The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 19, No. 1, pp.97-102, Feb. 28, 2019. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2019.19.1.97
JIIBC 2019-1-13

Big 5 성격 요소와 머신 러닝 알고리즘을 통한 창의적인 사람들의 특징 연구

Feature Selection for Creative People Based on Big 5 Personality traits and Machine Learning Algorithms

김용준*

Yong-Jun Kim*

요 약 창의적인 사람에 대한 정확한 기준이나 수치화를 사용하여 체계적인 분류와 분석 방법이 없었기에 정의하는 데에 어려움이 많다. 이 문제를 해결하기 위하여 본 연구에서는 창의적인 사람을 어떻게 구분 지을 수 있을지에 대한 것과 어떤 유사한 성격이 있는지 분석한다. 본 연구에서 우선 Big 5 성격 특성 기법을 이용하여 설문조사를 진행하고, 그 설문조사로 얻은 데이터 세트를 가지고 데이터 마이닝 도구인 WEKA를 이용하여 데이터 세트를 분류하고 분석한 뒤, 창의적인 사람들과 연관성 있는 성격 특징들을 다양한 머신 러닝 기법을 이용하여 분석하는 것을 목표로 진행하였다. 7개의 특징 선택 알고리즘을 활용하고, 특징 선택 알고리즘들로 분류된 특징 집단을 선택하여 머신 러닝 알고리즘에 적용하여 정확도를 알아냈고, 서포트 벡터 머신을 통해 나온 특징이 가장 높은 분류 결과를 도출하였다.

Abstract There are many difficulties to define because there is no systematic classification and analysis method using accurate criteria or numerical values for creative people. In order to solve this problem, this study attempts to analyze how to distinguish creative people and what kind of personality they have when distinguishing creative people. In this study, I first survey the Big 5 personality trait, classify and analyze the data set using the data mining tool WEKA, and then analyze the data set related to the creativity The goal is to analyze the features using various machine learning techniques. I use seven feature selection algorithms, select feature groups classified by feature selection algorithms, apply them to machine learning algorithms to find out the accuracy, and derive the results.

Key Words : Big 5, WEKA, Datamining, Machine Learning, Select attributes

Ⅰ. 서 론

기술 산업 분야의 발전으로 인하여 인간을 필요로 하여 인간들이 처리했던 수 많은 일들에서 컴퓨터가 처리에 도움을 주어 대체하고 있는 추세이다. 따라서 현재 기술 산업 분야의 발전에 필요한 사람들은 새로운 지식을 잘 이해하여 수행하는 것도 중요하지만, 새로운 지식을 창의적으로 창조하는 것이 요구되고 있는 사회이다. 미국의 심리학자 Guilford의 말은 “기발하고 새로운 답이나 아이디어를 제시 할 줄 아는 발산형 사람을 창의적인 사고 특성을 가진 사람이다.”라고 정의하였다[1]. 이러한 사람들을 일컬어 창의적인 사람이라고 말한다

*준회원, 아주대학교 컴퓨터공학과
접수일자: 2019년 1월 6일, 수정완료: 2019년 2월 8일
게재확정일자: 2019년 2월 8일

Received: 6 January, 2019 / Revised: 8 February, 2019
Accepted: 8 February, 2019
*Corresponding Author: yongjun615@gmail.com
Dept. of Computer Engineering, Ajou University, Korea

사회에서는 창의적인 사람을 뽑는다, 창의적인 사람을 양성하길 원한다 라는 슬로건을 붙인 회사들이 늘어가는 추세이다. 하지만, 이 창의적인 사람에 대한 정확한 기준이나 수치화를 사용하여 체계적인 분류와 분석 방법이 없었기에 정의하는 데에 어려움이 많다.

이 문제를 해결하기 위하여 본 연구에서는 창의적인 사람을 어떻게 구분 지을 수 있을지에 대한 것과 창의적인 사람을 구분했을 때 어떠한 성격들을 가지고 있는지를 분석하였다.

