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七月

**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

天際線演算法為大數據資料分析上最常做為使用者偏好的一種技術之一，但良好的天際線卻又仰賴於輸入資料集的完整性，因此本論文透過基於k鄰近填補法的鄰近點概念上改善資料完整性，本研究所提出的方法盡可能地找到可參考鄰近點以對缺失值填補新值，使得當鄰近點不足或是在缺失值高比率下改以採樣法取無缺失值鄰近點作為填補機制。本研究與k鄰近填補法比較，評測各填補法則以原天際線為基準。實驗結果顯示，本論文方法在低缺失值比率時具有與k鄰近填補法近乎相同，而在缺失值比率20%到70%之間也普遍表現比k鄰近法好30%至50%，在缺失值比率高達80%以上時填補後效果也能高於k鄰近法3到6倍，(最後對本論文做個結論)。

關鍵字：天際線問題，缺失資料，k鄰近填補法，採樣法

# Abstract

The skyline algorithm is one of the most commonly used techniques for user analysis in big data analysis, but a good skyline rely on data completeness of the input data set, this paper make some improvement based on the k-NN imputation. Under the case of small missing rate, proposed method impute the new values for the missing position as k-NN does. If there is insufficient neighbors to be referenced, we sample other neighbors who has non-NaN value. In this study, we compared with k-NN imputation, and benchmark by the original skyline set. The experimental results show that the proposed method has almost the same as the original k-neighbor filling method at a low missing value ratio, and it generally performs 30% to 50% better than the k-neighbor method between 20% and 70% of the missing value ratio. Furthermore, when missing rate reaches above 80%, the accuracy of proposed method is 3 to 6 times better than k-NN imputation.This remainder of this paper proceeds as follows: In Section I, we briefly introduce the skyline problem and dependency on completeness of input data set. Section II provides an overview of the related work, included types of missing values and differences between modern techniques on missing value handling. Further, we described the terminology, problem formulation and analysis. After that, we give proposed algorithm and a measurement of quality between different imputation strategy are illustrated in Section III. This is followed by experiments and analysis in Section IV. Finally, the paper is concluded in Section V.

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

# 目次

[摘要 1](#_Toc45197479)

[Abstract 2](#_Toc45197480)

[目次 3](#_Toc45197481)

[表目次 5](#_Toc45197482)

[圖目次 6](#_Toc45197483)

[第一章 簡介 7](#_Toc45197484)

[第二章 相關研究 8](#_Toc45197485)

[2.1天際線問題概述與完整資料集對天際線問題之影響 8](#_Toc45197486)

[2.2缺失資料類型與缺失值處理技術 8](#_Toc45197487)

[2.2.1資料缺失類型 8](#_Toc45197488)

[2.2.2缺失值的處理技術 9](#_Toc45197489)

[2.3填補法 10](#_Toc45197490)

[2.4 k鄰近填補法 10](#_Toc45197491)

[第三章 問題與方法 12](#_Toc45197492)

[3.1符號定義 12](#_Toc45197493)

[3.2研究動機 14](#_Toc45197494)

[3.3問題定義 14](#_Toc45197495)

[3.4問題分析 14](#_Toc45197496)

[3.5 sk-NN imputation 演算法 16](#_Toc45197497)

[3.6以skyline set作為填補法的表現優劣 18](#_Toc45197498)

[第四章 實驗結果與分析 20](#_Toc45197499)

[4.1實驗環境與資料來源 20](#_Toc45197500)

[4.2實驗一: k值大小對缺失值比例的影響 20](#_Toc45197501)

[4.2.1實驗目的與設計 20](#_Toc45197502)

[4.2.2實驗方法 20](#_Toc45197503)

[4.2.3實驗結果與分析 21](#_Toc45197504)

[4.3實驗二:比較各填補法填補後與原天際線結果之相似程度 22](#_Toc45197505)

[4.3.1實驗目的與設計 22](#_Toc45197506)

[4.3.2 實驗方法 22](#_Toc45197507)

[4.3.3實驗結果與分析 22](#_Toc45197508)

[4.4實驗一與實驗二結論 26](#_Toc45197509)

[第五章 第五章結論與未來方向 27](#_Toc45197510)

[5.1結論 27](#_Toc45197511)

[5.2 未來工作與方向 27](#_Toc45197512)

[參考文獻 28](#_Toc45197513)

# 表目次

[表 三.1 sk-NN imputation演算法符號定義表 13](#_Toc45197514)

[表 四.1 k=1各填補法比較表 23](#_Toc45197515)

[表 四.2 k=5各填補法比較表 24](#_Toc45197516)

