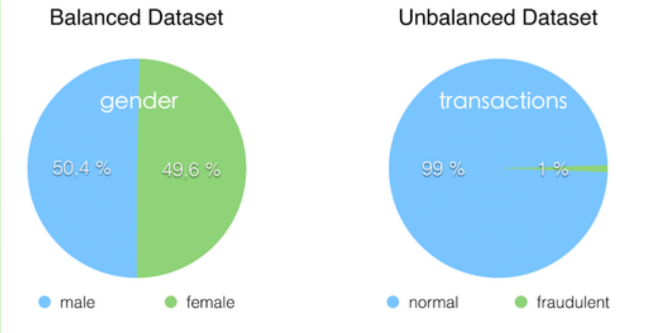
SMOTE：詐欺檢測之實例

1.不平衡資料集是什麼？ (Unbalanced dataset)

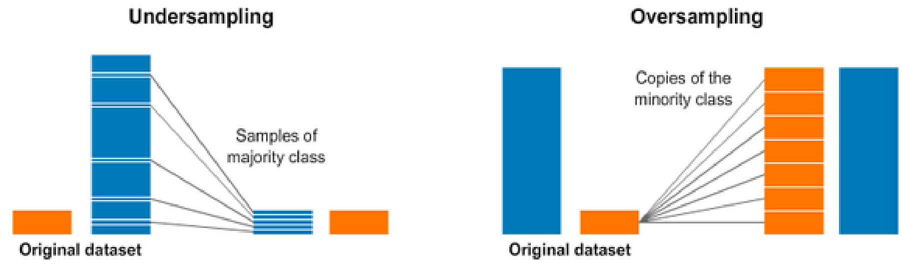


不平衡資料集之實例:

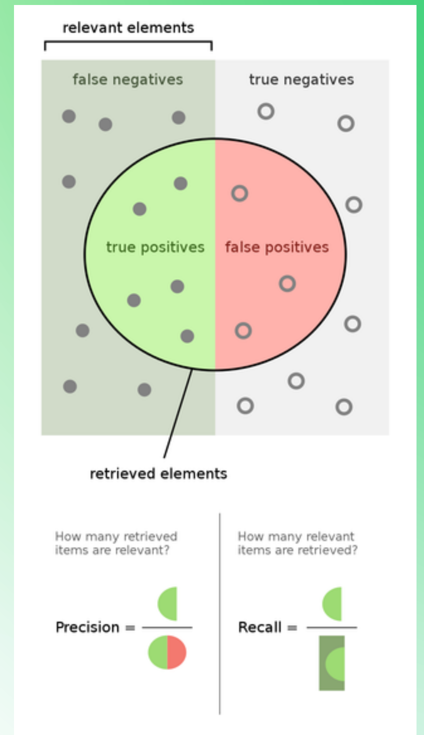
- 金融－詐騙檢測
- 垃圾郵件識別
- 醫療－疾病篩查
- 訂閱流失
- 廣告投放

2. 如何解決不平衡資料集之問題？

- 增加少數類別的樣本數量 (oversampling)
- 減少多數類別的樣本數量 (undersampling)
  - 但undersampling 可能導致總樣本數量過少



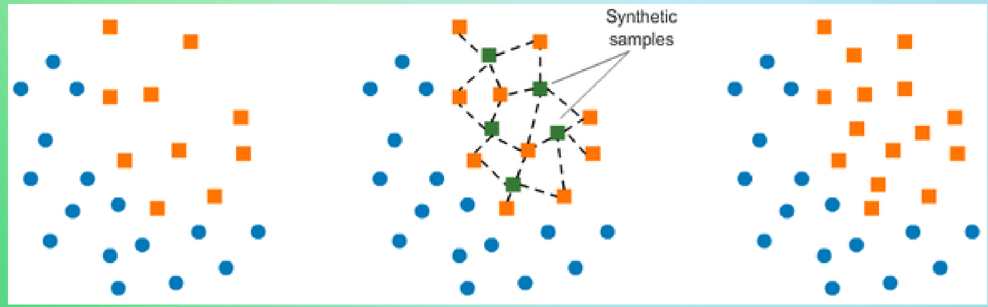
3. 最適合評估不平衡資料集的指標: recall



- recall
  - 在詐騙交易中真的被檢測出來是詐騙的比例
- why not accuracy ？
  - 詐騙交易的樣本數/總樣本數
  - 若全部都猜不是詐騙交易--> 99.8% accuracy

