The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 19, No. 4, pp.169-174, Aug. 31, 2019. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2019.19.4.169
JIIBC 2019-4-23

머신러닝 알고리즘 분석 및 비교를 통한 Big-5 기반 성격 분석 연구

A Study on Big-5 based Personality Analysis through Analysis and Comparison of Machine Learning Algorithm

김용준*

Yong-Jun Kim*

요 약 본 연구에서는 설문지를 이용한 데이터 수집과 데이터 마이닝에서 클러스터링 기법으로 군집하여 지도학습을 이용하여 유사성을 판단하고, 성격들의 상관 관계의 적합성을 분석하기 위해 특징 추출 알고리즘들과 지도학습을 이용하는 것을 목표로 진행한다. 연구 수행은 설문조사를 진행 후 그 설문조사를 토대로 모인 데이터들을 정제하고, 오픈 소스 기반의 데이터 마이닝 도구인 WEKA의 클러스터링 기법들을 통해 데이터 세트를 분류하고 지도학습을 이용하여 유사성을 판단한다. 그리고 특징 추출 알고리즘들과 지도학습을 이용하여 성격에 대해 적합한 결과가 나오는지에 대한 적합성을 판단한다. 그 결과 유사성 판단에 가장 정확도 높게 도움을 주는 것은 EM 클러스터링으로 3개의 분류하고 Naïve Bayes 지도학습을 시킨 것이 가장 높은 유사성 분류 결과를 도출하였고, 적합성을 판단하는데 도움이 되도록 특징추출과 지도학습을 수행하였을 때, Big-5 각 성격마다 문항에 추가되고 삭제되는 것에 따라 정확도가 변하는 모습을 찾게 되었고, 각 성격 마다 차이에 대한 분석을 완료하였다.

Abstract In this study, I use surveillance data collection and data mining, clustered by clustering method, and use supervised learning to judge similarity. I aim to use feature extraction algorithms and supervised learning to analyze the suitability of the correlations of personality. After conducting the questionnaire survey, the researchers refine the collected data based on the questionnaire, classify the data sets through the clustering techniques of WEKA, an open source data mining tool, and judge similarity using supervised learning. I then use feature extraction algorithms and supervised learning to determine the suitability of the results for personality. As a result, it was found that the highest degree of similarity classification was obtained by EM classification and supervised learning by Naïve Bayes. The results of feature classification and supervised learning were found to be useful for judging fitness. I found that the accuracy of each Big-5 personality was changed according to the addition and deletion of the items, and analyzed the differences for each personality.

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

*준회원, 아주대학교 컴퓨터공학과
접수일자 2019년 6월 14일, 수정완료 2019년 7월 14일
게재확정일자 2019년 8월 2일

Received: 14 June, 2019 / Revised: 14 July, 2019 /
Accepted: 2 August, 2019
*Corresponding Author: yjkim615@ajou.ac.kr
Dept. of Computer Engineering, Ajou University, Korea

Ⅰ. 서 론

현재 많은 대학교에서는 신입생들과 취업을 나가길 원하는 학생들을 대상으로 설문조사를 통한 본인 성격에 대한 궁금증 해소 목적인 성격 테스트 및 진로에 대한 궁금증 해소를 위하여 적성 검사를 많이 시행하고 있다. 하지만 기존의 성격 테스트 및 적성 검사를 검증하거나 확인할 때에는 정해져 있는 하나의 분석 방법을 이용하거나, 전문가의 의견을 꼭 필요로 하는 추세의 검사가 대부분 이었다.

기존에 사람의 성격에 대한 수치화나 체계화된 분석이 없었기에 사람의 성격에 대해 정의하는 것이 어려웠다. 그리고 테스트에 참여한 학생들이 알 수 있는 기존 성격 분석에서는 각 문항과 테스트에 맞는 분석 결과만을 알 수 있었고, 전문가들이 어떠한 이유로 결과가 나온지에 대해서는 알 수 없고, 전문가의 의견만을 알 수 있었다. 본인 성격 외에 성격 분포에 대해 연관성과 유사성에 대해선 알 수 없었다.

