사람의 여러 특징들과 자주 방문하는 장소 간 상관관계 분석

송하윤, 윤지선
홍익대학교 컴퓨터공학과
e-mail:hayoon@hongik.ac.kr
e-mail:yjsun_@naver.com

Correlation between various characteristics of people and their favorite location

Ha Yoon Song, Jiseon Yun
Department of Computer Engineering, Hongik University

요 약

이 논문은 사람의 성격과 나이, 직업, 결혼 유무, 최종학력, 전공, 종교, 월수입, 통근 수단 등 총 14가지로 이루어진 사람의 특징 데이터와 자주 방문하는 선호 장소에 대한 데이터의 상관관계를 분석하고, 어떠한 요인이 가장 크게 영향을 미치는가에 대해 분석하였다. 분석에는 17명의 실험자가 참여하였고, 분석 방법으로는 Boosting 기법을 사용하였다. 성격 데이터는 Big Five Inventory (BFI)를 통해 얻었고, 나머지 특징들에 대한 데이터는 직접 만든 설문지를 통해 얻었으며, 장소 데이터는 Swarm 애플리케이션을 통해 얻었다.

1. 서론

사람의 성격이나 연령대, 직업 등과 같은 특징은 선호하는 장소 방문에 영향을 끼친다. 이번 연구에서는 사람의 성격과 나이, 직업, 결혼 유무, 최종학력, 통근 수단, 여행 빈도 등에 대한 데이터와 자주 방문하는 장소에 대한 상관관계를 Boosting 기법을 사용해 분석하였다. 그리고 사람의 어떤 특징이 선호 장소 방문에 가장 영향을 끼치는지에 대한 feature importance를 분석하였다.

2장에서는 상관관계 분석을 위해 사용될 Boosting 기법에 대해서 기술할 것이다. 3장에서는 분석에 사용한 사람의 특징 데이터와 장소 데이터에 대한 설명을 할 것이다. 4장에서는 feature importance 결과를 분석하고, 5장에서는 이 연구의 결론 및 앞으로의 연구방향에 관해 기술할 것이다.

2. 분석 기법

이 연구에 쓰일 분석 기법은 앙상블 기법 중 하나인 Boosting이다. Boosting은 여러 개의 weak learner들을 순차적으로 훈련시켜, 잘못 예측한 데이터에 가중치를 더해 학습하여 최종적으로 생성된 learner를 이용해 예측하는 기법이다. Boosting 알고리즘 중에서도 이 연구에서 쓰일 알고리즘은 XGBoost 이다. 이는 feature importance 분석, 즉 모델이 어떤 요인에 얼마나 의존하고 있는지를 시각화해주는 알고리즘이다. 따라서 선호하는 장소 방문에 사람의 여러 가지 요인 중 가장 유효한 요인을 분석해주는데 적합한 알고리즘이다.

3. 실험 데이터

1) 성격 데이터

성격 데이터는 Big Five Inventory (BFI) 설문지를 통해 5가지의 성격 유형으로 수치화해서 나타내었다. 5가지 유형은 각각 개방성(O), 성실성(C), 열정성(E), 동조성(A), 신경성(N) 으로 구성되어있다. <표1>은 BFI 설문지를 통해 5가지의 성격 유형을 나타낸 실험자 17명 중 6명의 성격 데이터이다.

<표1> BFI를 사용한 실험자 6명의 성격 데이터

	O	C	E	A	N
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

- 개방성 O : Openness
- 성실성 C : Conscientiousness
- 열정성 E : Extraversion
- 동조성 A : Agreeableness
- 신경성 N : Neuroticism

<표2> 설문지를 통해 얻은 사람의 여러 가지 특징 데이터

	Age	Job	Marriage	Edu	Major	Religion	Salary	Vehicles	Comm T	Travel	SNS	SNS2	Culture
실험자1	2	1	2	2	4	1	2	4	3	2	1	3	3
실험자2	2	1	2	2	4	3	2	4	3	2	2	0	3
실험자3	3	3	2	4	4	2	5	2	2	2	2	0	2
실험자4	2	1	2	4	4	4	2	4	3	3	1	2	2
실험자5	2	1	2	3	4	1	1	4	3	1	1	3	1
실험자6	4	3	1	5	4	1	5	4	1	2	1	2	1

2) 성격 외 특징 데이터

사람의 성격 외 특징들에 대한 데이터는 직접 만든 설문지를 통해 수집하였다. 그리고 각각의 요인들에 대한 범주를 수치화해서 나타내었다. <표2>는 설문지를 통해 얻은 실험자 17명 중 6명의 특징 데이터이다.

