南　華　大　學

資訊管理學系

碩士論文

應用屬性值切割與基因分群技術以推估遺漏值

Applying Attribute Values Partitioning and GA Clustering Technique for Estimating Missing Values

研 究 生：吳文盛

指導教授：邱宏彬

中華民國 九十八 年 六 月

南　華　大　學

資訊管理學系

碩 士 學 位 論 文

應用屬性值切割與基因分群技術以推估遺漏值

研究生：吳之盛

經考試合格特此證明

口試委員：

李翔詔

指導教授：

系主任(所長)：

口試日期：中華民國 98 年 6 月 20 日

南華大學資訊管理學系碩士論文著作財產權同意書

立書人：吳文盛 之碩士畢業論文

中文題目：應用屬性值切割與基因分群技術以推估遺漏值

英文題目：Applying Attribute Values Partitioning and GA Clustering Technique for Estimating Missing Values

指導教授：[illegible] 博士

學生與指導老師就本篇論文內容及資料其著作財產權歸屬如下：

☑ 共同享有著作權

☐ 共同享有著作權，學生願「拋棄」著作財產權

☐ 學生獨自享有著作財產權

學　　生：吳文盛 （請親自簽名）

指導老師：[illegible] （請親自簽名）

中　華　民　國　98 年　6 月　20 月

南華大學碩士班研究生

論文指導教授推薦函

資訊管理 系碩士班 吳文盛 君所提之論文

應用屬性值切割與基因分群技術以推估遺漏值

係由本人指導撰述，同意提付審查。

指導教授 邱宏彬

98 年 6 月 20 日

誌　　謝

兩年的碩士生涯隨著論文的結束而落幕，往事歷歷猶在眼前，能夠順利的完成論文與學位，最感謝的是我的指導教授-邱宏彬老師，其細心教導與不厭其煩的給予學術上的指導和建議，讓我獲益良多；老師在做人處世與心靈層面上的提點更是我在碩士生活中最寶貴的資產，也會永遠謹記在心；另外感謝李翔詣老師與謝昆霖老師對於論文的寶貴意見與指導，讓論文疏漏之處能夠加以修正，讓論文更趨完整。也感謝在兩年碩士生活中所有教導我的老師們，謝謝您們的教導與解惑。

身在每天充滿歡笑的 204 實驗室，也讓我的研究壓力減輕不少；感謝伊汝姐在求學中的幫忙與協助，感謝意純在精神上的支持、照顧與陪伴，讓我的研究之路不孤單。感謝初進研究所時學長姐建源、育弘、宣均等人給予我的幫忙，謝謝我的研究夥伴們花君、錦忠、佳蓉、泡泡等人和學弟妹宜德、美秀、威志、山田、子祥等多位好友的幫忙，感謝我的所有好友們給予我永難忘懷的回憶。

最後要感謝我最親愛、也是最偉大的家人，謝謝您們在我的求學路上給予無限支持與鼓勵。謝謝爸爸、媽媽、奶奶、姑姑與姐姐對我的照顧與關懷，讓我能夠順利完成學業，謹以此論文獻給我最愛的家人。

吳文盛 謹誌於 嘉義

南華大學資管所

民國九十八年六月

應用屬性值切割與基因分群技術以推估遺漏值

學生：吳文盛　　指導教授：邱宏彬

南華大學　資訊管理學系碩士班

摘　　要

資料探勘是由大量資料中挖掘出隱藏知識的重要技術，目前企業或政府各方面決策幾乎是以歷史資料探勘結果分析為基礎，故資料庫的完整性則十分的重要。若是資料庫中出現過多的遺漏值，則容易影響資料分析結果的有效性。我們以群集分析為基礎來建立一個遺漏值推估模組，將物以類聚、群內同質、群間異質的特性應用在遺漏值推估上。再利用屬性值切割法來找出屬性之間的關聯，讓分群後的資料關聯與特性更為緊密；另外基因演算法具備隨機多點搜尋與演化過程的特性，可經由不斷演化找出較佳的分群結果。所以本研究嘗試結合屬性值切割法與基因分群技術，來進行遺漏值的推估，讓使用者可以在使用資料探勘方法時仍可保有最大的資訊量，期望探勘出的結果更具意義。本研究將此推估模式應用到四個真實資料集上，以驗證本研究方法之可行性與推估效能。

關鍵詞：資料探勘、遺漏值、群集分析、屬性值切割法、基因演算法。

Applying Attribute Values Partitioning and GA Clustering Technique for Estimating Missing Values

Student：Wen-Sheng Wu　　Advisors：Dr. Hung-Pin Chiu

Department of Information Management
The M.I.M. Program
Nan-Hua University

ABSTRACT

Data mining is a vitally important technique to uncover hidden information from a set of raw data. The managers can exploit the mining results to make effective decisions. However, missing data significantly distort data mining results. Therefore, data preprocessing of missing values is very critical in successful data mining. Data clustering techniques is the partitioning of a dataset into subsets so that the data in each subset share common pattern. The shared pattern can be utilized to estimate the missing values. In this study, we propose an attribute values partitioning technique to preserve the relationships between attributes for estimating missing values. In addition, genetic algorithm is a powerful population-based stochastic search process for finding the robust clustering result. Therefore, we also propose a genetic clustering-based approach to estimate the missing data. Furthermore, we integrate the attribute values partitioning with the genetic clustering techniques to improve the estimation performance. Effectiveness of the proposed approaches is demonstrated on four datasets for four different rates of missing data. The empirical evaluation shows the integrated missing data processing approach provides competitive results or performs well compared with the existing methods.

Keywords: data mining, missing value estimation, clustering analysis, attribute values partitioning, genetic clustering algorithms.

目錄

表目錄

圖目錄

第一章 緒論

本章將描述本研究的研究背景與動機、研究目的、研究程序、論文架構分佈。

第一節 研究背景與動機

壹、 研究背景

隨著網際網路的發達和資訊科技使用的普及化，讓使用者在資料和資訊的獲取更為方便與容易，而在這個資料量與資訊量超載的環境中，如何快速且有效率的取得所需的資料與資訊，並且能夠轉化為自身的知識，更是各界所關心的議題。

以企業經營管理的角度來看，企業為了永續經營發展，並且能夠創造獲利。必須持續收集大量有關企業營運相關的資料與資訊，例如每位會員的銷售記錄、產品每個月份的銷售記錄等等，都蘊藏著許多足以幫助企業決策者、管理者決定管理決策與公司營運方針的重要依據。

而資料探勘技術為近年來常被企業界應用於提昇企業整體的營運績效與加強產業之間競爭優勢上，其特色為可從龐大或不同平台的資料庫中尋找出有參考價值的資料，再經由人工智慧或統計分析等方法，從大量的資料中挖掘有利於決策參考的資訊與知識，並且能夠依照企業問題類型不同的需求建立不同的決策模型。

根據資料庫內知識發掘（Knowledge Discovery in Database，簡稱KDD）的過程[19]可以發現，首先瞭解所要應用的領域、熟悉其相關背景知識，接著建立目標資料集，並專注所選擇之資料子集；再從目的

資料中作前置處理，去除錯誤或不一致的資料；然後做資料簡化與轉換工作；再經由資料探勘的技術程序成為樣式、進行迴歸分析或找出分類型態；最後經過解釋與評估成為有用的知識。由此可知，在進行資料探勘時必須先進行資料前置處理，否則就可能影響到之後的探勘分析結果。若企業內關鍵性的資料裡有遺漏值的產生，而造成資料探勘結果的不準確，偏離真實狀況的情形將會提供錯誤的決策參考，也讓整個探勘過程失去意義。

貳、 研究動機

由於電腦與網路科技的技術快速發展，使我們透過資料庫可以快速大量地收集各式各樣的資料，並將資料經由統計分析轉化成有用的資訊與知識，因此其資料的完整性就顯得十分重要。而遺漏值的問題就是資料前置處理中最重要的一個議題，高品質的資料探勘結果，其主要因素往往是因為高品質與正確的資料所造成，所以如何有效的推估遺漏值問題即為本文的主要研究方向。

在過去遺漏值議題的相關研究中，傳統的處理方式如表1-1所示。例如只採取忽略並刪除遺失值的作法會使得資料中有用的資訊無法突顯出來，影響到分析的效果；而不當猜測遺失值的作法更可能造成探勘結果的偏差，例如以屬性平均值來填補遺漏值是最簡易的方法，但並未利用到屬性之間的關聯性。因此，藉由保留屬性之間的關係以獲得最佳的填補值便成為一種常用的務實方法。

而群集分析是最常用來當作遺漏值的推估基礎，利用其群間同質、群間異質的特性，將資料集中特徵相符的資料歸類成群。再依照

其群內的平均值回填到遺漏值中，分群的品質好壞往往影響著遺漏值推估的正確率，啟發式演算法(Heuristics algorithms)是近年來最常被應用於尋找出最佳的分群結果，啟發式演算法在搜尋的過程中，可以知道目前狀態距離初始狀態多遠，同時可以藉由啟發函數估計目前狀態和目標狀態還有多少距離，進而增進搜尋的效率。其方法具有不斷修正與尋找最佳化結果的特性，例如基因演算法(Genetic algorithms)[14]、蟻群最佳化演算法(Ant Colony Optimization)[21]與粒子群最佳化演算法(Particle swarm optimization)[5][7]等等。

不過當資料集龐大或複雜時，就可能讓運算的時間複雜度提高，演算法的整體執行效能就會下降。所以建立精確度高、估計方法簡便、運算效能更佳與運算速度更快的遺漏值推估方法則成為許多研究者所努力的目標。

表1-1 傳統遺漏值處理方式[7]

	特 色	缺 點
刪除法	直接刪除存有遺漏值的資料，以確保資料完整性。	易將造成資料量縮減，導致可分析和挖掘的資訊變少。
人工比對	以相關測量或題目類比的邏輯推理法，將遺漏值以最有可能出現的答案填補。	人易受時.地.心情等內外在因素影響，因此所估出的值並不客觀。
固定填補	從不含有遺漏值的資料中，將相同屬性欄位以機器式學習方法，找出一個固定值填補。	雖省時省力，但無法確保值的正確性，可能存有某種程度上的偏差(bias)。
眾數填補法	以資料庫同屬性全部資料中出現次數最多的值當作遺漏值的回填。	若屬性資料出現不具有高度重覆性的話，在分析應用上將有其限制。
平均數填補法	以資料庫中不含有遺漏值的同屬性資料計算出平均回填。	易受極端值或資料型態分佈影響，導致所求出的平均值有所偏差。

第二節 研究目的

由於之前大部份所提出的遺漏值推估法並沒有考慮到屬性之間的關聯與特性，容易導致推估出的遺漏值會有所偏差。所以我們以分群技術為基礎來建立本研究的遺漏值推估模組，將物以類聚、群間資料特性同質、群間資料特性異質的概念應用在遺漏值的估計上；依照資料分佈做空間上的切割，找出屬性之間的關聯讓分群後的資料關聯與特性更為緊密，並且利用基因演算法全域隨機多點搜尋的特性來尋找最佳的分群結果，作為遺漏值填補的基礎。

根據上述相關的研究背景與研究動機後，在此節介紹本研究其研究目的如下：

一、 提出新的遺漏值推估方法，以分群的概念針對不同大小與特性的資料集進行遺漏值回填。

二、 應用屬性值切割方法與基因演算法來改善分群的品質，使分群後的資訊能讓遺漏回填值更接近真實值。

三、 透過 Crude Oil、Iris plants、Glass 和 Vowel 四個真實資料集驗證本研究所提出的遺漏值推估方法之效能與可行性，並且與其它遺漏值推估方法的實驗數據做分析與探討。

第三節 研究程序

本論文研究流程如下，如圖 1-1 所示：

一、 研究動機(發現問題)：在資料探勘過程中，在資料前置處理時常會遇到收集的資料中含有遺漏值，即不完整的資料。而如果採取忽略或是以平均值回填，則容易影響其分析結果的真實性。

二、確立研究主題：提出一個新的遺漏值推估方法，期望其推估效果能優於其它方法。

三、文獻收集與探討：收集遺漏值處理方法、基因演算法、群集分析與屬性值切割方法相關文獻資料，並且探討其特色與優缺點。

四、建構遺漏值推估模組：進行相關文獻探討後，依其優缺點建立本研究所提出的新遺漏值推估模組，應用屬性值切割法與基因分群演算法進行遺漏值的回填。

五、個案實例驗證：利用真實資料集來驗證我們的方法，將遺漏值推估模組所得到的回填結果與其它方法比較分析，並探討其可行性與適用情況。

六、結論與建議：說明遺漏值推估模組的特色與優缺點，並且探討未來還有那些問題可以改良與修正，以讓整個推估模組更趨完整。

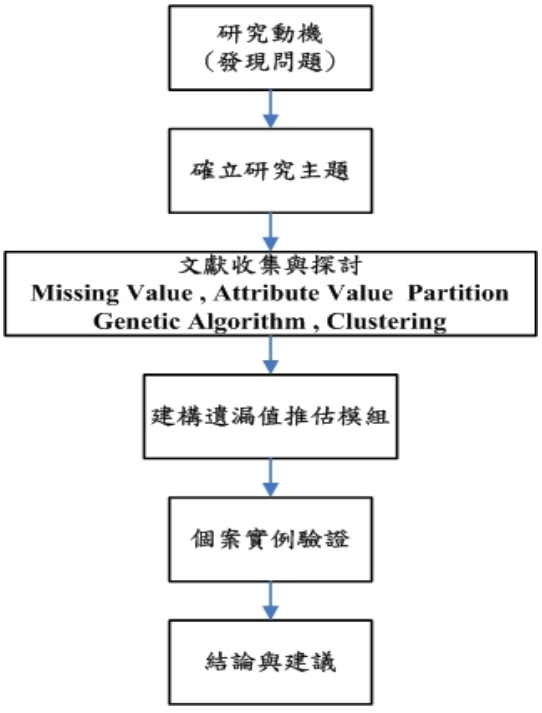

圖 1-1 研究流程圖

第四節 論文架構

全文共分為五章，各章內容說明如下：

第一章 緒論：介紹本論文之研究背景、動機與目的。

第二章 文獻探討：針對遺漏值處理、群集分析、屬性值切割演算法與基因演算法等相關文獻進行研究與分析。

第三章 研究方法：說明本論文所提出的遺漏值推估方法(透過屬性值的切割與基因分群技術來推估遺漏值)。

第四章 實驗結果與討論：透過真實資料集來驗證本論文提出的方法其效率與可行性。

第五章 結論與未來研究方向：說明本論文的總結與貢獻，另外闡述本論文的研究限制與未來的研究方向。

第二章 文獻探討

第一節 遺漏值

在資料探勘的過程中，不論是準備樣本資料或是建立探勘模型時，無時無刻都會面臨到大量資料的處理程序，為了讓資料更適合進行探勘的動作，產生高品質的探勘結果。必須對資料做資料前置處理的動作，以建立高品質的資料[8]。

