

**資訊工程系碩士班**

**碩士學位論文**

**SHAP-CF-SMOTE：結合SHAP與反事實插值的資料增補方法**

**SHAP-CF-SMOTE: A data augmentation method combining SHAP and counterfactual interpolation**

**研究生：陳勝誠**

**指導教授：郭忠義 博士，王秉豐 博士**

**中華民國一百一十五年七月**

摘 要

ABSTRACT

目 錄

[摘 要 i](#_Toc208793243)

[ABSTRACT ii](#_Toc208793244)

[目 錄 iii](#_Toc208793245)

[第一章 緒論 1](#_Toc208793246)

[1.1 研究動機與目的 1](#_Toc208793247)

[第二章 文獻探討 2](#_Toc208793248)

[2.1 相關研究背景介紹 2](#_Toc208793249)

[2.1.1 Imbalanced Learning 2](#_Toc208793250)

[2.1.2 SHAP 2](#_Toc208793251)

[2.1.3 反事實解釋 2](#_Toc208793252)

[2.1.4 DICE 3](#_Toc208793253)

[2.1.5 SMOTE 3](#_Toc208793254)

[2.1.6 Deep Learning Meets Oversampling 3](#_Toc208793255)

[2.2 相關研究比較探討 3](#_Toc208793256)

[2.2.1 SMOTE-XGBoost Classifier Model 3](#_Toc208793257)

[2.2.1 Counterfactual-based minority oversampling 4](#_Toc208793258)

[第三章 研究方法 4](#_Toc208793259)

[3.1 研究議題 4](#_Toc208793260)

[3.2系統架構 4](#_Toc208793261)

[3.3方法簡介 5](#_Toc208793262)

[3.4 問題與解決方法定義 5](#_Toc208793263)

[3.5 實驗方法 5](#_Toc208793264)

[3.5.1 資料集介紹 5](#_Toc208793265)

[3.5.2 比較方法 6](#_Toc208793266)

[3.5.3 分類器 6](#_Toc208793267)

[3.5.4資料擴增流程 6](#_Toc208793268)

[3.5.5評估指標 6](#_Toc208793269)

[第四章 實驗與討論 7](#_Toc208793270)

[參考文獻 8](#_Toc208793271)

第一章 緒論

## 1.1 研究動機與目的

在許多現實世界的分類任務中，例如醫療風險評估，高風險樣本往往佔比極低，導致訓練資料呈現嚴重不平衡。這樣的資料不平衡會導致模型偏向多數類別，進而忽略對少數類別（如高風險病患）的正確分類，降低模型在關鍵場合下的實用性與可信度，而傳統的特徵重要性方法，如 SHAP，雖能揭示模型對各特徵的依賴程度，卻無法具體指出「要如何改變特徵」，才能讓少數類樣本被正確分類。反事實樣本生成正是解決此問題的有效策略之一，其目標是針對某筆資料，產生一筆與其相似、但會被模型判斷為目標類別（如高風險）的樣本，並確保特徵變動合理。

以醫療應用為例，醫師可能想知道：「如果某位病患目前被模型判定為低風險，是否能透過提升某個基因的表現或是降某個指標，使模型改判為高風險？」若能生成這類與目標類別一致的反事實樣本，將有助於強化模型在少數類別上的理解與分類能力。

本研究提出結合生成對抗網路、反事實解釋與 SHAP，不僅可產生具備目標預測結果的反事實樣本，還能在 SHAP 特徵貢獻層面上與目標類別擁有相似的特徵貢獻。這樣的樣本不僅有助於模型學習關鍵決策邊界，也能在訓練資料中補足少數類樣本的多樣性與代表性，進一步緩解資料不平衡所帶來的預測偏誤問題。

第二章 文獻探討

## 2.1 相關研究背景介紹

### 2.1.1 Imbalanced Learning

### 2.1.2 SHAP

SHAP（SHapley Additive exPlanations）是一種以賽局理論為基礎的模型解釋方法，透過計算每個特徵對模型預測結果的邊際貢獻值，來量化特徵對預測結果的重要性。其理論基礎來自 Shapley Value，這是一種源自合作賽局理論的公平分配方式，用以分配多方共同造成的效益

