빅데이터 분석방법을 활용한 제조업 혁신성과예측 방법에 대한 연구 : 딥 러닝 알고리즘을 중심으로

Forecasting Innovation Performance via Deep Learning Algorithm:
A Case of Korean Manufacturing Industry

황정재(Hwang Jeong-jae)*, 김재영(Kim Jae Young)**,
박재민(Park Jaemin)***

목 차

논문 요약

기술혁신에는 본질적인 어려움이 따르는데, 이는 상당부분 기술이 지닌 불확실성에 기인한다. 따라서 혁신 추구의 어려움을 경감에는 혁신 예측 방법론이 큰 도움이 될 수 있다. 한편 최근 빅데이터와 인공지능에 큰 관심이 이어지며 특히 알파고의 알고리즘 중 하나인 딥 러닝이 뛰어난성능을 보이고 있다. 이에 본 연구는 혁신성과 예측에 있어 딥 러닝을 이용한 방법론을 접목하여 연구를 진행하였다.. 모델 구축 및 학습에 있어 KIS 2016 데이터를 이용하였으며, 투입 요인으로는 정보 원천의 사용도와 혁신 목적을 사용하였고 산출 요인으로는 혁신 성과 지표를 구성하여 사용하였다.

Keyword : 혁신예측, 빅 데이터, 인공신경망, 한국기술혁신조사, 딥 러닝

* 건국대학교 기술경영학과 석사과정, ohnhop@naver.com, 주저자
** 건국대학교 기술경영학과 석사과정, miracle013@naver.com
*** 건국대학교 기술경영학과 교수, jpark@konkuk.ac.kr, 교신저자

Ⅰ. 서 론

기술은 본질적으로 동적이며 복잡하기 때문에 이에 대한 예측은 매우 어려운 일이며, 특히 기술의 동적인 측면을 강조하는 혁신의 경우 그러한 어려움이 더욱 두드러지게 된다(정선양, 2016). 따라서 이를 수행하는 데에는 상당한 시간과 자원이 필요하나 이는 충분히 들일 만한 가치가 있는 일인데, 기술과 혁신을 예측하고 이에 대응하는 데에 자원을 들일수록 불확실성에서 오는 손실을 줄여 불확실성을 계산된 위험(calculated risk)로 만들 수 있기 때문이다(Tidd & Bessant, 2013).

이에 전문가들은 수많은 기술예측 방법을 만들어 왔으며, 이는 지금도 정량적 방법과 정성적 방법 양쪽에서 진화해 나가고 있다. 그리고 최근 빅데이터와 인공지능을 필두로 한 새로운 예측 방법론들이 등장하고 있는데, 특히 2016년 3월 구글의 알파고와 이세돌 9단의 바둑대결 이후 이에 대한 관심이 점점 증대되고 있다. 특히 알파고의 알고리즘은 기존의 바둑 인공지능인 몬테카를로 트리 서치 알고리즘에 딥러닝 알고리즘을 조합하여 구성된 것으로 알려져 있는데, 이에 딥러닝에 대한 관심은 더욱 증대되고 있다(이지훈, 2016).

이렇게 뜨거운 관심을 받고 있는 딥 러닝 알고리즘은 이미 금융 분야나 마케팅 분야에서는 널리 사용되고 있으나(이우식, 2017) 기술경영 분야에서는 타 분야에 비해 그 연구가 미진한 편이다. 이에 본 연구에서는 뛰어난 예측 능력을 지닌 것으로 알려진 딥 러닝 알고리즘을 혁신 예측에 적용하여 불확실성을 극복하는 데에 기여하고자 한다. 이를 위하여 과학기술정책연구원에서 조사한 한국기업혁신조사 2016년 데이터(KIS 2016)을 활용하여 연구를 진행하였다.

