國立中興大學資訊科學與工程學系

碩士學位論文

基於k鄰近填補法在不完整資料集中

找尋近似天際線

Finding Approximate Skyline Set by

k-NN Based Imputation under Incomplete Data Set

(初稿)

指導教授︰ 賈坤芳 Kuen-Fang Jea

研 究 生︰ 凌政楠　 Cheng-Nan Ling

中華民國一百零九年八月

**誌謝辭**

說明：表達對師長、受訪者、同學、家人等感謝之意，以一頁為原則，最多不超過兩頁。(非必備，由各系所自行決定)

# ­摘要

現今大數據資料分析有一類是考量使用者偏好的相關應用，而天際線查詢演算法是最常被使用於此應用的技術之一。良好的天際線查詢演算法仰賴於完整的輸入資料集，因此解決輸入資料集中因缺失資料而造成資料不完整就成為一個關鍵議題。本研究提出一個基於k鄰近點填補缺失資料的方法，儘可能地找到可參考的鄰近點以對缺失值填補新值，當鄰近點不足或是在缺失率高的情況下，則使用採樣法以參考該維度其他無缺失值的鄰近點作為填補新值的依據。本研究以與原天際線的相似程度作為評測我們的方法與原始k鄰近點填補法的填補效果比較標準。實驗結果顯示，本研究方法在低缺失率時與原始k鄰近點填補法的填補效果相近；當缺失率介於20%到70%間，其填補效果較原始k鄰近點填補法好30%至50%；即使在缺失率高達80%以上時，與原天際線相似度也高於原始k鄰近點填補法3到6倍。針對解決缺失資料集完整性的議題，本方法面對不同缺失率均具有良好的填補效果。

關鍵字：天際線查詢演算法，缺失資料，k鄰近點填補法，採樣法

# Abstract

In big data analysis, the skyline query algorithm is one of the most commonly used techniques to find optimal decisions satisfying user’s preference. A good algorithm for skyline queries relys on the completeness of input data set. Solving the missing data problem, which results in data incompleteness, is however a critical issue. A new imputation method is proposed in this study, which is based on the concept of k-nearest neighbor imputation and consideration of different missing situations simultaneously. The proposed method finds out the nearest neighbors to impute missing data as much as possible. When the avaliable neighbors are too insufficient to be referenced, or at a high rate of data missing, a sampling technique is used to select from the neighbors without missing data. To compare with the original k-nearest neighbor imputation, we adopt the closeness of the skyline set calculated from the imputed data to the original skyline set as the metric of measuring imputation quality. The experiments show that the proposed method has an approximate result to the original k-nearest neighbor imputation at a low missing rate. Furthermore, it outperforms the original k-nearest neighbor imputation from 30% to 50% at the missing rate between 20% and 70%. Even if the missing rate is higher than 80%, the imputation quality of the proposed method can also outperform 3 to 6 times than that of the original k-nearest neighbor imputation. Finally, under any kind of missing situations, the proposed method shows at least 50% approximation of the original skyline set. In sum, the proposed method is effective in solving the missing data problem for skyline query algorithms.

Keywords: skyline query algorithm, missing data, k-nearest neighbor imputation, sampling

# 目次

[摘要 1](#_Toc49302217)

[Abstract 2](#_Toc49302218)

[目次 3](#_Toc49302219)

[表目次 5](#_Toc49302220)

[圖目次 6](#_Toc49302221)

[第 1 章 簡介 7](#_Toc49302222)

[第 2 章 相關研究 9](#_Toc49302223)

[2.1資料缺失類型 9](#_Toc49302224)

[2.2缺失值處理方法 9](#_Toc49302225)

[2.2.1丟棄法 10](#_Toc49302226)

[2.2.2填補法 12](#_Toc49302227)

[2.2.3 k鄰近點填補法 14](#_Toc49302228)

[第 3 章 問題與方法 16](#_Toc49302229)

[3.1研究動機 16](#_Toc49302230)

[3.2問題定義 17](#_Toc49302231)

[3.3問題分析 17](#_Toc49302232)

[3.4 sk-NN imputation演算法 18](#_Toc49302233)

[3.5以原天際線評斷填補法的表現優劣 22](#_Toc49302234)

[第 4 章 實驗結果與分析 24](#_Toc49302235)

[4.1實驗環境 24](#_Toc49302236)

[4.1.1實驗平台 24](#_Toc49302237)

[4.1.2實驗資料來源 24](#_Toc49302238)

[4.2實驗一: k值與缺失值比例對天際線的差異 25](#_Toc49302239)

[4.2.1實驗目的 25](#_Toc49302240)

[4.2.2實驗方法 25](#_Toc49302241)

[4.2.3實驗結果與分析 25](#_Toc49302242)

[4.3實驗二: 各填補法產生的天際線與原天際線之相似度 28](#_Toc49302243)

[4.3.1實驗目的 28](#_Toc49302244)

[4.3.2實驗方法 28](#_Toc49302245)

[4.3.3實驗結果與分析 28](#_Toc49302246)

[4.4實驗結論 32](#_Toc49302247)

[第 5 章 結論與未來方向 33](#_Toc49302248)

[5.1結論 33](#_Toc49302249)

[5.2未來研究方向 33](#_Toc49302250)

[參考文獻 34](#_Toc49302251)

# 表目次

[表 2.1 含缺失值資料集M 11](#_Toc49302207)

[表 2.2 刪除資料列 11](#_Toc49302208)

[表 2.3 刪除維度 11](#_Toc49302209)

[表 2.4 過度刪除資料列 12](#_Toc49302210)

[表 2.5 過度刪除維度 12](#_Toc49302211)

[表 3.1 sk-NN imputation演算法符號定義表 19](#_Toc49302212)

[表 4.1 UCI Machine Learning Repository輸入資料集資訊 25](#_Toc49302213)

[表 4.2 各填補法相似度比較表 (k=1) 29](#_Toc49302214)

[表 4.3 各填補法相似度比較表 (k=5) 30](#_Toc49302215)

[表 4.4 各填補法相似度比較表 (k=13) 30](#_Toc49302216)

# 圖目次

[圖 3.1 NaN-Euclidean distance 18](#_Toc49302197)

[圖 3.2 sk-NN imputation演算法 21](#_Toc49302198)

[圖 3.3 Procedure Impute\_Process() 22](#_Toc49302199)

[圖 4.1 不同缺失值比例下之hit ratio (k=1) 26](#_Toc49302200)

[圖 4.2 不同缺失值比例下之hit ratio (k=2) 26](#_Toc49302201)

[圖 4.3 不同缺失值比例下之hit ratio (k=3) 27](#_Toc49302202)

[圖 4.4 不同缺失值比例下之hit ratio (k=4) 27](#_Toc49302203)

[圖 4.5 各填補法相似度比較圖 (k=1) 29](#_Toc49302204)

[圖 4.6 各填補法相似度比較圖 (k=5) 30](#_Toc49302205)

[圖 4.7 各填補法相似度比較圖 (k=13) 31](#_Toc49302206)

# 簡介

在一個具有n個d維度資料點的資料集中，若其中某一資料點p在所有維度依據某種指標都比另一資料點q好時，我們稱該資料點p支配資料點q[12]。在一個資料集中，所有不被任何其他資料點支配的點所形成的集合[4]，被稱為天際線(skyline set)。從資料集中找出天際線的演算法就稱為天際線查詢演算法(skyline query algorithm)。

在現今的大數據資料分析中，天際線查詢演算法在最佳化問題範疇中最常被廣泛地應用在路徑規劃、策略選擇、使用者偏好、多條件排程、多偏好分析與多準則決策等問題上，其中最典型為Borzsony於2001年的論文內的飯店例子[2]。在生活當中購買房屋時，欲找到的房屋，價格愈低愈好且房屋坪數越大越好。但現實上同時滿足上述兩個條件的房屋並不多，原因是通常坪數大的房屋價格也不低。藉由天際線查詢演算法計算後，最終結果不管是在價格上或是在房屋坪數上都能符合購屋者的期待。

