랜덤 포레스트를 사용하여 다양한 요인과 방문하는 장소 사이의 관계 분석

김영명*, 송하윤*
*홍익대학교 컴퓨터공학과
e-mail:hayoon@hongik.ac.kr
dudaud0205@gmail.com

Analysis of the relationship between Various Factors and Visiting Places using Random Forest Technique

Young-Myoung Kim*, Ha Yoon Song*
*Department of Computer Engineering, Hongik University

요 약

기존에는 Big Five Factor (BFF)를 이용하여 사람의 성격과 방문하는 장소 간의 관계를 분석하는 연구들이 많이 진행되었다. 본 연구에서는 성격뿐 아니라 sns 사용, 취미, 성별, 나이, 종교 등 다양한 요인을 추가하여 방문하는 장소에 영향을 미치는 요인을 찾고자 한다. 성격 데이터는 BFF 설문지로, 그 외 요인들은 본 연구팀이 직접 만든 설문지로 수집하였다. 방문하는 장소는 스마트폰 애플리케이션 SWARM을 이용하여 수집한 뒤 카테고리별로 분류하여 사용하였다. 총 17명의 참가자들이 약 3달간 모은 데이터를 사용하였다. 분석에는 앙상블 기법인 랜덤 포레스트를 사용하였다.

1. 서론

기존 연구들에서 사람의 성격과 방문하는 장소 사이에 일부 상관관계가 있다는 것은 이미 입증이 되었다. 그러나 자주 방문하는 장소에는 성격뿐 아니라 다른 요인도 영향을 미칠 것이라고 예상하였다. 이러한 생각을 입증하기 위하여 성격과 장소 외에 영향을 미칠 수 있다고 예상되는 다른 요인들을 설문으로 수집하였다. 수집한 항목으로는 성별, 나이, 결혼, 종교, 수입, 교통수단, sns 사용 여부 등이 있다. 성격 데이터는 Big Five Inventory(BFI)의 설문지를 이용하여 Big Five Factor(BFF)로 구성된 데이터를 수집하였다. 총 17명의 참가자들이 설문과 장소 데이터 수집에 참여하였다. 장소 데이터는 SWARM이라는 스마트폰 애플리케이션을 통해 약 3개월 간 수집하였다. 데이터 분석에는 앙상블 학습 방법인 랜덤 포레스트를 사용하였다. 랜덤 포레스트는 일반화 성능이 좋고 정확도가 높다. 많은 입력 변수들을 다룰 수 있고 노이즈에 강하다. 또한 입력 변수가 어느 정도 영향을 미치는지 importance 값을 알 수 있어 이번 실험에 사용하였다.

2장에서는 수집한 성격을 포함하는 다양한 데이터값들과 장소 데이터를 보여주고, 실험에 적용한 장소 카테고리 분류에 대하여 설명할 것이다. 3장에서는 랜덤 포레스트로 회귀 분석한 결과를 통해 방문하는 장소에 영향을 미치는 요소에 대해 말할 것이다. 4장에서는 결론 및 앞으로의 연구 방향을 기술할 것이다.

2. 입력 데이터와 레이블

2.1 입력 데이터

성격을 나타내는 척도로 많은 연구에서 McCrae와 Costa가 주장한 BFF를 사용하고 있다. 다섯 가지 성격 인자는 개방성(O: Openness), 성실성(C: Conscientiousness), 열정성(E: Extraversion), 동조성(A: Agreeableness), 신경성(N: Neuroticism)이다. BFF는 각 인자에 해당하는 값이 수치화되어 쉽게 학습에 적용할 수 있다. 아래의 <표 1>은 BFF로 표현된 실험 참가자 17명의 성격 데이터이다. 이를 통해 성격을 유추할 수 있다. 먼저 각 요인의 특징을 살펴보면, 개방성이 높은 사람은 창의적이고 감성적이며 예술에 대한 관심이 많다. 성실성이 높은 사람은 책임감이 있고 성취욕이 있으며 절제력이 있다. 외향성이 높은 사람은 따뜻하고 사교적이며 자신감이 있고 긍정적이다. 반면에 신경성이 높은 사람은 스트레스에 민감하고 충동적이며 적대적이고 우울한 기질이 있다. 이를 바탕으로 실험자들의 데이터를 살펴보면, 실험자 4는 창의적이고 감성적이면서도 책임감이 있고 절제력이 있다. 신경성이 낮은 것으로 보아 충동적이지 않고 스트레스에 강하다. <표 1>의 성격 데이터는 설문으로 수집한 다른 요인들과 함께 입력 데이터로 사용 된다.