본 연구에서 우선 Big 5 성격 특성 기법을 이용하여 설문조사를 진행하였고, 그 설문조사로 얻은 데이터 세트를 가지고 오픈 소스 기반의 데이터 마이닝 도구인 WEKA를 이용하여 데이터 세트를 분류하고 분석한 뒤, 창의적인 사람들과 연관성 있는 성격 특징들을 다양한 머신 러닝 기법을 이용하여 분석하는 것을 목표로 진행하였다. 본 연구에서는 기존 연구에서 하지 못한 다양한 사람들의 데이터 세트를 가지고 7개의 특징 선택 알고리즘을 활용하고, 특징 선택 알고리즘들로 분류된 특징 집단을 선택하여 머신 러닝 알고리즘에 적용하여 정확도 알아내고, 검증하였다.

이를 통하여 창의적인 사람들이 가지고 있는 성격 들과 그 유사한 성격들의 연관성과 공통점을 체계화하여 분석하였다.

II. 관련 연구

1. 기존 연구

본 연구의 기초연구로 진행한 미국 다트머스 대학교에서 개발한 "Student Life" 어플리케이션을 통해 정신건강, 학업 성취도를 평가하기 위해 10주동안 48명의 학생에게 수집된 데이터를 바탕으로 연구를 진행하였고. 공개된 데이터 세트에서, 창의적인 사람임을 나타내는 속성을 클래스 레이블로 정하고 나머지 속성들을 인자로 설정하여 분석하였고, 이 후 다양한 특징 선택 알고리즘을 통해 특징 집단을 선택하고, 다양한 머신 러닝 알고리즘을 활용하여 정확도를 검증하였으며 그 결과 최고 우선 검색 알고리즘을 통해 도출된 특징이 가장 높은 분류 결과를 도출한 연구였다. 추후 추가적으로 데이터 세트를 다른 설문조사 및 관련된 다른 데이터 값을 추가하여 더 정확한 분석 연구로 발전시키기 위해 본 연구를 진행하였다[2].

2. 관련연구

가. 장재영, 이병준, 조세진, 한다혜, 이규홍의 연구 [3]에서는 외식 블로그들 중에서 광고 블로그들을 수집하고 공통적으로 나타나는 특징을 분석한 후, 나이브 베이즈 분류 알고리즘과 신경망 분류 알고리즘을 활용한 자동 분류 알고리즘을 이용하여 분류 정확도를 실험하고, 최적의 알고리즘과 특징 조합을 탐색하였다.

나. 김완섭의 연구[4]에서는 데이터 마이닝 기법을 적용한 심층적인 분석 기법을 제시하였다. 대용량의 데이터에 숨겨져 있는 지식 또는 의미를 추출하여 설문 분석에 이용하였다. 분류 방법으로는 SPSS의 Clementine이라는 데이터 마이닝 도구에서 지원하는 C5.0 의사결정 트리 분류 알고리즘을 사용하였으며 학교 내 재학생에 대한 설문자료의 분석을 수행하였다. 결과로서 성적과 다른 항목들과의 연관성을 계층적으로 분석할 수 있었다.

다. Pratama의 연구[5]에서는 소셜미디어에서 설문지의 응답과 게시물을 분석하여 인간의 성격에 대해 분석 및 분류를 진행하였다. Twitter와 Facebook에서 데이터 수집 과정을 통해 텍스트 데이터 세트를 구성하였다. 나이브 베이즈, K-최근접 이웃 알고리즘 및 서포트 벡터 머신을 사용하였고, 이 중 나이브 베이즈 분류가 가장 좋은 결과를 보였다.

3. WEKA

웨카(Weka, Waikato Environment for Knowledge Analysis)는 자바로 개발된 기계 학습 소프트웨어 제품군으로, 데이터 마이닝 작업을 위한 기계 학습 알고리즘 모음이고, 뉴질랜드 와이카토 대학교에서 개발되었다.

Weka는 GNU General Public License 하에서 사용 가능한 자유 소프트웨어이고, 알고리즘은 데이터 세트에 직접 적용하거나 자신의 Java 코드에서 호출 할 수 있다. Weka 는 데이터 사전 처리, 분류, 회귀, 클러스터링, 연결 규칙 및 시각화를위한 도구를 포함 한다 .

Weka 워크 벤치는 데이터 분석 및 예측 모델링을위한 시각화 도구 및 알고리즘 모음과이 기능에 쉽게 액세스 할 수있는 그래픽 사용자 인터페이스를 포함한다[6].