資料在實際演算過程中必須妥善處理原始資料的內容，才不會使資料本身去影響到統計分析的結果，而導致結果產生嚴重的偏差。

壹、 遺漏值的定義與產生原因

資料探勘過程中，如果發生資料不完整的情況，容易造成最後探勘出來結果的偏差，進而影響到決策者的決策分析，發生錯誤的決策，而整個探勘過程就會顯得不具意義。資料不完整的情況最常見的便是資料中某些屬性有遺漏。例如某顧客填寫會員資料表時，可能漏填了年齡這一欄，當我們想要經過資料探勘工具，瞭解顧客年齡與購買商品種類的關係時，便會發現有資料遺失，造成探勘上的困難。

遺漏值的發生可解釋為，在現實生活中確實存在這筆資料，可能在收集資料或建立檔案過程中遺失了，以致資料集中會有遺漏值的現象。例如便利商店的 POS 系統，每小時會固定地從全省各地的加盟店傳回店內的銷售記錄與庫存量，統一彙整到總部主機的資料庫內。但若在某一時刻發生傳輸中斷的情形，例如電腦當機或是停電時，使得該筆資料無法寫入到資料庫中，雖然加盟店確實收集過該筆資料，但

實際上總部資料庫卻遺漏掉了某些資訊，造成資訊的不同步。因此我們可以定義遺漏值為：「在真實世界中實際存在，但是因為設備或人為等其它因素，而無法確實得到的資料」[6]。

遺漏資料值在關聯式資料庫中以null來表示，但是發生的原因可能會不一樣，其發生原因大致有下列幾種情形：(1).空數值；(2).不存在的數值；(3).資料的不完整；(4).資料未收集到。

貳、 遺漏值的型態

遺漏值的型態，可分為下列兩種型態：

一、 數值性資料：資料中的屬性為一連續性可計算的數值，如薪資可紀錄為 30000 元。

二、 類別型資料：資料中的屬性值可分出幾大類別稱之，如星座資料可記錄為雙魚座或射手座。

參、 遺漏值處理技術

一、 傳統處理方法：

(一)、直接刪除法：直接將存有遺漏值的資料刪除，只留下沒有遺漏的資料，來確保分析資料的完整性。不過這種方法的缺點是如果遺漏值存在過多時，其刪除的資料量過多，造成分析的資料量不足，導致最後資料探勘分析出來的準確率也會隨著下降。

(二)、人工比對法：以相關測量或題目類比的邏輯推理法，將遺漏值以最有可能出現的答案填補。其缺點是容易受人、事、時、地、心情等內外因素影響，因此此方法所

估計出的值並不客觀。

(三)、眾數填補法：以資料庫同屬性全部資料中出現次數最多的值當作遺漏值的回填。其缺點是若屬性資料出現不具有高度重覆性的話，意即資料分佈平均，在分析應用上將有其限制之處。

(四)、平均數填補法：以資料庫中不含有遺漏值的同屬性資料計算出平均值回填，也是傳統較常用的填補方法。不過容易受極端值或資料型態分佈影響，導致所求出的平均值有所偏差。

二、 近年來所提出的方法

(一)、多重插補法[15]：是一種資料擴充和統計分析方法，主張找一個和遺失值最相似的數值來取代遺漏值。再針對每個填補資料集合都以完整資料集的統計方法來進行統計分析，最後綜合每個填補資料集的結果，得到最終的遺漏推估值。不過其需產生大量的插補值與計算過程繁複，因此大幅增加估計的時間與成本。

(二)、類神經網路[4]：利用自組織映射圖網路(self-organize mapping , SOM)其學習演算法的特性，透過不斷的訓練找出輸入及期望輸出的差異所在，並調整加權值來達到學習的效果，讓推估模式在推估過程中記取經驗以達到推估最佳化的目的，不過其缺點在於當資料量大時，需要一定的學習時間，所以無法反應在即時的估計上。

(三)、群集分析方法：以分群技術為基礎，將物以類聚的觀念應用在遺漏值推估問題上，主要目的在於分析資料彼此之間的相似程度，讓群集之內的相似度高，群集與群集之間的相似度低。藉由分析所找到的分群結果，將遺漏值所屬的群集中其它未含有遺漏值的屬性平均回填到該遺漏值，最常見的分群推估法為 K-means 分群演算法[1]、自動分群法[2][3]等等。為了讓其分群結果更接近於原本資料集中資料的分佈，所以啟發式演算法常被用來改善分群的品質，藉由不斷修正與尋找最佳化結果的特性，來找到最適合的分群結果，以利較準確的回填遺漏值。其方法包括有粒子群演算法[7]、基因演算法[3][5][6]。

第二節 群集分析

壹、 定義與用途

分群(Clustering)又稱為資料切割、非監督式分類。其方法主要是將資料集合中的資料紀錄，又稱資料點，加以分群成數個群集(clusters)，使得每個群集中的資料點間相似程度高於其它群集中資料點的相似程度。意即主要目的在於分析資料彼此之間的相似程度，讓某群集之內相似度高，群集與群集之間相似度低。藉由分析所找到的分群結果，推論出有用、隱含、令人感興趣的特性與現象。在群集分析的過程中，並沒有事先指定好其類別資訊，也沒有任何資訊可以表示資料集中哪些資料是有相關性的，所以被視為一個非監督式學習的應用[20]。

貳、 分群程序

分群的程序可以分為四大步驟，選取資料中適當的特徵值，作為分群的依據，接著必須選取適合資料型態的分群演算法，最後透過評估的準則決定分群的結果，並且提供給專家進行解釋，如圖 2-1 所示

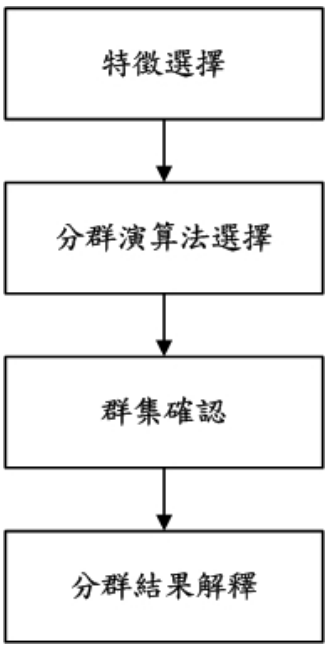

圖 2-1 群集分析流程圖[5]

參、 分群方式

群集分析的方法主要分為兩類，第一類為監督式群集分析，提供群集問題的輸入以及輸出資料，依據給定的輸入與輸出資料之間的映射規則將資料分成數個群集。例如分類(classification)就是屬於監督式群集分析方法；第二類為非監督式群集分析，不必事先給定分群資訊就自動將資料分成適當的群集，而分群就是採用非監督式的群集分析方法。其兩種最具代表性的群集分析方法，分別是階層式分群法(hierarchical clustering)與分割式分群(partitional clustering)，而分割式分群法中最為熟知、發展最久的一種作法就是 K-means 分群演算法，本

研究就是以 K-means 分群演算法為基礎進行分群的流程。以下將介紹 K-means 分群演算法的基本概念。

肆、 K-means 分群演算法

由於 K-means 分群演算法其概念與實作上較簡易，且處理上所需的時間與空間成本都相當低。K-means 使用群集中的質量中心(mean)作為群集中心(群心)，根據使用者事先定義之群集數量 k，K-means 首先隨機從資料集合中選擇任 k 個資料點當作起始 k 群集的群心；接著透過尤拉距離相似度計算公式，把資料集合中的每個資料點歸屬到離它最近之群心所屬的群集中，當作是同一群集。依照其群內所有資料點，計算每個新群集的群心。K-means 除了一開始需指定 k 個資料點當作 k 群集之群心外，其它回合都產生新的群心。反覆這樣的步驟，直到各群群心不再變動為止。

伍、自動分群

所謂的「群」，是指具有相似性質的元素所組合而成的集合。同一群的元素，其性質相近；不同群的元素，其性質則相異。若組成「群」的元素為數值或向量，則其最直接的性質就是「距離」。距離越接近我們說它們的性質最相近；反之，距離越遠，它們的性質就愈不相近。而自動分群演算法[2][12]，利用資料點之間的距離和群集間的距離來作自動的分群，改善了以往非自動分群演算法需要使用者事先決定好群集數的缺點。

第三節 基因演算法

壹、基因演算法簡述與特性

基因演算法(Genetic Algorithms , GA)是由學者John Holland於1975年提出的，它的原理主要參考物競天擇，適者生存的原理，其演算的流程是經由模仿生物演化特性，讓母代隨著每一代的變動來進行演化，其中主要包括選擇(Selection)、交配(Crossover)及突變(Mutation)這三種運算。經由保留較佳之染色體，淘汰較差之染色體的方法，使較好的染色體能擁有較高的機率將其基因遺傳至子代中。因為GA具有簡單運算與平行處理的特性，使其成為解決多項式困難(NP-Hard)最佳化問題的最佳方法之一。

在限制上，GA 是屬於隨機的(Stochastic)最佳化搜尋技術，因此，即使在運作參數完全不變的情形下，仍然會找出不同的解答。此外，雖然 GA 在搜尋的過程不易陷入局部最佳解，卻無法保證找出真正的全域最佳解，而是找出近似最佳解，因此比較適合解決多項式困難的複雜問題。GA 的主要有選擇(Selection)、交配(Crossover)、突變(Mutation)等三種運算。而在執行運算之前，除了要決定染色體的編碼方式以外，還有適應函數以及族群大小、執行代數等參數，以下依序說明各步驟其內容。

貳、基因演算法基本流程

一、編碼

就GA來說，設計編碼方式是最重要的第一步，它要能適當的描述出問題的內容與結構，另一方面也要考慮如何讓GA在上面運算。如果沒有兼顧這兩個方面的話，可能會造成演算法的效能低落。一般而言，

編碼方式以二進位(Binary)編碼、整數(Integer)編碼、實數(Real)編碼為主，其中以二進位編碼最常使用。

二、適應函數

適應函數是用來評估每一條染色體的好壞，進而引導下個族群的答案趨勢，在設計上必須考量問題本身的需求與特性，可能需要讓答案最大化或是最小化。

三、選擇

在選擇這個階段，主要是挑出適當的母代以進行交配的運算。最常使用的是「輪盤法」，它主要是使用適應函數的數值高低，來決定這個染色體被選中的機率，數值越高選中的機率也就越高。此外，還有期望值法、菁英法等挑選方式，為的是將比較好的母代保留到下一代。

四、交配

交配運算的目的，是希望藉由這個運算來創造出同時兼具母代雙方優點的子代，然而不可避免的，也可能遺傳到品質較差的基因，所以並不一定保證會產生更好的下一代。在GA中，針對二進位編碼，常用的交配方法主要有三個類型，即單點交配(One-Point Crossover)、兩點交配(Two- Point Crossover)和均勻交配(Uniform Crossover)等三種，依問題的特性，與不同的編碼方式，來選擇或設計適合的交配運算。

五、突變

突變的涵意如同自然界的基因突變，其目的在利用隨機改變基因，以創造出父母代所未具備的新特徵。過程是將交配後產生的子代，根據預設的突變機率選擇突變點進行突變。如果運算方式以二進位編碼方式而言即為反相運算(即 $0 \rightarrow 1$，$1 \rightarrow 0$)。

貳、基因分群演算法[14]

在基因分群演算法中，以K-means演算法為分群的基礎，由資料集內隨機選取其中的資料當成初始的染色體資訊，每一條染色體就是代表一組可能的分群法則。染色體的長度定義為：$L = N \times d$；其中L代表染色體長度，N代表分群數，d代表資料屬性維度。

一、 初始化族群

舉例來說，設分群數為 3，資料維度為 2。圖 2-2 為基因分群演算法初始母體中的一條染色體，圖中的染色體資訊則代表了 3 個群心[(0.26,0.56),(1.9,5.5),(4.6,9.8)]。

0.26	0.56	1.9	5.5	4.6	9.8

圖 2-2.一條隨機染色體

二、 計算適應值

當初始染色體產生之後，我們利用適應函式來計算適應值，來當作分群結果的依據。而K-means演算法的適應函式為各群體中心與同一群的資料點距離加總，適應值越佳則代表分群越成功。

三、 選擇

根據每條染色體的適應值來進行複製的動作，適應值較佳的染色體將被保留下來，適應值較差的染色體將被拋棄。這些經由複製過程被選擇的染色體將成為下一代的族群，繼續進行繁衍。其最常用的輪盤法是依照適應值的高低，賦與不同的選取機率，如公式(2-1)。$f(x_i)$代表第i條染色體的適應值，n表示族群中染色體的總數，而$P(i)$表示第i條染色體被挑選到的機率：

$$P(i) = f(x_i) / \sum_{i=1}^{n} f(x_i) \qquad (2\text{-}1)$$

四、 交配

依照輪盤法從族群中選取兩條染色體，隨機產生交配率來判斷染色體是否要進行交配。一般藉由產生從0到1之間的數值來判斷，若小於交配率則進行交配動作。透過此方式，能將親代的優良特性由子代延續下去，以期望產生更優良的後代。

