Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Economics and Finance 00 (2012) 000–000

Procedia
Economics and Finance

www.elsevier.com/locate/procedia

資料科學與機器學習報告

不平衡資料集之介紹與相關分析

呂雅琴[a], 蕭宸欣[b]

a 國立中央大學資訊管理學系 碩士班 108423013@cc.ncu.edu.tw
b 國立中央大學資訊管理學系 碩士班 108423025@cc.ncu.edu.tw

摘要

資料不平衡在資料科學領域有著相當程度的影響力，當人們關注的類別樣本遠少於其他類別樣本時，就會產生這個問題。本篇文章主要會透過介紹何謂不平衡資料集，以及舉例幾種常見的不平衡資料集處理方式和常見的演算法、衡量指標，進一步了解資料不平衡並加深對機器學習的認識。

關鍵字: 不平衡資料集、類別不平衡、SMOTE、SVM、Cost-Sensitive Learning

1. What is an imbalanced dataset?

一般而言，以二元分類的理想平衡資料集來說，兩類的樣本數應該差不多，而不平衡資料集則是指在一個資料集中，某一類的資料數遠大於另外一類，舉例來說，對於一個二元分類的資料集，若某一類的樣本百分比為 5%，而另一類的百分比為 95%，那我們在進行預測時，多數的演算法會為了提高準確率，而忽略了小類別的樣本，使得整體的準確率仍能高達 95%，但是若我們今天關注的是小類別樣本資料時，這項預測就會缺乏參考性。

現實生活中，大多數的資料都是屬於不平衡資料集，以自然災害為例，因為是屬於出現頻率低的現象，因此，若我們今天在乎的資料是屬於小類別(例如：森林大火)，那麼我們便無法對此資料進行精準的預測。因此在建構分類模型之前，需要先對資料不平衡的問題進行處理。

2. One imbalanced dataset example

2.1 資料集資訊(Dataset Information)

該資料集來自於 Kaggle 平台，紀錄了電信公司的客戶資料，共 7043 筆資料。資料集有 21 個屬性，包含客戶背景、電信服務、帳戶訊息等資料，目的是用這些屬性來預測客戶是否流失。

2.2 屬性資訊(Attribute Information)

變數:

- customerID
- 客戶背景資訊
  (例:gender, partner, ..., tenure 等 5 個欄位)
- 電信服務
  (例:PhoneService, InternetService, …, StreamingMovies 等 9 個欄位)
- 帳戶訊息
  (例:Contract, PaymentMethod, …, TotalCharges 等 5 個欄位)

目標變數:

- Churn： 上個月客戶是否流失。

V7046 =COUNTIF(U:U,"No")

