Imbalanced Dataset

朱家輝，余知諺，向九順

(104403050, 104403039, 104403041)

Abstract

在資料不平衡的情況下大多數分類器不能很好地預測。許多的研究人員提出了有關於減少大類別的數量或增加小類別的數量的解決方案或是修改訓練模型來提高分類器的可用性。但仍然不夠好，因為某些特定類型的資料是比較稀少，或者一些新類型的資料可能仍然沒有足夠的數據進行訓練，對於不平衡資料集的使用和處理上研究人員的一個目標是提升對少數的辨識能力而非對整體判別的準確度，以上述 class A 和 class B 分類比為 9:1 為例，分類器只要將所有的樣本都分類為 class A 即可獲得近 90 的準確率，但這樣的分類器並沒有任何價值，並沒有真正的訓練到和學習到東西。本篇我們將以不平衡資料集和我們所蒐集到目前常用的處理方式作為探討方向。

1. Introduction

一般來說，大多數的設備故障和缺陷檢測案例都有一個主要的問題為資料不平衡。何謂資料不平衡?當數據中的樣本為兩個或多個類別時若類別的樣本數量不相等就稱作不平衡，不平衡的現象一定存在只有分布比例相差懸殊才會有影響。以二分類為例，理想情況為 class A 和 class B 數量比是 1:1 但實際情況分布不會如此均勻，若兩類別的樣本比例數為 9:1 或者差距更大例如 97:3，就是一種嚴重的不平衡資料集。

2. Different kinds of classification

不平衡資料集在現實世界中非常容易出現，如(1)生活上：火災風險預測、地震在某時段預測，數據龐大，但實際上災害發生的機率在母體或抽樣機率都是偏低的(2)金融上：詐欺預測、信用卡欠款預測，銀行透過用戶的多項依據評估，而一般的交易情形高達 99%，而詐欺的情形都是過少。(3)醫療上：疾病預測、醫療併發症的研究、生理現象，而真正病發的狀況都是較為少見的。

若兩類別樣本數差異巨大，訓練後的模型對於詐欺交易的判別會非常的虛弱。而這樣的資料集又可再區分成二分類和多分類，以下將以實際資料集作為介紹。

2.1 Multi-class classification(以 UCI glass2 dataset 舉例)

Data Set Characteristics:	Multivariate	**Number of Instances:**	214	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	10	**Date Donated**	1987-09-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	280978

- ✓ Distribution: Total:45211, False:39922, True:5289
- ✓ Brief summary: 此資料集在研究玻璃的物理性質來分析最後玻璃的屬性，學者所蒐集的數據包含了折射率和八種元素(鈉、鎂、鈣、鐵…)的含量來做為期屬性，而其輸出的結果則分為玻璃的七種應用類型，如容器、餐具、窗戶…等，透過上述的資料可發現類型 1 和類型 2 的資量較多而類型 4 的資料最少是不平衡的狀態。

2.2 Binary classification（以 UCI bank-marketing dataset 舉例）

Data Set Characteristics:	Multivariate	**Number of Instances:**	45211	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	17	**Date Donated**	2012-02-14
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	825290

- ✓ Distribution: Total: 214, typeI:70, typeII:76, typeIII:17, typeIV:0, typeV:13, typeVI:9, typeVII:29
- ✓ Brief summary: 兩類別的數量比約為 13:87，此數據及與葡萄牙銀行機構的營銷活動有關蒐集的數據包含 age、job、education、house、loan、default、marital 等銀行客戶相關資訊和多項行銷相關、社會背景相關或是其他屬性，並透過這些屬性來判斷客戶是否會訂購定期存款。

2.3 Relationship between binary and multi-class

多分類可視為是一種二分類的延伸，因此我們可以著重考慮二分類的不平衡狀況處理，因為解決了二分類中的數據不平衡問題後，推廣之後就能解決多分類情況下的不平衡問題。以下將於 section 3, section 4, section 5 介紹面對不平衡資料集之常見處理方式，section 6 介紹對於不平衡之觀點。