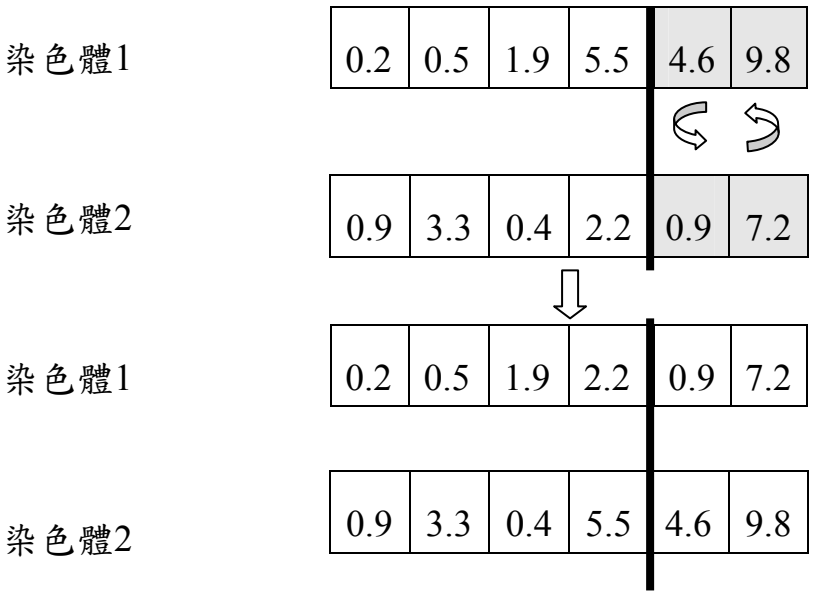

圖2-3. 單點交配圖示

五、 突變

族群中的每條染色體根據突變率來決定是否對染色體中的基因進行突變。藉由隨機產生0到1之間的數值來跟突變率做比較，若小於突變率則進行基因的突變。目的是希望將新的基因引入染色體內，可防止基因演算法提早收斂，也可改善搜尋空間無法被有效的搜尋，避免陷入局部最佳解的現象。而突變的規則是採用公式(2-2)，v為隨機挑選基因值，δ為一隨機產生值，範圍是0到1[14]。

$$v \pm 2 * \delta * v \quad , v \neq 0, \qquad (2\text{-}2)$$

$$v \pm 2 * \delta \quad , v = 0.$$

第三章 遺漏值推估方法

本研究應用屬性值切割方法與基因分群技術進行遺漏值的推估，其主要目的在於將資料中含有遺漏值的部份，透過本研究所提出的方法，填入適當的值。在本章以真實的半導體銅製程實驗資料集來協助瞭解整個系統的流程與實作方法。本章有四小節，第一節介紹本研究所提出的基於群集分析之遺漏值推估模組，分別敘述屬性值切割法回填遺漏值模組、基因分群法回填遺漏值模組與結合此兩種遺漏值模組之優點的遺漏值推估法；第二節說明屬性值切割法回填遺漏值模組的流程與作法；第三節說明基因分群法回填遺漏值模組的流程與作法；第四節說明結合屬性值切割法回填遺漏值模組與基因分群法回填遺漏值模組的流程與作法。

第一節 群集特性引導之遺漏值推估模組

在之前的文獻有探討到，遺漏值處理方法中的初始值大多都是以資料庫中未含有遺漏值的同屬性資料所計算出的平均值來回填，並沒有考慮到屬性之間的關聯與特性。其缺點在於易受極端值或資料型態分佈影響，導致推估出的遺漏值會有所偏差。所以我們以分群技術為基礎來建立本研究的遺漏值推估模組，將物以類聚、群內資料特性同質、群間資料特性異質的概念應用在遺漏值的估計，期望能獲得更優良的回填效果。

本研究所使用到的遺漏值推估模組可分為兩大模組，其推估模組皆是以分群技術為基礎，如圖 3-1 所示。第一個模組是屬性值切割法回填遺漏值模組，其主要精神在於利用動態的屬性值切割方法把含有遺漏值的地方回填適當的數值，事先不必給定其切割的資訊，只藉由資料點之間的距離資訊將資料集中的資料切割成數個區間，區間之內的資料點其特性是最相近的。由於距離計算條件範圍設定較嚴苛，其切割內的資料特性關聯是最靠近的，再加上其計算流程與參數設定較為簡便，所以將其切割結果用來回填遺漏值可望提升遺漏值推估的效率與品質。在本章第二節會介紹屬性值切割法回填遺漏值模組的演算流程與實例說明，並且與其它文獻[4][7]所提出的方法簡單比較其遺漏值推估效果之優劣。

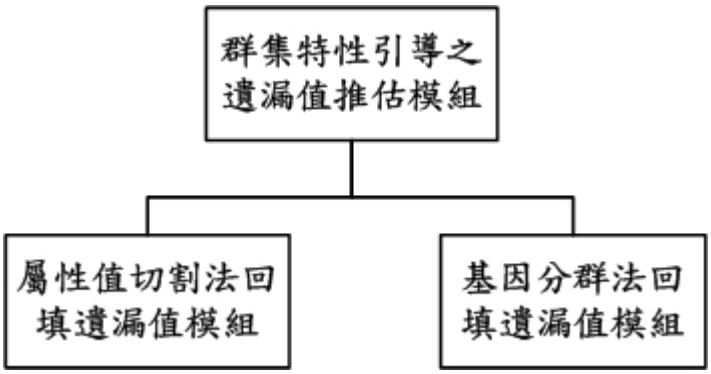

圖 3-1 群集特性引導之遺漏值推估模組

第二個模組是基因分群法回填遺漏值模組，由於分群是屬於最佳化的問題，其分群品質的好壞會影響到整體遺漏值估計的效能，基因演算法具有演化的特性，能夠經由交配、突變等演化過程，不斷的找出適應性最好的分群結果，期望改善分群的品質，讓遺漏值的推估更加準確。

在本章第三節會介紹基因分群法回填遺漏值模組的演算流程與實例說明，並且與其它文獻[4][7]所提出的方法簡單比較其遺漏值推估效果之優劣。

在本研究中嘗試結合上述兩種遺漏值推估模組，由於傳統的遺漏值處理方法中的初始值大多都是以資料庫中未含有遺漏值的同屬性資料所計算出的平均值來回填，易造成回填的誤差產生。所以我們利用屬性值切割法回填遺漏值模組的初始值，獲得比以屬性平均值回填更好的效果，以便之後的基因分群方法也能夠得到更好的分群結果，提升遺漏值推估的準確率，並且能夠縮短整體演算法的運算時間，加快基因演算法的收斂速度。再來將屬性值切割法回填遺漏值模組所回填的初始值資料集結果，利用基因分群的技術將遺漏資料集根據其演化出的分群結果，分別判斷每筆遺漏值分別屬於哪個群集，再將該群集的遺漏值屬性平均回填到遺漏值欄位中，計算出平均絕對誤差(Mean Absolute Error ，簡稱 MAE)，來評估我們的遺漏值推估效果之準確率。在本章第四節會介紹本研究所提出的整合式遺漏值推估模組之演算流程與實例說明，並且與其它文獻[4][7]所提出的方法簡單比較其遺漏值推估效果之優劣。

我們在本章節中使用半導體銅製程實驗資料集來說明我們的遺漏值推估方法之流程與實際運作，其實驗資料集共有 18 筆資料，每筆資料有 3 個屬性，分別是研磨速、均勻度與選擇性，如表 3-1 所示。以下兩小節我們將分別以屬性值切割回填初始值模組與基因分群回填遺漏值模組來推估此實驗資料集所隨機產生的遺漏值，以利瞭解本研究的實作流程。

表 3-1 半導體銅製程原始資料集

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
1	294	14.3	4
2	289	15.7	4.3
3	314	23.2	5.6
4	375	12.1	3.7
5	437	8.7	4.9
6	498	6.5	6.1
7	481	8.99	4.2
8	588	11.8	4.3
9	660	12.4	5.3
10	242	16.2	4.6
11	268	26.9	4.1
12	340	10.5	5.3
13	377	16.9	3.9
14	434	5.06	4.7
15	494	7.08	5.4
16	483	8.76	5.2
17	580	15.1	4.6
18	651	5	5.8

第二節 屬性值切割回填遺漏值模組

屬性值切割方法以分群的概念為基礎，在相同的屬性性質中，同一群的元素性質愈相近，而不同群的元素性質則差異愈大。在我們的方法中，其衡量的準則為「距離」，資料點與資料點之間的距離愈接近，則其性質愈相近，被歸類為一群的機會就愈大。屬性值切割方法事先不必決定切割為幾群，完全以資料點之間的距離與群集之間的距離作為判斷依據。

在本節中我們以前一節所介紹的半導體銅製程資料集當作屬性值切割法回填遺漏值模組的流程範例，再將最後的回填結果跟利用類神經網路[4]與粒子群演算法[7]推估遺漏值方法的實驗結果進行比較分析，來探討屬性值切割法回填遺漏值模組的效益與可行性。首先將資料集產生與文獻[4] [7]相同位置的遺漏值，讓實驗數據的分析比較具備公平性，框線部份為我們產生的遺漏值位置，其 3 筆遺漏資料集如表 3-2 所示。

表 3-2 半導體銅製程遺漏資料集

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
1	294	14.3	4
2	289	15.7	4.3
3	314	23.2	5.6
4	375	12.1	3.7
5	437	8.7	4.9
6	498	6.5	6.1
7	481	8.99	4.2
8	588	11.8	4.3
9	660	12.4	5.3
10	242	16.2	4.6
11	268	26.9	4.1
12	340	10.5	5.3
13	377	16.9	3.9
14	434	5.06	4.7
15	494	7.08	5.4
16	483	8.76	5.2
17	580	15.1	4.6
18	651	5	5.8

屬性值切割法回填遺漏值模組流程圖如圖 3-2 所示，並分成以下幾個步驟來描述說明：

current：屬性中目前欲被切割的數值。

preceding：current 的前一個數值。

total_average_dist：屬性中所有數值的平均差。

Partition _average_dist：屬性中某切割內所有數值的平均差。

步驟 1：將欲切割屬性內非遺漏值欄位的全部數值利用氣泡排序法(Bubble soft)由小到大排列。以表 3-2 的選擇性(Tan/Cu)屬性來看，其排列後的順序如表 3-3 所示：

表 3-3 選擇性(Tan/Cu)屬性排序

3.7	3.9	4	4.2	4.3	4.3	4.6	4.6	4.7	4.9	5.2	5.3	5.3	5.4	5.6	5.8	6.1

步驟 2：利用公式(3-1)計算排列後數值的平均差(total_average_dist)，並且將相同數值視為同一個計算值，避免重複計算造成偏差。

舉例來說：屬性中排序後的數值為 $x_a = x_b < x_c < x_d = x_e < x_f$，由於 $x_a = x_b$、$x_d = x_e$，所以只有四個不同的數值來計算平均差，從最大的數值開始減掉前一個數值，依此類推將之間的差相加並且除以計算的次數。

$$\text{total\_average\_dist}=\frac{(x_f - x_d)+(x_d - x_c)+(x_c - x_a)}{3} \qquad (3\text{-}1)$$

以表 3-3 來看，其計算出來的平均差為：

total_average_dist=(6.1-5.8)+(5.8-5.6)+(5.6-5.4)+(5.4-5.3)+(5.3-5.2)+(5.2-4.9)+(4.9-4.7)+(4.7-4.6)+(4.6-4.3)+(4.3-4.2)+(4.2-4)+(4-3.9)+(3.9-3.7)/13=0.18462

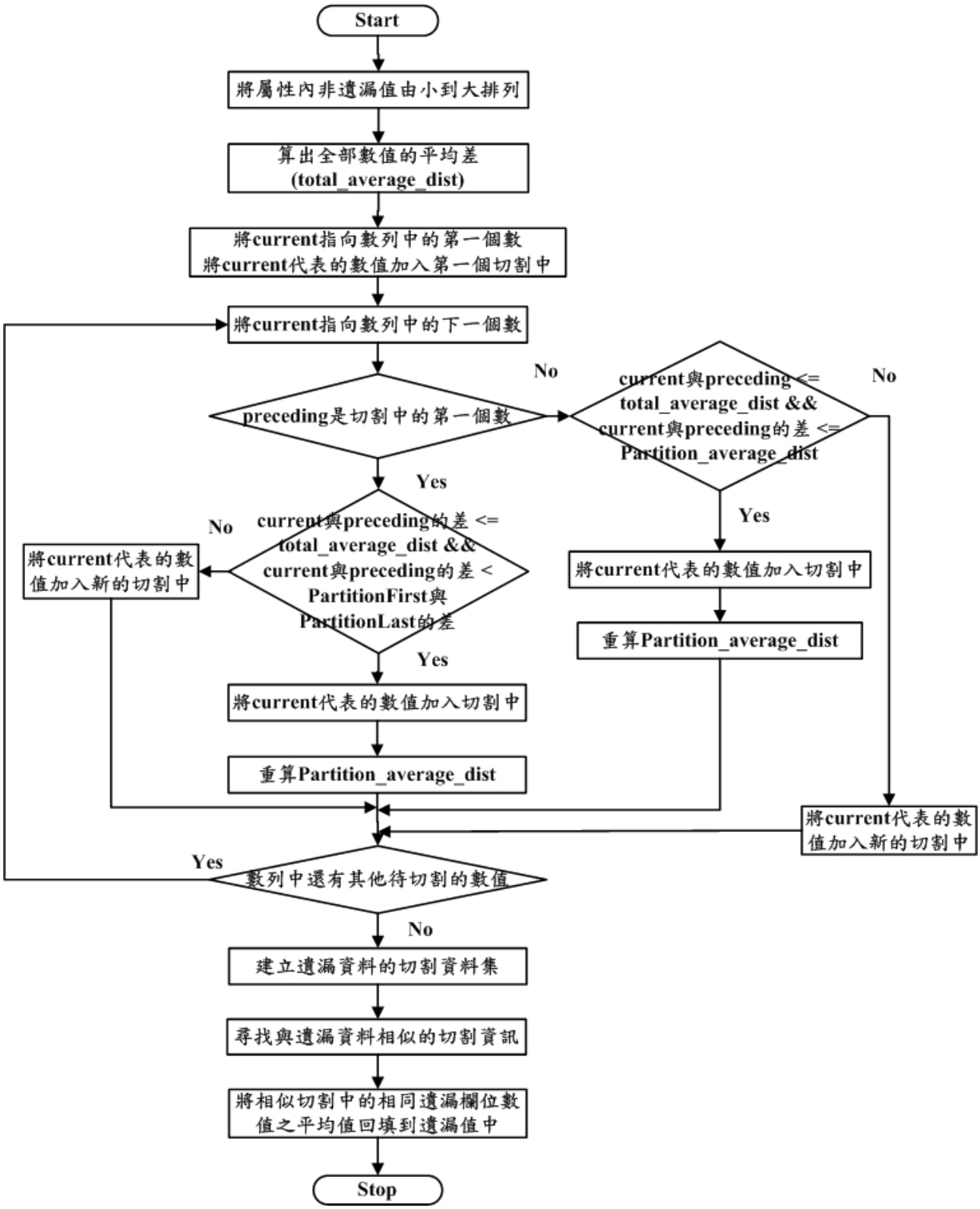

圖 3-2 屬性值切割法回填遺漏值模組流程圖

步驟 3：將 current 指向數列中的第一個數，亦即從排序後最小的數值開始進行切割的動作，並且將第一個 currnet 指向的數值先加入第一個切割中，記錄其切割資訊，如圖 3-3 所示。另外在此設定下列進行切割時會使用到的儲存資訊，說明如下：

PartitionNumber：記錄目前的切割數。

PartitionFirst：記錄目前切割中第一個數值。

PartitionLast：記錄切割中最後一個數值，由於第一個切割時並沒有前一個切割最後一個數值的記錄，所以一開始我們先預設為-99999 的數值，以方便我們之後的計算。

