JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 9, pp. 37-44, September 2020
https://doi.org/10.9708/jksci.2020.25.09.037

Proposal of a Hypothesis Test Prediction System for Educational Social Precepts using Deep Learning Models

Su-Youn Choi*, Dea-Woo Park*

*Graduate Student, Dept. of Convergence Engineering, Hoseo Graduate School of Venture, Seoul, Korea
*Professor, Dept. of Convergence Engineering, Hoseo Graduate School of Venture, Seoul, Korea

[Abstract]

AI technology has developed in the form of decision support technology in law, patent, finance and national defense and is applied to disease diagnosis and legal judgment. To search real-time information with Deep Learning, Big data Analysis and Deep Learning Algorithm are required. In this paper, we try to predict the entrance rate to high-ranking universities using a Deep Learning model, RNN(Recurrent Neural Network). First, we analyzed the current status of private academies in administrative districts and the number of students by age in administrative districts, and established a socially accepted hypothesis that students residing in areas with a high educational fever have a high rate of enrollment in high-ranking universities. This is to verify based on the data analyzed using the predicted hypothesis and the government's public data. The predictive model uses data from 2015 to 2017 to learn to predict the top enrollment rate, and the trained model predicts the top enrollment rate in 2018. A prediction experiment was performed using RNN, a Deep Learning model, for the high-ranking enrollment rate in the special education zone. In this paper, we define the correlation between the high-ranking enrollment rate by analyzing the household income and the participation rate of private education about the current status of private institutes in regions with high education fever and the effect on the number of students by age.

▸ **Key words**: AI, Big data Analysis, Deep Learning, Python, Programming, RNN

[요 약]

AI 기술은 법률, 특허, 금융, 국방의 의사결정지원 기술 형태로 발전하여 질병 진단과 법률 판정 등에 적용되고 있다. Deep Learning으로 실시간 정보를 검색하려면, Big data Analysis과 Deep Learning Algorithm이 필요하다. 본 논문에서는 Deep Learning 모델인 RNN(Recurrent Neural Network)을 이용하여 상위권 대학 진학률을 예측하고자 한다. 우선, 행정구역 사설학원 현황과 행정구역 연령별 학생 수를 분석하고 교육열이 높은 지역에 거주하는 학생이 상위권 대학 진학률이 높다는 사회 통념의 가설을 설정했다. 예측된 가설과 정부의 공공데이터를 활용하여 분석된 자료를 토대로 검증하고자 한다. 예측 모델은 2015년부터 2017년까지의 데이터를 활용하여 상위권 진학률을 예상하도록 학습하고, 학습된 모델은 2018년 상위권 진학률을 예측한다. 교육특구지역의 상위권 진학률을 Deep Learning 모델인 RNN을 이용하여 예측 실험을 수행했다. 본 논문은 교육열이 높은 지역의 사설학원 현황, 연령별 학생 수에 미치는 영향에 대해서 가구소득, 사교육의 참여 비율을 분석하여 상위권 진학률의 상관관계를 정의한다.

▸ **주제어**: 인공지능, 빅 데이터 분석, 딥 러닝, 파이썬, 프로그래밍, 순환신경망

• First Author: Su-Youn Choi, Corresponding Author: Dea-Woo Park
*Su-Youn Choi (mibm400@hanmail.net), Dept. of Convergence Engineering, Hoseo Graduate School of Venture
*Dea-Woo Park (prof_pdw@naver.com), Dept. of Convergence Engineering, Hoseo Graduate School of Venture
• Received: 2020. 07. 24, Revised: 2020. 08. 27, Accepted: 2020. 08. 31.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인공지능(AI; Artificial Intelligence)은 인간이 가진 지각, 학습, 추론, 자연어 처리 등의 능력을 컴퓨터가 실행할 수 있도록 하는 기술로 Machine Learning, Deep Learning, 자연어 처리(Natural language processing), 음성인식(Speech recognition), 시각 인식(Visual recognition) 등이 이에 속한다[1].

4차 산업혁명을 발전시키는 대표적 기술인 AI가 발전하려면 많은 비용이 소요되는 학습용 데이터가 필요하다. 데이터의 산업적, 사회적 가치가 인정되면서 데이터의 활용은 급진적으로 늘어나고 있다.

실시간으로 수많은 데이터가 쏟아지고 있는 상황에서 데이터는 마케팅, 의료 분야는 물론 관광이나 복지 등 공공 서비스 차원에서도 Big data 활용 범위가 넓어지고 있다. 또한 혁신적으로 발전한 Algorithm, Big data, Cloud, Computing power 등이 서로 융·복합되면서 실제 구현을 통해 산업전반에 적용되어 다양한 현실 세계의 문제를 해결하고 있다. 하지만 우리나라는 데이터에 대한 체계적 연구나 중요성에 대한 인식이 부족하다.

과학기술정보통신부는 2020년 7월 14일에 '한국판 뉴딜 계획 종합계획'을 발표하고 디지털 뉴딜의 중심축인 데이터, AI, Cloud 분야에 집중적으로 지원하여 데이터 경제를 가속화시키려 하고 있다. AI 기술의 빠른 발전과 서비스 적용을 위해서는 정부차원의 데이터 공유와 확산을 위한 노력이 필요하다[2].

