# 회귀분석을 통한 BFI 성격 데이터와 위치 데이터의 관계 분석

김승연*, 송하윤
홍익대학교 컴퓨터공학과
e-mail : brdosa@naver.com, hayoon@hongik.ac.kr

# Analysis of the relationship between BFI Personality Data and Location Data through Regression

SeungYeon Kim* , Ha Yoon Song
Department of Computer Engineering, Hongik University

요 약

심리학 연구에 따르면, 인간은 각자의 성격에 따라 이동패턴이 변화한다고 한다. 하지만 실험적 근거가 아닌, 어디까지나 가설로만 사용되어 왔다. 우리의 연구에서는 이런 가설을 증명하기 위해 실제 실험 참가자를 모집하였고, 각 참가자들의 GPS데이터와 BFI성격 데이터를 수집하였다. 그리고 BackProgagation Network를 이용하여, 새로운 위치 데이터를 추론하고, 이렇게 추론된 결과를 바탕으로 회귀분석을 하여, 실제 사람의 성격과 위치 데이터간의 관계를 통계적인 방법에 의해서 보여줄 것이다.

논문의 내용 중 첫 번째로 우리가 지금까지 한 선행 연구에 대해서 설명한다. 여기서 어떻게 참가자를 모집했으며, 각 GPS정보와 BFF성격 정보를 BPN에 학습시키는지 보여줄 것이다. 두 번째로 선행 연구에서 만든 BPN을 바탕으로 어떻게 회귀분석을 하는지 보여줄 것이며, 세 번째로 회귀분석을 통해 나온 통계적인 데이터를 분석하고, 거기에서 의미를 해석할 것이다.

## 1. 서론

선행 연구에서 실제 실험자의 GPS 위치 데이터와 성격 데이터를 이용한 사람의 성격 데이터를 사용했다. 그리고 성격 데이터와 위치 데이터를 패턴화 시키기 위해 Back Propagation Network알고리즘을 이용했으며, 새로운 위치 데이터를 추론하는 방법론을 보였다[1].

이 논문에서는 선행연구에서 가정으로 사용하던, '인간의 이동 패턴은 개인의 성격에 영향을 받는다.' [2]를 통계적인 방법에 의해 증명하고자 한다. 예를 들어, '개방적인 성향인 사람이 개방적이지 않은 사람에 비교하여 학교에 자주 있다.' 라는 가정을 실제 수치를 들어 보여줄 것이다. 그리고 회귀분석이라는 통계를 통해 우리가 수집한 데이터에서의 성격 데이터와 위치 데이터간의 어떤 관계가 있는지 보일 것이다.

인간의 성격을 나타내는 심리학적 표현 방법은 많이 있지만, 우리는 성격 표현방법에 관하여 가장 일반적으로 많이 쓰이는 BFI(Big Five Inventory)[3]를 이용할 것이다. 그리고 위치 데이터를 나타내기 위해 GPS장치로 수집한 정보를 사용한다.

이 연구가 다른 위치 예측 방법론과의 차이점은. 첫째로 우리는 예측 알고리즘을 다른 방법론들은 주로 Markov Model을 사용한데 반해[3][4], 이 논문에서는 Back Propagation NetWork(이하 BPN)를 사용한데 있다. BPN은 신경 네트워크의 일종으로 입력 데이터의 패턴을 학습하여, 학습된 정보를 바탕으로 새로운 입력 데이터의 패턴을 분석하는 알고리즘이다. Markov Model이 아닌 BPN을 사용하는 의미는 시간에 따른 위치 이동을 Sequence로 본 게 아니라 연속적인 흐름으로 본다는 것을 나타낸다. 즉 Markov Model과 비교하여, 좀 더 유연한 입력 데이터 처리가 가능하다. 두 번째로 우리는 예측 하는 방법에 대해 성격 데이터를 사용하는데, 다른 예측 방법론들은 성격 데이터를 사용하지 않는다. 이는 이 논문은 '위치 데이터의 이동에는 사람의 성격이 영향을 준다.' 라는 가정을 했기 때문이다.

