東　　海　　大　　學

工業工程與經營資訊研究所

碩士論文

在小樣本情況下智慧參數搜尋之研究

研 究 生：張雅君
指導教授：王偉華 副教授

中 華 民 國 九 十 九 年 七 月

A research on intelligent parameters searching in small dataset

By
Ya-Chun Chang

Advisor: Prof. Wei-Hua Wang

A Thesis
Submitted to the Institute of Industrial Engineering and
Enterprise Information at Tunghai University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in
Industrial Engineering and Enterprise Information

July 2010
Taichung , Taiwan , Republic of China

在小樣本情況下智慧參數搜尋之研究

學生：張雅君　　指導教授：王偉華 副教授

東海大學工業工程與經營資訊研究所

摘要

實務上，對於參數設定範圍的尋找在研發實驗階段多以實驗方法行之，但實驗次數多寡常受限於成本、時間和人力上的考量。理論上要找出許多關鍵因素的參數設定範圍需要大量的實驗次數與且消耗許多研發經費與時間，這些都受到現實條件的限制。

本研究旨在發展一個在小樣本情況下，能快速找出參數設定範圍的搜尋機制。以一目標導向的尋求有效實驗點的方法，在有少量的實驗資料時，從中找出對於釐清參數設定範圍有幫助的實驗區域。研究中根據小樣本資料集，使用區間化核密度估計產生其虛擬資料，再以支援向量機建立分類模型。

本研究發展三個有效方法。一、將區間化核密度估計結合支援向量機建立分類模型；二、方法在相同的有效分類程度下，減少虛擬樣本產生的數量，以減少運算時間；三、使用輪盤法分散虛擬資料選擇的區域，使方法在相同的有效分類程度下，更具有收斂性。三個方法其試驗結果皆優於隨機選擇；且第三個方法具有收斂特性。

關鍵字詞：小樣本、分類、區間化核密度估計、支援向量機、輪盤法

A research on intelligent parameters searching in small dataset

Student: Ya-Chun Chang Advisor: Prof. Wei-Hua Wang

Department of Industrial Engineering and Enterprise Information
Tunghai University

ABSTRACT

Practically, the experiment is the major methodology in the R&D stage of searching for the right parameter settings of a new product development. However, the searching procedure is very much consuming the cost, time and manpower. That is, a method in enhancing the speed and quality of the searching process will be very much benefit in the product development process.

This research is focused on the developing a searching mechanism under the small size datasets to achieve a better quality of the region of parameter settings in a faster way. A goal-oriented method is developed in effectively using the previous experiments information to limit the further explore region. This research adopted Intervalized Kernel Density Estimation (IKDE) method to generate the virtual dataset based on the existed real small dataset. And then, Support Vector Machine (SVM) is used to find the classifier.

In this research, three improved methods have been developed: 1) purely IKDE combined with SVM to construct a classifier, 2) limited the generation of virtual dataset and achieve an equal quality of the classifier which showed the efficiency in computation time, 3) using roulette wheel method in exploring the region of virtual dataset but without losing the quality of the classifier and showed a better convergence property. All the methods showed a better quality than the general random methods. And, the last method showed a convergence property in out run all methods.

Keywords: Small dataset, Classifier, IKDE, SVM, Roulette wheel method

致謝

兩年研究生的生活，經過了碩士論文口試之後即將在此告一段落。對於自己能夠完成論文，現在想來還是覺得有點難以置信。

能夠順利的完成碩士學位，首先要感謝研究所的指導教授王偉華老師的指導，無論是在課業、研究或是待人處事，老師總是不厭其煩的提醒與叮嚀，讓我更靠近這個世界；不擅言詞的我，要謝謝老師的耐心聆聽及一次次的示範與說明，給予我許多練習的機會；平常漫不經心的我偶爾出的一些大小包，更是要感謝老師的包容。此外，特別感謝丁兆平老師及黃欽印老師於口試期間所給予的建議與指正，使這篇論文能夠更加完備。

感謝研究室的彣昕學姐、滿靜學姐與俊良學長，在作研究時，給予我許多寶貴的意見、想法及信心，有你們的幫忙我的論文才能夠順利的完成。也要謝謝研究所的同學們及朋友們，特別是同處一間研究室的豬妹、小羊以及大我一屆的硬漢，豐富了我兩年的生活。

最後，有家人的支持與叔叔嬸嬸的熱情贊助，才能讓我無憂慮的過完最後的校園生活。

兩年的時間很快就過去了，轉眼間又是一段新的開始。感謝大家的協助，在此與大家一同分享此份小小成果與喜悅。

張雅君 謹誌於

東海大學工業工程與經營資訊研究所

智慧型系統研究室

2010 荷月

目錄

中文摘要.....i
英文摘要.....ii
致謝..... iii
目錄.....iv
圖目錄.....vi
表目錄.....vii
第一章　緒論.....1
1.1 研究背景與動機.....1
1.2 研究目的.....1
1.3 研究方法與步驟.....2
1.4 論文架構.....3
第二章　文獻探討.....5
2.1 小樣本分類.....5
2.2 小樣本學習&虛擬樣本.....6
2.2.1 區間化核密度估計.....7
2.3 支援向量機器.....10
2.3.1 支援向量機基本概念.....10
2.3.2 特徵空間學習.....12
2.3.3 LIBSVM.....13
第三章　研究方法.....14
3.1 研究資料範圍.....14
3.2 研究工具.....14
3.3 研究機制建立.....15
3.3.1 小樣本問題與虛擬樣本的建立.....15
3.3.2 資料分類.....16
3.3.3 Support vector selection mechanism.....18
3.3.4 Local virtual selection mechanism.....24
3.3.5 Global virtual selection mechanism.....27
第四章　實例驗證.....31
4.1 資料集一.....31
4.1.1 試驗說明.....32
4.1.2 試驗結果記錄.....33
4.1.3 資料集一－試驗結果討論.....37

4.2 資料集二 40
4.2.1 試驗說明 42
4.2.2 試驗結果記錄 43
4.2.3 資料集二－試驗結果討論 47
4.3 本章小結 51
第五章 結論及未來研究方向 53
5.1 結論 53
5.2 未來發展方向 54
參考文獻 55

圖目錄

圖 1.1 研究架構.....4
圖 2.1 直方圖估計法每個區間使用相同的 h 值.....7
圖 2.2 使用不正確的 h 對估計分配的影響.....8
圖 2.3 核密度估計與區間化核密度估計估計結果之差異.....9
圖 2.4 SVM 說明圖一.....11
圖 2.5 空間轉換圖.....13
圖 3.1 研究架構.....15
圖 3.2 修正機制示意圖一.....21
圖 3.3 修正機制示意圖二.....22
圖 3.4 SVS 參數範圍邊界找尋流程.....23
圖 3.5 LVS 參數範圍邊界找尋流程.....27
圖 3.6 GVS 參數範圍邊界找尋流程.....30
圖 4.1 資料集一分類情況.....32
圖 4.2 以 LVS 機制增加 50 個新實驗於資料集一之正確率.....36
圖 4.3 資料集二分類情況.....41
圖 4.4 資料集二測試資料分佈區域.....42
圖 4.5 以 LVS 機制增加 200 個新實驗於資料集二之正確率.....46
圖 4.6 LVS 實驗多次結果.....52

表目錄

表 3.1 SVS 實驗參數選擇步驟一 19
表 3.2 實驗參數選擇步驟一結果 19
表 3.3 群聚中心計算範例 20
表 3.4 SVS 實驗參數選擇步驟二 20
表 3.5 LVS 之 RX 計算範例 24
表 3.6 LVS，RX 最近距離計算 25
表 3.7 LVS，RX 最近距離計算結果 25
表 3.8 LVS 實驗參數選擇 26
表 3.9 GVS 實驗參數選擇步驟一 28
表 3.10 適應函數 29
表 3.11 GVS Roulette wheel selection 29
表 4.1 資料集一基本資料表 31
表 4.2 資料集一，Method O 結果記錄 34
表 4.3 資料集一，SVS 結果記錄 34
表 4.4 資料集一，LVS 結果記錄 35
表 4.5 資料集一，GVS 結果記錄 35
表 4.6 資料集一，GVS 實驗次數結果記錄 37
表 4.7 資料集一受試者間因子 38
表 4.8 資料集一變異數分析 39
表 4.9 資料集一多重比較分析結果 39
表 4.10 資料集一同質子集結果 40
表 4.11 資料集二基本資料表 41
表 4.12 資料集二，Method O 結果記錄 44
表 4.13 資料集二，SVS 結果記錄 44
表 4.14 資料集二，LVS 結果記錄 45
表 4.15 資料集二，GVS 結果記錄 45
表 4.16 資料集二，GVS 實驗次數結果記錄 47
表 4.17 資料集二受試者間因子 48
表 4.18 資料集二變異數分析 49
表 4.19 資料集二多重比較分析結果 49
表 4.20 資料集二同質子集結果 50
表 4.22 試驗結果列表 51

第一章　緒論

1.1 研究背景與動機

近年來，由於全球市場環境競爭日益激烈，以及顧客需求習慣變遷，各項產品的生命週期越來越短，搶先上市以搶佔關鍵通路與顧客，是產品能否在市場上取得競爭優勢的重要因素之一。為了取得市場先機，新產品從研發到上市的間隔時間變短，如何將新產品從試作迅速導入量產，搶先獲得較高的市場佔有率，將是研發人員的挑戰，造成研發人員面臨著開發以及量產時程縮短的壓力，故縮短產品開發到量產的時程，成為重要的研究議題。

為了兼顧成本及縮短產出時程，或是基於某些因素的限制與考量只能抽取少量的樣本時，都會期望能透過少量的樣本進行分析及建立模型，得到不失其準確性結果。而實務上，對於參數設定範圍的尋找在研發實驗階段多以實驗方法行之，但實驗次數多寡常受限於成本、時間和人力上的考量。理論上要找出許多關鍵因素的參數設定範圍需要大量的實驗次數與且消耗許多研發經費與時間，這些都受到現實條件的限制。

故本研究期望能建立一目標導向的尋求有效實驗點的方法，在有少量的實驗資料時，根據已知的資料及其類別屬性來建立資料的分類模型，從中找出對於釐清參數設定範圍有幫助的實驗區域，利用先前實驗的少量實驗樣本提供的資訊來選擇後續的實驗參數。期望此方法能有效的使用先前少量實驗樣本提供的資訊，亦能利用後續選擇參數的實驗結果攜帶的新訊息來確認或修正先前建立的模型，透過後續實驗的參數釐清與修正，快速找到關鍵因素之參數設定範圍。此舉不僅可減少實驗的次數，更可有效降低研發的成本與縮短研發時程，進而加速新產品開發上市的時間。

此構想源於研讀軟性顯示器研發階段的相關資料時，期望此參數範圍搜尋機制能實際應用於各產業的研發實驗單位如：藥品研發、電子產品研發、材料研發等，增進其競爭優勢。

1.2 研究目的

研發初期，關鍵製程的參數設定一般透過多次實驗試誤獲得，在研發經費與時間有限的限制下，縮短製程配方(recipe)尋找的時間，能有效的降低整

體研發的成本與時程。故本研究主要目的期望以系統化方法建立研發階段實驗參數範圍的縮小機制，以較少實驗次數下快速的找出關鍵製程參數組合之範圍。

本研究發展出三個修正方法，並以二個分類類別數量為三的資料集來驗證效果，運用蒐集之數據及所屬分類，在有限實驗次數下，以少量資料產生虛擬樣本，輔助做為下一製程實驗參數選擇的依據，並藉由系統化方法縮小參數範圍，提升參數設定範圍與分類正確率之關係，最終找出參數設定的邊界範圍。

在此將本研究所面臨到的問題，加以條列說明如下：

1. 當樣本數極少時，其他研究如何處理？（於第 5 頁）
2. 一般的參數範圍尋找的方法當中，對於少量樣本的處理方法為何？（於第 6 頁與第 7 頁）以及本研究欲建立的參數範圍搜尋機制是否適用？
3. 待參數範圍搜尋機制建立後，是否因使用此機制讓參數範圍變得明確而有助於原本的分類方法增加其分類的正確率？

1.3 研究方法與步驟

本研究將以下列四個階段進行，分別為文獻探討、機制建立、實例驗證及結論，詳細說明如下：

1. 文獻探討：

   首先探討小樣本問題及虛擬樣本之產生，以及使用區間化核密度估計(intevalized kernel density estimation, IKDE)之方法。接著探討支援向量機(support vector machine, SVM)的適用性。

2. 研究方法：

   支援向量選擇機制(support vector selection mechanism, SVS)：以少量資料開始，利用產生虛擬資料之方法，增加資料量以建構分類模型，並以支援向量機做為分類的工具，進而從中選擇對於提升分類正確率有較多訊息的實驗目標，在得到實驗結果後，利用此訊息修正其分類模型。

   區域虛擬點選擇機制(local virtual selection mechanism, LVS)：修正支援向

量選擇機制，以少量資料開始，但減少虛擬資料的產生，僅於關鍵資料產生虛擬資料，再由關鍵資料產生的虛擬資料中選擇一個對於提升分類正確率有較多訊息者為下一個實驗目標，得到實驗結果後，利用此訊息修正其分類模型；並與支援向量選擇機制比較虛擬資料之有無對於分類正確率是否有影響。全域虛擬點選擇機制(global virtual selection mechanism, GVS)：於區域虛擬點選擇機制中增加分散機制，給予此方法能得到全域解的可能，以此建構出一完整參數範圍邊界搜尋機制。

3. 實例驗證：

   經過實驗設計，來觀察研究方法中所建構之參數範圍邊界尋找機制（支援向量選擇機制、區域虛擬點選擇機制及全域虛擬點選擇機制），對於分類的正確率的提升是否有著相關性；以及全域虛擬點選擇機制趨近全域解所需的實驗次數。

4. 結論：

   經過前面幾個步驟，最後運用系統化的方式，討論其結果以及未來研究方向。

1.4 論文架構

本篇論文架構共分為五章，如圖 1.1。第一章為緒論，說明本研究期望建立尋找參數邊界的修正機制之研究背景與動機、研究目的及範圍，並概要的說明本研究方法。第二章為文獻探討，針對於小樣本分類與學習、虛擬樣本建立、支援向量機進行文獻探討回顧。第三章則為研究方法，說明本研究的研究方法及架構。第四章為實證研究，將以實例來驗證本論文設計出來的尋找參數邊界的修正機制對於分類的關聯性。第五章為結論與未來發展方向。

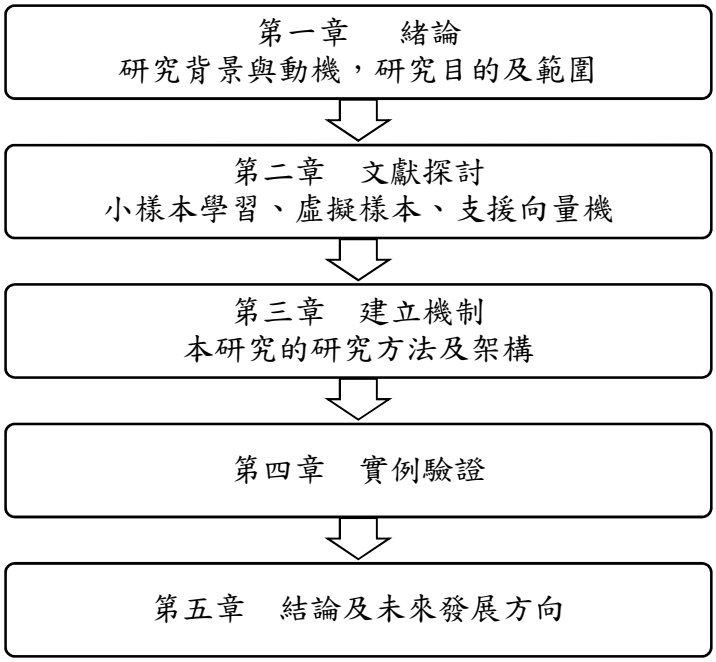
第一章　緒論
研究背景與動機，研究目的及範圍
第二章　文獻探討
小樣本學習、虛擬樣本、支援向量機
第三章　建立機制
本研究的研究方法及架構
第四章　實例驗證
第五章　結論及未來發展方向