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	customer	gender	Senior Ci	Partner	Depende	tenure	PhoneSe	Multiple	InternetS	OnlineSe	OnlineB	DeviceP	TechSup	Streamin	Streamin	Contract	Paperles	Payment	Monthly	TotalCha	Churn		
7029	0550-DC	Male	0	No	No	13	Yes	No	DSL	No	Yes	No	Yes	Yes	Yes	Month-to	No	Mailed c	73.35	931.55	No		
7030	9281-CE	Female	0	Yes	No	68	Yes	No	DSL	No	Yes	No	Yes	Yes	No	Two yea	No	Bank tra	64.1	4326.3	No		
7031	2235-DV	Female	1	No	No	6	No	No phon	DSL	No	No	No	No	Yes	Yes	Month-to	Yes	Electron	44.4	263.05	No		
7032	0871-OF	Female	0	No	No	2	Yes	No	No	No inter	No inter	No inter	No inter	No inter	No inter	Month-to	Yes	Mailed c	20.05	39.25	No		
7033	3605-JIS	Male	1	Yes	No	55	Yes	Yes	DSL	Yes	Yes	No	No	No	No	One yea	No	Credit ca	60	3316.1	No		
7034	6894-LF	Male	1	No	No	1	Yes	Yes	Fiber op	No	No	No	No	No	No	Month-to	Yes	Electron	75.75	75.75	Yes		
7035	9767-FF	Male	0	No	No	38	Yes	No	Fiber op	No	No	No	No	No	No	Month-to	Yes	Credit ca	69.5	2625.3	No		
7036	0639-TS	Female	0	No	No	67	Yes	Yes	Fiber op	Yes	Yes	Yes	No	Yes	No	Month-to	Yes	Credit ca	102.95	6886.3	Yes		
7037	8456-QD	Male	0	No	No	19	Yes	No	Fiber op	No	No	No	No	Yes	No	Month-to	Yes	Bank tra	78.7	1495.1	No		
7038	7750-EY	Female	0	No	No	12	No	No phon	DSL	No	Yes	Yes	Yes	Yes	Yes	One yea	No	Electron	60.65	743.3	No		
7039	2569-WC	Female	0	No	No	72	Yes	No	No	No inter	No inter	No inter	No inter	No inter	No inter	Two yea	Yes	Bank tra	21.15	1419.4	No		
7040	6840-RE	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	No	Yes	Yes	Yes	Yes	One yea	Yes	Mailed c	84.8	1990.5	No		
7041	2234-XA	Female	0	Yes	Yes	72	Yes	Yes	Fiber op	No	Yes	Yes	No	Yes	Yes	One yea	Yes	Credit ca	103.2	7362.9	No		
7042	4801-JZ	Female	0	Yes	Yes	11	No	No phon	DSL	Yes	No	No	No	No	No	Month-to	Yes	Electron	29.6	346.45	No		
7043	8361-LT	Male	1	Yes	No	4	Yes	Yes	Fiber op	No	No	No	No	No	No	Month-to	Yes	Mailed c	74.4	306.6	Yes		
7044	3186-AJ	Male	0	No	No	66	Yes	No	Fiber op	Yes	No	Yes	Yes	Yes	Yes	Two yea	Yes	Bank tra	105.65	6844.5	No		
7045																							
7046																						5174	
7047																						1869	

圖 1.電信公司之顧客資料 資料來源：https://www.kaggle.com/blastchar/telco-customer-churn

2.3 資料集不平衡的狀況

由上圖(圖 1)可知，7043 筆資料中，結果為 No 的資料有 5174 筆，Yes 的資料有 1869 筆，若我們在預測時，都將結果預設為"No"，準確率能高達 73.46%，但在現實生活中，電信業者關注的是客戶流失(Churn = Yes)，因此從這個現象我們可以發現，面對一個不平衡資料集，準確率並無法提供高度參考性。

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

3.1 調整抽樣方法 Resampling

3.1.1 oversampling

透過隨機複製少數類別中的樣本個數，新增進訓練集，來增加少數類別的代表性。
缺點 ： 1. 可能過度複製少數特徵，造成模型過度擬合(overfitting)。
2. 需要更多空間來儲存和訓練新增的資料。

- SMOTE (Synthetic Minority Oversampling Technique)

SMOTE 利用特徵空間中，對每一個小類別樣本 Xi，找出最近的 K 個小類別樣本，於 K 個鄰近點中隨機選取一個小類別樣本 Xj，並與 Xi 合成新樣本 X_{new}[2]。

$$X_{new} = Xi + (Xj – Xi) * \delta，\delta 為[0,1]隨機變數 \quad (1)$$

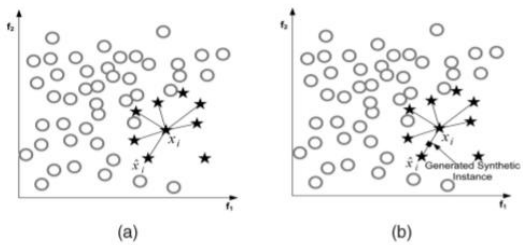

圖 2. 資料參考自[2]

重複上述步驟(圖 2)，透過小類別樣本之間的相似性來建立人工資料，讓資料集盡可能達到均衡的狀態，此方法能夠改善上述 oversampling 的過度擬合問題，並且提升機器學習的效能。
但是 SMOTE 仍有一些侷限性，最大的缺點是搜尋鄰近 K 點時，有可能會重複選取到某一個點 Xj，造成類別重疊度很高，而在這部分可以透過 Borderline-SMOTE 與 Adaptive Synthetic Sampling (ADASYN)改善。

