불균형 데이터 처리

1. 불균형 데이터란?

- 일반적으로 분류 문제에서 클래스들이 균일하게 분포하지 않은 문제를 의미

2. 불균형 데이터 다루는 전략

2.1 데이터 수집

2.2 평가 기준을 바꾸기

- accuracy는 비대칭 문제에서 사용하면 안되는 평가 기준
- f1 score, roc curve, mcc, kappa 등 사용

2.3 데이터 셋 RE-샘플링

- Over-sampling
- Under-sampling

2.4 가짜 데이터 샘플 생성

- SMOTE : 부족한 클래스의 모조 샘플을 생성, 2개 이상의 비슷한 객체들을 선택해 거리를 재고 사이 사이 새로운 데이터 생성

2.5 다른 Alogirhtms 사용

2.6 모델에 제한은 준다

- penalized-SVM

2.7 다른 관점으로 시도

- Anomaly Detection
- Change Detection