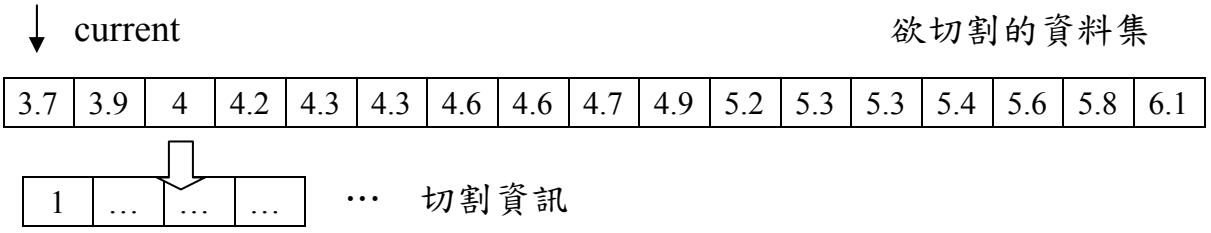

圖 3-3 屬性值切割方法

步驟 4：將 current 指向數列中的下一個數值。

步驟 5：判斷 preceding 是否為目前切割中的第一個數值。

```
if(preceding = PartitionFirst)
{
        if(current - preceding  ≤  total_average_dist &&
           current - preceding  ≤  PartitionFirst - PartitionLast)
        {
```

將 current 代表的數值加入此切割中；

重算 Partition _average_dist；

}

else

{

將 current 代表的數值加入新的切割中；

}

}

else if(current - preceding ≤ total_average_dist &&

current - preceding ≤ Partition _average_dist)

{

將 current 代表的數值加入此切割中；

重算 Partition _average_dist；

}

else

將 current 代表的數值加入新的切割中；

步驟 6：檢查數列中是否還有其它等待切割的數值；

if(current < 數列個數)

重複執行步驟 4~6；

else

終止屬性值切割演算法；

步驟 7：輸出最終切割完的切割資訊，如表 3-4 所示，其資訊從左到右表示第 1 個數值被分到第 1 個切割，第 2 個數值被分到第 2 個切割，第 4 個數值被分到第 3 個切割，以此類推。

表 3-4 切割資訊

1	2	2	3	3	3	4	4	4	5	6	6	6	6	7	8	9

步驟 8：將切割資訊還原到未排序的資料集中，意即建立遺漏資料的切割資料集，如表 3-5 所示。

表 3-5 屬性值切割資料集

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
1	1	5	2
2	1	5	3
3	2	6	7
4	4	4	1
5	5	missing	5
6	6	2	9
7	6	3	3
8	7	4	3
9	8	4	6
10	1	5	4
11	1	7	missing
12	3	4	6
13	4	5	2
14	5	1	4
15	6	2	6
16	missing	3	6
17	7	5	4
18	8	1	8

步驟 9：進行遺漏值回填動作，利用上一步驟建立的切割資料集作為遺漏值回填的參考基礎。其回填方式是將含有遺漏值欄位的第 N 筆資料中非遺漏值的切割資訊與切割資料集的全部資料來作比對，如果其切割資訊相近，就將其相近的資料筆數中與遺漏值欄位相同的原始資料數值相加，並取其平均值回填到遺漏值欄位中，以此類推就完成我們屬性值切割法回填遺漏值模組的程序。

在本研究中，經由 try and error 方式來分析實驗中的數據，我們最終找出判斷其切割資訊是否相近的最佳準則為：如果含有遺漏值欄位的第 N 筆資料中非遺漏值的切割資訊與切割資料集中相同屬性欄的數值相差在±1 的範圍之內，我們就判斷為相近的切割。不過為了避免設立太嚴苛的標準會造成資訊不足或是偏差，我們在挑選相似的切割時，將每筆資料中非遺漏值的切割資訊與切割資料集中相同屬性欄的數值相差在±1 範圍之內的個數記錄起來。當推估遺漏值時，我們選擇與遺漏值切割資訊相似最多的資料來估計，經由 try and error 方式觀察實驗中的數據，找出其最佳挑選準則為直到符合切割數中至少要有 3 筆資料的條件才被挑選，如未符合條件則往下一個符合的切割個數中尋找；等條件相符時，再將以上所找到的相似切割資料當作遺漏值回填的資訊，表 3-6 為與每筆資料與第 5 筆遺漏資料相似的切割個數。

表 3-6 與第 5 筆遺漏資料相似的切割個數

資料編號	符合切割個數
1	0
2	0
3	0
4	1
5	missing
6	1
7	1
8	0
9	1
10	1
11	0
12	1
13	1
14	2
15	2
16	1
17	1
18	0

以表 3-5 中第 5 筆含有遺漏值的資料來說，其非遺漏值欄位的切割資訊皆為 5，與切割資料集相同屬性欄位比對後，尋找符合 5±1 切割範圍的資料，並符合我們所設定的相似切割個數條件，符合最多切割數的資料是第 14、15 筆，不過並沒有符合我們所設定的至少 3 筆資料，所以我們往下修正到下一個符合切割資料集中，將其對應的原始數值相加後，取其平均值回填到 missing 的欄位中。表 3-7 是與第 5 筆遺漏資料相近的切割資訊，將第 4、6、7、9、10、12、13、14、15、16、17 筆的

均勻度欄位原始數值加總起來取其平均值(12.1+6.5+8.99+12.4+16.2+10.5+16.9+5.06+7.08+8.76+15.1)/11=10.8718，所以我們就完成一次遺漏值回填動作，其它屬性的遺漏值也依序按照這樣的流程來進行回填，如表 3-8 所示。

表 3-7 與第 5 筆遺漏資料相似的切割資訊與原始數值

資料編號	研磨速 切割編號/(原始數值)	均勻度 切割編號/(原始數值)	選擇性 切割編號/(原始數值)
5(遺漏資料)	5 (437)	missing	5(4.9)
4	4 (375)	4(12.1)	1(3.7)
6	6(498)	2(6.5)	9(6.1)
7	6(481)	3(8.99)	3(4.2)
9	8(660)	4(12.4)	6(5.3)
10	1(242)	5(16.2)	4(4.6)
12	3(340)	4(10.5)	6(5.3)
13	4(377)	5(16.9)	2(3.9)
14	5(434)	1(5.06)	4(4.7)
15	6(494)	2(7.08)	6(5.4)
16	missing	3(8.76)	6(5.2)
17	7(580)	5(15.1)	4(4.6)

表 3-8 半導體銅製程遺漏值回填結果-以屬性值切割法

遺漏資料編號	研磨速(原始數值)	均勻度(原始數值)	選擇性(原始數值)
5	437	10.8718	4.9
11	268	26.9	4.625
16	498	8.76	5.2

步驟 10：將遺漏值回填完畢後，我們針對遺漏值欄位與原始完整資料來做誤差的計算。在本研究中是以絕對平均誤差(MAE)作為實驗評估準則，經由 MAE 值計算出的大小，可以瞭解以屬性值切割模組回填遺漏值與真實值之離散程度。MAE 數值愈小代表其離散程度愈小，其回填的效果亦較佳。利用公式(3-2)來計算 MAE 數值，計算回填估計值($\hat{e}_i$)減掉真實值(O_i)後的絕對值之平均，其 N 為回填估計值數量。表 3-9 為屬性值切割法回填遺漏值模組、屬性平均值回填、K-means 分群演算法、類神經網路演算法(NBEMS)與粒子群演算法(RKPSO)推估相同的遺漏值樣本其回填效果之誤差比較。由表 3-9 的結果可得知，經過屬性值切割過程能將遺漏值回填效果有效的提昇，僅略差於 RKPSO 回填方法。

$$\text{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|\hat{e}_i-O_i\right| \qquad (3\text{-}2)$$

表 3-9 5 種遺漏值回填方法之誤差比較

	研磨速	均勻度	選擇性	MAE
屬性平均值	52.2941	4.0347	0.7176	19.0155
K-means	53.5000	0.7863	0.5250	18.2704
NBEMS	32.3333	1.0967	0.3857	11.2719
RKPSO	7.3333	1.55	0.25	3.0444
屬性值切割	15	2.1717	0.525	5.8989

第三節 基因分群推估遺漏值模組

先利用傳統的方式，以遺漏值屬性中非遺漏值欄位的平均值回填到遺漏值欄位中，建立分群的初始值資訊，如表 3-10 所示。再利用 K-means 分群演算法將資料集中具有相似特性的資料歸類在一起,經由基因演算法的染色體編碼、適應值計算、選擇、交配和突變，迭代演化出最佳的分群結果。根據其分群結果作為回填遺漏值的依據。圖 3-4 為基因分群法推估遺漏值模組流程圖。

表 3-10 半導體銅製程以屬性平均值回填初始值資料集

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
1	294	14.3	4
2	289	15.7	4.3
3	314	23.2	5.6
4	375	12.1	3.7
5	437	12.735	4.9
6	498	6.5	6.1
7	481	8.99	4.2
8	588	11.8	4.3
9	660	12.4	5.3
10	242	16.2	4.6
11	268	26.9	4.818
12	340	10.5	5.3
13	377	16.9	3.9
14	434	5.06	4.7
15	494	7.08	5.4
16	430.706	8.76	5.2
17	580	15.1	4.6
18	651	5	5.8

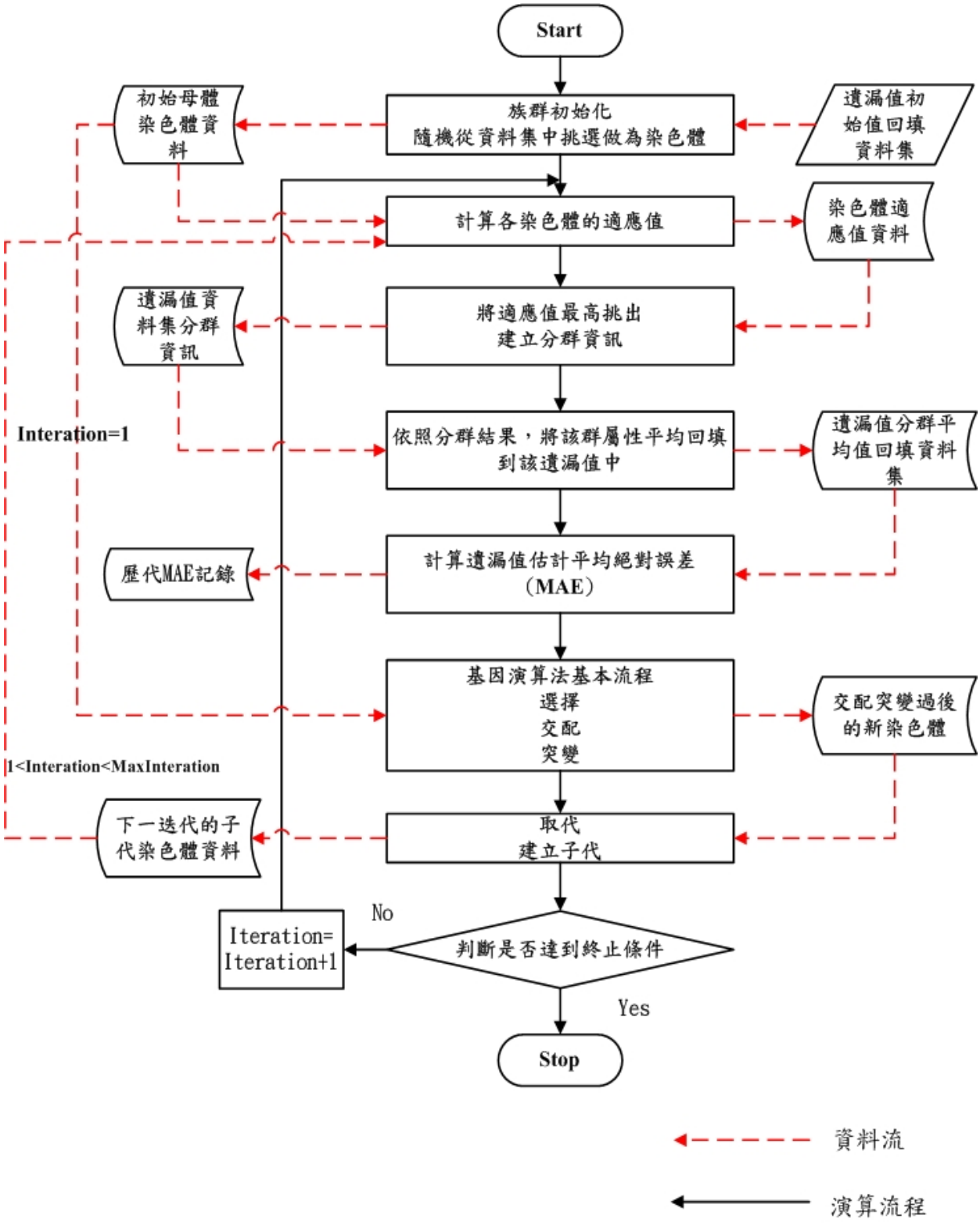

圖 3-4 基因分群法推估遺漏值模組流程圖

以下分別介紹基因分群技術推估遺漏值的作法。

步驟1：族群初始化，在基因分群演算法中，由遺漏值初始值資料集(表3-10)內隨機選取其中的一筆資料當成初始的染色體資訊，染色體個數與輸入的族群大小參數相等，每一條染色體就是代表一組可能的分群法則。染色體的長度定義為：$L = N \times d$，其中L代表染色體長度，N代表分群數，d代表資料屬性維度。

由於遺漏值初始值回填資料集屬於數值型態，因此染色體的編碼採用實數編碼。以表3-10資料集為例，分群數為3，資料維度為3，族群大小為5，表3-11為基因分群演算法隨機初始母體中的染色體資訊，圖中的第一條染色體資訊則代表了3個群心，其群心分別為：

第1群群心：(651 , 5 , 5.8)

第2群群心：(434 , 5.06 , 4.7)

第3群群心：(481 , 8.99 , 4.2)

表3-11 染色體資訊範例

染色體編號	染色體資訊								
1	651	5	5.8	434	5.06	4.7	481	8.99	4.2
2	294	14.3	4	437	12.7	4.9	588	11.8	4.3
3	377	16.9	3.9	651	5	5.8	660	12.4	5.3
4	498	6.5	6.1	242	16.2	4.6	340	10.5	5.3
5	314	23.2	5.6	580	15.1	4.6	340	10.5	5.3

步驟 2：計算每條染色體的適應值，在染色體資訊初始化之後，必須設定適應函數來計算每條染色體的適應值，以便估計每條染色體的好壞。在本研究中基因分群技術是以 K-means 分群演算法為

