

南 華 大 學

資訊管理學系

碩士論文

基於屬性值分割與蟻群分群技術之遺漏資料處理

Missing data processing based on
attribute values partitioning and ant colony clustering

研究生：李進鴻

指導教授：邱宏彬 教授

中華民國 九十八 年 六 月 二十日

南 華 大 學
資 訊 管 理 學 系
碩 士 學 位 論 文

基於屬性值分割與蟻群分群技術之遺漏資料處理

研究生： 李進鴻

經考試合格特此證明

口試委員： 謝忠豪
李翔詣
邱宏彬

指導教授： 邱宏彬

系主任(所長)： 鍾國貴

口試日期：中華民國 98 年 6 月 20 日

基於屬性值分割與蟻群分群技術之遺漏資料處理

Missing data processing based on
attribute values partitioning and ant colony clustering

研究生：李進鴻

Student: Chin-Hung Li

指導教授：邱宏彬 博士

Advisor: Dr. Hung-Pin Chiu

南 華 大 學

資 訊 管 理 學 系

碩 士 論 文

A Thesis

Submitted to Department of Information Management

College of Science and Technology

Nan-Hua University

in partial Fulfillment of the Requirements

for the Degree of

Master of Information Management

June 2009

Chaiyi Taiwan, Republic of China.

中華民國 九十八 年 六 月 二十日

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： 李進鴻 之碩士畢業論文

中文題目：基於屬性值分割與蟻群分群技術之遺漏資料處理

英文題目：Missing data processing based on attribute values partitioning
and ant colony clustering

指導教授： 邱宏彬 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： 李進鴻 (請親自簽名)

指導老師： 邱宏彬 (請親自簽名)

中 華 民 國 98 年 6 月 20 日

南華大學碩士班研究生

論文指導教授推薦函

資訊管理學系碩士班李進鴻君所提之論文
基於屬性值分割與蟻群分群技術之遺漏資料處理
係由本人指導撰述，同意提付審查。

指導教授



18年6月20日

誌 謝

在南華經歷了六個寒暑，終於到了學生時代的最後一個盛夏，雖然早已過了鳳凰花開的季節，但對於這六年來所遇見的所有人、事、物仍有著濃濃的不捨之情…

感謝家人的奧援，讓我的心靈與物質生活能夠不虞匱乏；

感謝邱老師的指點，讓我能在漫無頭緒之際找出論文的著力點；

感謝伊汝姐的器重，讓我知道自己也能承擔某些重責大任；

感謝研究夥伴毛利的砥礪，讓我得以掌握論文進度並如期完成；

感謝宜德的激發，讓我得到了許多創意發想與生活調劑；

感謝美秀的詢問，讓我感受到了教學相長的可貴；

感謝 204 所有成員，讓我在單調的研究生活中增添了不少樂趣；

感謝六年來的所有人、事、物，讓我相信每次的際遇都是奇蹟！

漫長的人生遊戲之學生時代關卡至此即將告一段落，過程中偶有艱辛、偶有挫折，但也相對地真切踏實！謝謝陪伴我渡過這段時光的所有人，相信這段時光絕對會是我生命旅程中最為青春、珍貴而且難忘的一頁！

進鴻 2009 年盛夏

基於屬性值分割與蟻群分群技術之遺漏資料處理

學生：李進鴻

指導教授：邱宏彬 博士

南 華 大 學 資 訊 管 理 學 系 碩 士 班

摘 要

資料探勘的過程中，資料的遺漏或缺失可能會使得探勘的結果產生異常與偏誤，導致組織決策判斷錯誤，進而造成企業經營績效的損失。因此該如何精準地估計並填補這些遺漏資訊，正是資料探勘的前置處理作業中相當重要的課題之一。本研究基於屬性值分割方法以及蟻群分群演算法，發展出一套完整的遺漏值推估模組，藉以估計資料屬性中的遺漏資訊。我們利用蟻群分群方法為基礎並且加以改良，提出改良式蟻群分群演算方法，並且將其應用於遺漏值推估問題上。

利用分群演算法所發展的遺漏值推估方法，主要是以物以類聚的資料特性，藉由資料分群取得群集相關資訊，用以作為遺漏值估計回填的依據。而從我們的實驗結果顯示，改良式蟻群分群方法確實有助於提昇解決分群問題的效率，此外在遺漏值推估部份，不論是屬性值分割或是蟻群分群方法之遺漏值推估模式，遺漏值的估計回填效果都要較傳統利用屬性平均值回填來的準確。

關鍵字：遺漏值、群集分析、蟻群演算法、屬性值分割

Missing data processing based on attribute values partitioning and ant colony clustering

Student: Chin-Hung Li

Advisors: Dr. Hung-Pin Chiu

Department of Information Management

The M.I.M. Program

Nan-Hua University

Abstract

Data mining is an important technique to extract useful knowledge from a set of raw data. The managers can exploit the mining knowledge to make right decisions. However, missing data significantly distort data mining results. Therefore, data preprocessing of missing values becomes extremely critical in successful data mining. Data clustering techniques is the partitioning of a dataset into clusters so that the data records in each cluster possess common characteristics. The shared characteristics can be utilized to predict the missing values. In this study, we propose an attribute values partitioning technique to preserve the relationships between attributes for estimating missing values. On the other hand, ant colony optimization (ACO) algorithm was recently proposed by few researchers to solve data clustering problems. In this study, we propose an improved ACO clustering approach, and employ the ant clustering as a basis to estimate the missing data. Furthermore, we integrate the attribute values partitioning with the ant 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 improved ant clustering algorithm outperforms the previous methods in clustering quality, and the integrated missing data processing approach provides competitive results or performs well compared with the existing methods.

Keyword: data mining, missing value estimation, clustering analysis, attribute values partitioning, ant colony optimization algorithm.

目 錄

論文口試合格證明.....	i
書名頁.....	ii
著作財產權同意書.....	iii
論文指導教授推薦書.....	iv
誌謝.....	v
摘要.....	vi
目錄.....	viii
表目錄.....	x
圖目錄.....	xii
第一章 緒論	1
第一節、研究背景	1
第二節、研究動機與目的	2
第三節、研究流程	4
第四節、論文架構	4
第二章 文獻探討	6
第一節、遺漏值處理	6
第二節、群集分析問題	9
第三節、蟻群最佳化演算法	13
第三章 改良式蟻群分群演算法	22
第一節、改良式蟻群分群演算法	22
第二節、改良式蟻群分群演算法演算流程	25
第四章 遺漏值推估方法	34
第一節、基礎遺漏值推估模組	34
第二節、基於屬性值分割方法之遺漏值推估模組	37
第三節、基於蟻群分群方法之遺漏值推估模組	45
第四節、基於屬性值分割與蟻群分群方法之遺漏值推估模組	51
第五章 實驗設計與結果分析	53
第一節、實驗環境	53
第二節、實驗參數設定	53
第三節、實驗資料描述	54
第四節、實驗設計	58
第五節、實驗結果分析	59
第六章 結論與未來展望	70
第一節、結論	70

第二節、未來展望	71
參考文獻	72

表 目 錄

表 3-1 分群資料範例.....	26
表 3-2 費洛蒙矩陣範例.....	27
表 3-3 分群結果表示範例.....	28
表 3-4 分群結果 S1 的群心位置.....	29
表 3-5 加入距離重新建構的分群結果 S1.....	29
表 3-6 分群結果與 fitness 值表示.....	30
表 3-7 Local Search 後的分群結果表示.....	31
表 3-8 解答編碼排序範例.....	32
表 3-9 表 3-8 中解答 S1 更新費洛蒙矩陣範例.....	33
表 4-1 半導體銅製程原始資料集.....	37
表 4-2 半導體銅製程遺漏資料集.....	38
表 4-3 半導體銅製程個別屬性排序結果.....	40
表 4-4 半導體銅製程之均勻度屬性分割結果.....	41
表 4-5 屬性值分割資訊資料集.....	42
表 4-6 屬性值分割資訊.....	43
表 4-7 屬性值分割後結果的資料原始值範例.....	44
表 4-8 屬性值分割遺漏值回填結果.....	44
表 4-9 遺漏值推估回填方法之誤差比較.....	45
表 4-10 半導體銅製程遺漏資料集-以屬性平均值回填初始值.....	46
表 4-11 首次迭代分群結果.....	48
表 4-12 首次迭代分群結果群心位置範例.....	49
表 4-13 首次迭代後遺漏值回填結果範例.....	49
表 4-14 遺漏值回填結果.....	50
表 4-15 遺漏值推估回填方法之誤差比較.....	51
表 4-16 遺漏值推估回填方法之誤差比較.....	52
表 5-1 參數設定內容.....	54
表 5-2 實驗資料集內容.....	55
表 5-3 Crude Oil 資料集屬性資訊.....	55

表 5-4 Crude Oil 資料集群集資訊.....	55
表 5-5 Iris Plants 資料集屬性資訊.....	56
表 5-6 Iris Plants 資料集群集資訊.....	56
表 5-7 Glass 資料集屬性資訊.....	56
表 5-8 Glass 資料集群集資訊.....	57
表 5-9 Vowel 資料集屬性資訊.....	57
表 5-10 Vowel 資料集群集資訊.....	58
表 5-11 Crude Oil 分群之實驗結果比較.....	59
表 5-12 Iris Plants 分群之實驗結果比較.....	60
表 5-13 Glass 分群之實驗結果比較.....	60
表 5-14 Vowel 分群之實驗結果比較.....	61
表 5-15 Crude Oil 5%遺漏值比例之實驗結果比較.....	62
表 5-16 Crude Oil 10%遺漏值比例之實驗結果比較.....	63
表 5-17 Crude Oil 15%遺漏值比例之實驗結果比較.....	63
表 5-18 Crude Oil 20%遺漏值比例之實驗結果比較.....	63
表 5-19 Iris Plants 5%遺漏值比例之實驗結果比較.....	64
表 5-20 Iris Plants 10%遺漏值比例之實驗結果比較.....	64
表 5-21 Iris Plants 15%遺漏值比例之實驗結果比較.....	65
表 5-22 Iris Plants 20%遺漏值比例之實驗結果比較.....	65
表 5-23 Glass 5%遺漏值比例之實驗結果比較.....	66
表 5-24 Glass 10%遺漏值比例之實驗結果比較.....	66
表 5-25 Glass 15%遺漏值比例之實驗結果比較.....	66
表 5-26 Glass 20%遺漏值比例之實驗結果比較.....	67
表 5-27 Vowel 5%遺漏值比例之實驗結果比較.....	67
表 5-28 Vowel 10%遺漏值比例之實驗結果比較.....	68
表 5-29 Vowel 15%遺漏值比例之實驗結果比較.....	68
表 5-30 Vowel 20%遺漏值比例之實驗結果比較.....	68

圖 目 錄

圖 1-1 研究流程圖.....	5
圖 2-1 階層式分群方法示意圖.....	11
圖 2-2 K-means 演算法執行步驟.....	12
圖 2-3 費洛蒙機制概念示意圖.....	14
圖 2-4 蟻群分群結果編碼方式示意圖.....	19
圖 2-5 蟻群分群演算法流程圖.....	21
圖 3-1 Local Search 更新示意圖.....	23
圖 3-2 分群結果編碼示意圖.....	24
圖 3-3 解答編碼順號示意圖.....	24
圖 3-4 改良式蟻群最佳化演算法流程圖.....	25
圖 4-1 基礎遺漏值推估模組架構圖.....	35
圖 4-2 屬性值分割方法之遺漏值推估模組流程圖.....	39
圖 4-3 蟻群分群遺漏值推估模組架構圖.....	46
圖 4-4 蟻群分群方法之遺漏值推估模組流程圖.....	47
圖 4-5 蟻群分群遺漏值推估模組架構圖.....	52
圖 5-1 Glass 資料集分群之 fitness 收斂圖.....	61

第一章 緒論

此章節主要描述本論文的研究背景、動機與目的、研究方法流程以及本論文的整體架構。

第一節、研究背景

隨著資訊科技演進，傳統以書面紙本形式來紀錄大量資料的作法，也逐漸被資料庫管理系統所取代。近年來，由於資料庫管理系統的廣泛應用，人們從原先資料管理的需求，漸漸延伸為資訊的過濾與萃取，資料探勘（Data Mining）技術也就此應運而生。

對於組織企業而言，資料庫管理系統的使用不僅能大幅增進日常交易與管理資料的紀錄及處理效能，達到快速且有效的資訊整合效果，輔以資料探勘技術的應用，更能協助組織從大量且龐雜的資料中，挖掘出潛藏在既有資料紀錄裡，具有其重要價值卻尚未被發現的關鍵資訊，並提供組織管理人員作為決策參考之用，藉以提昇企業組織之經營管理績效。

而在資料探勘的過程中，資料的遺漏或缺失可能會使得探勘的結果產生異常與偏誤，導致組織決策判斷錯誤，進而造成企業經營績效的損失。但在資料搜集的階段，卻往往因為人為疏忽或是軟硬體系統運作異常，以致所紀錄的資料產生誤差、甚至是缺失，我們通稱為遺漏值問題（Missing Value Problem），該如何精準地估計並填補這些遺漏資訊，正是資料探勘的前置處理作業中相當重要的課題之一。

第二節、研究動機與目的

高品質的探勘結果通常來自於高品質的資料[11]，因此我們在正式進行資料探勘之前，必須先將資料經過一番整合、清理與轉換，排除多餘或不完整的資料，並將資料格式作一致性的統整，經由這些資料前置處理的進行來提昇整體資料的品質，才得以期待資料探勘的結果能夠具有其參考價值。倘若所搜集的資料內容出現了遺漏，有可能會影響整體探勘的品質與結果，因此遺漏值填補方法一直是資料探勘領域中相當常見的討論與應用之一。

而面對遺漏值的處理，傳統上有幾種常見的作法：

1. 直接忽略

直接忽略指的是直接將含有遺漏值的資料整筆移除，這是最直接也最快速的處理方式，但是倘若具有遺漏值的資料筆數大於一定比例時，採用直接忽略方法將會造成資料大量流失，使得所能挖掘的資料量縮減，以致於無法正確且詳實的探勘出有用的參考資訊。因此直接忽略方法較為適用在探勘資料數量龐大，且具有遺漏值之資料所佔比例較少之情況。

2. 人工填補

有時候我們也會利用人工方式，藉由個人主觀判斷來填補遺漏的資料，但此方法牽涉到個人觀感與眾多外在因素之影響，容易使得遺漏值回填的標準不一致，導致遺漏值的估計效果不佳；而且人工填補方式不僅耗費人力、曠日費時，難以快速處理大量遺漏資料也是主要缺點之一。

3. 以平均數或眾數回填

以完整資料之平均數或是眾數來回填遺漏值是日前較為常見的作法，利用既有的完整資訊作為參考，並且估計出適當的填補資訊用以回填。這類作法由於參考了完整資料的資訊，因此所估計之遺漏值通常不致於出現過大的偏誤，也由於明確的回填值估計方式，相較於人工填補方式而言也會來的穩定許多。