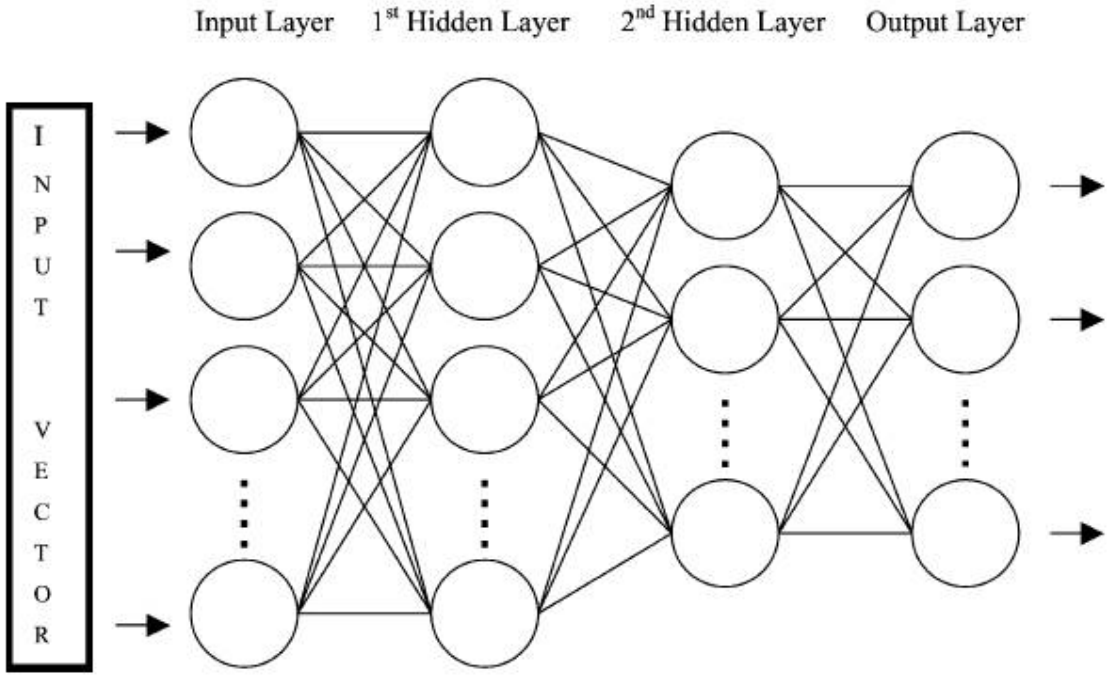
본 논문은 다음과 같이 구성된다. 2장에서는 선행연구 분석을 통하여 사회과학 분야에 딥 러닝 알고리즘이 어떻게 적용되는지를 파악할 것이며, 3장에서는 이를 바탕으로 예측 모델을 구성하고 변수를 설계한다. 그리고 4장에서는 KIS 2016 데이터를 이용한 모델 학습 및 실증분석을 진행하며 5장에서는 결론과 시사점을 도출하고자 한다.

Ⅱ. 선행연구 분석

1. 딥 러닝

딥 러닝은 인공신경망(artificial neural network) 구조를 이용한 모형으로 다층의 은닉층을 지닌 인공신경망 알고리즘을 의미한다(정한웅, 2016). 딥 러닝 모형은 크

게 입력층, 출력층, 은닉층으로 구성되며 각각의 층에는 노드들이 존재하고 이 노드들을 가중치(weight)가 연결하는 형태이다. 일반적인 딥 러닝 모형은 다음 (그림 1)과 같이 표현된다.



(그림 1) 일반적인 딥 러닝 모형

위의 (그림 1)은 두 개의 은닉층을 지닌 인공신경망 모형이다. 이렇듯 다수의 은닉층을 지닌 신경망을 심층 신경망(Deep neural networks, DNN)라고 부르며, 딥 러닝은 다수의 데이터를 통해 신경망의 가중치를 최적화시키는 방향으로 학습을 진행하게 된다(김인중 외, 2017). 이러한 최적화를 진행하는 방법으로는 경사 하강법(Gradient Descent), 모멘텀, AdaGrad, Adam 등 수많은 알고리즘이 존재한다(사이토 고키, 2017). 이렇게 학습된 딥 러닝 모델은 분류나 예측에 있어 뛰어난 성능을 보여 영상 인식이나 무인 자동차, 신용 등급 분류, 기업 부도 예측, 주가 예측 등 다양한 분야에 활용되고 있다.

2. 딥 러닝 모델의 활용 사례

심층 신경망을 이용한 딥 러닝 모델은 여러 분야에서 활용되고 있으나 본 연구에서는 사회과학분야에서 해당 모델이 어떻게 사용되는지를 살펴보고 이를 기반으로 하여 혁신 성과 예측 모델을 구축하고자 한다.