Age는 나이를 나타내며, 1은 10대, 2는 20대, 3은 30대, 4는 40대 이상에 해당된다. Job는 직업을 나타내며, 국제표준직업분류(ISCO) 기준에 '학생'을 더하여 범주를 정하였다. 1은 학생, 2는 관리자, 3은 전문가, 4는 기술공, 5는 사무 종사자, 6은 서비스업 및 판매 종사자, 7은 기능원, 8은 장치 및 기계 조작 종사자, 9는 단순 노무 종사자에 해당된다. Marriage는 결혼 유무를 나타내며 1은 기혼, 2는 미혼에 해당한다. Edu는 최종학력을 나타내며 1은 고등학교 졸업 미만, 2는 고등학교 졸업자, 3은 대학교 졸업자, 4는 석사, 5는 박사에 해당된다. Major는 전공을 나타내며 1은 인문 계열, 2는 사회 계열, 3은 교육 계열, 4는 공학 계열, 5는 자연 계열, 6은 의약 계열, 7은 예체능 계열에 해당된다. Religion은 종교를 나타내며 1은 무교, 2는 기독교, 3은 가톨릭교(천주교), 4는 불교에 해당된다. Salary는 월수입을 나타내며 1은 50만 원 이하, 2는 50만 원 이상 100만원 미만, 3은 100만 원 이상 200만원 미만, 4는 200만 원 이상 300만 원 이하, 5는 300만 원 이상에 해당된다. Vehicle은 통근 수단을 나타내며 1은 도보, 2는 자전거 이용, 3은 자차 이용, 4는 대중교통 이용에 해당된다. comm T는 통근 시간을 나타내며 1은 30분 이내, 2는 30분 이상 1시간미만, 3은 1시간 이상 2시간미만, 4는 2시간 이상에 해당된다. Travel은 여행 빈도를 나타내며 1은 1회 이하, 2는 2회 이상 4회 미만, 3은 4회 이상 6회 미만, 4는 6회 이상에 해당된다. SNS는 SNS 사용 유무를 나타내며 1은 사용, 2는 미사용에 해당된다. SNS2는 SNS 하루 사용 시간을 나타내며 1은 30분 이하, 2는 30분 이상 1시간 미만, 3은 1시간 이상 3시간 미만, 4는 3시간 이상에 해당된다. 마지막으로 Culture는 문화생활을 나타내며 1은 정적인 활동, 2는 동적인 활동, 3은 정적인 활동과 동적인 활동을 모두 하는 혼합형에 해당된다.

3) 장소 데이터

장소 데이터 수집에는 SWARM 애플리케이션을 이용하였다. SWARM은 사용자가 장소를 방문할 때, 방문 위치를 기록해주는 애플리케이션이다. 장소의 이름과 위치, 방문 횟수는 기록된 데이터로부터 크롤링을 통해 얻어냈다. <표3>은 실험자 1이 수집한 방문 데이터 중 일부이다.

<표3> 실험자 1의 방문 데이터 일부

장소 이름	장소 위치	방문 횟수
홍익대학교 문헌관	서강동	5
할리스 커피	서교동	1
써브웨이	아현동	1
영등포 구청	영등포구	1

<표4> 실험자 5명의 최종 장소 데이터

	외국기관	대형 종합 소매업	서비스업	일반 음식점업	주점업	비알콜 음료점업	영화 및 비디오물 상영업	고등 교육기관	병원	사적지 관리 운영업
실험자1	0.01705	0.05634	0.02965	0.19496	0.02743	0.19422	0.01705	0.43662	0.00741	0.01927
실험자2	0.00551	0.6725	0.00162	0.0762	0.01232	0.07847	0.00551	0.14008	0.00292	0.00486
실험자3	0.13559	0.04237	0.0226	0.4096	0.00847	0.0791	0.00565	0.27401	0.0226	0
실험자4	0.25833	0.01667	0.00333	0.15167	0.02	0.06167	0.01	0.47333	0	0.005

장소 데이터는 10개의 업종분류표에 각각의 방문 데이터를 적용해 카테고리 별로 방문 횟수를 합산하여 만들었다. 10개의 업종분류는 외국기관, 대형 종합 소매업, 서비스업 등에 해당한다. 최종적으로, 장소 데이터는 이 방문 횟수를 전체 방문횟수 대비 해당 카테고리 방문 횟수로 비율을 계산하여 완성하였다. <표4>는 최종적으로 얻은 실험자 17명 중 5명의 장소 데이터 중 일부이다.