3.1.2 undersampling

透過減少多數類別中的樣本個數，讓資料盡可能達成平衡。
缺點 ： 可能在刪除樣本的過程中，將具有代表性的資料刪掉。

- Tomek links

目的：刪除邊界上的大類別樣本
假設 Xi 與 Xj 分別屬於不同的類別，d(Xi, Xj)為兩點之間的距離，若找不到第三點 Z，使得 d(Xi, Z) < d(Xi, Xj)且 d(Xj, Z) < d (Xi, Xj)，便可稱(Xi, Xj)為 Tomek links，而將 Tomek links 中的大類別樣本刪除，最後就會得到所有小類別樣本的最近鄰點都是小類別樣本(圖 3)。
Tomek links 會將大類別中，過於靠近小類別的樣本視為雜訊並且刪除，目的在於清楚的劃清兩類別中的界線，讓我們較容易進行分類。

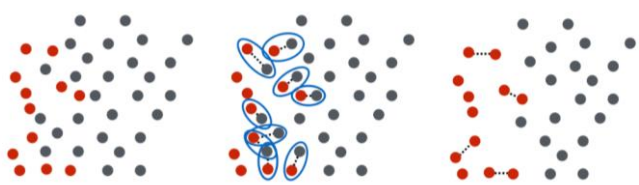

圖 3. 引用自 https://taweihuang.hpd.io/2018/12/30/imbalanced-data-sampling-techniques/

- Edited Nearest Neighbor (ENN)
  類似 Tomek links，若某一點鄰近的資料類別大多數都與自己的類別不同，便可將此點刪除。

- EasyEnsemble
  為一種非監督式演算法，透過隨機獨立採樣 N 次大類別中的樣本(採後放回)，採樣後的子集樣本數需相當於小類別樣本數，再將這 N 個子集與小類別樣本的數據組合成訓練集後，個別生成 N 個分類器，透過整合多個分類器的結果，取得一個最終的分類器。

- BalanceCascade
  為一種監督式演算法，作法類似於 EasyEnsemble，從大類別樣本中取樣，取樣後的個數也需相當於小類別樣本數量，將兩者組合產生第一個訓練集，生成第一個分類器(H1)後，對所有大類別中的樣本(X*)用 H1 進行分類，將預測正確的 X 刪去，將刪除完後的大類別資料再次取樣與小類別樣本生成 H2，以此類推，直到大類別樣本數<=小類別樣本數。

3.1.3 小結

我們可以利用 oversampling 與 undersampling 的互相搭配，生成更好的資料，例如使用 SMOTE + ENN，而值得注意的是，在做 undersampling 時，為了確保沒有移除到重要的資訊，要仔細的做「交叉驗證」(Cross Validation)。

3.3 GAN (Generative Adversarial Network)

GAN 為一種非監督式學習，簡單概念如下圖(圖 4)，可以想像成兩個角色之間互相的角力，一個是偽造者(Generator Network)，會不斷的製作假鈔，而另一個角色是警察(Discriminator Network)，會不斷的從偽造者那拿到假鈔判斷真偽，然後偽造者就會根據警察判斷的結果進行改良，雙方互相切磋最後製造出以假亂真的鈔票。

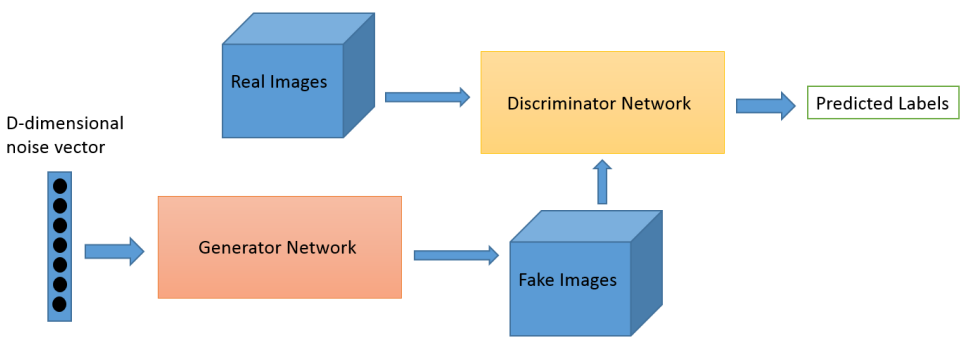

圖 4. 引用自 https://github.com/jonbruner/generative-adversarial-networks/blob/master/gan-notebook.ipynb