目前天際線查詢演算法中[15], [18]，以block-nested-loops(BNL)為例，每一個資料點必須與其他資料點逐一比較每一個維度(屬性)值的大小，在確定各點之間的支配關係後，方能決定哪些資料點可被納入天際線。任何天際線查詢演算法都有共同的假設：輸入資料集不能有缺失值的存在。若在確定支配關係的過程中，某資料點在某個或某些維度具有缺失值，造成該維度的值無法被比較進而導致無法確定該點與其他點的支配關係，這種情形將使得天際線查詢演算法無法執行。然而在現實生活上中，蒐集到資料難免會面臨到資料不齊全的狀況[11], [26]，例如在蒐集過程中不慎或某些因素致使資料遺失[5]，導致蒐集到的資料不完整。

為了解決天際線查詢演算法無法運用於不完整資料集的問題[27]，本論文針對不完整資料集提出新的填補缺失值技術，使所有缺失值都有可參考的新值，形成一個新的完整資料集，讓天際線查詢演算法可以順利執行。在過去填補缺失值填補研究中， k鄰近點填補法是最常見且簡單的方法。該填補法針對具有缺失值(假設在維度i)的資料點尋找其k個鄰近資料點，以這k個點在維度i的平均值作為填補的新值。k鄰近點填補法具有兩個缺點：其一是在填補過程中計算含有缺失值的資料點與其他點的歐氏距離時，尋找到的距離最短的k鄰近點可能不是真正的k鄰近點，並且不同鄰近點均給予相同的權重值也不一定合理；其二是若鄰近點不足k個，原始k鄰近點填補法並不會積極地尋找剩下可參考的鄰近點，使得填補效果趨近於單一數值的填補法，這種情況會隨著缺失值比例愈大而愈趨嚴重。過去研究對k鄰近點填補法並沒有針對不同缺失值(佔原資料集的)比例之填補效果有太多深入的分析。

為了解決上述k鄰近點填補法所面臨的問題，本研究提出了一個sk-NN imputation演算法賦予不同鄰近點合理權重值的方法以及挑選鄰近點的採樣機制。為彰顯愈鄰近的點對填補缺失值有較高的影響力，我們改以鄰近點與含有缺失值資料點距離的倒數作為填補缺失值的新權重值，藉此改善k鄰近點填補法中對不同鄰近點皆賦予相同權重值之不合理性。我們並且透過適當的採樣機制儘可能地挑選出足夠的鄰近點以計算填補值，改善了k鄰近點填補法面臨鄰近點不足時退化為以單一數值填補的問題。

本研究進行兩個模擬實驗。首先觀察原始k鄰近點填補法在不同k值與缺失值比例下對原天際線所造成的乖離狀況，然後比較本方法與各填補法在填補缺失資料後所產生的天際線與原天際線的相似度。實驗結果顯示，在k值足夠大時，本研究方法在缺失值比例低時所填補的效果幾乎與原始k鄰近填補法相同甚至更好；在缺失值比例較高時，本研究方法所得到天際線與原天際線相似度至少維持50%，表示超過一半的天際線經過本研究的填補方法被找回來。

本研究的主要貢獻在於改善原始k鄰近點填補法對鄰近點不合理的權重分配，並且提出鄰近點的採樣方法以解決無法找到足夠可參考的鄰近點進行填補的問題。這兩點改進使得針對高缺失值比例的資料集執行天際線查詢演算法時，仍然可以找到與原天際線相似的近似天際線。

本論文後續的章節如下: 第二章敘述相關研究，第三章描述問題與方法、第四章顯示實驗結果與分析，以及第五章是結論與未來研究方向。

# 相關研究

本論文相關研究有兩個面向：資料缺失類型(types of missing value)與缺失值處理方法(missing value handling)。

## 2.1資料缺失類型

資料集中常見的缺失值類型分別有隨機缺失類型(missing at random，MAR)、完全非隨機缺失類型(missing not at random，MNAR)以及完全隨機缺失類型(missing completely at random，MCAR)三種[8], [27]。

首先，若缺失資料發生的機率與資料集內的其他變數有關，但與缺失值本身的數值大小無關，則稱此一類型的缺失值為隨機缺失(MAR)類型。例如有關心情沮喪議題的問卷較常見到缺少男性族群的資料，使得男性族群與該問卷議題看似有相關性，但這只是男性較不願意填寫此類型問卷，事實上資料缺失因素與男性族群毫無關係，此一類型的缺失值就被歸類為隨機缺失值。若缺失資料類型屬於MAR類型時，則對該資料集內所含有的缺失值可以被納入考量或者經過缺失值的相關處理方法解決缺失值的問題。可選擇的方法包括將具有缺失值的資料紀錄刪除或是針對該缺失值賦予一個合理可供參考新的值，其他原本無缺失的數值仍可被拿來作為後續資料分析用途。

完全隨機缺失(MCAR)類型是指出現缺失值的變數其缺失機率與資料集內其它變數都沒有任何相關性，例如問卷上的回答錯誤、忘記填值、資料遺失等皆屬此一類型。若缺失資料類型屬於MCAR，可以在無缺失值的資料筆數足夠多時刪除部分具有缺失值的資料。

若缺失情形不屬於其他兩者，則被歸屬於完全非隨機缺失(MNAR)類型。此一類型的缺失值與某一維度具有一定程度的相關性，屬於此一類型缺失資料會表現出某一種資料特性，故此缺失類型完全不可忽略，也不宜用任何方式異動缺失值。例如薪資調查問卷時，高薪資與低薪資族群因為不想透漏實際薪資而拒絕填寫，造成資料缺失情況，進而導致影響資料集真實性。這類缺失類型完全不可以忽略含有缺失值的任何一筆資料，且不建議擅自刪除那些含有缺失值的資料點，以防止對資料集的特徵做出作錯誤判斷與誤導。

## 2.2缺失值處理方法

一般在做資料處理的時候，難免會遇到手上的資料集含有若干個資料欄位是不存在的，這樣的不存在任何資料的情況被稱為資料缺失，針對這些沒有資料的欄位，我們統一稱之為缺失值。

面臨資料集內含有缺失值時，為了讓整體資料集可以做後續的統計、分析、計算等工作。我們必須先對資料集內的缺失值做一些處理，無論是忽視不管、把含有缺失值的資料移除，或者是賦予這些缺失值一些替代的值，都可以是處理缺失值的其中一種方法。這些方法稱為缺失值處理方法，而缺失值處理方法可分為三種策略，丟棄法(dropout)與填補法(imputation)[3], [6], [14]，而k鄰近點填補法是效果較好的填補法，以下分別說明之。

### 2.2.1丟棄法

[介紹丟棄法]

首先介紹丟棄法，顧名思義，此類型的方法面對缺失值所採取的行為，是對缺失值所在的位置，根據目的不同以及適合的資料缺失類型。將缺失值從原資料集當中移除，以達到移除後剩下資料全部為完整資料(每個維度皆有值)且資料集為完整資料集(不含缺失值的資料集)。

[說明丟棄法種類]

即使缺失值的位置相同，我們也能以移除的方式不同，使得移除後有不同的結果。若存在有一筆資料點含有缺失值，將該資料點從原資料集中整筆刪除，這種移除缺失值的方式稱為刪除資料列。同理，若存在有一個維度含有缺失值，將整個維度從資料集內刪除，這種移除缺失值的方式稱為刪除維度。以下分別介紹刪除資料列與刪除維度兩種。表2.1為一個含有若干缺失值的資料集，表中以橫線顯示缺失值欄位。

[分別說明刪除資料列]

刪除資料列，也稱為名單去除法(listwise deletion)。根據刪除資料列的定義，將具有缺失值的資料點刪除，經刪除後會剩下id為2與id為4的資料點，如表2.2所示。假設資料集內的缺失值的缺失資料類型為完全隨機缺失類型(MCAR)，且缺失值比例不大時，此時比較適合採用刪除資料列來移除資料集的缺失值。理由是在上述假設中，將含有缺失值的資料點整筆刪除。而刪除後剩下的完整資料集相較於刪除前的資料集，等同於從一群資料母群體內取其中一部分的子集合 ，而 就是蒐集那些在所有維度中都沒有缺失值資料點的集合。且子集合 的大小會受到缺失值比例增加而減少。理論上 的分佈會與母群體(原資料集)的一致，換句話說，在MCAR的情況下刪除資料列並不會造成資料集的偏差(bias)。然而，刪除資料列也有其缺陷。當缺失值比例較高的時候，同時也意味著資料集內的完整資料較少，凸顯出資料集內可用的完整資料點個數不足的問題。

