國立中興大學資訊科學與工程學系

碩士學位論文

評估資料缺失值處理方法在深度神經網路分類準確率之研究

An Empirical Study of Missing-value Imputation Methods on The Accuracy of DNN-based Classification

指導教授：賈坤芳 Kuen-Fang Jea

研 究 生：李侃澄 Yi-Cheng Li

中華民國一百零八年八月

國立中興大學 資訊科學與工程學系

碩士學位論文

題目：評估資料缺失值處理方法在深度神經網路分類準確率之研究

An Empirical Study of Missing-value Imputation Methods on The Accuracy of DNN-based Classification

姓名：李倞澄　　學號：7106056102

經 口 試 通 過 特 此 證 明

論文指導教授 賈坤芳

論文考試委員

黃泓傳

游明宗

賈坤芳

中華民國 108 年 7 月 30 日

中文摘要

缺失資料(missing data)是統計分析中的常見問題。在缺失資料研究的領域發現，如果資料缺失的比率上升，對資料分析的影響嚴重，大量缺失資料將導致分析的結果與事實不符，故資料的缺失值處理極為重要。以往缺失值的研究皆著重於傳統機器學習的方法，而本研究欲使用深度神經網路（deep neural network）作為分類模型，探討資料的缺失處理對深度神經網路分類準確率的影響。

本研究挑選八種處理缺失值的方法做為探討對象。模擬實驗分為三部分，第一部分針對訓練資料有缺失值，第二部分針對測試資料有缺失值，第三部分為訓練資料與測試資料兩者皆有缺失值。依據不同缺失值比例，分別執行實驗比較八種填補法的準確率。

模擬實驗結果顯示，深度神經網路使用 KNN 填補法在不同的缺失值比例與資料集，都比其他填補法具有較佳的準確率。當資料缺失值比例在 5%到 40%之間，使用 KNN 填補法在三個實驗中的準確率平均提高了 6.81%。

關鍵字：缺失資料，填補法，機器學習，類神經網路，分類

Abstract

Missing data is a common problem in statistical analysis. In the field of missing data research, if the proportion of missing data increases, its effect on data analysis can be very serious. A large amount of missing data may make the analysis result deviate from the facts, so missing value handling is extremely important. Previous studies of handling missing values focused on traditional machine learning methods. This study however intends to use the deep neural network (DNN) as a classification model and investigates the effect of various missing data imputation methods on the accuracy of DNN-based classification.

In this study, eight different imputation methods for handling missing values are selected for comparisons. The simulation experiment includes three parts; missing values occurs in the training data, missing values occurs in the test data, and missing values occurs in both of the training and test data. According to the above three cases, simulation experiments are conducted to observe the classification accuracy of the eight imputation methods under different missing ratios.

Experimental results show that the deep neural network using KNN imputation have the best classification accuracy among the eight imputation methods in different missing ratios and data sets. When the missing ratio is between 5% and 40%, the accuracy of KNN imputation in the three experiments exceeds that of other imputation methods by an average of 6.81%.

Keywords: missing data, imputation, machine learning, neural network, classification

目錄

表目錄

圖目錄

第一章 簡介

在巨量資料分析之蒐集資料時，可能因為不同因素導致資料產生缺失。缺失資料是巨量資料分析中的常見問題，資料的缺失值比例低於 1%通常被認為是微不足道的，1-5%是可以控制的，然而介於 5-15%則需要複雜的方法來處理，超過 15%就可能會嚴重影響任何類型的解釋[1]。資料缺失值最常見於問卷調查的資料蒐集中，可能原因有受試者不想回答或是題目描述不清等。而在醫療資料中缺失值的問題也非常嚴重，例如一個病患的身體檢查，可能是不全面的，多項身體檢查中，只挑了部分項目做檢驗，其他項目因而產生缺失值，這就造成了資料分析的難度。在缺失資料研究的領域發現，如果缺失值的比例上升，對資料分析的影響嚴重，以完整資料與含有缺失資料分別作資料分析的結果也相差甚遠。缺失資料越多，對資料所造成的誤差和偏差就越大，分析的結果與事實越不符合，故資料的缺失值處理極為重要。資料的品質是機器學習和其他相關領域中的主要研究課題，由於大多數機器學習的演算法嚴格地從資料中獲取知識，所獲取的知識的品質很大程度上取決於原始資料的品質，而資料品質的一個重要議題就是缺失資料。

針對資料缺失的處理目前已有一些方法被提出。最簡單的處理方法是刪除有缺失值的資料，但此方法的缺點也很明顯，如果資料有較多缺失值，會使大量資料被刪除。有一類為補值法，利用不同的方法對缺失值進行填補，例如使用傳統統計學的估算技巧來填補缺失值，但此法會忽略了資料欄位間的相關性。另一類使用預設模型來填補缺失值，此類方法要先假設資料的分布，再透過統計方法依據資料的分布建構模型。此類方法的缺點為資料的分布通常不會是簡單的分布，在假設資料的分布時，如果與資料實際分布不同，會造成模型的偏差，再以此模型填補缺失值，顯然會產生較差的效果，在第二章相關研究中將較詳細說明這類方法。

Saar-Tsechansky 和 Provost[15]在處理缺失值時，假設資料的缺失只發生在測試資料集，也就是在分類階段資料才有缺失，而訓練資料集是完整的，亦即建立分類模型時原始資料沒有缺失。但在實際情況下，兩種資料集是皆可能有缺失值的，故本研究分成針對訓練資料集有缺失值、測試資料集有缺失值以及兩者皆有缺失值三種情況，來探討在不同階段有缺失值時會對分類模型的準確率造成何種影響。

得力於硬體設備的改善，人工智慧高速的發展，近年來深度學習成為非常熱門的議題。深度學習其實是機器學習的一個分支，在傳統的機器學習中，特徵提取（feature extraction）必須藉由不同演算法來達成，然後再進行分類，其間往往需要專業領域人員來針對資料開發演算法。異於傳統機器學習需經過特徵提取，深度學習透過不同類神經網路層的運算將資料做線性或非線性轉換，自動提取資料的特徵。由於深度學習可以自動做特徵提取，處理資料時可以減少特徵提取的成本，並且與傳統機器學習相比，深度學習透過不斷的訓練類神經網路，能達到更好的效果，突破以往的準確率。

本研究使用深度類神經網路作為分類模型，並針對類神經網路在做分類時，探討資料缺失對其分類的影響程度。以往研究基本上都是基於傳統機器學習的方法，本研究挑選八種常見的處理缺失值的方法，使用類神經網路做分類器，觀察這八種方法對深度類神經網路進行分類時準確率的變化。

本論文後續章節結構如下，第二章為缺失值的相關研究，第三章描述本研究的問題、使用的方法與深度神經網路架構，第四章為實驗結果及分析，第五章則是結論與未來的研究方向。

第二章 相關研究

本章回顧過去的相關研究，包括資料缺失值處理技術以及深度學習的類神經網路架構。

2.1資料缺失類型與處理技術

Allison對於資料缺失值的發生，歸納為以下三種類型[2]:

1.完全隨機缺失（missing completely at random，簡稱MCAR）:資料出現缺失值的原因與現有資料及其他缺失的資料都是獨立無關的，換句話說，資料出現缺失值完全是無規則的。在這種隨機性下，可以運用任何缺失值的處理方法而不存在對數據引入偏差的風險。

2.隨機缺失（missing at random，簡稱MAR）: 缺失資料值的發生與資料中其他變數有關，但是與缺失值本身的數值大小無關，亦即造成特定變數缺失的原因可能與其他因素有關。例如在各年齡層的智商IQ調查資料集中，只有年輕人缺少IQ資料，在這種情況下，缺少IQ資料的機率與年齡有關。MAR是屬於可忽略的缺失類型，即可以將缺失資料直接丟棄而不影響資料分析的結果。

3.不隨機缺失（missing not at random，簡稱MNAR）: 缺失資料的發生與缺失資料本身的值有關。例如在各年齡的IQ調查資料集中，只有IQ較低的部分出現缺失值，這違反了缺失資料是隨機出現的條件。簡而言之，如果違反MAR，即為此類型，是不可被忽略的（non-ignorable）缺失，亦即資料發生缺失時，如為不隨機缺失，不可將缺失資料直接丟棄。

此三種類型與缺失資料的處理關係很大，例如資料為不隨機缺失類型則會使資料的處理非常困難，故本研究假設缺失值的發生為完全隨機缺失類型（MCAR），即缺失值是隨機出現的。

當上述不同類型的缺失值發生時，對資料的分析，會造成一定程度的偏差，也增加分析的難度。在使用有缺失的資料時，我們首先要做的就是處理資料中的缺失值，而在處理缺失值的方法上，有幾種不同處理方法。Rubin將缺失值的處理方式大致分為三種 [14] :忽略並丟棄數據（ignoring and discarding data）、估計（parameter estimation）、補值（imputation）。此三種方式詳述如下:

1.忽略並丟棄數據（ignoring and discarding data）: 簡單地丟棄具有缺失值的資料是研究人員經常採用的方式。如果缺失值為完全隨機缺失類型，則這種處理方法是合適的；但如果缺失值為不隨機缺失類型，則不可使用這種處理方法。若資料有較多的缺失值時，將有缺失值的資料捨棄，會造成大量資料被丟掉，原本的訓練資料集會減小，造成分析效果嚴重損失[2]。

2.估計（parameter estimation）: 這種處理方式就是選定一種分布（例如常態分布、二項式分布、Gamma 分布等），找到適當的參數（例如平均數、變異數、形狀參數），盡量去符合無缺失值的資料。McLachlan 和 Krishnan 提出的 EM 插補法[8]是使用 Dempster 和 Rubin 提出的最大概似估計法（maximum likelihood, ML）[4]先將缺失資料模型化，透過 E 階段與 M 階段，再找出最精確與最合理的估計值，其中 E 階段是做缺失值最佳的估計值，M 階段則求出最大概似估計值再進行取代缺失值，直到估計值的變化可以被忽略為止。這種處理方法在缺失值發生的類型為隨機缺失時，才能使用這種處理方法。

3.補值（imputation）: 此類方法的目的為使用估計值來填補缺失值，利用給定的資料來預測缺失值的可能值。在補值的各方法中有不同策略被用來填補缺失值，本研究的重點是對缺失資料的補值，接下來描述關於這類處理方法的更多細節。

熱卡填補法(hot-deck imputation)[6]是最簡單的補值方法之一，熱卡填補法的基本概念就是依照輔助變數(instrumental variable)不同條件，將未出現缺失值的資料分成若干的插補空格（imputation cell)，然後將每一筆出現缺失的資料，依據其輔助變數的條件，從相對應的插補空格中找尋一筆資料，以其觀測所得的數值來代替缺失值。簡單來說就是觀察完整無缺失值的資料，透過選定的相關條件，將有缺失值資料的特徵值對應到有相同特徵的無缺失值資料，把無缺失資料的特徵值直接填入有缺失值的資料。此種方法常應用在社會科學調查。

另有一些基於統計學技巧的填補方法，如平均值填補法(mean imputation)[7]和中值填補法(median imputation)[1]等。此類處理方法雖然簡單但卻有其缺點，因為這些方法忽略了資料特徵(欄位)間的相關性[16]，並且將缺失值都以同一值填入，造成填補的缺失值都為相同值。又如迴歸填補法(regression imputation)是根據完整資料的迴歸分析來填補缺失值，此法的缺點為所有的估算值都遵循單一的迴歸曲線，並不能表示所有資料的既有差異性。

而同樣還有不考慮資料特徵(欄位)間的相關性，但考慮到同個特徵中的分布的插值填補法（interpolation imputation）。M.N. Noor 等人[11]研究指出，在資料為 Gamma 分布時使用線性插值法（linear interpolation）的效果很好，與平均值填補法和中值填補法不同的是，插值填補法會依據缺失值前後的值，來對缺失值的分布做預測，故此方法也適合應用在時間序列的資料。

隨著機器學習技術發展，基於傳統統計技巧來填補缺失值所帶來的缺點，也逐漸被克服。例如利用決策樹(decision tree)[12]或關聯法則(association rule)[13]以及 KNN (K-nearest neighbors imputation)[3]填補缺失值的方法等。決策樹會遞迴的根據資料的特徵，將特徵分支以建構一棵樹，並利用 information gain 或 gini index 來計算要以何特徵來當作決策樹的節點，此步驟一直重複至整棵樹建構完成為止。不過嚴格來說，此種處理方法不能歸類為填補法，因資料有缺失值時，是透過計算機率來判斷要分類的類別，將有缺失值的情況做分支，而不是填補缺失值，故本研究不討論此方法。

關聯法則是找出資料項目集間的出現關係所建立出來的規則。假設給定兩個項目集X和Y，若X出現時Y通常也會出現，則以關聯法則表示為X ⇒Y。在計算項目集間的關聯度時會用到兩個參數：支持度（support）與置信度（confidence），支持度(support)定義為同時包含X和Y的資料數／資料總數，置信度(confidence)