본 논문에서는 행정구역 사설학원 현황과 행정구역 연령별 학생 수를 분석하고 Deep Learning 모델인 RNN을 이용하여 상위권 대학 진학률을 예측한다. 예측 모델은 2015년부터 2017년까지 데이터를 활용하여 상위권 진학률을 예측하도록 학습하고, 학습된 모델은 2018년 상위권 진학률을 예측한다. 교육열이 높은 지역의 사설학원 현황, 연령별 학생 수에 미치는 영향에 대해서 가구소득, 사교육의 참여 비율을 분석하여 상위권 진학률의 상관관계를 정의한다.

II. Related Works

1. Types of prediction models based on Deep Learning

Deep Learning은 사람의 신경세포를 추상적으로 표현한 AI를 기반으로 하고 정보의 입력, 가공, 출력으로 구성된다. 하나의 세포는 같은 계층에 여러 형태의 입력으로 존재할 수 있고 가공된 정보는 Hidden 계층으로 여러 개의 입력을 다양한 형태로 받아들여 이를 조합하고 분석하는 정보에 따라 여러 개의 Hidden 계층을 둘 수 있다. 출력은 분석 결과를 가지고 결과의 정확도가 높으면 해당 학습 모델을 사용하여 분석과 예측을 할 수 있다.

Deep Learning 기반의 예측 모델 중에 이미지 처리와 인식 분야에서 많이 사용되는 CNN(Convolutional Neural Network)은 합성곱(Convolution) 필터를 이용한 신경망 네트워크이다. 신경망 네트워크 앞에 여러 계층의 합성곱 계층을 붙여 사용하고자하는 이미지의 특징을 추출하고 추출된 특징을 기존의 신경망 네트워크를 이용하여 분류하는 과정을 거친다.

RNN은 일반적인 신경망 네트워크와 달리 Hidden 노드가 방향을 가진 엣지로 연결되어 순환구조(Circulation Structure)를 이루는 인공신경망의 한 종류이다. 음성, 문자 등 순차적으로 등장하는 데이터 처리에 적합한 모델이다.

다음 그림 1은 RNN의 기본 구조를 나타낸 것이다. Hidden 노드의 순환 구조를 풀어서 나열한 그림으로 시퀀스 구조라고 한다. 시퀀스의 길이는 정해져 있지 않으며 입력(X)과 출력(h)은 시간에 따라 발생하며, Hidden 노드 A는 이전 출력과 현재 입력에 영향을 받는 구조로 되어 있다.

RNN은 Hidden 노드에서 사용하는 활성함수로 하이퍼볼릭탄젠트(tanh)를 사용한다. A는 이전 단계의 A정보와 현재 입력인 Xt의 영향을 받아 ht를 출력으로 보여준다[3].

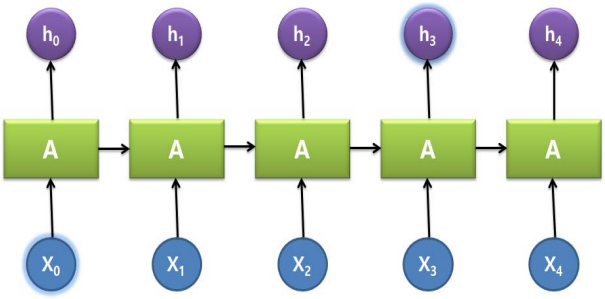

Fig. 1. RNN Basic Structure

LSTM(Long Short Term Memory) Algorithm은 RNN의 단점을 해결하기 위한 Algorithm으로 RNN은 계층이 깊어질수록 과거의 데이터를 학습하지 못하고 최근 데이터만을 기억하여 학습이 제대로 되지 않는 Vanishing Gradient(기울기 0) 문제가 발생한다[4].

인공신경망에서 학습 시 기울기 문제는 중요한 요소이다. 짧은 기간을 학습 할 때는 문제가 없지만 긴 기간을 학습하다보면 기울기가 0에 가까워지면서 학습이 느려지게 된다.

모든 단계의 학습에서는 상승과 하락이 존재하며 학습이 고정된 기울기를 가지면 기능을 하지 못하는 것으로 판단한다. 기울기는 학습에 대한 변화량을 의미하고 이 변화량이 매우 작으면 신경망 네트워크가 효과적으로 학습하지 못하고 학습이 완료되지 않은 채 수렴하는 문제가 발생한다.

Deep Learning 모델에서는 입력 데이터에 대한 특정 추출과 문제 해결을 위한 복잡한 함수를 학습하기 위해 다수의 층을 갖는 신경망 구조를 사용한다. 또한 이런 복잡한 구조의 신경망을 학습시키기 위해 많은 데이터와 컴퓨팅 자원이 필요하다[5].

본 논문에서는 Deep Learning Algorithm 중에서 시계열 데이터 예측 및 분석에 유용한 RNN을 이용하여 상위권 진학률을 예측하고자 한다.

