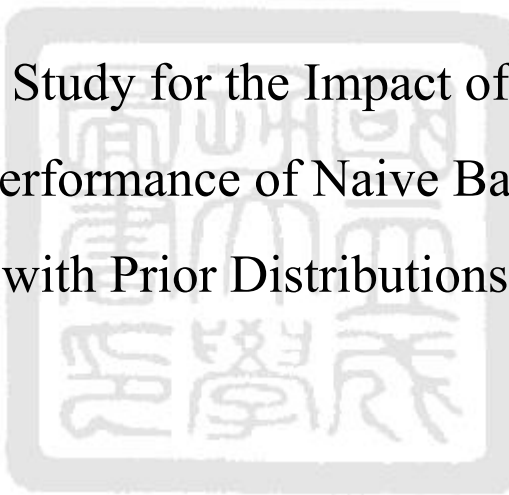


國立成功大學
工業與資訊管理學系碩士在職專班
碩士論文

不同離散化方法對於具先驗分配的簡易貝氏分類器
之影響評估

An Evaluation Study for the Impact of Discretization
Methods on the Performance of Naive Bayesian Classifiers
with Prior Distributions



研究生：楊乃玉

指導教授：翁慈宗 博士

中華民國九十九年六月

國立成功大學

碩士在職專班論文

不同離散化方法對於具先驗分配的簡易貝氏分類器之影響評估

An Evaluation Study for the Impact of Discretization Methods on the Performance of Naive Bayesian Classifiers with Prior Distributions

研究生：楊乃玉

本論文業經審查及口試合格特此證明

論文考試委員：翁榮宗

陳肇泰

王惠嘉

王維聰

指導教授：翁榮宗

系(所)主管：謝中奇

中華民國 99 年 5 月 30 日

摘要

由於簡易貝氏分類器具有使用簡便、運算速度快以及分類正確率佳等優勢，目前已廣泛應用在許多分類任務上。惟簡易貝氏分類器的運作係以離散型態資料為主，倘若欲進行連續型態資料預測，大多數會使用離散化方法進行資料的前置處理。再者，為提升簡易貝氏分類器之分類正確率，一般會假設屬性之先驗分配服從狄氏分配或廣義狄氏分配。所以過去曾經有學者採用不同的離散化方法測試簡易貝氏分類器之先驗分配為狄氏分配的分類正確率，然而研究結果發現，不同離散化方法對於分類正確率的影響並無顯著差異，推測可能原因為狄氏分配的條件限制在實務上過於嚴苛，進而對分類結果造成影響；反之，廣義狄氏分配放寬了狄氏分配的條件假設，使其在實務上能更廣泛地應用。是故，本研究採用常見的四種離散化方法：等寬度、等頻率、比例和最小 entropy，透過實證 UCI 資料存放站中 23 個含有連續型態屬性的資料檔，評估簡易貝氏分類器之先驗分配為最佳狄氏分配或最佳廣義狄氏分配的情況下，使用不同的離散化方法對於分類正確率之影響性。最終，研究結果顯示，等寬度、等頻率與比例離散化方法搭配先驗分配為最佳廣義狄氏分配時，對於分類正確率的提升較有助益；而最小 entropy 離散化方法搭配最佳廣義狄氏分配時的分類正確率，相較於最佳狄氏分配之差異並不大。因此本研究建議，當資料檔類別值個數與離散化後最大屬性可能值個數兩者皆偏多時，最小 entropy 離散化方法才考慮搭配先驗分配為最佳廣義狄氏分配，否則僅需採用最佳狄氏分配即可。

關鍵字：狄氏分配、離散化、廣義狄氏分配、簡易貝氏分類器、先驗分配

Abstract

Naïve Bayesian classifiers are widely employed for classification tasks, because of their computational efficiency and competitive accuracy. Discretization is a major approach for processing continuous attributes for naïve Bayesian classifiers. In addition, the prior distributions of attributes in the naïve Bayesian classifier are implicitly or explicitly assumed to follow either Dirichlet or generalized Dirichlet distributions. Previous studies have found that discretization methods for continuous attributes do not have significant impact on the performance of the naïve Bayesian classifier with noninformative Dirichlet priors. Since generalized Dirichlet distribution is a more appropriate prior for the naïve Bayesian classifier, the purpose of this thesis is to investigate the impact of four well-known discretization methods, equal width, equal frequency, proportional, and minimization entropy, on the performance of naïve Bayesian classifiers with either noninformative Dirichlet or noninformative generalized Dirichlet priors. The experimental results on 23 data sets demonstrate that the equal width, the equal frequency, and the proportional discretization methods can achieve a higher classification accuracy when priors follow generalize Dirichlet distributions. However, generalized Dirichlet and Dirichlet priors have similar performance for the minimization entropy discretization method. The experimental results suggest that noninformative generalized Dirichlet priors can be employed for the minimization entropy discretization method only when neither the number of classes nor the number of intervals is small.

Keywords: Dirichlet distribution, discretization, generalized Dirichlet distribution, naïve Bayesian classifier, prior distribution

誌 謝

本論文之付梓，承蒙恩師 翁慈宗博士的悉心指導。一直以來，其秉持著嚴謹的研究態度，無私地傳授學生從事研究的方法與精神。即使學生沒有良好的基礎，仍然耐心地教導與包容，而且總是帶著和藹親切的笑容，不厭其煩的為學生解惑。無論在課業、生活或感情上，皆會適時地給予關心與建議，促使學生更有動力繼續向前邁進。恩師的諄諄教誨與用心付出，學生永誌難忘，謹在此獻上最誠摯的敬意與謝忱。

此外，論文口試期間，特別感謝陳榮泰博士、王惠嘉博士和王維聰博士百忙之中撥冗審閱本論文，並給予許多寶貴的指正與建議，讓本論文更臻完善。其次，亦對於所有教導與協助過學生的師長們，一併致上萬分謝意。

兩年的研究所生涯轉眼即逝，白天工作晚上求學的辛勞與壓力，在畢業的前夕，似乎頓時消失無蹤。此時此刻心中除了即將畢業的喜悅之外，卻也有著些許的不捨。回顧過往種種，同窗好友間彼此的鼓勵與夥伴機制，以及良豪學長的不吝協助與課後輔導，讓繁忙的課業與報告能夠順利完成。期間有歡笑也有淚水，而這一切即將成為最美好的回憶，感謝大家的陪伴，謝謝你們。

最終，謹將此論文與碩士學位的榮耀，獻給我最愛的父母親和家人，以及鼓勵我不斷進步的朋友和華醫的師長，因為有你們的關心與支持，使我能夠順利完成學位；同時亦對於曾經幫助過我的每個人，表達由衷的感謝，並與大家分享這份喜悅。

楊乃玉 謹誌於

國立成功大學工業與資訊管理學系研究所

中華民國九十九年六月

目 錄

摘 要	I
Abstract	II
誌 謝	III
表目錄	VI
圖目錄	VII
第一章 緒論	1
1.1 研究背景與動機	1
1.2 研究目的	3
1.3 研究流程與架構	3
第二章 文獻探討	5
2.1 貝氏分類器	5
2.1.1 簡易貝氏分類器	6
2.1.2 簡易貝氏分類器相關應用	7
2.2 先驗分配	8
2.2.1 狄氏分配	9
2.2.2 廣義狄氏分配	11
2.3 離散化方法	15
2.3.1 非監督式離散化方法	15
2.3.2 監督式離散化方法	17
2.4 屬性排序方法	21

第三章 研究方法	24
3.1 離散化之前置處理	25
3.2 簡易貝氏分類器	30
3.3 屬性排序	32
3.4 先驗分配之參數調整	35
3.4.1 狄氏分配	35
3.4.2 廣義狄氏分配	37
3.5 結果評估方式	40
第四章 實證研究	42
4.1 資料檔特性	42
4.2 離散化之屬性可能值個數	44
4.3 離散化之先驗分配測試結果	47
4.4 小結	52
第五章 結論與建議	55
參考文獻	57

