KIPS Trans. Softw. and Data Eng.
Vol.11, No.10 pp.427~436
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.10.427

Predicting the Future Price of Export Items in Trade Using a Deep Regression Model

Kim Ji Hun† · Lee Jee Hang††

ABSTRACT

Korea Trade-Investment Promotion Agency (KOTRA) annually publishes the trade data in South Korea under the guidance of the Ministry of Trade, Industry and Energy in South Korea. The trade data usually contains Gross domestic product (GDP), a custom tariff, business score, and the price of export items in previous and this year, with regards to the trading items and the countries. However, it is challenging to figure out the meaningful insight so as to predict the future price on trading items every year due to the significantly large amount of data accumulated over the several years under the limited human/computing resources. Within this context, this paper proposes a multi layer perception that can predict the future price of potential trading items in the next year by training large amounts of past year's data with a low computational and human cost.

Keywords : KOTRA, BigData, Ministry of Trade Industry and Energy, Deep Learning, Multi Layer Perception

딥러닝 기반 무역 수출 가격 예측 모델

김 지 훈† · 이 지 항††

요 약

산업통상자원부에서 제공하는 KOTRA 무역 데이터는 해당 품목과 해당 국가에 대하여 GDP, 관세율, 비즈니스 점수, 과/차년도 수출금액 등을 제공한다. 그러나 무역 수출품목은 수없이 많을뿐더러 그에 따른 대량의 데이터를 매년 수작업 기반 분석을 통해 유의미한 결과를 이끌어내는 것은 상당히 큰 시간과 비용을 요구한다. 따라서 이번 연구에선 대량의 데이터를 학습하여 단기간에 저비용으로 결과 예측이 가능한 다층 퍼셉트론 모델을 구현하고 성능을 평가하였다. 먼저 딥러닝 기반 무역 수출 가격 예측 모델을 일반적 다변량 회귀 모델과 비교하였을 때, 예측 오류와 학습 시간 측면에서 통계적으로 우수한 성능을 보였다. 수출 가격 데이터는 시계열 속성이 있을 것으로 예상하는 바, 은닉 노드들이 모두 연결된 다층 퍼셉트론과 순환 신경망을 이용하여 수출 가격 데이터를 예측하였다. 그 결과 새로운 데이터에 대해 수출 가격 예측을 위한 일반화 능력은 순환 신경망이 우수한 성능을 보였으나, 다층 퍼셉트론이 무역 수출 가격 예측에서 더 뛰어난 성능을 보였다. 추후 장기간 데이터를 확보한다면, 순환 신경망 혹은 트랜스포머 기반 딥러닝 모델을 이용하여 더 뛰어난 수출 가격 예측이 가능할 것으로 사료된다.

키워드 : KOTRA, 빅데이터, 산업통상자원부, 딥러닝, 다층 퍼셉트론

1. 서 론

대한민국의 무역 수출입 규모는 2011년 최초 1조 달러를 넘어 2021년, 역대 최고치인 1조 2천억 달러의 무역 규모를 달성했다[1]. 이처럼 무역 규모가 증가함과 동시에, 4차 산업혁명으로 인해 무역 시장은 기존의 특정 물품의 일괄적 공급 시스템에서 전 세계 개개의 소비자를 대상으로 생산하여 판매하는 형태로 변화하고 있다[2]. 이로 인해 무역 데이터는 수많은 품목과 종류로 구분되어 기하급수적으로 축적되고 있다. 이를 분석할 수 있다면, 우리나라의 경제 성장을 예측하는데 중요한 변수가 되는 수출 관련 지표[3]에 대한 통찰을 얻을 수 있을 것이다. 하지만 이런 방대한 양의 데이터를 사람의 힘으로 분석하기에 시간과 비용이 상당히 요구된다. 기계학습을 이용하여 이러한 대용량 정보 처리를 자동화하고 분석한다면, 시간과 비용을 절약할 수 있을 것으로 보인다.

산업통상자원부에서 제공하는 KOTRA 무역 데이터는 무역 품목과 해당 국가에 대하여 과년도와 차년도 수출금액을 GDP, 관세율, 비즈니스 점수와 함께 제시하고 있다[4]. 이러

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1G1A1102683). 본 연구는 삼성미래기술육성센터의 지원을 받아 수행하였음(No. SRFC-TC1603-52).
※ 이 논문은 2021년 한국정보처리학회 ACK 2021의 우수논문으로 "딥러닝 기반 과년도 무역 데이터를 이용한 차년도 품목별 수출가 예측 모델 구현"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 상명대학교 휴먼지능정보공학전공 석사과정
†† 비 회 원 : 상명대학교 휴먼지능정보공학과 조교수
Manuscript Received : December 28, 2021
First Revision : March 10, 2022
Accepted : April 9, 2022
* Corresponding Author : Lee Jee Hang(jeehang@smu.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

한 여러 종류의 특성(feature)을 가진 데이터에 대해 분석 및 수치를 예측하는 문제에서는, 많은 특성을 허용하여 단순 선형 회귀 모델을 일반화하는 다변량 회귀 분석[5]이 사용되었다. 다변량 회귀 분석은 어느 한 특성이 결과에 미치는 영향력을 측정할 수 있는 이점을 가지고 있지만, KOTRA 데이터와 같이 빅데이터의 경우, 그에 따른 모델 성능 향상에 한계가 있다. 그러므로 특성 개수와 데이터 증가에 유연한 심층 신경망[6]을 접목하여 모델의 KOTRA 데이터의 차년도 수출금액을 예측하는 정확도를 높여보고자 한다.

2. 사전 연구

과거의 방대한 무역 데이터를 기반으로 국가별 미래의 수출금액을 예측하기 위해 기계학습을 사용하여 정확도를 높이고 계산 비용을 줄이는 연구가 진행되어왔다.

