IEE-CIS Fraud Detection

1. Resampling

- 불균형 데이터의 문제점

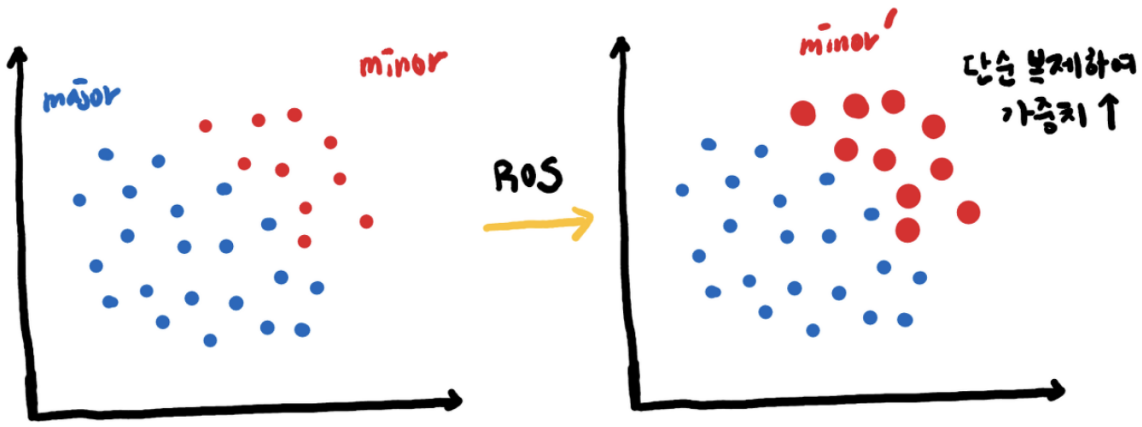
  불균형한 데이터 세트는 이상 데이터를 정확히 찾아내지 못하는 문제가 있음.

- 데이터를 조정해서 불균형 데이터를 해결하는 resampling 기법들이 있음.
  - under-sampling, over-sampling

2. Under-sampling

1. Random Sampling

   다수 범주에서 무작위로 샘플링 하는 것

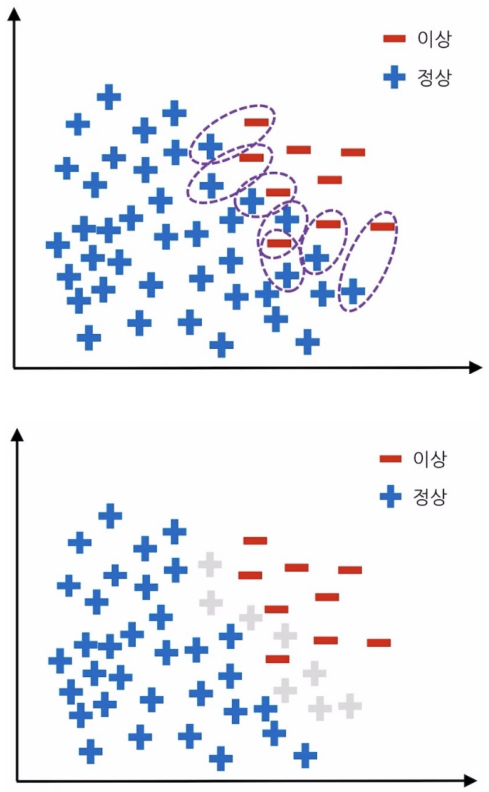


   기존에 존재하는 소수의 클래스를 단순 복제하여 비율을 맞춰주는 것.

   분포는 변하지 않지만, 숫자가 늘어나기 때문에 더 많은 가중치를 받게 되는 원리.

   똑같은 데이터가 증식되어 오버피팅의 위험이 있음.