圖 1.1 研究架構

第二章　文獻探討

本章節將以三個主題來探討與研究主題相關及使用到的文獻。首先討論小樣本分類相關文獻。接著探討產生虛擬樣本之方法。最後探討在高維度少量資料表現優異的分類方法－支援向量機。

2.1 小樣本分類

一般在尋找最佳參數時，多半利用田口方法(Taguchi method)做參數設計，應用直交表(orthogonal array)配置實驗，分析直交表實驗所得的資料研究出關鍵因子(key factors)與水準(levels)，找出最佳製程參數的可能方向，求出製程可能的最佳參數，本研究嘗試將部分實驗資源延遲，先配置部分實驗求得小樣本資料，以小樣本資料進行分類，找出參數可能的方向，以利後續實驗的投入。

線性區別分析(linear discriminant analysis, LDA) (Fukunaga, 1990)，使用平均數以及共變異數，被稱為是一種參數型的特徵萃取方法。在線性區別分析常見討論小樣本分類問題 (Huang, Liu, Lu, & Ma, 2002)，因傳統分類技術在統計樣式辨認中的假設大多都是基於有足夠的樣本可供使用，但是高維度資料(high dimensional data)所需要的訓練樣本點數比傳統資料多，更容易出現資料不足的情況，傳統的特徵萃取(feature extraction)或特徵選取(feature selection)技術被用於克服這個問題。

以統計為基礎之分類器在遇到高維度資料分類時，若訓練樣本點數少，則會出現 Hughes phenomena 之情況 (Hughes, 1968)。若資料的維度數提高而資料的數量無適當增加時，有可能會因為 Hughes phenomena 而造成辨識正確率下降。張光佑 (2006) 發現以無參數形式的分散矩陣(nonparametric scatter matrices)、RFE 正規化技術(regularized feature extraction regularization)、特徵值分解法(eigenvalue decomposition)，和 ML(maximum likelihood)分類器的組合，將有助提升特徵萃取在小樣本情況下的效能。

劉巧雯 (2010) 以擴充屬性資訊以提升小樣本分類之效果，提出一個處理資料屬性的方法來提升小樣本分類問題的正確率，方法包括建立屬性類別可能值，其藉由模糊裡論的基礎建立資料其各屬性於各類別的可能值，利用這

些各類別的可能值來增加小樣本資料的分析資訊以提升分類正確率；方法另一策略為屬性建構，目的在分析既有的屬性間關係，以發掘隱藏的有效屬性，且為避免因資料屬性過多而干擾模式學習的成果，其研究利用資料間各屬性的相關程度，將高度相關的屬性作合併以整合出指標屬性來提昇學習的正確度。

2.2 小樣本學習&虛擬樣本

小樣本問題泛指因資料量不足，導致分析績效不佳與錯誤推論的情況。當樣本數量很少時，不論使用何種學習機制，都不會有很好的成效，因為其缺乏樣本所屬的母體資訊，欲解決此類問題，收集更多的資料，顯然是最直接而有效的方式，但在研發初期階段，要獲得大量且充足的資料是相當困難的，除此之外，尚有許多其他的因素，如成本考量或實驗樣本的難易程度等，都會直接或間接影響最終得到的樣本數量。雖然資料探勘的技術在資訊的萃取上已經有相當廣泛的應用，但是當使用小樣本建模時，學習的正確率往往會受到樣本數不足的影響，因此在有限的樣本下，學習將會是一項困難的工作，在此觀點下，若能適當的增加學習樣本，將會是一個提升學習正確率的有效方法。

Niyogi, Girosi, and Poggio (1998) 利用先前知識(prior knowledge)藉由已取得的小樣本資訊建立虛擬樣本，進而增加有效的訓練資料(training set)；其針對目標識別的辨識率進行改善並應用在兩種識別的問題上；對於圖形識別的應用是在已取得的圖像上，以數學運算取得不同角度的圖像，利用這些計算所得到的虛擬樣本改善識別率；另一種應用是聲音辨識，將不同聲音特質的人，對相同母音的發聲特性結合起來產生一個新的人聲樣本，用於改善聲音的辨別率；其証實透過建立的虛擬樣本能提升學習後的正確率。

Li, Chen, and Lin (2003) 提出功能性虛擬母體(functional virtual population, FVP)的方法，將虛擬樣本的概念應用，在以類神經網路學習動態製造排程的知識上。實驗結果顯示，類神經網路學習正確率，可由原先使用舊排程知識時的 32.75%提升至 62.5%，有顯著的改善效果。但是，因為 FVP 是依照特定的製造標準而發展的，並無法直接適用於其它的狀況，不過此研究亦顯示出採用虛擬樣本確實是小樣本學習的一個可行方向。

2.2.1 區間化核密度估計

在樣本資料有限的情況下，我們常用無母數的密度估計方法來估計一個可能的機率密度函數。一般常見的密度估計方法包含了直方圖、簡單密度估計(naïve density estimator)、核密度估計(kernel density estimator)、級數法(series method)、懲罰式最大概似估計(the maximum penalized likelihood estimator)及類神經網路密度估計(the artificial neural network based methods)等多種不同的理論方法，詳見 Silverman (1986) 之文，其中又以核密度估計最常使用。

使用直方圖來觀察資料集的分佈趨勢以估計其樣本分配，是最古老、簡易的方法，假設一組資料有n個觀察值$\{x_k\}$,$1<k<n$，而且區間寬度(bin width)為$2h$，$f(x_k)$為在一群資料下的未知密度函數，如下圖 2.1 所示，每一個觀察值在區間(x_k-h, x_k+h)的機率為$2h$ 乘$f(x_k)$，可用以下數學式表示：

$$P(x_k-h<X<x_k+h)\approx 2hf(x_k). \qquad \text{(式 2.1)}$$

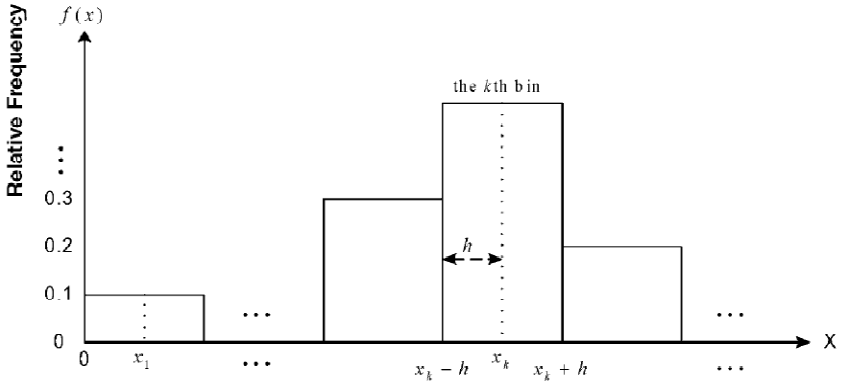

圖 2.1 直方圖估計法每個區間使用相同的 h 值

由式 2.1 推導，可用以下數學通式來表示一群資料的密度函數：

$$f(x)=\lim_{h\to 0}\frac{1}{2h}P(x-h<X<x+h). \qquad \text{(式 2.2)}$$

又觀察值落在區間(x_k-h, x_k+h)的比例可以用來估計$P(x-h<X<x+h)$，故式 2.2 可估計為

$$\hat{f}(x) = \frac{1}{2h} \times \frac{\text{觀察值} x_1,\ldots,x_n \text{落入區間} (x_k-h, x_k+h) \text{的數量}}{n}. \qquad \text{(式 2.3)}$$

式 2.3 即為簡單密度估計法的估計原理，常用一個指標函數 w(・)來簡化式 2.3，使其成為簡單密度估計的通式，即

$$\hat{f}(x) = \frac{1}{n} \times \sum_{i=1}^{n} \frac{1}{2h} w\left(\frac{x-X_i}{h}\right), \qquad \text{(式 2.4)}$$

where

$$w\left(\frac{x-X_i}{h}\right) = \begin{cases} 1, & |(x-X_i)/h| < 1, \\ 0, & \text{otherwise}. \end{cases}$$

在之後有學者提出以核函數$K(\cdot)$取代指標函數$w(\cdot)$的做法，稱為核密度估計(Parzen, 1962)。

$$\hat{f}(x) = \frac{1}{n} \times \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x-X_i}{h}\right). \qquad \text{(式 2.5)}$$

在核密度估計中，h 被稱為平滑參數(smoothing parameter)，針對同一個資料集，h 值越大則估計出來的分配則會越趨於平滑，反之則會越崎嶇，若選擇錯誤的 h 值，可能會有不佳的估計結果，如圖 2.2 所示。有別於簡單密度估計中使用的指標函數$w(\cdot)$之值只有 0 與 1 兩種，核函數$K(\cdot)$的可能值可以是任意的連續實數。採用不同的核函數會產出不一樣的估計分配，現有艾氏(Epanechnikov)核函數、雙權重(Biweight)核函數、三權重(Triweight)核函數、高斯(Gaussian)核函數、三角(Triangular)核函數和矩形(Rectangular)核函數等多種核函數可供選擇，其中高斯核函數最為廣泛的被應用。

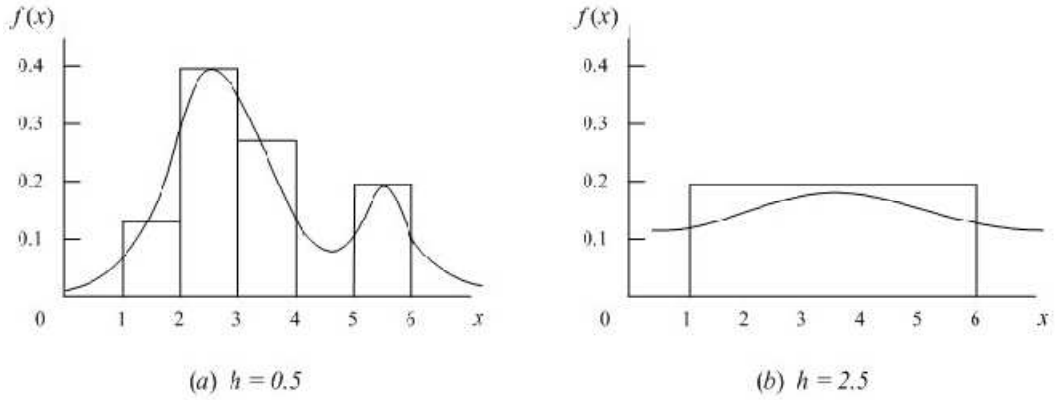

圖 2.2 使用不正確的 h 對估計分配的影響

區間化核密度估計是一種以核密度估計為基礎,且專為小樣本資料集(small size dataset)設計的方法 (Li and Lin, 2006),其主要的想法是認為在處理小樣本資料集時，應該以資料的分佈特性去劃分區間（使用多個 h 值），來估計個別區間所對應的樣本分配，而不是使用固定的區間（固定的 h 值）。

以圖 2.3 作說明，(a)為現有的小樣本資料集，以直方圖的形式呈現。核密度估計的做法是使用固定的 h 值，當 $h = 1$ 時估出來的樣本分配為多峰(multi-modal)分配；區間化核密度估計則根據此樣本資料的特性使用兩個不同的 h 值（h'=1 與 h"=2.5），其估出來的樣本分配則為雙峰(twin-modal)分配。以估計結果來看，於 (c)中，因區間化核密度估計採用不同長度的區間，以最後兩筆資料去估計一個分配，在統計上顯然比 (b)個別去估計一個分配的做法，還要來的合適。

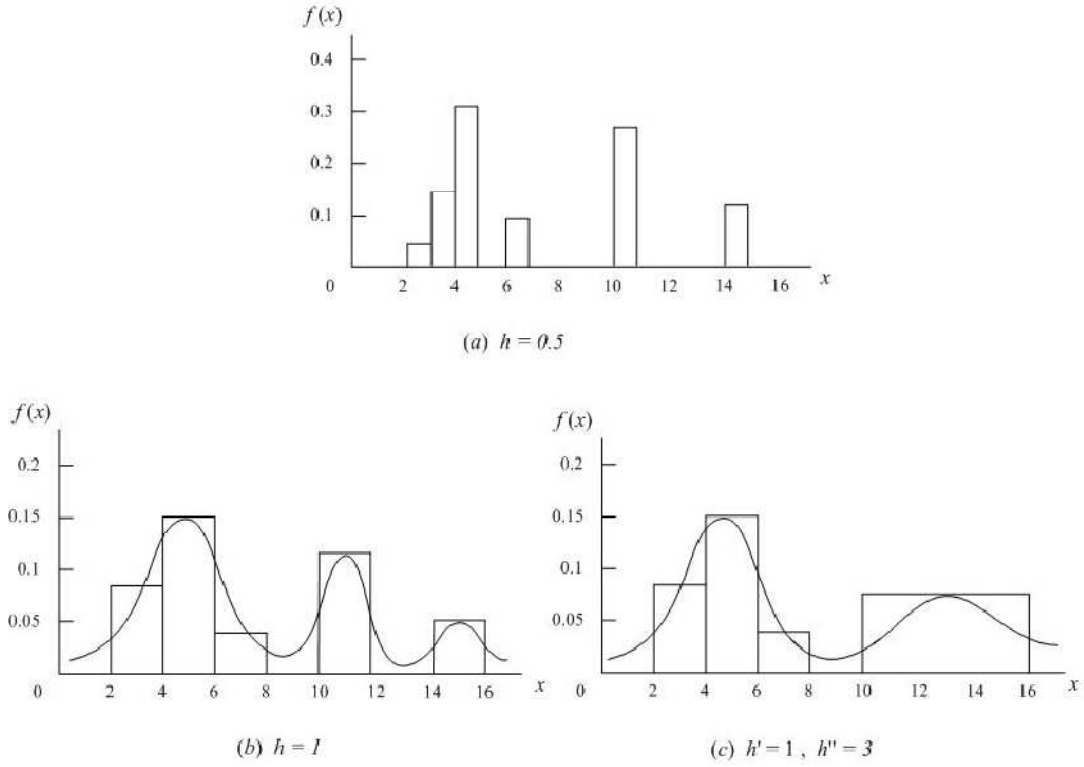

圖 2.3 核密度估計與區間化核密度估計估計結果之差異

綜合以上論述，區間化核密度估計的提出是為了增進核密度估計在小樣本學習中的效能表現。由式 2.5 加入多重區間概念的區間化核密度估計法其公式如下：

$$\hat{f}(x) = \sum_{i=1}^{m} \frac{1}{n_i}\frac{1}{h_i} K_i\left(\frac{x-c_i}{h_i}\right), \quad \text{(式 2.6)}$$

其中 $n = \sum_{i=1}^{m} n_i$.