(Keck et al., 2009.2)는 주요 국가의 연간 수입 추이를 중심으로 국제 무역의 성장을 예측하기 위해 순수 시계열 분석하였고, 추가 예측변수를 포함한 정보를 활용하여 연구를 진행하였다. GARCH 추정을 사용하며 추가 예측변수로 GDP를 사용하는 자기시차분포(ADL) 모델은 2분기 예측에 뛰어난 성능을 보였다. 이 모델은 시계열 데이터에 대해 우수한 성능을 나타내는 것으로 판단된다[7].

(Veenstra et al., 2001)는 해상 무역의 흐름에 대해 장기적 예측이 가능한 다변량 자기 회귀 시계열 모델을 소개하고 있다. 이 모델은 건화물 및 원유 상품에 대한 30년간의 거래 데이터를 기반으로 학습되었으며 비교적 작은 에러율을 갖는 예측 모델임을 보였다[8].

(Silva et al., 2015)는 경제 불황인 2008년 전후, 미국의 무역 예측을 위한 단일 스펙트럼 분석(SSA)을 제시하고, 아리마(ARIMA), 지수평활법(ETS), 신경망 모델과 비교하였다. 제시된 단일 스펙트럼 분석 모델은 앞선 비교 대상 모델에 비해 2008년 경제침체와 같은 구조적 붕괴(structural break)에 덜 민감하여 더 정확히 모형화하고 예측이 가능함을 보여주었다[9].

(Lu et al., 2001)에서 제안한 접근은 우리가 실험하고자 하는 방식과 유사하다. 6개의 기계학습 모델을 제시하여 중국의 여덟 개 시장의 일별 탄소섬유 가격과 거래량을 예측하는데, 특히 데이터 잡음 제거 방법인 CEEMDAN(complete ensemble empirical mode decomposition with adaptive noise)[10]을 사용하였다. 중국 여덟 개의 시장에서 가져온 데이터는 시계열에 따라 가격과 거래량 분포 특성이 다르고 데이터 간 불균형이 심한데, CEEMDAN 방법은 이를 부드럽게(smoothing) 해주는 기능을 수행한다. 이 방법을 통해 탄소섬유 시장에 더 적합한 모델을 얻을 수 있었다[11]. 본 연구에서는, 데이터 불균형 문제를 보완하기 위해 배치 정규화(Batch Normalize)[12] 작업을 진행하여 극복하고자 하였다.

앞서 언급한 바와 같이, 앞선 연구들에선 무역 수입/수출에 대해 예측하기 위해 다양한 기계학습 모델이 사용되었으며, 방대한 양의 시계열 데이터를 통해 학습하여 성능을 향상시킴을 보였다. 또한 임베딩(Embedding) 및 특이값 분해, 재구성(reconstruction)을 사용한 예측 모델을 제안하고, 기존 ARIMA, ETS, 신경망 기반 모형보다 더 우수한 성능을 보였다.

기존 연구들은 특히 장기간 데이터를 이용하여 무역 추이를 분석하고 국가별 총 무역액을 예측하였다. 다만, KOTRA에서 제공한 데이터는 지난 1년 전의 데이터만 제공한 바, 기존 연구와는 다른 접근을 통해, 좀 더 세밀한 부분인 무역 품목별 수출 가격을 1년 단위로 예측하는 모델을 구상해 보았다. 이와 동시에 임베딩 및 특이값 분해등의 과정을 내포하는 종단 간 모델을 다변량 회귀와 신경망이 접목된, 다층 퍼셉트론(MLP) 모델로 구상하여 진행해 보았다. 이후, 다변량 회귀, 순환 신경망 모델과 성능을 비교해 보고자 한다.

3. 방 법

3.1 사용 데이터 예시

Table 1에서 볼 수 있듯, KOTRA 데이터는 총 16개의 특성이 존재하며 해당 특성의 단위에 맞게 각 데이터가 입력되어 있다. KOTRA에서 제공하는 데이터의 각 특성들은 타 국 기준으로 설명되어 있다. 따라서 수입금액은 타국이 우리나라로부터 수입한 가격 정보를 말한다.

3.2 데이터 전처리 과정

전처리 과정에 앞서 이 실험에서 사용한 KOTRA 데이터는 2017년 단년 데이터로써, 모든 아이템은 동일한 연도 정보를 갖고 있다. 따라서, 사실상 구분 의미가 사라진 특성인 연도 정보 UNC_YEAR은 삭제하였다. 추후 다년간의 KOTRA 데이터가 제공된다면, 해당 특성은 시간 정보를 제공하는 특징이 되는 바, 장단기 메모리 (LSTM)[13]와 같은 순환 신경망 모델로 새로 구성하였을 때, 입력 정보로 사용할 수 있으며, 이를 통해 더 높은 성능을 기대할 수 있다.

다음으로 COUNTRYNM은 COUNTRYCD로 대체 가능하여 삭제되었다. HSCD와 COUNTRYCD은 각각 물품과 국가에 대한 고유 코드이기 때문에 각 번호는 유니크한 값으로 치환되어 1부터 오름차순으로 부여된다. 거리 관련 정보 특성인 SNDIST와 KMDIST는 단위가 매우 크다 판단하여 기존 1km이던 단위를 100km로 변환하였다. 인구 추정치에 해당