사회과학 분야 중 특히 경영 분야에 있어 최근 딥 러닝이 많이 활용되어 온 것으로 보인다. 특히 2016년 이후 국내에서 이를 활용한 연구들이 많이 시행되고 있는

데, 김인중 외(2017)은 딥러닝을 이용하여 T-커머스의 매출 예측을 시도하였다. 입력층에는 상품 정보, 방송 시간대, 날씨 정보 등이 포함된 623차원 벡터가 투입되었으며, 출력값은 매출액이었다. 이는 효과적인 매출 예측을 한 것으로 평가되었다.

이외에도 주가나 환율 예측 등에 딥 러닝 알고리즘이 사용된 예도 다수 존재한다. 이지훈(2016)은 딥러닝을 활용하여 주가 예측 모델을 구축하였으며 이우식(2017)은 딥러닝 분석을 활용하여 코스피 주가지수의 방향성을 예측한 바 있다. 또한 이우식·전희주(2016)은 중국 역내·외 위안화의 변동성을 예측하는 모델을 구축한 바 있다.

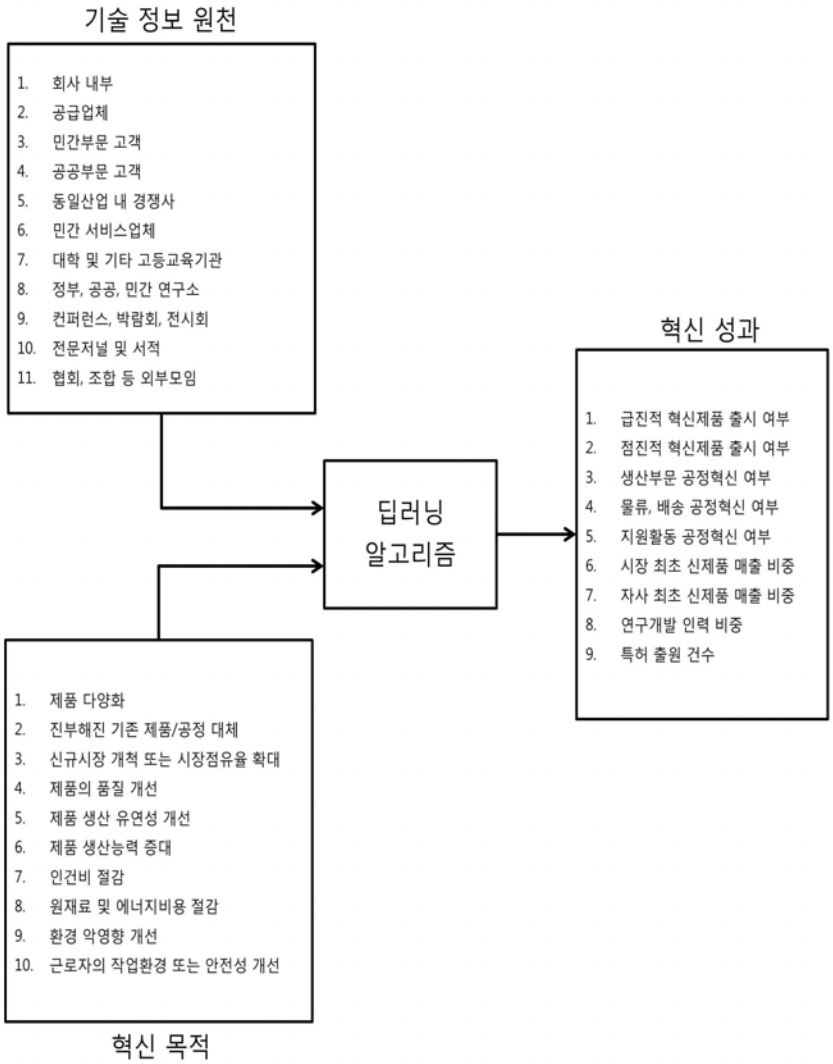
다만 기술경영 분야에서는 이러한 시도가 많지 않았다. 그런데 Wang & Chein (2006)은 대만 제조업 분야를 중심으로 인공신경망 모형을 활용한 혁신 성과 예측 모델을 제시한 바 있다. 해당 연구에서는 OECD Oslo Manual에 따라 대만 제조업 기업 53개를 대상으로 설문을 진행하였고, 여기에서 투입 요인을 기술 정보 원천의 활용도와 혁신 목적으로, 산출 요인을 혁신 성과로 하여 인공신경망 모형에 적용시켰다. 본 연구에서는 이를 참고하여 딥 러닝 모형을 구성하였다.

