DNN-LSTM 기반 다중 입력 모델을 이용한 놀이공원 입장객 수 예측

이현주, 곽명섭, 강찬혁, 박현재, 최영준
아주대학교 소프트웨어학과
{dlguswn7175, kwakms123, rkdcksgur, estancia, choiyj}@ajou.ac.kr

Prediction of the Number of Visitors to Amusement Park Using the Multiple Inputs Model based on DNN-LSTM

Hyun-Ju Lee, Myung-Sub Kwak, Chan-Hyuk Kang, Hyun-Jae Park, Young-Jun Choi
Dept. of Software, Ajou University

요 약

국내 주요 놀이공원의 연간 총 방문객은 천만명을 훨씬 넘는다. 하지만 놀이공원의 일일 입장객이 매일 동일하게 몰리는 것은 아니기 때문에 입장객들에게 놀이공원을 방문하고자 하는 날짜의 예상 입장객 수를 알지 못하는 것은 큰 불안요소이다. 이러한 불안요소를 줄이기 위해 본 논문에서는 기상 요소 및 공휴일 등의 다양한 요인과 일별 놀이공원 입장객 수를 기반으로 놀이공원의 향후 일별 방문객수를 예측한다. 예측을 위해 입장객 수를 결정하는 다양한 요인과 시계열화 한 과거의 입장객수 변화 추이를 이용한다. 그리고 다중 입력 인공신경망을 통해 학습을 하여 모델을 생성하고 결과를 도출해 정확도를 검증해본다.

1. 서 론

놀이공원을 방문함에 있어서 가장 중요한 요소는 그 날의 방문객 수이다. 방문객이 몰리게 되면 각종 편의 시설 이용에 불편함을 겪으며 놀이기구 이용을 위한 대기시간이 길어지고 방문객들의 놀이공원 이용 만족도가 떨어지게 된다.[1] 매년 40%가 넘는 많은 사람들이 국내 주요 놀이공원을 방문하고 주로 동행하는 사람들은 가족 또는 연인이다. 소중한 시간을 할애하여 방문한 놀이공원의 이용 만족도는 높아야 한다.[2] 한편, 당일 방문객 수가 놀이공원의 이용 만족도를 낮추는 큰 요인으로 작용함에도 불구하고 아직까지 일별 놀이공원의 방문객 수를 예측하는 방법은 없다.

놀이공원 방문객들에게 당일 방문객 수를 알지 못하는 것은 만족도에 큰 영향을 끼칠 것이다.[1] 이러한 문제를 개선하기 위해 놀이공원의 일별 방문객 수를 예측하여 방문객들의 만족도를 높이고자 한다.

놀이공원에 방문할 때 방문객들의 주요 고려 요인은 날씨와 이용료이다.[2] 따라서 본 연구에서는 서울시의 기상요소와 지난 15 년간의 서울대공원 일별 입장객 수를 분석하여 미래의 특정 날짜의 입장객 수를 예측하는 모델을 제안한다. 모델은 인공신경망 모델을 기반으로 하고 기상요소 및 공휴일을 이용하여 예측하는 일반적인 DNN 모델과 과거의 입장객수 변화 추이를 시계열 데이터로 이용하여 예측하는 LSTM 모델을 합한 구조인 다중 입력 인공신경망 모델을 생성한다. 지도학습으로 모델을 만들고 놀이공원의 일별 방문객 수의 정확도를 검증한다.

본문의 구성은 다음과 같다. 2 절에서는 데이터의 수집과 전처리 과정을 담았다. 3 절에서는 선택한 데이터를 사용하여 본 논문에서 제안하는 예측 방법을 비교 분석하여 담았다. 4 절에서는 결론 및 향후 연구 방향을 담았다.

2. 관련 연구

최근 놀이공원에서의 서비스 품질과 만족도 관련 연구는 많이 이루어지고 있지만 놀이공원에서의 입장객 예측에 대한 연구는 많지 않다. 본 연구와 관련해 다양한 분야에서 선행되었던 수요 예측 기법을 정리했다.

관광, 문화 분야에서는 시계열 분석 기법을 이용해 월별 해외 관광객 수를 단기적으로 예측하기에 적합하다는 것을 밝혔고[3] TV 프로그램 시청률을 선형 모형에 시간적 요인을 추가해 예측하기도 했다.[4]

산업 분야에서는 전력 수요와 상수도 운영 예측을 위해 CNN-LSTM 결합 모델 기반 분석을 제안했다.[5][6]

3. 데 이 터

3.1. 데이터 수집

모델 학습을 하기 위한 데이터로 서울 열린 데이터 광장에서 제공하는 2004.7.1 ~ 2019.4.30 서울대공원 일일 입장객 정보 데이터를 수집했다. 그리고 기상 자료 개방 포털에서 제공하는 2004.7.1 ~ 2019.4.30 서울시 종관 기상 관측 데이터를 수집했다. 마지막으로 전국의 초, 중, 고등학교 홈페이지에서 제공하는 연도별 공지사항을 통해 136 개의 놀이공원 현장체험학습 날짜 정보를 직접 수집했다. 연도별 공휴일 날짜도 직접 수집했다.

