國立臺灣大學生農學院生物產業機電工程學系

碩士論文

Department of Bio-Industrial Mechatronics Engineering
College of Bioresources and Agriculture
National Taiwan University

Master Thesis

基於分群集成技術的非平衡學習應用於預測非編碼區變異的致病性

Clustering Ensemble Based Imbalanced Learning for Predicting Pathogenic Non-coding Variants

莊凱文

Kai-Wen Chuang

指導教授：陳倩瑜 博士

Advisor: Chien-Yu Chen Ph.D.

中華民國 108 年 7 月

July 2019

doi:10.6342/NTU201903179

國立臺灣大學碩士學位論文
口試委員會審定書

基於分群集成技術的非平衡學習應用於預測非編碼區變異的致病性

Clustering Ensemble Based Imbalanced Learning for Predicting Pathogenic Non-coding Variants

本論文係莊凱文君(R06631033)在國立臺灣大學生物產業機電工程學系、所完成之碩士學位論文，於民國 108 年 07 月 03 日承下列考試委員審查通過及口試及格，特此證明

口試委員：陳倩瑜 (簽名)

(指導教授)

吳君泰　　蔡懷寬

系主任、所長 陳林祈 (簽名)

doi:10.6342/NTU201903179

致謝

回想當初從電機系的控制實驗室跨到生物資訊的 C4 Lab，沒有生物背景的我在剛加入實驗室時連分子生物的中心法則是什麼都不知道，但在經歷暑期課程的淬鍊後，開始漸漸能夠聽得懂一些生物的專有名詞，後來加入了翊安學長主持的讀書會，除了對機器學習有更多的認識之外，也練習使用一些生物的軟體工具，並在我每次有問題時總是不厭其煩的與我討論和分享，除了讀書會外，實驗室的學長姊和同學們在各方面總能互相扶持，不管是在課堂、研究、生活上都給予莫大的幫助，感謝東祈和蔡醫師在生物知識上給予的幫助，當我看論文有瓶頸時能替我解答，感謝文策在運算資源上的幫助，讓我有資源和時間能跑更多的程式，感謝文廷和王賀平常的打屁聊天以及一起在課堂上的奮鬥，讓研究的生涯不無聊不孤單，能加入 403 這一個大家庭是一件很幸運的事，喔不，現在已經變成 304 了。而最需要感謝的還是倩瑜老師了，在這兩年間，我能不斷感覺到老師對於研究的熱情，並且有耐心的與我討論，從最初的題目發想、實驗設計到最後的結果，老師總是能給予適當的建議引導我走在正確的方向上,才能順利的完成這篇論文，謝謝老師。

另外我也要謝謝家人的支持，沒有催促我趕快進入職場，而是讓我自己選擇我未來的道路，同時也讓我在求學期間不會有經濟上的負擔，能將心力全部專於學習和研究上，順利的完成碩士的學業，也謝謝一直陪伴著我的女友，在我心情低落的時候，總能逗我開心，在我畏懼時，總能讓我鼓起繼續向前的勇氣，最後再次謝謝一路上曾經幫助過我的人，謝謝大家!

doi:10.6342/NTU201903179

摘要

在次世代定序以及全基因組定序漸漸普及的情況下，已經在全人類的基因組中發現了數千萬個基因變異，其中大部分的基因變異集中在非編碼，這些發生於非編碼區的基因變異可能會導致基因的調控機制產生改變，進而導致疾病產生。然而，實際上會影響人體基因功能進而造成疾病的變異僅佔非常少數，所以要如何在這麼大量的變異中去找出與疾病有相關聯的變異是個很大的挑戰。

近年來已經有許多機器學習的方法用於預測人類基因組中的致病變異，但當非致病變異數量上升時，意味著資料集的正/負(致病/非致病)樣本間的比例變大，分類器的精確率和召回率明顯下降，為了讓分類器在不平衡資料集下的預測效果能有效的提升，本研究開發出一種基於分群集成 (Clustering Ensemble，CE)採樣技術和 Hyper-ensemble 集成方法的機器學習框架：CE-SMURF，改善一般機器學習演算法在學習不平衡資料集時效果不佳的問題，並應用於預測非編碼區的致病變異。

關鍵字：分群集成、非平衡資料、非編碼區變異、致病性、機器學習

doi:10.6342/NTU201903179

Abstract

With the help of Next Generation Sequencing (NGS) and whole-genome sequencing (WGS), many variants in the non-coding regions were found in the human genome, but the ensured pathogenic variants were only a minority. It is a challenge to find pathogenic variants from such a large number of non-coding variants. Recently, a method, HyperSMURF, was previously proposed to tackle this problem by using both sampling and over-sampling techniques to balance the data. Through reproducing the analytic results of HyperSMURF, we observed that this approach might generate samples that did not help with training in minority or reduced the samples that might benefit training in majority. In this regard, this study aims at presenting a machine learning framework, CE-SMURF. The CE-based (Clustering Ensemble-based) method is used to find the samples of the center in majority and the samples of the boundary in minority, and then use the resampling technique to balance the ratio of data. Moreover, in order to improve the learning performance, we used the ensemble method to build multiple models, and computed the final scores by averaging the probability of variants in each model. It is found that CE-SMURF can significantly improve the performance of the predicting non-coding pathogenic variants.

Keywords: clustering ensemble、imbalanced data、non-coding variant、pathogenic、machine learning

doi:10.6342/NTU201903179

目錄

doi:10.6342/NTU201903179

doi:10.6342/NTU201903179

圖目錄

doi:10.6342/NTU201903179

表目錄

doi:10.6342/NTU201903179

第一章 研究目的

次世代定序(Next Generation Sequencing，NGS)的出現使得定序價格開始下降，並且讓全基因組定序(Whole Genome Sequencing，WGS)變的更加普及，科學家們對於人類體內微小卻複雜的基因世界有更進一步的認識。在過去探討疾病與變異相關的研究中，大部分的研究著重於探討基因組中的編碼區域，因為編碼區的序列能夠藉由轉錄和轉譯的過程變成蛋白質，兩者之間存在著直接的關係，然而藉由全基因組關聯分析（Genome-wide association study，GWAS）所找到的變異中，絕大多數都位於基因組中與調控功能相關的非編碼區域[1]，表示其實非編碼區與疾病之間也存在著很大的關連性，這些發生於非編碼區內的變異可能會造成調控功能的改變，進而導致疾病的發生。直到目前已經在全人類的基因組當中找到上千萬個位於非編碼區內的變異，但在這麼多的非編碼區變異之中，僅會有相當少數的變異會真正造成功能性的影響[2]，因此要如何在這麼大量的變異中去找出與疾病有相關聯的變異是個很大的挑戰。

近年來，在預測致病變異相關的研究中，已有許多文獻提出了各種不同的機器學習方法來預測可能的致病變異，像是 CADD [3]、DANN [4]、Eigen [5]、GWAVA [6]和 HyperSMURF [7]等，然而隨著非致病性變異數量的上升，訓練集間的正負樣本比例上升，分類器的精準率和召回率明顯下降，表示一般的機器學習演方法在學習不平衡資料時有很大的限制與阻礙，在上述提到的預測致病變異方法中，僅有 GWAVA 和 HyperSMURF 有針對不平衡資料集的情況提出改進的方法與討論，GWAVA 使用了 Random sampling 的方式來減少負樣本，而 HyperSMURF 除了使用 Random sampling 之外，也使用 SMOTE [8]演算法增加正樣本的數量，藉此平衡資料集的比例，這種利用採樣技術來平衡資料集的方法使得分類器不再偏好預測數量較多的負樣本，但同時也產生了一些問題，像是在使用 Random sampling 時，可能會刪除負類分群邊界樣本，縮小負類樣本的區域，而在使用 SMOTE 增加正類的

doi:10.6342/NTU201903179

樣本時，可能會產生對於訓練沒有幫助的正類分群中心樣本，雖然平衡了正負樣本之間的數量，但卻可能也因此改變資料集的資料特性，影響到分類器的學習效果。為了改善這個問題，Chen Si et al. 提出了分群集成採樣(Clustering Ensemble Sampling) [9]，此採樣方法利用分群集成的概念來找出位於分群的中心和邊界區域，接著對負類的分群中心樣本做 CE-Under 欠採樣減少負類樣本的數量，另外對正類分群邊界的樣本做 CE-SMOTE 過採樣增加正類樣本的數量，這樣不但能有效平衡資料集正負樣本的數量，同時也能保有類似於原本資料集的資料特性。

參考上述文獻所提到的方法，本研究開發了 CE-SMURF 的機器學習框架，此框架使用了分群集成採樣技術和 Hyper-ensemble 集成方法來改善一般機器學習演算法在學習不平衡資料集所遭遇的問題，藉由分群集成採樣來縮小資料間的比例，再透過 Hyper-ensemble 的概念，訓練多個 Random Forest [10]的分類器並結合這些分類器的預測結果，為了證明 CE-SMURF 在預測表現上有所提升，將同有使用採樣技術的 GWAVA 和 HyperSMURF 進行預測表現的比較。

doi:10.6342/NTU201903179

第二章 文獻探討

本章節將分成四個部分，第一部分將介紹非編碼區變異，第二部分介紹本研究使用到的變異資料庫，第三部份介紹 HyperSMURF 機器學習框架以及所用到的技術，第四部分介紹分群集成採樣的概念。

2.1 非編碼區變異 (Non-coding variant)

非編碼區(Non-coding region)指的是 DNA 序列中那些不包含製造蛋白質訊息的序列或是產生不能轉譯成蛋白質的 RNA 的序列，且非編碼區在人體基因組中佔了超過 90%的部分。過去的非編碼區 DNA 被視為是垃圾 DNA，但隨著時間的推移，科學家慢慢發現到非編碼區內的 DNA 能夠透過調控機制影響轉錄或是轉譯的過程，並將這些能夠調控 DNA 的非編碼區序列稱為調控元件(Regulatory element)，包括：啟動子(promoter)、增強子(enhancer)、沉默子(silencer)和絕緣子(insulator)，啟動子可以被 RNA 聚合酶辨認，並與轉錄因子(Transcription factor)結合，開始基因的轉錄，增強子能夠強化基因的表現，而沉默子則相反，可以降低基因的表現，絕緣子能阻止附近調控元件對於基因的調控，如圖 2-1 所示，透過這些調控元件的交互作用，使得不同的細胞在相同的基因下能夠有不同的表現。