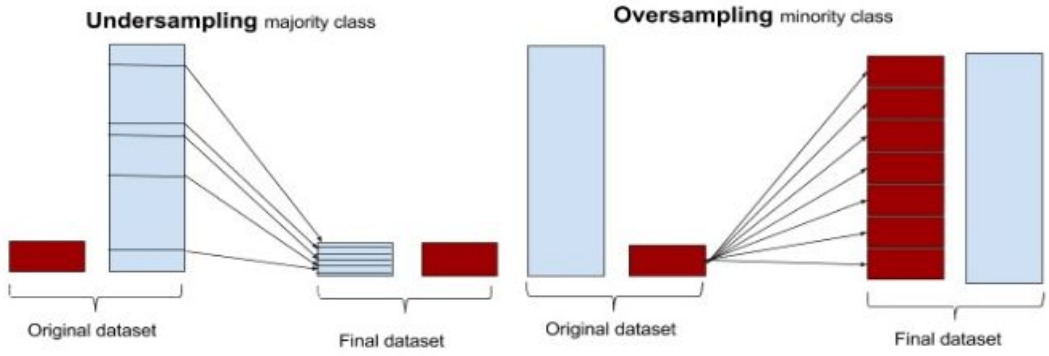
3. Data-level approaches(Resampling techiques)

處理不平衡的數據集最直觀的做法就是對兩比例差異極大的數據類進行重採樣，我們可以透過增加少數群體的頻率或降低多數群體的頻率來達到，這樣做是為了使得兩個類達到大約相同數量的比例，但採樣法不是只單純的從數據角

度改變了模型閾值，還改變了模型優化收斂等一系列過程，那以下開始介紹幾種 resampling 常見的方法

3.1 Under-sampling

對 majority class 進行 under-sampling，透過 under-sampling，解決不平衡問題，並提高了模型的靈敏度。在進行 under-sampling 時，可能會丟失有用的信息(information loss)。



(fig.1)(fig.2)

3.1.1 Randomized

隨機刪除 majority 來平衡類分佈。

Pros and Cons:當訓練數據集很大時，它可以通過減少訓練數據樣本的數量來幫助改善運行時間和存儲問題。但可能丟棄很重要的潛在有用信息且在採樣下隨機選擇的樣本可以是有偏差的樣本。它不能準確代表人口。因此，導致實際測試數據集的結果不準確。

3.1.2 Border cleaning: clean Tomek link

對於 minority 的點，如果它的鄰居(1NN)剛好是 majority，那麼這就構成了一個 Tomek Link，若認為這個 majority 離的太近了，就會刪除這個 majority 點

3.1.3 Cluster Centroid based minority under-sampling(CCMUT)

找到 majority 的集群質心後，距集群質心最遠 unimportant(可捨去)，愈接近集群質心 important(可保留)

3.2 Over-sampling(fig.2)

對 minority class 進行 over-sampling。因為是對少數類進行過採樣，會導致訓練器有過度擬合(overfitting)的情形可能性。

3.2.1 Randomized

隨機複製它們來增加 minority 中的實例數量。

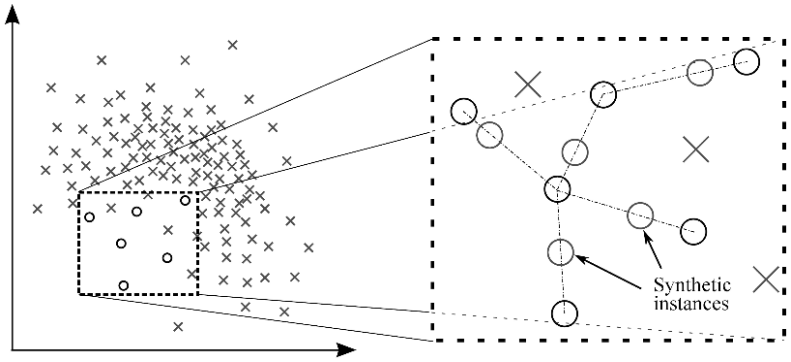
Pros and Cons: 不會導致信息丟失、優於抽樣。但增加了過度擬合的可能性，因為它複製了少數類事件。