定義為同時包含X和Y的資料數／包含X資料數。支持度（support）與置信度（confidence）為自定義參數，給定最小支持度（support）與最小置信度（confidence）後，在資料中找出超過最小支持度（support）的項目集，這些項目集稱為頻繁項目集（frequent itemset），再從頻繁項目集中選出高過最小置信度的項目集以產生關聯法則，但此方法通常用在資料為類別型或是需先將數值資料離散化後方可使用。

Batista [3]提出 KNN 填補法，從完整資料中以距離函數計算出與有缺失資料最隣近的 K 筆資料，稱為 K 個最近鄰居（nearest neighbors）。依這 K 筆資料的相對應於缺失值的特徵，計算這 K 筆資料的特徵值作平均，再將此平均值填回缺失值中，這裡也可以用中值代替平均值。在 KNN 填補法中距離函數常使用歐氏距離（Euclidean distance）與皮爾遜積矩相關係數(Pearson product-moment correlation coefficient)，此方法簡單且有效，但缺點是計算量大。

Saar-Tsechansky and Provost[15]提出了針對測試資料使用簡化特徵模型。在此研究中假設訓練資料室完整而無缺失值的，資料的缺失只出現在測試資料，具體處理方法為針對測試資料集有缺失值的特徵，對應到訓練資料集中的特徵，將訓練資料集中相對應特徵刪除。例如測試資料集中，總共有六個特徵，其中性別與身高兩個特徵有缺失值，因此把測試資料集中這兩個含有缺失值的特徵刪除而保留其餘四個特徵。反過來，在訓練資料集中也把性別與身高兩個特徵刪除，將資料從原本的六個特徵變為四個特徵，然後依此建立分類模型。他們的實驗結果顯示此方法處理缺失值有不錯的表現，但此方法會減少特徵數目，可能會對類神經網路的學習不利，不適用於類神經網路的分類。在訓練類神經網路時，通常希望資料量與特徵越多越好，若減少特徵則可能造成準確率下降。

以上缺失值的處理方法各有優缺點，每種方法的效果也有差異。在探討缺失值處理方法的良窳時，可以引入分類模型，來觀察其對分類模型的準確率所造成的影響。Acuña 與 Rodriguez [1]使用了線性判別分析（linear discriminant analysis)和 KNN 分類器兩個分類器，在處理缺失值時使用忽略並丟棄數據的方法、統計的平均值填補法和中值填補法以及 KNN 填補法這四種方法，來比較兩種分類模型的準確率，藉以評斷這四種處理缺失值的好壞。他們實驗的結果顯示，平均值填補法和中值填補法的準確率差異不大，忽略並丟棄數據的方法效果最差，KNN 填補法的效果表現最佳。

在醫療資訊的大數據分析中，缺失值一直是一個大問題，因為大多數醫療資料的缺失是非常嚴重的。Mundfrom 和 Whitcomb[10]使用心臟病患者的資料，先以完整的資料分別使用線性判別分析和邏輯回歸模型（logistic regression model）來當做分類器，接著再使用有缺失值的資料去分別比較使用平均值填補法、線性回歸填補法和熱卡填補法後的效果。他們實驗的結果顯示，使用線性判別分析分類時，平均值填補法的效果最佳，而使用邏輯回歸模型分類時，熱卡填補法效果最佳，由此可看出不同缺失值填補法對不同分類器的效果不盡相同。

2.2 深度學習的類神經網路架構

以往研究所採用的分類器，都是使用傳統的統計分類方法或是機器學習的方法，然而其準確率都不如深度學習。Eldem 等人[5]使用深度類神經網路來分類三種不同類型的鳶尾花科植物。在該研究中，以花瓣和萼片的寬度和長度作為輸入特徵，藉由不同的激活函數（activation function）和不同的訓練次數來對鳶尾花資料集進

行訓練，選出對鳶尾花資料集效果最佳的激活函數，最後實驗結果顯示對鳶尾花資料集的分類準確率可高達 96%。

在影像處理的分類應用中，現今主流的深度學習架構通常是使用卷積神經網路(Convolutional neural network, 簡稱 CNN)，將影像資料輸入卷積神經網路自動對圖片做特徵提取、降維等運算，訓練網路完成後進行分類，可以達到很高的精確度，但是卷積神經網路的計算量非常大，訓練的時間非常長。Seetha [17]等人針對腦部腫瘤的影像，使用了卷積神經網路，透過 CNN 卷積神經網路做特徵提取及降維，得到特徵後再進行分類，準確率可達到 97.5%。有趣的是，Mohsen[9]等人使用了深度神經網路加上傳統影像處理方法，來對腦部腫瘤做分類。然而因為深度神經網路在影像方面無法如同卷積神經網路自動做特徵提取及降維，所以在前處理方面他們採取了傳統影像處理方法，使用了離散小波變換(discrete wavelet transform)來提取特徵，並且使用主成分分析(principal components analysis)來降維，將處理後得到的特徵向量，再輸入深度神經網路訓練。此方法比卷積神經網路節省了大量時間，且最後他們得到的準確率為 96.97%，與卷積神經網路相比差異太大，由此可看出深度神經網路具有一定的優勢。

以往研究在處理缺失值時，大多是使用傳統的分類器。然而本研究使用上述的深度神經網路作為分類模型，探討不同缺失值處理方法對深度神經網路分類準確率的影響。深度神經網路的基本架構為一層輸入層，兩層隱藏層以及一層輸出層，依據不同的資料在隱藏層中的節點數會隨之調整。圖 2.1 為本研究所使用的類神經網路基本架構圖。

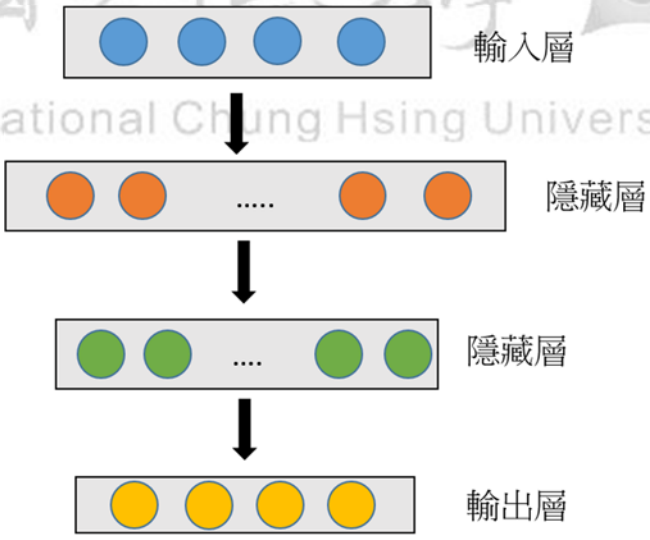

圖 2.1 深度神經網路架構圖

此架構中，不同層中的節點數都不相同，輸入層的節點數為輸入資料的特徵數。假設資料有三個特徵，輸入層的節點數為三，而隱藏層中的節點數可自行調整，輸出層則是資料的類別數，不同資料的隱藏層適合的節點數皆不同，透過實驗分別以不同節點數訓練類神經網路，找到較佳的組合。不同層之間為全連接，通常在資料數較少時，全連接的類神經網路容易造成過適(Overfitting)的情況。過適可以使用 dropout[18]來防止過適，其基本概念將原本為全連接的節點，隨機挑選一些節點忽

略，變成不完全連接，在訓練時不會更新所有節點的權重，如此便可以防止過適的情況。

在訓練類神經網路時，每個節點之間有一個權重值，權重值由訓練類神經網路得到。權重值最初為隨機值，每次訓練都會更新節點的權重值，透過倒傳遞(backpropagation)演算法，從輸出層的節點開始，使用梯度下降(gradient descent)來逐漸修改權重值，最後便可得到整個類神經網路中每個節點的權重值。梯度(gradient)定義為：

梯度= -2 *輸出值* (實際值－預測值)

輸出值為前一層節點的輸出值，實際值為輸入資料中要預測欄位的實際值，預測值為輸入資料輸入類神經網路得到的預測值。更新的權重值=原權重值-(學習速率*梯度)。學習速率為自定義參數，範圍為 0 到 1，學習速率越大，權重值下降越快。每層可以選定不同的激活函數（activation function），將前一層節點的輸出乘上節點的權重值後，代入不同的激活函數得到一個值，這個值就為下一個節點的輸入值，依同樣方法向後傳遞到輸出層。

每層中使用的激活函數可以自行定義，常用的有 sigmoid、tanh 和 ReLU。輸出層在分類時，通常使用 softmax。

圖 2.2 為 sigmoid 函數。圖中可以看出，代入函數後的值範圍在 0 到 1，當 x 越小，其值越趨近於 0；當 x 越大，其值越趨近於 1。圖 2.3 為 tanh 函數圖，y 的範圍為-1 到 1，當 x 越小，其值越趨近於-1；當 x 越大，其值越趨近於 1。圖 2.4 為 ReLU 函數圖，當 x 小於 0 時，輸出值為 0；當 x 大於 0 時，則輸出 x。此三種為常見的激活函數。

sigmoid 公式： $S(x) = \frac{1}{1+e^{-x}}$

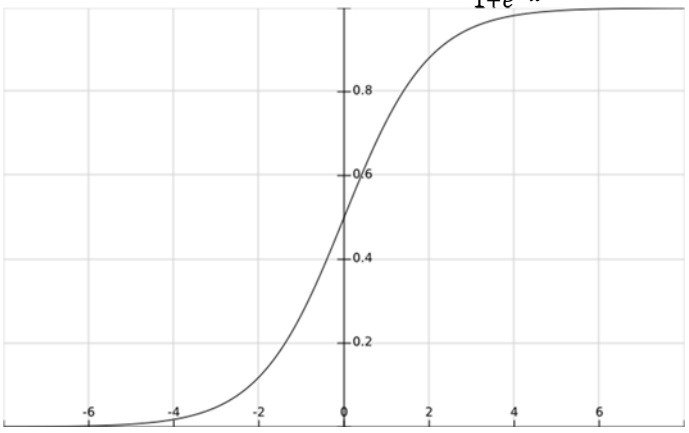

圖 2.2 Sigmoid 函數圖

tanh 公式：$T(x) = \frac{2}{1+e^{-2x}} - 1$

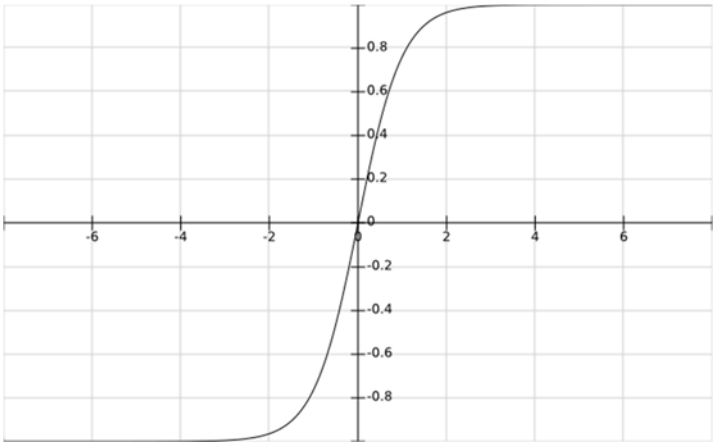

圖 2.3 tanh 函數圖

ReLU 公式：$R(x) = 0, \mathrm{x} < 0; \mathrm{x}, \mathrm{x} > 0$

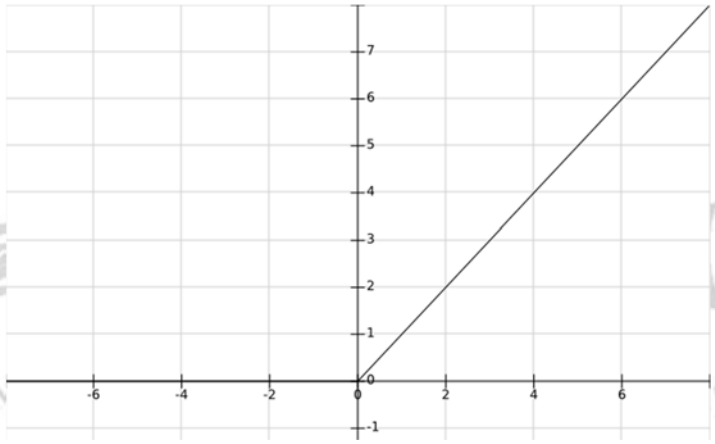

圖 2.4 ReLU 函數圖

Softmax 公式：$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$ for $j = 1, \ldots, K.$ K 為向量維度

在以類神經網路作分類的應用上，softmax 為輸出層所使用，此函數將輸出層前面一層的值做正歸化（normalized），也就是將所有值轉化成其加總值等於 1。類神經網路在分類時，輸出層裡的節點相對應於資料的類別，每個節點經過 softmax 函數後，所有數值的加總值為 1，每個值可視為機率，從其中選出最大的值，表示為該類別的機率最大，類神經網路將把此筆資料判定為該類別。