表 2.1 含缺失值資料集M

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | age | gender | height(cm) | weight(kg) |
| 1 | - | Female | - | 53 |
| 2 | 18 | Male | 180 | 75 |
| 3 | - | - | 174 | 67 |
| 4 | 21 | Female | 158 | 47 |

表 2.2 刪除資料列

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | age | gender | height(cm) | weight(kg) |
| 2 | 18 | Male | 180 | 75 |
| 4 | 21 | Female | 158 | 47 |

[分別說明刪除維度]

假設資料集內缺失值的缺失類型為不完全隨機缺失類型(MNAR)，且在某個維度d缺失的情況非常嚴重時，此時可以考慮將維度d整個刪除。也就是丟棄法中另一種移除缺失值的方法，刪除維度。如表2.3所示，經刪除後會剩下weight與id兩個維度。刪除維度在資料處理的角度就相當於將原本的資料集降維，好處是在對很多演算法而言可以減少執行時間。但是過度刪除維度也會有其衍生的問題，由於刪除後所剩的維度不足或是被刪除的維度對於資料集本身有很大的代表性，讓剩下的維度無法完整地表現資料集的特徵。例如：資料集內的分類標籤(Label)性質集中於某一維度時，刪除該維度將使對應到的分類資料大量失去特徵訊息，導致刪除後的分類不準確的問題。

表 2.3 刪除維度

|  |  |
| --- | --- |
| id | weight(kg) |
| 1 | 53 |
| 2 | 75 |
| 3 | 67 |
| 4 | 47 |

[說明丟棄法優缺點]

採用丟棄法的優點是可以透過簡單移除缺失值的方式，同時達到保有完整資料集且不會增加原資料集的偏差。然而，丟棄法的缺點是，在丟棄法的過程中可能會因為過度刪除產生額外的問題。

[過度刪除資料列缺點]

以刪除資料列來說，過度刪除資料列會導致資料點的數量不足。如表2.4表示由表2.2過度刪除資料列的結果，若是id為2的資料點任何維度也存在著缺失值，則原資料集剩下一筆id為4的資料點。

[過度刪除維度缺點]

以刪除維度來說，過度刪除維度會喪失資料集表現的特徵。如表2.5為由表2.3因過度刪除維度後，失去維度id後喪失原資料集表現特徵。失去兩者都會損失因為移除缺失值而喪失資料集內的資訊。

表 2.4 過度刪除資料列

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | age | gender | height(cm) | weight(kg) |
| 4 | 21 | Female | 158 | 47 |

表 2.5 過度刪除維度

|  |
| --- |
| weight(kg) |
| 53 |
| 75 |
| 67 |
| 47 |

### 2.2.2填補法

[說明填補法]

接著介紹填補法，不同於移除缺失值，另一種面對缺失值的處理方法，透過觀察資料集內的資料點，參考那些不含缺失值的資料點後，賦予缺失值一個合理的參考值。這樣計算出一個合理的參考值給予缺失值的過程，稱之為填補法。填補法與丟棄法相同的地方是，填補後仍然可以保持資料集的完整性；不同的是，經填補法填補後的資料個數與填補前一致。

[填補法種類，如何區別]

根據填補後可以產生一組或是多組的資料集數量來看，填補法可分為單一填補法(single imputation)與多重填補法(multiple imputations，MI)[1], [6], [23]兩類。兩者最明顯的差別在於，單一填補法經單次填補完後只會產生一組填補後的完整資料集，之後便以這組資料集進行後續的統計與分析。而多重填補法則可以視為以不同的單一填補法，經過多次地填補後，產生多組不同的完整資料集，再將這些多組填補結果合併後做後續的統計與分析。

[說明那些屬於單一填補法]

單一填補法[3], [14]主要的原理根據某一套尋找參考資料點的規則，根據該規則去計算或找到合理的數值，再以找到的合理值填補某筆資料點的缺失值。假設計算過程中產生多個候選值，單一填補法也會根據不同演算法的機制挑選其中一個填補該缺失值。

根據單一填補法的定義，下列每種填補法都有各自的一套尋找參考資料點或參考值的規則，屬於單一填補法的有：固定數值填補法、熱卡填補法(hot deck imputation)[9]、冷卡填補法(cold deck imputation)[21]、平均值填補法(mean imputation or mean substitution)、迴歸填補法(regression imputation)與最鄰近填補法(nearest neighbor method)[10]，分別於下面依序介紹。

[固定數值填補法]

首先最簡單的填補方式就是固定數值填補法。此填補法的規則是對相同維度的缺失值賦予同一固定數值，例如以缺失值相同維度下的眾數、平均值、中位數、極大值、極小值等等[14]。此方法最明顯的優點就是計算量即少，填補過程最簡單且不受任何缺失種類與資料集分布。但填補固定數值的問題在於，當缺失值數量更多的時候，由於在相同維度下的缺失值也都填補了相同的值，使得降低了相異資料點之間的差異性，造成填補的效果很差。除非缺失值數量極少，注重差異性的資料集不建議以固定數值填補缺失值。

[熱卡、冷卡填補法]

第二種與第三種單一填補法參考的規則大致相同，所以一起介紹。熱卡填補法(hot deck imputation)也稱為last observation carried forward (LOCF)，會依照每個維度的特徵屬性條件不同，資料集中的變數根據所設定的條件做排序或觀測頻率，以最常或者最近被觀測到資料點的數值填補缺失值。而冷卡填補法(cold deck imputation)與熱卡填補法作法完全相反，以最不常或者最少被觀測到資料點的數值來填補缺失值。至於觀測頻率與排序方式須是應用層面而定義。

[平均值值填補法]

先前提到的填補固定數值中的填補平均數一例，就是平均值填補法(mean imputation or mean substitution)，分別以各別維度中所有已知的數值去計算其平均值，將所計算出來的平均值作為填補缺失值的參考值。當缺失值比例變高時，此方法雖然不會改變整體資料集中的平均值，但經過填補後的維度會因為填補的值均相同，使得變異數會變小，因此平均值填補法也會間接影響資料的分布方式。

[迴歸填補法]

接著談到連續型資料上具有缺失值的填補方法。假設資料點的類型為連續型資料，此連續型資料中含有缺失值並且與資料集內其他的維度具有一定的相關性。與其他維度有相關性的缺失值類型為MAR，缺失值可以經由迴歸計算後得出預測的結果並取代該缺失值，這種方法稱為迴歸填補法(regression imputation)。此種填補法會讓填補值之間的變異數降低，因此還需要再加上一個隨機誤差作為最終缺失值的填補值。迴歸填補法與平均填補法相同地方在於，當缺失值比例很高時容易影響資料集的分布，因此這兩種填補法只適合在缺失值比例較小時使用。

[最鄰近值填補法]

最後一種單一填補法為，最鄰近填補法(nearest neighbor method)。由Fix與Hodges在1951提出[22]，其運作原理如下，資料集有n筆資料點，其中某一個 為資料集中第 筆資料，如果有一筆資料為 ，則可以運用那n筆資料中的已知訊息將該筆資料點 進行歸類。在資料歸類的過程中，會牽涉到任意兩筆資料距離(distance)的遠近或是相異程度的大小。採該資料集適合計算的距離方式，來測量出兩筆資料之間的距離，若距離越小則可以推測兩者的值越相近，因此兩資料點的值也會愈接近。

[介紹多重填補法，步驟]

上述介紹完各種規則下的單一填補法後，接下來說明可以經多次填補後，產生多組完整資料集的多重填補法。多重填補法首次在1978年被Rubin提出來[20]，比起單一填補法，更著重於分析與解決問題上。多重填補法利用蒙地卡羅法(Monte Carlo method)的概念，將缺失資料填補m個可能的數值，產生m個可能的資料集並進行分析。多重填補法主要有三個過程：第一步，對每一個缺失值被重複填補m次以產生m個完整資料集。第二步，將m個完整資料集都使用針對完整資料集進行統計分析。第三步，將m個來自各完整資料集的結果，以評分函數選擇一個最合理的值或合併所有值，產生最終的結果作為填補缺失值。根據Rubin的建議是，當m過大時並不會有更好的填補效果，因此建議m的範圍落在3到10之間。

[多重填補法優點]

相較於單一填補法，多重填補法也有些好處。其中因為填補的次數很多，因此對於相同缺失值，每次填補都是從候選填補值的集合中挑選出來，也會有比較多種的選擇，因此採隨機抽取其中一種作為填補值會比單一填補法更有多樣性，避免落入填補的值總是可以預測的。