表目錄

表 3.1 溫度屬性資料表.....	26
表 3.2 等寬度離散化之結果.....	26
表 3.3 遞增排序後的溫度屬性資料.....	27
表 3.4 等頻率離散化之結果.....	27
表 3.5 比例離散化之結果.....	28
表 3.6 最小 entropy 離散化之結果.....	30
表 4.1 資料檔特性.....	43
表 4.2 使用三種非監督式離散化方法產生之屬性可能值個數.....	45
表 4.3 使用最小 entropy 離散化方法產生之屬性可能值個數.....	46
表 4.4 離散化後使用拉普拉斯估計的簡易貝氏分類器之分類正確率.....	47
表 4.5 離散化後於最佳狄氏分配之分類正確率.....	48
表 4.6 離散化後於最佳廣義狄氏分配之分類正確率.....	49
表 4.7 離散化後使用最佳狄氏分配相較於僅加入拉普拉斯估計的簡易貝氏分類器 之分類正確率差異值.....	50
表 4.8 離散化後使用最佳廣義狄氏分配相較於最佳狄氏分配之分類正確率差異值	51

圖目錄

圖 1.1 研究流程圖	4
圖 3.1 研究方法步驟流程圖	24
圖 3.2 使用最小 entropy 離散化處理溫度屬性資料	30
圖 3.3 使用貝氏屬性挑選法進行屬性排序流程圖	34
圖 4.1 使用最小 entropy 離散化方法搭配最佳廣義狄氏分配相較於最佳狄氏分配之 分類正確率差異值與資料檔類別值個數關連圖	52
圖 4.2 第一類群組資料檔於最佳廣義狄氏分配和最佳狄氏分配之分類正確率差異值	53
圖 4.3 第二類群組資料檔於最佳廣義狄氏分配和最佳狄氏分配之分類正確率差異值	53
圖 4.4 四種離散化方法搭配不同先驗分配模式之分類正確率平均值	54

第一章 緒論

隨著資訊科技發展，資料庫的技術日趨成熟，無論在學術單位、政府機關或企業界對於資料庫的應用越來越倚重。其本身除了具有建立與存放資料之基本功能外，更重要的是，資料庫經過長時間的累積，內部儲存了大量的資料，倘若欲充分運用這些資料，一般會透過資料探勘 (data mining) 的方式進行處理，藉此找出隱藏於資料背後有價值的資訊與知識。資料探勘主要分為四大領域，包含：資料分類 (data classification)、資料分群 (data clustering)、資料關聯 (data association) 以及數值預測 (numeric prediction) 等。其中，資料分類領域具有監督式學習 (supervised learning) 的性質，可藉由分類演算法建構分類模型，對資料進行預測，因此經常應用於實務上。例如：以醫療領域而言，透過收集許多患者的歷史醫療記錄，並依照患者所屬的條件和其症狀出現與否，即能預測患者屬於罹患疾病之高風險族群或低風險族群。此外，在分類演算法中，由於簡易貝氏分類器 (naïve Bayesian classifier; NB) 具有使用簡便、運算速度快且分類正確率佳等優勢，目前已經被廣泛應用。

1.1 研究背景與動機

簡易貝氏分類器之運作係以貝氏定理 (Bayes' theorem) 為基礎，藉由訓練樣本的學習，根據貝氏定理以事後機率取代事前機率的方式，再配合各資料屬性間彼此條件獨立的假設，進行分類預測。其次，亦由於簡易貝氏分類器主要是透過計算訓練樣本中，資料屬性值 (attribute value) 分佈於各類別 (class) 的機率，並依此判斷資料所屬類別值。是故，其處理的資料屬性型態以離散型態 (discrete) 資料為主。

反觀於真實世界中，存在著大量連續型態 (continuous) 的資料。因此，簡易貝氏分類器在進行連續型態資料學習前，必須先將連續型態的屬性資料進行前置處理 (preprocess)。一般而言，常見的處理方式有兩種，第一種方式為假設該資料服從常態

分配 (normal distribution)，此即為透過計算該資料的平均值及標準差進行分類預測；第二種方式是先將資料以離散化 (discretization) 的方法處理後，再進行分類 (Hsu et al., 2003; Yang and Webb, 2009)。然而，倘若資料原本不屬於常態分配的情況，則藉由第一種方式進行分類學習的預測結果較不理想。是故，大多數會建議於處理連續型態資料時，採用第二種方式會得到較佳的分類正確率。再者，有研究結果發現，不同離散化方法對於簡易貝氏分類器之分類正確率的影響並無顯著差異 (Dougherty et al., 1995)。因而在進行連續型態資料的離散化處理時，一般會選擇使用簡便且效率佳的離散化方法，例如目前大部分所選用的離散化方法，係採用將資料離散化成十個等區間 (10-bin) 的方式，亦即把原本連續型態的資料，離散化成十個屬性可能值後，再利用簡易貝氏分類器進行分類學習。

除此之外，為了提升簡易貝氏分類器之分類正確率，大多數會以假設屬性之先驗分配 (prior distribution) 服從狄氏分配 (Dirichlet distribution) 的情況進行預測。直至近幾年有研究提出藉由廣義狄氏分配 (generalized Dirichlet distribution) 或羅氏分配 (Liouville distribution) 取代狄氏分配當成先驗分配之方式進行分類學習。所以後續有學者更進一步對簡易貝氏分類器之先驗分配從事研究。其中，曾有學者假設簡易貝氏分類器之先驗分配符合狄氏分配的情況下，以不同離散化方法進行測試比較 (Hsu et al., 2003)，然而研究結果顯示，不同離散化方法對於使用狄氏分配的簡易貝氏分類器之分類正確率並無顯著影響。

綜觀上述，雖然先前已有學者藉由不同的離散化方法進行簡易貝氏分類器之分類正確率的研究，但是其主要係以假設簡易貝氏分類器之先驗分配符合狄氏分配，或者不使用先驗分配的情況下所進行之研究，目前尚未有學者使用不同的離散化方法搭配其他先驗分配進行測試。是故，本研究欲採用其他的先驗分配方式，探討簡易貝氏分類器於不同離散化方法的應用下，其分類正確率之影響情形，並藉此瞭解是否不同的離散化方法有所適合之先驗分配模式。

1.2 研究目的

一般使用簡易貝氏分類器時，會先針對連續型態資料進行離散化處理，而且大部分會採用 10-bin 的方式處理。主要為其方法簡單且相較於其他更複雜的離散化方法之分類結果並無顯著差異。原因可能為簡易貝氏分類器一般會假設狄氏分配為其先驗分配；然而，狄氏分配的條件限制為變數之間必須符合等信賴需求(equal-confidence requirement) 與負相關需求 (negative-correlation requirement)。由於以應用層面而言，此條件限制太過嚴苛，所以對於簡易貝氏分類器之分類正確率可能會造成影響，進而導致不同離散化方法的選擇，對其分類結果無法產生顯著差異。

是故，本研究期望透過文獻探討，以不同的離散化方法搭配其他先驗分配之方式進行分類學習，測試其是否對於簡易貝氏分類器之分類正確率的提升有所助益。綜合以上，本研究之目的如下所述：

- (1) 當先驗分配為廣義狄氏分配時，測試不同離散化方法是否對於簡易貝氏分類器之分類正確率造成影響。
- (2) 瞭解是否有特定的離散化方法，特別適用於簡易貝氏分類器之先驗分配為狄氏分配或廣義狄氏分配的情況。

1.3 研究流程與架構

本論文共分為五章，依序為緒論、文獻探討、研究方法、實證研究，以及結論與建議。第一章緒論係描述本研究之背景與動機，並闡明研究目的；第二章為相關文獻探討，藉此可瞭解目前大部分簡易貝氏分類器所使用的離散化方法，並介紹簡易貝氏分類器、先驗分配與離散化的運作方式；第三章研究方法，主要說明本研究所選用的各種離散化方法如何搭配簡易貝氏分類器之先驗分配進行分類預測；再者，第四章會

透過 VB6.0 撰寫離散化及簡易貝氏分類器之程式，實證 23 個資料檔；最後第五章則對於實證研究結果進行詮釋與總結，並提出未來可能的研究方向與建議。在此彙整本研究之流程圖，如圖 1.1 所示。

