**應用深度學習模型於心肌梗塞致死併發症預測之探討**

**翁鼎傑 賴俊鳴 林祝興**

**東海大學資訊工程學系**

**jasonjason12570@gmail.com; chlin@thu.edu.tw**

摘要

一直以來醫學問題都是具有挑戰性的，並且醫學相關問題的解決需要謹慎與重要性兼具。在近期現代醫學當中雖然使用了現代治療方法，但仍有許多疾病的住院死亡率仍不樂觀。

本篇論文主要是針對心肌梗塞(Myocardial Infarction, MI)患者，在此次住院期間預測是否會有嚴重併發症產生[1]。在本文也利用特徵特性進行缺值填補(MICE)、運行各個演算法包含機器學習(Machine Learning)與深度學習(Deep Learning)來進行預測與比較評測，最終調適出達到97%準確率的深度學習模型。

**關鍵詞**：特徵醫學、深度學習、機器學習、Myocardial Infarction

Abstract

Medical problems have always been challenging, and the resolution of medical-related problems requires more caution and importance. In the recent modern medical, although modern treatment methods have been used, the hospital mortality rate of many diseases is still not optimistic.

This paper is mainly aimed at patients with myocardial infarction (Myocardial Infarction, MI), predicting whether there will be serious complications during the hospitalization period [1]. In this article, we also use feature characteristics to fill in missing values (MICE), various algorithms including machine learning and deep learning, for prediction and comparison evaluation, and finally build a 97% accuracy deep learning model.

**Keywords**: Medical, Deep Learning, Machine learning, Myocardial Infarction

## 緒論

隨著醫療設備與衛生環境進步，傳統致命疾病已由傳染病相關改變為非傳染性疾病。在世界衛生組織(World Health Organization, WHO)公布之西元2019全球十大死因中，主要以三個大主題有關「心血管疾病 (缺血性心臟病、中風)」、「呼吸系統疾病 (慢性阻塞性肺病、下呼吸道感染)」、「新生兒疾病」，尤其以心血管疾病為死亡首因。

自西元2000年以來，缺血性心臟病是全世界最大的死亡主因，並且死亡人數是攀升最急遽的疾病，於2019以來與2018相比增加200萬人次，目前已達890萬人次死亡。心臟病發作與中風通常屬於急病重症，主要是由於堵塞血管導致無法正常流通血液或是腦部，常見原因則是因為過多脂肪堆積於血管與腦部的血管內壁。

心血管疾病內細分為許多疾病，以冠心病(Coronary Syndrome, CS)為首，尤其急性心肌梗塞(Acute Myocardial Infarction, AMI)不管是對非住院病患抑或是住院期間的病患更是會讓其處於一個高危狀態。在以往的病例當中，經過各個研究數據急性心肌梗塞之住院中死亡率平均落至4-20%[2]，並且伴隨著更多的嚴重併發症，包括心源性休克、肺水腫、心肌破裂、血栓栓塞與心室顫動等等，這些急性和亞急性病症通常是致死的主要原因結果。

## 研究目的

針對心肌梗塞病患，大約一半的患者會出現併發症，導致病情惡化甚至死亡，即使是經驗豐富的專家也不能總是預見這些並發症的發展。對此，預測心肌梗死的並發症以便及時採取必要的預防措施是一項重要任務。因此即便有了設備之進步以外，針對預防惡化與預測併發症發生是尤其重要，本篇論文將利用機器學習與深度學習建構出良好的預測模型。

本論文之研究目的為構建出「針對心肌梗塞病患的致死併發症發生預測」模型。利用進院生理數據與患者信息，將每個數據進行分析與拆解，將其作為模型的訓練與驗證，最後製作出適合的預測模型並且對比不同結構的模型。

## 研究方法

本節介紹研究架構、數據集、數據處理、模型探討，這裡會詳細介紹本論文所使用的細節。

### 研究架構

### 數據集

本次選用的資料集為UCI的公開資料集，收集於1992至1995年，並公開於西元2020年12月。這次的這個資料集主要是為了研究與探討有關於「Myocardial Infarction，MI」的相關病症研究，在這次的數據集當中總共有1700筆數據與124個特徵欄位，其中涵蓋的數據有包括說入院時間所有的生理參數數據、第一天、第二天與第三天結束時所疼痛次數與使用具有嗎啡作用的藥物統計，依照所述特徵集合而成的數據集。