III. 연구설계

1. 연구모형 및 변수설계

본 연구에서는 과학기술정책연구원(STPEI)이 실시하는 '한국기업혁신조사(KIS)'의 2016년 제조업 데이터를 활용하였다. 한국기업혁신조사는 OECD Oslo Manual에 기초하여 실시되며, 조사 모집단은 2016년 이전 3년간 활동한 기업 중 상시종사자수 10인 이상의 제조업체이다. 이를 통해 얻어진 4000개 기업의 데이터 중 본 연구에서는 분석 대상 요인에 있어 결측치가 많은 자료와 이상치(outlier)를제외한 2079개 데이터를 활용하여 분석을 진행하였다.

앞서 살펴본 Wang & Chein (2006)에서는 혁신성과를 예측하기 위한 투입(input) 요인으로 기술 정보의 원천 활용정도와 혁신 목적을 활용하였으며 산출(output) 요인으로는 혁신성과와 관련된 요인들인 신제품 및 공정의 매출 비중, 특허의 수 등을 편집하여 사용하였다. 또한 각 요인들은 OECD Oslo Manual에 기초하여 조사되었으므로 본 연구에서 활용한 KIS 2016 데이터와도 유사한 점이 많다. 이를 참고하여 본 연구에서 혁신성과 예측을 위해 사용한 모형은 다음 (그림) 과 같다.



(그림2) 연구모형

본 모형에서 투입요인은 기술 정보 원천 관련 11개 변수와 혁신 목적 관련 10개 변수로 설정하였으며, 각 변수들은 0, 1, 2, 3의 4점척도로 조사되었다. 다만 Wang & Chein (2006)은 신경망에 한번에 많은 변수가 들어갈 경우 다중공선성의 문제가 생길 수 있음을 지적하였는데, 이를 피하기 위해 요인분석을 거쳐 투입 변수의 수를 감소시키는 방법을 사용하였다. 본 연구에서도 투입 변수들에 대한 요인분석을 실시하였다.

산출요인인 혁신성과의 경우에는 투입요인에 비해 더 많은 전처리 과정을 거쳐야 했는데, 투입 요인들의 경우 일괄적으로 4점척도에 따라 조사된 반면 혁신 성과는 0과 1로 이루어진 더미 변수와 비율 척도가 혼재되어 있으며 비율척도의 경우에도 0과 1 사이의 값인 비중과 자연수인 건수가 혼재되어 있어 이러한 차이를 완화해야 했기 때문이다. 각 변수들에 대한 자세한 내용은 다음 <표>와 같다.

<표 1> 변수의 조작적 정의

변수명	변수 설명	변수의 형태
pd_1	급진적 혁신제품 출시 여부	출시하였음 = 1 출시하지 않았음 = 0
pd_2	점진적 혁신제품 출시 여부	출시하였음 = 1 출시하지 않았음 = 0
pc_1	생산부문 공정혁신 여부	공정혁신이 있었음 = 1 공정혁신이 없었음 = 0
pc_2	물류, 배송 공정혁신 여부	공정혁신이 있었음 = 1 공정혁신이 없었음 = 0
pc_3	지원활동 공정혁신 여부	공정혁신이 있었음 = 1 공정혁신이 없었음 = 0
pdi_per_1	시장 최초 신제품 매출 비중	비중을 소수로 입력
pdi_per_2	자사 최초 신제품 매출 비중	비중을 소수로 입력
rnd_h	연구개발 인력 비중	비중을 소수로 입력
patent	특허 출원 건수	건수를 그대로 입력

이에 산출 요인은 변수를 하나로 합치는 작업을 실시하였는데 가장 차이가 큰 특허 출원 건수의 경우 min max scaling을 통해 0과 1 사이의 범위로 정규화하였으며, 변수들의 평균을 사용하였다.

Ⅳ. 실증분석 결과

1. 변수의 기술 통계량

본 연구에서 사용한 변수들의 기술 통계량은 다음 <표 2>와 같다.

