Part. 06

Class Imbalanced Problem

Hybrid resampling기법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

I Hybrid resampling 기법

- Class Imbalanced Problem을 해결하기 위한 방법
  - Resampling method

    - Over sampling : 소수의 데이터를 부풀리는 방법

    - Under sampling : 다수의 데이터를 줄이는 방법

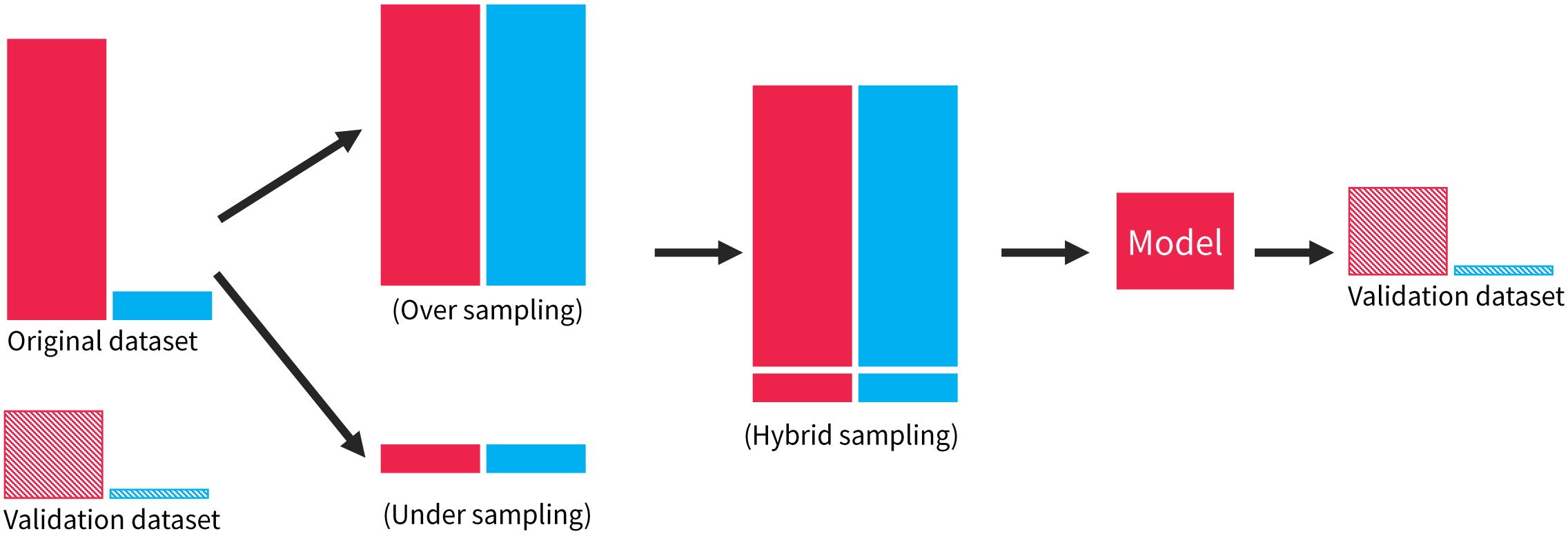
    - Hybrid resampling : Over & Under sampling을 결합해서 사용하는 방법