除了上述幾項傳統作法以外，近年來也陸續出現了以各種不同技術為基礎的遺漏值估計方法，如群集分析 (Cluster Analysis) [12][10]、類神經網路 (Artificial Neural Networks) [9]、回歸分析 (Regression Analysis)、模糊規則 (Fuzzy Rules)、灰關聯分析 (Gray Relational Analysis) [8]、關聯法則 (Association Rules) …等，藉由這些技術所發展出的遺漏值估計方法，相較於傳統作法而言通常都具有更好的估計成效。

本研究的主要目的，正是利用群集分析中物以類聚的概念，來解決遺漏值估計與填補問題。我們利用蟻群最佳化 (Ant Colony Optimization) 演算法為基礎概念所發展之群集分析方法，將原始資料區分為多個群集，並根據各個群集的屬性特徵以作為遺漏值資料的回填參考，發展出一套完整的遺漏值推估方法，並且結合了屬性值分割方法 (attribute values partitioning) 之概念，利用原始資料的個別屬性進行初步的分割，將屬性分割的結果匯集形成個別群集，並以屬性值分割的分群結果作為蟻群遺漏值推估模組的初始回填資訊，期待能以更有效率且更加準確的方式，達到資料前置處理作業中遺漏值估計之目的。

第三節、 研究流程

我們首先確立本研究的主要問題方向與解題目的，決定以蟻群分群方法以及屬性值分割方法為研究主軸，用以解決遺漏值填補問題。由於蟻群最佳化演算法具有正向回饋（Positive Feedback）與貪婪啟發式演算法（Greedy Heuristic）等特性，在搜尋解答的過程中能夠不斷的累積資訊，提昇整體求解效率，因此經常被應用在求解組合最佳化問題之上，而我們正是利用這樣的特性，並且將其方法延伸用以解決遺漏值估計的問題。而屬性值分割方法是根據個別維度屬性的分佈資訊作為分割之參考，無須指定分割數量，因此具有動態分割的特性。我們在研究中企圖藉由兩項方法的結合，提昇遺漏值推估模組的求解效率與品質。

在文獻回顧階段，根據我們所採用的研究方法來彙整並探討相關文獻，包含了遺漏值估計填補問題、群集分析方法與蟻群最佳化演算法三大主軸。之後透過多個資料集與不同演算方法的實驗，藉以分析比較本研究方法在分群收斂效果與遺漏值回填之成效。最後彙整本研究的研究結論，並提出未來展望與研究建議，如圖 1-1 所示。

第四節、 論文架構

本論文章節架構如下，第二章為文獻探討部份，我們將回顧並探討有關遺漏值處理、群集分析方法以及蟻群最佳化演算法的相關文獻。第三章為本研究的研究方法介紹，我們利用蟻群分群演算法為基礎並將之改良，以及基於屬性值分割之遺漏值推估模組，發展出我們的遺漏值推估方法。第四章則為實驗設計與結果分析，以四個不同資

料庫作為實驗目標，分別比較我們所提出的方法與傳統估計方式之差異，以及不同的遺漏值推估模組組合之應用成效。第五章將討論本研究的結論以及未來展望，並提出日後相關研究的發展建議。

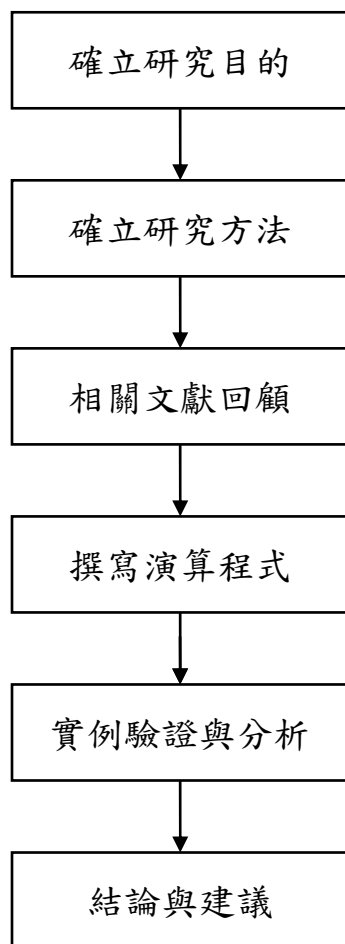


圖 1-1 研究流程圖

第二章 文獻探討

第一節、遺漏值處理

壹、遺漏值定義與類型

正如前述所說，高品質的探勘結果通常來自於高品質的資料，而高品質的資料則來自於妥善的資料前置處理程序，因此資料前置處理結果的好壞對於探勘結果有著一定程度的影響。資料前置處理包含了資料的整合（Integration）、清理（Cleaning）與轉換（Transformation）三項工作[11]，而其中資料清理階段主要是為了確保資料本身的完整性，以確保資料探勘過程能夠順利進行，而遺漏值的推估方法一直是資料清理過程中相當重要的主題之一。

在資料搜集的過程中，難免會遇到由於人為因素、各種軟硬體設備異常或是各種外在因素所造成的資料遺漏與缺失，若我們直接以這些包含著各項不確定遺漏資訊的原始資料來進行資料探勘活動，則很有可能會使得探勘結果產生偏誤，連帶影響了組織企業的決策制定，進而導致經營管理上的損失。主要的資料遺漏可分為以下幾種類型[9]：

1. 空數值（Empty Value）

空數值的產生通常是在資料搜集的過程中原本就未取得的部分。例如涉及到個人隱私的相關資訊，通常較為容易出現遺漏。

2. 不存在的數值（Nonexistent Value）

而數值不存在的情況，是由於所分析問題之特性所造成。例如想要得知公司各部門最近五年的人事營運成本，但是今年才成立的

部門並沒有過去的歷史紀錄，因而產生了遺漏值，導致無法達到資料探勘應有的效果。

3. 不完整的資料 (Incomplete Data)

當分析問題與取得資料無法配合時，則可能會發生資料不完整的狀況。譬如想要分析所有客戶的信用卡與現金卡兩者的使用比例，但並非所有客戶都有使用這兩項服務，使得無法正確取得探勘資料。

4. 未被搜集的資料 (Uncollected Data)

這是指在資料搜集過程中從未被紀錄的資料，其原因多半是因為在資料搜集設計時就沒有將其加入紀錄考量。像是在圖書借閱資料中通常僅有借閱日期，而沒有詳細借閱時間的紀錄。

貳、遺漏值處理方法

我們在第一章已經提到了幾種傳統的遺漏值處理方式，以下將介紹幾種有別於傳統遺漏值處理的應用技術。

1. 類神經網路 (Artificial Neural Networks)

類神經網路的主要精神為模擬生物神經網路的訊息傳導機制，以大量且簡單的神經元來傳遞資訊，並且藉由類神經網路的學習機制，利用範例資料作為學習目標，進而達到能夠計算複雜問題的能力。

在文獻[9]中的作法，則是利用類神經網路技術對範例資料作群集分析之學習，以分群後的結果作為回填之遺漏值，並且重複直到所估計之回填值穩定為止。嚴格說來這樣的作法仍然屬於群集分析的概念應用之一，也由於類神經網路方法需要透過範例資料來作

為學習目標的特性，因此在學習資料取得不易的情況下，可能較難以靈活應用於一般遺漏值問題之上。

2. 群集分析 (Clustering Analysis)

利用物以類聚之概念對資料進行群集分析，其後在以分群之結果作為具有遺漏值資料之回填數值參考。相較於傳統回填整體資料平均值的方式，利用資料之間的相似程度作為分群依據，並以分群結果中個別群集之群心值作為回填值，通常會具有比較良好的遺漏值估計效果。

例如文獻 [10][12] 中分別利用了基因演算法 (Genetic Algorithms, GA) 與粒子群最佳化演算法 (Particle Swarm Optimization, PSO) 對原始資料作群集分析處理，並以分群結果估計可能的遺漏值；由上述兩者的實驗結果可以發現，以分群為主的遺漏值估計方式確實要比傳統估計方法來的更加精確。此類以群集分析為基礎的遺漏值估計方法，對於具有明顯群集分佈現象的資料會有較佳的處理成效，但對於分群效果不顯著的資料而言，其遺漏值估計效果可能會因為群集分佈不夠明確，以使得無法從分群結果中取得精確的遺漏值估計結果。因此如何取得良好的分群結果，並藉以作為回填遺漏估計值，是此類方法的重要目標之一。

3. 灰關聯分析 (Grey Relational Analysis)

以灰關聯分析來估計遺漏值的作法，其精神主要是將具有遺漏值的資料與其他完整無遺漏的資料相互作比較，分別計算出與遺漏資料之灰關聯係數，其係數越高就代表該資料與包含遺漏值的資料越接近，並取灰關聯係數較高的資料作為遺漏值估計的參考，藉以

作為回填數值。

除了以傳統的灰關聯分析方法來估計遺漏值以外，在文獻[8]中亦將灰關聯分析加以延伸應用，提出了利用灰關聯分析進行資料分群，進而作為遺漏值估計回填基礎之方法，用以解決原始灰關聯分析方法中，每估計一次遺漏值就要重新計算的缺點；其實驗結果也證明在遺漏值處理之問題上，該方法要較傳統的灰關聯分析方式要來的優良。

第二節、群集分析問題

壹、群集分析定義

群集分析 (Cluster Analysis) 又稱為資料分群 (Data Clustering) 問題，最主要的目的就是將資料在未指定分群準則的情況下，藉由比較各個資料點彼此之間的相似程度，將原始資料區分為多個不同的群集 (Cluster)；由於群集分析無須事先指定分群準則，因此也被稱為非監督式分類 (Unsupervised Classification) [11]。

在群集分析問題中，形成群集的主要依據是根據資料之間的相似程度，資料之間的相似程度越高，則代表彼此歸屬在相同群集的可能性越大，相似程度越低則越不可能被分到同樣的群集中，因此如何計算相似度在分群問題中就顯得相當重要。

常見的相似度測量方法有歐幾里得距離 (Euclidean Distance，又稱為尤拉距離) 與曼哈頓距離 (Manhattan Distance) 兩種，簡單介紹如下：

1. 歐幾里得距離

歐幾里得距離為兩個資料點之間的最短直線距離，利用資料中各個維度的距離之平方總和並開平方根計算而得。

$$d_2(x_i, x_j) = \left(\sum_{d=1}^k |x_{id} - x_{jd}|^2 \right)^{1/2} \quad (2-1)$$

公式為資料點 x_i 與資料點 x_j 之距離，其中 k 表示資料維度數。

2. 曼哈頓距離

不同於歐幾里得距離，曼哈頓距離則是將兩個資料點之間，所有維度距離作加總計算而得之。

$$d_M(x_i, x_j) = \sum_{d=1}^k |x_{id} - x_{jd}| \quad (2-2)$$

由於以上兩種相似度計算公式的定義簡單明確，因此廣泛被應用於群集相似度的計算中。

利用上述計算公式比較資料之間的相似度，將相似程度較高的資料歸類為同一群集，相似度越低則歸屬於不同群集，進而形成同群內資料相似程度高、異群之間資料相異程度高的群集分析結果。

貳、分群問題解決方法

前面已經討論過群集分析的基本定義與精神，接著我們將分別介紹階層式 (Hierarchical) 分群與分割式 (Partitional) 分群兩種不同的分群演算方法。

1. 階層式分群法 (Hierarchical)

階層式分群方法的主要作法，是根據資料彼此之間的相似程度來決定是否聚合 (Agglomerative) 成為同一群或是分裂 (Divisive)

為兩個不同群集。圖 2-1 為階層式分群方法示意圖。

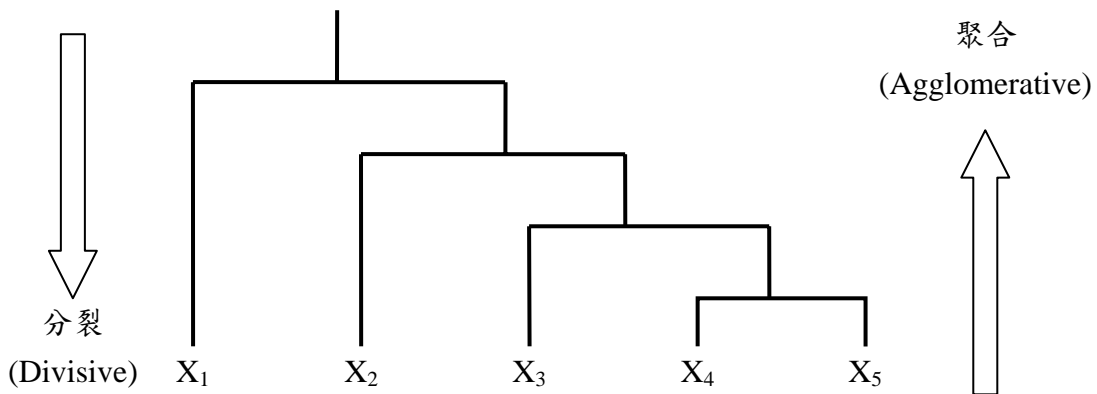


圖 2-1 階層式分群方法示意圖

聚合式的作法是先將個別資料點視為獨立一個群集，根據資料相似程度逐漸將兩筆最相似的資料合併為一群，直到所有資料形成一個群集，或是滿足群集分析之群數為止，是一種由下而上 (bottom-up) 的分群作法。而分裂式的作法正好與聚合式相反，是由上而下 (top-down) 的分群方式，先將所有資料視為同一群集，再依據資料之間的相似程度，將最不相似的資料切割成為另一個群集，直到滿足指定群集數量或是所有資料各自成為一群為止。

2. 分割式分群法 (Partitional)

不同於階層式分群的作法，分割式分群方法的精神，則是利用群集之群心 (Cluster Center) 位置作為相似度判斷的依據，透過反覆迭代更新不斷修正群集分析結果，其目的在於使得每一筆資料與該所屬群心能擁有最高的相似度，藉此形成多個群集並且滿足指定的分群條件。以下我們以分割式演算法中相當著名的 K-means 分群方法[1]為例，介紹分割式演算法的主要運作流程。

而由於分群的結果可視為是資料點的配對組合，具有組合最佳化問題之特性，因此群集分析問題亦屬於 NP-hard 問題的一類。此類 NP-hard 問題的複雜度會隨著選項的增加，使得求解組合呈現指數型的成長，一般而言難以在合理的時間範圍之內找到最佳的解答，因此近年來的相關研究大多採用啟發式（Heuristic）方法來求解此類問題，以期能在合理的時間內求得可以接受的較佳結果。

K-means 分群演算法的概念其實相當單純，最主要的精神是以群集之群心位置作為基準，比較每一筆資料與各個群心的距離，並且將資料歸屬到距離最近的群集中；當所有資料皆分群完畢之後重新計算各群集的群心位置，再次檢查各資料點與群心的距離，重新進行分群。根據上述步驟反覆執行，直到群心位置不再變動為止，即可得到最終的分群結果。

K-means

輸入：給定一個包含 n 個資料點的資料集合，與目標分群數 k 。

步驟一：從資料集中隨機選取 k 個資料點作為起始群心。

步驟二：計算所有資料與個別群心之距離，並將資料歸屬到距離最近之群中。

步驟三：重新計算分群完畢之後的群心位置。

步驟四：重複步驟二與步驟三，直到群心位置不再變動為止。

輸出：各資料分群結果。

圖 2-2 K-means 演算法執行步驟

由於 K-means 分群方法十分簡潔明瞭，分群效率也相當快速，因此這樣的分群概念常被應用於不同分群方法中，作為方法改良與延伸之參考。但 K-means 分群方法仍然存在著部份缺點，像是由於分群收斂速度過快，且僅以資料和群心距離作為分群的判斷依據，