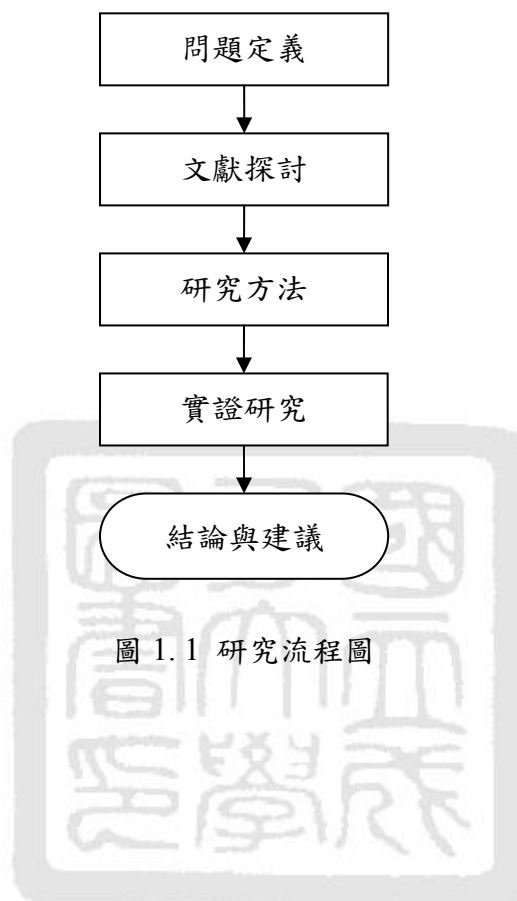


圖 1.1 研究流程圖

第二章 文獻探討

簡易貝氏分類器是由貝氏分類器 (Bayesian classifier) 所發展出的資料探勘分類方法，以貝氏定理為基礎，透過事件發生的機率進行分類預測，因此，處理的資料以離散型態資料為主。若欲進行連續型態資料預測時，大部分會採用離散化的方式進行資料的前置處理，以提升分類正確率。除此之外，簡易貝氏分類器會假設資料檔中的各個屬性值出現的機率服從某先驗分配，例如：狄氏分配或廣義狄氏分配，而且在進行屬性先驗分配之參數調整時，由於先調整的屬性具有較大之影響力，所以必須事先透過屬性排序的方法，決定屬性參數調整之先後順序。

是故，本章第一部分會先藉由貝氏分類器的概述，說明簡易貝氏分類器之運作原理，以及介紹近年來簡易貝氏分類器之相關應用；其次，在第二部分，主要進行狄氏分配與廣義狄氏分配兩種先驗分配的文獻探討；然後第三部分，即針對各種離散化方法做介紹，並透過文獻探討進行分析與比較；最後一個部分，則著重於屬性排序方法之相關文獻介紹。

2.1 貝氏分類器

貝氏分類器是由 Good (1950) 所提出，其透過貝氏定理與貝氏分類法則的結合，推導出貝氏分類器。因此，根據貝氏定理，假設目前有 n 個屬性 X_1, X_2, \dots, X_n ，其中一筆資料 $x=(x_1, x_2, \dots, x_n)$ 屬於第 j 個類別值 C_j 的機率為：

$$p(C_j | x) = \frac{p(C_j \cap x)}{p(x)} = \frac{p(x | C_j)p(C_j)}{p(x)} \quad (2.1)$$

上述等式 (2.1) 中的 $p(C_j | x)$ 係表示在給定資料 x 的條件下，此筆資料屬於類別值 C_j 的機率，即所謂的事後機率 (posterior probability)；其次， $p(x | C_j)$ 稱之為

概似函數 (likelihood function)；然後 $p(C_j)$ 則是代表類別值 C_j 在資料檔中出現的機率，亦稱為事前機率 (prior probability)；而 $p(x)$ 則為資料 x 出現的機率。由於在計算同一筆資料時， $p(x)$ 是固定的，因此等式 (2.1) 可以簡化為：

$$p(C_j | x) \propto p(x | C_j) p(C_j) \quad (2.2)$$

之後，透過計算資料 x 在各類別值的機率，並找出擁有最大事後機率之類別值 C^* ，將資料 x 判定為 C^* 。

2.1.1 簡易貝氏分類器

簡易貝氏分類器為貝氏分類器的一種，其運作原理係將貝氏分類器加入條件獨立的假設，亦即表示在給定某類別值的情況下，各屬性之間必須彼此條件獨立。因此，以等式 (2.2) 為基礎，再加入條件獨立之假設，則可將式子展開成：

$$\begin{aligned} p(C_j | x) &\propto p(x_1 | C_j) \times p(x_2 | C_j) \times \cdots \times p(x_n | C_j) \times p(C_j) \\ &= \prod_{i=1}^n p(x_i | C_j) \times p(C_j) \end{aligned} \quad (2.3)$$

由此可知，簡易貝氏分類器在給定資料 x 的條件下，可藉由等式 (2.3) 計算出該筆資料分佈在各個類別值的機率，並找出具有最大事後機率之類別值 C^* ，將資料 x 預測為類別值 C^* 。然而，大部分實務上的資料無法符合屬性條件獨立之假設，所以有學者 (Domingos and Plazzani, 1997) 提出即使在此假設不成立的情況下，簡易貝氏分類器依然能夠表現良好，其主要原因在於簡易貝氏分類器的損失函數 (loss function) 是採用 0-1 損失函數 (zero-one loss function) 的觀念，而非平方損失函數 (square error loss function)。是故，在只考量分類結果對與錯的情況下，簡易貝氏分類器屬性條件獨立之假設，對於分類正確率的表現影響並不大。

2.1.2 簡易貝氏分類器相關應用

一般而言，實務上在選擇資料探勘的分類工具時，所考量的基本要素包括：使用簡便、正確性高以及效率佳等。由於簡易貝氏分類器具有上述之優勢，目前已經廣泛應用在許多分類任務上。例如：以生物科技領域而言，運算資料量龐大，若能利用簡易貝氏分類器之條件獨立假設，將會大幅地降低分類複雜度。其中，在預測小分子 RNA (micro RNA) 基因方面，先前大多數係透過序列保存或尋找小分子 RNA 同源性的技術，直到 Yousef et al. (2006) 提出一種預測小分子 RNA 基因的新技術，且適用於多物種的預測，其主要作法是結合各物種中已知的小分子 RNA 序列結構，藉由簡易貝氏分類器進行訓練與學習，研究結果顯示，使用簡易貝氏分類器預測小分子 RNA 基因的技術，相較於先前的預測方法具有更高的正確率與敏感度。

再者，倘若能瞭解蛋白質與 RNA 之交互作用機制，亦能成為破解許多重要生物過程的關鍵，因此 Terribilini et al. (2007) 使用簡易貝氏分類器建立一個伺服器網站，透過識別在蛋白質資料庫中的 RNA 結合殘基 (RNA-binding residues)，分析與預測結合 RNA 複合物的蛋白質序列結構。其次，Chen et al. (2007) 提出一種新的矽預測系統，藉由簡易貝氏分類器訓練蛋白質資料庫中的胺基酸對排列方式，預測能夠使蛋白質結晶化的序列。其透過將蛋白質序列轉化為一個固定大小的向量，之後從結晶化與非結晶化的序列中找出46個特徵當成分類屬性，再與其他分類方法進行比較，例如：支援向量機、多變量迴歸以及 C4.5決策樹等。最後研究發現，簡易貝氏分類器在正確率與敏感度方面，顯著優於其他分類方法。

然而，簡易貝氏分類器除了在上述生物科技領域應用廣泛外，亦在其他實務上的分類預測受到重視。以工程材料缺陷偵測領域而言，先前大部分是採用類神經網路的方式進行預測。Addin et al. (2007) 首先將簡易貝氏分類器應用在工程材料缺陷偵測方面，其藉由收集複合材料單層板 (laminated composite materials; LCM) 的振幅波進

行缺陷預測，並透過 f-fold 屬性挑選法及 k-means 分群演算法進行資料前置處理，以簡化屬性個數，之後找出每群中的平均值、最大值及最小值，放入簡易貝氏分類器中進行分類學習。其結果顯示，簡易貝氏分類器能有效提升工程材料缺陷偵測之正確率。