2. Big data Analysis Case

디지털 경제의 확산으로 데이터와 정보가 폭발적으로 증가하고 있다. 데이터 기반 사회에서 Big data는 국가경쟁력을 좌우하는 중요한 '자원'이므로 Big data 활용을 위한 국가 전략이 필요하다. 미국과 영국 등은 BIg data 시대를 맞이하여 공공 정보의 전면적인 개방과 데이터 활용을 통한 가치 창출을 국가 전략으로 내걸고 새로운 혁신을 도모하고 있다[6].

AI 기술의 발전으로 Big data에 대한 중요성이 확대되고, AI, IoT, Cloud와 더불어 발전했다. Big data는 다양한 니즈에 부합할 수 있는 실시간성 확보가 중요하여 데이터 처리가 가속화되고 데이터 품질이 AI 기술의 가치에 좌우되므로 대량 데이터의 품질이 강화되고 있다.

방대하고 다양한 데이터로부터 가치 있는 정보를 추출해 낼 수 있는 가능성 때문에 민간 기업은 Big data 활용에 주목하고 있다. 민간 분야의 Big data 사례를 살펴보면 Google은 데이터 기반 혁신 기업으로 효과적인 검색엔진 개발을 위해 고가의 장비 대신 저렴한 PC를 대량으로 이용하는 하드웨어 전략과 함께 대용량 데이터를 분산 병렬처리하기 위해 MapReduce와 Google File System 등을 개발했다.

Google이 운영하는 YouTube는 개인맞춤형 동영상 추천 시스템을 가동하여 이용자의 실제 이용 데이터를 분석하여 가장 선호할 것으로 추정되는 동영상을 추천 해주는 방식이다.

미국 메이저리그는 Big data를 활용하여 Trackman이라는 야구공 추적 레이더 장비와 선수들의 움직임을 추적하는 카메라(ChyronHego)로 다양한 데이터를 측정하고 분석한다.

Netflix는 이용자 데이터 분석에 기초한 추천시스템으로 선호하는 동영상 서비스를 제공한다.

AI 분야의 Big data 활용사례로는 Google의 DeepMind Technologies Limited사가 개발한 AI 바둑프로그램인 AlphaGo가 있다[7].

Big data는 Deep Learning의 학습데이터로 많이 활용되고 있기 때문에 정확한 분석이 필요하다. AI의 높은 정확도와 실제사회 적용을 위해서는 Deep Learning Algorithm에 대한 연구가 필요하다.

III. Big data Analysis using Python

1. Analysis of public data for hypothesis testing

본 논문에서는 교육열이 높은 지역에 거주하는 학생이 상위권 대학 진학률이 높다는 사회 통념의 가설을 설정했다. 예측된 가설과 정부의 공공데이터를 활용하여 분석된 자료를 토대로 검증하고자 한다.

가설 검증 방법 중 Big data Analysis 방법에 사용할 데이터는 정부의 공공데이터 공개자료를 사용한다. 공개자료 중 '주민등록 연령별 인구 현황'에서 서울시 자치구별로 데이터를 추출하고 가설 검증을 위하여 교육특구로 불리는 강남구, 노원구, 양천구의 연령별 인구를 분석한 후 주변 사설학원 현황을 분석하여 서울 상위권 대학 진학률에 대한 연관성을 분석하고자 한다.

다음 표 1을 보면 행정구역 중에서 양천구 등 5개 구로 전입한 인구수가 총 2,203명으로 인구가 집중되고 있는 현상을 볼 수 있다. 일반적으로 자녀가 있는 가정의 전입은 자녀의 초등학교 입학 전, 후를 기점으로 많이 나타난다. 2019년 1, 2월 서울 초등학교 1학년 전입/전출현황을 보면 서울시내에서 이동했거나 타 시도에서 서울로 전입한 전체 숫자가 4,939명이다. 전체 수 4,939명과 강남구, 노원구, 서초구, 송파구, 양천구로의 전입 수 2,203명과 비교하면 절반에 가까운 수치로 매우 높음을 알 수 있다[8].

Table 1. Seoul Elementary School 1st Grade Transfer/Transfer Status

Division	Moving in					Move out				
	Total Moving in	In Seoul			Transfer to another city	Total move out	In Seoul			Transfer to another city
		Town subtotal	within the zone	Other gu Moving in			Town subtotal	within the zone	Other gu Moving out	
Songpa-gu	787	580	337	243	207	528	423	337	86	105
Gangnam-gu	468	350	163	187	118	297	235	163	72	62
Yangcheon-gu	362	280	139	141	82	241	181	139	42	60
Seocho-gu	323	278	159	119	45	285	236	159	77	49
Nowon-gu	263	185	107	78	78	209	148	107	41	61
Dongjak-gu	233	196	125	71	37	286	215	125	90	71
Eunpyeong-gu	230	170	126	44	60	233	184	126	58	49
Gangseo-gu	225	186	136	50	39	307	208	136	72	99
Gangdong-gu	206	157	111	46	49	224	160	111	49	64
Seongbuk-gu	195	166	102	64	29	220	172	102	70	48
Mapo-gu	164	123	73	50	41	206	173	73	100	33
Seodaemun-gu	160	128	79	49	32	175	140	79	61	35
Guro-gu	144	113	72	41	31	201	127	72	55	74
Gwangjin-gu	137	109	76	33	28	145	119	76	43	26
Yeongdeungpo-gu	124	78	50	28	46	209	136	50	86	73
Dongdaemun-gu	123	111	78	33	12	166	134	78	56	32
Jungnang-gu	113	100	85	15	13	161	120	85	35	41
Yongsan-gu	111	92	68	24	19	138	112	68	44	26
Dobong-gu	107	92	69	23	15	131	105	69	36	26
Seongdong-gu	104	81	54	27	23	169	143	54	89	26
Gwanak-gu	100	83	61	22	17	198	133	61	72	65
Geumcheon-gu	84	58	47	11	26	121	74	47	27	47
Gangbuk-gu	79	66	49	17	13	118	93	49	44	25
Jongno-gu	73	62	29	33	11	75	56	29	27	19
Jung-gu	24	19	8	11	5	45	36	8	28	9
total	4939	3863	2403	1460	1076	5088	3863	2403	1460	1225