[表 四.3 k=13各填補法比較表 25](#_Toc45197517)

# 圖目次

[圖 三.1 NaN-Euclidean distance 15](#_Toc45197518)

[圖 三.2 sk-NN imputation演算法 17](#_Toc45197519)

[圖 三.3 Procedure Impute\_Process() 17](#_Toc45197520)

[圖 四.1 k=1時hit ratio versus miss rate圖 21](#_Toc45197521)

[圖 四.2 k=4時hit ratio versus miss rate圖 21](#_Toc45197522)

[圖 四.3 k=1各填補法比較圖 23](#_Toc45197523)

[圖 四.4 k=5各填補法比較圖 24](#_Toc45197524)

[圖 四.5 k=13各填補法比較圖 25](#_Toc45197525)

# 簡介

現實生活中，在蒐集資料的過程當中難免會面臨到資料不齊全的狀況，可能在蒐集過程中不慎讓內容流失，亦有可能是蒐集到的資料值很難完備，而此一類資料集在資料科學領域中被稱之為「不完整資料集」，即資料當中不全具有值。

在多維度資料集當中，資料難免會因為諸多因素造成資料集內有缺失的情形從而影響結果，而針對缺失值填補的演算法其中又以k鄰近填補法對不同缺失狀況表現較好，然而，過往的k鄰近填補法都是以一個自定義常數k作為對所有缺失值找到該維度鄰近點的方法，但卻少有研究針對不同缺失情形下做個別處理，一般而言，在各個維度當中無論缺失值的分布狀況較如何，一旦採取k鄰近填補法機制填補法，在所有的維度中，缺失個數、某一缺失值與該維度所佔有比例等等，均予以同一機制搜尋k個鄰近點後，並在此方法下所找到的鄰近點也予以相同權重，將值填補回該缺失值欄位。

根據上述之情形，在以往經過傳統k鄰近填補法填補後的效果可能不會那麼好，因此，本研究想觀察在對於缺失值分布不同的情況以及不同缺失情形下，基於原k鄰近填補方法上做改善，藉以達到可以修正考慮某些缺失分布情形下，填補後的值可以更加合理且完善，並提出一個改進的方法。

本論文的章節結構 : 第一章簡介、第二章相關研究、第三章問題與方法、第四章實驗結果與分析以及最後第五章結論與未來方向。

# 相關研究

本論文所牽涉到的相關議題有四個面向 :天際線問題(skyline problem)、缺失資料類型與缺失值處理技術(missing value type and handling)、填補法 (imputation)、k鄰近填補法(k-nearest neighbor imputation)。

## 2.1天際線問題概述與完整資料集對天際線問題之影響

天際線問題主要目的是要找出所有不會被任何一點支配的資料點集合，其決定被支配與否的關鍵是在比較點p的所有維度下的值均不小於或等於點q，稱之為點p支配點q，但由於天際線問題必須比較所有的維度方能決定支配性與否，故一個完整資料集對天際線問題具有舉足輕重的影響力，在缺失值的資料集下欲比較所有維度是一項重大的困難所在[9]。

## 2.2缺失資料類型與缺失值處理技術

本節主要討論範圍包含常見的資料集缺失類型以及缺失值處理的技術。

### 2.2.1資料缺失類型

首先在處理缺失值之前，必須先了解即將面對的缺失類型，才能對不同類型的缺失值施予適當地處裡技術，本節將資料集中常見的缺失值類型做簡單的分類，分別有隨機缺失值類型(Missing at Random，縮寫為MAR)、完全隨機缺失類型(Missing Completely at Random，縮寫為MCAR)以及完全非隨機缺失類型(Missing not at Random，縮寫為MNAR)共三種 :

隨機缺失值類型 :

在分析當中控制了某些變數，某一變數缺失的機率與觀察到的數據本身有關而與未觀測到的數據無關，則此一類型缺失成為隨機缺失。例如:

男性比較不願意填寫關於沮喪或失敗的調查問卷(觀察到的)，並非因為該族群沮喪與否有關(非觀察值)，因此男性在該類型問卷當中會比較容易具有缺失值以至於在資料表現上看起來男性族群所表現出來似乎比較不會沮喪，而事實上是因為該族群相較女性更不願意填問卷而顯示出來的資訊，此一類型缺失方式就被歸類為隨機缺失值。

完全隨機缺失類型 :

假設有一變數有缺失數據，若該缺失數據的機率與該維度本身的值或該數據中任何其他變數的值都完全獨立並沒有任何相關性，則可稱此一缺失類型為完全隨機缺失，本論文實驗中採取資料集後所模擬的隨機缺失模式則屬於此種，以確保缺失值並不會與其他因素與變數有任何相關性，作為填補效果優劣之依據。例如:回答錯誤、忘記填值、資料遺失……等等。