Table 1. KOTRA Data Overview

	Column	Description	Unit	Item1	Item2	Item3
1	UNC_YEAR	Base Year	YYYY	2017	2017	2017
2	HSCD	Item Code	Digit Code	820210	820210	820210
3	COUNTRYCD	Country Code (ISO)	Digit Code	818	826	842
4	COUNTRYNM	Country Name	Character	Egypt	United Kingdom	USA
5	TRADE_ COUNTRYCD	The amount of the country imported for all items in the year.	US$	66338749061	641332000000	2405280000000
6	TRADE_HSCD	The amount of world imports for the item in the year.	US$	344968106	344968106	344968106
7	TARIFF_AVG	The average tariff rate applied to the item.	%	2	0	0
8	SNDIST	Average distance between that country and the importing country.	km	3964.230828	3799.545271	7938.630665
9	NY_GDP_ MKTP_CD	GDP	US$	235734000000	2666230000000	19519400000000
10	NY_GDP_ MKTP_CD_1Y	GDP of the previous year	US$	332442000000	2694280000000	18715000000000
11	SP_POP_ TOTL	Population (Yearly estimate)	Number	96442593	66058859	324985539
12	PA_NUS_ FCRF	Official exchange rate.	US$	17.78253352	0.776976682	1
13	IC_BUS_EASE_ DFRN_DB	The score for ease of business.	Score (0~100)	55.47428	83.34108	83.35255333
14	KMDIST	Distance between the country and Korea.	km	8497.368164	8875.389648	11065.70215
15	TRADE_HSCD_ COUNTRYCD	The amount the country imported for the item in the year.	US$	1708002	35559546	48108516
16	KR_TRADE_ HSCD_ COUNTRYCD	The amount the country will import for the item in the next year.	US$	7043	97397	3294503

하는 특성인 SP_POP_TOTL은 이후 설명할 비율로의 치환에 대하여, 표현하기 위한 관련 데이터가 존재하지 않기 때문에 해당 특성에 대해서 0~1 사이의 값으로 정규화를 진행하였다.

데이터의 스케일(scale)이 동일하지 않은 경우, 수렴 속도에 영향을 미칠 수 있다[14]. 우리가 모델의 입력으로 사용할 데이터가 이러한 성격을 띄고 있기 때문에 관련있는 특성들을 결합하여 비율 값으로 표현하고자 하였다. 따라서 해당 연도 해당 국가의 전체 품목 수입 금액인 TRADE_COUNTRYCD와 해당 연도 해당 국가의 해당 품목 수입금액인 TRADE_HSCD_COUNTRYCD을 결합하여 해당 국가가 수입한 전체 품목 중 해당 품목의 수입비율로 변환하였으며, 해당 연도 해당 품목의 전 세계 총 수입금액인 TRADE_HSCD와 해당 연도 해당 국가의 해당 품목 수입금액인 TRADE_HSCD_COUNTRYCD을 결합하여 해당 국가가 전 세계 기준 해당 품목을 수입한 비율로 변환하였다. GDP 데이터 또한 현재와 과거에 대한 데이터가 존재하기 때문에 NY_GDP_MKTP_CD_1Y, NY_GDP_MKTP_CD를 묶어 증감비율로 표현했다. 반면에 11번 SP_POP_TOTL특성은 앞서 언급했듯, 독립적인 데이터이기 때문에 0~1사이의 값으로 정규화 과정을 거쳤다.

마지막으로 우리가 예측하고자 하는 차년도 수입금액은 품목별로 범위 차이가 크기 때문에 과년도 수입금액과 차년도 수입금액의 증감비율로 변환하였는데, 차년도 수입 금액이 음수로 표현되지 않아 비율로 표현하더라도 오류가 없으므로 이처럼 구성하였다.

3.3 예측 모델

Table 1에서 보인 데이터 중, 제외 또는 인덱싱(Indexing), 변환을 통해 최종적으로 전 처리된 12개 특성 데이터를 Table 2에 정리하였다. 이 중 마지막 특성을 제외한 11개 특성이 모델 학습의 입력으로 사용된다. 결과적으로 다층 퍼셉트론 기반 예측 모델의 출력값은 마지막 특성인 수입 금액의 증감 비율 값이 된다.

본 연구에서 제시하는 다층 퍼셉트론 모델의 성능 평가를 위해 다변량 회귀, 순환 신경망 모델을 제시하고 비교하였다. 다변량 회귀 모델의 구조는 $AX+b$ 형태로써, A는 가중치, X는 입력데이터, b는 편향이다. 다음으로, 순환 신경망 모델은 히든 사이즈가 512인 장단기 메모리를 2개 쌓아 학습을 진행하였다. 시계열 데이터에 적합한 모델이지만 비교를 위해 학습을 진행해 보았다. 만약 학습 데이터가 시계열 데이터 성향을 띈다면 더 높은 성능을 기대할 수 있을 것이다. 사실 본 연구에서 사용하는 KOTRA 데이터는 한 해 동안의 무역 데이터이기 때문에 이번 연구에서 평가로 사용되는 순환신경망 모델은 일반적인 다층 퍼셉트론의 기능과 유사할 것으로 판단된다.

기존 다변량 회귀분석 모델은 데이터의 차원이 증가할수록 예측 정확도가 낮아지는 한계가 존재한다. 따라서 전처리되어 입력 차원이 11개인 KOTRA 데이터의 이점을 취하기 위해 다층 퍼셉트론으로 모델 내부를 구성하였다. 입출력 포함 총 3개의 층(Layer)을 쌓고, 각 층은 드랍아웃(Dropout)[15]을 포함한 512노드(node)로 구성하였다. 3개의 층은 전결합 레이어(Fully Connected Layer)로 구현하였다. 최종적으로 출력층은 1개 노드이며, 이는 모델의 예측값을 나타낸다. Table 3에서 본 연구에서 사용한 순환 신경망과 다층 퍼셉트론의 구조를 보였다.

