은 퍼지 연관 유도 알고리즘(fuzzy association induction algorithm)으로부터 파생된 주가 예측을 위한 새로운 기술을 제시하여, 변화하는 시장 동학(market dynamics)이 실시간으로 고려되면서 진화하는 규칙기반 전문가 시스템(evolving rule based expert system)[77]을 개발할 수 있도록 했다. Cheng 등(2009)은 주가예측을 위해 다중 주식 변동성 인과관계(multi-stock volatility causality)에 기초한 새로운 모형, 융합 적응 네트워크 기반 퍼지 추론 시스템(fusion adaptive-network-based fuzzy inference system, fusion ANFIS)을 제안하였다. Atsalakis와 Valavanis (2009)도 neuro-fuzzy system을 이용하여 여러 주식 시장의 단기 추세를 예측하였는데, 다른 방법 대비 탁월한 성능을 보였다. Bagheri 등(2014)은 ANFIS membership functions를 튜닝하기 위하여 QPSO (Quantum-behaved Particle Swarm Optimization)을 사용한 모형을 제안하였다. Yu(2005)는 대만 주식시장 예측을 위해 가중

76) HMM은 시스템이 은닉된 상태와 관찰 가능한 결과의 두 요소로 이루어졌다고 보는 모형으로 통계적 마르코프 모형의 일종이다. 즉, 관측 결과의 직접적 원인인 관측될 수 없는 은닉 상태들이 마르코프 과정을 통해 도출되고 이 결과들만이 관찰될 수 있다고 보는 은닉 마르코프 모형은 동적 베이지안 네트워크(Dynamic Bayesian network)로 간단히 나타낼 수 있다.

77) Rule-Based System은 미리 정의된 약속들을 기반으로 다양한 입력 자료들을 분석하여 결과를 도출하는 System을 의미하는데, 특정 영역이나 자료 형식에 국한되지 않고 rule 영역만을 전문적으로 개발하고 적용하는 System이라 할 수 있으며, rule engine 이라고도 한다. 활용 예로 전문가 시스템(expert system)을 들 수 있는데, 결과를 내기까지 적용되는 rule이 복잡하거나 고정적이지 않고 자꾸 바뀌는 경우에도 일반 시스템보다 빠른 성능 향상을 기대할 수 있다.

퍼지 시계열 모형(weighted fuzzy time series model)을 제안하는데, 이것은 퍼지 시계열 예측에서의 두 가지 이슈, 즉 recurrence와 weighting을 처리하기 위한 모형이다. Hadavandi 등(2010)은 주가 예측 시스템 구축을 위해 genetic fuzzy system (GFS) 와 artificial neural networks (ANN)의 통합 접근법을 제시하였다. 우선, 주가에 가장 큰 영향을 미치는 요인을 결정하기 위해 단계별 회귀 분석 (stepwise regression analysis, SRA)을 사용하고, 다음 단계에서 자기 조직화 지도(self-organizing map, SOM)[78]을 통해 원시 데이터(raw data)를 k개의 클러스터로 나눈다. 마지막으로, 모든 클러스터는 rule base extraction 및 data base tuning 능력을 가진 독립적인 GFS(Genetic Fuzzy System)모형으로 이송(feed)된다. 이러한 접근법은 주가예측 문제에 적합한 도구로 간주될 수 있을 정도로 이전의 다른 방법들보다 우수한 성과를 보여 주었다. Afolabi 등(2007)은 일일 주가 예측의 정확도를 높이기 위해 backpropagation, Kohonen 자기 조직화 지도(self organizing map, SOM)[79], hybrid Kohonen SOM을 이용하였는데, 다른 기법과 비교하여 hybrid Kohonen SOM이 더 나은 예측 성능을 보여주었다. Kuo 등(1996)은 정량적 요인(quantitative factor)과 정성적 요인을 (qualitative factor) 동시에 고려한 지능형 주식시장 예측 시스템 (intelligent stock market forecasting system)을 제안하였는데, 요인 수집, 정량적 모형 (인공신경망, artificial neural network), 정성적 모형(퍼지 델파이, fuzzy Delphi) 및 decision integration(인공신경망)으로 구성되었다. 대만 주식시장 데이터로 테스트한 결과, 우수한 성과를 보여주었다.

78) 자기 조직화 지도(self-organizing map, SOM, or self-organizing feature map, SOFM)은 인공 신경망(ANN)에 기초한 자율 학습(unsupervised learning)의 한 방법으로, 지도(map)이라고 불리는 훈련 샘플의 입력 공간에 대한 저차원 (일반적으로 2차원)의 이산화된 표현(discretized representation)을 생성하기 위해 자율 학습(unsupervised learning)을 사용하여 훈련된다.

79) SOM (self-organizing map)은 인공 신경망 (ANN)의 한 유형으로, 비지도 학습(unsupervised learning)을 사용하여 지도(map)라 불리는 훈련 샘플 입력 공간의 저 차원 (일반적으로 2 차원) 이산 표현(discretized representation)을 생성한다. SOM은 다른 신경망과 달리 error-correction learning이 아닌 경쟁학습(competitive learning)을 한다.

Huang과 Jane (2009)는 grey system theory[80], rough set(RS) theory 및 the moving average autoregressive exogenous (ARX) 예측 모형을 결합하여 자동 주식 시장 예측 메커니즘(automatic stock market forecasting mechanism)을 만들었다. 제안된 접근법에서 매 분기마다 자동으로 수집된 데이터는 ARX 예측 모형에 입력되어 다음 분기 또는 반기 후의 추세를 예측한다. 예측 데이터는 K-means clustering algorithm과 GM(1, N)[81]및 일련의 의사 결정 규칙을 적용하여 RS(Rough Set) classification module에 제공된다. 이러한 하이브리드 방식은 GM (1,1)[82] 방식보다 예측 정확도가 우수하고 더 높은 수익률을 산출하는 것으로 나타났다. Wang (2002)은 주어진 시간에 주가를 즉시 예측하기 위하여 fuzzy grey prediction system을 개발하였다. 주가를 예측할 때 두 개의 연속적인 데이터 세트에 크고 작은 차이가 있거나 데이터의 양이 너무 커서 사용하는데 영향을 주는 문제를 해결하기 위해 주식 데이터의 크기를 줄일 수 있는 data mart를 구성했으며, grey theory와 fuzzification technique를 결합하여 시스템에서 가능한 답을 즉시 예측하는 예측 기능의 하나로서 fuzzy grey prediction을 만들었다. 이러한 예측 시스템을 사용하여 주식 데이터를 분석하고 특정 시점에 신속하게 주가를 예측한 결과, 주식 거래자들이 데이 트레이딩(day trading)[83]을 효과적으로 처리하도록 도움을 줄 수 있음을 입증하였다. Kayacan 등(2010)은 GM(1,1), Grey Verhulst 모형, Furier Series를 사용하는 modified grey model 등 등 다양한 grey model의 정확도를 조사했다. 시뮬레이션 결과, modified grey model이 model fitting뿐만 아니라 예측에서 더 높은 성능을 가짐을 보여

80) grey system theory는 1982년에 시작되는데, 정보와 관련하여 구조 메시지(structure message), 작동 메커니즘(operation mechanism)및 행동 문서(behaviour document)와 같은 정보가 부족한 시스템을 grey system이라고 한다. grey model은 알려지지 않은 시스템의 동작을 추정하기 위해 제한된 데이터양만을 요구한다.

81) multi-variable grey model

82) single factor grey model

83) 하루에도 몇 번 씩 주식이나 파생 상품을 반복적으로 사고파는 것을 말한다. 주식 보유 기간에 따라 scalping과 swing 등의 전략이 있다.

주었다. 그 중에서도 특히 시간에 따른 Fourier Series를 사용하는 modified GM (1,1)이 model fitting 및 예측에서 가장 좋았다.

Lotka-Volterra 모형을 주식시장 예측에 사용하는 연구도 있다. Lotka-Voltera 모형은 잘 알려진 경쟁 확산 모형 (competitive diffusion model) 이다. Lee 등(2005)은 Lotka-Volterra 모형을 이용한 동적 경쟁 분석(dynamic competition analysis)를 주식시장의 기술적 예측(technological forecasting)에 적용하였다. generalized Lotka-Voltera(GLV)모형 power-law probability distribution을 따르는 다양한 현상을 시뮬레이션하고 분석 및 예측을 할 수 있는 일반적인 수단을 제공한다. Solomon (2000)은 이러한 GLV모형을 주식시장에 어떻게 적용할 수 있는지 보여주었다.

Kim은 시퀀스 정렬 알고리즘(sequence sort algorithm)[84]을 응용한 주식 가격 예측 기법을 제안하였다(Kim & Cho, 2010). Majhi (2009)는 효율적인 주가지수 예측을 위해 BFO(Bacterial Foraging Optimization)과 ABFO(Adaptive Bacterial Foraging Optimization) 기반의 기법을 도입하여 GA(geneticl algorithm)와 PSO(particle swarm optimization) 기반의 모형 대비 더 빠르고 정확한 성과를 보여주었다. Grosan 등(2005)은 NASDAQ지수와 NIFTY 지수의 예측을 위하여 유전 프로그래밍 기법(genetic programming technique)인 MEP (multi-expression programming)을 적용하였는데, SVM, Levenberg-Marquardt algorithm[85] 및 Takagi-Sugenoneuro-fuzzy

84) 정렬 알고리즘(sort algorithm)이란 원소들을 일정한 순서대로 열거하는 알고리즘인데, 효율적인 정렬은 다른 알고리즘을 최적화하는데 중요하며, 정렬 알고리즘은 데이터 정규화와 의미 있는 결과물 생성에 흔히 유용하게 쓰인다. 데이터를 새로운 순서(sequence)로 재배열하는 데는 동일한 결과를 얻을 수 있는 많은 솔루션이 있으며 다른 것보다 빠르게 데이터를 재배열 할 수 있는 정렬 알고리즘이 존재한다. 출력은 보통 비 내림차순이며 입력을 재배열한 순열이다. 정렬 알고리즘은 그 특징에 따라 비교 정렬, 제자리 정렬, 온라인 정렬 등으로 분류될 수 있다.

85) Levenberg-Marquardt Algorithm (LMA)은 비선형 최소 제곱 문제(non-linear least squares problems)를 해결하는 데 사용되며, DLS (damped least-squares) method로도 알려져 있다. 이러한 최소화 문제는

inference system 과 비교하여 좋은 성능을 보여 주었다. Pai는 주식 가격 시계열 데이터를 입력으로 하여 전통적 시계열 예측인 ARIMA[86]를 사용한 후 SVM을 사용하여 주가 방향에 대한 예측을 수행하였다(Pai & Lin, 2005). Wu는 결정 나무(decision tree)를 이용하여 주식 가격 예측 방법을 제안하였다(Wu, Lin & Lin, 2006). Lai 등(2009)은 주식에 대하여 퍼지 결정 나무(Fuzzy Decision Tree, FDT)를 진화시키고 클러스터링하여 새로운 금융 시계열 예측 모형을 확립한다. 이 예측 모형은 데이터 클러스터링 기술(data clustering technique), 퍼지 결정 나무(Fuzzy Decision Tree) 및 유전 알고리즘(GA)를 통합하여 과거 데이터 및 기술적 지표를 기반으로 의사 결정 시스템(decision-making systems)을 구성한다. k-평균 알고리즘(k-means algorithm)[87]을 사용하여 과거 데이터 집합을 k개의 서브 클러스터로 나눈 다음 GA를 적용하여 FDT에서 각 입력 지수에 대한 퍼지 항목의 수를 늘림으로써 모형의 예측 정확도를 향상시켰다. 이러한 GAFDT모형을 TSEC(Taiwan Stock Exchange Corporation)에 있는 주식들에 대해 적용한 결과 82%의 높은 예측성공률을 나타냈다. Dey는 주식 시장의 추세를 예측하는 효과적인 모형을 만들기 위해 XGBoost(eXtreme Gradient Boosting)[88]을 도입하여 60일과 90일 기간 동안 무려 87%에 이르는 정확도를 보여주었다. 이로써 XGBoost 모형이 전통적인 비앙상블 학습기법(non-ensemble learning technique)에 비해 훨씬 나은 모형임을 증명하였다(Dey et al.).

특히 최소 제곱 커브 피팅(least squares curve fitting)에서 발생한다.

86) ARIMA모형은 AR모형과 MA모형이 결합(I)된 모형으로 시계열분석에서 널리 쓰이고 있는 모형이다. ARIMA 모형은 안정적 시계열(stationary time series)이라는 가정이 필요한데 비안정적 시계열(nonstationary time series)일 경우에는 차분(difference), 로그변환(log transform) 등을 통해 안정적 시계열로 만들어주어야 한다.

87) k-평균 알고리즘(k-means algorithm)은 주어진 데이터를 각 cluster와 거리 차이의 분산을 최소화하도록 k개의 cluster로 묶어 주는 알고리즘이다. 이것은 label이 없는 입력 데이터에 label을 달아주는 역할을 하는 unsupervised learning의 일종이다.

88) XGBoost는 Friedman의 “Greedy Function Approximation: A Gradient Boosting Machine”에서 제안된 Gradient Boosting 모형에 기초하며 지도 학습(Supervised Learning)에 사용된다.

Han은 다른 기법들과 달리 재무 정보를 기반으로 주가 방향을 SVM을 통해 예측하였다(Han & Chen, 2007). Kazem은 Support Vector Regression(SVR)을 이용하여 주식 가격에 대한 예측 모형을 제안하였다(Kazem, Ahmam, et al., 2013). 이것은 chaotic mapping[89], firefly algorithm[90], 그리고 SVR을 기반으로 한 예측모형이다.[91] Wen은 SVM을 사용하여 주가 방향에 대한 예측과 자동으로 매수/매도 결정을 내리는 시스템을 제안하였다(Wen, Qinghua, et al., 2010). Yu(2005)는 유전자 알고리즘 기반 서포트 벡터 머신(Genetic Algorithm based Support Vector Machine, GASVM) 모형을 제안하였다. 유전자 알고리즘(GA)은 SVM의 모형 복잡성(complexity)을 감소시키고 SVM의 속도를 향상시키기 위한 변수 선택(variable selection)에 이용된다. 이러한 GASVM 모형의 예측성과는 통계모형, 시계열모형, 신경망 모형을 모두 능가하였다. Zhiqiang 등(2013)은 주식 가격 시계열 데이터를 입력으로 하여 SVM을 이용한 주가 예측 모델을 제안하는데, 다른 연구와 달리 Zhiqiang은 PSO(particle swarm optimization)[92]을 이용하여 SVM 파라미터를 결정하였다. 기계학습에서 제어 파라미터 (control parameter)를 최적화하는 것이 중요하므로 연구자는 효율적인 최적화 기법을 찾아야 하는데, Mustaffa 등(2014)은 최소제곱 SVM(Least Square Support Vector

89) chaotic map은 일종의 chaotic behavior를 나타내는 진화 함수(evolution function)이다.