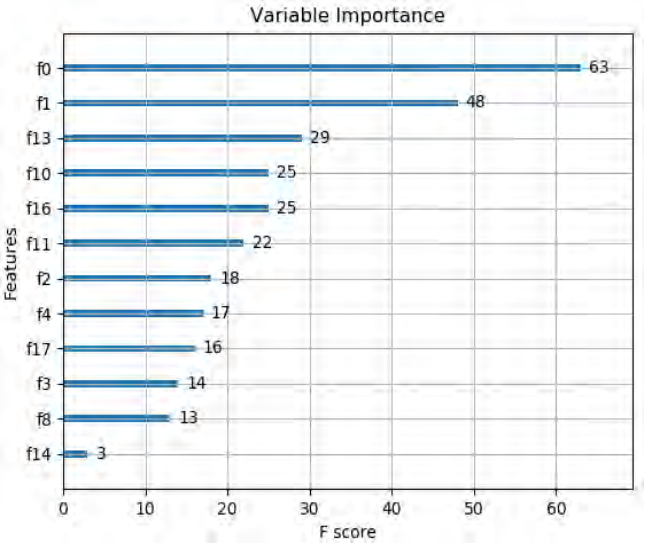
4. 실험 및 결과 분석

독립변수에 들어갈 특징 데이터는 <표1>에서와 같이 BFI를 사용하여 얻은 성격 데이터와, <표2>에서와 같이 설문지를 통해 얻은 사람의 여러 가지 특징 데이터를 병합하여 만들었다. <표5>는 실험자 17명 중 3명의 특징 데이터이다. 종속 변수에는 <표4>와 같은 장소 데이터를 넣었다. 그 다음, 성격을 포함한 사람의 여러 가지 특징 중 어떠한 요인이 장소 데이터에 가장 유효한지를 분석하였다. 분석 기법으로는 2장에서 언급한 xgboost를 사용하였다.

<표5> 실험자 3명의 특징 데이터

	실험자1	실험자2	실험자3
O (f0)	3.3	2.7	4.3
C (f1)	3.9	3.2	3.1
E (f2)	3.3	3.2	2.3
A (f3)	3.7	2.7	3.2
N (f4)	2.6	2.8	2.9
Age (f5)	2	2	3
Job (f6)	1	1	3
Marriage (f7)	2	2	2
Edu (f8)	2	2	4
Major (f9)	4	4	4
Religion (f10)	1	3	2
Salary (f11)	2	2	5
Vehicles (f12)	4	4	2
Comm T (f13)	3	3	2
Travel (f14)	2	2	2
SNS (f15)	1	2	2
SNS2 (f16)	3	0	0
Culture (f17)	3	3	2

(그림 1)은 xgboost로 feature importance 분석을 실행한 결과들 중, 외국기관에 대한 독립변수의 영향 정도를 나타낸 것이다. x축(feature)은 사람의 특징 데이터에 포함된 각각의 요인들을 나타내고, y축(F score)은 종속변수에 대한 독립변수의 유효도를 나타낸다. f0부터 f17은 <표5>에 나와 있는 요인 순으로 나타낸 것이다. (그림1) 결과를 보면, 외국기관으로 분류된 장소 데이터에 f0(O, 개방성)가 가장 영향을 끼친 것을 알 수 있다. 그 다음으로, f1(C, 성실성), f13(통근 시간), f10(월수입) 등의 순서로 영향을 미친다고 해석할 수 있다. 이러한 방법으로, 어떠한 요인이 각각의 장소 데이터에 가장 영향을 미치는지에 대한 결과값을 얻을 수 있다.



(그림 1) 외국기관에 대한 xgboost 실행 결과

5. 결론

이 연구에서는 사람의 여러 가지 특징과 선호하는 장소 방문 간의 상관관계를 boosting을 통해 분석하였다.

연구 결과, 사람의 특징 중 어떠한 요인이 선호 장소 방문에 가장 영향을 끼치는지 확인하였다. 보다 정확한 결과를 얻기 위해서는 xgboost의 여러 가지 매개변수를 좀 더 조정하고, 더 많은 실험자들의 데이터가 필요하다. 따라서 다음 연구에서는 실험자의 수를 늘리고, 분석기법 xgboost의 매개변수를 더 세밀히 조정하는 과정을 거쳐, 보다 정확한 결과를 얻을 수 있도록 하겠다.

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] 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
[6] 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.
[7] 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.