基礎，計算資料集中每筆資料與每條染色體中的各群群心的離散程度，亦即以距離空間距離計算公式，在此以尤拉距離(公式3-3)當作適應函數來衡量資料點間彼此的相似程度。

$$d_2(X_i, X_j) = \left(\sum_{d=1}^{k} \left| X_{id} - X_{jd} \right|^2 \right)^{\frac{1}{2}} = \left\| X_i - X_j \right\| \qquad (3\text{-}3)$$

資料點 X_i=< $X_{i1}, X_{i2}, \ldots, Xi_k$ >，X_j=< X_{j1}, X_{j2} ,…, X_{jk} >為資料集合中的任兩筆資料記錄，其中 k 為資料表示所採用的資料維度。由於以距離來衡量資料點間彼此的相似程度，所以距離愈短其相似程度愈高，所以在此適應值算法為 1/距離，距離愈短其適應值愈高。其公式如下：

$$\text{Fitness Function} = \frac{1}{\left(\sum_{d=1}^{k} \left| X_{id} - X_{jd} \right|^2 \right)^{\frac{1}{2}}} \qquad (3\text{-}4)$$

以表 3-11 染色體資訊為例，我們計算第一條染色體的適應值為：

$$\frac{1}{\sqrt{(294-651)^2 + (14.3-5)^2 + (4-5.8)^2 + \ldots.. + (4-4.2)^2}} = 0.000752$$

依序計算出每條染色體適應值後，記錄其適應值資訊，如表 3-12 所示。

表 3-12 適應值資訊範例

染色體編號	染色體資訊									適應值
1	651	5	5.8	434	5.06	4.7	481	8.99	4.2	0.000752
2	294	14.3	4	437	12.7	4.9	588	11.8	4.3	0.000778
3	377	16.9	3.9	651	5	5.8	660	12.4	5.3	0.000608
4	498	6.5	6.1	242	16.2	4.6	340	10.5	5.3	0.000752
5	314	23.2	5.6	580	15.1	4.6	340	10.5	5.3	0.000772

步驟 3：根據上一步驟建立的適應值資訊，將其適應值最高的染色體挑選出，當作資料集分群的主要依據。以表 3-12 來看，其適應值最高的是第 2 條染色體的 0.000778，所以我們將第 2 條染色體當作我們的分群準則，計算每筆資料與此染色體中的 3 群群心之距離，來判斷每筆資料應該被分到哪個群集之中。以表 3-10 中的第 2 筆資料來說明該筆資料會被分到哪一群集中，如表 3-13 所示。

表 3-13 遺漏值資料集第 2 筆資料分群範例

第 2 筆資料數值								
289			15.7			4.3		
第 1 群群心			第 2 群群心			第 3 群群心		
294	14.3	4	437	12.7	4.9	588	11.8	4.3
第 2 筆資料與第 1 群群心之距離			第 2 筆資料與第 2 群群心之距離			第 2 筆資料與第 3 群群心之距離		
5.201			148.032			299.025		

由表 3-13 可得知第 2 筆資料與第 1 群群心之距離最近，所以將第 2 筆資料分到第 1 個群集中，依序將每筆資料分到最相似的群集之中，建立每筆資料的分群資訊，如表 3-14 所示。

表 3-14 半導體銅製程遺漏值初始值回填資料集分群資料

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)	所屬群集
1	294	14.3	4	1
2	289	15.7	4.3	1
3	314	23.2	5.6	1
4	375	12.1	3.7	2
5	437	12.735	4.9	2
6	498	6.5	6.1	2
7	481	8.99	4.2	2

8	588	11.8	4.3	3
9	660	12.4	5.3	3
10	242	16.2	4.6	1
11	268	26.9	4.818	1
12	340	10.5	5.3	1
13	377	16.9	3.9	2
14	434	5.06	4.7	2
15	494	7.08	5.4	2
16	430.706	8.76	5.2	2
17	580	15.1	4.6	3
18	651	5	5.8	3

步驟 4：進行遺漏值回填動作，利用上一步驟建立的分群資訊作為遺漏值回填的參考基礎。其回填方式是將含有遺漏值欄位的第 N 筆資料其所屬群集與每筆資料所屬群集作比對，當資料集中的資料所屬群集與含有遺漏值欄位的資料所屬群集相同時，就把符合條件的資料其跟遺漏值欄位相同的原始資料數值相加，並取其平均值回填到遺漏值欄位中。以表 3-10 中第 5 筆含有遺漏值的資料來說，對照表 3-14 後，其分群資訊為第 2 群，與每筆資料所屬群集比對後，尋找符合第 2 群集範圍的資料，將其遺漏值欄位對應的原始數值相加後，取其平均值回填到遺漏值的欄位中(均勻度屬性)的欄位中。表 3-15 是與第 5 筆遺漏資料相同群集的資料，將其均勻度欄位數值加總起來取其平均值為(12.1+6.5+8.99+16.9+5.06+7.08+8.76)/7=9.3414，將其值取代第 5 筆資料遺漏欄位原本的數值，所以我們就完成一個迭代的遺漏值回填動作，其它含有遺漏值的資料也依序按照這樣的流程來進行回填，如表 3-16 所示。

表 3-15 與第 5 筆遺漏資料相同群集的資料集

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)	所屬群集
4	375	12.1	3.7	2
6	498	6.5	6.1	2
7	481	8.99	4.2	2
13	377	16.9	3.9	2
14	434	5.06	4.7	2
15	494	7.08	5.4	2
16	430.706	8.76	5.2	2

表 3-16 半導體銅製程屬性值切割法遺漏值回填結果

遺漏資料編號	研磨速	均勻度	選擇性
5	437	9.3414	4.9
11	268	26.9	4.9
16	451	8.76	5.2

步驟 5：將遺漏值回填完畢後，我們針對遺漏值欄位與原始完整資料來做誤差的計算。利用公式(3-2)來計算 MAE 數值，評估基因分群法推估遺漏值模組此一迭代回填遺漏值的效果優劣。

步驟 6：執行完一個迭代的遺漏值回填動作後，接著進行基因演算法的演化流程，以下為演化流程說明。

(1) 選擇：在選擇這個階段，主要是挑出適當的母代以進行交配的運算。我們使用的是輪盤法它主要是使用適應函數的數值高低，來決定這個染色體被選中的機率，數值越高選中的機率也就越高，一次挑選兩條染色體，以供之後的交配執行。

(2) 交配：我們利用單點交配法，將之前依照輪盤法挑選出

的兩條染色體，先產生一個範圍 0~1 的交配亂數值，如果亂數值小於或等於我們所設定的交配率，我們就進行交配的動作。在本研究中我們交配率依照文獻[14]訂定為 0.8。接著隨機挑選一個交配點，將交配點之後的基因值進行互換，其交配點必須為不同群集之間的切割位置，以表 3-12 的範例來看，其被挑選到的機率較高的染色體為第 2 條與第 5 條，其交配方式如圖 3-5 所示。

染色體編號	基因值								
			交配點						
2	294	14.3	4	437	12.7	4.9	588	11.8	4.3
5	314	23.2	5.6	580	15.1	4.6	340	10.5	5.3
2	294	14.3	4	580	15.1	4.6	340	10.5	5.3
5	314	23.2	5.6	437	12.7	4.9	588	11.8	4.3

圖 3-5 交配示意圖

(3) 突變：持續進行染色體交配的動作，直到其產生的子代大小跟母代大小相同時即停止，之後進行突變的流程。以大自然角度來看，突變其發生機率並不會很高。所以本研究中依照文獻[14]將突變率設定為 0.001，產生一個範圍 0~1 的突變亂數值，如果亂數值小於或等於我們所設定的突變率，我們就進行突變的動作。突變的方法我

們參考文獻[14]的突變方法加以修正，針對染色體上其中一個基因值，利用公式(2-2)的方式產生一個值來取代原來的基因值，達到突變的效果。其突變方法如圖 3-6 所示。

染色體編號	基因值								
突變點									
2	294	14.3	4	580	15.1	4.6	340	10.5	5.3
$v \pm \delta * v \quad , v \neq 0$ v 為隨機挑選基因值 ±機率各為 50% δ 為一隨機產生值，範圍是 0 到 1					15.1+0.45*15.1=21.9				
突變後									
2	294	14.3	4	580	21.9	4.6	340	10.5	5.3

圖 3-6 突變示意圖

步驟 7：建立新子代：經過基因演化流程之後，將之前交配和突變過後所產生的新族群，建立新的子代，來當作我們下一迭代基因分群法推估遺漏值模組新的分群樣本資料集。

步驟 8：判斷是否到達終止條件：在本研究中，基因分群法推估遺漏值模組設定的執行迭代為 200 代，如果還未完成 200 代的終止條件，就繼續重複步驟 2~步驟 8，直到執行迭代到達 200 次即終止整個推估模組的演算，表 3-17 為最終的遺漏值回填結果。

表 3-18 為基因分群法回填遺漏值模組、屬性平均值回填、K-means 分群演算法、類神經網路演算法(NBEMS)、粒子群演算法(RKPSO)與上一節的屬性值切割法推估相同的遺漏值樣本其回填效果之誤差比較。由表 3-18 的結果可得知，經過基因分群技術能將遺漏值回填效果有效的提昇，僅略差於 RKPSO 回填方法。

表 3-17 半導體銅製程基因分群法遺漏值回填結果

遺漏資料編號	研磨速	均勻度	選擇性
5	437	7.278	4.9
11	268	26.9	4.4857
16	468.8	8.76	5.2

表 3-18 6 種遺漏值回填方法之誤差比較

	研磨速	均勻度	選擇性	MAE
屬性平均值	52.2941	4.0347	0.7176	19.0155
K-means	53.5000	0.7863	0.5250	18.2704
NBEMS	32.3333	1.0967	0.3857	11.2719
RKPSO	7.3333	1.55	0.25	3.0444
屬性值切割	15	2.1717	0.525	5.8989
基因分群法	14.2	1.422	0.3857	5.3359

第四節 本研究提出的遺漏值推估模組

經由前兩節的兩種推估模組之流程介紹與實例說明，再經由初步的和傳統回填方法與之前所提出的遺漏值填補方法之誤差比較，可知其兩種推估模組皆具有不錯的推估效果，所以我們嘗試將兩種推估模組結合，預期能改善整體遺漏值推估演算法的效能和效率。

屬性值切割法回填遺漏值模組在本研究的主要目的在於將含有遺漏值的資料集，先初步回填初始值，提供基因分群回填遺漏值模組執行時的資料來源。其用意是藉由屬性值切割方法先獲得比屬性平均值回填、K-means 分群演算法回填和類神經網路演算法回填還要優良的推估結果，再利用基因分群演算法的演化特性來估計遺漏值，期望能獲得比之前研究更佳的推估結果。

本研究之推估模組流程圖如圖 3-7 所示，並分成以下步驟來說明：

步驟 1：將含有遺漏值的資料集匯入至屬性值切割法回填初始值模組執行遺漏值初始值的回填，得到遺漏值初始值回填後的資料集。

步驟 2：將遺漏值初始值回填資料集匯入至基因分群法回填遺漏值模組中，利用基因分群將資料集進行分群，建立資料集的分群資訊。並判斷含有遺漏值的資料分屬於哪一群，將資料集中同一群屬性平均回填到遺漏值中，得到遺漏值回填資料集後，如表 3-19 所示，再與原始完整資料集計算 MAE。

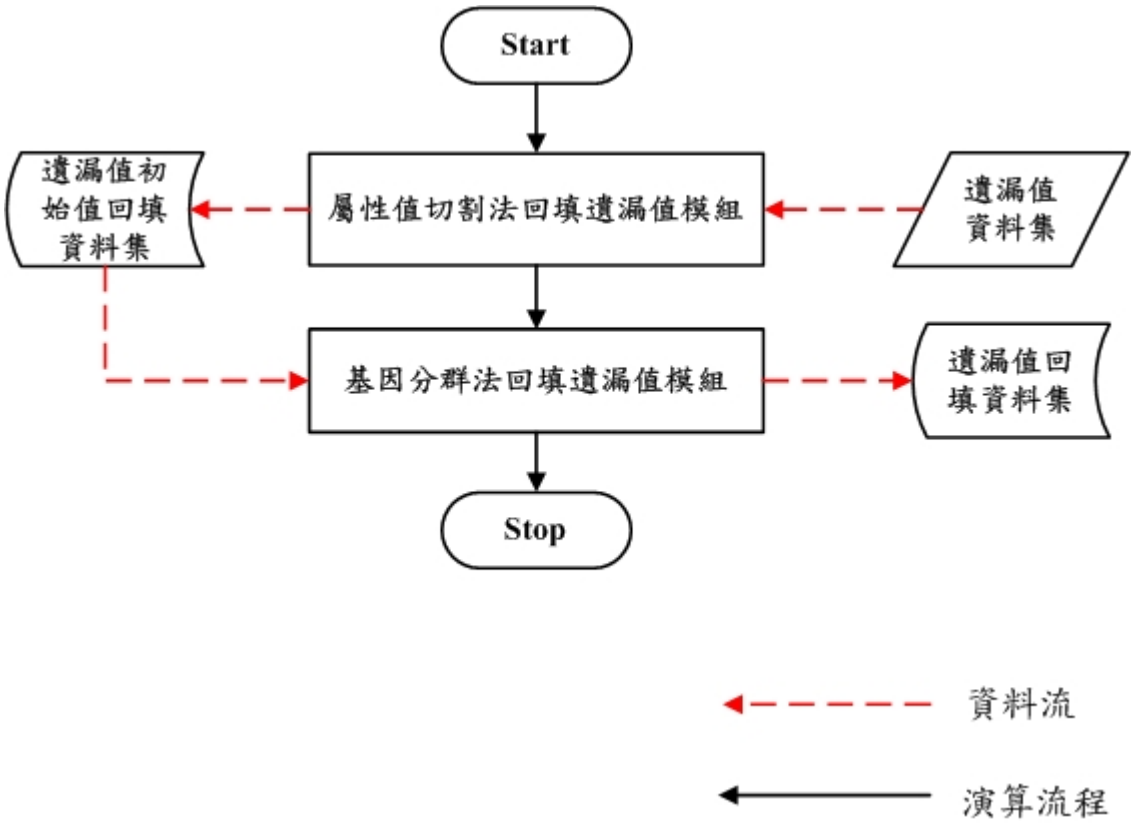

圖 3-7 本研究推估模組流程圖

表 3-20 為本研究遺漏值推估模組、基因分群法回填遺漏值模組、屬性平均值回填、K-means 分群演算法、類神經網路演算法(NBEMS)、粒子群演算法(RKPSO)與上一節的屬性值切割法推估相同的遺漏值樣本其回填效果之誤差比較。由表 3-20 的結果可得知，在半導體銅製程資料集中，本研究所提出的遺漏值推估模組回填效果雖然與基因分群法的回填效果相同，不過其收斂速度有明顯的加快，如圖 3-8、3-9 所示。在第四章將會探討不同性質、大小的資料集其遺漏值推估效果與收斂效果。