  - Cost-sensitive learning

    - Class의 오 분류에 대한 cost의 가중치를 조절하여 학습하는 방법

I Hybrid resampling 기법

- Class Imbalanced Problem을 해결하기 위한 방법
  - Resampling method

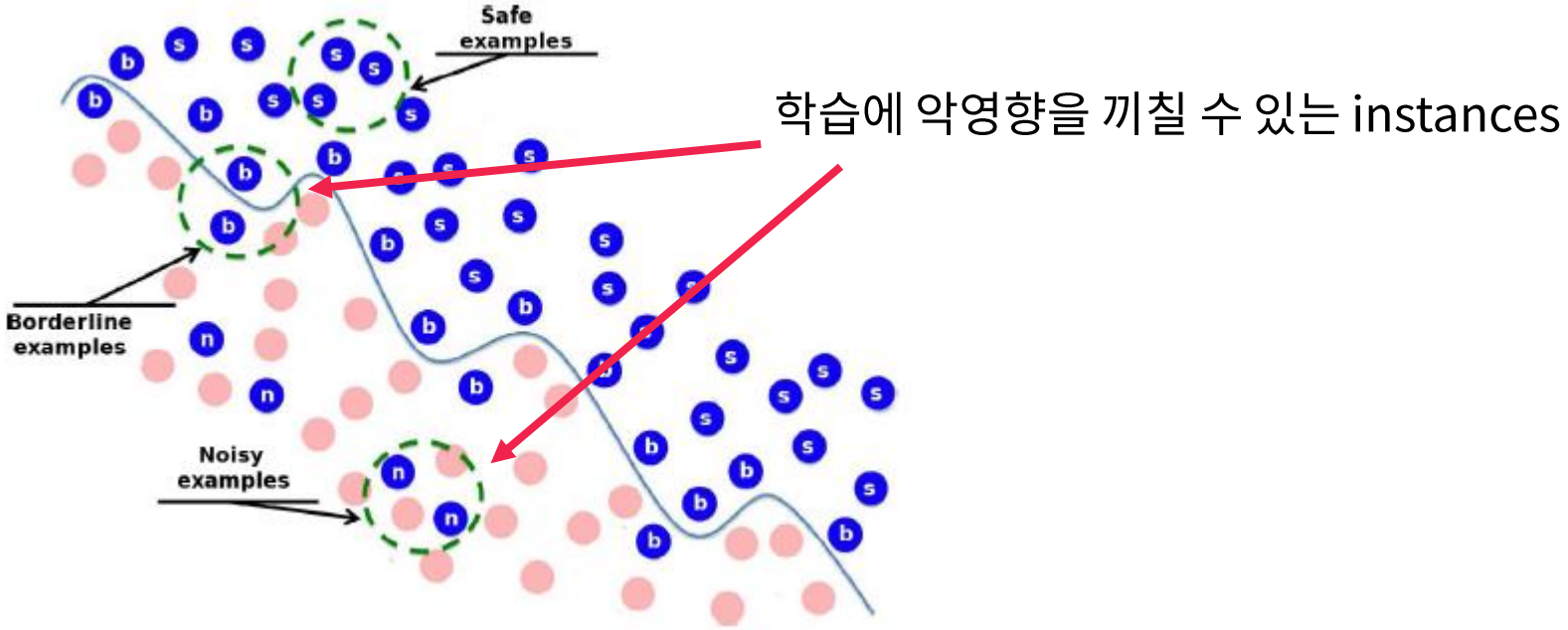


I Hybrid resampling 기법

- SMOTE + Tomek Link
  - SMOTE를 이용하여 minority 데이터를 oversampling
  - Tomek Link를 이용하여 majority 데이터를 undersampling