第三章 問題與方法描述

3.1 問題定義

本研究探討在使用深度神經網路(DNN)作分類時，若資料發生缺失，使用各種缺失值處理方法對深度神經網路的分類準確率所造成的差異。本研究探討的缺失值處理方法分別為：特殊值填補法、資料均值填補法、資料類別均值填補法、資料中值填補法、資料類別中值填補法、內插值填補法、KNN 填補法和簡化特徵模型等八種。

異於以往研究大多探討傳統機器學習的分類，本研究針對數值資料的分類選用深度神經網路為分類模型來探討資料缺失的問題，此模型廣泛應用在分類上，並且分類的表現相較傳統機器學習更佳。

本研究假設資料缺失的發生類型為完全隨機缺失，即資料出現缺失值是隨機發生的，不會因資料的特性、資料本身數值的大小，或資料中的特徵等而出現特定缺失。

3.2 研究方法

資料缺失可以分為三種情況：（1）在訓練階段時訓練資料有缺失值（2）在分類階段時測試資料有缺失值（3）在訓練階段與分類階段時的訓練資料與測試資料皆有缺失值。本研究依上述三種情況，分別使用不同處理缺失值的方法填補缺失資料，再使用 DNN 建立分類模型來預測類別。DNN 的架構如第二章所示，使用了二層隱藏層，每個隱藏層中使用 ReLU 作為激活函數，輸出層使用 softmax 作為分類層，並分析每種缺失值處理方法對 DNN 的分類所造成的影響。

（1）在訓練階段時訓練資料有缺失值：

輸入資料含有缺失值，因此在訓練類神經網路前，需先處理有缺失值的資料。在預先處理階段，我們分別使用了幾種不同的填補方法，將缺失值填補完後，輸入 DNN 中，建立出分類模型。最後分析各種不同填補方法對 DNN 分類準確率的影響程度。

（2）在分類階段時測試資料有缺失值：

因為輸入資料無缺失值，故直接將資料輸入 DNN 中，建立出分類模型。但在此情況下必須對測試資料中的缺失值做處理，才能進行分類。我們同樣使用不同方法來填補測試資料中的缺失值，最後分析這些填補方法對 DNN 分類準確率的影響程度。

（3）訓練階段的訓練資料與分類階段的測試資料皆有缺失值：

輸入資料含有缺失值，所以在訓練類神經網路前，需先填補有缺失值的資料。我們在預處理階段分別使用了幾種不同的填補方法，將填補完整的資料輸入 DNN

中，建立出分類模型。測試資料的缺失值也必須處理，因此同樣使用不同填補法填補缺失值，最後分析在此情況下各種填補方法對 DNN 分類準確率的影響程度。

在上述三種情況下，我們探討在各種不同的資料缺失比率，分別為 5%、10%、20%、30%、40%，透過模擬實驗觀察、分析 DNN 分類的準確率變化。

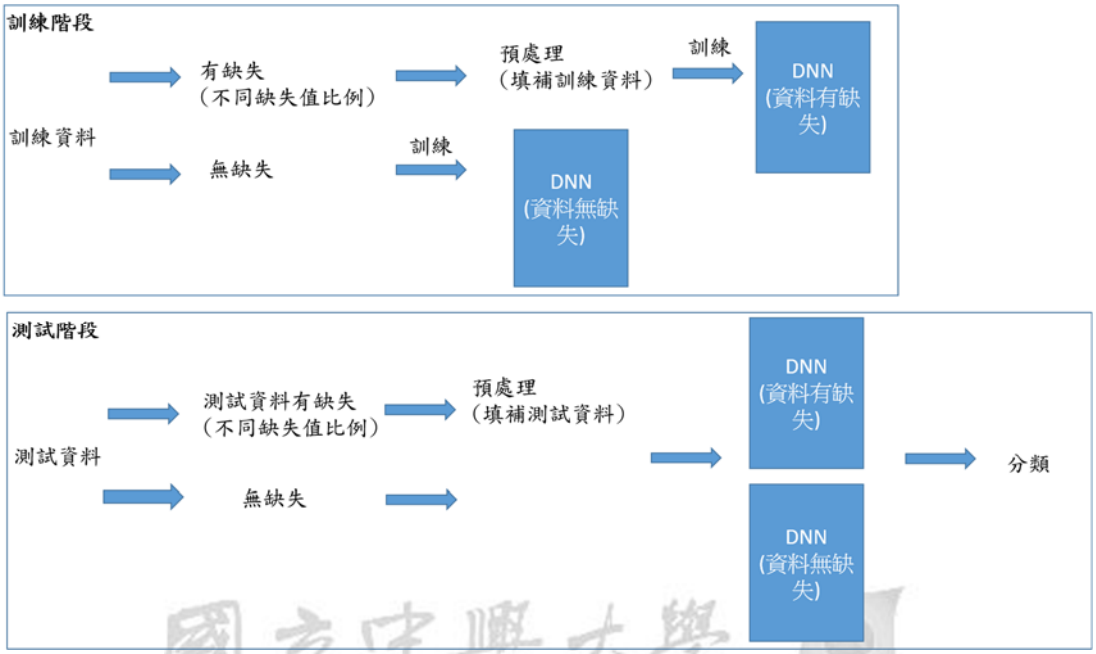

圖 3.1 研究方法架構圖

3.3 處理缺失值方法

根據第二章相關研究中常見的處理缺失值方法，本研究選定其中八種方法作為研究對象。我們先以特殊值來填補缺失值，實驗 DNN 是否會將特殊值當成特例而忽略缺失值，不會去修改類神經網路中節點的權重值。接著我們以統計的技巧，針對有缺失值欄位以其他無缺失值資料的平均值（mean）與中值（median）分別來填補缺失值。我們再將上述資料均值填補法與中值填補法做延伸，依據缺失值欄位資料的類別，計算同類別資料的平均值與中值，然後以此值填補同類別資料的缺失值。

此外，本研究針對同一特徵分布選定內插值填補法（interpolation-value imputation）來填補缺失值。異於資料均值與中值填補法，內插值填補法為預估特徵的連續分布，缺失值填補是依據缺失特徵的前後幾筆資料，利用已有的資料點求出多項式函數來預估缺失值。

本研究也選定 KNN 填補法來填補缺失值，此方法與其他方法最大的差異就是它考慮了每筆資料之間的關係。KNN 填補法找到與缺失資料最接近的 K 筆資料，將這些資料的特徵值做平均再填入缺失值中，此方法簡單有效，但計算量大、耗時較久。

最後本研究針對測試資料有缺失值採用了簡化特徵模型(feature reduction model)[15]，此方法實際上不是填補法，它將測試資料中有缺失的特徵刪除，並將訓練資料相對應特徵刪除，因此會改變資料的維度導致深度神經網路可以用來訓

練的特徵變少，有可能造成準確率下降。然而 Saar-Tsechansky 等人[15]的研究指出簡化特徵模型有好的準確率表現，與填補方法不相上下，因此本研究納入此方法做為比較對象。以下詳細說明各個選定的方法

方法一：特殊值填補法（unique-value imputation）

此法以一個不在正常取值範圍內的數值，比如-999 或 0 等，來表示缺失值，將缺失值視為一種特殊的值來處理。它不同於此特徵的任何值，而是將缺失值本身視為一個新值。

方法二：資料均值填補法（mean imputation）[7]

此法為處理缺失值常用的方法之一，將有缺失值的欄位用該欄位其他無缺失值的欄位值做平均，取代缺失值。

方法三：資料類別均值填補法

此法依據訓練資料中某些欄位的各類別分別取資料均值，然後針對有缺失值的資料依其類別以類別均值來填補缺失值。由於在訓練階段的資料有類別欄位，以其類別作為依據，將相同類別資料各自分別計算平均值，例如若此資料總共有三個類別，則將資料分為三份，再依各類別資料的欄位平均值來填補缺失值。

方法四：資料中值填補法(median imputation)[1]

此法與**資料均值填補法**類似，將有缺失值的欄位，使用該欄位其他無缺失值的欄位值，取這些值的中間值來填補資料中的缺失值。

方法五：資料類別中值填補法

此法與資料類別均值填補法類似，依據訓練資料中某些欄位的的類別分別計算該類別的中值，然後針對有缺失值的資料依其類別以類別中值來填補該資料的缺失值。

方法六：內插值填補法(interpolation-value imputation) [11]

此法藉由無缺失值的離散資料點，在其範圍內推求一方程式，依此方程式計算一推估值，然後以此值來填補缺失值[19]。假設有兩資料點（x0，y0），（x1，y1），y=f（x）為一條直線，通過已知的此兩點。若某資料點(x’，y’)中的 y’是缺失值，將 x’代入 f（x），即可得到 f（x’）來填補缺失值。此法可以讓缺失值與資料分布較相近，但與**資料均值填補法**、**資料中值填補法**相比，卻不會將缺失值都以相同的固定值代入。f（x）不一定是直線方程式，可以為多項式。

方法七：KNN 填補法(K-nearest neighbors imputation)[3]

實際上 KNN (K-Nearest Neighbors)方法沒有訓練模型，而是使用每筆資料的特徵來計算與其他筆資料的距離，找出 K 筆最鄰近的資料來推求某測試資料的特徵值，此演算法的概念為相近的資料通常會有類似的特徵。K 為自定義變數，表示 K 筆鄰近的資料。在 K-Nearest Neighbors 演算法中常使用的距離函數包括 Euclidean distance metric 和 Pearson correlation metric 等。假設有兩個 n 維的資料點 X 與 Y，

點 X=(X1, …, Xn)，μX 為 X1, …, Xn 的平均數，其標準差為 σX，點 Y =(Y1, …,Yn)，μY 為 Y1, …, Yn 平均數，其標準差為 σY。
兩個資料點 X 與 Y 的 Euclidean distance 定義如下：

$$D(X，Y) = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2}$$

兩個資料點 X 與 Y 的 Pearson correlation 定義為:

$$P(X，Y) = \frac{\sum_{i=1}^{n}(Xi - \mu X)(Yi - \mu Y)}{\sigma X \sigma Y}$$

本研究使用 Euclidean distance 來計算距離。此方法將缺失值當作要預測的特徵，透過其他筆無缺失值的資料，使用距離公式計算出 K 筆相近的資料，將這 K 筆資料相對應的特徵值計算平均值或中值，再將此值填補缺失值。

方法八：簡化特徵模型(feature reduction models)[15]

此方法為針對測試資料有缺失值時使用，其作法是將測試資料中發生缺失值的特徵刪除，並把訓練資料相對應的特徵刪除，藉以達到測試資料變成無缺失值的情況，但此方法會減少特徵數目。

3.4 改變深度神經網路架構

深度神經網路的架構可以隨不同情況調整，本研究使用的網路架構為壹層輸入層、兩層隱藏層和壹層輸出層，隱藏層中的節點會隨資料的不同而調整。類神經網路架構越複雜，層數與節點數越多，神經網路能學習到的權重值就越多，對於準確率的提昇可能會有幫助。

本研究亦探討當資料有缺失值時，改變神經網路的架構對準確率所造成的影響，是否可能透過較多的隱藏層與節點數而提升準確率，研究過程如圖 3.2。我們針對更改後的 DNN 與使用原本架構的 DNN 進行實驗比較，分析兩者分類的準確率差異。

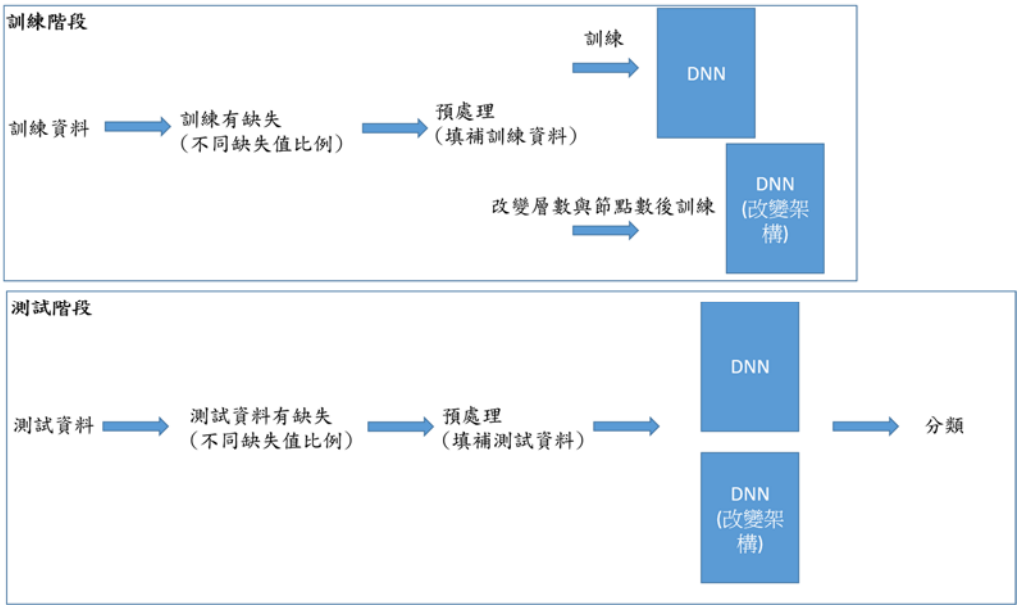

圖 3.2 改變深度神經網路架構圖

第四章 實驗與結果分析

本章於 4.1 節說明實驗的環境以及實驗所使用到的資料來源。4.2 節以實驗觀察訓練資料有缺失值時，七種處理缺失值方法對 DNN 分類準確率的影響情形。4.3 節以實驗觀察測試資料有缺失值時，五種處裡缺失值方法對 DNN 分類準確率的影響情形，並比較填補法與簡化特徵模型對 DNN 分類準確率的影響情形。4.4 節以實驗觀察訓練資料與測試資料皆有缺失值時，五種處理缺失值方法對 DNN 分類準確率的影響情形。4.5 節以實驗觀察若改變 DNN 架構對有缺失值資料的分類準確率影響情形。4.6 節為本章的實驗小結。