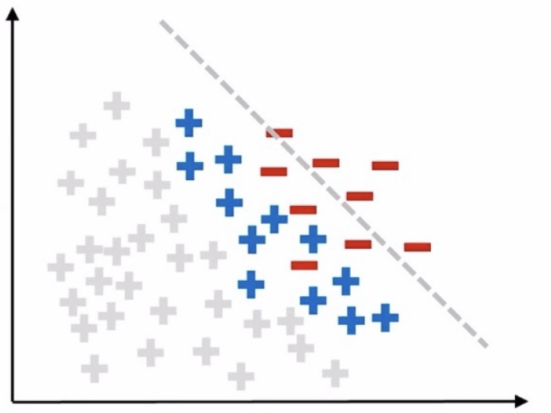
2. Tomek Links



서로 다른 클래스의 데이터 두 점을 연결을 한다. 이 주변에 다른 임의의 데이터 k가 있다고 하자. 이 때 두 점을 연결한 데이터 모두 데이터 k와 각각 연결했을 때의 거리가 두 점(초록색 동그라미 내부의 두 점)을 연결한 거리보다 길다면 이 때 이 두 점간의 링크를 Tomek Link라고 함.

이렇게 Tomek Link에 해당하는 초록색 동그라미들을 모두 색출해내고 이 데이터 쌍들 중 다수의 클래스에 속해있는 데이터를 삭제한다. 즉, 위 예시 그림에서는 초록색 동그라미 안에서 파란색 데이터들을 모두 삭제해주어 언더샘플링을 수행하는 것.

3. CNN Rule



소수 클래스에 속하는 데이터 전체 ‘A’와 다수 클래스에 속하는 데이터들 중 무작위로 하나 선택한 데이터 하나 ‘B’ 의 합 ‘A’ + ‘B’로 구성된 Sub-data를 구성.

그리고 다수 범주에 속하는 나머지 데이터들 중 하나씩 K=1 인 1-NN 방식을 이용하여 해당 데이터가 처음에 선택한 다수 범주 데이터와 가까운지 아니면 소수 범주와 가까운지 확인하여 가까운 범주로 임시로 분류시킨다.

이 과정이 끝나면 정상 분류된 다수 범주 관측치를 제거하여 언더 샘플링한다.

K가 1이 아닌 K-NN 방식을 사용할 경우 모든 데이터가 이상 범주의 데이터로 분류되기 때문에 K는 1이어야 한다.

4. 언더 샘플링의 장단점

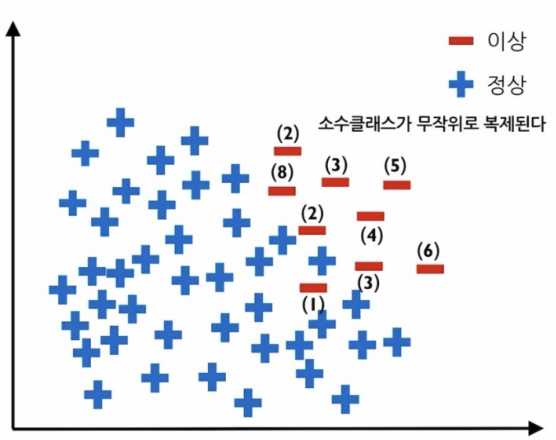
   1) 장점 : 다수 범주 데이터의 제거로 계산 시간 감소.

   2) 단점 : 데이터 제거로 인한 정보 손실 발생 가능.

3. Over-sampling

소수 범주의 데이터를 다수 범주의 데이터 수에 맞게 늘리는 샘플링 방식.

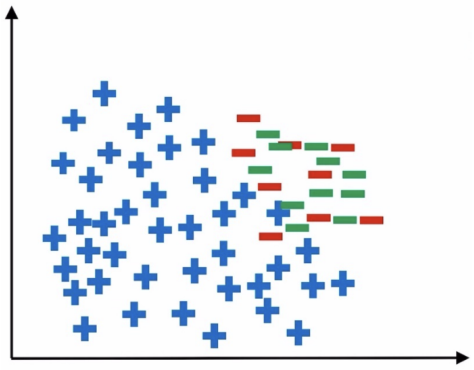
1. Resampling



Resampling 방법은 소수 범주의 데이터 수를 다수 범주의 데이터 수와 비슷해지도록 증가시키는 방법.

소수 범주의 데이터는 무작위로 복제되고 소수 범주에 과적합이 발생할 수 있다는 단점이 있다. 이를 보완하기 위해 나온 방법이 SMOTE 방법이다.