Shapley value公式:

(1)

* N:特徵集合
* S:特徵子集
* :子集S的總貢獻
* :特徵i的Shepley value，代表特徵i的平均貢獻

### 2.1.3 反事實解釋

反事實解釋（Counterfactual Explanation）是一種可解釋人工智慧（Explainable Artificial Intelligence, XAI）技術，其核心概念為找出能夠使模型輸出結果改變的最小輸入變化，藉此揭示模型決策的關鍵因素。此類解釋方式基於反事實推理（counterfactual reasoning）的邏輯，即「若輸入資料在某些特徵上略有不同，模型的預測結果將隨之改變」。

在實務應用中，反事實樣本（counterfactual instances）通常用於回答具有行動導向的問題，例如：「要使模型的預測結果由類別 A 變為類別 B，需要對輸入資料做出哪些具體調整？」此特性使反事實解釋特別適用於決策支援場景。

反事實解釋方法的設計一般包含以下核心原則：（1）最小變化性（Minimality）：生成的反事實樣本應與原始樣本在特徵空間中保持最小差異；（2）有效性（Validity）：反事實樣本必須能夠使模型預測結果產生目標變化；（3）可行性（Feasibility）：生成的變更應符合現實可行性。

因此反事實解釋不僅有助於揭示模型的決策邏輯，也能在保持解釋性與操作性的同時，為使用者提供可落實的行動建議。

### 2.1.4 DICE

### 2.1.5 SMOTE

### 2.1.6 Deep Learning Meets Oversampling

## 2.2 相關研究比較探討

### 2.2.1 SMOTE-XGBoost Classifier Model

Prasher 等人（2023）針對卵巢癌檢測提出結合 SMOTE 與 XGBoost 的方法，以解決小樣本與類別不平衡的問題。SMOTE（Synthetic Minority Over-sampling Technique）透過在少數類別樣本與其鄰近樣本之間進行插值，生成新的合成資料點，相較於單純複製樣本，更能有效增加少數類別的代表性並改善分類器對少數類別的識別能力。在該研究中，作者整合來自 Kaggle 的六個卵巢癌微陣列資料集（GSE6008、GSE9891、GSE18520、GSE38666、GSE66957、GSE69428），共計 529 筆樣本與 20 個基因特徵，先透過 SMOTE 進行資料平衡，再利用 XGBoost 的特徵重要性分數選取關鍵基因（如 EIF1、SPON1、ATG4B、SFN、AP3S1），最後以 XGBoost 分類器進行建模與評估。實驗結果顯示，該方法在測試集上達到 94.14% 的準確率，表現優於 CNN（約 93%）、SVM（約 90%）及其他集成式機器學習方法（約 85%），顯示資料增強與特徵選取結合傳統機器學習模型，能在醫學基因表達資料的分類任務中展現出高效能與良好泛化能力