4. Big 5

5가지 성격 특성 요소(Big Five personality traits)는 심리학에서 경험적인 조사와 연구를 통하여 정립한 성격 특성의 다섯 가지 주요한 요소 혹은 차원을 말한다. 신경성, 외향성, 친화성, 성실성, 경험에 대한 개방성의 다섯 가지 요소가 있으며, Costa & McCrae에 의해서 집대성된 모델로 다양한 나라들에서 그 유효성이 확인된 바 있다. 현대 심리학계에서 가장 널리 인정받고 있는 성격이론이다.

수많은 연구 결과 성격 5요인 이론이 개인의 행복, 신체적·정신적 건강, 종교성, 정체성뿐 아니라 가족, 친구, 연인 사이에서의 각종 관계적 결과들 및 직업 선택, 직무 만족도, 수행, 사회 참여, 범죄 행동, 정치적 입장 같은 요소들을 잘 예측한다는 것이 밝혀졌다.

이 이론을 토대로 한 검사로는 NEO - PI-R 성격 검사지가 있다. 이러한 Big 5 모델은 다양한 자료에서 신뢰성과 타당성을 가진다[7].

Ⅲ. 연구 방법

본 연구에서는 문제 인식 및 해결을 위한 데이터 마이닝 프로세스 명료화와 데이터 탐색을 이해하는 것에 도움을 주고, 타당한 일관성과 반복 가능성, 객관성을 필요로 하기 때문에 데이터 마이닝을 위한 교차 산업 표준 절차(Cross Industry Standard Process for Data Mining, CRISP-DM)를 기반으로 수행하였다[8]. 해당 방법론 모델의 도식은 그림 1과 같다.

첫 번째는 업무 이해로, 프로젝트 및 연구목적의 이해와 데이터 마이닝 문제 정의를 수행하는 단계이다. 본 논문에서는 설문조사를 진행하기 전 Big 5에 대해 분류 및 분석하고, 창의적인 사람들의 특징과 성격을 일반화하여 분류 및 분석하는 과정에 해당한다.

두 번째 데이터 이해는, 초기 데이터 수집을 시작으로 데이터를 파악하기 위한 활동, 데이터의 품질 확인, 데이터에서 통찰력 발견 및 숨겨진 정보 탐색을 모두 포함한다. 본 논문에서는 데이터를 수집하기 위한 설문조사 준비 및 정보 탐색하는 과정에 해당한다.

세 번째 데이터 준비 단계에서는 원본 데이터부터 최종 데이터 세트를 구성하기 위한 모든 과정을 의미한다. 이때, 데이터 변환 및 정제, 속성등의 과정이 포함된다.

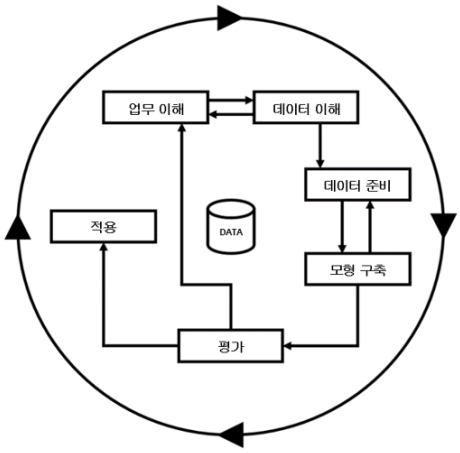

그림 1. 교차 산업 표준 절차 모델 도식
Fig. 1. Cross Industry Standard Process for Data Mining, CRISP-DM

본 논문에서는 시행한 Big5성격 특성에 기반한 설문조사의 원본 데이터를(Raw Data) 숫자로 변환하여 데이터마이닝 분석 툴인 WEKA의 분석에 사용할 데이;터의 준비 과정을 나타낸다.

네 번째는 모형 구축단계로, 모형 구축을 통하여 다양한 기법을 선택 및 적용하는 단계이다. 만약, 어떠한 알고리즘을 적용할 때 전처리 과정이 필요하다면 다시 데이터 준비단계로 회귀하여 다시 시행한다.