3.2.2 Cluster-Based Over-Sampling

K-means 聚類算法獨立地應用於少數類和多數類實例，這是為了識別數據集中的聚類。對每個集群進行 Over Sampling，使得同一類的所有集群具有相同數量的實例，並且所有類具有相同的大小

3.2.3 Synthetic Minority Over-sampling Technique(SMOTE-Kneighbors)

合成少數類: 從少數類中獲取數據子集作為 instance，然後創建新的合成類似實例。然後將這些合成實例添加到原始數據集中。新數據集用作訓練分類模型的樣本。



3.3 Hybrid over and under sampling

可以混合 SMOTE 和 Under-sampling 技術來達到。一邊合成 minority 的數據，一邊刪除 majority 的數據，更快達到均衡。

4. Evaluation

一般來說判斷某個分類器的優劣，大多數都以準確率作為判斷依據，但在某些特殊狀況下（如：不平衡資料集），準確率無法判斷出是否實質上對應到分類器的鑑別力。於 Section 4.1 開始介紹其他評估標準。

- ✓ 準確率（Accuracy）：The fraction of predictions that a classification model got right. In multi-class classification, accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Number Of Examples}}$$

- ✓ Multiclass classification 可透過 binary classification 簡化，以便此主題討論。

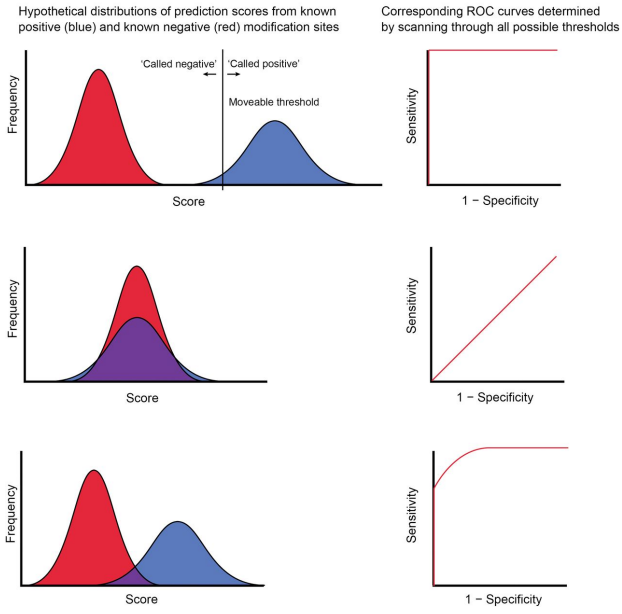
4.1 混淆矩陣（Confusion Matrix）：一種可視化工具，特別用於監督式學習，在非監督式學習一般來說稱為匹配矩陣。矩陣的每一列代表一個類的實例預測。可以更容易判斷機器是否將兩個類別混淆（wikipedia）

	(predicted)Class +	(predicted)Class -
(Actual)Class +	TP	FN
(Actual)Class -	FP	TN

- ✓ (TP)True positive/(TN)True Negative：An example in which the model **correctly** predicted the positive/negative class.
- ✓ (FP)False positive/(FN)False negative：An example in which the model **mistakenly** predicted the positive/negative class.
- ✓ Metrics
  - ■ Precision: identifies the frequency with which a model was correct when predicting the positive class.
  - ■ Recall: Out of all the possible positive labels, how many did the model correctly identify?
  - ■ F1-score:Precision 與 Recall 值的調和平均。

