빅데이터와 인공지능 기법을 이용한 기업부도예측연구

2017년 12월 26일
최정원, 오세경,장재원
발행처 한국금융연구원

나의 부도 일지

팀소개

나의 부도일지

LEADER 황윤재
FOLLOWER 김태환
FOLLOWER 박희연
FOLLOWER 장동민

발표 목차

논문 요약

Word2vec

뉴스 텍스트 → 조기 경보 모형

논문 요약

본 연구는 기업 부도 예측 과정에서 새로운 정보 원천으로 비정형 데이터인 뉴스 텍스트 데이터를 계량화하여 활용할 수 있도록 인공지능 기법인 'Word2vec' 방법으로 측정하는 방법을 제시한다. 또한 인공지능 기반의 예측 방법론을 제시하고 기존의 방법론과 예측력을 비교 분석하였다. 연구 결과, 우선 연간 모형에서는 인공지능 기법인 Random forests 기법이 가장 우수한 예측력이 나타나는 것으로 분석되었다. 또한 인공지능을 이용한 다른 방법론들도 전반적으로 기존의 전통적인 예측 방법보다 예측력이 우수한 것으로 나타났다. 뉴스 텍스트를 추가적인 정보 원천으로 추가한 효과는 연간 예측 모형에서는 다소 미미하였다. 하지만 월간 예측 모형에서는 텍스트 정보 기반의 예측 모형이 시장 정보 기반의 예측 모형인 KMV 모형과 유사한 결론을 도출할 수 있어 기업 부도 예측 과정에서 조기 경보 모형으로 충분히 활용이 가능함을 실증하였다.

논문 연구 배경

재무정보는 분기 혹은 연단위로 작성되고 기업의 결산시점이후 공시 되는데 까지 일정기간이 소요되므로 **적시성이 떨어지는 한계점**이 있다.

과거 활용이 어려웠던 텍스트 형태의 **비정형 정보인 뉴스** 정보를 활용하여 기업의 부도 예측 수준이 향상될수 있는지 연구하였다.
뉴스정보는 기업에 관한 가장 빠른 정보 중 하나이다.

논문 선정 이유

재무 비율만으로는 빠르게 변화하는 기업경영환경을 반영하기에 어려움이 있다.

적시성 문제를 해결하기 위해 뉴스 기사활용

비상장 기업에도 사용가능한 비재무 정보활용

논문 연구 배경

재무정보는 분기 혹은 연단위로 작성되고 기업의 결산시점이후 공시 되는데 까지 일정기간이 소요되므로 **적시성이 떨어지는 한계점**이 있다.

과거 활용이 어려웠던 텍스트 형태의 비정형 정보인 뉴스정보를 활용하여 기업의 부도 예측 수준이 향상될수 있는지 연구하였다.
(뉴스정보는 기업에 관한 가장 빠른 정보 중 하나이다.)

논문 선정 이유

재무 비율만으로는 빠르게 변화하는 기업경영환경을 반영하기에 어려움이 있다.

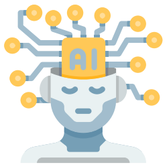
적시성 문제를 해결하기 위해 뉴스 기사활용

비상장 기업에도 사용가능한 **비재무 정보활용**

분야별
선행 연구

기업 부도 예측 연구

빅데이터 기법을 활용한 관련 분야 연구



인공지능 기법을 활용한 관련 분야 연구

기업부도 예측연구

1.알트만 모형- Financial Ratios,Discriminant Analysis and the Prediction of Corporate Bankruptcy(Altman,1968)

2.KMV 모형 -A Comment on Markent vs Accounting-Based Sentiment analysis in financial news (McQuown,1993)

3.해저드 모형 - Forecasting bankrupty more accurately: A simple hazard model(Shumway,2001)

빅데이터 기법을 활용한 관련 분야 연구

1. 텍스트마이닝 기법의 기술정보분석 적용 가능성 연구(배상진,박철균,2003)
- 텍스트 마이닝 과정을 4단계(문서 수집,문서 전처리, 텍스트 분석 그리고 결과 해석 및 정제 과정)로 나누어 설명

2.뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형(김유신,김남규,정승렬,2012)
- 뉴스 키워드 감성분석을 통해 시장대비 초과 수익률 얻을 수 있는 투자의사결정 모형 구축

3. 텍스트마이닝 방법론을 활용한 기업 부도예측연구(최정원,한호선,이미영,안준모,2015)
- 부도가 발생한 기업의 뉴스 텍스트 데이터를 텍스트마이닝기법으로 분석하여 기업 부도예측의 가능함을 시도하였음

인공지능 기법을 활용한 관련 분야 연구

1.인공신경망을 이용한 중소기업 도산 예측에 있어서의 비재무정보의 유용성 검증(이재식,한재홍,1995)
-재무정보로만 활용했을 때의 단점을 보완하기 위하여 비재무정보를 활용한 **인공신경망 기반**의 부도예측 모형을 제시

2. 기업신용등급 예측을 위한 랜덤포레스트의 응용(김성진,안현철,2016)
- 전통적인 기업 신용등급예측 방법론과 비교하여 **랜덤포레스트 방법론**의 성능이 우수함을 실증 분석함

데이터 소개

부도사건의 정의

유가증권시장에서 '**상장폐지**'가 결정된 기업들 중 부도에 관련된 공시가 발생한 기업들을 부도 발생기업으로 인식하였음.