3.2 데이터 전처리

서울대공원 일일 입장객 정보 데이터에서 전체 입장객 수에 대한 모델을 만들고자 하기 때문에 동식물원, 돌고래쇼, 테마가든, 자연학습교실, 캠프에서의 입장객 변수는 제거했다. 그리고 국적이 대한민국인 입장객에 한해 데이터를 선정하기 위해 외국인계 변수를 제거했고 마지막으로 한글 텍스트로 이루어진 날씨 변수를 제거하여 각 날짜에 해당하는 국내 입장객 데이터를 얻었다.

서울시 기상 관측 데이터에서는 풍속, 운량, 이슬점, 기압, 지중온도, 증발량 등의 변수 등은 모델 생성에 있어 불필요한 변수로 판단해 제거했고 최저, 최고, 평균 기온과 강수 계속 시간, 일 강수량 변수를 사용했다. 마지막으로 놀이공원의

입장객 수에 큰 영향을 주는 요인인 단체 방문객인 현장체험학습 데이터는 각 데이터의 요일별, 월별 현장체험학습 비율을 계산해 가중치를 주어 새로운 변수로 추가해 최종적으로 총 11 개의 feature 를 사용했다.

최종 데이터를 연구에 사용하기 전에 missing values 는 시계열 분석을 위해 제거하지 않고 이전 해의 데이터의 값으로 채웠다.

4. 설 계

4.1. 특성 선택

전처리한 데이터를 기반으로 다중 입력 인공신경망 모델의 입력 값으로 사용할 값들을 결정해야 한다. 먼저, feature 들을 기반으로 예측을 하는 DNN 모델에 입력 값으로 들어갈 데이터를 결정하기 위해 연관성이 높은 변수를 얻기 위한 과정을 진행했다. 날짜 변수에 대해서 월별, 요일별 의미를 부여하기 위해 one-hot-encoding 을 적용해 카테고리화를 했다. 년도와 일 변수는 특성을 반영하지 않는다고 판단해 제거하고 월과 요일 변수만 남겼다. 최종적으로 26 개의 feature 로 이루어진 데이터를 얻었다.

그리고 과거의 입장객수 변화 추이가 미래의 입장객수를 결정할 수도 있다고 판단해 입장객 수만 따로 추출하여 시계열 데이터셋으로 만들었다. 최근 한달 정도의 짧은 범위(14~41 일 전)의 입장객 수의 경향성을 담는 일별 데이터셋과 최근 8 개월 정도의 넓은 범위(2~32 주 전)의 입장객 수의 경향성을 담는 주별 평균 시계열 데이터셋을 만들었고 각각 28 개와 31 개의 시계열 feature 로 이루어진 데이터를 얻었다. 2 주 전부터의 데이터를 시계열 데이터화를 하는 이유는 실제로 놀이공원 방문 계획을 통상 2 주 정도 전에 계획하기 때문이다. 따라서 모델의 입력, 출력 변수들을 '표 1'과 같이 구성했다.

표 1. 입력변수와 출력변수

변수	요소	
Dense 입력 변수	시간 요소	월
		요일
		공휴일(0,1)
	기상 요소	평균 기온
		최저 기온
		최고 기온
		강수 계속 시간
		일 강수량
	기타 요소	현장체험학습
LSTM(day) 입력 변수	과거 입장객 수	14 ~ 42 일 전
LSTM(week) 입력 변수	과거 주별 입장객 수 평균	2 ~ 33 주 전
출력변수	일별 입장객 수	

4.2. 모델링

본 연구에서는 '그림 1'과 같이 모델을 구성하기 위해 데이터셋을 특성 데이터와 시계열 데이터로 나누어 입력 값으로 사용하고 예측하고자 하는 입장객 수를 하나의 출력 값으로 두어 지도학습으로 모델을 구성했다.