[多重填補法缺點]

多重填補法仰賴於資料集上模擬分布模型[19]，在遇到缺失資料時根據模型的分布給予一群可能為該缺失值的候選解集合，並在陸續填補過程中調整資料集的分布、變異數以及信賴區間等，因此在填補過程中會需要極大量的計算需求且在實務上不容易進行模擬。加上生成多重填補值的過程與所花費時間比起單一填補法複雜許多，也是不常採用多重填補法的原因之一。

[填補法優、缺點]

填補法的優點在於，不影響輸入統計模型、資料集分布以及避免喪失資料集的表現特性，參考無缺失值的資料點以填補缺失值較有可參考性且預期會有比較好的效果，也比較接近完整資料集的特徵。然而，填補法也其缺點，就是無法保證或是確定填補缺失的值是否正確，且填補缺失值後一定會影響原本資料集的內容，造成資料的失真。因此沒有最好的填補法，只有最適合的填補法。

### 2.2.3 k鄰近點填補法

[介紹k鄰近點填補法]

k鄰近點填補法(k-nearest neighbor imputation，k-NN imputation)[8], [16]是很實用的一種填補缺失值的方法。k鄰近點填補法核心概念為：假設鄰近的資料應該很相近，因此在相同維度上，鄰近資料的值也會很接近。今已知存在一個缺失值，找尋與含有該缺失值資料點最接近的資料點，作為鄰近點。蒐集這些鄰近點的值後，並且加這些值合併後為一個新值，最後將此新值填補該缺失值。

[ k鄰近點填補法過程]

k鄰近點填補法的填補過程敘述如下，假設資料集中存在某資料點p，在任意維度d上含有缺失值。在填補值之前，首先計算所有資料點與點p的距離值，接著將這些距離值由小到大排序後，蒐集那些距離p最近的k個點，這k個點也被稱為資料點p的k個鄰近點。接著取k個鄰近點同樣在維度d上的值，將這k個值相加之後求其算術平均值作為最終填補資料點p在維度d上的缺失值。

[k鄰近點填補法優點]

由於上述說明k鄰近點填補法的填補過程中，並無牽涉到維度大小限制、資料類型。在不考慮計算量下，k鄰近點填補法可以輕易地推廣至任意維度且幾乎適用於各種資料類型[27]，如連續型資料(continuous data)、離散型資料(discrete data)、有序型資料(ordinal)甚至是分類型資料(categorical data)[28]。k鄰近點填補法是最為常見且被認為效果比較好的補值法。且比起填補固定數值，例如平均數、中位數、極值、眾數等，k鄰近點填補法來的更準確許多。原因是k鄰近點填補法會同時參照其他與該缺失值相鄰點去預測一個更合理的值。

[k鄰近點填補法缺點]

然而，k鄰近點填補法也存在某些缺點。其中一個是，由於k鄰近點填補法參考鄰近點的值，在無資料預處理的前提下，如果挑選到不好的鄰近點或者鄰近點含有偏差值，都無法對缺失值填補較合理的值。另一個缺點是，k鄰近點填補過程會計算任意兩資料點之間的距離，計算距離的過程會被維度大小與資料點的個數影響，而需要大量的計算量。尤其在多維度資料集當中含有大量的資料點時，此計算量更是可觀。今以n筆資料為例，新增一筆資料點時就必須多n次的距離計算。此時若再新增一筆資料點時，必須再增加n+1次的距離計算(目前總共有n+1筆資料)，以此類推。使得資料點之間的距離計算量會隨著資料點的增加而呈現指數成長。

# 問題與方法

本章3.1節說明研究動機，3.2節敘述問題定義，3.3節提出問題分析，3.4節提出sk-NN imputation演算法，最後3.5節闡述在不完整資料集中如何以原天際線評斷各填補法的表現優劣。

## 3.1研究動機

尋找天際線時需要針對每二筆資料比對所有維度的值，也就是說每一筆資料的每一個維度都必須有值存在方能比對，因此資料集的完整性在天際線查詢演算法中就成為了必要條件。由於現實生活中有許多不可抗拒之因素使得取得資料集的過程難免會遇到欄位裡的值無法完備，期待蒐集到的資料每一維度都沒有缺失值是不切實際的。

為了保證資料集的完整性，解決缺失值最直覺採用的策略為丟棄法。丟棄法有兩種刪除資料的方式，分別為刪除資料列以及刪除維度。如果不完整資料集中缺失值的比例越高，刪除缺失值所在的資料列會讓所剩資料數量不足，而刪除維度則會喪失原資料集所表現的特徵。反之，採取填補法不僅不會喪失原資料集特徵，還可以保證資料點個數與原始資料集一致。在考量資料集的完整性與天際線查詢演算法的適用性，填補法會比丟棄法來得更適合。

針對不完整資料集如何執行天際線查詢演算法的問題上，本研究採取填補法填補缺失值。從填補後的效果來看，k鄰近點填補法是填補法當中表現比較好的[24]。k鄰近點填補法首先針對維度d有缺失值的資料點找出其k個鄰近資料點，取得這些鄰近點的維度d之值，再計算這些d維度值的算術平均作為新填補值，這使得被填補值相較於以一般填補固定數值的方法更具有參考性。然而k鄰近點填補法也有其缺點，其一是當計算含有缺失值之資料點與其他資料點的歐氏距離(必須先忽視含有缺失值之維度)會不準確而導致找錯鄰近點，其二是k鄰近點填補法遇到可參考的鄰近點不足時會退化為填補固定數值[7]。

對一個含有缺失值的資料點p而言，尋找鄰近點的過程中，k鄰近點填補法先計算所有資料點與p的距離，再根據距離值由小到大排序。排序的意義代表著各資料點分別對p的影響程度，但是k鄰近點填補法卻將p的所有k個鄰近點視為相同的權重值[15]。此做法與先前排序的意義，期待找出最有參考價值的鄰近點以填補該缺失值的概念相違背且不合理[27]，也就是未考慮鄰近程度對填補值的影響力。因此本研究針k鄰近點填補法的缺點以及未考慮鄰近程度的影響力問題分別提出鄰近差別權重分配與新的選擇鄰近點機制以改善k鄰近點填補法。

## 3.2問題定義

本研究要解決的問題定義如下: 在不完整資料集中，如何改善原始k鄰近點填補法填補缺失值，使填補後的完整資料集具有最近似的天際線?

本研究假設不完整資料集中，缺失值的缺失類型為2.1節中提到的完全隨機缺失類型(MCAR)，即含有缺失值的屬性與其他欄位屬性無相關性。為驗證填補效果，我們將填補後的完整資料集計算近似天際線，並與原先無缺失資料集的天際線比較其差異，以此差異作為衡量近似天際線的相似程度。若相似程度愈高，代表該填補法的填補效果愈好。

## 3.3問題分析

所有計算距離公式[8], [16]中，最普遍常見的歐氏距離計算方法[13]是採資料集中兩兩資料點相對應維度的差值平方和再取平方根。若是至少一個維度具有缺失值，則在計算歐氏距離時並不會採計具有缺失值的維度，此計算方式是最廣為主流的算法[24]。由此計算方式可看出一個潛在的問題 : 具有缺失值的兩資料點其距離計算所得之值可能會誤導此二資料點之間的實際距離，今舉例說明之。

如圖3.1所示，A、B、C三個二維資料點，其座標分別為(1, 1)、(2, 8)、(3, 3)，。在沒有任何缺失值情況下，按照傳統歐氏距離的計算: AB距離 應該為 且 應該為 ，故 。但若將點A中的y座標設為缺失值時，則含有缺失值後的歐氏距離會將有缺失值的維度捨棄而不列入計算，新的AB距離 為 ，而新的AC距離 則為，使得 。此時新的距離會讓原本為了避免誤算不納入缺失值的機制反而錯估了距離的實際值，間接導致了大小順序上誤判的結果，這就是k鄰近點填補法在有缺失情況下只單依靠距離大小決定鄰近參考點所可能會陷入的誤區，最終與其原目的相違背。



圖 3.1 NaN-Euclidean distance

k鄰近點填補法的另一個問題在於，當缺失值愈高時， k值很大意味著鄰近點仍存有非缺失值的機會並不大，而此時k鄰近點填補法在無法找到滿足k個鄰近點情況下，選擇不從剩下的鄰近點補足並從缺，這樣的現象尤其當存在非缺失值很稀少時更為嚴峻，導致k鄰近點填補法會幾乎用同一數值填補回去，如此便會與只填補單一數值(同一維度的平均數、眾數、極大值、極小值)無異，填補後找尋天際線時又會因為該維度幾乎都是同一數值，更容易形成有如該維度直接被刪除一樣而產生無意義地比較結果。