### 2.2.1 Counterfactual-based minority oversampling

第三章 研究方法

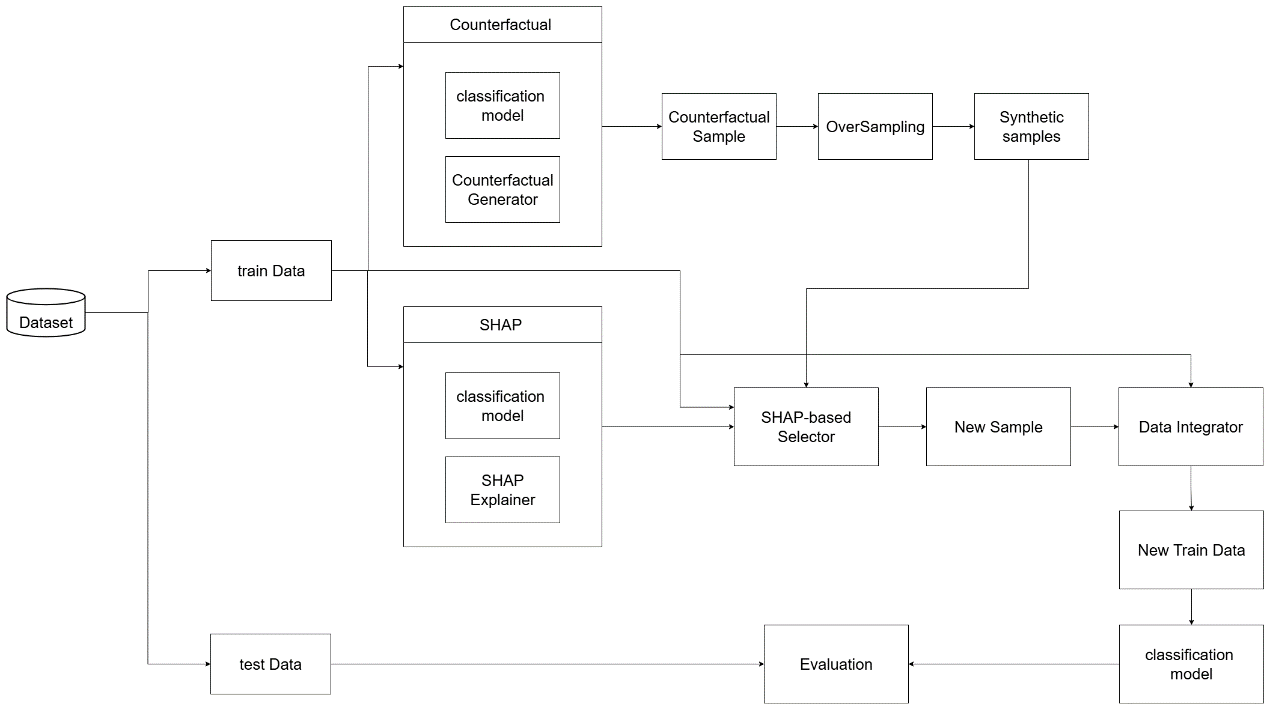
## 3.1 研究議題

在處理類別不平衡資料時，傳統的過採樣方法如 SMOTE 雖被廣泛應用，但其基於 KNN 插值的生成策略存在明顯限制。由於 SMOTE 僅考量樣本在特徵空間中的距離進行線性插值，缺乏模型決策邊界的參考，其生成的合成樣本往往落在類別分界模糊或語意不清的區域，甚至可能與多數類樣本重疊，造成模型誤判，再加上 SMOTE 無法保證生成樣本在語意上與原始樣本一致，容易導致擴增資料無效甚至干擾模型學習。

為解決上述問題，本研究提出一種結合反事實解釋（Counterfactual Explanation）與SHAP 解釋方法的改良式過採樣策略。本研究以反事實樣本作為插值的端點，透過跨越模型邊界的生成機制，產生更具區別性的候選樣本，避免 SMOTE 插值落於模糊決策區域，接著本研究進一步利用 SHAP 所計算的特徵重要度，從多個候選合成樣本中挑選與原始樣本 SHAP 分布最接近者，以確保生成樣本在語意與模型解釋上具一致性與可信度。透過此方法，本研究期望在不平衡資料情境下提升模型對少數類的識別能力，同時保留樣本的合理性。

## 3.2系統架構

本研究提出一套結合反事實生成與 SHAP 解釋值的資料擴增系統，其整體架構如圖一所示。系統主要可分為四個模組：反事實生成模組（Counterfactual Module）、SHAP 解釋模組（SHAP Module）、樣本篩選與擴增模組（SHAP-based Selector & Oversampling），以及資料整合與分類模組（Data Integrator & Classification Module）。首先，訓練資料用於訓練初始的分類模型，此模型將作為 SHAP 解釋器與反事實生成器的基礎。反事實模組針對少數類別樣本生成潛在的反事實樣本，接著經由 SHAP 解釋模組計算原始樣本與候選樣本的特徵貢獻差異，透過 SHAP-based Selector，選出與原始樣本在 SHAP 分布上最相似的前k個合成樣本，作為資料擴增的樣本。最後，這些新樣本與原始訓練資料整合後重新訓練分類模型，並利用測試資料於評估模組( Evaluation )中進行效能驗證，藉此評估資料擴增策略對分類表現的貢獻。