這對一些因為隱私考量而無法獲得大量資料的領域來說非常重要，因為 GAN 可以填補缺失的資料，將樣本數較少的資料作補償，解決資料不平衡的問題。

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

4. 1 Cost-Sensitive Learning(CSL)

CSL 的目的是建立最小總成本模型，著重在替大類別和小類別的錯誤分類分配不同的成本，透過使用不同的成本矩陣(表 1)來學習資料不平衡的問題，使模型能關注小類別樣本。

而使用成本矩陣的概念可以幫助我們理解何謂錯誤分類成本，使用成本矩陣為每個單元分配成本，如下表所示，其中 C(i,j)表示對實際例子進行分類的錯誤分類成本，「i」為預測分類,「j」為真實分類：

表 1. 成本矩陣

Cost Matrix		**True Answers**	
		Positive	**Negative**
Prediction	**Positive**	**C(1,1) True Positive**	**C(1,0) False Positive**
	Negative	**C(0,1) False Negative**	**C(0,0) True Negative**

一般而言，雖然小類別樣本數量較少，但是小類別樣本錯誤分類的代價比較高，因此可以透過調整成本的加權，例如賦予小類別樣本更大的錯誤分類成本迫使模型更關注小類別樣本的分類準確率，以處理資料不平衡的問題。

4. 2 SVM (Support Vector Machine)

常見的二元分類器，為監督式學習，目的是希望能在一個混合的資料集中，找到一個最佳的超平面(hyper plane)，將兩個不同類別的資料分開來，然而，SVM 通常會傾向於追求準確率最佳化，這對不平衡資料集來說，是比較缺乏參考性的，但是可以透過以下方法來修正：

- one-class SVM

  只利用其中一類的樣本來訓練，透過這些樣本特徵尋找出一個合適的超球面，再去判斷測試資料是否與訓練數據相似，若超出邊界則為不同類資料。假設今天我們利用大類別進行訓練，得到一個超球面能將我們的樣本大致上都包在其中，在測試時，若預測結果落在圈內，便是大類別，落在圈外，便是小類別。

  在 one-class classification 中，只有一類的資料會做為訓練集，另一類的資料可以視為 outlier，但嚴格上來說，這並不是一種 outlier detection，而是一種 novelty detection 方法，目的在找出與大多數資料有偏差行為的個體，例如，在詐欺檢測中，信用卡公司會判斷消費者的購物習慣，若今天消費者的購物行為有偏差時，信用卡公司便能立即通知用戶，確認消費者身分是否被盜用。

- CS – SVM (Cost-Sensitive SVM)

  利用 CS-SVM 分類時，依據前面提到的 Cost-Sensitive Learning(CSL)賦予少數類樣本更大的錯誤分類成本，以調整權重的方式來處理資料不平衡的問題。

$$\min \frac{1}{2}\| w \|^2 + C^+ \left(\sum_{y_i=+1} \xi_i \right) + C^- \left(\sum_{y_i=-1} \xi_i \right) \qquad (2)$$

4.3 Random Forest

隨機森林是以隨機的方式建構一個森林，而森林是由多個決策樹所組成，隨機森林的每一顆決策樹彼此之間沒有關聯，其目的為藉由結合多個決策樹創造出好的預測能力。也因此隨機森林的結果會是以多數決的概念呈現，在追求高準確度的過程中容易傾向大類別樣本，這對不平衡資料集的狀況，是比較有疑慮的，不過可以透過以下方法來修正：

