Journal of Science and Engineering Technology, Vol. 17, No. 1, pp. 31-40 (2021)

運用多模型對缺失值預測之研究

邱紹豐* 馮晉易 徐尉庭
大葉大學 資訊工程學系
51591 彰化縣大村鄉學府路 168 號
*schiou@mail.dyu.edu.tw

摘　要

在大數據分析的過程中，資料的完整性與一致性往往是影響分析結果正確性的很重要因素。因此在分析的程序開始之前，要對所收集的資料來源進行資料清理的工作，以確保後續分析不會因為資料的異常而造成結果的錯誤，因此在資料清理中維持資料的完整性是一項相當重要的工作。造成資料不完整的原因之一是所收集的資料中含有缺失值，而缺失值的出現源自於資料收集過程中人為疏失、儀器故障等因素。目前對於處理缺失值的常見方式為以下幾種：將有缺失值的值組直接忽略、或是使用缺失值屬性的集中趨勢量測（如均值、中位數等方式）進行缺失值的填補。這些方法可能會造成將該值組的原有特徵性的流失，對於後續的資料分析、應用的產出造成影響，而導致結果的不正確。針對此問題，本研究針對單一欄位缺失值使用機器學習方法來進行填補。我們以不包含缺失值的資料作為訓練資料，以 K-Means 分群方式將資料分為多個群集以捕捉資料之間不易見的關聯，每個群集再以多重迴歸以及類神經網路建立預測模型。對需要預測的缺失值首先以 KNN 演算法求得該資料所屬的群集，再套用該群集的模型來計算預測值。在實驗中證明本研究所提出的多模型填補的方式，在以均方根誤差來統計精準度的結果中，均優於現有的填補演算法。

關鍵詞：缺失值，多重迴歸分析，類神經網路，k-平均分群演算法

Predicting Missing Values Using Multiple Models

SHAO-FONG CHIOU*, CHING-YANG FENG and WEI-TING HSU
Department of Computer Science and Information Engineering, Da-Yeh University
No. 168, University Rd., Dacun, Changhua, 51591, Taiwan, R.O.C.
**schiou@mail.dyu.edu.tw*

The integrity and consistency of data substantially influence the results of big data analytics. Data cleansing is often performed prior to the start of analyses to maintain these qualities in input data and ensure the results are not distorted by data anomalies. A key goal of data cleansing is to preserve data integrity. Missing values in the collected data are the main factor undermining such integrity and often result from human negligence or machine malfunction during data collection. Methods for addressing this problem include ignoring data that contain missing values or substituting the missing values with measures of central tendency, such as means or medians. These methods may result in incorrect predictions of missing values because of an inability to detect relationships among the input data. As a

result, outcomes of subsequent analyses may also be incorrect. In this study, we used machine learning techniques to manage data containing missing values for a single attribute. We used a data set without missing values as the training data and clustered it using the k-means algorithm. Prediction models were built for each cluster using the resulting data. The k-nearest neighbor algorithm was used to determine the clusters of data, and models of the clusters were used to compute the missing values. We compared the results of the root-mean-square error of our models with that of other models commonly used in simulations, and the results revealed that our models were more accurate.

Key Words**:** Missing Value, Multiple Regression, Artificial Neural Network, K-Means Clustering

一、前言

大數據分析的技術在資訊量成長迅速的現代，有著相當顯著的進步。這個新的技術藉由分析過往資料所建立的模型，用以預測未來的行為與事件。在軟硬體效能不斷的提升之下，這項技術已經應用在相當多的領域，如銷售額預測、顧客行為模式預測等。這些用以建立模型的資料來自相當多元的管道，但是常會因為資料蒐集的過程中設備或是人為的因素而產生資料異常，進而造成模型準確度的降低。因此資料清理（Data Cleansing）是大數據分析技術中很重要的前期處理，以確保後續所建立模型的正確性。一般蒐集的資料以值組（tuple）的形式呈現，其中包含了多個值已表示該筆資料的各種屬性（attribute）。例如，學生的資料包含了如「學號」、「姓名」、「地址」等屬性。資料清理主要的目的在於維持資料的一致性（Consistency）以及完整性（Integrity）。維持一致性的基本工作是統一資料中屬性值的格式，如性別「男性」統一為「M」值等，或是確保所有蒐集的資料有相同的值組架構。完整性的維護工作主要確保資料中各個屬性均有值並處理所蒐集資料中有缺失的部分。

資料中有缺失的部分稱為缺失值（Missing Value），發生缺失值常見原因如下：人工輸入資料時所造成的疏失、收集問卷時填寫者因某些原因刻意跳過題目以及收集資料的機器發生問題等。Little 等學者的研究指出，造成缺失值可以分為以下三種類型 [7]：

1. 完全隨機缺失（Missing Completely At Random, MCAR），該存在有缺失值的資料與本身或在資料集中與其他欄位屬性的值毫無關係
2. 隨機缺失（Missing At Random, MAR），該類型之缺失值可能取決其他欄位的相關性
3. 非隨機缺失（Not Missing At Random, NMAR），該類型缺失值可能與已發現到或未發現到的資訊有關係

缺失值對於後續資料探勘以及統計分析結果的正確性有相當大的影響，例如 K Nearest Neighbor（KNN）、k-Means 等演算法因需要計算資料點特徵的距離，若資料中存在缺失值則無法進行運算，因此缺失值的處理在大數據分析的前置工作中佔有非常重要的地位。一般在前置工作中缺失值常見的處理方式如下：

