資料科學與機器學習 作業一

Dealing with imbalanced datasets

,

*a. 鄭筠叡, 108423034, 資訊管理學系碩士班一年級*

*b. 張德芳, 108423029, 資訊管理學系碩士班一年級*

**Abstract**

不平衡資料集為現實中很常見的現象，當某一類事件發生的機率很低時，蒐集到的資料類別就會相對較低。本篇作業就不平衡資料集的定義與影響作介紹，以及不平衡資料集常見的幾個處理方法，最後再介紹適用與不適用於不平衡資料集的演算法，還有常用的衡量方法。

*Keywords: imbalanced datasets, re-sampling, cost sensitive learning, ensemble learning, decision tree, random forest, kNN, confusion metric, MCC, AUC - ROC Curve*

**1. What is an imbalanced dataset**

*1.1 不平衡資料集的定義*

不平衡資料集通常是指在資料中分類標籤的分布不均衡的狀況，在有兩種或兩種以上的分類中，某些分類的佔比遠高於其他分類，以兩種分類標籤為例，如圖1所示，若兩種分類標籤的占比皆為50%，則屬於平衡資料集；若為90%與10%則屬於不平衡資料集。如果占比為60%與40%，則屬於輕微不平衡資料集。[1]

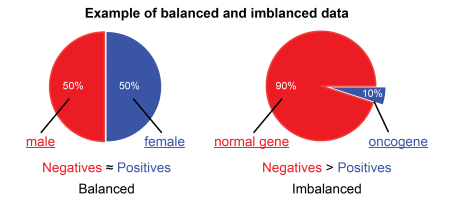
****

圖1.Example of balanced and imbalanced data [2]

*1.2 不平衡資料集的影響*

不平衡資料集會造成模型訓練結果的偏差，導致結果預測偏向佔多數的類別，而在這樣的情況下，預測結果的準確度衡量效果就會下降，因為只要預測為主要類別，整體的準確度會最高。但如果少樣本類別為研究的主要目標，那麼對少樣本類別的預測就會不準確。

像是在病理上，大部分的罕見疾病發生率在10%以內，如果照正常情況蒐集到的醫療資料會是陰性結果佔大多數，以這樣的不平衡資料集來訓練模型，結果會傾向為陰性，而陽性病患的被誤判為陰性的機率極高，造成無法即早救治。

**2. Find ONE imbalanced dataset and give a brief intro/description on it**

*2.1 資料集資訊(Dataset Information)*

我們在Kaggle上找到一個屬於UCI Machine Learning的資料集[3]，該資料集為信用卡會員的繳費資訊(Default of Credit Card Clients Dataset)，資料蒐集時間為2005年4月至9月，總共有3萬筆資料，25個欄位(包含會員ID)。

*2.2 屬性資訊(Attribute Information)*

* 類別屬性：default.payment.next.month，表示下個月是否繳費，屬性值為二元變數，分為有繳費(yes=1)、未繳費(no=0)
* 其他屬性：
* LIMIT\_BAL(信用卡額度)
* SEX(性別)
* EDUCATION(教育程度)
* MARRIAGE(婚姻狀況)
* AGE(年齡)
* PAY\_0～6(每月還款狀況)
* BILL\_AMT1～6(每月帳單金額)
* PAY\_AMT1～6(前個月付款金額)

*2.3 資料不平衡狀況描述(Imbalanced Description)*

我們將資料集的類別屬性做統計，如圖2所示，在3萬筆資料中下個月有繳分者(n1)的比率為22% (6,636筆)，而未繳費者(n0)的比率為78%(23,364筆)，比例差距極大，因此為不平衡資料集。