2장에서는 선행 연구에 대해서 설명할 것이고, 주로 어떤 데이터를 사용하고, 어떻게 수집했는지, 그리고 BPN을 어떻게 설계하고 사용했는지 나타낼 것이다. 그리고 3장에서는 성격과 위치 데이터간의 관계를 통계적으로 나타내기 위해 회귀분석을 어떻게 사용했는지 말할 것이며, 4장에서는 3장에서 언급한 방법으로 회귀분석이 있을 것이다.

## 2. 선행 연구

지금까지 했던 연구는 사람의 과거의 위치 데이터 정보를(Historical Data)를 이용하여, 현재의 또는 특정 시간에서의 사람의 위치 데이터를 추론하기 위한 방법론이었다[3][4]. 또한 다른 연구와는 달리, 위치 데이터 이외에 사람의 성격 데이터 BFF(Big Five Factor)를 사용하였다. BFF는 사람의 성격을 5가지 요소로 분류하는 심리학 모델로서[5], 사람의 성격을 분석할 때 많이 사용된다. 또한 각 성격들은 직교성(orthogonality)[6]을 만족하기 때문에 BPN같은 패턴 알고리즘에 적용하기 적합하다.

실험에서는 총 5명의 참가자가 참가하였다. 각 참가자는 BFF 설문조사를 통하여 성격데이터를 구했으며[5], 6계월 동안 GPS 수집장치(Garmin) 또는 스마트폰 어플을(Sport Tracker) 이용하여 위치 데이터를 수집했다. 그림 1은 실제 실험 참가자의 위치 정보를 시간당 각 주요 위치에 있을 가중치를 나타낸다. 그림 1에 대한 간략한 설명을 하자면, 참가자 1은 14시에는 학교와 집의 값이 약 0.5로 비슷한 것을 알 수 있다. 따라서 참가자 1은 14시 에는 학교와 집에 있을 확률이 비슷하다는 걸 알 수 있다. 그리고 15시쯤에 집에 해당하는 값이 증가하고, 학교에 해당하는 값이 감소하는 것을 알 수 있는데, 이는 참가자 1이 약 15시에 학교에서 집으로 온다는 것으로 해석 할 수 있다.

그림 1과 같은 양식의 위치 데이터를 각 참가자들에 대하여, 성격과 시간의 정보를 패턴의 단서로 제공하고, 위치 데이터를 패턴의 결과하게 BPN을 학습시켰다. 즉 하나의 BPN에 각 참가자들의 정보가 반영시킨 것이다. 그림 2는 그림 1에 해당하는 참가자의 정보를 BPN에 입력했을 때의 결과를 나타낸다. 그림 2는 그림 1하고 같은 방법으로 해석할 수 있는데. 그림 2에서 약 14시에 학교에 해당하는 그래프의 값이 증가하는 것을 알 수 있다. 즉 참가자 1은 약 14시 전후로 해서 학교에서 집으로 갈려는 성향이 있다는 걸 알 수 있다[8][9].

그림 1과 그림 2를 비교해 보면, 그래프의 세밀한 면은 다소 차이가 있으나, 그래프의 전반적인 계형이 상당히 유사한 것을 알 수 있다. 앞에서 예로 든 것처럼 그림 1과 그림 2는 둘 다 비슷하게 약 15시 전후로 학교에서 집으로 이동한다고 해석할 수 있다. 즉 이런 비슷한 계형의 그래프를 가지는 것으로 보아, 다양한 사람의 정보가 입력된 BPN을 통해 실제 데이터와 유사한 위치 정보를 예측할 수 있음을 알 수 있다.

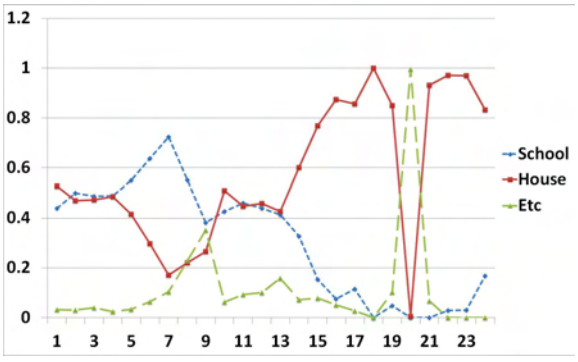
지금까지 이전의 실험에 대해서 언급했다. 이제 부터는 이 실험의 목적인 학습된 BPN에서 성격 데이터와 위치 데이터간의 관계를 경험에 따른 가정이 아닌 수학적인 방법으로 증명할 것이다. 이를 통해, 단순히 사람의 성격과 위치 데이터의 관계가 가정이 아닌, 실제 관련이 있음을 보일 수 있게 된다.