而式 2.6 中c_i為核函數$K_i(\cdot)$的中心位置，n_i表示第 i 個區間的資料個數，h_i則代表第 i 個區間寬度的一半。其研究結果顯示，在樣本數受限的情況下，以區間化核估計產生虛擬樣本，使用類神經網路學習排程知識，減少原先因樣本數不足而產生的誤差。結果顯示，虛擬樣本確實可以增進學習正確率；但是當虛擬樣本數過多時，則會產生過度延伸的效應，反而使得學習正確率因虛擬樣本數的增加而減少。

2.3 支援向量機器

支援向量機又稱為支撐向量機，是由 Vapnik 及 AT&T Bell Labs 的研究小組，基於統計學習理論(statistical learning theory)所提出的一個機器學習理論，屬於監督式學習(supervised learning)的一種，廣泛的應用於統計分類以及迴歸分析中 (Vapnik, 1995, 1998)，其理論依據主要是來自於統計學習理論中的結構風險最小誤差法(structural risk minimization, SRM)，利用分隔超平面(separating hyperplane)的方法，找尋最大的邊界(margin)，進而將資料區分成兩類以上的類別。

SVM 對於解決小樣本、非線性以及高維模式的識別問題，有許多特有的優勢 (Burges, 1998)。目前 SVM 已應用於手寫體識別、三維目標識別、人臉識別、文本圖像分類等實際問題，性能優於已有的學習方法，表現出良好的學習能力，從有限訓練樣本得到的決策規則對獨立的測試集仍能夠得到較小的誤差。

2.3.1 支援向量機基本概念

支援向量機的基本概念如下，假設有 n 筆訓練樣本 x_i，i =1, … , n，$x_i \in R^d$（R^d 為高維度空間），對應的期望輸出為 y_i，$y_i \in \{1, -1\}$，其中 1 和-1 分別代表兩種類別的類別標籤。我們希望能夠在 R^d 中找出一個超平面，將這 n 筆訓練樣本分為兩類，讓屬於同一類的數據均在超平面的同側。如圖 2.4 所示，

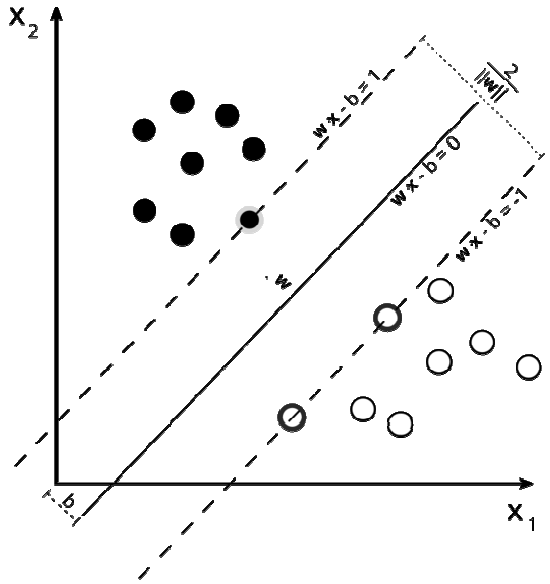

圖 2.4 SVM 說明圖一

最佳超平面的數學形式為$w \cdot x - b = 0$（在此 · 為向量內積），其中 w 為超平面之法向量(normal vector)、x 為輸入向量、b 為偏移量。為了讓此超平面有最大邊界，我們需要知道支援向量以及與最佳超平面平行並且離支援向量最近的超平面，此兩條離支持向量最近的超平面方程式為$w \cdot x - b = 1$ 與 $w \cdot x - b = -1$，$f(x) = w \cdot x - b$ 稱為決定函數(decision function)，若 $w \cdot x - b \geq 1$，則將該筆資料歸類為+1；$w \cdot x - b \leq -1$，則將該筆資料歸類為−1。而距離超平面最近、具決定性的資料點就是所謂的支援向量。

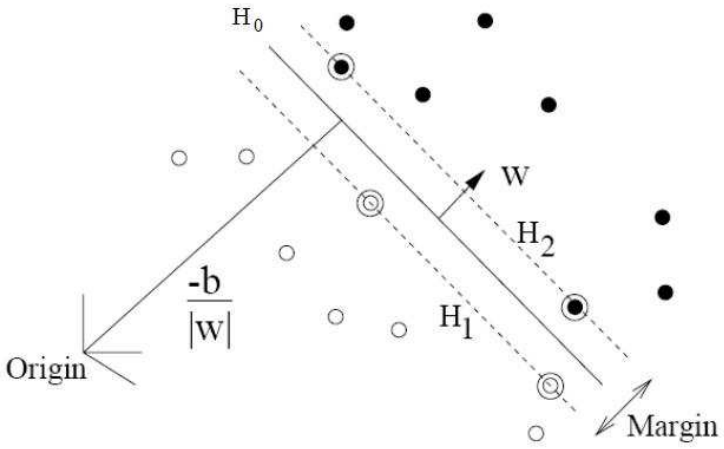

圖 2.5 SVM 說明圖二

支援向量機就是希望可以在不同類別的資料中，找出最大邊界的區分超

平面。我們可以從圖 2.5 中發現超平面H_0、H_1、H_2都可以達到區分類別的效果，而H_0是最好的，因為H_0與邊界 A 和邊界 B 的距離最大。

兩個超平面之間的距離是 2/|w|，因此我們需要最小化|w|，其中必須保證對於所有的 x_i 滿足以下其中的一個條件，$w \cdot x_i - b \geq 1$或$w \cdot x_i - b \leq -1$，合併上列兩式成為 $c_i\ (w \cdot x_i - b) \geq 1, 1 \leq i \leq n.$

將尋找最佳超平面問題簡化如下：

$$\text{minimize } \frac{||w||^2}{2}, \text{ subject to } c_i\ (w \cdot x_i - b) \geq 1, 1 \leq i \leq n.$$

2.3.2 特徵空間學習

在大部分的情況下，資料沒有辦法被線性的分類，因此要把資料映射(map)到特徵空間，見圖 2.6，意味著若資料無法在所在的維度下明確的分類，SVM 會將資料轉換至高維度之後再做分類，轉換公式如下：

$$\Phi : R^n \rightarrow R^m, m > n. \quad \text{(式 2.7)}$$

觀察對偶問題，發現在對偶問題中，資料的處理都是用到向量內積(inner product)，因此若要在特徵空間學習，只要能計算出資料在特徵空間中的內積值就可以了，並不需要直接把資料映射到特徵空間。

而將資料轉換至高維度之後，在式 2.8 時就必須耗費時間來做內積的運算，因此 SVM 便會定義核函數來簡化內積運算，以加快運算的速度，核函數定義如下：

$$K(x, y) = \Phi(x) \cdot \Phi(y) \quad \text{(式 2. 8)}$$

而 SVM 中所定義的核函數有下列幾種：

Simple dot : $K(x, y) = x \cdot y$ (式 2. 9)

Polynomial : $K(x, y) = (x \cdot y)^p$ (式 2. 10)

Radial basis function : $K(x, y) = \exp(\frac{-||x-y||^2}{2\delta^2})$ (式 2. 11)

Sigmoid kernel : $K(x, y) = \tanh(K(x \cdot y)\text{-}\Theta)$ (式 2. 12)

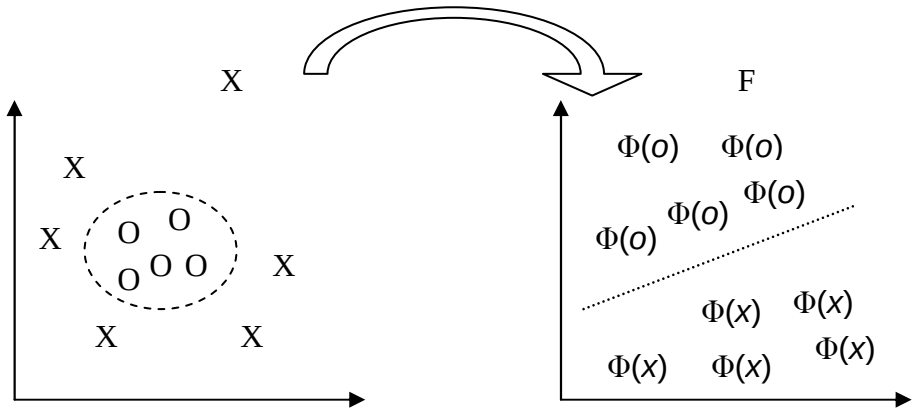

圖 2.5 空間轉換圖

2.3.3 LIBSVM

LIBSVM (Chang and Lin, 2001)全名為 A Library for Support Vector Machines，是一套完整的 SVM 軟體，可以解決分類問題(包括 C-SVC、ν-SVC)、迴歸問題（包括 ϵ-SVR、ν-SVR）以及分配估計等問題，本研究將以 LIBSVM 做為分類工具，做為分類結果的依據。

第三章　研究方法

在面對參數範圍邊界找尋的問題上，本章節將以三個主題來分別敘述所使用的研究方法與相關的假設條件。首先在 3.1 節定義資料的範圍以及假設條件。接著在 3.2 節說明所使用的研究工具。最後在 3.3 節說明尋找參數範圍的流程，以及流程中各步驟所使用的方法。

3.1 研究資料範圍

本研究探討如何在少量的分類資料中有效的尋找正確的參數範圍，規劃建立一個系統化的流程，能夠快速而有效率的尋找到不同分類的參數範圍。

假設一樣本空間 S，$\{S \mid S = (X_1, X_2)\}$，$X_1, X_2 \in \{1, 2 \ldots, 100\}$，$Y \in \{1, 2, 3\}$。研究所使用的資料集有 Real data set（以 RX 表示）、virtual data set（以 VX 表示）及 test data set（以 TX 表示）；RX 為由樣本空間中隨機抽出的少數幾組參數，而 VX 為 RX 產生的虛擬資料，TX 為相關的測試資料。

其相對於資料的假設條件如下所示：

1. 兩參數資料X_1, X_2獨立。
2. X_1, X_2兩個資料的資料型態為數值型態。
3. 資料結構為線性及凸集合。
4. 資料的收集都是來自一個靜態的系統。
5. 資料組視為隨機樣本。
6. 表示位數為小數點後四位，小數點後第五位採用四捨五入法。

3.2 研究工具

本研究原始資料為文字檔格式及 Microsoft Excel 的資料。在此以 Microsoft Excel 做為資料整理、計算及轉換格式的工具。

支援向量機的分類部分將以 LIBSVM 2.91 版 (Chang & Lin, 2001)做為分類工具，最後將其結果記錄之。

3.3 研究機制建立

本研究由小樣本問題為基礎，逐步建立參數範圍邊界找尋之機制，產生虛擬樣本之方式於 3.3.1 小節說明，3.3.2 小節說明建立分類模型的方法，3.3.3 說明實驗參數的選擇機制 SVS，3.3.4 說明修正後的實驗參數選擇的機制 LVS，3.3.5 則說明增加了分散機制的實驗參數選擇的機制 GVS。

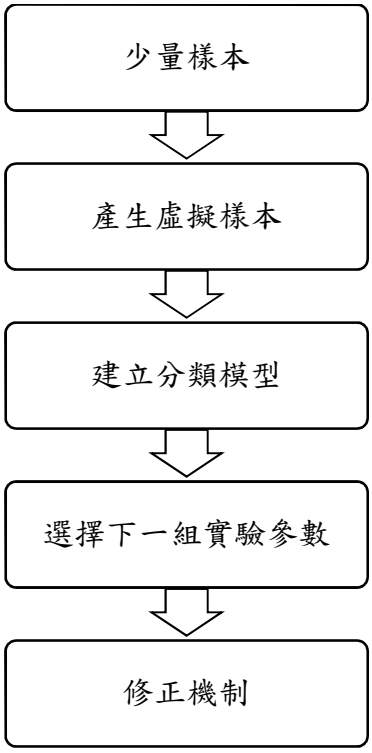

圖 3.1 研究架構

3.3.1 小樣本問題與虛擬樣本的建立

小樣本問題之所以較難以處理，原因在於由於樣本小以致難以利用統計學的方法推估母體的分配，故以往應用於大樣本的方法常在小樣本的條件下產生相當大的預測誤差。若在小樣本情況下要推估母體則需使用重複取樣法 (resampling method)，方法包括 Bootstrap、Jackknife 等。經由資料的重新抽樣 (re-sampling)藉以估計統計量的分配，但過程十分繁瑣複雜，因而使用只假設密度函數為連續且型式沒有任何的限制的無母數方法。在眾多的無母數密度估計方法中，核密度估計是密度估計的常用工具，它是利用適當的核心函數去估計母體的密度函數，雖然當資料數很少時，核密度估計可能會產生可信

度低的估計器，但區間化核密度估計擁有區間函數可以依照樣本數去調整平滑參數及資料的位置，克服了資料數很少的問題，因此本研究使用區間化核密度估計器，從原始的小樣本產生更多的資料。

3.3.2 資料分類

在擁有部分資料後，接踵而來的問題是該如何選擇下一組實驗參數能夠達到最大的實驗效益，故需要建立一個模型來幫助我們從資料中了解目前資料的分類邊界，目前在分類上所發展出來的分類器有支援向量機、類神經網路、k 最近鄰(k- nearest neighbor, kNN)、線性最小平方(linear least-squares Fit, LLSF)、貝氏分類(Bayesian classifier)等，皆有其適用範圍。而支援向量機屬於監督式學習方法的一種，能經由給予一群已經分類好的資料，可以經由訓練獲得一組模型。爾後，有尚未分類的資料時，支援向量機可以依據利用先前建構的模型去預測這筆資料所屬的分類，雖然類神經網路也有相同功能，但支援向量機在經由訓練後會得到支援向量，本研究將支援向量視為提供分類邊界訊息的一組參數，因而選擇支援向量機做為本研究建立模型的工具。

本研究使用 LIBSVM 作為分類工具，其功能相當完整且可支援多維分類器的建立。下面將對 LIBSVM 的主程式部分及本研究所使用到的資料加以說明 (Hsu, Chang, and Lin, 2004)。

1. 資料型態

   LIBSVM 的檔案格式必須為如下
   [label] [Index1]:[value1] [Index2]:[value2] ...
   [label] [Index1]:[value1] [Index2]:[value2] ...

