Clustering Analysis

DASS(Depression Anxiety Stress Scales Responses) 분석

데이터마이닝이론및응용
에이쁠원하조

2020147024 김우영
2020251009 김혜리
2020195053 노가원
2020147018 조윤영

목차

연구 목적 및 데이터

- 연구 목적

Depression Anxiety Stress Scales (DASS) dataset은 우울증, 불안, 스트레스에 대한 설문조사 데이터이다. PCA/FA를 이용해 분석 및 정리한 DASS data를 clustering을 통해 군집화함으로써 피설문자들을 특징에 따라 분류하고자 한다. 이와 같은 군집 분석은 군집별로 더 세분화된 2차 설문조사나 연구에 이용되거나, 군집별 특성에 따른 심리치료 등에 활용될 것으로 기대할 수 있다.

- 데이터

저번 주차에서 DASS dataset을 대상으로 FA(수직회전)한 결과 데이터를 사용하였으며,
이 때, Factor은 “motivation, physical, emotional”이다.

다만, 데이터의 개수가 총 39775개로 clustering을 확인하기에는 개수가 다소 많아,
그 중 임의로 3000개의 데이터를 추출하여 분석을 진행했다.

```
[6]  import random
     random.seed(123)
     drop_random = list(range(39775))
     for i in range(3000):
       drop_choice = random.choice(drop_random)
       drop_random.remove(drop_choice)
     for i in drop_random:
       data = data.drop(i,axis=0)
```

	motivation	physical	emotional
7	-1.066920	-0.482416	-0.965835
22	1.107883	1.142016	-0.167817
23	-0.531137	-0.733275	1.141257
53	-1.213510	1.319785	0.464762
64	-0.398582	-0.909878	0.667486
...	...	...	...
39734	1.275523	-0.158417	1.706220
39739	1.385161	0.531276	0.883788
39757	1.346693	-1.557512	0.200226
39763	-0.043217	-0.056196	-0.328070
39764	-0.118283	-1.102917	-0.908024

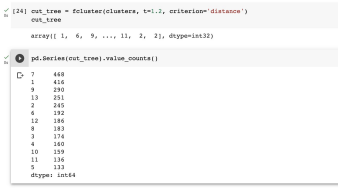
3000 rows × 3 columns

Hierarchical clustering

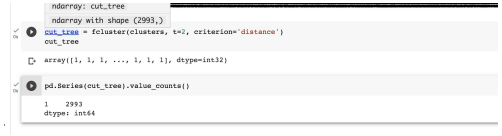
1. Complete linkeage(non-euclidean-cosine)

step1. 군집 수 결정(t값 결정)

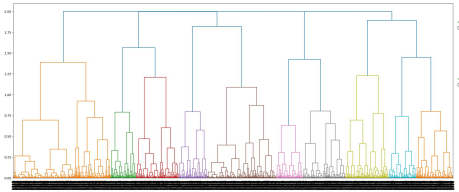
t=0.5일 때 너무 많은 cluster이 생성

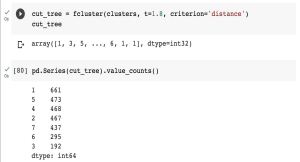


t=1.2일 때 0.5보다는 적지만
여전히 너무 많은 cluster이 생성



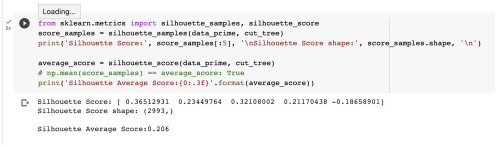
t=2일 때 유의미한 결과가 나오지 않음



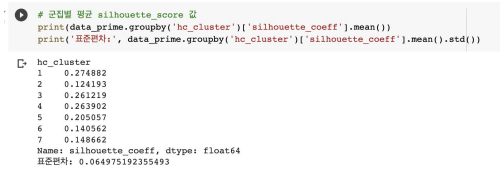


cophenetic distance를 고려하였을 때 t=1.8 에서 값이 급격이 바뀐다고 판단.
따라서 **t값을 1.8로 설정**

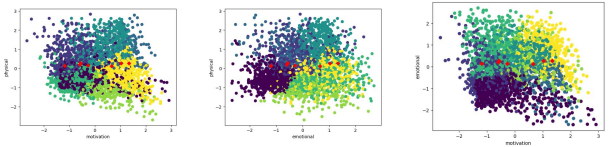
step2. Evaluation



실루엣 계수는0.206으로
No substantial structure has been found라고 볼 수 있다



1 0.274882 The structure is weak and could be artificial
2 0.124193 No substantial structure has been found
3 0.261219 The structure is weak and could be artificial
4 0.263902 The structure is weak and could be artificial
5 0.205057 No substantial structure has been found
6 0.140562 No substantial structure has been found
7 0.148662 No substantial structure has been found에 해당함으로
클러스터링이 잘 군집화 되었다고 보기 어렵다. 그나마 cluster 1이 0.27수준으로 그중에서는 군집이 잘 되었다. 다만 편차가 0.064수준으로 편차가 적은 것을 보아 군집화가 잘 진행되었다는 것을 보여준다고 볼 수 있다.