부도 O : '부도발생', '화의절차개시신청', '회사정리절차개시신청', '감사인의 의견 거절' , '은행 거래정지' 등

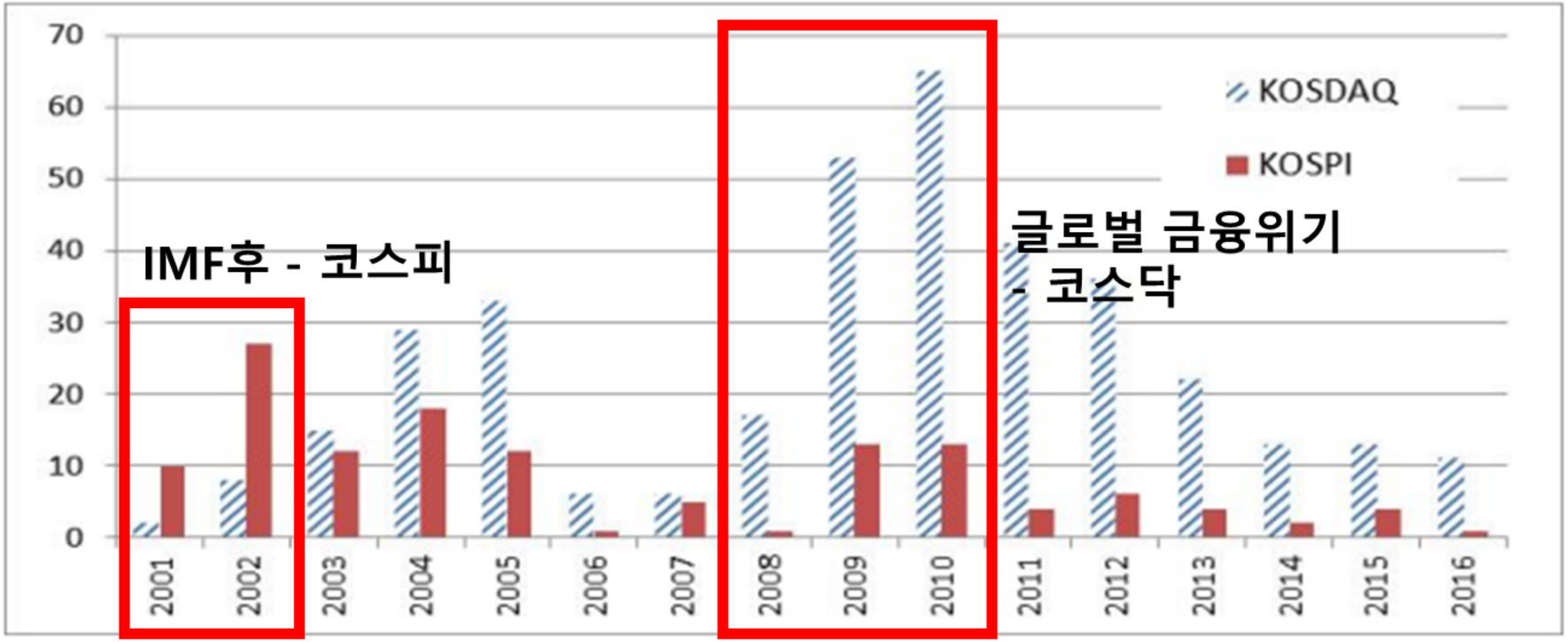
부도 X: '신규/변경 상장', '특수 목적에 의한 상장폐지', '기업 피인수' 등

분석기업데이터

시장구분	건실기업	부도기업	Total
KOSPI	678	133	**811**
KOSDAQ	1108	370	**1478**
Total	**1786**	**503**	**2289**

<표 5> 분석대상 기업

코스피 비율 83:17
코스닥 비율 75:25
총 비율 78:21



<그림 5> 연도별 부도기업 추이

분석에 활용한 데이터 종류 및 특징

구분	의의	활용가능 데이터
1. 재무 정보	기업 공시(재무제표) 정보 결산(연/분기)기준 재무비율	- 수익성: 자산(자본)대비 수익률 등 - 성장성: 매출증가율, 자산증가율 등 - 건정성: 부채비율, 이자보상배율 등 - 기타재무지표, 주주비율 등 기업정보
2. 시장 정보	상장 기업의 주식 거래 관련 정보	- 시장지표: 주가, 시가총액, 주가수익률, 거래량 - 재무비율 혼합지표 시장가 대비 장부가 비율 시장조정부채비율, 시장조정 등
3. 거시경제지표	주요 기관에서 집계&발표하는 거시경제지표	- 거시경제지표: 국가총생산(GDP), 통화량, 물가지수(PPI ,CPI), 기업경기실사지수(BSI) 등 - 금융시장지표: 금리, 종합주가지수, 변동성 지수 등
4. 비정형 정보	전통적인 방법으로 활용하기 어려웠던 비정형(텍스트) 데이터	- 뉴스 정보뉴스 및 방송잡지 등), - 공시자료, SNS(인터넷 사이트) 등의 정보 - 주로 텍스트 형태의 데이터로 확보

<표 1> 기업 부도예측을 위한 원천 정보 구분 및 특성

뉴스 데이터 수집매체

구 분	언론 매체
종합	경향신문, 국민일보, 뉴시스, 동아일보, 로이터, 문화일보, 서울신문, 세계일보, 연합뉴스, 조선일보, 중앙일보, 한겨레, 한국일보, JTBC, KBS, MBC, SBS, YTN
경제	뉴스토마토, 매일경제, 머니투데이, 서울경제, 아시아경제, 이데일리, 조선비즈, 파이낸셜뉴스, 한국경제, 한국경제TV, 헤럴드경제, MBN, SBSCNBC
온라인/인터넷	데일리안, 오마이뉴스, 쿠키뉴스

<표 2> 뉴스 텍스트 수집 대상 언론 매체

데이터 수집 및 정제

수집된 연도별 뉴스 기사

구분	Total	2010	2011	2012	2013	2014	2015	2016
기사 수	2,506,080	110,213	339,040	390,764	394,128	402,792	426,991	442,152
기업당 평균	1,402	62	190	219	220	225	239	247

<표 6> 총 뉴스 기사 수 연간 추이 및 합계

(a) 2010 년 전의 부도가 일어난 기업: **기사를 확보할 수 없음(기사확보 X)**

(b) Sample 수 부족: 분석대상기간(2010 년 ~ 2016 년) 동안 기사 수 **100 건 이하**

(c) 기업의 이름이 **일상적인 용어**와 같은 경우 (Ex: 전방, 청구, 부흥, 진도 등)

(d) 기타 해당 기업의 기사인지 정확하게 **확인할 수 없는 기업**

Word embedding

워드임베딩이란 쉽게 말해서 자연어를 컴퓨터가 처리할 수 있게 벡터화 하는 것

워드 임베딩을 (분산 표현: distributed representation)이라고도 부름

단어를 벡터로 변환해야하는 이유? 벡터로 바꾸어야 유사도 같은 계산이 가능함

단어를 벡터화 하는 방법

1. NNLM, RNNLM, Word2Vec, FastText : 예측 기반의 벡터화 방법
2. LSA, HAL : 카운트 기반의 벡터화 방법
3. GloVe : 예측 기반과 카운트 기반 두 가지 방법을 모두 사용하는 방법

Word2vec(Word to Vector)

Word2Vec은 2개의 히든레이어를 가지고 있는 뉴럴 네트워크(Neural Network) 모델이다.

