사람 성격 요소에 따른 위치 방문 선호도 예측의 자동화 시스템

송하윤 ,정지현
hayoon@hongik.ac.kr, rachel618@g.hongik.ac.kr

The Automated System for Location Visiting Preference Prediction with Personality Factors

Ha Yoon Song , Ji Hyun Jung
Dept. of Computer Engineering, Hongik University

요 약

데이터 베이스에 저장된 사용자의 위치, 성격정보를 자동으로 받아서 머신러닝으로 회귀분석하여 방문 장소에 대한 선호도를 예측한다. 사람의 성격 요소로는 BFF 와 다른 기본 요소들을 사용하였다. 이를 위하여 자동화된 시스템을 구성하였고 위치 방문 선호도를 예측하기 위한 머신러닝 기법으로는 앙상블기법을 사용하였다. 예측 결과는 장소 카테고리별로 방문 선호도가 나타나고 이를 사용자 별로 나누어 저장할 예정이다. 데이터의 양이 많아지면서 나타나는 문제들을 해결하여 향후 연구에 도움이 될 것이다.

1. 서론

기존 연구에서는 사용자의 위치정보를 머신러닝으로 회귀분석하여 특정 장소 카테고리에 대한 방문 비율을 예측하였다. 세 가지 앙상블 기법인 랜덤 포레스트, XGBoost, Stacking 으로 분석하였다. 또한 특징 선택과 하이퍼파라미터 최적화 과정을 통해 입력 데이터의 차원은 줄이고 예측 정확도는 높였다.

이번 연구에서는 사용자의 데이터가 추가될 때마다 반복적으로 실행되는 일련의 과정들을 자동화 하고자 한다. 먼저, 수치화 된 성격 검사 결과, gpx 데이터를 기반으로 장소 카테고리별로 분류한 데이터를 병합하여 하나의 데이터로 만들어 데이터베이스에 업로드 한다. 그 데이터를 받아 위치 방문 선호도를 예측하고 결과를 연구용 데이터베이스에 업로드하는 과정까지 연결하고자 한다. 데이터의 양이 증가하였으므로 파라미터를 조정하여 정확도를 높이는 과정도 필요할 것이다.

2. 관련 연구

2.1 앙상블기법

앙상블 학습은 여러 개의 결정 트리(Decision Tree)를 결합하여 하나의 결정트리 보다 더 좋은 성능을 내는 머신러닝 기법으로, Decision Tree 알고리즘의 과적합(overfitting) 문제를 보완한 기법이다.

앙상블 알고리즘은 학습 방식에 따라 배깅(Bagging), 부스팅(Boosting), 스태킹(Stacking)으로 나눌 수 있다. 배깅은 Bootstrap Aggregating 의 약자로, 샘플을 여러 번 뽑아(Bootstrap) 각 모델을 학습시켜 그 결과를 결합(Aggregation)하는 방법이다. 본 연구에서는 배깅기법에서 랜덤포레스트(Random Forest)를 사용하였다. 부스팅은 반복적으로 모델을 업데이트 해 나가는 방법이다. 이전 반복(iteration)의 결과에 따라 데이터셋 샘플에 대한 가중치를 부여하여, 반복할 때 마다 각 샘플의 중요도에 따라 다른 분류기가 만들어지게 된다. 분류하기 힘든 관측 값에 효과적이고 데이터의 분포를 고르게 하는 효과가 있으며 대표적인 예로는 XGBoost 가 있다[1].

2.2 랜덤포레스트

랜덤 포레스트는 전통적인 결정트리 기법을 하나가 아닌 여러 개의 트리로 확장 시킨 결정트리의 메타학습(meta-learning) 형태를 갖고 있는 기계학습 기법이다. 트리 생성 시 각 노드에서는 학습데이터의 모든 특징 값을 고려하지 않고 임의로 선택된 차원의 특징 값만 고려하는 랜덤 결정트리를 생성한다. 이 경우 각 개별 랜덤 결정트리의 정확도는 떨어질 수 있으나 랜덤포레스트는 이를 종합하여 예측을 수행하므로 과적합은 줄이면서 예측 성능은 향상시킬 수 있다[2]. 랜덤포레스트의 성능에 영향을 미치는 매개변수로는 랜덤결정트리의 깊이와 수, 트리의 노드 최적화 시 선택하는 차원의 수 등이 있다. 본 연구에서는 하이퍼파라미터 최적화를 통해 랜덤포레스트의 정확도를 높였다.

2.3 XGBoost

XGBoost 는 Extreme Gradient Boosting 의 약자이다. Boosting 기법을 이용하여 구현한 알고리즘은 Gradient Boost 가 대표적인데 이 알고리즘을 병렬학습이 지원되도록 구현한 라이브러리가 XGBoost 이다.

