기후요소에 따르는 관광객 관심정보 예측 모델[1)]

박세린*, 이연지**, 이정민***, 이소희****, 이정훈****
*제주대학교 사대부고, **신성여자고등학교, ***대정고등학교,
****제주대학교 컴퓨터교육과, *****제주대학교 전산통계학과
{serin4103,2yeon_g}@naver.com, 01222020118@onedu.jje.go.kr,
{lsh5044,jhlee}@jejunu.ac.kr

Prediction model of tourists' interest according to the climate condition

Serin park*, Younji Lee**, Jungmin Lee***, Sohee Lee****, Junghoon Lee****
*JNU Highschool, **Shinsung Girls' Highschool, ***Daejung Highschool,
****Jeju National University

요 약

관광관련 광고, 상품판매 촉진, 추천 등을 위해 제주도 관광객의 관심 정보에 있어 기후요소가 끼치는 영향을 분석하고 이를 토대로 예측모델을 개발한다. 예측모델은 입력으로 기온, 강수량, 풍속, 습도, 일사량 및 전운량, 출력으로 가장 관심도가 높은 관광지 유형을 가지며 TMAP의 검색순위 이력 데이터와 기상청의 기후이력 데이터를 다운로드하여 학습패턴을 생성한다. 예측모델은 Sklearn 인공신경망 라이브러리를 이용하여 구현하였으며, 81.8 %의 정확도를 보인다.

1. 서론

본 논문에서는 기후에 따르는 관광객들의 관심사를 예측모델로 구축하여, 이를 바탕으로 기상 예보에 따라 관광객이 많이 모일 장소나 유형을 미리 예상하여 광고나 상품판매 등에 활용하도록 한다. 최근 공개 데이터의 활성화로 다양한 데이터셋이 모델 구축에 이용될 수 있는데, 학습 데이터를 생성하기 위해 SKT 데이터 허브 TMAP에서 공개한 제주도내 검색순위 키워드와[1] 기상청에서 공개한 기후정보를[2] 다운로드 받는다.

관련 데이터들은 CSV 파일의 형태를 가지며 파이썬에서 제공하는 다양한 데이터 처리 기능을 통해 이 파일들을 읽어 들여 원하는 형태로 변경하거나 필요한 정보들을 추출할 수 있다. 또 파이썬과 연계된 sklearn 라이브러리는 인공신경망, 의사결정나무, kNN 등의 다양한 예측 모델을 생성할 수 있도록 하기 때문에 이 라이브러리 함수들이 요구하는 데이터 포맷을 생성하면 자동으로 모델이 생성된다[3].

2. 예측 모델의 구축

먼저 관광객들의 관심사를 파악하는데 있어서 SK Telecom에서 제공하는 TMAP 방문 빅데이터를 이용하는데, 지역마다 매일 상위 검색순위 30개의 레코드가 생성되며 하나의 레코드에는 다음과 같이 타임스탬프, 지역(시도), 지역(시군구), 검색지명, 검색지유형1 (7종), 검색지유형2 (22종), 검색지유형3 (44종), 검색지랭킹 등 8개의 필드가 속한다.

우선적으로 전국 데이터 중 제주시와 서귀포시에 해당하는 데이터를 추출하였고 관광객 관심 정보는 검색지명, 검색지 유형과 검색지 랭킹에 나타난다. 검색지 유형 1, 2, 3은 각각 대분류, 중분류, 소분류에 해당하며 대분류는 공공편의, 교통편의 등의 목록을, 중분류는 관광명소, 유통점, 숙박, 시장 등의 수준의 목록을, 또 소분류는 공원, 공항, 음식점 수준의 목록을 포함하고 있다. 검색지 랭킹은 일별로 30위까지 제공되는데 순위만 제공될 뿐 검색 수까지는 제공되지 않아 절대적인 평가척도는 되지 않는다.

예측모델 개발의 입력을 위해 선별된 기후요소는 평균기온, 일강수량, 평균풍속, 상대습도, 일사량, 전운량 등이며 기상청 사이트에서는 제주지역 4개의 기상 포스트에서 관측된 데이터를 공개하고 있다. 이 과정에서 각 기후요소와 관심관광지와의 상관관계 분석을 수행하였으나 두드러진 선형적인 상관관계는 나타나지 않아 어떤 형태든 영향력이 있을 것으로 보이는 요소들을 선택하였다.

기계학습을 위한 y=F(x)의 입력 x와 출력 y로 구성

본 연구는 R-WeSET사업의 지원을 받아 수행되었음.

된 레코드가 생성되고 결과적으로 <그림 1>에서 보는 바와 같은 인공신경망 모델을 구축하였다[3]. 은닉 계층의 노드 수는 10에서 15개로 변화시켰는데 거의 성능이 유사하다.