1. 檢查每組值組中是否有缺失值，若有發現缺失值，則將該值組忽略不進行分析。此方法雖較為簡單，但該方法可能導造成數據偏差，忽略太多資料本身所提供的特徵性，導致影響分析結果，例如：對於 MAR 缺失類型的資料而言，刪除會導致特定群集的特徵消失，對於分析時被刪除的資料無法提供原有的特徵性。
2. 使用屬性的集中趨勢量測（measure of central tendency）（如均值、中位數等方式）進行缺失值的填補。因只使用中位數或均值的單一數值進行填補，造成某特定值數量提升，可能改變其特徵原有的分布，影響後續分析結果。
3. 使用機器學習方式（如迴歸分析、決策樹、K Nearest Neighbor 等方式）進行填補 [6, 12]，此方法是目前較為多人使用的方式。使用機器學習方式進行資料填補，可以利用現有資料的特性進行分析，進一步預測缺失欄位值。機器學習方法相較於上述兩種方式可以替資料找出特徵性並對缺失資料預測出較合理的數據，而非僅用一值填補或刪除。

上述缺失值填補方式中若採用忽略或是以單一模型來預測缺失值，未考慮資料的屬性常常是非單一模型所能描述，因此預測的結果往往會與真實數值有所差異。在考慮資料屬性之間的異質性，在本研究中提出了一個框架，採用了建立多模型的方式將資料歸類於不同屬性的模型，在預測缺失值時則套用最適合的模型計算，以提高預測的準確度。在研究的實驗中，也以本論文所提出的多模型預測方式與其他常用的單一模型的預測方式，比較預測值與實際值的均方根誤差（Root Mean Squared Error, RMSE）來判斷預測的精準

度，而實驗結果顯示本研究所提出的預測方式明顯的做出較正確的預測。本論文的架構如下：第二章為在相關研究領域的文獻，介紹學者所分類的缺失值類型、目前較常使用的缺失值理方法以及本驗中所利用到的演算法。第三章為本實驗中所提出的架構及內容。第四章為本研究之架構與目前常使用之方法進行比較。第五章為結論以及本實驗限制與未來發展。

二、文獻探討

缺失值填補在資料清理的前置作業中相當的重要，若在此步驟中僅忽略存有缺失值的資料，往往會造成後續數據分析的誤差，並產生不正確的結果。因此為了提高後續資料分析與應用的準確性，通常在進行資料分析及應用前需對資料進行資料清理的步驟。針對缺失值的處理方式分為刪除方法以及填補方法， 在此領域的研究中已經有許多學者提出貢獻，相關的研究成果在本章中會一一的探討。此外，多個較常用的數據分析的技術也會在本章中介紹，由此建構本研究方法的基礎。

（一）缺失值的分類與判斷方式

缺失值在數據收集過程中是常發生的問題，而處裡缺失值的首要任務則是瞭解缺失值的類型，方可對於各類型的缺失型態進一步的處裡。Little 等學者將其分類為完全隨機缺失（MCAR）、隨機缺失（MAR）、以及非隨機缺失（NMAR）等三種類型 [1]，說明如下：

1. 完全隨機缺失（Missing Completely At Random, MCAR）：該缺失的資料與本身或在資料集中與其他變數的值毫無關係，其缺失機率性屬於隨機，並無一定關聯。在此類型之下，缺失值對於全部資料而言屬於隨機分佈的，也就是與資料各自的屬性無關。以表 1 問卷調查的結果為例，該問卷為填寫者實際年齡與薪水的調查，各種類型的缺失值以 X 表示。在 MCAR 類型中，問卷填寫者所缺漏的答案與資料中其他欄位並無關係，屬於完全隨機產生的缺失值。
2. 隨機缺失（Missing At Random, MAR）：此類型之缺失值的產生可能取決其他欄位之間的相關性。以表 1 的問卷為例，在回答薪水問題時，對於剛出社會就業的特定群族（如，年齡小於 26 歲者）而言，可能會因工作資歷不足而不願意透露目前薪資，進而導致資料缺失。
3. 非隨機缺失（Not Missing At Random, NMAR）：此類型的缺失值可能與已發現到或未發現到的資訊有關係。以表 1 的問卷為例，可以發現在於年齡越高但薪水相對較低的群族，可能會因個人因素而不願意透漏目前薪資，導致資料缺失。

表 1. 缺失值類型示意表

年齡	實際薪水/月	MCAR	MAR	NMAR
20	30000	30000	X	30000
24	28000	X	X	X
25	50000	50000	X	50000
25	23000	X	X	X
26	55000	X	55000	55000
26	27000	27000	27000	X
27	50000	50000	50000	50000
30	60000	60000	60000	60000
30	30000	X	50000	X
37	70000	70000	70000	70000

