딥러닝 기반 과년도 무역 데이터를 이용한 차년도 품목별 수출가 예측 모델 구현

김지훈 [1], 이지항 [1,+]
[1] 상명대학교 지능데이터융합학부 휴먼지능정보공학전공
neti2207@gmail.com, jeehang@smu.ac.kr

Predicting the Future Price of Export Items using a Deep Neural Network with Past Year's Trade Data

Ji-Hun Kim[1], Jee Hang Lee[1,+]
[1]Department of Human-Centered AI, Sangmyung University, Seoul, South Korea
[+]Corresponding author: Jee Hang Lee

요 약

산업통상자원부에서 제공하는 KOTRA 무역 데이터는 해당 품목과 해당 국가에 대하여 GDP, 관세율, 비즈니스 점수, 과/차년도 수출금액 등을 제공한다. 그러나 무역 수출품목은 수 없이 많을 뿐더러 그에 따른 대량의 데이터를 매년 인간의 분석을 통해 유의미한 결과를 이끌어내는 것은 상당히 큰 시간과 비용을 요구한다. 따라서 이번 연구에선 대량의 데이터를 학습하여 단기간에 저비용으로 결과를 예측할 수 있는 심층신경망 모델을 구현해 보았다.

1. 서론

KOTRA 무역 데이터는 무역 품목과 해당국가에 대하여 과년도와 차년도 수출금액을 GDP, 관세율, 비즈니스 점수와 함께 나열하고 있다 [1]. 한 해에 제공되는 무역 데이터 수는 2 만개 이상이지만 매년 쌓이게 되면 사람의 힘으로 분석하기엔 시간과 비용이 상당히 많이 요구된다.

기계학습을 이용한 데이터 분석 기법은 이러한 대용량 정보 처리를 자동화하여 시간과 비용을 절약할 수 있는 좋은 방법이다. 데이터 분석에서 수치를 예측하는 문제에서는 대표적 분석 기법인 다변량회귀분석이 널리 사용되고 있지만, 데이터가 기하급수적으로 증가하면, 그에 따른 성능향상엔 한계가 있다. 그러므로 데이터 증가와 컬럼 증가에 유연한 심층신경망 [2]을 추가하여 예측 정확도를 높여보고자 한다.

2. 데이터 속성 및 전처리

우선 <표 1>에서 볼 수 있듯, KOTRA 에서 제공하는 데이터의 각 컬럼들은 타 국가 기준으로 설명되어 있다. 따라서 수입금액의 의미는 타 국가가 수입을 한 가격 정보를 말한다. 총 16 개의 컬럼이 존재하는데 기준 연도는 고정값이기 때문에 삭제하였다. 국가명은 국가코드로 대체하였다. 국가간 평균거리와 환율은 가격 예측에 의미가 없다 판단하여 제외하였다. 인구데이터는 추정치가 포함돼있기 때문에 제외하였다. 또한 해당연도 해당 국가의 전체 품목 수입금액과 해당연도 해당 국가의 해당품목 수입금액을 결합하여 해당 국가가 수입한 전체 품목중 해당품목의 수입비율로 변환하였으며 해당 연도 해당 품목의 전세계 총 수입금액과 해당연도 해당 국가의 해당품목 수입금액을 결합하여 해당 국가가 전세계 기준 해당품목을 수입한 비율로 변환하였다. 또한 GDP 데이터는 증감비율로 변환하였다. 우리가 구해야 하는 차년도 수입금액은 품목별로 범위차이가 크기 때문에 과년도 수입금액과 차년도 수입금액의 증감비율로 변환하였다.

<표 1> KOTRA 무역 데이터 개요

1	UNC_YEAR	기준연도	YYYY
2	HSCD	HS Code (품목코드)	6자리 숫자코드
3	COUNTRYCD	ISO 국가코드	숫자코드
4	COUNTRYNM	영문 국가명	Character
5	TRADE_COUNTRYCD	해당 연도 해당 국가의 전체 품목 수입금액	US$
6	TRADE_HSCD	해당 연도 해당 품목의 전세계 총 수입금액	US$
7	TARIFF_AVG	해당 국가에서 해당 품목에 적용되는 평균 관세율	%
8	SNDIST	해당 국가와 수입 국가 간 평균 거리	km
9	NY_GDP_MKTP_CD	GDP	US$
10	NY_GDP_MKTP_CD_1Y	이전년도 GDP	US$
11	SP_POP_TOTL	인구 (연중 추정치)	명
12	PA_NUS_FCRF	공식 환율	US$
13	IC_BUS_EASE_DFRN_DB	비즈니스 용이성 점수	점수 (0~100)
14	KMDIST	해당 국가와 한국과의 거리	km
15	TRADE_HSCD_COUNTRYCD	해당 연도 해당 국가의 해당 품목 수입금액	US$
16	KR_TRADE_HSCD_COUNTRYCD	내년 해당 국가가 해당 품목을 한국으로부터 수입한 금액	US$