因此一旦落入局部最佳解時，將會難以跳脫出來，導致無法取得更加優良的解答結果。此外，由於在 K-means 中，各個群集的起始群心是以隨機選擇的資料點來決定，假若挑選到不佳的資料點作為起始群心，例如離群值，很可能會導致整體分群結果產生極大的偏差，因此起始群心的選擇會連帶影響最終分群結果的好壞，這也是 K-means 分群的主要缺點之一。

第三節、 蟻群最佳化演算法

由 Dorigo 等學者所發展出的蟻群最佳化 (Ant Colony Optimization) 演算法[5][3][2][4]，其主要靈感是源自於螞蟻的特殊行為模式，與粒子群最佳化演算法同樣是屬於群體智慧 (Swarm Intelligence) 方法的一種。從過去昆蟲學家的研究中發現，螞蟻在尋找食物的過程中，在行進間會分泌一種叫做費洛蒙 (Pheromone) 的化學物質，螞蟻能夠據此作為行經路徑的紀錄，除此之外，陸續到達的螞蟻也會藉由地面所殘留的費洛蒙氣味，作為移動路徑選擇的參考依據。

隨著螞蟻行經此路徑的次數越來越頻繁，路徑上費洛蒙的濃度也會逐漸增強，費洛蒙氣味越強烈，則螞蟻選擇該路徑的機率也越高，反之氣味越淡，該路徑被挑選的機率也相對較低。蟻群演算法的主要精神，正是藉由此費洛蒙機制來達到彼此溝通與訊息交換的目的，並且搜尋出距離目標的最短行進路徑，進而驅使整個蟻群朝向正確的方向移動。

圖 2-3 為簡單的概念示意圖[3]，在這裡我們用虛線來表示螞蟻的費洛蒙資訊。當蟻群發現在原先的行經路徑上出現了某項障礙物，使得原本單一的路徑被分隔為兩條路線，如圖 2-3(A)所示。由於螞蟻在此時並沒有任何路徑選擇的參考依據，因此在左右兩側的螞蟻會根據隨機機率作分佈，隨機選擇往上或往下前進，如圖 2-3(B)。隨著螞蟻的前進，在這兩條路線上分別會遺留濃度不等的費洛蒙資訊，而從圖 2-3(C)中我們可以發現，由於下方的路徑相較於上方路徑要來的短，因此費洛蒙濃度也會隨著螞蟻通過而快速的累積。費洛蒙累積資訊越多，將使得螞蟻有比較高的機會朝向該路徑前進，逐漸形成圖 2-4(D)所顯示的費洛蒙濃度分佈情形。藉由此費洛蒙機制持續不斷的運作，最終將會突顯出一條具有相當高濃度費洛蒙資訊的最佳行經路徑。

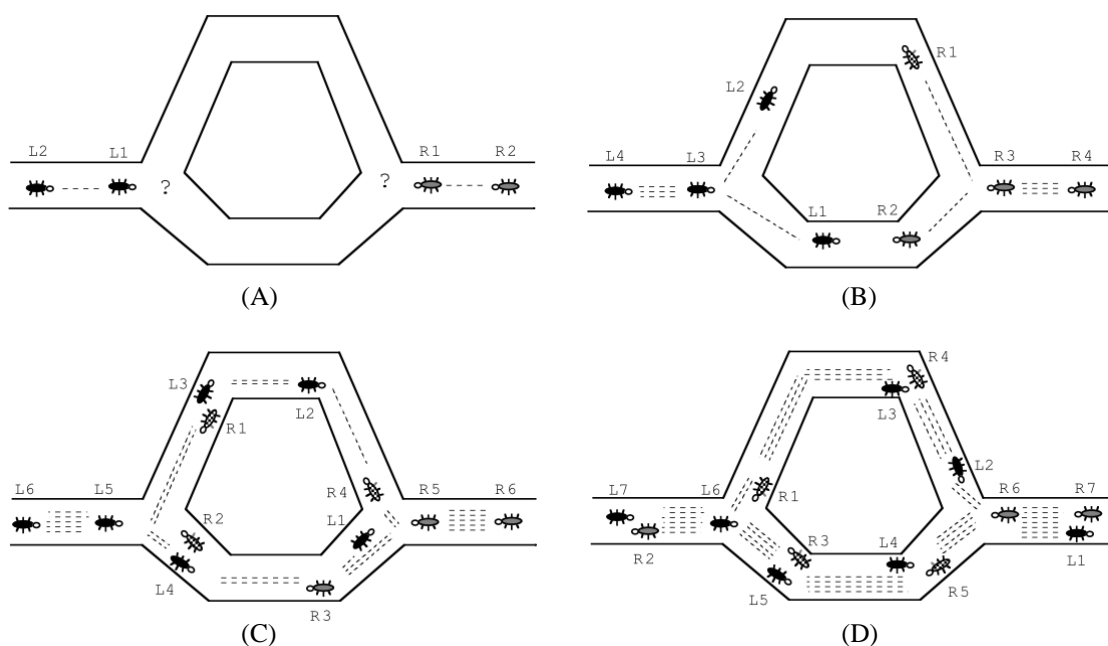


圖 2-3 費洛蒙機制概念示意圖[3]

利用這樣的概念，Dorigo 等學者陸續發展出了螞蟻系統（Ant System, AS）[5]、螞蟻群系統（Ant Colony System, ACS）[3]與螞蟻群最佳化（Ant Colony Optimization, ACO）[2][4]演算法，起初是用來解決旅行售貨員問題（Traveling Salesman Problem, TSP）這一類具有明顯節點與連線的組合最佳化問題，而之後也被延伸應用在解決各種不同最佳化問題上，例如二次分配（Quadratic Assignment Problem, QAP）、工作排程（Job-shop Scheduling Problem, JSP）、貝式網路（Bayesian Networks）與群集分析等問題。

壹、螞蟻系統演算法（Ant System）

螞蟻系統[5]由 Dorigo 等學者在 1996 年所發表，是最早將螞蟻費洛蒙概念應用於解決最佳化問題之演算法，研究中並揭櫫了三項螞蟻群演算法的主要特性，分別是正向回饋機制（Positive Feedback）、分散式運算（Distributed Computation）以及具有貪婪啟發式演算法（Greedy Heuristic）之特性。以下將簡述此三大特性：

1. 正向回饋機制

正向回饋機制意指螞蟻群在行為過程中會不斷的累積資訊並且逐漸自我強化，正如同我們在前面所解釋的費洛蒙機制一樣。透過費洛蒙機制的運作，使得具有良好效果的行為能夠累積較多的費洛蒙資訊，而螞蟻也會偏向選擇費洛蒙濃度較高的行為，費洛蒙濃度資訊經由如此反覆持續地累積，形成強者恆強、弱者越弱的態勢，使得效果最佳的行為選項能夠逐漸地突顯，最終將可得到所求解問題的最佳解答。

2. 分散式運算

分散式運算則是由於在蟻群演算法的問題求解過程中，蟻群內的每一隻螞蟻皆是獨立進行著搜尋最佳解答的任務，彼此之間除了透過費洛蒙資訊相互溝通以外，螞蟻並不會對彼此行為產生直接的影響與改變，因此螞蟻系統亦具有分散式運算的特性。

3. 具貪婪啟發式演算法特性

除此上述兩點特性之外，螞蟻系統還包含了貪婪啟發式演算法的特徵。由於螞蟻在求解過程中，是採取逐步建構解答的策略，而且螞蟻具有根據目前所在的位置或情況，選擇目前已知的最佳行為選項的傾向，這樣的特徵與貪婪啟發式演算法的作法相當雷同，差別僅在於螞蟻的行為主要是根據費洛蒙濃度所決定，因此作者認為螞蟻系統確實具有貪婪啟發式方法之特性。

在螞蟻系統中，螞蟻的行為選擇主要是根據費洛蒙濃度與選項之能見度所構成的轉換機率（Transition Probability）為基準，如公式 2-3 所示。

$$P_{ij}^k(t) = \frac{[\tau_{ij}(t)]^\alpha \times [\eta_{ij}]^\beta}{\sum_{k \in allowed_k} [\tau_{ij}(t)]^\alpha \times [\eta_{ij}]^\beta} \quad (2-3)$$

當螞蟻位於節點選項 i 欲選擇下一個節點選項 j 時，則會計算所有能夠可供選擇的項目之轉換機率 P_{ij} ，而其中的 τ 意指節點選項 i 與節點選項 j 之間的費洛蒙濃度，而 η 則是指兩個選項之間的能見度，在研究中 Dorigo 等學者將能見度定義為兩點間距離之倒數，距離越近則能見度越高，反之則能見度越低。所計算出來的轉

換機率越大，則代表螞蟻選擇該選項的機率也越大，但這並不代表螞蟻只會選擇轉換機率最大的項目，螞蟻每一次的選擇都必須根據上述轉換機率來決定，而這樣的概念也一直延續到之後的蟻群最佳化演算法之中。

除了轉換機率以外，螞蟻系統中還有另外一項重要的機制，就是費洛蒙濃度更新（Pheromone Updating）機制，利用螞蟻所求得之解答結果，將其轉換為費洛蒙資訊，並反饋至每一個選項之間的連結關係中，使得費洛蒙資訊得以累積，用以提供下一個迭代螞蟻作為轉換機率計算的依據。這樣的費洛蒙更新機制，也就是我們前面所闡述的正向回饋特性的體現。

而在費洛蒙濃度的更新累積機制之外，在螞蟻系統中也考量到了費洛蒙的散失效果。當路徑距離太遠導致所建構解答的效果不佳時，該路徑不僅無法快速累積費洛蒙濃度，就連路徑上原有的費洛蒙資訊也會隨著時間而逐漸散失，進而使得其他螞蟻行經該路徑的頻率逐漸降低，形成了弱者越弱的惡性循環。透過上述費洛蒙累積與散失的機制，螞蟻能夠適時的淘汰效果不佳的選項，並趨向較佳的選項做選擇，藉以建構出具有良好求解效果之優良解答。

貳、蟻群系統演算法（Ant Colony System）

Dorigo 等學者在 1997 年所發表之蟻群系統[3]，同樣延續了螞蟻系統中費洛蒙機制的精神。而不同於螞蟻系統，蟻群系統主要有以下三項改變：採用全域最佳解進行費洛蒙全域更新（Global Updating）、新增費洛蒙局部更新（Local Updating）機制，並新增了轉換規則（Transition Rule）機制。

1. 費洛蒙全域更新

在原本的螞蟻系統中，費洛蒙更新機制是以該迭代中的最佳結果作為更新依據，而各個迭代都會取得各自的最佳解答，這樣的更新機制設計可能會導致先前迭代中優良的解答資訊無法被保留，反而被後來迭代中相對較差的結果所取代，降低了早先迭代中優良解答對於費洛蒙資訊的累積效果；因此在蟻群系統中，作者將費洛蒙更新機制改良為利用所取得之歷史最佳解當作更新依據，讓在不同迭代中的最佳解答得以反應在費洛蒙資訊上，讓螞蟻搜尋能夠更有效率。

2. 費洛蒙局部更新

費洛蒙局部更新機制則是讓螞蟻在建構解答的過程中，能夠即時更新所行經路徑上之費洛蒙，提供給同一迭代中的螞蟻作為參考，讓費洛蒙資訊能夠隨著蟻群的選擇而即時更新與反應，無需等到該迭代完成才進行更新，這樣的作法是為了避免螞蟻受限於全域最佳解答所累積的費洛蒙資訊影響，以期能跳脫出局部最佳解的陷阱中。

3. 轉換規則

轉換規則是由螞蟻系統的轉換機率為基礎而延伸，當螞蟻必須選擇下一個項目時，則透過隨機機率來決定是要以何種方式進行選擇。可分為有兩種不同的項目選擇策略，分別是以傳統轉換機率來選擇、或是直接選擇轉換機率中機率值最大的項目。

而轉換規則的設計是為了讓螞蟻能夠藉由參數的設定，根據機率來調整求解策略，能夠藉由求解過程中開拓 (Exploitation) 與探

索 (Exploration) 兩者的比例之考量，兼顧求解效率與效能，避免由於過快收斂而導致陷入局部最佳解的情形。

參、 蟻群分群演算法

由 Shelokar 等學者在 2003 年所發表的研究中，將螞蟻費洛蒙資訊的概念應用到求解資料分群問題上，發展出以蟻群概念進行群集分析的演算方法[6]。

與 Dorigo 等學者用蟻群系統來求解旅行售貨員 TSP 問題的方式相同，在此研究中，費洛蒙同樣也是用來紀錄求解過程中所需的資訊，最大的差別在於將原本儲存在節點連線上的費洛蒙，轉換成為較為抽象的費洛蒙矩陣 (Pheromone Matrix)，藉以紀錄每一筆資料與個別群集之間的費洛蒙濃度，螞蟻會利用此費洛蒙矩陣來決定個別資料的分群，建構出最後的群集分析結果，而該分群演算法也會進一步評估每一隻螞蟻所建構之分群結果的優劣，作為費洛蒙矩陣更新的依據，如此反覆不斷，直到滿足指定迭代次數為止，最後得到所有迭代螞蟻所建構之解答中最佳的資料分群結果。

而在此研究中 Shelokar 等學者將分群結果以編碼化方式呈現，將分群結果定義為一串長度為 N 的字串，用以紀錄完整的資料分群結果，如圖 2-4 所示：

2	1	3	2	2	3	2	1
---	---	---	---	---	---	---	---

圖 2-4 蟻群分群結果編碼方式示意圖[6]

圖 2-4 中的字串紀錄了將 8 筆資料分為 3 群的分群結果，字串內的數字代表著個別資料所歸屬之群集編號，1 為第一群、2 為第

二群…以此類推，分群編號若相同則表示資料被分在同樣的群集中，例如圖 2-4 中的第 2 筆與第 8 筆資料都被區分到第 1 群內。

每一隻螞蟻所求得的解答正是以上述的字串結構紀錄在解答陣列中。在這裡我們沿用 Shelokar 等學者在螞蟻分群方法[6]中所使用的 fitness 公式，利用個別螞蟻的分群結果，分別計算出各解答的權重矩陣（Weight Matrix）以及各個群集的群心位置，並且以此兩者為依據計算出分群結果之 fitness 值，用以判斷分群解答的品質優劣程度。Fitness 值計算方式如公式 2-4 所示。

$$\text{Min } F(w, m) = \sum_{j=1}^K \sum_{i=1}^N \sum_{v=1}^n w_{ij} \|x_{iv} - m_{jv}\|^2 \quad (2-4)$$

公式 2-4 中的 j 為群數、 i 為資料筆數，而 v 則為個別屬性維度，其中 w_{ij} 即為權重矩陣，是由個別螞蟻的分群結果轉換而來，所紀錄的內容為個別資料 i 與不同群集 j 的分群結果，若資料被分到群集中則標示為 1，若無則為 0，為稀疏矩陣形式，在方法中是以該權重矩陣作為是否分到該群中的參考資訊。 x_{iv} 表示資料 i 的個別屬性維度 v 的值，而 m_{jv} 則表示群集 j 群心之屬性維度 v 的數值。透過這三個矩陣中所紀錄的資訊，計算出個別螞蟻之分群結果 fitness 值。而 fitness 值越小，表示群集內資料越集中，也就代表分群結果越優良，相反地，若 fitness 值越大則象徵著群集分析結果越差。