4.2 ROC (receiver operating characteristic) Curve：在不同的 discrimination thresholds 下，基於 TPR 和 FPR 計算呈現的點，進而形成一條曲線。

- ✓ Y 軸：True positive rate(TP/(TP+FN)=recall)---**ideally is 1**
- ✓ X 軸：False positive rate(FP/(FP+TN))---**ideally is 0**
- ✓ AUC (Area under the ROC Curve)：ROC 曲線底下所為出來的面積。
  - ■ 可以很清楚、客觀判斷此模型是否具備鑑別力之重要性。
  - ■ 若 ROC 為 y=x（**AUC=0.5**）則表示此方法**無**分類能力。（fig .2）
  - ■ AUC 值越大（接近左上方）表示 xy 接近理想，為分類能力強的方法。



4.3 成本矩陣（Cost Matrix/Risk Matrix）

套用演算法到現實世界時，很常出現 imbalanced data 的狀況，但同時也會伴隨 cost imbalanced。這時需要了解現實世界對於預測錯誤的成本，改變對分類結果的評估。（也可將此概念應用至演算法中 e.g. cost-efficient learning）。

Predicted Class	True Outcome : Customers Default or Not	
	Positive (or Good)	Negative (Bad)
Positive (or Good)	0	7
Negative (Bad)	1	0

Now, the cost matrix suggests that there is no cost of correct classification. Cost of approving a potential defaulter is 7times the cost of rejecting a good customers. But in some of the scenarios, assigning cost of Type I and Type II is not very easy especially in healthcare.

5. Algorithm

5.1 good performance

Tree classifier

Radom Forest

One-class SVM

5.2 bad performance

Logic regression

Naive Bayes

5.3 others

5.3.1 類神經網路中參數調整：Activation function: e.g. sigmoid 右移

5.3.2. Ensemble learning: e.g. Bagging-base/Boosting-based

6. Perspective

6.1 Cost sensitive learning

權重調整，使演算法進行時，存在類似考慮現實世界成本的因素，以影響分類的能力。

6.2 Penalize model

利用調整比例和參數，對於演算法模式學習時，讓少數的類別的錯誤降低演算法分數的權重高一些，以影響分類能力。e.g. penalized-SVM/LDA

6.3 Outlier detection (anomaly detection)

更改對分類問題的想法，將某極度多數的類別當作正常分佈，而將預測分類之問題更改成偵測異常的演算法。e.g. clustering, isolation forest, one-class SVM

7. summary

- ✓ Classification problems where the classes are NOT represented equally.
- ✓ Multiclass case is the extension from binary classification
- ✓ Algorithm with several resampling approach (e.g. SMOTE)
- ✓ Over-sampling Preferred: GAN, SMOTE, ADASYN perform well
- ✓ Evaluation approach could be significant: confusion matrix, ROC curve
- ✓ Algorithm: Tree classifier have fewer impact with imbalanced dataset(especially Random Forest). While Logic Regression is poor.
- ✓ Different thinking: weight balance(cost-efficient learning), One-class approach, outlier detection.

8. Reference

- https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/(How to handle Imbalanced Classification Problems in machine learning?)
- https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation(Dealing with imbalanced data: under-sampling, over-sampling, and proper cross validation)
- https://www.zhihu.com/question/269698662(undersampling 和 oversampling 會對模型帶來怎樣的影響)
- https://imbalanced-learn.readthedocs.io/en/stable/under_sampling.html (under-sampling 技術實作)
- https://bigdatafinance.tw/index.php/tech/data-processing/353-2017-03-28-11-36-54(二分類與多分類處理關係)
- https://data.world/data-society/bank-marketing-data (Bank Marketing 預測)
- http://archive.ics.uci.edu/ml/datasets/Bank+Marketing(銀行訂閱預測 UCI)
- https://archive.ics.uci.edu/ml/datasets/glass+identification(玻璃分類 UCI)
- https://zh.wikipedia.org/wiki/ ROC 曲線
- http://dni-institute.in/blogs/confusion-matrix-and-cost-matrix/ (Cost Matrix)
- http://essays.biochemistry.org/content/52/165.figures-only (AUC)
- https://towardsdatascience.com/working-with-highly-imbalanced-datasets-in-machine-learning-projects-c70c5f2a7b16 (cost-efficient)
- https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation (CV should follow)

- https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/ (ensemble learning)