doi:10.6342/NTU201903179

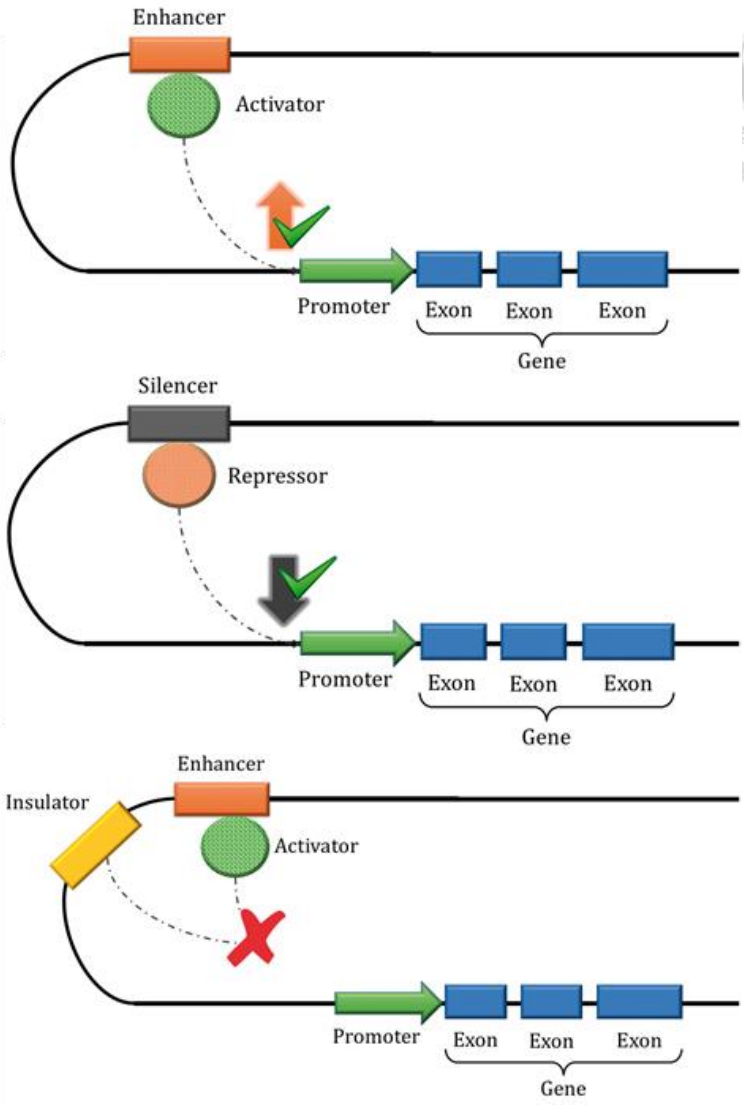

圖 2-1 不同調控元件的交互作用，摘自文獻[11]

在同一生物族群的基因組中，每一個體的基因都會有些微的不同，這些不同來自於基因組中序列的改變普遍稱之為變異(variant)，變異的原因非常多，可能是所在環境誘導發生，也可能是自然產生。基因的變異是演化發生的推手，變異的產生使得每一個生物個體的表型(phenotype)有所差別，這種差別促使著演化的發生，某些變異能夠對生物體帶來幫助，相反的，也存在對生物體造成傷害的變異，原因是變異若發生於編碼區內的話可能導致改變或是喪失遺傳訊息，而若是發生於非編碼區內的話可能會影響調控功能。

在人類的基因組中，通常會有數百萬個變異發生，且絕大部分位於非編碼區，

doi:10.6342/NTU201903179

發生於非編碼區內的變異可能會影響調控元件的調控功能，進而導致調控元件的目標基因的表現能力有所變化，但在這麼多的變異之中，僅會有非常少數比例的變異會造成調控功能的影響[2]，這些少數變異在剖析疾病產生的原因中佔有很重要的角色，若是能夠在大量的非編碼區變異中篩選出可能會致病的變異，對於未來疾病的治療或是預防都會有很大的進展和幫助。

2.2 資料庫

2.2.1 HGMD (Human Gene Mutation Database)

HGMD [12]是於 1996 年所建立的一個人類基因突變資料庫，其致力於蒐集與疾病相關的已知人類基因病異，且提供公開版和專業版兩種版本，公開版的資料主要免費提供給學術或非營利機構做研究，但公開版的資料僅提供已收錄超過 3 年的數據。截至目前為止，公開版的 HGMD 總共收錄了 171,397 個基因變異，其種類包括：錯異 (Missense)/非錯異(Non-Missense)、調控(Regulatory)、小片段刪除 (Micro-deletion)、小片段插入 (Micro-insertion)、重複序列 (Repeat variations) 以及複雜基因重組 (Complex rearrangements)等。

2.2.2 ClinVar

ClinVar [13]資料庫是人類基因變異和臨床醫療資訊間的橋梁，收錄了疾病與人類基因組中變異的關係，紀載了非常多臨床變異的詳細資訊。特別的是，ClinVar 中的資料是能夠供不同的研究者自行上傳的，為了能夠識別資料的可信度，除了透過系統自動比較同一個變異中不同上傳者的結果之外，ClinVar 也擁有自己組成的委員會來驗證這些不同資料來源的可信度，並將其分成不同等級的 review status，等級越高則代表資料的可信度越高，其詳細資訊如表 2-1：

doi:10.6342/NTU201903179

表 2-1 ClinVar review status

Number of gold stars	Review status
* * * *	practice guideline
* * *	Reviewed by expert panel
* *	criteria provided, multiple submitters, no conflicts
*	criteria provided, conflicting interpretations
	criteria provided, single submitter
None	no assertion for the individual variant
	no assertion criteria provided
	no assertion provided

2.2.3 1000 Genomes Project

1000 Genomes Project [14]是個於 2008 年 1 月啟動的跨國際研究計畫，由來自世界各國的研究團隊共同執行，該計畫旨在尋找位於人體基因組中族群攜帶率超過 1%的變異，且因為定序技術的進步，大大降低了定序的價格，使得 1000 Genomes Project 能在有限資源下對更多各地區不同人種的基因組進行定序，並分析這些定序資料，找出位於人體中與疾病相關聯的未知遺傳變異，蒐集並建立一個公開的人類基因變異資料庫，使研究者能夠在短時間內大量取得人類基因中常見或稀有的變異，對於人類變異相關研究有著非常大的貢獻。

doi:10.6342/NTU201903179

2.3 HyperSMURF

過去在利用機器學習方法預測致病變異時，大部分都是使用平衡的訓練集來做訓練，意味著正負樣本之間有著差不多的數量，但隨著越來越多非致性的變異被找到，使得能夠用於訓練的負樣本數量大幅增加，正負樣本之間的數量差距越來越大，在這樣的情況下，導致一般的機器學習在訓練分類器時效果不如預期，為了改善這個問題，Schubach et al. 在 2017 年提出了一種基於不平衡感知的機器學習框架：HyperSMURF (Hyper-ensemble of SMOTE Undersampled Random Forests) [7]，其整體架構如圖 2-2 所示，下面將分別介紹 HyperSMURF 所使用到的技術以及運算流程。

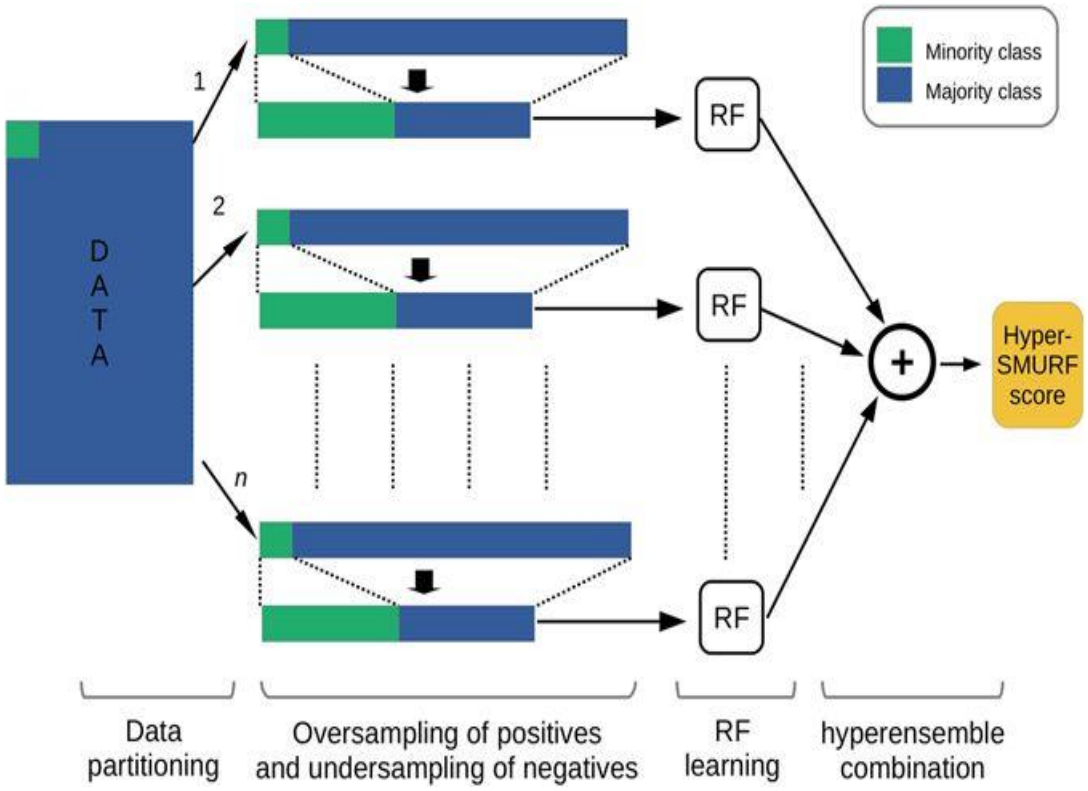

圖 2-2 不平衡感知機器學習框架，摘自文獻[7]

doi:10.6342/NTU201903179

2.3.1 採樣技術

在 HyperSMURF 中為了平衡正負樣本的數量，分別使用欠採樣(Under-sampling)減少負樣本的數量以及過採樣(Over-sampling)增加正樣本的數量。在欠採樣的部分，使用 Random sampling 來減少負樣本數量，Random sampling 是一種常見的欠採樣方法，因為其容易實現，僅需透過隨機挑選樣本的方式便能來達到減少樣本的目的，而在過採樣的部分，使用 2002 由 Chawla 提出的 SMOTE [8] 過採樣算法來增加正樣本的數量，SMOTE 與一般的重複採樣技術不同的地方在於，一般重複採樣是產生重複性的樣本，而一味的產生重複性的樣本可能會造成樣本間的多樣性不足，雖然增加了樣本的數量，但卻沒有增加資料的廣度，而 SMOTE 改善了這個問題，不但能夠產生與原樣本不同但卻相似的新樣本，同時也能增加樣本數量，其原理是透過選定一個樣本，接著在選定樣本的附近 k 個樣本中隨機選取一個樣本，並在兩個樣本的特徵向量直線上隨機選取一個點當作新樣本，如圖 2-3，不斷的重複上述的步驟，以產生指定的樣本數量，藉由 SMOTE 算法能夠讓正樣本數量接近負樣本的數量，這樣的過採樣方式能夠讓正樣本更多樣性，使得模型在做預測時能考慮更多樣與更多數量的正樣本。