이러한 문제를 해결하기 위해 데이터 마이닝을 이용한 연구를 진행하였으며 데이터 마이닝은 인공 지능 및 텍스트 마이닝과 결합 된 빅 데이터 분석 방법이다[1]. 본 연구에서는 기계학습과 특징 선택 알고리즘을 활용하고 지식베이스 기반 접근법이 아닌, 데이터 중심적 접근법을 사용하여 성격 분석을 할 수 있도록 특징 집단을 선택하고 심리학 분야로 국한되어 있는 것을 공학 분야의 기계학습과 인공지능 기법을 이용하여 유사성과 적합성을 판단하여, 전문가가 여러 가지 분석을 할 때 도움이 되도록 보조하는 보다 효율적인 알고리즘 및 접근 방식을 제안한다.

본 연구에서는 우선 Big-5 성격 검사 기반으로 만들어진 설문 테스트를 이용하여 설문지를 만들고, 그 후 학생들에게 설문조사를 진행했고, 설문조사를 이용하여 얻은 데이터들을 우선 데이터 세트화를 진행하여, 데이터 마이닝 도구인 WEKA로 수행할 수 있도록 작업을 진행한다. 그리고 WEKA를 이용하여 군집화 기법을 이용하여 우선 분류를 하고, 군집 된 데이터 세트를 1차 분석으로 특징 추출 알고리즘들을 이용하여 분석하고, 머신 러닝 기법을 이용하여 2차 분석을 진행한다. 2차 분석에 사용된 머신 러닝 기법은 지도학습과 특징 추출 알고리즘을 같이 사용하여 분석을 진행한다.

Ⅱ. 본 론

1. 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[2].

2. 지도학습

지도학습(Supervised Learning)은 훈련 데이터(Training Data)로부터 하나의 함수를 유추해내기 위한 기계 학습(Machine Learning)의 한 방법이다. 훈련 데이터는 일반적으로 입력 객체에 대한 속성을 벡터 형태로 포함하고 있으며 각각의 벡터에 대해 원하는 결과가 무엇인지 표시되어 있다. 이렇게 유추된 함수 중 연속적인 값을 출력하는 것을 회귀분석(Regression)이라 하고 주어진 입력 벡터가 어떤 종류의 값인지 표식하는 것을 분류(Classification)라 한다.

지도 학습기(Supervised Learner)가 하는 작업은 훈련 데이터로부터 주어진 데이터에 대해 예측하고자 하는 값을 올바로 추측해내는 것이다[3].

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[4].

Ⅲ. 연구 방법

기본 연구 진행의 토대는 교차 산업 표준 절차 방법론을 따른다.

먼저, 연구 절차 중, 비즈니스 이해 단계에서 본 논문의 서론에서 언급한 모든 내용을 토대로, Big-Five가 무엇인지에 대해 분석 후 데이터 마이닝으로 해결해야 하는 문제를 정의하고, 연구의 목적을 구분한다.

두 번째, 데이터 이해 과정에서는 설문 조사로 얻은 설문지 데이터를 분석한다. 해당 과정은 Big-Five 분석과 분류를 위한 기반 지식 및 데이터를 수집 및 추출하고 분석하여 이해하는 단계이다.

다음 과정은 데이터 준비로, 지식 기반의 접근을 통해 원본 데이터를 가용 데이터 세트로 새롭게 구성하기 위하여 행하는 모든 전처리 과정을 포함하며, 데이터 세트를 구축하여 지식화 작업을 수행한다. 설문지로 얻은 데이터를 데이터 마이닝의 클러스터링을 이용하여 군집화를 적용 및 정제하고, 특징 선택 알고리즘을 연구 및 적용하는 과정이다.

네 번째 단계는 전처리 과정을 통해 생성한 데이터 세트를 바탕으로, Big-Five에 적용 가능한 알고리즘을 연구한다. 나이브 베이즈와 다층 퍼셉트론 구조를 이용하여 분류작업을 한다. 이 때, 그 수행 결과를 비교 및 분석하고, 인과 관계를 파악한다.

다음, 평가 과정에서는 모델링을 통해 나온 자료들과 처음 비즈니스 단계에서 분석했던 Big-Five와의 비교 분석 및 테스팅을 검증하고 평가한다. 이를 통해 전 처리 방식과 그 결과, 분석 내용, 구축한 데이터 세트의 형식 등 전반적인 내용을 판단하여 유효성을 검증한다. 마지막, 전개 단계에서는 데이터 세트를 조금 더 체계화하고 가용 알고리즘을 선택하여 진단 및 처방에 적용하고, 사용되는 알고리즘의 개선과 테스트 및 검증을 지속적으로 수행한다.