Table 2. Columns of KOTRA Data After Preprocessing

	Column	Description	Unit	Item1	Item2	Item3
1	HSCD	Item Code	Digit Code	497	497	497
2	COUNTRYCD	Country Code(ISO)	Digit Code	33	34	35
3	TARIFF_AVG	The average tariff rate applied to the item.	%	2	0	0
4	SNDIST	Average distance between that country and the importing country.	100km	39.64231	37.99545	79.38631
5	SP_POP_TOTL	Population (Yearly estimate)	Number	0.067469	0.045504	0.232687
6	PA_NUS_FCRF	Official exchange rate.	US$	17.78253	0.776977	1
7	IC_BUS_EASE_ DFRN_DB	The score for ease of business.	Score (0~100)	55.47428	83.34108	83.35255
8	KMDIST	Distance between the country and Korea.	100km	84.97368	88.7539	110.657
9	HSCD_IMPORT_ PERCENTAGE	The ratio of the amount of import of the relevant item among all items.	%	0.002575	0.005545	0.002
10	COUNTRY_IMPORT_P ERCENTAGE	The ratio of the country to the amount of imports of the relevant item from around the world.	%	0.495119	10.30807	13.94579
11	GDP_PERCENTAGE	GDP change ratio	%	70.90981	98.95891	104.2982
12	TRADE_HSCD_ PERCENTAGE	Increase/decrease ratio of income amount	%	0.412353	0.273898	6.848066

Table 3. LSTM and MLP Model Architecture

LSTM		MLP	
Layer	Output Shape	Layer	Output Shape
LSTM (with dropout)	Batch Size, 512	BatchNorm1d	Batch Size, 11
		Linear	Batch Size, 512
		Dropout	Batch Size, 512
Linear	Batch Size, 1	Linear	Batch Size, 1

KOTRA 데이터는 학습 데이터가 선형적인 구조를 띄고 있지 않아 모델의 학습을 어렵게 만들기 때문에 [12]안정적 학습과 성능 향상을 위해 모든 층에서 배치 정규화를 진행하였다. 또한 각 층에서 활성화 함수로 렐루(ReLU)를 적용하였으며, 예측 증감률과 실제 증감률 간의 차이 값을 줄이는 방향으로 학습하기 위해 로스(Loss)는 평균 제곱 오차(Mean Squared Error)와 평균 절댓값 오차(Mean Absolute Error) 두 개로 정의하여 진행하였다. 역전파를 통해 학습되도록 하였고, 이를 위해 최적화 알고리즘으로 아담 옵티마이저(Adam Optimizer)를 사용하였다.

모델 선정에 있어, 입력데이터를 고려하여 두 가지 경우를 실험하였다. 첫째, 신경망 기반 모델의 특성(characteristic)을 유지하여 입력 특성에 대해 사람의 개입을 제거한 전체 데이터 사용 모델 (Full-Data Model)을 사용하였다. 이와 대조적으로, 특성의 중요도를 계산[16]할 수 있다고 판단하여 입력 특성이 선별된 특정 데이터 사용 모델 (Selected-Data Model) 또한 구현하고 분석하였다. 각 모델의 세부사항을 살펴보면, 전체 데이터 사용 모델(Full-Data Model)은 Table 2의 12개 특성을 모두 사용하여 학습되며, 특정 데이터 사용 모델 (Selected-Data Model)은 이번 실험에서 12개 특성의 중요도를 계산하지 않고, 단순히 불특정 국가와 관련된 SNDIST와 추정치가 포함된 SP_POP_TOTL, 임의로 선택한 PA_NUS_FCRF 을 삭제하여 총 9개 특성으로 학습되었다.

3.4 선험적 실험 결과

중요하지 않은 특성에 대해 섞기(shuffling)를 진행하더라도 학습 모델의 에러(error)는 증가하지 않는다[17]. 하지만 다수의 특성 중 특정 특성만이 모델의 성능에 직접적인 연관이 있다는 것에 의거[18], 학습 속도와 모델의 질을 향상하기 위해 특성을 선택하는 데에 있어 특성의 중요도를 파악하지 않고 임의로 제거한다면, 모델의 에러는 증가하게 된다. 이는 임의로 제거할 특성을 선택할 때, 실제 중요도가 있는지 판단할 필요가 있다는 의미이다.

앞서 무역 관련 전문 지식 없이 제거할 특성을 임의로 선택하였기 때문에 학습 결과에 영향을 미칠 것이라 판단하여 여기서는 전체 데이터 사용 모델 (Full-Data Model)로 실험을 진행하였다. 이상치에 대한 방안으로, 평균 절댓값 오차로 학습을 진행한 결과와 로스값을 기준으로 John Tukey가 제안한 Tukey's fences[19]에 따라 1사분위수와 3사분위수를 구한 뒤 IQR을 계산하여 1사분위수에서 IQR의 1.5배를 뺀 값과 3사분위수에서 IQR의 1.5배를 더한 값의 사이에 분포하지 않는 값을 이상치(outlier)로 설정하고 해당 데이터를 제거한 뒤 다시 평균 절댓값 오차를 도출하였다. Fig. 1에서 볼 수 있듯, Full-Data Model이 Selected-Data Model보다 이상치(outlier) 제거 후 로스값이 약 6% 낮은 것으로 확인할 수 있다.

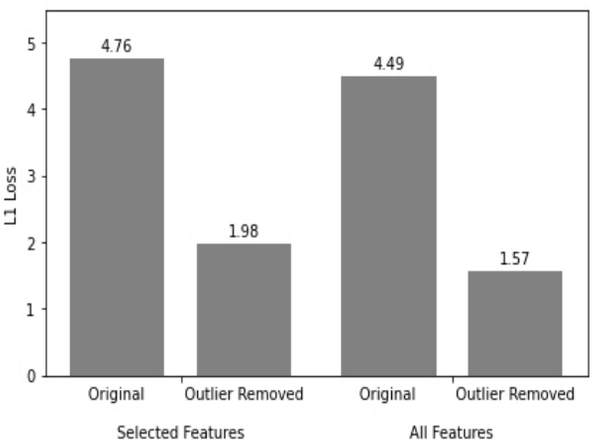


Fig. 1. Loss of Each Model

또한, 모델의 크기가 증가할수록 학습 및 테스트 시간이 길어지고, 과적합(Overfitting)을 처리하기 어렵기 때문에 모델 크기를 작게 유지하고 드랍아웃 층을 추가하여 과적합을 방지하고자 하였다[15].