鑒於以上分析，本論文提出新演算法除了在缺失值比例不高時填補效果能與原始k鄰近點演算法相近，並且在缺失值比例較高時也能夠改善原始k鄰近點填補法的不足。有別於k鄰近點填補法對鄰近點不足而選擇從缺不補，本研究方法在缺失值所在的維度上，尋找其他沒有缺失值的點，從這些點採樣其中k個點，計算它們在該維度的平均值，最後以該平均值填補原缺失值。其目的是不讓鄰近點的不足而計算不準確，導致填補值後產生的天際線與原天際線乖離太大。

## 3.4 sk-NN imputation演算法

本節先說明本論文提出的演算法內會運用到的符號與其定義，如表3.1所示，然後提出本研究的演算法sk-NN imputation。

已知一個不完整資料集 含有若干個缺失值，常數 為可參考鄰近點的上限個數，n與m分別為 的資料點個數與 的維度個數。 表示 中第 筆資料點，且 為該資料點 於維度 的值。

為不完整資料集 中任兩資料點 與 之間距離的距離矩陣，且其中 為任兩資料點 與 之歐氏距離。 是一個資料點 在每一個維度 的值 相對其他任意相異資料點 的權重矩陣，代表兩資料點 與 之間的權重值。 是一個決定權重值的函數，用來決定某一缺失值的被參考維度值 的權重值。若 值為 ，則所有權重值 均被設為1 ; 若 值為 ，則 會以 中任兩資料點 與 之間距離 的倒數作為該權重值。

NN list是一個由元素 有序串列所構成的串列。每一個 記錄k個鄰近 的資料點 的索引值(即 中的第j筆資料點)，而索引值j的順序則是依照資料點 與 相距的距離值 ，由小至大排序。 記錄 所有的鄰近點之中， 不為0的且與 第h個最接近資料點 的索引值j。

最後說明填補過程中使用到的符號，若輸入資料集 中某 為缺失值，則本論文填補法將會賦予該缺失值一個新值 填入原缺失欄位。mask為一個長度同為輸入參數k值的陣列，記錄某 最鄰近k個參考點在出現缺失值的維度 是否有缺失值若被參考的鄰近點也為缺失值，則標註為True，否則標註為False。最終，演算法輸出一個原始缺失值均已被填補的資料集 。

表 3.1 sk-NN imputation演算法符號定義表

|  |  |
| --- | --- |
| 符號 | 說明 |
| n | 資料點個數，共n筆資料點 |
| m | 資料點維度個數，共m個維度 |
| incomplete data set ,  = | 輸入不完整資料集，大小為 |
| , k | 常數k，k屬於自然數，決定參考鄰近點個數。 |
|  | 表示資料集內第i筆資料點維度d的值 |
|  | 的第i筆資料 |
|  | 為一個對稱的距離矩陣，記錄任意兩資料點之間的歐氏距離。 |
|  | 資料點 與 之間的距離，記作 且 。 |
|  | 權重矩陣，記錄任兩資料點 與 , 之間的權重值，其中i |
| = | 兩資料點 與 之間的權重值 |
|  | 決定權重值函數 |
| NN list | 由元素 有序串列所構成的串列 |
|  | 記錄k個鄰近 的資料點 的索引值j (即 中的第j筆資料點) |
|  | 記錄 所有鄰近點中， 不為0的且與 第h個最接近資料點 的索引值j。 |
|  | 被填補的新值 |
| mask | 長度為k的陣列，記錄在維度d上含有  缺失值的資料點 ， 最鄰近的k個  鄰近點，這些鄰近點在維度 上是否  也具有缺失值。  若是則標註為True，否則標註為False。 |
| imputed data set | 填補缺失值後的填補資料集 |

本論文所提出的sk-NN imputation演算法顯示於圖3.2，其輸入參數為一不完整資料集、常數k、以及決定權重值的函數。

執行過程依序為：step1與step2載入不完整資料集並初始化填補資料集。step3先初始化距離矩陣，step3-1與step3-2計算任兩資料點之間包含相對應維度有缺失值的歐氏距離。step4初始化權重值矩陣後，根據 來決定計算任兩資料點之間的權重值，若 為uniform則如同k鄰近點填補法給予相同權重，若 為distance則給予差別權重方式為兩點之間的距離倒數，其背後意義為兩點距離愈大對彼此的影響力愈小。step5列出每一筆資料點其所有鄰近點並儲存於 中，並且依照與該點 的歐式距離值由小到大排序。step6遍歷輸入資料集中所有缺失值，並分別檢視其鄰近點以填補新值，填補新值是透過執行副程式Impute\_Process()(示於圖3.3)所完成。step7回傳填補後的新完整資料集，結束sk-NN imputation演算法。

|  |
| --- |
| Algorithm sk-NN imputation () {  Input : incomplete data set , constant k, weighting type  Output : imputed data set  Method :  step 1. load incomplete data set  step 2. initialize imputed data set to  step 3. *// compute distance matrix*  3-1. **for** each in distance matrix **do**  3-2. Euclidean distance of data samples and  **end for**  step 4. *// compute weight matrix*  4-1. **for** each in weight matrix W **do**  4-2. =  **end for**  step 5. *// store all nearest neighbors of into a nearest neighbor list*  5-1. create and initialize an empty NN list, with size n  5-2. **for** each in **do**  5-3. retrieve all between any pair of the and from  5-4. sort in ascending order, record index j of  5-5. **repeat**  5-6. **if** 0 **then**  5-7. append index j of to  **end if**  5-8. **until** size of or all are inserted into  **end for**  step 6. *// search all missing values in , then impute new value back into the missing position in*  6-1. **for** each in **do**  6-2. **if** is missing **then**  6-3. **call** procedure Impute\_Process(i, d, k)  **end if**  **end for**  step 7. **return** the imputed data set  } |

圖 3.2 sk-NN imputation演算法

接下來說明Procedure Impute\_Process()的填補過程，圖3.3為計算填補值的pseudocode。首先step1初始化代表鄰近點缺失狀況的mask array。step2檢查每一個缺失值的鄰近點在該維度的值是否也為缺失值，若為缺失值則在mask array中標記True，否則標記為False。step3利用mask array檢視所有在維度d有缺失值的資料點其所有鄰近點在維度d是否全部都是缺失值；若mask的值皆為True，則判定可參考的鄰近點在相同維度也都是缺失值。此時就會觸發採樣機制去參考相同維度非缺失值的點(不一定為鄰近點，因為鄰近點已經不具參考性)。step4將採樣點在該維度的平均值填補回原缺失位置，至此填補其中某一缺失值結束。

|  |
| --- |
| Procedure Impute\_Process(i, d, k) {  Input : indices of missing value found in data sample at dimension d, constant k  Output : imputed value  Method :  step 1. create a mask array and initialize all elements in mask array to False  step 2. *// check all elements in with respect to the value of*  *// at column d, , is missing or not*  2-1. **for** each index from 0 to k **do**  2-2. r  2-3. **if** is missing **then**  2-4. mask[r] True  **end if**  2-5. **else**  2-6. mask[r] False  **end else**  **end for**  step 3. *// retrieve values in which index in mask array assigned to False*  3-1. **for** each index h in mask **do**  3-2. **if** all elements in mask array are not all True **then**  3-3. retrieve all value in  **end if**  3-4. **else**  3-5. reset all elements in mask array to False  3-6. sampling the rest of not missing value at column d  **end else**  3-7. evaluate mean or weighted mean for imputed value  3-8. assign imputed value to  step 4. **return** imputed value  } |

圖 3.3 Procedure Impute\_Process()

## 3.5以原天際線評斷填補法的表現優劣

執行模擬實驗時，為了觀察填補效果[25]對原天際線所造成的影響，本論文採用填補缺失值後的天際線與原天際線兩者之間的漢明距離(hamming distance)作為評斷兩者相似度之標準。漢明距離主要是用於計算兩個字串相對應的位置具有不同字符的個數，換句話說，將一個字串變換成另外一個字串所需要替換多少個字符的總數即為漢明距離。例如：兩等長(老師，我這裡的例子擺錯位置，應該放在原本漢明距離的後面，因此在這裡還是舉等長的例子)二進位字串1011101 與 1001001 由左向右第3與第5個位元相對位置值不同，故計算此字串的標準漢明距離為2；同理，toned 與 roses之間的漢明距離為3，以此類推。

