

**資訊工程系碩士班**

**碩士學位論文**

**XAI-latent-SMOTE：結合SHAP與反事實插值的資料增補方法**

**XAI-latent-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](#_Toc211198943)

[ABSTRACT ii](#_Toc211198944)

[目 錄 iii](#_Toc211198945)

[第一章 緒論 1](#_Toc211198946)

[1.1 研究動機與目的 1](#_Toc211198947)

[1.2研究貢獻 1](#_Toc211198948)

[1.3章節編排 2](#_Toc211198949)

[第二章 文獻探討 3](#_Toc211198950)

[2.1 相關研究背景介紹 3](#_Toc211198951)

[2.1.1 Imbalanced Learning 3](#_Toc211198952)

[2.1.2 SHAP 3](#_Toc211198953)

[2.1.3 反事實解釋 3](#_Toc211198954)

[2.1.4 SMOTE 4](#_Toc211198955)

[2.1.5 DiCE 4](#_Toc211198956)

[2.1.6 Tomek Link 5](#_Toc211198957)

[2.2 相關研究比較探討 5](#_Toc211198958)

[2.2.1 ISMOTE 5](#_Toc211198959)

[2.2.1 Counterfactual-based minority oversampling 6](#_Toc211198960)

[第三章 研究方法 7](#_Toc211198961)

[3.1系統架構 7](#_Toc211198962)

[3.2方法簡介 8](#_Toc211198963)

[3.3 問題與解決方法定義 9](#_Toc211198964)

[3.4 實驗方法 9](#_Toc211198965)

[3.4.1 資料集介紹 9](#_Toc211198966)

[3.4.2 比較方法 9](#_Toc211198967)

[3.5.3 分類器 9](#_Toc211198968)

[3.4.4資料擴增流程 9](#_Toc211198969)

[3.4.5評估指標 9](#_Toc211198970)

[第四章 實驗與討論 10](#_Toc211198971)

[參考文獻 11](#_Toc211198972)

1. 緒論

本章將說明本研究之整體背景、動機、研究目的與貢獻。首先，將闡述實務中常見的資料不平衡問題，並介紹本研究提出的結合反事實解釋（Counterfactual Explanation）與 SHAP（SHapley Additive exPlanations） 特徵貢獻之過採樣方法的研究動機與目的，接著介紹本研究的方法與實驗上的主要貢獻與特色。最後，將簡述本論文各章節之安排與內容概要。

## 1.1 研究動機與目的

隨著人工智慧在醫療、金融、製造等領域的廣泛應用，某些領域的資料常呈現嚴重的類別不平衡的問題，導致模型的決策偏向多數類別，而少數類判斷能力不佳，但獲取更多資料的成本過於昂貴，為緩解此問題，過採樣（oversampling）方法被廣泛採用，其中以 SMOTE（Synthetic Minority Oversampling Technique）為代表。雖然 SMOTE為過採樣的經典方法，然而其生成機制以近鄰線性插值為主，容易產生不合理的樣本；在決策邊界或高重疊區插值時，可能會擴大類別重疊——亦即不同類別的資料在特徵空間中重合，進而提高錯誤分類的風險。

近期的改良方法（例如本研究採用為比較基準的 ISMOTE [[7]](#reference7)）嘗試從「幾何生成空間」補強，該方法除了在線段內插值之外，亦可在樣本周邊產生新的樣本，並以隨機與局部分佈特性自動地的微調生成位置，藉此降低高密度區過度擴增與過擬合的風險，使生成分佈更貼近原始樣本分佈。然而，該方法仍可能缺乏特徵合理性上的約束，生成樣本會在與決策無關的特徵上產生偏移、或未能對齊模型對於少數類別的特徵貢獻結構。為此，本研究提出結合「反事實解釋」與「SHAP」之過採樣框架：以「最小變動即可翻轉預測」之反事實觀點，保留樣本的決策邊界意義；再以 SHAP 對齊篩選與約束，使合成樣本的特徵貢獻結構與真實樣本相似，降低傳統插值法造成的不合理樣本與類別重疊。