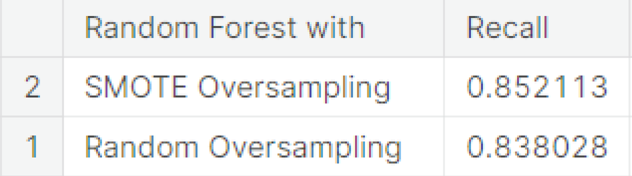
4. SMOTE 是什麼？

- Synthetic Minority Over-sampling Technique
- 選取少數類樣本然後在該樣本與最鄰近樣本之間隨機生成新的樣本。



5. 為什麼要用SMOTE 處理不平衡資料集?

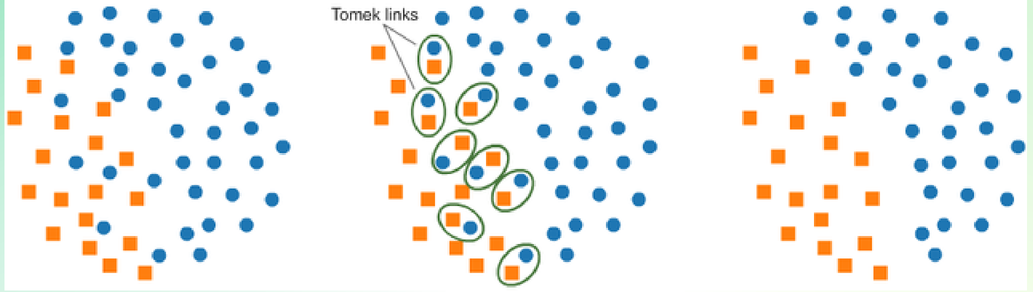
- Random Oversampling 可能導致過擬合，因為模型可能會學會記憶這些重複的樣本。
- SMOTE 選取少數類樣本然後在該樣本與最鄰近樣本之間隨機生成新的樣本。有助於創建更多樣化的資料集，減少過擬合的風險。



	Random Forest with	Recall
2	SMOTE Oversampling	0.852113
1	Random Oversampling	0.838028

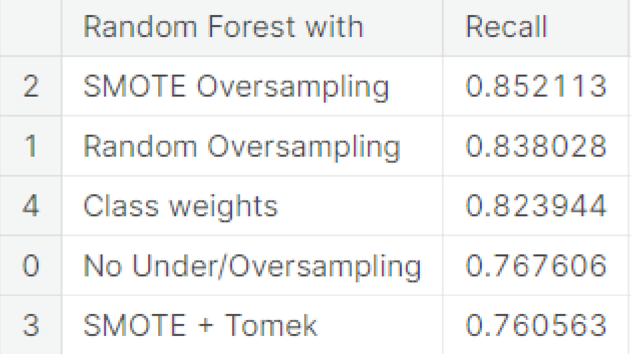
6. 其他處理不平衡資料集的方法

- Class weights
  - 讓少數類別擁有更高的權重，使其對loss function有更大的影響。
- SMOTE+TOMEK
  - TOMEK :找出邊界那些鑑別度不高的樣本，並將其剔除