다음 표 2는 서울특별시교육청에서 발표한 2019년 사설학원 현황이다. 행정구역별 사설학원 수를 비교한 결과 강남구, 양천구, 송파구, 서초구, 노원구 순으로 나타났다. 교육특구로 불리는 이 지역은 좋은 학군과 사교육을 받을 수 있는 학원가가 형성되어 있어 사교육 환경이 잘 조성되어 있다. 교육 환경이 우수한 지역을 선호하므로 그 결과 교육특구인 강남구, 양천구, 노원구 등 주거선호도가 높은 지역으로 집중되고 있다.

Table 2. Private School Status in 2019

Administrative area	Academy number	Average monthly teaching time	Monthly average tuition
Gangnam-gu	1,153	28	383,511
Yangcheon-gu	721	25	275,893
Songpa-gu	651	26	270,890
Seocho-gu	539	24	331,538
Nowon-gu	485	26	245,895
Gangseo-gu	422	23	236,715
Gangdong-gu	421	24	238,887
Eunpyeong-gu	359	25	230,884
Seongbuk-gu	274	27	246,037
Mapo-gu	257	33	250,595
Gwangjin-gu	241	33	269,920
Gwanak-gu	229	25	228,453
Guro-gu	226	30	222,348
Dongjak-gu	221	22	210,612
Dobong-gu	216	23	220,531
Dongdaemun-gu	204	26	237,954
Seodaemun-gu	204	24	235,119
Yeongdeungpo-gu	192	25	246,377
Jungnang-gu	173	28	238,444
Seongdong-gu	152	26	247,060
Gangbuk-gu	141	24	218,034
Geumcheon-gu	121	25	219,921
Yongsan-gu	67	27	263,859
Jongno-gu	63	29	253,330
Jung-gu	38	26	248,272

2. Verification and Analysis of hypotheses according to social norms and variables

사교육비 지출의 원인을 분석하는 연구에서는 사교육비 지출의 주체가 학부모라는 것에 근거하여 사교육은 자녀 교육을 위한 재정적 지원이기 때문에 사교육비 지출 연구에 있어서 학부모의 사회경제적 지위(Socio-economic status)가 중요한 변인으로 고려되었다. 학부모의 학력, 거주지, 직업, 월평균 소득과 같은 학부모의 배경변인들에 따른 사교육비 지출의 차이를 부모의 사회경제적 지위가 높을수록 사교육비 지출이 많아진다는 결론이 도출되었다. 부모의 지적수준이 높을수록, 경제적 수준이 높을수록, 부

모가 자녀에게 보다 많은 물질적 자원을 투입할수록 자녀의 학업성취도에 영향을 미친 것으로 나타났다[9].

국가통계포탈에서 2019년 초중고 사교육비를 조사한 결과 가구 소득구간별 사교육비는 월평균 소득 700-800만원 미만에서 464,000원, 800만 원 이상에서 539,000원으로 전년대비 각각 9.7%, 6.6% 증가했다[10].

통계청에서 2019년 초중고 사교육비를 조사한 결과 2019년 사교육비 총액은 약 21조원으로 전년도 19조 5천억원에 비해 1조 5천억원으로 7.8% 증가했고 주당 참여시간도 6.5시간으로 전년대비 7.8%로 증가했다. 전체 학생 수는 전년대비 감소하였으나 참여율과 주당 참여시간은 증가했다.

다음 그림 2는 국가통계포탈에서 조사한 결과로 가구 소득수준별 1인당 월평균 사교육비 및 참여율을 나타낸 그래프이다. 가구의 월평균 소득수준이 높을수록 사교육비 지출과 참여율이 높은 양상을 나타내고 있다[11].

