Information Systems Review
Vol. 18, No. 2 June 2016 http://dx.doi.org/10.14329/isr.2016.18.2.173

데이터마이닝 기법을 이용한 기업부실화 예측 모델 개발과 예측 성능 향상에 관한 연구

Development of Prediction Model of Financial Distress and Improvement of Prediction Performance Using Data Mining Techniques

김 량 형 (Raynghyung Kim) 국립한밭대학교 경영학과 박사과정
유 동 희 (Donghee Yoo) 경상대학교 경영정보학과 조교수
김 건 우 (Gunwoo Kim) 국립한밭대학교 경영학과 부교수, 교신저자

요 약

본 연구의 목적은 비즈니스 인텔리전스 연구 관점에서 기업부실화 예측 성능을 향상키시는 것이다. 이를 위해 본 연구는 기존 연구들에서 미흡하게 다루어졌던 1) 데이터셋을 구성하는 과정에서 발생하는 바이어스 문제, 2) 거시경제위험 요소의 미반영 문제, 3) 데이터 불균형 문제, 4) 서술적 바이어스 문제를 다루어 경기순환국면을 반영한 기업부실화 예측 프레임워크를 제안하고, 이를 바탕으로 기업부실화 예측 모델을 개발하였다.

본 연구에서는 경기순환국면별로 각각의 데이터셋을 구성하고, 각 데이터셋에서 의사결정나무, 인공신경망 등 단일 분류기부터 앙상블 기법까지 다양한 데이터마이닝 알고리즘을 적용하여 실험하였다. 또한 본 연구는 데이터불균형 문제를 해결하기 위해, 오버샘플링 기법인 SMOTE(synthetic minority over-sampling technique) 기법을 통해 초기 데이터 불균형 상태에서부터 표본비율을 1:1까지 변화시켜 가며, 기업부실화 예측 모델을 개발하는 실험을 하였고, 예측 모델의 변수 선정 시에 선행연구를 바탕으로 재무비율을 추출하고, 여기서 파생된 IT 산출물인 재무상태변동성과 산업수준상태변동성을 예측 모델에 삽입하였다. 마지막으로, 본 연구는 각 순환국면에서 만들어진 기업부실화 예측 모델의 예측 성능 비교와 경기 확장기와 수축기에서의 기업부실화 예측 모델의 유용성에 대해 논의하였다.

본 연구는 비즈니스 인텔리전스 연구 측면에서 기존 연구에서 미흡하게 다루어졌던 4가지 문제점을 검토하고, 이를 해결할 프레임워크를 제안함으로써 기존 연구 대비 기업부실화 예측률을 10% 이상 향상시켰다는 점에서 연구의 의의를 찾을 수 있다.

키워드 : 기업부실화, 경기순환, 데이터마이닝, 불균형 데이터, 서술적 바이어스, IT 산출물

Ⅰ. 서 론

기업부실화(financial distress)는 기업이 재무적 의무(financial obligation)를 다하지 못하는 상태, 즉 계약상의 채무를 상환할 수 없는 상태에서부터 법률적 파산에 이르는 전 과정을 포함하는 개념이다(Beaver, 1966; Zmijewski, 1984). 기업부실화는 주주, 채권자 등 기업의 여러 이해관계자들에게 막대한 경제적 손실을 입힐 뿐만 아니라, 특히 경제위기 때마다 부실기업 구제를 위한 정부의 공적자금 투입 사례를 보면, 거시경제 측면에 있어서도 큰 비용을 초래한다는 점에서 매우 중요한 이슈이다(김성규, 이화득, 2012; Abbasi *et al*., 2012; Suarez and Sussman, 2007).

기업부실화와 관련된 선행연구들을 살펴보면, 주로 기업의 재무데이터를 활용하여 기업부실화를 예측하고 있다. 주요 국외 선행연구에서 제시한 기업부실화 예측 모델의 예측률은 평균 80~90% 대를 기록하고 있다(Abbasi *et al*., 2012; Altman, 1968; Chuang, 2013; Sánchez-Lasheras *et al*., 2012; Tsakonas *et al*., 2006; Wang *et al*., 2014). 국내 선행연구의 경우 대부분 국외 선행연구를 기반으로 진행되었고, 보다 진보된 알고리즘을 활용하여 기업부실화를 예측하고 있다. 그럼에도 불구하고, 국내 선행연구에서 제시된 기업부실화 예측 모델의 예측률은 70~80% 대로 국외 선행연구에 비해 저조한 예측률을 보이고 있다(김나라 등, 2013, 김명종, 강대기, 2009; 민성환, 2012; 배재권, 2010; 신태수, 홍태호, 2011; 최소윤, 안현철, 2015).

이러한 현상에 대한 논의는 여러 각도에서 접근해 볼 수 있겠으나, 본 연구는 기업부실화 예측을 위한 방법론적 측면에서, 특히 기존 연구에서 미흡하게 다루어졌던 1) 데이터셋을 구성하는 과정에서 발생하는 바이어스(bias) 문제, 2) 거시경제위험(systematic risk) 요소의 미반영 문제, 3) 데이터 불균형(imbalanced data) 문제, 4) 서술적 바이어스(declarative bias) 문제를 위 현상의 원인으로 인식하고, 이 문제를 해결할 방안을 제안하고자 한다. 본 연구에서는 경기순환국면을 반영한 기업부실화 예측 프레임워크를 해결방안으로 제안하고, 이를 바탕으로, 기존의 연구보다 발전된 기업부실화 예측 모델을 개발하고자 한다. 또한 각 순환국면에서 만들어진 기업부실화 예측 모델의 예측 성능을 비교하고, 경기 확장기와 수축기에서의 기업부실화 예측 모델의 유용성을 논의하고자 한다. 이를 통해, 본 연구는 비즈니스 인텔리전스 연구 측면에서 기존 연구에서 미흡하게 다루어졌던 4가지 문제점을 검토하고, 기업부실화 예측 성능을 대폭 향상시키는 프레임워크를 제안하는 점에서 연구의 의의를 찾고자 한다.

본 연구의 구성은 다음과 같다. 제Ⅱ장에서는 선행연구를 검토하고, 선행연구의 한계점에 대해 논의한다. 또한 논의를 바탕으로 도출된 해결방안과 경기순환국면을 반영한 기업부실화 예측 프레임워크를 살펴본다. 제Ⅲ장에서는 본 연구에서 제안한 프레임워크를 기반으로 한 연구방법을 설명한다. 제Ⅳ장에서는 실험 및 연구결과를 제시한다. 마지막으로 제Ⅴ장에서는 본 연구를 요약하고, 연구의 한계점과 향후 연구에 대해 언급한다.

Ⅱ. 선행연구 검토

2.1 기업부실화 예측 연구의 흐름

전통적으로 학계와 산업계에서는 기업의 재무데이터를 활용하여 다양한 방법을 통해 기업부실화 예측 모델을 개발해 왔다. <표 1>에 제시된 기업부실화 예측 모델에 관한 주요 선행연구들을 살펴보면, 재무비율의 평균차이를 분석한 Beaver (1966)의 단변량 분석(univariate analysis), Altman (1968)의 판별분석(multivariate discriminant analysis), Ohlson(1980)의 로지스틱 분석(logistic analysis), Zmijewski(1984)의 프로빗 분석(probit analysis) 등이 대표적이다. 이러한 선행연구들의 공통적 특징은 정규분포와 확률모형에 기반하여 모수를 추정하는 통계적 접근 방법을 사용한다는 것이다.

〈표 1〉 통계적 접근 방법 선행연구

연구	변수	분석방법	데이터셋	예측률
Beaver(1966)	6개	단변량 분석	부실: 79 건전: 79	78.00~87.00%
Altman(1968)	5개	판별분석	부실: 33 건전: 33	74.00~95.00%
Ohlson(1980)	9개	로지스틱	부실: 105 건전: 2058	92.84~96.12%
Zmijewski(1984)	5개	프로빗 분석	부실: 129 건전: 1681	99.20~99.70%

통계적 접근 방법은 입력변수에 대한 엄격한 통계적 제약조건이 충족되어야만 하고 이를 만족시킬 경우 뛰어난 예측률을 얻을 수 있지만, 그렇지 않는 경우에는 저조한 예측률을 보이는 한계점이 존재한다(Fedorova *et al*., 2013; Wang *et al*., 2014).

이와 관련하여, Fedorova *et al*.(2013)은 기존 연구에서 제안한 기업부실화 예측 모델에 유가상장된 러시아 기업의 재무데이터(표본 크기: 888개)를 적용하여 기업부실화를 예측하는 연구를 진행하였다. 그 결과 기존 연구의 예측률 보다 낮은 70~80% 대의 예측률(<표 2> 참조)을 기록하며, 통계적 접근 방법이 가진 한계점을 지적하였다. 통계적 접근 방법에서는 자료, 기간, 표본의 크기가 달라지면 예측 모델의 강건성 또는 확장성에 문제가 발생하며, 결국 예측 성능에 부(-)의 영향을 미치는 원인이 된다. 이는 정규분포의 변화, 등분산 조건의 불충족, 큰 표준편차(또는 분산)의 발생 등과 같은 이유로 나타난다(Hsieh, 2005; Tsai and Hsu, 2013).

이러한 통계적 접근 방법의 한계점을 극복하기 위해, 1980년대 후반부터 Artificial neural network, Decision tree, Bayesian network, SVM(Support Vector Machine) 등과 같은 데이터마이닝 기법이 기업부실화 예측 연구에 활용되고 있다(안현철, 2014). 최근 데이터마이닝 분야에서는 하나의 분류 알고리즘을 사용하는 것에 더하여 앙상블 기법 및 혼합 기법을 활용하여 예측률을 높이는 연구가 진행되고 있다(Abbasi *et al*., 2012; Hsieh, 2005; McKee and Lensberg, 2002; Sánchez-Lasheras *et al*., 2012; Tsai and Hsu, 2013; Tsakonas *et al*., 2006; Wang *et al*., 2014). <표 3>과 <표 4>는 데이터마이닝 기법을 활용한 국외 및 국내의 선행연구들을 보여준다. 국외 선행연구에서는 85.2~96.9%의 예측률이 제시되고 있으며, 국내 선행연구에서는 71.0~80.9%의 예측률이 제시되고 있다.

〈표 2〉 Fedorova *et al*.(2013)의 연구 결과

(단위: %)

Model	Overall Accuracy	Precision	Sensitivity	Specificity	F-measure
Altman's model	77.5	71.2	92.3	62.6	80.4
Fulmer's model	82.0	85.0	77.7	86.3	81.2
Springate's model	77.2	70.7	93.2	61.3	80.4
Taffler's model	73.9	66.7	95.5	52.3	78.5
Zmijewski's model	78.9	72.4	93.7	64.2	81.6

자료: Fedorova *et al*.(2013).