2. SMOTE



초록색이 생성된 가상 데이터

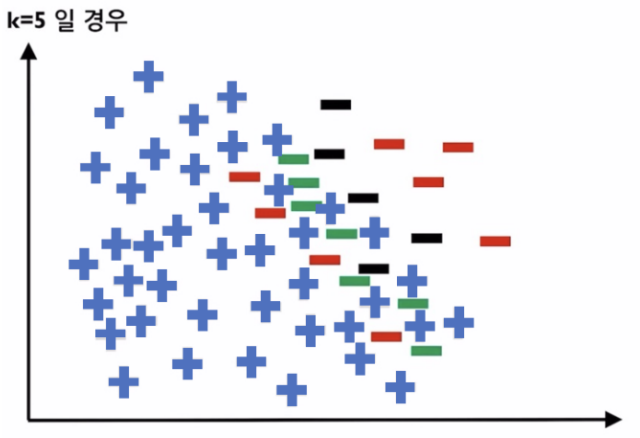
소수 범주에서 가상의 데이터를 생성하는 방법.

K 값을 정한 후 소수 범주에서 임의의 데이터를 선택한 후, 선택한 데이터와 가장 가까운 K 개의 데이터 중 하나를 무작위로 선정해 Synthetic 공식을 통해 가상의 데이터를 생성하는 방법.

K가 1일 경우 데이터가 이상한 형태로 늘어나므로, K값은 무조건 2 이상의 값을 가져야 한다.

3. Boderline SMOTE

   Borderline 부분에 대해서만 SMOTE 방식을 사용하는 것.



검은색이 Danger 관측치, 초록색이 생성된 가상 데이터

   Borderline 을 찾는 것은 임의의 소수 범주의 데이터 한 개에 대해 주변의 K개 데이터를 탐색하고 그 중 다수 범주의 데이터 수를 확인한다.

   이 때 다수 범주 데이터의 수가 K와 같을 경우, 소수 범주의 데이터를 Noise 관측치라고 하며, 다수 범주 데이터의 수가 K/2~K에 속할 경우 Danger 관측치. 0~K/2에 속할 경우 Safe 관측치라고 한다.

   이 중 Danger 관측치에 대해서만 SMOTE를 적용하여 오버 샘플링을 진행한다.

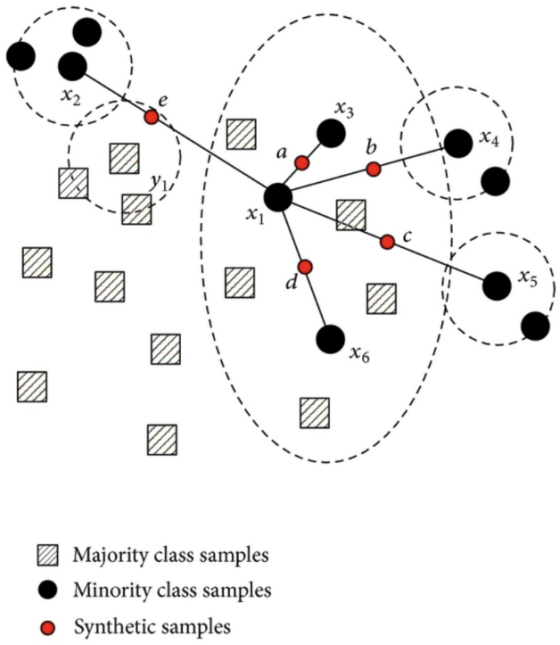
4. ADASYN

   Borderline SMOTE 방법과 비슷하지만 샘플링 개수를 데이터 위치에 따라 다르게 설정하는 것이 차이점.