단어가 많아져도 저차원벡터를 가지고 다차원 공간에 벡터화해서 유사성을 표현할 수 있다.

Ex) 강아지 = [0 0 0 0 1 0 0 0 0 0 0 0 ... 중략 ... 0]
(원핫 인코딩)

Ex) 강아지 = [0.2 0.3 0.5 0.7 0.2 ... 중략 ... 0.2]
(워드투벡터)

Word2vec(Word to Vector)

1. CBOW(Continuous Bag of Words)

주변에 있는 단어들을 입력하여 중간에 있는 단어들을 예측하는 방법

2. Skip-gram

중간에 있는 단어들을 입력하여 주변 단어들을 예측하는 방법

유사도 기준 상위 20 개 단어

Rank	'부도' 기준		'부도' & '상장폐지' 기준	
	word	유사도	word	유사도
1	도산	0.74	퇴출	0.63
2	파산	0.63	관리종목	0.62
3	경영난	0.60	파산	0.62
4	외환	0.60	도산	0.61
5	자금난	0.60	분식회계	0.60
6	법정관리	0.57	법정관리	0.57
7	어음	0.57	원리금	0.56
8	연체	0.55	잠식	0.56
9	워크아웃	0.54	연체	0.55
10	대출금	0.53	자금난	0.55
11	원리금	0.53	손실	0.54
12	폐업	0.53	매매거래	0.53
13	부실화	0.53	워크아웃	0.53
14	부실	0.52	부실	0.53
15	채무	0.50	기업회생	0.52
16	손실	0.49	감사보고서	0.52
17	몰락	0.48	대출금	0.52
18	제때	0.48	회생	0.52
19	기업회생	0.48	부실기업	0.51
20	속출	0.47	정지	0.51

<표 7> 'Word2vec' 유사도 산출 결과

부도 유사 단어

'부도' 와 특정 기준이상의 유사도를 나타나거나, 유사도 기준으로 순위(rank)를 부여하여 상위 단어들을 '부도 유사 단어'로 선정할 수 있다. 이후 선정된 부도 유사 단어 중 1 개라도 포함된 기사를 '부도 관련 기사'로 판별할 수 있다

- 부도 유사 단어(1): '부도' 단어와 'Word2vec' 유사도 상위 20 개 단어
- 부도 유사 단어(2): '부도' 와 '상장폐지' 단어와 동시 'Word2vec' 유사도 상위 20 개 단어
- 부도 유사 단어(3): 이전 연구에서 '부도'와 연관이 있음을 밝힌 단어 '회생', '공시', '자금', '횡령', '증자', '채권단', '워크아웃'

변수 – 부도 기사 비율

$$\text{부도 기사 비율}_{it} = \frac{\textit{부도 관련 기사 수}}{\textit{총 정상 기사 수}},$$

i= 기업, t= 분석 기간(월간, 부도발생 기준 직전 각 12 개월)

본 연구에서는 이러한 현상을 **계량적**으로 분석하기 위하여 '**부도 관련 기사 비율**'을 측정

기간 별로 전체 기사 중 부도와 관련된 기사의 비중을 산출하고, 이 비율이 높게 나타날 경우 이를 사전적인 '부도'의 징후로 판단하여 부도 예측에 활용하기로 함.

변수
부도 기사 비율

구분		Total	2010	2011	2012	2013	2014	2015	2016
부도연관 단어(1)	부도연관 기사 수	380,673	16,586	48,636	59,214	65,863	60,729	59,473	70,172
	부도기사 비율(1) 평균	15.19%	15.05%	14.35%	15.15%	16.71%	15.08%	13.93%	15.87%
부도연관 단어(2)	부도연관 기사 수	389,952	14,496	46,398	59,157	69,142	64,457	61,718	74,584
	부도기사 비율(2) 평균	15.56%	13.15%	13.69%	15.14%	17.54%	16.00%	14.45%	16.87%
부도연관 단어(3)	부도연관 기사 수	221,523	11,616	29,269	31,948	37,553	35,198	35,176	40,763
	부도기사 비율(3) 평균	8.84%	10.54%	8.63%	8.18%	9.53%	8.74%	8.24%	9.22%

'부도' 유사단어

'부도' +'상장폐지' 유사단어

'선행연구 단어'

<표 8> 부도연관기사 및 부도기사비율 연간 추이

변수- 부도 유사도

구분	Total	2010	2011	2012	2013	2014	2015	2016
부도유사도 ('부도')	0.0206	0.0124	0.0216	0.0276	0.0279	0.0296	0.0247	0.0206
부도유사도 ('부도' & '상장폐지')	0.0546	0.0309	0.0609	0.0730	0.0728	0.0749	0.0695	0.0546

<표 9> 부도연관기사 및 부도기사비율 연간 추이

- 부도 유사도(1):특정월의 해당기업의 기사를 구성하고 있는 모든 단어의 유사도 평균수준(단어 유사도 총합/단어 수)

- 부도 유사도(2):특정월의 해당 기업의 기사단위 유사도 평균(단어 유사도 총합/기사 수)

변수

후에 비교 및 선정

NumNeg_1	(뉴스 텍스트)	부도기사비율_1	연간 부도(w2v-부도) 기사수 / 연간 기사수
NumNeg_2		부도기사비율_2	연간 부도(w2v-부도&상폐) 기사수 / 연간 기사수
NumNeg_3		부도기사비율_3	연간 부도(선행연구 단어) 기사수 / 연간 기사수
w2v1_1		부도 유사도_1	연관도평균(w2v-부도)
w2v1_2		부도 유사도_2	연관도합계 (w2v-부도) / 기사 수
w2v2_1		<부도+상장폐지>유사도_1	연관도평균(w2v-부도&상장폐지)
w2v2_2		<부도+상장폐지>유사도_2	연관도합계 (w2v-부도&상장폐지) / 기사 수