〈표 3〉 데이터마이닝 기법을 활용한 국외 선행연구

연구	변수	분석방법	데이터셋	예측률
Tsakonas *et al.*(2006)	12개	혼합 기법	부실: 118 건전: 118	93.30%
Sánchez-Lasheras *et al.*(2012)	7개	혼합 기법	부실: 204 건전: 50,485	85.22%
Abbasi *et al.*(2012)	12개	Staking	부실: 815 건전: 8,191	93.10%
Chuang(2013)	26개	혼합 사례 기반추론	부실: 42 건전: 279	92.20~96.90%
Wang *et al.*(2014)	24개	앙상블 기법	부실: 66 건전: 66	86.79%

〈표 4〉 데이터마이닝 기법을 활용한 국내 선행연구

연구	변수	분석방법	데이터셋	예측률
김명종, 강대기(2009)	7개	Boosting 인공신경망	부실: 729 건전: 729	71.02~75.10%
배재권(2010)	11개	혼합 인공신경망	부실: 944 건전: 944	74.42~80.89%
민성환(2012)	25개	앙상블 SVM	부실: 609 건전: 609	71.35~75.76%
신태수, 홍태호(2011)	6개	AdaBoost SVM	부실: 74 건전: 135	73.24~74.66%
김나라 외 2인(2013)	31개	앙상블 인공신경망	부실: 1,472 건전: 1,472	73.24~74.66%
최소윤, 안현철(2015)	9개	GA Fuzzy SVM	표본: 1548	75.48%

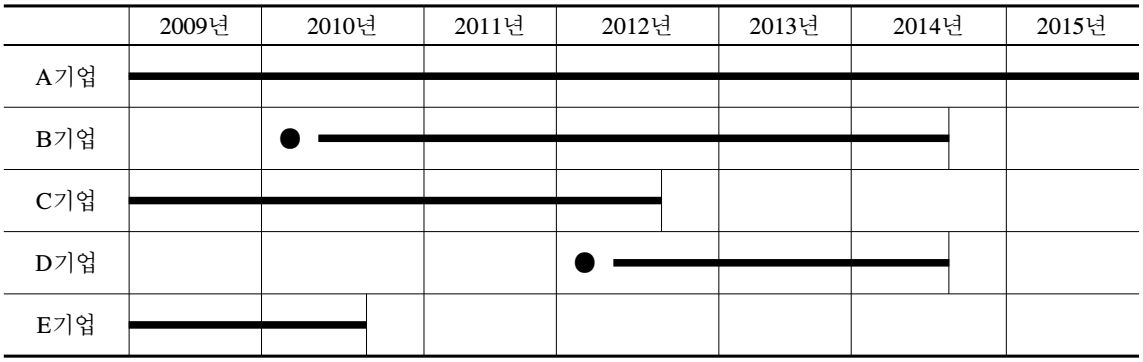
주) 최소윤, 안현철(2015)의 연구의 경우, 표본의 수는 나와 있으나, 부실 및 건전기업의 수에 대한 구체적인 언급은 없음.

이와 같은 현상에 대해서 본 연구는 예측 모델의 강건성과 확장성에 문제에 있다고 판단하였으며, 이를 저해하는 요인을 찾고 해결 방안을 제시하고자 한다. 다음 절에서는 기존 연구에서 미흡하게 다루어진 문제점들과 이에 관한 해결 방안에 대해 살펴보고자 한다.

2.2 기존 연구에서 미흡하게 다루어진 문제점과 해결 방안

2.2.1 데이터셋을 구성하는 과정에서 발생하는 바이어스 문제

기업부실화 예측 모델은 기업의 재무데이터를 활용한다. 따라서 재무데이터를 활용하여 연구 설계를 할 경우 데이터의 특성이 함께 고려되어야 한다. 일반적으로 재무데이터는 횡단면(cross section)과 종단면(time series) 특성을 동시에 갖추고 있어 패널데이터와 유사한 측면이 있다. 재무데이터는 우리나라 상법 제447조와 제579조에 의해 기업의 탄생에서부터 파산 직전까지 존재하게 되는데, 이것은 데이터 분석 관점에서 보았을 때, 사례(instance)의 생성과 소멸을 의미한다. 문제는 개별 사례들의 생성 시점과 소멸 시점이 서로 다르다는 점이다. 즉, 재무데이터는 <그림 1>에서와 같이 중도절단(censoring)의 문제가 발생하여 패널데이터와는 완전히 다른 구조를 가지게 되



주) ●은 IPO 상장 시점을, ⊣ 는 IPO 퇴출 시점을 의미함.

〈그림 1〉 재무데이터에서 중도절단의 예시

며, 데이터셋을 구성하는 과정에서 중도절단으로 인한 바이어스 문제가 발생된다.

그러나 대부분의 기존 연구들에서는 단순히 회계연도(fiscal year) 단위인 1년 단위로 데이터셋을 구성하거나(Altman, 1968; Beaver, 1966; Ohlson, 1980; Zmijewski, 1984), 임의로 선정한 기간 내에 모든 부실기업과 건전기업을 하나의 데이터셋으로 구성하여 정태적 분석(static analysis)을 하였다(Hsieh, 2005; McKee and Lensberg, 2002; Tsakonas *et al*., 2006; West *et al*., 2005). 여기에서 단순히 회계연도 단위인 1년 단위로 데이터셋을 구성하는 전자의 경우, 데이터는 횡단면 형태로 존재하고 시계열 특성은 없다. 이는 스냅샷(snapshots) 관점에서의 자료수집과 분석이라 할 수 있는데, 해당 1년치의 데이터셋에서 구조적 패턴을 만들게 되고, 예측 로직(prediction logic)이 그 1년에만 과적합(overfitting) 되는 문제가 발생하게 한다. 다만, 이 경우에는 시계열 특성이 통제되었기 때문에 중도절단의 문제가 크게 제기되지 않는다. 한편, 임의로 선정한 기간 내에 모든 데이터를 하나의 데이터셋으로 구성한 후자의 경우, 일정 기간 관점에서의 자료수집과 분석이라 할 수 있으며, 데이터에는 횡단면과 종단면이 모두 존재한다. 그러나 개별 사례들의 생성 시점과 소멸 시점이 다르기 때문에 이 경우에는 왼쪽 중도절단(left censoring) 또는 오른쪽 중도절단(right censoring) 문제가 발생하게 된다. 따라서 연구자는 데이터 전처리 과정에서 개별 사례를 제한적으로 사용할 수밖에 없으며, 정보의 손실 또한 감안해야 한다.

결국 연구자는 데이터셋을 구성하는 과정에서 발생하는 바이어스 문제와 부딪칠 수밖에 없으며, 이 바이어스 문제는 예측 모델의 강건성과 확장성을 떨어뜨려 예측 성능에 부(-)의 영향을 주게 된다.

2.2.2 거시경제위험 요소의 미반영

기업부실화는 기업 내부 자산가치의 변동위험인 기업고유위험(idiosyncratic risk)과 시장전체의 변동위험인 거시경제위험(systematic risk)에 의해 결정된다(김성규, 이화득, 2012; 김성태 등, 2010; 이치송, 2005; Pederzoli and Torricelli, 2005). 이 중 거시경제적 측면에서 기업부실화가 발생하는 현상의 특징을 살펴보면 기업부실화는 경기불황이나 수축기에 집중적으로 발생하는 현상을 보인다(김성태 등, 2010).

<표 5>는 우리나라의 경기순환국면별 경제동향을 보여주며, 경기순환국면이 수축기일 때 부도율이 높은 반면, 확장기일 때는 부도율이 낮은 것을 볼 수 있다(참고 <부록 1>). 이러한 현상은 선행연구에서 제시된 결과와도 일치한다(김성규, 이화득, 2012; 김성태 등, 2010; 이치송, 2005; 장영민, 변재권, 2010; 조성표, 류인규, 2007; Altman, 1984; Altman

〈표 5〉 경기순환국면별 경제동향

구 분	순환국면	연도	부도율
제7순환	수축기(2000. 8~2001. 7)	2000	0.40
		2001	0.39
제8순환	확장기(2001. 7~2002. 12)	2002	0.11
	수축기(2002. 12~2005. 4)	2003	0.17
		2004	0.18
		2005	0.14
제9순환	확장기(2005. 4~2008. 1)	2006	0.11
		2007	0.11
	수축기(2008. 1~2009. 2)	2008	0.15
제10순환	확장기(2009. 2~2011. 8)	2009	0.14
		2010	0.15
		2011	0.11
	수축기(2011. 8~미확정)	2012	0.12
		2013	0.14
		2014	0.19

자료: 국가통계포털(KOSIS)(http://kosis.kr).

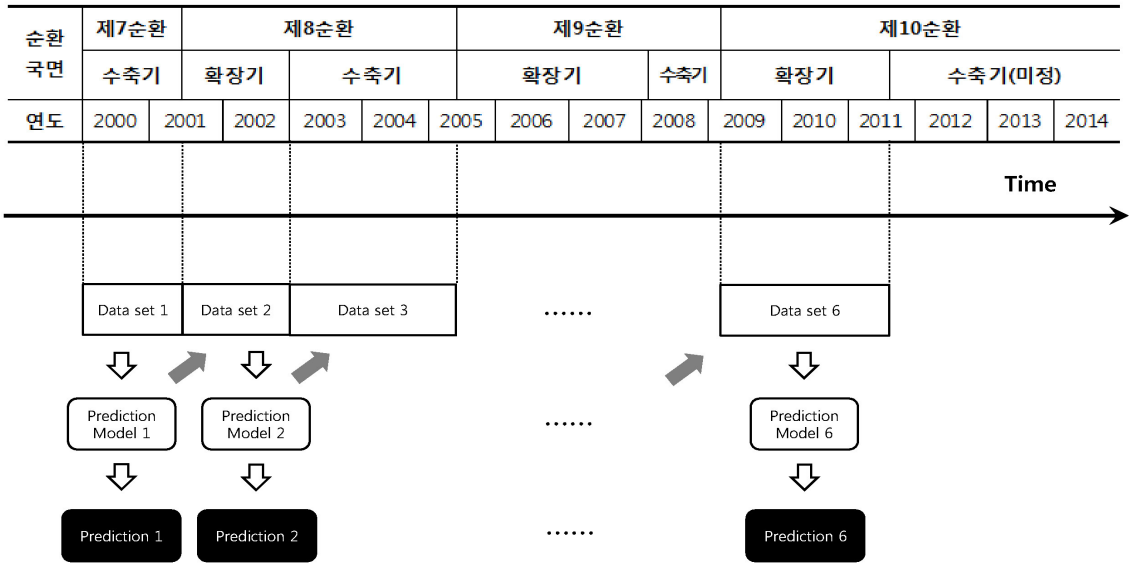
et al., 2005; Pederzoli and Torricelli, 2005).