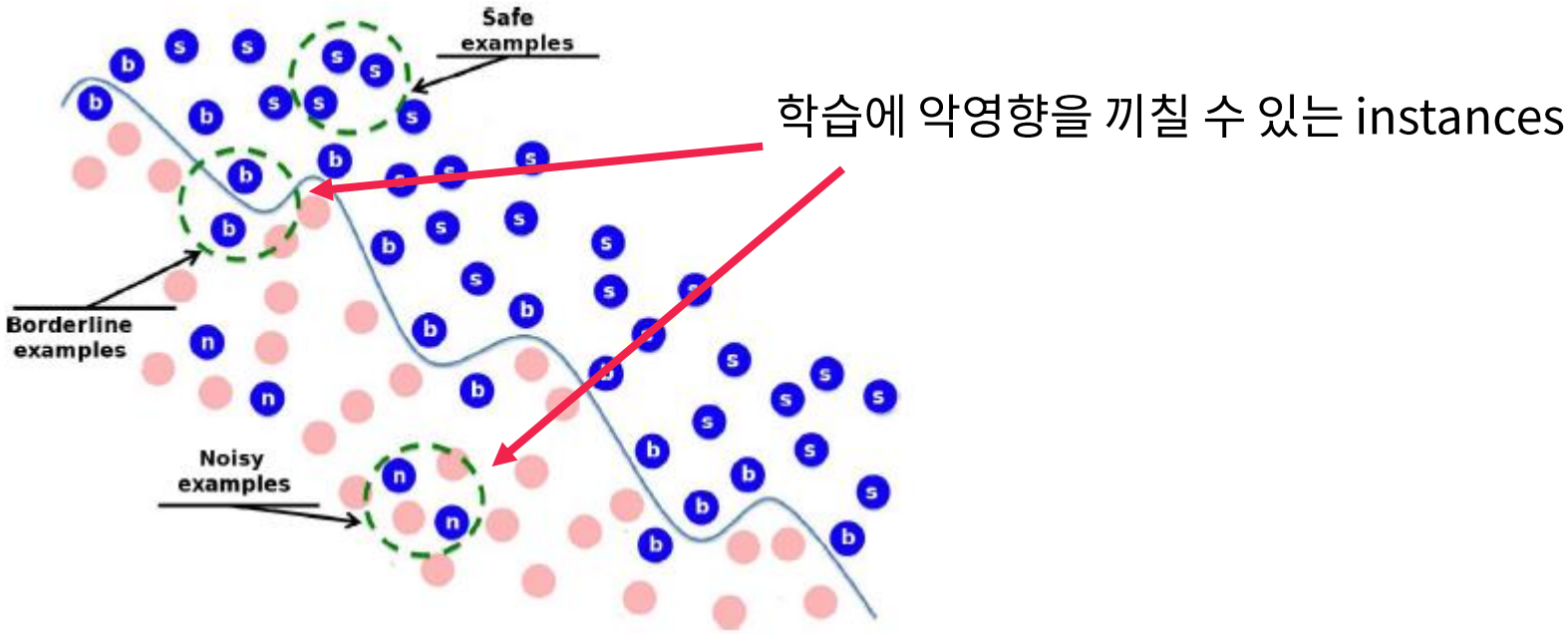
I Hybrid resampling 기법

- SMOTE-IPF
  - SMOTE의 단점을 지적(noise data와 borderline instance가 classifier의 성능을 저하시킴)
  - Iterative Partitioning Filter(IPF)라는 ensemble기반 noise filter를 SMOTE와 결합



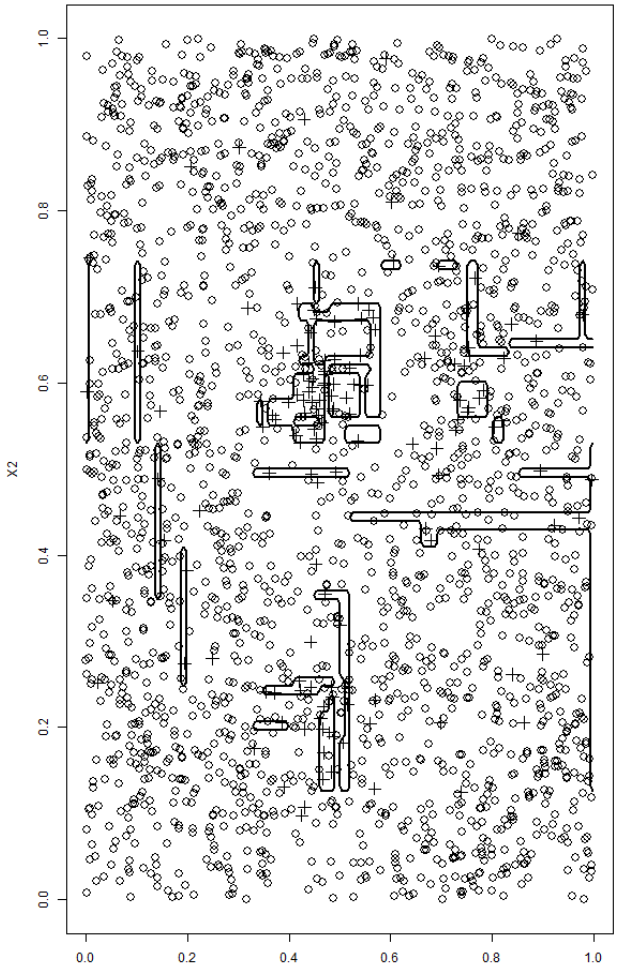
I Hybrid resampling 기법

- SMOTE-IPF
  - Minority 데이터에 대하여 SMOTE적용
  - Cross validation을 통해 여러 모델을 학습하고 평가하여 여러 모델이 오분류하는 데이터를 noise로 구분(voting)
  - Noise데이터를 제거하고 이 과정을 반복