이 과정에서 만든 변수 → 7개

데이터 셋

방법론	SET1	SET2	SET3	SET4
적용정보(Source)	재무정보	재무정보 +거시경제	재무정보 +거시경제+시장정보	재무정보 +거시경제+시장정보+텍스트정보
데이터수집가능기간	1998~2015년(연간)	1998~2015년(연간)	1998~2015년(연간/월간)	2010~2015년(연간/월간)
변수정보	31개 변수(21개 재무변수+10개 기업특성)	42개 변수(SET1+거시 11개)	49개변수(SET2+시장 7개)	53개 변수(SET3+텍스트4개)
이용가능 데이터 수	결측제외 총 33621개->2291기업(부도502개)	결측제외 총 30268개->2291기업(부도502)	결측제외 총 21402개->2291기업(부도502)	결측제외 총 9706 개 ->1586기업(부도258)

1) Set A: 재무, 시장, 거시경제 정보(2001~2016 년). 총 2291 개 (부도 502 개) 기업 대상
[SetA_1] / [SetA_2] / [SetA_3]

2) Set B: 재무, 시장, 거시경제 정보(2010~2016 년). 총 1586 개 (부도 258 개) 기업 대상
[SetB_1] / [SetB_2] / [SetB_3] / [SetB_4]

평가 지표

모형 평가 지표

		예측 범주		합 계
		1	0	
실제 범주	1	n_{11}	n_{10}	n_{1+}
	0	n_{01}	n_{00}	n_{0+}
합 계		n_{+1}	n_{+0}	n_{++}

정확도(Accuracy, 정분류율) = $(n_{11}+n_{00}) / n_{++}$

민감도(Sensitivity) = n_{11} / n_{1+}

특이도(Specificity) = n_{00} / n_{0+}

<표 4> 이진분류 모형의 예측 정확도 지표 산출방법

민감도,특이도 증가

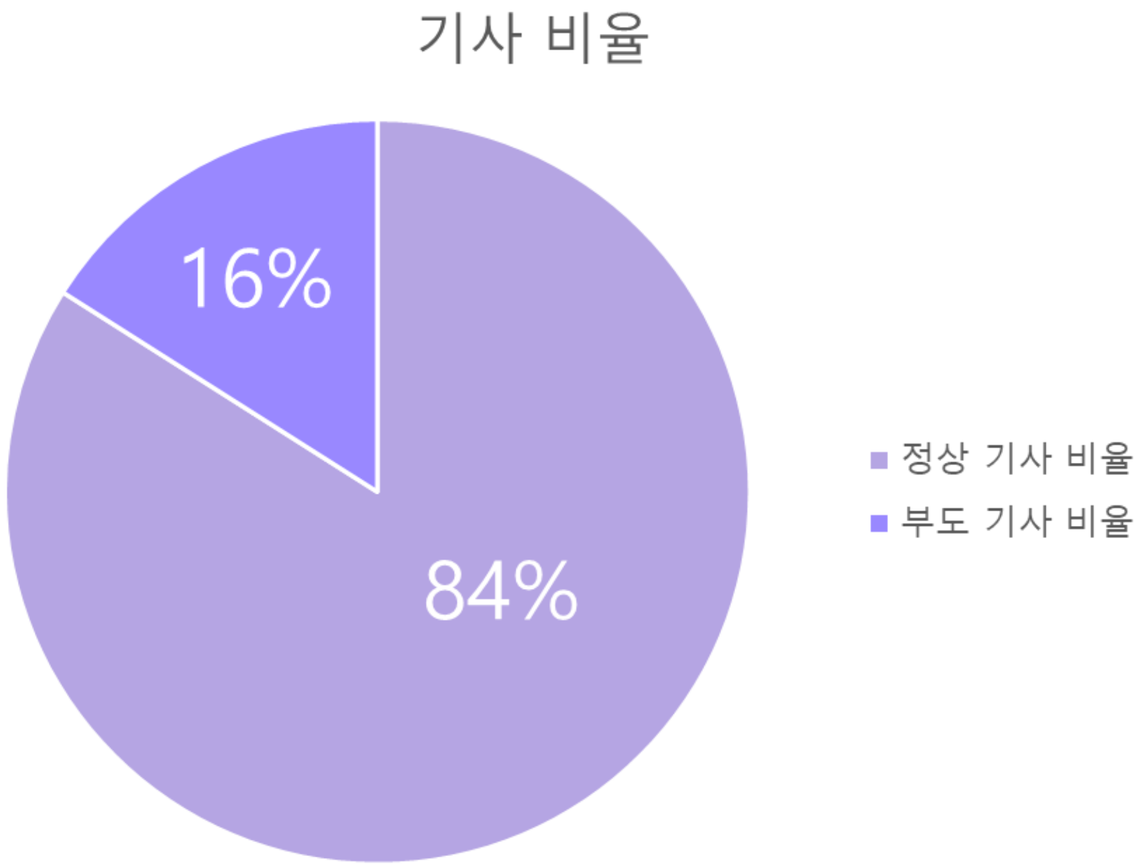
↓

AUC 증가

↓

정확도 증가

모형 평가 강건성 증대 방안



Random Under Sampling
(50% vs 50%)

100회 반복

모든 평가 세트
정확도 평균 산출

연간 예측 모형

연간 예측 모형

재무정보를 포함하는 기업 부도 예측 모형은
연간 단위로 예측을 수행하여야 한다.