〈표 2〉 변수의 기술통계량

변수명	N	평균	표준편차	변수명	N	평균	표준편차
ir_1	2079	2.52	0.75	ip_6	2079	2.30	0.77
ir_2	2079	2.32	0.86	ip_7	2079	2.51	0.76
ir_3	2079	2.45	0.83	ip_8	2079	2.89	0.78
ir_4	2079	1.83	1.06	ip_9	2079	1.97	0.79
ir_5	2079	2.19	0.88	ip_10	2079	2.01	0.79
ir_6	2079	1.52	0.95	pd_1	2079	0.18	0.39
ir_7	2079	1.31	0.97	pd_2	2079	0.55	0.50
ir_8	2079	1.48	0.99	pc_1	2079	0.33	0.47
ir_9	2079	1.55	0.97	pc_2	2079	0.23	0.42
ir_10	2079	1.37	0.93	pc_3	2079	0.16	0.36
ir_11	2079	1.35	0.91	pdi_per_1	2079	0.04	0.14
ip_1	2079	2.50	0.73	pdi_per_2	2079	0.13	0.22
ip_2	2079	2.58	0.69	rnd_h	2079	0.11	0.11
ip_3	2079	2.57	0.71	patent	2079	1.93	6.53
ip_4	2079	2.75	0.55	patent_mm	2079	0.01	0.05
ip_5	2079	2.45	0.73	dep	2079	0.19	0.14

ir_1부터 ir_11까지는 기술 정보의 원천에 대한 문항이며 ip_1부터 ip_10까지는 혁신 목적에 대한 문항이다. 이들은 모두 4점척도로 조사되었다. pd_1과 pd_2는 각각 급진적 혁신제품과 점진적 혁신제품의 시장 출시 여부이며, pd_2의 평균값이 pd_1보다 높은 것으로 보다 급진적 혁신제품보다 점진적 혁신제품이 시장에 더 많이 출시되고 있음을 확인할 수 있다. pc_1부터 pc_3까지는 각각 제조, 물류, 지원 관련 공정혁신의 여부이다. pdi_per_1은 시장 최초 혁신제품의 매출 비중이며 pdi_per_2는 자사 최초 혁신제품의 매출 비중이다. rnd_h는 연구개발인력 비중이며 patent는 특허 출원 개수이다. patent_mm은 특허 출원 개수를 0과 1 사이로 정규화한 수치이며 dep은 산출 요인에 해당하는 변수들의 평균값이다.

2. 요인분석

신경망에 변수를 투입하기에 앞서 투입변수의 수를 줄이기 위한 요인분석을 실시하였으며 요인 분석 결과는 다음 <표>와 같이 나타났다.

〈표 3〉 요인분석 결과

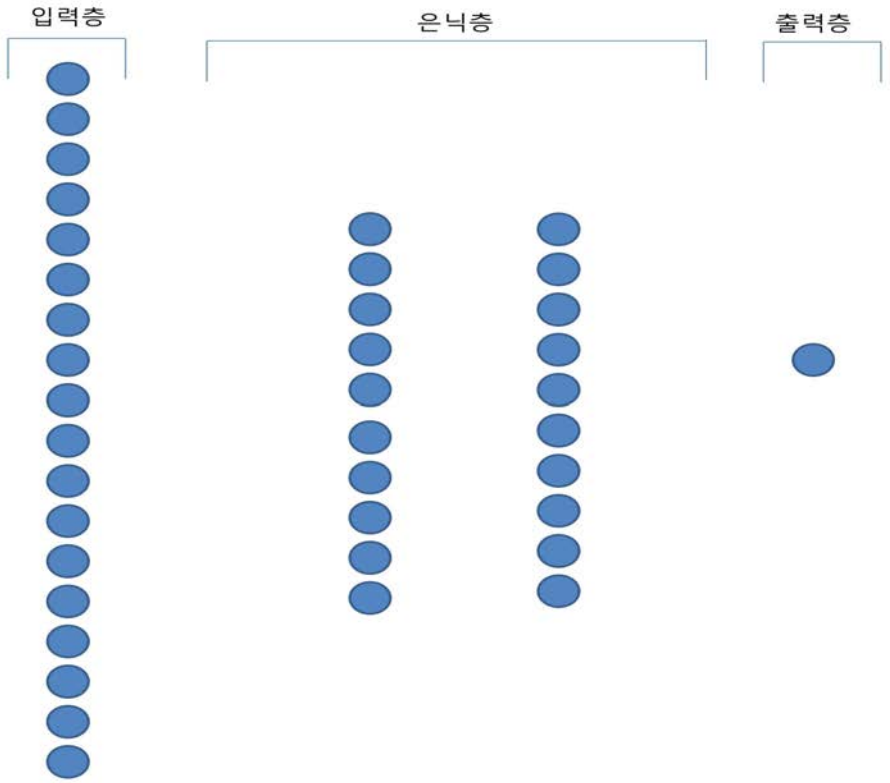
Variable	Factor1	Factor2	Factor3
ir_6	0.6399	0.1964	0.1126
ir_7	0.6855	0.1496	0.0993
ir_8	0.7071	0.1703	0.0748
ir_9	0.7648	0.1641	0.1219
ir_10	0.7675	0.0835	0.0334
ir_11	0.6811	0.1215	-0.0929
ip_1	0.0633	0.1749	0.7386
ip_2	0.0582	0.2495	0.7109
ip_3	0.0823	0.241	0.6859
ip_7	0.1397	0.7336	0.2048
ip_8	0.1269	0.791	0.1815
ip_9	0.1526	0.8703	0.1359
ip_10	0.1068	0.8606	0.1059
아이겐 값	1.1016	2.9515	1.6788
설명량(%)	0.4074	0.3876	0.2205
누적 설명량	0.4074	0.7950	1.0155
cronbach's alpha	0.8651	0.9086	0.8204