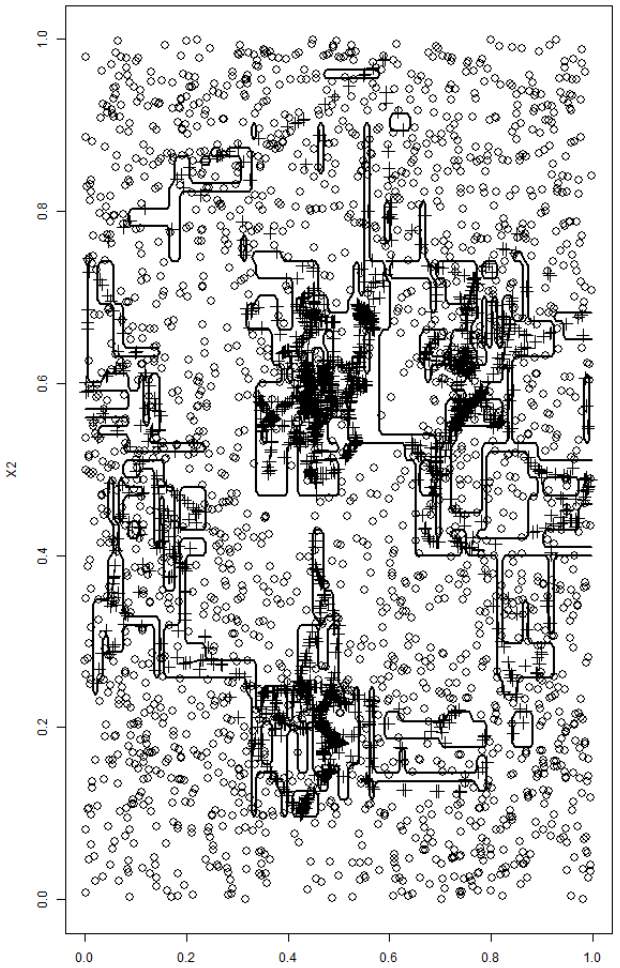


I Hybrid resampling 기법

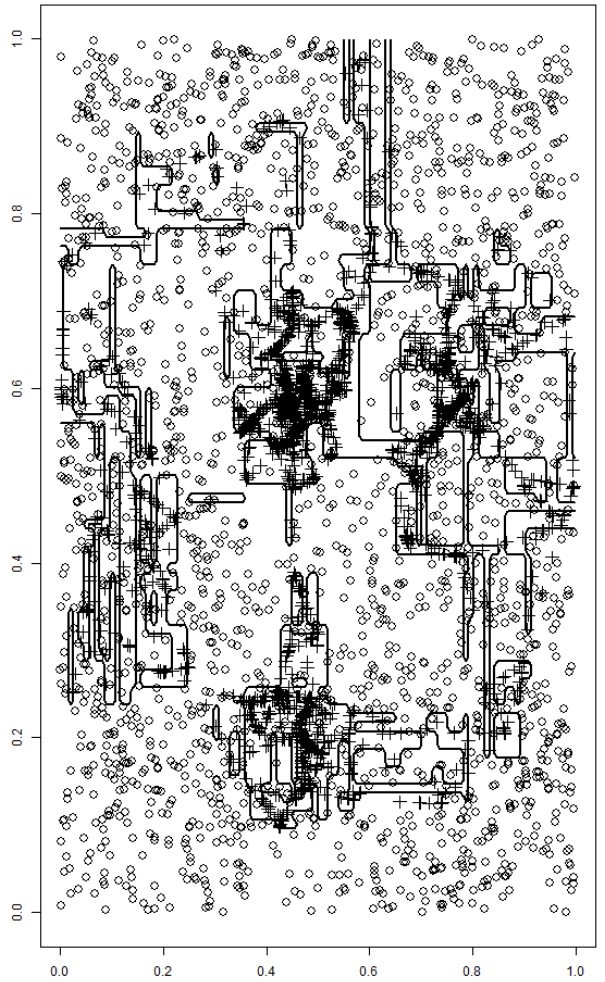
- SMOTE-IPF



SMOTE-IPF적용 전 (0.19)



SMOTE적용 후 (0.28)