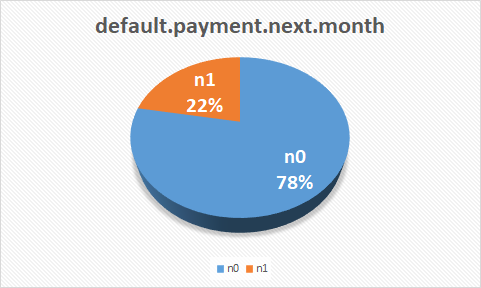


圖2. 繳費統計

**3. How to handle/deal with an imbalanced dataset**

*3.1 Re-sampling*

要使不平衡資料變得較平衡，將較少的資料放大或較多的資料縮小，使各類別屬性數量平衡，最常見的技術為Re-sampling，其分為兩種，over-sampling與under-sampling，以下分別介紹這兩種採樣技術與其常用的一些方法。

*3.1.1 Over-sampling*

Over-sampling是將資料集中少樣本類別的資料複製來增加少樣本類別資料的比重。以下是Over-sampling的幾個實作方法：

*3.1.1.1 Random oversampling*

隨機挑選少樣本類別中的樣本，並且複製樣本至資料集中，以增加類別比重。優點為不會造成任何訊息遺失，以及實務操作上較簡單；缺點為，該方法會放大部分重複資料而造成overfitting的後果。[4][5]

以下為示意圖：

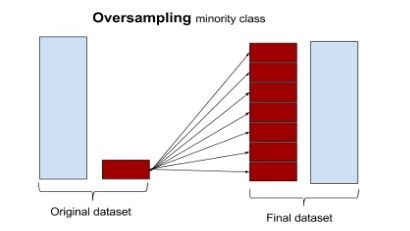


圖3. Random oversampling [2]

*3.1.1.2 Synthetic Minority Oversampling Technique (SMOTE)*

該方法為透過人工的方式增加少樣本類別的數量，基於特徵空間向量的概念創造新數據，實作的方法如下：

(1) 先選定一個少樣本類別的點，在空間中找出其最近的幾個點，並算出他們之間的向量差。

(2) 再將這些向量差乘上從0到1的隨機亂數。

(3) 將成出來的向量加回原來的點。

(4) 最後得到新的合成樣本，再將這些合成樣本標註為少樣本類別，最後丟回資料集中。[4][5]

過程如下圖所示：

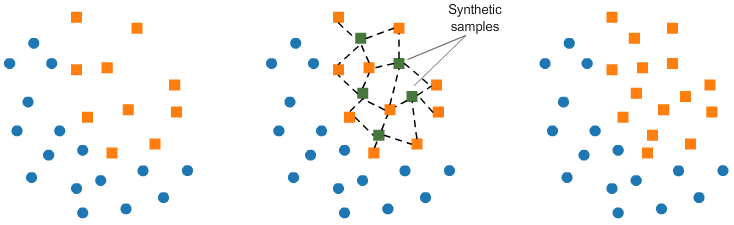


圖4. SMOTE [4]

由於產生的資料點為人工合成的新資料，其優點為改善Random oversampling複製原有資料而造成overfitting的問題。

*3.1.2 Under-sampling*

Under-sampling是將資料集中多樣本類別的資料刪除，以減少多樣本類別資料的比重來平衡資料集。以下是Under-sampling的幾個實作方法：

*3.1.2.1 Random undersampling*

從資料集的多數樣本類別中，隨機選取並丟棄樣本，以此來降低多數樣本類別所佔的比重。優點為降低訓練資料量，降低訓練模型時間、減少記憶體空間；缺點為將資料刪除會造成一些關鍵性資料的遺失。[4][5]

以下為示意圖：

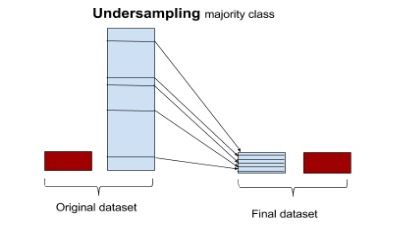


圖5. Random undersampling [2]

*3.1.2.2 Tomek link*

Tomek link的定義為兩個相鄰但是不同類別的點所產生的集合，這些點經常會造成分類的演算困難，而該方法是將Tomek link上的多樣本類別直接刪除，以增加兩個類別的邊界距離。這方法雖然不會使資料集變平衡，但是可以清理資料集，使得分類模型的計算效能提升。[4][5]