부도 여부(1: 부도, 0: 정상)을 목표(Target) 변수로 하여
각 방법론을 활용하여 예측 모형을 구성하였다

방법론별 최적 예측 모형 도출

방법론	세부 적용 방법론 및 가정	산출(fitting) 및 모형 평가 방법
1. 로지스틱 (Logit)	다중회귀분석 모형 (Stepwise) Engine: R (glm)	• Cross-section 형태의 분석 방법이므로 시점별(t-1,2,3) 변수를 모두 설명변수로 각각 적용 • 변수가 많아 과다 적합 문제 발생 가능 → Stepwise 로 변수 선택 적용 • F-value(P-value) 및 R^2 로 모형 평가
2. Cox-PH Hazard (Cox)	Cox PH 모형(다중회귀, 층화, Stepwise) Engine: R (survival)	• 주가, 거시경제, 비정형정보 등 Hazard 함수 설명 변수로 반영 가능 • 산업별 생존함수를 추정하여 산업별 특성 반영 • 변수 선택(Stepwise) 필요 • F-value(P-value) 및 R^2 로 모형 평가
3. Decision Tree (Dtree)	Max maxsurrogate(노드 수): 3 단계 Engine: R (Dtree)	• 비교 모형으로 활용 • Accuracy 로 사후적 모형평가
4. Random-Forest (RF)	Sampling 을 통한 paramenter 최적화 Engine: R (e1071)	• 다양한 설정 값 시뮬레이션 • Accuracy 로 사후적 모형평가
5. SVM	Sampling 을 통한 paramenter 최적화 Engine: R (e1071)	• 다양한 설정 값 시뮬레이션 • Accuracy 로 사후적 모형평가
6. 인공신경망 (DNN)	Deep 구조: 512 EU * 8 Layer Activation Function: ReLU 초기값 설정: Xavier initializer[15] Engine: Python (TensorFlow)	• Cost 함수(평균예측오차): $\frac{(실제값-예측값)}{평가\ 횟수}$ → 학습횟수 2 만 or Cost 기준 0.1 이하 까지
7. 인공신경망 (RNN)	Deep 구조: 3 기간(LSTM Cell) 적용 Activation Function: ReLU 초기값 설정: Xavier initializer Engine: Python (TensorFlow)	• Cost 함수(평균예측오차): $\frac{(실제값-예측값)}{평가\ 횟수}$ → 학습횟수 2 만 or Cost 기준 0.1 이하 까지

<표 11> 각 모형의 세부 적용 방안 및 산출 모형 적합도 평가 방법

DNN 체계의 은닉층 구조 – 1 열 8 개층 (layer)중첩 구조

RNN – 3 기간 10 개층(layer)구조

연간 예측 모형 성과 분석

SET A 결과 (분석기간 2001 년~2016 년 적용)

방법론	SET A_1	SET A_2	SET A_3	평균
logit	0.9258 0.0146	0.9208 0.0153	0.9272 0.0142	**0.9246**
Cox	0.7798 0.0183	0.7033 0.0237	0.7115 0.0199	**0.7315**
Dtree	0.8998 0.0183	0.8984 0.0179	0.8956 0.0180	**0.8979**
R.F	0.9357 0.0133	0.9350 0.0127	0.9381 0.0125	**0.9363**
SVM	0.9217 0.0153	0.9082 0.0179	0.9212 0.0226	**0.9170**
DNN	0.8533 0.0200	0.8584 0.0184	0.9052 0.0148	**0.8723**
RNN	0.8867 0.0210	0.9065 0.0232	0.9046 0.0279	**0.8992**
평균	**0.8861**	**0.8758**	**0.8862**	

<표 12> 모형별 예측 정확도 산출 결과(SET A)[16]

연간 예측 모형 성과 분석

SET B 결과 (분석기간 2010 년~2016 년 적용)

방법론	SET B_1	SET B_2	SET B_3	SET B_4	평균
logit	0.8651 0.0427	0.8804 0.0410	0.8989 0.0383	0.9093 0.0338	**0.8884**
Cox	0.8280 0.0312	0.8235 0.0335	0.8473 0.0335	0.8745 0.0282	**0.8433**
Dtree	0.8910 0.0293	0.8895 0.0288	0.8868 0.0274	0.8862 0.0271	**0.8884**
R.F	0.9369 0.0224	0.9373 0.0226	0.9381 0.0225	0.9392 0.0222	**0.9379**
SVM	0.9217 0.0273	0.9148 0.0263	0.9271 0.0278	0.9178 0.0282	**0.9203**
DNN	0.9071 0.0285	0.9053 0.0282	0.9215 0.0286	0.9317 0.0299	**0.9164**
평균	**0.8916**	**0.8918**	**0.9033**	**0.9098**	

<표 13> 모형별 예측 정확도 산출 결과(SET B)

인공신경망(RNN) 의 경우 분석 과정에 3 개년 연속된 데이터가 필요한 데, 이럴 경우 Set B 는 Data Sample 수의 손실이 너무 심해서 유효한 분석이 어렵다. 따라서 <Set B> 분석에서는 인공신경망-RNN 은 제외하고 분석하였다.

연간 예측 모형 성과 분석

분석결과 인공지능 중 RANDOM FORESTS 방법론이 두 데이터 SET 모두 가장 높은 수준의 예측력을 보여주었다. 특히 데이터 수가 상대적으로 적은 SET B에서도 우수한 예측력을 유지 함으로서 인공지능 기법이 강건하게 기업의 부도에 대한 예측을 잘 수행할 수 있음을 실증하는 결과이다

재무정보를 제외한 다른 데이터의 추가가 예측 정확도에 큰 영향을 주지 못했다.
이는 상장 기업의 경우 다양한 공시 요구 및 규제에 의하여 기업의 정보가 재무정보에 이미 충분히 반영되어 나타나는 결과라 판단된다.

인공지능(DNN)을 적용한 결과를 보면 SAMPLE 데이터 수가 많은 SET A에 비하여 SET B의 예측 정확도가 오히려 높게 나오는 현상이 발생하였다. 이는 과적합으로 인해 오히려 예측력이 떨어지는 현상이 나타난 것으로 추정된다.