Motivation, Physical, Emotional 3개의 성분을 이용해 그린 2차원 플롯

step3. EDA/Interpretation

hc_cluster	motivation	physical	emotional	silhouette_coeff
1	0.863136	0.434421	0.472236	0.182307
2	0.512564	0.689313	0.691987	0.159349
3	0.548054	0.621556	0.621365	0.141867
4	0.589269	0.617922	0.430851	0.158498
5	0.467665	0.472366	0.821267	0.148259
6	0.637748	0.451565	0.768142	0.141017
7	0.532122	0.583516	0.720935	0.147860

Cluster 1: 동기부여가 낮고, 신체적 문제가 적으며 감정 기복이 매우 심하지 않은 사람
Cluster 2: 동기부여가 낮고 신체적 문제가 있으며 감정 기복이 그다지 심하지 않은 사람
Cluster 3: 동기부여가 높고 신체적 문제 또한 많으며 감정 기복이 매우 심하지 않은 사람
Cluster 4: 동기부여가 그다지 높지 않고, 신체적 문제가 매우 많으며 감정 기복이 다소 심한 사람
Cluster 5: 동기부여가 아주 낮고 신체적 문제가 없으며 감정 기복이 매우 심한 사람
Cluster 6: 동기부여가 다소 높고 신체적 문제가 거의 없으며 감정 기복이 심하지 않은 사람
Cluster 7: 동기부여가 아주 높고 신체적 문제가 거의 없으며 감정 기복이 다소 있는 사람

Hierarchical clustering
2. Ward linkeage(euclidean)

step1. 군집 수 결정(t값 결정)

```
cut_tree = fcluster(clusters, t=10, criterion='distance')
cut_tree

array([ 5, 13, 21, ..., 16,  2,  1], dtype=int32)

pd.Series(cut_tree).value_counts()

5     298
3     237
11    209
1     202
16    166
21    164
2     162
8     162
9     161
17    143
6     142
19    128
13    120
14    116
20    112
```

t=10일 때 너무 많은 cluster이 생성

```
cut_tree = fcluster(clusters, t=40, criterion='distance')
cut_tree

array([1, 2, 3, ..., 3, 1, 1], dtype=int32)

pd.Series(cut_tree).value_counts()

5    732
2    661
3    660
4    473
1    467
dtype: int64
```

t=20일 때 10보다는 적지만 더욱 유의미한 clustering이 가능할 것이라고 생각 됨

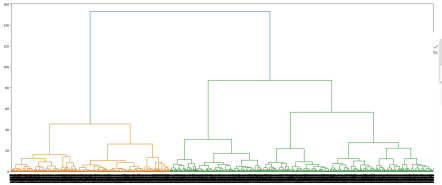
```
cut_tree = fcluster(clusters, t=60, criterion='distance')
cut_tree

array([1, 2, 2, ..., 2, 1, 1], dtype=int32)

pd.Series(cut_tree).value_counts()

2    1989
1    1004
dtype: int64
```

t=60일 때 유의미한 결과가 나오지 않음



```
cut_tree = fcluster(clusters, t=40, criterion='distance')
cut_tree

array([1, 2, 3, ..., 3, 1, 1], dtype=int32)

pd.Series(cut_tree).value_counts()

5    732
2    661
3    660
4    473
1    467
dtype: int64
```

cophenetic distance를 고려하였을 때 t=40 에서 값이 급격이 바뀐다고 판단.
t=20일 때보다 더 좋은 clustering을 제공. **따라서 t값을 40로 설정**

step2. Evaluation

```
from sklearn.metrics import silhouette_samples, silhouette_score
score_samples = silhouette_samples(data_prime, cut_tree)
print('Silhouette Score:', score_samples[:5], '\nSilhouette Score shape:', score_samples.shape, '\n')

average_score = silhouette_score(data_prime, cut_tree)
# np.mean(score_samples) == average_score: True
print('Silhouette Average Score:{0:.3f}'.format(average_score))

Silhouette Score: [ 0.46383682  0.40159063  0.53838022  0.46843616 -0.1784919 ]
Silhouette Score shape: (2993,)

Silhouette Average Score:0.399
```