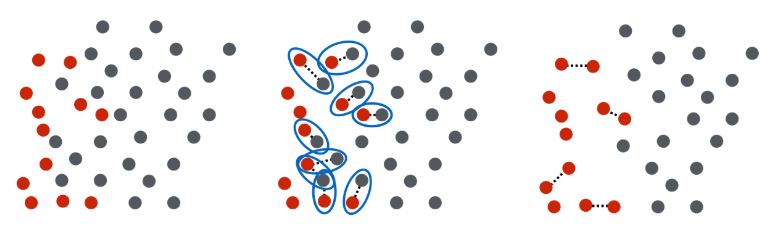


圖6. Tomek link [6]

*3.2 Cost sensitive learning*

大部分的分類會將錯誤分類的成本視為相同，但該方法評估錯誤分類所產生的成本來修正預測結果。此方法不會使資料集趨向平衡，而是利用各種情境下錯誤分類所產生的成本矩陣來觀察與調整不平衡學習的情況。[7]下圖為cost matrix的圖示：

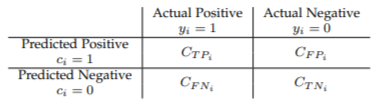


圖7. cost matrix [7]

*3.3 Algorithmic Ensemble Techniques*

該方法是利用不平衡資料集訓練多個符合資料集的分類器，再將多個訓練出來的分類器的分類結果集成起來，以提高最後的分類能力[8]，概念如下圖所示：

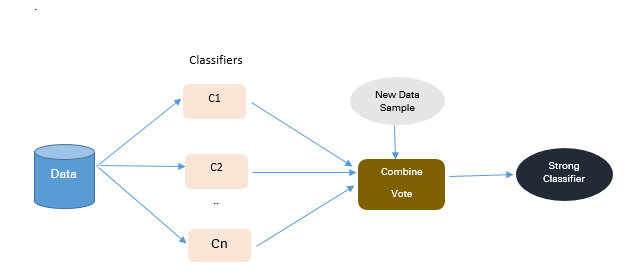


圖8. Approach to Ensemble based Methodologies [8]

Algorithmic Ensemble Techniques的實作方法4有以下兩種：

*3.3.1 Bagging-Based techniques*

將不平衡資料集隨機抽樣得到多個小的資料集，並訓練出不同的weak learners，最後將全部的weak learners進行組合得到strong learners。優點為減少overfitting發生的機率、提高機器學習算法的穩定性和準確性；缺點為如果一開始的learner不好的話，有可能連帶影響最後的結果不好。[8]

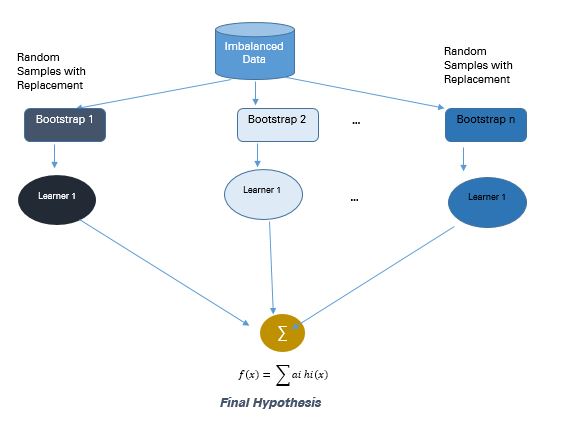


圖9. Approach to Bagging Methodology [8]

*3.3.2 Boosting-Based techniques*

該方法透過兩次學習的改善，最後在得出一個強分類器。先進行第一次學習，此次學習結果只比平均強一點，而第二次學習的weak learner會將第一次學習所誤分類的資料進行重點加強學習，最後再將二次學習的learners集結成一個strong learner。優點為加強Bagging-Based techniques的分類效果；缺點為，在noisy的環境下，二次學習的效果較容易偏頗，因為二次學習會加強對noisy的學習。[8]