90) 수학적 최적화에서, 반딧불이 알고리즘(firefly algorithm)은 반딧불이의 반짝이는 행동에서 영감을 받아 Xin-She Yang이 제안한 메타 발견적 학습(metaheuristic)이다.

91) 예측 모형은 세 단계로 구성된다. 첫 번째 단계에서는 보이지 않는 위상 공간 동역학(unseen phase space dynamics)을 재구성하기 위해 coordinate embedding method를 사용하고, 두 번째 단계에서는 SVR 하이퍼 파라미터를 최적화하기 위해 chaotic firefly 알고리즘을 적용한다. 마지막으로 세 번째 단계에서는 최적화된 SVR이 주식 시장 가격을 예측하는데 이용된다.

92) PSO(Particle Swarm Optimization)는 Kennedy, Eberhart (1995)에 기인하며 처음에는 사회적 행동을 시뮬레이션하기 위한 것이었다. 주어진 quality measure와 관련하여 candidate solution을 반복적으로 개선함으로써 문제를 최적화하는 computational method이다. 즉. 입자라고 불리는 candiadte solution 집단을 만들고 입자의 위치와 속도에 대한 간단한 수학 공식에 따라 검색 공간에서 이 입자를 움직여서 문제를 해결한다.

Machine, LS-SVM)[93]의 파라미터를 최적화하는데 Swarm Intelligence 접근법, 즉 artificial bee colony (ABC)[94]를 사용하여 BP-NN, GA 대비 높은 예측 정확도를 보여주었다. Shen 등(2009)도 LS-SVM을 Dynamic Inertia Weight Particle Swarm Optimization (W-PSO)[95]을 사용하여 최적화시킨 결과, 주식 수익률 예측에서 BP-NN을 초과하는 정확도를 시현하였다.

최근의 많은 연구들은 성공적인 feature 선택방법이 주식 시장 예측의 정확도를 향상시킬 수 있음을 보여주고 있다. 피쳐 선택(feature selection)은 효과적인 예측을 위해 주어진 데이터 세트로부터 대표성이 없는 변수를 필터링하는 것을 목표로 한다. 학습 과정(learning process)을 실행하기 전에 의미 있는 특징(feature)을 선택하는 것은 가설 공간(hypothesis space)의 크기를 줄여 데이터의 차원수(dimensionality)를 감소시킬 수 있을 뿐만 아니라, 결과적으로 해석이 더 쉬운 압축된 표현을 얻을 수 있다. Ni (2011)는 일일 주가 지수의 방향을 예측하기 위하여 프랙탈(fractal)[96]과 SVM을 결합한다. 프랙탈 피쳐 선택 방법(fractal feature selection method)은 비선형문제를 풀기에 적합하며 몇 개의 중요 피쳐(features)를 선택해야 하는지 정확하게 파악할 수 있게 해준다. Ni (2011)의 연구는 프랙탈 피쳐 선택 방법(fractal feature selection method)이 비교적 적은 수의 피쳐(features)를 선택하고 최

93) 표준 SVM보다 더 빨리 회귀하는 향상된 알고리즘이다.

94) ABC 알고리즘은 Basturk과 Karaboga가 제안한 새로운 메타 - 휴리스틱 접근법(meta-heuristic approach)인데, 저자들은 검색전략(searching strategy)을 풍부히 하고 과적합(over fitting)을 방지하는 것과 같은 중요한 문제를 고려하여 original ABC에 대해 두 가지 수정 사항을 도입한다.

95) 파라미터 선택에서 standard PSO 보다 좋은 성과를 나타낸다.

96) 프랙탈은 모든 규모로 표시되는 반복 패턴을 나타내는 수학적 집합이다. 확장 대칭(expandig symmetry) 또는 진화 대칭(evolving symmetry)라고도 한다. 복제가 모든 규모에서 정확히 동일하면 자기 유사 패턴(self-similar pattern)이라고 한다. 또한 프랙탈은 다른 레벨들에서 거의 동일 할 수 있는데, 이러한 패턴은 Mandelbrot 집합의 작은 확장들로 예시된다. 아울러 프랙탈은 자체 반복되는 상세한 패턴의 아이디어를 포함한다.

선의 평균 예측 정확도를 달성한다는 것을 보여준다. Tsai (2010)은 더 나은 예측을 위한 대표 변수 선택을 위해 여러 피쳐 선택 방법(multiple feature selection method)을 결합하였다. 특히, 주성분분석(Principle Component Analysis, PCA)[97], GA(Genetic Algorithm) 및 분류 회귀 나무(Classification and Regression Tree, CART)[98]와 같은 세 가지 잘 알려진 피쳐 선택방법을 사용하고, 결합 방법으로 통합(union), 교차(intersecton) 및 다중 교차(muti-intersection) 전략을 기반으로 대표성이 없는 변수를 필터링한 후 각각의 성과를 비교 분석하였다. 예측 모형으로는 BP-NN을 사용하였는데, PCA와 GA의 교차 전략, PCA와 GA 및 CART의 다중교차 전략이 최상의 결과를 나타냈다.

Sun 등(2001)은 홍콩 일일 지수에 대해 다중프랙탈 분석(multifractal analysis)를 수행하여 예측가능성(predictability)을 검증하였다. 하지만, Wei와 Huang(2005)는 상하이 지수의 고빈도(5분) 데이터를 다중프랙탈(mutifracta)l로 분석한 결과 Sun 등(2001)과는 상이한 결과를 얻었다. 주가지수 일일 수익률에 따르는 다중프랙탈 스펙트럼(multifractal spectrum) 매개 변수들의 의존성(dependence)에 대한 보편적인 규칙이 없다고 가정하고, Sun 등(2001)보다 철저한 방법으로 다중프랙탈 스펙트럼(multifractal spectrum)에 기초한 시장위험 측정을 구성하고 주가지수 변동을 예측하는 능력을 시험하였다.

---

97) 주성분 분석 (Pricipal Component Analysis, PCA) 은 상관관계가 있는 변수들을 선형 결합하여 변수를 축약하는 기법으로 요인 분석의 한 종류이다. 여러 변수들 간에 내재하는 연관성을 이용해 소수의 주성분으로 차원을 축소한 후 분석을 시행하면 연산속도와 결과를 향상시킬 수 있다.

98) 결정 나무(Decision Tree, DT)의 두 주요 유형인 분류 회귀 나무 (CART)는 분류 나무(Classification Tree)와 회귀 나무(Regression Tree)를 둘 다 지칭하기 위해 사용하는 포괄적인 용어로 Breiman 등(1984)이 처음 도입하였다. 분류 나무(Classification Tree)와 회귀 나무(Regression Tree)는 몇 가지 유사점을 가지고 있지만, 또한 분열(split) 위치를 결정하는 데 사용되는 절차와 같은 몇 가지 차이점도 있다.

Hsieh 등(2010)은 주가 예측을 위해 ABC-RNN과 웨이블릿 변환(wavelet transformation)[99]이 결합된 통합 시스템을 도입하였다. 웨이블릿 변환(wavelet transformation)은 시간영역 입력을 웨이블릿 영역에 매핑(mapping)하여 재무 특성을 명확하게 식별 할 수 있다. 주식 가격 시계열을 분해하여 잡음을 제거하기 위해 Haar wavelet[100]을 이용한 웨이블릿 변환(wavelet transformation)을 적용하고, 다양한 기본적, 기술적 지표를 이용한 RNN을 적용하여 단계적 회귀-상관관계 선택(Stepwise Regression-Correlation Selection, SRCS)을 통해 선택된 input features를 구성하였다. Artificial Bee Colony 알고리즘 (ABC)은 파라미터 공간 설계(parameter space design) 하에서 RNN 가중치(weights)와 바이어스(biases)를 최적화하는 데 사용하였다. 이처럼 wavelet기반 전처리(preprocessing)된 ABC-RNN모형을 여러 주식시장(Dow Jones Industrial Average Index (DJIA), London FTSE-100 Index (FTSE), and Tokyo Nikkei-225 Index (Nikkei), Taiwan Stock Exchange (TAIEX))에 시뮬레이션한 결과 다른 방법들보다 우수한 성과를 거두었다. Chang과 Fan (2008)은 주가예측을 위해 웨이블릿(wavelet)과 TSK(Takagi-Sugeno-Kang) 퍼지법칙 기반(fuzzy-rule-based) 시스템을 통합한 접근법을 개발하였다. 웨이블릿 변환에 의해 얻어진 시계열 데이터의 계층적 스케일 분해(hierarchical scalewise decomposition)로부터 흥미로운 표현을 여러 개 선택한 다음, TSK fuzzy-rule-based system을 사용하여 선택된 기술적 지표에 기초한 주가 예측을 진행한다. rule explosion을 피하기 위해, 데이터를 클러스터를 위한 k-means algorithm이 적용되고 각 클러스터 퍼지 규칙이 생성된다. 마지막으로 KNN(K nearest neighbor)을 sliding window로 적용하여 TSK 모형의 예측결과를 더욱 세밀하게 조정한다. 이러한 방법으로 시뮬레이션을 한 결과, 주식 가격의 변

99) 웨이블릿 변환 (wavelet transformation)은 웨이블릿 기저함수를 사용해 데이터를 변환하는 것을 말한다. 웨이블릿 기저 함수는 적분하면 0이 되고 진동하면서 진폭이 0으로 수렴하는 함수이다.

100) Haar Wavelet은 가장 단순하며 대표적인 웨이블릿(wavelet) 기저함수이다. 웨이블릿 기저함수는 다양하게 구성할 수 있는데 활발한 연구가 이루어지고 있다.

동을 매우 정확하게 예측할 정도로 성공적이었다. 기존 예측 모형과의 비교 연구를 통해 TSK 모형이 주가 예측을 위한 실시간거래 시스템에서 구현될 수 있을 정도로 유망한 모형임을 입증하였다. Huang(2011)은 주가지수 예측을 위해 웨이블릿(wavelet)과 커널(kernel) 부분 최소 제곱(partial least square, PLS) 회귀 분석을 결합하였다. 입력 데이터의 높은 차원(dimensionality)과 다중공선성(multicollinearity)때문에 웨이블릿 커널 PLS regressor를 사용하여 입력과 출력 간의 최대 공분산을 유지하고 최종 예측을 수행하는 가장 효율적인 subspace를 만들었다. 분석 결과, 제안된 모형이 전통적인 신경망, SVM, GARCH 모형보다 우수함을 보여 주며, 예측 오차를 크게 줄였다.

주식 시장에 에이전트 기반 모델링(Agent Based Modeling, ABM)[101]을 도입하여 시뮬레이션하려는 시도도 있다. 금융시장의 에이전트 기반 모델링은 주가 수익률의 정형화된 패턴에 대한 행동적인 기원을 다루었다. 전통적인 효율적 시장 관점과 달리 최근의 행동 금융 모형(behavioral finance model)은 금융 시장에서의 보편적인 상호 작용 패턴의 궤적(imprint)으로 수익률의 특성을 설명한다. Chen과 Liao(2005)는 주식 시장의 에이전트 기반 모델링을 바탕으로 주식 수익률과 거래량 간의 인과 관계를 설명한다. Feng 등(2012)과 Alfarano 등(2005)도 주식시장에 에이전트 기반 모델링을 적용하여 가격 패턴과 수익률 변동을 분석한다. Feng의 분석은 주가 수익률의 장기 기억에 대한 행동적인 해석(behavioral interpretation)을 제공한다(Feng et al., 2012). Alfarano는 주가 수익률 분포의 정형화된 사실(fat tails, volatility clustering)이 거래자들 간 상호작용으로부터 나타나는 속성이라는 것을 에이전트 기반 모델링을 도입하여 설명한다(Alfarano et al., 2005). Lee (2006)은 HRBFN (Hybrid Radial Basis-Function recurrent Network)을 이용한 에이전트 기반 주가예측 시스템인 iJADE Stock Advisor를 도입하여 RTT (round

101) ABM (agent-based model)은 시스템에 미치는 영향을 평가하기 위해 자율적인 에이전트 (개별 또는 집단 모두)의 동작 및 상호 작용을 시뮬레이션하기 위한 모형이다.

trip time) 분석, window-size evaluation test (장기 추세 및 단기 예측), 주가 예측 성능 평가(stock prediction performance test)등을 수행한 결과 효율성, 정확성, 유연성 등에서 모두 좋은 결과를 보였다.

소수 게임 (Minority Game, MG)[102]은 한정된 리소스를 위한 이상적인 상황에서 에이전트의 집단행동(collective behavior of agents)을 이해하기위한 간단한 모형인데, 그것은 통계 역학의 관점에서 무질서한 동적 복잡계(complex dynamical disordered system)로 간주되어왔다. Ma 등(2010)은 주식시장예측에 소수 게임 데이터 마이닝 (Minority Game Data Mining, MGAM)이라는 프레임워크를 제안하였다. 소수 게임을 따르는 다양한 에이전트 그룹의 행동을 결합하여 집단 데이터를 생성한 다음 MGAM 프레임 워크를 실제 주식 시계열 데이터 분석에 적용한 결과 모형의 승률이 random walk 보다 통계적으로 우수함을 입증하였다. Chen 등(2008)은 주식시장 예측에 에이전트 기반 혼합 게임 모형(agent-based mix-game model)을 수정한 형태인 진화 혼합 게임 모형(evolutionary mix-game model)을 이용하였다. 원래의 혼합 게임 모형(original mix-game model)에 에이전트들의 전략 진화 능력을 추가하여 향상시킨 진화 혼합 게임 모형을 주가지수 예측에 적용한 결과 적절한 매개 변수를 선택하면 예측의 정확도를 크게 향상시킬 수 있음을 보여 주었다.

데이터 마이닝 기술의 발전과 더불어 숫자형의 재무 데이터뿐만 아니라 뉴스 같은 텍스트 기반의 데이터도 주가 예측에 활용되고 있다. 뉴스 기사 기반 주가 변동 예측처럼 텍스트 문서와 시계열을 동시에 마이닝(mining)하는 것은 데이터 마이닝에서 새로운 주제라 할 수 있다. 이전 연구는 이미 뉴스 기사와 주가의 관계가 존재함을 시사했다. 금융 텍스트 마이닝

102) 소수 게임(minority game)은 Fribourg 대학의 Yi-Cheng Zhang과 Damien Challet이 제안한 El Farol Bar 문제의 한 변형이다. 소수 게임에서, 홀수의 플레이어는 각 턴마다 두 개의 선택 중 하나를 독립적으로 선택해야 한다.