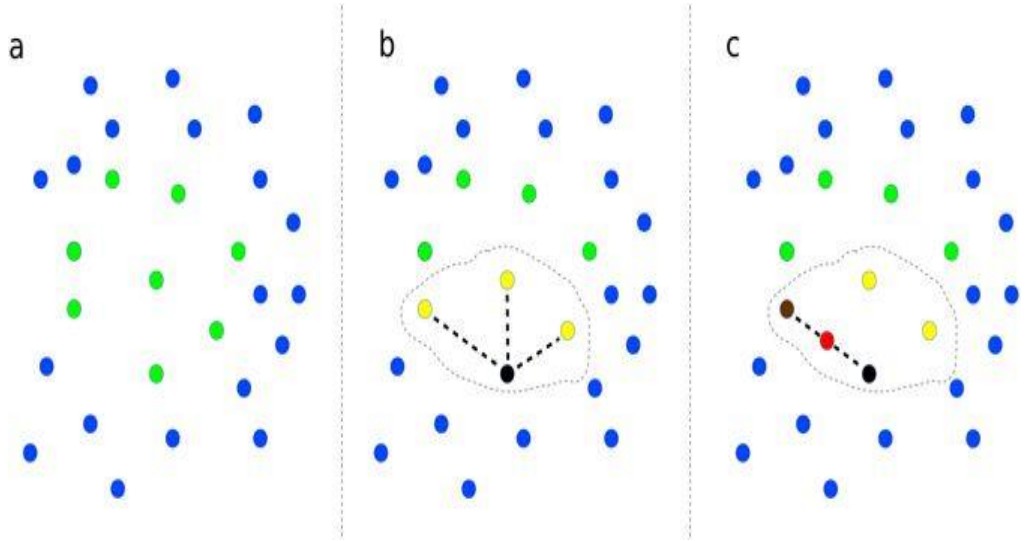

圖 2-3 SMOTE 演算法，摘自文獻[7]

doi:10.6342/NTU201903179

2.3.2 Hyper-ensemble

在分類器的部分，HyperSMURF 選擇使用 Random Forest 來做為分類器的學習演算法，Random Forest 是 2001 年由 Breiman 提出的一種機器學習演算法[10]，因其在算法的兩處採用隨機的方式而得其名。Random Forest 擁有許多的優點，不需要對資料做正則化，訓練速度快，能夠處理大量的輸入變數，並在分類時，評估這些變數的重要性，以選取最適合的變數來進行分類，且因為 Random Forest 是使用 Ensemble 的概念由多個 Decision Tree 所組成，透過多個決策數的判斷來減少單一 Decision Tree 造成的過擬和，並增加整體模型預測的準確度。

而在正負樣本數量差距很大的情況下，使用欠採樣方法來平衡訓練集時，可能會導致負樣本喪失與原本樣本類似的資料特性，為了彌補這個問題，作者利用 Ensemble 的方式，將訓練集的樣本平均分到 n 個不同的子訓練集，並分別訓練不同的 Random Forest 模型，最後在平均每個模型的結果，且因為 Random Forest 的演算法原本就是使用到 Ensemble 的概念，將多個決策樹合在一起成為一個森林，所以作者將這個 Random Forest 的 Ensemble 方法稱為 Hyper-ensemble。

2.3.3 Pseudocode

圖 2-4 為 HyperSMURF 的 pseudocode，*P* 為正樣本集合，*N* 為負樣本集合，並假設 N 數量遠大於 P，首先將負樣本集合 *N* 平均分配到 *n* 個獨立的子資料集，並在每個子資料集中加入一樣的正樣本集合 *P*，此時的每個子集合中的正負樣本還是處於數量不平衡的狀態，所以 HyperSMURF 使用 SMOTE [8]演算法增加正樣本的數量，同時用 Random sampling 的方式減少負樣本的數量，使得正負樣本數量平衡，而在 SMOTE 的算法部分，HyperSMURF 能透過設定 *k* 值來決定 SMOTE 演算法要尋找的最近 k 個樣本以及設定 *f* 值來決定要增加多少的正樣本，接著將得到的 n 個平衡的子資料集分別丟入不同的 Random Forest 模型進行訓練，最後將 n 個模型的預測結果分數做加權平均，得到一個介於 0 到 1 的致病可能性分數。

doi:10.6342/NTU201903179

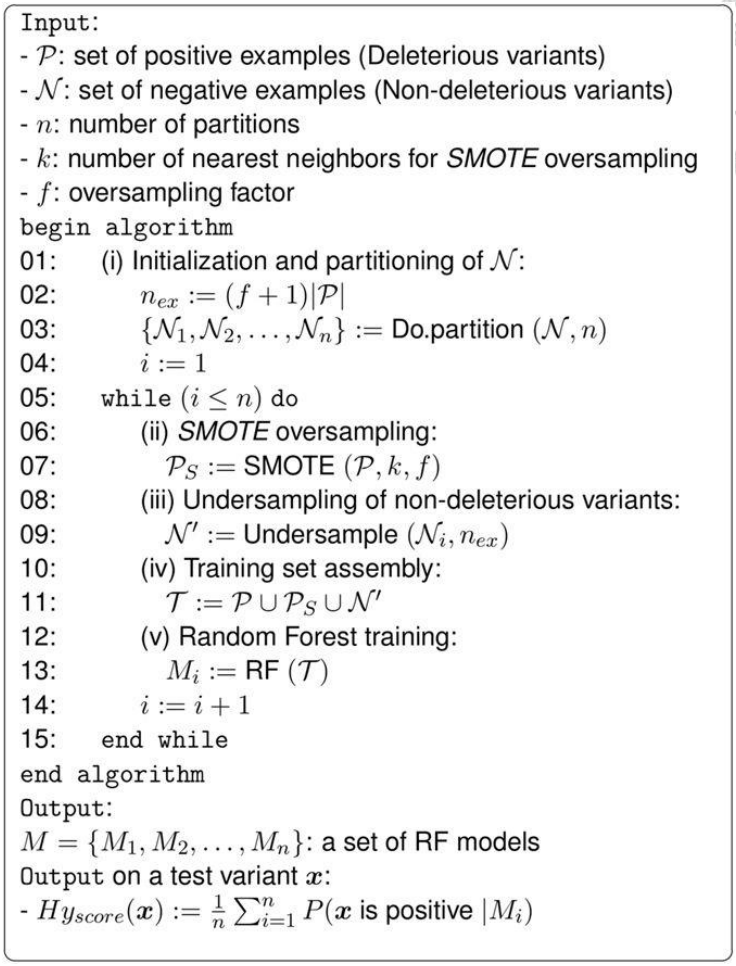

圖 2-4 HyperSMURF pseudocode，摘自文獻[7]

2.4 不平衡資料的分群集成採樣技術

一般的機器學習演算法在學習不平衡資料集時，會傾向於將資料預測為負(多數)類，來達到較好的準確度，進而忽略了正(少數)類，但其實使用者真正關注的卻是那些正類的樣本，為了改善這個問題，Chen Si et al. 在 2010 年提出了分群集成採樣技術[9]，一種基於分群集成概念的採樣演算法，其中包括了 CE-SMOTE 和 CE-Under，分群集成採樣技術在平衡正負樣本間的數量比例的同時，也能保有類似於原樣本的資料特性，藉此改善一般機器學習演算法在學習不平衡資料集時面臨的問題。

doi:10.6342/NTU201903179

2.4.1 分群集成

分群集成(Clustering Ensemble，CE)衍生於傳統分群演算法，目的為改善傳統分群演算法的不足之處，分群集成旨在探討要如何產生有效的多個分群，並合併這些分群來試圖產生更好的分群結果。在獲得多個不同分群的方法部分，可以選擇透過混和不同類型的分群演算法、調整分群演算法的初始值及參數或是隨機選擇在不同的特徵子空間及特徵投影下進行分群運算等，讓多個分群之間擁有適當的差異性以提升後續集成的結果表現，其中在分群演算法的部分，Chen Si et al. 使用 K-means [15]分群演算法來產生分群，而在合併分群方面的方法有 Multiple combiners [16]、Majority voting [17]和 Hypergraph [18]等，但在這篇研究中沒有使用到合併分群的步驟，所以在這篇參考文獻內並沒有多做介紹。

然而，在後續分析多個分群時會遇到標記匹配的問題，原因是在這些分群當中，相同性質的群可能會被賦予不同的群標記，例如，5 個樣本分別在兩次分群計算下所產生的兩個結果{1,0,1,0,1}和{0,1,0,1,0}，仔細觀察這兩個分群，可以發現第 1、3、5 個樣本都被分類到同一個群，但卻是不同的標記。為了解決標記的問題，Chen Si et al. 參考了[17][19]兩篇研究中使用的標記匹配方法來計算不同群之間的相似性。首先將樣本進行編號，並將 h 個分群中的各個標記做 one-hot encoding，成為一個二維的矩陣 C，矩陣中的第 i 行第 j 列元素 $C_{i,j}$ 若為 1 則表示第 i 個樣本在第 j 個群中，反之，若為 0 則表示第 i 個樣本不在第 j 個群中，接著計算 Jaccard 係數用以判斷不同分群之間的最佳相似群，並給予新的標記，Jaccard 係數的計算方法如公式 2.1。在 h 個分群之中，能夠選擇任意一個分群作為對照組，並將其他分群與之匹配，藉此得到一個匹配完成的 h 個分群。

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.1}$$

仔細觀察匹配完後的所有分群，可以發現到部分樣本在多次的分中時常被分配到同一個群，而部分樣本則時常被分配到不同的群，因此 Chen Si et al. 引入了

doi:10.6342/NTU201903179

分群一致性係數的概念[20][21]，計算單一樣本被分配到多數群的次數與分群次數 h 的比值，分群一致性係數的計算方式如公式 2.2：

$$CI(x) = \frac{1}{\mathrm{h}} \max \left\{ \sum_{i=0}^{n} \delta(\pi_i(x), C) \right\}_{C \in \text{cluster labels}} \tag{2.2}$$

其中

$$\delta(a, b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

$\pi_i(x)$表示樣本 x 在第 i 個分群中的標記，當分群一致性係數越高，表示樣本在多次分群中較常被分配到同一群，定義樣本位於分群的中心區域，相反的，分群一致性係數越低，表示樣本常被分配到不同的群中，定義樣本位於分群的邊界區域，故計算分群一致性係數能夠判斷樣本在多個分群間的穩定性，並透過設定閾值來判斷哪些樣本是位於分群中心或分群邊界。