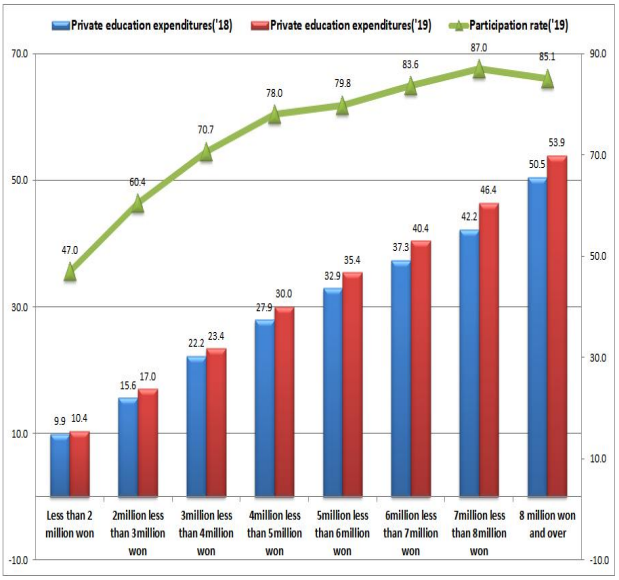

Fig. 2. Average monthly private education expenditures per student and private education participation rate by household income

국가통계포탈에서 2019년 초중고 사교육비조사 보고서를 보면 사교육 참여율은 가구의 월평균 소득 700-800만원 미만에서 87.0%로 가장 높게 나타났고 200만원 미만에서 47.0%로 가장 낮게 나타났다. 가구의 월평균 소득수준이 높을수록 사교육비 지출과 참여율이 높게 나타났다.

표 3은 성적 구간별 전체 고등학생 1인당 월평균 사교육비 및 참여율을 나타낸 것이다. 사교육비는 상위 10% 이내 학생은 475,000, 하위 20%이내 학생은 248,000으로 차이가 있으며 전년대비 23.7%, 17.5% 증가되었다. 학생 성적 구간별 사교육비 및 참여율을 보면 성적이 상위일수록 1인당 월평균 사교육비 지출과 참여율이 높게 나타났다.

Table 3. Average monthly private education expenditures per high school student and participation rate by school performance

Classification	Private education expenditures (10 thousand won, %)			Participation rate (%, %p)		
	2018	2019	Percent change	2018	2019	Change from the previous year
Total	32.1	**36.5**	13.6	58.5	**61.0**	2.4
Within top 10%	38.4	**47.5**	23.7	65.8	**72.3**	6.5
11 ~ 30%	38.7	**43.0**	11.1	64.9	**67.8**	2.9
31 ~ 60%	34.8	**38.5**	10.8	61.0	**62.8**	1.8
61 ~ 80%	29.0	**32.6**	12.7	55.5	**57.2**	1.7
81 ~ 100%	21.1	**24.8**	17.5	47.4	**48.9**	1.5

* From 2018, data by school performance cover only high school students.

서울특별시 가구를 대상으로 현 거주지로 이주한 이유를 조사한 결과 '자녀 교육 여건 때문에'라는 응답이 14.5%로 교통, 직장 변동에 이어 세 번째로 높은 비중을 차지했다. 특히 교육 환경이 우수한 강남구, 서초구, 양천구의 경우 각각 29.5%, 23.1%, 26.8%가 현 거주지로 이주한 이유가 자녀 교육 여건 때문이라고 응답했다[12].

이처럼 특정 지역에서 자녀 교육 여건을 더욱 중시하는 경향을 보이는 이유는 교육과 거주지가 밀접한 관계를 맺고 있기 때문이다.

3. Analysis of RNN algorithm for hypothesis verification

교육열이 높은 지역에 거주하는 학생이 상위권 대학 진학률이 높다는 사회 통념의 가설은 있다. 사실 이 가설은 한국사회의 일반적인 통념으로 자리하고 있다. 하지만 이 가설은 맞는 것일까?

다음 그림 3은 가설 검증을 위한 Big data Analysis 방법과 Deep Learning 모델인 RNN을 이용한 예측 방법론이다.

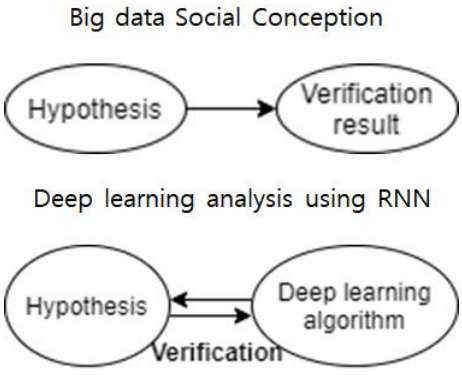

Fig. 3. Big data and Deep Learning Analysis using RNN

다음 그림 4는 가설 검증을 수행하기 위한 Deep Learning 모델인 RNN을 이용한 Flow chart다. 사회적 통설과 변수에 따른 가설의 검증 및 분석을 위해 RNN

Algorithm을 이용하여 예측한다.