<표 1 > 성격 데이터

	O	C	E	A	N
	개방성	성실성	외향성	우호성	신경성
실험자1	3.3	3.9	3.3	3.7	2.6
실험자2	2.7	3.2	3.2	2.7	2.8
실험자3	4.3	3.1	2.3	3.2	2.9
실험자4	4.2	4.3	3.5	3.6	2.6
실험자5	4	3.7	4	3.9	2.8
실험자6	3.8	4	3.1	3.8	2.3
실험자7	3.2	3.2	3.5	3.3	3.5
실험자8	2.8	3.8	3.8	3.3	2.3
실험자9	3.4	3.6	3.5	3.6	3.1
실험자10	3	3.6	2.5	3	3
실험자11	4.1	3.8	3.8	2.8	3
실험자12	3.1	3	2.8	3	2.8
실험자13	3.3	3.2	3.5	2.6	2.6
실험자14	3.7	3.3	3.6	3.8	3.5
실험자15	2.4	3.7	3	2.8	2.6
실험자16	3.4	3.2	3	3	2.6
실험자17	3.9	3.3	3.5	2.9	2.8

<표 2> 실험자 1의 설문 응답

요인	응답
나이(Age)	2
직업(Job)	1
결혼(Marriage)	2
최종학력(Edu)	2
전공(Major)	4
종교(Religion)	1
월수입(Salary)	2
교통 수단(Vehicles)	4
통학/통근 시간(CommT)	3
1년간 여행 빈도(Travel)	2
SNS 사용 여부(SNS1)	1
SNS 1일 사용 시간 (SNS2)	3
문화생활(Culture)	3

<표 2>에서 응답에 해당하는 숫자가 나타내는 바는 다음과 같다. 나이(1: 10대, 2: 20대, 3: 30대, 4: 40대 이상), 직업(1: 학생, 2: 관리자, 3: 전문가, 4: 기술공, 5: 사무 종사자, 6: 서비스, 판매, 7: 기능원, 8: 장치, 기계 조작 및 조립 종사자, 9: 단순 노무 종사자), 결혼(1: 기혼, 2: 미혼), 최종학력(1: 고졸 미만, 2: 고졸, 3: 대졸, 4: 석사, 5: 박사), 전공(1: 인문, 2: 사회, 3: 교육, 4: 공학, 5: 자연, 6: 의약, 7: 예체능), 종교(1: 무교, 2: 기독교, 3: 천주교, 4: 불교), 월수입(1: 50만원 이하, 2: 50~100만원, 3: 100~200만원, 4: 200~300, 5: 300만원 이상), 통근수단(1: 도보, 2: 자전거, 3: 자동차, 4: 대중교통), 통근시간(1: 30분이내, 2: 30분 ~1시간, 3: 1~2시간, 4: 2시간 이상), 여행 빈도(1: 1회 이하, 2: 2~3회, 3: 4~5회, 4: 6회 이상), SNS 사용 여부(1: 사용, 2: 사용 안함), SNS 사용 빈도(1: 30분 이하, 2: 30분~1시간, 3: 1~3시간, 4: 3시간 이상), 문화생활(1: 정적인 활동, 2: 동적 활동, 3: 정적, 동적 모두). 따라서 실험자 1은 20대, 학생, 미혼, 고졸, 공학, 무교, 월수입은 50~100만원, 대중교통, 1~2시간 통학, 1년간 2~3회 여행, 하루 평균 SNS를 1~3시간 사용, 문화생활은 정적, 동적 활동을 모두 한다.

2.2 Target Data

지도학습인 랜덤 포레스트에 이용할 데이터 중 레이블(target data)로는 장소 데이터를 사용하였다. 장소 데이터는 SWARM 애플리케이션을 이용해 방문한 장소에 체크인하였다. 그 후 웹 크롤링을 이용하여 방문한 장소와 방문횟수를 파악하였다. 수집한 실험자 16의 일부 데이터는 다음과 같다.