另外，在軟體缺陷檢測方面，大部分的研究係藉由線性迴歸、判別式分析、決策樹、類神經網絡和簡易貝氏分類器等方法進行分析比較，結果顯示簡易貝氏分類器優於其他方法 (Menzies et al., 2007)。但是由於簡易貝氏分類器的屬性具有條件獨立和相同重要性之假設，因此，Turhan and Bener (2009) 透過與先前研究相同的 NASA 公用軟體缺陷資料檔來分析簡易貝氏分類器這些假設。在屬性條件獨立假設的部分，採用主成份分析法 (principal component analysis; PCA) 與子集合挑選法 (subset selection) 做為資料前置處理的方式；而在相同重要性的假設方面，則使用 8 種啟發式的屬性排序方法以加權屬性。最後結果顯示，透過主成份分析的前置處理方式，以及使用 gain ratio 與 information gain 兩種加權屬性方法，能有效提升簡易貝氏分類器之分類正確率。再者，Keren (2003) 係將簡易貝氏分類器應用在圖像風格及視訊動作之辨識方面，藉由圖像特徵的二維屬性預測，鑒別出該畫作之藝術家身份，同時將此應用延伸到三維屬性的視訊動作偵測，則可辨識視訊動作之順序。

2.2 先驗分配

簡易貝氏分類器加入先驗分配的目的，主要係為了在進行分類預測時，能有更接近原始資料概念 (concept) 的訓練樣本。藉由使用者對於資料背景的瞭解，將收集樣本後的資料檔，加入人為的調控，使資料的分佈更趨近於真實情況，以提升預測的正確率。再者，簡易貝氏分類器是透過計算屬性可能值出現的機率進行分類預測，亦即表示在實驗結果為衡量機率值的情況下，其所對應的分配必須符合變數非負與總和為

一的性質。由於多變量分配中的狄氏分配與廣義狄氏分配擁有單位體 (unit simplex) 的特性 (變數非負且總和為一)，因此可將其假設為簡易貝氏分類器之先驗分配。

2.2.1 狄氏分配

在單位體性質的多變量分配中，經常會使用狄氏分配做為先驗分配 (Aitchison, 1985)，主要原因為狄氏分配的一般動差函數 (general moment function) 計算十分簡單，且具有共軛性質，而狄氏分配之定義如下所示：

定義 2.1 隨機向量 $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ 滿足 $\theta_1 + \theta_2 + \dots + \theta_k \leq 1$ 與 $\theta_j > 0, (j=1, 2, \dots, k)$ ；

參數 $\alpha_j > 0, (j=1, 2, \dots, k+1)$ 且 $\alpha = \alpha_1 + \alpha_2 + \dots + \alpha_{k+1}$ ，其機率密度函數為：

$$f(\Theta) = \frac{\Gamma(\alpha)}{\prod_{j=1}^{k+1} \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{\alpha_j-1} (1 - \theta_1 - \theta_2 - \dots - \theta_k)^{\alpha_{k+1}-1} \quad (2.4)$$

上述等式 (2.4) 中的隨機向量 Θ 服從 k 維的狄氏分配，以 $\Theta \sim D_k(\alpha_1, \alpha_2, \dots, \alpha_k; \alpha_{k+1})$ 表示。由此可知，狄氏分配為一多變量分配，並且具有變數非負且總和為一的單位體性質，所以其適合做為簡易貝氏分類器之先驗分配。

再者，由於狄氏分配為貝他分配 (beta distribution) 的擴充，因此當狄氏分配只有一維時，即會退化成貝他分配。另外，Wilks (1962) 證明出狄氏分配的一般動差函數為：

$$m_D(r_1, r_2, \dots, r_k) = E(\theta_1^{r_1} \theta_2^{r_2} \dots \theta_k^{r_k}) = \frac{\Gamma(\alpha) \prod_{j=1}^k \Gamma(\alpha_j + r_j)}{\Gamma(\alpha + r) \prod_{j=1}^k \Gamma(\alpha_j)} \quad (2.5)$$

其中， r 為一般動差函數之參數，且 $r = r_1 + r_2 + \dots + r_k$ 。是故，藉由等式 (2.5) 可得知各變數的期望值、變異數與共變異數，分別如下所示：

$$\text{期望值： } E(\theta_j) = \frac{\alpha_j}{\alpha} \quad (2.6)$$

$$\text{變異數： } Var(\theta_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)} \quad (2.7)$$

$$\text{共變異數： } Cov(\theta_i, \theta_j) = -\frac{\alpha_i \alpha_j}{\alpha^2(\alpha + 1)}, i \neq j \quad (2.8)$$

透過等式 (2.8) 可得知，在服從狄氏分配的隨機向量中，任兩變數之間的共變異數為負相關。因此，使用狄氏分配的其中一項限制為，變數彼此之間必須為負相關，亦即所謂的負相關需求 (negative-correlation requirement)。

其次，針對定義於區間 $[0,1]$ 之變數 W ，Bier and Yi (1995) 定義其正規化變異數為：

$$NV(W) = \frac{Var(W)}{E(W)[1 - E(W)]} \quad (2.9)$$

之後，將等式 (2.6) 與 (2.7) 代入等式 (2.9) 中，即可得到 θ_j 的正規化變異數為：

$$NV(\theta_j) = \frac{1}{\alpha + 1} \quad (2.10)$$

由上述等式 (2.10) 中，可發現 $NV(\theta_j)$ 與 j 的值並無關連，所以無論 j 為任何值，其正規化變異數皆會相同。意即在服從狄氏分配的隨機向量中，藉由正規化變異數做為判斷依據時，其信賴水準都會一樣，在此稱為等信賴需求(equal-confidence requirement) (Wong, 2009)。

另一方面，若將拉普拉斯估計 (Laplace's estimate) (Cestnik and Bratko, 1991)應用於簡易貝氏分類器中，則相當於係假設屬性之先驗分配服從狄氏分配，且參數設定為1的情況，即 $D_k(1, 1, \dots, 1; 1)$ 。此外，狄氏分配具有共軛性質，係指當先驗分配為狄氏分配，且概似函數為多項式分配時，事後分配亦會服從狄氏分配的情況。假設 $y = (y_1, y_2, \dots, y_{k+1})$ 為資料屬性的 $k+1$ 個可能值分別出現的次數，且隨機向量 $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ 為該屬性第 $1 \sim k$ 個可能值出現的機率，其服從 k 維的狄氏分配 $\Theta \sim D_k(\alpha_1, \alpha_2, \dots, \alpha_k; \alpha_{k+1})$ ，然後概似函數 $(y|\Theta)$ 服從多項式分配，而 Θ 在給定 y 的機率密度函數 $f(\Theta|y)$ 會與 $p(y|\Theta) \times p(\Theta)$ 成正比。由於狄氏分配具有共軛性質，因此 Θ 在給定 y 的條件下，仍服從狄氏分配，即 $(\Theta|y) \sim D_k(\alpha_1 + y_1, \alpha_2 + y_2, \dots, \alpha_k + y_k, \alpha_{k+1} + y_{k+1})$ (Wong, 1998)。如果 θ_j 為 Θ 的其中一個變數，則 θ_j 在給定 y 的條件下，期望值為：

$$E(\theta_j|y) = \frac{y_j + \alpha_j}{\sum_{i=1}^{k+1} (y_i + \alpha_i)} \quad (2.11)$$

因此藉由等式 (2.11) 可計算出在給定類別值 C_j 的情況下，某屬性可能值 x_i 發生的機率 $p(x_i|C_j)$ ；之後，再透過等式 (2.3) 找出具有最大事後機率的類別值，當作該筆預測資料的類別值。

2.2.2 廣義狄氏分配

以實務應用層面而言，由於狄氏分配之條件限制過於嚴苛，因此 Connor and Mosimann (1969) 提出廣義狄氏分配，透過完全中立 (complete neutrality) 的概念，將狄氏分配一般化，推導出廣義狄氏分配。其主要條件限制為第一個變數與其他變數之間必需為負相關，而其他變數則無此限制；再者，相較於狄氏分配，廣義狄氏分配並

無正規化變異數相等之限制。是故，整體而言，廣義狄氏分配放寬了狄氏分配的條件限制，雖然增加了一些運算的複雜度，但是在實務上可更廣泛的被應用。其定義如下所示：

定義 2.2 隨機向量 $\Theta=(\theta_1, \theta_2, \dots, \theta_k)$ 滿足 $\theta_1+\theta_2+\dots+\theta_k \leq 1$ 與 $\theta_j > 0, (j=1, 2, \dots, k)$ ；

參數 α_j 、 β_j 、 λ_j 符合 $\alpha_j > 0, (j=1, 2, \dots, k)$ 、 $\beta_j > 0, (j=1, 2, \dots, k)$ 、 $\lambda_k = \beta_k - 1$ 與