本論文使用漢明距離定義中須置換字符次數的觀念，因此並沒有要求兩字串必須等長之限制。本論文所採用的是集合上的漢明距離概念，換句話說，任意兩個集合之間必須插入或刪除多少元素才能使兩集合相同。例如：已知三個集合，。P、Q需要置換三個元素才能使兩集合相同，因此P、Q之間的在本論文上所定義的漢明距離為3，Q、R的漢明距離為2，而P、R之間的漢明距離則為5。使用這樣的觀念原因有二：

1. 集合內元素不具有順序性，只能檢查某元素存在與否，此性質在字串問題上即為對應位置是否具有相同值。
2. 兩集合相同的充分且必要條件為兩集合具有相同元素且相異元素個數相同，此性質對應到字串問題上則為兩字串長度必須相同。

判斷原天際線 與近似天際線 的相似程度之計算方法如下:

* 1. 天際線 與 中相異元素總個數為該集合size。
  2. 與 集合具有相同元素個數稱為hit count。
  3. 之中有的元素但 中沒有的元素之個數以及 之中有的元素但 中沒有的元素之個數，將其加總之和，稱為miss count。miss count就是本論文所定義的集合上的漢明距離。

4. hit ratio =

今舉例說明:

original skyline set: size為5

estimated-1 skyline set : ，size為7

estimated-2 skyline set : ，size為4

estimated-1 skyline set與original skyline set具有2個元素相同C、H，hit count = 2，且A、B、D、G、F、R並沒有猜中故miss count = 6，hit ratio為 = 0.2。

estimated-2 skyline set與original skyline set具有3個相同元素B、C、E，hit count = 3，且D、G、H沒猜中故miss count = 3，hit ratio為 = 0.5。

由上述例子可知， size愈大並不能保證hit ratio一定愈好，亦即猜得多不如猜得精準。天際線為一個不被其他點支配的資料點所構成的集合，如果經填補後所找到的天際線集合與原天際線集合之相似度愈高，則可推斷該填補法對天際線所填補效果愈好。本論文用上述相似度指標hit ratio來評斷各填補法填補效果之優劣。

# 實驗結果與分析

本章依序於4.1節說明實驗環境、平台與所使用的資料來源。接著4.2節觀察k值的大小與缺失值比例對原始k鄰近點填補法填補缺失值後產生天際線的差異。4.3節探討不完整資料集在不同k值與缺失值比例下，各填補法(包含原始k鄰近點填補法、權重型k鄰近點填補法與本論文所提出的sk-NN imputation填補法)填補缺失資料後所計算出的天際線與原天際線的相似度。

## 4.1實驗環境

### 4.1.1實驗平台

本實驗的硬體設備包括處理器為Intel® Core™ i7-6700 CPU @ 3.40GHz，記憶體為16.0GB，作業系統為Microsoft Windows 10 Profession version 2004 64bits。開發環境主要使用的程式語言為Python 3.8.2版本，並以Anaconda整合開發環境(IDE)。實驗程式架設內建於Anaconda的Jupyter Lab與Notebook虛擬環境中，並引用包含處理資料流的pandas套件、數學與矩陣函式相關的numpy套件、機器學習與資料挖掘所需要的sklearn套件與數據視覺化的matplotlib套件。本研究利用Office Professional Plus 2019 Excel來輔助實驗結果分析。

### 4.1.2實驗資料來源

本研究使用的資料來源為UCI Machine Learning Repository[29]中純數值資料類型的資料集，輸入資料集名稱分別為Bike Sharing dataset、Real Estate Valuation dataset、Real-time Election Results Portugal 2019 dataset三個資料集。資料集資訊、來源與內容特徵呈現於表4.1[29]。

Bike Sharing dataset之資料性質屬於單變量(univariate)，共有17389筆資料，屬性特徵(attribute characteristics)均為整數(integer)與實數(real)，特徵欄位(attributes)總共有16個特徵。Real Estate Valuation dataset之資料性質屬於多變量(multivariate)，共有414筆資料，屬性特徵均為整數與實數，特徵總共有7個。Real-time Election Results Portugal 2019 dataset之資料性質屬於多變量，共有21643筆，屬性特徵均為整數與實數，總共有29個特徵。

表 4.1 UCI Machine Learning Repository輸入資料集資訊

|  |  |  |  |
| --- | --- | --- | --- |
|  | Bike Sharing dataset | Real Estate Valuation dataset | Real-time Election Results Portugal 2019 dataset |
| Data Set Characteristics | univariate | multivariate | multivariate,  time-series,  text |
| Number of Instances | 17389 | 414 | 21643 |
| Attribute Characteristics | integer, real | integer, real | integer, real |
| Number of Attributes | 16 | 7 | 29 |

## 4.2實驗一: 原始k鄰近點填補法對天際線的相似度

### 4.2.1實驗目的

本實驗目的是針對不同鄰近點k值與缺失值比例(missing rate)，原始k鄰近點填補法對填補缺失資料後的天際線與原始天際線之相似度差異。

### 4.2.2實驗方法

本實驗使用Real Estate Valuation dataset作為輸入資料集，由於特徵數量只有7個，因此取k值時只採1到4作為觀察對象。本實驗將同一缺失資料集於不同缺失值比例下，隨著k值增加，觀察原始k鄰近點填補法是否可以得到更高的天際線相似度。本實驗採用原始完整資料集中所得出天際線作為比較基準，hit ratio (計算方式詳見3.5節)愈高則相似度愈高，表示填補結果愈準確。

### 4.2.3實驗結果與分析

實驗結果如圖4.1、4.2、4.3、4.4所示。圖4.1顯示，當k=1且缺失值比例尚未達到20%時，原始k鄰近點填補法之hit ratio已經降至約50%左右，且缺失值比例提高至40%時，hit ratio只剩下40%左右。圖4.2顯示，k=2且缺失值比例為20%時，原始k鄰近點填補法之hit ratio比k=1時稍微上升約10%；不過缺失值比例為40%時，hit ratio與k=1的時候一致，推測可能是因為鄰近點變多，使得hit ratio稍微提升。



圖 4.1 不同缺失值比例下原始k鄰近點填補法之hit ratio (k=1)



圖 4.2 不同缺失值比例下原始k鄰近點填補法之hit ratio (k=2)

圖4.3顯示，k=3且缺失值比例為20%時，此時即使可參考的鄰近點數增加，原始k鄰近點填補法之hit ratio仍然未超過60%，且缺失值比例為40%時hit ratio降至約30%。圖4.4顯示，k=4且缺失值比例為20%時，可參考的鄰近點為圖4.1的四倍，但是此時原始k鄰近點填補法之hit ratio大約為40%，比k=1時的hit ratio為50%還低；且缺失值比例為40%時，hit ratio也沒超過k=1時的40%。由此可知k=4並沒有比k=1時相似度更好高。



圖 4.3 不同缺失值比例下原始k鄰近點填補法之hit ratio (k=3)



圖 4.4 不同缺失值比例下原始k鄰近點填補法之hit ratio (k=4)

觀察圖4.1到圖4.4可知，隨著缺失值比例在資料集當中逐漸增加，原始k鄰近點填補法的準確率(即相似度)並沒有因為參考更多的鄰近點而有明顯地改善填補效果。原始k鄰近點填補法期待透過增加鄰近點的數量，以提升填補值的品質，但是這將使得計算錯誤的填補值因為參考更多無效的鄰近點反而更嚴重。隨著缺失值比例逐漸上升，加上原始k鄰近點填補法中對鄰近點不足k個卻從缺不補的問題，致使即使計算鄰近點值的平均值也會逐漸失效，這同時也代表參考鄰近點值之可靠度會因高缺失值比例而降低。

## 4.3實驗二: 各填補法產生的天際線與原天際線之相似度

### 4.3.1實驗目的

本實驗的目的是針對在固定鄰近點k值且隨著缺失值比例上升，各填補法(包含原始k鄰近點填補法、權重型k鄰近點填補法與本論文所提出的sk-NN imputation填補法)填補後所產生的天際線與原天際線的相似度。