월간 예측 모형

월간 예측 모형

미디어의 뉴스 기사는 시장 정보(주가)와 마찬가지로 실시간으로 공개되는 정보

1) 대상기간: 2010 년~2016 년 (텍스트 DB 확보 가능 기간)

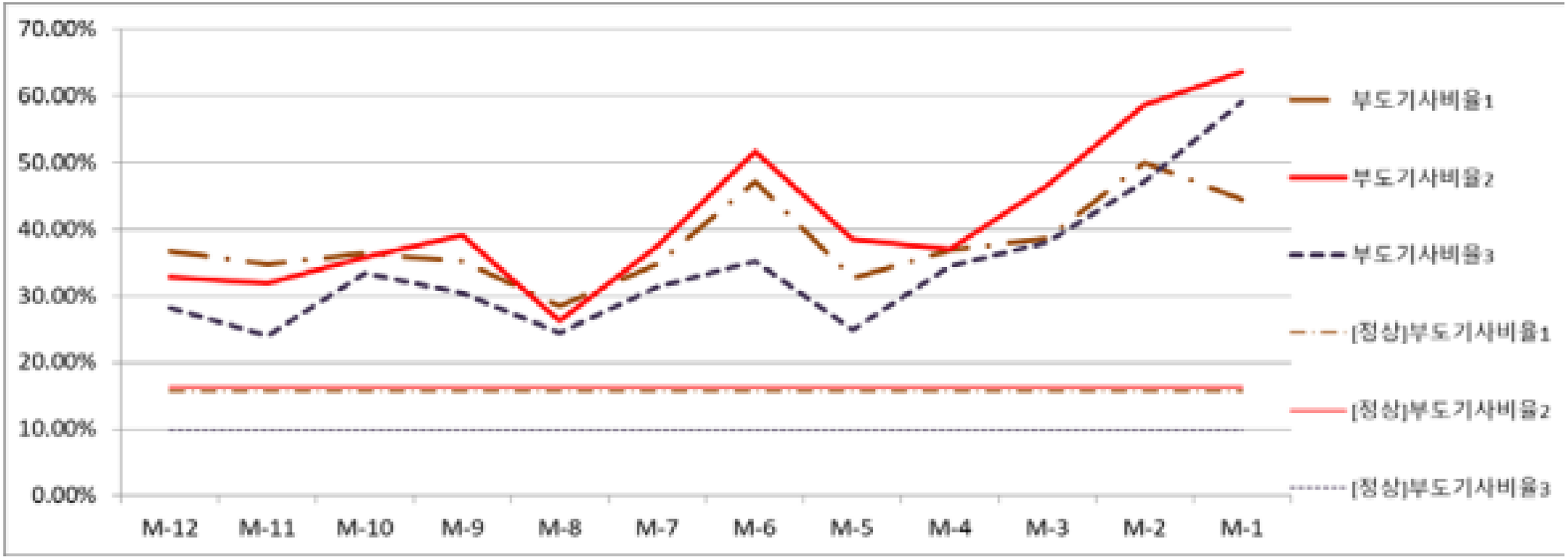
2) 기사 수 기준: 대상기간 동안 총 기사 수 합계 100 건 이상

비교대상 : KMV모형

KMV 모형 산출 결과

<그림 6> 부도 발생 12 개월 전 D.D. 평균 추이 [18]

텍스트 정보기반 예측 모형 산출 결과

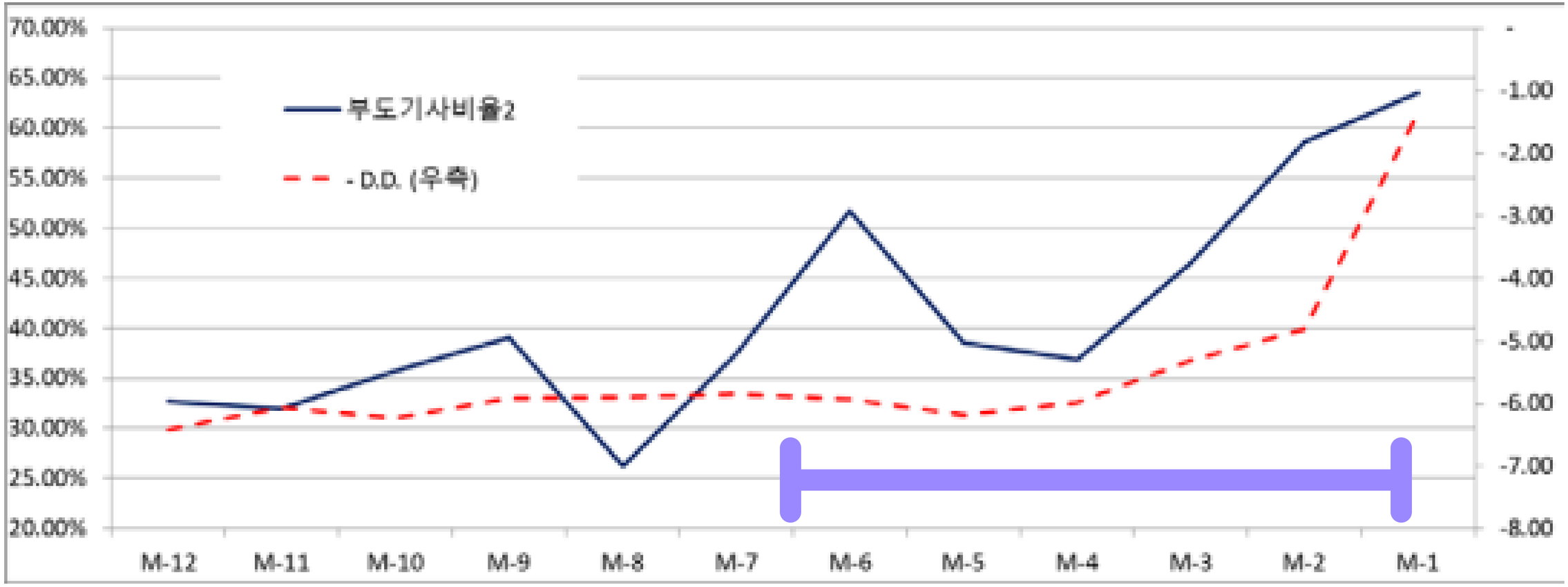


<그림 7> 부도 발생 12 개월 전 부도 기사 비율 추이

부도기사 비율 중에는 '부도'와 '상장폐지'를 동시에 'Word2vec'을 활용하여 상위 20 개 단어가 포함된 기사를 부도 기사로 간주한 [부도기사 비율 2] 가 정상 수준에 대비하여 가장 유의한 차이를 보이고 있다