(financial text mining)에 관한 대부분의 기존 문헌은 특정 용어가 사용되는 뉴스 기사에 표현 기법을 적용하고 주가가 움직이는 방향에 기반 하는 용어에 가중치를 할당한다. 그런 다음 이러한 가중치를 새로운 기사에 적용하여 가능한 주가 이동 방향을 예측한다. 최근의 일부 연구는 구글 검색 및 온라인 소셜 미디어 (블로그, 페이스북, 트위터 피드, etc.) 등 에서 다양한 경제 지표의 변화를 예측할 수 있는 초기 지시자(indicator)를 추출할 수 있다고 제안한다.

Schumaker & Chen(2010)은 언어, 재무 및 통계 기법을 통합한 텍스트 마이닝 기반의 AZFinText (Arizona Financial Text system)을 만들고 개별 주식 가격 예측 가능성을 조사하였다. AZFinText는 금융 뉴스 기사와 주가 시세표의 조합을 기반으로 이산 수치 예측(discrete numeric prediction)을 하는 데 중점을 두는 시스템으로 텍스트 분석(textual analysis)과 SVR(Support Vector Regression)을 통합하여 quant fund 중 수위권의 성과를 거두었다. Fung 등(2003)도 실시간 뉴스를 이용한 text mining과 시계열 데이터를 통합하여 주가예측을 수행하는 방법을 제안하였다. Ding 등(2015)은 이벤트 중심 주식 시장 예측(event-driven stock market prediction)을 위한 deep learning method[103]를 제안하였다. 먼저 뉴스 텍스트에서 이벤트를 추출하고 neural tensor network를 사용하여 훈련된 고밀도 벡터(dense vectors)로 표현한 다음, 회선 신경망(convolutional neural network, CNN)[104]을 이용하여 주가 움직임에 대한 장단기 영향을 모델링하였다. 실험결과, S&P 500 지수예측 및 개별주식 예측에 대해 기준방법보다 거의 6% 개선될 수 있음을 보여주었으며, 시뮬레이션 결과는 이전에 보고된 시스템들에 비해 높은 수익 창출 가능

103) 신경망을 여러 층 쌓아 올린 다층 구조(multi-layer structure) 형태의 신경망(neural network)을 기반으로 하는 기계학습(machine learning)의 한 분야이다. CNN, RNN, LSTM, Deep Belief Network, Deep Auto-encoder 등의 알고리즘을 사용하여 다량의 데이터로부터 학습을 통해 높은 수준의 모형을 구축하고자 하는 기법이다.

104) convolutional neural network (CNN, or ConvNet)은 피드포워드 인공신경망(feed-forward artificial neural network)의 한 유형이다. 뉴런간의 연결 패턴이 동물의 시간 피질과 유사하게 형성된다.

성을 보여주었다.

Bollen 등(2010)은 트위터 피드(Twitter feed)[105]의 텍스트 콘텐츠를 분석하여 트위터 무드(Twitter mood)[106]와 다우지수(Dow Jones Industrial Average, DJIA)간의 높은 상관관계를 발견하였고, 이것은 주식시장의 상승 하락 예측에 높은 적중률(86.7%)을 보여 주었다. Zhang 등(2010) 역시, 수집된 트위터 피드로부터 emotional tweet percentage가 다우 존스(Dow Jones) 지수, 나스닥(NASDAQ) 지수, S&P 500 지수와는 유의미한 음의 상관관계를, VIX(Volatility Index)[107]와는 유의미한 양의 상관관계를 나타냄을 밝혔다. 이는 트위터 포스트 (twitter post) 분석을 통해 주식 시장 지수를 예측할 수 있음을 시사한다. Choudhury 등(2010)은 블로그 영역(blogosphere)에서 커뮤니케이션 동학 (communication dynamics)을 분석하는 간단한 모델을 개발하고 이러한 blog communication dynamics와 주식 시장 활동 간에 유의미한 상관관계가 있음을 밝혔다. 이를 바탕으로 주식 시장 변동의 크기를 예측하는 데 약 78%의 정확도와 방향을 예측하는 데 약 87%의 정확도를 보여주었다. Antweiler & Frank(2004)는 야후 파이낸스 등의 인터넷 주식 메시지 보드(internet stock message board)에 포스팅된 메시지를 분석하여 주식 메시지(stock message)가 시장 변동성을 예측하는데 도움을 줄 수 있음을 발견하였다. Gilbert와 Karahalios (2010)도 웹 로그로부터 emotion을 추정하는 것이 미래의 주식 시장 가격에 관한 새로운 정보를 제공한다는 것을 증명하였다. Preis 등(2013)은 Google Trends를 이용하여 주식 시장의 움직임을 예측할 수 있는 패턴을 발견하였다. Mao 등(2011)은 다양한 온라인 데이터 세트 (twitter feeds, news headlines, google search queries)및 감정 추적 방법(twitter investor

105) OpinionFinder 와 Google-Profile of Mood States (GPOMS) 와 같은 mood tracking tool을 사용한다.

106) 뉴스가 주식 시장 가격에 확실하게 영향을 미치지만, 공공의 분위기 상태 (public mood states) 또는 정서(sentiments)도 또한 똑같이 중요한 역할을 할 수 있다.

107) 변동성 지수를 말하는데, S&P 500 지수 옵션과 관련해 향후 30일간의 변동성에 대한 시장의 기대를 나타내는데 증시에 참여하는 투자자들의 심리를 반영하는 지표라는 의미에서 일명 공포지수라고도 불린다. 보통 VIX지수와 주가지수는 반대로 움직이는 경향이 있다.

sentiment, negative news sentiment, tweet & google search volumes of financial terms)을 조사하고 다우 지수, 거래량, 변동성 지수(VIX)등과 같은 시장 지표의 예측에 대한 가치를 비교 평가하였다.

III. 문제 기술과 데이터 설명

3.1 기업 파산 예측

기업 파산 예측은 경영에서 중요한 문제이며 이 문제의 목표는 건실한 기업과 향후 파산 가능성이 높은 기업을 가려내는 것이다. 즉, 기업의 위기 시점을 예측하는 모델을 세우고, 이 모델을 통해 적절한 의사 결정을 내리는 것이다. 파산 예측을 위해서 기업의 현재 재무 상황을 아는 것이 필수적이다. 이 정보는 기업의 규모에 따라 값의 차이가 매우 크기 때문에 보통 파산 예측 모델에서는 비율을 사용한다(Alfaro et al., 2008). 예를 들어 유동 자산(current assets)의 경우 총자산(total assets)으로 나눈 비율을 사용한다.

기업의 파산 예측 중요성에 대해서 누구나 동의하지만 기업의 파산 시점을 어떻게 정의하느냐는 이견이 있을 수 있다. 본 논문에서는 워크아웃, 법정관리, 파산의 경우를 모두 파산 시점으로 정의하였다. 기업이 재무적 곤경에 처하면 주주, 채권자, 피고용인 등 여러 이해당사자들이 큰 고통에 직면한다. 위 세 가지 형태의 재무적 곤경 모두 경제적 이해당사자들에게 큰 손실을 주게 되므로 이를 예측을 위한 시점으로 정의하는 데 무리가 없을 것으로 판단된다.

기업의 파산 예측 모델에 사용할 재무 정보는 다양한 변수들이 사용될 수 있다. 예측이란 것은 현재의 자료를 토대로 미래의 상태를 판단하는 것으로, 본 논문의 예측 문제에서는 현재 재무 정보를 토대로 1년 후 기업의 파산 여부를 예측하도록 한다. 모델을 설정하고 모델의 예측력을 확인하기 위해서 파산 기업이 파산 시점에서 1년 전 재무 정보와 정상 기업의 1년 전 재무 정보를 사용하였다.

기업 파산 예측 모델에 적용할 기업의 재무 정보는 최근 5년 간 기업들의 재무제표를 참고로 하였다. 대한민국 모든 건설 기업에 대해 다음과 같은 재무 정보를 수집하였다. 이 데이터 ㈜나이스디앤비(NICE, 2013)에서 보유하고 있는 자료에서 원하는 재무 정보만 추출한 것이다. ㈜나이스디앤비는 국내 최대 기업신용정보 제공업체로 기업 신용 평가와 정보 서비스를 제공하고 있다.

기업을 파산 기업과 정상 기업으로 나누고 파산 기업은 2008년부터 2012년 사이에 워크아웃, 법정관리, 파산한 기업으로 정의하고, 정상 기업은 2012년 12월 기준으로 파산하지 않은 기업으로 정의하였다. 결과적으로 파산 기업은 총 1,381개, 정상 기업은 총 28,481개를 선정하였다. 그리고 파산 기업과 정상 기업의 재무 정보를 수집하였다. 이 때 파산 기업 재무 정보는 파산 기업의 파산 시점에서 직전 년도의 재무 정보를 사용하였고, 정상 기업의 재무 정보는 정상 기업의 2011년도 재무 정보를 사용하였다. 즉, 5년 간 파산한 기업의 파산 직전 1년 전 재무 정보와 정상 기업의 1년 전 재무 정보를 사용하여 1년 후 파산 유무를 예측하는 모델에 사용하였다.

파산 기업의 파산 직전 년도의 재무 정보에 대한 평균값과 최솟값, 최댓값은 〈표 1〉과 같다. 정상 기업의 2011년도 재무 정보에 대한 평균값과 최솟값, 최댓값은 〈표 2〉와 같다. 〈표 1〉과 〈표 2〉에서 알 수 있듯이 기업간 재무 항목별 값의 범위가 매우 차이가 큰 것을 확인할 수 있다. 따라서 재무 정보 값을 그대로 사용할 경우 예측 모델이 적절하지 않을 수 있다. 따라서 모델에 사용할 재무 정보, 즉 모델 변수들은 기업의 규모나 매출과 무관하게 적용할 수 있도록 비율로 계산하였다. 예를 들어 유동 자산(Current assets)과 관련한 변수는 총자산(Total assets)이나 유동부채(Current liabilities)로 나눈 값을 사용하였다.

〈표 1〉 파산 기업의 파산 직전 년도 재무 정보 (단위: 백만)

재무 항목	평균값	최솟값	최댓값
Capital	6,577	0	604,023
Cash	1,657	-827	143,406
Current Liabilities	38,538	0	2,183,073
Current Assets	25,833	0	901,609
Earnings before taxes	-127,120	-1,167,191	72,482
Earnings before interest and taxes	-11,824	-1,057,752	87,278
Liabilities	62,819	0	4,485,978
Sales	41,915	3	2,495,217
Total Debt	62,818	0	4,485,978
Total Assets	60,392	5	4,184,630
Working Capital	-12,600	-1,305,085	250,370

〈표 2〉 정상 기업의 2011년도 재무 정보 (단위: 백만)

재무 항목	평균값	최솟값	최댓값
Capital	11,360	-2,696	15,000,000
Cash	4,250	-848	2,718,731
Current Liabilities	34,360	-2	26,969,912
Current Assets	40,346	-38	39,496,344
Earnings before taxes	3,591	-10,964,144	11,518,274
Earnings before interest and taxes	5,172	-2,049,748	11,770,716
Liabilities	115,174	-244	290,107,362
Sales	72,643	-21,079	120,815,977
Total Debt	115,174	-244	290,107,362
Total Assets	161,198	0	305,270,506
Working Capital	6,404	-16,521,547	13,727,831

수집한 기업의 재무 정보를 이용하여 〈표 3〉과 같은 변수를 생성하여 모델에 적용하였다. 기업의 재무 정보에서 매출과 이익, 유동성, 채무, 자산을 설명할 수 있는 변수를 기존 유사연구에서 사용한 변수를 바탕으로 선정하였다(Wilson & Sharda, 1994; Altman, 2000; Min & Lee, 2005; Shin et al., 2005; Verikas et al., 2010; Sun et al., 2013). 또한 변수들 간에 상관계수가 낮은 변수를 선정함으로써, 서로 독립 가능성이 높은 변수를 채택하였다. 이는 상관계수가 높은 두 변수 중 하나를 제거해도 파산 예측 모델에 영향을 주지 않고 모델을 더 단순화 할 수 있기 때문이다. 각 변수는 기업의 규모나 매출과 무관하게 적용할 수 있도록 비율로 계산하였다. 예를 들어 유동 자산(Current assets)과 관련한 변수인 CA/TA와 CA/CL의 경우 각각 총자산(Total assets)과 유동부채(Current liabilities)로 나눈 값을 사용한다.

〈표 3〉 모델 변수

변수	설명
EBIT/TA	Earnings before interest and taxes/Totalassets
EBT/CAP	Earnings before taxes/Capital
WC/TA	Working capital/Total assets
WC/S	Working capital/Sales
CA/TA	Current assets/Total assets
CA/CL	Current assets/Current liabilities
C/TA	Cash/Total assets
C/CL	Cash/Current liabilities
lnTA	Natural logarithm value of total assets
S/CAP	Sales/Capital
S/CA	Sales/Current assets
S/TA	Sales/Total assets

〈표 4〉와 〈표 5〉에 전체 건설 기업의 모델 변수들에 대한 기술적 분석과 상관관계 행렬을 표시하였다. 앞서 말했듯이 〈표 5〉을 통해 변수들 간에 상관계수가 낮음을 확인할 수 있다. 〈표 4〉에서 WC/S나 C/CL, WC/TA, EBT/CAP, S/CAP과 같은 변수들의 편차가 평균의 10배가 넘는 매우 큰 편차를 보이고 있다. 이는 건설 기업들의 재무 구조가 서로 매우 상이함을 나타낸다고 볼 수 있고, 결국 Z-score와 같은 통계적 방법의 예측력을 낮게 만든다.