表 3-19 半導體銅製程遺漏值回填結果

遺漏資料編號	研磨速	均勻度	選擇性
5	437	7.278	4.9
11	268	26.9	4.4857
16	468.8	8.76	5.2

表 3-20 七種遺漏值回填方法之誤差比較

	研磨速	均勻度	選擇性	MAE
屬性平均值	52.2941	4.0347	0.7176	19.0155
K-means	53.5000	0.7863	0.5250	18.2704
NBEMS	32.3333	1.0967	0.3857	11.2719
RKPSO	7.3333	1.55	0.25	3.0444
屬性值切割	15	2.1717	0.525	5.8989
基因分群法	14.2	1.422	0.3857	5.3359
屬性值切割+基因分群法	14.2	1.422	0.3857	5.3359

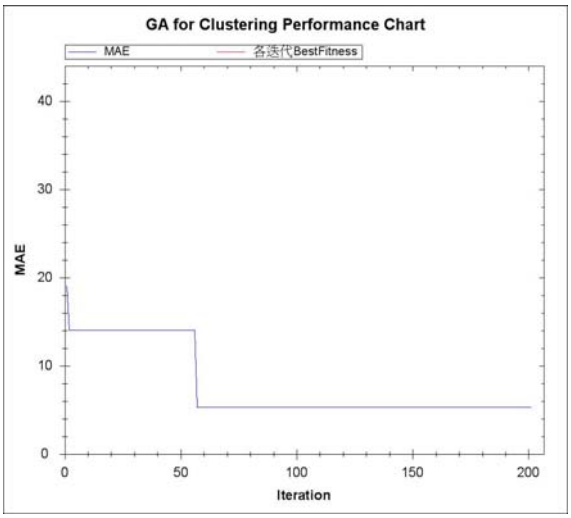

圖 3-8 屬性值平均+基因分群收斂圖

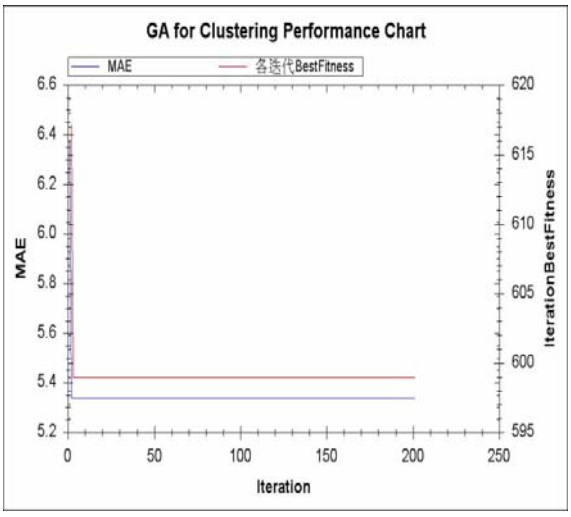

圖 3-9 屬性值切割+基因分群收斂圖

第四章 實驗結果

依照第三章的研究方法將本研究所提出的遺漏值推估模組，以樣本較小的半導體銅製程實驗來說明及驗證本方法的可行性與初步推估結果，並且與之前所提出的遺漏值推估方法進行比較分析。在本章將遺漏值推估模組應用到資料量更大的真實資料庫上，在第一節介紹實驗環境，本研究架設於何種平台上與使用之程式開發工具；第二節介紹遺漏值推估模組演算的參數設定；第三節介紹用來實驗的四個真實資料集其資料格式、維度與資料量大小，包含 crude oil、iris、glass 與 vowel 資料集；第四節說明本研究的實驗設計、流程與評估準則；第五節呈現實驗結果，並且分析其數據結果加以討論。

第一節 實驗環境

本研究所提出的遺漏值推估模組建構於個人桌上型電腦上；其軟硬體環境為：CPU：Intel Core2 1.86GHz；RAM：DDR 400 2GB；作業系統(OS)：Microsoft Windows XP Professional；開發工具：Microsoft visual studio 2008；程式語言：Visual C# 2008；實驗資料集檔案格式：逗號分隔型取值格式(Comma Separated Values，簡稱 CSV)。

第二節 參數設定

在本節介紹遺漏值推估模組其實驗參數設定，屬性值切割法回填遺漏值模組中多皆以資料之間的距離做判斷，自動切割成適當的切割數，所以不必設定參數值。在基因分群法回填遺漏值模組中，我們需設定演化的迭代次數即基因演算法的執行迴圈；設定族群大小即染色

體樣本個數；設定交配率與突變率以決定進行交配與突變的機率，上述基因分群參數皆參考文獻[14]所設定；而四個真實資料集其遺漏值比率設定為 5%、10%、15%與 20%等四種條件；由於基因分群演算法以 k-means 演算法為基礎，所以必須事先給予實驗資料集的分群數量，依照資料集不同其分群數量也有所不同，如表 4-2 所示；而評估本研究的推估效能之準則為 MAE 數值，MAE 愈低代表遺漏值推估效果愈好。相關的實驗參數設定如表 4-1 所示。

表 4-1 實驗參數設定

參數名稱	本實驗設定值
迭代次數	200
族群大小(染色體)	100
交配率	0.8
突變率	0.001
遺漏值比率	5%、10%、15%、20%
分群群集數	依各資料集而訂
評估準則	MAE

第三節 實驗資料集

本論文使用四個真實資料庫作為本研究的實例驗證，其資料庫筆數呈現遞增狀態，其屬性特徵也有所差異。主要是藉此測試本研究所提出的遺漏值推估模組隨著資料量的變化，其推估的效果會有何變化，來驗證本方法的可行性與效能評估，以求能適用在多樣化的資料

集中。

本研究四組真實資料集名稱為天然石油(Crude Oil)、鳶尾花(Iris plants)、玻璃(Glass)與母音(Vowel)，其詳相關資訊如表 4-2 所示。

表 4-2 真實資料庫之相關資訊

資料集名稱	資料筆數	屬性維度	分群數目
Crude Oil	56	5	3
Iris plants	150	4	3
Glass	214	9	6
Vowel	871	3	6

壹、天然石油(Crude Oil)

天然石油資料集中共有56筆資料，依其資料屬性分為Wilhelm、Sub-Mulinia與Upper三個群集，其中每筆資料皆有5個屬性維度，分別為釩(Vanadium)、鐵(Iron)、鈹(Beryllium)、飽和碳氫化合物(Saturated Hydrocarbons)與芳香族碳氫化合物(Aromatic Hydrocarbons)[23]，其資料格式如表4-3所示。其它詳細資料如表4-4與表4-5所示。

表4-3 Crude Oil資料格式範例

ID	Vanadium	Iron	Beryllium	Saturated Hydrocarbons	Aromatic Hydrocarbons
1	3.9	51	0.2	7.06	12.19
2	2.7	49	0.07	7.14	12.23
3	2.8	36	0.3	7	11.3
4	3.1	45	0.08	7.2	13.01
5	3.5	46	0.1	7.81	12.63

表4-4 Crude Oil資料集數值分佈資訊

屬性維度	最大值	最小值	平均值	標準差
Vanadium	11	1.2	6.1804	2.4026
Iron	52	5.6	27.0464	11.5014
Beryllium	1.5	0	0.3414	0.3111
Saturated Hydrocarbons	9.25	3.06	5.2911	1.3678
Aromatic Hydrocarbons	13.01	2.22	6.4336	3.1263

表4-5 Crude Oil資料集群集資訊

群集名稱	所佔資料筆數	所佔資料比例
Wilhelm	7	12.5%
Sub-Mulinia	10	17.86%
Upper	39	69.64%

貳、鳶尾花(Iris plants)

鳶尾植物資料庫共有150筆資料，依其資料屬性分為Iris Setosa、Versicolour與Virginica三個群集，其每筆資料皆有萼片長度(Sepal Length)、萼片寬度(Sepal Width)、花瓣長度(Petal Length)與花瓣寬度(Petal Width)4個屬性維度[22]，其資料格式如表4-6所示。其它詳細資料如表4-7與表4-8所示。

表4-6 Iris plants資料格式範例

ID	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2

表4-7 Iris plants資料集數值分佈資訊

屬性維度	最大值	最小值	平均值	標準差
Sepal Length	7.9	4.3	5.84	0.8253
Sepal Width	4.4	2.0	3.05	0.4321
Petal Length	6.9	1.0	3.76	1.7585
Petal Width	2.5	0.1	1.02	0.7606

表4-8 Iris plants資料集群集資訊

群集名稱	所佔資料筆數	所佔資料比例
Iris Setosa	50	33.3%
Versicolour	50	33.3%
Virginica	50	33.3%

參、玻璃(Glass)

Glass資料庫共有214筆資料，由7種不同的玻璃組成，每筆資料包含9種化學元素[22]，其資料格式如表4-9所示。其它詳細資料如表4-10與表4-11所示。

表4-9 Glass資料格式範例

ID	Ri	Na	Mg	Ai	Si	K	Ca	Ba	Fe
1	1.52101	13.64	4.49	1.1	71.78	0.06	8.75	0	0
2	1.51761	13.89	3.6	1.36	72.73	0.48	7.83	0	0
3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0	0
4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0	0
5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0	0

表4-10 Glass資料集數值分佈資訊

屬性維度	最大值	最小值	平均值	標準差
Ri	1.5339	1.51	1.51	0.0030
Na	17.38	10.73	13.40	0.8147
Mg	4.49	0	2.68	1.4390
Ai	3.5	0.29	1.44	0.4981
Si	75.41	69.81	72.65	0.7727
K	6.21	0	0.49	0.6507
Ca	16.19	5.43	8.95	1.419
Ba	3.51	0	0.175	0.4961
Fe	0.51	0	0.57	0.0972

表4-11 Glass資料集群集資訊

群集名稱	所佔資料筆數	所佔資料比例
Float Building Windows	70	32.71%
Float Vehicle Windows	17	7.94%
Non-Float Building Windows	76	35.5%
Non-Float Vehicle Windows	0	0%
Containers	13	6.07%
Tableware	9	4.20%
Headlamps	29	13.55%

肆、母音(Vowel)

母音資料庫是具有3種音頻屬性的印地安語言之母音，共有871筆資料，將分類為{δ , a , i , u , e , o }六種母音[24]，其資料格式如表4-12所示。其它詳細資料如表4-13與表4-14所示。

表4-12 Vowel資料格式範例

ID	F1	F2	F3
1	700	1500	2600
2	550	1550	2400
3	700	1500	2600
4	700	1600	2700
5	550	1600	2600

表4-13 Vowel資料集數值分佈資訊

屬性維度	最大值	最小值	平均值	標準差
F1	900	250	470.4822	129.148
F2	2550	700	1514.6842	507.2898
F3	3200	1800	2561.0218	244.4014

表4-14 Vowel資料集群集資訊

群集名稱	所佔資料筆數	所佔資料比例
δ	72	8.27%
a	89	10.22%
i	172	19.57%
u	151	17.34%
e	207	23.77%
o	180	20.67%

第四節 實驗設計

本研究的實驗方法是將上一節所介紹的四個真實資料集當作遺漏值推估模組的實驗資料集，以人工方式產生與文獻[7]相同的遺漏值欄位。其四個資料集資料遺失比率分別為 5%、10%、15%與 20%，以 Vowel 資料集舉例，其資料筆數有 871 筆，產生 5%的遺漏值筆數為 44 筆，而這 44 筆中含有遺漏值的屬性欄位皆與文獻[7]中資料集的遺漏屬性欄位相同，一筆資料中只有一個欄位有遺漏值；藉此測試本研究所提出的遺漏值推估模組在不同資料集大小與遺漏值個數下其推估的效果與效率。再來將各實驗資料集匯入推估模組中。

其模組所使用的演算參數設定如表 4-1 所示，首先利用本遺漏值推估模組中的屬性值切割法回填測試資料集中的遺漏值，當作是之後基因分群演算法的分群基礎，再將基因分群最終的分群結果當作是回填遺漏值的重要依據。

遺漏值回填方式可分為 Iteration-base 和 Run-base 的方式來回填，Iteration-base 方式是以演算法每執行一個迭代就進行遺漏值回填的動作；Run-base 方式是以整體演算法趨於收斂之後再進行遺漏值回填的動作。

依照文獻[7]的實驗結果發現，以 Iteration-base 的方式估計遺漏值的回填效率與準確率比以 Run-base 的方式回填要來得優良。所以本研究採用 Iteration-base 的方式進行遺漏值的回填，意謂基因分群演算每執行一個迭代後，依照其分群結果即進行一次的遺漏值回填動作，並計算這次估計的回填誤差。

對於遺漏值推估的準確度其評估準則，在本研究中是採用 MAE 來做為評估的標準。意即將最終遺漏值的回填值與原始完整未遺漏的資料計算其誤差，MAE 數值愈小代表遺漏值估計的結果與真實值的平均差異度也相對愈小，愈接近真實值其遺漏值推估模組的估計效能也就愈佳。

在之前所提到的 MAE 在本文是用來評估遺漏值推估準確度的準則，在本研究的實驗中我們以原始完整的資料集來人工產生遺漏值，最後再將回填後的遺漏值與原始完整的資料集來計算其 MAE，判斷其誤差。不過在真實世界的遺漏值推估程序中，我們只擁有含遺漏值的資料集來估計，並不會得知原始完整的資料為何，所以不能以最終計

算出的 MAE 來評估效能。

而我們能夠掌握到的是分群資訊，例如適應值、收斂結果等等。為了能夠瞭解要以何種估計方式得到的 MAE 才是效果較佳的，所以在實驗程序中我們採用兩種方式來記錄遺漏值推估模組的 MAE 數值，藉以探討以那種方式取得的 MAE 數值較小，即推估效果較佳。第一種方式是記錄基因分群法迭代中最佳的適應值，並擷取當代遺漏值回填的 MAE 數值，來判斷適應值好壞對於 MAE 是否有所影響；第二種方式是記錄基因分群演算法收斂時所得到的 MAE 數值，再來探討其實驗數據與其它方法的分析比較與其各自被採用的時機。

本研究所分析與探討的遺漏值推估模組效能之實驗設計如下：

一、 進行 10 次獨立實驗，分析本遺漏值推估模組與屬性值平均、k-means 分群法、RKPSO、屬性值切割法與屬性值平均+基因分群法等遺漏值推估方法的回填效果，並且探討推估模組其適用情況與影響估計結果的因子為何。