랜덤포레스트와 달리 앙상블 기법으로 부스팅을 사용하여 약한 학습기를 순차적으로 생성하여 강한 학습기를 만드는 것이다. 이전 분류기의 에러가 작아지는 방향으로 다음 분류기에 가중치(weight)를 부여하면서 학습과 예측을 진행한다. XGBoost 는 다수의 하이퍼파라미터가 존재하며 일반 파라미터, 부스터 파라미터, 학습과정 파라미터 세가지로 나뉜다. 일반 파라미터에서 부스터를 결정하는 파라미터가 성능에 큰 영향을 준다. XGBoost 에서 사용하는 부스터에는 gbtree, gblinear, dart 가 있다. gbtree 는 약한 학습기로 회귀트리를 사용하는 것이고, gblinear 는 약한 학습리고 선형 회귀모델을 사용하는 것이다. dart 는 신경망 학습에 사용되는 드롭아웃을 회귀트리에 적용한 것이다.

2.4 Stacking

스태킹은 서로 다른 종류의 단일 모델들을 독립적으로 학습시켜 그 예측 값들을 기반으로 예측하는 기법이다. 스태킹의 경우 일반 알고리즘과 다르게 2 단계로 학습을 진행한다.1 단계에서는 다양한 모델의 예측값을 뽑아낸다. 2 단계에서는 1 단계에서 도출한 예측값들을 다시 훈련 데이터으로 적용하여 meta 학습기 또는 블렌더 라고 불리는 최종 모델에서 최종 예측 값을 만들어 낸다. 이 때, 과적합을 방지하기 위해 학습용 데이터를 생성할 때는 k-fold 교차검증을 사용한다.

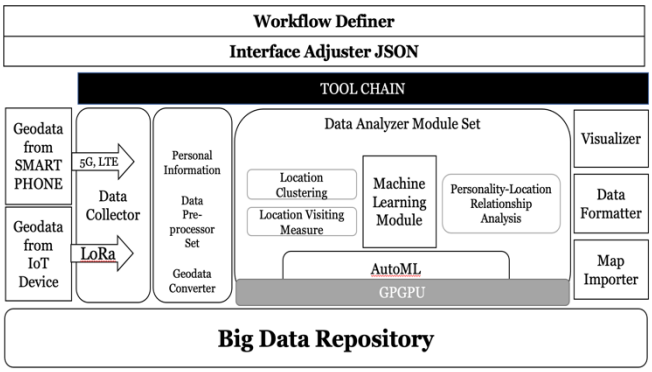
2.5 Database 디자인

마리아 데이터베이스(MariaDB)는 오픈소스의 관계형 데이터베이스 관리 시스템(RDBMS)이다. MySQL 과 동일한 소스코드를 기반으로 하기 때문에 MySQL 의 기본 아키텍처와 동일하다. pluggable Storage Engine 의 종류만 조금 다를 뿐 명령어나 사용방법이 같아 MySQL 과 호환성이 뛰어나다. 또한 마리아 데이터베이스에는 새로운 저장엔진인 아리아(aria)뿐만 아니라 InnoDB 를 교체할 수 있는 XtraDB 저장엔진을 포함하고 있다. 이러한 장점으로 본 연구에서는 마리아데이터베이스를 채택하였다.

2.6 BFF(Big Five Factor)

BFF 는 P.T Costa 와 R.R McCrae 에 의해 제안된 성격모델로, 현대 심리 학계에서 가장 널리 인정받고 있는 성격이론이다[3][4]. 다섯 가지 요인으로는 경험에 대한 개방성(Openness to experience), 성실성(Con scientiosness), 외향성(Extraversion), 친화성(Agre eableness), 신경성(Neuroticism)이다. 각 특성에 대한 점수는 사용자의 답변으로 얻어지며 1 에서 5 점사이의 값으로 표현된다. BFF 는 추상적인 사람의 성격을 수치화할 수 있어 많은 연구에서 사용되고 있다[5][6][7][8][9][10].

3. 연구과정



(그림 1. 워크플로우)

3.1 연구 계획

스마트폰과 IoT Device 를 통해 먼저 사용자의 위치 정보를 수집한다. 처음에 raw data 형태로 수집된 위치 데이터를 gpx data 로 변환하고 연구할 수 있는 형태로 바꿔준다. 위도, 경도 기준으로 나와 있던 위치 정보를 클러스터링하여 방문 장소를 알아내고 장소 방문도를 계산한다. 성격 검사 결과는 코드에서는 csv 파일 형태로 입력되지만

데이터베이스에 저장할 때는 json 파일로 변환하여 저장한다. 머신러닝에 사용할 수 있는 형태로 데이터를 모두 변환하고 나서 모델에 입력 데이터로 넣어 위치 방문 선호도를 예측한다. 결과값은 다시 데이터베이스에 업로드한다.