〈표 4〉 건설 기업의 모델 변수들에 대한 기술적 분석

	WC/S	C/CL	S/CA	EBIT/TA	CA/TA	WC/TA
Mean	-13.06	0.58	3.45	-0.14	0.52	-0.29
Median	-0.02	0.03	2.34	0.03	0.52	-0.02
Std. Dev	298.03	5.45	6.01	0.96	0.27	6.65
	S/TA	EBT/CAP	S/CAP	lnTA	C/TA	CA/CL
Mean	1.40	-1.67	116.41	9.46	0.06	2.78
Median	1.02	0.08	9.85	9.44	0.02	0.96
Std. Dev	1.62	75.77	4671.31	1.67	0.10	16.81

〈표 5〉 건설 기업의 모델 변수들에 대한 상관관계

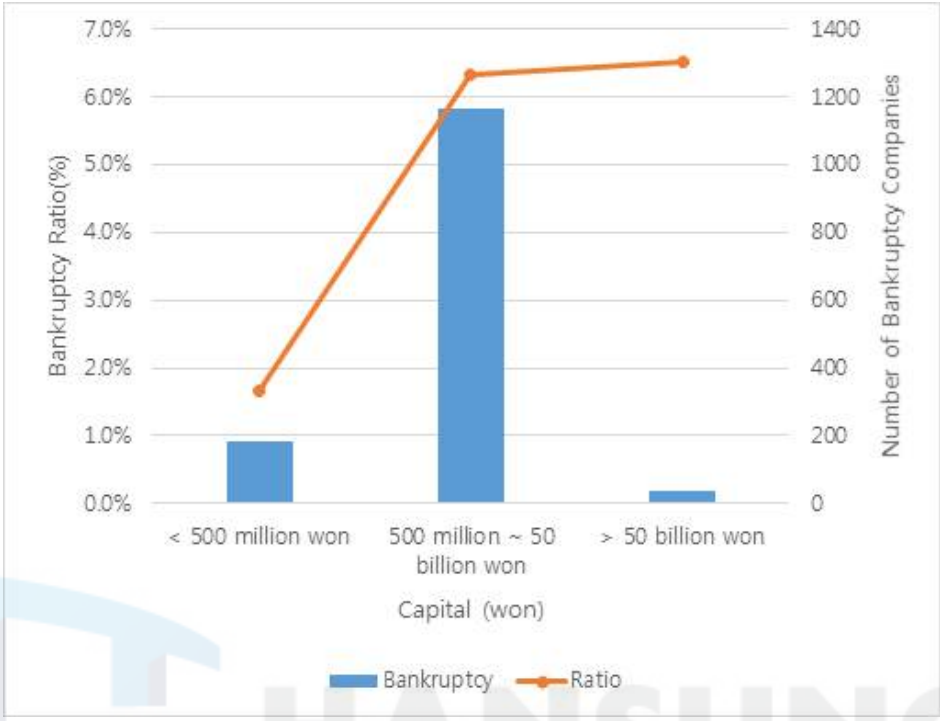
	WC/S	*C/CL*	*S/CA*	*EBIT/TA*	*CA/TA*	*WC/TA*
WC/S	1.00	0.01	0.02	0.08	0.05	0.02
C/CL	0.01	1.00	-0.03	0.02	0.06	0.01
S/CA	0.02	-0.03	1.00	0.14	-0.24	-0.58
EBIT/TA	0.08	0.02	0.14	1.00	0.06	-0.51
CA/TA	0.05	0.06	-0.24	0.06	1.00	0.07
WC/TA	0.02	0.01	-0.58	-0.51	0.07	1.00
S/TA	0.04	-0.02	0.52	0.05	0.30	-0.32
EBT/CAP	0.26	0.00	0.00	0.03	0.03	0.01

S/CAP	0.00	0.00	−0.01	0.00	−0.01	0.00
lnTA	−0.03	−0.11	−0.09	−0.01	−0.18	0.01
C/TA	0.04	0.32	−0.03	0.08	0.32	0.01
CA/CL	0.01	0.68	−0.04	0.03	0.10	0.02
	S/TA	*EBT/CAP*	*S/CAP*	*lnTA*	*C/TA*	*CA/CL*
WC/S	0.04	0.26	0.00	−0.03	0.04	0.01
C/CL	−0.02	0.00	0.00	−0.11	0.32	0.68
S/CA	0.52	0.00	−0.01	−0.09	−0.03	−0.04
EBIT/TA	0.05	0.03	0.00	−0.01	0.08	0.03
CA/TA	0.30	0.03	−0.01	−0.18	0.32	0.10
WC/TA	−0.32	0.01	0.00	0.01	0.01	0.02
S/TA	1.00	0.01	−0.01	−0.29	0.19	−0.01
EBT/CAP	0.01	1.00	0.82	−0.01	0.01	0.00
S/CAP	−0.01	0.82	1.00	0.05	−0.01	0.00
lnTA	−0.29	−0.01	0.05	1.00	−0.27	−0.11
C/TA	0.19	0.01	−0.01	−0.27	1.00	0.14
CA/CL	−0.01	0.00	0.00	−0.11	0.14	1.00

규모별 분류를 위해 건설 기업의 자본금에 따라 소형, 중형 대형으로 분류하였다. 소형 건설 기업은 자본금 5억 미만, 중형 건설 기업은 자본금 5억 이상 500억 미만, 대형 건설 기업은 자본금 500억 이상인 기업으로 선정하였다. 〈표 6〉와 〈그림 1〉은 기업 규모에 따른 파산 회사의 비율을 보여준다.

〈표 6〉 파산 기업 비율

자본금	파산 기업 수	정상 기업 수	파산 기업 비율
5억 미만	181	10,762	1.7%
5억 이상 500억 미만	1,164	17,203	6.3%
500억 이상	36	516	6.5%
전체	1,381	28,481	4.6%



〈그림 1〉 규모에 따른 파산 기업의 수와 비율

〈표 6〉과 〈그림 1〉에서 알 수 있듯이 대형 규모 건설 기업의 경우 전체 숫자는 많지 않으나, 파산 비율은 상대적으로 높은 것으로 확인되었다. 또한 소규모 건설 기업의 경우 반대로 기업의 수는 많으나 파산 비율은 매우 낮은 것을 알 수 있다. 대형 건설 기업의 경우 파산 비율도 높고 파산으로 인한 그 피해도 훨씬 크기 때문에 대형 건설 기업의 파산 예측이 특히 중요함을 알 수 있다.

3.2 주가 예측

분기별로 발표되는 기업의 재무 정보를 바탕으로 주식 가격 상승/하락 여부를 기계학습을 이용하여 예측한다. 기업의 분기별 재무 정보 발표가 주식에 얼마나 영향을 미치게 되며 시간이 지남에 따라 그 영향이 얼마나 차이가 있는지 알아보기 위함이다.

기업의 재무 정보에서 기업의 주식 가격과 밀접한 주당순이익과 주당순자산, 당기순이익을 입력으로 하여 한 달과 두 달 후 주식 가격이 상승/하락할지 예측한다. 주당순이익과 주당순자산, 당기순이익은 기존 연구에 따르면 일반적으로 주식 가격 예측에 유의한 것으로 알려져 있다(Han & Chen, 2007; Kim & Kim, 2004). 예측의 성능을 비교하기 위해 전문가 투자의견점수와 비교하여 재무정보를 활용한 기계학습 예측력을 평가한다. 그리고 한 달과 두 달 후의 주식 가격 상승여부 예측을 통해 재무 정보에 따른 주식 예측이 기간에 따라 얼마나 차이가 나는지 평가한다.

주당순이익과 주당순자산의 경우 기업의 규모에 따라 상대적으로 결정되는 값이지만, 당기순이익의 경우 기업 규모에 따라 매우 큰 차이를 보인다. 따라서 당기순이익을 그대로 입력으로 사용하지 않고 직전 분기의 당기순이익과 비교하여 비율을 사용한다. 즉, 당기순이익증가율을 사용한다.

본 논문에서 사용한 기업 데이터는 2013년 기준으로 KOSPI 200에 속한 200개 기업에 대해 2010년 1분기부터 2013년 3분기까지 분기별로 수집한 재무 데이터이다. 이 재무 데이터에서 주당순이익(EPS, Earning Per Share)과 주당순자산(BPS, Book-value Per Share), 당기순이익을 선택하고, 증권회사에서 발표하는 투자의견점수(5점부터 1점)를 수집하였다. 그리고 각 분기마다 분기 마지막 일을 기준으로 한 달 후, 두 달 후 주식 가격을 수집하였다. 위 데이터는 모두 FnGuide에서 제공하는 DataGuide에서 조회하였다. 200개 회사에 대해 총 15분기의 정보를 조

회하면 총 3000개의 데이터가 나와야 하나, 경우에 따라 필요한 정보가 빠져 있는 경우가 있어서 총 2913개의 샘플만 확보할 수 있었다. 〈표 7〉은 2013년 3분기 KOSPI 200 기업의 당기 순이익과 EPS, BPS, 주가 등에 대한 평균값과 최솟값, 최댓값을 나타낸 것이고, 〈표 8〉은 삼성전자의 샘플을 예로 보인 표이다.

〈표 7〉 2013년 3분기 KOSPI 200개 기업 데이터

	평균값	최솟값	최댓값
Net Profit(천원)	71,732,967	-732,439,459	4,627,864,000
EPS(원)	1,395	-48,650	46,227
BPS(원)	100,751	-1,667	2,048,262
Stock Price(원)	130,067	2,010	1,693,000
Experts(5-1)	3.8	3	4.05

〈표 8〉 삼성전자의 재무 정보와 주가

분기	2010-03-31	2010-06-30
Net Profit(천원)	3,167,036,000	3,155,614,000
EPS(원)	21,370.00	21,125.00
BPS(원)	445,444.69	460,047.68
Stock Price(원)	814,000	792,000
1M Later(원)	825,000	827,000
2M Later(원)	778,000	776,000
Experts(5-1)	4 점	4 점

EPS와 BPS는 기업의 규모에 상관없이 주식 가격을 기준으로 이익과 자산을 계산한 것이므로 기계학습에 사용하는데 문제가 없다. 하지만 당기순이익의 경우 기업의 규모에 따라 매우 큰 차이를 보이기 때문에 당기순이익율(NPGR, Net Profit Growth Rate)로 계산하여 사용하였다. 당기순이익율은 재무 데이터에서 당기순이익을 분기별로 수집하고 분기마다 증가/감소율을 계산하였다.

당기순이익율(%) = (당기순이익 – 직전분기당기순이익)/직전분기당기순이익 X 100

분기 한 달 후와 두 달 후 주식 가격은 증가/감소 여부에 따라 +1과 −1로 계산하였다. 예를 들어 2010년 1분기 한 달 후(4월 말일) 주가가 1분기(3월 말일)보다 상승하였다면 +1로 하고, 두 달 후(5월 말일) 주가가 1분기보다 감소하였다면 −1로 하였다. 투자의견점수는 기계학습 예측과 비교하기 위한 비교군으로 5점과 4점, 3점을 주가 상승(+1)으로 하고, 나머지는 주가 하락(−1)으로 가정하였다. 이렇게 나눌 때 가장 좋은 예측 결과가 나왔기 때문이다. 〈표 9〉은 〈표 8〉의 삼성전자 데이터를 전처리한 데이터를 보여준다.

〈표 9〉 삼성전자 데이터 전처리 결과

Quarter	2010-03-31	2010-06-30
EPS	21,370.00	21,125.00
BPS	445,444.69	460,047.68
NPGR	3.71	-0.36
1M Later	+1	+1
2M Later	-1	-1
Experts(5-1)	+1	+1

정리하면, 200개 기업에 대한 〈표 10〉과 같이 재무 데이터와 주식 가격을 수집하여 SVM의 학습과 테스트에 사용하였다.

〈표 10〉 학습/테스트 데이터

EPS	Earnings Per Share
BPS	Book-value Per Share
NPGR	Net Profit Growth Rate
1M Later (Target_1)	+1 if the price rises after one month later. -1 if drops.
2M Later (Target_2)	+1 if the price rises after two month later. -1 if drops.
Experts	+1 if experts' score is 3~5. -1 if experts' score is 1~2.

IV. 실험 및 결과

4.1. 파산 예측 모델

4.1.1 적응형 부스팅을 이용한 파산 예측 모델

적응형 부스팅은 기계 학습 알고리즘의 하나로 Freund와 Schapire가 수식화 한 것이다(Freund & Schapire, 1995)[108]. 다른 학습 알고리즘과 결합하여 사용하는 형태로, 보다 개선된 학습 알고리즘을 만들 수 있다. 적응형 부스팅[109]은 약 분류기(weak classifier)를 결합하여 보다 강한 분류기 학습 알고리즘을 만든다. 결합 방식은 가중 평균 방식이고, 가중치 값을 정하는 알고리즘과 분류 방법을 본 논문에서 사용한 경우에 적용하여 설명하면 다음과 같다.

약 분류기 알고리즘으로는 결정 트리를 사용하였고, 깊이(depth)는 1로 하였다. 즉 12개의 결정 트리 알고리즘은 각각 변수에 대해 분류기 학습을 수행한다. 각각의 결정 트리 알고리즘은 하나의 변수만 사용하고 깊이가 1이기 때문에 파산 예측력은 매우 낮다. 적응형 부스팅은 이 약한 예측력을 갖는 결정트리를 조화시켜 높은 예측력을 보이게 된다. 이 12개의 약 분류기의 집합을 H라고 한다.

---

108) 이들은 Adaboost를 개발한 공로를 인정받아 2003년 괴델상을 받았다. Adaboost는 이전 분류기의 잘못된 분류를 이어지는 약 학습기로 수정 가능하다는 점에서 다양한 상황에 적용할 수 있는 장점이 있다(adaptive). 다만 이로 인해 잡음(noise)이 많은 데이터와 이상값(outlier)에 취약하다는 단점도 있지만 과적합(overfitting)에 덜 취약한 면도 있다. 개별 학습기의 에러율이 0.5보다 낮다면 최종 모델은 강한 학습기로 수렴하게 된다는 것은 이론적으로 증명 가능하다. 약 학습기로 결정트리를 사용한 Adaboost는 종종 발군의 분류기로 불릴 만큼 좋은 성능을 보인다. 이때 AdaBoost 알고리즘의 각 단계에서 얻을 수 있는 훈련 샘플의 상대적 난이도 정보가 트리 성장 알고리즘에 반영된다.

109) Boosting에는 AdaBoost, Adacost, Arcing, LogitBoost, BrownBoost와 같은 많은 버전이 있지만, AdaBoost가 가장 널리 사용된다.

훈련 샘플 m개를 $(x_1, y_1), \ldots (x_m, y_m)$이라고 하자. 여기에서 x는 분류 대상의 특징을 나타내고, y는 클래스로 -1 또는 1의 값을 가질 수 있다. 본 논문에서는 한 기업의 모델 변수 집합이 x가 되고 파산 기업은 -1, 정상 기업은 1로 분류 된다. 즉 x=(EBIT/TA, EBT/CAP, WC/TA, WC/S, CA/TA, CA/CL, C/TA, C/CL, lnTA, S/CAP, S/CA, S/TA)인 특징 벡터가 된다. 각각의 약 분류기는 특징 x에서 하나의 값만으로 분류를 시도한다.

$D_1(i) = \frac{1}{m}$로 가중치의 분포를 초기화한다.

t=1부터 T까지 총 T회에 대하여 다음을 반복 수행한다.

- 오차가 가장 적은 약 분류기를 골라서 h_t라고 한다. 여기에서 오차는 가중치 분포에 따라 결정되는 가중 오차이다.