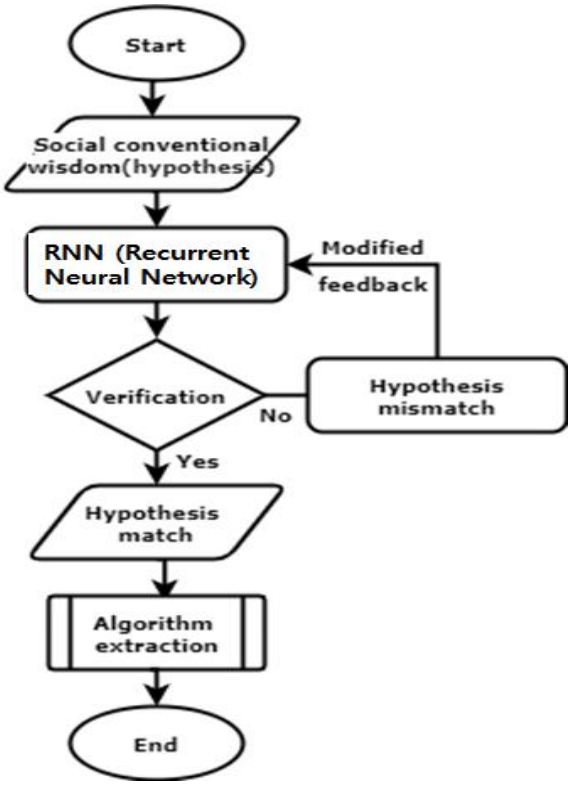

Fig. 4. Design of RNN Algorithm for hypothesis verification

4. School District Demand Analysis Using Python

교육부의 2018년 교육기본통계에 따르면 서울의 학령(유치원-고등학교) 인구는 995,587명으로 전년 1,037,349명보다 4.0% 감소하였다. 하지만 강남구는 10-14세 기준 1,182명이 순유입되었고 양천구도 465명이 늘었다. 강남구 대치동, 양천구 목동과 함께 서울 교육 특구로 손꼽히는 중계동이 속한 노원구는 13명 순유입되는데 그쳤으나 증가세를 기록했다. 따라서 인구가 학군수요를 쫓아 이동하고 있는 것으로 분석되고 있다.

인구가 학군수요에 따라 집중되고 있다는 것을 검증하기 위해 공공 Big data에서 추출한 가설과 검증을 위해 Python 3.8로 구현하였다.

다음 그림 5는 Python 소스코드를 실행한 결과로 2010년부터 2019년까지 교육특구 지역인 노원구, 강남구, 양천구의 연령별 인구수가 감소하고 있다는 것을 알 수 있다. 우리나라 저출산으로 인해 인구가 감소하고 있지만 교육특구 지역은 학군수요가 있는 한 크게 감소하지는 않을 것이다. 또한 교육특구 지역 중에서도 노원구 중계2,3동, 강남구 대치2동, 양천구 목5동이 다른 구에 비해 인구수가 2배 이상 집중되어 있었다. 인구가 학군수요에 따라 집중되고 있다는 것을 말한다.

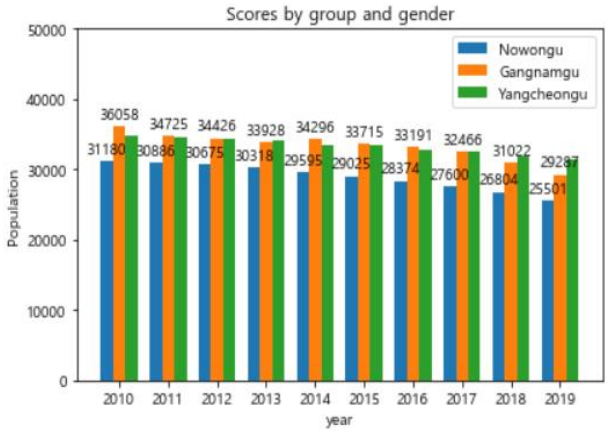

Fig. 5. Population Status by Year in Special Education Zones

IV. Prediction and Result Analysis

1. Python validation for data prediction

행정안전부에서 교육특구로 불리는 노원구, 강남구, 양천구의 연령별 인구수를 비교분석한 결과 10대 청소년과 그 부모세대인 40-50대의 인구 비중이 높게 나타났다[13].

교육특구의 인구 패턴을 분석하기 위해 공공 Big data에서 추출한 데이터를 사용하여 Python으로 coding한 결과 중계동, 대치동, 목동의 인구분포가 비슷한 결과가 나왔다[14].

그림 6와 같이 교육특구인 중계동, 대치동, 목동의 연령별 인구분포는 10대와 40대가 가장 높게 나타났고 각 구의 인구분포도는 비슷한 양상을 보이고 있다.

2016년에서 2018년까지 일반고에서 평균 10명 이상을 상위권 대학에 보낸 자치구는 모두 6개로 강남구, 양천구, 송파구, 노원구, 서초구, 강서구이다[15]. 교육특구지역의 상위권 진학률을 Deep Learning 모델인 RNN을 이용하여 예측 실험하여 비교 분석하고자 한다.