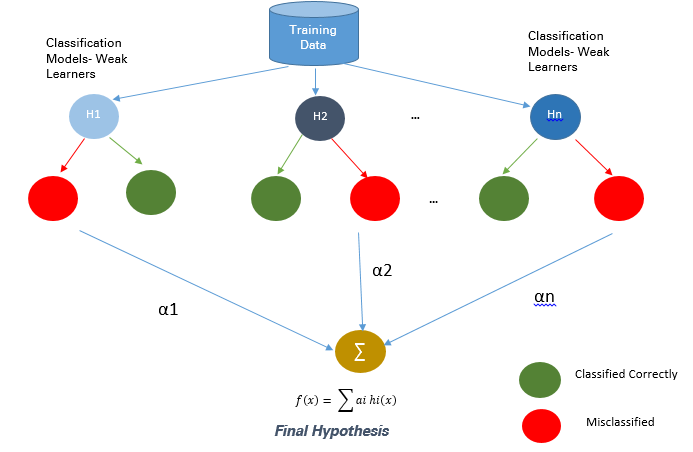


圖10. Approach to Boosting Methodologies [8]

**4. Quick survey on which algorithm(s) perform well/poorly on imbalanced datasets**

*4.1 Algorithm(s) perform well on imbalanced datasets*

Decision Tree對於imbalanced datasets有較佳的預測結果，因為Decision tree會將每一筆資料結果納入訓練的過程，分支到最後每一筆資料都會通向一個分類，對於資料集會有overfitting的結果，因此對一些少數資料類別比較容易分類出來，所以在imbalanced datasets會有較佳的訓練結果。[9]

Random forest是以Decision Tree為基礎的Ensemble learning演算法。Random forest集結多個Decision Trees而成，能夠更細微的看到每個特徵的重要程度，可以降低Decision Tree的誤差，提高準確度。[10]

*4.2 Algorithm(s) perform poorly on imbalanced datasets*

KNN是從一個類別未知的點與鄰近的點距離中，找出距離小的k個資料點，從k個資料點的類別中佔大多數者，預測為未知點的類別。因此，如果資料集為不平衡的話，會造成類別未知的點附近可能皆為類別樣本佔比較大的資料點，而影響最後預測的結果偏向佔比大的類別。[11]

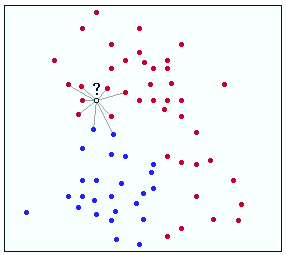


圖11. KNN

**5. Evaluation metrics for imbalanced datasets**

*5.1 Confusion Metric*

由Precision=TP/(TP+FP)可得知，所有預測為Positive(1)的正確比例；由Recall=TP/(TP+FN)可得知，所有真實資料為Positive(1)被正確預測的比例。F1-Score(F1) = 2\*Precision\*Recall/(Precision + Recall)，將Precision與Recall一起綜合考慮的加權平均值。[12]

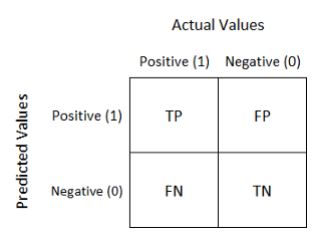
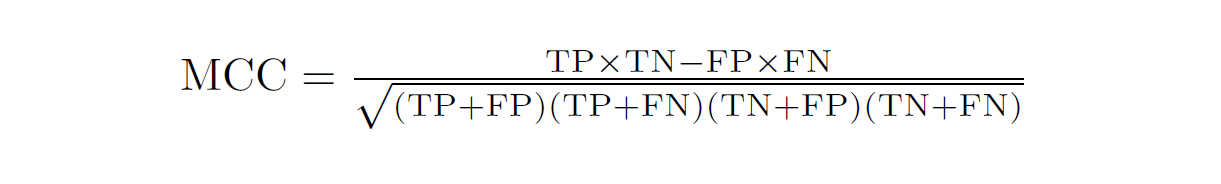