4.1 實驗環境

4.1.1 實驗資料來源

本研究的模擬實驗使用 UCI Machine Learning Repository 中的 abalone data set[20](共 4,177 筆，8 個特徵)、 landsat satellite data set[21](共 6435 筆，36 個特徵)以及 heart disease data set[22](共 303 筆，13 個特徵)，並取 70%的資料作為訓練資料，其餘 30%的資料作為測試資料。

4.1.2 實驗平台

模擬實驗的硬體設備包括：處理器為 Intel(R) Core(TM) i7-7700 CPU @ 3.6GHz，記憶體為 16GB；作業系統為 Microsoft Windows 10 企業版 64bit。本研究所有演算法皆採用 python 程式語言編寫，透過 spyder 進行編譯執行，並利用 Microsoft Office Professional Plus 2010 Excel 來輔助實驗結果分析。

4.2 實驗一：訓練資料有缺失值

4.2.1 實驗設計與實作

本實驗針對訓練資料集有缺失值時，在不同的資料缺失值比例下使用七種填補法填補訓練資料集，以之訓練 DNN，然後輸入測試資料集觀察其分類準確率。本實驗為訓練資料集有缺失值，所以未使用簡化特徵模型進行實驗，簡化特徵模型是訓練資料有缺失值時的處理方法。

本實驗假設訓練資料集有缺失值，所以我們以隨機函數挑選訓練資料集的資料刪除其特徵值(使該筆資料出現缺失值)。由於是隨機挑選資料產生缺失值，所以

符合資料為完全隨機缺失類型。我們使用隨機函數 rand()控制資料缺失值出現的比例分別為 5%、10%、20%、30%和 40%。

4.2.2 實驗結果與分析

表 4.1、4.2、4.3 分別為訓練資料集在不同資料缺失值比例下，使用七種補值法與未填補缺失值對 DNN 的分類準確率。

由表 4.1、4.2、4.3 中可以看出當缺失值未填補時，DNN 中節點的權重值不會更新，所有的權重值皆為一開始的隨機值，因此 DNN 在不同缺失值比例下分類的準確率皆相同且非常低。而缺失值以特殊值取代時，DNN 的準確率皆非常低，因為 DNN 不知該特殊值為特例，只是將它視為一般特徵值。

表 4.1 abalone 訓練資料集在不同缺失值比例下七種填補法與未填補缺失值對 DNN 的準確率

	5%	10%	20%	30%	40%
未填補缺失值	36.58	36.58	36.58	36.58	36.58
特殊值填補法	61.05	55.69	54.55	52.06	50.05
資料均值填補法	83.83	75.36	64.79	59.81	54.35
資料類別均值填補法	83.64	76.6	65.07	58.56	54.55
資料中值填補法	83.75	76.65	65.36	60.57	53.11
資料類別中值填補法	82.01	77.22	65.55	58.56	54.83
內插值填補法	85.74	79.85	70.72	64.02	60.48
KNN 填補法	87.37	83.73	73.59	65.84	60.67

表 4.2 landsat satellite 訓練資料集在不同缺失值比例下七種填補法與未填補缺失值對 DNN 的準確率

	5%	10%	20%	30%	40%
未填補缺失值	24.17	24.17	24.17	24.17	24.17
特殊值填補法	86.1	82.8	65.95	56.4	53.25
資料均值填補法	91	90.35	89.7	88.7	86.5
資料類別均值填補法	90.8	90.5	88.85	88.05	86.3
資料中值填補法	90.55	89.3	89.5	88.05	85.1
資料類別中值填補法	90.95	90.45	89.05	87.65	85.8
內插值填補法	90.7	90.8	90.4	89.3	89.15
KNN 填補法 (K=7)	90.15	90.05	90.25	89.95	89.5

表 4.3 heart 訓練資料集在不同缺失值比例下七種填補法與未填補缺失值對 DNN 的準確率

	5%	10%	20%	30%	40%
未填補缺失值	54.73	54.73	54.73	54.73	54.73
特殊值填補法	78.38	79.05	79.73	79.05	75.68
資料均值填補法	83.1	82.44	83.28	82.43	81.08
資料類別均值填補法	82.09	82.44	82.26	82.09	81.42
資料中值填補法	82.78	81.59	81.01	81.08	80.41
資料類別中值填補法	82.43	82.78	83.11	82.77	82.27
內插值填補法	83.78	83.45	82.77	82.43	81.08
KNN 填補法 (K=7)	82.1	83.11	82.43	82.43	82.09

圖 4.1 為 abalone 資料集使用七種填補法填補缺失資料後，分別訓練 DNN 後所得到的準確率。經過實驗後，我們將 KNN 填補法的 K 值選定為 7，因為在 K 值為 7 時，DNN 分類的準確率最高。為了公平比較，在實驗二與實驗三中 KNN 填補法的 K 值也皆為 7。在各填補法中，KNN 填補法使得 DNN 的準確率最高，內插值填補法相較於資料均值、中值填補法對 DNN 有較高的準確率。在資料缺失值比例為 10%時，使用 KNN 填補法的準確率還維持在 80%以上。此資料集隨著缺失值比例上升，準確率下降劇烈，此資料集受到缺失值比例的影響較嚴重。以效果較佳的 KNN 填補法來觀察，缺失值比例為 5%與 40%時，其準確率差了大約 27%。

值得注意的是資料類別均值、中值填補法與資料均值、中值填補法從實驗結果顯示，它們的效果是差不多的，針對類別將缺失資料分別取均值和中值填補缺失值，其功效是微乎其微的。

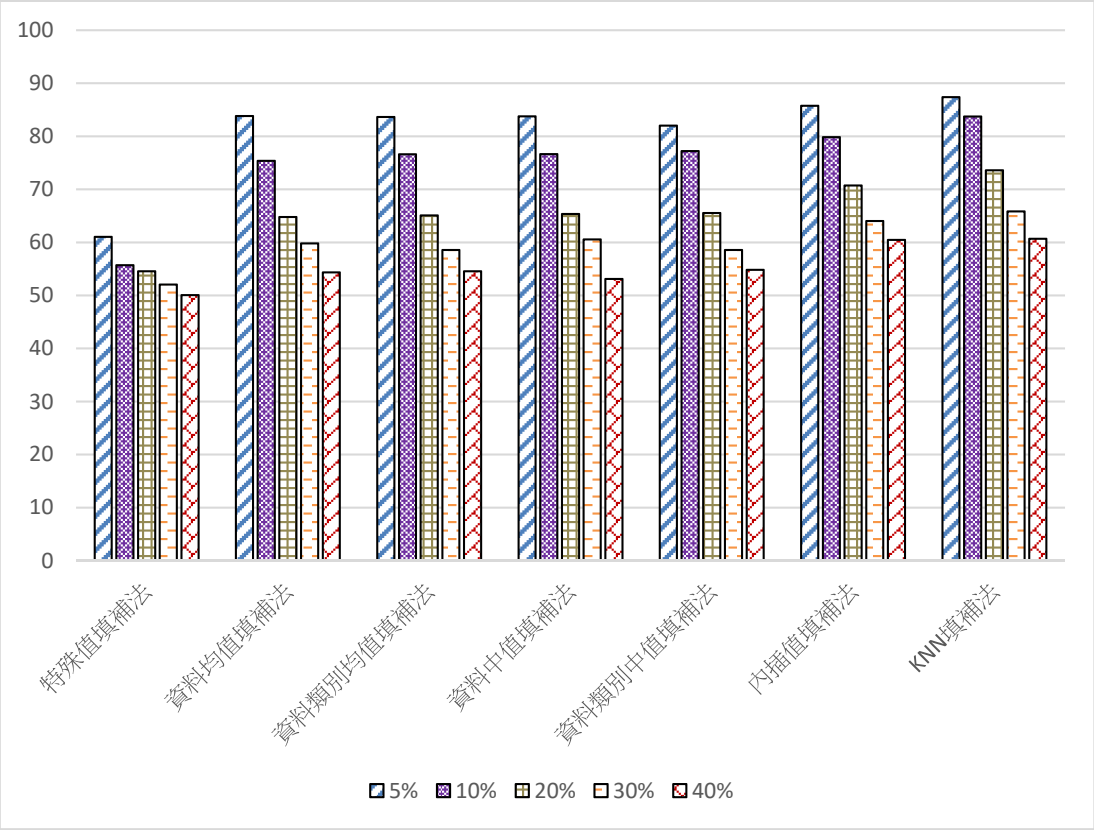

圖 4.1 在不同缺失值比例下各填補法填補 abalone 訓練資料集訓練 DNN 所得之準確率

由圖 4.2 可以看出 landsat satellite 資料集在訓練資料有缺失值時，DNN 準確率的差異不大，使用不同方法填補資料，分類資料的準確率都維持在 90%左右，準確率只有在缺失值比例 40%時會稍微降低到 80%。landsat satellite 資料集隨缺失值比例增加，準確率的下降較平緩，此資料集與 abalone 資料集受缺失值影響大不同。

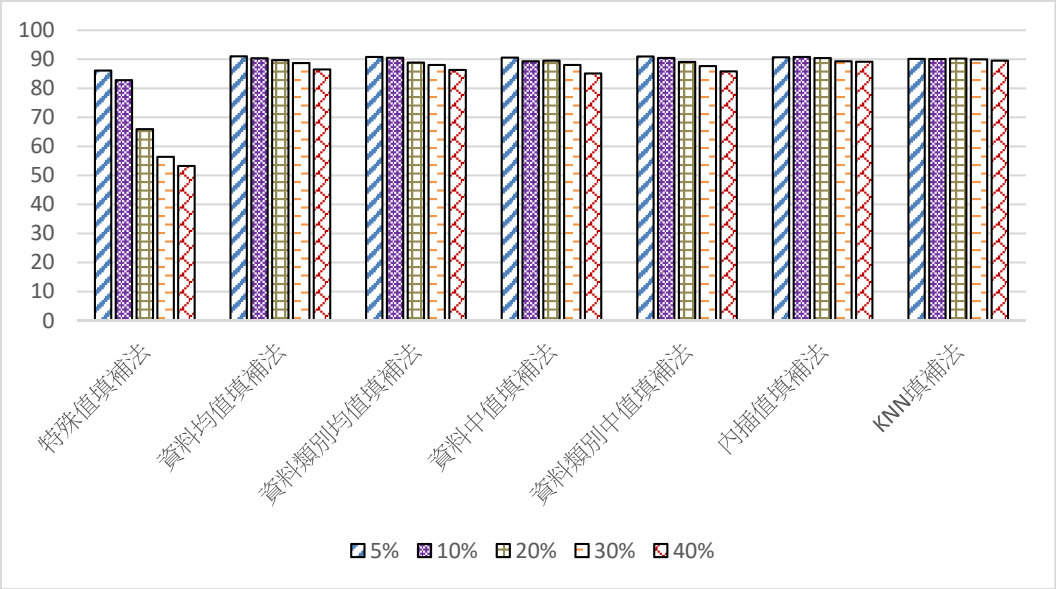

圖 4.2 在不同缺失值比例下各填補法填補 landsat satellite 訓練資料集訓練 DNN 所得之準確率

圖 4.3 為 heart 資料集使用七種填補法填補缺失資料後，分別訓練 DNN 後所得到的準確率，此資料集筆數較少，在訓練資料有缺失情況下，與 landsat satellite 資料集類似，使用不同方法填補不同比例缺失的訓練資料，對準確率的影響不大，基本都維持在 81%到百分之 83%之間，較無法看出不同方法間的差別。