$\lambda_j = \beta_j - \alpha_{j+1} - \beta_{j+1}, (j=1, 2, \dots, k-1)$ ，且機率密度函數為：

$$f(\Theta) = \prod_{j=1}^k \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j-1} (1-\theta_1-\dots-\theta_j)^{\lambda_j} \quad (2.12)$$

在上述等式 (2.12) 中的隨機向量 Θ 服從 k 維的廣義狄氏分配，表示為 $\Theta \sim GD_k(\alpha_1, \alpha_2, \dots, \alpha_k; \beta_1, \beta_2, \dots, \beta_k)$ 。透過定義 2.2 可得知，廣義狄氏分配與狄氏分配相同，兩者皆為多變量分配，並且具有單位體的特性；是故，同樣適合做為簡易貝氏分類器之先驗分配。

除此之外，Wong (1998) 也證明廣義狄氏分配之一般動差函數為：

$$m_{GD}(r_1, r_2, \dots, r_k) = E(\theta_1^{r_1} \theta_2^{r_2} \dots \theta_k^{r_k}) = \prod_{j=1}^k \frac{\Gamma(\alpha_j + \beta_j)\Gamma(\alpha_j + r_j)\Gamma(\beta_j + \delta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)\Gamma(\alpha_j + \beta_j + r_j + \delta_j)} \quad (2.13)$$

其中， r 為一般動差函數之參數，且 $\delta_j = r_{j+1} + r_{j+2} + \dots + r_k, (j=1, 2, \dots, k-1)$ ， $\delta_k = 0$ 。

同時，藉由等式 (2.13) 可得知各變數的期望值、變異數與共變異數，茲如下述：

當 $j=1, 2, \dots, k$ 時，

$$\text{期望值： } E(\theta_j) = E\left[Z_j \prod_{i=1}^{j-1} (1-Z_i)\right] = \frac{\alpha_j}{\alpha_j + \beta_j} \prod_{i=1}^{j-1} \frac{\beta_i}{\alpha_i + \beta_i} \quad (2.14)$$

$$\text{變異數： } Var(\theta_j) = \frac{\alpha_j(\alpha_j + 1)}{(\alpha_j + \beta_j)(\alpha_j + \beta_j + 1)} \prod_{i=1}^{j-1} \frac{\beta_i(\beta_i + 1)}{(\alpha_i + \beta_i)(\alpha_i + \beta_i + 1)} - E(\theta_j)^2 \quad (2.15)$$

此外，若為第 $k+1$ 個變數時，

$$\text{期望值： } E(\theta_{k+1}) = E\left[\prod_{i=1}^k (1-Z_i)\right] = \prod_{i=1}^k \frac{\beta_i}{\alpha_i + \beta_i} \quad (2.16)$$

$$\text{變異數： } Var(\theta_{k+1}) = \prod_{i=1}^k \frac{\beta_i(\beta_i + 1)}{(\alpha_i + \beta_i)(\alpha_i + \beta_i + 1)} - E(\theta_{k+1})^2 \quad (2.17)$$

以下為三種情況之共變異數：

首先，當 $j = 2, 3, \dots, k+1$ 時，第1個與第 j 個變數之共變異數為：

$$Cov(\theta_1, \theta_j) = -\frac{E(\theta_j)}{E(1-\theta_1)} Var(\theta_1) \quad (2.18)$$

其次，當 $j = 2, 3, \dots, k-1$ 時，除了第1個變數以外的任意兩個相鄰變數之共變異數為：

$$Cov(\theta_j, \theta_{j+1}) = E(Z_{j+1})E\left[Z_j(1-Z_j)\right] \prod_{i=1}^{j-1} E\left[(1-Z_i)^2\right] - E(\theta_j)E(\theta_{j+1}) \quad (2.19)$$

然後在 $1 < j < m < k$ 的情況下，任意兩個變數之共變異數為：

$$Cov(\theta_j, \theta_m) = \left[\frac{E(Z_m)}{E(Z_{j+1})}\right] \left[\prod_{i=j+1}^{m-1} E(1-Z_i)\right] Cov(\theta_j, \theta_{j+1}) \quad (2.20)$$

透過等式 (2.18)、(2.19) 與 (2.20) 可得知， θ_1 與其他變數皆為負相關，但除了 θ_1 以外，

任意兩個變數之間可為正相關或負相關。再者，Wong (2009) 證明出廣義狄氏分配的隨機變數並無等信賴需求，亦即其變數的正規化變異數不一定相同；然而，倘若其參數符合 $\beta_j = \alpha_{j+1} + \beta_{j+1}$, ($j=1, 2, \dots, k-1$) 時，所有變數的正規化變異數皆會相同，則隨機向量 Θ 將退化為狄氏分配。是故，相對於狄氏分配之負相關需求與等信賴需求，廣義狄氏分配的條件限制較為寬鬆，也更能符合實務上之應用。

除此之外，若將廣義狄氏分配當作簡易貝氏分類器之先驗分配時，假設 $y = (y_1, y_2, \dots, y_{k+1})$ 為資料屬性的 $k+1$ 個可能值分別出現的次數，且隨機向量 $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ 服從 k 維的廣義狄氏分配，即 $\Theta \sim GD_k(\alpha_1, \alpha_2, \dots, \alpha_k; \beta_1, \beta_2, \dots, \beta_k)$ ，然後概似函數 $(y | \Theta)$ 服從多項式分配，而 Θ 在給定 y 的機率密度函數 $f(\Theta | y)$ 會與 $p(y | \Theta) \times p(\Theta)$ 成正比。由此可知，廣義狄氏分配亦具有共軛性質，因此 Θ 在給定 y 的條件下，仍然會服從廣義狄氏分配，即 $(\Theta | y) \sim GD_k(\alpha_1 + y_1, \alpha_2 + y_2, \dots, \alpha_k + y_k; \beta_1 + \sum_{i=2}^{k+1} y_i, \beta_2 + \sum_{i=3}^{k+1} y_i, \dots, \beta_k + y_{k+1})$ (Wong, 1998)。

再者，倘若 θ_j 為 Θ 的其中一個變數，令 $n_i = y_i + y_{i+1} + \dots + y_k + y_{k+1}$, ($i=1, 2, \dots, k+1$)，則 θ_j 在給定 y 的條件下，當 $j=1, 2, \dots, k$ 時，期望值為：

$$E(\theta_j | y) = \frac{y_j + \alpha_j}{\alpha_j + \beta_j + n_j} \prod_{i=1}^{j-1} \frac{\beta_i + n_{i+1}}{\alpha_i + \beta_i + n_i} \quad (2.21)$$

其次，若是第 $k+1$ 個變數，則期望值為：

$$E(\theta_{k+1} | y) = \prod_{i=1}^k \frac{\beta_i + n_{i+1}}{\alpha_i + \beta_i + n_i} \quad (2.22)$$

藉由等式 (2.21) 或 (2.22) 即可計算在類別值 C_j 的情況下，某屬性可能值 x_i 發生的

機率 $p(x_i|C_j)$ 。之後，再透過等式 (2.3) 找出擁有最大事後機率的類別值，做為簡易貝氏分類器預測資料之類別值。

2.3 離散化方法

由於簡易貝氏分類器的運作方式，主要是透過計算資料屬性值分佈於各類別的機率進行分類預測，所以處理的資料屬性以離散型態的資料為主。反觀在真實世界中，存在著大量連續型態的資料；是故，簡易貝氏分類器在預測連續型態的屬性資料時，大部分會藉由離散化的方式進行資料的前置處理，或者假設資料分佈符合常態分配的情況下進行預測。然而有研究發現，倘若資料屬性不符合常態分配時，選用離散化的方式進行資料的前置處理，會有更好的結果(Dougherty et al., 1995; Kohavi and Sahami, 1996)。

此外，Hsu et al. (2003) 提出狄氏分配具有完美聚集 (perfect aggregation) 的特性，所以在假設簡易貝氏分類器之先驗分配服從狄氏分配的情況下，採用離散化的方法進行資料處理，會有較佳的分類表現。再者，Dougherty et al. (1995) 將離散化方法區分為：非監督式 (unsupervised) 與監督式 (supervised)、整體 (global) 與局部 (local)，以及靜態 (static) 與動態 (dynamic) 等三種構面。其中，非監督式與監督式的離散化方法，係以離散化時有無考量樣本的類別作為區隔。因此，本研究將採用非監督式與監督式的區分方式，針對簡易貝氏分類器經常使用的離散化方法，以及基於簡易貝氏分類器所設計之離散化方法進行探討。