最後，如果缺失情形均不屬於上述兩者，則被歸屬於完全非隨機缺失類型，此一類型的缺失資料值與某維度具有一定程度的關係或傾向，換句話說即與該缺失值有很高的相關性，屬於此一類型缺失資料會表現某一種資料特性，故又被稱作不可忽略缺失類型。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，造成該資料缺失情況進而導致影響資料集真實性。

缺失資料若屬於MCAR或者是控制相關變數下少數缺失的MAR類型時，對該缺失情形是可以被納入考量並接受缺失值相關技術處裡方式解決缺失問題，但若缺失資料屬於MNAR時，因缺失類型為非隨機缺失，故此情形完全不可忽略且不建議受以任何缺失處理機制，以防止對資料集判斷與誤導。

### 2.2.2缺失值的處理技術

根據以往經驗當中，為了使具有缺失資料集具有完整性，一般在處理缺失資料時不外乎最常見的分別為丟棄法 (dropout) 與填補法 (imputation) 兩種主要處裡技術 :

丟棄法技術其中包含了刪除資料列以及刪除特徵欄位兩種。如果一個資料列裡面的欄位當中有觀察到具有缺失值，則將整筆資料列直接刪除，此種刪除方式可以在整體資料集中含有少量缺失值時採用移除具有該缺失值之資料列，以達到完整性資料集之目的，但此處理技術並不適用於本實驗中模擬的資料集缺失類型。其中主要的原因是，在模擬缺失值佔資料集整體比例越來越高且完全隨機缺失類型模型下，刪除資料列的方法會讓原資料集當中剩下堪用的資料筆數非常少，也意味著可參考點也很少，致使最終填補法效果不彰。

若整體資料集的某一維度或某一特徵欄位缺失筆數的數量太多，甚至遠多於其他特徵欄位缺失狀況時，則可選擇直接放棄該特徵欄位。遇缺失值時無論就上述採用刪除資料列或者是刪除特徵欄位，雖然可能不需要面對因具有缺失值所造成的資料不完整性，然而在另一方面，無論是丟棄資料列或是丟棄整個特徵欄位，仍然需要面臨因丟棄法所產生的額外問題 : 喪失資料表現特徵性質以及資料筆數來源不足。

填補法最大的好處則是可以用簡單的統計方法來計算最終需要被填入的填補值，最簡易補值的方法有對所有缺失值欄位只賦予單一固定數值，如 : 眾數、平均值、中位數、極大值、極小值......等等。在不影響資料集分佈模型下，採取單一固定數值填補法比丟棄法來得合理。考量不影響整體輸入資料集分佈、統計模型以及避免遺失了資料點可能在輸入資料集所表現的特性下，參考資料集中非缺失的數值來作為缺失值填補依據比較具有可參考性且預期會有比較好的效果，本論文所採用的研究方向亦是屬於填補法的技術來處理缺失值。

## 2.3填補法

眾多填補法當中又可以分為單一填補法(single imputation)與多重填補法(multiple imputation)兩者 :

單一填補法主要目的是填補某一特定欄位中遺失的值。即使計算出可能具有不只一種可參考解，也會依照不同演算法的機制來挑選其中一個值作為最後填補該缺失欄位。當缺失資料在整體資料集佔有很少的數量時，單一填補法則不失為一種簡易卻又實用的方式。且又因單一填補法計算容易，所以也最常被用來當作填補缺失值的首選方法。另外，單一填補法在面臨一定程度的缺失情形時，可能會面臨到一個嚴重的問題。一旦被填補進缺失欄位後，便會被當作真正的資料值而無法再分辨出原始資料集的值與被填補值之間的真偽。此一疑慮甚至可能造成日後分析結果時被誤導。相較於單一填補法，多重填補法更著重於分析與解決問題上。多重填補法仰賴於資料集上模擬分佈模型，在遇到缺失資料時會根據模型的分佈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有大量的計算需求。

## 2.4 k鄰近填補法

k-nearest neighbor imputation (往後簡稱kNN 或 k鄰近填補法)至今仍是一個很實用且頗為主流的填補缺失值方法，k鄰近填補法的核心概念即在多維度空間資料集中某參考點p，藉著找尋其他k個與p點相鄰最接近的點在p點缺失的維度上作為填補點p的參考值，k鄰近填補法可以被應用於的資料種類不僅僅是連續型資料(continuous data)上，也可以被應用於離散型資料(discrete data)、有序型資料(ordinal)甚至是分類型資料(categorical data)，幾乎均適用在各式各樣的資料種類。k鄰近填補法為上述所有補值方法當中，最為常見也是普遍被認為效果比較好的補值法，以下概述其優、缺點 :