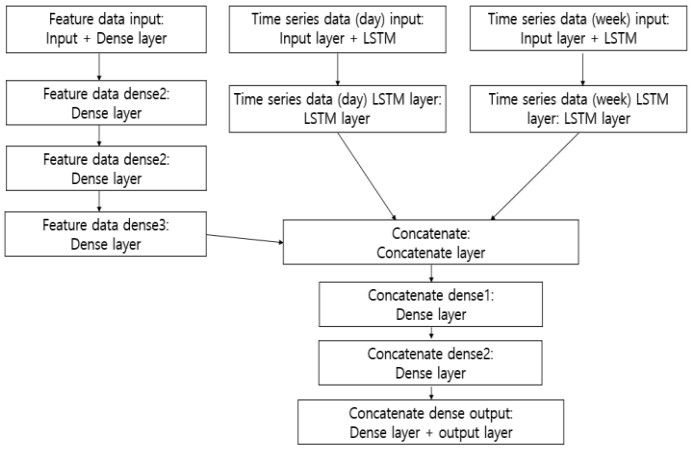

그림 1. 모델링 구조

다양한 종류의 데이터셋을 사용하기 때문에 Deep Neural Network, Long Short-Term Memory 두 알고리즘을 각각의 데이터셋에 적용시키기 위해 서브모델을 병합할 수 있도록 하는 다중 입력 모델을 사용했다. Concatenate 를 사용하여 모델을 병합했고 여러 조합의 모델들로 병합해 구성하고 비교해 가장 낮은 오차과 가장 낮은 val_loss 를 보이는 모델을 선택했다. 각 모델의 loss function 과 optimizer 는 regression 에 적합한 Mean-Square-Error 와 Adam 으로 동일하게 사용했고 Activation function 은 relu 로 동일하게 사용했다. 각각의 모델들의 레이어와 노드의 개수는 만족스러운 값을 얻을 때까지 반복 수행해보면서 '표 2'와 같이 적절한 값을 얻었다.

표 2. 각 모델의 구성 요소 및 하이퍼파라미터

모델	구분	설정
DNN(Dense)	입력층 노드 수	26
	은닉층 개수	3
	은닉층 노드 수	500,250,100
	출력층 노드 수	50
LSTM(day)	입력층 노드 수	28
	은닉층 개수	2
	은닉층 노드 수	28, 28
	출력층 노드 수	28
LSTM(week)	입력층 노드 수	31
	은닉층 개수	2
	은닉층 노드 수	31,31
	출력층 노드 수	31
Concatenate 이후 Dense	입력층 노드 수	109(50+28+31)
	은닉층 개수	2
	은닉층 노드 수	100, 10
	출력층 노드 수	1
미니 배치 수		50
Epochs		70

3 개의 단일 모델과 4 개의 다중 입력 모델에 대한 성능 비교를 위해 RMSE, MAE, val_loss 를 구해 '표 3', '표 4'로 정리했고 각각 6 회 이상 학습하여 val_loss 가 가장 작은 모델을 저장하여 나온 오차의 평균과 최저 오차를 구했다. training set 은 2005.2.10 ~ 2018.4.11, 4809 개의 데이터로 구성하고 test set 은 2018.4.10 ~ 2019.4.29, 385 개의 데이터로

구성했다. 평가표에 나온 오차값은 최대 입장객수인 115002 명, 최소 입장객수인 2 명으로 스케일링한 후 구한 값이다.

표 3. 모델 평가표(평균값)

	RMSE	MAE	val_loss
DNN	0.05339	0.03759	0.002351
LSTM(day)	0.05930	0.03930	0.003787
LSTM(week)	0.07631	0.05840	0.004668
LSTM (day+week)	0.06718	0.04577	0.003910
DNN + LSTM(day)	0.04955	0.03396	0.002197
DNN + LSTM(week)	0.04626	0.03112	0.002111
DNN + LSTM (day+week)	0.04349	0.02899	0.001981

표 4. 모델 평가표(최저값)

	RMSE	MAE	val_loss
DNN	0.04355	0.02633	0.00193
LSTM(day)	0.05744	0.03666	0.00364
LSTM(week)	0.07098	0.05426	0.00457
LSTM (day+week)	0.06190	0.04090	0.00356
DNN + LSTM(day)	0.04296	0.03046	0.00200
DNN + LSTM(week)	0.04236	0.02745	0.00184
DNN + LSTM (day+week)	0.03907	0.02474	0.00180

좋은 모델일수록 실제 입장객 수와 예측한 입장객 수의 오차가 작다. 생성한 7 개의 모델 중에서 단일 모델 3 개를 모두 결합한 다중 입력 모델이 평균값과 최저값에서 RMSE 와 MAE 가 모두 가장 작았고 가장 좋은 결과를 나타냈다.