- Balanced Random Forest
  若只針對大類別樣本去採 undersampling 或是對小類別樣本採 oversampling 的方式，雖然可能會有所幫助，但也很有可能會因為 undersampling 而刪除一些重要的資料，因此平衡隨機森林的演算法對此做了一些修正，演算法的步驟如下：
  1. 對隨機森林每次的迭代會從小類別樣本中生出一個 boostrap sample，再從大類別樣本中多次隨機獨立抽取相同數量的樣本
  2. 讓分類樹長到最大的大小後再使用 CART 修剪樹
  3. 重複上述兩個步驟直到合併出最終的分類器便可對資料預測。

- Weighted Random Forest
  使用 Weighted Random Forest 分類時，會依據前面提到的 Cost-Sensitive Learning(CSL)賦予小類別樣本更大的錯誤分類成本，以調整權重的方式來處理資料不平衡的問題。

5. Evaluation metrics for imbalanced datasets

5.1 衡量指標

如果以準確度作為分類不平衡資料集的衡量指標，模型會為了追求高準確度，在學習時傾向多數樣本，導致模型預測結果雖然具有高準確度，卻無法學習到少數樣本、不太實際的狀況，稱之為「準確度悖論(Accuracy Paradox)」。

因此在不平衡資料集的分類問題中通常會改用別的衡量指標，例如:Precision、Recall、F-measure、ROC 曲線/AUC、PR 曲線等。介紹常用的衡量指標前，要先對混淆矩陣(Confusion Matrix)有個初步的認識，混淆矩陣是分類模型中很重要的衡量機制，如下表(表 2)所示:

表 2.混淆矩陣

Confusion Matrix		**True Answers**	
		Positive	**Negative**
Prediction	**Positive**	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

- $\text{Precision} = \frac{TP}{(TP + FP)}$, 代表預測為 positive 的樣本中，真實為 positive 的比例
- $Recall = \frac{TP}{(TP + FN)}$, 代表真實為 positive 的樣本被預測中的比例

通常在不平衡資料集中，我們關注的是小類別樣本是否真正被識別出來，因此與 Accuracy 比起來，Recall 值與 Precision 值若能越高，代表分類器的效能越符合我們的需求，所以我們會選擇使用 F-measure 作為衡量指標。

- F-measure 算法如下：

$$F - measure = \frac{(\alpha^2 + 1) * precision * recall}{(\alpha^2) precision + recall} \quad (3)$$

α = 1 時代表 P、R 的重要程度相當(此時 F 值為 P、R 的調和平均)，α > 1 時 (通常使 α = 2)，代表 R 的重要性比較高，反之，α < 1 (通常使用 α = 0.5)時代表 P 的重要性較高。

5.2 模型效能評估

5.2.1 ROC 曲線

其中 ROC 曲線以 FPR 為 X 軸，TPR 為 Y 軸，紀錄在各決策門檻下（decision threshold），FPR(False Positive Rate) 與 TPR(True Positive Rate)的變化，最後形成一條曲線(圖 5)。

- FPR = FP / (FP + TN) ，若 FPR 越低，代表模型能夠正確判斷 Negative 樣本，表現越好
- TPR = TP / (TP + FN) ，若 TPR 越高，代表模型能夠正確判斷 Positive 樣本，表現越好

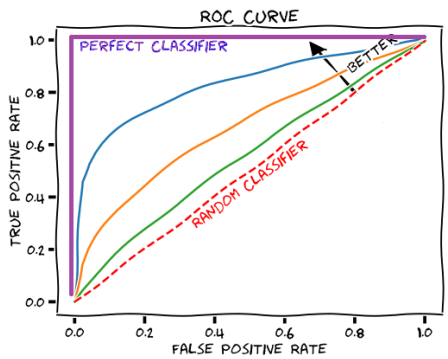

圖 5. 引用自 https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-

因此，若 ROC 曲線越接近左上方，效能越好。

- AUC(Area under curve)
  為 ROC 曲線下所覆蓋的面積，若 AUC 面積越大，表示模型預測效果越好。

5.2.2 PR 曲線

在各決策門檻下（decision threshold），紀錄 Recall 與 Precision 的變化，最後形成一條曲線(圖 6)。和 ROC 相同，PR 曲線下的面積可代表模型在各 decision threshold 下的表現，若 Precision = Recall = 1，我們視他為完美預測，因此我們的 PR 曲線如果能越靠近右上角，代表整體的模型表現較佳。