2.4.2 分群集成採樣

由於位於分群邊界的樣本對於分類器在學習時的影響大於分群中心的樣本，且分群邊界的樣本較容易被誤判，因此 Chen Si et al. 利用分群集成的概念來改良 SMOTE 過採樣演算法和 Random sampling 欠採樣演算法，提出 CE-SMOTE 和 CE-Under 的採樣技術，CE-SMOTE 忽略正類分群中心的樣本，只針對正類分群邊界的樣本進行 SMOTE 過採樣處理，目的是增加正類樣本在分群邊界區域的數量，藉此強化分類器對於正類分群邊界區域的學習，而 CE-Under 則是相反的概念，忽略負類分群邊界的樣本，只針對負類分群中心的樣本進行 Random sampling 的欠採樣處理，目的是減少負類樣本在分群中心區域的數量，且因為只針對分群中心做 Random sampling，在減少整體負樣本數量的同時，也能保留與原本類似的資料特性，讓分類器在學習時能有更好的效果。

doi:10.6342/NTU201903179

第三章 研究方法

本章節將介紹整體的研究方法，包括訓練集和測試集的非編碼區變異資料來源以集資料處理流程、訓練用的資料特徵、所用之機器學習框架和模型表現的評估指標。

3.1 訓練集

為了取得非編碼區內的變異做為訓練的樣本，本研究參考 Liu et al. 於 2019 的研究[22]提供的致病性非編碼區變異資料集做為訓練集的正樣本，並另外從 1000 Genomes Project [14]取得非致病性的非編碼區變異作為訓練集的負樣本，且因為 Liu et al. 提供的致病性非編碼區變異均為單點突變的 SNP（Single Nucleotide Polymorphism）變異，因此在本研究中後續使用的所有變異均為單點突變的類型，下面將分別介紹致病性和非致病性的資料來源以及篩選步驟。

3.1.1 致病性

本研究參考 Liu et al. 於 2019 年研究[22]提供的變異資料集，此資料集是從 2015 版本的 HGMD [12]資料庫整理而來，首先從 HGMD 收集與調控相關的 2,037 個致病的非編碼區變異，然而在這些變異中可能包含了被誤判為致病的變異，因此參考了 ACMG(American College of Medical Genetics and Genomics) [23]所提出的方針來篩選掉這些潛藏的錯誤致病變異，第一步首先移除了在 HGMD 資料庫中被標示為 DM (disease-causing mutation)和 DP (disease-associated polymorphism)的變異，表示這些變異在影響功能表現的證據不足，不能有效證明它與疾病之間的關係，第二步移除了目前已知的調控元件區域以外的變異，第三步移除了在 1000 Genomes Project 中 MAF (minor allele frequency)超過 1%的變異，經過上述三個步驟的篩選，最後留下 764 個致病性的非編碼區變異作為訓練集的正樣本。

doi:10.6342/NTU201903179

3.1.2 非致病性

本研究從 1000 Genomes Project [14]的 phase1 release 下載了染色體 1~22、X 和 Y 的 vcf 格式的變異資料，並從這些檔案內篩選出位於非編碼區內且 MAF(minor allele frequency)超過 1%的常見 SNP 變異，藉此減少得到稀有變異的機會，原因是當變異在人群中的頻率越低時，稀有變異導致疾病發生的機會比較大，而常見的變異因為在多個基因組樣本中都有出現，所以較能排除產生疾病的可能性，接著移除與正樣本內有位置重複的變異，經由上述的篩選步驟後得到了 12,278,392 個位於非編碼區內的常見變異，並隨機挑選正樣本 1,000 倍的數量，最後留下 764,000 個非致病性的非編碼區變異作為訓練集的負樣本。

3.2 測試集

為了測試模型是否能夠預測除了訓練集之外的非編碼區致病變異，本研究另外準備了一組測是用的變異資料集，此資料集分別從 ClinVar [13]蒐集致病性的非編碼區變異以及 1000 Genomes Project [14]蒐集非致病性的非編碼區變異，下面將分別對兩者的資料處理做介紹。

3.2.1 致病性

本研究從 ClinVar 資料庫下載了名為 clinvar_20190513.vcf 的檔案，此檔案內的資料為 ClinVar 在西元 2019 年 5 月 13 日之前蒐集的所有變異，首先從檔案篩出可信度較高 (review state ≥ 2)，且判斷為是有致性病的 SNP 變異(INFO 欄位，CLNSIG=Pathogenic)，接著移除了位於編碼區內的變異以及和資料集正樣本內位置重複的變異，避免測試集的變異在訓練時已經出現過分析過，最後留下 370 個致病性的非編碼區變異作為測試集的正樣本。

doi:10.6342/NTU201903179

3.2.2 非致病性

此處使用和訓練集中蒐集非致病性非編碼區變異相同的資料來源和步驟，從 1000 Genomes Project 的 phase1 release 下載了染色體 1~22、X 和 Y 的 vcf 格式的變異資料，篩選出位於非編碼區內等位基因頻率(minor allele frequency，MAF)超過 1%的常見 SNP 變異，並排除了和訓練集負樣本以及測試集正樣本有位置重複的變異後隨機挑選正樣本 10 倍的數量，最後留下 3,700 個非致病性的非編碼區變異作為測試集的負樣本。

3.3 特徵選取

為了將得到的變異資料轉變為能訓練和預測的資料，必須賦予每個變異位置對應的基因訊息。本研究使用和 GWAVA [6]研究相同的特徵資訊，表 3-1 列出了不同類別的基因特徵與數量，所使用到的特徵總共有 174 個，其中每個類別內的各個特徵紀錄於附錄 1。

表 3-1 變異特徵種類和數量

Type	Number of feature
TF binding	126
Histone modifications	12
Open chromatin	3
RNA polymerase binding	3
CpG islands	1
Genome segmentation	7

doi:10.6342/NTU201903179

Type	Number of feature
Conservation	2
Human variation	2
Genic context	13
Sequence context	5

3.4 CE–SMURF 機器學習框架

為了在正負樣本數量不平衡的情況下能夠有效的提升正樣本的預測準確度，本研究提出基於分群集成採樣技術和 Hyper-ensemble 方法的機器學習框架：CE-SMURF (Clustering Ensemble of SMOTE Undersampled Random Forests)，其整體的架構如圖 3-1，下面將分別介紹 CE-SMURF 的各個組成部分，以及整體演算法的 Pseudocode。

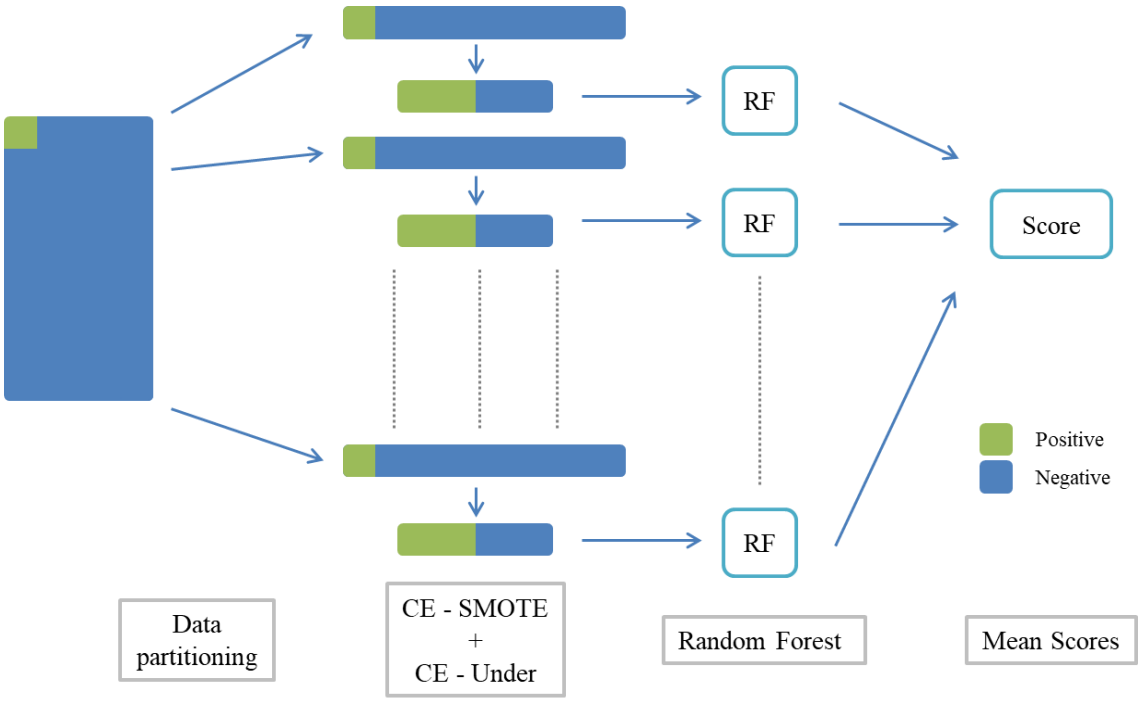

圖 3-1 CE-SMURF 機器學習框架，參考自文獻[6]

doi:10.6342/NTU201903179

3.4.1 分群集成採樣

本研究使用 Chen Si et al. 提出的 CE-SMOTE 和 CE-Under 分群集成採樣技術[9]來平衡資料集，CE-SMOTE 演算法與 SMOTE 演算法不同的地方在於只針對位於正類分群邊界的樣本做 SMOTE 過採樣處理，並不是對所有正類樣本做過採樣處理，CE-Under 演算法與 Random sampling 演算法不同的是只針對負類分群中心樣本做欠採樣處理，並不是對所有負類樣本做欠採樣處理。

在 CE-SMOTE 和 CE-Under 的計算部分，需要先利用分群集成方法分別找出正負樣本的中心和邊界區域，圖 3-2 為分群集成方法的 Pseudocode，並在下面對分群集成方法作詳細說明，首先對輸入的資料集 S 做正則化運算得到 S'集合，接著計算 h 次的 K-means 分群演算法得到 h 個不同的分群，用集合 K 表示，而為了讓每次分群的結果有一定的差異性，在每次的 K-means 分群過程之前會隨機挑選 10 到 15 個的特徵組成特徵子空間，並在這樣的特徵子空間之下來做 K-means 分群運算。做完 K-means 分群後便能獲得群標記，且為了匹配多個分群間的標記，需要將 h 個分群的標記先經由 one-hot enocoding 轉換成一個二維的陣列，再利用公式 2-1 來計算 Jaccard 係數，匹配不同分群間的最佳標記，匹配完的集合為 K'，在匹配標記時能夠使用任何一個固定的分群來作為其他分群的固定匹配對象，通常使用第一個分群當做其他分群匹配的標準。有了正確的標記後，便能計算樣本的分群一致性係數，並同時計算資料集 S 中樣本的分群一致性係數的平均值來作為判定邊界或中心的閾值，若是分數大於閾值則表示樣本在多個分群中常被分配到同一個分群，定義其位於分群的中心 S_c，若是分數小於閾值則表示樣本在多個分群結果中多次被分配到不同的分群，則定義其位於分群的邊界 S_b。

利用上述的分群集成方法分別找出正負樣本的分群中心和邊界樣本後，把正類分群邊界的樣本做 SMOTE 的過採樣處理，並保留正類分群中心的樣本，同時也把負類分群中心的樣本做 Random sampling 的欠採樣處理，並保留負類分群邊界的樣本，最後合併經過分群集成採樣處理的正樣本和負樣本得到一個新的資料集，

doi:10.6342/NTU201903179

新的資料集不但縮小了正負樣本之間的數量差距，也能保有類似於原本資料集的資料特性。

Clustering Ensemble (CE)

S : Sample set
c : number of cluster in k-means
h : number of clustering results in CE
Input : S
Start algorithm
01. // Normalize computation
02. S' = Normalize(S)
03. for i in (1..h) :
04. // K-menas Clustering
05. K_i = kmeans(S', c)
06. $K = \{K_1, K_2, \ldots, K_n\}$
07. // Matching different clustering
08. K' = match_cluster(K)
09. // Calculating CI scores and matching samples
10. $\{S_b, S_c\}$ = CI(S, K')
End algorithm
Output :
S_b : boundary samples of sample set
S_c : central samples of sample set