在 Schlomer 等學者所提出的論文中，判斷缺失類型的步驟如下[12]：

1. 針對 MCAR 及 MAR 的判斷方法：根據已往分析經驗評估所收集到的變量和缺失值之間的關係，由以下方式判斷：
   (1) 使用虛擬變量（dummy variable）來表示：利用虛擬變量 0 或 1 來表示欄位是否缺失。
   (2) 使用統計方法來測試此變量與其他變量之間的關係
      A. 如果虛擬變量與其他變量無相關性，則可以判定此缺失值的型態為 MCAR 或 NMAR。
      B. 如果虛擬變量與其他變量具有關聯，可以判定為 MAR 或 NMAR。
2. 使用學者 Little 的 MCAR 測試，其方法如下
   (1) 利用統計工具，如 SPSS 軟體中的 Little 測試模組進行測試。
   (2) 當 p-value 的值如果不顯著，則資料屬於 MCAR。
   (3) 該方法屬於綜合測試，將資料視為一體而非單變量。

當類型不屬於 MCAR 或 MAR，則可以將該缺失視為 NMAR，最終對於缺失的資料進行缺失類型分類。在上述的三種缺失值類型中，第二及第三類型缺失值(MAR 與 NMAR)的判別常需要領域專家的解答，較難以客觀的數據模型分析，因此在本研究中主要以解決 MCAR 類型的缺失值為主。

（二）缺失值的處理方式

有許多學者致力於探討缺失值處裡的方法，根據 Houari 等學者的研究以刪除及填補方法 [6]，最受到專家學者的重

視，其方法說明如下：

1. 刪除方法（Deletion Method）

刪除對於處裡缺失值是最常見的方式，但是刪除缺失欄位可能遺失大多數的資料，進而造成分析的正確性發生錯誤。一般常見的刪除方法有列刪除（Listwise Deletion）以及成對刪除（Pairwise Deletion）等。列刪除的方法刪除所有包含缺失值的資料列，僅保留所有欄位都有完整值的資料列。該方法對於非 MCAR 類型之缺失數據，可能會將某類型的特徵完全刪除，導致在進行數據分析時該類型之資料無法提供應有的資訊，造成分析結果的偏差[2]。以表 1 中的資料為例，若是某群族（如，26 歲以下填答者），對於問卷中之特定題目的資訊不願透漏，若使用列刪除之方法，可能會遺失該群族之特徵，無法取得年齡 26 歲以下群族的資料，可能造成某些特定特徵的資料被刪除，而無法對後續的分析提供有效的資料。

成對刪除方法又稱為可用案例分析（Available Case Analysis），並不會直接刪除整列資料，而是利用統計方式對於缺失欄位進行 t-test 或皮爾森積差（Pearson Correlation）進行相關分析。若該缺失欄位與其他欄位有存在關聯，則在分析時利用其他欄位做為代替。該方法對於非時間序列的資料而言，可以利用未缺失的資料進行分析，但對於時間序列的資料會導致嚴重的偏差 [2]。

2. 填補方法（Imputation Method）

主要是利用預測模型所預測的值或統計等方式對於缺失值進行填補。目前較為常用的方法，分別為平均值填補、回歸方法針對單一缺失值的預測填補、機器學習方法對於資料進行分析後給予填補的值以及多欄位缺失中使用的多重填補方法。平均值填補使用缺失欄位之平均值進行填補，雖較為便利但藉由平均值填補之數據可能導致與自身欄位方差減少以及減少與其他欄位之方差 [12]。Houari 等學者在研究中也指出了均值填補的方法會造成如樣本大小被高估、降低資料間的差異性進而導致屬性方差減少、以及相關性具有負偏差等缺失 [6]，因此其他使用統計的方法也漸漸被重視，以提高正確度。

回歸方法將含有缺失值的欄位視為依變數，依據自變數（非缺失值欄位）的參數調整進行缺失欄位的預測。在 Tabasi 等學者的研究中，使用了回歸模型對於能源的預測，利用回歸模型將缺失欄位視為依變數，並利用其他非缺失欄位作為自變數進行預測，最終將預測值進行缺失填補 [13]。

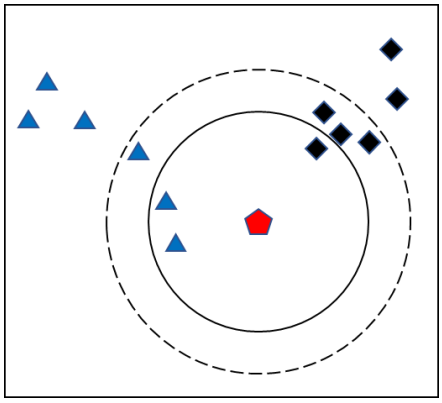

圖 1. KNN 分類示意圖

K-近鄰演算法（K-Nearest Neighbor, KNN）是由 T. Cover 等學者於 1967 年所提出，主要利用資料點映射至向量空間中，並計算其距離進行分類[4]。KNN 演算法的主要核心是根據預測點周遭 K 個最近距離的樣本標籤做為依據，若周遭 K 個點中某一類別之標籤較多，則將預測點是為該類別之分類。如圖 1 中的紅色點為預測點，在實線圓圈內（K = 3），因淺藍色標籤數量較多，則預測點會分類為淺藍色；在虛線圓圈中（K = 5），因深藍色數量較多，預測點會歸類為深藍色之類別。

Beretta 等學者對使用以 KNN 方法進行缺失值填補，並指出以 KNN 填補具備有以下特點 [3]：

1. KNN 填補之特徵為待測點周圍 K 個點所出現過的值，而非模擬所產生的值。
2. 依據缺失欄位其他欄位所提供之資訊進行預測，從而保留了原始數據的結構。
3. KNN 為非參數及不需要明確的模型來確認依變數 Y 與自變數 X 之間的關係。