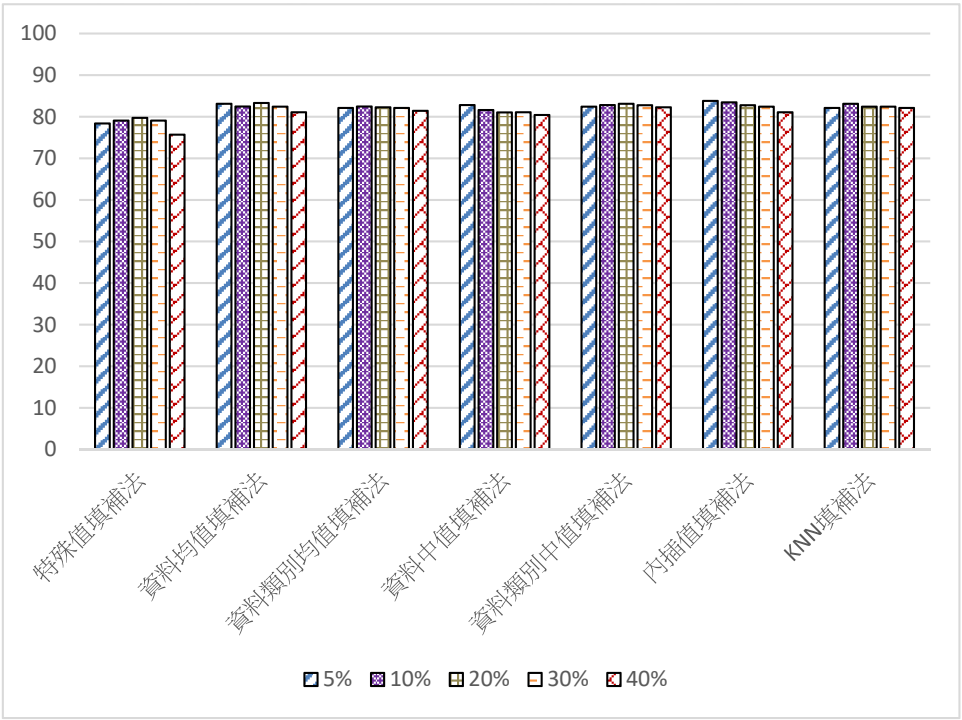

圖 4.3 在不同缺失值比例下各填補法填補 heart 訓練資料集訓練 DNN 所得之準確率

由以上三個資料集實驗結果顯示 KNN 填補法相較其他方法有較佳的效果，因為 KNN 填補法考慮了資料之間的關係。較不受缺失值影響的 landsat satellite 資料集與 heart 資料集，使用不同方法填補缺失值都可以維持一定準確率。使用統計方法如資料均值與中值填補法填補缺失值，準確率也維持在百分之九十。

表 4.4 為 abalone 訓練資料集在不計入特殊值填補法時，KNN 填補法對其他五種方法的準確率差異，表中數字為 KNN 填補法對 DNN 所提升的準確率。從表 4.4 可以看出 abalone 訓練資料集在 5%、10%、20%、30%和 40%的缺失值時，KNN 填補法使 DNN 的準確率平均分別提高了 3.58、6.59、7.29、5.54、5.21 個百分點。landsat satellite 資料集與 heart 資料集使用各填補法對 DNN 的準確率差異不大，因此不特別列出。

表 4.4 abalone 訓練資料集在不同缺失值比例下 KNN 填補法(K=7)對其他五種方法的準確率差異

	5%	10%	20%	30%	40%
資料均值填補法	3.54	8.37	8.8	6.03	6.32
資料類別均值填補法	3.73	7.13	8.52	7.28	6.12
資料中值填補法	3.62	7.08	8.23	5.27	7.56
資料類別中值填補法	5.36	6.51	8.04	7.28	5.84
內插值填補法	1.63	3.88	2.87	1.82	0.19
平均	3.58	6.59	7.29	5.54	5.21

4.3 實驗二：測試資料有缺失值

4.3.1 實驗設計與實作

本實驗針對測試資料集有缺失值時，在不同的資料缺失值比例下使用五種填補法填補訓練資料集，以之訓練 DNN，並使用簡化特徵模型與填補法比較，然後輸入測試資料集觀察其分類準確率。

實驗一使用七種方法做比較，由於資料類別均值填補法與資料類別中值填補法是適用於訓練資料有缺失值的情況，所以本實驗沒有使用這兩種方法，而是只使

用其他五種填補法。本實驗另外加入了簡化特徵模型，此方法適於測試資料集有缺失值時使用。

本實驗假設測試資料集有缺失值，所以我們以隨機函數挑選測試資料集的資料刪除其特徵值(使該筆資料出現缺失值)。由於是隨機挑選資料產生缺失值，所以符合資料為完全隨機缺失類型。我們使用隨機函數 rand()控制資料缺失值出現的比例，分別為 5%、10%、20%、30%和 40%。

4.3.2 實驗結果與分析

表 4.5、4.6、4.7 分別為針對測試資料集在不同資料缺失值比例下，使用五種補值法補足測試資料與未填補缺失值所測得 DNN 的分類準確率。

由表 4.5、4.6、4.7 中可以看出本實驗與實驗一相同，當缺失值以特殊值取代時，DNN 的準確率皆非常低，因為 DNN 不知該特殊值為特例，只是將它視為一般特徵值。而測試資料未填補缺失值與訓練資料未填補缺失值不同，訓練資料如果有缺失值未填補，DNN 中節點的權重值無法更新，但測試資料有缺失時，DNN 可以做分類，但分類的準確率皆非常低

表 4.5 abalone 測試資料集在不同缺失值比例下五種填補法與未填補缺失值對 DNN 的準確率

	5%	10%	20%	30%	40%
未填補缺失值	69.6	59.04	47.08	40.19	38.09
特殊值填補法	60	52.15	47.27	41.05	37.7
資料均值填補法	80.29	73.68	63.25	52.34	47.66
資料中值填補法	80.48	71.48	62.01	51.96	45.36
內插值填補法	82.2	74.74	67.46	51.1	47.85
KNN 填補法 (K=7)	88.33	84.5	73.59	64.02	55.89

表 4.6 landsat satellite 測試資料集在不同缺失值比例下五種填補法與未填補缺失值對 DNN 的準確率

	5%	10%	20%	30%	40%
未填補缺失值	23.25	11.39	9.4	9.36	9.36
特殊值填補法	60.6	50.85	38.5	34.05	33.05
資料均值填補法	88.55	84.9	78.95	76.25	74.25
資料中值填補法	88.45	85.35	79.45	75.75	73.9
內插值填補法	89.85	88.5	81.55	78.75	74.3
KNN 填補法 (K=7)	90.6	90.5	89.10	89.85	89.4

表 4.7 heart 測試資料集在不同缺失值比例下五種填補法與未填補缺失值對 DNN 的準確率

	5%	10%	20%	30%	40%
未填補缺失值	64.19	53.38	49.32	44.59	45.27
特殊值填補法	76.35	79.73	72.97	70.95	70.27
資料均值填補法	83.78	84.46	83.78	82.44	81.08
資料中值填補法	82.43	81.08	82.43	81.76	79.73
內插值填補法	83.78	84.46	81.76	79.73	79.03
KNN 填補法 (K=7)	83.11	83.79	82.43	83.45	82.44

圖 4.4 為 abalone 資料集使用五種填補法填補測試資料後，所測得 DNN 的準確率。在此實驗中，KNN 填補法相較於其他方法可以維持較高的準確率，雖然 abalone 測試資料集隨著缺失值比例的增加，呈現了準確率急劇下降的情形，但是在測試資料缺失值比例為 5%與 10%時，使用 KNN 填補法的準確率維持在百 80%以上，比其他補值方法高了約十個百分點。

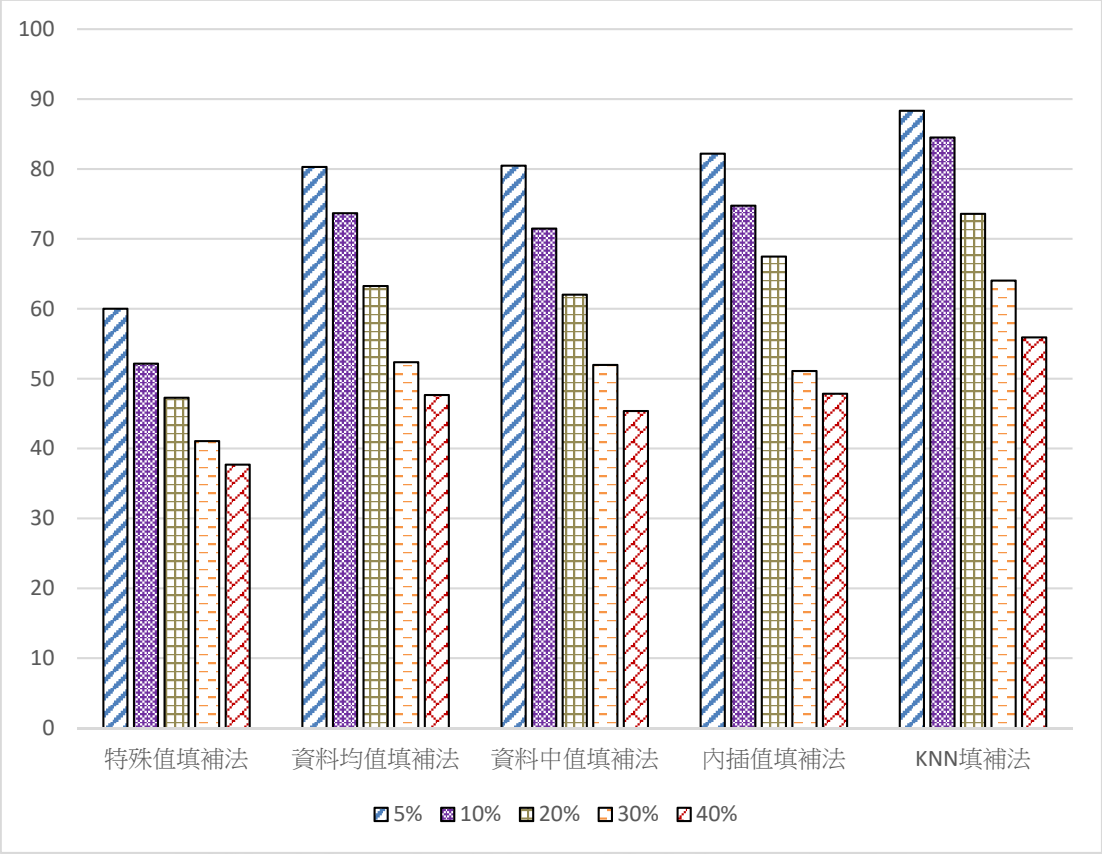

圖 4.4 在不同缺失值比例下各填補法填補 abalone 測試資料集對 DNN 之準確率

圖 4.5 為 landsat satellite 測試資料集使用五種填補法填補後所測得 DNN 的準確率。由圖 4.5 可以看出 landsat satellite 測試資料集有缺失值時，除了使用 KNN 填補法以外，其他的填補方法在缺失值比例增加時，準確率會隨之下降，但 KNN 填補法的準確率仍可維持在 90%。在 landsat satellite 測試資料集缺失值比例為 5% 時，使用不同缺失值處理方法對分類準確率造成的差異不明顯，因為在缺失值比例較低時，資料分類的準確率受到缺失值的影響較小。由此資料集的實驗結果可以看出 KNN 填補法效果較其他為佳。

圖 4.6 為 heart 資料集使用五種填補法填補測試資料後，所測得 DNN 的準確率。heart 測試資料集有缺失值時，在不同缺失值比例下，各填補法的準確率差異不大。但在缺失值比例較高時，KNN 會稍微優於其他方法。由實驗一與實驗二可以看出此資料集在不同缺失值比例下，使用不同填補方法對準確率造成的差異不大。