本研究的目的在於實現並驗證上述的過採樣框架，檢驗其在各種不平衡比例的資料集與多種分類模型（例如隨機森林、CART、KNN）下的預測能力。本研究聚焦於兩個層面：其一是方法論層面，評估本方法是否能在不犧牲多數類表現的前提下提升少數類預測；其二是可解釋性層面，檢驗 SHAP 對齊是否能穩定維持合成樣本的特徵貢獻結構。為此，本研究將在多個不平衡資料集上，與 SMOTE 及其常見變形 ISMOTE [[7]](#reference7)進行比較，並以少數類關鍵指標（如 F1、G-mean 與 AUC ）進行評估，以此確認本方法在不平衡的資料集中的優勢。

## 1.2研究貢獻

本研究提出一個兼具「貼近決策邊界的最小必要變動」與「特徵可解釋性」的過採樣框架，透過反事實解釋與 SHAP 對齊產生合理的合成樣本，同時有效降低傳統插值法容易造成的類別重疊與不合理樣本，提升在不平衡資料情境中模型對於少數類別的判斷能力，並建立一套可重現的評估系統，包含合成樣本品質與特徵貢獻對齊之量化指標與流程，為後續研究與比較提供一致的基準。

## 1.3章節編排

本論文分為五個章節，第二章節為背景知識與文獻探討，介紹可解釋人工智慧的技術、過採樣的方法與其變形。第三章節為研究方法，針對本研究所設計的過採樣方法進行說明。第四章節為對多個資料集與不同模型的評估結果與分析。第五章節為研究結論與未來研究方向。

1. 文獻探討

本章將深入介紹本研究所使用的主要技術與相關文獻。首先說明可解釋人工智慧（XAI）中的 SHAP 與反事實解釋，接著介紹過採樣方法 SMOTE 及其優缺點，再說明本研究生成反事實解釋的方法。最後，在第二節將比較相關改良方法（如 ISMOTE [[7]](#reference7)），以作為後續研究設計的比較基準。

## 2.1 相關研究背景介紹

### 2.1.1 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

### 2.1.2 反事實解釋

反事實解釋（Counterfactual Explanation）是一種可解釋人工智慧技術，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，藉此找出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有些微不同，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。

### 2.1.3 SMOTE

SMOTE（Synthetic Minority Oversampling Technique） 由 Chawla 等人於 2002 年提出，該方法是在特徵空間中以插值的方式合成新的少數類樣本，以緩解類別不平衡對模型造成的影響。具體作法是：對每一筆少數類樣本 ，先找到其 k 個少數類近鄰，隨機挑一個近鄰 ，再於兩者之間做線性插值，插值公式如下:

, (2)

SMOTE 的優點在於：能夠平衡資料集，且由於生成的樣本是線性插值所產生，相較於單純複製既有樣本，通常較不易因重複而導致過擬合。然而缺點在於：線性插值容易在類別重疊或靠近決策邊界時生成含雜訊或不合理樣本，導致模型分類能力下降。

為改善上述問題，後續衍生出多種變體：如 Borderline-SMOTE 專注於邊界樣本以強化分類困難區；SMOTE-Tomek 與 SMOTE-ENN 結合清理法以移除重疊與雜訊；ADASYN 依學習困難度自適應地偏向稀疏區域；SMOTE-NC 處理含類別與連續特徵的資料。這些方法在生成位置與候選樣本選擇策略上更精細，但仍擁有 SMOTE 的插值本質與相關風險。

### 2.1.4 DiCE

Mothilal 等人於 2020 年提出的 Diverse Counterfactual Explanations（DiCE）是一套具代表性的反事實生成框架，其目的是在維持模型預測可翻轉的前提下，同時生成多組合理且多樣化的反事實樣本。傳統的反事實方法往往只能產生單一可行樣本，缺乏多樣性，且部分方法依賴於模型的梯度資訊，難以應用於黑箱模型。DiCE 則透過多目標優化設計，兼顧生成樣本與原樣本之間的距離、可行性、多樣性與預測翻轉性，從而克服上述限制。

DiCE 將反事實的生成視為一個多目標優化問題，在最小化樣本變動的同時，確保生成樣本能改變模型預測結果，並保持資料分佈的合理性。該方法同時引入多樣性正則項，以鼓勵生成多組互不重疊的反事實樣本。DiCE 提供兩種主要的生成策略：其一是基於梯度的優化方法，適用於可微分模型，透過反向傳遞梯度以最小化整體損失；其二是基於資料的搜尋策略，適用於黑箱模型，透過鄰近樣本與啟發式搜索產生合理的候選反事實樣本。