二、 分析本遺漏值推估模組與屬性值平均+基因分群法遺漏值推估方法的演算法收斂效果，並且探討影響收斂效果的因子為何。

第五節 實驗結果

本節將實驗結果分為兩個部份，第一部份探討遺漏值推估實驗結果，即分析比較各種遺漏值方法與本研究提出的方法之 MAE 與其相關數據；第二部份探討遺漏值推估模組的整體演算法收斂效果，而詳細實驗資料集遺漏資料個數如表 4-15 所示。

表 4-15 實驗資料集遺漏值個數

資料集名稱	資料筆數	5%遺漏值	10%遺漏值	15%遺漏值	20%遺漏值
Crude Oil	56	3	6	9	12
Iris plants	150	8	15	23	30
Glass	214	11	22	33	43
Vowel	871	44	88	131	175

壹、遺漏值推估實驗結果

此部份為分析比較各種遺漏值方法與本研究提出的方法之 MAE 與其相關數據，另外說明屬性值平均+基因分群與屬性值切割法+基因分群的欄位中各有兩個數值，第一列記錄演算法收斂之後的 MAE 數值；第二列記錄基因分群法迭代中最佳的適應值，並擷取當代遺漏值回填的 MAE 數值，以下分別為四個真實資料集其 5%、10%、15%與 20%遺漏值比率的 MAE 比較，框線部份表示該值為最佳。

一、Crude Oil 資料集實驗結果

表 4-16 Crude Oil 5%遺漏值比率之實驗結果

Crude Oil 5%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	1.1783	0.7148	0.6747	1.0544	0.4717	0.5464
					0.5464	0.5464
平均 MAE	1.1783	0.7148	0.7408	1.0544	0.4866	0.5464
					0.5464	0.5464
MAE 標準差	-	0	0.1045	-	0.03149	0
					0	0

表 4-17 Crude Oil 10%遺漏值比率之實驗結果

Crude Oil 10%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.9924	0.5549	0.7690	0.9147	0.3959	0.4461
					0.4461	0.4461
平均 MAE	0.9924	0.5549	0.7911	0.9147	0.421	0.4461
					0.4461	0.4461
MAE 標準差	-	0	0.0386	-	0.02645	0
					0	0

表 4-18 Crude Oil 15%遺漏值比率之實驗結果

Crude Oil 15%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	1.2876	0.9656	0.6731	1.0075	0.9722	1.0177
					1.0177	1.0177
平均 MAE	1.2876	1.0334	0.7644	1.0075	0.9858	1.0177
					1.0177	1.0177
MAE 標準差	-	0.0448	0.0619	-	0.02197	0
					0	

表 4-19 Crude Oil 20%遺漏值比率之實驗結果

Crude Oil 20%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	1.754	1.0081	0.909	1.6843	1.1134	1.1134
					1.0942	1.1134
平均 MAE	1.754	1.0360	1.3275	1.6843	1.1134	1.1134
					1.1057	1.1134
MAE 標準差	-	0.0588	0.2488	-	0	0
					0.00991	0

由表 4-16 至 4-19 實驗結果可看出，屬性值平均+基因分群與屬性值切割法+基因分群兩種方法在 Crude Oil 5%與 10%遺漏值比率得到比其它方法還要優良推估效果，而屬性值平均+基因分群略優於屬性值切割法+基因分群方法，不過當遺漏值比率提高時，屬性值切割法+基因分群其推估效果並沒有其它方法來得優良，另外屬性值切割法+基因分群在其估計誤差時，比其它方法要來得穩定。

二、Iris plants 資料集實驗結果

表 4-20 Iris plants 5%遺漏值比率之實驗結果

Iris plants 5%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.7615	0.3554	0.2797	0.3774	0.6112	0.2810
					0.6112	0.2810
平均 MAE	0.7615	0.4815	0.2885	0.3774	0.6129	0.2810
					0.6112	0.2810
MAE 標準差	-	0.1330	0.0254	-	0.00089	0
					0	0

表 4-21 Iris plants 10%遺漏值比率之實驗結果

Iris plants 10%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.6569	0.3293	0.1992	0.2771	0.3933	0.2192
					0.3933	0.2192
平均 MAE	0.6569	0.3724	0.2403	0.2771	0.3943	0.2192
					0.3933	0.2192
MAE 標準差	-	0.0454	0.023	-	0.00076	0
					0	0

表 4-22 Iris plants 15%遺漏值比率之實驗結果

Iris plants 15%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.7026	0.3961	0.264	0.3106	0.3949	0.2774
					0.3949	0.2774
平均 MAE	0.7026	0.4217	0.2971	0.3106	0.3949	0.2774
					0.3951	0.2774
MAE 標準差	-	0.0188	0.0137	-	0	0
					0.00014	0

表 4-23 Iris plants 20%遺漏值比率之實驗結果

Iris plants 20%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.7675	0.3899	0.2656	0.3000	0.4542	0.2656
					0.4542	0.2656
平均 MAE	0.7675	0.4770	0.2895	0.3000	0.4542	0.2656
					0.4542	0.2656
MAE 標準差	-	0.0459	0.0185	-	0	0
					0	0

由表 4-20 至 4-23 實驗結果可看出，屬性值切割法+基因分群方法在 Iris plants 遺漏值資料集中，其平均 MAE 估計效果皆來得比其它方法優良，在整體的估計效果來說，屬性值切割法+基因分群其推估效果優於 k-means 和屬性值平均+基因分群，最佳 MAE 只略差於 RKPSO 方法，但其差距很小；不過屬性值切割法+基因分群在其估計誤差時，比其它方法要來得穩定。

三、Glass 資料集實驗結果

表 4-24 Glass 5%遺漏值比率之實驗結果

Glass 5%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.5608	0.3496	0.3656	0.2375	0.3338	0.2699
					0.2674	0.2699
平均 MAE	0.5608	0.4190	0.4823	0.2375	0.3887	0.2705
					0.2702	0.2705
MAE 標準差	-	0.0755	0.0971	-	0.0874	0.0004
					0.001	0.0004

表 4-25 Glass 10%遺漏值比率之實驗結果

Glass 10%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.6621	0.4992	0.5201	0.3716	0.4969	0.4524
					0.4582	0.4524
平均 MAE	0.6621	0.5352	0.5609	0.3716	0.5313	0.4767
					0.5047	0.4767
MAE 標準差	-	0.0322	0.021	-	0.02732	0.0321
					0.02591	0.0321

表 4-26 Glass 15%遺漏值比率之實驗結果

Glass 15%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.5772	0.3674	0.3786	0.3337	0.3978	0.3859
					0.3757	0.3859
平均 MAE	0.5772	0.4368	0.4446	0.3337	0.4687	0.4175
					0.4029	0.4175
MAE 標準差	-	0.0268	0.0298	-	0.04698	0.0439
					0.02246	0.0439

表 4-27 Glass 20%遺漏值比率之實驗結果

Glass 20%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	0.6430	0.4242	0.3736	1.0228	0.4165	0.3734
					0.3818	0.3734
平均 MAE	0.6430	0.4422	0.4393	1.0228	0.4599	0.4375
					0.3954	0.4375
MAE 標準差	-	0.0139	0.0612	-	0.02076	0.09392
					0.00708	0.09392

由表 4-24 至 4-27 實驗結果可看出，在 Glass 資料集中以屬性值切割法皆能獲得良好的推估效果，屬性值切割法+基因分群方法在 Glass 20%遺漏值比率其估計效果比其它方法來得優良，當遺漏值比率提高時，屬性值切割法+基因分群方法也能維持不錯的遺漏值推估效果，只略差於屬性值切割法，但其差距很小。

四、Vowel 資料集實驗結果

表 4-28 Vowel 5%遺漏值比率之實驗結果

Vowel 5%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	309.7194	275.9669	132.2301	219.0735	233.4366	187.484
					233.6706	187.484
平均 MAE	309.7194	287.7374	185.3032	219.0735	254.3585	211.9554
					256.2705	211.9554
MAE 標準差	-	8.0933	26.3104	-	10.95915	20.4631
					14.78089	20.4631

表 4-29 Vowel 10%遺漏值比率之實驗結果

Vowel 10%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	259.1569	231.9669	157.156	182.2576	182.8824	170.8564
					198.94	170.8564
平均 MAE	259.1569	244.0729	177.501	182.2576	211.8574	183.6081
					215.8794	183.6081
MAE 標準差	-	11.1217	9.3793	-	20.70619	8.6288
					18.21878	8.6288

表 4-30 Vowel 15%遺漏值比率之實驗結果

Vowel 15%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	240.9259	216.9966	163.2652	175.8178	183.6568	153.0797
					183.3428	153.0797
平均 MAE	240.9259	229.8197	179.7402	175.8178	209.9701	172.967
					211.3239	172.967
MAE 標準差	-	10.5983	10.0421	-	14.41303	9.1362
					16.55878	9.1362

表 4-31 Vowel 20%遺漏值比率之實驗結果

Vowel 20%	屬性值平均	k-means	RKPSO	屬性值切割法	屬性值平均+基因分群	屬性值切割+基因分群
最佳 MAE	240.3617	205.7502	157.1018	170.8998	175.3971	160.1485
					183.7597	160.1485
平均 MAE	240.3617	212.5961	185.0526	170.8998	198.1884	166.7859
					201.0487	166.7859
MAE 標準差	-	8.8563	14.8001	-	17.88915	4.1802
					16.73448	4.1802

由表 4-28 至 4-31 實驗結果可看出，屬性值切割法+基因分群方法在不同遺漏值比率下其估計效果皆比 k-means 和屬性值平均+基因分群方法來得優良，只略差於 RKPSO 方法；當 Vowel 15%遺漏值比率時，屬性值切割法+基因分群方法其遺漏值推估效果皆優於其它方法；而屬性值切割法+基因分群在其估計誤差時，比其它方法都要來得穩定。

貳、演算法收斂效果

以下為(A)、屬性平均值+基因分群法推估遺漏值；(B)、屬性值切割法+基因分群法推估遺漏值其 5%遺漏值比率收斂效果之比較：

一、Crude Oil 資料集收斂效果

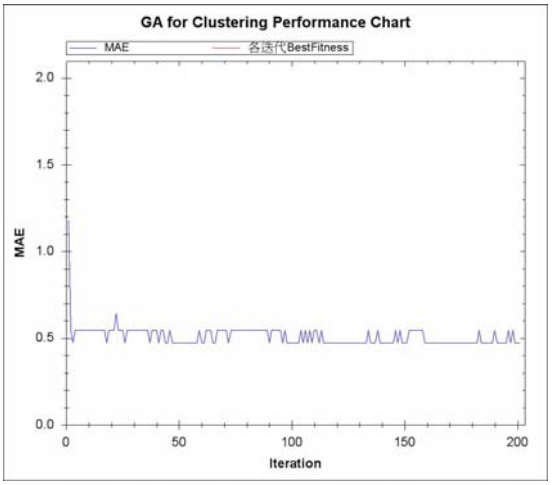

圖 4-1 Oil 5%遺漏值(A)收斂圖

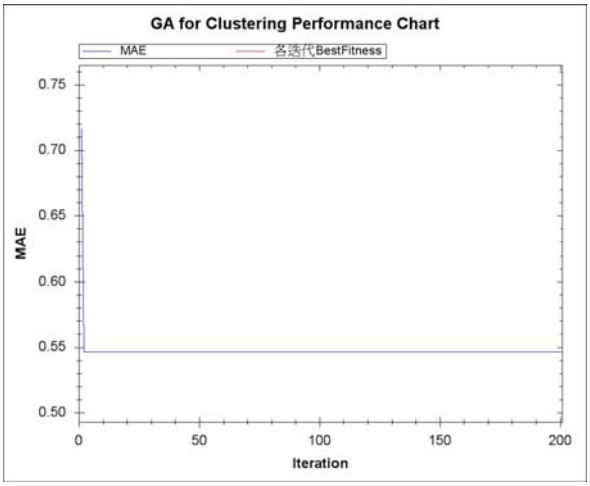

圖 4-2 Oil 5%遺漏值(B)收斂圖

由圖 4-1 與 4-2 可看出，(A)屬性值平均+基因分群方法在迭代演化過程中不會有明顯的收斂效果，可能在執行 200 迭代之後還未有收斂的現象；反觀(B)屬性值切割+基因分群方法在 10 個迭代之內即呈現收斂的現象，而其估計的誤差值並無太大的差異，意即我們提出的方法能夠在較少的迭代內就能得到不錯的估計值，節省演算時間。

二、Iris plants 資料集收斂效果

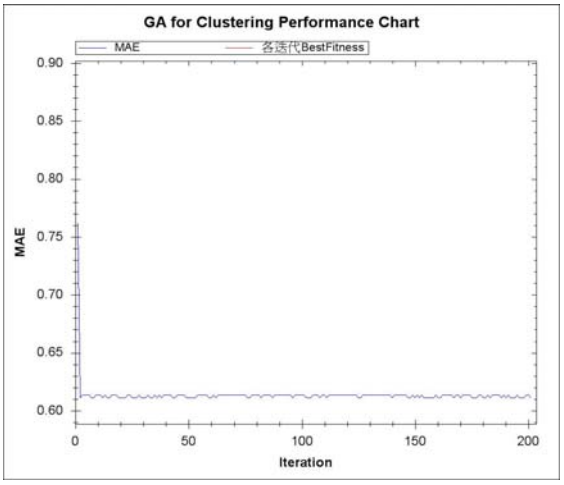

圖 4-3 Iris 5%遺漏值(A)收斂圖

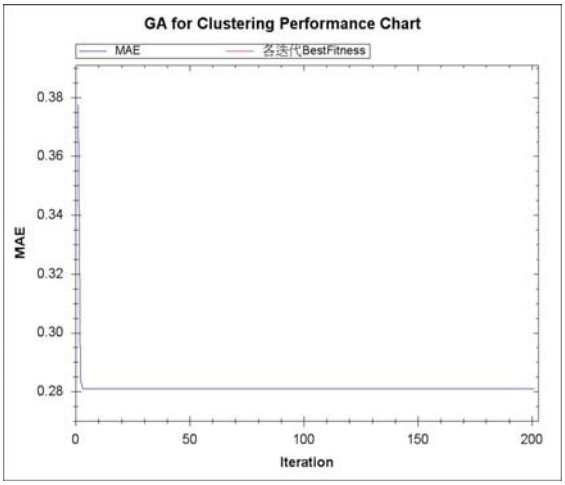

圖 4-4 Iris 5%遺漏值(B)收斂圖

由圖 4-3 與 4-4 可看出，(A)屬性值平均+基因分群方法在迭代演化過程中不會有明顯的收斂效果，可能在執行 200 迭代之後還未有收斂的現象；反觀(B)屬性值切割+基因分群方法在 10 個迭代之內即呈現收斂的現象，而其估計的誤差值並優於(A)，意即我們提出的方法能夠在較少的迭代內就能得到不錯的估計值，節省演算時間。