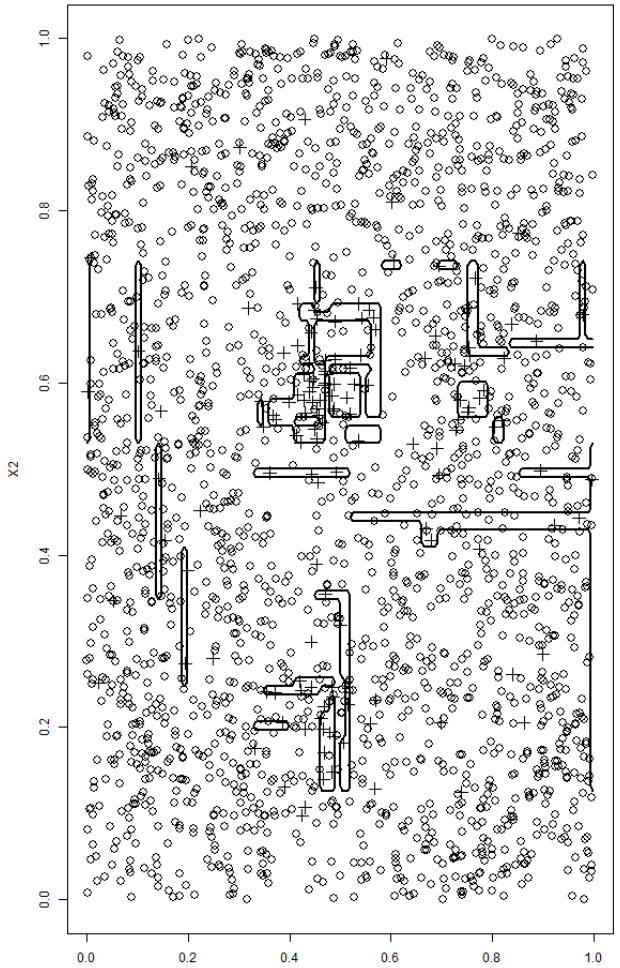
SMOTE-IPF적용 후 (0.29)

I Hybrid resampling 기법

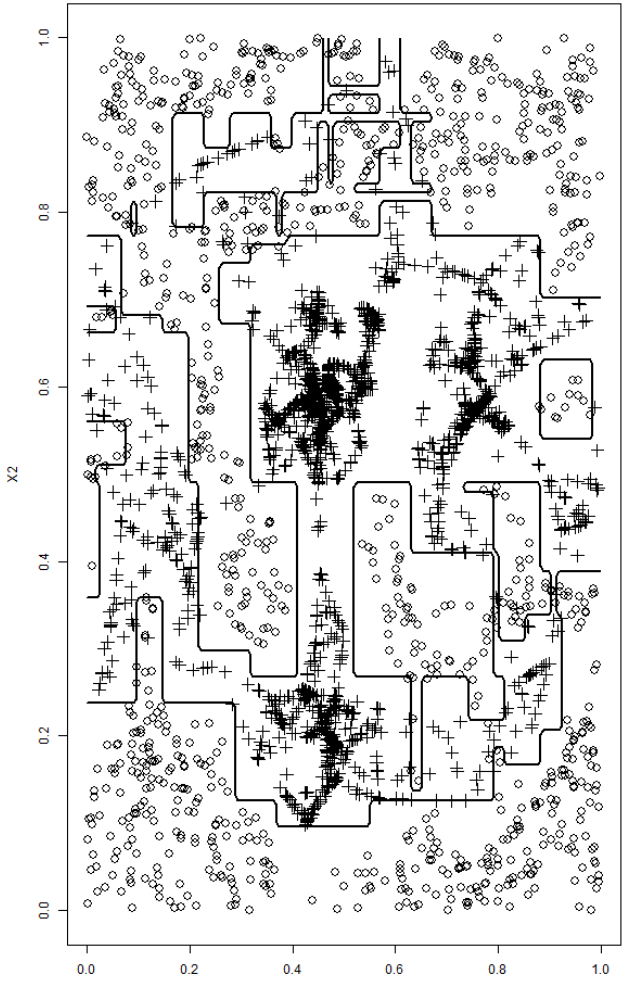
- DBSM
  - DBSCAN을 clustering 수행, 생성된 cluster중 majority data로 이루어진 cluster와 majority data와 minority data가 섞여있는 cluster에서 majority data를 undersampling
  - Minority data는 SMOTE 적용

I Hybrid resampling 기법

- DBSM



DBSM적용 전 (0.19)



DBSM적용 후 (0.22)