다음 그림[1]은 연구 구조, 연구 전략, 그리고 연구 방법론을 통합하여 나타낸 것으로, 연구의 전반적인 절차 및 단계를 표현한다[5].

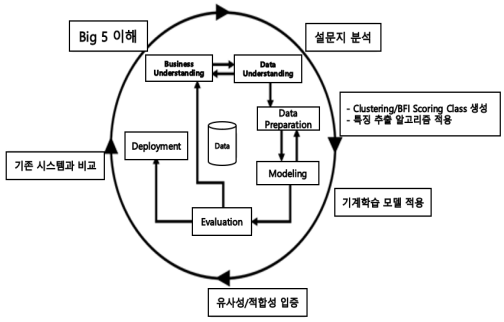


그림 1. 연구 절차 및 방법 도식
Fig. 1. Overview of research procedures and methods visualization.

Ⅳ. 실험 및 결과

1. 연구 내용

가. 유사성 판단

Big-5 성격 분석 기반으로 제작한 설문지를 이용하여 학생들에게 데이터를 수집하고, 데이터를 데이터 세트화 하여 준비한다. 그 후 클러스터링 군집화 방법을 이용하여 EM 클러스터링, X-means 클러스터링으로 Class Label을 생성하여 지도학습을 하기 위한 준비를 한다. 그 후 지도학습을 적용하기 위하여 인공지능과 데이터 마이닝 분야에서 가장 많이 사용되고, 머신 러닝에서 지도학습으로 가장 많이 사용되는 Naïve Bayes, MLP, SVM 을 이용하여 분류된 데이터 세트를 분석한다. 결과 중에 가장 높은 정확도와 정밀도를 확인하여 유사성을 판단하는 데에 도움이 되도록 클러스터링과 지도학습을 정해준다.

나. 적합성 판단

Big-5 성격 분석 기반으로 제작한 설문지를 이용하여 학생들에게 데이터를 수집하고, 데이터를 데이터 세트화 하여 준비한다. 그 후 BFI Scoring을 이용하여 분류하기 위한 데이터 세트 작업을 통하여 데이터 세트들을 정리한다. 그 후 특징 추출 알고리즘과 지도학습을 적용하고, 각 문항과 결과의 연결성을 확인하여 각 문항의 성격이 적합한 지에 대한 판단에 도움이 되도록 분석한다.

2. 연구 결과

가. 유사성 판단

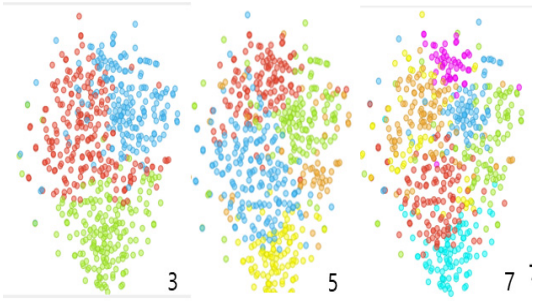


그림 2. 군집화 성공 시각화
Fig. 2. Clustering success visualization.

클러스터링을 이용하여 유사성을 예측하기 위하여 데이터 세트를 시각화 하였다. 홀수로 진행한 이유는 EM 클러스터링에서 아무런 작업을 하지 않은 기준으로 수행하였을 때, 7개의 클러스터링으로 분류가 최적화된 것으로 나왔기 때문에, EM 클러스터링과 X-means 클러스터링 두 개 다 7을 기준으로 클러스터링을 진행하였고, 그렇기에 홀수로 분류를 진행하였다. 하지만, 그림[2]에서 나타나듯, 3개와 5개, 7개로 분류한 그림에서는 구역이 나뉘는 것이 확실히 나타나는 것을 확인할 수 있다.

표 1. EM, X-means 결과
Table 1. EM, X-means results.

EM		K-means	
0	186 (32%)	0	182 (31%)
1	160 (28%)	1	191 (33%)
2	233 (40%)	2	206 (36%)
0	114 (20%)	0	113 (20%)
1	94 (16%)	1	135 (23%)
2	105 (18%)	2	114 (20%)
3	141 (24%)	3	125 (22%)
4	125 (22%)	4	92 (16%)
0	72 (12%)	0	65 (11%)
1	51 (9%)	1	45 (8%)
2	40 (7%)	2	123 (21%)
3	103 (18%)	3	127 (22%)
4	106 (18%)	4	82 (14%)
5	100 (17%)	5	79 (14%)
6	107 (18%)	6	58 (10%)