   格式說明：

   label 或稱之 class，分類的種類。

   Index 為有順序的索引，通常是連續的整數。

   value 就是用來訓練的資料內容值。

   舉例來說，1 1:0 2:3 4:3。這表示此筆訓練資料分類為 1，第一個特徵值值為 0，第二個特徵值值為 3，第三個特徵值值為 0，第四個特徵值值為 3。

2. 主程式說明

(1) svm-scale

svm-scale 是用來調整資料內容值的大小範圍，以免有某一項資料內容值太大，在算距離時主導結果。通常將取值的範圍訂在 0~1 或 -1~1， training data 與 test data 都必須作相同程度的 scaling。

(2) svm-train

svm-train 在訓練過程中會接受特定格式的輸入，產生一個 "Model" 檔。這個 model 可以想像成 SVM 的內部資料，因為預測(predict)要有 model 才能 predict，不能直接讀取原始資料。在 model 檔內會提供以下訊息：svm_type、kernel_type、gamma、nr_class （分類種類數量）、total_sv（所有支援向量的數量）、rho（decision function(s) wx + b 的 b 值）、label（分類的種類）、nr_sv（分類種類其對應的支援向量數量）以及 SV（支援向量）。

(3) svm-predict

依照已經訓練完成的 model，再加上給定的輸入，輸出預測(predict)新值所對應的類別(class)於一 predict 檔，若測試資料內含分類結果，則程式會將分類結果與預測結果比較，計算 model 分類的正確率。

(4) grid.py

利用二個參數：cost、gamma，來對於使用 C-SVC (RBF kernel function)尋找最佳參數。cross validation 方式預設為 5。

(5) easy.py

將資料格式處理成 libsvm 的檔案格式後，執行 easy.py，便能自動執行 svm-scale、grid.py（尋找最佳參數）、svm-train、svm-predict，最後會產生預測的結果檔案及分類的正確率(accuracy)。

3. 本研究使用 LIBSVM 說明

透過 Microsoft Excel 將 Microsoft Excel 中的資料轉換成 LIBSVM 能使用的格式，做為訓練資料，訓練 SVM 分類器建立 model，libsvm 訓練時主要參數設定皆為預設值（svm type : C-SVC 及 kernel type : radial basis function），完成訓練後可在 model 檔中取得支援向量，另有一組 test data set，在經過預測後可以得到此 test data 分類的正確率。

3.3.3 Support vector selection mechanism

於樣本空間中隨機抽取 n 個樣本為本研究之 RX，以 RX 為產生虛擬樣本的基礎，本研究以區間化核密度估計之方法產生虛擬樣本，本研究產生虛擬樣本之平滑參數 h 以 RX 之變異數(σ_x)取代，SVS 的修正機制中，虛擬樣本之產生以 RX 之$\sqrt{\sigma_{x|Y}}$做為平滑參數 h（原因於第 20 頁-修正機制內說明）。將 RX 及產生的 VX 以 SVM 建立分類模型，接續下一組參數選擇步驟。

1. 下一組參數選擇

   在眾多候選參數中，SVS 以最近鄰分類法(nearest neighbor rule, NNR)及 K-means 分群法(k-means clustering)的概念選出下一個實驗參數。

   最近鄰分類法所根據的基礎是「物以類聚」，也就是同一類別的物件其歐基里德距離(Euclidean distance)較短。若以高度空間中的點來表示，則這些點的距離應該會比較接近。因此，那麼對於一個未知類別的一筆資料，只要找出在訓練資料中和此筆資料最接近的點，就可以判定此筆資料的類別應該和最接近的點的類別較易一致。所以本研究以此概念將經過訓練後的 SVM 模型列出的支援向量，VX 中不同分類的支援向量相互計算其距離，取其不同分類結果最短的兩個支援向量，提供做為下一個實驗參數的參考依據。

   以 6 個 VX 舉例說明如下：表 3.1 實驗參數選擇 A 步驟一，Y=1 與 Y=2 之間的最短距離即為 VX 1、VX 2 與 VX 3、VX 4 各計算之間的距離取最小值；Y=2 與 Y=3 之間的最短距離即為 VX 3、VX 4 與 VX 5、VX 6 各計算之間的距離取最小值；Y=1 與 Y=3 之間的最近距離即為 VX 1、VX 2 與 VX 5、VX 6 各計算之間的距離取最小值，結果如表 3.2，因 11.71 為所有分類間最近距離，11.71 為 VX 2 與 VX 3 之距離，故取 VX 2 與 VX 3 為可能的下一個實驗參數。因實驗參數於分類 1 與分類 2 之間做選擇，故下一次的實驗參數間的選擇僅為分類 1 與分類 3 之間或分類 2 與分類 3 間，擇分類間最短距離的支援向量做為可能的下一個實驗參數。

表 3.1 SVS 實驗參數選擇步驟一

VX	Y	(X1 , X2)	VX 1	VX 2	VX 3	VX 4	VX 5	VX 6
VX 1	1	(35 , 50)			39.9	35.4	**19.4**	21.8
VX 2		(62 , 71)			**11.7**	17.9	46.1	47.0
VX 3	2	(72 , 65)	39.9	**11.7**			47.4	47.5
VX 4		(70 , 55)	35.4	17.9			39.2	**38.9**
VX 5	3	(39 , 31)	**19.4**	46.1	47.4	39.2		
VX 6		(41 , 29)	21.8	47.0	47.5	**38.9**		

：代表分類一與分類二的最近距離

：代表分類二與分類三的最近距離

：代表分類一與分類三的最近距離

表 3.2 實驗參數選擇步驟一結果

	VX	Y	X1	X2	不同分類間的最短距離
Y=1 與 Y=2	**VX 2**	1	62	71	11.7**
	VX 3	2	72	65	
Y=2 與 Y=3	**VX 4**	2	70	55	38.9
	VX 6	3	41	29	
Y=1 與 Y=3	**VX 1**	1	30	50	19.4
	VX 5	3	39	31	

**為所有不同分類間最短距離

而 K-means 分群法其主要目標是要在大量高維的資料點中找出具有代表性的資料點，目的是希望盡量減小每個群聚中，每一點與群中心的距離平方誤差(square error)，以此概念將上一段以最近鄰分類法找出的兩個支援向量，個別計算其群聚中心，視其離群聚中心較遠之支援向量為對其分類結果不確定性較高的參數，故選擇離群聚中心最遠的支援向量做為下一次的實驗參數。

舉例說明如下：VX 2 與 VX 3 為上一階段選擇出的支援向量，此階段要

計算支援向量與其群聚中心的距離，群聚中心的計算方法以 RX 為計算基準（如表 3.3 群聚中心計算範例），在計算出群聚中心後，計算該分類與其群聚中心之距離，取距離最大者為下一個實驗參數（如

表 3.4 SVS 實驗參數選擇步驟二，下一個實驗參數為 VX 2 (62 , 71)）。

表 3.3 群聚中心計算範例

Real data set	Y	X1	X2	群聚中心
RX 1	1	41	57	（ 49.33 , 73.67 ）
RX 2	1	50	95	
RX 3	1	57	69	
RX 4	2	78	63	（ 81.00 , 66.77 ）
RX 5	2	80	82	
RX 6	2	85	55	
RX 7	3	68	6	（ 38.25 , 10.75 ）
RX 8	3	41	4	
RX 9	3	11	13	
RX 10	3	33	20	

表 3.4 SVS 實驗參數選擇步驟二

	Y	X 1	X 2	與群聚中心之距離	下一個實驗參數選擇
VX 2	1	62	71	12.98	v
VX 3	2	72	65	9.16	

2. 修正機制

當選出實驗參數後，經實驗得到其分類結果其分類結果有以下二種可能：與原分類相同或與原分類相異。

當實驗結果與原分類相同時，依此組參數與結果產生其虛擬樣本，將此次實驗參數與產生的虛擬樣本加入訓練資料，重新讓分類器學習，修正原先產生的分類模型，如圖 3.2 修正機制示意圖一，說明如下：（a）箭頭所指處為下一組實驗參數，（b）此組參數實驗結果為 X，（c）依此實驗參數與結果產生之虛擬樣本。

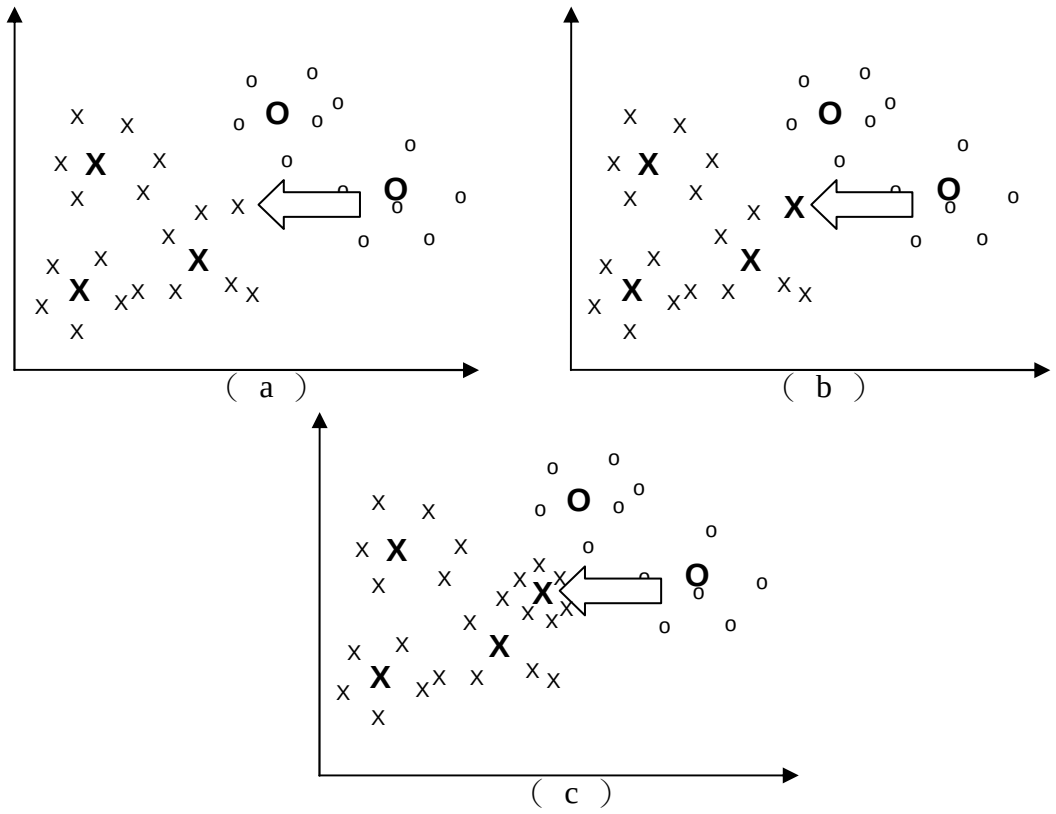

圖 3.2 修正機制示意圖一

當實驗結果與原分類相異時，除依此組參數與結果產生其虛擬樣本外，尚需找出此組實驗參數背後的真實實驗資料，重新產生此筆實驗資料的虛擬樣本，縮小虛擬樣本產生的範圍，最後將此組參數與結果產生其虛擬樣本加入訓練資料，並將重新產生的前實驗樣本虛擬資料於訓練資料中取代其上一次產生的虛擬樣本。如圖 3.3，說明如下：(a) 箭頭所指處為下一組實驗參數，(b) 此組參數實驗結果為 O，(c) 依此實驗參數與結果產生之虛擬樣本，找出此組實驗參數背後的真實實驗資料，重新產生此筆實驗資料的虛擬樣本（在此以 X 表示），將此組參數與結果產生其虛擬樣本加入訓練資料。

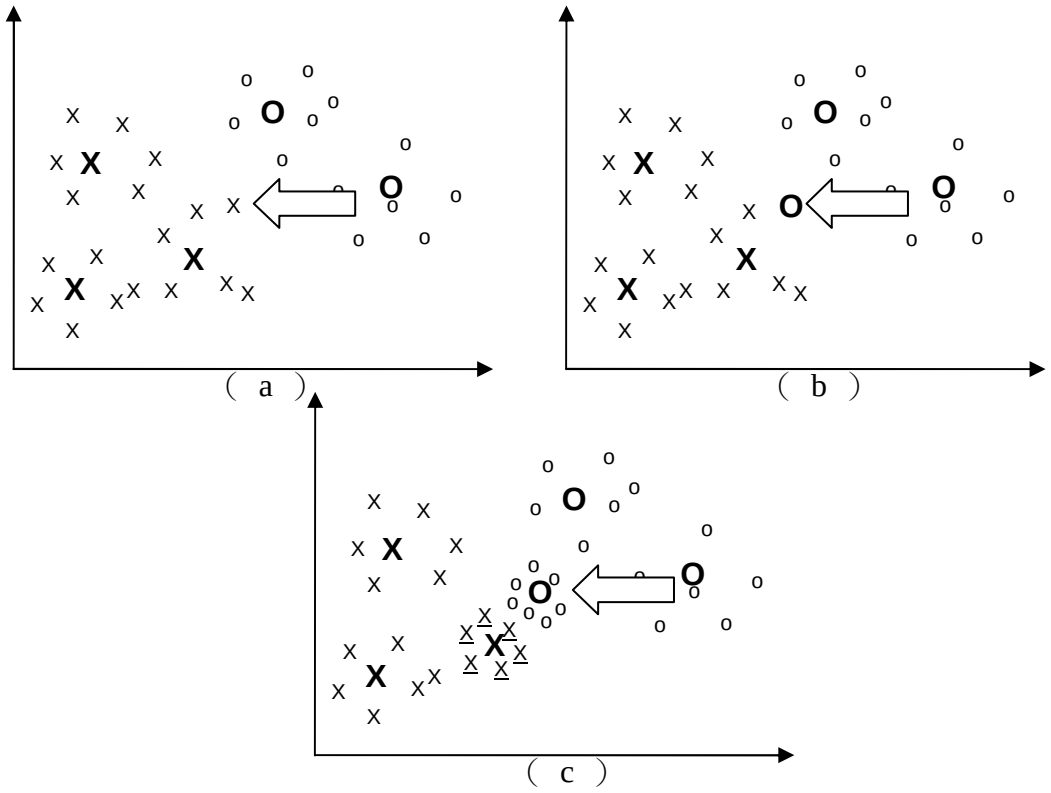

圖 3.3 修正機制示意圖二

SVS 之步驟方法如圖 3.4 SVS 參數範圍邊界找尋流程。

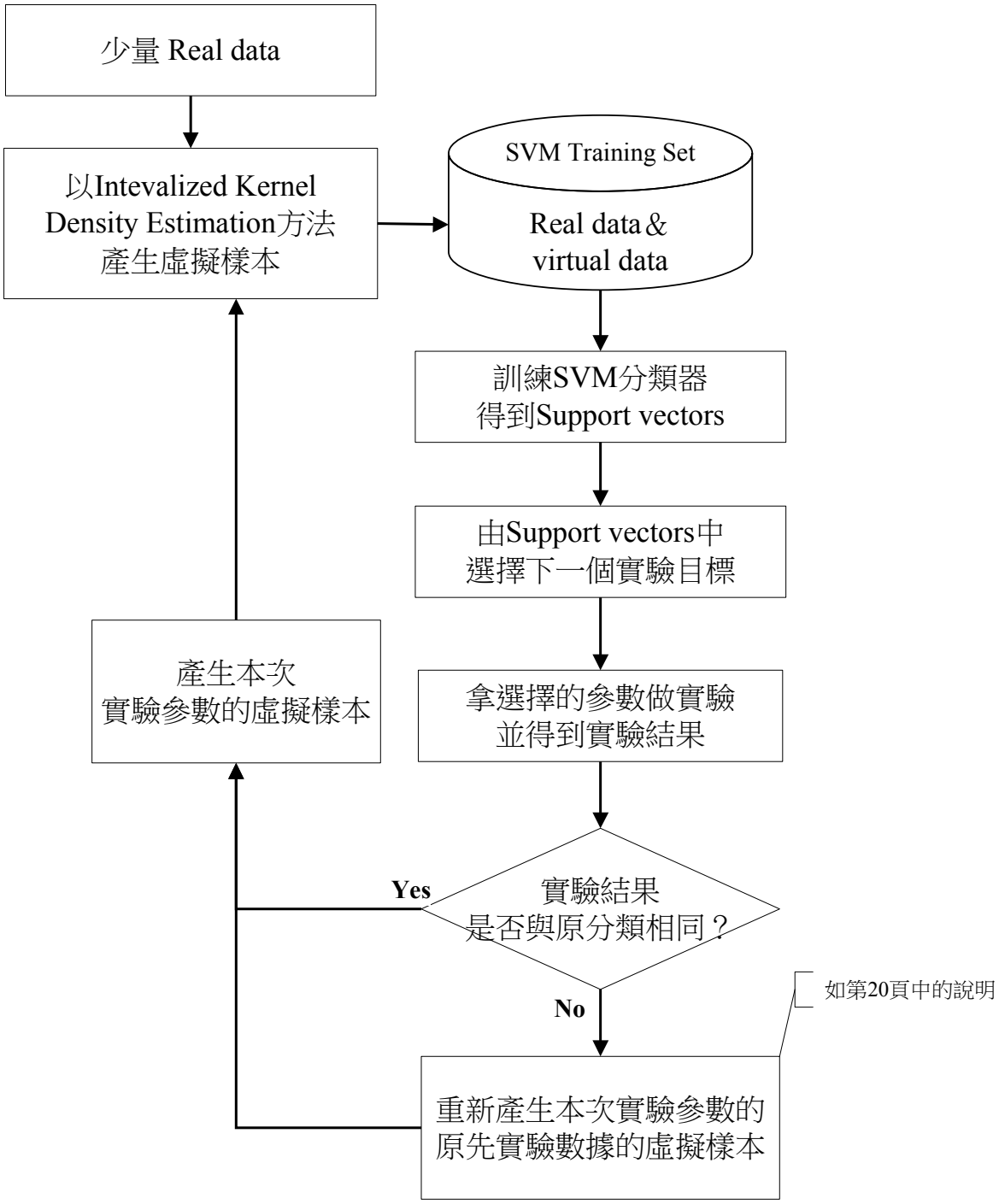

圖 3.4 SVS 參數範圍邊界找尋流程

3.3.4 Local virtual selection mechanism

在 SVS 中，每個 RX 皆先產生虛擬資料，經分類後才由虛擬資料中選擇下一個實驗點，此時有許多非位於分類交界處之 RX 所產生的虛擬資料對於分類器之分類可能無太多助益，在 LVS 中將修改此部分，減少產生對於分類正確率提升無幫助的虛擬資料。