조성표, 류인규(2007)는 경기를 불황기(하락기, 위기기, 지속기)와 호황기로 구분하고 기업부실화의 요인이 경기시점별로 어떻게 달라지는지를 실증분석하였다. 이 연구에서는 1993~1998년 사이의 부실기업과 건전기업을 1:3 비율로 구성하여 로지스틱 분석을 실시하였다. 그 결과, 하락기에는 부채비율, 회전률, 수익성이, 위기기에는 부채비율, 현금흐름, 기업규모, 회전율이, 그리고 지속기에는 부채비율과 현금흐름이 유의적이었음을 제시하면서 경기변동에 따라 기업부실화에 유용한 재무변수에 차이가 있음을 주장하였다.

김성태 등(2010)은 부도상관관계를 측정하기 위해 국내 상장기업의 재무데이터(1993~2005년)를 이용하여 거시위험 민감도를 반영한 해저드 모형을 소개하였다. 이 연구에서는 새롭게 제안된 해저드 모형의 해저드율(hazard rate)을 사용하여 부도상관관계를 측정하였으며, 그 결과 기존의 해저드 모형에서 측정한 값보다 높은 예측 값을 얻을 수 있었다. 이를 통해 경기변동과 부도상관관계가 서로 관련성이 있음을 보여주었다.

이와 같이 선행연구들에서는 거시경제위험을 반영하지 않은 재무변수만을 이용하여 기업부실화를 예측하거나 거시경제변수와 기업부실화 간의 관계를 규명하는 실증적 연구가 대부분이었다. 따라서 기존 연구에는 기업고유위험과 거시경제위험을 통합적으로 반영되지 못한 한계점이 있음을 알 수 있다(김성규, 이화득, 2012; 김성태 등, 2010; 조성표, 류인규, 2007; Pederzoli and Torricelli, 2005). 그러나 선행 연구들을 통해 경기변동에 따라 기업부실화 예측에 유용한 재무변수가 무엇인지를 파악할 수 있으며, 어떤 변수를 사용하는가에 따라 기업부실화 예측 모델의 유용성이 달라짐을 알 수 있다.

따라서 본 연구에서는 데이터셋을 구성하는 과정에서 발생하는 바이어스 문제를 해결하고, 기업고유위험과 거시경제위험을 통합적으로 반영한 기업부실화 예측 모델을 개발하기 위해 경기순환국면을 반영한 기업부실화 예측 프레임워크를 제안하고자 한다. 여기에서 경기순환국면에



〈그림 2〉 경기순환국면을 고려한 데이터셋 구성과 예측

따른 기업부실화 예측 모델의 유용성을 반영하고 중도절단 문제를 해결하기 위해 데이터셋을 경기순환국면[1]을 기준으로 구성하고자 한다.

<그림 2>는 본 연구에서 제안한 프레임워크의 진행과정을 보여주며, 시간의 흐름에 따라 계속적으로 발생되는 데이터와 이를 분석 및 예측하는 방법을 포함하고 있다. 그 진행과정을 요약하면 다음과 같다. 1) 확장기 및 수축기 국면별로 데이터셋을 구성한다. 가령, 제7순환 수축기가 데이터셋 1이 되며, 제8순환 확장기가 데이터셋 2가 된다. 2) 각 데이터셋의 70%는 훈련셋(training set), 나머지 30%는 테스트셋(test set)으로 나누어 훈련셋을 통해 예측 모델을 생성하고, 테스트셋을 이용하여 생성된 예측 모델의 예측률을 측정한다. 3) 특정 순환국면(t시점)에서 생성된 'Prediction Model t'를 다음 순환국면(t+1시점)의 데이터셋에 적용시킨다. 그리고 예측 결과인 'Prediction t+1'을 얻는다. 즉, 제7순환 수축기에서 만들어진 예측 모델 1을 제8순환 데이터셋 2에서 테스트하여 예측 결과 2를 얻게 된다. 4) 그리고 전자와 후자의 예측률을 비교한다. 5) 이와 같은 과정을 마지막 데이터셋까지 반복하며 예측률을 얻는다.

2.2.3 데이터 불균형 문제

일반적으로 기업부실화는 매우 희소한 사건이다. 따라서 부실기업의 수는 건전기업과 비교해 보았을 때 그 수가 현저히 적다. 여기에서 소수 클래스와 다수 클래스 간에 표본 크기의 차이에서 발생하는 데이터 불균형 문제가 발생된다(Zhou, 2013). 실제로 부실기업과 건전기업간의 비율은 데이터셋의 구성에 따라 1:100, 1:1000, 심지어 1:10000까지도 존재한다(Chawla *et al.*, 2002).

데이터 불균형 상황에서 기업부실화 예측 모델을 개발하면 다음과 같은 문제점이 제기된다. 먼저, 대부분의 기계학습 알고리즘은 오분류를 줄임으로써, 전반적인 예측 성능을 최적화시키려는 경향이 있기 때문에 소수 클래스에 속한 데이터들을 다수 클래스로 분류하게 되는 문제가 나타난다(Sundarkumar and Ravi, 2015). 또한 예측

1) 통계청(http://kostat.go.kr)의 산업활동동향 보도자료에서 매마다 경기종합지수와 경기순환국면을 공시한다.

로직(prediction logic)이 데이터가 많은 다수 클래스를 중심으로 패턴을 인식하게 됨으로써, 다수 클래스에 대한 precision과 recall은 높아지지만 소수 클래스에서 대한 precision과 recall은 낮아지게 된다. 따라서 지표상 전체 예측률이 높게 나왔다고 하더라도, 이는 다수 클래스인 건전기업에 대한 분류가 잘 이루어진다는 것을 의미하며, 소수 클래스인 부실기업에 대한 분류는 잘 이루어지지 않는 문제가 발생된다. 그러나 많은 선행연구들에서는 데이터 불균형 문제를 고려하지 않고 기업부실화 예측 모델을 개발하였다(Zhou, 2013).

본 연구는 보다 나은 기업부실화 예측 모델을 개발하기 위해 샘플링 방법을 적용하여 데이터 불균형 문제를 해결하고자 한다. 일반적으로 샘플링 방법에는 언더샘플링(undersampling) 방법과 오버샘플링(oversampling) 방법이 있다(김태훈, 안현철, 2015; 허준, 김종우, 2007). 언더샘플링은 다수 클래스의 샘플 크기를 소수 클래스의 샘플 크기만큼 줄여서 데이터 균형화 시키는 방법이다. 그러나 전체 데이터의 양이 줄어들어 유용한 정보가 줄어드는 단점이 있다. 반면, 오버샘플링은 다수 클래스의 샘플을 기준으로 소수 클래스의 샘플을 중복 발생시켜 그 양을 늘리는 방법이며 소수 클래스의 샘플 크기가 매우 작을 때 효과적이다.

현실적으로 소수 클래스인 부실기업의 수가 다수 클래스인 건전기업의 수보다 절대적으로 적기 때문에 오버샘플링이 필요하다(김경민 등, 2014; 김은나 등, 2011; 오상훈, 2009; 정한나 등, 2010). 따라서, 본 연구에서는 오버샘플링 방법을 통해 데이터 불균형 문제를 해결하고자 하며, Chawla *et al.*(2002)가 제안한 synthetic minority over-sampling technique(SMOTE)을 예측 모델에 적용하였다.

2.2.4 서술적 바이어스 문제

현상을 반영하기 위해 만들어지는 연구모형은 본질적으로 서술적 바이어스(declarative bias)가 존재한다(Brazdil *et al.*, 2008). 과학적 연구는 연구자가 연구목적에 부합하는 가설을 세우고, 이를 검증하기 위한 연구모형을 만든다. 이때, 이론으로부터 도출된 가설이나 연구모형에 사용되는 개념들은 구성개념(construct)이라고 하며, 이 구성개념은 측정이 가능하도록 변수화 시키게 된다. 이때의 변수를 잠재변수(latent variable)라고 한다. 조작적 정의(operational definition)는 이 잠재변수를 어떻게 측정할 것인가에 대하여 명확하게 기술하는 것을 말한다(Witten *et al.*, 2011).

기업부실화 예측 모델에서 데이터셋은 기업이라는 사례들(instances)과 재무비율이라는 속성(attributes)으로 구성되어 있다. 데이터마이닝에서 속성은 잠재변수로 인식되며 학습된 기계는 구조적 패턴을 찾아 서술(declare)하게 된다. 그러나 구조적 패턴을 통해 표현되는 서술문이 실제 개념을 파악할 수 있을 만큼 충분히 표현되지 못하거나, 사례들에 잡음(nosiy)이 존재할 경우, 서술적 바이어스 문제가 나타난다(Witten *et al.*, 2011).

Abbasi *et al.*(2012)과 Brazdil *et al.*(2008)에 따르면, 서술적 바이어스는 가설을 구체화하는 과정에서 발생하며, 이를 변수의 유형과 양을 조정하여 다룰 수 있다고 하였다. Abbasi *et al.*(2012)의 연구에서는 기업부실화 예측에서 발생되는 서술적 바이어스 문제를 해결하기 위해, 설계과학(design science) 패러다임을 적용하여 새로운 IT 산출물(IT artifacts)을 생성해 예측 모델의 변수에 삽입하였다. 설계과학 패러다임은 Simon(1996)의 “The Sciences of the Artificial”에 뿌리를 두고 있다. 설계과학은 일반적 이론 수립(theory building)에 초점을 둔 행동 연구(behavioral research)와 함께, 이를 기반으로 산출물(artifacts)을 만들어 적용함으로써, 문제 해결을 목적으로 하는 패러다임이다(Hevner *et al.*, 2004; Baskerville *et al.*, 2015). 일반적으로 정보시스템 분야에서 설계과학 패러다임이 적용될 경우, 이때의 산출물들은 “IT artifacts”로 불리어지며, 이는 정보시스템을 개발하거나 사용할 때 적용되는 구성개념, 모델, 방법 등을 의미한다(Hevner *et al.*, 2004; March and Smith 1995).

본 연구에서는 구성개념의 관점에서 새롭게 생성된 변수를 IT 산출물로 정의하며, 서술적 바이어스 문제를 다루기 위해 기존의 기업재무변수에서 파생된 재무상태변동성 변수와 산업수준상태변동성 변수를 IT 산출물로 사용하고자 한다.

(1) 재무상태변동성(Financial context variability)

기업부실화와 관련된 선행 연구들에서는 정량적 분석 방법을 사용하기 위해 기업 재무제표를 사용해왔다(Kaminski *et al.*, 2004; Lin *et al.*, 2003; Summers and Sweeney, 1998). 기업 재무재표를 통해 기업의 경영성과와 재무상태에 관한 정보를 확인할 수 있다. 이러한 재무정보는 기업 내부 및 외부환경을 시계열로 반영하는 특성을 가지고 있기 때문에 시간 흐름에 따른 변화가 고려되어야 한다(Wang *et al.*, 2003).

따라서, 본 연구는 이러한 시계열 특성, 즉 경영성과와 재무상태의 변화를 반영하기 위해 당해 시점과 이전 시점의 재무변동성을 반영한 재무상태변동성을 기존 재무변수에서 유도하여 생성하고자 한다(Abbasi *et al.*, 2012). 재무상태변동성은 당해 시점 재무비율과 이전 시점 재무비율의 변동량을 의미하며, 아래와 같이 각 시점 재무 비율의 차(-) 또는 비(/)로 산출된다.