4. 실 험

다층 퍼셉트론을 사용한 예측 모델의 학습을 위해 총 20713개의 2017년 KOTRA 데이터셋을 6:4의 비율로 학습 데이터와 테스트 데이터로 분리하여 학습을 진행했으며, 총 100 에포크(epoch)를 수행하였다. Table 4에 본 실험을 위해 적용한 하이퍼 파라미터(hyper-parameter)를 정리하였다.

앞서 제안한 다층 퍼셉트론 기반 예측 모델의 출력은 수입금액의 증감비율이다. 이렇게 예측된 수입금액 증감비율을 과년도 우리나라가 해당 품목을 수출한 금액에 곱하면, 차년도 우리나라의 해당 품목 수출금액을 예측할 수 있다.

Table 5는 다층 퍼셉트론 모델로 테스트 데이터를 분석한 결과이며, 예측 증감률에 대해 실제와의 차이를 기준으로 정렬하여 상위 세 케이스와 히위 세 케이스로 정리하였다. 1번 데이터를 보면, Item Code가 840734인 품목을 과년도 Australia가 우리나라로부터 70,528,230달러 수입하였고, 차년도에는 1,458,527달러 수입하였다. 해당 데이터에 대한 실제 증감률은 약 2.068%이며, 우리 모델은 이를 2.075%로

Table 4. Hyper-parameters

Hyper-parameter	Size
Learning rate	1e-5
Batch size	32
epoch	100

Table 5. Prediction Result

	Percentage			Trade Price		Country	Item Code
	Ground Truth	Prediction	Diff	Previous	Next		
1	2.068004608	2.075022936	0.007018328	70,528,230	1,458,527	Australia	840734
2	4.616559982	4.609478474	0.007081509	39,207,961	1,810,059	Philippines	840999
3	0.403601289	0.396152377	0.007448912	73,649,172	297,249	Chile	841391
4	92.46986389	3.792806149	88.67705774	118,502,639	109,579,232	India	842890
5	102.6585083	2.59561491	100.0628934	1,833,623	1,882,370	Egypt	830230
6	113.8865128	8.586242676	105.3002701	4,900,496	5,581,004	Viet Nam	830710

예측하였다. 약 0.007%의 오차로 예측했다는 의미이다.

이처럼 근사치로 예측한 경우도 있었지만 반대로, 예측에 크게 벗어난 경우도 확인할 수 있었다. 4번 데이터를 보면 Item Code 842890 품목에 대해 India가 우리나라로부터 과년도 수입한 금액인 118,502,639달러와 차년도 수입한 금액인 109,579,232달러에 따라 실제 증감률은 92.469%이지만, 모델의 예측결과는 3.792%이다. 이는 약 88.67%의 오차로 예측했다는 의미이다.

Fig. 2에서 다층 퍼셉트론 모델로 테스트 데이터를 예측하여 나온 오차값을 막대그래프와 분포도로 표현하였다. 앞선 부분에서 오차가 크게 벗어난 경우가 관찰되지만, 이는 분포도 x축 기준 최우측과 같이 오차값이 매우 큰 이상값에 의한

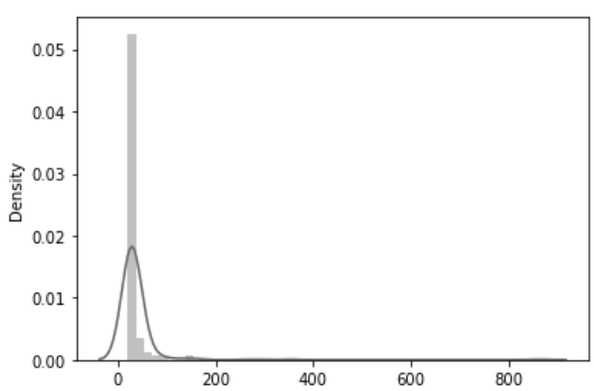


Fig. 2. Distribution of Loss

소수 케이스에 불과한 상황이다. 이에 대한 정확한 분석을 위해 추가 실험을 진행하였다.

앞서 3.3절에서 밝힌 것과 같이, 본 실험에서는 다변량 회귀, 다층 퍼셉트론, 순환 신경망 총 세 모델을 사용하였다. 학습에 대한 결과는 Table 6에서 확인할 수 있다.

결과적으로, 로스는 평균 절댓값 오차 기준, 다층 퍼셉트론과 순환 신경망은 유사하고, 다변량 회귀 모델이 가장 높다. 예외사항으로, 절댓값 오차로 학습된 다변량 회귀 모델은 양수의 값을 예측하지 못하고 음수를 예측하는 결과가 소수 존재하는 바, 성능 비교를 위한 모델로 평균 절댓값 오차로 학습된 다변량 회귀 모델을 사용하였다.

세 모델 모두 이상치에 영향을 적게 받는 평균 절댓값 오차의 수치가 평균 제곱 오차값에 비해 더 낮기 때문에 이상치가 존재함을 확인했으며, 두 로스 모두 상대적으로 높은 값이 도출되어, 오차들을 사후분석 해보았다.

Fig. 3의 다변량 회귀(A), 다층 퍼셉트론(B), 순환 신경망(C) 세 그래프는 모델의 학습 로스와 테스트 로스를 그래프로 출력한 것이다. 다변량 회귀 모델(Fig. 3-A)의 경우, 수렴시간이 다른 모델에 비해 느린 것을 확인할 수 있다. 이를 통해 다층 퍼셉트론(MLP), 순환 신경망(LSTM) 모델이 다변량 회귀 모델에 비해 초기성능이 더 높고, 수렴 지점까지 소요되는 시간과 비용이 더 적음을 알 수 있다.