LVS 步驟說明如下：由 RX 中，先選擇不同分類的兩個最近的 RX，由這兩個 RX 產生虛擬資料(VX)，經計算不同分類的 VX 與 RX 之距離後，取距離最短的為下一次的實驗參數。表 3.5 為 LVS 計算範例，表 3.6 為不同分類的 RX 之間最近距離的計算，結果列於表 3.7，分類 Y=1 與分類 Y=2 之最靠近之 RX 為 RX 3 及 RX 6，距離為 27.1 單位，分類 Y=2 與分類 Y=3 最近之 RX 為 RX 5 及 RX 8，距離為 19.3 單位，分類 Y=1 與分類 Y=3 最近之 RX 為 RX 4 及 RX 10，距離為 19.1 單位，故先選擇分類 Y=1 與分類 Y=3 的 RX 4 及 RX 10 產生 VX。

產生 VX 後，計算 VX 對其不同分類的 RX 之距離，擇其距離最近之 VX 為下一個實驗參數，如表 3.8，VX 4 分類為 1，但其距離分類為 3 的 RX 10 僅有 7 單位，故選擇 VX 4 (54 , 37) 為下一個實驗參數。修正後的研究 LVS 之步驟方法如圖 3.5 LVS 參數範圍邊界找尋流程。

表 3.5 LVS 之 RX 計算範例

RX	Y	X1	X2
RX 1	1	35	89
RX 2	1	31	82
RX 3	1	37	63
RX 4	1	42	42
RX 5	2	96	29
RX 6	2	64	65
RX 7	2	71	97
RX 8	3	78	22
RX 9	3	12	4
RX 10	3	61	40

表 3.6 LVS，RX 最近距離計算

		RX 1	RX 2	RX 3	RX 4	RX 5	RX 6	RX 7	RX 8	RX 9	RX 10
Y=1	RX 1					85.6	37.6	36.9	79.6	88.1	55.5
	RX 2					83.9	37.1	42.7	76.2	80.3	51.6
	RX 3					68.1	27.1	48.1	58.0	64.1	33.2
	RX 4					55.5	31.8	62.2	41.2	48.4	19.1
Y=2	RX 5	85.6	83.9	68.1	55.5				19.3	87.6	36.7
	RX 6	37.6	37.1	27.1	31.8				45.2	80.2	25.2
	RX 7	36.9	42.7	48.1	62.2				75.3	110.1	57.9
Y=3	RX 8	79.6	76.2	58.0	41.2	19.3	45.2	75.3			
	RX 9	88.1	80.3	64.1	48.4	87.6	80.2	110.1			
	RX 10	55.5	51.6	33.2	19.1	36.7	25.2	57.9			

：代表分類一與分類二的最近距離

：代表分類二與分類三的最近距離

：代表分類一與分類三的最近距離

表 3.7 LVS，RX 最近距離計算結果

	RX	**Y**	**X1**	**X2**	**不同分類間的最短距離**
Y=1 與 Y=2	**RX 3**	1	37	63	27.1
	RX 6	2	64	65	
Y=2 與 Y=3	**RX 5**	2	96	29	19.3
	RX 8	3	78	22	
Y=1 與 Y=3	**RX 4**	1	42	42	19.1**
	RX 10	3	61	40	

**為所有不同分類間最短距離

表 3.8 LVS 實驗參數選擇

	No.	Y	X1	X2	與 RX 10 之距離
RX 4 產生的 VX	VX 1	1	41.9104	42.8208	19.3
	VX 2	1	32.369	44.4479	29.0
	VX 3	1	47.3139	52.3355	18.4
	VX 4	1	54.4995	37.46	7.0 *
	VX 5	1	47.5596	45.5137	14.5
	VX 6	1	49.6849	37.3414	11.6
	VX 7	1	43.1003	47.6029	19.4
	VX 8	1	35.3583	48.6836	27.1
	VX 9	1	47.3634	44.6653	14.4
	No.	**Y**	**X1**	**X2**	**與 RX 4 之距離**
RX 10 產生的 VX	VX 10	3	64.1017	43.7924	22.2
	VX 11	3	59.9889	48.3237	19.1
	VX 12	3	66.4660	43.8345	24.5
	VX 13	3	66.5009	36.9122	25.0
	VX 14	3	64.3730	43.7552	22.4
	VX 15	3	67.0208	37.3720	25.4
	VX 16	3	63.4862	34.5417	22.7
	VX 17	3	58.5445	43.0106	16.6
	VX 18	3	58.0399	47.7273	17.0

*為 LVS 中最後選擇之 VX 與不同分類間的 RX 最近距離

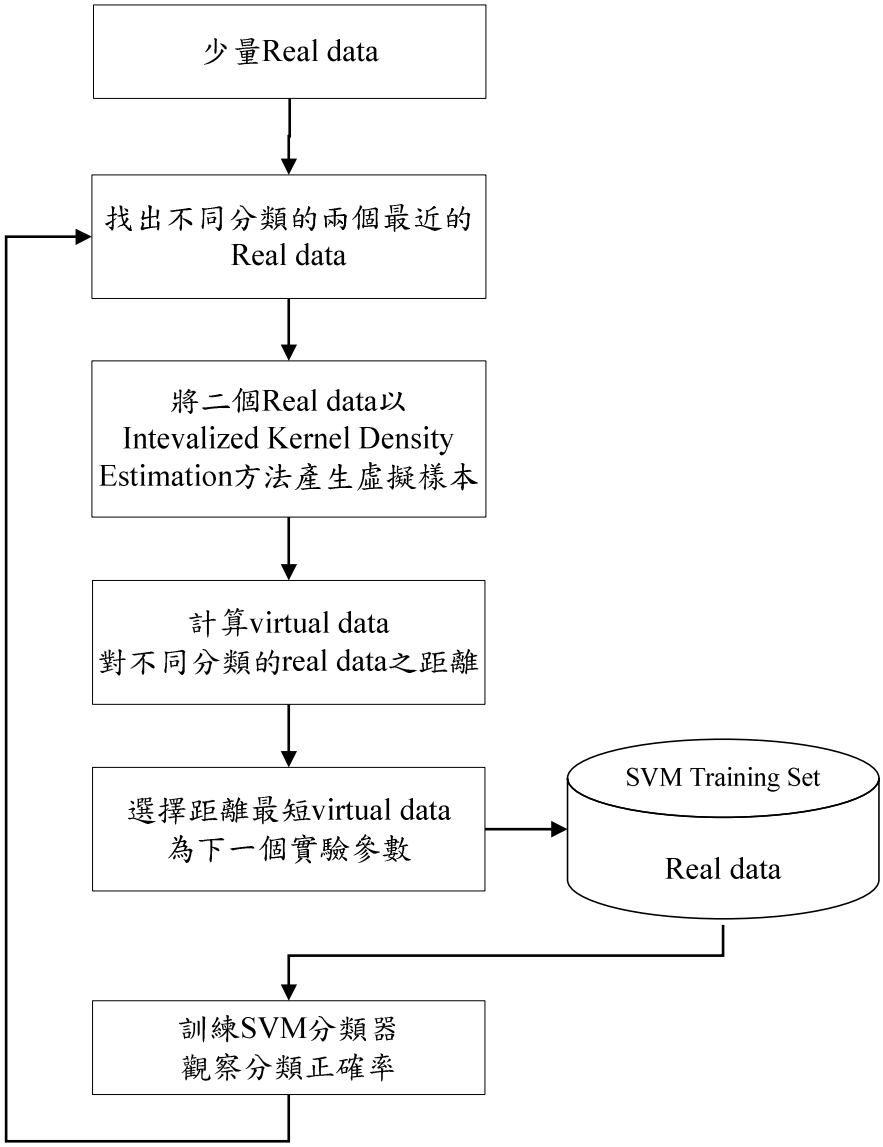

圖 3.5 LVS 參數範圍邊界找尋流程

3.3.5 Global virtual selection mechanism

在 LVS 中，當 RX 數量越來越多時，若有兩個極為相近卻不同分類的 RX 時，LVS 的實驗點選擇方法僅能選擇 VX 對其不同分類的 RX 距離最近之參數，易停滯在該區，導致最終的分類正確率受到限制，增進幅度有限，故 GVS 增加一分散機制，讓下一個實驗參數的選擇有機會跳脫該區域，解決了 LVS 中陷入區域最佳化而無法跳脫的問題。

GVS 步驟說明如下：由 RX 中，先選擇不同分類的兩個最近的 RX，由這兩個 RX 產生 VX，與表 3.6、表 3.7 相同。

產生 VX 後，計算 VX 對其不同分類的 RX 之距離（與表 3.8 相同），各

分類的 virtual data 各取距離最近的前 5 個（如表 3.9），經適應函數的評估產生適應值(Fitness)（適應函數如表 3.10，以輪盤法(Roulette wheel selection)選擇下一個實驗參數，如表 3.11。加入分散機制之研究 GVS 的步驟方法如圖 3.6 GVS 參數範圍邊界找尋流程。

表 3.9 GVS 實驗參數選擇步驟一

	No.	Y	X1	X2	與 RX 10 距離	Rank
RX 4 產生的 VX	VX 1	1	41.9104	42.8208	19.3	6
	VX 2	1	32.369	44.4479	29.0	9
	VX 3	1	47.3139	52.3355	18.4	5*
	VX 4	1	54.4995	37.46	7.0	1*
	VX 5	1	47.5596	45.5137	14.5	4*
	VX 6	1	49.6849	37.3414	11.6	2*
	VX 7	1	43.1003	47.6029	19.4	7
	VX 8	1	35.3583	48.6836	27.1	8
	VX 9	1	47.3634	44.6653	14.4	3*
	No.	**Y**	**X1**	**X2**	**與 RX 4 距離**	**Rank**
RX 10 產生的 VX	VX 10	3	64.1017	43.7924	22.2	4*
	VX11	3	59.9889	48.3237	19.1	3*
	VX 12	3	66.466	43.8345	24.5	7
	VX 13	3	66.5009	36.9122	25.0	8
	VX 14	3	64.373	43.7552	22.4	5*
	VX 15	3	67.0208	37.372	25.4	9
	VX 16	3	63.4862	34.5417	22.7	6
	VX 17	3	58.5445	43.0106	16.6	1*
	VX 18	3	58.0399	47.7273	17.0	2*

***為 VX 對其不同分類的 RX 之距離各選出距離最近的前 5 個**

表 3.10 適應函數

距離 d	Fitness value
d < 1	0.1
1 < d ≤ 5	2
6 < d ≤ 10	4
11 < d ≤ 15	8
16 < d ≤ 20	16
21 < d ≤ 25	8
26 < d ≤ 30	4
31 < d ≤ 35	2
36 < d ≤ 40	1
d > 40	0.5

表 3.11 GVS Roulette wheel selection

	Y	X1	X2	與不同分類之 RX 的距離	Fitness value	機率	累積機率
VX 4	1	54.4995	37.46	7	4	0.037	0.037
VX 6	1	49.6849	37.3414	12	8	0.074	0.111
VX 9	1	47.3634	44.6653	14	8	0.074	0.185
VX 5	1	47.5596	45.5137	15	8	0.074	0.259
VX 3	1	47.3139	52.3355	18	16	0.148	0.407
VX 17	3	58.5445	43.0106	17	16	0.148	0.556
VX 19	3	58.0399	47.7273	17	16	0.148	0.704
VX 11	3	59.9889	48.3237	19	16	0.148	0.852
VX 10	3	64.1017	43.7924	22	8	0.074	0.926
VX 14	3	64.373	43.7552	22	8	0.074	1.000

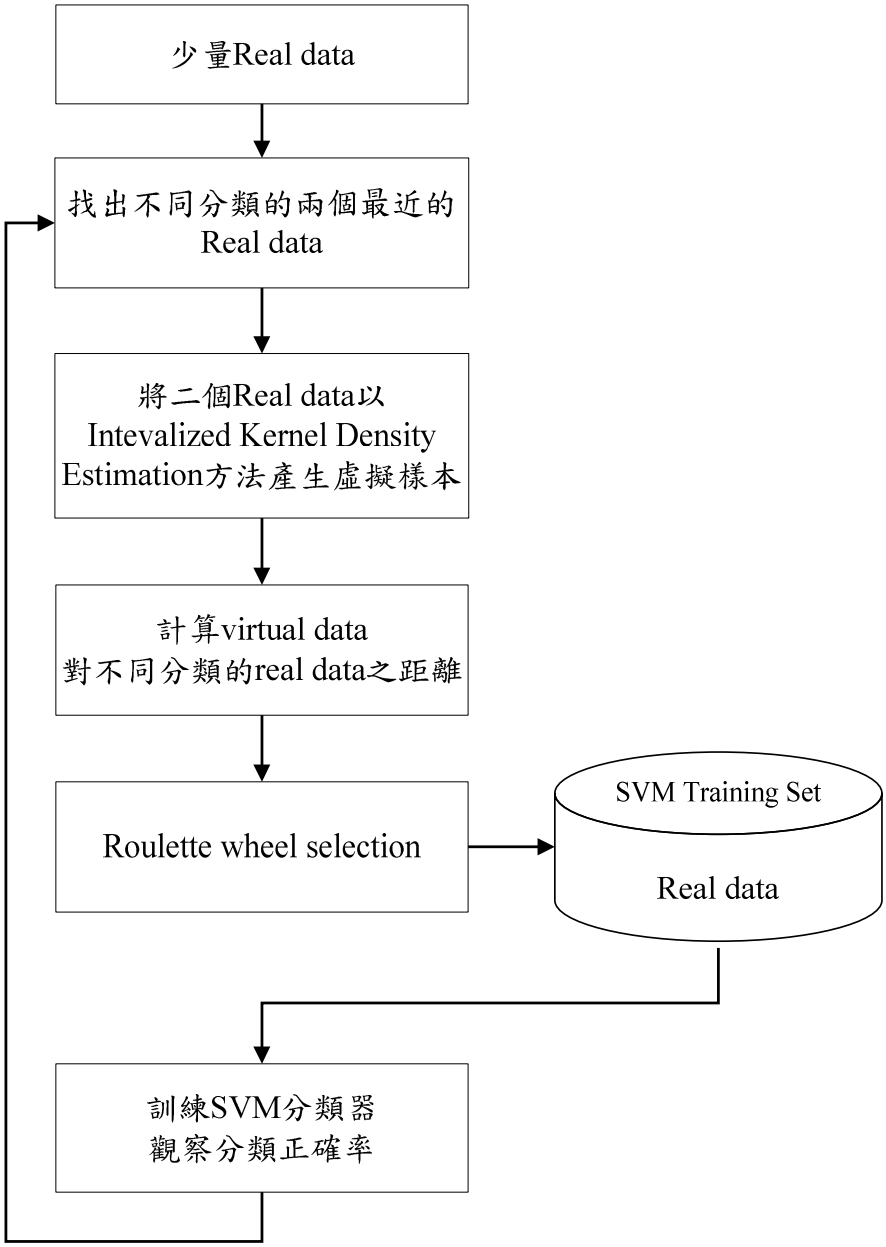

圖 3.6 GVS 參數範圍邊界找尋流程

第四章　實例驗證

本章將以兩個完整的完整資料集來進行研究方法的試驗步驟，用來測試本研究方法之合理性。4.1 節利用一個已知分類的資料集一，從中隨機抽取少量樣本做為 RX，比較本研究的方法在未知分類情形時，在小樣本的情況下是否可以少量的分類資料尋找參數分類邊界。4.2 節利用另一個較複雜的已知分類資料集二，從中隨機抽取少量樣本做為 RX，比較本研究的方法在未知分類情形時，在小樣本的情況下是否依舊可以少量的分類資料尋找參數分類邊界。