當所有螞蟻皆求解建構完畢，且將全部解答之 fitness 值計算完成之後，接著以 fitness 值為依據，取出前 L 個最佳的解答用以進行重新分群。此步驟稱之為 local search，主要作法是利用重新分配機率參數來決定是否要對解答中的個別資料進行重新分群，並且計

算重新分群前後之 fitness 值，若重新分群後的結果要較原本分群結果來的佳，則以新的分群結果取代原始分群解答。重新分群的主要目的是為了避免螞蟻所建構的解答由於費洛蒙資訊的過度影響，而陷入了局部最佳解中，因此設計了此 local search 機制，用來改善螞蟻因為求解過程過快收斂，導致落入局部最佳解的情形。

最後以前 L 個最佳解答之 fitness 值倒數，作為費洛蒙更新依據。持續以上的步驟，直到滿足指定迭代次數為止，最後輸出全域 fitness 值最小之分群結果，完成螞蟻分群演算法流程，如圖 2-5 所示。

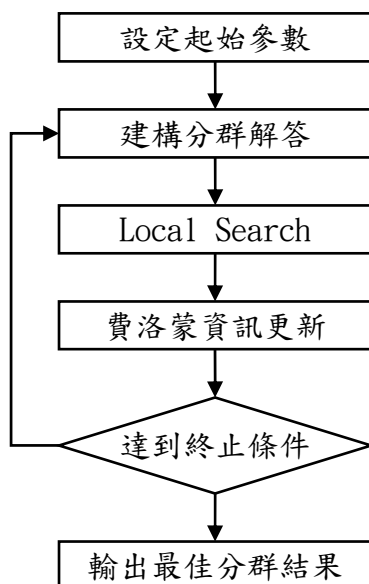


圖 2-5 螞蟻分群演算法流程圖

第三章 改良式蟻群分群演算法

第一節、改良式蟻群分群演算方法

我們所提出的改良式蟻群分群方法是以 Shelokar 等學者所發展之蟻群分群方法[6]為基礎加以修正，希望能在取得同等甚至是更好的解答品質的前提下，以更快速且有效的方式改善原先蟻群分群方法中需要大量迭代運算的缺陷。我們主要的改良重點如下：

1. 加入資料與群心距離資訊考量並重新建構分群解答

由於在最原始的蟻群分群方法中，螞蟻在解答建構的過程中僅以費洛蒙矩陣資訊作為參考依據，並未將資料點與群心之間的距離納入求解考量，但在該方法中費洛蒙資訊的累積必須仰賴各迭代所取得之優良解答來作更新，因此無法即時且有效的反應出整體分群求解趨勢，使得求解效率不佳，必須要經由大量演算迭代才能累積足夠的費洛蒙資訊，藉以建構出更加優良的分群結果。

而在原始的蟻群最佳化演算法中，轉換機率是根據費洛蒙資訊以及個別選項之間的能見度計算而得[5][3]，由先前學者的實驗也可以發現除了螞蟻費洛蒙所構成的啟發式（Heuristic）訊息以外，選項之間的幾何距離關係也會對螞蟻求解效率造成相當程度的影響[7]。因此在我們的作法中，加入了資料點與群心的幾何距離考量，在螞蟻根據費洛蒙資訊建構解答完成之後，我們會檢查各個資料點與個別群心之距離，並且重新修正先前的分群結果，將資料分配至距離最近之群心群集中，以期能大幅增進整體螞蟻的求解效能。

2. 改良 Local Search 機制

在最原始的蟻群分群方法中，是以數個優良的分群結果來進行 local search 重新分群，此機制的設計主要用意是希望能夠將具有良好效果的分群結果加以改善，透過機率控制決定是否將原有的資料群集分析結果重新分群，希望能藉由這樣的方式取得更加優良的分群解答。

但這樣的作法卻存在著幾項問題，在這裡我們就兩個角度來分析，首先是對於原先就已經相當良好的分群結果而言，local search 機制的目的是要讓原先的分群效果能夠進一步提昇，但卻無法確保經過重新分群之後能夠出現比原本解答還要好的結果。而對於其他效果不佳的分群結果，也無法透過 local search 機制重新進行分群，進而達到改良的效果。此外，蟻群求解範圍也由於原有之 local search 機制偏重於優良結果的改良，反而容易因此陷入局部最佳解的窠臼中。

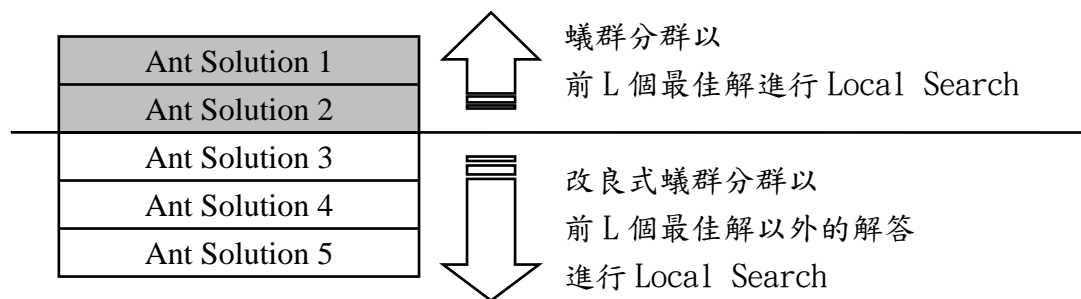


圖 3-1 Local Search 更新示意圖

因此在我們的改良方法中，將 local search 的目標從原本的前 L 個優良分群結果，修正成為利用非優良解答之分群結果作改良，期待能藉由 local search 目標的調整，讓原先就相當優良的結果不會因為 local search 而變差，而較差的結果能因為 local search 而得到改善，使得整體求解品質能得到有效的改善，提昇分群求解效能。如圖 3-1 所示。

3. 加入分群解答字串排序機制

由於蟻群分群方法的分群結果是透過字串編碼的形式來紀錄，因此可能會出現同樣的分群解答卻有著不同編碼字串的情形。如圖 3-2 中所示，我們可以看出這兩組字串具有相同的分群結果，但是由於在蟻群分群方法中，由於解答編碼方式的限制，使得同樣的分群結果可能會出現不同的編碼順序，造成費洛蒙資訊無法有效的聚集，反而延緩了螞蟻費洛蒙累積的效率。正如我們在前面所解釋的，分群解答的建構與費洛蒙資訊具有相當重要的關聯，而費洛蒙資訊的累積更是從優良的螞蟻分群結果而來，但倘若原本具有良好效果的分群解答，卻由於編碼順序的不同，反而會使得費洛蒙在更新時無法將良好的分群結果正確地反應在費洛蒙資訊上，導致分群效率無法有效的提昇，分群求解的品質也會因為費洛蒙資訊的不當累積而造成不佳的影響。

S_1	1	1	2	2	2	3	3	3
S_2	2	2	3	3	3	1	1	1

圖 3-2 分群結果編碼示意圖

為了避免此問題產生，我們將每一組分群解答的編碼字串進行順號，如圖 3-3，讓具有相同分群結果之解答能藉由順號動作，得到相同的解答編碼字串，使費洛蒙資訊能更有效的累積，提昇求解效率。

S_2 順號前	2	2	3	3	3	1	1	1
↓								
S_2 順號後	1	1	2	2	2	3	3	3

圖 3-3 解答編碼順號示意圖

第二節、改良式蟻群分群演算法演算流程

綜合以上改良點，我們提出了改良式蟻群分群方法，如圖 3-4 所示。

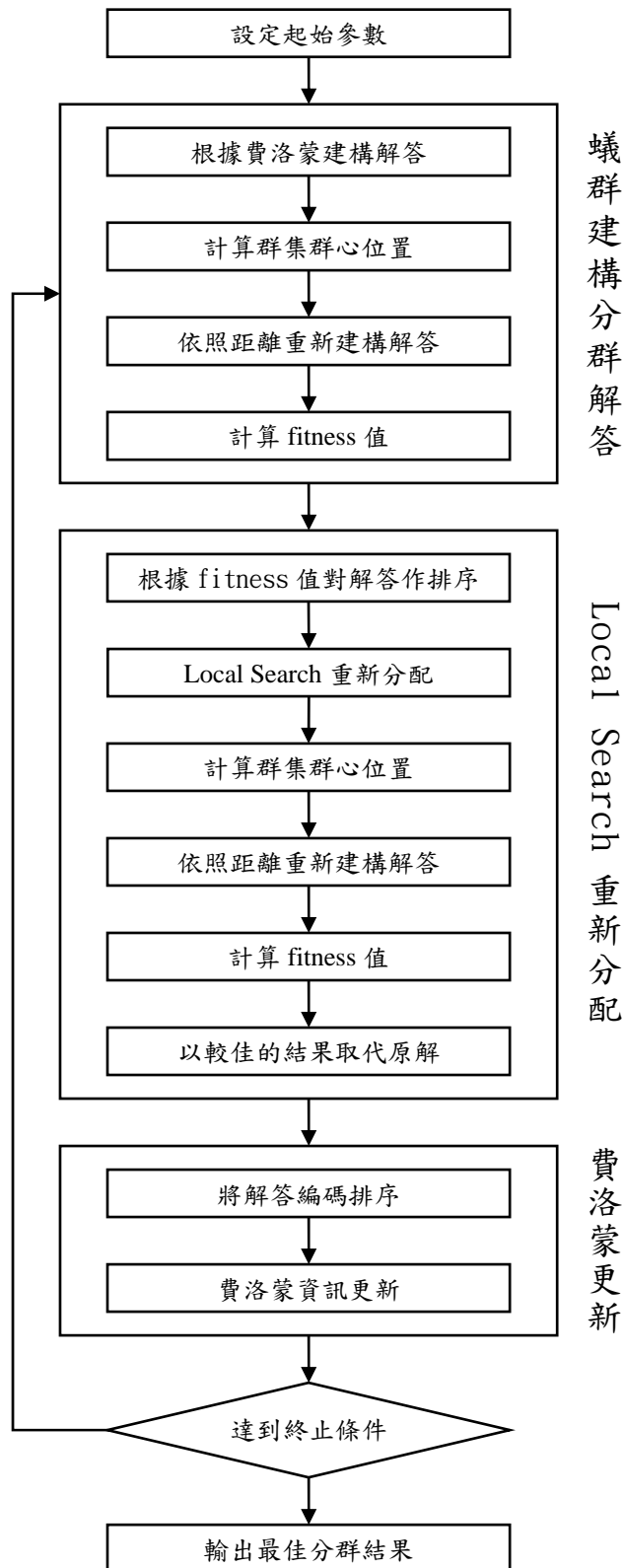


圖 3-4 改良式蟻群最佳化演算法流程圖

我們所提出的方法主要可以分為蟻群建構分群解答、Local Search 與費洛蒙資訊更新三大階段，以下將以一個簡單的範例資料來詳細介紹各個步驟的運作方式，並且比較原始蟻群分群方法與我們所提出之改良式蟻群分群法的分群求解效果。

表 3-1 分群資料範例

No.	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
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2

我們沿用文獻[6]中的範例，以 Iris 資料集的前 8 筆資料為例，其中分別包含了 4 個維度數值，內容如表 3-1 所示。我們將以此資料作為範例，示範蟻群分群方法如何將這 8 筆資料分為 3 個群集，詳細步驟分別解釋如下：

壹、蟻群建構分群解答階段

步驟一：設定起始參數

起始參數的設定有以下幾項，包括螞蟻個數 R 、分群數 K 、改良解答個數 L 、最大迭代次數、費洛蒙散失率、轉換規則機率 q_0 、local search 機率 P_{ls} 。

步驟二：建構分群解答

螞蟻在要開始建構分群解答之前，會先根據費洛蒙矩陣內的資訊作為資料分群選擇的依據。

表 3-2 費洛蒙矩陣範例

資料	第 1 群	第 2 群	第 3 群
1	0.00015	0.00003	0.00037
2	0.00018	0.00041	0.00001
3	0.00017	0.00012	0.00038
4	0.00036	0.00021	0.00007
5	0.00003	0.00049	0.00021
6	0.00011	0.00030	0.00018
7	0.00006	0.00024	0.00035
8	0.00007	0.00024	0.00017

費洛蒙矩陣的形式如表 3-2 所示，費洛蒙矩陣中紀錄了個別資料分到不同群集的費洛蒙濃度，其濃度越高也就表示資料分至該群集中的效果越好，而被螞蟻選擇分到該群的機率也越高。在整個螞蟻群分群方法中，就是利用這樣費洛蒙矩陣的概念來紀錄分群資訊，並且作為選擇資料分群的主要依據。

在演算一開始時，我們會將費洛蒙矩陣初始化為一個極小的隨機亂數值，主要目的是由於在螞蟻建構分群解答的過程中，是根據費洛蒙矩陣資訊作為依據，費洛蒙濃度越高，則螞蟻將資料劃分給該群集的機率也越大，但在尚未開始分群之前我們並不能確定資料應該要分到那個群集中，因此我們藉由一個微小的亂數作為螞蟻建構解答的依據，並用以驅動螞蟻開始進行分群解答的建構與搜尋。

至於螞蟻選擇資料分群的機率，是以我們先前所提到的轉換規則[3]作為分群選擇的依據，如公式 3-1 所示。

$$s = \begin{cases} \arg \max \{ \tau_{ij}, j = 1, \dots, K \} & \text{if } q \leq q_0 \\ p_{ij} = \frac{\tau_{ij}}{\sum_{k=1}^K \tau_{ik}}, j = 1, \dots, K & \text{otherwise} \end{cases} \quad (3-1)$$

每當螞蟻在決定資料要分到哪一群時，系統會產生一個隨機亂數值 q ，用以判斷採用何種規則來決定分群結果。若該亂數值小於或等於我們所設定的轉換規則機率參數 q_0 時，螞蟻就直接將資料分到具有最大費洛蒙濃度的群集中。若該亂數大於參數設定，則根據原始的轉換機率來計算分群機率，並且根據機率隨機決定該筆資料的分群結果。

表 3-3 分群結果表示範例

結果	資料							
	1	2	3	4	5	6	7	8
S_1	3	3	3	3	1	2	3	1
S_2	1	2	2	2	1	3	2	1
S_3	1	2	2	2	1	3	1	1
S_4	2	1	3	1	2	2	3	2
S_5	3	1	1	1	3	2	1	3

分群結果如表 3-3 所示，表中分別紀錄了 5 隻螞蟻所建構出的 5 個分群解答。

步驟三：計算群心位置

得到分群結果之後，接著我們就可以利用個別螞蟻所求得之結果來計算群心位置。群心位置的計算方式是將同一群集內的資料依照個別維度加總並平均，最後所得到的值即為分群結果的群心資料。由螞蟻所建立的每一組分群結果都能夠計算出一組群心位置，有多少組解答就會有多少組群心位置，而我們就是利用群心來計算個別資料與群心的距離，用以重新建構分群結果，並且計算其 fitness 值。