요인분석 결과 투입 요인은 크게 세 가지 요인으로 묶을 수 있었다. 첫 번째 요인에 해당하는 변수들은 정보의 원천 중 민간 서비스업체, 대학 및 기타 고등교육기관, 정부, 공공, 민간 연구소, 컨퍼런스, 박람회, 전시회, 전문저널 및 서적, 협회, 조합 등 외부모임에 해당하는 변수들이었으며, 나머지 변수들은 유의미한 요인부하량을 가지지 못하여 삭제하였다.

두 번째 요인에 해당하는 변수들은 혁신 목적 중 인건비 절감, 원재료 및 에너지 비용 절감, 환경 악영향 개선, 근로자의 작업환경 또는 안전성 개선이었으며 세 번째 요인은 혁신 목적 중 제품 다양화, 진부해진 기존 제품/공정 대체, 신규시장 개척 또는 시장점유율 확대였다. 나머지 변수들은 유의미한 요인부하량을 가지지 못하여 삭제하였다. 세 요인의 cronbach′s alpha 값은 모두 0.8 이상으로 높게 나타나 유의미한 요인으로 해석할 수 있었다.

3. 딥러닝을 사용한 예측 모형

변수들을 투입하여 혁신성과를 예측하기 위한 모델의 뉴런은 다음 (그림 3)와 같이 구성되었다. 모델 구축에는 Python 3.60 아나콘다 배포판과 Tensorflow 1.3 버전이 사용되었다.