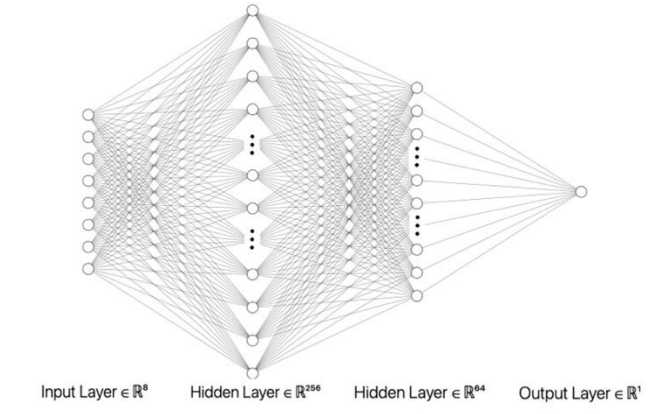
<표 2>는 예측을 위한 심층신경망 모델의 입력데이터를 보여준다. <표 1>에서 보인 데이터 중, 제외되거나, 인덱싱 처리하거나 변환을 통해 전처리된 8 개 컬럼의 데이터가 심층신경망 모델의 입력으로 사용된다. 이에 대한 정답 (Label)으로 수입금액의 증감비율을 설정하였다. 결과적으로, 심층신경망 기반 예측 모델의 출력 값은 수입 금액의 증감 비율 값이 된다.

<표 2> 전처리 후 KOTRA 데이터 컬럼

1	HSCD	HS Code (품목코드)	HS Code (품목코드)
2	COUNTRYCD	ISO 국가코드	ISO 국가코드
3	TARIFF_AVG	해당 국가의 해당 품목 평균 관세율	%
4	IC_BUS_EASE_DFRN_DB	비즈니스 용이성 점수	점수 (0~100)
5	KMDIST	해당 국가와 한국과의 거리	km
6	HSCD_IMPORT_PERCENTAGE	해당 국가의 해당 품목 수입금액 비율	%
7	COUNTRYCD_IMPORT_PERCENTAGE	해당 품목의 해당 국가 수입금액 비율	%
8	GDP_PERCENTAGE	1년간 증감한 GDP 비율	%
9	TRADE_HSCD_PERCENTAGE	수입금액의 증감 비율	%

3. 모델 구조 및 학습결과

앞서 전처리된 데이터인 수입금액의 증감비율을 제외한 총 8 개 컬럼값이 예측 모델의 입력데이터로 들어간다. 기존 다변량 회귀분석 모델은 데이터의 차원 증가에 따른 성능 향상에 한계가 존재한다. 따라서 KOTRA 데이터와 같은 빅데이터 성향의 이점을 가져가기 위해 심층 신경망 모델을 구성하였다. 모델은 입출력 포함 총 4 개의 Layer 로 구성되어 있으며 입력으로 표 2 에서 제시한 8 개 값을 사용한다. 심층 신경망의 은닉층은 2 개로, 각각 256 노드, 64 노드로 구성되었으며, Fully Connected Layer 로 구현하였다. 최종적으로 출력층은 1 개 노드이며, 예측 값을 나타낸다.

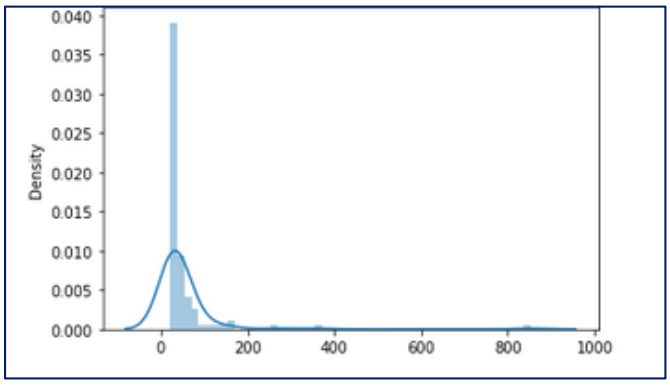


(그림 1) 모델 구조

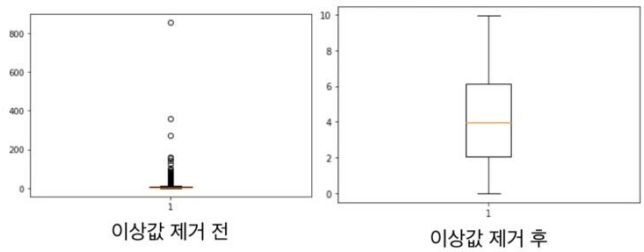
안정적 학습과 성능 향상을 위해 모든 층에서 Batch 정규화를 진행하였다. 각 층에서 활성화 함수로 ReLU 를 적용하였다. Loss 는 Mean Squared Error 로 정의하여 역전파를 통해 학습이 진행되도록 하였다. 최적화 알고리즘으로 Adam Optimizer 를 사용하였다.

