資料前處理 - 不均勻類別問題

Class Imbalance Problem

國立東華大學電機工程學系 楊哲旻

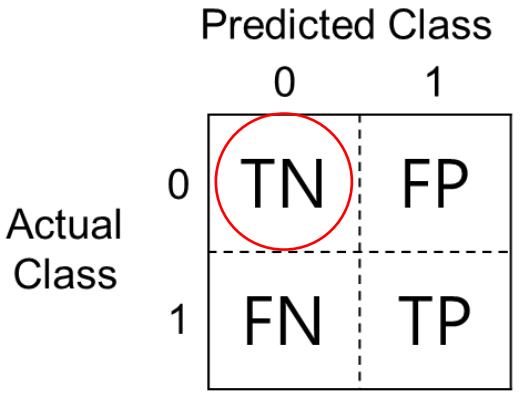
Outline

1. 不均勻類別問題
2. 欠取樣
3. 過取樣
4. SMOTE

01. 不均勻類別問題

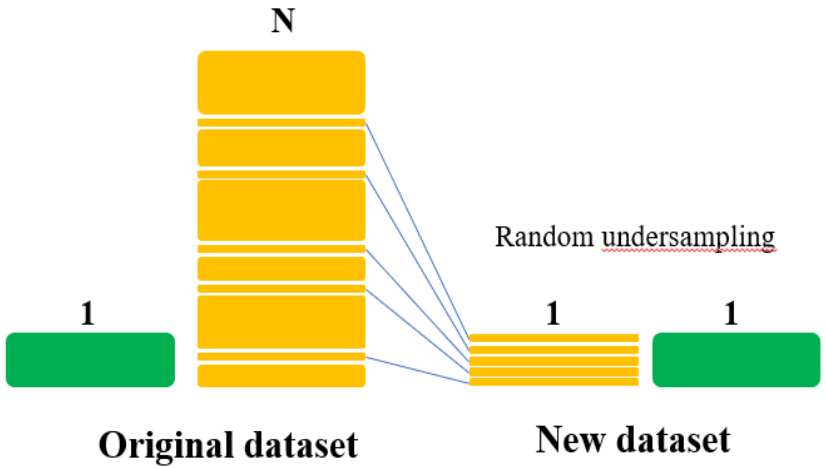
不均勻類別問題(Class Imbalance Problem)

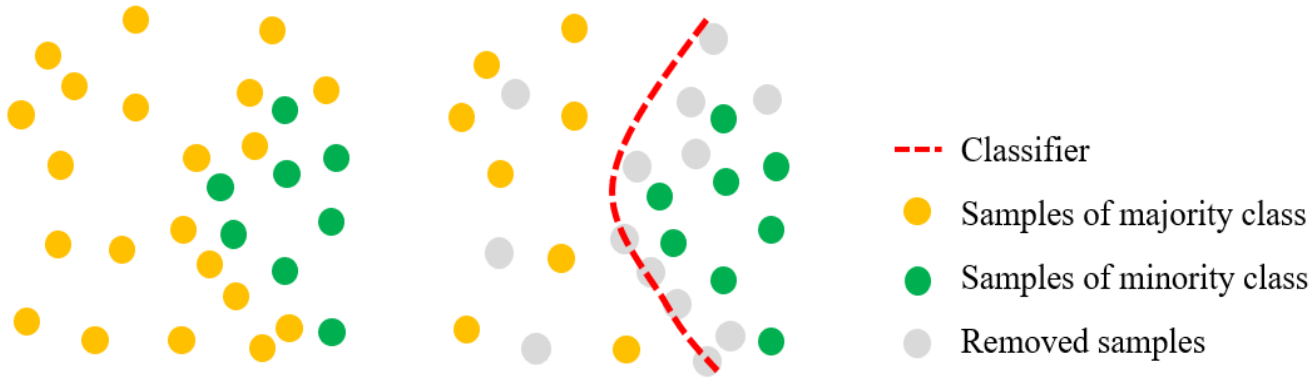
現實的資料大多都為不均勻類別資料，比如：產品瑕疵檢測、疾病檢測、信用卡盜刷、天災預測等，由於多數分類器針對所有資料分類錯誤的誤差是相同的，若多數類別的資料分錯易導致整體的誤差偏高，所以訓練過程都會專注於多數類別資料分類正確