### 4.3.2實驗方法

本實驗所使用的資料集為Bike Sharing dataset，特徵欄位共有16個特徵，因此k值最大範圍可以到15，分別取三種不同k值分別做三次比較，並觀察原始k鄰近點填補法、權重型k鄰近點法以及本研究提出的方法填補後的值所能夠找回原天際線的程度。

x軸為缺失值占整體資料集中的比例(缺失值比例)，y軸為各填補法填補所有缺失值，再分別跑同一支天際線查詢演算法的程式(以BNL為例)，比較各填補法填補後所產生出的天際線與無缺失值的原天際線。以3.5節的評測方法計算填補前後兩個天際線的相似程度。產生的天際線與原天際線若相似度越高，則y軸的值越接近1.0，也表示該填補效果越好。

### 4.3.3實驗結果與分析

實驗二結果，如表4.2所示。當k=1時，我們可以觀察到原始k鄰近點填補法與權重型k鄰近點填補法在缺失值比例(missing rate)由20%提高至30%間，原始k鄰近點填補法與權重型k鄰近點填補法所產生的天際線與原天際線的相似程度(後簡稱相似度)都從原本的70%驟降至53.8%，而本研究所提出的sk-NN imputation填補法仍然可以維持在81.8%，直到缺失值比例為50%時相似度才降到63.6%。甚至當缺失值比例達到70%時，原始k鄰近點填補法與權重型k鄰近點填補法產生的天際線相似度分別只剩下38.4%與28.5%。sk-NN imputation填補法還可以維持63.6%。

由圖4.5更可以看出來缺失值比例大於75%以上時，sk-NN imputation填補法的已經與其他兩者有明顯的差距。由此可知，當鄰近點參考數量稀少且較高缺失值比例時，本論文提出的方法所產生的天際線比較接近原天際線。

表 4.2 各填補法相似度比較表 (k=1)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| k-NN | 0.8 | 0.7 | 0.538 | 0.6 | 0.421 | 0.5 | 0.384 | 0.25 | 0.1 |
| weighted  k-NN | 0.8 | 0.7 | 0.538 | 0.6 | 0.421 | 0.5 | 0.285 | 0.25 | 0.1 |
| sk-NN | 0.8 | 0.8 | 0.818 | 0.909 | 0.636 | 0.75 | 0.7 | 0.636 | 0.545 |

圖 4.5 各填補法相似度比較圖 (k=1)

如表4.3所示，在k=5時，各個填補法所產生的天際線與原天際線的相似度。可以看出來當鄰近點比較多，原始k鄰近點填補法與權重型k鄰近點填補法比k=1時的表現更好一些。可以觀察出在缺失比例介於40%至50%間，原始k鄰近點填補法有機會擁有較好的填補效果是因為缺失程度不高，原始k鄰近點填補法還能夠以足夠的k與鄰近點計算平均後填回。即使缺失值比例已達70%，原始k鄰近點填補法與權重型k鄰近點填補法都還有60%與66.6%的相似度。

圖4.6顯示，當k=5時，sk-NN imputation填補法除了在缺失值比例為40%與50%時，相似度略低於其他兩者。在超過80%的缺失情形下，sk-NN imputation填補法所產生天際線的相似度至少與其他兩種填補法相同，甚至高於其他兩種填補法。甚至在缺失值比例90%時候，sk-NN imputation填補法所產生天際線的相似度為k鄰近點填補法的2倍。

表 4.3 各填補法相似度比較表 (k=5)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| k-NN | 0.9 | 0.666 | 0.666 | 1 | 0.625 | 0.533 | 0.6 | 0.454 | 0.181 |
| weighted  k-NN | 0.9 | 0.625 | 0.6 | 1 | 0.529 | 0.692 | 0.666 | 0.1 | 0.25 |
| sk-NN | 0.9 | 0.833 | 0.909 | 0.818 | 0.5 | 0.8 | 0.6 | 0.5 | 0.363 |

圖 4.6 各填補法相似度比較圖 (k=5)

如表4.4所示，當k=13時，各填補法所產生的天際線與原天際線的相似度。原始k鄰近點填補法與權重型k鄰近點填補法在缺失值比例從30%到50%的區間以及從50%到90%的區間，兩者的相似度分別從75%與81.8%下降至18.1%。反觀本研究提出的sk-NN imputation填補法，除了40%缺失值比例相似度45.4%以外，其他情形下幾乎都維持在70%以上的相似度。表示高缺失值比例時所觸發的採樣機制有其效果。

表 4.4 各填補法相似度比較表 (k=13)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| k-NN | 0.9 | 1 | 0.75 | 0.333 | 0.727 | 0.583 | 0.615 | 0.384 | 0.181 |
| weighted  k-NN | 0.9 | 0.833 | 0.818 | 0.4 | 0.277 | 0.529 | 0.428 | 0.352 | 0.181 |
| sk-NN | 0.9 | 1 | 0.833 | 0.454 | 0.818 | 0.7 | 0.727 | 0.75 | 0.7 |

圖4.7顯示，在缺失值比例30%以下時，sk-NN imputation填補法幾乎與原始k鄰近填補法無異。原因是此時可以參考的鄰近點足夠多且缺失值比例不高，加上k值夠大因此不易觸發採樣機制，使得sk-NN imputation填補法所計算值與原始k鄰近填補法大部分都一樣。並且由圖4.7可以發現，缺失值比例大於55%之後sk-NN imputation填補法與其他兩者開始有顯著的差異，原始k鄰近填補法與權重型k鄰近點填補法的相似度幾乎呈現嚴格下降趨勢，但sk-NN imputation填補法仍然維持在60%至80%之間。由此可知即使在對於原始k鄰近填補法最有優勢的情形下(k=13)，在高缺失值比例下本研究所提出sk-NN imputation填補法對天際線的填補效果明顯具有優勢。

圖 4.7 各填補法相似度比較圖 (k=13)

本論文的方法之中有採取採樣的機制，此機制在k值不大的時候且缺失值比例不高時會比較容易啟動，但隨著缺失比率增加下，增加k值所帶來的益處會越來越不明顯。而本論文中適時觸發採樣的方式，比增加k值有更大的機會找到有效的鄰近點。從表4.2與4.3中可以看出原始k鄰近點填補法從缺失比例約為30%時，相似度度就開始急遽下降，雖然中間偶有小幅度的上升，但可看出下降程度在缺失值比例大於75%以後下降幅度加劇，表4.4更是提前在缺失值達55%時就開始大幅下滑，這都顯示出一現象，原始k鄰近點填補法參考鄰近點的機制，在缺失率高下無法具有穩定的填補效果。

## 4.4實驗結論

[已重寫4.4全部段落]

實驗結果如圖4.5、4.6、4.7所示。可以知道無論是原始k鄰近點填補法或是權重型k鄰近點填補法，在k值無論為1、5或是13時，當缺失值比例上升，原始k鄰近點填補法的填補效果都不好，分析其原因有二。

原因之一是以目前主流計算含有缺失值的距離公式，明確地定義出，對應相同的維度上兩資料點都必須有值才能做計算。若兩資料點在相對應維度(假設維度d)上至少有一者的數值為缺失值，則維度d那一項就不會被納入歐氏距離公式當中計算。這樣的結果使得維度d在兩資料點距離計算上的影響力會直接被無視，也是原始k鄰近點填補法在找尋最接近鄰近點時會被誤判鄰近關係的主要原因之一。

另一個原因則是，在尋找k個鄰近點的過程中，可能無法滿足k個鄰近點在維度d上都沒有缺失值。原始k鄰近點填補法在遇到此種情況時，會選擇從缺不補。如此看似正常的步驟，探討在計算填補值的過程就會發現到，剩下不足k個的鄰近點，這些點在維度d值之權重其實正在無形地上升。說明如下，原始k鄰近點填補法如果找得到滿足k個鄰近點，在維度d上都有值的話。此情形下平均每個鄰近點所分配到的權重值為 ，假設今天找不到滿足k個鄰近點，例如找到 個，其中 ，此時每個點的權重值將為 。而 值會隨著缺失值比例上升而降低， 與缺失值比例呈現負相關。換句話說，缺失值比例增加，可以找得到滿足k個鄰近點在維度d上不含有缺失值的機會就會降低。因此當缺失值比例越高的時候，權重值被迫上升的狀況也就愈嚴重。導致原始k鄰近點填補法所填補的新值雖然表面上是看似公平的平均值，但事實上填補的新值幾乎退化為固定數值填補法(因為此時)，如2.2.2節中所提及。因此填補效果不好也是可預期的。

