[LG U+ Why Not SW Camp 1기]

IPTV 채널 추천 서비스 -3팀

모델 정의서 및 평가서

프로젝트명: 우리 TV가 달라졌어요

팀명: 우티달

팀장: 박종현

팀원: 권정인, 신새봄, 이중찬, 정연진

목차

1. 모델 개요

1.1 배경 및 모델 선정 고려 사항

1) 비지도 학습

- 사전에 레이블이 없는 사용자 데이터를 클러스터링 필요
- 데이터셋의 명확한 타겟 변수(예: 선호 장르 레이블) 불분명

2) 해석 용이성

- 각 클러스터는 유사한 시청 패턴을 가진 사용자 그룹으로 정의
- 클러스터링별 특성 분석을 통해 추천 시스템과의 통합에 용이해야함

3) 효율성

- 계산 속도가 빠르고, 대규모 데이터셋으로 확장 가능
- 데이터가 적은(초기 사용자)도 활용가능한 모델

1.2 모델 비교

알고리즘	장점	단점	적용 가능성
K-means	간단하고 계산 속도가 빠름	구형 클러스터에만 적합, 이상치에 민감	대규모 데이터셋 및 초기 분석에 적합
GMM	클러스터 간 중첩 허용, 유연성 높은 모델링 가능	계산 비용이 높음, 초기값에 민감	복잡한 행동 패턴 분석에 유리
DBSCAN	이상치 제거 가능, 비구형 클러스터 적합	밀도 차이가 크면 부적합	이상치가 많은 데이터 분석에 유리
Hierarchical Clustering	계층 구조 제공, 클러스터 개수 사전 설정 불필요	계산 비용 높음, 대규모 데이터셋에 부적합	소규모 데이터셋 또는 계층적 관계 분석에 적합

1.3 모델 선정

- 초기 데이터셋 및 대규모 데이터셋 모두 활용 가능한 K-means 알고리즘으로 선정

2. 데이터셋 구성 및 전처리

2.1 데이터셋 구성

1) 데이터셋: IPTV 사용자 시청 이력 및 음성 데이터 (생성 데이터)
2) 형식: csv 파일
3) 구성
    - 학습 데이터: 312개(80%)
    - 테스트 데이터: 78개(20%)
    - 총 390개
4) 주요 특징
    - ID: 사용자 ID(1,2로 구분)
    - Mean Pitch: 사용자 평균 음성 피치
    - Voiced Duration / Total Duration: 음성 지속 시간 대비 전체 시간 비율
    - 시청 프로그램명

2.2 데이터 전처리 과정

1) 결측값 처리: 클러스터별 특성이 중요하므로 결측값은 분석 제외
2) 이상치 제거: IQR(사분위수 범위) 방법으로 이상값 탐지 및 제거
3) 정규화:
    - Min-Max 정규화를 통해 데이터를 [0, 1] 범위로 스케일링

- Min-Max 정규화 산식:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

3. 모델 알고리즘 및 평가 결과

3.1 알고리즘 개요

1) 알고리즘: K-means 클러스터링

2) 특징:

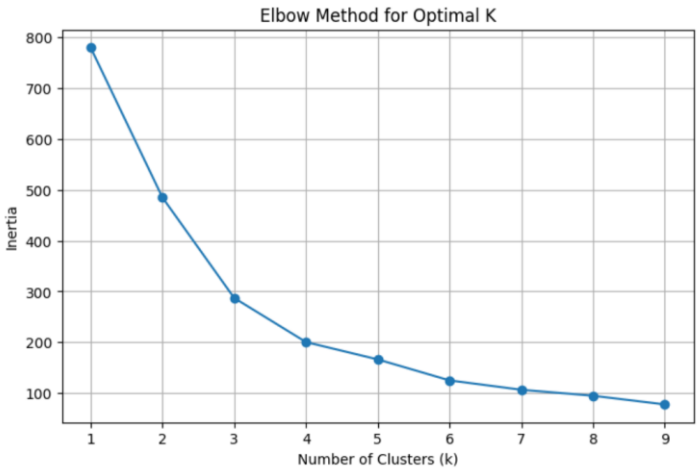
- 비지도 학습을 통해 데이터 그룹화 수행
- 사용자 데이터를 기반으로 유사한 행동 패턴을 가진 그룹 도출

3) Feature Selection:

- 총 8 개 feature 중 상호 관계성이 높은 Feature 들은 제거하여 최종 3 개의 Feature 로 모형 개발

4) 파라미터 설정:

- 클러스터 개수(K): 4 -> Elbow Method 로 결정



- 거리 측정: 유클리드 거리

5) 학습 방식:

- 학습 데이터(80%)를 기반으로 클러스터링 모델 생성
- 테스트 데이터를 통해 모델 성능 평가
- 총 4 번 학습/테스트로 검증하고, 이 중 2 번 모형으로 진행

3.2 성능 평가 결과

번호	학습 데이터 정확도(%)	테스트 데이터 정확도(%)
1	72.69	79.46
2	83.44	84.21
3	80.42	81.58
4	73.33	86.67

- 평균 정확도(테스트 데이터): 약 82.98%

4. 향후 개선 방향

4.1 데이터

1) 데이터 부족 문제 해결:

다양한 사용자 그룹의 데이터를 추가적으로 수집하고, 외부 데이터(SNS, 추가 사용자 조사 데이터 등)를 통합하여 데이터 양과 질 개선

2) 추가 변수 도입:

제한적인 현재 변수에 위치 데이터, 시즌별 시청 패턴, 디바이스 유형 등 새로운 변수를 포함하여 정밀도 향상

4.2 모델 성능 개선

1) 클러스터링 품질 향상:

PCA(주성분 분석)로 차원을 축소하거나, GMM(Gaussian Mixture Model) 같은 대안 클러스터링 알고리즘 적용 검토

2) 동적 클러스터링 도입:

실시간 데이터 피드백 시스템 구축과 k 값을 동적으로 조정하는 방법 적용.