(당해 시점 재무비율-이전 시점 재무비율) 또는
(당해 시점 재무비율/이전 시점 재무비율)

(2) 산업수준상태변동성(Industry-level context variability)

Porter(1998)는 산업에서 기업의 치열한 경쟁은 기업의 수익을 줄어들게 만드는 위협이라고 하였다. 특히 기업이 난립하고 있는 산업(fragmented industry)[2]과 합병정리된 산업(consolidated industry)[3]에서 서로 경쟁할 경우, 기업은 기회보다는 위협 상황에 직면하게 되어 파산을 겪을 수 있다고 경고하고 있다(Hill and McShane, 2008). 이러한 맥락에서 경쟁에서 뒤쳐진 부실기업들의 경영성과와 재무상태는 우량기업에 비해 상대적으로 저조할 뿐만 아니라, 산업의 평균과도 비교할 경우 차이가 있을 것이다.

본 연구는 이러한 배경에서 산업수준상태변동성을 재무변수에서 파생시켜 생성하고자 한다(Abbasi *et al.*, 2012; Brazdil *et al.*, 2008). 본 연구는 부실기업과 우량기업의 격차를 확인하기 위해, 데이터셋에서 제조업과 비제조업으로 산업을 구분한 후, 각각의 산업에서 총자산순이익률이 가장 높은 상위 10개 기업을 추출하고 이들 10개 기업의 각 속성의 평균을 산출하였다. 총자산순이익률(net income to total assets)은 기업의 세차감후 순이익으로, 최종적인 성과인 당기순이익을 타인자본과 자기자본의 합인 총자산으로 나누어 계산되는 기업경영의 총괄적인 성과지표로서 기업의 경영성과를 종합적으로 판단하는데 널리 활용된다(김복만 등, 2010).

따라서 본 연구에서는 총자산순이익률을 기준으로 해당 산업의 상위 10개 기업을 추출하였다. 앞서 추출된 상위 10개 기업의 속성 평균을 기준으로 각 사례들과의 차(-)와 비(/)를 통해 산업수준상태변동성: Top-10 Gap Model 변수를 생성하였으며, 또한 산업 평균과의 격차도 패턴인식 하기 위해, 해당 산업의 평균과 각 사례들의 차(-)와 비(/)를 통해 산업수준상태변동성: Average Gap Model 변수를 생성하였다.

<표 6>은 본 연구에서 사용한 변수에 대한 정보를 보여주고 있다. 사용된 총 변수의 수는 105개이며, 기본 재무비율과 재무변수로부터 파생된 재무상태변동성 및 산업수준상태변동성으로 구분된다.

2) 난립하고 있는 산업(fragmented industry): 산업 내에 많은 중소업체들로 구성된 산업.

3) 합병정리된 산업(consolidated industry): 몇몇 대형업체들에 의해 지배되는 산업.

〈표 6〉 본 연구에서 사용된 변수 정보

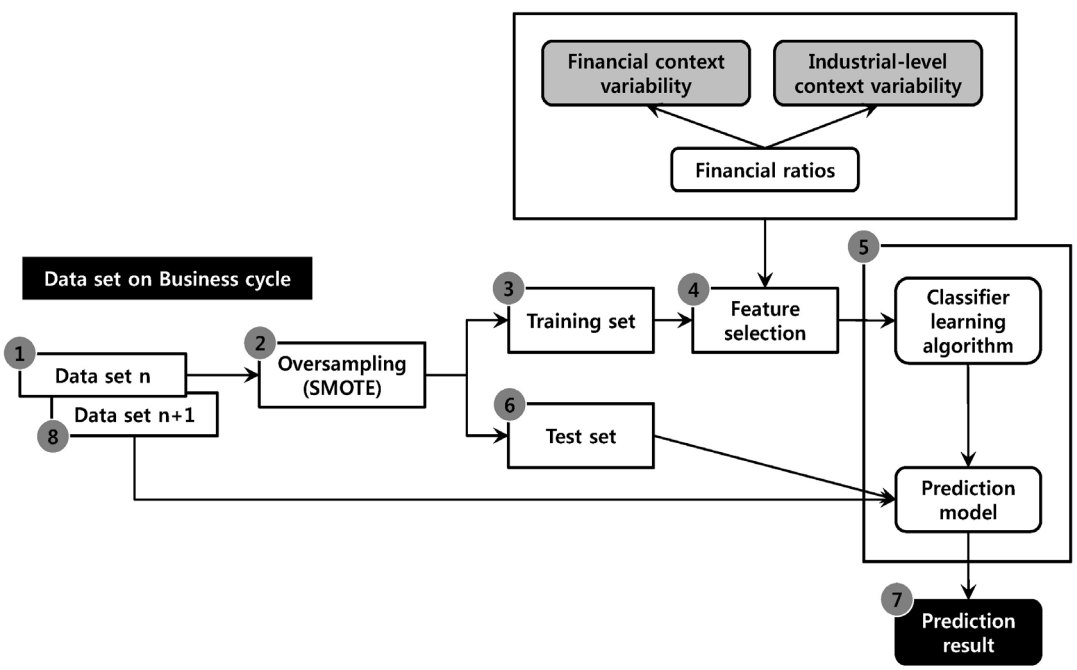
구 분	변 수	개 수
기본 재무비율(연간 보고서)	R1, R2, …, R15	15개
재무상태변동성	R1-P1, R2-P2, …, R15-P15	15개
	R1/P1, R2/P2, …, R15/P15	15개
산업수준상태변동성 : Top-10 Gap Model	R1-T1, R2-T2, …, R15-T15	15개
	R1/T1, R2/T2, …, R15/T15	15개
산업수준상태변동성 : Average Gap Model	R1-A1, R2-A2, …, R15-A15	15개
	R1/A1, R2/A2, …, R15/A15	15개
Total		105개

Ⅲ. 연구방법

3.1 실험 설계

본 연구는 경기순환국면을 반영한 동태적 기업부실화 예측을 위해 <그림 3>과 같이 실험을 설계하였다. 실험의 진행과정을 요약하면 다음과 같다.

1단계) 경기순환국면을 기준으로 데이터셋을 구성한다. 2단계) 소수 클래스와 다수 클래스 간의 샘플 크기를 균형화 시키기 위해 오버샘플링 기법인 SMOTE를 적용한다. 3단계) '데이터셋 1'부터 기업부실화 예측 모델을 개발한다. 여기서 데이트셋의 70%는 훈련셋으로 사용하고, 나머지 30%는 검증을 위한 테스트셋으로 사용한다. 4단계) 변수선정(feature selection) 시에 기존의 재무변수와 IT artifact인 재무상태변동성 변수와 산업수준상태변



〈그림 3〉 경기순환국면을 반영한 기업부실화 예측 모델

동성 변수를 추가한다. 5단계) 분류 학습 알고리즘을 통해 예측 모델을 생성하고, 6단계) 훈련셋에서 만들어진 예측 모델을 테스트셋을 통해 검증한다. 7단계) 이후 예측 결과를 얻는다. 8단계) 그리고 다시 돌아와서, '데이터셋 1'에서 만들어진 예측 모델을 다음 데이터셋 2에 적용한다. 그리고 예측 결과를 얻는다. 7단계)에서의 예측 결과와 8단계)에서의 예측 결과를 비교하고, 마지막 데이터셋까지 1단계)에서 8단계)의 과정을 반복한다.

추가로, 본 연구에서는 확장기(혹은 수축기)에서 수축기(혹은 확장기)로 넘어갈 때의 예측성과들과 확장기(혹은 수축기)에서 확장기(혹은 수축기)로 넘어갈 때의 예측성과들을 비교하여 경기순환국면에 따른 기업부실화 예측 모델의 유용성을 검토하고자 한다.

3.2 분석 데이터

3.2.1 목표 변수

부실기업은 재무적 의무를 다하지 못한 기업, 즉 계약상의 채무를 상환할 수 없는 기업에서부터 법률적 파산에 이른 기업을 의미한다. 본 연구에서는 한국증권거래소에서 상장폐지한 기업을 부실기업으로 정의하고, 상장폐지 여부를 목표변수로 사용하였다. 여기에서 상장폐지기업을 부실기업으로 정의한 이유는 한국증권거래소의 상장폐지기준인 매출액 미달, 감사의견 부적정, 자본잠식, 최종부도 등의 내용이 기업부실화의 구성개념과 부합하기 때문이다.

3.2.2 독립 변수

본 연구에서는 기업부실화 예측에 영향을 주는 독립 변수를 선정하기 위해 선행연구를 참고하여 가장 많이 사용된 재무비율 51개를 변수로 추출하였다(김건우, 1987; 김경재, 한인구, 2001; 김명종, 2009; 김명종, 2010; 김명종, 2012; Abbasi *et al.*, 2012; Altman, 1968; Beaver, 1966; Fedorova *et al.*, 2013; Lin *et al.*, 2003; Ohlson, 1980; Wang *et al.*, 2014; Zmijewski, 1984). 이 51개 재무비율은 성장성, 수익성, 생산성, 안정성, 활동성으로 구분된다(<부록 2> 참고).

본 연구에서는 선행연구로부터 추출한 51개 변수 중 예측 모델에 유의한 핵심변수를 선정하기 위해 변수 선정을 실시하였고, 변수 선정 방법에는 Information gain과 Gain ratio를 사용하였다[4] (Abbasi *et al.*, 2010; Abbasi *et al.*, 2012). 이 때, 각각 순환국면별 데이터셋에서 51개 재무비율 변수에 대해 Information gain과 Gain ratio값을 산출하고, 산출 값이 높은 순을 기준으로 1위부터 51위까지 순위화하였다. 그 후 모든 순환국면의 데이터 셋에서 Information gain과 Gain ratio값의 순위가 20위 내에 있는 재무비율 변수들만을 선정하였다. 이와 같은 방법을 통해 추출된 변수는 총 15개로 <표 7>과 같다.

〈표 7〉 Information gain과 Gain ratio를 사용하여 선정된 재무비율

구 분	변수
성장성	매출액증가율
수익성	매출액순이익률
	총자산영업이익률
	총자산순이익률
	자기자본순이익률
	유보율
생산성	총자본투자효율
	설비투자효율
안정성	자기자본구성비율
	유동비율
	부채비율
	Cash Flow 대 부채비율
	Cash Flow 대 총자본비율
활동성	총자본회전률
	자기자본회전률

4) Information gain과 Gain ratio는 주어진 타겟 변수에 영향을 미치는 각 속성들의 순위를 산출하고, 가장 좋은 값을 갖는 속성이 주어진 데이터를 분리 할 수 있는 분리 속성(splitting attribute)으로 선택된다.