即使試圖只依靠權重法計算加權平均值來試圖彌補填補值計算錯誤，也無法解決鄰近點不足的問題。這也是為何無法單純以權重型k鄰近點填補法就可以解決的。由實驗結果可知，在高缺失值比例下，挑選出更具參考性鄰近點的影響力比挑選出更多鄰近點更大。

# 結論與未來方向

本本章分為兩節，5.1節總結本研究，5.2節探討未來可能的研究方向與工作。

## 5.1結論

[5.1、5.2已重寫過]

本研究提出一個概念基於原始k鄰近點填補法的sk-NN imputation演算法，提出一個以距離值作為鄰近差別權重分配。在缺失值比例較低時，此時排序後有效的鄰近點較多，因此針對較近的鄰近點賦予較大的權重值，反之賦予較小的權重值。隨著缺失值比例上升，鄰近點個數的影響力已經不及於挑選適當的鄰近點。本研究再提出一個新的選擇鄰近點機制以改善原始k鄰近點填補法在缺失值比例較高時無法找到有效鄰近點的缺點。

在資料集內的缺失值為完全隨機缺失類型(MCAR)的假設下，本研究所提出的sk-NN imputation演算法填補後所產生天際線的相似度在80%的缺失情形下，都比k鄰近點填補法好。本研究的方法不僅可以在低缺失值比例時與k鄰近點填補法不分軒輊，尤其在高缺失值比例的情況時，仍可以保持一半以上的天際線。若是再加上足夠的k值，與原天際線的相異性更可以控制在20%以內。所以本研究提出的差別權重分配與新的鄰近點選擇機制，確實能夠改善天際線查詢演算法在不完整資料集中找到近似原始的天際線。

## 5.2未來研究方向

在實驗的過程中，我們也注意到當k值很小(例如k=1)且缺失比例低時，在填補缺失值的過程容易觸發對鄰近點採樣的機制。因此我們希望能夠更進一步的去平衡k值與缺失值比例之間的關係，降低在缺失值比例低時啟動採樣機制的頻率，並多利用鄰近點填補缺值。使觸發條件隨著缺失值比例的增加而降低觸發門檻。例如可以利用4.4節提到的 值，針對觸發條件設一個門檻值。在未來的研究方向上，我們認為可以根據不同缺失類型，找出分別適合隨機缺失(MAR)以及完全非隨機缺失類型(MNAR)的填補法。觀察缺失值與其他維度的相關性，改善尋找鄰近點的規則，以提升填補後近似天際線的相似度。

# 參考文獻

[1] A. A. Alwan, H. Ibrahim, N. Udzir, and F. Sidi, “Missing Values Estimation for Skylines in Incomplete Database,” *The International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 66–75, 2018.

[2] S. Borzsony, D. Kossmann, and K. Stocker, “The Skyline Operator,” *Proceedings of the 17th International Conference on Data Engineering*, pp. 421–430, 2001.

[3] S. Deepa Kanmani, E. Kirubakaran, R. E. Blessing Vinoth, and A. S. Ebenezer, “An Effective Imputation Technique for Improving The Performance of Skyline Queries for Incomplete Database,” *Proceedings of the International Conference on Data Science and Communication (IconDSC)*, pp. 1–5, 2019.

[4] G. B. Dehaki, H. Ibrahim, N. I. Udzir, F. Sidi, and A. A. Alwan, “Efficient Skyline Processing Algorithm over Dynamic and Incomplete Database,” *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*, pp. 190–199, 2018.

[5] Y. Gulzar, A. A. Alwan, N. Salleh, I. F. A. Shaikhli, and S. I. M. Alvi, “A Framework for Evaluating Skyline Queries over Incomplete Data,” *Procedia Computer Science*, vol. 94, pp. 191–198, 2016.

[6] Y. Gulzar, A. A. Alwan, and S. Turaev, “Optimizing Skyline Query Processing in Incomplete Data,” *IEEE Access*, vol. 7, pp. 178121–178138, 2019.

[7] C. Hasler and Y. Tille, “Balanced k-Nearest Neighbor Imputation,” *Statistics*, vol. 50, no. 6, pp. 1310–1331, 2016.

[8] J. Huang, J. W. Keung, F. Sarro, Y.-F. Li, Y. T. Yu, W. K. Chan, and H. Sun, “Cross-Validation Based k Nearest Neighbor Imputation for Software Quality Datasets: An Empirical Study,” *Journal of Systems and Software*, vol. 132, pp. 226–252, 2017.

[9] D. W. Joenssen, “Hot Deck Methods for Imputing Missing Data,” *Machine Learning and Data Mining in Pattern Recognition*, vol. 7376, pp. 63–75, 2012.

[10] H. Kang, “The Prevention and Handling of The Missing Data,” *Korean Journal of Anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[11] M. E. Khalefa, M. F. Mokbel, and J. J. Levandoski, “Skyline Query Processing for Incomplete Data,” *Proceedings of the IEEE 24th International Conference on Data Engineering*, pp. 556–565, 2008.

[12] J. Lee, H. Im, and G. You, “Optimizing Skyline Queries over Incomplete Data,” *Information Sciences*, vol. 361, pp. 14–28, 2016.

[13] J. Lee, G. You, S. Hwang, J. Selke, and W.-T. Balke, “Interactive Skyline Queries,” *Information Sciences*, vol. 211, pp. 18–35, 2012.

[14] R. Malarvizhi and D. A. S. Thanamani, “K-Nearest Neighbor in Missing Data Imputation,” *International Journal of Engineering Research and Development*, vol. 5, no. 1, pp. 5–7, 2012.

[15] X. Miao, Y. Gao, G. Chen, and T. Zhang, “k -Dominant Skyline Queries on Incomplete Data,” *Information Sciences*, vol. 367–368, pp. 990–1011, 2016.

[16] X. Miao, Y. Gao, G. Chen, B. Zheng, and H. Cui, “Processing Incomplete k Nearest Neighbor Search,” *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 6, pp. 1349–1363, 2016.

[17] T. A. Myers, “Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data,” *Commun. Methods Meas.*, vol. 5, no. 4, pp. 297–310, 2011.

[18] W. Ren, X. Lian, and K. Ghazinour, “Skyline Queries over Incomplete Data Streams,” *The VLDB Journal*, vol. 28, no. 6, pp. 961–985, 2019.

[19] P. Royston, “Multiple Imputation of Missing Values,” *The Stata Journal*, vol. 4, no. 3, pp. 227–241, 2004.

[20] D. B. Rubin, “Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse,” *Proceedings of the Survey Research Methods Section of the American Statistical Association*, vol. 1, pp. 20–34, 1978.

[21] J. Shao, “Cold Deck and Ratio Imputation,” *Survey Methodology*, vol. 26, no. 1, pp. 79–86, 2000.

[22] B. W. Silverman and M. C. Jones, “An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951),” *International Statistical Review*, vol. 57, no. 3, p. 233, 1989.

[23] G. Tonini, M. Ricerche, S. Scartoni, M. Ricerche, C. Paoli, and M. Ricerche, “Missing Data For Repeated Measures: Single Imputation VS Multiple Imputation,” *Proceedings of PharmaSUG Conference*, p. 10, 2015.

[24] G. Tutz and S. Ramzan, “Improved Methods for The Imputation of Missing Data by Nearest Neighbor Methods,” *Computational Statistics & Data Analysis*, vol. 90, pp. 84–99, 2015.

[25] J. Van Hulse and T. M. Khoshgoftaar, “Incomplete-Case Nearest Neighbor Imputation in Software Measurement Data,” *Information Sciences*, vol. 259, pp. 596–610, 2014.

[26] Y. Wang, Z. Shi, J. Wang, L. Sun, and B. Song, “Skyline Preference Query Based on Massive and Incomplete Dataset,” *IEEE Access*, vol. 5, pp. 3183–3192, 2017.

[27] K. Zhang, H. Gao, X. Han, Z. Cai, and J. Li, “Modeling and Computing Probabilistic Skyline on Incomplete Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 7, pp. 1405–1418, 2019.

[28] S. Zhang, “Nearest Neighbor Selection for Iteratively kNN Imputation,” *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012.

[29] “UCI Machine Learning Repository,” 2013. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.