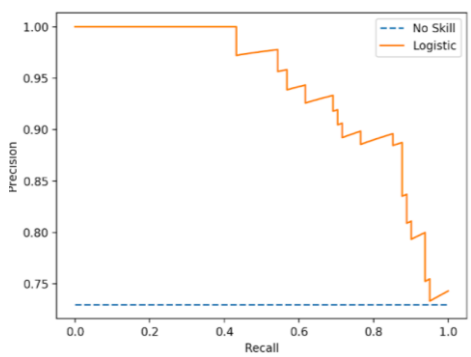

圖 6. 引用自 https://medium.com/nlp-tsupei/roc-pr-%E6%9B%B2%E7%B7%9A-f3faa2231b8c

5.2.3 ROC curve 與 PR curve 比較

由於 PR 曲線只有考慮 Positive 是否被正確預測，而 ROC 涵蓋了 Positive 和 Negative，因此在不平衡資料的情況下，PR 曲線會是較好的選擇，而 ROC 曲線在類別較平均的情況下較適合。

6. 結語

現實生活中有很大的機率會出現資料不平衡的狀況，而在面對這些資料時需要格外小心，才不會在處理資料時造成過度擬合或是刪除重要資料，而每一種演算法都各有優缺，我們也發現若只用單一個演算法，在訓練過程中並沒有辦法很好的處理這種狀況，即使是 SVM 也需加入一些條件才能更好的建立模型，因此若能結合幾種演算法或是設定好條件、權重，根據資料集選擇合適的資料前處理方式，如此一來便能透過良好的訓練得到更佳的結果。

References

[1]Kaggle, Telco Customer Churn。檢自 https://www.kaggle.com/blastchar/telco-customer-churn, (Mar.18, 2020)
[2] Haibo He, Member, IEEE, and Edwardo A. Garcia, “Learning from Imbalanced Data”, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009*
[3] 不平衡資料的二元分類 2：利用抽樣改善模型品質。檢自 https://taweihuang.hpd.io/2018/12/30/imbalanced-data-sampling-techniques/, (Mar. 18, 2020)
[4] 不平衡資料分類演算法介紹與比較。檢自 https://www.itread01.com/content/1541252823.html, (Mar.18, 2020)
[5] 何謂生成對抗網路? 聽聽頂尖研究員怎麼說。檢自 https://blogs.nvidia.com.tw/2017/05/generative-adversarial-network/, (Mar.19, 2020)
[6] Cost-Sensitive Learning for Imbalanced Classification。檢自 https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/,
[7] Zhang Sheng*, Wang Wei, Huang Xiuli , Shang Xiuyu, “Optimizing the Classification Accuracy of Imbalanced Dataset Based on SVM”, *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*
[8] Zhuang, L., & Dai, H. (2006). “Parameter Estimation of One-Class SVM on Imbalance Text Classification.” Lecture Notes in Computer Science”, 538–549. Advances in Artificial Intelligence
[9] Python 機器學習筆記——One Class SVM。檢自 https://www.itread01.com/content/1557561483.html, (Mar.20, 2020)
[10] Chen, Chao, Andy Liaw, and Leo Breiman. "Using random forest to learn imbalanced data." University of California, Berkeley 110.1-12 (2004): 24.
[11] 机器学习性能评估指标。檢自 http://dingby.site/2018/03/07/%-E6%9C%BA%E5%99%A8%E5%AD%A-6%E4%B9%A0%E6%80%A7%E8%83%BD%E8%AF%84%E4%BC%B0%E6%8C%87%E6%A0%87/, (Mar.20, 2020)
[12] 深入介紹及比較 ROC 曲線及 PR 曲線。檢自 https://medium.com/nlp-tsupei/roc-pr-%E6%9B%B2%E7%B7%9A-f3faa2231b8c。, (Mar.20, 2020)
[13] 不平衡資料的二元分類 1：選擇正確的衡量指標。檢自 https://taweihuang.hpd.io/2018/12/28/imbalanced-data-performance-metrics/, (Mar.20, 2020)