Houari 等學者提出多重填補方式，由預測的分布中抽取一組 *m* 個（通常 5~10 之間）合理的替換值進行缺失值的填補方法 [6]。該方法避免了單值插補中因僅預測出一個數值即進行填補所造成之不確定性，但受限於必須滿足資料常態分佈的要求，常態分佈為常見的連續機率分布。學者 Pedersen 等中將多重填補分為三個步驟 [10]：

1. 選擇可能對於計算缺失欄位有幫助的變數，並建立多個預測的數據集，每個數據集當中是由觀察數據中給出的缺失數據分布中提取的。
2. 使用所選擇的方法，如分類和回歸樹（Classification and Regression Tree, CART）以及迴歸（Regression）等，在每

個被填補的的數據集中計算較具有較高關聯度的關聯。可以計算出係數的標準差，做為計算數據集中的關聯標準。

3. 進行所有預測集的合併，該步驟減少了不同預測集之間的標準差且考慮不同填補間的變化。

因使用重複抽樣的概念，多重填補的數據相較於單值填補的方法可以降低其不確定性。而對於多缺失欄位的資料集，多重填補可以對多欄位的缺失值進行填補。

（三）K-Means 分群演算法

K-means 的演算法於 1967 年由 James MacQueen 所提出的非監督式分群演算法[8]，後續由多位學者如 Hartigan 等提出更高效率的改良版本[5]。此方法利用使用者給予的分群數，進行反覆的迭代運算，且將各群資料點之平均視為該群群心，最終將各群群新周圍的點視為同一群集（Cluster, 或簇）。計算原理為首先由使用者選定 K 個群心，計算 *n* 個資料點至群心的距離，距離群心較近的資料點則歸類到該群集。重新計算該群心資料點的中心（群集均值）位置，並將原始群心移至新計算的群心，經過多次迭代運算後，將其群心附近點視為同一個群集。

（四）類神經網路（Artificial Neural Network, ANN）

類神經網路是由人類模擬出的神經元(Neuron)所組成。其網路架構分成輸入層（Input Layer）、隱藏層（Hidden Layer）、輸出層（Output Layer），架構如圖 2 所示。神經元經過與神經元所連接之權重（Weight）與偏差（Bias）並透過激勵函式(activation function)以決定該神經元是否啟動，經過前向傳導（Feed Foreword）方法傳遞至隱藏層，經過相同的方式運算後在傳遞至輸出層。所計算出的輸出值與實際標籤進行比較，在經過反向傳導（Backpropagation）方法進行類神經網路中權重與偏差之修正，最後建立模型 [1, 11]。

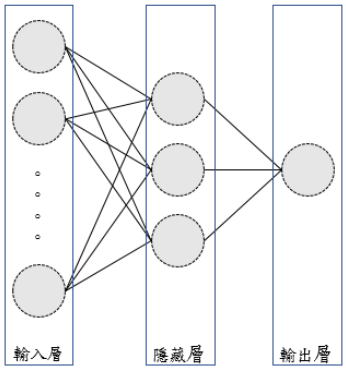

圖 2. 類神經網路示意圖

（五）生成對抗網路

生成對抗網路（Generative Adversarial Nets, GAN）是用來計算圖形或是文字中缺失值常用的方法，經過正確的訓練之後可以很正確的產生缺失的部分。但是針對時序性的資料（如具有時間戳記的資料），GAN 就無法有效的計算出缺失值。雖然有學者提出 GAN 的變形 GAN-2-Stage，整合了 GAN 以及循環神經網路（RNN）來預測在具有時序性資料中的缺失值。但是 GAN-2-Stage 有較高運算成本以及錯誤遞延並放大的問題。為解決此問題，學者 Zhang 等提出 end-to-end GAN with RF（E2GAN-RF）的模型 [15]，整合 Encoder Network 於 GAN-2-Stage 模型中，解決原方法須有一個獨立的步驟處理生成的測試資料所產生的成本。學者們使用了一些公開資料集，如 PhysioNet、KDD CUP 2018、Air Quality in Northern Taiwan 等，與其他缺失值預測演算法，如 RDA（Recurrent Denoising Autoencoder）、GAIN-2-Stage 等，比較在效能以及正確度的差異，都顯示作者所提出的方法具有相當大的優勢。

三、研究方法

在一般資料集中不容易看出資料與資料之間是否存在關聯，因此無法以單一模型來描述資料的屬性。因此本研究以建立多模型的方法，將資料集分群並為各群建立預測模型。在預測有缺失值的資料時，先判斷該筆資料群集的歸屬，並使用該群集的模型並進行預測。本研究使用公開資料來訓練模型，在實驗中藉由機器學習的方法找出資料中非缺失欄位的特性，以分群的演算法將資料分割成多個群集，並為每個群集建立預測模型。而對於待測的測試資料進行預測前，先找出非缺失欄位所歸屬的群集，並利用不同群集所建立之模型進行預測。

（一）假設條件

以下為符合本研究所設計之多模型的預測模式所做的假設：

1. 用以訓練以建立預測模型的資料，以及須預測缺失值的資料均有相同的固定結構。
2. 資料中各欄位的值均為量化數值。
3. 所要預測的缺失值在資料中為單一且固定欄位。