3.2 데이터 수집

그림 1 에서와 같이 먼저 사용자의 스마트폰을 이용하여 위치 데이터를 수집한다. 스마트폰 어플리케이션 SWARM, Sports Tracker 를 사용하여 수집되었다.

성격 검사결과는 BFF 값으로 수치화 되어 나타난다.

	O	C	E	A	N
사용자 1	3.4	3.222222222	3.375	3.333333333	3.125
사용자 2	3.1	3.555555556	3.5	2.777777778	2.5
사용자 3	2.7	3.22222	3.25	2.66667	2.75
사용자 4	2	2.888888889	3.875	2.777777778	2.875
사용자 5	3	3.333333333	3.25	2.555555556	2.875
사용자 6	3.8	3.555555556	2.75	3.111111111	3.375
사용자 7	4	3.66667	4	3.88889	2.75

(표 1. BFF 데이터 예시)

3.3 데이터 전처리 과정

위치 데이터는 raw data 에서 gpx data 로 변환하여 사용하고 방문한 장소를 장소 카테고리 별로 분류하여 방문 횟수를 비율로 환산한다.

기존 연구에서 만든 설문조사결과, BFF 값과 위의 위치데이터를 모두 병합하여 하나의 파일로 만들어 다시 데이터베이스에 업로드 할 예정이다. 새로운 사용자의 BFF 값과 설문조사값은 모두 데이터 전처리가 완료되었지만 위치데이터를 전처리하는 작업이 필요하다.

다른 연구에서도 활용할 수 있도록 위치 방문 선호도 예측 결과도 데이터베이스에 다시 업로드 하는 과정까지 진행할 것이다.

4.기대효과

기존에 있던 정보에 더해 새로운 사용자와 기존 사용자의 위치정보가 수정되는 것이 더 용이해져 최신의 데이터를 유지할 수 있을 것이다.

또한 개별적으로 진행중인 연구들을 데이터베이스를 통해 하나로 통합하고자 한다. 위치데이터를 클러스터링 한 결과, 장소 방문도를 측정하여 수치화한 데이터 등을 데이터베이스에 업로드하여 관리한다면 본 연구에서도 사용자가 추가될 때 더 편리하게 위치 방문 선호도를 예측할 수 있을 것이다. 위와 같이 본 연구의 입력 데이터가 다른 분야의 결과 데이터가 되고 그 반대가 되기도 하므로 데이터의 수정이 반영되면 더 수월하게 연구가 진행될 것으로 예상된다.

이 연구는 정부(교육과학기술부)의 재원으로
한국연구재단의 지원을 받아 수행됨
(NRF-2019R1F1A1056123)

참고문헌

[1] Biau, Gérard, and Erwan Scornet. "A random forest guidedtour." Test 25.2 (2016): 197-227.

[2] Breiman, Leo. "Random forests." Machine learning 45.1 (2001):5-32.

[3] Costa Jr, Paul T., and Robert R. McCrae. "Four ways five factors are basic." Personality and individual differences 13.6 (1992): 653-665.

[4] Kim, Young Myung, and Ha Yoon Song. "Analysis of Relationship between Personal Factors and Visiting Places using Random Forest Technique." 2019 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE, 2019.

[5] J. Hoseinifar, M. M. Siedkalan, S. R. Zirak, M. Nowrozi, A. Shaker, E. Meamar, and E. Ghaderi, “An investigation of the relation between creativity and five factors of personality in students,” Procedia - Social and Behavioral Sciences, vol. 30, pp. 2037-2041,

[6] D. Jani, J.-H. Jang, and Y.-H. Hwang, “Big

five factors of personality and tourists' internet search behavior," Asia Pacific Journal of Tourism Research, vol. 19, no. 5, pp. 600-615, 2014.

[7] D. Jani and H. Han, "Personality, social comparison, consumption emotions, satisfaction, and behavioral intentions," International Journal of Contemporary Hospitality Management, vol. 25, no. 7, pp. 970-993, sep 2013.

[8] O. P. John, S. Srivastava et al., "The big five trait taxonomy: History, measurement, and theoretical perspectives," Handbook of personality: Theory and research, vol. 2, no. 1999, pp. 102-138, 1999.

[9] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," Computers in Human Behavior, vol. 26, no. 6, pp. 1289-1295, nov 2010.

[10] M. J. Chorley, R. M. Whitaker, and S. M. Allen, "Personality and location-based social networks," Computers in Human Behavior, vol. 46, pp. 45-56, 2015.