Fig. 4에 다층 퍼셉트론(A)과 순환 신경망(B)에 대한 로스값을 확대해 보았다. 학습 데이터에 대한 예측 로스와 테스트

Table 6. Train Results

	Linear Regression		MLP		LSTM	
Loss Function	L1	L2	L1	L2	L1	L2
Loss (Test Data)	7.1	7.5	4.5	7.1	4.5	7.3
Loss (Outlier Removed)	2.6	4.0	1.6	3.3	1.7	3.3
Number of Outlier	456	415	488	388	406	470
Convergence Time	Slowest		Medium		Fastest	
Significant	negative number	-	-		-	

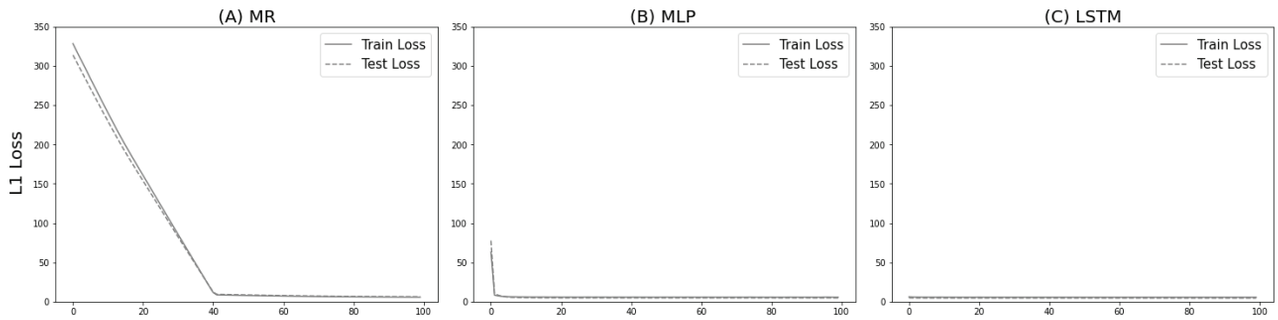


Fig. 3. Loss of Each Model

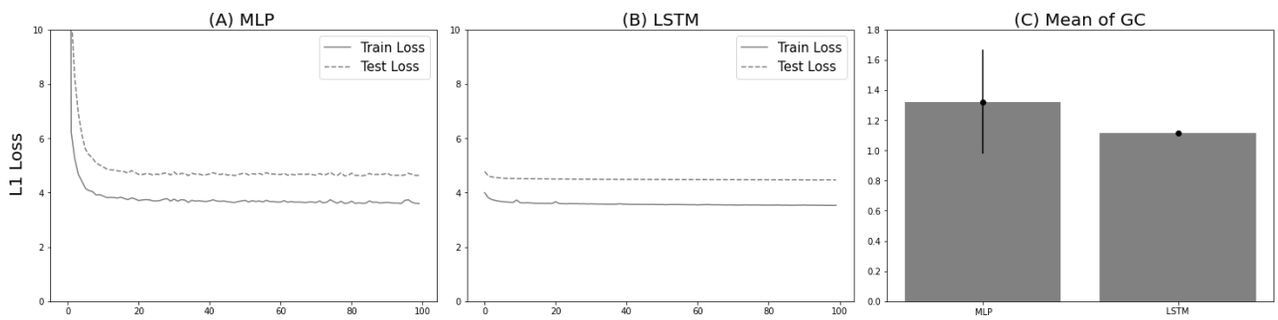


Fig. 4. Magnified Loss Plots of MLP and LSTM

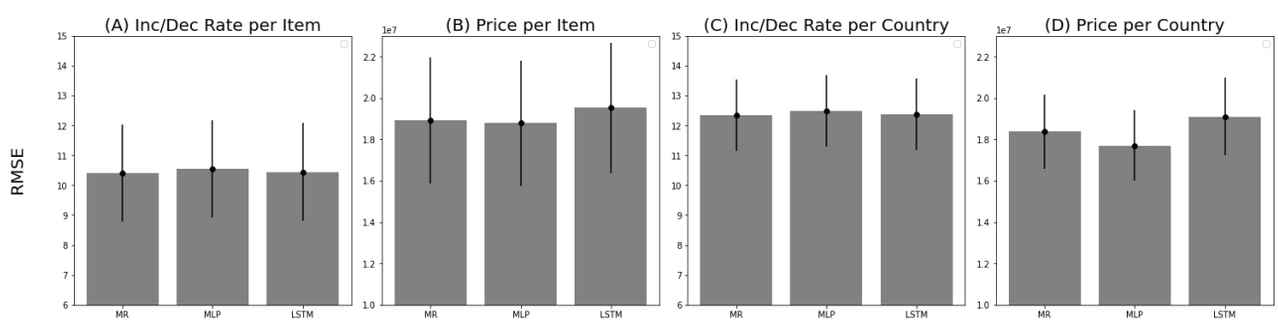


Fig. 5. RMSE of Each Model

데이터에 대한 예측 로스 사이에 차이가 존재하는 것이 관찰되었다. 이 두 로스의 차이를 일반화 능력의 척도로 간주하고 (Eq. 1), 이 척도를 이용하여 다층 퍼셉트론과 순환 신경망의 일반화 능력을 비교해 보았다.

$$GC = \ | loss_{test} - loss_{train}| \tag{1}$$

각 모델 수렴 후 40 에포크 동안 GC(Generalization Capacity) 값들을 표본으로 선택하고, 두 모델 간 일반화 능력 차이를 t-test를 이용하여 비교하였다. 그 결과, N=40, $p<0.0001$로써, 모델의 일반화 능력은 Fig. 4-C에서 볼 수 있듯 평균 GC값이 더 낮은 순환 신경망 모델이 다층 퍼셉트론 모델보다 통계적으로 더 높다고 해석할 수 있다. 모델별 평균 GC 값(Fig. 4-C) 에러바는 표준오차(Standard error of the mean; SEM)이다.

평균 제곱근 오차 (Root mean square error; RMSE)를 사용하여, 각 모델의 예측 성능을 추가로 진행하였다.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y} - y)^2}{n}}, \tag{2}$$

여기서 n은 전체 데이터 개수, y는 실제 측정값, $\hat{y}$은 예측값을 나타낸다.