7. 性能比較

- SMOTE 的performance 最佳
- SMOTE ＋Tomek 的 performance 最差



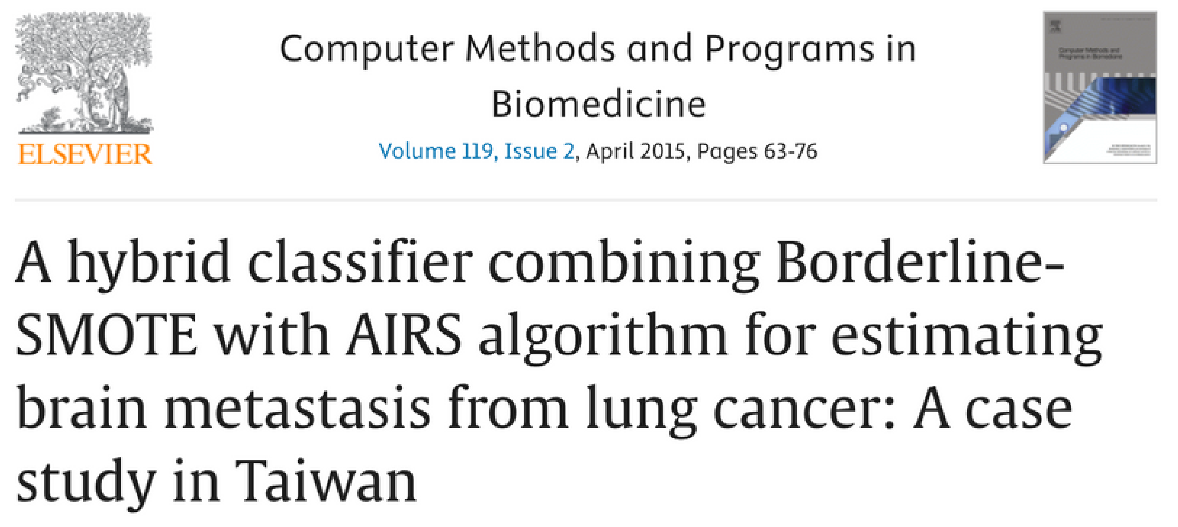
	Random Forest with	Recall
2	SMOTE Oversampling	0.852113
1	Random Oversampling	0.838028
4	Class weights	0.823944
0	No Under/Oversampling	0.767606
3	SMOTE + Tomek	0.760563

8. 使用缺點

- **放大雜訊**：如果原始數據中包含雜訊或異常值，生成新樣本時會放大這些雜訊
- **模糊類別邊界**：在類別邊界附近創建樣本，可能導致模型對邊界的劃分變得不清晰，而降低分類性能

9. 應用實例

- 醫療診斷和健康照護
- 能源和工業
- 金融和風險管理
- 社會科學和政策研究



ELSEVIER

Computer Methods and Programs in Biomedicine

Volume 119, Issue 2, April 2015, Pages 63-76

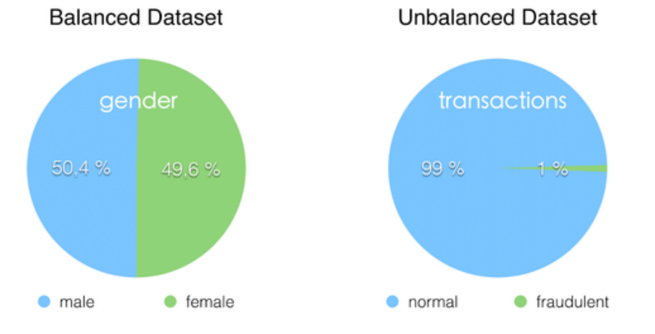
A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan

Take-away

SMOTE的目標是幫助模型學習並預測罕見但關鍵的事件

SMOTE: An Example in Fraud Detection

1.What is imbalanced dataset

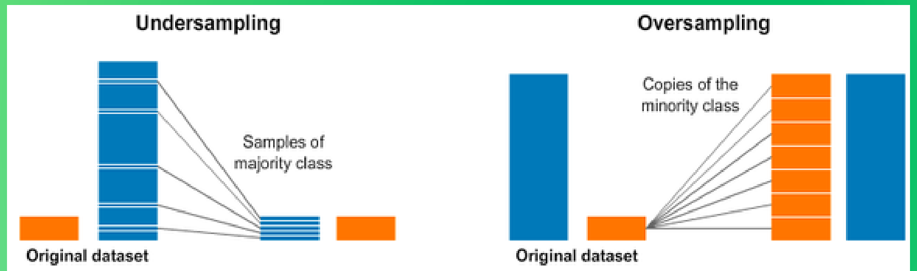


Examples:

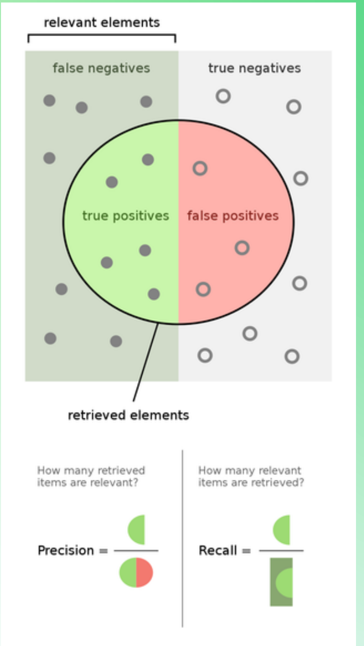
- Fraud Detection
- Spam Identification
- Disease Screening
- Subscription Churn
- Ad Placement

2. How to solve imbalance issue?

- Increase sample number of minority (oversampling)
- Decrease sample number of majority (undersampling)
  - undersampling may cause the sample become too few



3. Most suitable metric for evaluating imbalanced datasets: Recall



recall:

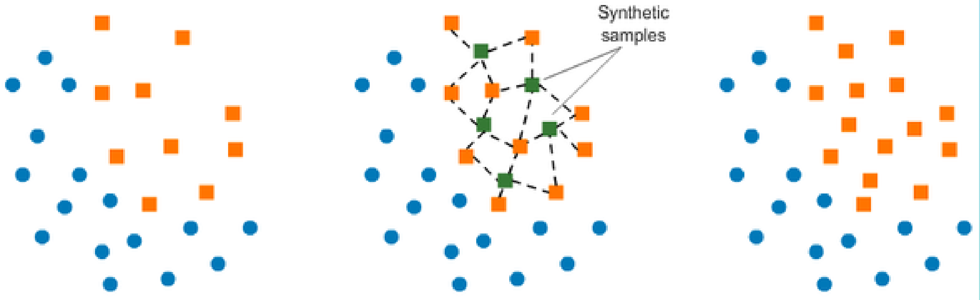
- The proportion of actual fraudulent transactions that are correctly identified as fraud.

why not accuracy ?

- Number of fraudulent transaction samples / Total number of samples.
- If all are guessed as non-fraudulent transactions --> 99.8% accuracy.

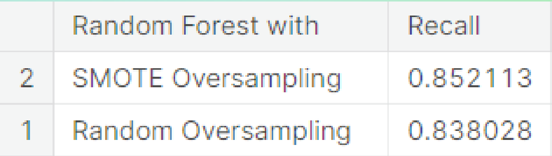
4. What is SMOTE?

- Synthetic Minority Over-sampling Technique
- Select minority sample and randomly generate new samples between this sample and the nearest neighbor sample.



5. Why SMOTE?

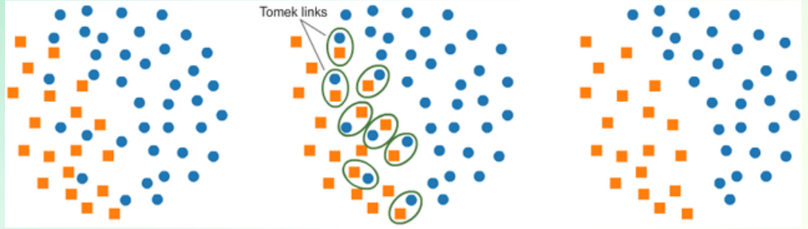
- Oversampling may lead to overfitting.
- SMOTE selects samples from the minority class and then randomly generates new samples between that sample and its nearest neighbors. This helps in creating a more diversified dataset, reducing the risk of overfitting.