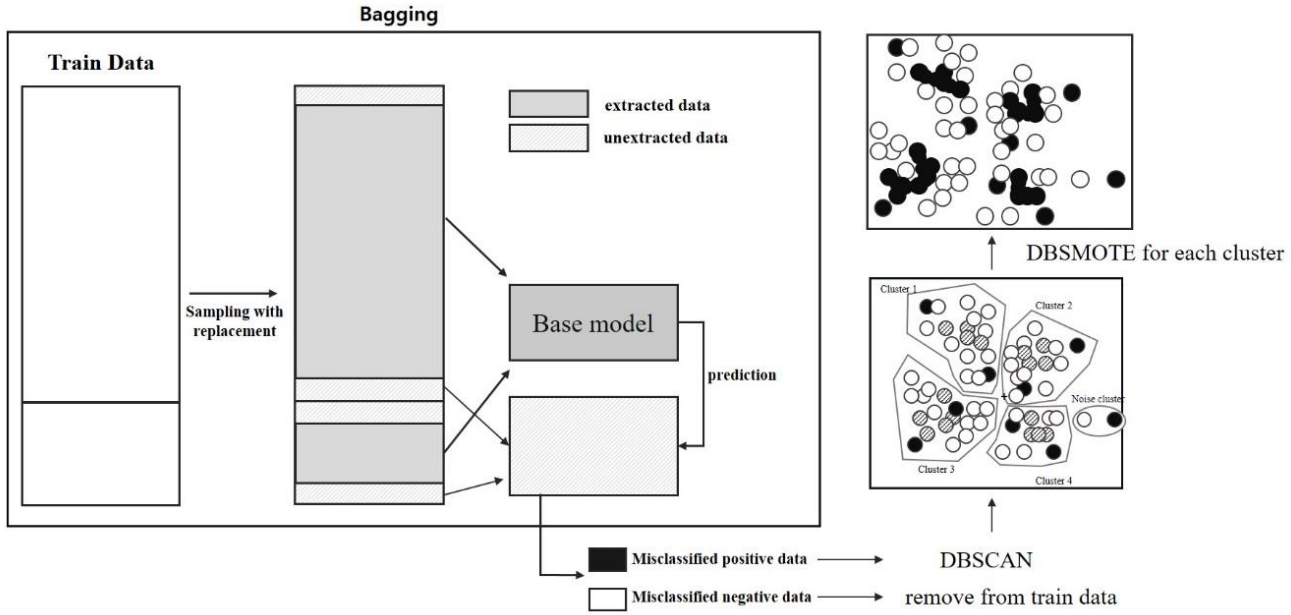
I Hybrid resampling 기법

- Classifier based hybrid resampling method
  - 기존의 SMOTE 와 같은 oversampling기법들은 전체 minority 데이터에 대해서 oversampling 시키는 것이기 때문에, 의사 결정 경계가 과하게 커지는 경향이 있으며, 기존의 데이터 분포를 과하게 왜곡시키는 경향이 있음.
  - 기존의 데이터의 분포를 최대한 유지하면서 더 좋은 의사 결정 경계를 만들 수 있는 방법이 필요.
  - 의사 결정 경계 부근에 있는 데이터만을 조정함으로서, 데이터 분포를 해치지 않으면서 더 좋은 의사 결정 경계를 만들도록 함.

I Hybrid resampling 기법

- Classifier based hybrid resampling method

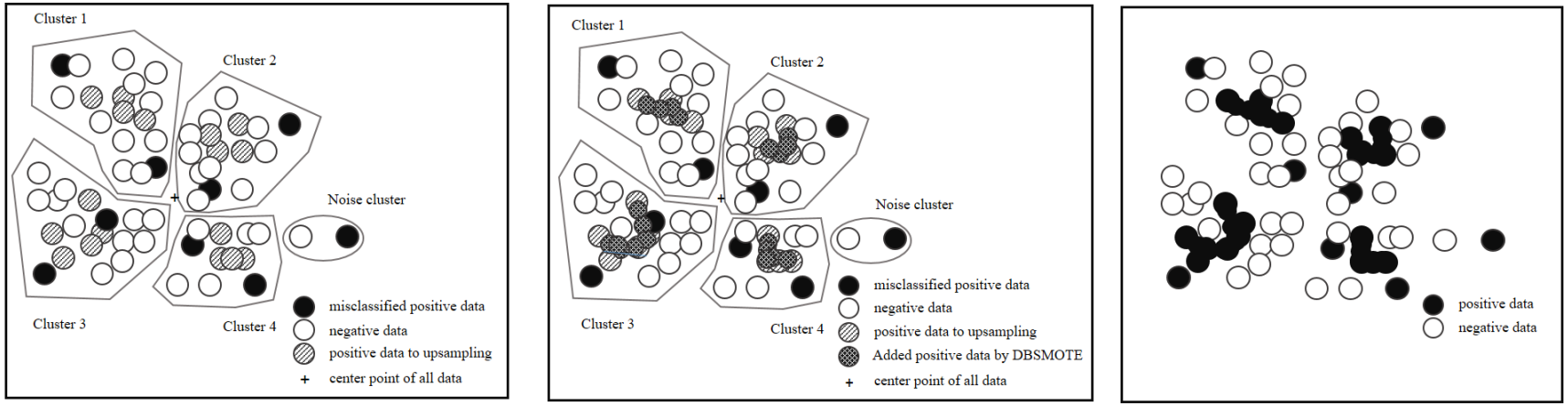
1. 학습데이터와 같은 크기로 복원 추출을 통해 새로운 데이터셋을 만들어 모델을 학습시키고, 이 과정에서 추출되지 않은 데이터를 가지고 학습한 모델로 분류를 수행함.
2. 분류를 했을 때 오 분류된 majority 데이터와 minority 데이터를 저장함.
3. 1번과 2번 과정을 k번 반복하는 과정에서 오 분류된 majority 데이터와 minority 데이터를 모음. (본 실험에서는 30번 수행)