表 3-4 分群結果 S_I 的群心位置

群集	sepal length	Sepal width	petal length	petal width
1	5.00	3.50	1.45	0.20
2	5.40	3.90	1.70	0.40
3	4.78	3.24	1.40	0.22

表 3-4 中所顯示的是由表 3-3 中分群解答 S_I 所計算出之個別群心位置，其中第 2 群只有一筆資料，因此群心就等同於該資料點。

步驟四：根據距離重新建構分群解答

在計算出群心位置後，我們將進一步利用歐幾里得距離公式來計算分群結果中個別資料與群心的距離，並且修正其分群結果。以 S_I 為例，在重新建構解答之後的分群結果顯示在表 3-5。

表 3-5 加入距離重新建構的分群結果 S_I

結果	資料							
	1	2	3	4	5	6	7	8
$S_{I\ old}$	3	3	3	3	1	2	3	1
$S_{I\ new}$	1	3	3	3	1	2	3	1

在表 3-5 中以粗框標示的部份為 S_I 在考量資料與群心距離之後的分群結果改變情形。我們可以發現在加入了幾何距離的考量之後，有可能會造成部份資料的分群結果改變。而藉由資料點與群心的距離考量，使得蟻群在建構解答的過程中，除了以費洛蒙矩陣作為分群選擇的參考依據之外，還能夠藉由幾何距離來修正解答的品質，並且希望能夠透過這樣的方式來改善蟻群求解收斂速度。

步驟五：計算個別分群結果之 fitness 值

在所有螞蟻皆將解答建構完成之後，利用群心位置資訊我們即可計算出個別解答的 fitness 值。如同我們前面所提過的蟻群分群演

算法一樣，在這裡我們也是利用 fitness 的計算來判斷分群結果的好壞。我們將分群結果與 fitness 值放在一起比較，結果如表 3-6。

表 3-6 分群結果與 fitness 值表示

結果	資料								Fitness
	1	2	3	4	5	6	7	8	
S_1	1	3	3	3	1	2	3	1	1.102214
S_2	1	2	2	2	1	3	2	1	1.102214
S_3	1	2	2	2	1	3	1	1	1.311676
S_4	2	1	3	1	2	2	3	2	1.634727
S_5	3	1	1	1	3	2	1	3	1.102214

我們之前曾經解釋過，若分群結果的 fitness 值越低，表示其結果中的資料與分群群心距離越接近，整體的分群效果越佳，反之若 fitness 值越高，則代表資料與群心越離散，分群效果也相對較差。從表 3-6 中可以看出個別螞蟻所求得之分群結果與其所計算出之 fitness 值，在這裡我們將 fitness 值視為是分群結果品質的參考依據，並且作為之後 local search 重新分群的依據。

貳、Local Search 重新分配階段

步驟六：根據 fitness 值對解答作排序

在計算出 fitness 值後，系統會根據 fitness 值對所有的分群結果作排序，接著以排序之後的結果來進行 local search 重新分配。

步驟七：進行 Local Search 重新分配

Local search 的主要重點是以隨機機率來決定某筆資料是否要重新進行群集分配。隨著演算迭代的推進，費洛蒙資訊也會逐漸顯出某些較佳的分群結果，而我們在蟻群分群方法中可以利用 local search 機制，讓蟻群所建構的分群結果能有跳脫局部最佳解的機

會，並且藉由解答的變異，增添了得以建構出相較於目前更好的分群解答之可能性。

在這個階段中，是否要對結果進行重新分配是透過重新分配機率參數 P_{ls} 與隨機亂數來決定。在演算流程中系統會檢視特定的分群結果，並且對於結果中的每一筆資料產生一個隨機亂數值，若當該隨機亂數大於重新分配機率參數 P_{ls} 時，則隨機將該筆資料重新分配到與原先分群結果不同的群集之中；相反地，若亂數小於參數值則不做任何改變。在這裡我們利用參數設定中的改良解答個數 L 為依據，對於前 L 個優良解答以外的分群結果進行 local search，以期能夠讓較差的解答能有改善的機會。完成 local search 之後的分群結果表示如表 3-7。

表 3-7 Local Search 後的分群結果表示

結果	資料							
	1	2	3	4	5	6	7	8
S_1	1	3	3	3	1	2	3	1
S_2	1	2	3	2	1	3	2	1
S_3	1	2	2	2	1	3	3	1
S_4	2	1	1	1	2	3	3	2
S_5	3	1	1	1	3	2	1	3

表 3-7 是進行 local search 之後的分群結果，由於重新分群與否是藉由隨機亂數來決定，因此亦有可能出現分群結果毫無改變的情況。而在完成 local search 之後，由於分群結果產生變化，使得群集中心位置有所改變，而 fitness 值也會連帶受到影響，所以我們必須重複進行步驟三到步驟五的計算群心、重新建構解答以及 fitness 值計算，直到所有進行 local search 的結果皆計算出 fitness 值為止。

步驟八：比較解答在重新分配前後之 fitness 值，並擇優取代

待 local search 後分群結果之 fitness 值計算完成之後，系統會比較解答在進行重新分群前後的 fitness 值，並且選擇兩者之中具有較佳 fitness 值的結果作為最後的分群解答。

參、費洛蒙資訊更新階段

步驟九：將解答編碼排序

正如我們在前面所提到的，由於本研究所沿用的蟻群分群方法 [6] 在解答編碼方式上之限制，使得一樣的分群結果但有著不同編碼順序的情形，導致同樣分群結果的費洛蒙資訊無法有效累積，因此我們經由分群編碼排序的動作，讓費洛蒙資訊的累積更有效率。

表 3-8 解答編碼排序範例

結果	資料							
	1	2	3	4	5	6	7	8
S_1	1	2	2	2	1	3	2	1
S_2	1	2	3	2	1	3	2	1
S_3	1	2	2	2	1	3	3	1
S_4	1	2	2	2	1	3	3	1
S_5	1	2	2	2	1	3	2	1

排序後的解答編碼如表 3-8 所顯示，從表中可以發現原先編碼結果不相同的解答 S_1 與 S_5 ，在經過編碼排序之後呈現了一樣的編碼結果，這樣的處理將有利於蟻群費洛蒙資訊的更新與累積。

步驟十：更新費洛蒙矩陣

在解答編碼排序之後，我們以前 L 個具有最佳 fitness 值的分群結果作為費洛蒙矩陣更新依據，以螞蟻將某筆資料分到某群集為費洛蒙更新目標，而更新的費洛蒙量為該解答的 fitness 值之倒數。

表 3-9 表 3-8 中解答 S_l 更新費洛蒙矩陣範例

資料	第 1 群	第 2 群	第 3 群
1	0.90741	0.00003	0.00037
2	0.00018	0.90767	0.00001
3	0.00017	0.90738	0.00038
4	0.00036	0.90747	0.00007
5	0.90729	0.00049	0.00021
6	0.00011	0.00030	0.90744
7	0.00006	0.90750	0.00035
8	0.90733	0.00024	0.00017

我們以表 3-8 中的解答 S_l 來作費洛蒙更新範例，結果如表 3-9 所示。利用同一個迭代中具有良好求解效果的前 L 個分群結果的 fitness 值之倒數，將其累加至費洛蒙矩陣中，據此作為費洛蒙更新的依據，隨著迭代經過可以明顯的看出費洛蒙的累積效果。而除了費洛蒙更新機制以外，還加入了費洛蒙散失機制的考量，我們利用散失參數作為費洛蒙機制的抑制效果，避免由於費洛蒙資訊累積過快導致落入局部最佳解中。

在蟻群最佳化方法中就是藉由這樣的費洛蒙資訊達到建構解答、並且累積資訊相互傳遞的作用，由於費洛蒙的正向反饋效果，使得強者越強、弱者漸弱，好的解答能夠隨著蟻群搜尋而逐漸突顯，較差的結果則慢慢地從每一次的迭代中散失，最後所留下的費洛蒙資訊即為在搜尋過程中最佳求解解答。

步驟十一：輸出最佳分群結果

在每一個迭代的過程中，系統會紀錄各個迭代螞蟻所求得之最佳分群結果，當到達指定運算迭代次數後，輸出整個回合中最佳的分群解答，完成蟻群分群求解過程。

第四章 遺漏值推估方法

本章節將介紹如何透過屬性值分割方法以及蟻群分群演算法來解決遺漏值估計問題。本章分為四個小節，我們將利用半導體製程的真實資料為例，介紹本研究所使用之基礎遺漏值推估模組。在第一節中我們將分別介紹屬性值分割方法之遺漏值推估模組、蟻群分群方法之遺漏值推估模組以及將此兩種方法結合之遺漏值推估模組的主要特性，並說明將兩者結合的原因。第二節敘述利用屬性值分割方法回填遺漏值推估模組的詳細流程與作法。第三節則介紹我們所提出之改良式蟻群分群方法，用於遺漏值估計回填的流程說明。第四節將說明如何結合屬性值分割以及蟻群分群方法，並用以回填遺漏值推估模組的運作流程。

第一節、基礎遺漏值推估模組

正如先前所談到的，傳統面對資料內遺漏值的處理方式通常是採用整體資料之平均值進行回填，但這樣的作法並未考量到整體資料中個別維度屬性之間的關係，亦容易受到資料中極端數值所影響，導致所估計之遺漏值有所偏誤。有鑑於此，在本研究中我們採用群集分析技術為基礎，利用物以類聚的群聚概念為出發點，藉由分群結果中群內資料同質、群間異質的群集關係，讓遺漏值的推估結果能夠更加貼近真實數值，建構出合適的遺漏值推估模組，取得更好的回填效果。

本研究中所使用的遺漏值推估模組有二，分別是基於屬性值分割方法之遺漏值推估模組以及基於蟻群分群方法之遺漏值推估模組，此兩種方法的概念皆以群集分析技術為基礎發展而成。如圖 4-1 所示，屬性值分割方法之遺漏推估模組的主要精神，是以資料中個別維度的距離分佈情形作為分割依據，無須事先指定分割群數，方法中會自動根據資料維度計算出每個維度的切割位置，形成個別維度的分割結果，是一種具有動態分割概念的演算方法。此方法利用個別資料維度所計算出的平均差來作為分割點的判斷依據，由於分割點的計算條件限制嚴格，因此可以確保在同一個屬性分割內的資料距離是最為相近的。在我們的研究中以屬性值分割方法應用於遺漏值推估回填問題之上，期望能藉以提昇遺漏值推估的效率與品質。在本章第二節中我們介紹基於屬性值分割方法之遺漏值回填模組的演算流程與實例說明。

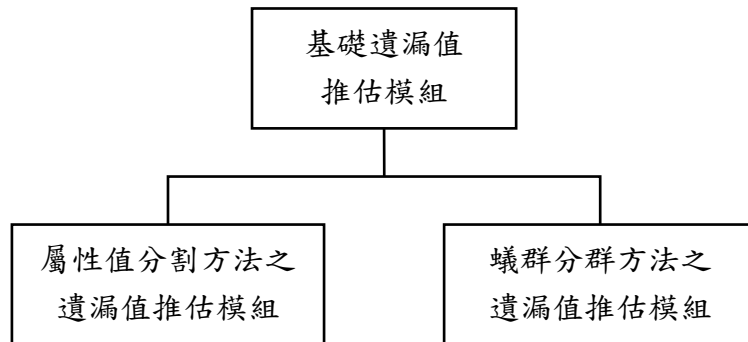


圖 4-1 基礎遺漏值推估模組架構圖

另一個遺漏值推估模組則是由我們在上一節中所提出的改良式蟻群分群方法所延伸而來，利用蟻群不斷建構解答與改良比較的過程，逐漸取得更好的分群結果，並且希望能藉由分群品質的改善，進而推估出更為準確的估計值。本章第四節將介紹蟻群分群方法遺漏值推估模組的演算流程與實例說明。

在本研究中我們嘗試將上述兩種遺漏值推估模組作結合，主要是由於在傳統的作法中多半是利用未包含遺漏值之屬性平均值回填，用以作為遺漏值估計的初始數據，但文獻資料中我們可以發現以屬性平均值回填的作法通常會產生較大的偏誤[9][12]，因此我們利用屬性值分割遺漏值推估模組所估計之遺漏值來回填，取代以往利用整體資料屬性平均值的回填方式，作為之後蟻群分群方法的資料初始值，以期能夠取得更接近原始資料的分群結果，提昇遺漏值估計的準確率與求解效率，並加速蟻群演算法的求解收斂速度。而在蟻群分群演算收斂之後，由最後所求得之分群結果取得具有遺漏值資料的分群位置，並以該群集之屬性平均值作為最後回填的遺漏值資訊，最後利用平均絕對誤差（Mean Absolute error, MAE）來計算我們的遺漏值估計回填結果準確率。

在以下章節中我們以半導體銅製程的實驗資料集來說明遺漏值推估方法之流程與實際運作說明，該資料集共有 18 筆資料，每筆資料有 3 個屬性，分別是研磨速度、均勻度與選擇性，如表 4-1 所示。以下我們將以此資料集為例，分別說明以屬性值分割遺漏值推估模組以及蟻群分群遺漏值推估模組的實做流程。

表 4-1 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、基於屬性值分割方法之遺漏值推估模組

屬性值分割方法與分群的概念類似，都是利用資料彼此之間的距離作為分割的主要依據，資料之間的距離越接近，代表兩者的相似程度越高，也就越容易被分到同一群集中，反之距離越遠則越相異。不同於分群方法以資料整體屬性維度距離為相似度的計算方式，在於屬性值分割方法是以個別屬性維度作為參考來進行資料分割，因此每一項屬性都會有一個分割結果，我們就是利用個別屬性分割後的資訊來產生資料的分群結果。由於屬性值分割方法完全根據資料屬性值的分

佈來計算分割點，無須事先指定分群群數，因此屬性值分割方法具有基礎的動態分群概念。

以下我們同樣利用半導體銅製程資料集作為蟻群分群遺漏值推估模組的流程範例，並將最後的回填結果和類神經網路[9]與粒子群演算法[12]的遺漏值推估方法作比較，探討屬性值分割遺漏值推估模組的執行效益與可行性。我們利用與文獻[9][12]相同之遺漏值資料集進行實驗，並且以相同的評估指標來判斷遺漏值推估結果的優劣。遺漏資料集中包含 3 筆遺漏屬性值，如表 4-2 中框線標注部份所示。