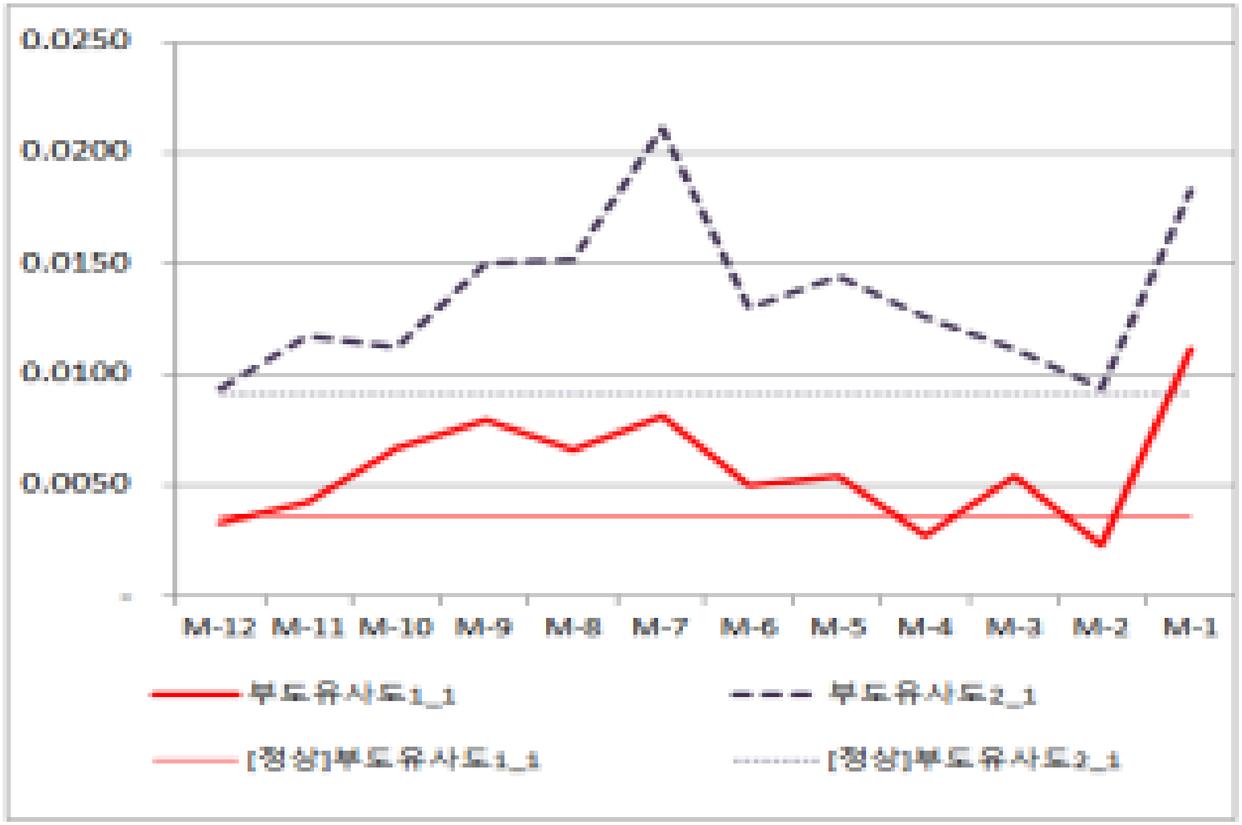
KMV 와 텍스트 정보기반 예측 모형 비교 결과

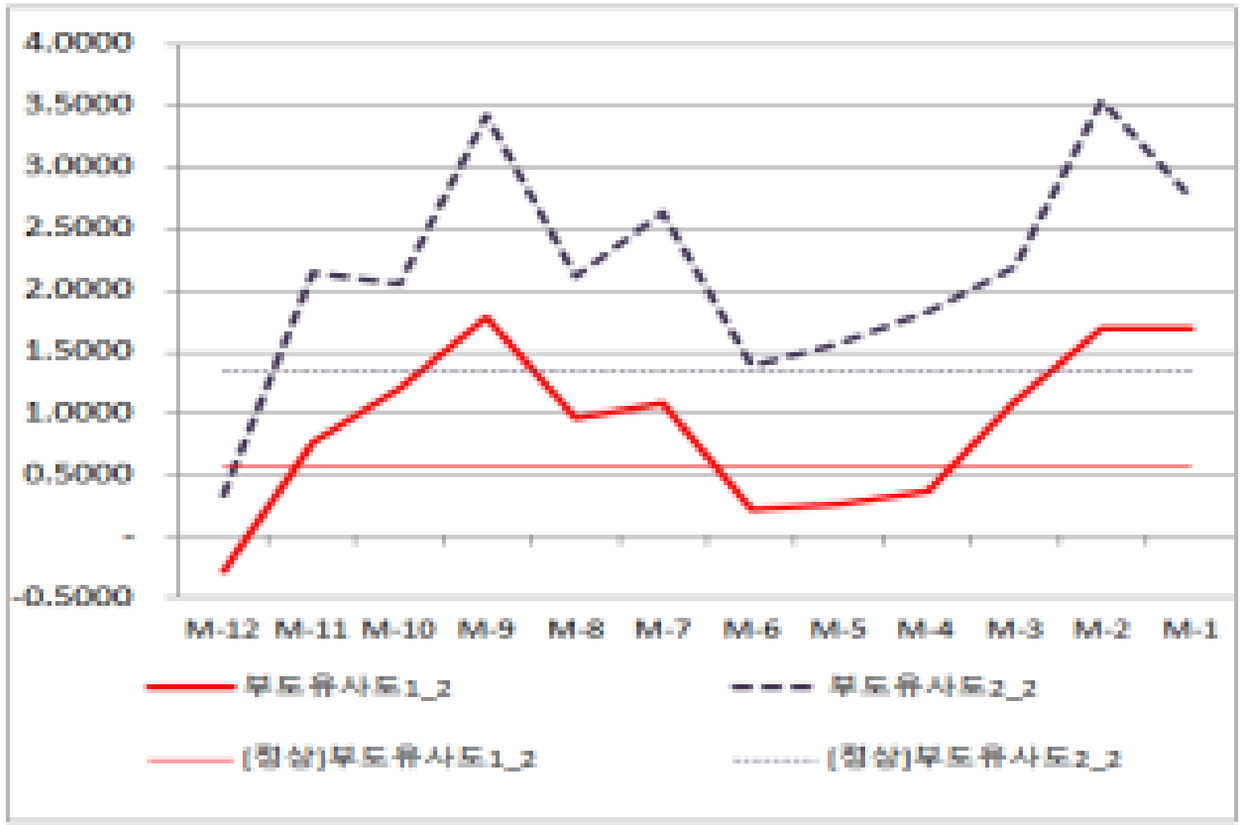
부도 기사 비율은 KMV 모형의 결과인 D.D. 와 비슷한 형태로 부도 가능성에 대한 신호를 주고 있는 것을 볼 수 있다. 특히 부도 발생 6 개월 이전 시점 부터는 지속적으로 KMV 모형보다 다소 높은 수준으로 부도 기사 비율이 나타난다