본 논문에서는 특징 선택으로 개미(Ant Algorithms), 최우선(Best First Algorithms), 뻐꾸기(Cuckoo Search), 코끼리(Elephant Search), 유전자(Genetic Algorithms), 탐욕(Greedy Step Wise Algorithms), 랭커(Ranker Algorithms) 알고리즘들로 특징 선택 알고리즘들을 사용하였으며, 특징 선택 알고리즘을 통해 도출된 결과들을 검증하기 위한 분류 방법으로는 머신 러닝 분야에서 널리 사용되는 나이브 베이즈 분류(Naive Bayes), 다층 퍼셉트론(Multilayer Perceptron, MLP), 서포트 벡터 머신(Support Vector Machine, SVM), 결정 트리(Decision Tree)를 활용하여 검증하였다.

다섯 번째는 평가 단계로, 데이터 분석을 통한 높은 품질을 갖는 모형을 구축하고, 최종 단계 이전에 평가하여 재검토하는 작업을 수행한다.

마지막 적용 과정에서는 지식 데이터를 체계화하고, 모든 과정들을 통하여 도출된 모형 및 결과를 제시한다.

Ⅳ. 실험 및 결과

1. 연구 내용

이전 연구에서는 미국 다트머스 대학교에서 개발한 “Student Life” 애플리케이션을 통해 정신건강, 학업 성취도를 평가하기 위해 10주동안 48명의 학생에게 수집된 데이터를 바탕으로 연구를 진행했다. 본 연구에서는 수원 아주대학교 IT 전공학생들 580명을 기준으로 설문조사를 진행하였다. 설문조사를 완료한 데이터들을 분류 및 분석하기 위하여 데이터 세트화 하였고, 그 데이터 세트에서, 창의적인 사람임을 나타내는 속성을 클래스 레이블로 선정하고, 데이터 세트의 나머지 속성을 인자로 설정하여 클래스 레이블에 어떤 영향을 미치는지 분석하였다.

먼저 데이터 마이닝 도구인 WEKA를 통해 데이터 세트를 분석하고 창의적인 사람들에 대한 특징을 관찰하고 상대적으로 클래스 레이블과 관련이 높고 우선순위가 높게 판별된 속성을 선별한다. 해당 속성은 새로운 데이터 세트 생성 작업의 일환으로 우선순위가 높은 속성만을 선택하여 새로운 데이터 세트로 정교화 한다.

데이터 마이닝 작업에서 첫 단계인 데이터 세트 분석 결과, Big 5 성격 특성 데이터 세트 내 “Is original, comes up with new ideas”라는 속성을 클래스 레이블로 정의한다. 다음으로, 이 클래스 레이블에 대하여 각 데이터를 나열하고, 43개의 나머지 속성에서 주요한 속성만을 선별해낸다. 이 과정에서 앞서 설명한 7개의 특징 선택 알고리즘을 활용하여 각각 10개에서 20개의 속성을 추출한 다음 개별적으로 새로운 데이터 세트를 생성한다.

다음으로 이렇게 만들어진 새로운 데이터 세트를 앞에서 설명한 3가지 알고리즘에 적용하여 10배수 교차검증(10-fold cross validation)을 통해 검증한다.

마지막으로 이에 대한 결과를 바탕으로 가장 좋은 결과를 나타내는 특징 선택 알고리즘을 선별한다.

2. 연구 결과

WEKA를 이용하여 Big 5 설문조사 기반으로 도출해낸 데이터 세트 내 “Is original, comes up with new ideas”라는 “나는 독창적이고, 새로운 것을 제안하길 좋아한다.” 창의적인 사람의 기준이 되는 것을 클래스 레이블을 기준으로 7가지 특징 선택 알고리즘의 결과들의 정확도를 나타낸 것은 표 1과 같다.