  $h_t = \arg max_{h_t \in H} |0.5 - e_t|$, 여기에서 $e_t = \sum_{i=1}^{m} D_t(i) I(y_i \neq h_t(x_i))$함수 I는 인자가 참일 경우 1, 거짓일 경우 0을 리턴 한다.
- $|0.5 - e_t| \leq \beta$ 이면 루프를 종료한다. 상수 β는 미리 정해 둔 임계값으로 오차가 이 값 이하로 낮아지면 훈련을 종료한다.
- α_t를 다음과 같이 계산한다. $\alpha_t = \frac{1}{2} ln \frac{1 - e_t}{e_t}$
- 그리고 m개의 가중치 분포 D를 다음과 같이 변경한다.

  $$D_{t+1}(i) = \frac{D_t(i) e^{\alpha_t (2I(y_i \neq h_t(x_i)) - 1)}}{\sum_i D_t(i) e^{(\alpha_t I(y_i \neq h_t(x_i)))}}$$

최종 분류기는 다음과 같이 계산된다.

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right)$$

여기에서 sign 함수는 인자가 양수일 경우 +1(정상 기업), 음수일 경우 -1(파산 기업)을 리턴 한다. 본 논문에서 T는 1000으로 설정하였다. 즉 12개의 약 분류기가 1000개의 가중치 합으로 계산되어 정상(+1)과 파산(-1)으로 분류된다.

4.1.2 실험 방법

정상 기업의 수가 파산 기업의 수에 비해 절대적으로 다수이기 때문에 둘의 수를 비슷하게 조정할 필요가 있다. 다만, 정상 기업에서 어떤 방식으로 표본을 추출하는 가에 따라 모델의 파산 예측이 달라질 가능성이 있기 때문에 공정한 방법이 필요하다. 따라서 정상 기업에서 무작위 표본 추출을 하여 파산 기업의 수와 일치하도록 하였다.

정상 기업과 파산 기업을 자본금 규모에 따라 분류한 후 파산 기업의 수만큼 정상 기업을 무작위로 표본 추출하여 파산 기업의 수와 동일하게 하였다. 전체 기업 샘플 수는 정상 기업 샘플 수와 파산 기업 샘플 수가 각각 1381로, 둘을 합하면 2762개이고, 이 중 80%인 2208개를 훈련 샘플로 사용하고 나머지 20%인 554개를 테스트 샘플로 사용하였다. 훈련 샘플과 테스트 샘플에는 각각 정상 기업과 파산 기업의 수가 동일하다. 기업 규모에 따른 실험에서도 80%를 훈련 샘플로 사용하고 20%를 테스트 샘플로 사용하였다.

적응형 부스팅(AdaBoost)의 상대적 성능 비교를 위해 다른 기계 학습을 사용한 모델로 인공 신경망과 SVM, 결정 트리를 실험하였다. 그리고 전통적 모델로 Z-score를 실험하였다.

적응형 부스팅의 약 분류기로 결정 트리(Decision Tree)를 사용하였고, 단순 결정 트리와 비교를 위해 결정 트리를 사용한 모델에 대한 실험도 함께 수행했다. 인공 신경망 모델은 변수를 모두 입력으로 하고, 히든 레이어의 노드 수는 10개, 출력은 1개로 하였다. SVM은 모델 변수를 특징(feature)로 하고, 커널 함수는 RBF를 사용하였다. 모든 기계 학습 기반 모델에서 타겟 값(target value)은 파산을 −1, 정상 기업을 1로 하였다.

Z-score는 개인 회사(private firm)를 위한 모델을 사용하였다. 전체 기업 샘플에 비상장 기업도상당수 포함되어 있기 때문이다. 이 경우 Z값은 다음 식으로 계산된다.

$$Z\text{-}score\text{-}private = 0.717T1 + 0.847T2 + 3.107T3 + 0.420T4 + 0.998T5$$

여기에서 T1=(유동자산-유동부채)/총자산, T2=유보이익/총자산, T3=세전영업이익/총자산, T4=장부가치/총부채, T5=매출액/총자산 이다. T4에서 장부 가치는 수집한 데이터에서 자본금(Capital)을 대신 사용하였다. 계산 결과 Z-score-private값이 2.9 초과이면 정상 기업, 1.23 미만이면 파산 기업, 그리고 중간 범위의 경우 예측 보류로 하였다.

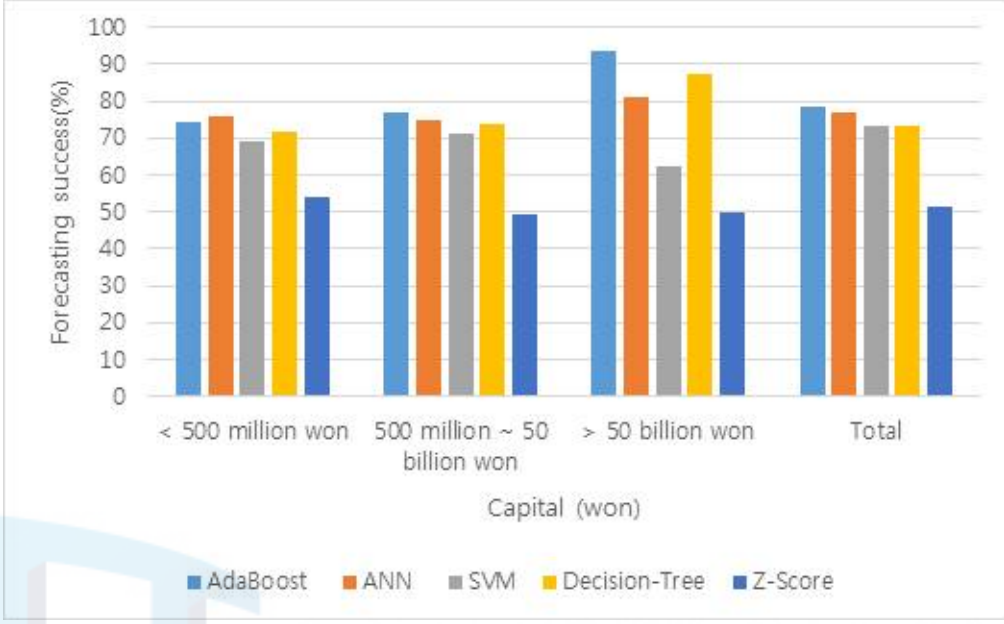
4.1.3 실험 결과

각 기업 규모별, 모델별 예측 결과는 〈표 11〉과 같다. 예측 보류는 Z-score에서 정상, 파산 판단이 되지 않는 범위에 속하는 것으로 성공이나 오류 어디에도 속하지 않는 것이다. 정상 기업의 샘플은 파산 기업의 샘플과 수를 동일하게하기 위해 실험 때마다 무작위로 표본 추출을 하기 때문에 각 기업 규모별 실험의 결과 합계가 전체 규모의 합과 일치 하지 않을 수 있다. 이를 보완하기 위해 실험을 10차례 반복하여 각 모델에 대해 중간값에 해당하는 결과를 선별하였다.

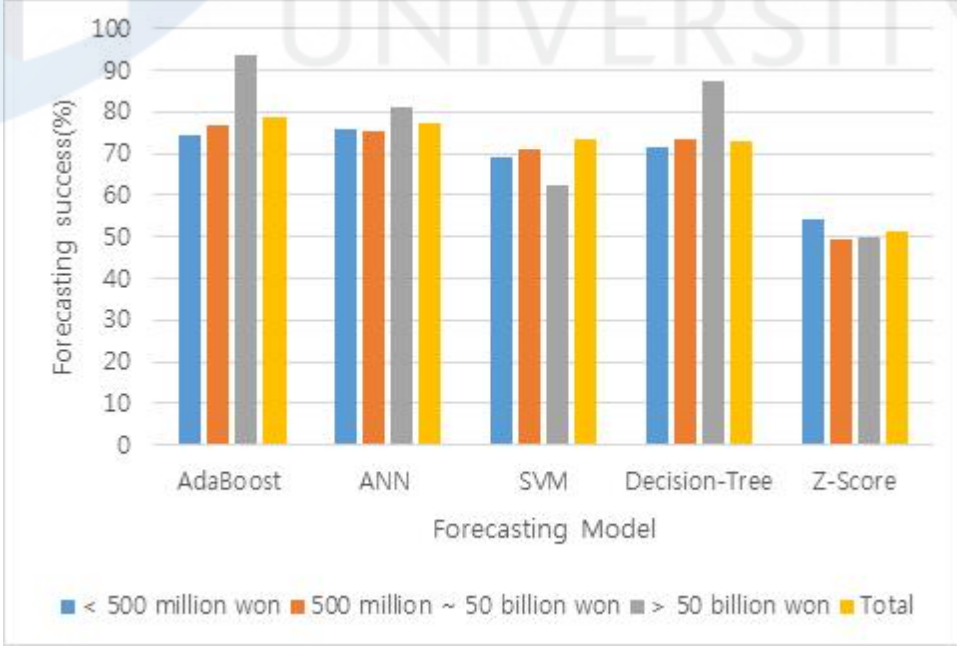
〈표 11〉의 결과를 바탕으로 모델별, 기업 규모별 예측 성공률을 계산하여 〈그림 2〉와 〈그림 3〉 에 표시하였다. 예측 성공률은 예측 성공 수를 테스트 샘플 수로 나눈 것이다.

〈표 11〉 파산 예측 결과

	성공	오류	미분류	오류 유형	
				Type-I	Type-II
자본금 5억 미만 기업 (Training samples:288, Test samples:74)					
AdaBoost	55	19	0	5	14
ANN	56	18	0	9	9
SVM	51	23	0	6	17
Decision-Tree	53	21	0	7	14
Z-Score	40	7	27	3	4
자본금 5억 이상 500억 미만 (Training samples: 1862, Test samples: 466)					
AdaBoost	358	108	0	43	65
ANN	350	116	0	57	59
SVM	332	134	0	61	73
Decision-Tree	343	123	0	57	66
Z-Score	230	96	140	78	18
자본금 500억 이상 기업 (Training samples: 56, Test samples: 16)					
AdaBoost	15	1	0	1	0
ANN	13	3	0	1	2
SVM	10	6	0	0	6
Decision-Tree	14	2	0	2	0
Z-Score	8	5	3	5	0
전체 기업 (Training samples: 2208, Test samples: 554)					
AdaBoost	435	119	0	48	71
ANN	427	127	0	57	70
SVM	406	148	0	59	89
Decision-Tree	405	149	0	60	89
Z-Score	284	102	168	80	22



〈그림 2〉 기업 규모별 모델의 예측 성공률



〈그림 3〉 모델별 예측 성공률

전통적 Z-score의 경우 예측 보류 구간으로 인해 성공률이 기계 학습 기반의 모델에 비해 현저히 떨어진다. 게다가 보류 구간으로 인해 예측 오류가 발생하지 않는 것은 아니다. 즉, 기계 학습 기반에 비해 Z-score의 예측력은 매우 낮다고 할 수 있다.

〈그림 2〉와 〈그림 3〉에서 보듯이 기업 규모에 따른 모델의 예측 성공률은 적응형 부스팅(AdaBoost)이 가장 우수함을 알 수 있다. 특히 기업 규모가 500억 이상인 대형 건설 기업의 경우 특히 상대적으로 더 우수함을 알 수 있다. 인공신경망(ANN) 모델의 경우 소형 규모의 경우 AdaBoost보다 약간 더 예측력이 좋은 것으로 보이나, 의미를 부여할 만큼은 아닌 것으로 판단된다. SVM의 경우 실험 결과로 볼 때 파산 예측 모델로 적합하지 못한 것으로 판단된다. 결정 트리는 단순한 알고리즘에 비해 예측력이 어느 정도 우수하며 특히 대형 규모 건설 기업에서 눈에 띄는 예측력을 보인다. 적응형 부스팅은 이 결정 트리를 약 분류기로 하여 더 정확한 파산 예측이 가능했던 것으로 보인다.

4.2 주가 예측 모델

4.2.1 실험 방법

수집한 2913개의 샘플을 무작위로 섞은 후 절반을 나누어 학습용(1457개)과 테스트용(1456개)으로 사용하고, 이 둘을 바꾼 후에 다시 학습과/테스팅을 수행하였다. 그리고 이 과정을 10회 반복하여 평균 예측률을 계산하였다. 데이터를 무작위로 섞고, 상호검증을 반복적으로 함으로써 학습과 테스팅 데이터의 우연에 의해 나타날 수 있는 SVM 예측 결과를 상쇄하도록 하였다.

각 기업 분기 재무 정보에서 수집한 EPS, BPS, NPGR 중 어떤 조합이 예측력이 더 우수한지 비교하기 위해 모든 가능한 조합에 대하여 실험을 하였다. 즉, {EPS}, {BPS}, {NPGR}, {EPS, BPS}, {EPS, NPGR}, {BPS, NPGR}, {EPS, BPS, NPGR} 각각을 입력으로 하는 예측 모델을 만들었다.

〈표 12〉 파라미터에 따른 SVM 예측 정확도(%) 입력={EPS, BPS}

	γ =1	0.1	0.5	0.01	0.001
C=1	57.3	57.4	57.4	57.3	57.02
5	57.3	57.3	57.4	57.5	56.64
10	57.3	57.3	57.36	57.32	57
Average=57.30%, Standard Dev.=0.22					

SVM의 비선형 분류를 위해 커널 함수를 RBF(Radial Basis Function)를 사용하였다. SVM 구현은 LibSVM(Chang & Lin, 2001)을 파이썬(Python)에서 쓸 수 있도록 만든 mlpy(machine learning py)를 사용하였다(Albanese, Visintainer, Merler, Riccadonna, Jurman & Furlanello, 2012). 이때 파라미터 gamma($\gamma = 1/2\sigma^2$)는 0.01, C는 5를 사용하였다.

이 파라미터를 선정하기 위해 다양한 gamma와 C값에 대해 〈표 12〉와 같은 실험 결과를 구하였다. 그 중 가장 좋은 결과를 낸 파라미터 값을 선택하였으나, 예측 정확도가 파라미터에 따라 차이가 크지 않았다. 비교를 위한 인공신경망은 neurolab을 사용하였고(NeuroLab, 2011), 결정트리와 적응형부스팅은 scikit learn 모듈을 사용하였다(scikit learn, 2010). 인공신경망은 입력 층의 노드 수는 입력 변수의 수와 동일하게 하고, 출력 층의 노드 수는 1개, 히든 층의 노드 수는 10개로 하였다. 적응형부스팅의 약분류기(weak classifier)로는 결정트리를 사용하였다.