<그림 9>. 부도 기사비율과 D.D.의 비교 [19]

유사도 수준의 산출 결과



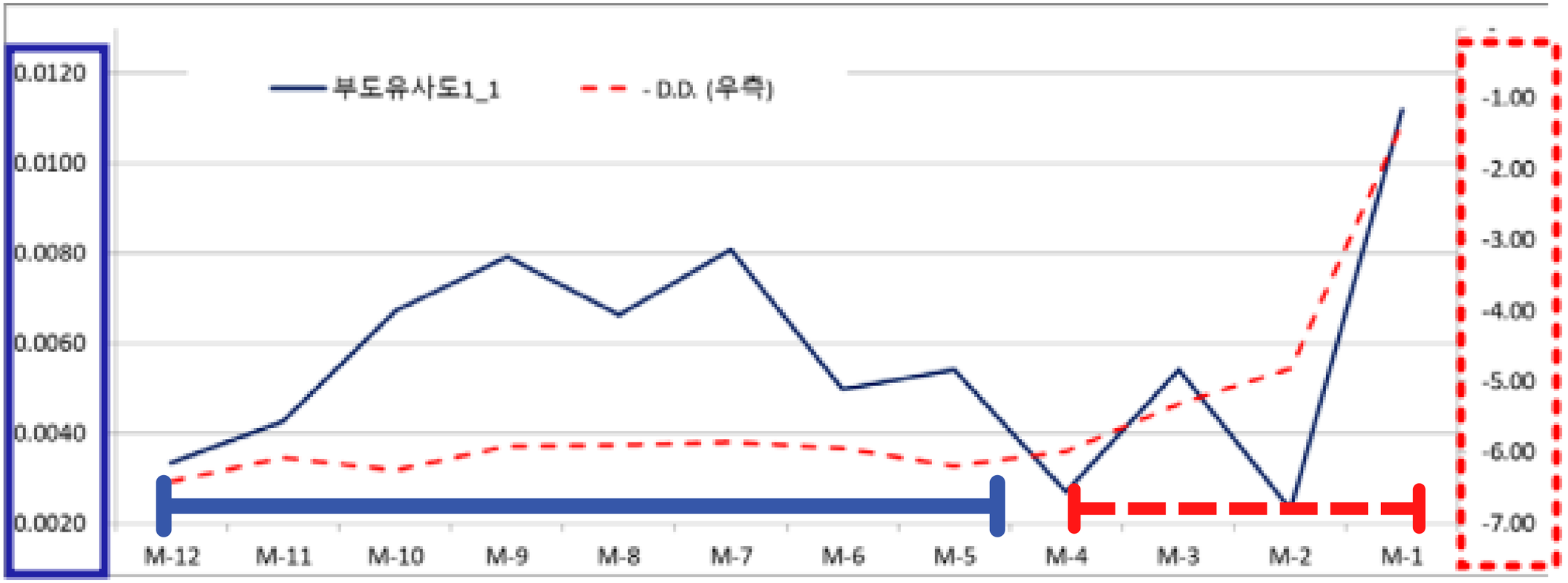


<그림 8> 부도 발생 12개월 전 부도 유사도(평균, 기사단위 평균) 추이

부도유사도 1_1:연관도평균(w2v-부도), 부도 유사도1_2:연관도합계 (w2v-부도) / 기사 수

부도유사도2_1:연관도평균(w2v-부도&상장폐지) ,부도유사도 2_2:연관도합계 (w2v-부도&상장폐지) / 기사 수

KMV 와 텍스트 정보기반 예측 모형 비교 결과



<그림 10>. 부도 유사도과 D.D.의 비교

기사의 부도 유사도 수준은 D.D. 에 비하여 부도 발생 7~10 개월 전에 매우 큰 차이를 보인다. 다만, 부도 발생 2~4 개월 기간은 D.D. 보다 낮은 수준으로 부도 가능성을 예측하고 있다. 이러한 현상은 부도 기업의 경우 실제 부도가 나타나기 오래 전부터 부도 관련 단어가 기사에서 많이 나타나는 현상을 실증한다.

월간 예측 모형 성과 분석

부도 기사비율과 **부도 유사도**를 활용할 경우 KMV 모형과 유사한 형태로 부도 예측이 가능함을 알 수 있다.

부도 발생 시점을 기준으로 KMV 모형 보다 이전 기간에 부도 유사도가 상승하여 기업 부도에 대한 조기경보 지표로서 기사 정보를 이용한 **텍스트 기반의 모형 결과가 활용될 수 있는 충분한 가능성**을 보여주었다.

텍스트 정보 기반의 부도예측은 주가 정보가 없는 비상장기업에도 활용이 가능하다는 점에서 **KMV 의 단점(재무,주가정보 만 포함)을 보완**하는 방법론으로 더욱 의미가 있다.

결론

한계점

연간 예측 모형에서의 **과적합**

데이터 수가 많음에도 데이터 SET A의 과적합

기업관련 뉴스의 **편중 문제**이다.

대기업 관련 뉴스 기사 수 > 중소기업 관련 뉴스 기사 수

논문의 아쉬운 점

Set A의 재무, 시장,
거시경제 정보 수집 기간
2001~2016 년

→ 금융위기 사건을 고려하지 않음

평가지표에 선정에 대한
의문점

→ 하나의 평가지표를 정하려
고 했다면 PR곡선 이나
F1score

논문의 좋았던 점

- 다양한 데이터 셋 구성

- 새로운 변수들을 만들어서 분석과정에 활용한 점

질문이
있으신가요?