I Hybrid resampling 기법

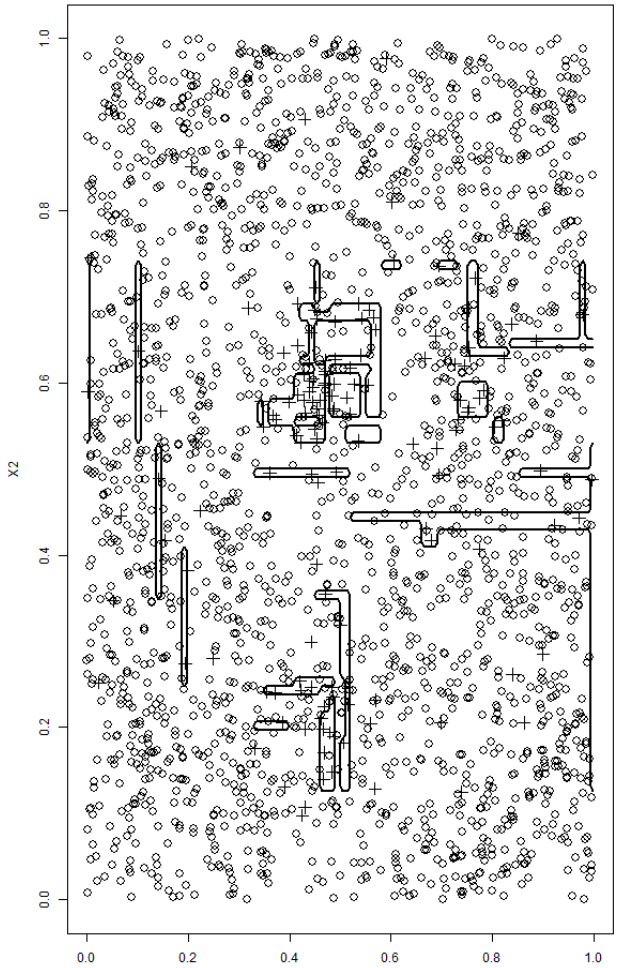
- Classifier based hybrid resampling method

4. 오 분류된 majority 데이터는 모두 제거하는 undersampling을 수행함.
5. 오 분류된 minority 데이터에 대해서는 DBSCAN clustering 방법을 이용하여 군집분석을 수행하고 noise 군집을 제외한 각 군집 안에서 DBSMOTE를 이용하여 oversampling 을 수행함. (이 때 각 군집에서 중심점에서 가장 멀리 있거나 가장 가까이에 있는 data 의 10% 를 제외하고 DBSMOTE를 적용함)