표[1]에서 확인할 수 있듯이. 7개의 분류를 기준으로 작업을 진행하였다. 표에서 나타나는 결과로 확인했을 때, 확실히 EM에 비해 X-means로 작업한 것이 결측 값이 발생한 개체와 가장 가까운 거리에 있는 K개의 이웃 개체와 거리를 계산하여 다수결로 결측 값을 대체하는 방법이기 때문에 7개의 분류에서 확실하게 중앙 집중형 구조를 보이는 것을 확인할 수 있다. EM 클러스터링을 이용하여 분류 한 결과 중 가장 높은 정확도를 보인 것은 각 분류 개수 마다 Naïve Bayes가 가장 높은 정확도를 보였고, 특히 3개로 분류하고 Naïve Bayes를 이용하여 작업한 것이 가장 높은 정확도를 보였다. 이 결과로 EM 클러스터링에서는 3개로 분류하고 지도학습은 Naïve Bayes를 이용하여 유사성을 분석하는데 도움을 주는 것으로 판단할 수 있다.

X-means를 이용하여 분류 한 결과 중 가장 높은 정확도를 보인 것은 이 또한 3개로 분류하고, 지도학습으로는 Naïve Bayes와 MLP를 이용하여 분류한 것이 가장 좋은 결과를 보였다. 이 결과로 X-means 클러스터링에서는 3개로 분류하고, 지도학습으로는 Naïve Bayes와 MLP를 이용하여 유사성을 분석하는데 도움을 주는 것으로 판단할 수 있다.

더 정확히 확인해 본다면, 정확도와 kappa 일치도, 정밀도를 전체적으로 보았을 때, 가장 높은 정확도를 확인할 수 있는 것은 EM 클러스터링을 이용하여 분류하고 Naïve Bayes를 이용한 결과가 95.6822라는 높은 정확도를 보이면서 가장 좋은 결과라는 것을 볼 수 있고, 그 결과로 전체 설문지 데이터 세트의 유사성 분석에 도움을 줄 수 있는 것으로 판단할 수 있다.

나. 적합성 판단

(Big-5 성격분석 중 OCEAN에서 E로 분석한 것을 결과에 넣어 설명하겠다.) 표[2]에서 확인할 수 있듯이, 설문지 항목에 따라 분석한 E성격에 해당하는 항목 중 1 (나는 말하는 것을 좋아한다.), 4 (나는 우울한 사람이 아니라고 생각한다.), 6 (나는 내성적이지 않은 사람이라 생각한다.), 11 (나는 에너지가 넘치는 사람이다.), 16 (나는 열정이 많은 사람이다.), 21 (나는 조용한 경향이 없는 사람이다.), 23 (나는 게으르게 보이지 않는 사람이라 생각한다.), 26 (나는 적극적인 성격을 가진 사람이다.), 27 (나는 차갑고, 냉담한 사람이라 생각하지 않는다.), 30 (나는 예술적, 미적인 경험을 가치 있게 여기는 사람이다.), 31 (나는 억압받는 사람이라 생각하지 않는다.), 34 (나는 긴장된 상황에서 침착함을 잘 유지한다.), 36 (나는 외향적이고, 사교성이 좋은 사람이다.), 42 (나는 다른 사람들과 협력하는 것을 좋아한다.)

Pratama의 연구[6]에서 이용한 나이브 베이즈, K-최근접 이웃 알고리즘 및 서포트 벡터 머신 사용 이 3가지를 사용한 것을 기반으로 분류 하면 표[3]에서 확인할 수 있듯이, Naive Bayes에서 개미 탐색 알고리즘과 최우선 탐색 알고리즘, 탐욕 알고리즘 3개의 결과가 같게 나왔

표 2. Big-5 중 E 성격에 관한 결과
Table 2. Results for Big-5 E personality

Class E		Correctly	Incorrectly	kappa 통계치
Naivebayes	Ant	92.5734	7.4266	0.8516
	BestFirst	92.5734	7.4266	0.8516
	Cuckoo	91.8826	8.1174	0.8389
	Genetic	91.019	8.981	0.8228
	GSW	92.5734	7.4266	0.8516
MLP	Ant	96.7185	3.2815	0.931
	BestFirst	94.9914	5.0086	0.893
	Cuckoo	94.3005	5.6995	0.8768
	Genetic	95.8549	4.1451	0.9118
	GSW	94.9914	5.0086	0.893
SVM	Ant	96.2003	3.7997	0.9201
	BestFirst	95.8549	4.1451	0.9128
	Cuckoo	96.0276	3.9724	0.9163
	Genetic	93.9551	6.0449	0.8712
	GSW	95.8549	4.1451	0.9128