4.1 資料集一

使用本研究建立之資料集一共 10000 筆資料，做為小樣本資料集的為原始資料，下表 4.1 為資料集一的基本項目，其分類情形如圖 4.1 資料集一分類情況，共有三個分類。計算分類正確率的測試資料集為整體 10000 筆資料。從中隨機抽取 10 筆資料為本研究小樣本資料集，每分類至少有一筆資料。

表 4.1 資料集一基本資料表

資料集名稱	資料集一
資料筆數	10000 筆資料
類別數	三類，Y=1、Y=2、Y=3
維度	二維，X_1、X_2

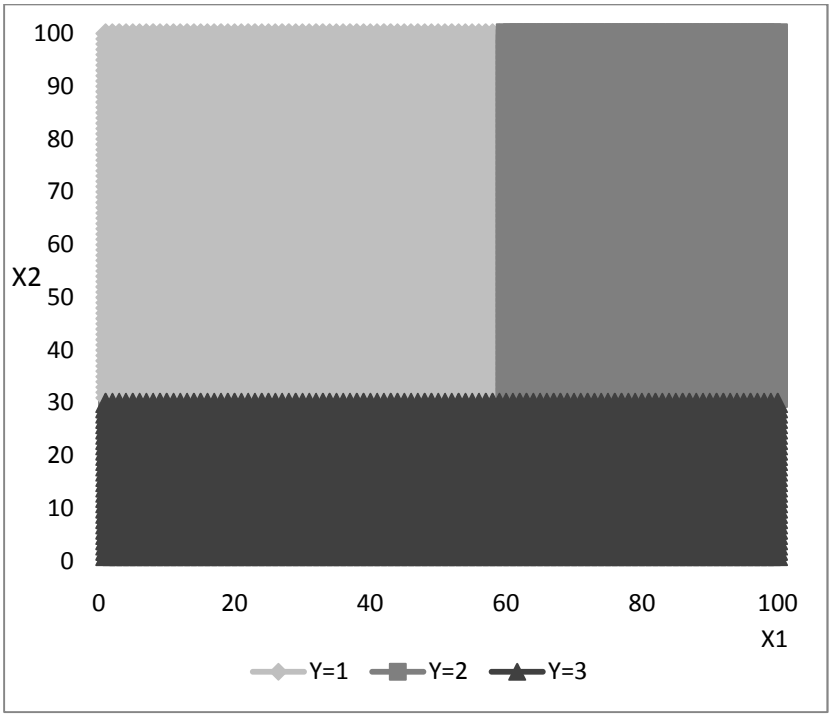

圖 4.1 資料集一分類情況

4.1.1 試驗說明

對於資料集一，做了以下幾種不同的實驗，測試經過挑選方法後加入的 Real data 對於分類的正確率是否能提升。

1. Original method（Method O）：先隨機抽取 10 筆資料為小樣本資料集，再隨機選擇 3 筆資料進行分類。

   先隨機抽取 10 筆資料為視為 Real data，再隨機選擇 3 筆資料，使用 SVM 分類，試驗次數 40 次。

2. 使用與 Method O 抽取的 10 筆相同資料，以 SVS 再選出 3 筆資料，進行分類。

   欲得知 SVS 機制是否有效，使用與 Method O 相同的隨機抽取的 10 筆資料視為 Real data，以 SVS 的步驟，陸續選出 3 個對於分類正確率有幫助的 Virtual data 並於資料集一中查詢其分類屬性，查詢後將此筆完整資料視為新的 Real data，使用 SVM 測試分類正確率，試驗次數 40 次。

3. 使用與 Method O 抽取的 10 筆相同資料，以 LVS 再選出 3 筆資料，進行分類。

欲得知 LVS 機制是否有效，使用與 Method O 相同的隨機抽取的 10 筆資料視為 Real data，以 LVS 的步驟，陸續選出 3 個對於分類正確率有幫助的 Virtual data 並於資料集一中查詢其分類屬性，查詢後將此筆完整資料視為新的 Real data，使用 SVM 測試分類正確率，試驗次數 40 次。

4. 使用與 Method O 抽取的 10 筆相同資料，以 GVS 再選出 3 筆資料，進行分類。

   欲得知 GVS 機制是否有效，使用與 Method O 相同的隨機抽取的 10 筆資料視為 Real data，以 GVS 的步驟，陸續選出 3 個對於分類正確率有幫助的 Virtual data 並於資料集一中查詢其分類屬性，查詢後將此筆完整資料視為新的 Real data，使用 SVM 測試分類正確率，試驗次數 40 次。

5. 隨機抽取 10 筆資料，以 LVS 選出下一次實驗資料進行分類，直到分類正確率達到 90%即停止。

   為了觀察 LVS 最終是否能夠收斂，先隨機抽取 10 筆資料視為 Real data，以 LVS 的步驟，陸續選出對於分類正確率有幫助的 Real data，使用 SVM 分類，分類正確率達到 90%即停止，試驗次數 40 次。

6. 隨機抽取 10 筆資料，以 GVS 選出下一次實驗資料進行分類，直到分類正確率達到 90%即停止。

   為了觀察 GVS 最終是否能夠收斂，先隨機抽取 10 筆資料視為 Real data，以 GVS 的步驟，陸續選出對於分類正確率有幫助的 Real data，使用 SVM 分類，分類正確率達到 90%即停止，試驗次數 40 次。

4.1.2 試驗結果記錄

根據以上不同方法，使用資料集一試驗。結果如下：

試驗一（Method O：先隨機抽取 10 筆資料為小樣本資料集，再隨機選擇 3 筆資料進行分類）：

表 4.2 資料集一，Method O 結果記錄

試驗一　Method O 分類正確率				單位：%
77.29	87.23	84.25	83.93	82.70
79.20	86.80	73.23	90.40	79.80
74.80	87.06	84.08	83.28	85.45
70.33	84.74	74.00	85.68	83.19
82.04	83.42	72.86	75.14	68.55
82.21	77.54	85.67	80.76	82.60
81.23	86.67	85.68	77.76	68.33
76.60	82.57	80.84	79.46	88.31
Average : 80.89			Standard Deviation : 5.49	

試驗二（使用與 Method O 抽取的 10 筆相同資料，再以 SVS 選出 3 筆資料，進行分類）：

表 4.3 資料集一，SVS 結果記錄

試驗二　SVS 分類正確率				單位：%
94.08	92.73	86.77	90.79	90.99
89.20	93.58	89.04	92.33	88.73
90.46	95.97	94.01	86.83	91.41
92.14	91.34	86.69	91.11	89.53
95.13	97.21	73.68	77.27	90.92
62.44	82.37	96.42	92.40	91.07
89.13	90.98	87.72	91.86	87.02
76.78	84.96	91.66	89.09	91.40
Average : 88.93			Standard Deviation : 6.59	

試驗三（使用與 Method O 抽取的 10 筆相同資料，再以 LVS 選出 3 筆資料，進行分類）：

表 4.4 資料集一，LVS 結果記錄

試驗三　LVS 分類正確率				單位：%
89.47	91.30	90.18	88.53	86.36
87.48	92.83	86.08	91.18	88.67
91.52	92.67	90.97	87.79	89.52
88.04	90.52	85.94	91.33	87.30
91.46	96.51	85.35	79.65	77.27
87.47	77.60	82.71	87.81	88.71
89.12	86.73	86.68	93.64	87.32
82.45	85.45	87.79	92.88	92.27
Average : 88.16			Standard Deviation : 4.10	

試驗四（使用與 Method O 抽取的 10 筆相同資料，再以 GVS 選出 3 筆資料，進行分類）：

表 4.5 資料集一，GVS 結果記錄

試驗四　GVS 分類正確率				單位：%
88.87	89.76	92.22	88.59	86.38
87.73	93.24	88.22	93.30	88.13
91.88	92.20	90.19	87.86	91.80
89.28	95.08	89.13	90.61	90.45
93.39	91.34	77.86	76.53	74.87
81.23	78.46	88.85	86.94	90.33
84.02	87.13	86.00	83.94	88.36
82.41	91.72	84.46	86.73	91.82
Average : 87.78			Standard Deviation : 4.80	

試驗五(隨機抽取 10 筆資料，以 LVS 的步驟選出資料直到正確率達 90%以上)：

當兩分類間的距離小於 1 時，因 LVS 選擇方式之問題（僅選擇 VX 與不同分類的 RX 之間最近距離)，其分類正確率不會再提升，其正確率無法每次都能達到 90%。圖 4.2 是以 LVS 的步驟陸續選出 50 筆資料之正確率，實驗次數 40 次。

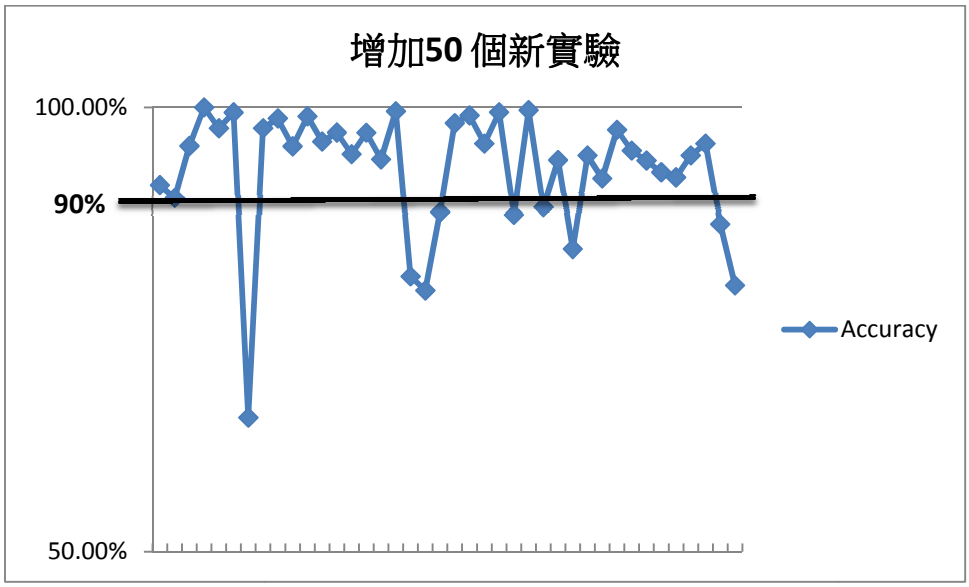

圖 4.2 以 LVS 機制增加 50 個新實驗於資料集一之正確率

試驗六(隨機抽取 10 筆資料，以 GVS 的步驟選出資料直到正確率達 90%以上)：

表 4.6 資料集一，GVS 實驗次數結果記錄

試驗六　GVS 分類正確率達 90%所需實驗次數				
15	17	15	13	6
8	17	14	9	8
6	11	12	6	4
15	18	12	2	17
20	13	8	10	15
3	16	9	13	11
2	15	11	15	9
6	6	2	7	11
Average : 10.675			Standard Deviation : 4.84	

4.1.3 資料集一－試驗結果討論

本試驗結果將以隨機化區集設計(雙因子變異數分析)來做為驗證工具。多因子變異數分析是在探討，兩個以上（包含兩個）的自變數對應變數之分析，可用來檢定自變數對應變數的影響效果和相互作用。

利用 SPSS 檢定三個不同的機制影響是否顯著,如表 4.7 資料集一受試者間因子，每組試驗資料 RX 會經過 4 種不同的機制實驗；有 40 組不同的起始 RX，因此每種機制均有 40 筆資料。

為驗證本研究設計之機制於資料集一中是否有效，以虛無假設 H_0 為 $\mu_{Original} = \mu_{SVS} = \mu_{LVS} = \mu_{GVS}$，其對立假設 H_1 為 $\mu_{Original}$、μ_{SVS}、μ_{LVS} 與 μ_{GVS} 中至少有一個不相等，顯著水準 $\alpha = 0.05$ 來檢定。統計分析結果如表 4.8 所示（表 4.8 中 accuracy 代表最後實驗後的分類正確率）：

表 4.7 資料集一受試者間因子

受試者間因子

		個數
Mechanism	GVS	40
	LVS	40
	Original	40
	SVS	40
Number of test	Test 1	4
	Test 10	4
	Test 11	4
	Test 12	4
	Test 13	4
	Test 14	4
	Test 15	4
	Test 16	4
	Test 17	4
	Test 18	4
	Test 19	4
	Test 2	4
	Test 20	4
	Test 21	4
	Test 22	4
	Test 23	4
	Test 24	4
	Test 25	4
	Test 26	4
	Test 27	4
	Test 28	4
	Test 29	4
	Test 3	4
	Test 30	4
	Test 31	4
	Test 32	4
	Test 33	4
	Test 34	4
	Test 35	4
	Test 36	4
	Test 37	4
	Test 38	4
	Test 39	4
	Test 4	4
	Test 40	4
	Test 5	4
	Test 6	4
	Test 7	4
	Test 8	4
	Test 9	4

表 4.8 資料集一變異數分析

受試者間效應項的檢定

依變數: ACCURACY

來源	型 III 平方和	自由度	平均平方和	F 檢定	顯著性
校正後的模式	.479[a]	42	1.141E-02	10.247	.000
截距	119.557	1	119.557	107385.140	.000
MECHANIS	.167	3	5.568E-02	50.011	.000
NUMBER_O	.312	39	8.002E-03	7.188	.000
誤差	.130	117	1.113E-03		
總和	120.166	160			
校正後的總數	.609	159			

a. R 平方 = .786 (調過後的 R 平方 = .710)

從表 4.8 中可以得知，機制間的檢定結果 p 值為 0.000 小於 0.05，認定分析結果達到顯著差異之標準，$\mu_{Original}$、μ_{SVS}、μ_{LVS}與μ_{GVS}中至少有一個不相等，但無法得知何者較佳，因此再進行多重比較法(multiple comparisons)進一步分析，在此使用 Tukey's Test (Tukey's Honestly Significant Difference Test, Tukey's HSD)，多重比較分析結果如表 4.9 所示：

表 4.9 資料集一多重比較分析結果

多重比較

依變數: ACCURACY

Tukey HSD

(I) Mechanism	(J) Mechanism	平均數差異 (I-J)	標準誤	顯著性	95% 信賴區間	
					下限	上限
GVS	LVS	-3.81E-03	7.461E-03	.956	-2.33E-02	1.564E-02
	Original	6.891E-02*	7.461E-03	.000	4.946E-02	8.835E-02
	SVS	-1.15E-02	7.461E-03	.418	-3.09E-02	7.964E-03
LVS	GVS	3.810E-03	7.461E-03	.956	-1.56E-02	2.326E-02
	Original	7.272E-02*	7.461E-03	.000	5.327E-02	9.216E-02
	SVS	-7.67E-03	7.461E-03	.733	-2.71E-02	1.177E-02
Original	GVS	-6.89E-02*	7.461E-03	.000	-8.84E-02	-4.95E-02
	LVS	-7.27E-02*	7.461E-03	.000	-9.22E-02	-5.33E-02
	SVS	-8.04E-02*	7.461E-03	.000	-9.98E-02	-6.09E-02
SVS	GVS	1.148E-02	7.461E-03	.418	-7.96E-03	3.093E-02
	LVS	7.673E-03	7.461E-03	.733	-1.18E-02	2.712E-02
	Original	8.039E-02*	7.461E-03	.000	6.094E-02	9.984E-02