圖 3-2 分群集成方法 pseudocode

3.4.2 Hyper−ensemble

Hyper-ensemble 是一種基於 Random Forest [9]分類器的 Ensemble 方法，因為 Random Forest 已經是在 Ensemble 概念下由多個 Decision Tree 結合而成的機器學習演算法，故將此方法稱為 Hyper-ensemble。Hyper-ensemble 首先把輸入的訓練集樣本平均分配到 n 個獨立的子訓練集，接著每個子訓練集單獨做 CE-SMOTE 和 CE-Under 的採樣以及訓練 Random Forest 分類器，最後在平均每個分類器的結果，其中平均分配的動作能確保之後每個做完採樣的子訓練集都互相獨立，除了增加所有子集合對於原資料集的覆蓋率，同時提升多個分類器之間的差異性。

doi:10.6342/NTU201903179

3.4.3 CE-SMURF 參數

表 3-2 列出了 CE-SMURF 中的參數，且分別描述每個參數的用途，另外本研究藉由調整與採樣部分相關的係數 f 和 r 來探討分群採樣技術對於訓練表現的影響，*f* 為 CE-SMOTE 的過採樣係數，能夠用其決定正類樣本的分群邊界過採樣比例，f 為大於等於 0 的正整數，且當 *f* 值為 0 時，表示沒有使用 CE-SMOTE，r 為 CE-Under 的欠採樣係數，能夠用其決定負類樣本的分群中心欠採樣比例，*r* 為大於 0 且小於等於 1 的浮點數，當 *r* 值為 1 時，表示沒有使用 CE-Under。

表 3-2 CE-SMURF 參數介紹

Parameter	Description
n	Number of partitions
h	Number of Clustering times
c	Number of clusters in K-means
f	Over-sampling ratio of CE-SMOTE
r	Under-sampling ratio of CE-Under
k	SMOTE k-nearest neighbor
t	Number of trees in Random Forest

3.4.4 Pseudocode

CE-SMURF 整體框架的 Pseudocode 如圖 3-3，*P* 為正樣本集合，*N* 為負樣本集合，且假設 *N* 集合數量遠大於 *P* 集合，首先將 *N* 平均分成 *n* 個子集合$\{N_1, N_2, \ldots, N_i\}$，接著利用 for 迴圈(2-12 行)迭代 *n* 次，在每次的迭代中，分別將 *P* 集合和 N_i(*i* 為第 *n* 次迭代)集合做分群集成得到各自的分群邊界樣本和分群中心樣本，接著將 *P* 集

doi:10.6342/NTU201903179

合的分群邊界樣本 P_b 做 CE-SMOTE 過採樣演算法得到 P_b'，將 N_i 集合的分群中心樣本 N_c 做 CE-Under 欠採樣演算法得到 N_c'，藉此縮小正負樣本間的數量。

做完分群採樣後，把保留下來的樣本 P_c 和 N_b 以及經過採樣處理的 P_b' 和 P_b' 作聯集得到新的樣本集合 T，並將 T 做為 Random Forest 機器學習演算法的訓練資料，訓練完的模型以 M_i 表示，經過 n 次迭代後，便能獲得一個由 n 個模型所組成的 M 集合，最後將 n 個模型的預測分數做加權平均，得到一個介於 0 到 1 的 CE-SMURF 致病可能性分數。

CE-SMURF

P : positive set
N : negative set
n : number of partitions
k : number of nearest neighbors for SMOTE
f : over-sampling factor
r : under-sampling factor
Input : P, N
Start algorithm
01. // Partition of N
02. $\{N_1, N_2, \ldots, N_i\}$ = Partition(N, n)
03. // Hyper-ensemble method
04. for i in (1..n) :
05. // Clustering Ensemble
06. $\{P_b, P_c\}$ = CE(P)
07. $\{N_b, N_c\}$ = CE(N)
08. // CE-SMOTE over-sampling
09. P_b' = CE-SMOTE(P_b, f, k)
10. // CE-Under under-sampling
11. N_c' = CE-Under(N_c, r)
12. $T = N_c' \cup P_b' \cup N_b \cup P_c$
13. // Random Forest training
14. M_i = RandomForest(T)
End algorithm
Output :
$\text{Score}(x) = \frac{1}{n}\sum_{k=1}^{n} P(x \text{ is positive}|M_i)$

圖 3-3 CE-SMURF pseudocode

doi:10.6342/NTU201903179

3.5 模型表現評估指標

一般在評估機器學習的訓練表現時，會將訓練資料依比例分為訓練集和驗證集，訓練集用來訓練模型，而驗證集用來檢視模型的訓練效果，但在基因變異的測試上，沒辦法保證哪些變異最能代表整體的資料，因此本研究使用 K-fold 的交叉驗證方法來評估模型的訓練結果，因為 K-fold 較能避免隨機選擇某一部分資料當作驗證集造成的偏差，K-fold 概念如圖 3-4 所示，首先將資料集切成 *K* 等份，並進行 *K* 次迭代，每次迭代都選擇其中 *K*-1 份當作訓練集，剩下那份當做驗證集，迭代完 *K* 次之後，把每次驗證集的預測結果分數作平均，另外在迭代次數 *K* 的部分，本研究選擇 *K* 等於 10 的 10-fold，將資料切為 10 等份並做 10 次的交叉驗證來評估訓練的結果。

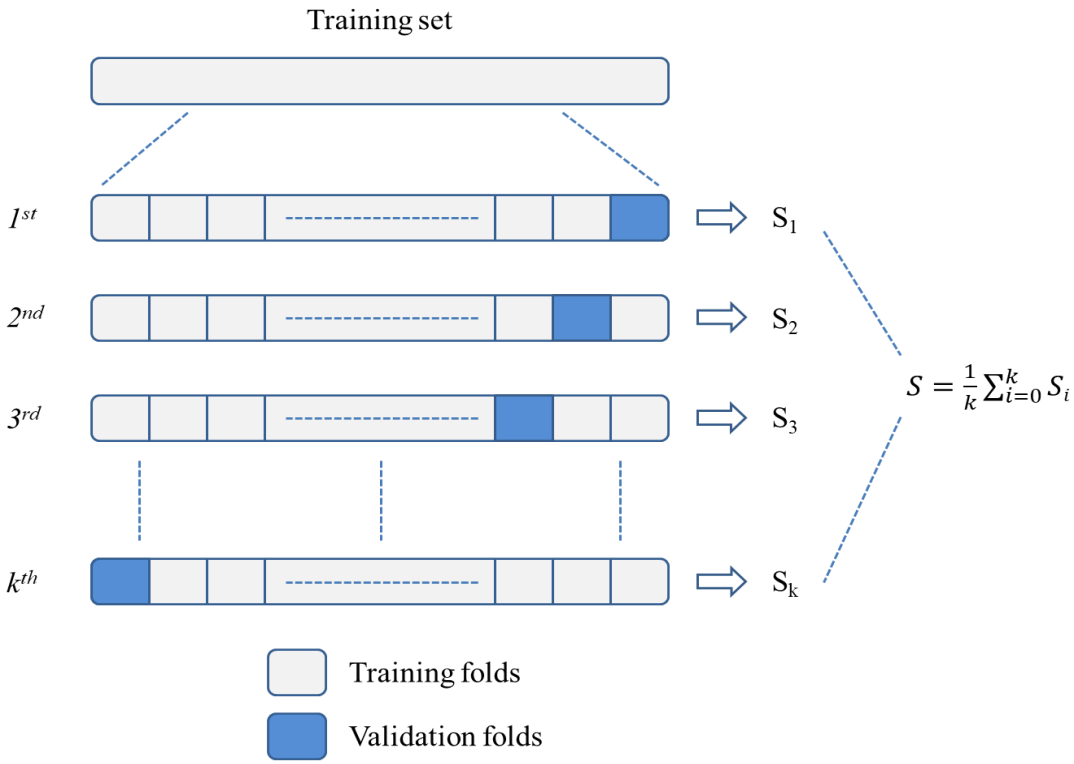

圖 3-4 K-fold 交叉驗證

在機器學習的二分類問題當中，樣本會有正類(positive)和負類(negative)兩種標記，而在分析結果時，透過設定特定的閾值(threshold)，樣本的分類結果可以歸

doi:10.6342/NTU201903179

納為 4 種不同的情況：(1) 正類預測成正類，稱之為真正類（true positive，TP）；(2)負類預測成正類，稱之為假正類（false positive，FP）；(3) 負類預測成負類，稱之為真負類(true negative，TN)；(4) 正類預測成負類，稱之為假負類(false negative，FN)，利用上面 4 種不同的分類結果，可以使用公式 3.1 來計算真陽性率(true positive rate，TPR)和偽陽性率(false positive rate，FPR)。

$$TPR = \frac{TP}{TP + FN} \quad ; \quad FPR = \frac{FP}{FP + TN} \tag{3.1}$$

藉由設定不同的閾值來得到多組 TPR 和 FPR，並利用這些數值作為點座標畫在 XY 座標圖中，連結這些點座標後便能夠得到 ROC 曲線(receiver operating characteristic curve)，計算曲線下的面積可得到 AUROC(area under the curve of ROC)，一個好的分類器 TPR 要趨近於 1，而 FPR 要趨近於 0，正常情況下，隨著閾值的下降，當 TP 上升時，FP 也會同時上升，導致 TPR 和 FPR 之間會互相牽制，但在正負資料量不平衡的情況下，負樣本數量遠大於正樣本的緣故，當閾值下降，TP 和 FP 同時上升時，因為 FP + TN 過大的緣故會導致 FP 的變化在 FPR 上沒辦法呈現，使得在 ROC 指標上會有模型表現的很好的錯覺，為了能更有效辨別各個模型間的優劣，本研究改由使用 PRC(precision-recall curve)曲線和 AUPRC(area under the curve of PRC)作為主要的指標，PRC 曲線是在不同閾值下的精準率(precision)和召回率(recall)構成的曲線，兩者的計算方式如公式 3.2：

$$precision = \frac{TP}{TP + FP} \quad ; \quad recall = \frac{TP}{TP + FN} \tag{3.2}$$