고, 가장 높은 정확률을 보이고 있다. 뻐꾸기 탐색에서 20번 문항인 "나는 활발한 상상력을 가진 사람이라 생각한다." 이것과 40번 문항인 "나는 내 생각을 남들에게 말하는 것을 좋아한다."이 추가되어 정확도가 낮은 것으로 판단. 유전 탐색 알고리즘에서는 5번 문항인 "나는 독창적이고, 새로운 것을 제안하길 좋아한다." 이것과 10번 문항인 "나는 호기심이 많은 사람이다.", 24번 문항인 " 나는 정서적으로 안정되어 쉽게 화를 내지 않는다."이 추가되었고, 30번 문항인 "나는 예술적, 미적인 경험을 가치 있게 여기는 사람이다." 이것이 삭제되어 정확도가 낮은 것으로 판단. MLP와 SVM 경우에는 개미탐색 알고리즘으로 하였을 때 가장 높은 정확률을 확인할 수 있다. 최우선 탐색 알고리즘에서는 10번 문항인 "나는 호기심이 많은 사람이다." 이것이 추가되어 정확도가 낮은 것으로 판단. 뻐꾸기 탐색 알고리즘에서는 20번 문항인 "나는 활발한 상상력을 가진 사람이라 생각한다." 이것과 40번 문항인 "나는 내 생각을 남들에게 말하는 것을 좋아한다."이 추가되었고, 30번 문항인 "나는 예술적, 미적인 경험을 가치 있게 여기는 사람이다."이것이 삭제되어 정확도가 낮은 것으로 판단. 유전 탐색 알고리즘에서는 5번 문항인 "나는 독창적이고, 새로운 것을 제안하길 좋아한다." 이것과 24번 문항인 "나는 정서적으로 안정되어 쉽게 화를 내지 않는다."이 추가되었고, 30번 문항인 "나는 예술적, 미적인 경험을 가치 있게 여기는 사람이다." 이것이 삭제되어 정확도가 낮은 것으로 판단. 탐욕 알고리즘에서는 10번 문항인 "나는 호기심이 많은 사람이다." 이것이 추가되어 정확도가 낮은 것으로 판단한다.

표 3. 같게 나온 결과 확인
Table 3. Confirm the same result.

Class E	Ant Search		BestFirst	
	Number of fold(%)	Attribute	Number of fold(%)	Attribute
	10 (100%)	1 1	10 (100%)	1 1
	0 (0%)	2 2	0 (0%)	2 2
	0 (0%)	3 3	0 (0%)	3 3
	7 (70%)	4 4	7 (70%)	4 4
	0 (0%)	5 5	0 (0%)	5 5
	10 (100%)	6 6	10 (100%)	6 6
	0 (0%)	7 7	0 (0%)	7 7
	0 (0%)	8 8	0 (0%)	8 8
	1 (10%)	9 9	1 (10%)	9 9
	0 (0%)	10 10	0 (0%)	10 10
	10 (100%)	11 11	10 (100%)	11 11
	0 (0%)	12 12	0 (0%)	12 12
	0 (0%)	13 13	0 (0%)	13 13
	0 (0%)	14 14	0 (0%)	14 14
	0 (0%)	15 15	0 (0%)	15 15
	10 (100%)	16 16	10 (100%)	16 16
	0 (0%)	17 17	0 (0%)	17 17