在本篇論文所選的資料集當中，含有不少缺少的數據。在所有的特徵值中有110個特徵含有缺失值，大約是88%的欄位皆有缺失值。在個別特徵部分(Serum CPK content (KFK\_BLOOD)、Heredity on CHD (IBS\_NASL))在缺失數值部分高達90%以上，並且有多個特徵值(Diastolic blood pressure according to Emergency Cardiology Team (D\_AD\_KBRIG)、Systolic blood pressure according to Emergency Cardiology Team (S\_AD\_KBRIG)、Use of NSAIDs by the Emergency Cardiology Team (NOT\_NA\_KB)、Use of lidocaine by the Emergency Cardiology Team (LID\_KB)、Use of opioid drugs by the Emergency Cardiology Team (NA\_KB))共計五個缺失率皆大於33%，其餘缺失部分狀態則較為輕微，**表 1**為前三大缺失數據之數值表格。

**表 1 前三大缺失數據**

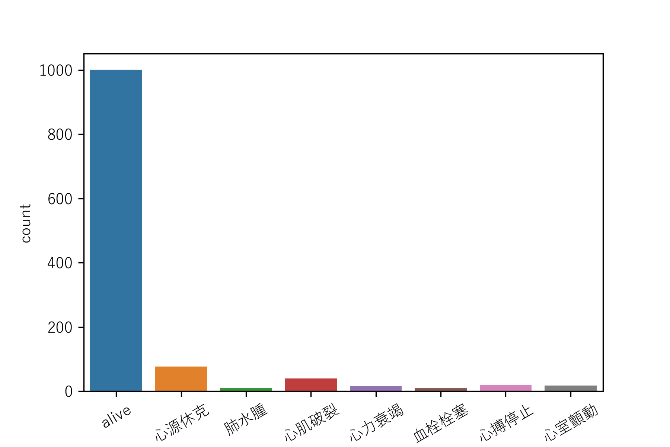
|  |  |  |  |
| --- | --- | --- | --- |
| 特徵欄位 | IBS\_NASL | KFK\_BLOOD | D\_AD\_KBRIG |
| 缺失率 | 99.7% | 95.7% | 63.2% |

在數據集中，我們的目標為構建出「針對心肌梗塞病患的致死併發症發生預測」模型，在採用的目標特徵欄位是「Lethal outcome (cause)」，0的標註為「存活」，其餘則是各項致死併發症，包括「cardiogenic shock」、「pulmonary edema」、「myocardial rupture」、「progress of congestive heart failure」、「thromboembolism」、「asystole」、「ventricular fibrillation」。

**表2** 為Lethal outcome (cause)欄位的原始數據分布，**圖1** 為此特徵欄位的數據分布。

**表 2 Lethal outcome欄位的原始數據分布**

|  |  |
| --- | --- |
| **Lethal outcome (cause)** | **sample counts** |
| Alive | 1001 |
| cardiogenic shock | 70 |
| pulmonary edema | 40 |
| myocardial rupture | 20 |
| progress of congestive heart failure | 17 |
| thromboembolism | 15 |
| asystole | 10 |
| ventricular fibrillation | 10 |



**圖 1 目標特徵欄位數據分布**

### 數據處理

數據缺失是數據集常有的問題，時常因為各種因素與環境，導致無法蒐集完全所有數值，因此這部分是數據處理第一個必須要解決的問題，而在這些丟失的數據當中，又有分為「完全隨機丟失 (MCAR, Missing Completely at Random)」、「非隨機丟失 (MNAR, Missing not at Random)」。這部分將會導致我們後續處理數據集的方式，包括說是否能夠直接丟棄抑或是需要將其作為另一種標示變量。以完全隨機丟失(MCAR, Missing Completely at Random)，數據集丟失的數據與其丟失之機率以及其他特徵值都完全無關；非隨機丟失(MNAR, Missing not at Random)，有些可能的情況以醫學部分來說，可能這個病人病情並不需要插管治療，則他的插管數據則會是缺失的狀態，因此這部分的特徵缺失則可以變成另一種的特徵狀態而去進行預測。所以說以這些情況與數據集，對於數值的有效識別是極為重要的。