Ⅳ. 실험 및 결과

4.1 실험 데이터

한국거래소의 홈페이지[5]에 등록된 상장폐지 공시는 2001년부터 시작하였기 때문에, 본 연구에서는 공시가 시작된 제7순환 수축기(2000년 8월~2001년 7월)부터 경기순환국면이 확정된 제10순환 확장기(2009년 2월~2011년 8월)까지의 자료를 수집하였다. 또한 본 연구에서는 공공데이터를 제공하는 한국신용정보의 KIS-Value[6]와 한국상장협의회의 TS-2000[7]을 이용하여 제7순환 수축기부터 제10순환 확장기까지 KOSPI, KOSDAQ에 상장된 기업을 대상으로 자료를 수집하였으며, 한국표준산업분류를 기준으로 분류된 제조업 및 비제조업 1691개 기업을 대상으로 재무비율을 추출하였다.

본 연구에서는 제조업 및 비제조업을 제외한 신용금고업, 보험업, 여신금융업, 은행업, 종합금융업, 증권업, 기타금융업에 속한 기업은 그 수가 매우 적을 뿐만 아니라, 제조업 및 비제조업의 경영활동과 재무상태와도 상이하기 때문에 이상치(outliers)로 인식될 가능성이 있어, 이를 제외하였다. 제외된 기업의 표본은 다음과 같다. 신용금고업(1개), 보험업(12개), 여신금융업(8개), 은행업(13개), 종합금융업(1개), 증권업(23개), 기타금융업(35개)이다.

재무데이터는 회계연도 1년간의 기업 경영활동을 내포하므로 순환국면에 해당하는 연도에서 (+1)년을 하여 순환국면별 데이터셋을 구성하였다. 예를 들면, 제7순환 수축기의 경우에는 2000년 8월부터 2001년 7월까지이다. 이때의 데이터셋은 2001년 8월부터 2002년 7월까지의 재무제표로 구성된다.

5) http://www.krx.co.kr.
6) www.kisvalue.com.
7) www.kocoinfo.com.

본 연구에서는 부실기업에 대한 재무자료는 부실직전연도를 중심으로 수집하였다. 재무데이터 특성상 이상치가 존재할 수 있으며 너무 크거나 작은 극단값은 분석에 영향을 미칠 수 있기 때문에 표본자료에서 모든 변수를 양극단 1%에서 윈저라이징(winsorizing)하였다.

4.2 표본의 특성

각 경기순환국면별 데이터셋의 표본 특성은 <표 8>과 같다. <표 8>의 각 데이터셋별 건전기업과 부실기업의 표본비율을 살펴보면 건전기업이 부실기업보다 최소 17배 이상 많음을 알 수 있다.

<표 9>는 제7순환 수축기의 기술통계량 및 t-test 결과이다. <표 9>를 보면, 자기자본회전률과 설비투자효율을 제외하고는 모든 변수가 유의수준 0.05이내에서 집단 간(건전기업과 부실기업 간) 유의한 차이를 보였다. 또한 제8순환 확장기부터 최근 제10순환 확장기까지 각각을 t-test한 결과 모두 집단 간 유의한 차이를 보였다.

4.3 실험 결과

본 연구에서 데이터마이닝 기법을 이용한 기업부실화 예측 모델을 개발하기 위하여 단일 분류학습 알고리즘과 앙상블 기법을 사용하였다. 단일 분류 학습 알고리즘으로는 Bayesian network, Logistic regression, Artificial neural networks, J48, SVM-RBF, RandForest, JRip을 사용하였고, 앙상블기법으로는 AdaBoost, Bagging, RotationForest, ThresholdSelector를 사용하였다. 기업부실화 예측 모델 개발에 사용된 데이터마이닝 프로그램은 오픈 소프트웨어인 WEKA ver.3.6.11을 사용하였다.

본 연구는 제Ⅲ장의 실험 설계를 토대로 실험을 진행하였으며, 데이터 불균형 상태에서 예측 모델을 만들 시에 AdaBoost-J48이 모든 순환국면 데이터셋에서 가장 높은 예측성과를 나타내었다. 따라서 SMOTE 기법을 통해 데이터를 균형화 시키는 단계

〈표 8〉 경기순환국면별 표본의 수

경기순환국면		표본의 수		합 계
제7순환	수축기 (2000. 8~2001. 7)	제조업	건전: 1,560개	건전: 2,130개 부실: 86개 (표본비율 1:25)
			부실: 58개	
		비제조업	건전: 570개	
			부실: 28개	
제8순환	확장기 (2001. 7~2002. 12)	제조업	건전: 1,683개	건전: 2,322개 부실: 90개 (표본비율 1:26)
			부실: 62개	
		비제조업	건전: 639개	
			부실: 28개	
	수축기 (2002. 12~2005. 4)	제조업	건전: 1,883개	건전: 2,616개 부실: 63개 (표본비율 1:42)
			부실: 50개	
		비제조업	건전: 733개	
			부실: 13개	
제9순환	확장기 (2005. 4~2008. 1)	제조업	건전: 2,964개	건전: 4,179개 부실: 251개 (표본비율 1:17)
			부실: 151개	
		비제조업	건전: 1,215개	
			부실: 100개	
	수축기 (2008. 1~2009. 2)	제조업	건전: 1,105개	건전: 1,587개 부실: 68개 (표본비율 1:23)
			부실: 47개	
		비제조업	건전: 482개	
			부실: 21개	
제10순환	확장기 (2009. 2~2011. 8)	제조업	건전: 3,328개	건전: 4,851개 부실: 220개 (표본비율 1:22)
			부실: 130개	
		비제조업	건전: 1,523개	
			부실: 90개	

주) 표본비율은 부실기업기준으로 산출함.

와 IT 산출물을 삽입하는 단계에서 AdaBoost-J48을 이용하여 예측 모델을 개발하였다. 또한 기본적인 재무비율(15개)과 IT 산출물(90개)을 변수에 추가하여 예측 모델을 개발하였는데, 부실기업에 대한 예측성과가 40% 미만으로 너무 저조하여 Information gain[8]을 산출한 후, Information gain의 값이 "0"인 변수들을 제거하였다(참고 <부록 3>). 그 결과, 본 연구에서는 최종적으로 78개의 변수를 구성하여 기업부실화 예측 모델을 개발하였다.

8) Information gain은 변수의 유용성을 제시할 수 있어, 변수 선정에 널리 쓰인다(Witten *et al*., 2011).

본 연구의 실험결과를 요약하면 <표 10>~<표 15>와 같다.

<표 10>은 제7순환기의 기업부실화 예측 모델과 표본비율에 따른 예측성과를 보여준다. 여기서 좌측 셀의 Baseline은 기본 재무비율만을 사용하여 만든 예측 모델의 성과를 나타내고, 우측 셀의 IT 산출물은 기본 재무비율과 여기서 파생된 IT 산출물을 변수에 추가하여 만든 예측 모델의 성과를 나타낸다. 제7순환 수축기 데이터셋을 이용하여 개발된 예측성과를 통해 3가지 측면의 시사점을 파악할 수 있다. 첫째, 데이터 불균형 상태에서 기본 재무비율만을 사용하여 예측 모델을

만들었을 경우, 전체 accuracy는 약 96%로 나타났다. 이는 선행연구에 비해 상당히 높은 예측 성능을 나타낸다. 그러나 건전기업의 경우, accuracy, precision, recall, AUC가 모두 높게 나타나는 예측 성과를 보였으나, 부실기업은 대조적으로 상당히 낮은 것을 볼 수 있다. 이는 학습 알고리즘이 건전기업을 건전기업으로는 잘 분류하지만 부실기업의 경우에는 그렇지 못함을 의미한다. 둘째, 예측 모델이 표본 크기에 많은 영향을 받는 것을 알 수 있다. <표 10>을 보면, 표본비율이 1:1에 가까워질수록 즉, 데이터가 균형화 될수록 건전기업뿐만 아니라 부실기업의 예측성과도 높아지는 것을

〈표 9〉 제7순환 수축기: 기술통계량 및 t-test

변수	Class	N	평균	표준편차	평균차	t값
매출액증가율	건전	2130	10.98	28.237	9.01	2.905**
	부실	86	1.97	27.199		
매출액순이익률	건전	2130	1.49	18.621	45.75	17.762**
	부실	86	-44.25	74.839		
총자산영업이익률	건전	2130	6.60	8.995	9.47	9.485**
	부실	86	-2.86	10.929		
총자산순이익률	건전	2130	3.50	11.453	24.21	17.598**
	부실	86	-20.71	28.129		
자기자본순이익률	건전	2130	5.14	23.551	52.72	18.413**
	부실	86	-47.57	61.286		
유보율	건전	2130	510.53	610.204	401.36	6.090***
	부실	86	109.17	160.138		
자기자본구성비율	건전	2130	54.10	19.437	23.68	10.988***
	부실	86	30.42	23.176		
유동비율	건전	2130	218.02	188.386	67.84	3.319***
	부실	86	150.18	103.148		
부채비율	건전	2130	125.47	151.433	-208.22	-11.596***
	부실	86	333.69	346.161		
Cash Flow 대 부채비율	건전	2130	17.14	28.183	22.58	7.397***
	부실	86	-5.43	12.876		
Cash Flow 대 총자본비율	건전	2130	5.71	8.273	10.02	10.935***
	부실	86	-4.30	9.715		
총자본회전률	건전	2130	1.06	.577	0.16	2.484**
	부실	86	.91	.579		
자기자본회전률	건전	2130	2.41	1.919	0.33	1.557
	부실	86	2.09	1.266		
총자본투자효율	건전	2130	22.59	16.430	23.68	12.737***
	부실	86	-1.08	26.067		
설비투자효율	건전	2130	134.64	250.738	46.43	1.645
	부실	86	88.21	373.949		

주) $^{***}p < 0.01$, $^{**}p < 0.05$

확인할 수 있다. 셋째, IT 산출물을 추가할 경우, 기본 재무비율만을 사용했을 때보다 예측성과가 비교적 높은 것을 볼 수 있다.

본 연구는 나머지 제8순환 확장기부터 제10순환 확장기까지도 같은 방식으로 기업부실화 예측 모델을 만들었다. 그 결과, 제7순환 수축기에서 나타났던 결과와 마찬가지로 대부분의 건전기업이 부실기업보다 예측성과가 높은 것으로 나타났으며, 표본비율이 1:1에 가까워질수록 부실기업의 예측성과도 점점 높아지는 것을 보였다. 뿐만 아니라, IT 산출물을 추가로 삽입하였을 때, 앞선 결과와 동일하게, 기본 재무비율만을 사용했을 때 보다 예측성과가 비교적 높아지는 것을 알 수 있었다.