表 4-2 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

屬性值分割方法遺漏值推估模組演算流程如圖 4-2，以下將以個別步驟進行說明。

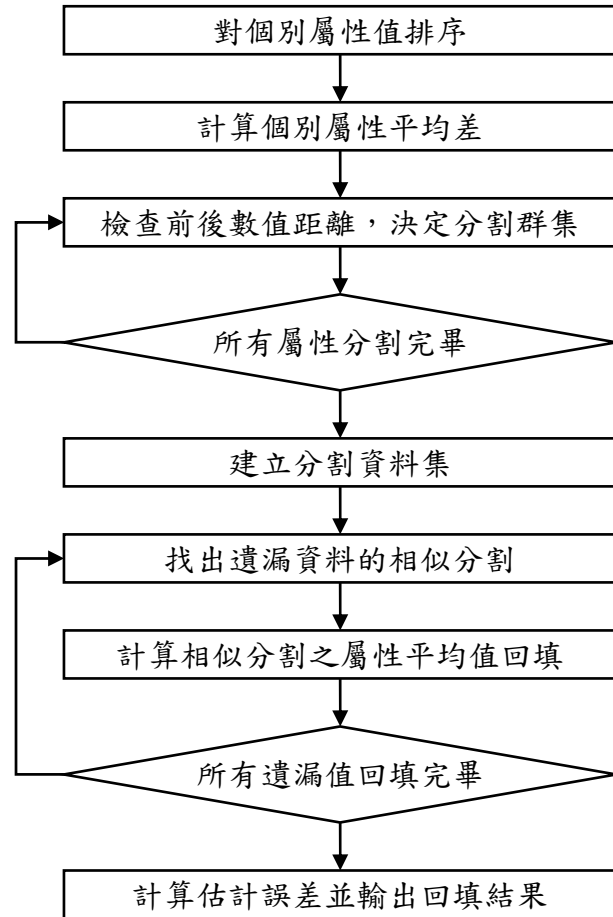


圖 4-2 屬性值分割方法之遺漏值推估模組流程圖

步驟一：以個別屬性值做排序

利用排序演算法對個別屬性維度內非遺漏值的資料由小到大進行排列。表 4-3 為半導體銅製程中的三個屬性進行排序之後的結果。

表 4-3 半導體銅製程個別屬性排序結果

研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
242	5	3.7
268	5.06	3.9
289	6.5	4
294	7.08	4.2
314	8.76	4.3
340	8.99	4.3
375	10.5	4.6
377	11.8	4.6
434	12.1	4.7
437	12.4	4.9
481	14.3	5.2
494	15.1	5.3
498	15.7	5.3
580	16.2	5.4
588	16.9	5.6
651	23.2	5.8
660	26.9	6.1

步驟二：計算個別屬性維度之平均差

在數列排序完成之後，我們必須先將數列中具有相同數值的屬性資料排除，避免在平均差的計算過程中造成偏誤。平均差的計算方法是將特定屬性維度中，具有不重複屬性值的距離間隔做加總並求其平均值，公式如 4-1 所表示。

$$total_average_dist = \frac{\sum_{i=1}^{N-1} |x_i - x_{i+1}|}{N-1} \quad (4-1)$$

以表 4-3 中的均勻度為例所計算之平均差為：

$$\frac{|5-5.06|+|5.06-6.5|+|7.08-8.76|+\dots+|16.2-16.9|+|16.9-23.2|+|23.2-26.9|}{16} = 1.36875$$

之後我們將利用個別屬性的平均差值，作為判斷屬性維度分割位置的主要依據。

步驟三：檢查前後數值距離，決定分割至何分割中

計算出平均差之後，接著利用排序完成的資料進行個別屬性值分割。我們依序檢查除了遺漏值以外的數值，先將第一個數值資料分到第一群中，接著比較第一筆數值與第二筆數值的距離，若該距離小於平均差，則將第二筆數值分到與第一筆數值相同的群中，若大於平均差則將第二筆數值區分到新的群集內。以這樣的方式逐步檢查所有的數值，最後可以得到個別屬性的分割結果，表 4-4 為均勻度分割結果。

表 4-4 半導體銅製程之均勻度屬性分割結果

均勻度(NU；%)	屬性數值距離	分割結果
5	-	1
5.06	0.06	1
6.5	1.44	2
7.08	0.58	2
8.76	1.68	3
8.99	0.23	3
10.5	1.51	4
11.8	1.3	4
12.1	0.3	4
12.4	0.3	4
14.3	1.9	5
15.1	0.8	5
15.7	0.6	5
16.2	0.5	5
16.9	0.7	5
23.2	6.3	6
26.9	3.7	7
平均差	1.36875	

步驟四：建立分割資料集

重複步驟三直到所有屬性皆分割完畢後，接著將個別屬性分割資訊還原到原本的資料集中，建立完整的分割資料集。如表 4-5 所示。

表 4-5 屬性值分割資訊資料集

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 Value	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 Value
12	3	4	6
13	4	5	2
14	5	1	4
15	6	2	6
16	Missing Value	3	6
17	7	5	4
18	8	1	8

步驟五：找出遺漏值資料的相似分割

利用建立完成之屬性值分割資料，來找出和具有遺漏值資料分割結果相似的完整資料，並將這些資料視為同一群集。檢查遺漏值資料中非遺漏值屬性之分割結果，與資料集中其餘完整資料做比對，利用該屬性分割結果之前後群集作為該遺漏資料的相似分割結果。而為了避免具有相同屬性分割結果的資料數量過少，因此我們以具有相似屬

性分割結果的資料開始搜尋，若過濾出的資料少於三筆，則逐漸減少屬性分割數量繼續往下比對，直到過濾出具有同樣相似屬性數量達到三筆以上的資料，藉以取得更多的遺漏值填補參考資訊。

以表 4-5 中的第 16 筆為例，在該筆資料中除了遺漏值以外的均勻度與選擇性屬性分割結果分別為第 3 分割與第 6 分割，而我們的資料過濾範圍根據此屬性分割的前後一個分割來決定，均勻度的部份搜尋分割結果為 2、3、4 的資料，而選擇性屬性則搜尋分割結果為 5、6、7 的資料，並且將這兩個搜尋結果取聯集組合，資料過濾結果如表 4-6。

表 4-6 屬性值分割資訊

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
16(遺漏資料)	Missing Value	3	6
9	8	4	6
12	3	4	6
15	6	2	6

由表 4-6 中可以看出與第 16 筆遺漏值資料相似的完整資料分別有 9、12、15 三筆，我們將這三筆資料與該遺漏資料視為是同一群集，並且以此作為資料回填的參考。

步驟六：計算相似分割之屬性平均值回填

根據前一步驟中所過濾出的資料內容來計算出遺漏值回填數值，計算方式是將所有相似資料群集的數值加總並平均，並將該平均值作為遺漏值推估回填之結果。表 4-7 為表 4-6 中的資料原始數值。

表 4-7 屬性值分割後結果的資料原始值範例

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
16(遺漏資料)	Missing Value	8.76	5.2
9	660	12.4	5.3
12	340	10.5	5.3
15	494	7.08	5.4

延續上個步驟中的第 16 筆資料為例，表 4-7 為過濾後的相似群集資料原始數值，我們利用相似群集中的屬性平均值來回填遺漏值，而根據表 4-7 中可得知數值所求得之研磨速度的屬性平均值為 $(660+340+494)/3=498$ ，因此我們以 498 作為該遺漏資料的回填值。並且重複上述步驟五與步驟六，直到所有遺漏數值皆估計填補完畢為止。所有遺漏資料回填結果在表 4-8。

表 4-8 屬性值分割遺漏值回填結果

資料編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)
5	437	10.8718	4.9
11	268	26.9	4.625
16	498	8.76	5.2

步驟七：計算估計值誤差並輸出遺漏值回填結果

當遺漏值回填完成之後，最後利用原始資料與回填值來計算的遺漏值推估的誤差。延續文獻[9][12]中的作法，我們也採用平均絕對誤差 (Mean Absolute error, MAE) 作為遺漏值回填結果的評估準則，MAE 的計算方式如公式 4-2，將所有的遺漏估計值 \hat{e}_i 與真實資料 O_i 的誤差做加總，並且依照其遺漏值個數 N 計算平均，即可得到遺漏值推估結果的 MAE 值。

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{e}_i - O_i| \quad (4-2)$$

以我們在表 4-8 回填的結果為例，其 MAE 值為：

$$\frac{1}{3} \times (|468.75 - 483| + |10.871818 - 8.7| + |4.625 - 4.1|) = 5.898939$$

若 MAE 值越小，就表示估計值與真實資料越接近，遺漏值推估的效果也就越好，反之則代表遺漏值估計效果不佳。在表 4-9 中我們以屬性平均值、K-means、類神經網路(NBEMS)[9]、粒子群演算法(RKPSO)[12]、與屬性值分割遺漏值推估方法來進行遺漏值推估的結果比較。

表 4-9 遺漏值推估回填方法之誤差比較

方法	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)	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	1.5500	0.2500	3.0444
屬性值分割	15	2.1718	0.5250	5.8989

由表中可以發現，相較於其他遺漏值推估方法，利用屬性值分割來估計遺漏值更能有效的提高遺漏值估計的準確率，但略遜於粒子群最佳化的遺漏值估計方法。

第三節、基於蟻群分群方法之遺漏值推估模組

在第三章中我們已經介紹過改良式之蟻群分群演算法，該演算法主要延續自 Shelokar 等學者所發展之蟻群分群方法[6]而來，並且針對該分群方式的求解效率與品質作探討。而在此章節中我們將利用此改良式蟻群分群方法作為遺漏值推估模組的基礎架構，主要可分為兩大部份，分別是改良式蟻群分群模組與遺漏值估計回填模組，遺漏值推估架構如圖 4-3 所示。

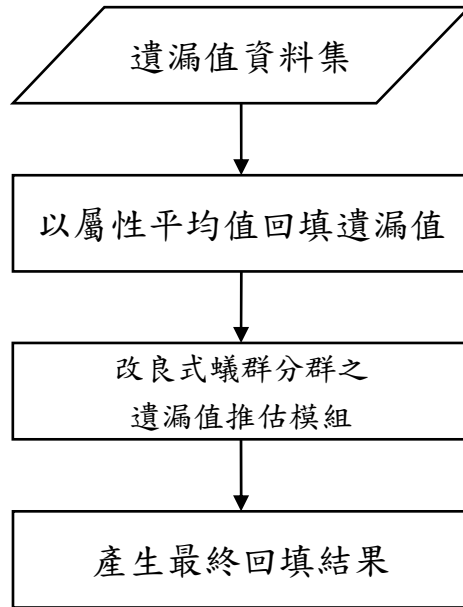


圖 4-3 蟻群分群遺漏值推估模組架構圖

在開始進行遺漏值推估之前，系統會先將資料中的遺漏值部份以傳統屬性平均值方式回填，讓蟻群分群模組得以進行分群。初始回填資料內容顯示如表 4-10。

表 4-10 半導體銅製程遺漏資料集-以屬性平均值回填初始值

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

利用屬性平均值作為初始回填資料後，即可開始進行蟻群分群演算流程。由於在第三章中已經詳細介紹過改良式蟻群分群演算法的運作過程，因此在這裡我們將直接說明遺漏值推估模組的運作方式。蟻群分群遺漏值推估模組流程如圖 4-4。

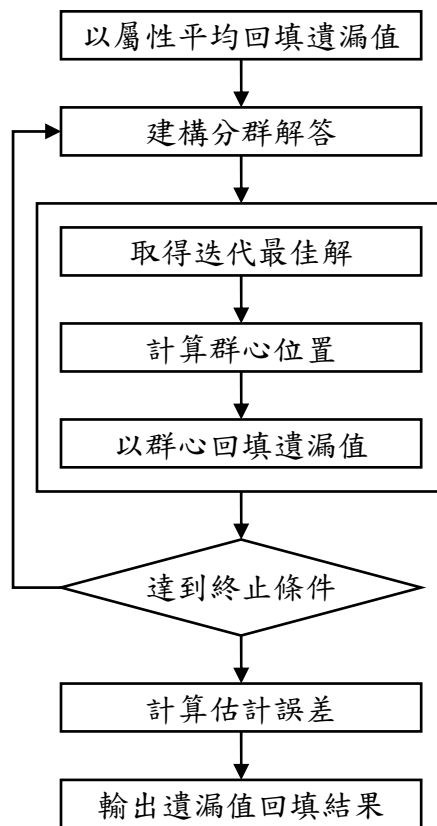


圖 4-4 蟻群分群方法之遺漏值推估模組流程圖

步驟一：取得迭代最佳分群結果

首先我們先將填入屬性平均值的初始回填資料作分群，經由蟻群分群結果得到該演算迭代中具有最低 fitness 值的分群解答，並且輸出

該解答至遺漏值推估模組。如表 4-11 為第一個迭代中最佳的螞蟻分群結果，我們利用這個分群解答來進一步計算該分群結果的群心位置。

表 4-11 首次迭代分群結果

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	1
5	437	12.735	4.9	3
6	498	6.5	6.1	3
7	481	8.99	4.2	3
8	588	11.8	4.3	2
9	660	12.4	5.3	2
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	1
14	434	5.06	4.7	3
15	494	7.08	5.4	3
16	430.706	8.76	5.2	3
17	580	15.1	4.6	2
18	651	5	5.8	2

步驟二：計算分群結果群心位置

群心位置的計算如同我們在蟻群分群方法中所描述的，將同一個群集中的所有資料，依照個別屬性加總並求其平均值，計算完群內所有屬性之平均值後即可得到該分群之群心位置。表 4-12 紀錄了表 4-11 的分群結果群心位置。

表 4-12 首次迭代分群結果群心位置範例

群集編號	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)	資料數
1	312.375	16.975	4.527	8
2	619.750	11.075	5.000	4
3	462.451	8.188	5.083	6

步驟三：根據群心位置回填遺漏值

接著我們利用分群結果所計算出的群心位置來作為遺漏值回填依據，檢查分群結果中具有遺漏值的資料被分配到那一個群集中，就以該群集群心值來回填遺漏值。回填結果如表 4-13 所示。

表 4-13 首次迭代後遺漏值回填結果範例

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	8.188	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	4.527
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	462.451	8.76	5.2
17	580	15.1	4.6
18	651	5	5.8

步驟四：計算回填結果與原始資料誤差

根據迭代最佳分群結果計算群心，並用以作為遺漏資料回填值，再將回填之後的結果重新進行分群，如此重複直到滿足演算迭代次數為止，我們以最後所回填之遺漏值作為的最後的估計結果，並且計算回填結果之 MAE 值。最後的遺漏值回填結果如表 4-14 所示。

表 4-14 遺漏值回填結果

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	7.278	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	4.486
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	468.8	8.76	5.2
17	580	15.1	4.6
18	651	5	5.8

以我們在表 4-14 回填的結果為例，其 MAE 值為：

$$\frac{1}{3} \times (|468.8 - 483| + |7.278 - 8.7| + |4.485714 - 4.1|) = 5.335905$$

表 4-15 中顯示了原始蟻群分群以及改良式蟻群分群方法的遺漏值推估誤差與 MAE 值，從表中可得知在遺漏值推估問題上，相較於原先的蟻群分群方法，我們所提出的改良式蟻群分群方法更能有效的提高遺漏值估計的準確率，而且也比屬性值分割推估方法要來的好，僅略遜於粒子群遺漏值估計方法。

表 4-15 遺漏值推估回填方法之誤差比較

方法	研磨速(RR)	均勻度 (NU；%)	選擇性 (Tan/Cu)	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	1.5500	0.2500	3.0444
屬性值分割	15	2.1718	0.5250	5.8989
蟻群分群	41.0000	1.4656	0.8500	14.4385
蟻群分群改良	14.2000	1.4220	0.3857	5.3359