而當數據不能以捨棄進行時，勢必會有缺失數據，而缺失數據會造成大量偏差使數據的處理與分析困難，於是「插補」被視為避免使用捨棄欄位的首要方法。目前有一些眾所周知的方法，包括說hot deck imputation、cold deck imputation、mean imputation、regression imputation、stochastic imputation、multiple imputation…等等插補方法，在這篇論文中主要以三種插補方法進行探討。

在本篇論文中，首先針對捨棄動作部分，我們將較高比率之缺失值欄位進行捨棄動作，在這其中捨去的特徵值經觀察為隨機丟失 (MAR, Missing at Random) ，因為過高的缺失若是進行強行補值動作，將會造成整體模型偏移，因此我們這邊將缺失率高於33%之特徵進行捨棄動作，並且我們將以116項特徵進行訓練與預測。**表3** 為本論文捨棄之特徵欄位，**表4** 為數據集可能的並發症（輸出）列，位於特徵欄位第 113-124 列。

**表 3** 本論文捨棄之特徵欄位

|  |  |
| --- | --- |
| **Droped Feature** | **Lost rate** |
| Serum CPK content (KFK\_BLOOD) | 99.7% |
| Heredity on CHD (IBS\_NASL) | 95.7% |
| Diastolic blood pressure Emergency Cardiology Team (D\_AD\_KBRIG) | 63.2% |
| Systolic blood pressure Emergency Cardiology Team (S\_AD\_KBRIG) | 63.2% |
| Use of NSAIDs by the Emergency Cardiology Team (NOT\_NA\_KB) | 40.3% |
| Use of lidocaine by the Emergency Cardiology Team (LID\_KB) | 39.8% |
| Use of opioid drugs by the Emergency Cardiology Team (NA\_KB) | 38.6% |

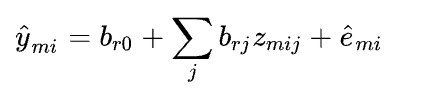
**表 4** 數據集可能的並發症（輸出）列

|  |  |  |
| --- | --- | --- |
| **Output Feature** | **Ratio** | **sample counts** |
| Atrial fibrillation (FIBR\_PREDS) | 10.00% | 170 |
| Supraventricular tachycardia (PREDS\_TAH) | 1.18% | 20 |
| Ventricular tachycardia (JELUD\_TAH) | 2.47% | 42 |
| Ventricular fibrillation (FIBR\_JELUD) | 4.18% | 71 |
| Third-degree AV block (A\_V\_BLOK) | 3.35% | 57 |
| Pulmonary edema (OTEK\_LANC) | 9.35% | 159 |
| Myocardial rupture (RAZRIV) | 3.18% | 54 |
| Dressler syndrome (DRESSLER) | 4.41% | 75 |
| Chronic heart failure (ZSN) | 23.18% | 394 |
| Relapse of the myocardial infarction (REC\_IM) | 9.35% | 159 |
| Post-infarction angina (P\_IM\_STEN) | 8.71% | 148 |

接下來針對缺值補值部分，缺失數據需要進行資料填補，而「插補」是用替換值去與缺失的數值做一個填補 。在本文中我們主要探討的是以下三種方法進行「均值插補 (mean imputation)」、「KNN (k-nearest neighbors)」、「多重插補

(Multiple Imputation)」。

並沒有最正確的，只有最適合的；對於插補的數據來說，能夠符合其原有分佈與數值才是最重要的。在這次所選的三種方法有各自的優點與缺點，並且有各自適合的使用環境。以「均值插補 (mean imputation)」來說，這個方法的核心技術是將其變量的均值替換掉任何缺失值，這個方法的好處是可以讓該變量的樣本均值不會因為插補而造成改變；反之其缺點會造成減弱其插補變量的相關性，造成多變量分析時發生問題。



**圖 2** 均值插補公式

若是以「KNN (k-nearest neighbors)」來說，K-近鄰演算法的核心技術是將計算每一個樣本點的距離，根據距離測量來識別相鄰點，並且可以使用相鄰觀測值的完整值來估算缺值，簡單來說即是以下步驟 ;