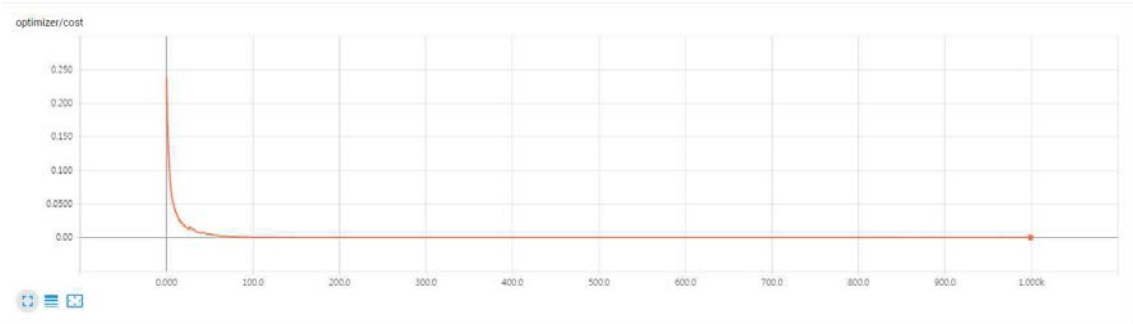


(그림 3) 뉴런의 구성

입력층의 경우 18개로 구성되었는데, 이는 요인분석 결과 유의한 부하량을 가진 변수가 18개이기 때문이다. 그리고 출력층은 하나로 구성하였는데, 예측하고자 하는 혁신 성과를 하나의 변수로 만들었기 때문이다. 그리고 은닉층은 총 두 개로 구성하였으며 각각 10개의 뉴런을 배치하여 점점 줄어드는 형태의 신경망을 구현하였다. 데이터는 학습 데이터(training data)와 시험 데이터(test data)로 구분하여 학습에는 학습 데이터를, 예측 정확도 파악에는 시험 데이터를 사용하였다. 학습에는 Adam Optimizer를 사용하였으며 학습과정에는 과적합(overfitting)을 피하기 위해 30%의 drop out을 적용하였다. 본 모델은 오차를 최소화하는 방향으로 가중치를 변화시키며 학습해 나가는데, 오차로는 오차제곱평균(mean squared error)을 활용하였다. 이러한 알고리즘을 Tensorboard를 이용하여 다음 (그림 4)과 같은 그래프로 나타낼 수 있다.

(그림 4) 신경망 그래프

학습을 위해 학습 데이터를 1000회 학습시켰으며, 그 결과 오차는 다음 (그림 5)과 같이 줄어들고 있음을 확인할 수 있었다.



(그림 5) 학습에 따른 오차 감소

학습 결과 오차를 0.005 미만으로 감소시킬 수 있었으며, 이는 시험 데이터에 있어서도 비슷한 결과를 나타냈다.

V. 결론 및 시사점

본 연구에서는 빅데이터 분석방법을 활용하여 기업의 혁신 성과를 예측하는 알고리즘을 제시하고자 하였고 그 중 딥 러닝을 활용한 알고리즘을 통해 이를 달성하고자 하였다. 이에 딥 러닝 모델에 KIS 2016 데이터를 학습시켜 예측모델을 구축하였다. 또한 학습된 알고리즘의 시험 결과 학습 데이터와 시험 데이터간 오차의 차이가 크게 존재하지 않아 비교적 정확한 예측을 기대할 수 있었다.

하지만 본 연구는 다음과 같은 한계점도 가진다. 빅 데이터 분석 방법의 경우 빅

데이터가 방대하고, 빠르게 축적되며 비정형성을 지닌다는 특성 때문에 데이터의 전처리를 생략할 수 있는 알고리즘을 필요로 하나 본 연구에서는 투입 요인과 산출 요인에 있어 둘다 전처리 이후 모델에 적용하였다. 이 부분에 있어서는 모델의 지속적인 개량이 필요할 것이다.

하지만 이러한 한계점에도 불구하고 본 연구는 혁신 예측 분야에 잘 적용되지 않던 빅데이터 및 인공지능을 접목시키고자 했다는 데에 의의가 있다. 혁신과 기술에 대한 예측은 본질적으로 어려운 분야이지만, 이러한 시도가 지속된다면 어려움을 경감하는 데에 상당한 도움을 줄 수 있을 것으로 기대한다.

참 고 문 헌

김인중·나기현·양소희·장재민·김윤종·신원영·김덕중 (2010), “딥러닝과 통계 모델을 이용한 T-커머스 매출 예측”, 「정보과학회논문지」, 44(8): 803-812.

이우식(2017), “딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측”, 「한국데이터정보과학회지」, 28(2), 287-295.

이우식·전희주 (2017), “딥러닝 분석을 이용한 중국 역내·외 위안화 변동성 예측”, 「한국데이터정보과학회지」, 27(2): 327-335.

이지훈 (2016), “딥러닝을 이용한 주가 예측 모델”, 숭실대학교 대학원 석사학위 논문.

정선양 (2016), 「전략적 기술경영」, 제4판, 박영사.

정한웅 (2016), “딥러닝 알고리즘에 기반한 기업부도 예측”, 한양대학교 대학원 석사학위 논문.

Tidd, J. and Bessnat, J. (2013), Managing Innovation: Integrating Technological, Market and Organizational Change, 5th edition, Chichester: Wiley.

Wang, T. Y. and Chien, S. C. (2006), "Forecasting innovation performance via neural networks: a case of Taiwanese manufacturing industry", Technovation, 26(5): 635-643.