第四節、基於屬性值分割與蟻群分群方法之遺漏值推估模組

在前面兩個章節中分別介紹了以屬性值分割以及蟻群分群為基礎的遺漏值推估方式，透過簡單的實例資料計算，我們可以發現這兩種作法都能取得比傳統屬性平均值方式還要精確的回填值，因此我們希望能夠藉由屬性值分割以及蟻群分群兩種方式的結合，改善整體遺漏值估計之效能。利用屬性值分割方法所估計之遺漏值作為蟻群分群方法中的初始分群資料，並且以此資料進行蟻群分群遺漏值推估，主要目的是想藉由屬性值分割能得到比傳統屬性平均值回填更好的回填結果之特性，作為蟻群分群模組的前導資料，以期能達到相輔相成的遺漏值推估效果。圖 4-5 為本研究所提出之遺漏值推估模組流程圖。

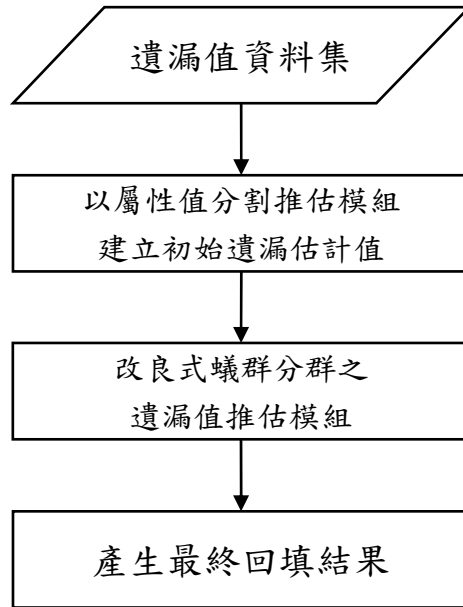


圖 4-5 蟻群分群遺漏值推估模組架構圖

從表 4-16 中可以發現在半導體銅製程遺漏資料集的估計上，改良式的蟻群分群推估方法與結合屬性值分割之改良式蟻群分群推估方法兩者所回填的估計值並無差別，在第五章中我們將測試四個不同資料集在四種不同遺漏比例的遺漏值推估結果比較，驗證各種推估方法組合在各種不同資料規模的估計成效。

表 4-16 遺漏值推估回填方法之誤差比較

方法	研磨速(RR)	均勻度(NU；%)	選擇性(Tan/Cu)	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	1.5500	0.2500	3.0444
屬性值分割	15	2.1718	0.5250	5.8989
蟻群分群	41.0000	1.4656	0.8500	14.4385
蟻群分群改良	14.2000	1.4220	0.3857	5.3359
屬性值分割+ 蟻群分割改良	14.2000	1.4220	0.3857	5.3359

第五章 實驗設計與結果分析

在第四章中我們已經利用半導體銅製程遺漏值估計回填問題作為例子，並與其他遺漏值推估方法做比較，說明並證明了本研究方法對於遺漏值推估問題的可行性。本章節將對我們所提出的方法做進一步的探討，分別透過真實資料的群集分析以及遺漏值推估實驗，觀察改良式蟻群分群方法的改善效果，並且瞭解不同的遺漏值估計方法，對於不同規模的資料庫與遺漏值數量有著何種影響。在第一節中我們將簡述實驗環境與使用工具。第二節介紹實驗中所使用的演算法參數設定。第三節描述我們所採用之四個真實資料集的相關資訊。第四節則說明本研究中的實驗設計與實驗結果之評估依據。第五節為實驗結果彙整介紹，我們將在此節中討論各項實驗與使用方法之結果分析。

第一節、實驗環境

本研究所使用的演算程式是以 C# 語言撰寫，開發工具為 Microsoft Visual Studio 2005，實驗環境平台規格如下：

CPU：Intel Pentium 4 3.00GHz

RAM：DDR400 1GB

OS：Microsoft Windows XP Professional

第二節、實驗參數設定

在之前的章節中我們主要介紹了兩種不同遺漏值推估方法，分別是屬性值分割遺漏值推估模組以及蟻群分群遺漏值推估模組，其中由

於屬性值分割方法僅需利用原始資料個別屬性的距離資訊，即可完成分割演算程式，因此在屬性值分割方法部份我們無須給定任何參數設定值。而在蟻群分群方法中，由於不同的分群數量、螞蟻數量、演算迭代次數等相關參數設定，將會對演算結果造成重大的影響，因此我們必須明定實驗中所使用的參數設定組合，以確保實驗的公平性與結果的穩定性。參數設定內容如表 5-1 所示。

表 5-1 參數設定內容

螞蟻數量	50
改良解答個數	5
最大迭代次數	1000 (分群問題) / 500 (遺漏值推估)
費洛蒙散失率	0.1
轉換規則機率	0.8
重新分配機率	0.01
分群群數	依資料集內容調整

第三節、實驗資料描述

本研究的實驗主要分為兩大部份，分別是改良式蟻群分群方法與傳統蟻群分群的比較，以及屬性值分割、蟻群分群和其他演算方法對於遺漏值推估之成效實驗分析。在本研究中我們採用與文獻[12]相同的四個資料集作為實驗結果的比較目標，分別為天然石油 Crude Oil、鴛尾花 Iris、玻璃 Glass 與母音 Vowel，此四個規模與性質各異的真實資料集，並分為 5%、10%、15%與 20%這四個不同比例的遺漏程度，我們藉此測試在不同資料規模與遺漏程度比例時，各種方法的遺漏值推估效果。實驗資料內容如表 5-2，以下將介紹個別資料集的主要特徵。

表 5-2 實驗資料集內容

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

1. 天然石油資料集 (Crude Oil)

此資料庫為本次實驗中筆數最小的資料集，僅有 56 筆資料，每筆資料皆有 5 個屬性內容，分別有 Vanadium、Iron、Beryllium、Saturated Hydrocarbons 以及 Aromatic Hydrocarbons，共分為 Wilhelm、Sub-Mulinia 與 Upper 這 3 個群集，屬性與群集資訊如表 5-3 與表 5-4。

表 5-3 Crude Oil 資料集屬性資訊

屬性名稱	最大值	最小值	平均值	標準差
Vanadium	11.00	1.20	6.1804	2.4243
Iron	52.00	5.60	27.0464	11.6055
Beryllium	1.50	0.00	0.3414	0.3139
Saturated Hydrocarbons	9.25	3.06	5.2991	1.3872
Aromatic Hydrocarbons	13.01	2.22	6.4336	3.1546

表 5-4 Crude Oil 資料集群集資訊

群集名稱	群內資料筆數	群內資料比例
Wilhelm	7	12.50%
Sub-Mulinia	10	17.86%
Upper	39	69.64%

2. 鸞尾花資料集 (Iris Plants)

鸞尾花資料集中共有 150 筆資料，資料分別有 4 個屬性，分別為花萼長度 (Sepal Length)、花萼寬度 (Sepal Width)、花瓣長度 (Petal Length) 與花瓣寬度 (Petal Width)，根據鸞尾花的屬性可以分為 Iris Setosa、Versicolour 與 Virginica 三個品種，如表 5-5 與表 5-6 所示。

表 5-5 Iris Plants 資料集屬性資訊

屬性名稱	最大值	最小值	平均值	標準差
Sepal Length	7.9	4.3	5.8433	0.8281
Sepal Width	4.4	2.0	3.0540	0.4336
Petal Length	6.9	1.0	3.7587	1.7644
Petal Width	2.5	0.1	1.1987	0.7632

表 5-6 Iris Plants 資料集群集資訊

群集名稱	群內資料筆數	群內資料比例
Iris Setosa	50	33.33%
Versicolour	50	33.33%
Virginica	50	33.33%

3. 玻璃資料集 (Glass)

Glass 資料庫中有 214 筆資料，屬性中分別紀錄了 9 種不同化學元素的成份數值，總共分為 6 群，詳細內容如表 5-7 及表 5-8 所示。

表 5-7 Glass 資料集屬性資訊

屬性名稱	最大值	最小值	平均值	標準差
Ri	1.5339	1.5112	1.5184	0.0030
Na	17.3800	10.7300	13.4079	0.8166
Mg	4.4900	0.0000	2.6845	1.4424

Ai	3.5000	0.2900	1.4449	0.4993
Si	75.4100	69.8100	72.6509	0.7745
K	6.2100	0.0000	0.4971	0.6522
Ca	16.1900	5.4300	8.9570	1.4232
Ba	3.1500	0.0000	0.1750	0.4972
Fe	0.5100	0.0000	0.0570	0.0974

表 5-8 Glass 資料集群集資訊

群集名稱	群內資料筆數	群內資料比例
Float Building Windows	70	32.71%
Float Vehicle Windows	17	7.94%
Non-Float Building Windows	76	35.51%
Containers	13	6.07%
Tableware	9	4.21%
Headlamps	29	13.55%

4. 母音資料集 (Vowel)

母音資料集則是印第安語言的母音音頻資料，共有 871 筆紀錄，根據 3 種音頻區分為 6 種母音分類，資料屬性與群集資訊如表 5-9 及表 5-10 所示。

表 5-9 Vowel 資料集屬性資訊

屬性名稱	最大值	最小值	平均值	標準差
F1	900	250	470.4822	129.2222
F2	2550	700	1514.6840	507.5812
F3	3200	1800	2561.0220	244.5418

表 5-10 Vowel 資料集群集資訊

群集名稱	群內資料筆數	群內資料比例
δ	72	8.27%
a	89	10.22%
i	172	19.75%
u	151	17.34%
e	207	23.77%
o	180	20.67%

第四節、實驗設計

本研究中的實驗設計主要分為兩個階段，我們首先比較改良式蟻群分群方法與傳統蟻群分群方法的群集分析求解效率，利用第三節中所介紹的四個資料集作為分群對象，並進行 10 次獨立實驗，並且紀錄求解過程中最好的 fitness 值作為分析數據，用以比較兩種方法在分群問題的求解時間與求解品質上是否有所差異。

至於遺漏值推估問題部份，我們則分別利用此四個原始資料集與文獻[12]中用以實驗的遺漏資料集作為實驗對象，測試本研究所提出的遺漏值推估模組與屬性值平均、K-means 分群法、粒子群最佳化方法以及屬性值分割方法，在不同資料規模與遺漏值比例之下遺漏值推估的效率與品質。在這裡我們以遺漏值估計過程最後收斂的 MAE 值作為遺漏值估計品質的判斷依據，對於各個資料集的四種不同遺漏比例均進行 10 次的獨立測試，藉以分析探討各種遺漏值推估方法的演算效果與品質，以及方法的適用時機。

第五節、實驗結果分析

本節主要分為兩個階段，分別為群集分析實驗以及遺漏值推估之實驗結果與分析，並藉由實驗結果的觀察來探討各種方法的適用時機。數據部份我們用粗體字型表示該實驗中的最佳結果，而表格中灰底部份則是我們提出方法之實驗結果。

壹、群集分析實驗結果

在這個階段我們利用分別利用 K-means、基因分群方法、傳統蟻群分群方法以及我們所提出的改良式蟻群分群方法做比較，分別進行 10 次的獨立實驗，K-means 與基因分群皆執行 1000 個迭代，蟻群分群參數則以表 5-1 的內容做設定。實驗結果彙整如下：

表 5-11 Crude Oil 分群之實驗結果比較

Crude Oil	K-means	基因分群	蟻群分群	改良式蟻群分群
最佳 fitness	279.2710	278.9652	277.5911	278.5295
平均 Fitness	279.3182	279.2534	277.9024	278.6574
Fitness 標準差	0.1493	0.2495	0.1984	0.1530

由表 5-11 的實驗數據中可發現，加入幾何距離所改良之蟻群分群效果稍稍遜於蟻群分群方法，但兩者之間的差異並不大，而且其分群結果也僅略優於基因分群與 K-means 分群方法，改良式蟻群分群方法並不具有明顯的演算優勢。

表 5-12 Iris Plants 分群之實驗結果比較

Iris Plants	K-means	基因分群	蟻群分群	改良式蟻群分群
最佳 fitness	97.3259	97.2221	97.2221	96.8116
平均 Fitness	115.7323	97.2771	97.7185	96.8530
Fitness 標準差	12.7009	0.0516	0.4208	0.0311

相較於 K-means 的平均分群 fitness 結果，改良式蟻群分群方法的 fitness 值顯得相當良好，而整體而言亦優於基因分群與蟻群分群方法，其中就標準差來看也較蟻群分群方法來的穩定許多。

表 5-13 Glass 分群之實驗結果比較

Glass	K-means	基因分群	蟻群分群	改良式蟻群分群
最佳 fitness	244.9772	227.8953	260.4070	212.8757
平均 Fitness	289.0779	269.5492	270.4087	215.3186
Fitness 標準差	44.7782	17.9908	6.8780	1.6302

在表 5-13 的 Glass 資料集分群結果中可發現，改良式蟻群分群方式能取得相對最佳的演算結果，而且由標準差數值可以發現改良式蟻群分群顯得相當穩定，是以上幾種演算法中求解結果最佳的方法。

表 5-14 Vowel 分群之實驗結果比較

Vowel	K-means	基因分群	蟻群分群	改良式蟻群分群
最佳 fitness	149489.1867	149417.1344	326036.9558	149828.3901
平均 Fitness	151862.1662	157618.7685	344154.9849	151580.8489
Fitness 標準差	4197.3769	10137.8415	16732.3586	2117.9317

而在 Vowel 資料集中，雖然改良式蟻群所取得的最佳 fitness 值略差於基因分群與 K-means 的結果，但就平均 fitness 與標準差而言，改良式蟻群分群方法仍具有相當穩定的求解品質。

加入了幾何距離考量以及 local search 改良之後的蟻群分群方法，確實具有提高分群求解收斂速度的特性。圖 5-1 顯示了個別演算迭代所求得的分群 fitness 值，在 1000 個演算迭代過程中，改良式蟻群分群方法在演算初期就能找到夠好的分群結果，而傳統蟻群分群在同樣的參數設定之下，直到滿足迭代限制之前都還沒有出現與改良式方法相當的分群結果，因此我們認為改良式方法有助於改善分群求解的品質與速度，數據如表 5-13 所示。

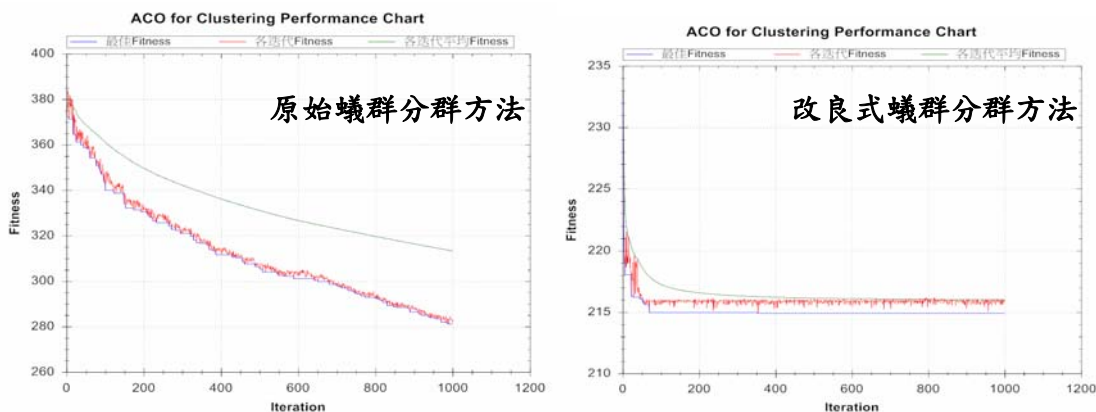


圖 5-1 Glass 資料集分群之 fitness 收斂圖

此外，在群集分析問題上改良後的方法相較於原始的蟻群分群能取得更好的分群求解結果，因此我們認為本研究所提出的蟻群分群改良方式確實有助於提昇蟻群在求解資料分群問題的演算效能。