제안한 심층신경망 기반 예측 모델의 출력은 수입금액의 증감 비율이다. 이렇게 예측된 수입 금액 증감 비율에 과년도 우리나라가 해당 품목을 수출한 금액에 곱하면, 차년도 해당 품목의 수출 금액을 예측할 수 있다.

제안한 예측 모델 학습을 위해 총 1000epoch 를 수행하였다. 그 결과 L1 loss 는 5.97, L2 loss 는 5.80 로 관측되었다. 상대적으로 높은 loss 값이 도출되어, 오차들을 사후 분석 해보았다. 그림 2 는 오차들의 분포를 보인다. 오차들의 분포를 살펴보았을 때, x 축 기준 최우측과 같이 오차값이 매우 큰 이상 값이 존재 하는 것을 확인하였다.



(그림 2) 오차 분포도



(그림 3) 오차에 대한 Box Plot

오차가 매우 큰 데이터를 제거한 뒤 다시 L2 loss 를 구한 결과, 4.50 으로 소폭 줄어든 것을 확인할 수 있었다. 이는 (그림 3)과 같이, 이상값을 제거하기 전 Box Plot 에 다수 존재하는 특이점에 의한 것임을 알 수 있다. 따라서 이상값을 제거한 뒤 다시 그려본 Box Plot 을 보면 중앙값 약 4.0 이며 특이값 없이 0~10 사이에 정상정으로 분포하고 있음을 확인할 수 있다.

4. 결론

수출 금액과 그에 따른 속성을 가진 KOTRA 데이터를 다변량 회귀모델을 기반으로 한 심층신경망 모델로 학습 시키고, 차년도 수출할 금액을 예측한 후, 성능 평가를 진행하였다.

심층신경망 기반 예측 모델을 통한 예측결과는 에러율 4%대로 계산되었으며, 이는 실제 값과 예측값의 차이가 4% 차이 난다는 것을 의미한다. 무역 데이터 비전문가의 입장에서 학습 결과로 나온 오차율이 적

정한 수치인지 판별하기 위해서는 더 엄밀한 기준을 확립할 필요가 있다. 그럼에도 불구하고, 무역데이터를 통해 차년도 수출금액을 실제값에 가까이 예측할 수 있는 가능성은 충분히 확인하였다.

또한 오차가 매우 큰 값이 존재했던 것처럼 모든 데이터에 대하여 균일한 예측이 불가능 했던 이유는 모델이 학습할 수 없는 데이터 외적인 요인이 존재하기 때문으로도 생각해 볼 수 있다. 예를 들어, 이는 세계 정세, 전염병, 경제상황 등 무역환경에 변화를 줄 수 있는 외부 요인들이 있었다는 의미로 해석될 수 있다. 이에 따라 예측과 실제의 차이가 매우 큰 사례가 다수 존재할 수밖에 없었다.

따라서 추후 국제 정세, 전염병, 경제 상황 등 무역환경의 변화에 영향을 줄 수 있는 인자를 개발하고 예측 모델에 적용할 수 있을 것이며 이번 연구에서 제시된 일반 심층신경망 모델과 달리 진보된 모델들, 예를 들어 CNN [3], lenet [4], LSTM [5]와 같이 새로운 모델을 통해 예측정확도를 더 높일 수 있을 것이다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1G1A1102683). 본 연구는 삼성미래기술육성센터의 지원을 받아 수행하였음 (No. SRFC-TC1603-52). 본 결과물은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 사회맞춤형 산학협력 선도대학 (LINC+) 육성사업의 연구결과임.

참고문헌

[1] 제 9 회 공공데이터활용 BI 공모전, http://www.datacontest.kr(retrieved 20210926)
[2] Plieninger, Andreas, STEUERUNGS-und REGELUNGSTECHNIK, and J. Conradt. Deep Learning Neural Networks on Mobile Platforms. Neurocomputing Systems, 14p, 2016.
[3] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. “A convolutional neural network for modelling sentences.” arXiv preprint arXiv:1404.2188 (2014).
[4] LeCun, Yann, et al. “Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.
[5] Li, Xinyi, et al. “DP-LSTM: Differential privacy-inspired LSTM for stock prediction using financial news.” arXiv preprint arXiv:1912.10806 (2019).