다만, 여기서 앞선 결과와 다른 흥미로운 결과는 기본 재무비율만을 사용한 예측 모델인 baseline의 경우, 예측성과가 표본비율이 1:1에 가까워질수록 accuracy가 점차 떨어지는 것을 볼 수 있다. 이러한 이유는 데이터가 균형화 됨에 따라 baseline의 건전기업에 대한 precision과 recall이 점점 떨어지기 때문이다. 결과적으로 전체적인 예측성과가 높아지는 이유는 데이터가 균형화됨에 따라 부실기업에 대한 precision, recall의 증가폭이 건전기업에 대한 precision과 recall의 감소폭보다 더 크기 때문이다. 그러나 IT 산출물을 추가한 예측모델의 예측성과를 보면, 표본비율이 1:1에 가까울수록 accuracy이 높아질 뿐만 아니라, 건전기업과 부실기업에 대한 precision과 recall도 높아짐을 알 수 있다.

〈표 10〉 제7순환 수축기의 표본비율에 따른 예측성과

제7순환 수축기								
표본비율	① 1 : 25				AdaBoost-J48			
	Baseline		IT artifacts					
Acc.	96.2545		96.435					
	건전	부실	건전	부실				
Pre.	0.969	0.543	0.971	0.585				
Rec.	0.992	0.221	0.992	0.279				
AUC	0.832	0.823	0.862	0.876				
표본비율	② 1 : 12				③ 1 : 6			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	94.9175		95.6125		96.0792		96.2813	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.961	0.727	0.963	0.82	0.975	0.871	0.973	0.896
Rec.	0.985	0.512	0.991	0.529	0.98	0.843	0.985	0.828
AUC	0.942	0.943	0.931	0.943	0.976	0.977	0.981	0.985
표본비율	④ 1 : 3				⑤ 1 : 1.5			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	96.4159		96.6998		97.3474		97.8608	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.978	0.922	0.983	0.92	0.988	0.952	0.989	0.963
Rec.	0.975	0.932	0.973	0.948	0.968	0.982	0.976	0.983
AUC	0.991	0.991	0.993	0.994	0.996	0.996	0.997	0.997

주) ACC.: accuracy, Prec.: precision, Rec.: recall, AUC: area under the ROC curve.

〈표 11〉 제8순환 확장기의 표본비율에 따른 예측성과

제8순환 확장기								
표본비율	① 1 : 26				AdaBoost-J48			
	Baseline		IT artifacts					
Acc.	96.1857		98.3831					
	건전	부실	건전	부실				
Pre.	0.989	0.493	0.985	0.932				
Rec.	0.971	0.733	0.998	0.611				
AUC	0.952	0.952	0.883	0.904				
표본비율	② 1 : 13				③ 1 : 6			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	94.2846		97.7218		92.3565		97.651	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.979	0.581	0.978	0.956	0.963	0.696	0.978	0.966
Rec.	0.959	0.733	0.997	0.717	0.948	0.764	0.995	0.856
AUC	0.957	0.956	0.946	0.959	0.96	0.96	0.985	0.988
표본비율	④ 1 : 3				⑤ 1 : 1.6			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	90.3024		98.7508		88.9155		98.9367	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.936	0.796	0.988	0.984	0.892	0.883	0.992	0.985
Rec.	0.937	0.794	0.995	0.963	0.933	0.819	0.991	0.987
AUC	0.964	0.964	0.996	0.996	0.969	0.969	0.996	0.998

〈표 12〉 제8순환 수축기의 표본비율에 따른 예측성과

제8순환 수축기								
표본비율	① 1 : 42				AdaBoost-J48			
	Baseline		IT artifacts					
Acc.	99.4774		99.3281					
	건전	부실	건전	부실				
Pre.	0.995	0.98	0.994	0.979				
Rec.	1	0.794	1	0.73				
AUC	0.938	0.938	0.919	0.907				
표본비율	② 1 : 21				③ 1 : 10			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	98.9059		98.9788		98.1172		98.9191	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.99	0.962	0.99	0.99	0.982	0.971	0.989	0.991
Rec.	0.998	0.794	1	0.786	0.998	0.81	0.999	0.885
AUC	0.973	0.973	0.955	0.975	0.972	0.972	0.989	0.994
표본비율	④ 1 : 5				⑤ 1 : 1.3			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	96.6346		99.1667		94.3709		99.6137	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.967	0.965	0.991	0.998	0.936	0.967	0.996	0.996
Rec.	0.994	0.821	1	0.95	0.989	0.825	0.998	0.99
AUC	0.977	0.977	0.995	0.995	0.98	0.98	0.998	0.999

〈표 13〉 제9순환 확장기의 표본비율에 따른 예측성과

제9순환 확장기								
표본비율	① 1 : 17				AdaBoost-J48			
	Baseline		IT artifacts					
Acc.	98.1264		98.4424					
	건전	부실	건전	부실				
Pre.	0.983	0.938	0.985	0.974				
Rec.	0.997	0.717	0.999	0.745				
AUC	0.924	0.924	0.908	0.915				
표본비율	② 1 : 8				③ 1 : 4			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	96.2401		97.3937		93.2279		97.5111	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.967	0.916	0.974	0.973	0.934	0.92	0.972	0.989
Rec.	0.992	0.715	0.997	0.779	0.985	0.712	0.998	0.881
AUC	0.931	0.931	0.956	0.965	0.935	0.935	0.987	0.99
표본비율	④ 1 : 2				⑤ 1 : 1			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	89.5588		98.3352		85.1007		99.0726	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.879	0.947	0.982	0.986	0.787	0.959	0.988	0.993
Rec.	0.981	0.719	0.993	0.963	0.97	0.727	0.994	0.988
AUC	0.938	0.938	0.996	0.997	0.94	0.94	0.999	0.999

〈표 14〉 제9순환 수축기의 표본비율에 따른 예측성과

제9순환 수축기								
표본비율	① 1 : 23				AdaBoost-J48			
	Baseline		IT artifacts					
Acc.	99.1541		99.3958					
	건전	부실	건전	부실				
Pre.	0.995	0.909	0.994	0.983				
Rec.	0.996	0.882	0.999	0.868				
AUC	0.981	0.981	0.954	0.956				
표본비율	② 1 : 12				③ 1 : 6			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	99.0714		99.3035		98.3862		99.5697	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.994	0.948	0.993	0.992	0.99	0.948	0.996	0.993
Rec.	0.996	0.934	0.999	0.919	0.991	0.941	0.999	0.978
AUC	0.989	0.989	0.971	0.973	0.993	0.993	0.994	0.994
표본비율	④ 1 : 3				⑤ 1 : 1.4			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	97.5598		99.343		97.4206		99.3271	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.983	0.954	0.994	0.991	0.972	0.978	0.994	0.992
Rec.	0.984	0.95	0.997	0.983	0.985	0.959	0.994	0.992
AUC	0.995	0.995	0.997	0.997	0.995	0.995	0.998	0.998

〈표 15〉 제10순환 확장기의 표본비율에 따른 예측성과

제10순환 확장기								
표본비율	① 1 : 22				AdaBoost-J48			
	Baseline		IT artifacts					
Acc.	98.4618		98.6196					
	건전	부실	건전	부실				
Pre.	0.989	0.882	0.987	0.969				
Rec.	0.995	0.745	0.999	0.705				
AUC	0.946	0.946	0.907	0.922				
표본비율	② 1 : 11				③ 1 : 6			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	96.8059		97.9021		94.3291		98.1155	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.977	0.858	0.979	0.977	0.954	0.87	0.98	0.989
Rec.	0.989	0.739	0.998	0.766	0.98	0.742	0.998	0.888
AUC	0.948	0.948	0.969	0.976	0.945	0.945	0.992	0.994
표본비율	④ 1 : 3				⑤ 1 : 1			
	Baseline		IT artifacts		Baseline		IT artifacts	
Acc.	91.4083		98.5025		88.1257		99.1399	
	건전	부실	건전	부실	건전	부실	건전	부실
Pre.	0.916	0.906	0.983	0.992	0.853	0.935	0.989	0.995
Rec.	0.972	0.756	0.997	0.952	0.961	0.772	0.996	0.985
AUC	0.947	0.947	0.997	0.998	0.949	0.949	0.999	0.999

한편, 본 연구는 경제위기시 기업 이해관계자들에게 더욱 정확하고, 합리적인 의사결정을 내릴 수 있는 통찰을 제공하기 위해 순환국면이 수축기일 때의 데이터셋을 활용하여 예측 모델을 만들고, 다가올 순환국면의 데이터셋을 테스트셋으로 가져와 예측하였다. 본 연구에서는 IT 산출물을 추가한 예측 모델의 표본비율이 1:1에 가까울수록 예측성과가 가장 높아지는 결과를 바탕으로, 제7순환 수축기에서 예측 모델을 만들어 제8순환 확장기 데이터셋과 제8순환 수축기 데이터셋을 테스트셋으로 가져와 예측하였다. 예측 모델은 AdaBoost-J48을 사용하여 만들었으며, 표본비율은 1:1.5로 하였다. <표 16>에서 '제7순환 수축기→제8순환 확장기'와 '제7순환 수축기→제8순환 수축기'의 accuracy는 각각 97.2637%와 97.9097%를 보여주며, '제7순환 수축기→제8순환 수축기'에서의 accuracy가 0.646% 더 높게 나타났다. 이는 건전기업의 precision과 부실기업의 recall이 비교적 높게 나타났기 때문이다.

〈표 16〉 제7순환 수축기→제8순환 확장기/수축기 예측결과 비교

AdaBoost-J48	제7순환 수축기 → 제8순환 확장기		제7순환 수축기 → 제8순환 수축기	
표본비율	1 : 1.5		1 : 1.5	
Acc.	97.2637		97.9097	
	건전기업	부실기업	건전기업	부실기업
Pre.	0.979	0.714	0.989	0.559
Rec.	0.993	0.444	0.990	0.524
AUC	0.909	0.919	0.895	0.926

〈표 17〉 제8순환 수축기→제9순환 확장기/수축기 예측결과 비교

AdaBoost-J48	제8순환 수축기 → 제9순환 확장기		제8순환 수축기 → 제9순환 수축기	
표본비율	1 : 1.3		1 : 1.3	
Acc.	91.4221		95.1662	
	건전기업	부실기업	건전기업	부실기업
Pre.	0.982	0.368	0.997	0.457
Rec.	0.926	0.717	0.953	0.926
AUC	0.855	0.855	0.983	0.983

<표 17>은 제8순환 수축기에서 예측 모델을 만들어, 제9순환 확장기 데이터셋과 제8순환 수축기 데이터셋을 테스트셋으로 가져와 예측한 결과이다. 예측 모델은 AdaBoost-J48를 사용하여 만들었으며, 표본비율은 1:1.3으로 하였다. <표 17>에서 '제8순환 수축기→제9순환 확장기'와 '제8순환 수축기→제9순환 수축기'의 accuracy는 각각 91.4221%와 95.1662%를 보여주며, '제8순환 수축기→제9순환 수축기'의 accuracy가 3.7441% 더 높게 나타났다. 이는 건전기업의 precision, recall 및 AUC 뿐만 아니라, 부실기업의 recall 역시 0.926로 상당히 높았기 때문에 accuracy가 높게 나타난 것으로 보인다.