## 3. 실험 설계

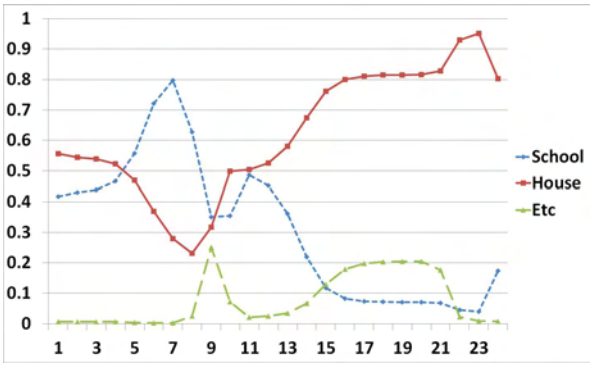
3장에서는 2장에서 학습 시킨 BPN을 이용하여, 성격 데이터와 위치 데이터간의 관계를 회귀 분석을 통해 보여줄 것이다[7].

회귀분석[7]에 대해서 간단하게 설명하면, 통계모델의 한 종류로서, 독립 변수와 종속 변수 간의 관계의 정도를 파악할 때 사용된다. 이때 독립 변수는 1개 일수도 있고, 두 개 이상일 수도 있다. 분석결과는 독립 변수와 종속 변수가 선형 함수나 로그 함수같이 어떤 함수에 따르는지 그리고 얼마 정도로 특정 함수와 유사한지를 나타내는 방식으로 사용된다. 회귀분석에 관한 예로서, '아이스커피의 판매량과 일일 최고 기온과의 관계.' 가 있다.

2장에서 학습된 BPN은 사람의 성격을 패턴으로 분석한 것이다. 따라서 BPN의 입력 값인 '성격' 과 출력 값인 '위치' 에 대해 회귀분석을 하면, '성격' 과 '위치' 의 관계의 정도나, 그리고 어떤 그래프 형태로 관련이 있는지 알 수 있을 것이다. <표 1>은 회귀분석에서 독립변수로 사용할 성격 데이터를 나타내는 것이다. <표 1>에서 '보통' 에 해당하는 값은 BFF 성격 분포에서 '평균적인 성격' 값을 나타난다. 예로 들어 개방성 값이 3.21이 나온 사람은 BFF 분포에서 '평균 적인 개방성' 을 가진다는 걸 나타낸다. 마찬 가지로, '많이 낮음' ,과 '많이 높음' 은 BFF성격 분포표에서 하위 10%와 상위 10%에 해당하는 값을 나타내는데, 예로 들어 개방성이 2.71이 나온 사람이면, 이 사람은 분포상 개방성이 하위 10%에 해당한다고 보면 된다. 따라서 독립변수에 사용될 데이터의 개수는 총 78125(5^7)개의 성격 데이터 조합이 나온다. 또한 회귀분석에 사용



<그림 1> 참가자 1의 위치 데이터 그래프



<그림 2> 참가자 1의 학습된 위치 데이터 그래프

| | 많이 낮은 | 낮은 | 약간 낮은 | 평균 값 | 약간 높은 | 높은 | 많이 높은 |
|---|---|---|---|---|---|---|---|
| 개방성 O | 2.71 | 2.876 | 3.043 | 3.21 | 3.376 | 3.54 | 3.71 |
| 성실성 C | 2.866 | 3.026 | 3.186 | 3.346 | 3.506 | 3.666 | 3.826 |
| 외향성 E | 2.574 | 2.76 | 2.947 | 3.134 | 3.32 | 3.507 | 3.694 |
| 협동성 A | 3.227 | 3.393 | 3.56 | 3.727 | 3.893 | 4.06 | 4.227 |
| 예민성 N | 2.31 | 2.48 | 2.66 | 2.84 | 3.016 | 3.193 | 3.37 |

<표 1> 성격 데이터의 분포표

될 종속변수 값은 <표 1>의 성격 데이터 값을 BPN에 입력시 출력되는 값들을 사용할 것이다. 즉 우리가 알고자하는 종속 변수의 종류는 '학교', '집', '기타' 이고, 독립변수와 마찬가지로 총 78125개의 데이터를 사용하여 분석할 것이다.