以觀察的平均數為基礎。

*. 在水準 .05 上的平均數差異顯著。

由表 4.9 中得知，本研究設計出的 SVS、LVS、GVS 與無經過任何挑選機制的 Original 方法比較之下，其統計結果皆有顯著差異，均優於 Original 方法。

表 4.10 資料集一同質子集結果

ACCURACY

Tukey HSD[a,b]

Mechanism	個數	子集 1	子集 2
Original	40	.808920	
GVS	40		.877828
LVS	40		.881637
SVS	40		.889310
顯著性		1.000	.418

同質子集中組別的平均數已顯示。
以型 III平方和為基礎
平均平方和 (誤差) = 1.113E-03 中的誤差項。

a. 使用調和平均數樣本大小 = 40.000

b. Alpha = .05

從表 4.10 中的結果顯示出，機制明顯分成兩群，表中顯示出 SVS、LVS、GVS 較為顯著不同，使用 SVS、LVS 與 GVS 機制可以得到較好的分類正確率。但 SVS、LVS 與 GVS 三種機制間無明顯優劣差異。

在資料集一中，SVS、LVS 與 GVS 三種機制間無明顯優劣差異的情況下，因 LVS 可減少大量於 SVS 中產生的虛擬資料，考量實際操作的便利性及運算資源的節省，LVS 機制是優於 SVS 機制的；若比較 LVS 與 GVS，於資料集一的試驗五中，在指定的試驗迭代次數中，LVS 有可能無法達到正確率 90% 之收斂門檻值，而試驗六中，因 GVS 加入分散機制，能做更多的實驗次數而不會受困於某區域（因分類邊界單純，所需實驗次數平均為 10.675 次就能使分類正確率達 90%），故 GVS 機制又優於 LVS 機制。

4.2 資料集二

使用本研究建立之資料集一共 10000 筆資料，做為小樣本資料集的為原

始資料，下表 4.11 為資料集二的基本項目，其分類情形如圖 4.3 資料集二分類情況，共有三個分類。從中隨機抽取 10 筆資料為本研究小樣本資料集，每分類至少有一筆資料。為突顯分類結果，資料集二計算分類正確率的測試資料集為 3080 筆資料，分布區域如圖 4.4。

表 4.11 資料集二基本資料表

資料集名稱	資料集二
資料筆數	10000 筆資料
類別數	三類，Y=1、Y=2、Y=3
維度	二維，X_1、X_2

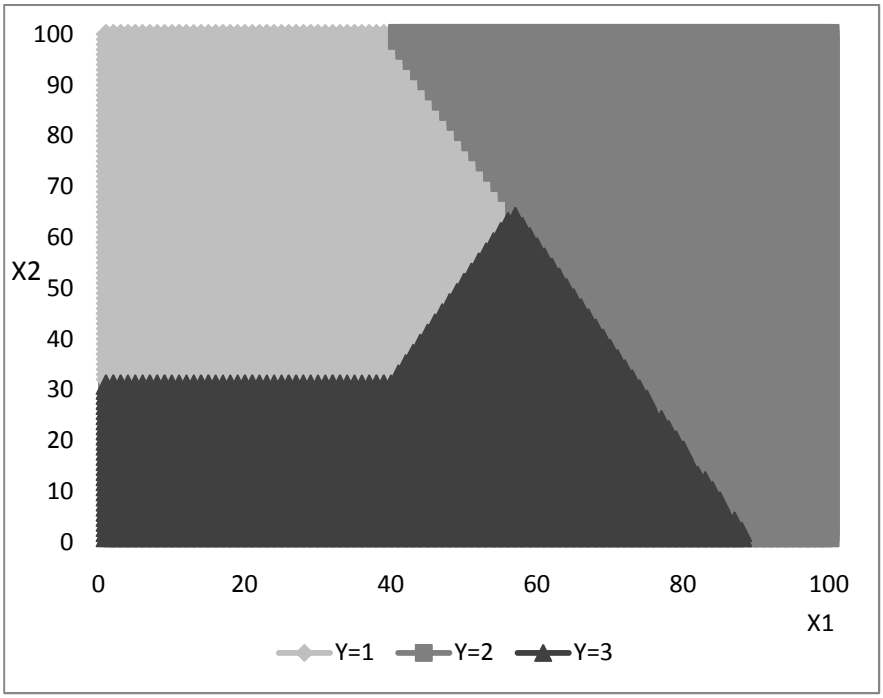

圖 4.3 資料集二分類情況

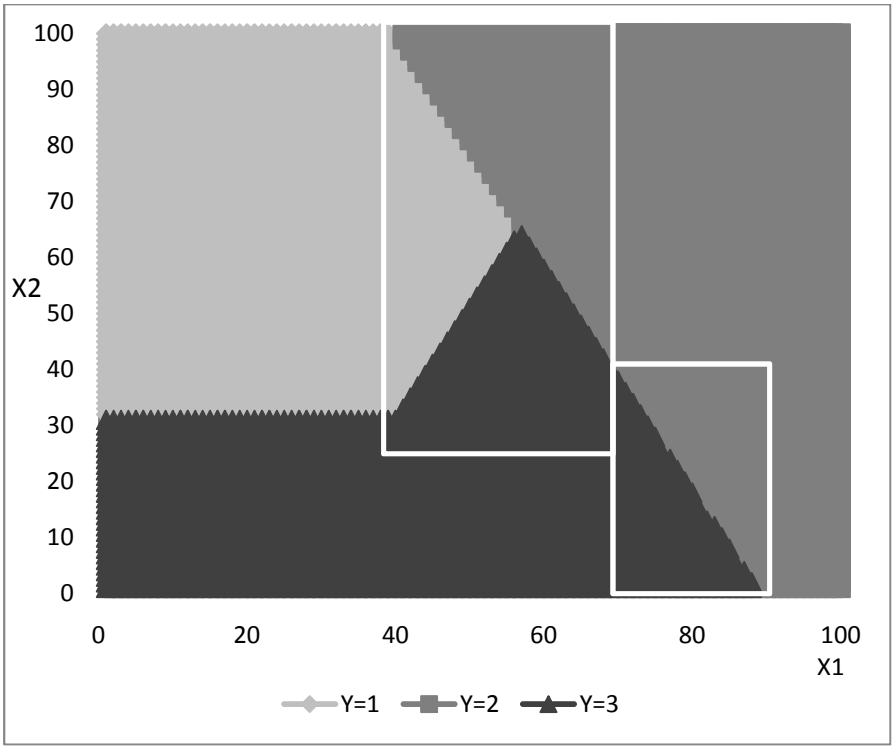

圖 4.4 資料集二測試資料分佈區域

4.2.1 試驗說明

對於分類邊界較複雜的資料集二，做了以下幾種不同的實驗，測試經過挑選方法後加入的 RX 對於分類的正確率是否能提升。

1. Original method（Method O）：先隨機抽取 10 筆資料為小樣本資料集，再隨機選擇 3 筆資料進行分類。

   先隨機抽取 10 筆資料為視為 Real data，再隨機選擇 3 筆資料，使用 SVM 分類，試驗次數 40 次。

2. 使用與 Method O 抽取的 10 筆相同資料，以 SVS 再選出 3 筆資料，進行分類。

   欲得知 SVS 機制是否有效，使用與 Method O 相同的隨機抽取的 10 筆資料視為 Real data，以 SVS 的步驟，陸續選出 3 個對於分類正確率有幫助的 Virtual data 並於資料集二查詢其分類屬性，查詢後將此筆完整資料視為新的 Real data，使用 SVM 測試分類正確率，試驗次數 40 次。

3. 使用與 Method O 抽取的 10 筆相同資料，以 LVS 再選出 3 筆資料，進行分類。

   欲得知 LVS 機制是否有效，使用與 Method O 相同的隨機抽取的 10 筆資料視為 Real data，以 LVS 的步驟，陸續選出 3 個對於分類正確率有幫助的 Virtual data 並於資料集二查詢其分類屬性，查詢後將此筆完整資料視為新的 Real data，使用 SVM 測試分類正確率，試驗次數 40 次。

4. 使用與 Method O 抽取的 10 筆相同資料，以 GVS 再選出 3 筆資料，進行分類。

   欲得知 GVS 機制是否有效，使用與 Method O 相同的隨機抽取的 10 筆資料視為 Real data，以 GVS 的步驟，陸續選出 3 個對於分類正確率有幫助的 Virtual data 並於資料集二查詢其分類屬性，查詢後將此筆完整資料視為新的 Real data，使用 SVM 測試分類正確率，試驗次數 40 次。

5. 隨機抽取 10 筆資料，以 LVS 選出下一次實驗資料進行分類，直到分類正確率達到 90%即停止。

   為了觀察 LVS 最終是否能夠收斂，先隨機抽取 10 筆資料視為 Real data，以 LVS 的步驟，陸續選出對於分類正確率有幫助的 Real data，使用 SVM 分類，分類正確率達到 90%即停止，試驗次數 40 次。

6. 隨機抽取 10 筆資料，以 GVS 選出下一次實驗資料進行分類，直到分類正確率達到 90%即停止。

   為了觀察 GVS 最終是否能夠收斂，先隨機抽取 10 筆資料視為 Real data，以 GVS 的步驟，陸續選出對於分類正確率有幫助的 Real data，使用 SVM 分類，分類正確率達到 90%即停止，試驗次數 40 次。

4.2.2 試驗結果記錄

根據以上不同方法，使用資料集二做試驗。結果如下：

試驗一（Method O：先隨機抽取 10 筆資料為小樣本資料集，再隨機選擇 3 筆資料進行分類）：

表 4.12 資料集二，Method O 結果記錄

試驗一　Method O 分類正確率				單位：%
71.27	67.18	71.30	62.86	60.97
70.78	67.05	68.86	68.47	52.47
63.18	68.34	68.64	64.71	79.68
62.01	68.57	69.71	72.18	73.83
62.99	67.37	73.73	69.12	58.93
63.51	62.34	63.02	59.81	49.71
66.69	67.47	46.69	59.74	63.12
51.43	51.49	67.44	67.50	62.01
Average : 64.65			Standard Deviation : 7.01	

試驗二（使用與 Method O 抽取的 10 筆相同資料，再以 SVS 選出 3 筆資料，進行分類）：

表 4.13 資料集二，SVS 結果記錄

試驗二　SVS 分類正確率				單位：%
83.73	79.06	76.33	70.16	48.28
74.90	70.45	73.99	73.70	61.95
64.38	73.90	64.16	73.21	83.44
70.06	80.65	63.21	78.86	72.21
68.77	53.38	54.81	65.65	66.46
86.46	59.64	75.58	64.87	53.57
64.58	70.55	72.50	65.23	49.64
76.30	61.92	59.35	80.45	64.19
Average : 68.76			Standard Deviation : 9.39	

試驗三（使用與 Method O 抽取的 10 筆相同資料，再以 LVS 選出 3 筆資料，進行分類）：

表 4.14 資料集二，LVS 結果記錄

試驗三　LVS 分類正確率				單位：%
86.56	78.67	75.23	69.16	61.20
78.99	86.95	75.78	73.31	74.35
64.58	82.82	56.62	66.07	77.99
72.86	66.62	75.29	71.88	63.47
69.25	73.41	86.27	67.34	47.31
70.32	64.55	67.11	71.59	69.48
69.22	73.96	79.16	62.63	73.83
67.86	58.93	65.16	63.77	68.57
Average : 70.70			Standard Deviation : 8.18	

試驗四（使用與 Method O 抽取的 10 筆相同資料，再以 GVS 選出 3 筆資料，進行分類）：

表 4.15 資料集二，GVS 結果記錄

試驗四　GVS 分類正確率				單位：%
80.65	72.66	75.75	70.03	63.12
81.07	77.79	88.67	70.52	66.79
63.44	83.02	70.00	69.12	76.17
66.23	69.42	72.53	69.61	70.97
70.52	70.71	83.73	75.88	73.60
71.75	65.94	61.07	65.49	77.01
86.53	72.47	73.54	64.97	51.17
66.04	57.76	59.58	61.53	52.69
Average : 70.49			Standard Deviation : 8.37	

試驗五(隨機抽取 10 筆資料，以 LVS 的步驟選出資料直到正確率達 90%以上)：

與資料集一試驗五相同，當兩分類間的距離小於 1 時，因 LVS 選擇方式之問題（僅選擇 VX 與不同分類的 RX 之間最近距離)，其分類正確率不會再提升，其分類正確率有時無法達到 90%。圖 4.5 以 LVS 機制增加 200 個新實驗於資料集二之正確率是以 LVS 的步驟陸續選出 200 筆資料之正確率，實驗次數 40 次。

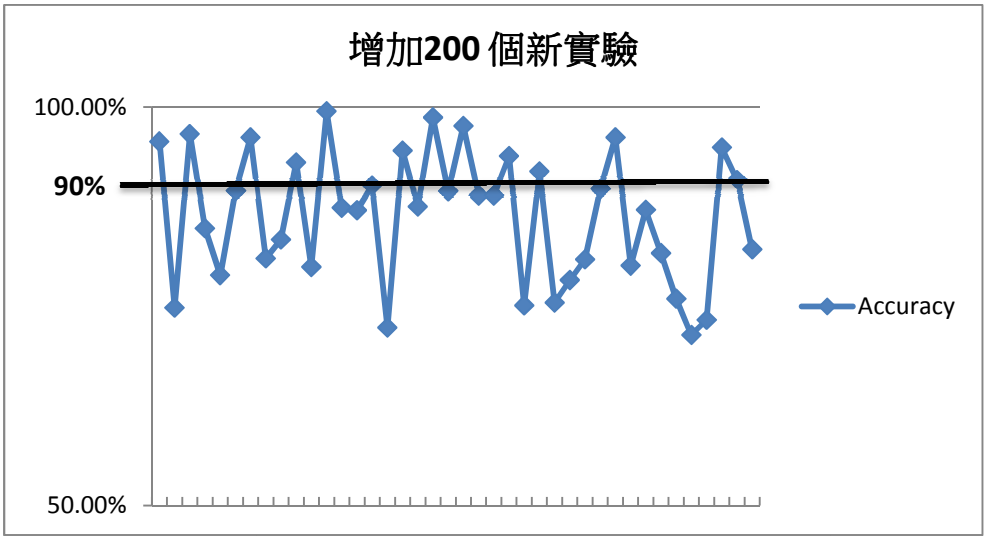

圖 4.5 以 LVS 機制增加 200 個新實驗於資料集二之正確率

試驗六(隨機抽取 10 筆資料，以 GVS 的步驟選出資料直到正確率達 90%以上)：

表 4.16 資料集二，GVS 實驗次數結果記錄

試驗六　Medhod C 分類正確率達 90%所需實驗次數				
15	58	17	33	36
90	16	47	79	45
15	37	54	53	8
63	60	57	74	26
21	80	36	50	40
87	57	11	59	78
55	23	25	25	46
53	30	14	55	60
Average : 44.7			Standard Deviation : 22.48	