표 1. 특징 선택 알고리즘 정확도
Table 1. Feature Selection Algorithm Accuracy

	나이브 베이즈	다중 퍼셉트론	서포트 벡터 머신	결정 트리
개미	66%	63%	69%	61%
최우선	68%	65%	69%	62%
뻐꾸기	69%	66%	71%	61%
코끼리	66%	64%	70%	59%
유전자	65%	63%	68%	60%
탐욕	68%	65%	69%	62%
순위	62%	64%	70%	56%

4가지 알고리즘들로 수행하였으며 정확도는 10배수 교차검증을 통해 검증하였다. 결과적으로 이전 연구에서는 최고 우선 검색 알고리즘이 다층 퍼셉트론과 서포트 벡터 머신 두가지에서 가장 높은 정확도를 보였지만, 본 논문에서는 48명이 아닌 580명에게 설문조사를 시행하여 얻은 데이터를 가지고 실험하였기 때문에 결과가 다르게 도출되었다. 다층 퍼셉트론과 새로 추가한 결정 트리의 값이 가장 낮은 정확도를 나타내었고, 지난 연구에서도 높은 정확도를 보인 서포트 벡터 머신이 이번엔 가장 좋은 정확도를 보인 것을 볼 수 있다. 다른 알고리즘 들에 비하여 정확한 것을 알 수 있다.

뻐꾸기 알고리즘을 통하여 선택된 특징 집단에 대한 선택 결과를 보여준다. 표에서 음영이 있는 부분은 창의적인 사람의 클래스 레이블과 연관 있는 특징을 의미하며 퍼센트는 얼마나 많은 경우에 연관성을 보여주었는지를 의미한다. 검증 과정에서 10배수 교차검증을 활용하였기에 전체 10개의 경우 중 최대 10개에서 최소 3개까지의 연관이 있었던 특징들을 선택하여 최종 특징 집단을 생성한 것을 나타낸 것은 표 2와 같다.

표 2. 뻐꾸기 알고리즘 결과
Table 2. Cuckoo Algorithm results

배수	속성	배수	속성
10(100%)	1 1	0(0%)	23 23
0(0%)	2 2	0(0%)	24 24
4(40%)	3 3	10(100%)	25 25
0(0%)	4 4	10(100%)	26 26
0(0%)	5 5	0(0%)	27 27
0(0%)	6 6	0(0%)	28 28
0(0%)	7 7	0(0%)	29 29
0(0%)	8 8	5(50%)	30 30
10(100%)	9 9	0(0%)	31 31

5(50%)	10 10	0(0%)	32 32
7(70%)	11 11	10(100%)	33 33
0(0%)	12 12	0(0%)	34 34
0(0%)	13 13	0(0%)	35 35
0(0%)	14 14	0(0%)	36 36
10(100%)	15 15	0(0%)	37 37
0(0%)	16 16	0(0%)	38 38
0(0%)	17 17	0(0%)	39 39
0(0%)	18 18	90(0%)	40 40
0(0%)	19 19	0(0%)	41 41
10(100%)	20 20	0(0%)	42 42
0(0%)	21 21	0(0%)	43 43
0(0%)	22 22	0(0%)	44 44

마지막으로 빼꾸기 알고리즘으로 도출해낸 창의적인 사람과 연관성이 높은 문항에 대한 설명과 번역이다. 총 12개로 구성되어 있으며, 창의적인 성격을 가진 사람들이 어떤 유사한 성격이 있는지에 대해서 알 수 있는 결과가 표 3에 나타나있다.

표 3. 결과와 연관성 있는 성격과 Big 5 결과
Table 3. Big 5 results with relevant personality

문항	내용	Big 5
1	나는 말하는 것을 좋아한다	E
3	나는 어떠한 일이건 꼼꼼하게 하려고 한다	C
9	나는 남에게 편안한 사람이고, 스트레스를 잘 처리한다	N
10	나는 호기심이 많은 사람이다	O
11	나는 에너지가 넘치는 사람이다	E
15	나는 깊은 생각을 자주 하는 사람이다	O
20	나는 활발한 상상력을 가진 사람이다.	O
25	나는 발명하는 것을 좋아하는 사람이다	O
26	나는 적극적인 성격을 가진 사람이다	E
30	나는 예술적, 미적인 경험을 가치있게 여기는 사람이다	O
33	나는 일을 효율적으로 하는 사람이다	C
40	나는 내 생각을 남들에게 말하는 것을 좋아한다	O

Big 5로 적용했을 때, Big 5 Inventory에 적용하여 문항들을 정해보았을 때, 표 3의 가장 우측에 나오는 결과를 나타낸다. OCEAN 중에 O가 가장 많이 나온 것을 확인 할 수 있다. 그 의미는 경험에 대한 개방성(Openness to experience). 즉, 개인의 심리 및 경험의 다양성과 관련된 것으로, 지능, 상상력, 고정관념의 타파, 심미적인 것에 대한 관심, 다양성에 대한 욕구, 품위 등과 관련된 특질을 포함한다.