<표 3> 실험자 16의 장소 데이터 일부

장소	위치	방문 횟수
홍익대학교 와우관	서강동	19
홍익대학교 정보통신관	서강동	7
카네마야제면소	서강동	3
스타벅스	서강동	4
홍익대학교 중앙도서관	서강동	8
커피스미스	서교동	2
다이소	서교동	3

이렇게 수집된 데이터를 10개의 카테고리로 분류하였다. <표 4>는 실험자 16의 데이터를 카테고리로 분류한 것이다.

<표 4> 카테고리 분류

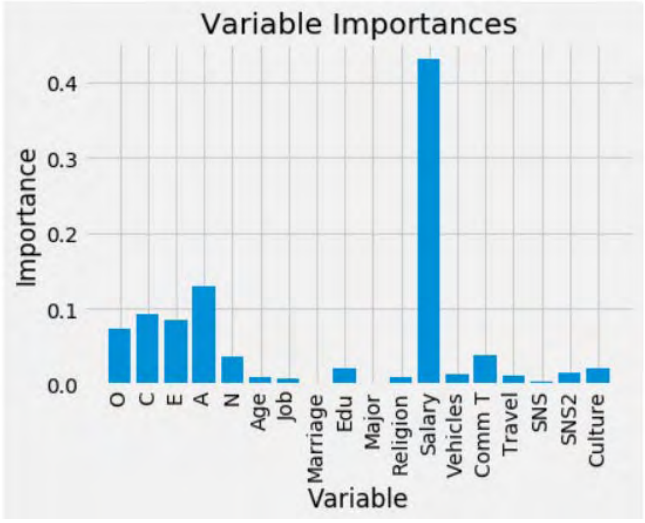
카테고리	방문 횟수	비율
국제 및 외국기관	0	0
대형 종합 소매업	6	0.04
서비스업	6	0.04
일반 음식점업	29	0.193333
주점업	2	0.013333
비알콜 음료점업	26	0.173333
영화 및 비디오, 공연 상영업	4	0.026667
교육 기관	62	0.413333
병원	6	0.04
박물관 및 사적지	9	0.06

3. 실험 결과

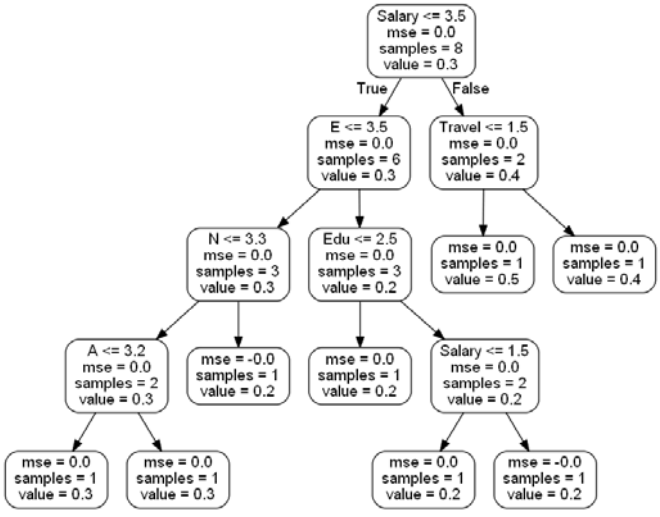
랜덤 포레스트를 이용하여 데이터를 분석하면 각 입력변수가 결과에 영향을 미치는 정도인 Variable importance(변수 중요도)를 알 수 있다. 실험 결과로 만들어진 트리와 Value Importance 그래프를 모두 첨부할 수 없어 표로 작성하였다. <표 5>에서는 실험에 대한 평가를 위해 Symmetric Mean Absolute Percentage Error (SMAPE) 값과 Accuracy, 가장 영향이 큰 상위 세 입력

<표 5> 실험 결과

	국제 및 외국기관	대형 종합 소매업	서비스업	일반 음식점업	주점업	비알콜 음료점업	영화 및 비디오, 공연 상영업	교육 기관	병원	박물관 및 사적지
SMAPE (%)	43.88	37.14	54.63	25.59	24.4	22.64	51.07	29.69	53.09	39.43
Accuracy (%)	56.12	62.86	45.37	74.41	75.6	77.36	48.93	70.31	46.91	60.57
Variable 1	O	A	O	월수입	O	월수입	월수입	O	O	여행
	0.46	0.19	0.27	0.43	0.16	0.19	0.19	0.2	0.16	0.26
Variable 2	종교	문화생활	여행	A	E	C	문화생활	A	E	N
	0.11	0.14	0.19	0.13	0.13	0.17	0.17	0.15	0.15	0.17
Variable 3	C	통근시간	N	C	A	O	교통수단	최종학력	N	A
	0.1	0.13	0.16	0.09	0.09	0.12	0.14	0.11	0.15	0.15