## 4. 실험 및 결과

4장에서는 3장에서 언급한 회귀분석에 따라 실제 BPN알 회귀분석에 적용하고, 회귀분석의 결과 값을 볼 것이다.

<표 2>는 <표 1>의 성격 데이터와, 그 성격 데이터를 BPN에 입력 했을 때 출력되는 위치 데이터 간의 회귀 분석을 했을 때의 결과이다. 논문의 내용이 제한적이기 때문에 위치 데이터는 '학교' 만을 사용하였고, 전체 시간에서 0~ 3시까지의 데이터만을 사용하였다. 열값부터 설명하자면, Y절편은 말 그대로 독립 변수와 종속 변수 간의 관계를 함수식으로 구할 때 함수에서의 상수 값을 나타낸다. 즉 성격과 위치데이터간의 관계를 파악할 때에는 딱히 중요한 값이 아니다. 다음 값으로 순서대로 개방성, 성실성, 외향성, 협동성, 예민성이 있는데, 각 행에 해당하는 성격을 나타낸다. 행의 값들에 대해서 언급하면, 계수 값은 해당하는 성격 값이 종속변수에 대해서 어느정도 영향을 주는지 나타낸다. 즉 계수 값이 클수록 해당하는 성격 값이 증가 할수록 종속 변수의 값이 크게 변화한다. 반대로 계수의 값이 작을수록 해당하는 독립변수는 종속변수에 낮은 영향을 준다는 것이다. 또한 계수가 양수면 '긍정적인' 영향을 주고, 반대로 음수이면 '부정적인' 형향을 준다. 또한 'P-값' 이 있는데 'P-값' 은 해당하는 독립변수가 얼마나 신뢰성이 있게 종속변수에 영향을 주는지 나타내는 것이다. 보통은 0.05% 미만이면 해당하는 독립변수가 종속변수에 영향을 준다고 해석한다. 따라서 <표 2>를 해석하자면, 'P-값' 에 따라 각 독립 변수들이 종속 변수 값에 충분한 영향을 준다고 생각할 수 있다. 또한 '학교' 에 대해 '개방성' 과 '성실성' 은 '긍정적' 인 영향을 주고, 반대로 '외향성', '협동성', '예민성' 은

| | | 계수 | 표준 오차 | t 통계량 | P-값 |
|---|---|---|---|---|---|
| 0시 | Y 절편 | 0.493938 | 0.012515 | 39.46921 | 0 |
| | 개방성 O | 0.060183 | 0.001744 | 34.4986 | 4.8E-252 |
| | 성실성 C | 0.166806 | 0.001817 | 91.79387 | 0 |
| | 외향성 E | -0.04882 | 0.001558 | -31.3407 | 1.3E-209 |
| | 협동성 A | -0.15724 | 0.001744 | -90.136 | 0 |
| | 예민성 N | -0.13217 | 0.001646 | -80.3077 | 0 |
| 1시 | Y 절편 | 0.401541 | 0.009864 | 40.70609 | 0 |
| | 개방성 O | 0.078541 | 0.001375 | 57.11768 | 0 |
| | 성실성 C | 0.190687 | 0.001432 | 133.1272 | 0 |
| | 외향성 E | -0.06924 | 0.001228 | -56.3966 | 0 |
| | 협동성 A | -0.13658 | 0.001375 | -99.3247 | 0 |
| | 예민성 N | -0.14275 | 0.001297 | -110.042 | 0 |
| 2시 | Y 절편 | 0.379067 | 0.009932 | 38.16535 | 0 |
| | 개방성 O | 0.059091 | 0.001385 | 42.67966 | 0 |
| | 성실성 C | 0.195112 | 0.001442 | 135.2865 | 0 |
| | 외향성 E | -0.05641 | 0.001236 | -45.6342 | 0 |
| | 협동성 A | -0.13967 | 0.001385 | -100.879 | 0 |
| | 예민성 N | -0.11564 | 0.001306 | -88.5336 | 0 |