Ⅴ. 결 론

본 연구에서는 그동안 기존 연구에서 미흡하게 다루어졌던 데이터셋을 구성하는 과정에서 발생하는 바이어스 문제, 거시경제위험 요소의 미반영 문제, 데이터 불균형 문제, 그리고 서술적 바이어스 문제가 예측 모델의 강건성과 확장성을 저해할 수 있음을 제기하고, 이에 대한 해결방안에 대해 논의하였다. 이를 위해 본 연구에서는 경기순환국면을 반영한 기업부실화 예측 프레임워크를 해결 방안으로 제시하였으며, 이를 토대로 기업부실화 예측 모델을 개발하였다.

그 결과, 다음과 같은 시사점을 얻게 되었다. 첫째, 본 연구는 기업고유위험과 거시경제위험을 통합적으로 반영한 기업부실화 예측 모델로서 중도절단 문제의 해결과 경기순환국면에 따른 예측 모델의 유용성을 보여주었다. 둘째, 본 연구는 데이터 불균형 문제를 오버샘플링 기법인 SMOTE를 활용하여 다루었으며, 표본비율이 1:1에 가까워질수록 예측성과가 점차 향상됨을 보여주었다. 셋째, 기본 재무비율에서 파생하여 만든 IT 산출물인 재무상태변동성과 산업수준상태변동성을 예측 모델의 변수로 추가하여 서술적 바이어스 문제를 다룸으로써 예측성과를 향상시켰다. 마지막으로, 본 연구는 순환국면이 수축기일 때의 데이터셋을 활용하여 예측 모델을 만들고, 다가올 순환국면의 데이터셋을 테스트셋으로 가져와 예측하였다. 그 결과, 전체적인 예측률이 90% 이상을 기록하였으며, 특히, 수축기에서 수축기로의 예측성과가 비교적 높게 나타남을 보여주었다. 이와 같이, 본 연구는 비즈니스 인텔리전스 연구 측면에서 기존 연구에서 미흡하게 다루어졌던 4가지 문제점을 검토하고, 이를 해결할 프레임워크를 제안함으로써 기존 연구 대비 기업부실화 예측률을 10% 이상 향상시켰다는 점에서 연구의 의의를 가진다.

본 연구는 다음과 같은 한계점을 갖는다. 본 연구에서는 경기순환국면을 기업부실화 예측 모델에 반영하기 위해 경기순환국면이란 지표를 기준으로 데이터셋을 구성하였다. 예를 들어, 제7순환 수축기는 2000년 8월부터 2001년 7월까지이다. 재무보고가 2001년 12월이라면, 본 연구에서는 다음 순환기인 제8순환 확장기 데이터셋으로 구성하였다. 그러나 보고일이 12월인 재무제표는

전년 11월부터 금년 11월까지의 경영활동을 담고 있다. 특히, 유가상장 된 기업들의 재무보고일은 3월, 9월, 12월에 집중되어 있기 때문에 경기순환 국면 기간과 재무보고일을 정확하게 일치시키지 못한 한계점이 발생하였다. 또한 산업수준상태변동성을 만드는 단계에서 한국표준산업분류에 따라 산업군을 제조업과 비제조업으로 구분하였다. 그러나 실제로 기업의 활동은 산업과 시장에 따라 더욱 세분화되어 있기 때문에 향후 이를 고려한 산업수준상태변동성을 새롭게 만들어 변수에 추가하는 작업이 진행되어야 한다.

참 고 문 헌

[1] 김건우, "재무비율로 판단한 기업 부실 징후와 예측", *경영학연구*, 제16권, 제2호, 1987, pp. 263-316.

[2] 김경민, 장하영, 장병탁, "불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법", *정보과학회 컴퓨팅의 실제 논문지*, 제20권, 제10호, 2014.

[3] 김경재, 한인구, "퍼지신경망을 이용한 기업부도예측", *한국지능정보시스템학회논문지*, 제7권, 제1호, 2001, pp. 135-147.

[4] 김나라, 신경식, 안현철, "부도예측을 위한 확신 기반의 선택 접근법에서 앙상블 멤버 사이즈의 영향에 관한 연구", *지능정보연구*, 제19권, 제2호, 2013, pp. 55-71.

[5] 김명종, "기업부실화 예측데이터의 불균형 문제 해결을 위한 앙상블 학습", *지능정보연구*, 제15권, 제3호, 2009, pp. 1-15.

[6] 김명종, "유전자 알고리즘을 이용한 분류자 앙상블의 최적 선택", *지능정보연구*, 제16권, 제4호, 2010, pp. 99-112.

[7] 김명종, "회사채 신용등급 예측을 위한 SVM 앙상블학습", *지능정보연구*, 제18권, 제2호, 2012, pp. 29-45.

[8] 김명종, 강대기, "부스팅 인공신경망학습의 기업부실예측 성과비교", *한국정보통신학회 논문지*, 제14권, 제1호, 2010, pp. 63-69.

[9] 김복만, 박철용, 오종근, 윤석철, *New 경영분석*, 대경, 2010.

[10] 김성규, 이화득, "경기변동에 따른 비상장 중소기업 신용위험을 설명하는 회계변수와 부도예측에 관한 실증연구", *회계저널*, 제21권 제6호, 2012, pp. 287-323.

[11] 김성태, 강충오, 이필상, "기업별 거시위험 민감도를 반영한 부도상관관계의 측정", *한국증권학회지*, 제39권, 제1호, 2010, pp. 31-57.

[12] 김은나, 이성건, 최종후, "경영경제논문: 목표범주가 희귀한 자료의 과대표본추출에 대한 연구", *응용통계연구*, 제24권, 제3호, 2011, pp. 477-484.

[13] 김진화, 홍광헌, 민진영, "지식 누적을 이용한 실시간 주식시장 예측", *지능정보연구*, 제17권, 제4호, 2011, pp. 109-130.

[14] 김태훈, 안현철, "부도예측 개선을 위한 하이브리드 언더샘플링 접근법", *지능정보연구*, 제21권, 제2호, 2015, pp. 173-190.

[15] 민성환, "부도 예측을 위한 앙상블 분류기 개발", *한국산업정보학회*, 제17권, 제7호, 2012, pp. 139-148.

[16] 박정식, 신동령, *제4판 경영분석*, 다산출판사, 2006.

[17] 배재권, "Voting 알고리즘과 인공신경망을 이용한 부도예측을 위한 통합알고리즘", *한국비즈니스리뷰*, 제3권, 제2호, 2010, pp. 79-101.

[18] 신택수, 홍태호, "AdaBoost 알고리즘 기반 SVM 을이용한 부실 확률분포 기반의 기업신용평가", *지능정보연구*, 제17권, 제3호, 2011, pp. 25-41.

[19] 안현철, "유전자 알고리즘을 이용한 다분류 SVM의 최적화: 기업신용등급 예측에의 응용", *Information Systems Review*, 제16권, 제3호, 2014, pp. 161-177.

[20] 오상훈, "다층퍼셉트론에 의한 불균형 데이터

의 학습 방법", *한국콘텐츠학회논문지*, 제9권, 제7호, 2009, pp. 141-148.

[21] 이치송, "거시경제변수와 산업별 신용위험에 관한 연구", *산업경제연구*, 제18권, 제1호, 2005, pp. 79-99.

[22] 장영민, 변재권, "거시경제변수와 대출부도의 시간변화에 따른 상관관계 연구", *금융연구*, 제24권, 제1호, 2010, pp. 131-160.

[23] 정한나, 이정화, 전치혁, "불균형 이분 데이터 분류분석을 위한 데이터마이닝 절차", *대한산업공학회지*, 제36권, 제1호, 2010, pp. 13-21.

[24] 조성표, 류인규, "불황기에서 회계정보에 의한 기업 부실화 예측", *경영연구*, 제22권, 제1호, 2007, pp. 1-132.

[25] 최소윤, 안현철, "퍼지이론과 SVM 결합을 통한 기업부도예측 최적화", *디지털융복합연구*, 제13권, 제3호, 2015, pp. 155-165.

[26] 허 준, 김종우, "불균형 데이터 집합에서의 의사결정나무 추론: 종합병원의 건강 보험료 청구 심사 사례", *Information Systems Review*, 제9권, 제1호, 2007, pp. 45-65.

[27] Abbasi, A., C. Albrecht, A. Vance, and J. Hansen, "Meta-fraud: A meta-learning framework for detecting financial fraud", *MIS Quarterly*, Vol.36, No.4, 2012, pp. 1293-1327.

[28] Abbasi, A., Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker, "Detecting fake websites: The contribution of statistical learning theory", *MIS Quarterly*, Vol.34, No.3, 2010, pp. 435-461.

[29] Altman, E. I., *Corporate Financial Distress*, New York: John Wiley & Sons, 1983.

[30] Altman, E. I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The Journal of Finance*, Vol.23, No.4, 1968, pp. 589-609.

[31] Baskerville, R. L., M. Kaul, and V. C. Storey. "Genres of inquiry in design-science research: Justification and evaluation of knowledge production", *Mis Quarterly*, Vol.39, No.3, 2015, pp. 541-564.

[32] Beaver, W. H., "Financial ratios as predictors of failure", *Journal of Accounting Research*, Vol.4, 1966, pp. 71-111.

[33] Brazdil, P., C. Giraud-Carrier, C. Soares, and R. Vilalta, *Meta-learning: App.lications to Data Mining*, Berlin: Springer-Verlag, 2008.

[34] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, Vol.16, 2002, pp. 321-547.

[35] Chuang, C. L., "Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction", *Information Sciences*, Vol.236, 2013, pp. 174-185.

[36] Fedorova, E., E. Gilenko, and S. Dovzhenko, "Bankruptcy prediction for Russian companies: Application of combined classifiers", *Expert Systems with Applications*, Vol.40, No.18, 2013, pp. 7285-7293.

[37] Hevner, A. R., S. T. March, J. Park, and S. Ram, "Design science in information systems research", *MIS Quarterly*, Vol.28, No.1, 2004, pp. 75-105.

[38] Hill, C. W. and S. L. McShane, *Principles of management*, McGraw-Hill/Irwin, 2008.

[39] Hsieh, N. C., "Hybrid mining approach in the design of credit scoring models", *Expert Systems with Applications*, Vol.28, No.4, 2005, pp. 655-665.

[40] Kaminski, K. A., T. S. Wetzel, and L. Guan, "Can financial ratios detect fraudulent financial reporting?", *Managerial Auditing Journal*, Vol.19, No.1, 2004, pp. 15-28.

[41] Ligang, Z., "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods", *Knowledge Based Systems*, Vol.41, 2013, pp. 16-25.

[42] Lin, J. W., M. I. Hwang, and J. D. Becker, "A fuzzy neural network for assessing the risk of fraudulent financial reporting", *Managerial Auditing Journal*, Vol.18, No.8, 2003, pp. 657-665.