一般而言，kNN會比單一填補值，如:平均數、中位數、極值、或是眾數等填補法還要來得準確許多，原因是該方法會同時參照其他該缺失值相鄰點去預測其應該原有的合理值，尤其當最終必須找出skyline set的點時，單一值填補法對於尋找skyline point set不會有更好的幫助，反而會因此增加許多不必要的計算在比較相同維度上具有相同值，因缺失而被填補回去的相同值。類似這樣的問題在無論考慮單一維度或整體資料集缺失值密度，當missing rate愈趨增加時，其餘具有完整資料值的點也會越來越少。

縱然如此，kNN 也存在著某些潛在的缺點 :

身為非監督式填補法會參考其餘非缺失值，換句話說，此類型填補法並不會對資料內容做任何過濾或預處理，對於既有資料敏感度距甚，很容易受限於輸入資料集當中非缺失值的內容是否為noise value導致結果很有可能無法精準地將資料集內的原缺失部分填補回去。

另外也需考量的點是，現今實際層面常常處理資料量大而且高維度的巨量資料，k鄰近填補法必須先行儲存整個欲參考的鄰近點於記憶體中，並且逐一計算求出各點兩兩之距離，可想而知其計算量所費不貲。對某一個點p來說，在找尋其相鄰近點集合過程裡，其所有相鄰近點之中從第一個至第k個相鄰近點都具有相同的權重值，但是在計算點p與資料集當中其他點q的與點p距離同時，並且對所有距離值先行排序，如此距離值的排序行為已經代表著點p與點q的距離上的意義，我們應該利用這樣的排序關係設計出一個合理的權重比值，以表示被參考的其他點q在填補的演算過程中應該被參考的價值有多少分量，最常見也最簡易的權重表示方式即為兩點距離的倒數。

# 問題與方法

本章節會依照結構順序分別先說明敘述符號定義與研究動機，然後是問題定義以及問題分析，接著探討本論文所提出的sk-NN imputation演算法以採樣的機制填補缺失值，本章的最後一節再闡述為何決定以skyline set作為填補法的表現優劣。

## 3.1符號定義

這節主要目的是描述並說明所有本論文提出的演算法內會運用到的符號與定義，表3.1則為本論文演算法的符號定義表。已知一個不完整資料集 含有若干個缺失值與屬於自然數的常數 作為限制可參考鄰近點的上限個數，兩者皆為演算法的輸入參數。其中n與m分別為 的資料列個數與 的維度個數。 表示 中第 筆資料列，且 為該資料列 於維度 的值。

為不完整資料集 中記錄任兩資料列 與 之間距離的距離矩陣，且其中 為任兩資料列 與 之歐氏距離。 是一個記錄資料點 在每一個維度 的值 相對其他任意相異資料點 的權重矩陣， 則表示兩資料點 與 之間的權重值。 為一個包含 與 兩種類別的類別型集合，其目的是用來決定計算某一缺失值的被參考維度值 的權重值該如何分配。若 設為 ，則所有權重值 均被設為1 ; 若 設為 ，則 會依照 中任兩資料列 與 之間距離 的倒數作為該權重值。

NN list蒐集已排序串列 。每一個 內記錄所有資料點 按照與其相距的歐氏距離 由小至大排序好後，記錄該順序所對應鄰近點的index。 代表 所有相鄰近點中，經計算歐氏距離後 不為0的且與 第j個最接近資料點的index。

最後說明填補過程中會使用到的符號，若輸入資料集 當中某 為缺失值，則經過本論文填補法計算過後會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，去記錄某 最鄰近k個參考點在發現該缺失值所在的維度 是否為NaN。若被參考的鄰近點在與該缺失值同一維度值也為NaN，則標註為True，否則為False。最終，演算法輸出一個已被填補所有缺失值後的資料集 。