<표 2> 0시 ~ 2시 까지의 성격과 학교에 있을 가중치 값과의 회귀 분석 결과

학교에 있을 확률에 '부정적' 인 영향을 준다고 해석할 수 있다. 이때 '개방성' 보다 '성실성' 이 '학교' 의 위치 가중치 값에 좀 더 큰 영향을 주고, '협동성' 과 '예민성' 둘 다 비슷한 수치로 학교의 값에 '부정적' 인 영향을 준다고 볼 수 있다. 이런 해석에 의미를 부여하자면, 성격값에 대해서 성격이 '개방적' 이고 '성실성' 한 사람일수록 학교에 자주 있으며, 반대로 '외향성', '협동성', '예민성' 이 높을수록 학교 오랫동안 있지 않는다.

## 5. 결론 및 분석

우리는 이전 실험에서 실험자의 실제 성격 데이터와 위치 데이터를 사용하여, 시간에 따른 사람의 현재 위치를 추론하는 방법론에 대해 언급했다. 또한 이를 사용하여, 우리가 사용한 BPN에 대하여 성격변수와 위치 데이터 간의 관계를 파악하기 위해 회귀분석을 사용하였다. 이런 실험들을 통하여 실제 위치 데이터를 통한 새로운 위치 데이터를 추론하는 게 가능하다는 것을 보였고. 회귀 분석을 통하여, 성격 데이터와 위치 데이터간의 관계를 통계적인 방법을 통하여 분석 가능함을 보였다.

하지만 실험에서 사용된 변수가 너무 적다고 본다. 떨어지는 이유는 RawData의 부족하기 때문이라 생각한다. 따라서 우리의 다음 실험은 참가자를 더욱 모집하여 RawData를 더욱 수집할 것이며, 더욱 정확도 있는 실험을 하는 것이 목적이다.

그 외에도 우리는 심리데이터를 BFF로 사용하였는데, BFF이외의 심리 지표를 심리 데이터로 사용하는 방법에 대해 고려해 볼 것이다. 또한 현재 우리 연구실에서 HMM를 이용한 위치 클러스터 분석법이 있는데, 이런 분석 방법과 논문에서 언급한 예측 방법을 융합하여 한 단계 높은 결과를 내보려고 한다.

## 6. Acknowledgement

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No. 2012046473)

## 7. 참고문헌

[1] Jillian Anable, "'Complacement Car Addicts' or 'Aspring Enviromentalists'? Identifying travel behaviour segments using attitude theory", Transport Policy, Vol.12, pp.65-78, 2005

[2] Giuseppe Carrus, Paola Passafaro, Mirilia Bonnes "Emotions, habits and rational choices in ecological behaviours: The case of recycling and use of public transportation" Journal of Environmental Psychology,2008.

[3] Burbey, Ingrid E. Predicting future locations and arrival times of individuals. Diss. Virginia Polytechnic Institute and State University, 2011.

[4] G. F. Luger, "Artificial intelligence: Structures and strategies for complex problem solving," 2008.

[4] Walter Mischel ,손정락 역, 성격심리학, 시그마프레스 출판사, ch-3, pp.273-293, 2006

[5] Schmitt D. P., Allik J., McCrae R. R., Benet-Martinez V. "The geographic distribution of Big Five Personality Traits:patterns and profiles of human self-description across 56 nations", J. Cross Cult. Psychol., Vol.38, pp.73-212. 2007.

[6] Jack Block, "A Contrarian View of the Five-Factor Approach to Personality Description", University of California, Berkeley, Psychological Bulletin, Vol. 117, No. 2, 187-215, 1995.

[7] DRAPER, Norman Richard; SMITH, Harry. Applied regression analysis 2nd ed. 1981.

[8] Gokul Chittaranjan, Jan Blom and Daniel Gatica-Perez, "Whos who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones", the Proceedings of the Internatonal Symposium onWearable Computers, San Francisco, California, June 2011.

[9] Chang-Hyeon Joh , H. J.P. Timmermans & T. A. Arentze (2006): "Measuring and Predicting Adaptation Behavior In Multidimensional Activity-Travel Patterns", Transportmetrica, 2:2, 153-173.