	Random Forest with	Recall
2	SMOTE Oversampling	0.852113
1	Random Oversampling	0.838028

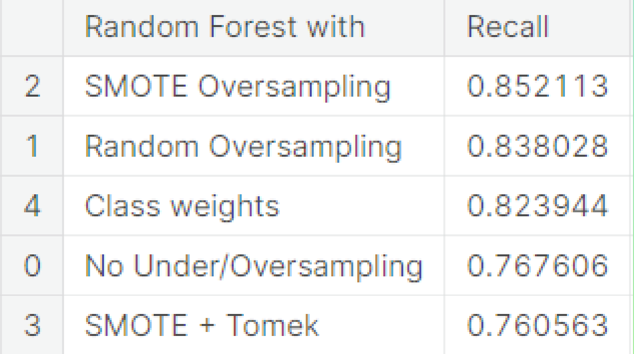
6. Other ways for Imbalanced Datasets

- Class weights
  - Assign higher weights to minority classes, allowing them to have a greater impact on the loss function.
- SMOTE+TOMEK
  - TOMEK: Identifies samples near the boundary with low discriminative power and removes them.



7. Performance Comparison

- SMOTE has the best performance.
- SMOTE + Tomek has the worst performance.



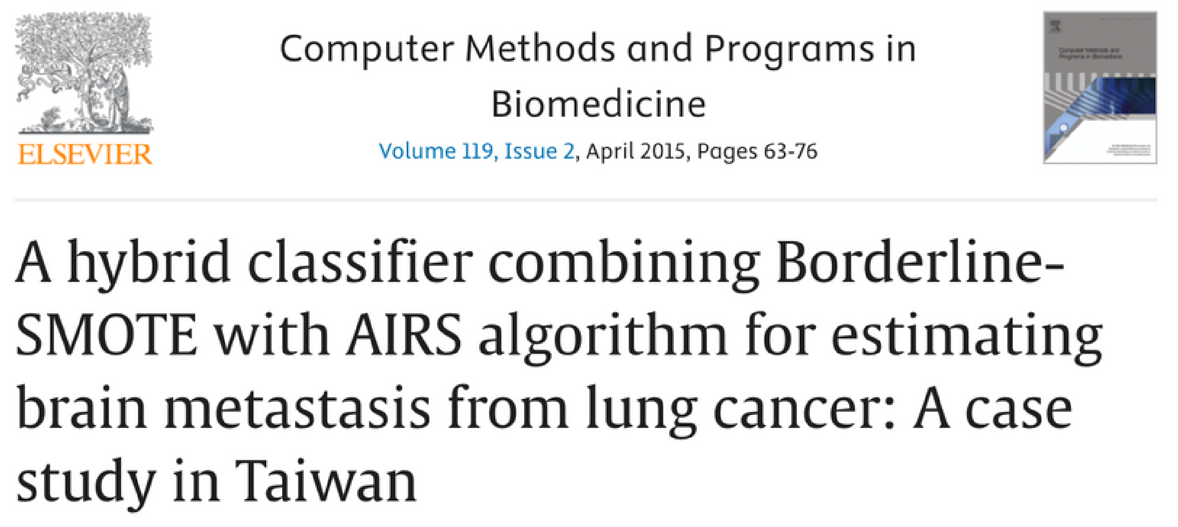
	Random Forest with	Recall
2	SMOTE Oversampling	0.852113
1	Random Oversampling	0.838028
4	Class weights	0.823944
0	No Under/Oversampling	0.767606
3	SMOTE + Tomek	0.760563

8. Cons

- **Amplification of Noise:** Generating new samples may amplify noises.
- **Blurring the Class Boundaries:** Creating samples near the class boundaries may lead to unclear divisions by the model at these boundaries

9. Application Examples

- Healthcare Diagnosis
- Energy and Industry
- Financial and Risk Management
- Social Science



ELSEVIER

Computer Methods and Programs in Biomedicine

Volume 119, Issue 2, April 2015, Pages 63-76

A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan

Take-away

The goal of SMOTE is to assist models in learning and predicting rare but crucial events.

應用實例

Estimating brain metastasis from lung cancer: A case study in Taiwan

Improving detection of COVID-19

Gas turbines diagnosis

Wind turbine blade icing diagnosis