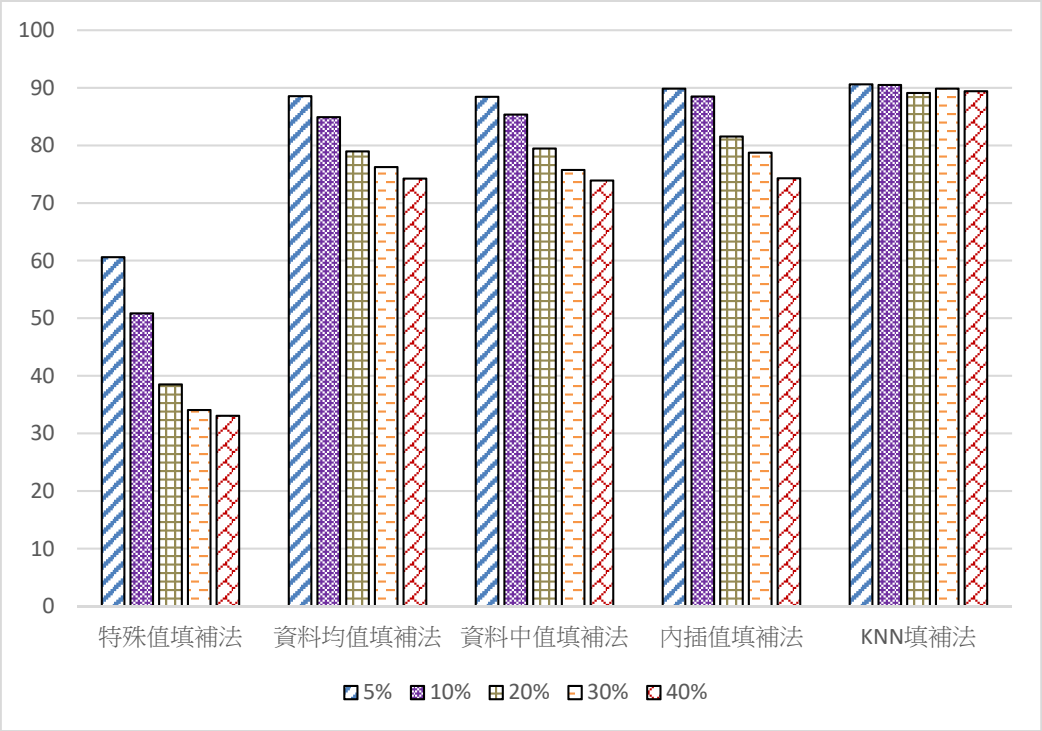

圖 4.5 在不同缺失值比例下各填補法填補 landsat satellite 測試資料集對 DNN 之準確率

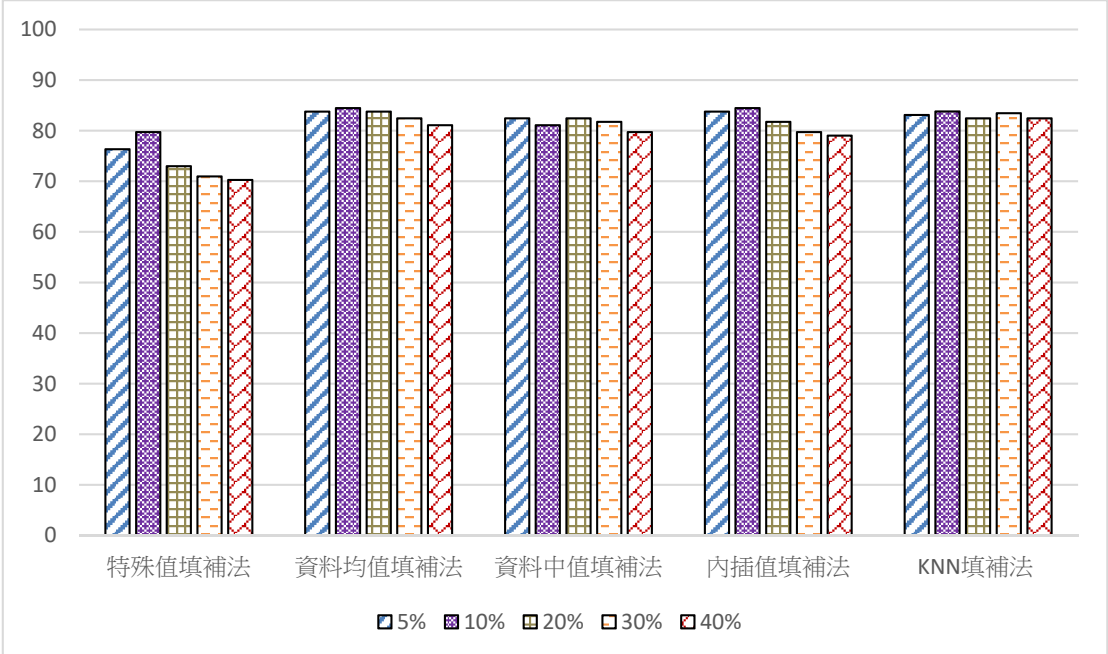

圖 4.6 在不同缺失值比例下各填補法填補 heart 測試資料集對 DNN 之準確率

表 4.8 為 abalone 測試資料集在不同缺失特徵數與缺失值比例下，KNN 填補法與簡化特徵對 DNN 的準確率。圖 4.7 為 abalone 資料集在不同缺失特徵數與缺失值比例下，KNN 填補法與簡化特徵模型分別訓練 DNN 所得之準確率，由於 abalone 資料集在缺失值比例增加時，準確率大幅下降，故選此資料集來做比較。

在此實驗中我們將 abalone 測試資料集分成 1 個、2 個、3 個、5 個和 7 個特徵有缺失值並將不同數目的特徵有缺失值時再分成不同的缺失值比例，分別觀察 KNN 填補法與簡化特徵法對 DNN 的準確率。因實驗一與實驗二結果顯示 KNN 填補法的準確率較高，故只選定 KNN 填補法與簡化特徵模型做比較。

有別於補值方法，簡化特徵模型此方法針對測試資料有缺失值時，直接將缺失的特徵(欄位)刪除。例如有一資料集含有八個特徵，當訓練資料集完整無缺失而測試資料集中有兩個特徵缺失時，使用簡化特徵模型會刪除測試資料集的兩個缺失特徵，並刪除訓練資料集相對應的兩個特徵，然後再利用訓練資料集剩餘的六個特徵來訓練 DNN。由圖 4.7 可以看出資料缺失值比例低於 20%時，不論測試資料缺失幾個特徵，KNN 填補法對 DNN 的準確率都較簡化特徵模型高。當資料特徵的缺失數量較多時，簡化特徵模型在不同缺失值比例的準確率會比 KNN 填補法低；當資料特徵的缺失數量較少時，簡化特徵模型在不同缺失值比例的準確率與 KNN 填補法差異不大，除了在缺失值比例為 40%時，當特徵的缺失數量為 2 和 3，簡化特徵模型的準確率會比 KNN 填補法高 5%，在其他情況下準確率都較低。

表 4.8 abalone 測試資料集在不同缺失特徵數與缺失值比例下 KNN 填補法(K=7)與簡化特徵對 DNN 的準確率

	1 特徵缺失	2 特徵缺失	3 特徵缺失	5 特徵缺失	7 特徵缺失
KNN 填補法資料缺失 10%	93.59	91.77	87.85	84.88	84.98
KNN 填補法資料缺失 20%	91.1	88.33	84.21	81.15	76.46
KNN 填補法資料缺失 30%	90.96	85.55	81.34	77.32	68.23
KNN 填補法資料缺失 40%	90.05	82.87	77.61	71.1	61.82
簡化特徵模型	93.01	87.18	82.2	67.56	52.92

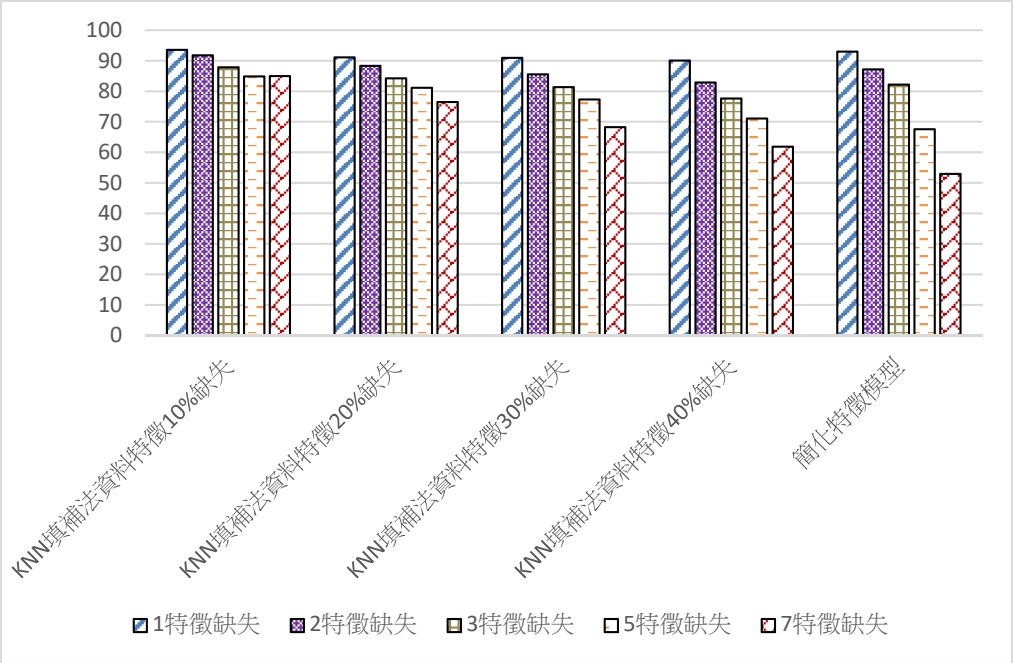

圖 4.7 在不同缺失特徵數與缺失值比例下 KNN 填補法(K=7)填補 abalone 測試資料集與簡化特徵模型訓練 DNN 所得之準確率

表 4.9 為 abalone 測試資料集在不計入特殊值填補法時，KNN 填補法對其他三種方法的準確率差異，表中數字為 KNN 填補法對 DNN 所提升的準確率。從表 4.9 可以看出，abalone 測試資料集在 5%、10%、20%、30%和 40%的缺失值時，KNN 填補法的準確率平均分別提高了 7.34、11.2、9.35、12.22、8.93 個百分比。

表 4.9 abalone 測試資料集在不同缺失值比例下 KNN 填補法(K=7)對其他三種方法的準確率差異

	5%	10%	20%	30%	40%
資料均值填補法	8.04	10.82	10.34	11.68	8.23
資料中值填補法	7.85	13.02	11.58	12.06	10.53
內插值填補法	6.13	9.76	6.13	12.92	8.04
平均	7.34	11.2	9.35	12.22	8.93

表 4.10 為 landsat satellite 測試資料集在不計入特殊值填補法時，KNN 填補法對其他三種方法的準確率差異，表中數字為 KNN 填補法對 DNN 所提升的準確率。從表 4.10 可以看出 landsat satellite 測試資料集在 5%、10%、20%、30%和

40%的缺失值時，KNN 填補法的準確率平均分別提高了 1.65、4.25、9.12、12.93、15.25 個百分比。heart 測試資料集使用各填補法對 DNN 準確率差異不大，因此不特別列出。

表 4.10 landsat satellite 測試資料集在不同缺失值比例下 KNN 填補法(K=7)對其他三種方法的準確率差異

	5%	10%	20%	30%	40%
資料均值填補法	2.05	5.6	10.15	13.6	15.15
資料中值填補法	2.15	5.15	9.65	14.1	15.5
內插值填補法	0.75	2	7.55	11.1	15.1
平均	1.65	4.25	9.12	12.93	15.25

4.4 實驗三：訓練資料及測試資料皆有缺失值

4.4.1 實驗設計與實作

本實驗為訓練資料集與測試資料集皆有缺失值時，在不同的資料缺失值比例下使用五種填補法填補訓練與測試資料集，以之訓練 DNN，然後輸入測試資料集觀察其分類準確率。

由於資料類別均值填補法與資料類別中值填補法只適用於訓練資料有缺失值的情況，簡化特徵模型只適用於測試資料有缺失值的情況，而本實驗為訓練資料與缺失資料皆有缺失，所以本實驗未納入這三種方法。此實驗沒列出未填補缺失值之準確率，因為訓練資料有缺失值時，無法更新權重，準確率不會有變化，而此處的訓練資料有缺失值。

本實驗假設訓練資料集與測試資料集皆有缺失值，使用的資料為實驗一與實驗二中的缺失資料集，當缺失值比例為 5%時，表示使用實驗一中有 5%缺失值的訓練資料集與實驗二中有 5%缺失值的測試資料集。

4.4.2 實驗結果與分析

表 4.11、4.12、4.13 分別為訓練資料集與測試資料集在不同資料缺失值比例下，使用五種補值法補足訓練資料與測試資料後所測得 DNN 的分類準確率。

由表 4.11、4.12、4.13 中可以看出，與實驗一和實驗二相同，當缺失值以特殊值取代時，DNN 的準確率皆非常低，因為 DNN 不知該特殊值為特例，只是將它視為一般特徵值。

表 4.11 abalone 訓練與測試資料集在不同缺失值比例下五種填補法對 DNN 的準確率

	5%	10%	20%	30%	40%
特殊值填補法	64.55	57.13	53.49	51.96	47.08
資料均值填補法	75.98	66.6	56.56	53.88	50.14
資料中值填補法	74.55	65.84	56.08	54.55	49.95
內插值填補法	74.69	69.38	56.75	52.34	48.9
KNN 填補法 (K=7)	82.49	78.09	64.21	61.05	54.07

表 4.12 landsat satellite 訓練與測試資料集在不同缺失值比例下五種填補法對 DNN 的準確率

	5%	10%	20%	30%	40%
特殊值填補法	82.3	79.55	75.35	74.25	70.8
資料均值填補法	89.75	87.75	85.6	84.35	84.25
資料中值填補法	88.7	87.7	85.9	84.6	84.05
內插值填補法	90.2	87.77	85.05	80.28	78.7
KNN 填補法 (K=7)	89.7	89.65	90.15	89.85	86.85