貳、遺漏值推估實驗結果

在此階段我們將比較不同資料集的遺漏值推估結果，分別比較了屬性值平均回填、K-means、RKPSO、屬性平均加上原始蟻群分群方法、屬性平均加上改良式蟻群分群方法、屬性值分割方法以及屬性值分割加上改良式蟻群分群…共七種不同的遺漏值推估方法，同樣進行 10 次個別獨立實驗，並以其結果作為比較數據。

首先是 Oil 資料集的結果比較，從下面四個表格的實驗結果可以發現，改良式蟻群分群遺漏值推估以及加上屬性分割後的蟻群遺漏值推估方法雖然無法取得最佳的遺漏估計值，但以遺漏值推估穩定度來比較，此兩種作法在此四種不同的遺漏比例中，皆能夠取得相對穩定的估計結果。而在 5% 與 10% 遺漏值部份，雖然遺漏值推估的最佳 MAE 值遜於蟻群分群方法，但相較於其他方法而言，這兩種方法能得到更為相對精確的遺漏估計值。

表 5-15 Crude Oil 5% 遺漏值比例之實驗結果比較

Oil 5% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	1.1783	0.7148	0.6747	0.2974	0.4465	1.0544	0.4466
平均 MAE 值	1.1783	0.7148	0.7408	0.8015	0.4664	1.0544	0.4689
MAE 值標準差	-	0	0.1045	0.2539	0.0105	-	0.0079

表 5-16 Crude Oil 10% 遺漏值比例之實驗結果比較

Oil 10% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.9924	0.5549	0.7690	0.3389	0.3717	0.9147	0.3717
平均 MAE 值	0.9924	0.5549	0.7911	0.7189	0.3935	0.9147	0.3911
MAE 值標準差	-	0	0.0386	0.1727	0.0077	-	0.0102

表 5-17 Crude Oil 15% 遺漏值比例之實驗結果比較

Oil 15% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	1.2876	0.9656	0.6731	0.7529	0.9722	1.0075	0.9722
平均 MAE 值	1.2876	1.0334	0.7644	1.0818	0.9722	1.0075	0.9722
MAE 值標準差	-	0.0448	0.0619	0.1722	0	-	0

表 5-18 Crude Oil 20% 遺漏值比例之實驗結果比較

Oil 20% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	1.7540	1.0081	0.9090	0.9845	1.0942	1.6843	1.0941
平均 MAE 值	1.7540	1.0360	1.3275	1.6579	1.1076	1.6843	1.1096
MAE 值標準差	-	0.0588	0.2488	0.5112	0.0092	-	0.0081

在 Iris 資料集的實驗中，可以發現雖然四種遺漏比例的最佳遺漏值估計結果皆以 RKPSO 為優，但以平均估計值而言，屬性值分割結合蟻群分群的遺漏值估計方法在這四項實驗中都能取得相當良好的回填結果，而且遺漏值估計回填的求解品質也十分穩定。而我們也觀察到藉由改良式蟻群分群方法的求解品質相當穩定之特色，再加上利用屬性值分割方法的良好推估結果作為遺漏值推估模組的初始資料，能使得此兩種方法結合的求解效果能夠有效且穩定的提昇，得到比原先單純蟻群分群推估方法或是屬性值分割方法還要優良的遺漏估計值。

表 5-19 Iris Plants 5% 遺漏值比例之實驗結果比較

Iris 5% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.7615	0.3554	0.2797	0.6058	0.6112	0.3775	0.2834
平均 MAE 值	0.7615	0.4815	0.2885	0.7997	0.6129	0.3775	0.2840
MAE 值標準差	-	0.1330	0.0254	0.1750	0.0009	-	0.0002

表 5-20 Iris Plants 10% 遺漏值比例之實驗結果比較

Iris 10% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.6569	0.3293	0.1992	0.4321	0.3925	0.2771	0.2184
平均 MAE 值	0.6569	0.3724	0.2403	0.6472	0.3935	0.2771	0.2186
MAE 值標準差	-	0.0454	0.0230	0.1315	0.0004	-	0.0003

表 5-21 Iris Plants 15% 遺漏值比例之實驗結果比較

Iris 15% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.7026	0.3961	0.2640	0.4402	0.3949	0.3106	0.2774
平均 MAE 值	0.7026	0.4217	0.2971	0.6215	0.3949	0.3106	0.2775
MAE 值標準差	-	0.0188	0.0137	0.1073	0	-	0

表 5-22 Iris Plants 20% 遺漏值比例之實驗結果比較

Iris 20% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.7808	0.3899	0.2656	0.4804	0.4521	0.3000	0.2662
平均 MAE 值	0.7808	0.4770	0.2895	0.7163	0.4521	0.3000	0.2662
MAE 值標準差	-	0.0459	0.0185	0.1156	0	-	0

而在 Glass 資料集中，單純屬性分割方法在 5%、10% 與 15% 的遺漏比例時能取得比 K-means、RKPSO 等方法還要準確的回填結果。但在此資料集中的 20% 遺漏值比例實驗中，利用屬性值分割推估方法所取得的估計結果反而是最差的，而在結合了蟻群分群推估方法之後則能夠取得與 RKPSO 方法差不多的估計結果。

表 5-23 Glass 5% 遺漏值比例之實驗結果比較

Glass 5% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.5608	0.3496	0.3656	0.3776	0.2622	0.2375	0.2699
平均 MAE 值	0.5608	0.4190	0.4823	0.5521	0.3073	0.2375	0.3067
MAE 值標準差	-	0.0755	0.0971	0.0863	0.0346	-	0.0282

表 5-24 Glass 10% 遺漏值比例之實驗結果比較

Glass 10% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.6621	0.4992	0.5201	0.6024	0.4644	0.3716	0.5289
平均 MAE 值	0.6621	0.5352	0.5609	0.6876	0.5689	0.3716	0.5833
MAE 值標準差	-	0.0322	0.0210	0.0677	0.1109	-	0.0375

表 5-25 Glass 15% 遺漏值比例之實驗結果比較

Glass 15% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.5772	0.3674	0.3786	0.5093	0.3763	0.3337	0.4677
平均 MAE 值	0.5772	0.4368	0.4446	0.5974	0.4197	0.3337	0.4841
MAE 值標準差	-	0.0268	0.0298	0.0541	0.0291	-	0.0108

表 5-26 Glass 20% 遺漏值比例之實驗結果比較

Glass 20% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	0.6430	0.4242	0.3736	0.5399	0.3987	1.0228	0.3819
平均 MAE 值	0.6430	0.4422	0.4393	0.6361	0.4467	1.0228	0.4459
MAE 值標準差	-	0.0139	0.0612	0.0573	0.0391	-	0.0520

至於在 Vowel 資料集的實驗中，不論是蟻群分群、改良式蟻群分群或是結合屬性值分割方法的推估結果皆遜於 RKPSO 方法，而屬性值分割方法的推估結果又優於蟻群分群與改良式蟻群分群。但以由改良式蟻群分群方法再加入了屬性值分割前後的結果看來，利用屬性值分割的推估結果作為蟻群分群的前導資訊，能夠略為降低遺漏值回填的變異，提高求解的穩定效果。

表 5-27 Vowel 5% 遺漏值比例之實驗結果比較

Vowel 5% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	309.7194	275.9669	132.2301	297.4949	225.9199	219.0725	224.1896
平均 MAE 值	309.7194	287.7374	185.3032	313.1793	274.7580	219.0725	261.5080
MAE 值標準差	-	8.0933	26.3104	10.4972	42.3474	-	32.7474

表 5-28 Vowel 10% 遺漏值比例之實驗結果比較

Vowel 10% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	259.1569	231.3289	157.1560	254.2526	188.7380	182.2576	201.1319
平均 MAE 值	259.1569	244.0729	177.5010	260.2245	253.6232	182.2576	219.9316
MAE 值標準差	-	11.1217	9.3793	4.6490	37.9051	-	16.6107

表 5-29 Vowel 15% 遺漏值比例之實驗結果比較

Vowel 15% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	240.9259	216.9966	163.2652	235.5387	185.7410	175.8178	182.5354
平均 MAE 值	240.9259	229.8197	179.7402	240.4132	226.1651	175.8178	198.9031
MAE 值標準差	-	10.5983	10.0421	3.7359	31.6498	-	11.2483

表 5-30 Vowel 20% 遺漏值比例之實驗結果比較

Vowel 20% Missing	屬性值平均	K-means	RKPSO	屬性平均 + 蟻群分群	屬性平均 + 改良式蟻群分群	屬性值分割方法	屬性分割 + 改良式蟻群分群
最佳 MAE 值	240.3617	205.7502	157.1018	235.9827	167.8388	170.8998	176.2866
平均 MAE 值	240.3617	212.5961	185.0526	240.1872	213.7489	170.8998	199.1224
MAE 值標準差	-	8.8563	14.8001	2.6768	29.6124	-	13.0057

綜合以上實驗結果，在 Oil 與 Iris 此類小樣本的遺漏值估計問題上，蟻群分群遺漏值推估方法以及結合屬性值分割遺漏值推估方法通常能夠取得十分穩定的估計結果，表示在每次的遺漏值估計都

能產生同等品質的解答，但在遺漏值估計準確度方面可能會因為分群結果改變、初始回填值的不同，而造成分群演算在進入收斂階段時落入了遺漏值估計的區域解中，導致最後收斂的分群結果不一定會是最理想的遺漏值回填值。

我們在研究過程中也發現分群結果與遺漏值的估計效果之間並不具有絕對的因果關係，分群效果良好並不代表所估計的回填值就比較準確。由於我們的方法是在每一次分群完成後，以該遺漏資料的所在群集之群心值作為回填的依據，再次進行下一個迭代的分群，因此在各個迭代中分群資料的初始值都不盡相同，而所回填的估計值又會影響整個資料分群之結果；加上遺漏值估計問題有著原始資料未知的最大前提，因此我們在遺漏值推估的過程中並沒有任何的資訊能夠作為估計的參考，僅能藉由實驗結果來判斷遺漏值估計的成效，間接證實不同遺漏值推估方法的適用時機與成效。

第六章 結論與未來展望

第一節、 結論

本研究以群集分析概念，結合蟻群分群演算法以及屬性值分割方法，提出了基於屬性值分割與蟻群分群方法之遺漏值推估模組，試圖藉由屬性值分割方法所求得之資料分割群集，計算出遺漏值估計的回填值，並且用以作為改良式蟻群分群方法的分群初始資料，加快群集分析的收斂速度，並且利用群集群心位置作為遺漏值估計之回填值，希望能利用兩種不同方法的結合來解決遺漏值推估問題。而經由我們的研究得到以下結論：

1. 在分群求解的部份，原始的蟻群分群演算法僅利用費洛蒙資訊作為求解建構參考，必須經由大量迭代才能累積足夠的求解資訊，使得求解收斂速度過於緩慢。因此我們加入了資料分佈的幾何距離考量，經由實驗證明在相同的求解品質之下，改良式蟻群分群方法確實有助於提昇分群求解的收斂速度。
2. 在原始的蟻群分群方法中以編碼字串形式來分群解答，但這樣的作法會使得費洛蒙資訊無法有效的累積，造成分群求解效率下降。而我們藉由分群編碼字串順號的作法，讓同樣的分群結果能藉由相同的分群編碼，有效地將求解資訊累積到費洛蒙矩陣中，藉以提高蟻群求解效率。
3. 利用群集分析結果作為遺漏值估計的主要依據，能夠取得比傳統屬性平均值回填的方法還要好的估計結果。根據所採用的分群技術不同，其遺漏值推估的結果也會有所差異，而結合屬性

值分割與蟻群分群技術為基礎之遺漏值估計方式，在樣本數量較小的遺漏資料中，能夠取得還不錯的遺漏值估計效果，而且其求解品質也相當穩定，雖然在部份資料集中可能無法求得夠好的估計結果，但仍優於傳統屬性平均值的回填方式。

第二節、未來展望

蟻群最佳化演算法雖然具有正向回饋的費洛蒙機制以及貪婪啟發式方法的演算特色，有助於讓求解資訊能隨著演算迭代而逐漸累積，但由於蟻群演算法必須依靠大量迭代以及螞蟻進行搜尋，使得整體求解過程需要經過長時間的搜尋、運算與累積，才能找出夠好的問題解答，因此如何促進費洛蒙資訊的有效累積，正是蟻群最佳化方法中相當重要的課題之一。

而在蟻群分群演算法中，由於分群結果的編碼方式限制，導致費洛蒙資訊無法快速有效的累積，未來可嘗試調整費洛蒙矩陣資訊的紀錄方式，跳脫原本利用資料與群集關係作為費洛蒙資訊紀錄的方式，直接將費洛蒙紀錄在資料與資料之間，藉以改善原有分群編碼方式的缺失，讓分群資訊累積的效果能夠更加突顯。

此外在遺漏值推估部份，遺漏值資料的初始回填值對於分群結果具有一定的影響，連帶造成了後續遺漏值估計的結果，因此日後的相關研究可以朝向觀察初始回填值對於遺漏值推估的影響，並進一步探討分群方法與遺漏值估計的主要影響因素。

參考文獻

- [1] J. MacQUEEN, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS", 1967.
- [2] M. Dorigo, G. Di Caro. "Ant colony optimization: a new meta-heuristic," IEEE Transactions on Evolutionary Computation, 1999.
- [3] M. Dorigo, and L. M. Gambardella. "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem," IEEE Transactions on Evolutionary Computation, 1(1), 1997, pp. 53–66.
- [4] M. Dorigo, M. Birattari, and T. Stutzle. "Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique," IEEE Computational Intelligence Magazine, November 2006, pp. 28-39.
- [5] M. Dorigo, V. Maniezzo, and A. Colorni. "The ant system: optimization by a colony of cooperating agents," IEEE Transactions on Systems, Man, and Cybernetics–Part B, vol. 26, No. 2, 1996, pp. 29–41.
- [6] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni. "An ant colony approach for clustering," Analytica Chimica Acta (509), 2004, pp. 187-195.
- [7] 李泰琳、張靖及卓裕仁，「調適型螞蟻演算法應用於旅行推銷員問題之研究」，運輸學刊，第十九卷第一期，民國 96 年 3 月，頁 89-120。
- [8] 沈永勝，「整合自動分群與加權式灰關聯技術於大型資料庫內遺失值之處理」，台灣科技大學電子工程學系碩士論文，民國 94 年 6 月。
- [9] 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碩士論文，民國 94 年 6 月。
- [10] 游裕昌，「運用基因群集技術於大型資料庫內遺失值之處理」，台灣科技大學電子工程學系碩士論文，民國 93 年 6 月。
- [11] 曾憲雄、蔡秀滿、蘇東興、曾秋蓉及王慶堯，「資料探勘」，旗標出版股份有限公司，台北，民國 96 年 3 月。
- [12] 魏岑甄，「基於反彈機 KPSO 分群之有效遺漏推估方法」，南華大學資訊管理學系碩士論文，民國 97 年 6 月。