2.3.1 非監督式離散化方法

非監督式的離散化方法係指在不考慮類別值的情況下，所進行的資料離散化處理。一般最常見的方法有兩種，分別是等寬度離散化 (equal width discretization; EWD)

與等頻率離散化 (equal frequency discretization; EFD) (Catlett, 1991; Kerber, 1992; Dougherty et al., 1995)。首先，等寬度離散化方法的步驟為，先找出資料的最大值與最小值（即全距），然後把所有資料依全距平均分成 t 個相同寬度的區間。其中， t 為一預設參數，大部分會假設 $t=10$ ，即所謂的 10-bin，此亦表示將該資料屬性離散化成 10 個屬性可能值。

再者，等頻率離散化方法，係指先將資料進行排序，再把排序後的屬性資料個數平均分成 t 個區間， t 同樣為預設參數，因此每個區間中會有相同數量的資料樣本。此外，等寬度與等頻率離散化方法經常被應用於簡易貝氏分類器之離散化處理，主要的原因為使用簡便、效率佳，且離散化的表現與其他方法並無顯著差異 (Hsu et al., 2003)。反之，此兩種離散化方法皆必須事先設定參數 t ，是故，倘若在小規模的資料檔進行離散化處理時，則可能造成每個區間的資料筆數呈現不足之情況。

另一方面，以簡易貝氏分類器而言，雖然透過離散化的方式處理連續型態的資料，會比假設為常態分配的情況有更好的表現。但是，由於離散化會造成在相同區間中，原本不一樣的屬性值有資訊損失的情形，因而產生離散化的偏差與變異。所以，最佳的離散化方法必須要兼顧離散化的偏差與變異，因為較大的區間樣本數通常存在著較低變異；而較大的區間數則會有較低的偏差。因此，當訓練資料增加時，區間數與區間內的樣本數皆需要增加；反之，在訓練資料減少時，亦應減少區間數與區間內樣本數。是故，有學者 (Yang and Webb, 2009) 針對改善離散化的偏差與變異，提出兩種非監督式離散化方法，分別為：比例離散化 (proportional discretization; PD) 與固定頻率離散化 (fixed frequency discretization; FFD)。

其中，比例離散化係指將區間數與區間內的樣本數設定為相同的值。假設 N 為樣本數，則其區間數 t 及區間內樣本數 s 為：

$$\sqrt{N} = s = t, \quad s \times t = N \quad (2.23)$$

由等式 (2.23) 可得知，只要透過增加訓練資料的樣本數，即可達到較低的離散化偏差與變異，藉此使簡易貝氏分類器的機率預測結果更為穩定與可靠。同時，為了確保每個離散化的區間內有足夠的樣本數，使分類錯誤率能保持在可容忍的範圍內，Yang and Webb (2009) 亦提出固定頻率離散化的方式。其主要是藉由設定足夠大小的區間樣本數 m ，例如： $m=30$ （統計推論中的最小樣本數），並將遞增排序後的屬性值，離散化到 m 個樣本數的區間內，藉此讓每個區間都包含足夠的 m 個樣本，而產生更可靠的分類預測結果。因此，固定頻率離散化與等頻率離散化最主要的差異在於，等頻率離散化的區間樣本數為一任意值，而固定頻率離散化的區間樣本數 m 值，為一個足以減少簡易貝氏分類器離散化變異之數值。

2.3.2 監督式離散化方法

相對於非監督式的離散化方法，在考慮資料類別值的情況下進行的資料離散化處理，稱為監督式離散化方法。首先，Fayyad and Irani (1993) 提出一個以熵(entropy) 為基礎的方法，稱為最小熵離散化方法 (minimization entropy discretization)。其步驟為先將連續型屬性資料進行排序，評估每個排序屬性二元分割後的候選切割點，並計算其 entropy 值。之後，選擇 entropy 值最小的切割點，將資料離散化成兩個區間。假設有一連續資料樣本 S ，包含 k 個類別值 C_1, C_2, \dots, C_k ，每個類別值在樣本 S 中的機率為 $P(C_i, S)$ ，則樣本 S 的 entropy 值計算方式如等式 (2.24) 所示：

$$Ent(S) = -\sum_{i=1}^k P(C_i, S) \log_2(P(C_i, S)) \quad (2.24)$$

然後，在屬性變數為 X_i 的情況下，切割點 T 會將樣本 S 之區間範圍離散化成兩個子區間，其中包含的樣本子集合分別為 S_1 與 S_2 ，此時 entropy 值 則為：

$$E(X_i, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \quad (2.25)$$

因此，在給定資料屬性 X_i 的情況下，藉由等式 (2.25) 的計算結果，可選出最小 entropy 值的切割點 T_{min} ，當成樣本 S 的二元離散化切割點，並重複此方式繼續尋找各個子區間中最小 entropy 值之切割點，使樣本 S 離散化成多元區間，直到符合停止切割條件為止。

再者，Fayyad and Irani (1993) 藉由最小化描述長度原則(minimum description length principle; MDLP) 做為最小 entropy 離散化方法的停止切割條件，如下所示：

$$Gain(X_i, T; S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(X_i, T; S)}{N} \quad (2.26)$$

其中， N 為樣本 S 的資料筆數， T 為最小 entropy 值的切割點；

$$Gain(X_i, T; S) = Ent(S) - E(X_i, T; S) \quad (2.27)$$

$$\Delta(X_i, T; S) = \log_2(3^k - 2) - [kEnt(S) - k_1Ent(S_1) - k_2Ent(S_2)] \quad (2.28)$$

在等式 (2.28) 中， k_i 代表樣本 S_i 中的相異類別值個數。因此，倘若切割點達到等式 (2.26) 的條件時，則可停止進行離散化的切割。同時，Fayyad and Irani (1993) 也證明透過最小化 entropy 離散方法所選出的切割點，必定會位於兩個不同類別值區間的交界處 (boundary)，而可藉此減少大量候選切割點的計算。

是故，有學者 (Dougherty et al., 1995) 透過簡易貝氏分類器與決策樹C4.5演算法的應用，比較非監督式離散化方法的10-bin 和 bin-log ℓ (Spector, 1990)，與監督式離散化方法的 entropy-based 和 one rule (Holte, 1993)。其中，bin-log ℓ 同樣為一種等寬度離散化方法，而bin的個數是在 ℓ 與 $2\log \ell$ (ℓ 為相異屬性可能值的個數) 中，取最大值當作 t 值；另外，one rule 則係指將排序後的資料樣本依類別值分 bin，每個 bin 至少須包含6筆資料，而且只能有一個主要類別值，並依此方式找出其分類規則進行預測。最後研究結果發現，使用 entropy-based 的離散化方法，能有效提升分類演算

法的表現，而且以簡易貝氏分類器離散化後的表現最佳。此外，Kohavi and Sahami (1996) 也藉由實驗比較 entropy-based 與 error-based 之離散化方法。其中，error-based 的離散化方法係指以找出最小錯誤率區間之方式，進行離散化切割。雖然方法較為簡單，但由於其不允許兩個相鄰區間有相同的類別值，因此，最後研究結果顯示，使用 entropy-based 方法離散化後的表現，優於 error-based 的離散化方法。

之後，Elomaa and Rousu (2003) 進一步延續最小化 entropy 離散方法的應用，提出在簡易貝氏分類器離散化的決策切割點，稱為分割交界 (segment borders)。主要作法是將排序後的相同屬性值合成一個 bin，再找出每個 bin 的主要類別值，使其錯誤率最小。其次，在相鄰的 bin 之間若有相同類別值，則繼續合併成一個區塊 (block)。因此，最後產生的切割點會落在不同類別區塊分佈的交界處。由於此離散方法只需檢查產生在不同類別區塊交界處的切割點，是故，也能藉此減少許多不必要的切割點檢查，進而提升簡易貝氏分類器的離散化效率。