在本研究中，將 DiCE 用於過採樣的架構中，生成接近決策邊界的合成樣本，並作為替代SMOTE 方法的鄰近樣本。

### 2.1.5 Tomek Link

## 2.2 相關研究比較探討

### 2.2.1 ISMOTE

### 2.2.1 Counterfactual-based minority oversampling

第三章 研究方法

## 3.1系統架構

本研究提出一套結合反事實解釋與 SHAP 的過採樣系統，其整體架構如圖一所示。系統主要可分為四個模組：反事實模組（Counterfactual Module）、教師與學生模組(Teacher & Student Model)、SHAP導向的過採樣與樣本篩選模組(SHAP-based Selector & SHAP-CVAE Oversampling)以及分類與評估模組(Classification Module & Evaluator)。首先，資料集切割成訓練集與測試集後，訓練集用於建立初始分類模型。此分類模型同時提供給教師模組與反事實模組使用。在教師模組中，分類器結合 SHAP 解釋器以得到每個特徵對模型決策的貢獻，並將這些 SHAP 向量蒸餾至學生模型 (Student Model)，使其能夠可微分並近似 SHAP 值，以便在後續的SHAP-CVAE中作為損失項使用。接著，反事實模組針對少數類別樣本產生反事實樣本，這些樣本再輸入 SHAP-CVAE 模型進行潛在空間插值與生成，以獲得候選的合成樣本集合。生成後的樣本會經過 SHAP-based Selector，挑選出最佳的k的樣本 ，最後，這些經過篩選的合成樣本與原始訓練資料合併，再由新的分類模型重新訓練。系統最終透過 Evaluator 模組，對測試集進行評估，以驗證本研究所提出之資料增補策略在分類效能上的提升。

Dataset

Training set

Test set

**Counterfacutal**

Classification model

Counterfactual Generator

**Teacher**

Classification model

SHAP Explainer

Counterfactual Sample

Oversampling

Synthetic Samples

SHAP-based Selector

Classification model

Evaluator

Student Model

SHAP-CVAE

#### 圖1 系統架構

### 3.1.1 反事實模組

本研究使用DiCE作為反事實解釋生成的工具，在生成階段中，使用Dice針對少數類別樣本尋找能使分類結果翻轉至其他類別的樣本​，並確保反事實樣本在特徵空間中與原樣本的距離最小化，以維持可解釋性與合理性。本研究將 DiCE 的生成方式設定為 "genetic" 模式，即利用基因演算法（Genetic Algorithm, GA）進行搜尋與優化。該方法以隨機生成的候選樣本為初始族群，透過交配（crossover）、突變（mutation）與適應度評估（fitness evaluation）等操作，不斷演化出同時滿足「分類翻轉」與「特徵擾動最小」條件的反事實樣本。

### 3.1.2 教師與學生模組

由於 SHAP 解釋器本身不可微分，無法直接整合進生成模型的損失函數，

因此本研究設計教師與學生模型，用以建立能夠近似SHAP 特徵貢獻向量的可微分模型。首先，教師模組（Teacher Module）由已訓練完成的分類模型與其對應的 SHAP解釋器所組成，教師模型會輸出SHAP特徵貢獻向量，以此作為學生模型的學習目標。學生模型（Student Model）是一個可微分的多層感知機（MLP）架構，目的是近似教師模組輸出的 SHAP 特徵貢獻向量。學生模型以原始樣本特徵作為輸入，預測對應的 SHAP 特徵貢獻分布。以此讓學生模型輸出的 SHAP 向量能夠加入至 SHAP-CVAE 模型的損失函數中，使生成樣本能夠與真實資料擁有相似的特徵貢獻。

## 3.2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，插值公式如公式(2)。

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 特徵貢獻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。本方法結合反事實解釋的與 SHAP 特徵貢獻，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。

## 3.3 問題與解決方法定義

## 3.4 實驗方法

### 3.4.1 資料集介紹

### 3.4.2 比較方法

### 3.5.3 分類器

### 3.4.4****資料擴增流程****

### 3.4.5評估指標

第四章 實驗與討論

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] [Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT\* '20). Association for Computing Machinery, New York, NY, USA, 607–617. https://doi.org/10.1145/3351095.3372850](https://doi.org/10.1145/3351095.3372850)

[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.

[7] Li, Y., Yang, Y., Song, P. *et al.* An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. *Sci Rep* **15**, 23521 (2025). https://doi.org/10.1038/s41598-025-09506-w