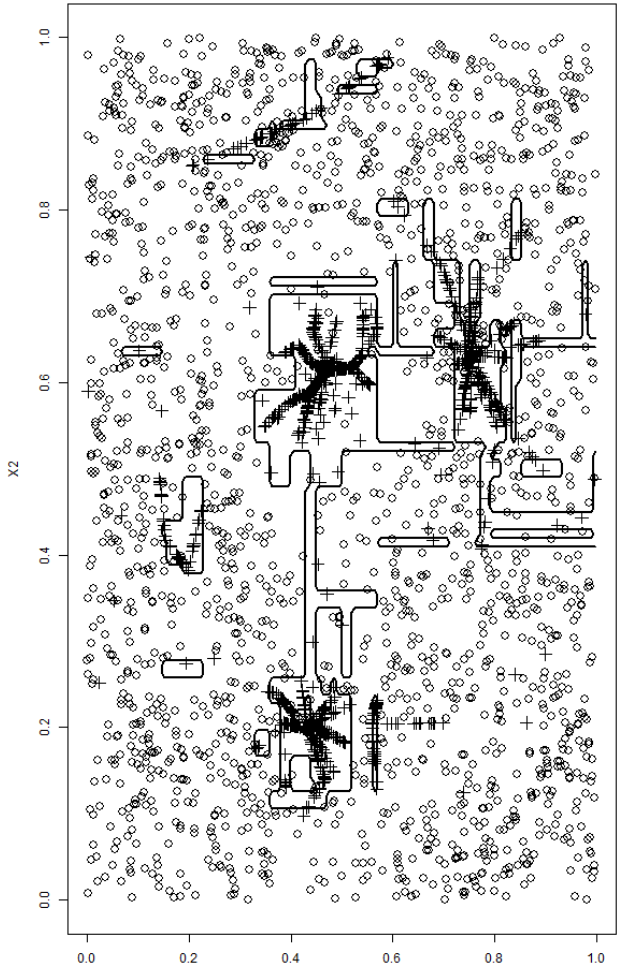


I Hybrid resampling 기법

- Classifier based hybrid resampling method



적용 전 (0.19)

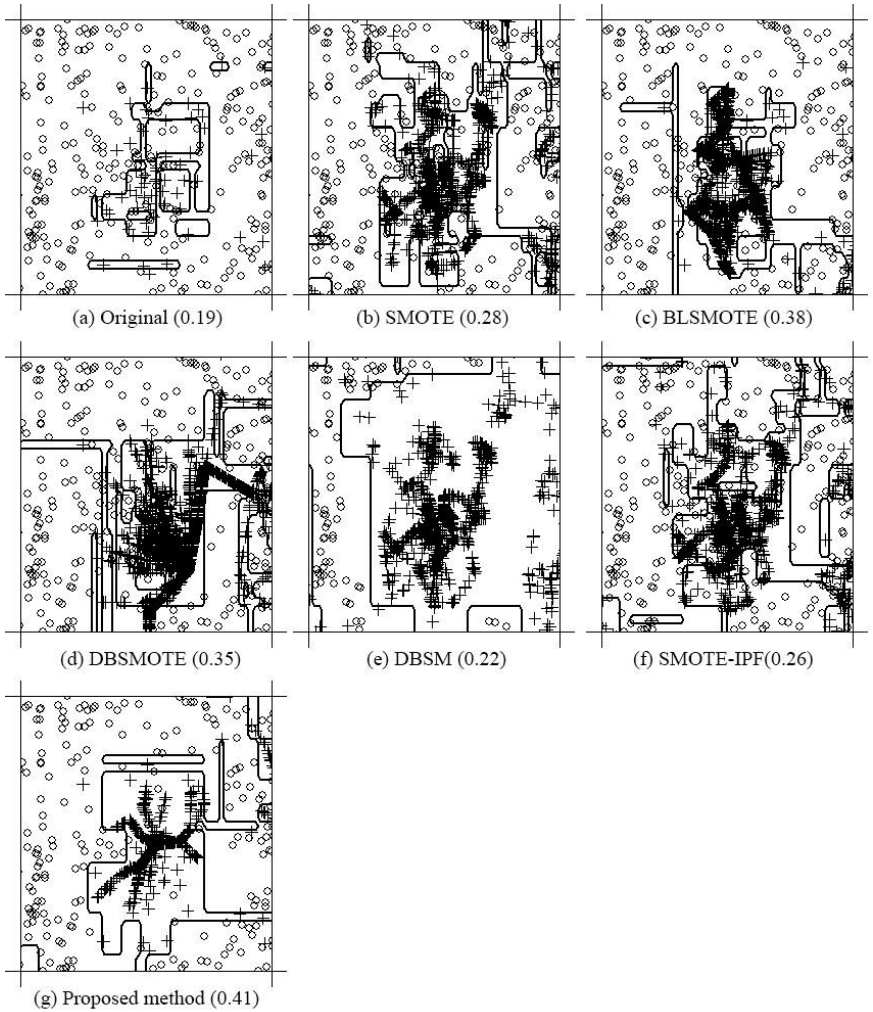


적용 후 (0.42)

FAST CAMPUS
ONLINE
이경택 강사.

I Hybrid resampling 기법

- Classifier based hybrid resampling method



Part. 07

빅콘테스트

| 빅콘테스트 후기

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택