유튜브 급상승영상 분석

이혜규
강남대학교 데이터사이언스 전공
E-mail: yhg3120@naver.com

Introduction 서론

- **주제 선정 이유**
  - 전세계 최대 규모 동영상 공유 플랫폼
  - 누구나 쉽게 접근할 수 있고 다양한 분야와 주제가 있음
  - 계속 성장하고 발전 중인 플랫폼

- **주제**
  - 급상승 영상 가능 분류 분석
  - 급상승 영상 군집화 분석
  - 급상승 영상 특징

- **데이터 소개**
  - 캐글데이터 : YouTube Trending Video Dataset
  - 2020년 8월 12일 ~ 2021년 9월 20일의 79554개의 인기 급상승 동영상 목록 데이터

Method 방법론

- **종속변수**
  - 5일 이상 급상승 동영상 목록에 있던 데이터를 1로, 아닌 데이터는 0으로 설정
  - 1 : 8098 / 0 : 2824 데이터 불균형 상태
    => oversampling의 방법인 SMOTE 알고리즘 사용하여 해결

- **분류 분석**
  - KNN : 과적합 발생
  - LDA : 분류가 거의 안됨
  - SVM : 그나마 괜찮았던 방법
  => 공통으로 정확도 뿐 아니라 오차행렬로 같이 확인.
  => 공통으로 파라미터를 넣지 않고 한번, 파라미터를 조정하고 한번씩 진행

- **군집 분석**
  - K-Means : 분산도가 크게 나옴
  - K-Means+PCA : 필요한 주성분이 10개의 피처가 나옴. 분산도는 약 1/3이 줄어들음
  - DB-SCAN : 밀도가 좁은 데이터라 115개 군집화됨
  - SOM : pca를 적용하지 않았을 때가 더 좋은 결과
  - GMM : 분포의 크기는 달라도 어느정도 군집화된 데이터가 나옴

- **특징 분석**
  - k-means분석으로 조회수와 관련되어 다른 변수들과의 관계/영향을 분석했지만, 상관관계를 제외한 다른 관계는 나오지 않음.

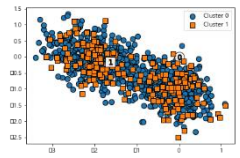
Result 결과

- **분류 분석**

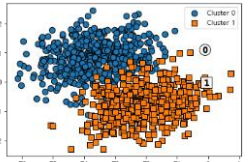
알고리즘	KNN	LDA	SVM
오차행렬	[[6005 0] [0 6140]] [[1379 713] [742 1215]]	[[3263 2850] [2440 3592]] [[1050 934] [866 1199]]	[[3893 2220] [1869 4163]] [[1200 784] [637 1428]]

- **군집분석**

K-Means



K-Means +PCA

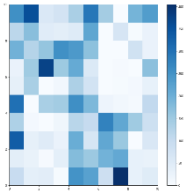


DB-SCAN

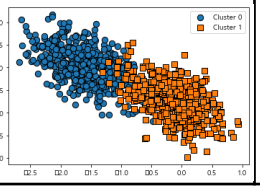
```
pd.crosstab(Y, df_dbs['predict'])
```

predict	-1	0	1	2	3	4	5	6	7	8	...	104
pred												
0	2481	3	6	13	16	2	9	11	15	0	...	0
1	6350	3	0	1	46	8	142	8	8	50	...	8

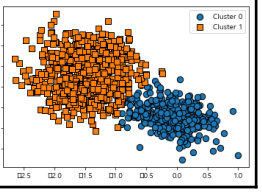
SOM



SOM +PCA



GMM



Discussion 토의

- **활용방안**
  - 좀 더 데이터 양이 많으면 과적합 문제가 해결되기 때문에 더 다양한 분석 가능
  - 유튜버라는 직업이 새로 나오는데 이를 서포터할 데이터 분석가가 있다면 큰 도움 가능
  - 현재 데이터는 많지만 분석에 뛰어드는 사람은 거의 없는 분야

- **아쉬운 점**
  - 관련 데이터의 종류가 적어서 직접 수집해야하는 부분이 많았다.
  - 좀 더 다양한 데이터가 있었으면 더 좋은 결과가 나왔을 것이다.
  - 프로젝트를 진행하면서 방향설정에 대해 고민하느라 시간을 많이 놓쳐서 아쉽다.

References 참고
1. https://www.kaggle.com/rsrishav/youtube-trending-video-dataset