表 三.1 sk-NN imputation演算法符號定義表

|  |  |
| --- | --- |
| notation | description |
| n | number of data instances within an input data set |
| m | dimensionality of input data set |
| Incomplete data set | an incomplete data set with certain missing values of size |
| Incomplete data set , = | an incomplete data set with certain missing values of size |
| , k | specified constant k to determine number of neighbors |
|  | data value of the data instance,  at index row and column, contains NaN value if there is missing |
|  | data instance at row, composed by |
|  | a distance matrix, which is symmetric and records the distance between two pair wised data instances |
|  | distance between and , denoted as , |
|  | a weight matrix record to pairwise any and , i |
| = | weighting value with respect to between and |
|  | type of weighting, determine mechanism of the way assigning weight in weight matrix |
| NN list | record a sorted nearest neighbor list between all pairs of and |
|  | a sorted list at index i in NN list to keep all neighbors of , where |
|  | nearest neighbor of element at index j in sorted list |
|  | value been imputed after imputation at index (i, j) only if is missing or value NaN before imputation, otherwise = |
| mask | a boolean type array of size k to label the candidate at column j having non-NaN value |
| Imputed data set | a completed data set with missing value all imputed, of size |

## 3.2研究動機

搜尋skyline set時會需要去比較所有特徵欄位的值，也就是說每一筆資料列的每一個特徵欄位都必須有值存在，因此，具有完整的資料集是 skyline problem中必要的先決條件。但在實際情況下，期待蒐集每一個特徵欄位都不能有缺失值似乎是不切實際地，現實生活當中具有太多不可抗拒因素使我們在取得資料集的過程難免會遇到欄位裡的值會有缺失，一般解決缺失值的方法不外乎刪除與丟棄法，但此一類型方式最大的缺點是當missing rate不高的時候，很容易遺失該筆表現在整體資料集所具有的資料特徵，此方法在missing rate低於某一門檻值時，本論文不建議貿然刪除該資料欄位，反之本論文方法為盡可能地保留具有資料特性的機會，缺失值則可以0、null或是NaN作為該缺失欄位的表示法。

而在眾多填補法，其中又以k鄰近填補法在大多數情況下填補後的結果普遍表現不錯，原因是k鄰近填補法會參考著其鄰近鄰居的再取算術平均值，使得其所填補後的值以相較於其他只以單一數值填補法更具有參考價值。

## 3.3問題定義

在不完整資料集中，如何改善k鄰近填補法填補缺失值，使填補後的完整資料集具有最近似skyline set的內容?

## 3.4問題分析

所有計算距離公式中，歐式距離的計算方法是採資料集中兩兩資料點相對應維度的差平方再取平方根，若是遇到相同維度下至少一個具有NaN值，則在計算歐式距離時並不會採計具有NaN值的數值差的平方和，此計算方式是最廣為主流的算法，在此論文中並不打算更改計算方式，但由此計算方式可看出潛在問題 :

具有缺失值的兩資料點計算可能會誤導兩資料點之間的實際距離。



圖 三.1 NaN-Euclidean distance

圖3.1所示A、B、C三點座標分別為(1, 1)、(2, 8)、(3, 3)，在沒有任何NaN值下，以往按照傳統歐氏距離的計算AB距離 應該為 且 應該為 ，故 ，若將點A中的y座標設為NaN表示缺失值時，則含有NaN值後的歐氏距離會因為相對座標不全有相對應維度值而不列入計算中，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目的為了避免誤算不納入NaN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鄰近填補法在具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。

k鄰近填補法的另一個問題在於，當缺失值越高的時候，當k值很大的同時也意味著鄰近點仍存有non-NaN value的機會並不大，而此時k鄰近點填補法則在無法找到滿足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non-NaN值很稀少時更為嚴峻，導致k鄰近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(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，更容易形成有如該欄位直接被刪除一樣無意義地比較的結果。

鑒於以上分析後，本論文的核心算法在解決缺失值比例不高的狀況下填補效果與k鄰近演算法相近，並且在缺失值比例升高的時候也能夠一定程度地解決當k鄰近點不足時適時地補足，對缺失值所在該維度其餘non-NaN值做採樣後再取平均值，目的致力於不讓NaN值的距離計算影響填補後找尋天際線的結果與原天際線相去甚多。

## 3.5 sk-NN imputation 演算法

|  |
| --- |
| Algorithm sk-NN imputation () {  Input : incomplete data set , constant k, weight type  Output : complete data set  Method:  step 1. load incomplete data set  step 2. make a copy of , complete data set to be imputed later  step 3. initialize all values of distance matrix into zero  3-1. **for** each row in distance matrix **do**  3-2. **for** each column in distance matrix **do**  3-3. = euclidean distance of pairwise data samples  3-4. **return** distance matrix  step 4. initialize all values of weight matrix into zero  4-1. **for** each row in weight matrix **do**  4-2. **for** each column in weight matrix **do**  4-3. according to weight type , alter the weighting mechanism  =  4-4. **return** weight matrix  step 5. establish a nearest neighbor list to store all nearest neighbors with respect to certain  5-1. initialize an empty NN list, with size n  5-2. **for** each in NN list **do**  5-3. check each in  5-4. retrieve all between any pair of the and from  5-5. sorting in ascending order, keep track of corresponding  5-6. **repeat** from step 5-6 to step 5-9  5-7. if 0 (assignment)  5-8. append into  5-9. **until** size of or all inserted into  5-10. update  5-11. **return** NN list  step 6. search all missing values among data set, then impute new value back into the missing position  6-1. **for** each row i in **do**  6-2. **for** each column j in **do**  6-3. **if** missing value or value NaN is found  6-4. Impute\_Process(i, j, k)  step 7. **return** an imputed complete data set  } |