4.2.3 資料集二－試驗結果討論

本試驗結果將以隨機化區集設計(雙因子變異數分析)來做為驗證工具。利用 SPSS 檢定三個不同的機制影響於資料集二中是否顯著，如表 4.17 資料集二受試者間因子，每組試驗資料 RX 會經過 4 種不同的機制實驗；有 40 組不同的起始 RX，因此每種機制均有 40 筆資料。

為驗證本研究設計之機制於資料集二是否有效，以虛無假設 H_0 為 $\mu_{Original} = \mu_{SVS} = \mu_{LVS} = \mu_{GVS}$，其對立假設 H_1 為 $\mu_{Original}$、μ_{SVS}、μ_{LVS} 與 μ_{GVS} 中至少有一個不相等，顯著水準 $\alpha = 0.05$ 來檢定。統計分析結果如表 4.18 所示（表 4.18 中的 accuracy 代表最後實驗後的分類正確率）：

表 4.17 資料集二受試者間因子

受試者間因子

		個數
Mechanism	GVS	40
	LVS	40
	Original	40
	SVS	40
Number of test	Test 1	4
	Test 10	4
	Test 11	4
	Test 12	4
	Test 13	4
	Test 14	4
	Test 15	4
	Test 16	4
	Test 17	4
	Test 18	4
	Test 19	4
	Test 2	4
	Test 20	4
	Test 21	4
	Test 22	4
	Test 23	4
	Test 24	4
	Test 25	4
	Test 26	4
	Test 27	4
	Test 28	4
	Test 29	4
	Test 3	4
	Test 30	4
	Test 31	4
	Test 32	4
	Test 33	4
	Test 34	4
	Test 35	4
	Test 36	4
	Test 37	4
	Test 38	4
	Test 39	4
	Test 4	4
	Test 40	4
	Test 5	4
	Test 6	4
	Test 7	4
	Test 8	4
	Test 9	4

表 4.18 資料集二變異數分析

受試者間效應項的檢定

依變數: ACCURACY

來源	型 III 平方和	自由度	平均平方和	F 檢定	顯著性
校正後的模式	.621[a]	42	1.478E-02	3.186	.000
截距	75.410	1	75.410	16253.606	.000
MECHANIS	9.432E-02	3	3.144E-02	6.777	.000
NUMBER_O	.526	39	1.350E-02	2.910	.000
誤差	.543	117	4.640E-03		
總和	76.574	160			
校正後的總數	1.164	159			

a. R 平方 = .533 (調過後的 R 平方 = .366)

從表 4.18 中可以得知，機制間的檢定結果 p 值為 0.000 小於 0.05，認定分析結果達到顯著差異之標準，$\mu_{Original}$、μ_{SVS}、μ_{LVS}與μ_{GVS}中至少有一個不相等，但無法得知何者較佳，在此使用 Tukey's HSD 進行多重比較，多重比較分析結果如表 4.19 所示：

表 4.19 資料集二多重比較分析結果

多重比較

依變數: ACCURACY

Tukey HSD

(I) Mechanism	(J) Mechanism	平均數差異 (I-J)	標準誤	顯著性	95% 信賴區間	
					下限	上限
GVS	LVS	-2.14E-03	1.523E-02	.999	-4.18E-02	3.755E-02
	Original	5.835E-02*	1.523E-02	.001	1.866E-02	9.805E-02
	SVS	1.725E-02	1.523E-02	.670	-2.24E-02	5.695E-02
LVS	GVS	2.143E-03	1.523E-02	.999	-3.76E-02	4.184E-02
	Original	6.050E-02*	1.523E-02	.001	2.080E-02	.100192
	SVS	1.939E-02	1.523E-02	.582	-2.03E-02	5.909E-02
Original	GVS	-5.84E-02*	1.523E-02	.001	-9.80E-02	-1.87E-02
	LVS	-6.05E-02*	1.523E-02	.001	-.100192	-2.08E-02
	SVS	-4.11E-02*	1.523E-02	.039	-8.08E-02	-1.41E-03
SVS	GVS	-1.72E-02	1.523E-02	.670	-5.69E-02	2.245E-02
	LVS	-1.94E-02	1.523E-02	.582	-5.91E-02	2.031E-02
	Original	4.110E-02*	1.523E-02	.039	1.407E-03	8.080E-02

以觀察的平均數為基礎。

*. 在水準 .05 上的平均數差異顯著。

由表 4.19 中得知，本研究設計出的 SVS、LVS、GVS 與無經過任何挑選機制的 Original 比較之下，其統計結果皆有顯著差異，均優於 Original 機制。

表 4.20 資料集二同質子集結果

ACCURACY

Tukey HSD[a,b]

Mechanism	個數	子集 1	子集 2
Original	40	.646534	
SVS	40		.687638
GVS	40		.704886
LVS	40		.707029
顯著性		1.000	.582

同質子集中組別的平均數已顯示。
以型 III平方和為基礎
平均平方和 (誤差) = 4.640E-03 中的誤差項。

a. 使用調和平均數樣本大小 = 40.000

b. Alpha = .05

從表 4.20 中的結果顯示出，機制亦明顯分成兩群，表中顯示出 SVS、LVS、GVS 較為顯著不同，使用 SVS、LVS 與 GVS 機制於資料集二中也可以得到較好的分類正確率。但 SVS、LVS 與 GVS 三種機制間無明顯優劣差異。

在資料集二中，SVS、LVS 與 GVS 三種機制間無明顯優劣差異的情況下，因 LVS 可減少大量於 SVS 中產生的虛擬資料，考量實際操作的便利性及運算資源的節省，LVS 機制是優於 SVS 機制的，與資料集一相同；若比較 LVS 與 GVS 兩機制之優劣，於資料集二的試驗五中，在指定的試驗迭代次數中，LVS 亦有可能無法達到正確率 90%之收斂門檻值，而資料集二試驗六中，因 GVS 加入分散機制，能做更多的實驗次數而不會受困於某區域，可求得全域解，故 GVS 又優於 LVS。因分類邊界較複雜之原因，分類正確率達 90%所需實驗次數平均為 44.1 次高於資料集一使用 GVS 之平均所需實驗次數 10.675 次，且達 90%正確率之次數差異性大，原因可能為起始抽樣參數的所在位置以及是否均勻分散會影響分類正確率達 90%所需實驗次數，此外分散機制內的適應函數的制訂，亦有影響實驗次數的可能。

4.3 本章小結

將第四章中 SVS、LVS 及 GVS 的各試驗比較結果如下：

表 4.21 試驗結果列表

驗證項目	資料集一	資料集二
SVS 與 Method Original 何者較佳	SVS 較佳	SVS 較佳
LVS 與 Method Original 何者較佳	LVS 較佳	LVS 較佳
SVS 與 LVS 何者較佳	無明顯差異	無明顯差異
LVS 是否能收斂	否	否
GVS 與 Method Original 何者較佳	GVS 較佳	GVS 較佳
LVS 與 GVS 何者較佳	無明顯差異	無明顯差異
GVS 是否能收斂	可	可

試驗討論條列如下：

1. 統計結果顯示，使用 SVS、LVS 及 GVS 其分類正確率皆較隨機選擇佳。
2. SVS 之方法在 10 個 VX 中，以虛擬資料增加其資料量（一個 RX 產生 9 個 VX，共 90VX）並從中選擇一個帶有較多訊息的 VX 為實驗目標，接續產生其虛擬資料（再增加一個 RX 與 9 個 VX），使資料量增加，而 LVS 之方法為僅於關鍵 RX 產生 VX 再從 VX 當中選擇一個為下一個實驗目標（一次實驗只增加一個 RX），可減少 VX 的增加（減少 VX 數量等於 RX 數量的九倍），且 SVS 與 LVS 之分類正確率於兩資料集皆無明顯差異的情況下，可得知非關鍵區域的虛擬資料對於 SVM 之正確率的提升成果有限，因此在 SVS 與 LVS 中選擇 LVS 做正確率達 90%所需實驗次數的試驗。

3. 因 LVS 的正確率達 90%所需實驗次數試驗結果得知未能收斂，如圖 4.6，後續的實驗點會太集中無法有效分散。故本研究使用 LVS 之方法增加發散機制形成 GVS，建立一個完整尋求有效實驗點的機制。

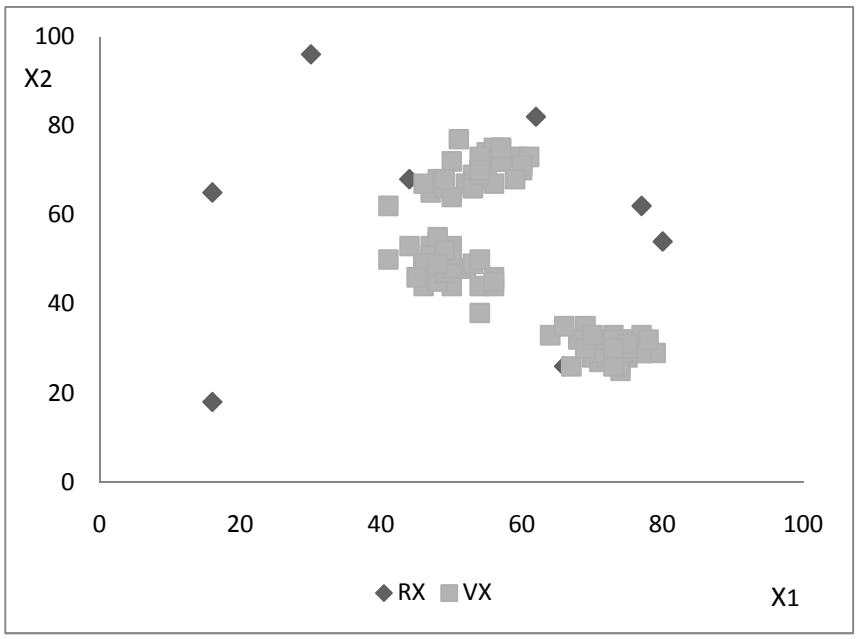

圖 4.6 LVS 實驗多次結果

4. 嘗試以 GVS 做正確率達 90%所需實驗次數試驗，於資料集一約 10.675 次、資料集二約 44.1 次正確率可達 90%。GVS 之試驗結果顯示，於資料集一與資料集二能提升的分類正確率與 SVS 及 LVS 無明顯差異，且 GVS 較 SVS 減少許多虛擬資料的產生，且較 LVS 能求得全域解，故 GVS 為一完整尋求有效實驗點的機制。

第五章　結論及未來研究方向

本章共分為兩部分，將提出本研究之結論，最後提出本研究後續可再深入探究之議題與建議。

5.1 結論

本研究以更有效的利用小樣本資料之方法－參數範圍的搜尋機制來探討在有限的實驗預算中，可能發生的實驗參數分配及選擇的問題，在此提出了一套參數範圍的搜尋機制。首先，在支援向量選擇機制中我們利用所有小樣本資料建立虛擬資料，將小樣本資料及虛擬資料以 SVM 建立模型，期望利用虛擬資料增加 SVM 建立之分類模型的正確率，並從支援向量中選擇一有助於釐清分類邊界的參數；區域虛擬點選擇機制先由小樣本資料中找出對於增加分類正確率有幫助的區域，再透過該區已實驗之參數建立虛擬樣本，由虛擬樣本中選擇一個有助於釐清分類邊界的參數。全域虛擬點選擇機制則是由區域虛擬點選擇機制增加分散機制，有助於全域解的取得。

本研究以兩種資料集之實例試驗來評估參數範圍搜尋機制的可行性。以簡單隨機抽樣之方式從資料集中抽取少量樣本視為小樣本資料，資料集一之分類為直線，資料集二之分類邊界較複雜，以兩資料集各以三種機制試驗，區域虛擬點選擇機制與支援向量選擇機制之差異為減少了大部分虛擬樣本建立的步驟，兩機制皆於統計結果中顯示出後期選擇的實驗參數皆對於分類正確率有幫助，但兩機制間無明顯優劣差異，可得知 SVM 在小樣本情況下已有不差的分類正確率，故虛擬資料對於 SVM 分類正確率提升有限。資料集一與資料集二其分類正確率比較之下可發現分類邊界較複雜時，所需後續實驗次數需要較多次才能有效釐清參數分類邊界，但還是有機會找出所有分類邊界。

本研究之貢獻為建立一參數範圍搜尋機制－全域虛擬點選擇機制，能尋求有效實驗點，減少實驗資源的浪費。在資料為二維、分類屬性值有三種的情況下，若僅有有限的實驗預算，在安排整體實驗參數之參數組合時，先將少量實驗平均分布，取得小樣本分類的分佈狀況，再依本研究之參數範圍的搜尋機制，可將後續僅有的實驗預算用於能有效釐清分類邊界之區域。

本研究仍有不足之處，如全域虛擬點選擇機制中表 3.10 適應函數之制訂，不同的系統特性需要透過不同適應函數之制定，才能夠設計出符合各種系統特性與性能要求，且適應函數的設定與收斂所需次數有關。

5.2 未來發展方向

本研究未來之發展方向條列說明如下：

1. 虛擬樣本平滑參數 h 的設定

   產生虛擬樣本時平滑參數 h 的設定，是否應依兩參數間的距離或是實驗次數而調整，以及其對於求得全域解之實驗次數的關聯性。

2. 抽樣的技術

   當實驗預算有限時，初期小樣本資料實驗參數的分配及選擇，對於分類正確率是否有影響？對於後續參數範圍的搜尋機制有無影響？

3. 分類屬性值的數量增加

   當分類屬性值的數量增加時，本研究之挑選方法必須改進，需另尋一方法選擇參數，計算參數間的距離尚有可發揮之空間，如：加入懲罰因子，亦為值得探討之方向。

4. 資料維度高

   高維度(high dimensional)的資料首先要面對的就是所謂「維度的詛咒」(curse of dimensionality)，當維度越來越高時，任二筆資料會越來越不相似，想要找到所有可能屬性組合的困難度逐漸提高，整個資料空間（資料量及資料複雜度）會成幾何倍數增加，高維度的資料分類特性要求必須有大量的訓練樣本數目，若僅有少量樣本，建立一分類精度高的分類模型更是困難重重，同時大批的資料量亦降低了計算效率，分類精度也未必會有顯著地提升，且維度很高時，一筆資料包含雜訊或記錄不精確的機會便提高，因此要處理高維度資料的方法，必須克服這一方面的問題，除此之外，高維度資料在產生虛擬樣本時，所需的虛擬樣本數量亦有待討論。

參考文獻

林耀三（2003）。應用密度函數估計法提升小樣本學習精確性。（碩士論文，國立成功大學，2003）。全國博碩士論文資訊網，091NCKU5041030。

張光佑（2006）。探討特徵萃取要素於小樣本分類問題。（碩士論文，國立台中教育大學，2003）。全國博碩士論文資訊網，094NTCTC629006。

劉巧雯（2010），擴充屬性資訊以提升小樣本分類之效果。未出版博士論文，國立成功大學，台南。

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 1-43.

Chang, C., and Lin, C. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2004). A Practical Guide to Support Vector Classification. Available at http://www.csie.ntu.edu.tw/ ~cjlin/libsvm

Huang, R., Liu, Q., Lu, H., and Ma, S. (2002). Solving the Small Sample Size Problem of LDA. *IN: PROCEEDINGS OF THE 16 TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, *3*, 29-32.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, *14*(1), 55-63.

Li, D., Chen, L., and Lin, Y. (2003). Using Functional Virtual Population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research*, *41*(17), 4011-4024.

Li, D., and Lin, Y. (2006). Using virtual sample generation to build up management knowledge in the early manufacturing stages. *European Journal of Operational Research*, *175*(1), 413-434.

Niyogi, P., Girosi, F., and Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, *86*(11).

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1076, 1065.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and

Hall.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York, Inc.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Springer, New York.