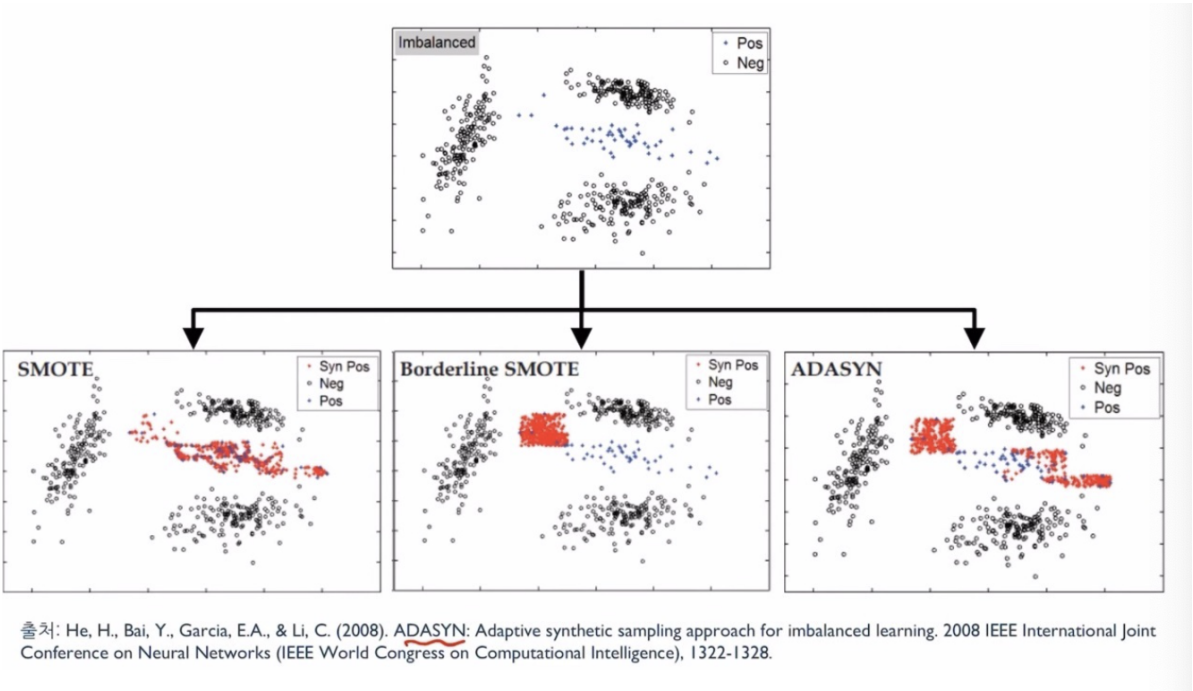
   먼저, 모든 소수 범주 데이터에 대해 주변의 K개의 데이터를 탐색하고 그중 다수 범주 데이터의 비율을 계산.

   → 계산된 각 비율들을 비율의 총합으로 나눠 스케일링을 진행

   → 그 후 (다수 범주 데이터 수 - 소수 범주 데이터 수) 를 스케일링이 진행된 비율에 곱해주고, 반올림된 정수의 값만큼 각 소수 범주 데이터 주변에 SMOTE 방식으로 가상 데이터를 생성



소수 범주 데이터 주변의 다수 범주 데이터의 수에 따라 유동적으로 생성이 가능하다는 장점이 있음.



출처: He, H., Bai, Y., Garcia, E.A., & Li, C. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322-1328.

5. GAN (Generative Adversarial Nets)

   생성자와 구분자로 구성되어 있고 모델은 딥러닝을 사용하는 최신 오버 샘플링 기법.

무작위로 노이즈를 생성하고 생성자를 통해 가짜 샘플을 만들어냄.
→ 그 후 구분자에서 진짜 샘플과 가자 샘플을 판별하고 너무 쉽게 판별될 경우 생성자에게 피드백을 준다.

→ 생성자는 더욱 진짜 샘플과 비슷한 가짜 샘플을 만들어내고 구분자에게 판별 시킨다.

이렇게 생성자와 구분자가 서로 경쟁하며 업데이트 되고 결국 가짜 샘플은 진짜 샘플과 매우 유사한 형태로 생성된다.

6. 오버 샘플링의 장단점

   1) 장점: 데이터를 증가 시키기 때문에 정보의 손실이 없다. 대부분의 경우 언더 샘플링에 비해 높은 분류 정확도를 보인다.

   2) 단점: 데이터 증가로 인해 계산 시간이 증가할 수 있으며 과적합 가능성이 존재한다. 노이즈 또는 이상치에 민감하다.

---

https://casa-de-feel.tistory.com/15

https://wyatt37.tistory.com/10

https://techblog-history-younghunjo1.tistory.com/123