本研究所所使用的各種分群、預測的方法，在建立預測模型時的訓練資料必須有相同的結構，而針對非結構式資料則須經由額外處理步驟，轉換成結構式資料。需要預測缺失

值的資料需與模型訓練所使用資料的結構相同，以確保預測的正確性。第二項假設中，因為各種分群以及預測的方法均以數值為主，針對非數值資料可藉由其他映射方式，轉換為數值型態，以利後續處理。此外，本研究所提出的模型是針對資料中單一欄位的缺失值進行預測，而此限制可以藉由反覆套用模型訓練的步驟，進行多重缺失值的預測。

（二）模型建立方式

令模型訓練資料集 S 中有 n 筆資料值組（tuple）t，每筆資料均有相同的結構 t: $<a_1, a_2, \ldots, a_m>$，a_i 為資料所包含的各個屬性（欄位），且均為數值型態。屬性 a_p 為需要預測缺失值的欄位，其餘則稱為非缺失值欄位。模型建立的方式為以 K-means 分群方式將訓練資料以非缺失欄位特徵建立 K 個群集，$C = \{C_1, C_2, \ldots, C_k\}$。對每一群集 $C_i \in C$ 建立多重回歸（Multiple Regression）模型以及類神經網路（Artificial Neural Network, ANN）模型，以進行準確度的評估，模型建立的步驟如圖 3 所示。

1. 資料標準化

由於原始資料數據值的差異不一，在分析、演算法運用時需要進行資料標準化，資料標準化可以將值進行比例性的縮放。以機器學中領域中使用歐幾里得距離計算兩點之間的距離，當點中一值特別大時，對於結果會造成很大影響。對於此問題，利用資料標準化將值進行比例縮放至特定區間內，可以減少影響。

本實驗中使用最大最小標準化（Min-Max Normalization）方法進行資料的標準化，且在實驗中亦發現經過標準化的資料相對於位經過標準化的資料進行本實驗模型可以得到較低的均方根誤差（Root Mean Square Error, RMSE）。Min-Max Normalization 的方法如式（1）所示，其中 x 代表原始資料值，y 代表經過標準化的值，x_{min} 為該欄位之最小值，x_{max} 為該欄位最大值。經過 Min-Max Normalization 計算後，值的區間會落於 0 ~ 1 之間。

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

2. 資料分群

整體資料集中的資料之間是否存在關聯性往往很難以單一模型來描述，若是將資料依某些特徵分成多個群組，各個群組中資料的關聯性就會較容易被發掘。在本研究中以 K-Means 分群方法對非缺失值欄位分群，利用這些非缺失欄位值映射至空間向量中，並利用每筆資料點之特性進行分群。經過資料屬性所提供的特徵性進行分群，找出資料中特徵相關性較為一致的資料，在模型建立過程中可以減少因不同特徵所造成誤差的問題。

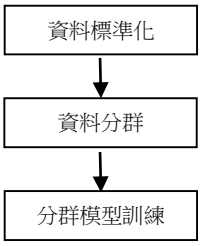

圖 3. 模型訓練流程

3. 分群模型訓練

在資料完成分群之後，則為各分群中的資料訓練並建立模型。在此階段中我們建立兩種模型用以發掘各群中資料間的關聯並且比較精準度，此兩種模型分別為多重迴歸模型（Multiple Regression）以及類神經網路（Artificial Neural Network）等。因多重迴歸模型中自變數與依變數之間的關係屬於線性關係，為發掘資料中所存在的非線性關係，所以利用類神經網路之 TanH（Hyperbolic Tangent）激勵函數轉換得到非線性之結果。

多重迴歸主要用來探討一個依變數與多個自變數之間的關係，並建立回歸方程式進行預測。回歸表示式如式（2）所示。

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \quad (2)$$

在此模型中，β_0 為常數項，$\beta_1 \ldots \beta_n$ 為迴歸係數，$X_1 \ldots X_n$ 為模型中的自變數，亦即資料集中的非缺失值欄位，ε 代表誤差，而 Y 為所需要預測的缺失值欄位。自變數的選擇使用順序搜尋法，並使用 F-test 檢定方式來檢測依變數和自變數之間是否有統計顯著性（Statistically Significant），使最後選出的自變數能夠解釋整個回歸模型。經過順序搜尋方法的回歸模型，因自變數的數量減少，在模型創建過程中可以減少計算量，對於數據量較大的資料可以明顯提升其計算速度。

類神經網路為一種模擬人類大腦架構的網路，藉由多個神經元與權重所組成，神經元透過激勵函數決定激勵程度並可轉換為非線性之關係，再經過隱藏層的傳遞進行預測，最

後針對預測與實際值的結果進行修改，以達到學習的效果。類神經網路模型的訓練以資料集中的非缺失欄位作為神經網路輸入層，並將目標欄位（缺失值欄位）作為神經網路之標籤。經過神經網路的預測與實際標籤值的誤差計算，並透過反向傳遞方法進行權重的修改方式替不同的群集建立不同的神經網路模型。