결과적으로 WEKA를 통한 하는 창의적인 성격을 지닌 사람의 특성을 알수 있는 결과이다.

V. 결 론

본 논문에서는 이전 연구에서 실행한 공개된 데이터 세트가 아닌 직접 모은 아주대학교 IT전공 학생 580명을 대상으로 한 설문조사를 기반으로 얻은 데이터를 데이터 세트화 하여 진행하였고, 창의적인 사람임을 나타내는 속성을 (“나는 독창적이고, 새로운 것을 제안하길 좋아한다.”) 클래스 레이블로 지정하고, 나머지 속성들을 인자로 설정하여 분석하였으며, 다양한 특징 선택 알고리즘들을 통하여 특징 집단을 선택하고, 다양한 알고리즘들을 활용하여 정확도를 검증하였고, 그 결과 서포트 벡터 머신을 통해 나온 특징이 가장 높은 분류 결과를 도출하였다. 다른 분류기보다 성능이 좋은 이유는, 본 연구의 데이터 세트는 설문지에서 추출한, 다양한 데이터보다 서로 연관 있는 데이터를 속성으로 많이 포함하고 있는데, 이러한 특징이 성능 순위에 영향을 준 것이라고 판단 할 수 있다. 다양하거나 서로 연관이 없는 작은 데이터 세트이면 나이브 베이즈가 더 성능이 좋았을 것이고, 다양하고 연관이 없는 대규모 데이터 세트이면 다층 퍼셉트론이 가장 성능이 좋았을 것으로 예상한다.

그 후 설문지에 기본이 된 Big 5 성격 특성 기법을 이용하여 분류 및 분석에 이용하여, 유사한 성격들에 대한 분석을 하여 결과를 도출하였다.

향후 연구로는 Big 5 성격 특성 기법을 이용하여 분류 및 분석에 이용한 본 연구를 적용할 수 있는 어플리케이션 등 연결할 수 있는 시스템을 개발 및 분석할 계획이다. 또, 본 연구에서 적용한 머신 러닝 알고리즘들 뿐만 아니라 클러스터링 기법 등 추가적으로 다른 알고리즘들을 사용하여 다양한 결과를 분석할 수 있을 것이라 기대한다.

References

[1] Guilford, Joy Paul. "Three faces of intellect."

American psychologist 14.8 (1959): 469.
DOI: https://doi.org/10.1037/h0046827

[2] Yong-Jun Kim, Seok-Won Lee, "Feature Selection for Creative People Using Machine Learning Algorithms", The Journal of Korean Institute of Information Scientists and Engineers, 1033-1035, 2018.

[3] Chang Jae-Young, Lee Byung-Jun, Cho Se-Jin, Han Da-Hye, Lee Kyu-Hong, "Automatic Classification of Advertising Restaurant Blogs Using Machine Learning Techniques",The Journal of The Institute of Internet, Broadcasting and Communication v16, n2, p55 - 62, 2016.
DOI: https://doi.org/10.7236/JIIBC.2016.16.2.55

[4] Wanseop Kim, Soowon Lee. "An In-depth Survey Analysis Applying Data Mining Techniques." Korean Society for Engineering Education & Technol 9.4 71-82. 2006.

[5] Pratama, Bayu Yudha, and Riyanarto Sarno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM." Data and Software Engineering (ICoDSE), 2015 International Conference on. IEEE, 2015.

[6] Wikipedia [Online]. Available: https://en.wikipedia.org/wiki/Weka_(machine_learning)

[7] Wikipedia [Online]. Available: https://en.wikipedia.org/wiki/Big_Five_personality_traits

[8] Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. "O'Reilly Media, Inc.", 2013.

저자 소개

김 용 준(준회원)

- 2016.02 : 고려대학교 컴퓨터정보학과 (공학사)
- 2016 ~ 아주대학교 컴퓨터공학과 석사과정 재학
- 관심분야 : Data Science, Machine Learning, and Software Engineering