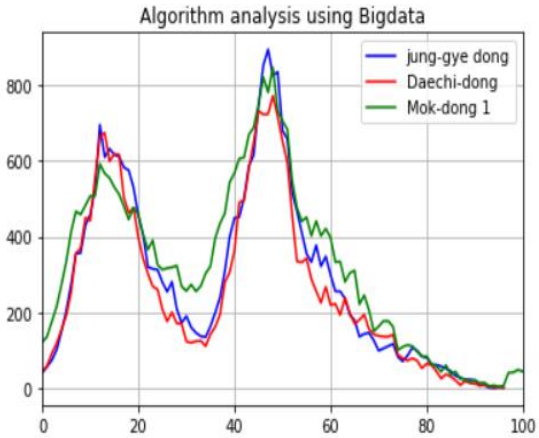

Fig. 6. Comparison of Population by Age in Special Education Zone

2015년부터 2019년까지 연령별인구수를 분석한 결과 강남구 대치동은 연령별 전체 인원수에 비해 10대는 24% - 25%, 그 부모세대인 40대는 26% - 27%의 비율이 나왔다. 양천구 목동은 연령별 전체 인원수에 대비해 10대는 20% - 21%, 그 부모세대인 40대는 25%의 비율이 나왔다. 노원구 중계동은 연령별 전체 인원수에 대비해 10대는 20% - 21%, 그 부모세대인 40대는 23%-34%의 비율이 나왔고 2015년부터 2019년까지 연령별 인구분포에 큰 변화가 없는 것을 알 수 있다.

서울특별시가 2017년 12월에 집계한 2018년 서울의 초등학교 학령인구(만 6-11세)는 435,106명으로 이 중 18.2%인 79,236명이 강남구, 노원구, 양천구에 살고 있다. 2018년 교육을 목적으로 노원구에 전입한 인구는 4,651명인데, 다른 시·도에서 온 경우가 2,645명으로 절반 이상이다. 이에 비해 양천구는 교육을 위해 전입한 3,203명 중 2,289명이 구 내에서 이동하거나 다른 구에서 전입해온 경우로 강남구도 비슷한 양상을 보였다[16].

비교육특구의 인구 패턴을 분석하기위해 공공 Big data에서 추출한 데이터를 분석하여 실험한 결과 그림 7과 같이 중계동, 대치동, 성북동의 인구분포가 다르게 형성되고 있다. 교육특구인 중계동, 대치동은 비슷한 패턴이지만 비교육특구인 성북동을 비교한 경우 다른 인구분포가 나온 것을 알 수 있다.

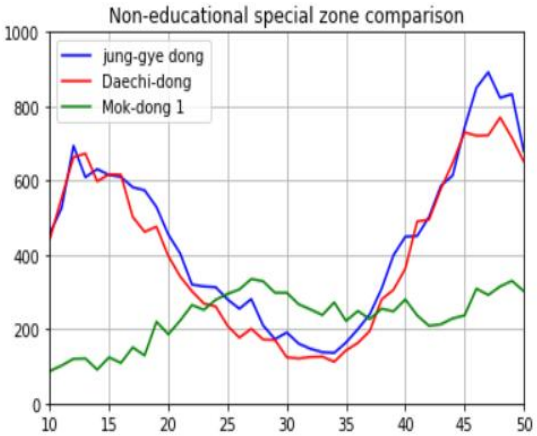

Fig. 7. Comparison of populations in non-educational area

2. RNN algorithm Analysis design

예측 지표로 사용한 데이터는 행정구역의 사설학원 수, 월평균교습시간, 월평균교습비, 행정구역 연령별 학생 수를 사용했고 상위권 진학률을 예측하기 위해 예측 모델은 2015년부터 2017년까지의 데이터를 사용하였다. 상위권 진학률을 예상하도록 학습하고, 학습된 모델을 예측하기 위해 Python 3.8, TensorFlow를 사용했다. 예측 모델은 TensorFlow 라이브러리를 기반으로 Deep Learning Algorithm인 RNN으로 구현하여 상위권 대학 진학률을 예측하였다. RNN은 입력 레이어와 출력 레이어, 1개의 은닉 레이어를 형성했다.

그림 8은 교육특구지역의 상위권 진학률을 Deep Learning 모델인 RNN을 이용하여 예측 실험을 수행한 결과로 데이터가 많지 않아서 정확도가 다소 낮지만 그래프의 패턴에 대한 학습이 이뤄진 것을 알 수 있다.

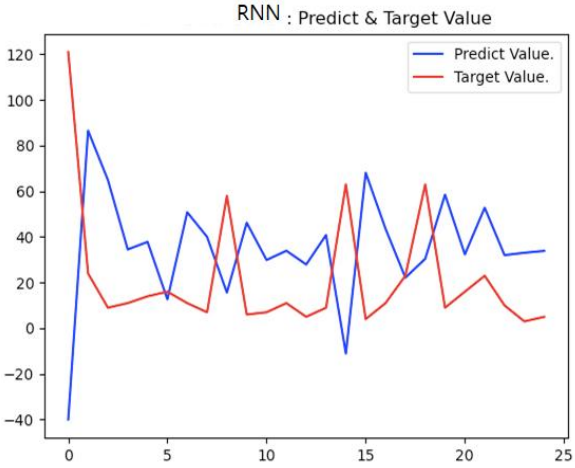

Fig. 8. Prediction of RNN model

V. Conclusions

교육열이 높은 지역에 거주하는 학생이 상위권 대학 진학률이 높다는 사회 통념을 가설로 세우고 예측된 가설과 정부의 공공데이터를 활용하여 분석된 자료를 토대로 검증하였다. 사설학원수가 많고, 교육열 높은 인구분포에 대한 Big data는 정부의 공공데이터 공개자료를 사용했다.