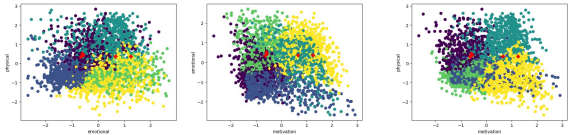
실루엣 계수는 0.399으로
the structure is week and could be artificial에 해당한다고
볼 수 있다

```
# 군집별 평균 silhouette_score 값
print(data_prime.groupby('hc_cluster')['silhouette_coeff'].mean())
print('표준편차:', data_prime.groupby('hc_cluster')['silhouette_coeff'].mean().std())

hc_cluster
1    0.364762
2    0.433927
3    0.369271
4    0.507388
5    0.345505
Name: silhouette_coeff, dtype: float64
표준편차: 0.06661806070228969
```

1 0.364762 The structure is weak and could be artificial
2 0.433927 The structure is weak and could be artificial
3 0.369271 The structure is weak and could be artificial
4 0.507388 The structure is weak and could be artificial
5 0.345505 The structure is weak and could be artificial

각 군집의 평균 실루엣 계수가 0.3~0.5인것으로 보아 아주 잘 군집화 된 특정 군집이 보이지 않고 있다. 그나마 cluster 4가 0.50으로 다른 clusters에 비해 군집이 잘 되었다고 볼 수 있다. 다만, 각 문집의 평균값의 편차가 0.66으로적은것을 보아 적절히 군집화가 진행되었다고 볼 수 있다.



Motivation, Physical, Emotional 3개의 성분을 이용해 그린 2차원 플롯

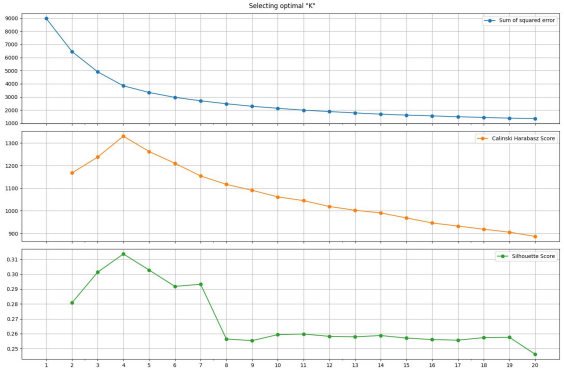
step3. EDA/Interpretation

hc_cluster	motivation	physical	emotional	silhouette_coeff
1	-0.937546	0.657896	-0.222973	0.364762
2	-0.367462	-0.567966	-1.192029	0.433927
3	0.624053	1.249414	0.279436	0.369271
4	-0.830384	-0.599195	0.707451	0.507388
5	0.903857	-0.646184	0.509573	0.345505

Cluster 1: 동기부여가 매우 낮고, 신체적 문제가 적으며 감정 기복이 심하지 않은 사람
Cluster 2: 동기부여가 낮고 신체적 문제가 적으며 감정 기복이 매우 심하지 않은 사람
Cluster 3: 동기부여가 낮지 않으며 신체적 문제가 많으고 감정 기복이 매우 심하지 않은 사람
Cluster 4: 동기부여가 매우 낮고 신체적 문제가 매우 적으며 감정 기복이 아주 심한 사람
Cluster 5: 동기부여가 매우 높고 신체적 문제가 없으며 감정 기복이 어느정도 있는 사람

K-means Clustering

step1.군집 수 결정



Elbow method를 사용했을 때 그래프는 한눈에 군집수를 결정하기 어렵지만, Silhouette score, Calinski Harabasz score 그래프에서는 모두 군집수가 4일때 값이 가장 크게 나타나므로 군집수를 최종적으로 4로 정한다.

step2. K-means clustering

centroids

	motivation	physical	emotional
0	0.396454	1.346348	0.242703
1	-0.789847	-0.415461	-0.799223
2	-0.336117	-0.332243	1.126591
3	1.245174	-0.668901	-0.222656

군집별 centroids 좌표값

```
# Observation and cluster
data_prime['cluster'] = pred
data_prime
```

	motivation	physical	emotional	cluster
0	-1.130394	-0.551188	-1.060694	1
1	1.175179	1.238131	-0.175303	0
2	-0.562395	-0.827510	1.277099	2
3	-1.285799	1.433943	0.526536	0
4	-0.421869	-1.022040	0.751456	2
...	...	...	...	...
2988	1.352899	-0.194302	1.903919	2
2989	1.469130	0.565398	0.991440	0
2990	1.428349	-1.735410	0.233037	3
2991	-0.045136	-0.081705	-0.353101	1
2992	-0.124716	-1.234673	-0.996554	1