#### 圖1 系統架構

## 3.3方法簡介

首先，對於一筆少數類別的樣本 x1，透過反事實生成器產生潛在的異類樣本x2，接著透過線性插值產生10個候選樣本，插值公式如下:

候選樣本 = (1−α)⋅x1+α⋅x2,α∈[0.1,0.9]

其次，對於每一筆候選樣本，使用 SHAP 解釋器計算其特徵貢獻向量，並與原始樣本 x1進行對比，挑選出前k個 SHAP 差距最小者作為增補資料，最後，將篩選後的樣本與原訓練資料合併後重新訓練模型，以強化少數類在決策邊界上的表現。整體方法結合了反事實樣本的模型導向特性與 SHAP 解釋值的語意解釋能力，在保留樣本多樣性的同時，提升資料的語意一致性與分類效能。

## 3.4 問題與解決方法定義

## 3.5 實驗方法

### 3.5.1 資料集介紹

* Glass
* Breast Cancer Wisconsin

### 3.5.2 比較方法

### 3.5.3 分類器

### 3.5.4****資料擴增流程****

### 3.5.5評估指標

第四章 實驗與討論

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Method** | **Precision** | **Recall** | **F1-score** |
| Glass | AutoSMOTESelf | 67.29 | 62.99 | 62.89 |
| AutoSMOTECohort | 70.95 | 68.58 | 67.87 |
| SHAP-CF-SMOTE | **75.55** | **71.03** | **72.19** |
| Wisconsin | AutoSMOTESelf | **98.00** | **97.88** | **97.92** |
| AutoSMOTECohort | **98.00** | **97.88** | **97.92** |
| SHAP-CF-SMOTE | **97.41** | **96.92** | **97.16** |
| Diabetes | AutoSMOTESelf | 73.42 | 71.64 | 72.10 |
| AutoSMOTECohort | **74.72** | 72.53 | 72.74 |
| SHAP-CF-SMOTE | **72.53** | **73.22** | **72.76** |
| Page-blocks | AutoSMOTESelf | 81.25 | 58.84 | 65.17 |
| AutoSMOTECohort | 60.61 | 72.42 | 60.12 |
| SHAP-CF-SMOTE | **81.31** | **90.96** | **85.29** |

參考文獻

[1] [S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions, "in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4768–4777, Dec. 2017.](https://dl.acm.org/doi/10.5555/3295222.3295230#bibliography)

[2] A. Odena, C. Olah, and J. Shlens, “Conditional Image Synthesis With Auxiliary Classifier GANs,” arXiv preprint arXiv:1610.09585, Jul. 2017

[3] [Cai, J., Lee, Z.-J., Lin, Z., & Yang, M.-R. (2025). A Novel SHAP-GAN Network for Interpretable Ovarian Cancer Diagnosis. Mathematics, 13(5), 882.](https://www.mdpi.com/2227-7390/13/5/882)

[4] [E. Albini, J. Long, D. Dervovic, and D. Magazzeni,“Counterfactual Shapley Additive Explanations,”in Proc. 2022 ACM Conf. on Fairness, Accountability, andTransparency(FAccT'22),pp.1054-1070,2022.doi:10.1145/3531146.3533168](https://dl.acm.org/doi/10.1145/3531146.3533168)

[5] [R. Richman and M. V. Wüthrich,"Conditional Expectation Network for SHAP,"](https://arxiv.org/abs/2307.10654v1)

[arXiv preprint arXiv:2307.10654, 2023.](https://arxiv.org/abs/2307.10654v1)

[6] X. Li, K. Mao, F. Lin, and Z. Feng,“Feature-aware conditional GAN for category text generation,”arXiv preprint arXiv:2308.00939, 2023.