1. 計算每一個點之間的距離
2. 利用K值決定鄰近之樣本點數目
3. 進行投票動作（若是在連續型資料當中，則是計算其所有樣本點的平均數）
4. 以投票結果決定類別 ; 以平均數作為最終結果

而距離部分該如何去計算，在存在缺失座標情況之下，KNN通過忽略缺失值並且放大非缺失座標之權重來計算歐幾里德距離（Euclidean distance），**圖3** 是在高維度的情況下的計算公式



**圖 3 歐幾里德距離高維度情況計算公式**

KNN是一個簡易易懂的模型，在多種類別分類時表現優異，然而缺點也顯而易見，所有變數變量的計算相當龐大，每個點之間的距離皆要計算，並且當資料樣本非平衡（Imbalanced Data）時,會加大預測干擾則情況。

最後本論文最後使用的一個插補演算法-「多重插補 (Multiple Imputation)」，在許多情況下我們會因為插補方法無法契合而造成有極多的雜訊，這個方法的核心是利用平均跨過多個插補數據集的結果來解決這個問題，遵循的三個步驟: 「插補」、「分析」、「合併」。本論文中採用的多重插補，是鏈式方程進行的多元插補(MICE， Multiple Imputation by Chained Equations)。在MICE方程底下，預設立場所有缺失的數值都是缺失的，利用其各變數之間缺失的特性，將缺失數值置入其他變數做回歸，並且重複循環直至所有的缺失值都被回歸模型的預測所替代。總而言之，MICE使用了分而治之方法去估算數據集變量中的缺失情況，一次僅觀察一個遍量，使用數據集所有便量來預測該便量的缺失。這樣子的預測模型，基於回歸並且形式取決於關注變量之性質。

另外針對數據集結果不平衡的部分，我們對其做了特別的處理。在這份資料集當中，我們可以明確的知道說有7種結果[**表 2**]，其中一種是活下來，而其於則是因為各種併發症而死亡。所以對總體分類來說，可以大致分為兩類: 死亡與非死亡。在這樣子的狀況下，可以大幅提高資料不平衡的狀況 ，已這樣的處理可以使得整體的資料平衡達到原有的3倍左右，使得整體資料集之目的結果更加明顯。**圖 4** 為總資料集合併預測項目的數據集樣本，**表 5** 為總資料集合併預測輸出樣本數。



**圖 4 總資料集合併預測輸出**

**表 5 總資料集合併預測輸出個數**

|  |  |
| --- | --- |
| **存活個數** | **死亡個數** |
| **1429** | **271** |

但是好景不常，這樣的資料雖然已相較以往提高了不少，但是對於整個模型預測來說仍然處於資料不平衡的狀態，目前的資料平衡比為1:5，目標是希望盡可能的平衡整體狀態。在資料處理中，有方法能夠對資料集數量進行重構，透過個體的複製與刪除，對兩者拉近彼此距離，分別為過取樣(Oversampling)、降採樣(Downsampling)，這些是基於抽樣方法。

本論文選擇採用過取樣方法，除了在經過混淆來打亂樣本順序以外，僅將訓練集之樣本進行過採樣提升目標訓練樣本數，這樣子的做法既可以避免過多的特異性下降並且可以使得整體可預測樣本數量達到平衡，並且沒有經過取樣的驗證樣本可以表示目前的模型是可行的。在本論文中的數據集，將以7:3的比率來分配訓練集與測試集，**表 6** 為實際訓練集與測試集樣本數。

**表 6** 為實際訓練集與測試集原有樣本數

|  |  |
| --- | --- |
| **訓練集總樣本數** | **測試集總樣本數** |
| **1190** | **510** |

**圖 5** 為經過取樣的訓練集，以樣本數看來特徵目標目前已達平衡。**表 7** 為實際預測輸出樣本數。



**圖 5 過取樣後之訓練集**

**表 7 測試集合併預測輸出個數**

|  |  |
| --- | --- |
| **存活個數** | **死亡個數** |
| **476** | **34** |

以這樣子的數據集進行訓練動作，可以大幅提升原始訓練集的訓練效果，讓整體較平衡。

### 模型探討

本論文所選之數據集特性與任務模型為分類型，故我們選擇使用classification 類型模型，採用DecisionTree、Randomforest、Logistic Regression、XGBoost模型進行比對與調整訓練。