[43] March, S. T. and G. Smith, "Design and natural science research on information technology", *Decision Support Systems*, Vol.15, No.4, 1995, pp. 251-266.

[44] McKee, T. E. and T. Lensberg, "Genetic programming and rough sets: A hybrid approach to bankruptcy classification", *European Journal of Operational Research*, Vol.138, No.2, 2002, pp. 436-451.

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

[46] Pederzoli C. and C. Torricelli, "Capital requirements and business cycle regimes: Forward-looking modelling of default probabilities", *Journal of Banking & Finance*, Vol.29, No.12, 2005, pp. 3121-3140.

[47] Porter, M. E., *Competitive Strategy*, New York, Free Press, 1998.

[48] Sánchez-Lasheras, F., J. de Andrés, P. Lorca, and F. J. de Cos Juez, "A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy", *Expert Systems with Applications*, Vol.39, No.8, 2012, pp. 7512-7523.

[49] Simon, H. A., *The Sciences of the Artificial* (3rd ed.), MIT Press, Cambridge, MA, 1996.

[50] Suarez, J. and O. Sussman, "Financial distress, bankruptcy law and the business cycle", *Annals of Finance*, Vol.3, No.1, 2007, pp. 5-35.

[51] Summers, S. L. and J. T. Sweeney, "Fraudulently misstated financial statements and insider trading: An empirical analysis", *The Accounting Review*, Vol.73, No.1, 1998, pp. 131-146.

[52] Sundarkumar, G. G. and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance", *Engineering Applications of Artificial Intelligence*, Vol.37, 2015, pp. 368-377.

[53] Tsai, C. F. and Y. F. Hsu, "A meta learning framework for bankruptcy prediction", *Journal of Forecasting*, Vol.32, No.2, 2013, pp. 167-179.

[54] Tsakonas, A., G. Dounias, M. Doumpos, and C. Zopounidis, "Bankruptcy prediction with neural logic networks by means of grammar-guided genetic programming", *Expert Systems with Applications*, Vol.30, No.3, 2006, pp. 449-461.

[55] Wang, G., J. Ma, and S. Yang, "An improved boosting based on feature selection for corporate bankruptcy prediction", *Expert Systems with Applications*, Vol.41, No.5, 2014, pp. 2353-2361.

[56] Wang, H., W. Fan, P. S. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers", In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 226-235.

[57] Witten, I. H., E. Frank, and A. H. Mark, *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed), The Morgan Kaufmann Series in Data Management Systems, Elsevier Inc, 2011.

[58] Zmijewski, M. E., "Methodological issues related to the estimation of financial distress prediction models", *Journal of Accounting Research*, Vol.22, No.1, 1984, pp. 59-82.

〈부록 1〉 경기순환국면별 경제동향

구 분	순환국면	경제동향
제7순환 (1998. 8~2001. 7)	수축기 (2000. 8~2001. 7)	- 세계 IT 경기 침체로 인한 해외수요의 감소로 IT 제품의 수출이 부진하여 경기 수축 - 설비투자의 급격한 위축과 소비 둔화도 경기 수축에 기여
제8순환 (2001. 7~2005. 4)	확장기 (2001. 7~2002. 12)	- 가계대출 확대 등 내수경기부양책으로 소비가 증가하여 경기회복 - 2002년 하반기에는 내수 증가세가 다소 둔화되었지만 수출의 회복으로 경기상승이 지속
	수축기 (2002. 12~2005. 4)	- 2003년은 소비와 투자의 침체, 신용 및 투신사 유동성 위기 등으로 경제성장 둔화 - 2004년 수출호조에도 불구하고 소비·투자 등의 내수부문이 계속 부진하여 경기회복이 지연
제9순환 (2005. 4~2009. 2)	확장기 (2005. 4~2008. 1)	- 민간소비 등의 내수가 침체에서 벗어나 수출호조와 함께 완만한 경기회복 - 2007년에는 미국발 서브프라임 사태에도 불구하고 수출상승이 지속되었으며 소비·설비투자도 호조
	수축기 (2008. 1~2009. 2)	- 2008년 상반기는 수출은 호조가 지속되었으나, 유가급등 등으로 인해 내수부문에서 경기둔화 가시화 - 글로벌 금융위기 발생에 따른 세계 경기침체 여파로 '08년 4/4분기이후 수출, 내수 모두 급락
제10순환 (2009. 2~)	확장기 (2009. 2~2011. 8)	- 정책당국의 경기부양책 등으로 2009년 2/4분기부터 내수를 중심으로 경기가 빠르게 회복되기 시작하였고 2009년 하반기 이후에는 수출도 상승세

자료: 통계청, 보도자료-산업활동동향(http://kostat.go.kr).

〈부록 2〉 기업부실화 예측 변수

변 수		변 수	
성장성	총자본증가율	활동성	총자본회전률
	영업이익증가율		자기자본회전률
	순이익증가율		타인자본회전률
	자기자본증가율		유동자산회전률
	매출액증가율		당좌자산회전률
	종업원수증가율		재고자산회전률
수익성	매출액총이익률		매출채권회전률
	매출액영업이익률		순운전자본회전률
	매출액순이익률		운전자본회전률
	총자산영업이익률		
	총자산순이익률	생산성	종업원 1인당 부가가치
	자기자본영업이익률		노동장비율
	자기자본순이익률		기계장비율
	금융비용부담률		자본집약도
	수지비율		총자본투자효율
	사내유보 대 자기자본비율		설비투자효율

변 수		변 수	
	1주당 매출액(원)		기계투자효율
	주당순이익		부가가치율
	주당현금흐름		노동소득분배율
	주당순자산가치		
	유보율		
안정성	자기자본구성비율		
	유동비율		
	당좌비율		
	현금비율		
	재고자산 대 순운전자본비율		
	매출채권 대 매입채무비율		
	부채비율		
	이자보상비율(이자비용)		
	Cash Flow 대 부채비율		
	Cash Flow 대 차입금비율		
	Cash Flow 대 총자본비율		
	Cash Flow 대 매출액비율		

〈부록 3〉 각 순환국면에서 Information gain값을 산출하여 순위화한 상위 10순위 변수들

<부록 3>에서 Information gain값을 산출하여 순위화한 상위 10순위 변수들을 보면, 재무비율 R2 = 매출액순이익률, R5 = 자기자본순이익률, R4 = 총자산순이익률, R7 = 총자본투자효율과 이들에서 파생된 IT 산출물(가령, 산업수준상태변동성: Top-10 Gap Model의 R2-T2나 산업수준상태변동성: Average Gap Model의 R5/A5 등)이 예측 모델에 유용한 변수로 나타난 것을 볼 수 있다.

Rank	제7순환 수축기	제8순환 확장기	제8순환 수축기	제9순환 확장기	제9순환 수축기	제10순환 확장기
1	R2-T2	R5/A5	R7-T7	R5	R2-A2	R2-A2
2	R2	R5	R7/T7	R5-A5	R2	R2-T2
3	R2-A2	R5/T5	R7/A7	R4	R2-T2	R2
4	R5-T5	R5/A5	R2-A2	R5/T5	R5	R5-A5
5	R4	R5-T5	R2-T2	R4-T4	R5-A5	R5
6	R4-A4	R2	R2/T2	R4/T4	R5-T5	R4-T4
7	R2/T2	R2-A2	R2	R5-T5	R2/A2	R5/T5
8	R5	R2/T2	R7-R7	R2-A2	R2/T2	R5-T5
9	R4/T4	R2-T2	R4-A4	R7-A7	R5/A5	R4/A4
10	R5/A5	R4	R4-T4	R2	R4/T4	R2/T2
…	…	…	…	…	…	…

주) R = 재무비율, A = 산업 평균, T = 산업 TOP 10 기업들의 평균.

Information Systems Review
Volume 18 Number 2
June 2016

Development of Prediction Model of Financial Distress and Improvement of Prediction Performance Using Data Mining Techniques

Raynghyung Kim* · Donghee Yoo** · Gunwoo Kim***

Abstract

Financial distress can damage stakeholders and even lead to significant social costs. Thus, financial distress prediction is an important issue in macroeconomics. However, most existing studies on building a financial distress prediction model have only considered idiosyncratic risk factors without considering systematic risk factors. In this study, we propose a prediction model that considers both the idiosyncratic risk based on a financial ratio and the systematic risk based on a business cycle. Ultimately, we build several IT artifacts associated with financial ratio and add them to the idiosyncratic risk factors as well as address the imbalanced data problem by using an oversampling technique and synthetic minority oversampling technique (SMOTE) to ensure good performance. When considering systematic risk, our study ensures that each data set consists of both financially distressed companies and financially sound companies in each business cycle phase. We conducted several experiments that change the initial imbalanced sample ratio between the two company groups into a 1:1 sample ratio using SMOTE and compared the prediction results from the individual data set. We also predicted data sets from the subsequent business cycle phase as a test set through a built prediction model that used business contraction phase data sets, and then we compared previous prediction performance and subsequent prediction performance. Thus, our findings can provide insights into making rational decisions for stakeholders that are experiencing an economic crisis.

Keywords: Financial Distress, Business Cycle, Data Mining, Imbalanced Data, Declarative Bias, IT Artifacts

* Department of Business Administration, Hanbat National University
** Department of Management Information Systems, Gyeongsang National University, BERI
*** Corresponding Author, Dept. of Business Administration, Hanbat National University

◐ 저 자 소 개 ◑

김 량 형 (krh419@gmail.com)

한밭대학교에서 경영학 석사를 전공하고, 동 대학원에서 박사 학위 과정 중이며, 현재 한국노동연구원 임금직무혁신센터 R.A로 일하고 있다. 연구 관심분야는 빅데이터 분석, 비즈니스 인텔리전스, 데이터마이닝, 통계학, 소셜네트워크 분석 등이다.

유 동 희 (dhyoo@gnu.ac.kr)

현재 경상대학교 경영정보학과 조교수로 재직하고 있다. 고려대학교 경영정보학과를 졸업하고, 고려대학교 경영학과에서 경영학 박사를 취득하였다. 주요 관심분야는 데이터마이닝, 시맨틱 웹, 온톨로지, 지능정보시스템 등이다.

김 건 우 (gkim@hanbat.ac.kr)

현재 대전에 소재한 국립한밭대학교에서 경영회계학과 부교수로 재직하고 있다. 연세대학교 공과대학에서 컴퓨터 사이언스를 전공하였으며 고려대 경영학과에서 석사를 졸업하고 동대학에서 박사 학위를 수여하였다. 현재 한국창업학회 부회장을 맡고 있으며 ICT플랫폼학회 빅데이터분과 위원장을 맡고 있다. 그 외 다수의 학회에서 편집위원 및 이사로써 활동하고 있다. 주요 관심분야는 비즈니스 온톨로지 모델, 빅데이터 분석 및 핀테크 기술 및 전략 등이다.

논문접수일 : 2016년 05월 17일 **게재확정일 : 2016년 06월 15일**
1차 수정일 : 2016년 06월 10일