選擇使用 PRC 曲線的原因是因為精準率和召回率不會受到不平衡資料集的影響，所以較能明顯的凸顯不同模型之間的表現差異。

為了證實分群集成採樣技術能夠提升不平衡資料的訓練表現，將同為預測非編碼區致病變異且同樣基於採樣技術的 GWAVA [6]以及 HyperSMURF [7]當作比較對象，與 CE-SMURF 和 HyperSMURF 不同的是 GWAVA 僅使用欠採樣的技術的來平衡數據集，並沒有使用 Hyper-ensemble 的方法來合併多個分類器的預測結果，

doi:10.6342/NTU201903179

除此之外，三者在機器學習分類器的部分都是使用 Random Forest 演算法來做學習，因此更能觀察出 Hyper-ensemble 和分群集成採樣對於結果的影響。然而，GWAVA 和 HyperSMURF 在使用的訓練資料(變異樣本、特徵選取)並不相同，必須在相同訓練資料的基準下才有辦法進行比較，基於這個原因，本研究使用了前面小節所蒐集的變異樣本和 174 個註釋特徵來重新訓練 GWAVA 及 HyperSMURF，此步驟能讓結果的差異單純來自於模型的不同，而避免訓練資料不同所造成的差異影響分數評估。

doi:10.6342/NTU201903179

第四章 結果與討論

4.1 採樣參數對訓練的影響

為了探討 CE-SMOTE 和 CE-Under 對於訓練的影響，本研究調整 CE-SMOTE 的過採樣係數：*f* 和 CE-Under 的欠採樣係數：*r* 來檢視兩種採樣技術對於訓練的影響，其結果如圖 4-1，從圖中可以明顯比較 CE-SMURF 在不同 *f* 和 *r* 下的訓練結果，在相同的 *f* 時，可以發現到隨著 *r* 的下降，CE-SMURF 的 AUPRC 值呈現上升的趨勢，此結果進一步證實了 CE-Under 在刪去負樣本時能有效的保持與原本訓練集類似的資料特性，且同時縮小了正負樣本間的數量比例，讓隨機森林模型在訓練時能更好的平衡正負樣本的誤差，而不會因此偏袒某一方。另一方面，隨著 *f* 的上升，CE-SMURF 的 AUPRC 值卻沒有與之上升，表示在準備的資料集當中透過 CE-SMOTE 增加的新正樣本對於隨機森林模型的樣本辨別能力沒有幫助。

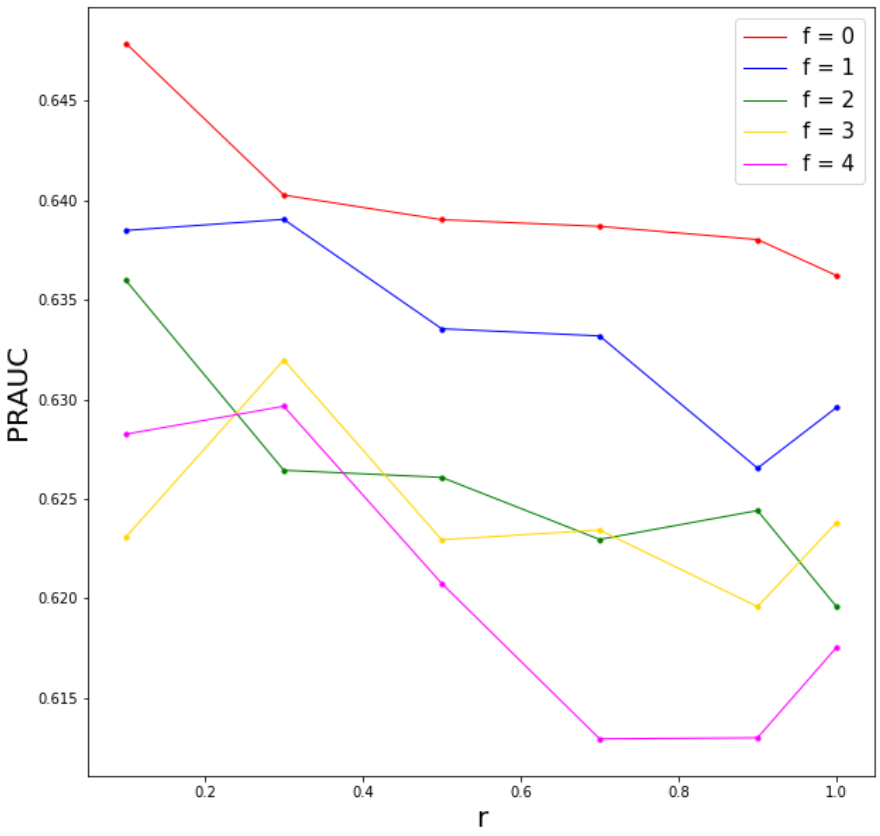

圖 4-1 不同採樣參數配對下的訓練結果

doi:10.6342/NTU201903179

本研究藉由圖 4-1 的結果選擇了 AUPRC 最好的採樣係數組合作為後續實驗的 CE-SMURF 代表模型，其 *f* 等於 0，*r* 等於 0.1，表示在 CE-SMURF 中只單獨使用 CE-Under 針對負樣本做欠採樣，而不使用 CE-SMOTE 對正樣本做過採樣，這樣的方式能在訓練上獲得最好的表現，除了 *f* 和 *r* 之外，表 4-1 列出了 CE-SMURF 機器學習框架內的其他預設參數。

表 4- 1 CE-SMURF 預設參數值

Parameter	**Description**	**Default**
n	Number of partitions	10
h	Number of Clustering times	10
c	Number of clusters in K-means	3
f	Over-sampling ratio of CE-SMOTE	0
r	Under-sampling ratio of CE-Under	0.1
k	SMOTE k-nearest neighbor	5
t	Number of trees in Random Forest	100

doi:10.6342/NTU201903179

4.2 不同方法間預測的比較

本研究將有使用採樣技術的 GWAVA [5]和 HyperSMURF [6]做為比較對象，利用相同的資料集重新訓練方法的模型後使用 10-fold 的交叉驗證方式來評估訓練表現，圖 4-2 為三種方法訓練表現的 ROC 曲線和 PRC 曲線，從右邊的 ROC 曲線可以觀察到 CE-SMURF 達到了 0.989 的 AUROC，而 HyperSMURF 和 GWAVA 則是分別達到 0.989 和 0.987，雖然三者的 AUROC 都趨近於 1，看似都有很好的訓練表現，但其實是因為在正負樣本間的數量比例極不平衡時 FPR 值會容易趨近於 0，所以在 ROC 曲線上會有三者都表現很好的錯覺，因此需要使用 PRC 曲線來做為衡量的指標，從左邊的 PRC 曲線可以觀察到 CE-SMURF 達到了 0.648 的 AUPRC，而 HyperSMURF 和 GWAVA 則是分別達到 0.540 和 0.396，三者之間有了明顯的差異，CE-SMURF 相較第二好的 HyperSMURF 在 AUPRC 的分數上有了 20%的提升，且不管是在精確率(precision)或是召回率(recall)的部分都較另外兩者來的好，從 CE-SMURF 和 HyperSMURF 的差異中能夠看出 CE-SMOTE 和 CE-Under 相對於 SMOTE 和 Random sampling 在樣本採樣上的改善，而從 GWAVA 和 HyperSMURF 的差異也能看出 Hyper-ensemble 方法對於預測的結果表現有一定的貢獻。

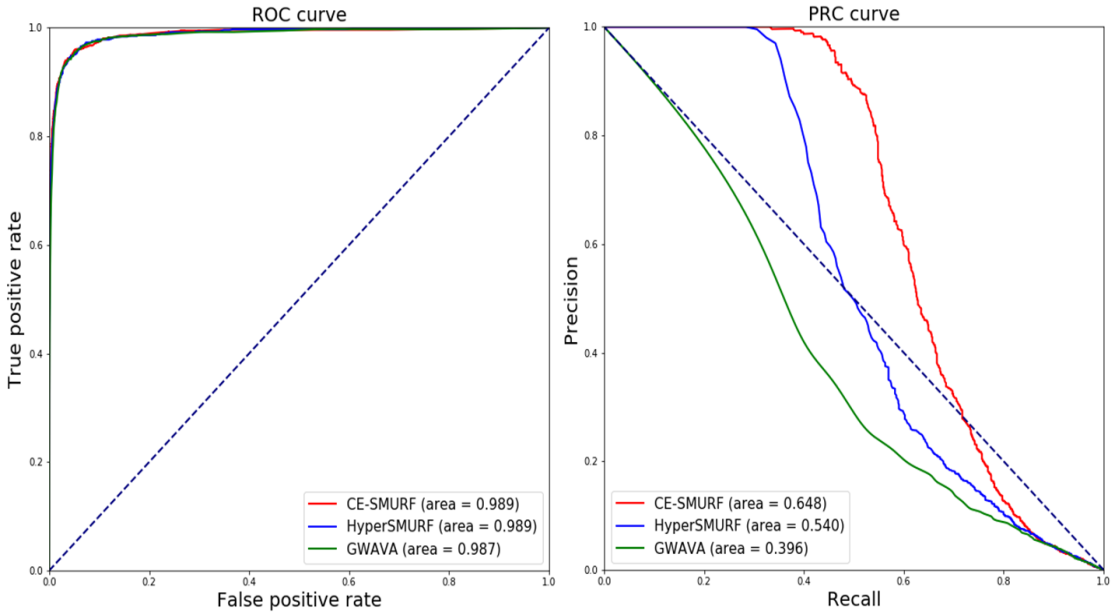

圖 4-2 訓練集的 ROC 曲線和 PRC 曲線比較

doi:10.6342/NTU201903179

做完訓練結果的比較後，本研究接著測試了三種方法對於尋找未知變異的能力，將訓練集訓練完的模型用來預測測試集的資料，三者的 ROC 曲線和 PRC 曲線如圖 4-3，從圖中可看出 CE-SMURF 相較另外兩種方法，不管是在 ROC 指標或是 PRC 指標都有最佳的表現，在 ROC 曲線圖中能夠觀察到與訓練表現結果相同的情況，三種方法在測試表現上都有相當高的 AUROC，但卻無法明顯區分三者的表現，而在 PRC 曲線圖中便能輕易將 CE-SMURF 和另外兩者明顯區分開，且在相同 recall 的情況下，CE-SMURF 的 precision 有很大的提升，表示 CE-SMURF 能夠有效減少模型預測 FP (false positive)的情況發生，但和預期不同的是，雖然 HyperSMURF 在訓練集上的 AUPRC 遠高於 GWAVA，但在測試集上的預測表現卻略低於 GWAVA，表示 HyperSMURF 在訓練集上的好並沒辦法代表它擁有比 GWAVA 好的變異預測能力。