三、Glass 資料集收斂效果

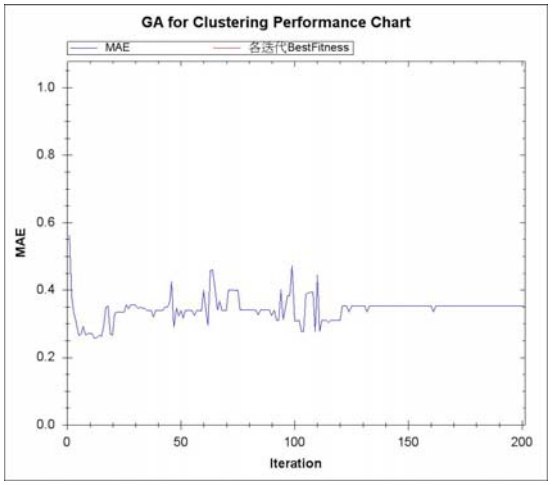

圖 4-5 Glass 5%遺漏值(A)收斂圖

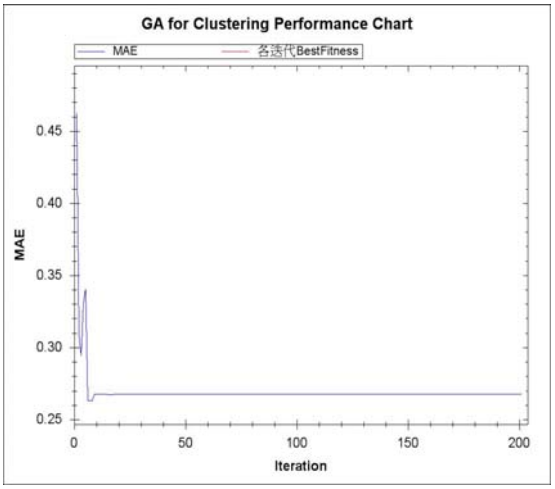

圖 4-6 Glass 5%遺漏值(B)收斂圖

由圖 4-5 與 4-6 可看出，(A)屬性值平均+基因分群方法在迭代演化過程中不會有明顯的收斂效果，而且其跳動明顯，可能在執行 200 迭代之後還未有收斂的現象；反觀(B)屬性值切割+基因分群方法在 10 個迭代之內即呈現收斂的現象，而其估計的誤差值並無太大的差異，並且更能得到比(A)還要優良的推估結果，意即我們提出的方法能夠在較少的迭代內就能得到較佳的估計值，節省演算時間。

四、Vowel 資料集收斂效果

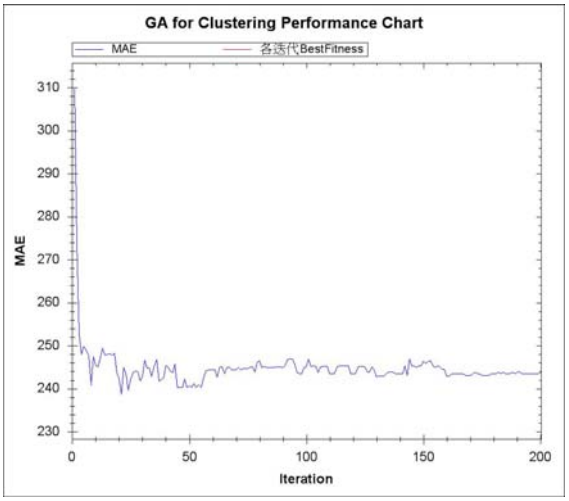

圖 4-7 Vowel 5%遺漏值(A)收斂圖

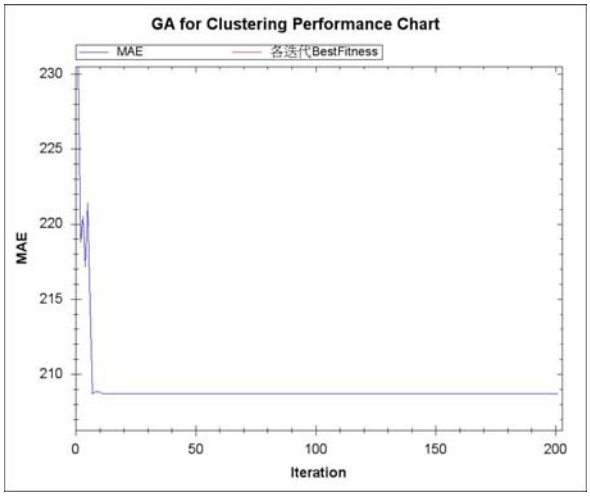

圖 4-8 Vowel 5%遺漏值(B)收斂圖

由圖 4-7 與 4-8 可看出，(A)屬性值平均+基因分群方法在迭代演化過程中不會有明顯的收斂效果，可能在執行 200 迭代之後還未有收斂的現象，造成其搜尋與運算的時間過長；反觀(B)屬性值切割+基因分群方法在 20 個迭代之內即呈現收斂的現象，而其估計的誤差值反而比(A)方法要來得優良，意即我們提出的方法能夠在較少的迭代內就能得到較佳的估計值，節省演算時間，增進遺漏值的估計效率。

第五節 實驗討論

經由觀察上述所呈現四個真實資料集其不同遺漏值比率的實驗結果後，可以發現資料集大小與其數值分佈特性容易影響各遺漏值推估方法的估計結果。在 Iris 資料集中，RKPSO 和屬性值切割+基因分群方法對於較大遺漏值比率資料集皆有不錯的推估效果；在 Crude Oil 和 Glass 資料集中，兩種初始值回填方法+基因分群方法皆有良好的估計效果，而且其特色在於每次估計遺漏值都能維持穩定的回填結果，在較大的 Vowel 資料集中，RKPSO 和我們提出的屬性值切割+基因分群方法皆能得到良好的推估結果，並無太大的差異性。而我們加入屬性值切割方法確實能比傳統以屬性平均值當作初始值的方法更能得到較佳的估計值，讓回填的資料更接近真實值。

另外在實驗中所記錄的兩種 MAE，經過實驗數據發現，其適應值的優劣對於 MAE 並無明顯的相對關係，意即適應值愈低其 MAE 不一定就會降低，而屬性值切割+基因分群方法中兩種 MAE 都是相同的數值，表示適應值最佳的迭代即代表著整體演算法已經趨於收斂，所以我們能夠利用最終演算法收斂的結果來得到較佳的估計效果，並且能適時縮短演算時間，而且能夠比傳統基因分群法更易取得正確的 MAE。

依屬性值平均+基因分群方法與屬性值切割+基因分群方法的收斂圖來看，屬性值平均+基因分群方法在我們設定的 200 個迭代內還未呈現收斂的現象；而加入屬性值切割方法後，平均不到 20 迭代即趨向收斂狀態，並且得到與屬性值平均+基因分群方法差不多的估計數值，甚至更佳的推估效果。其原因為屬性值切割法比以傳統屬性值平均回填初始值的效果來得更佳，讓之後的基因分群方法能夠快速的得到較佳

的分群結果，縮短運算時間。依實驗結果整體來看，加入屬性值切割法的基因分群方法其估計遺漏值的準確率普遍比未加入屬性切割法的基因分群方法要來得更好；加入此方法的優點在於能夠減少整體遺漏值推估模組的執行時間，執行較少的迭代數即能得到較佳的估計值。

第五章 結論與未來研究方向

第一節 研究結論

本論文利用群集分析和啟發式演算法的概念為基礎，應用屬性值切割法和基因分群技術來解決遺漏值估計的問題，期望以屬性值切割法能夠自動將資料的分佈作空間上的切割，找出屬性之間的關聯性來改善分群的品質，再來藉由基因演算法全域隨機多點搜尋的特色，找出最佳的分群結果來提昇遺漏值推估的準確率。

經過四個真實資料的實驗分析後，可歸納出以下幾點結論：

一、 屬性值切割法+基因分群技術能應用於真實資料庫的遺漏值推估問題上，並且其它遺漏值推估方法進行 MAE 比較，大致來說可獲得比 k-means 和屬性平均值+基因分群方法更佳的估計效果，並且在 Glass 資料集中其估計效果優於 RKPSO。

二、 實驗中基因分群為基礎的遺漏值估計法採取記錄兩種 MAE 的方式；一個是記錄迭代中最佳的適應值，並擷取當代遺漏值回填的 MAE 數值；另一種方式是記錄基因分群演算法收斂時所得到的 MAE 數值，經由實驗發現其適應值的優劣對於 MAE 並無明顯的相對關係，不過屬性值切割+基因分群方法中兩種 MAE 都是相同的數值，表示適應值最佳的迭代即代表著整體演算法已經趨於收斂，能適時縮短演算時間，並且能夠比傳統基因分群法更易取得正確的 MAE 數值。

三、 分析屬性值平均+基因分群技術與屬性值切割法+基因分群技術兩者的收斂效果變化，可以清楚的發現加入屬性值切割方法後，

平均不到 20 迭代即趨向收斂狀態，並且得到與屬性值平均+基因分群方法差不多的估計數值，甚至更佳的推估效果。依推估結果的 MAE 來看，加入屬性值切割法的基因分群方法其估計遺漏值的準確率普遍比未加入屬性切割法的基因分群方法要來得更好，能夠減少整體遺漏值推估模組的執行時間，執行較少的迭代數即能得到較佳的估計值。本研究提出之方法因為使用屬性值切割方法其應用到屬性值之間的關聯性較多，反之，屬性值平均+基因分群則應用較少，其簡易分析比較如表 5-1 所示。

表 5-1 兩種初始值回填方法+基因分群方法之比較表

	屬性值平均+基因分群技術	屬性值切割法+基因分群技術 (本研究提出之方法)
估計誤差	較高	較低
收斂速度	緩慢	快速
執行迭代	較多	較少
應用屬性值之間的關聯性	較少	較多

第二節 未來研究方向

壹、 加入動態分群技術

由於本論文提出的基因分群技術是以 k-means 為基礎，必須事先給定分群數，才能進行分群的動作，顯得沒有彈性；而且在現實生活中的資料庫中我們並不曉得其真實分群數為多少，容易造成遺漏值推估的困難。所以之後可以應用動態分群的技術，讓資料集依據其分群指標自動分成適當的群集，以符合多樣化的資料庫，保持其彈性且能改善分群的品質。

貳、 利用模糊分群概念

由於本論文分群方式是採取硬分群，一筆資料只可以屬於某一群集，不過在真實資料集中卻存在著模糊地帶，一筆資料可能分屬於 N 個群集，所以我們可以加入模糊理論的觀念，讓其資料可以被分到適合的群集之中，讓分群結果能夠更接近真實狀況。

參考文獻

中文部份

[1] 李建逸,「基於間隙法與K-means分群法之遺漏值推估模式」，南華大學資訊管理學系研究所碩士論文，94年6月。

[2] 沈永勝,「整合自動分群技術與加權式灰關聯技術於大型資料庫內遺失值之處理」，國立臺灣科技大學電子工程學系研究所碩士論文，94年6月。

[3] 邱宏彬、吳文盛及林宜德,「基於基因演算法分群技術之遺漏值推估模組」，八十五週年校慶暨第十六屆三軍官校基礎學術研討會，高雄，cs-141~cs148，98年5月。

[4] 林俊男,「應用類神經網路法於遺漏值問題之研究」，南華大學資訊管理學系研究所碩士論文，94年6月。

[5] 林如梅,「整合遺傳演算法和粒子群最佳化演算法於分群分析之研究」，國立臺北科技大學工業工程與管理系研究所碩士論文，97年6月。

[6] 游裕昌，「應用基因群集技術於大型資料庫內遺失值之處理」，南華大學資訊管理學系研究所碩士論文，93年6月。

[7] 魏岑甄,「基於反彈機制KPSO分群之有效遺漏值推估方法」，南華大學資訊管理學系研究所碩士論文，97年6月。

[8] 曾憲雄著，*資料探勘 (Data Mining)*，台北，旗標出版社，94 年。

[9] 蘇木春、張孝德著，*機器學習：類神經網路、模糊系統以及基因演算法則*，全華科技圖書，89 年。

西文部份

[10] C.H. Cheng and J.W. Wang ” A new approach for estimating null value in relational database”,*Soft Comput* 10*,* pp.104-114 (2006).

[11] S. Bandyopadhyay and U. Maulik, “Genetic clustering for automatic evolution of clusters and application to image classification”, *International Journal of Pattern Recognition*, 35 , pp.1197-1208 (2002).

[12] S.M. Chen and H.R. Hsiao ” A New Method to Estimate Null Values in Relational Database Systems Based on Automatic Clustering Techniques”, *Information Sciences* 169*,* pp.47-69 (2005).

[13] Y.T. Kao*,* Erwie Zahara and I-W. Kao ” A hybridized approach to data clustering”, Expert Systems with Applications ,34, pp.1754-1762 (2005).

[14] U. Maulik and S. Bandyopadhyay “Genetic algorithm-based clustering technique”, *International Journal of Pattern Recognition*, 33 , pp. 1455-1465 (2000).

[15] Little, R. J. A. and Rubin, D. B. Statistical analysis with missingdata. New York：Wiley. (1987).

[16] Y.T. Kao,Erwie Zahara and I.W. Kao ” A hybridized approach to data clustering”,*Expert Systems with Applications* 34*,* pp.1754-1762 (2008).

[17].M. Negnevitsky. Artificial Intelligence: A Guide to Intelligent Systems (Second Edition). Addison Wesley, New York (2005).

[18] H. R. Hsiao and S. M. Chen, "A new automatic clustering algorithm for fuzzy query processing," Proceedings of the 6th Conference on Artificial Intelligence and Applications, Kaohsiung, Taiwan, Republic of China, pp. 550-555, November 2001.

[19] Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery: An Overview,” *In: Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, (1996).

[20] Periklis, A., “Data Clustering Techniques”, March 2002.

URL: http://www.cs.toronto.edu/~periklis/pubs/depth.pdf.

[21] M. Dorigo, M. Birattari, and T. Stutzle. “Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique,” IEEE Computional Intellige- nce Magazine, November 2006, pp. 28-39.

[22] http://archive.ics.uci.edu/ml/.

[23] http://www.stat.ualberta.ca/~mizera/441/oils.d.

[24] http://www.isical.ac.in/~sushmita/patterns/vowel.dat.