V. 결 론

본 연구는 실제 설문 조사를 통해 얻은 설문지의 기록을 기반으로, 인공지능의 데이터 마이닝 기법을 적용하기 위한 높은 수준의 데이터 세트를 생성하고, 객관적인 평가 기준을 바탕으로 기계학습 알고리즘을 분류, 비교하였다. 이를 통해 Big-Five를 보다 정확한 예측하는 분류기를 선별하였고, 주요한 지표가 되는 5가지 성격들로 분류 분석 검증하였다. 전문적인 지식이 아닌 비지도 학습을 이용한 설문 조사를 한 사람들의 특성을 유사도를 통해 Big-Five로 나누어지는 것을 유추하였다. BFI Scoring을 이용하여 특징추출 알고리즘인 뻐꾸기 탐색으로 인해 나오는 결과가 적합성을 판단하는데 가장 높은 정확도를 보였고, 군집화의 EM 클러스터링 기법과 지도학습의 Naïve Bayes로 유사성을 판단하는데 가장 높은 정확도를 보인 것을 판단하였다. 심리학 전문가들 개인마다 의견과 관점이 여러 가지로 나뉠 때, 판단에 도움을 주도록 각 성격 마다 적합성을 판단하는 것에는 BFI Scoring을 이용하여 데이터 세트를 정리하고, 지도학습과 특징추출 알고리즘을 이용하여 각 성격의 적합성을 판단하는 것이 가장 높은 정확도를 보이는 것으로 적합하게 판단한다. 이를 통해, 본 연구의 궁극적인 의미를 간략하게 설명하면 다음과 같다. 앞서 다룬 연구의 필요성과 목적을 만족시키고 달성하기 위하여 심리학 분야인 Big-Five를 위해 데이터 마이닝 연구 방법론 및 지도학습 프로세스를 접목하였다는 것이다. 본 연구의 한계점은 크게 세 가지로 구분된다. 첫 번째는 어느 특정한 집단인 대학교 학생들만 가지고 데이터 화 시켜서 분류 및 분석을 했기 때문에

편향되어 있을 수도 있다. 특정 세대를 통해 성격을 분석을 했지만, 이것을 가지고서는 다른 세대와 다른 집단의 성격을 분석하거나 정의하는 것에 대해서는 구체적인 데이터를 추출하기 어렵기 때문에 신뢰성을 보장할 수 없다. 또한, 두 번째는 데이터 마이닝 기법을 바로 적용하는 데에 데이터의 형식과 일관성 면에서 적합하지 않기 때문에 발전 연구를 진행하기 위하여는 상당히 오랜 시간의 반복적인 분석 과정, 다양한 전처리, 그리고 철저하고 치밀한 데이터 세트 가공 작업이 필요하다는 것이다. 마지막, 본 실험으로 사용한 분류기는 관련 연구에서 가장 많이 사용한 세 가지를 선정하여 적용하였다. 그러나 추후 비교 연구 시, 다양한 설정과 선택 요소가 존재하기 때문에, 같은 분류기를 사용할지라도 다른 프로젝트 및 연구 결과와 차이가 있을 수 있다. 향후 관련 연구는 본 연구 한계점을 극복하는 방향으로 진행할 수 있다. 먼저, 데이터 세트를 보강하기 위하여 그래프 또는 관련된 다른 데이터 값을 추가할 수 있다. 이는 그래프에 대한 직접적인 분석 결과와 수치를 추출하여, 독립된 속성으로써 값을 삽입하는 것이다. 이러한 작업과 연구를 통해, 심리학자가 지표로 활용하는 가용 Big 5 기준의 항목과 점차 동일하게 데이터 세트를 생성할 수 있다. 또한, 현재 주어진 데이터 세트의 테스트 확인 기록에 대하여, 보다 더 복잡한 알고리즘을 통해 자연어 처리를 수행하여 세밀하게 속성 및 수치를 부여할 수 있다. 마지막으로, 본 연구에서 적용한 다양한 데이터 마이닝 알고리즘이 아닌 또 다른 알고리즘을 탐구 및 사용하여 다른 결과를 이끌어 낼 수 있을 것이다. Big-Five를 통한 성격분석을 위한 설문지뿐만 아니라, PHQ-9 및 여러 설문 방법들을 더 이용하여 확장성을 고려할 것이다. 적용한 다양한 머신 러닝 알고리즘들과 특징 선택 알고리즘이 아닌 또 다른 알고리즘들을 추가적으로 사용하여 다양한 결과를 분석할 계획이다.

References

[1] Ko, Sujeong. “Big Data Analysis in School Adjustment Factors Using Data Mining.” International Journal of Advanced Smart Convergence, vol. 8, no. 1, 한국인터넷방송통신학회,Mar.2019,pp.87-97, doi:10.7236/IJASC.2019.8.1.87.

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

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

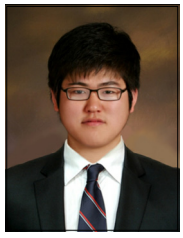
[4] Wikipedia [Online]. Available: https://en.wikipedia.org/wiki/Weka_(machine_learning)

[5] 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.

[6] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174. doi: 10.1109/ICODSE.2015.7436992

저 자 소 개

김 용 준(준회원)



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