圖12. Confusion Metric

*5.2 Matthews Correlation Coefficient (MCC)*

因為將所有值納入計算，所以類別的規模大小不同(不平衡資料集)也可以使用。MCC值是代表真實類別與預測類別之間的相關係數，介於-1與1之間，1代表完全預測準確，-1代表預測完全不準確，0代表跟隨機預測準確度差不多。MCC的計算會使用到Confusion Metric中的值，其公式如下：[12]



*5.3 AUC - ROC Curve*

AUC - ROC Curve也是利用Confusion Metric中的值來做計算，繪畫出一個ROC Curve，坐標軸分別為TPR= Recall與FPR = FP/(TN+FP)，而曲線下的面積為AUC，代表分類預測準確的機率，其值介於0到1之間[13][14]，關係圖如下：

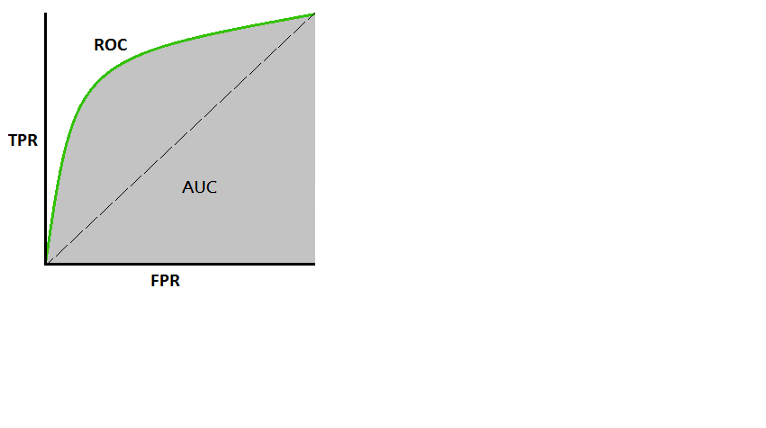


圖12. AUC - ROC Curve [14]

下面是AUC的幾個極端值是意圖，左邊的圖示TN、TP的分布，右邊的圖是AUC - ROC Curve。

* 這是AUC為1的時候，代表完全預測準確，TP、TN曲線完全不重疊。

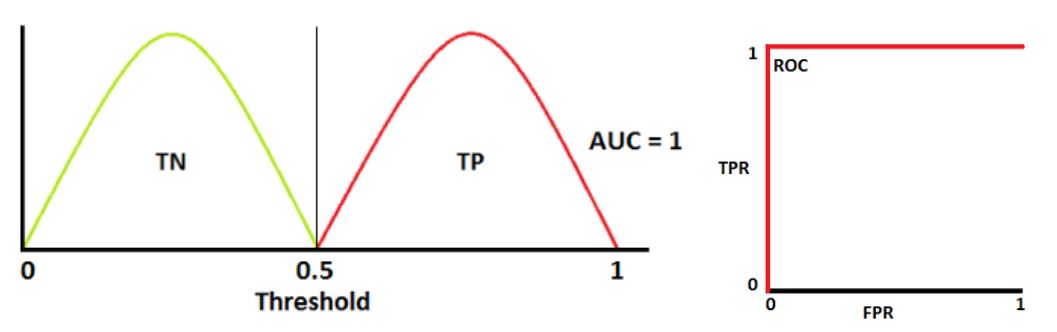


圖13. AUC - ROC Curve (AUG=1) [14]