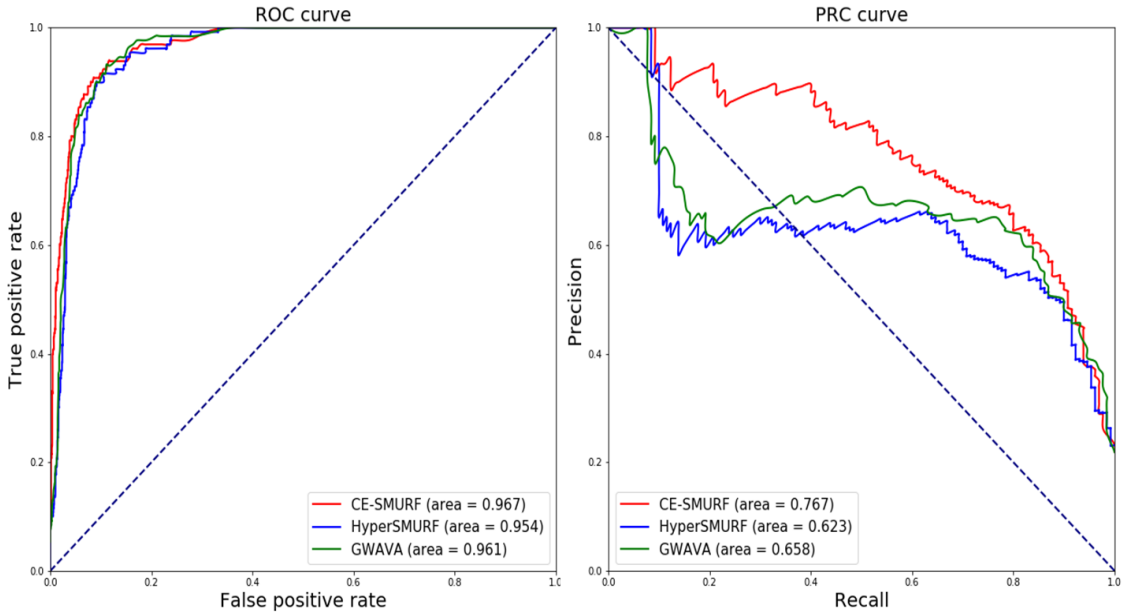

圖 4-3 測試集的 ROC 曲線和 PRC 曲線比較

doi:10.6342/NTU201903179

4.3 不平衡程度對預測的影響

為了比較訓練集資料的不平衡程度對於 CE-SMURF、HyperSMURF 以及 GWAVA 的影響，本研究透過改變訓練集中負樣本的數量來準備不同比例的訓練資料，其比例包括：1、10、100 和 1,000，利用這 4 筆不同的訓練資料來分別訓練三種方法的模型並做 10-fold 的訓練表現評估比較，其結果如圖 4-4，隨著不平衡程度的上升，三種方法的 AUPRC 都呈現下降的趨勢，表示資料量不平衡程度的提升會造成模型訓練表現變差。當資料比例為 1 時三種方法的 AUPRC 大致相同，但隨著資料量的上升不同方法間的差距逐漸明顯，而 CE-SMURF 在 4 種比例下，都能有最好的訓練表現，表示 CE-SMURF 在三種方法中對於不平衡程度的改變有最小的敏感度，能減少資料集中負樣本增加時模型訓練表現降低的現象。

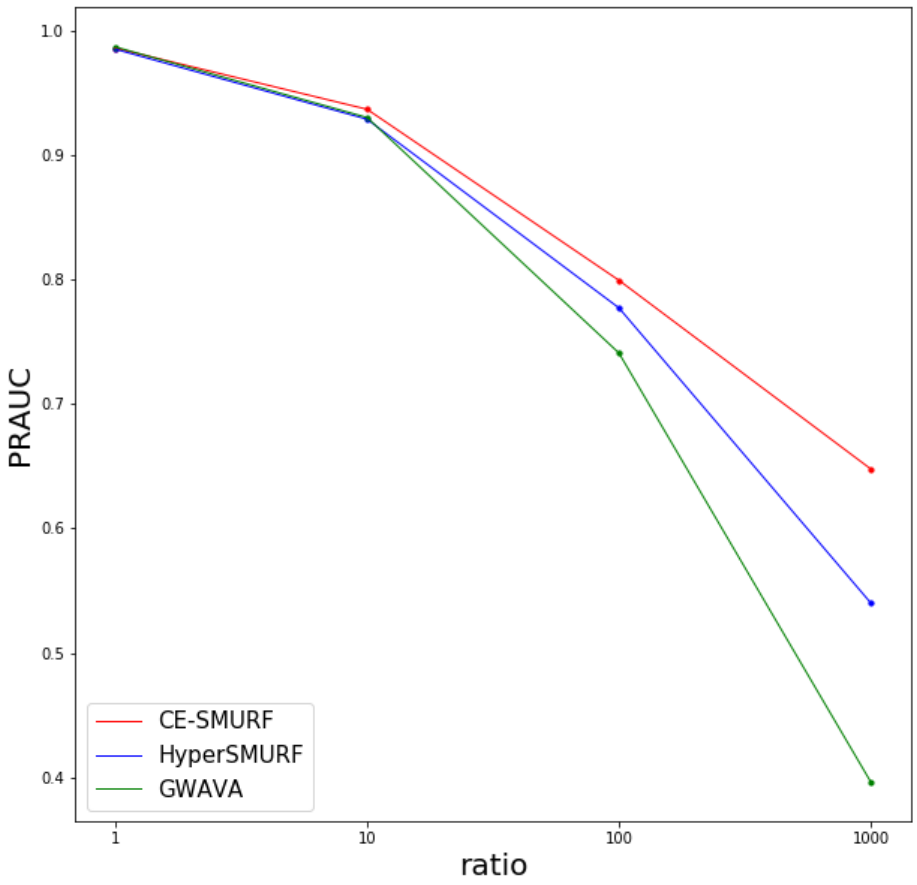

圖 4-4 不同訓練資料比例下的訓練表現

doi:10.6342/NTU201903179

除此之外，本研究也試圖探討不同比例的資料對於測試集預測的影響，因此利用上述準備的 4 筆資料來分別訓練三種方法的模型並做測試集的預測比較，其結果如圖 4-5，從圖中的結果發現到隨著不平衡程度的上升，測試集的預測結果是呈現上升的趨勢，這是與訓練結果呈現完全相反的情況，表示雖然增加資料集內的負樣本會導致訓練結果的表現變差，但卻能讓未來在尋找新變異時有更高的準確度。

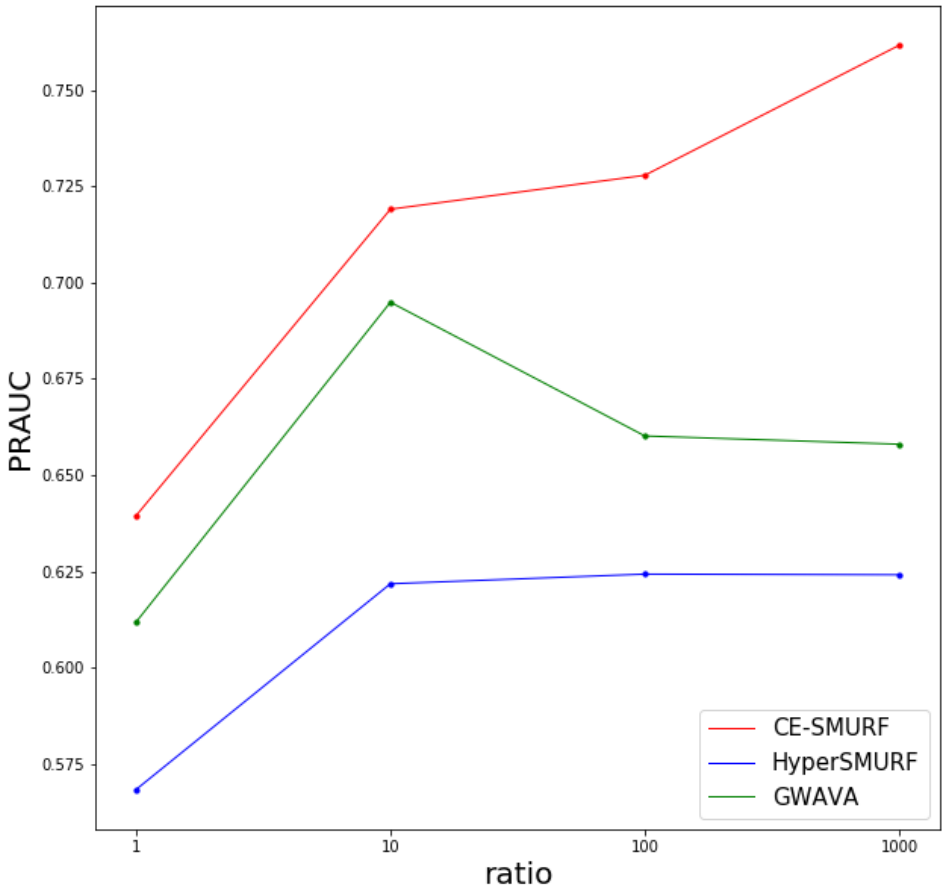

圖 4-5 不同訓練資料比例下的測試表現

doi:10.6342/NTU201903179

4.4 不同可信度變異資料對預測的影響

為了探討不同可信度的變異資料是否會影響預測的表現，本研究使用 Ritchie et al. 在 GWAVA [5]研究中提供的致病變異來替換訓練集中的正樣本，其為從 HGMD 資料庫 2013 年的公開版本中篩選出位於非編碼區的 1,614 個與疾病相關變異，GWAVA 提供的資料因為使用沒有移除掉潛藏的錯誤致病變異，所以相較於本研究原本使用 Liu et al. [21]研究中提供的致病相關變異具有較低的可信度。

首先，利用兩組不同的資料集分別訓練 CE-SMURF 並做 10-fold 表現評估，其結果如圖 4-6，圖例中的 Liu 為使用 Liu et al. 提供之正樣本，做為訓練資料之結果，Ritchie 為使用 Ritchie et al. 提供之正樣本做為訓練資料之結果，從結果可以發現到雖然 Liu 訓練集中的正樣本數量較少，但訓練結果不管是在 ROC 指標或是 PRC 指標都表現的比 Ritchie 來的好，AUROC 為 0.989，AUPRC 為 0.648。

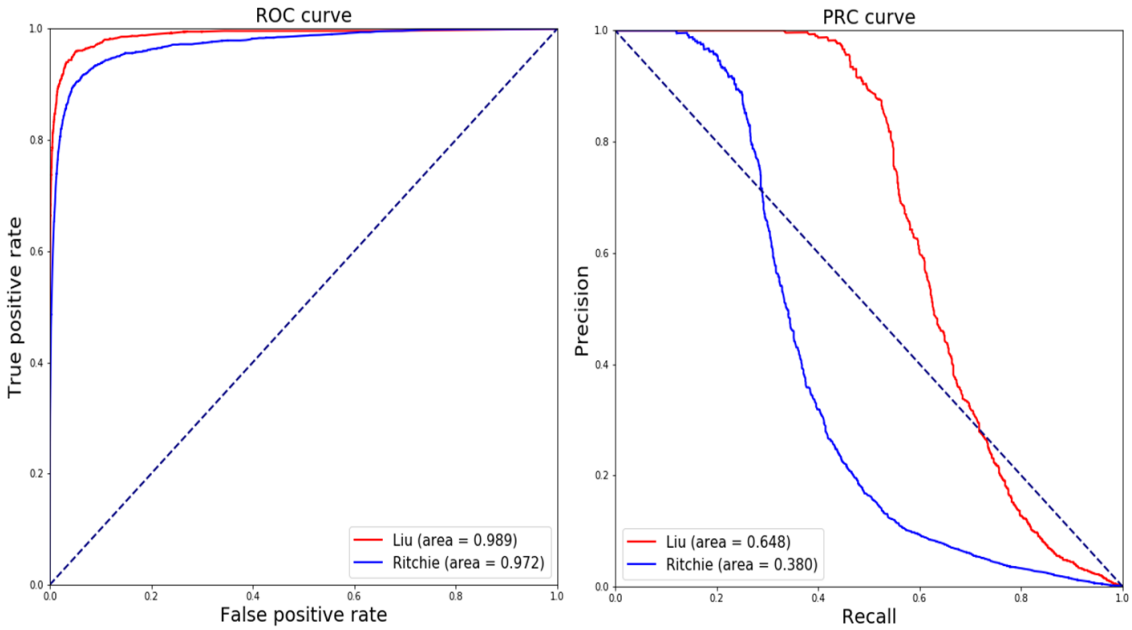

圖 4-6 不同可信度資料的訓練表現

接著用這兩組訓練資料分別做測試集的測試，其結果如圖 4-7，同樣也是用 Liu 做為訓練資料時在 ROC 指標和 PRC 指標都獲得較好的表現，AUROC 為 0.967，AUPRC 為 0.767。從訓練和測試的結果來看，可以發現到確實使用較高可信度的

doi:10.6342/NTU201903179

變異資料做為正樣本能夠有效的提升預測的表現，若是從數量上來看，用 Ritchie 資料集做為訓練資料應當表現的較好，但結果卻相反，原因可能是因為 Ritchie 資料集當中正樣本和負樣本有著一定的相似性，導致分類器的學習效果不佳，表示 HGMD 資料庫內具有一定數量的潛藏錯誤致病變異。