4.2.2 실험 결과

〈표 13〉은 실험 결과를 나타낸 표이다. 재무 정보 발표 한 달 후 주가 예측(SVM(1M))은 두 달 후 예측(SVM(2M))에 비해 어떠한 입력을 사용하더라도 예측력이 높음을 알 수 있다. 즉, 시간이 지남에 따라 재무 정보에 따른 주가 예측력이 떨어짐을 확인할 수 있다.

〈표 13〉에서 SVM(1M)과 Expert에 해당하는 결과를 비교해보면 쉽게 한 달 후 예측이 전문가 점수보다 뛰어남을 알 수 있다. 다만, 전문가 점수가 증권회사 등에서 발표하는 전문가 점수를 단순히 평균한 것이므로 이 결과만으로 SVM의 예측이 전문가보다 뛰어나다고 단정 짓기는 어렵다.

〈표 13〉에서 SVM(1M)과 ANN, D.Tree, Adaboost 결과를 비교해보면 SVM의 예측 결과가 다른 기계학습 방법에 비해 우수함을 알 수 있다. SVM의 예측력과 가장 근접한 결과를 보이는 인공신경망(ANN)의 경우 학습 시간을 최대 10분으로 하였고, 매번 학습 시간을 초과하여 학습이 종료되었다. 즉, 본 실험에서 사용한 데이터와 같이 인공신경망의 노드 가중

치가 빠르게 수렴되지 않는 경우 학습 시간이 너무 오래 걸리는 단점이 있다. SVM은 빠른 학습시간에도 불구하고 인공신경망보다 나은 예측 결과를 보였다.

〈표 13〉 예측 결과 – 예측 정확도(%) 평균과 표준편차

	eps, bps, npgr	eps, npgr	eps, bps
svm(1m)	57.1 (0.12)	54.6 (0.02)	57.5 (0.11)
svm(2m)	50.1 (0.21)	50.6 (0.22)	50.4 (0.23)
expert	48.6	48.6	48.6
ANN(1m)	57.0 (0.30)	53.0 (0.25)	56.5 (0.31)
D.Tree(1m)	53.9 (0.25)	51.0 (0.22)	55.4 (0.31)
Adaboost(1m)	56.0 (0.32)	52.5 (0.33)	55.9 (0.41)

	bps,npgr	eps	bps	npgr
svm(1m)	56.1 (0.34)	53.7 (0.44)	55.2 (0.53)	56.7 (0.45)
svm(2m)	49 (0.55)	49 (0.53)	47.7 (0.49)	52.9 (0.37)
expert	48.6	48.6	48.6	48.6
ANN(1m)	54.5 (0.33)	53.7 (0.23)	54.7 (0.45)	54.9 (0.43)
D.Tree(1m)	54.9 (0.25)	49.7 (0.21)	51.7 (0.19)	53.5 (0.22)
Adaboost(1m)	55.0 (0.30)	51.7 (0.25)	53.6 (0.23)	53.9 (0.31)

〈표 13〉에서 SVM(1M)의 결과를 중심으로 입력 변수에 따른 결과를 비교해보면 하나의 입력 변수로는 {NPGR}이 제일 예측력이 우수하지만, 둘 이상을 조합하는 것이 예측력 향상에 도움이 됨을 알 수 있다. 본 실험에서 사용한 데이터의 경우 {EPS, BPS}만으로 된 입력 조합이 가장 우수한 결과를 보여준다. NPGR까지 포함한 {EPS, BPS, NPGR} 입력 조합 보다 더 우수하다. 이는 입력 변수가 많다고 항상 좋은 결과가 나오지 않음을 의미한다.

예측 정확도가 57.5%로 기존의 주가 방향 예측 연구들과 유사한 성능을 나타내고 있다. 50% 초과 예측 정확도는 승률게임에서는 매우 유용할 수 있다. 특히 오랜 기간에 걸쳐 여러 번 게임이 이루어지는 경우 복리로 수익이 쌓이기 때문에 51%의 승률도 의미를 가질 수 있다.

4.2.3 결과의 통계적 검증

SVM의 예측 정확도에 대한 통계적 검증을 위해 다음 두 가지 귀무가설에 대해 p-value를 계산하고자 한다. 이때 〈표 13〉에서 결과가 가장 좋은 입력이 {EPS, BPS}인 경우만을 고려한다.

1번 귀무가설: SVM의 예측 정확도는 무작위 예측에 의해서도 우연히 발생할 가능성이 있다.

2번 귀무가설: SVM의 예측 정확도와 다른 기계학습 방법(ANN, D.Tree, Adaboost)의 정확도의 차이는 우연히 발생할 가능성이 있다.

먼저 1번 귀무가설이 발생할 확률(p-value)를 계산하기 위해 무작위로 주가 예측을 시도한다고 가정한다. 이때 예측이 맞을 확률은 반반, 즉 0.5이다. 실험에서 총 1456개의 테스트 데이터에 대해 예측을 하기 때문에 SVM(1M) 만큼 예측 정확도가 나오기 위해서는 57.5% (837개)에 대해

예측이 성공해야 한다. 이러한 무작위 예측은 베르누이 시행으로 볼 수 있다. 즉, 다음과 같이 계산된다.

$$P(X=837)=\binom{1456}{837}0.5^{1456}$$

여기에서 57.5% 이상의 정확도가 나올 확률을 계산하면 1번 귀무가설에 대한 p-value를 구할 수 있다.

$$\begin{aligned} P(X>836) &= 1-P(X<837) \\ &= 1-F_x(836) \\ &= 4.5\times10^{-9} \end{aligned}$$

1번 귀무가설에 대한 p-value는 0에 가까운 아주 작은 값으로 귀무가설을 기각할 수 있고, 대립가설인 SVM(1M)의 예측 정확도는 통계적으로 유의미한 결과라고 할 수 있다.

2번 귀무가설에 대한 검증을 위해 예측 정확도의 평균이 중심극한정리에 의해 정규분포에 근사함을 보인다. 10번의 실험 결과(각각에 대한 확률변수를 Xn이라 함)를 더한 것은 $Y=X_1+X_2+...+X_n$ 이고, 중심 극한정리에 의해 정규분포 $N(10\mu,10\sigma^2)$에 근사하게 따른다고 할 수 있다. 평균을 위해 10으로 나누게 되면, 즉 Y에 0.1을 곱하면 정규분포 성질에 따라 평균에 대한 확률변수 0.1Y는 정규분포 $N(\mu,\frac{\sigma^2}{10})$를 근사하게 따른다.

동일한 분포에서 샘플링된 두 그룹의 평균 차이에 대한 분포는 정규분포 성질에 따라 $Y_1-Y_2\sim N(0,\frac{\sigma^2}{5})$ 라고 할 수 있다. 이 결과를 이용하여 2번 귀무가설에 해당하는 확률 p-value를 구한다.

2번 귀무가설에 따라 SVM(1M)과 ANN의 결과가 동일 분포에서 나왔다면 두 측정된 평균의 차이는 앞에서 구한 평균 차이 분포(Y_1-Y_2)에 따라 정규분포 $N(0,0.022)$에 근사한다고 볼 수 있다. 이때 표준편차 σ는 SVM(1M)과 ANN의 결과 전체에 대해 계산한 값(0.33)을 사용한다.

SVM(1M)과 ANN의 결과 차이, 즉 평균 차이는 1이다. 평균 차이 1이상이 발생할 확률은 다음과 같이 구할 수 있다.

$$P(|Y_1 - Y_2| \geq 1)$$
$$= F_{Y_1 - Y_2}(-1) + (1 - F_{Y_1 - Y_2}(1))$$
$$= 3.5 \times 10^{-22}$$

2번 귀무가설에 대한 p-value는 0에 가까운 아주 작은 값으로 귀무가설을 기각할 수 있고, 따라서 SVM(1M)의 결과가 ANN의 결과 차이가 통계적으로 유의미한 결과라고 할 수 있다. 마찬가지로 D.Tree와 Adaboost의 결과에 대한 p-value도 0에 가까운 값이 나오기 때문에 SVM(1M)의 결과가 이 방법들에 비해 우수하다고 볼 수 있다.

V. 결 론

건설기업을 대상으로 다양한 파산예측모형의 결과와 적응형 부스팅을 이용한 파산예측모형의 결과를 비교 분석해보았다. 이러한 결과는 우리에게 실무적으로 중요한 시사점을 준다. 부동산 경기침체 여파로 건설기업의 파산이 급증하고 이로 인한 경제적, 사회적 문제가 심화되고 있는 현 시점에서, 건설기업의 파산이 주주, 채권자, 피고용인 등 여러 경제주체들에게 큰 고통을 주는 만큼 건설기업에 특화된 파산예측모형의 필요성은 매우 크다고 할 수 있다.

우리가 연구결과 알 수 있었던 것은 전통적인 통계적 회귀분석 방법의 일종인 알트만의 Z-score 방법에 비해 기계학습 방법이 훨씬 더 우수한 예측력을 보였다는 점이다. 기계학습에서 종종 발생하는 과적합(overfitting) 문제가 예측력 저하로 이어지는 경우가 있다는 점을 감안할 때 우리가 사용한 기계학습 방법 모두 Z-score 방법에 비해 우수한 예측력을 보였다는 점은 특기할 만하다. 아울러 적응형 부스팅 방법이 모든 모형 중 가장 우수한 예측력을 보여주었으며, 특히 기업 규모가 큰 경우 예측력이 더욱 우수하다는 점은 적응형 부스팅의 실무적 적용에 매우 유용한 시사점을 준다고 할 수 있다.

이러한 결과의 원인으로 건설 기업의 재무 비율 변수들의 분산이 큰 이유를 꼽을 수 있다. 이로 인해 Z-score의 예측력이 낮아지고, 마찬가지로 변동성에 영향을 받는 인공 신경망이나 SVM 또한 적응형 부스팅에 비해 상대적으로 예측력이 낮았다. 이번 연구를 통해 적응형 부스팅을 이용한 파산예측모형 – 특히 건설업 부문에서 – 이 적합함을 확인하였다. 이 결과를 통해 건설기업과 관련된 주주 및 경영진, 채권자, 피고용인 등이 이를 활용하여 파산가능성을 미리 예측하고 고통을 최소화하기 위한 합리적인 의사결정을 하는데 도움이 되기를 기대한다.

두 번째로, 본 논문에서 SVM과 같은 기계학습을 통하여 재무정보를 기반으로 주식 가격의 변화를 예측하였다. 기업의 재무정보를 SVM의 입력으로 사용하여 기본적 분석을 하고 이 결과로 주식의 향후 등락을 예측하는 것이다. 기본적 분석을 위한 재무 정보로 자산과 이익에 대한 정보를 활용하였다. 재무 정보에서 자산과 이익에 대한 정보는 대표적으로 해당 기업의 재무 상태를 설명해줄 수 있는 지표이다. 이 지표에 따라 주가의 등락을 예측할 수 있는지, 그리고 그 예측이 어느 시점까지 가능한지 SVM을 사용하여 평가하였다.

실험 결과 재무 정보를 활용한 SVM의 주가 예측력은 전문가 예측에 비하여 우수한 예측력을 보여주며, 기간이 지날수록 예측력이 떨어지게 된다. 이는 재무 정보를 기반으로 한 예측이 단 기간에는 우수하지만 일정 기간 후에는 재무 정보와 주식 가격의 부조화가 합리적 투자자들에 의하여 상쇄된다고 볼 수 있다. 또한 SVM은 다른 기계학습 방법인 인공신경망이나 결정나무, 적응형부스팅에 비해 수행 속도나 예측력이 더 뛰어남을 확인하였다. 이러한 연구 결과를 바탕으로 더욱 확장된 연구가 이어진다면, 투자자들의 수익률을 높이고 경영자 및 연관 산업 실무자들의 판단에 도움을 줄 수 있는 유의미한 예측 정보를 지속적으로 생산해낼 수 있을 것이다.

참 고 문 헌

Afolabi, M. O. & Olider, O. (2007). Predicting Stock Prices Using a Hybrid Kohonen Self Organization Map. *In Proceeding of the 40th Hawaii International Conference on System Science*.

Agarwal, V. & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8), 1541-1551.

Ahn, H. & Kim, K. J. (2009). Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing*, 9(2), 599-607.

Al Shalabi, L. & Shaaban, Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. *In 2006 International Conference on Dependability of Computer Systems* (pp. 207-214). IEEE.

Albanese, D., Visintainer, R., Merler, S., Riccadonna, S., Jurman, G. & Furlanello, C. (2012). *mlpy: Machine learning python*. arXiv preprint arXiv:1202.6548.

Alfaro, E., García, N., Gámez, M. & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems*, 45(1), 110-122.

Alfarano, S., Lux, T. & Wagner, F. (2005). Estimation of agent-based models: the case of an asymmetric herding model. *Computational Economics*, 26(1), 19-49.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589-609.

Altman, E. I. (2000). Predicting financial distress of companies: revisiting the

Z-score and ZETA models. *Stern School of Business, New York University*, 9-12.

Altman, E. I., Hartzell, J. & Peck, M. (1998). Emerging market corporate bonds: a scoring system. *In Emerging Market Capital Flows* (pp. 391-400). Springer US.

Antweiler, W. & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), 1259-1294.

Aoki, S. & Hosonuma, Y. (2004). Bankruptcy Prediction Using Decision Tree. *In The Application of Econophysics* (pp. 299-302). Springer Japan.

Armano, G., Murru, A. & Roli, F. (2002). Stock market prediction by a mixture of genetic-neural experts. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(05), 501-526.

Asadi, S., Hadavandi, E., Mehmanpazir, F. & Nakhostin, M. M. (2012). Hybridization of evolutionary Levenberg-Marquardt neural networks and data pre-processing for stock market prediction. *Knowledge-Based Systems*, 35, 245-258.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, 12(4), 929-935.

Atsalakis, G. S. & Valavanis, K. P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications*, 36(7), 10696-10707.

Back, B., Laitinen, T. & Sere, K. (1996). Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications*, 11(4), 407-413.

Bagheri, A., Peyhani, H. M. & Akbari, M. (2014). Financial forecasting using ANFIS networks with quantum-behaved particle swarm optimization. *Expert Systems with Applications*, 41(14), 6235-6250.

Bharath, S. T. & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3), 1339-1369.

Bharath, S. T. & Shumway, T. (2004). Forecasting default with the KMV-Merton model. *In AFA 2006 Boston Meetings Paper*.

Beaver, W. H., McNichols, M. F. & Rhie, J. W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting studies*, 10(1), 93-122.

Bettman, J. L., Sault, S. J. & Schultz, E. L. (2009). Fundamental and technical analysis: substitutes or complements?. *Accounting & Finance*, 49(1), 21-36.