表 4.13 heart 訓練與測試資料集在不同缺失值比例下五種填補法對 DNN 的準確率

	5%	10%	20%	30%	40%
特殊值填補法	78.28	79.73	76.35	69.59	70.95
資料均值填補法	80.75	81.42	82.43	78.38	79.05
資料中值填補法	81.09	82.77	81.53	78.05	77.7
內插值填補法	84.8	82.43	80.75	77.73	76.35
KNN 填補法 (K=7)	83.78	81.42	81.31	80.41	79.73

圖 4.8 為 abalone 資料集使用五種填補法填補缺失資料後，分別訓練 DNN 後所得到的準確率。abalone 資料集在訓練資料集及測試資料集皆有缺失值時與實驗一和實驗二相同，DNN 分類的準確率會隨著缺失值比例增加導致準確率下降。

abalone 資料集在此實驗中，使用 KNN 填補法的準確率也是最高的，而其他填補法之間的準確率差異不大。

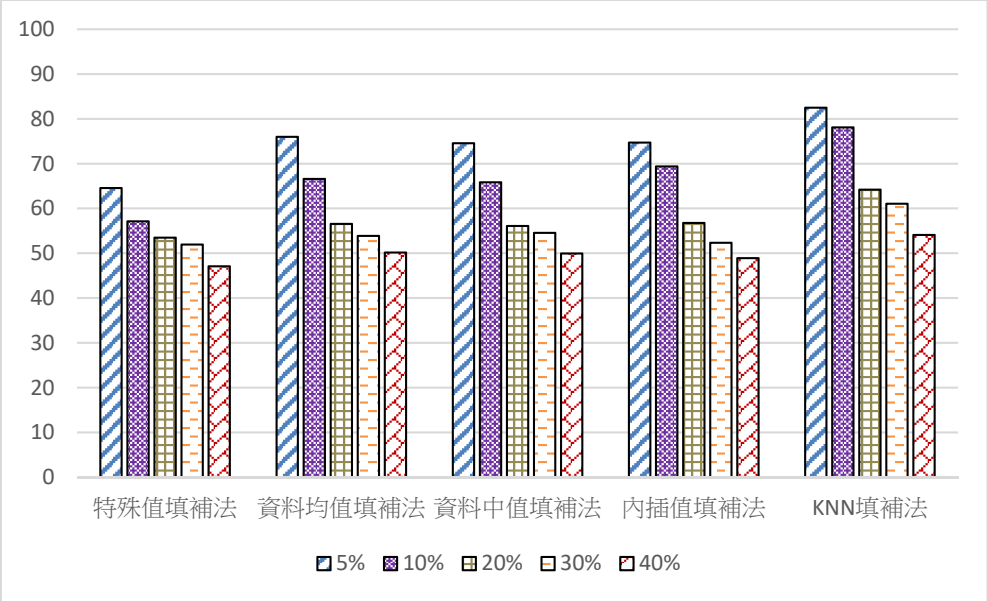

圖 4.8 不同缺失值比例下各填補法填補 abalone 訓練與測試資料集對 DNN 所得之準確率

圖 4.9 為 landsat satellite 資料集使用五種填補法填補缺失資料後，分別訓練 DNN 後所得到的準確率。由圖 4.9 可以看出 landsat satellite 訓練資料集與測試資料集皆有缺失值時，準確率皆維持在 80%以上，然而使用內插值填補法，在缺失值比例較大於 30%時，準確率不到 80%。此資料集使用各填補法的準確率會隨著缺失值比例上升隨之下降，只有使用 KNN 填補法能維持較高的準確率。

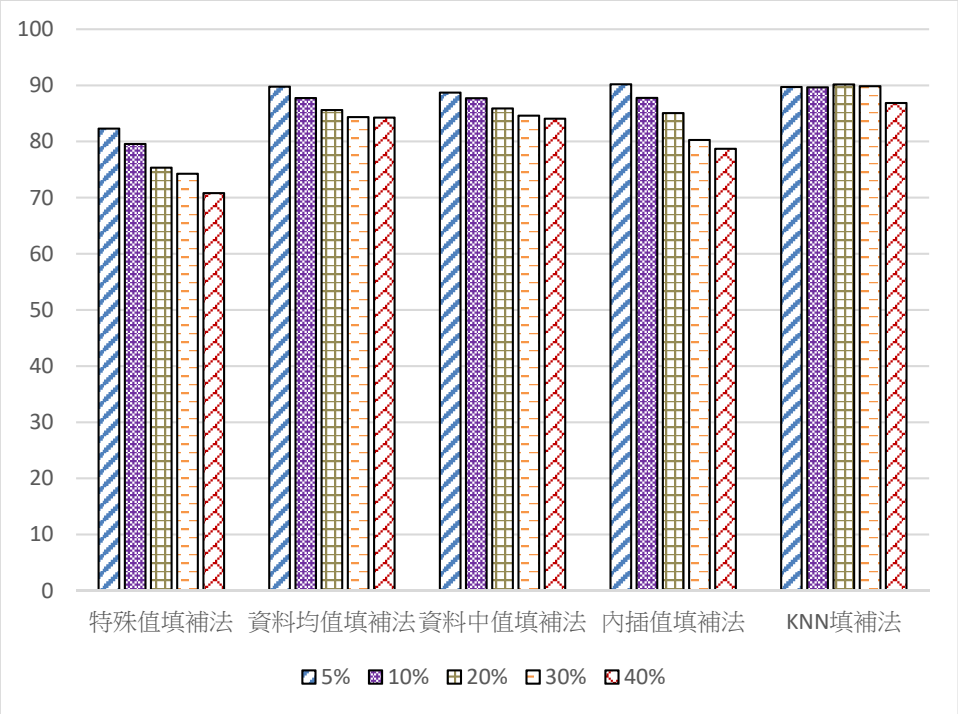

圖 4.9 在不同缺失值比例下各填補法填補 landsat satellite 訓練與測試資料集對 DNN 所得之準確率

圖 4.10 為 heart 資料集使用五種填補法填補訓練與測試資料後，所測得 DNN 的準確率。此資料集在本實驗與實驗一和實驗二相同，使用五種填補法的準確率沒有太大的差異。在缺失值比例小於 20%時，使用內插值填補法的準確率最高，但與 KNN 填補法差距不到 1 個百分點，與資料平均值與中值填補法相差不到 3 個百分點；在缺失值比例大於 30%時，KNN 填補法的準確率比其他方法高了 1 到 3 個百分點，各方法之間差異不大。

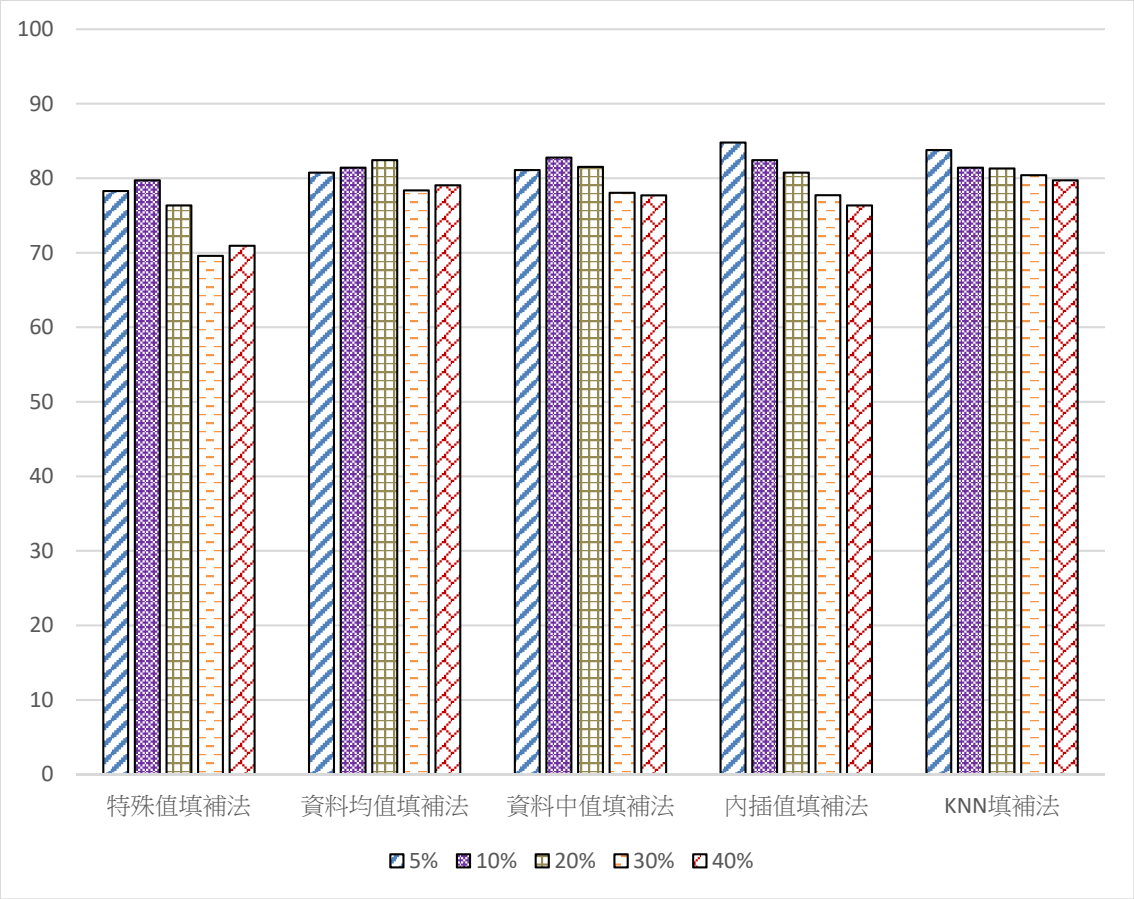

圖 4.10 在不同缺失值比例下各填補法填補 heart 訓練與測試資料集對 DNN 所得之準確率

表 4.14 為 abalone 訓練與測試資料集在不計入特殊值填補法時，KNN 填補法對其他三種方法的準確率差異，表中數字為 KNN 填補法對 DNN 所提升的準確率。從表 4.14 可以看出，abalone 訓練與測試資料集在 5%、10%、20%、30%和 40%的缺失值時，KNN 填補法的準確率平均分別提高了 7.42、10.82、7.75、7.46、4.41 個百分點。

表 4.14 abalone 訓練與測試資料集在不同缺失值比例下 KNN 填補法(K=7)對其他三種方法的準確率差異

	5%	10%	20%	30%	40%
資料均值填補法	6.51	11.49	7.65	7.17	3.93
資料中值填補法	7.94	12.25	8.13	6.5	4.12
內插值填補法	7.8	8.71	7.46	8.71	5.17
平均	7.42	10.82	7.75	7.46	4.41

表 4. 15 為 landsat satellite 訓練與測試資料集在不計入特殊值填補法時，KNN 填補法對其他三種方法的準確率差異，表中數字為 KNN 填補法對 DNN 所提升的準確率。從表 4.15 可以看出 landsat satellite 訓練與測試資料集在 5%、10%、20%、30%和 40%的缺失值時，KNN 填補法的準確率平均分別提高了 0.15、1.91、4.63、6.77、4.52 個百分比。heart 訓練與測試資料集使用各填補法對 DNN 準確率差異不大，因此不特別列出。

表 4.15 landsat satellite 訓練與測試資料集在不同缺失值比例下 KNN 填補法(K=7)對其他三種方法的準確率差異

	5%	10%	20%	30%	40%
資料均值填補法	-0.05	1.9	4.55	5.5	2.6
資料中值填補法	1	1.95	4.25	5.25	2.8
內插值填補法	-0.5	1.88	5.1	9.57	8.15
平均	0.15	1.91	4.63	6.77	4.52

4.5 實驗四：改變深度神經網路架構對準確率的影響

4.5.1 實驗設計與實作

本實驗的目的為觀察若改變 DNN 架構對有缺失值資料的分類準確率影響情形。類神經網路架構越複雜，層數與節點數越多，類神經網路能學習到的權重值就越多，對於準確率的提昇可能會有幫助。在本實驗中我們增加一層隱藏層，並將每個隱藏層中的節點數擴增為原本的兩倍。

本實驗選用缺失值比例越大，準確率下降越多的 abalone 資料集與 landsat satellite 資料集。此處使用的 abalone 資料集與 landsat satellite 資料集為實驗三中使用的缺失資料集，並使用 KNN 填補法填補缺失值。

4.5.2 實驗結果與分析

表 4.16、表 4.17、圖 4.11 與圖 4.12 為 abalone 資料集與 landsat satellite 資料集改變網路架構後的實驗結果。從圖 4.11 與圖 4.12 可以看出，改變類神經網路架構後整體準確率變化不大，準確率相差在兩個百分點之內；在缺失值比例較大時，準確率反而有所下降。由此實驗可以看出，使用有缺失值的資料作分類時，透過改變類神經網路架構來提升準確率的效果不佳，因此在前處理時應選擇更佳的缺失值處理方法以提升分類準確率。