(그림 1) 변수 중요도 그래프(label=일반 음식점)



(그림 2) 결정 트리 예시(label=일반 음식점)

변수와 importance 값을 표현하였다. 정확성이 높았던 일반음식점의 경우의 결정 트리와 변수 중요도 그래프를 예시로 첨부하였다.

일반 음식점을 label로 했을 때의 예측 정확도는 74.41(%)이다. 월수입의 중요도가 0.43으로 가장 많은 영향을 미치는 것으로 나타났다. 성격의 요인인 우호성과 성실성도 많은 영향을 미친다. 다만 성실성의 중요도는 0.09로 많은 비중을 차지하지 않는다. 실험 결과를 분석할 때 중요도의 값이 0.1 이상일 때 영향을 미친다고 생각하였다. 박물관 및 사적지의 경우 여행을 자주 가고 신경성과 우호성이 높은 사람이 방문한다는 것을 알 수 있다. 정확도가 낮은 장소 카테고리의 예시로서 서비스업의 경우는 해당하는 업종이 너무 다양해 예측 하기 어려울 것이라고 예상된다. 병원의 경우도 예측이 어렵다. 우발적 사고를 당해 몇 달간 꾸준히 병원을 방문하는 예외적 상황이 발생할 수 있기 때문이다. 실험 결과를 종합적으로 보면 대체적으로 성격뿐 아니라 월수입, 여행 빈도, 문화생활, 교통수단, 통근시간 등도 방문하는 장소에 영향을 미친다는 것을 알 수 있다.

4. 결론

이번 연구를 통해 성격뿐 아니라 다양한 요인들이 방문하는 장소에 영향을 미친다는 것을 알 수 있었다. 다양한 요인을 통해 방문하는 장소 예측을 하는 것은 위치 기반 시스템과 추천 시스템에 활용할 수 있을 것으로 기대한다. 다만 아직은 성능의 향상이 필요하다고 생각한다. 정확도가 낮았던 몇몇 장소의 경우 분류하는 카테고리의 조정을 통해 정확도를 높일 수 있을 것이다. 특히 서비스업의 경우 포함하는 장소들이 다양하여 더 세분화할 필요가 있다고 생각한다. 또, 17 명의 데이터로 방문하는 장소를 정확히 예측하는 것은 불가능하다고 생각된다. 다음 연구에서는 더 많은 데이터와 적절한 장소 카테고리 분류를 통해 더 나은 결과를 얻고자 한다.

Acknowledgement

이 연구는 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행되었다.(NRF-2017R1D1A1B03029788)

참고문헌

[1] E. B. Lee and H. Y. Song, “An Analysis of the Relationship between Human Personality and Favored Location” 2014
[2] Ha yoon Song, Hwa Baek Kang, Analysis of Relationship Between Personality and Favorite Places with Poisson Regression Analysis,
[3] https://www.ilo.org/, 국제표준직업분류표(ISCO)
[4] P. T. Costa and R. R. McCrae, "Four ways five factors are basic," Personality and individual differences, vol. 13, no. 6, 1992, pp. 653 - 665.
[5] Gérard Biau and Erwan Scornet, “A random forest guided tour” TEST (2016) 25:197 - 227 DOI 10.1007/s11749-016-0481-7
[6] J Hoseinifar, MM Siedkalan, SR Zirak et al., “An investigation of the relation between creativity and five factors of personality in students.” Procedia - Social and Behavioral Sciences. Volume 30, 2011, Pages 2037-2041
[7] Dev Jani, Jun-Ho Jang &Yeong-Hyeon Hwang, “Big Five Factors of Personality and Tourists' Internet Search Behavior”, Asia Pacific Journal of Tourism Research Volume 19, 2014 - Issue 5.
[8] Dev Jani, Heesup Han, “Personality, social comparison, consumption emotions, satisfaction, and behavioral intentions: How do these and other factors relate in a hotel setting?”, Internatoal Journal of contemprary Hosptality management volume 25, Issue 7.