（三）缺失值分群判斷

在以訓練資料分群並為各群建立預測模型之後，對缺失值預測需決定含缺失值的該筆資料的群組歸屬。此含缺失值資料的分類使用 KNN 的方法，將訓練資料的非缺失值欄位作為 KNN 的輸入資料，並將前述群集作為 KNN 的訓練標籤，最終對於測試資料進行分類。而 KNN 計算中 K 值的選擇，則使用交叉驗證（Cross-Validation）[9]方法進行評估，以找出 KNN 的最佳 K 值。確定含缺失值資料的分類之後，則以該群的模型進行缺失值的預測。完整模型訓練以及缺失值預測的流程如圖 4 所示。

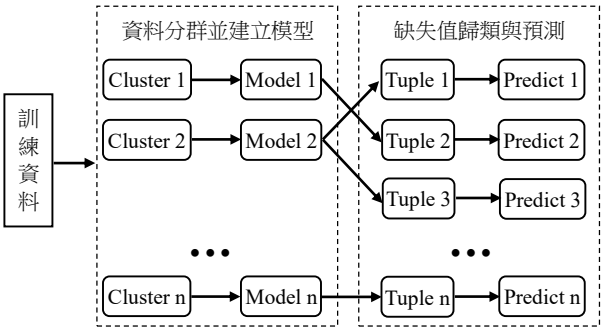

圖 4. 模型訓練與缺失值預測流程

四、實驗方法與結果

為驗證本研究所提出的塑模方式的正確性，在本實驗中將完整的資料集中部分資料作為測試資料，而其他資料則用來訓練模型。各個測試資料經分群歸類後以該群模型預測缺失值，並以預測值與實際值計算均方根誤差來評估模型的精確度。實驗步驟如下：

1. 將資料集中的資料隨機分為訓練資料以及測試資料
2. 使用訓練資料建立分群以及各群的多重迴歸與類神經網路進行模型
3. 使用訓練資料的分群結果作為標籤，將測試資料分類
4. 依測試資料分類後的群組模型進行預測
5. 使用 RMSE 比較模型所預測的目標值與原始值的差異

RMSE 誤差值計算的方式如式（3）所示

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2} \qquad (3)$$

n 為比對的資料數量，p_i 為所預測之值，r_i 為實際值。當 RMSE 值越小代表與預測結果越為相似。

（一）測試資料集

本實驗所使用之數據為加州大學 Irvin 分校所提供之葡萄酒品質數據 [14]，該數據包含兩個關於白酒與紅酒且具相同屬性的資料集。每個資料集中包含了 12 個屬性，如表 2 所示。在白酒資料集中包含了 4898 個值組，紅酒資料集中包含了 1599 個值組，兩資料集中皆不包含任何缺失值（NA）。在實驗中目標預測欄位為 fixed acidity，利用不同缺失比例將資料進行完全隨機缺失（MCAR）類型進行缺失值的替換。

表 2. 實驗資料集

英文名稱	中文名稱	英文名稱	中文名稱
fixed acidity	固定酸度	total sulfur dioxide	總二氧化硫
volatile acidity	揮發性酸度	density	密度
citric acid	檸檬酸	pH	酸鹼度
residual sugar	剩餘糖量	sulphates	硫酸鹽
chlorides	氯化物	alcohol	酒精濃度
free sulfur dioxide	游離二氧化硫	Quality	品質

表 3. 測試資料分割比率

比例（%）	1	3	5	10	15	20	25	30
白酒（筆數）	48	146	244	489	734	979	1224	1469
紅酒（筆數）	15	47	79	159	239	319	399	479

（二）訓練資料與測試資料比例

為比較訓練資料量是否會影響模型的精準度，在實驗中將實驗資料集依不同比率分割為 8 組的訓練資料以及測試資料，含有缺失值的測試資料比率如表 3 所示。

本研究所提出的塑模方法與其他常用填補做模擬比較，各種比較方法在下列統計圖表中的命名如下：

1. 均值填補：mean
2. 多重填補：mice_lr
3. KNN 填補：knn
4. 回歸模型：lm_x，其中 lm_org 表示以全部訓練資料建立多重迴歸模型，其他則表示將訓練資料分群後為各群建立多重迴歸模型，例如 lm_cluster3 表示將訓練資料分割

為 3 群之後為各群建立迴歸模型的預測方式

5. 類神經網路模型：ann_x，其中 ann_all 表示將全部訓練資料以類神經網路建立預測模型，其他則表示將訓練資料分群後為各群以類神經網路建立預測模型，例如 ann_cluster3 表示將訓練資料分割為 3 群之後為各群建立類神經網路的預測模型

（三）紅白酒資料比較結果

以紅白酒資料依上述各項預測模型的 RMSE 的統計圖如圖 5、6 所示。在兩個不同資料集所做的模擬分析結果中顯示，傳統「均值填補」以及「多重填補」的方式精準度皆低於其他模型的填補的結果，且以 KNN 針對所有資料進行資料填補所得到的 RMSE 也較其他以分群方式建立模型的方法的精準度低。以訓練資料佔資料集總資料量的比率觀察，在兩個不同的資料集均顯示，預測的精準度不會因為訓練資料量的提高而增加。在兩個資料集的模擬測試中均顯示，訓練資料量佔約 90%優於較多訓練資料量的 99%。