表 4.16 改變網路架構的準確率比較(abalone 資料集)

	5%	10%	20%	30%	40%
原本架構	82.49	78.09	64.21	61.05	54.07
增加一隱藏層並將節點數增為 2 倍	84.5	78.95	66.41	58.66	53.59

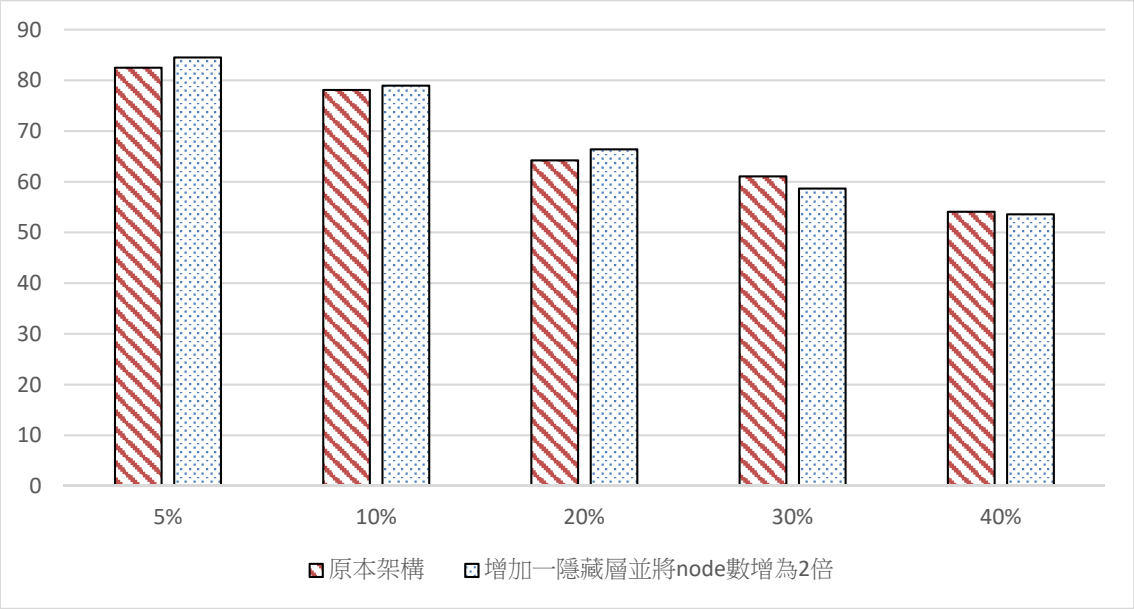

圖 4.11 改變網路架構的準確率比較(abalone 資料集)

表 4.17 改變網路架構的準確率比較(landsat satellite 資料集)

	5%	*10%*	20%	30%	40%
原本架構	89.7	89.65	90.15	89.85	86.85
增加一隱藏層並將節點數增為 2 倍	90	89.85	89.04	88.15	88.3

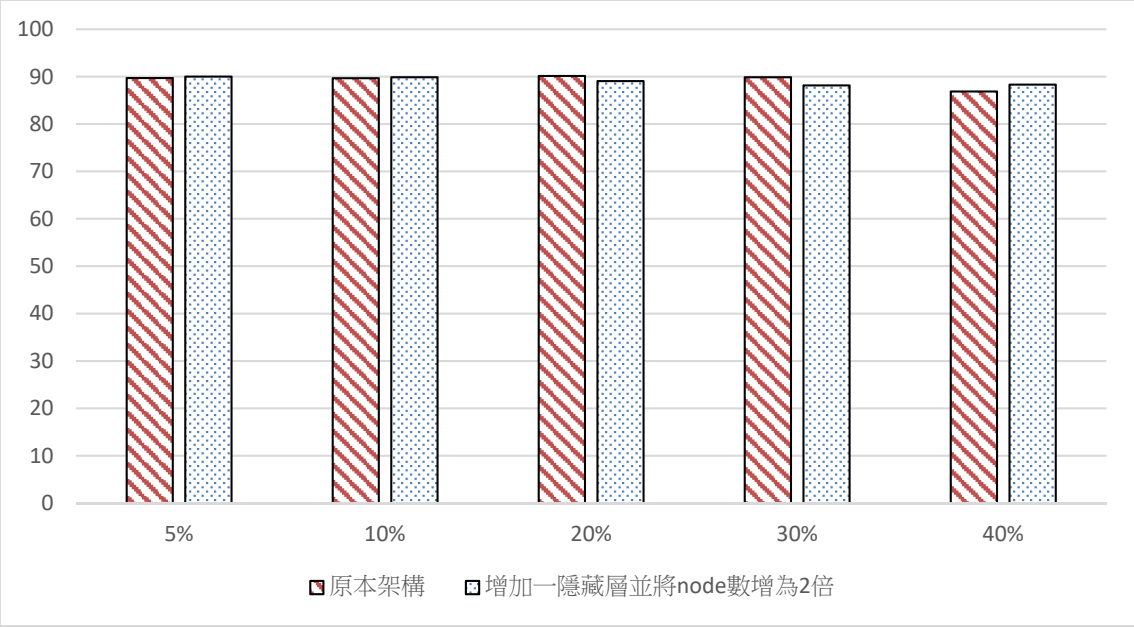

圖 4.12 改變網路架構的準確率比較(landsat satellite 資料集)

4.6 實驗小結

由以上 4.2 節至 4.4 節的三個實驗證明，當缺失值以特殊值取代時，因為 DNN 不知該特殊值為特例，只是將它視為一般特徵值，所以準確率皆非常低。而訓練資料有缺失值時，DNN 無法更新節點的權重值；測試資料有缺失值時，準確率非常低，因此當資料有缺失值必須做適當的處理。

實驗二的結果與實驗一的結果相同，考慮到資料關係的各填補法中，KNN 填補法整體準確率較高。原本在實驗一中 landsat satellite 資料集的分類準確率不受缺失值影響，但在實驗二測試資料有缺失值時，資料缺失值比例越大，不同填補方法會造成分類準確率下降，因此不管在訓練階段或測試階段資料有缺失值時，都應該對缺失值做適當的處理。

實驗三中 landsat satellite 資料集使用內插值填補法，在缺失值比例大時準確率最低；但在實驗一中，abalone 資料集使用內插值填補法的準確率比資料均值與中值填補法高。由此可知，同樣的方法在資料集不同或缺失值比例不同時，填補完的資料集在分類時的準確率會有所不同。

我們發現三個不同資料集在不同缺失值比例下受到缺失值的影響情況不盡相同。abalone 資料集隨著缺失值比例越大，DNN 分類的準確率大幅下降；landsat satellite 資料集與 heart 資料集在不同的缺失值比例下，DNN 分類的準確率差異較小，造成此情形的原因可能跟資料集的特徵數有關。abalone 資料集的特徵數總共有 8 個，而 landsat satellite 資料集與 heart 資料集分別有 36 和 13 個特徵。當資料

發生缺失時，特徵數越多，DNN 可以從資料中獲取的資訊越多，因此當資料有缺失值時，特徵數較多的資料在缺失值比例較高時準確率不會大幅下降。

由實驗一可得知在訓練資料集有缺失值時，資料缺失值比例在 5%到 40%之間，使用 KNN 填補法的準確率平均提高了 5.64%。在實驗二測試資料集有缺失值時，資料缺失值比例在 5%到 40%之間，使用 KNN 填補法的準確率平均提高了 9.2%。在實驗三訓練資料集與測試資料集兩者皆有缺失值時，資料缺失值比例在 5%到 40%之間，使用 KNN 填補法的準確率平均提高了 5.59%。總而言之，資料缺失值比例在 5%到 40%之間，使用 KNN 填補法在三個實驗中的準確率平均提高了 6.81%。

由以上三個實驗可以看出，KNN 填補法在實驗二測試資料集有缺失值時，準確率提高較多。在實驗二中的訓練資料集是完整無缺失的，訓練出的 DNN 模型較偏向真實情況，當測試資料的填補不考慮資料間的關係，填補完的資料可能會導致與真實情況不符，因此在實驗二中，KNN 填補法的準確率會提升較多。

第五章 結論與未來方向

5.1 結論

本研究針對不同缺失值處理方法分析其對深度神經網路在分類的準確率變化。由實驗結果得知缺失值會造成深度神經網路分類的準確率下降，深度神經網路與傳統分類器相比其演化能力較好，但資料有缺失值時，一樣會影響準確率。隨著資料集的不同，受到缺失值影響的程度也不同，在訓練資料集與測試資料集有缺失值時，準確率的高低也會有所不同。

我們在處理缺失值時，可以發現使用統計的方法如資料均值與中值填補法，會造成較低的分類準確率，因為統計的方法單純取未缺失特徵值的平均或中值，而忽略了每筆資料間的關係，即使類神經網路可以透過學習有較好的表現，但準確率還是較低的。使用 KNN 填補法得到的效果最佳，資料缺失值比例在 5%到 40%之間，使用 KNN 填補法在三個實驗中的準確率平均提高了 6.81%。

簡化特徵模型較不適合使用在深度神經網路的分類，因深度神經網路透過特徵調整每個節點的權重值，但使用簡化特徵模型會減少特徵，當特徵數量減少，調整權重值的效果會變差，所以當缺失特徵數較多時，準確率會大幅下降。當資料特徵的缺失數量較多時，簡化特徵模型在不同缺失值比例的準確率會比 KNN 填補法低；當資料特徵的缺失數量較少時，簡化特徵模型在不同缺失值比例的準確率與 KNN 填補法差異不大。

改變深度神經網路架構對缺失資料所造成分類準確率的提升也有限，準確率平均差距只有兩個百分點，且會有過適的可能，因此訓練出的分類模型較為不理想。欲透過改變模型架構使深度神經網路分類準確率提高的可行性較低，有缺失值的資料集在前處理階段使用較合適的方法則可以提高分類的準確率。

5.2 未來研究方向

未來研究將探討不同的缺失發生機制，由於缺失值的發生機制會影響缺失值的處理方式，在分類上可能會影響準確率。另外可以設計針對深度神經網路的缺失值處理方法或是建造一個可以輸入缺失值的類神經網路，這些都是未來可以研究的方向。

參考文獻

[1] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and Its Effect on Classifier Accuracy," *Classification, Clustering and Data Mining Applications*, Springer-Verlag Berlin Heidelberg, pp. 639-648, 2004.

[2] P. Allison, *Missing Data*, Sage, 2001

[3] G. Batista and C. Monard, "A Study of K-Nearest Neighbour as a Model-Based Method to Treat Missing Data," *Proc. of the Argentine Symposium on Artificial Intelligence (ASAI'03),* vol. 30, pp. 1-9, 2001.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via The EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[5] A. Eldem, H. Eldem, and D. ÜSTÜN, "A Model of Deep Neural Network for Iris Classification with Different Activation Functions," *Proc. of International Conference on Artificial Intelligence and Data Processing (IDAP)*, pp. 1-4, 2018.

[6] D. Joenssen and U. Bankhofer, "Hot Deck Methods for Imputing Missing Data," *Proc. of 8th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, vol. 7376, pp. 63-75, 2012.

[7] R. Little and D. Rubin, *Statistical Analysis with Missing Dat*a, John Wiley & Sons, 2002.

[8] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, 1997.

[9] H. Mohsen, E. El-Dahshan, E. El-Horbaty, and A. Salem, "Classification Using Deep Learning Neural Networks for Brain Tumors," *Future Computing and Informatics Journal*, vol. 3, pp. 68-71, 2018.

[10] D. Mundfrom and A. Whitcomb, "Imputing Missing Values: The Effect on The Accuracy of Classification," *Multiple Linear Regression Viewpoints*, vol. 25, pp. 13-19, 1998.

[11] M. Noor, A. Yahaya, N. Ramli, and A. Mustafa, "Filling Missing Data Using Interpolation Methods: Study on The Effect of Fitting Distribution," *Key Engineering Materials,* vols. 594-595, pp. 889-895, 2014.

[12] J. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[13] A. Ragel and B. Crémilleux, "Treatment of Missing Values for Association Rules," *Proc. of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining*, pp. 258-270, 1998.

[14] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, 1987.

[15] M. Saar-Tsechansky and F. Provost, "Handling Missing Values When Applying Classification Models," *The Journal of Machine Learning Research Archive*, vol. 8, pp. 1623-1657, 2007.

[16] J. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman and Hall, 1997.

[17] J. Seetha and S. Raja, "Brain Tumor Classification Using Convolutional Neural Networks," *Biomedical and Pharmacology Journal*, pp. 1457-1461, 2018.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[19] https://zh.wikipedia.org/wiki/%E6%8F%92%E5%80%BC

[20] https://archive.ics.uci.edu/ml/datasets/abalone

[21] https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29

[22] https://archive.ics.uci.edu/ml/datasets/heart+Disease