在傳統的機器學習模型當中，我們採取了一些比較常見的進行實作訓練與預測，可以看到以這些模型來預測的結果圖為以下:

|  |  |
| --- | --- |
| **Model** | **Precision** |
| DecisionTree | 82.7% |
| Randomforest | 88.2% |
| Logistic Regression | 80.5% |
| K-Nearest Neighbours |  |

## 研究結果

## 研究討論

## 結論

**主要內容**

論文撰寫中英文皆可，請選用標楷體之中文字型及Times New Roman之英文字型，論文標題為粗體字型，點數為16點，作者資料及章節標題亦為粗體字型，點數為12點，圖表說明請用粗體字型，點數為10點，其餘論文內容及參考文獻為標準體字型，點數為10點。中文稿件請附摘要與關鍵詞，英文稿件除附Abstract與Keywords外，尚須中文摘要及關鍵詞。定稿請勿插入頁碼。

標號請用半型阿拉伯數字，標題請靠左，並與前後保持一行的間隔，第一層標題(如前言**.**前言)使用粗體字型，點數12，第二層含以下之標題(如**邊界設定邊界設定**) 則請使用粗體字型，點數11。

* 1. **圖表及公式**

圖形、表格及公式請依先後次序標號，標號請用半型阿拉伯數字，並將圖說撰寫於圖形下方置中，表格說明撰寫於表格上方置中。所附圖表請務必清晰並註明正確來源。以下**圖 1** 為圖形及圖形說明之範例。**表 1** 為表格及表格說明之範例。



**圖 1 大會主題**

* 1. **邊界設定**

文章格式請以A4紙格式撰寫，每一頁請用兩欄格式，行距使用單行間距，上下及左右邊界留白各2.5cm，欄寬7.75cm，兩欄間距0.5cm。最後一頁內容請平均置於兩欄中。

**表 1 邊界設定格式**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 邊界留白(cm) | | | | 欄寬  (cm) | 欄距  (cm) |
| 上 | 下 | 左 | 右 |
| 2.5 | 2.5 | 2.5 | 2.5 | 7.75 | 0.5 |

* 1. **數學公式/方程式**

論文中之數學方程式以阿拉伯數字逐一按出現或引用順序編碼，並加小括號“()”表示之，例如，第四個方程式應表示成“(4)”。展列(display)之方程式應置於版面中間，各方程式編碼一律置於每式在欄之最右側切齊。

如下面之(4)式數學公式/方程式：

 , (4)

方程式應有標點。論文中提及方程式時可用：「第(4)式」、「(5)式」或「方程式(4)」等。本文中所使用之數學符號一律使用斜體字體，如 *x*、*y*、*µ* 或*ω*等，同一符號其大小高度、字體等應與所展列之方程式完全一致。

誌謝

本研究作者特別感謝 科技部贊助經費，使研究得 以順利進行，計畫編號：MOST 107-2221-E-029-005-MY3。

**參考文獻**

按照內文引用的順序，參考文獻需連續地編號，編號寫在中括號內，如 [1]。參考文獻格式如以下[1]-[6]例子所示。內文引用到參考文獻時，只需簡單地提到參考數字，如[3]，不需寫成“參考文獻 [3]”。

除非有六位作者或更多作者，所有作者的名字都要寫出來，而不要使用“…等”。尚未被發表的論文，即使已經投稿了，仍應標示為“未出版(unpublished)” [4]。而已經被接受了的論文應該標示為“即將出版(in press)” [5]。除了專有名詞和元素符號之外，文章題目的第一個字母需要大寫。

1. D. A. Rossiev, S. E. Golovenkin, V. A. Shulman and G. V. Matjushin, Neural networks for forecasting of myocardial infarction complications, The Second International Symposium on Neuroinformatics and Neurocomputers, Rostov on Don, Russia,1995, pp. 292-298, DOI: 10.1109/ISNINC.1995.480871
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.