以分群的方式建立多預測模型所得到的精準度優於其他預測模型，不同群組數量所得到的結果間的差異並不大，在詳細分析結果後發現訓練資料分群數量為 4 的時候在所有缺失率之下皆可獲得較低的 RMSE。在以分群建立模型的原則之下，以類神經網路建立的模型所得到的 RMSE 在兩個不同的資料集中均略優於以多重迴歸的模型，以白酒資料集的用類神經網路以及多重迴歸模型模擬結果如圖 7、8 所示。

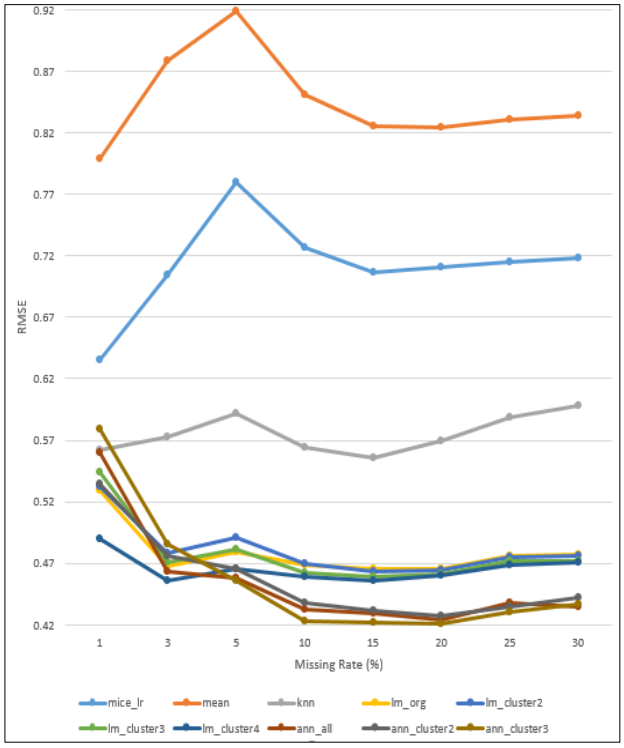

圖 5. 白酒資料於不同缺失率下 RMSE 比較

五、結論與未來發展

在資料收集過程中，缺失值的問題是不可避免的。而對於缺失值的處理方法，會造成後續分析及應用有很大的影響。在本實驗中利用資料中屬性的特徵進行分群，並為每個

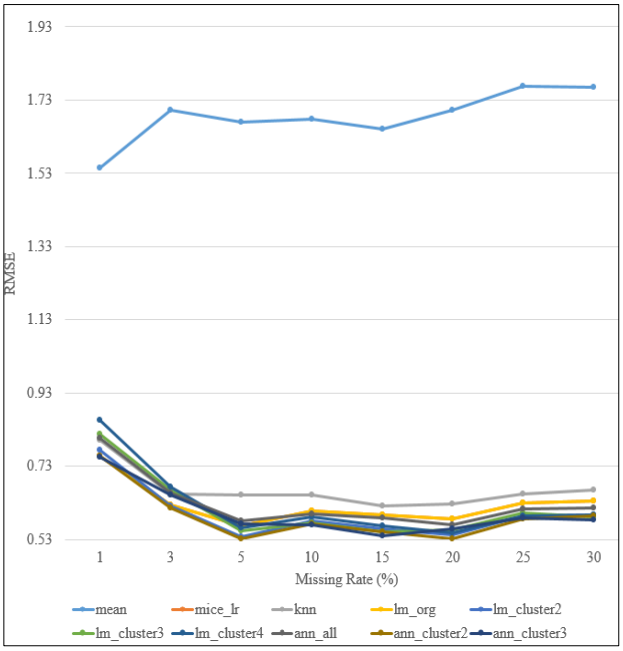

圖 6. 紅酒資料於不同缺失率下 RMSE 比較

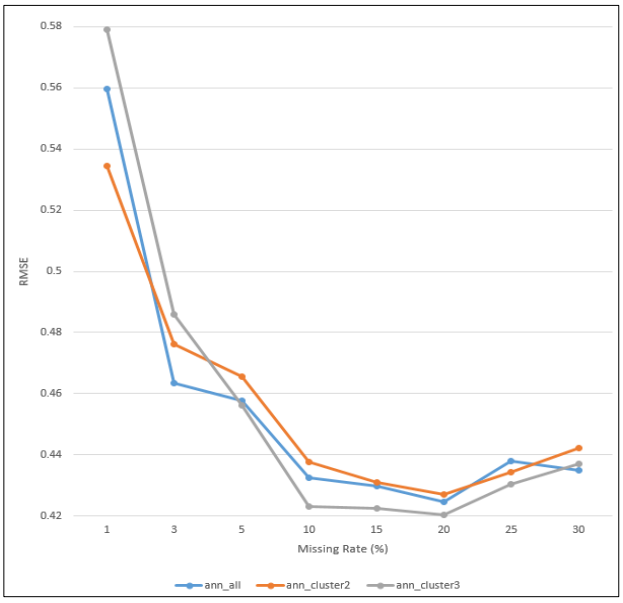

圖 7. 白酒資料於類神經模型之 RMSE 比較

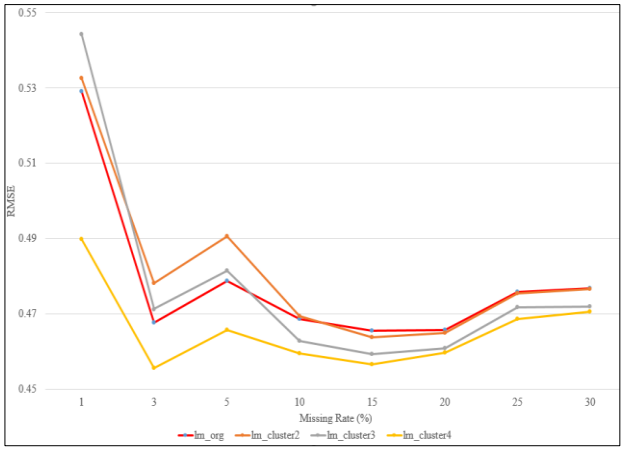

圖 8. 白酒資料於多重迴歸模型之 RMSE 比較

群集產生一個模型。當需進行預測時，先判斷資料中的特徵屬於哪種特徵群，再利用該群集所建立出的模型進行預測。在實驗結果中，利用本實驗框架可以得到較精準的預測。在實驗的結果顯示，以類神經網路所建立的模型在分群的模式下均略優於多重迴歸的模型，因此未來對缺失值預測以類神經網路所建立的模型可以提供較高的精準度。此外，經由實驗的結果顯示，利用不同的 K 值進行訓練資料的 K-means 分群，在結果亦發現經由不同 K 群所建立出的模型對於結果亦會造成影響，然而群集的數量所產生的誤差在實驗所使用的資料集與其他方法相較並不是特別的明顯。因此，實驗結果可以驗證以本研究所提出的以分群方式所建立的預測模型在精準度的比較上均優於其他傳統的缺失值預測方式。

在延續本研究的成果，未來的方向可以下列的方向來提升預測模型的精準度：

1. 在目前類神經網路的模型中採用了 TanH 的激勵函數，在若干的研究論文中提出因 ReLU 激勵函數具有解決梯度爆炸問題、計算數度相當快、收斂速度快等特性，能提供較佳的預測模型，未來可以不同的激勵函數實驗，以取得最佳的預測模型。
2. 在本實驗中無法明確指出資料應該分為的群集數量，而是利用實驗數據進行群集數量（K 值）的選擇，在後續研究中需要提出一個系統性的方式找出較好的群集數量。
3. 在本研究中，利用非缺失欄位進行分群並建立模型，對於單一欄位缺失可以獲得較精準的預測。但對於多欄位缺失的狀況之下，因無法明確建立各群的模型，導致無法的進行缺失值的預測 。針對此問題，未來可以在群集中利用非監督式學習的方法進一步的探討。

參考文獻

1. Basheer, Imad A. and Maha Hajmeer (2000) Artificial Neural Networks: Fundamentals, Computing, Design, and Application, *Journal of Microbiological Methods*, 43(1), 3-31.
2. Bennett, Derrick A. (2001) How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5), 464-469.
3. Beretta, Lorenzo and Alessandro Santaniello (2016) Nearest Neighbor Imputation Algorithms: a Critical Evaluation, *BMC Medical Informatics and Decision Making*, 16, 198-208.
4. Cover, Thomas M. and Peter E. Hart (1967) Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory*, 13(1), 21-27.
5. Hartigan, J. A. and M. A. Wong (1979) Algorithm AS 136: A k-Means Clustering Algorithm, *Journal of the Royal Statistical Society*, 28(1), 100-108.