2993 rows × 4 columns

각 관측치와 해당되는 군집

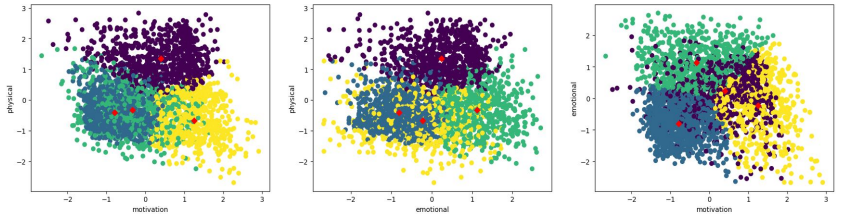
step3. Evaluation

```
Average Silhouette Score:0.477
```

클러스터의 average silhouette score가 0.477로 크게 높지 않음을 확인 할 수 있다

```
k_means_cluster
0    0.476731
1    0.551366
2    0.407482
3    0.433578
Name: k_silhouette_coeff, dtype: float64
표준편차: 0.06290605551590168
```

각 군집의 평균 silhouette coefficient값이 약 0.4-0.5인 것을 미루어 보아 군집들중에 아주 잘 군집화 된 특정 군집이 보이지지진 않았고, 각 군집이이 아주 좋게 군집화되진 않았음을 알 수 있다. 그나마 cluster2가 0.55로 4개의 군집들중 가장 잘 군집화 되었다. 다만, 각 군집들의 실루엣 계수 평균값의 편차가 작은걸보아아 적절히 군집화가 진행됐다고 보아도 좋겠다.



시각화를 한 결과 위에 내린 결론과 같이 보여짐을 알 수 있다

step4. EDA/Interpretation

```
# 군집별 통계
data_prime.groupby('k_means_cluster').mean()
```

k_means_cluster	motivation	physical	emotional	cluster	k_silhouette_coeff
0	0.396454	1.346348	0.242703	0.0	0.476731
1	-0.789847	-0.415461	-0.799223	1.0	0.551366
2	-0.331846	-0.334403	1.127200	2.0	0.407482
3	1.245753	-0.667572	-0.228088	3.0	0.433578

Cluster0: 신체적으로 특징이 나타나는 사람

Cluster1: 동기부여도 없으며 감정조절에 문제가 있는 사람

Cluster2: 감정조절에 문제가 있는 사람

Cluster3: 동기부여가 안되는 사람

clustering 최종비교

	Complete linkeage	Ward linkeage	K-means Clustering
(1) 전체 실루엣 계수의 평균값	0.206	0.399	**0.477**
(2) 개별 클러스터의 실루엣 계수 평균값의 표준편차 (소수점 아래 넷째자리에서 반올림)	0.065	0.067	**0.063**

(1) '전체 실루엣 계수의 평균값'의 경우 K-means clustering이 가장 높다.
(2) '개별 클러스터의 실루엣 계수 평균값의 표준편차'의 경우 세 가지 방법 모두 비슷하지만 K-means clustering이 아주 조금 더 낮다.

따라서
(1) '전체 실루엣 계수의 평균값'은 높을수록 좋고,
(2)'개별 클러스터의 실루엣 계수 평균값의 표준편차'은 낮을수록 좋으므로

최종적으로 K-means clustering을 채택한다.

한계점 및 개선점

- 한계점

우리는 기존 40000개에 가깝던 데이터에서 3000개만을 추출해서 사용했는데, 랜덤으로 추출하기는 했지만 그 과정에서 데이터가 왜곡되었을 확률도 존재한다. 또한 K-mean 클러스터링 결과 강하게 군집화를 보이는 그룹이 없었던 것이 이 연구의 가장 큰 한계이다. 우리 연구의 목적은 설문 응답자들이 겪고 있는 증상을 통해서 응답자들의 감정 상태나 증상을 통해 응답자들의 유형을 분석하는 것인데, 강한 군집화를 보이지 않아 결과 분석에 우리 연구자의 주관적인 해석이 크게 작용할 수 있다.

- 개선점

40000개의 데이터를 모두 활용해서 군집화를 진행한다면 더욱 신뢰할 수 있는 데이터를 얻을 수 있다. 또한 현재 연구에서는 군집 사이의 이질성이 강하지 않아서 시각적으로 명확히 군집이 구별된다고 해석하기 힘들다. 따라서 연구에 사용된 데이터를 다른 응답자들을 대상으로 수집해보는 것도 하나의 방법이 될 수 있다. 결과를 장담할 수는 없지만 우리 연구의 목적에 맞게 우울증 증상을 가진 환자들을 대상으로 데이터를 수집하면 강한 군집을 관찰할 가능성도 있다.

Reference

LUCAS GREENWELL. "Data Set." LUCAS GREENWELL