另一方面，Hsu et al. (2000, 2003) 藉由狄氏分配的特性推導出延遲離散化(lazy discretization; LD) 方法。由於此離散化方法在訓練資料時，尚未進行離散化處理，而必須等待測試資料出現後，才開始進行離散化，所以稱之為延遲離散化方法。其主要是透過測試資料產生一對切割點，亦相當於只創造一個區間，並將此區間定義為 I ，因此，在連續屬性變數 $X=x$ 的情況下，

$$I=[x-0.5\delta, x+0.5\delta] \quad (2.29)$$

在等式 (2.29) 中， δ 為一個常數區間，其寬度係藉由其他離散化方法產生之區間而定。例如：以 10-bin、bin-log ℓ 或 entropy-based 產生的離散化區間。此外，Hsu et al. (2000, 2003) 發現離散化後所產生的切割處，不只是一個切割點的值 (point value)，亦有可能是集合值 (set value) 或區間值 (interval value)。由於延遲離散化可以產生集合值或區間值的切割處，因此更適合大部分離散化之情況；反之，延遲離散化的效能

表現雖然與其他離散化方法近似，但卻因為其必須等待測試資料出現後才進行離散化處理，所以需要大量的記憶體空間儲存訓練資料；是故，延遲離散化方法在大型資料檔的表現並不理想。

此外，先前大部分的離散化技術，是將連續型屬性離散化到不相交的重疊區間。但是，以簡易貝氏分類器而言，在給定樣本的情況下，只需要考慮單一屬性區間在各個類別值的機率；是故，可將連續型屬性離散化到重疊的區間內。其次，位於離散區間交界點附近的屬性值，相較於位於區間中段處的屬性值，可能預測正確率較不可靠，倘若能藉由離散化方法形成重疊的區間，則可產生較靠近區間中段處之屬性，例如：延遲離散化亦是將屬性值設定於區間的中段處。因此，Yang and Webb (2002a) 藉由上述之概念，提出一種離散化方法，稱為重疊離散化 (non-disjoint discretization; NDD)。其透過將連續屬性離散化後產生重疊的區間，使屬性值皆能靠近離散化區間的中段處，而獲得更可靠的分類正確率。同時，更以重疊離散化的方法調整區間數量和區間內的樣本數，藉此尋求離散化偏差和變異之間的適當權衡。

其次，Yang and Webb (2002b) 比較 9 種離散化方法在簡易貝氏分類器的正確率，以及在大型資料檔的表現。研究結果顯示，透過延遲離散化、重疊離散化與加權比例 k 區間的離散化方法進行資料的前置處理，會比其他離散化方法更能降低簡易貝氏分類器之分類錯誤率。其中，重疊離散化可以減少資訊損失的情形；加權比例 k 區間離散化可以適當地調整離散化的偏差與變異；然而，使用延遲離散化方法雖然可獲得不錯的分類正確率，但在大型資料檔的表現並不理想。因此，Yang and Webb (2002b) 藉由結合加權比例 k 區間離散化和重疊離散化的優勢，推導出另一種新的離散化方法，稱為加權重疊離散化 (weighted non-disjoint discretization; WNDD) 方法，並透過實證研究與其他離散化方法進行分析比較。最後研究結果顯示，使用加權重疊離散化方法在簡易貝氏分類器的表現會優於其他離散化方法。

2.4 屬性排序方法

一般而言，為了提升簡易貝氏分類器之分類正確率，大部份會假設資料樣本各個屬性可能值出現之機率服從某先驗分配，例如：狄氏分配或廣義狄氏分配（如本章2.2節所述）。其次，在使用先驗分配時，必須針對各個屬性的先驗分配進行參數調整。由於優先調整參數的屬性具有較大之影響力，所以需要事先透過屬性排序的方法，決定各屬性先驗分配參數調整之先後順序，才能產生最佳先驗分配的參數設定模式。除此之外，大多數的屬性排序評估方式是從屬性挑選方法延伸而來，所以在決定屬性之重要程度時，亦可直接透過屬性挑選方法排列各屬性之優先順序 (Liang et al., 2008)。是故，本小節也將針對一些表現良好的屬性挑選方法進行探討。

首先，有學者 (John et al., 1994) 將屬性挑選演算法分為 filter 與 wrapper 兩種。主要的區別方式在於，使用 filter 方法挑選屬性時，係依據統計測度分析屬性間的相關性進行挑選，並不考慮特定分類器的影響；反之，wrapper 方法主要是以屬性集合在特定分類器的表現進行評估。因此，Langley and Sage (1994) 提出一種 wrapper 演算法，稱為貝氏屬性挑選法 (selective Bayesian classifier; SBC) 或稱簡易貝氏屬性挑選法 (selective naïve Bayesian algorithm; SNB)。其基於簡易貝氏分類器的概念，透過屬性條件獨立的假設，以向前搜尋的方式評估各屬性之分類正確率，且每次僅挑選一個最佳屬性放入原本的空集合中，直到屬性子集合的分類正確率下降或不再提升為止，藉此可排除相依屬性與冗餘的屬性，並由研究結果可發現，SNB 能有效提升簡易貝氏分類器之分類表現。

再者，張良豪 (2009) 藉由SNB與先驗分配結合的方式，提升簡易貝氏分類器之效能。其在資料前置處理的部分是採用10-bin做為離散化方法，並透過兩種模式進行分析比較，模式一係先透過 SNB 篩選出屬性群，再依被選出的屬性先後順序，設定各屬性之先驗分配；模式二則為每當 SNB 挑選出一個屬性時，即針對該屬性設定最

適合的先驗分配參數，之後才尋找下一個屬性，直到正確率不再提升為止。最後經由研究結果顯示，使用模式一的方式，並搭配先驗分配為廣義狄氏分配時，能產生較好且較穩定之分類正確率。

另一方面，於使用 filter 屬性挑選方法時，大多數所採用的屬性排序評估準則，係透過 entropy 或相互資訊 (mutual information; MI) 進行評估。因此，Battiti (1994) 提出一種藉由 MI 的評估方法，可改善 MI 計算複雜度高的缺點，其透過在給定選擇屬性的條件下，使用貪婪屬性選擇方法估算候選屬性與類別之間的相互資訊，稱為相互資訊屬性選擇方法 (mutual information feature selection; MIFS)，並利用微調冗餘屬性之參數 β 呈現。然而，此方式必須由使用者自訂參數 β ，所以可能無法準確地估計 MI。由於一個不恰當的 β 值，可能會嚴重影響屬性挑選的正確性。因此，之後有學者 (Kwak and Choi, 2002) 改良 MIFS 方法，稱為 MIFS-U。其主要係透過假設屬性符合均勻分配之情況進行估計，但卻仍無法提供一個與冗餘屬性有強烈相關的參數 β 。是故，Huang et al. (2008) 推導出一種無參數的 MI 方法，即在給定屬性選擇子集合的條件下，候選屬性與類別之間的 MI 估計方式，稱為第二相互資訊屬性挑選法 (second order MI based feature selection algorithm; SOMIFS)。其主要優勢在於不需任何參數設定，且可權衡類別和冗餘選擇屬性之間的相關性。最後實驗結果顯示，雖然 SOMIFS 的正確率在統計上與其他方法並無顯著的差異，但由於其不需設定參數，且在大型資料檔的表現也不差，所以整體而言 SOMIFS 優於其他的 MI 方法。

此外，由於屬性挑選方法是一種全面性的搜尋策略，可能會導致計算複雜度高的問題；同時，最佳的屬性挑選方法為了權衡計算複雜度與最佳化的問題，雖然提供了解決計算複雜度的方法，但卻無法保證能選出最佳的屬性子集合。因此，Liang et al. (2008) 提出一個新的 filter 屬性挑選方法，稱為基於距離辨別的屬性挑選法 (feature selection algorithm based on a distance discriminant; FDSS)。其不僅能解決高計算複雜度的問題，而且可以克服次佳方法的缺點。FDSS 主要是將屬性選擇的搜尋問題轉換

成屬性排序的方式，藉以降低計算的複雜度。其次，FDSS 在計算屬性之間的距離時，並非採用一般的歐氏距離 (Euclidean distance)計算，而是以 Liang et al. (2008) 提出的距離辨別 (distance discriminant)方式進行計算，並且透過設定參數 β 值，做為在相同類別間的屬性距離較大者之懲罰方式。最後研究結果顯示，FSDD 勝過 Robnik and Kononenko (2003) 提出的屬性排序方法 ReliefF，以及 Peng et al. (2005) 提出的 mrmrMID 方法。並且由於 FSDD 計算複雜度低，運算效率佳，因此也適用於高維度或大型資料檔的評估