6. Houari, Rima, Ahcène Bounceur, A. Kamel Tari and M. Tahar Kecha (2014) Handling Missing Data Problems with Sampling Methods, IEEE International Conference on Advanced Networking Distributed Systems and Applications. Bejaia, Algeria.
7. Little, Roderick J. A. and Donald B. Rubin (2019) *Statistical Analysis with Missing Data*, 3rd Edition, 4-11, John Wiley & Sons, Inc., NJ.
8. MacQueen, J. B. (1967) Some Methods for Classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 281-297.
9. Mullin, Matthew D. and Rahul Sukthankar (2000) Complete Cross-Validation for Nearest Neighbor Classifiers, Proceedings of the Seventeenth International Conference on Machine Learning, 639-646.
10. Pedersen, Alma B., Ellen M Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R. Kristensen, Tra My Pham, Lars Pedersen and Irene Petersen (2017) Missing Data and Multiple Imputation in Clinical Epidemiological Research, *Clinical epidemiology*, 9, 157-166.
11. Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams (1988) Learning Representations by Back-propagating Errors, *Nature*, 323, 533-536.
12. Schlomer, Gabriel L., Sheri Bauman and Noel A. Card

(2010) Best Practices for Missing Data Management in Counseling Psychology, *Journal of Counseling Psychology*, 5(1), 1-10.

13. Tabasi, Sanaz, Alireza Aslani, and Habib Forotan (2016) Prediction of Energy Consumption by using Regression Model, *Computational Research Progress in Applied Science & Engineering*, 2(3), 110 - 115.
14. Wine Quality Data Set, Machine Learning Repository, University of California, Irvine, Retrieved December 23, 2019, from https://archive.ics.uci.edu/ml/datasets/wine+quality.
15. Zhang, Y., B. Zhou, X. Cai, W. Guo , X. Ding and X. Yuan (2020) Missing Value Imputation in Multivariate Time Series with End-to-end Generative Adversarial Networks, *Information Sciences*, 551, 67–82.

收件：109.10.11　修正：109.12.24　接受：110.02.01