Fig. 5는 순서대로, 품목(Item)별 예측 증감률(Fig. 5-A), 품목별 예측 가격(Fig. 5-B), 나라별 예측 증감률(Fig. 5-C), 나라별 예측 가격(Fig. 5-D)을 실제 측정치와의 차이인 평균

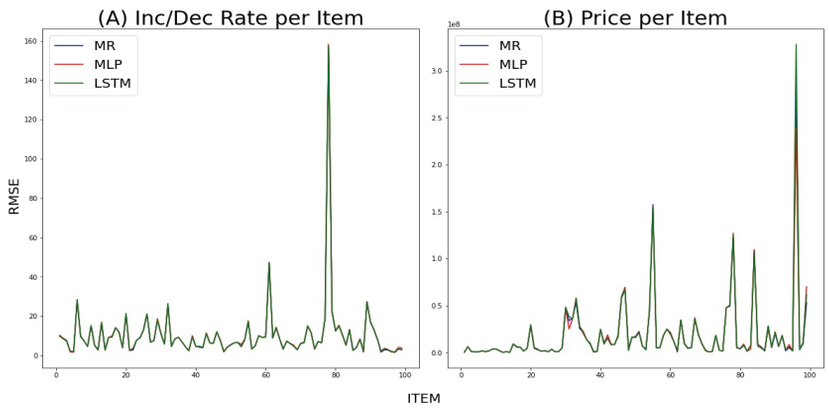


Fig. 6. RMSE of Each Item for Each Model

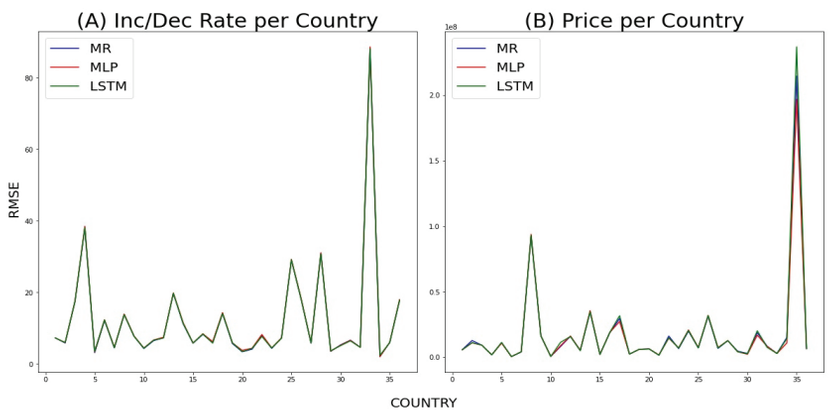


Fig. 7. RMSE of Each Country for Each Model

제곱근 오차로 나타낸 그래프이다. 에러바는 표준오차이다.

증감률(Fig. 5-A, C)에 대하여, 평균 제곱근 오차값이 다층 퍼셉트론 기준으로 다변량 회귀 모델은 +1.4%, 순환 신경망 모델은 +1.1% 차이를 보이는데, 모델별 차이가 크지 않다. 하지만 이런 작은 차이도 실제 값(Fig. 5-B, D)을 예측하게 된다면 값의 차이가 뚜렷해진다. Fig. 5-B와 Fig. 5-D에서 확인할 수 있듯, 다층 퍼셉트론의 평균 제곱근 오차 값이 가장 작음을 확인할 수 있다.

해당 평균 제곱근 오차 값들의 의미를 확인해보기 위해 Fig. 6에 각 품목별로 모델별 평균 제곱근 오차를 실제 가격, 증감비율 두 가지에 대하여 계산한 뒤 출력해 본 결과, 세 모델 모두 특정 품목에서 평균 제곱근 오차가 높은 것을 확인할 수 있다. 또한 (Fig 7)에 각 나라별로 나누어 평균 제곱근 오차를 출력해 보았다. 마찬가지로 세 모델 모두 특정 나라에서 평균 제곱근 오차가 공통적으로 높은 것을 확인할 수 있다.

Fig. 6과 Fig. 7을 통해 세 모델 모두 특정 품목과 특정 나라별로 실제 가격 예측과 증감률 예측에 크게 벗어난 결과가 동시에 존재함을 알 수 있다. 이는 KOTRA 데이터에 포함된 이상 값에 대해서 우리 모델이 대처하지 못한다는 점을 시사한다. 다시 말해, 이상 값이 존재하며, 이를 고려하여 예측 성능을 높이기 위해서는 추후 이에 대한 원인을 확인할 필요가 있다.

5. 고 찰

다변량 회귀와 신경망을 기반으로 한 다층 퍼셉트론 모델을 사용하여 과/차년도 우리나라의 수출 금액과 그에 관련된 특성을 가진 KOTRA 데이터를 통해 학습을 진행했고, 우리나라의 차년도 수출금액 증감 비율을 예측하여 성능평가를 진행하였다.

평균 제곱근 오차를 통해 다변량 회귀 모델, 순환 신경망 모델과 성능평가를 진행하였으며, 상위 두 모델에 비해 더 나은 정확도를 보인 다층 퍼셉트론 모델을 통하여 나온 예측 결과는 오차율 4%대, 이상치를 제거한 오차율은 1%대로 계산되었다. 이는 실제 값과 예측 값의 차이가 4% 이내 라는 것을 의미한다. 무역 데이터 비전문가의 입장에서 학습 결과로 나온 오차율이 적정한 수치인지 판별하기 위해서는 더 엄밀한 기준을 확립할 필요가 있다. 그럼에도 불구하고, 무역 데이터

를 통해 우리나라의 차년도 수출금액을 실제 값에 가까이 예측할 수 있는 가능성은 충분히 확인하였다.