Bliss, C. A., Frank, M. R., Danforth, C. M. & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 5(5), 750-764.

Bollen, J., Mao, H. & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40(3), 229-242.

Broomhead, D. S. & Lowe, D. (1988). *Radial basis functions, multi-variable functional interpolation and adaptive networks* (No. RSRE-MEMO-4148). ROYAL SIGNALS AND RADAR ESTABLISHMENT MALVERN (UNITED KINGDOM).

Brock, W., Lakonishok, J. & LeBaron, B. (1992). Simple technical trading rules and

the stochastic properties of stock returns. *The Journal of finance*, 47(5), 1731-1764.

Bryant, S. M. (1997). A case-based reasoning approach to bankruptcy prediction modeling. *Intelligent Systems in Accounting, Finance and Management*, 6(3), 195-214.

Bryll, R., Gutierrez-Osuna, R. & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291-1302.

Burrell, P. R. & Folarin, B. O. (1997). The impact of neural networks in finance. *Neural Computing & Applications*, 6(4), 193-200.

Butler, K. C. & Malaikah, S. J. (1992). Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia. *Journal of Banking & Finance*, 16(1), 197-210.

Carty, L. V. & Lieberman, D. (1996). Corporate bond defaults and default rates 1938-1995. *Moody's Investors Service, Global Credit Research*.

CERIK(Construction Economy Research Institute of Korea) (2013). Construction Market Outlook Report 2012

Chang, C. C. & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Chang, P. C. & Fan, C. Y. (2008). A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(6), 802-815.

Chang, P. C., Wang, Y. W. & Yang, W. N. (2004). An investigation of the hybrid forecasting models for stock price variation in Taiwan. *Journal of the*

Chinese Institute of Industrial Engineers, 21(4), 358-368.

Charitou, A. & Trigeorgis, L. (2000). Option-based bankruptcy prediction.

Chauhan, N., Ravi, V. & Chandra, D. K. (2009). Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Systems with Applications*, 36(4), 7659-7665.

Chava, S. & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, 8(4), 537-569.

Chen, A. S., Leung, M. T. & Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, 30(6), 901-923.

Chen, F., Gou, C., Guo, X. & Gao, J. (2008). Prediction of stock markets by the evolutionary mix-game model. *Physica A: Statistical Mechanics and its Applications*, 387(14), 3594-3604.

Chen, H. J., Huang, S. Y. & Lin, C. S. (2009). Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach. *Expert Systems with Applications*, 36(4), 7710-7720.

Chen, S. H. & Liao, C. C. (2005). Agent-based computational modeling of the stock price-volume relation. *Information Sciences*, 170(1), 75-100.

Chen, W. S. & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert systems with applications*, 36(2), 4075-4086.

Cheng, C. H., Wei, L. Y. & Chen, Y. S. (2009). Fusion ANFIS models based on multi-stock volatility causality for TAIEX forecasting. *Neurocomputing*, 72(16), 3462-3468.

Choi, H. & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88(s1), 2-9.

Chaudhuri, A. & De, K. (2011). Fuzzy support vector machine for bankruptcy prediction. *Applied Soft Computing*, 11(2), 2472-2486.

Cho, S., Hong, H. & Ha, B. C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. *Expert Systems with Applications*, 37(4), 3482-3488.

De Choudhury, M., Sundaram, H., John, A. & Seligmann, D. D. (2008). Can blog communication dynamics be correlated with stock market activity?. *In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (pp. 55-60). ACM.

Chung, C. H. & Kim, S. K. (2010). An Investigation on the Stock Return Predictability of Dividend Yield and Earning-Price Ratio. *The Korean Journal of Financial Engineering*, 9(3), 61-87.

Cielen, A., Peeters, L. & Vanhoof, K. (2004). Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, 154(2), 526-532.

Conrad, J. & Kaul, G. (1989). Mean reversion in short-horizon expected returns, *Review of Financial Studies*, 2(2), 225-240.

Das, S. R., Hanouna, P. & Sarin, A. (2009). Accounting-based versus market-based cross-sectional models of CDS spreads. *Journal of Banking & Finance*, 33(4), 719-730.

De A. Araújo, R. (2010). A quantum-inspired evolutionary hybrid intelligent approach for stock market prediction. *International Journal of Intelligent Computing and Cybernetics*, 3(1), 24-54.

De Andrés, J., Landajo, M. & Lorca, P. (2012). Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios.

Knowledge-Based Systems, 30, 67-77.

Dey, S., Kumar, Y., Saha, S. & Basak, S. Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting.

Drucker, H., Schapire, R. & Simard, P. (1993). Improving performance in neural networks using a boosting algorithm. *Advances in neural information processing systems*, 42-42.

Etemadi, H., Rostamy, A. A. A. & Dehkordi, H. F. (2009). A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Systems with Applications*, 36(2), 3199-3207.

Fama, E. F. (1991). Efficient capital markets: II. *The journal of finance*, 46(5), 1575-1617.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, 38(1), 34-105.

Fama, E. F., Fisher, L., Jensen, M. C. & Roll, R. (1969). The adjustment of stock prices to new information. *International economic review*, 10(1), 1-21.

Feng, L., Li, B., Podobnik, B., Preis, T. & Stanley, H. E. (2012). Linking agent-based models and stochastic models of financial markets. *Proceedings of the National Academy of Sciences*, 109(22), 8388-8393.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2), 368-378.

Patrick, P. (1932). A comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant*, 2, 598-605.

Freund, Y. (1995). Boosting a weak algorithm by majority. *Information and Computation*, 121(2), 256-285.

Freund, Y. & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *In European conference on*

computational learning theory (pp. 23-37). Springer Berlin Heidelberg.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Fujiwara, Y. (2004). Zipf law in firms bankruptcy. *Physica A: Statistical Mechanics and its Applications*, 337(1), 219-230.

Fung, G. P. C., Yu, J. X. & Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. *In Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering*, (pp. 395-402). IEEE.

Fung, G. P. C., Yu, J. X. & Lam, W. (2002). News sensitive stock trend prediction. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 481-493). Springer Berlin Heidelberg.

Garcia-Almanza, A. L., Alexandrova-Kabadjova, B. & Martinez-Jaramillo, S. (2013). Bankruptcy Prediction for Banks: An Artificial Intelligence Approach to Improve Understandability. *In Artificial Intelligence, Evolutionary Computing and Metaheuristics* (pp. 633-656). Springer Berlin Heidelberg.

Gilbert, E. & Karahalios, K. (2010). Widespread Worry and the Stock Market. *In ICWSM* (pp. 59-65).

Grice, J. S. & Ingram, R. W. (2001). Tests of the generalizability of Altman's bankruptcy prediction model. *Journal of Business Research*, 54(1), 53-61.

Grosan, C., Abraham, A., Ramos, V. & Han, S. Y. (2005). Stock market prediction using multi expression programming. *In 2005 portuguese conference on artificial intelligence* (pp. 73-78). IEEE.

Gruhl, D., Guha, R., Kumar, R., Novak, J. & Tomkins, A. (2005). The predictive power of online chatter. *In Proceedings of the eleventh ACM SIGKDD*

international conference on Knowledge discovery in data mining (pp. 78-87). ACM.

Gupta, A. & Dhingra, B. (2012). Stock market prediction using hidden Markov models. *In 2012 Students Conference on Engineering and Systems (SCES)*, (pp. 1-4). IEEE.

Hadavandi, E., Shavandi, H. & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8), 800-808.

Hamilton, W. P. (1922). *The Stock Market Barometer; a Study of Its Forecast Value Based on Charles H. Dow's Theory of the Price Movement*. Harper & Bros..

Han, S. & Chen, R. C. (2007). Using svm with financial statement analysis for prediction of stocks. *Communications of the IIMA*, 7(4), 63.

HassHassan, M. R. & Nath, B. (2005). Stock market forecasting using hidden Markov model: a new approach. *In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)* (pp. 192-196). IEEE.

Hassan, M. R., Nath, B. & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. *Expert systems with Applications*, 33(1), 171-180.

Hawawini, G. & Keim, D. B. (1995). On the predictability of common stock returns: World-wide evidence. *Handbooks in operations research and management science*, 9, 497-544.

Heo, J. & Yang, J. Y. (2014). Bankruptcy Forecasting Model using AdaBoost: A Focus on Construction Companies. *Journal of Intelligence and Information Systems*, 20(1), 35-48.

Heo, J. & Yang, J. Y. (2015). SVM based Stock Price Forecasting Using Financial Statements. *KIISE Transactions on Computing Practices*, 21(3), 167-172.

Hillegeist, S. A., Keating, E. K., Cram, D. P. & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of accounting studies*, 9(1), 5-34.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 832-844.

Hong, B. H., Lee, K. E. & Lee, J. W. (2007). Power law in firms bankruptcy. *Physics Letters A*, 361(1), 6-8.

Horta, I. M., Camanho, A. S. & da Costa, J. M. (2012). Performance assessment of construction companies: A study of factors promoting financial soundness and innovation in the industry. *International Journal of Production Economics*, 137(1), 84-93.

Hsieh, T. J., Hsiao, H. F. & Yeh, W. C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied soft computing*, 11(2), 2510-2525.

Huang, K. Y. & Jane, C. J. (2009). A hybrid model for stock market forecasting and portfolio selection based on ARX, grey system and RS theories. *Expert systems with applications*, 36(3), 5387-5392.

Huang, S. C. (2011). Forecasting stock indices with wavelet domain kernel partial least square regressions. *Applied Soft Computing*, 11(8), 5433-5443.

Hung, C. & Chen, J. H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert systems with applications*, 36(3), 5297-5303.

Jo, H., Han, I. & Lee, H. (1997). Bankruptcy prediction using case-based

reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, 13(2), 97-108.

Jolai, F. & Ghanbari, A. (2010). Integrating data transformation techniques with Hopfield neural networks for solving travelling salesman problem. *Expert Systems with Applications*, 37(7), 5331-5335.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263-291.

Karaboga, D. & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of global optimization*, 39(3), 459-471.

Kavussanos, M. G. & Dockery, E. (2001). A multivariate test for stock market efficiency: the case of ASE. *Applied Financial Economics*, 11(5), 573-579.

Kayacan, E., Ulutas, B. & Kaynak, O. (2010). Grey system theory-based models in time series prediction. *Expert systems with applications*, 37(2), 1784-1789.

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M. & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied soft computing*, 13(2), 947-958.

Kennedy, J. & Eberhart, R. (1995). Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks. IV*, pp. 1942 - 1948.

Kim, H. J. & Shin, K. S. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Applied Soft Computing*, 7(2), 569-576.

Kim, H. J. & Cho, H. G. (2010). Developing Stock Pattern Searching System using Sequence Alignment Algorithm. *Journal of KIISE: Computer Systems and Theory*, 37(6), 354-367.

Kim, K. Y. & Kim, Y. B. (2004). Testing the Predictability of Stock Return in the

Korean Stock Market. *Korean Journal of Industrial Economic*, 17(4), 1255-1271.

Kim, K. J. & Lee, W. B. (2004). Stock market prediction using artificial neural networks with optimal feature transformation. *Neural computing & applications*, 13(3), 255-260.

Kim, M. J. (2009). Ensemble Learning for Solving Data Imbalance in Bankruptcy Prediction. *Journal of Intelligence and Information Systems*, 15(3), 1-15.

Kim, M. J. & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications*, 37(4), 3373-3379.

Kumar, P. R. & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques－A review. *European journal of operational research*, 180(1), 1-28.

Huarng, K. & Yu, T. H. K. (2006). The application of neural networks to forecast fuzzy time series. *Physica A: Statistical Mechanics and its Applications*, 363(2), 481-491.

Kuo, R. J., Lee, L. C. & Lee, C. F. (1996). Integration of artificial neural networks and fuzzy delphi for stock market forecasting. *In IEEE International Conference on Systems, Man, and Cybernetics,* (Vol. 2, pp. 1073-1078). IEEE.

Lai, R. K., Fan, C. Y., Huang, W. H. & Chang, P. C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, 36(2), 3761-3773.

Laitinen, E. K. & Laitinen, T. (2001). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International review of financial analysis*, 9(4), 327-349.

Lam, M. (2004). Neural network techniques for financial performance prediction:

integrating fundamental and technical analysis. *Decision support systems*, 37(4), 567-581.

LeBaron, B. (1998). Technical trading rules and regime shifts in foreign exchange. *Advanced trading rules*, 5-40.

Lee, K. C., Han, I. & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, 18(1), 63-72.

Lee, R. S. (2006). iJADE Stock Advisor—An Intelligent Agent-Based Stock Prediction System Using the Hybrid RBF Recurrent Network. *Fuzzy-Neuro Approach to Agent Applications: From the AI Perspective to Modern Ontology*, 231-253.

Lee, S. & Choi, W. S. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications*, 40(8), 2941-2946.

Leng, G., Prasad, G. & McGinnity, T. M. (2004). An on-line algorithm for creating self-organizing fuzzy neural networks. *Neural Networks*, 17(10), 1477-1493.

Leng, G., Zeng, X. J. & Keane, J. A. (2009). A hybrid learning algorithm with a similarity-based pruning strategy for self-adaptive neuro-fuzzy systems. *Applied Soft Computing*, 9(4), 1354-1366.

Leshno, M. & Spector, Y. (1996). Neural network prediction analysis: The bankruptcy case. *Neurocomputing*, 10(2), 125-147.

Li, H. & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples－Evidence from the Chinese hotel industry. *Tourism Management*, 33(3), 622-634.

Li, M. Y. L. & Miu, P. (2010). A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based

information: A binary quantile regression approach. *Journal of Empirical Finance*, 17(4), 818-833.

Lim, K. P. & Brooks, R. (2011). The evolution of stock market efficiency over time: a survey of the empirical literature. *Journal of Economic Surveys*, 25(1), 69-108.

Lin, X., Yang, Z. & Song, Y. (2009). Short-term stock price prediction based on echo state networks. *Expert systems with applications*, 36(3), 7313-7317.

Lindsay, D. H. & Campbell, A. (1996). A chaos approach to bankruptcy prediction. *Journal of Applied Business Research*, 12(4), 1.

Liu, C. L. & Marukawa, K. (2004, October). Normalization ensemble for handwritten character recognition. *In Ninth International Workshop on Frontiers in Handwriting Recognition*, (pp. 69-74). IEEE.

Liu, Y., Huang, X., An, A. & Yu, X. (2007, July). ARSA: a sentiment-aware model for predicting sales performance using blogs. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 607-614). ACM.

Lu, N., Zhou, J., He, Y. & Liu, Y. (2009). Particle swarm optimization for parameter optimization of support vector machine model. *In Second International Conference on Intelligent Computation Technology and Automation*, (Vol. 1, pp. 283-286). IEEE.