* 這是AUC為0.7的時候，代表預測準確的機率為70%，TP、TN曲線部分有重疊。

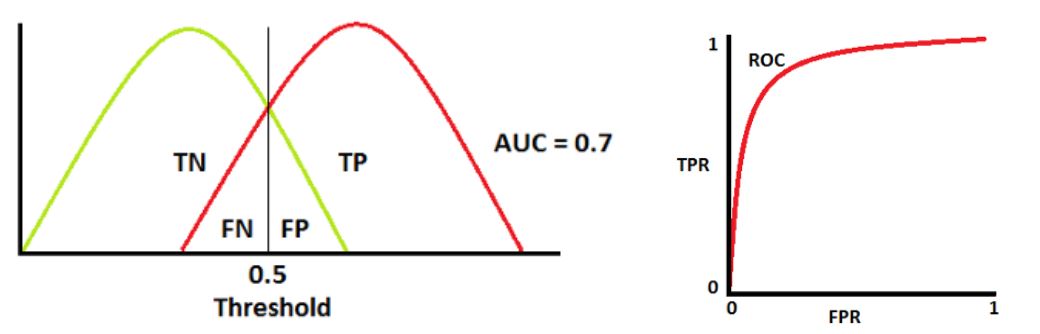


圖14. AUC - ROC Curve (AUG=0.7) [14]

* 若AUC為0.5，則TP、TN得曲線分布完全重疊，判斷正確的機率與隨機猜的一樣為50%，ROC曲線為TPR=FPR的一次方斜線。
* 若AUC為0，則TP、TN得曲線分布會相反，分布也完全不重疊，但預測正確的機率為0%，ROC曲線與AUC = 1時的ROC曲線鏡像相反(以TPR=FPR斜線為中心)。

**6. References / Citations**

[1] Pankaj Malhotra, 2015. What is an imbalanced dataset? Retrieve from: <https://www.quora.com/What-is-an-imbalanced-dataset> (March 19, 2020)

[2] Himanshu Tripathi, 2019. What Is Balanced And Imbalanced Dataset? Retrieve from: <https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5> (March 19, 2020)

[3] UCI Machine Learning, 2017. Default of Credit Card Clients Dataset. Retrieve from: <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home#UCI_Credit_Card.csv> (March 19, 2020)

[4]Prashant Banerjee, 2019. Data Preprocessing Project – Imbalanced Classes Problem. Retrieve from: <https://github.com/pb111/Data-Preprocessing-Project-Imbalanced-Classes-Problem> (March 19, 2020)

[5] [massquantity](https://home.cnblogs.com/u/massquantity/),2018. 机器学习之类别不平衡问题 (3) —— 采样方法。Retrieve from: <https://www.cnblogs.com/massquantity/p/9382710.html>

[6]David Huang, 2018. 不平衡資料的二元分類 2：利用抽樣改善模型品質。Retrieve from:

<https://taweihuang.hpd.io/2018/12/30/imbalanced-data-sampling-techniques/> (March 19, 2020)

[7]Alejandro Correa Bahnsen, Djamila Aouada, and Bjorn Ottersten, 2015. Ensemble of Example-Dependent, Cost-Sensitive Decision Trees.

[8]Guest Blog, 2017. Imbalanced Data : How to handle Imbalanced Classification Problems. Retrieve from: <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/> (March 19, 2020)

[9]Sebastian Raschka, Vahid Mirjalili(2018). Python Machine Learning, 2/e. (March 22, 2020)

[10]Yeh James, 2017.[資料分析&機器學習] 第3.5講 : 決策樹(Decision Tree)以及隨機森林(Random Forest)介紹。Retrieve from: <https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC3-5%E8%AC%9B-%E6%B1%BA%E7%AD%96%E6%A8%B9-decision-tree-%E4%BB%A5%E5%8F%8A%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-%E4%BB%8B%E7%B4%B9-7079b0ddfbda> (March 22, 2020)

[11] 佚名, 2018. KNN演算法（有監督學習演算法）。Retrieve from: <https://www.itread01.com/content/1546211528.html> (March 22, 2020)

[12] Igor Kuznetsov, 2019.Metrics for Imbalanced Classification. Retrieve from: <https://towardsdatascience.com/metrics-for-imbalanced-classification-41c71549bbb5> (March 19, 2020)

[13] 佚名, 2018.【機器學習】馬修斯相關係數（Matthews correlation coefficient）。Retrieve from: <https://www.itread01.com/content/1544982006.html> (March 19, 2020)

[14]Sarang Narkhede, 2018. Understanding AUC - ROC Curve. Retrieve from: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5> (March 22, 2020)