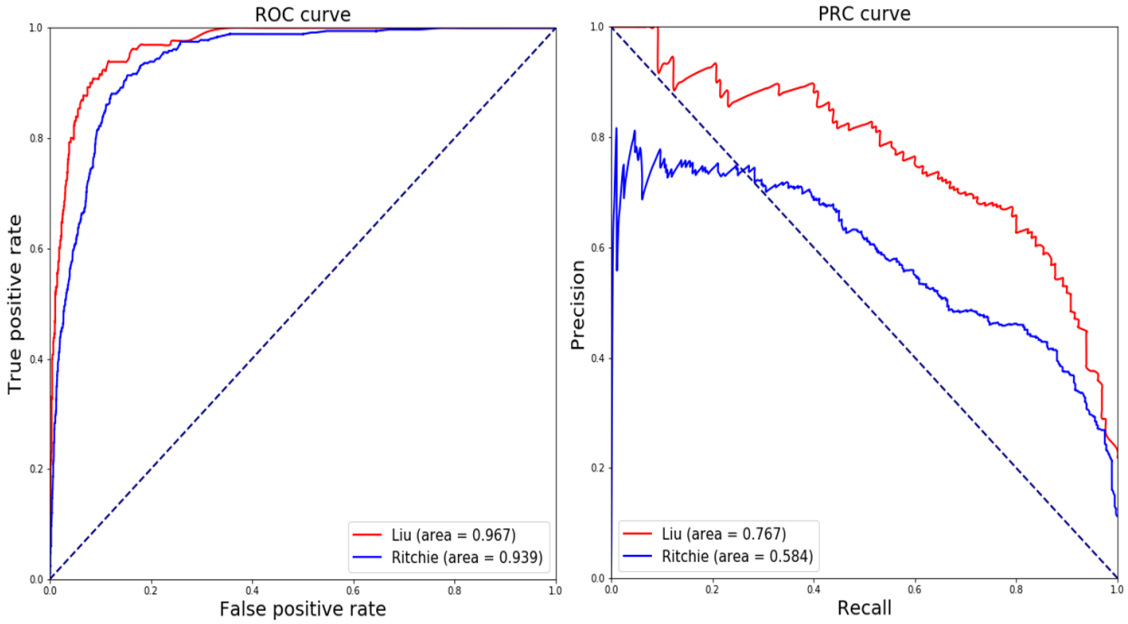

圖 4-7 不同可信度資料的測試表現

doi:10.6342/NTU201903179

第五章 結論

本研究基於分群集成採樣技術和 Hyper-ensemble 方法開發出 CE-SMURF 機器學習框架，並應用於預測非編碼區致病變異，在調整 CE-SMURF 採樣參數以獲取最佳化模型時，發現單獨使用 CE-Under 能夠有最好的表現，但這並不能直接否定 CE-SMOTE 的作用，或許在其他訓練集中同時使用 CE-SMOTE 和 CE-Under 會有更好的表現。在目前有使用到採樣技術的方法中，CE-SMURF 不管是在 ROC 指標或是 PRC 指標都能取得較高的分數，表示分群集成採樣和 Hyper-ensemble 能有效改善一般機器學習演算法在學習不平衡資料集時的限制，其中分群集成採樣能在平衡正負樣本數量的同時降低對於資料特性的影響，而 Hyper-ensemble 透過平均多個 Random Forest 分類器的結果，藉此得到比單一 Random Forest 分類器更好的預測結果。此外 CE-SMURF 對於訓練資料集的不平衡程度有較低的敏感度，隨著不平衡程度的上升，訓練的表現能有較小幅度的降低，特別的是雖然訓練的表現有下降的趨勢，但在測試集的預測表現上卻是大幅的上升。除此之外，本研究也發現移除資料庫中潛藏的錯誤致病變異，使用較高可信度的致病變異當作訓練資料，能讓正負樣本之間有更大的差異，進而提升預測的準確度。未來能夠使用較新的採樣技術或是 Ensemble 的方法來改良 CE-SMURF 的部分框架，且隨著更多臨床實驗資料的釋出，正負樣本的數量必然都會呈現上升的趨勢，適當的選擇可信度較高的樣本則能在預測致病變異的問題上有更好的表現。

doi:10.6342/NTU201903179

參考文獻

1. Edwards, S.L., et al., *Beyond GWASs: illuminating the dark road from association to function.* Am J Hum Genet, 2013. **93**(5): p. 779-97.
2. Smedley, D., et al., *A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease.* Am J Hum Genet, 2016. **99**(3): p. 595-606.
3. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.
4. Quang, D., Y. Chen, and X. Xie, *DANN: a deep learning approach for annotating the pathogenicity of genetic variants.* Bioinformatics, 2015. **31**(5): p. 761-3.
5. Ionita-Laza, I., et al., *A spectral approach integrating functional genomic annotations for coding and noncoding variants.* Nat Genet, 2016. **48**(2): p. 214-20.
6. Ritchie, G.R., et al., *Functional annotation of noncoding sequence variants.* Nat Methods, 2014. **11**(3): p. 294-6.
7. Schubach, M., et al., *Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants.* Sci Rep, 2017. **7**(1): p. 2959.
8. Chawla, N.V., et al., *SMOTE: synthetic minority oversampling technique.* J. Artif. Int. Res., 2002. **16**(1): p. 321-357.
9. 陈思，郭躬德，陈黎飞, *基于聚类融合的不平衡数据分类方法.* 模式识别与人工智能, 2010. **23**(6): p. 772-775
10. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
11. Rojano, E., et al., *Regulatory variants: from detection to predicting impact.* Brief Bioinform, 2018.
12. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.
13. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Res, 2018. **46**(D1): p. D1062-D1067.
14. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
15. MacQueen, J. *Some methods for classification and analysis of multivariate observations*. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. 1967. Berkeley, Calif.: University of California Press.

doi:10.6342/NTU201903179

16. Fred, A.L. and A.K. Jain. *Data clustering using evidence accumulation*. in *Object recognition supported by user interaction for service robots*. 2002. IEEE.
17. Fred, A. *Finding consistent clusters in data partitions*. in *International Workshop on Multiple Classifier Systems*. 2001. Springer.
18. Strehl, A. and J. Ghosh, *Cluster ensembles---a knowledge reuse framework for combining multiple partitions.* Journal of machine learning research, 2002. **3**(Dec): p. 583-617.
19. Zhou, Z.-H. and W. Tang, *Clusterer ensemble.* Knowledge-Based Systems, 2006. **19**(1): p. 77-83.
20. Topchy, A., et al. *Adaptive clustering ensembles*. in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* 2004. IEEE.
21. Chen, S., G. Guo, and L. Chen. *Semi-supervised classification based on clustering ensembles*. in *International Conference on Artificial Intelligence and Computational Intelligence*. 2009. Springer.
22. Liu, L., et al., *Biological relevance of computationally predicted pathogenicity of noncoding variants.* Nat Commun, 2019. **10**(1): p. 330.
23. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genet Med, 2015. **17**(5): p. 405-24.

doi:10.6342/NTU201903179

附錄 1 各類別內詳細特徵

TF binding

JUND, SP1, FOSL2, HNF4A, EP300, FOXA2, TCF12, TBP, HDAC2, HEY1, FOXA1, HNF4G, GATA1, SIN3A, GTF2F1, MYC, TCF7L2, CHD2, TAF1, STAT1, BCLAF1, MAX, CEBPB, MXI1, BATF, RDBP, BCL3, E2F4, POU2F2, SLC22A2, HMGN3, PAX5, YY1, NFKB1, NR3C1, USF1, STAT3, GATA2, TFAP2C, BHLHE40, TAL1, HSF1, TFAP2A, ELF1, GTF2B, USF2, FOS, CCNT2, E2F6, IRF4, CTCF, E2F1, ZEB1, STAT2, REST, SREBF2, MEF2A, SMARCB1, EGR1, RXRA, SPI1, ELK4, EBF1, PBX3, RFX5, BRCA1, SMC3, SMARCA4, SREBF1, NR2C2, TRIM28, TAF7, NFYA, RAD21, SRF, ZBTB7A, IRF1, SIRT6, NFE2, ZNF263, THAP1, CTBP2, MEF2_complex, GTF3C2, ATF3, BCL11A, BDP1, BRF1, BRF2, CTCFL, ERALPHAA, ESRRA, ETS1, ERALPHAA, FAM48A, FOSL1, GABPA, GATA3, HDAC8, IRF3, JUN, JUNB, KAT2A, MAFF, MAFK, NANOG, NFYB, NR4A1, NRF1, POU5F1, PPARGC1A, PRDM1, SETDB1, SIX5, SMARCC1, SMARCC2, SP2, SUZ12, WRNIP1, XRCC4, ZBTB33, ZNF143, ZNF274, ZZZ3, bound motif, pwm

Histone modifications

H3K4me3, H3K4me2, H3K9ac, H2AFZ, H3K4me1, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H3K9me3, H3K9me1, H4K20me1

Open chromation

DNase, FAIRE, dnase_fps

RNA polymerase binding

POLR2A, POLR2A_elongating, POLR3A

CpG islands

cpg_island

Genome segmentation

TSS, TRAN, ENH, WEAK_ENH, CTCF_REG, TSS_FLANK, REP

Conservation

gerp, avg_gerp

doi:10.6342/NTU201903179

Human variation

avg_daf, avg_het

Genic context

EXON, INTRON, CDS, UTR’5, UTR’3, DONOR, ACCEPTOR, START, STOP, tss_dist, ss_dist, GC, in_cpg

Sequence context

seq_A, seq_C, seq_G, seq_T, repeat

doi:10.6342/NTU201903179