또한 학습된 모델의 예측 오차를 Fig. 2와 Fig. 3에서 간단히 살펴보고, Fig. 6, Fig. 7에서 무역 품목별, 나라별로 자세히 확인하였다. 앞서 제시된 세 모델들에 대하여 이상값이 공통적으로 존재했던 것처럼 모든 데이터에 대하여 균일한 예측이 불가능했던 이유는 모델이 학습할 수 없는 데이터에 포함되지 않은 외적인 요인이 존재하기 때문으로도 생각해 볼 수 있다. 따라서 이는 세계정세, 전염병, 경제 상황 등 무역환경에 영향을 줄 수 있는 외부 요인들이 있었다는 의미로 유추해볼 수 있다. 이러한 요인들을 고려하여 모델링 한다면, 예측과 실제 차이를 줄일 수 있을 것으로 보인다. 따라서 추후 국제 정세, 전염병, 경제 상황 등 무역환경의 변화에 영향을 줄 수 있는 인자를 확인하고 예측 모델에 적용할 수 있는 연구가 필요하다.

추가로, 연구에서 제시된 일반 다층 퍼셉트론 모델과 달리 더 복잡한 구조의 모델들, 예를 들어 합성곱 신경망(CNN)[20], 르넷(LeNet)[21]과 같은 새로운 모델을 통해 데이터 특징 추출에 대해 인간의 개입을 최소화 하고, 더 많은 수의 특성 선정 및 모델이 스스로 특성을 유연하게 다루도록 시도해 볼 수 있다. 이는 각 특성간의 상관관계를 내재적으로 고려하여, 예측 정확도를 더 높일 수 있을 것이다.

또는 앞서 전처리 과정에서 언급했듯, 이번 실험에서는 한 해 동안의 품목별 무역 데이터를 사용하여 1년 단위의 예측을 하였지만, 추후 품목들에 대한 장기간의 데이터가 주어진다면 이를 예측하기 위해 시계열 데이터에 강인한 장단기 메모리 [13]모델이 모든 면에서 더 나은 결과를 보일 것이다. 또는 장단기 모델에서 더 진보된 트랜스포머(Transformer)[22]도 사용할 수 있을 것이다.

References

[1] Korea International Trade Association [Internet], https://stat.kita.net.

[2] B. M. Lee, H. J. Jeong, and K. S. Park, "An influence of the fourth industrial revolution on international trade and countermeasure strategies to promote export in Korea," *Korea Trade Review*, Vol.42, No.3, pp.1-24, 2017.

[3] S. H. Nam, "Comparison of long-term forecasting performance of export growth rate using time series analysis models and machine learning analysis," *Korea Trade Review*, Vol.46, No.6, pp.191-209, 2021.

[4] The 9th Public Data Utilization BI Contest [Internet], http://www.datacontest.kr (retrieved 20210926)

[5] S. Weisberg, "Applied linear regression," John Wiley & Sons, pp.47, 2005.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, Vol.521, No.7553, pp.436-444, 2015.

[7] A. Keck, A. Raubold, and A. Truppia. "Forecasting international trade: A time series approach," *OECD Journal: Journal of Business Cycle Measurement and Analysis*, Vol.2009, No.2, pp.157-176, 2010.

[8] A. W. Veenstra and H. E. Haralambides. "Multivariate autoregressive models for forecasting seaborne trade flows," *Transportation Research Part E: Logistics and Transportation Review*, Vol.37, No.4, pp.311-319, 2001.

[9] E. S. Silva and H. Hassani. "On the use of singular spectrum analysis for forecasting US trade before, during and after the 2008 recession," *International Economics*, Vol.141, pp.34-49, 2015.

[10] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2011.

[11] H. Lu, X. Ma, K. Huang, and M. Azimi, "Carbon trading volume and price forecasting in China using multiple machine learning models," *Journal of Cleaner Production*, Vol.249, pp.119386, 2020.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, PMLR, 2015.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol.9, No.8, pp.1735-1780, 1997.

[14] Analytics Vidhya, "Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization," [Internet], https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/(retrieved 20200403).

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, Vol.15, No.1, pp.1929-1958, 2014.

[16] R. Genuer, J. M. Poggi, and C. Tuleau-Malot. "Variable selection using random forests," *Pattern Recognition Letters*, Vol.31, No.14, pp.2225-2236, 2010.

[17] Christoph Molnar, "Permutation feature importance," [Internet], https://christophm.github.io/interpretable-ml-book/feature-importance.html (retrieved 20220217).

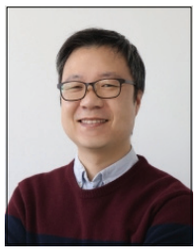
[18] K. Kira and L. A. Rendell, "A practical approach to feature selection," *Machine Learning Proceedings 1992*, Morgan Kaufmann, pp.249-256, 1992.

[19] J. W. Tukey, "Exploratory data analysis," Addison-Wesley. ISBN 978-0-201-07616-5. OCLC 3058187, 1977.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, Vol.60, No.6, pp.84-90, 2017.
[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, Vol.86, No.11, pp.2278-2324, 1998.
[22] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

김 지 훈

https://orcid.org/0000-0003-4176-5766
e-mail : neti2207@gmail.com
2017년 ~ 현 재 상명대학교 휴먼지능정보공학전공 석사과정
관심분야 : Deep Learning, Meta Control RL



이 지 항

https://orcid.org/0000-0002-4337-2774
e-mail : jeehang@smu.ac.kr
2015년 University of Bath(박사)
2000년 ~ 2005년 한글과컴퓨터 주임연구원
2005년 ~ 2010년 삼성전자 DMC연구소 책임연구원
2015년 ~ 2016년 University of Bath, Research Associate
2017년 ~ 2019년 한국과학기술원 KI-Postdoc
2019년 ~ 2020년 한국과학기술원 바이오및뇌공학과 연구조교수
2020년 ~ 현 재 상명대학교 휴먼지능정보공학과 조교수
관심분야 : 의사결정, 규범추론, 뇌기반인공지능