圖 三.2 sk-NN imputation演算法

|  |
| --- |
| Procedure Impute\_Process(i, j, k) {  Input : row index i, column index j, nearest neighbor k  Output : imputed value at index (i, j)  step 1. create a mask array with size k and initialize all elements to False  step 2. find all elements in with respect to non-NaN in  2-1. **for** each index from 0 to k **do**  2-2. r  2-3. **if** is missing or value NaN  2-4. mask[r] True  2-5. **else**  2-6. mask[r] False  step 3. retrieve values in which index in mask array assigned to False  3-1. **for** each index in mask **do**  3-2. **if** all elements in mask array are not all True  3-3. retrieve all value in  3-4. **else**  3-5. reset all elements in mask array to False  3-6. randomly sample the rest of non-NaN value at column j  3-7. evaluate mean or weighted mean of retrieved value  3-8. assign imputed value to  step 4. **return** imputed value  } |

圖 三.3 Procedure Impute\_Process()

## 3.6以skyline set作為填補法的表現優劣

由於最終目的是為了看出對skyline set結果的填補效果，因此本論文設計了一套以skyline set作為評斷各填補方法在缺失資料集的填補效果，所採用的基本概念為對原始skyline set 的 Hamming distance

Hamming distance 主要是用在計算兩個字串相對應的位置不同字符的個數，換句話說，將一個字串變換成另外一個字串所需要替換字符的總個數，即為Hamming distance，本論文中所使用的為Hamming distance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。例如 :

兩等長二進位字串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hamming distance為2，同理，toned 與 roses 之間的hamming distance為3，以此類推。本論文所採用的是對集合上的Hamming distance概念，換句話說，集合A必須插入或刪除多少元素才能使兩集合相同，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. 兩集合是否相同的充分必要條件為兩集合是否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是否相同。

在判斷skyline set 與新skyline set 的相似程度計算方法如下:

* 1. 集合skyline set中相異元素個數為該集合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 中沒有的元素個數以及 之中有的元素但 中沒有的元素的個數總和，稱為miss count，本論文也定義為Set Hamming distance。

1. hit ratio =

例子:

original skyline : size為5

estimated-1 skyline : ，size為7

estimated-2 skyline : ，size為4

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

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

由上述例子可知， size越大並不能保證hit ratio一定越好，所以猜得數量多不如猜的精準。天際線集合為一個蒐集滿足天際線條件各點的集合，故經填補後所找到的天際線集合如果與原天際線集合越香相似，則判斷該填補法對天際線所填補效果越好，本論文用上述方法來評斷各填補法缺失資料集填補後針對天際線集合作為填補效果優劣之依據。

# 實驗結果與分析

本章節結構依序先說明整體實驗環境與資料來源，接著實驗一為鄰近k值的大小在不完整資料集當中對不同缺失值程度的影響。實驗二為比較傳統k鄰近演算法、權重型k鄰近演算法與本論文所提出的sk-NN imputation填補法交叉分別在參考不同鄰近k值與不同缺失程度的資料集下，以與原skyline set比較相似度作為衡量各個填補法對缺失資料的效果依據。

## 4.1實驗環境與資料來源

本實驗總共有兩支程式分別為計算填補過程的Impute\_Process()以及找出缺失值並決定填補位置的主程式sk-NN Imputation。開發環境主要使用的程式語言為Python的3.8.2版本，所撰寫的整合開發環境工具則為Anaconda，虛擬環境架設在Jupyter Lab與Notebook內，並引用包含處理資料流的pandas套件、數學與矩陣函式相關numpy套件、以及機器學習與資料挖掘相關的sklearn以及數據視覺化所需要的matplotlib套件。本實驗所使用的輸入資料集為UCI Machine Learning Repository中純數值資料類型的資料集，資料集名稱為Bike Sharing dataset、Real estate valuation dataset、Real-time Election Results Portugal 2019 dataset。