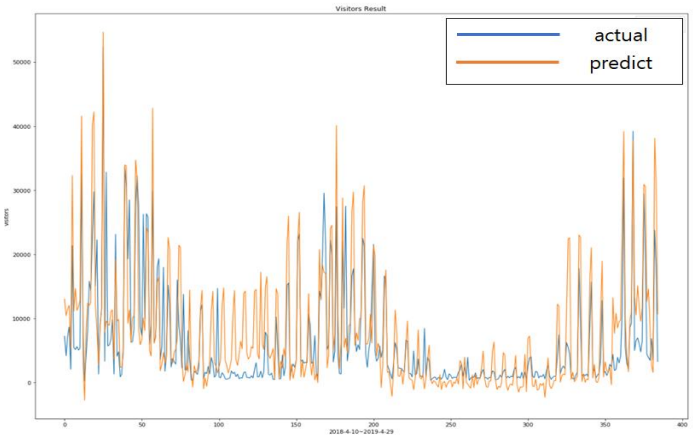

그림 2. 날짜에 따른 실제 입장객 수

'그림 2'는 가장 오차가 작은 마지막 다중 입력 모델에 대해서 test set 으로 예측한 결과값과 실제값이다. 경향성이 잘 예측되는 것을 확인했다.

5. 결 론

본 연구에서는 놀이공원의 일별 입장객 수를 예측하기 위해 Neural Network 와 다중 입력 모델을 이용하여 여러 모델을 생성하고 정확도를 비교 분석했다. 분석 결과, 7 개의 모델 중에서 DNN 모델과 일별, 주별 시계열 데이터셋을 학습한 2 개의 LSTM 모델을 모두 결합한 다중 입력 모델의 오차가 가장 작았다.

RMSE 와 MAE 의 값을 역스케일링을 통해 실제 입장객수를 기준으로 오차를 다시 구한 결과 3 개의 단일 모델을 결합한 마지막 다중 입력 모델의 MAE 는 평균적으로 3334 명이고 가장 낮을 때는 2845 명이었다. 실제 평균 입장객수인 6136 명에 비하면 오차가 굉장히 크다고 볼 수도 있다. 하지만 주어진 데이터 자체가 train set 에서 최대값이 115002 명, 최소값이 2 명, test set 에서 최대값이 52310 명, 최소값이 288 명으로 편차가 굉장히 큰 데이터인 것을 감안해야 한다. '그림 2'의 그래프처럼 오차는 크지만 사람이 몰리는 경향성 자체는 높은 확률로 예측을 할 수 있었다.

일별 입장객 수의 예측 방법이 본 연구에서는 여러 특성과 시계열의 영향을 모두 반영하는 다중 입력 모델이 가장 정확한 예측방법으로 분석되었다. 하지만 예측정도가 아주 정확한 예측이 아니었고 본다. 15 년간의 입장객 추이를 볼 때 최근 2, 3 년간의 입장객 수가 갑자기 급격하게 줄어들었는데 그 영향에 대한 학습이 제대로 이루어지지 않은 것의 큰 원인이라고 본다. 실제로 train set 의 입장객 수 평균값인 9484 명에 비해 test set 은 6136 명으로 급격히 떨어졌다. 정확한 예측을 위해 오차를 최소로 할 수 있는 연구가 이어져야 할 것이다.

본 연구의 한계점으로는 현장체험학습에 대한 집계방법이 데이터와 전문성의 부족으로 정확하지 않았다. 단체입장객이 놀이공원의 당일 혼잡도를 크게 좌우하는 요인이 되기에 이를 포함하는 신뢰할 만한 다양한 데이터가 있었다면 오차를 더욱 줄일 수 있을 것이다.

향후 연구과제에서는 날씨, 공휴일, 체험학습 외의 어떤 요인이 상관관계가 높은지 구하고 놀이공원 방문을 결정하는 많은 요인들을 적용하여 더 정확하게 예측하는 모델을 제시한다.

5. Acknowledgements

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2015-0-00908)

6. 참고문헌

[1] 김민주, 김세용. "테마파크 방문자의 혼잡 지각과 방문자 만족 간의 인과관계". 여가관광연구, vol. 16, pp.1-24, 2010.

[2] 엠브레인 트렌트모니터. "놀이공원 이용 및 전반적 U&A 관련 조사". pp.1-44, 2018.

[3] 최영문, 김사헌. "단변량 시계열 관광수요 예측모형의 적정성 비교평가". 관광학연구, 제 21 권, 제 2 호, 1998.

[4] Gensch, D. H. & Shaman, P. "Predicating TV rating. Journal of advertising Research". August, pp.85~92, 1980.

[5] 김도현, 김명수, 노재형, 박종배, 채명석. "딥러닝 기반 LSTM-CNN 을 활용한 단기 수요 예측에 관한 연구". 대한전기학회 학술대회 논문집, (), 281-282, 2018.

[6] Kerang Cao, Hangyung Kim, Chulhyun Hwang, Hoekyung Jung. "CNN-LSTM Coupled Model for Prediction of Waterworks Operation Data". , 14(6), 1508-1520, 2018.