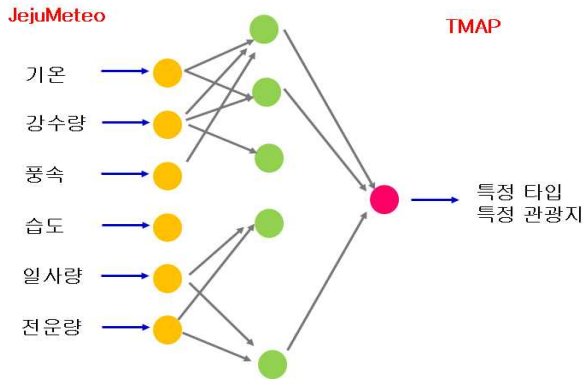


그림 1. 인공신경망 모델

출력으로 시도해본 관광객 관심사는 동문시장과 같은 특정 관광지가 5위 안에 드는지, 가장 많이 찾을 소분류 유형 등이다. 제주시의 경우는 제주공항이 거의 매일 1순위이므로 제외한다. 또 서귀포 지역 1순위, 최대관심 지역 유형 등을 고려할 수 있다. sklearn에서는 하나의 출력 노드만을 허용하기 때문에 예측하고자 하는 값이 달라지면 새로이 학습 레코드를 생성하여야 한다.

학습 패턴을 만드는데 있어서 <그림 2>에서 보는 바와 같이 각 인자마다 갖는 값의 범위가 온도는 최대 40, 전운량은 0-10 등 스케일이 달라서 기계학습에 주는 영향이 다르다는 점을 고려하여야 한다. 따라서 <그림 3>처럼 평균과 표준편차에 의해 값들을 평균 0을 중심으로 정규화하면 예측을 시도해봤던 모델마다 10~20% 정확도가 높아졌다.

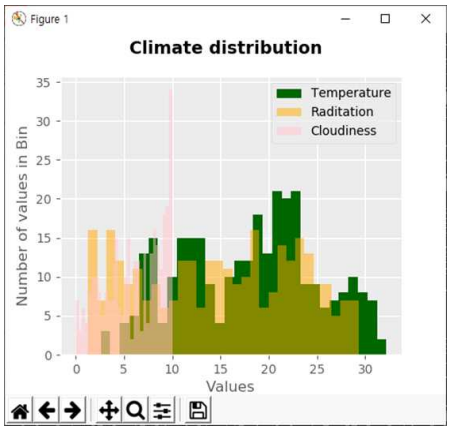


그림 2 정규화 이전의 데이터 값의 분포

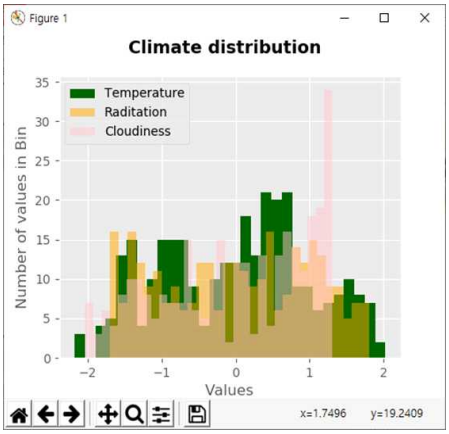
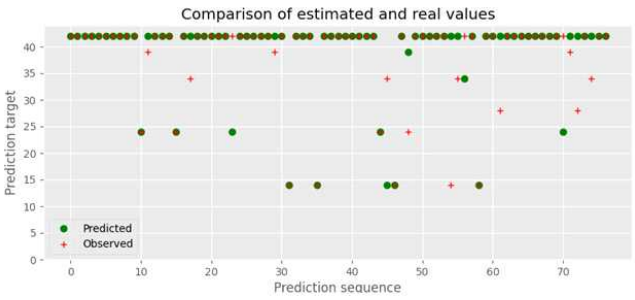


그림 3 정규화 이후 데이터 분포

제주시에서 제주공항외 관심유형을 예측해본 결과는 <그림 4>와 같다. 이때 준비된 레코드 셋을 sklearn 라이브러리는 자동으로 학습용과 테스트용으로 나누어주며 그 성능을 score로 나타내준다. <그림 4>에서 ●표시는 수집한 데이터를 토대로 만든 예측 모델을 통해 예측한 값이고, + 표시는 실제 해당일에 높은 순위를 차지하고 있는 데이터의 번호이다. 두 표시가 일치하면 예측이 맞는 것이며 결국 81.8 %의 예측 정확성을 보이고 있다. 번호는 각 유형마다 부여한 것으로 해수욕장이 가장 많이 검색된다.



3. 결론

공개 데이터와 인공지능 라이브러리를 이용하여 기후요소가 제주지역 관광객들의 관심사에 미치는 영향을 분석하고 이를 기반으로 예측모델을 구축한 결과 81.8 %의 정확도를 얻을 수 있었고 이 예상을 기반으로 관광 상품이나 액티비티의 품질을 개선할 수 있다. 추후 기후요소뿐만 아니라 교통상황, 이벤트 진행 상황 등을 모델에 포함시킬 수 있다.

참고문헌

[1] http://www.bigdatahub.co.kr/index.do
[2] https://data.kma.go.kr/cmmn/main.do
[3] https://scikit-learn.org/stable/