## 4.2實驗一: k值大小對缺失值比例的影響

### 4.2.1實驗目的與設計

本實驗目的是在探討參考鄰近點k值越大，對填補後的資料集找skyline set是否會有更好的填補效果。

### 4.2.2實驗方法

本實驗方法為將同一缺失資料集分別從k=1測試，觀察相同k值分別於不同缺失比例下情況之下，隨著k值增加是否可以得到更佳的準確度。

衡量此實驗效果，本論文採用原完整資料集 complete data set中所得出 skyline set 的點作為最終填補效果的依據。

### 4.2.3實驗結果與分析

根據圖4.1實驗一顯示結果，隨著缺失比率在資料集當中增加，k鄰近填補法的準確率並沒有因為找尋更多的鄰近點數量改善填補效果。

觀察圖4.2可知，k鄰近填補法只著重在將鄰近參考點的數量逐漸地增加，但因為缺失比例也逐漸增加，再配合k鄰近填補法中對可供參考點不足從缺的機制，致使即使計算鄰近點值的平均也會逐漸失效，這同時也意味著可供參考點數量以及參考值之可靠性嚴重不足。



圖 四.1 k=1時hit ratio versus miss rate圖



圖 四.2 k=4時hit ratio versus miss rate圖

## 4.3實驗二:比較各填補法填補後與原天際線結果之相似程度

### 4.3.1實驗目的與設計

在同一k值下，在不同missing rate程度下，本論文方法與k鄰近填補法即所能夠找回近似skyline set的程度。

### 4.3.2 實驗方法

本實驗所使用的資料集k值最大範圍可以到17，故分別取三種不同k值分別做三次比較，以觀察k鄰近填補法、權重型k鄰近法以及本論文方法填補後的值所能夠找回原skyline set的程度。x軸為缺失值佔整體資料集當中的比例，y軸為最填補所有缺失值之後，再分別跑同一支尋找skyline set的程式，並與缺失前的原skyline set做比較計算出相似程度。若越接近原skyline set則y軸的值越接近1.0。

### 4.3.3實驗結果與分析

實驗二結果顯示出，本實驗方法雖然在某些k值不大情形下準確度會略差，如表4.2與圖4.4中，當k=5且缺失比率0.3到0.55間，本論文方法準確度有下降，此時k鄰近填補法有機會擁有較好的填補效果是因為缺失程度不高下，k鄰近填補法還能夠以足夠的k與鄰近點計算平均後填回。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比例不高時會比較容易啟動，但隨著缺失比率增加下，k值的增加所帶來的益處會越不明顯，而此時本論文中採樣的方式反而可以起到更大機會能有效的找到鄰近參考點。

從表4.1與4.2中可以看出k鄰近填補法大約從缺失比率30%時準確度就開始急遽下降，雖然中間可能有小幅度的上升，但可看出下降程度在75%以後下降幅度又更加嚴重，表4.3更是提前在缺失值達55%時就開始大幅下滑，這都顯示出一個現象，即k鄰近填補法所參考鄰近點的機制，對於缺失值很多的情形下抗性不高。

表 四.1 k=1各填補法比較表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **缺失比例** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** |
| **kNN** | **0.8** | **0.7** | **0.538** | **0.6** | **0.421** | **0.5** | **0.384** | **0.25** | **0.1** |
| **weighted kNN** | **0.8** | **0.7** | **0.538** | **0.6** | **0.421** | **0.5** | **0.285** | **0.25** | **0.1** |
| **proposed method** | **0.8** | **0.8** | **0.818** | **0.909** | **0.636** | **0.75** | **0.7** | **0.636** | **0.545** |

圖 四.3 k=1各填補法比較圖

表 四.2 k=5各填補法比較表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **缺失比例** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** |
| **kNN** | **0.9** | **0.666** | **0.666** | **1** | **0.625** | **0.533** | **0.6** | **0.454** | **0.181** |
| **weighted kNN** | **0.9** | **0.625** | **0.6** | **1** | **0.529** | **0.692** | **0.666** | **0.1** | **0.25** |
| **proposed method** | **0.9** | **0.833** | **0.909** | **0.818** | **0.5** | **0.8** | **0.6** | **0.5** | **0.363** |

圖 四.4 k=5各填補法比較圖

表 四.3 k=13各填補法比較表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **缺失比例** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** |
| **kNN** | **0.9** | **1** | **0.75** | **0.333** | **0.727** | **0.583** | **0.615** | **0.384** | **0.181** |
| **weighted kNN** | **0.9** | **0.833** | **0.818** | **0.4** | **0.277** | **0.529** | **0.428** | **0.352** | **0.181** |
| **proposed method** | **0.9** | **1** | **0.833** | **0.454** | **0.818** | **0.7** | **0.727** | **0.75** | **0.7** |

