[A4] Clustering and Feature Engineering

2021312882 이준서

1. 데이터셋 설명 (data point 개수, feature 개수)

(1) 세 종류의 레이블이 있고, data point의 개수는 210개, float 속성 feature 7개를 가지고 있습니다.

2. 다양한 데이터 전처리 방법 비교

(1) log(x+1) transform: 피처의 분포를 확인하여, 오른쪽으로 꼬리가 있는 0,3,5,6 피처에 대해 데이터 전처리 수행

(2) 정규화: 모든 피처에 Standard Scaling을 이용하여 데이터 전처리 수행

(3) Polynomial Features & Interactions: 7개의 피처에 degree=2 적용, 28개의 피처 생성

(4) PCA를 이용한 feature extraction: 전체 분산의 95%를 설명하는 주성분 이용

(5) Filter, Wrapper, Embedded Method를 이용한 feature selection: Filter에서는 mutual_info_classif와 f_classif를 사용하여 비교, Wrapper 에서는 logistic regression을 피처 선택시 사용, Embedded에서는 Lasso, Random Forest 이용해서 피처 선택 후 비교

3. 클러스터 개수에 따른 평가지표 추이 확인, Feature engineering의 활용에 따른 성능 개선 확인

(1) 클러스터 수를 2~7개일 때로 나누어 평가 진행, 피처 선택에서는 2~6개 사용

	K	ARI(raw data)	ARI(log transform)	ARI(Scaled)	ARI(Poly)	ARI(PCA)
0	2	0.4648	0.4815	0.4805	0.2351	0.4805
1	3	0.7168	0.6699	0.7860	0.3490	0.7631
2	4	0.5467	0.6197	0.6545	0.3290	0.6191
3	5	0.5600	0.4745	0.5227	0.3494	0.5999
4	6	0.4699	0.4389	0.4717	0.2816	0.4745
5	7	0.4658	0.4268	0.3874	0.2401	0.4443

그림 1 2-(1) ~ 2-(4) 전처리 방법의 ARI 점수

	Clusters_K	Selected_Features	ARI(Filter1)	Clusters_K	Selected_Features	ARI(Filter2)
0	3	6	0.7039	3	6	0.7039
1	3	2	0.6595	3	2	0.6595
2	3	4	0.6562	3	5	0.6532
3	3	5	0.6532	3	4	0.6527
4	3	3	0.6364	3	3	0.6364
5	4	5	0.5709	4	4	0.6178

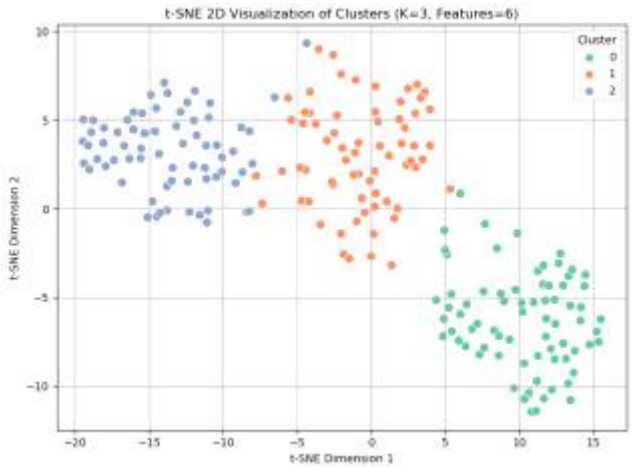
그림 2 2-(5) 전처리 방법의 ARI 점수(Filter Method)

Clusters_K	Selected_Features	ARI(Wrapper)	Clusters_K	Selected_Features	ARI(Embedded1)	Clusters_K	Selected_Features	ARI(Embedded2)
3	6	0.7871	3	7	0.7860	3	4	0.6327
3	5	0.7109	3	6	0.7521	4	4	0.6178
4	5	0.6723	3	5	0.7412	2	4	0.5183
3	4	0.6723	3	5	0.7412	5	4	0.5127
4	6	0.6619	4	5	0.6699	6	4	0.4650
3	3	0.6085	4	5	0.6699	7	4	0.4158

그림 3 2-(5) 전처리 방법의 ARI 점수(Wrapper, Embedded Method)

최적의 결과는 Wrapper Method를 이용하고, 클러스터의 개수가 3개, 피처의 개수가 6개인 경우 ARI가 0.7871로 가장 높았습니다.

4. 2차원 시각화를 통한 최적의 클러스터링 결과 확인



5. 추가 성능 개선 및 적절한 클러스터링 성능평가를 위한 방안 논의

(1) 데이터셋과 피처에 대한 이해도를 높여, 파생 피처 생성을 통해 성능을 개선할 수 있습니다.

(2) KMeans 외 다른 군집화 알고리즘을 사용해 성능 개선을 논의할 수 있습니다.

(3) ARI 이외에도, Silhouette, Davies-Bouldin Index 등을 이용해 클러스터링 성능을 평가할 수 있습니다.