Majhi, R., Panda, G., Majhi, B. & Sahoo, G. (2009). Efficient prediction of stock market indices using adaptive bacterial foraging optimization (ABFO) and BFO based techniques. *Expert Systems with Applications*, 36(6), 10097-10104.

Majhi, R., Panda, G. & Sahoo, G. (2009). Development and performance evaluation of FLANN based model for forecasting of stock markets.

Expert Systems with Applications, 36(3), 6800-6808.

Lo, A. W. (2007). Efficient Markets Hypothesis. The New Palgrave Dictionary of Economics.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, 17(1), 59-82.

Mao, H., Counts, S. & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*.

Mazzatorta, P., Benfenati, E., Neagu, D. & Gini, G. (2002). The importance of scaling in data mining for toxicity prediction. *Journal of chemical information and computer sciences*, 42(5), 1250-1255.

Mckee, T. E. (2000). Developing a bankruptcy prediction model via rough sets theory. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 9(3), 159-173.

McKee, T. E. (2003). Rough sets bankruptcy prediction models versus auditor signalling rates. *Journal of Forecasting*, 22(8), 569-586.

Min, J. H. & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, 28(4), 603-614.

Min, S. H., Lee, J. & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert systems with applications*, 31(3), 652-660.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. *In Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, (pp. 10-pp). IEEE.

Mustaffa, Z. & Yusof, Y. (2011). Optimizing LSSVM using ABC for non-volatile

financial prediction. *Australian Journal of Basic and Applied Sciences*, 5(11), 549-556.

Mustaffa, Z., Yusof, Y. & Kamaruddin, S. S. (2014). Enhanced artificial bee colony for training least squares support vector machines in commodity price forecasting. *Journal of Computational Science*, 5(2), 196-205.

Nanni, L. & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, 36(2), 3028-3033.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347-370.

NeuroLab (2011), https://pythonhosted.org/neurolab/

Ni, L. P., Ni, Z. W. & Gao, Y. Z. (2011). Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications*, 38(5), 5569-5576.

NICE, Credit Information Service, (2013). Available at http://www.nicednb.com

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3), 144-160.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129), 1-2.

Odom, M. D. & Sharda, R. (1990). A neural network model for bankruptcy prediction. *In 1990 IJCNN International Joint Conference on neural networks* (pp. 163-168).

Olson, D. & Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting*, 19(3), 453-465.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

Pai, P. F. & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, 33(6), 497-505.

Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In LREc* (Vol. 10, pp. 1320-1326).

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1-135.

Park, C. S. & Han, I. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications*, 23(3), 255-264.

Phua, P. K. H., Zhu, X. & Koh, C. H. (2003). Forecasting stock index increments using neural networks with trust region methods. *In Proceedings of the International Joint Conference on Neural Networks*, (Vol. 1, pp. 260-265). IEEE.

Prechter, R. R. (1999). *The wave principle of human social behavior and the new science of socionomics* (Vol. 1). New Classics Library.

Preis, T., Moat, H. S. & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, 3.

Qian, B. & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1), 25-33.

Quah, T. S. & Srinivasan, B. (1999). Improving returns on stock investment through neural network selection. *Expert Systems with Applications*, 17(4), 295-301.

Ravi, V. & Pramodh, C. (2008). Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy

prediction in banks. *Applied Soft Computing*, 8(4), 1539-1548.

Reisz, A. S. & Perlich, C. (2007). A market-based framework for bankruptcy prediction. *Journal of Financial Stability*, 3(2), 85-131.

Rendleman, R. J., Jones, C. P. & Latane, H. A. (1982). Empirical anomalies based on unexpected earnings and the importance of risk adjustments. *Journal of Financial Economics*, 10(3), 269-287.

Roh, T. H. (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications*, 33(4), 916-922.

Romahi, Y. & Shen, Q. (2000). Dynamic financial forecasting with automatically induced fuzzy associations. *In The Ninth IEEE International Conference on Fuzzy Systems*, (Vol. 1, pp. 493-498). IEEE.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2), 197-227.

Schumaker, R. & Chen, H. (2006). Textual analysis of stock market prediction using financial news articles. *AMCIS 2006 Proceedings*, 185.

Schumaker, R. P. & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2), 12.

Schumaker, R. P. & Chen, H. (2010). A discrete stock price prediction engine based on financial news. *Computer*, 43(1), 51-56.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101-124.

Scikit learn (2010). http://scikit-learn.org/

Sewell, M. (2011). History of the efficient market hypothesis. *RN*, 11(04), 04.

Shen, W., Zhang, Y. & Ma, X. (2009). Stock return forecast with LS-SVM and particle swarm optimization. *In International Conference on Business*

Intelligence and Financial Engineering, (pp. 143-147). IEEE.

Shin, K. S. & Lee, Y. J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, 23(3), 321-328.

Shin, K. S., Lee, T. S. & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1), 127-135.

Shin, T. S. & Hong, T. H. (2011). Corporate Credit Rating based on Bankruptcy Probability Using AdaBoost Algorithm-based Support Vector Machine. *Journal of Intelligence and Information Systems*, 17(3), 25-41.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101-124.

Solomon, S. (1999). Generalized Lotka-Volterra (GLV) models and generic emergence of scaling laws in stock markets. *arXiv preprint cond-mat/9901250*.

Song, D. S. (2011). A Study on the Relation Between the Financial Ratio and Earnings Quality. *Korea International Accounting Review*, 40(2011), 135-156

Standard & Poor's (1995). Corporate Defaults Level Off in 1994. CreditWeek

Sun, J., Liao, B. & Li, H. (2013). AdaBoost and bagging ensemble approaches with neural network as base learner for financial distress prediction of Chinese construction and real estate companies. *Recent Patents on Computer Science*, 6(1), 47-59.

Sun, L. & Shenoy, P. P. (2007). Using Bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, 180(2), 738-753.

Sun, X., Chen, H., Yuan, Y. & Wu, Z. (2001). Predictability of multifractal

analysis of Hang Seng stock index in Hong Kong. *Physica A: Statistical Mechanics and its Applications*, 301(1), 473-482.

Sung, T. K., Chang, N. & Lee, G. (1999). Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. *Journal of Management Information Systems*, 16(1), 63-85.

SwalesGeorge Jr, S. (1992). Applying artificial neural networks to investment analysis. *Financial Analysts Journal*, 48(5).

Tai, Q. Y. & Shin, K. S. (2010). GA-based Normalization Approach in Back-propagation Neural Network for Bankruptcy Prediction Modeling. *Journal of Intelligence and Information Systems*, 16(3), 1-14.

Tam, K. Y. (1991). Neural network models and the prediction of bank bankruptcy. *Omega*, 19(5), 429-445.

Timmermann, A. & Granger, C. W. (2004). Efficient market hypothesis and forecasting. *International Journal of forecasting*, 20(1), 15-27.

Tsai, C. F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22(2), 120-127.

Tsai, C. F. & Wang, S. P. (2009). Stock price forecasting by hybrid machine learning techniques. *In Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, No. 755, p. 60).

Tsai, C. F. & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34(4), 2639-2649.

Tsai, C. F. & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, 50(1), 258-269.

Tsakonas, A., Dounias, G., Doumpos, M. & Zopounidis, C. (2006). Bankruptcy

prediction with neural logic networks by means of grammar-guided genetic programming. *Expert Systems with Applications*, 30(3), 449-461.

Tseng, F. M. & Hu, Y. C. (2010). Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems with Applications*, 37(3), 1846-1853.

Tserng, H. P., Lin, G. F., Tsai, L. K. & Chen, P. C. (2011). An enforced support vector machine model for construction contractor default prediction. *Automation in Construction*, 20(8), 1242-1249.

Tumarkin, R. & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3), 41-51.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Verikas, A., Kalsyte, Z., Bacauskiene, M. & Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft Computing*, 14(9), 995-1010.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71-111.

Wang, H. & Zhang, J. (2009, June). Analysis of Different Data Standardization Forms for Fuzzy Clustering Evaluation Results' Influence. *In 2009 3rd International Conference on Bioinformatics and Biomedical Engineering* (pp. 1-4). IEEE.

Wang, J. L. & Chan, S. H. (2006). Stock market trading rule discovery using two-layer bias decision tree. *Expert Systems with Applications*, 30(4), 605-611.

Wang, Y. H. (2009). Nonlinear neural network forecasting model for stock index option price: Hybrid GJR-GARCH approach. *Expert Systems with*

Applications, 36(1), 564-570.

Wang, Y. F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications*, 22(1), 33-38.

Wang, Z. (2004). *Prediction of stock market prices using neural network techniques* (Doctoral dissertation, University of Ottawa (Canada)).

Wei, Y. & Huang, D. (2005). Multifractal analysis of SSEC in Chinese stock market: A different empirical result from Heng Seng index. *Physica A: Statistical Mechanics and its Applications*, 355(2), 497-508.

Wen, Q., Yang, Z., Song, Y. & Jia, P. (2010). Automatic stock decision support system based on box theory and SVM algorithm. *Expert Systems with Applications*, 37(2), 1015-1022.

Whaley, R. E. (2000). The investor fear gauge. *The Journal of Portfolio Management*, 26(3), 12-17.

Wilson, R. L. & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision support systems*, 11(5), 545-557.

Wu, D., Fung, G. P. C., Yu, J. X. & Pan, Q. (2009). Stock prediction: an event-driven approach based on bursty keywords. *Frontiers of Computer Science in China*, 3(2), 145-157.

Wu, M. C., Lin, S. Y. & Lin, C. H. (2006). An effective application of decision tree to stock trading. *Expert Systems with Applications*, 31(2), 270-274.

Yang, J. & Counts, S. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. *ICWSM*, 10, 355-358.

Yang, Z. R., Platt, M. B. & Platt, H. D. (1999). Probabilistic neural networks in bankruptcy prediction. *Journal of Business Research*, 44(2), 67-74.

Yu, H. K. (2005). Weighted fuzzy time series models for TAIEX forecasting. *Physica A: Statistical Mechanics and its Applications*, 349(3), 609-624.

Yu, L., Wang, S. & Lai, K. K. (2005, December). Mining stock market tendency using GA-based support vector machines. *In International Workshop on Internet and Network Economics* (pp. 336-345). Springer Berlin Heidelberg.

Zhang, X., Fuehres, H. & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, 26, 55-62.

Zhiqiang, G., Huaiqing, W. & Quan, L. (2013). Financial time series forecasting using LPP and SVM optimized by PSO. *Soft Computing*, 17(5), 805-818.

Zhu, X., Wang, H., Xu, L. & Li, H. (2008). Predicting stock index increments by neural networks: The role of trading volume under different horizons. *Expert Systems with Applications*, 34(4), 3043-3054.

Zięba, M., Tomczak, S. K. & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 58, 93-101.

ABSTRACT

Machine-learning-based Management Prediction Systems using the Corporate Financial Information

Yang, Jin-Yong
Major in Computer Engineering
Dept. of Information and Computer Engineering
The Graduate School
Hansung University

In the increasingly complex and sophisticated corporate management environment, proper management forecasting becomes more important than ever in the survival and development of companies. Machine learning techniques are developed to classify and analyze the pouring data appropriately and to derive and predict valuable information. This paper proposes business forecasting applications such as corporate bankruptcy forecasting and stock price forecasting by machine learning approaches using financial information and stock price data.

Recently, bankruptcy prediction research based on machine learning is active. Pattern recognition, which is a typical application field of machine learning, is applied to bankruptcy prediction. It creates a pattern based on the company's financial information and determines whether the pattern belongs to the bankruptcy risk group or not. Previous studies using traditional Z-score and machine learning techniques do not take industry-specific characteristics into account because they are targeted at general firms rather than specific industries. The construction

industry is a capital-intensive industry characterized by long business cycles, large-scale investments, and long payback periods. This may lead to different capital structures from other industries, and it may be difficult to apply the same criteria as those used to judge the corporate financial risk of other industries. This paper classifies the construction companies into three classes according to their size and compares the predictive power of each technique. Experimental results show that adaptive boosting provides better results than other techniques, especially for companies with capital stock of over 50 billion.

Stock market forecasting has attracted much attention in academia and business as an interesting and challenging research theme in financial time series forecasting. Efficient forecasting of the stock market is a very important issue for investors. Accurate forecasting algorithms can bring high profits and loss avoidance to investors. Forecasting stock market with traditional time-series analysis has proven to be difficult. Machine learning is emerging as an alternative. It is a technique that computers make classification and prediction by learning. SVM is a fast and reliable machine learning method widely used for classification and prediction. This paper examined the predictive power of SVM based on financial information. This can be used to assess how effective the financial information, which represents the intrinsic value of the company, is in estimating the stock price. Corporate financial statements are used as input into SVM, to predict whether the stock price will rise or not. The results have been compared with those obtained by experts' scores and machine learning methods such as artificial neural networks, decision trees, and adaptive boosting. As a result, the performance of SVM was superior to other techniques in terms of execution time and predictability.

KEYWORD : machine learning, forecasting system, bankruptcy prediction, stock price prediction

감사의 글

먼저 공학도의 길로 이끌어 주시고 이 논문이 나올 때 까지 지도해 주신 허준영 교수님께 진심으로 감사의 마음을 전합니다. 제게 컴퓨터 공학은 신세계를 발견한 어린 아이처럼 설레고 기쁜 영감의 원천이었습니다. 사방에 아름다운 꽃이 피어있는 길을 허준영 교수님과 함께 걸었던 경험은 앞으로의 제 연구 인생에 귀중한 자산이 될 것입니다.

아울러 자상하고 세심한 지도를 아끼지 않으신 김남윤 교수님, 최상영 교수님께도 이 자리를 빌려 깊은 감사를 드립니다. 귀한 시간을 쪼개어 논문을 심사해 주신 심사위원 선생님들 한분 한분께도 깊이 감사드립니다. 같이 공부했던 동료 학생들의 열정은 제게 매우 인상적이고 학업을 수행하는데 큰 자극이 되어 주었습니다. 이 분들께도 감사의 마음을 전합니다.

항상 변함없는 아내 최호세피나 박사와 사랑하는 아이들 수인, 민혁, 석준, 민서, 민준은 제 삶에서 가장 소중한 존재들입니다.

무엇보다도 전 생애에 걸쳐 모든 사랑을 다 주시고 가신 석정 양회연 선생과 김초연 선생님 두 분께 이 논문을 바칩니다. 두 분께서 가르쳐 주신 진리와 자유를 두 분 곁으로 가는 날 까지 깊이 간직하겠습니다.

"너희가 진리를 알게 될 것이니 진리가 너희를 자유롭게 할 것이다."

2016년 12월

양 진 용 드림