圖 四.5 k=13各填補法比較圖

## 4.4實驗一與實驗二結論

由實驗二中的圖可以知道無論是k鄰近填補法或是權重型k鄰近填補法，在k值分別為1、5與13時，隨著missing rate在整體資料集中升高，亦無法有很好的填補效果，經分析其原因有二 :

1. 在考慮最鄰近的k值作為參考該維度的值時，可能會遇到不足k個不為NaN的值，而傳統k鄰近填補法在遇到此種情況時，會選擇從缺不補，使得剩下的不足k個被參考鄰近點該維度值之權重無形中上升。

當缺失比例越高的時候，此狀況也就越趨明顯，最後k鄰近填補法所填補的新值雖然為平均，但也幾乎被簡化為單一值填補法的效果，最終效果如同只填補均值、眾數、或最大最小數結果一樣。

1. 遇到NaN值計算距離的機制，當兩兩資料點計算出距離時，若兩個資料點在相對應維度上其中一點至少有一數值為NaN，則在計算歐氏距離時該維度值之間的差平方並不會被納入歐氏距離的計算式中，使該維度對距離上的影響力被無視，也是k鄰近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。

上述兩個原因亦無法藉權重法求加權平均數來彌補此一現象的缺陷，因此亦可看出即使採用權重型k鄰近填補法也不會有太好的填補效果，從此可以得知，挑選可參考性鄰近點的值在高缺失值比例下，其影響力遠比參考更多鄰近點來的大。

# 第五章結論與未來方向

本章節分為兩部分，第一節簡單總結skyline set對資料完整性的仰賴程度，與各處理缺失值技術差異性，並相較於k鄰近演算法與本論文所提出的方法中對於不同缺失情形下填補的效果，第二節探討未來可探索解的方向與工作。

## 5.1結論

在面臨需要考量使用者喜好的應用問題時，例如根據使用者使用習慣或評分的推薦系統，每個特徵欄位均需要被考慮的情形下，通常以 skyline algorithm 是最為常見的，而 skyline algorithm 又仰賴於輸入資料集的完整性(completeness)，但若是無法順利取得完整的資料集、或是具有缺失值使得所能取得完整資料筆數不足，在大部分情況下某一缺失值是無法與該特徵欄位其他非缺失值相比較的，此時 skyline algorithm 就無法針對每一筆資料的所有特徵欄位做比較，導致最終無法得出skyline set points，有鑑於缺失值存在的問題，本論文提出的方法基於原k鄰近填補法之上同時考慮缺失情形與計算距離方式進而決定一個權重向量weight vector w對其k個鄰近點分配以新的權重值來做為原缺失值欄位被填補的新值，如此一來不僅不會因刪除而導致喪失關鍵的資料特徵，也可以填補比原本k鄰近填補法更具有參考性的值，因此填補後對於 skyline set 的影響也是本論文所關注的重點之一，其中也會對不同填補法所跑出來的 skyline set points 結果做評估。

本論文方法係鑒於k鄰近填補法概念當中的考量鄰近點的基礎之上，與傳統k鄰近填補法最大的不同是在於傳統k鄰近填補法對於某缺失值其考量所有的鄰近點在該特徵欄位上只單純取算術平均後以填補新值，而本論文所提出的方法則是對其周邊鄰近點根據缺失程度取加權平均，以此來改善原方法

本論文方法在 missing rate介於 20% ~ 80% 所填補值對於最終 skyline set 的結果均具有不錯的填補效果，尤其是當該欄位缺失很集中或是缺失比例偏高的狀況下，其所填補後結果以 hamming distance 測量出來的精準度都不錯，表示本論文方法在填回去 skyline set points 中具有一定的效果。

## 5.2 未來工作與方向

若是能夠透過資料集提供的資訊，綜合考量之後挑選出最有可能成為 skyline set points 的因素，考慮 skyline 與 non-skyline 佔其鄰近點比例也可能會對最終填補值的準確程度造成若干程度的影響，畢竟如果考慮的 skyline 的值越多，則具有該缺失值的點與在特徵 skyline set points 的相似程度可能越大。

# 參考文獻

[1] A. A. Alwan, H. Ibrahim, N. Udzir, and F. Sidi, “Missing Values Estimation for Skylines in Incomplete Database,” *Int. Arab J. Inf. Technol.*, vol. 15, no. 1, pp. 66–75, 2018.

[2] A. Choudhury and M. R. Kosorok, “Missing Data Imputation for Classification Problems,” *arXiv preprint arXiv:2002.10709*, 2020.

[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] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

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