資料前處理 02. 欠取樣

前處理常使用欠取樣與過取樣方法來平衡每個類別的數量



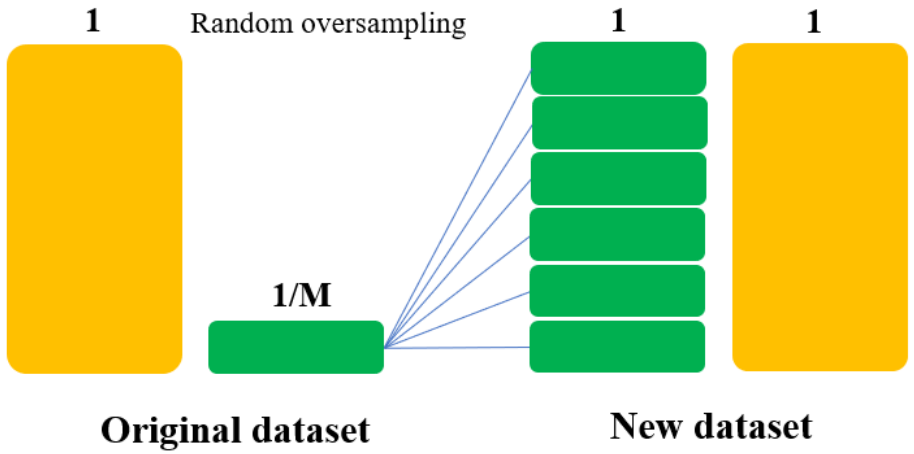


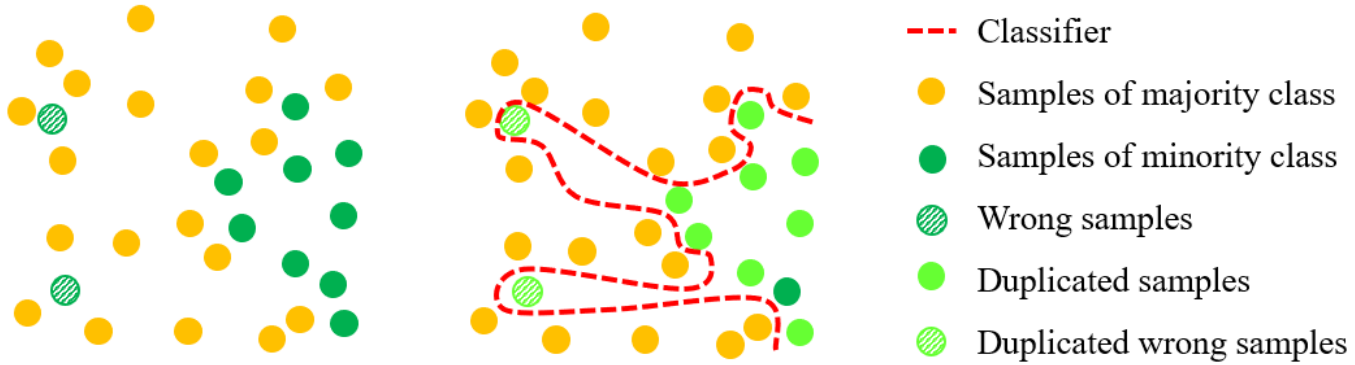
欠取樣(Undersampling)

- 欠取樣容易遺失多數類別資料的重要資訊造成分類器欠擬合

資料前處理 03. 過取樣

前處理常使用欠取樣與過取樣方法來平衡每個類別的數量



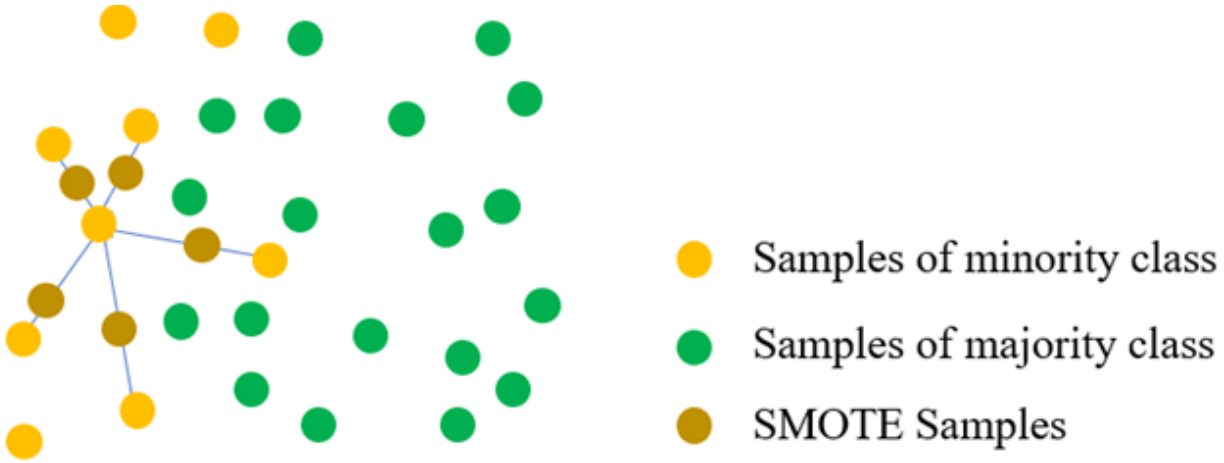


過取樣(Oversampling)

- 過取樣容易使分類器專注分對複製的少數類別資料，如果複製的資料又是異常值更容易造成分類器過度擬合

資料前處理 04. SMOTE

合成少數類別過取樣技術 (Synthetic Minority Oversampling Technique, SMOTE)



將少數類別的樣本計算彼此的鄰近點，該S_i樣本點挑選k個鄰近點，在S_i與S_j的樣本間產生新樣本S_{SMOTE}

$$S_{SMOTE} = S_i + rand(0,1) \times (S_j - S_i)$$