정부의 공공데이터를 사용하기 위해서 Big data의 분석을 위한 방법으로 Python 프로그래밍을 하였다. 또한 가설검증을 수행하기 위한 Deep Learning 모델인 RNN을 이용하여 예측 실험을 수행했다. 교육특구지역의 상위권 진학률을 Deep Learning 모델인 RNN을 이용하여 예측 실험을 수행한 결과 제시된 데이터가 많지 않아서 정확도의 신뢰성이 낮다는 문제점이 발견되었지만 그래프의 패턴으로 학습상황을 예측해볼 수 있다는 평가를 나타낼 수 있었다.

본 논문에서는 정확한 상위권 대학 진학률의 예측을 위해서 수반되어지는 기본요건으로 교육열이 높은 지역의 사설학원 현황, 연령별 학생 수의 예측, 예측한 상위권 대학 진학률과의 상관관계를 나타내는 기본적인 분석 자료의 정확도를 위해 학생이 거주하는 지역, 그 지역의 외부

지역과의 경제지표편차, 지역적 시설현황의 차이, 교육관심도에 반영된 경제적 관리능력, 거주 외 지역과의 교육비 차이 등이 있다. 예측한 데이터를 Deep Learning 모델인 RNN을 이용한 예측모델의 연구로 적용시켜서 도출하였고 절충해서 나온 결과의 신뢰도를 위해서 적용프로그램 구축의 필요성이 확인되었다.

가설을 예측해서 얻어진 결과 데이터를 중심으로 교육열이 높은 학생의 진학시스템, 교육자료에 충분조건의 활용으로 쓰일 수 있게 제시한 프로그래밍의 통찰적 쓰임의 요건을 위해서 본 논문에서 제시한 Deep Learning 모델인 RNN을 이용한 예측 모델의 연구가 직접적인 필요성을 제시한다는 결론에 도달하였다.

REFERENCES

[1] YWgugg, "Artificial Intelligence Technology and Industry Cases" ITFIND, No. 1888, pp. 3, 2019.

[2] Ministry of Science and Technology Information and Communi cation, "Korean New Deal Comprehensive Plan Announced", 2020

[3] Sijae, "Heat supply prediction system using Deep Learning", Seoul energy, Research Report 18-02, 2019.

[4] Christopher Olah, "Understanding LSTM Networks", *https://colah.github.io/posts/2015-08-Understanding-LSTMs*, 2015

[5] Chris Nicholson, "Data for Deep Learning", *https://pathmind.com/wiki/data-for-deep-learning*

[6] Ycjung, "Utilization Strategy of Big data for Official Statistics", Statistical Office, NO, 1-16, 2016.

[7] Ycjung, “Big data", Communication Books, pp. 6-14 2013.

[8] Sjkwon, "Transferred to a special education zone before entering elementary school", *http://www.veritas-a.com/news/articleView.html?idxno=148267*

[9] Sjkim, Kkryu, Sjson, "Parental Wealth, Children’s Ability and Entering Prestigious Colleges", Seoul National University Economic Research Institute, Economic Journal Vol. 54, No. 2, pp. 356~383, 2015

[10] Ministry of Education, “2019 elementary, middle and high school private education expenditure survey”, 2020

[11] Ypwoo, "So should I buy it now?" Korea Economic Daily, pp. 94-95, 2019.

[12] Jcpark, Jwjeong, Jwnoh, "A Study on the Choice of Mismatch between Housing in the Metropolitan Area", Collection of outstanding papers for graduate students' thesis contest, 2017

[13] Ministry of Public Administration and Security, “Population Status by Age of Resident Registration”, *http://27.101.213.4/index.jsp#*

[14] Seoul Open Data Plaza, "Population statistics by age in each autonomous district in Seoul", *https://data.seoul.go.kr/dataList/10837/S/2/datasetView.do*

[15] The Seoul Institute, "What is the status of private academies in Seoul?", *https://www.si.re.kr/node/60031*

[16] Seoul Open Data Plaza,, "Statistics of elementary school age population by autonomous district in Seoul", *https://data.seoul.go.kr/dataList/10830/S/2/datasetView.do;jsessionid=D5C8666004E3745C9BCD2EA9C63DB282.new_portal-svr-21, 2020*

Authors

Su-Youn Choi received the M.A. degree in Computer Science from Sungkyunkwan University Graduate School of Information and Communication, Korea, in 2009 and Among doctoral Dept. of Convergence Engineering from Hoseo Graduate School of Venture, Korea.

Su-Youn Choi is among doctoral Dept. of Convergence Engineering from Hoseo Graduate School of Venture, Seoul, Korea. She is interested in AI, Big data Analysis, Deep Learning, and Programming.

Dea-Woo Park received the P.D. degrees in Computer Science, from Soongsil University General Graduate School, Seoul, Korea. Dr. Park is currently a Professor in the Dept. of Convergence Engineering, Hoseo Graduate School of Venture, Seoul, Korea.

Dr. Park is interested in Hacking, CERT/CC, incident response, e-Discoverty, Forcentive, and cybersecurity, network security, smartphone security.