再者，Biesiada et al. (2005) 針對六個基於 entropy 的屬性排序方法，以及兩個基於統計指標之屬性排序方法進行分析比較，分別為：非對稱相依係數(asymmetric dependency coefficient) (Sridhar et al., 1998)、標準化 gain ratio (Setiono and Liu, 1995)、計算類別和屬性 entropy 之 gain ratio、基於 MI 的對稱不確定性(symmetrical uncertainty) (Press et al., 1988)、距離公理準則 D_{ML} (Lopez de Mantaras, 1991)、結合權重與 entropy 的指標 Chi、卡方值 (χ^2)，以及皮爾森線性相關係數(Pearson's linear correlation coefficient) 等，並藉由 k 最鄰近法 (k-nearest neighbors algorithm) ($k=1$)、簡易貝氏分類器與決策樹 C4.5 進行測試比較。最後研究結果顯示，每個評估指標在不同分類器與不同資料檔的表現並不一致。是故，Biesiada et al. (2005) 認為上述 8 種屬性排序評估方式，沒有單一最好的指標，而應依照不同的資料檔，採用不同的屬性排序指標，藉此提升分類正確率。

第三章 研究方法

透過第二章文獻探討之內容介紹，可瞭解各種不同離散化方法的特性、簡易貝氏分類器和其先驗分配方法的定義，以及常見的幾種屬性排序方法之運作情形。因此，後續本章將分為五個小節，闡述研究方法之詳細流程架構與評估方式。第一節主要為說明如何透過本研究選用的離散化方法，將連續型屬性資料進行離散化處理；其次，在第二節的部分會詳述本研究使用簡易貝氏分類器之處理細節；而第三節將針對如何應用屬性排序方法，找出各個屬性先驗分配之參數調整順序進行介紹；然後第四節係說明簡易貝氏分類器在使用先驗分配時，屬性參數之調整與設定方式；最後在第五節的部分，則會介紹本研究結果之評估方法。綜合上述，彙整本研究方法之步驟流程，茲如圖3.1所示。

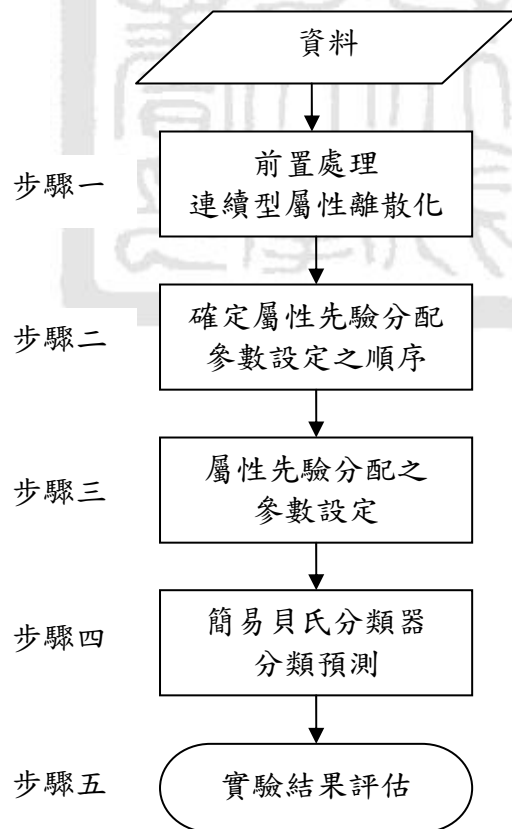


圖 3.1 研究方法步驟流程圖

3.1 離散化之前置處理

由於使用簡易貝氏分類器預測前，必須先將連續型態的屬性資料進行離散化之前置處理，為此在第二章的文獻探討已經針對目前經常使用在簡易貝氏分類器，或基於簡易貝氏分類器所設計的幾種離散化方法進行介紹。其中，以非監督式離散化方法而言，等寬度與等頻率兩種離散化方法由於使用簡便，且分類正確率相較於其他更複雜的離散化方法並無顯著差異，所以經常應用於簡易貝氏分類器之離散化處理；再者，Yang and Webb (2009) 提出的比例離散化方法，亦透過實驗證明離散化後的分類正確率優於其他方法。因此，本研究會將此三種非監督式的離散化方法納入研究範圍內。其次，於監督式離散化方法的部分，最常被使用與比較的離散化方法為 Fayyad and Irani (1993) 所提出的最小entropy離散化方法，雖然其計算方式較為複雜，但在提升分類正確率方面表現良好。是故，本研究將採用上述四種離散化方法進行測試比較，分析不同離散化方法在簡易貝氏分類器使用先驗分配時的正確率表現，並於此小節以某溫度屬性資料為例（如表 3.1），詳述四種離散化方法之使用方式與步驟。

首先，在非監督式離散化方法的部份，以等寬度離散化方法而言，主要係透過找出該連續型屬性資料的最大值 V_{max} 與最小值 V_{min} 之區間範圍，把此區間範圍分為 t 個等寬區間的 bin，如等式 (3.1) 所示：

$$W = (V_{max} - V_{min}) / t \quad (3.1)$$

因此會產生 $t-1$ 個切割點，分別為：

$$(V_{min} + W), (V_{min} + 2W), \dots, (V_{min} + (t-1)W) \quad (3.2)$$

在上述等式中， t 為一預設參數，大部分會假設 $t=10$ ，即一般常見的 10-bin，此亦表示可將連續型態的資料離散化成 10 個屬性可能值。

再者，以表3.1 溫度屬性資料表為例，其中包含33筆資料和2個類別值，倘若欲使用等寬度的10-bin進行離散化處理，則先要找出資料樣本中的最大值 85 與最小值 65，然後將此樣本離散化成 10 個等區間，每個區間寬度皆為 $(85-65)/10=2$ 。因此，透過10-bin離散化後的切割點為：67、69、71、73、75、77、79、81、83，而最後離散化的區間結果如表 3.2 所示。亦即可將原本 33 筆連續型態的屬性資料，轉化成10個離散型態之區間範圍。

其次，在等頻率離散化方法的部分，必須先將屬性以遞增方式排序，並把排序後的屬性可能值分成 t 個區間，且 t 亦為一預設參數。然而，等頻率與等寬度離散化方法最主要的差別在於，等頻率離散化方法的每個區間必須包含 N/t 個訓練樣本（ N 為樣本個數），此外如果有相同的屬性可能值，則會離散化到相同的區間內，所以每個區間會產生非常近似的樣本數，藉此使分類預測結果更可靠。倘若同樣以溫度屬性資料為例，進行等頻率離散化之處理時，亦需先將屬性資料以遞增方式排序，而經過排序後的屬性值及類別分佈如表 3.3 所示。由於溫度屬性資料的樣本數為33筆，假設預設參數 $t=8$ ，則每個區間會有接近 $33/8 \approx 4$ 筆資料。因此，使用等頻率離散化的結果，茲如表 3.4 所示。

表 3.1 溫度屬性資料表

類別	Temperature																
yes	83	65	81	80	70	65	81	75	81	85	65	83	85	81	66	79	80
no	70	75	69	72	71	72	68	75	75	69	72	72	68	75	70	70	

表 3.2 等寬度離散化之結果

區間	1	2	3	4	5	6	7	8	9	10
範圍	[65,67]	(67,69]	(69,71]	(71,73]	(73,75]	(75,77]	(77,79]	(79,81]	(81,83]	(83,85]
樣本數	4筆	4筆	5筆	5筆	5筆	0筆	1筆	6筆	2筆	2筆

表 3.3 遞增排序後的溫度屬性資料

類別	65	66	68	69	70	71	72	75	79	80	81	83	85
yes	3	1			1			1	1	2	4	2	2
no			2	2	3	1	4	4					

表 3.4 等頻率離散化之結果

區間	1	2	3	4	5	6	7	8
範圍	[65,66]	(66,69]	(69,71]	(71,74]	(74,78]	(78,80]	(80,82]	(82,85]
樣本數	4筆	4筆	5筆	4筆	5筆	3筆	4筆	4筆

是故，透過上述溫度屬性資料進行離散化處理的例子，除了能更清楚說明等寬度與等頻率離散化方法之運作方式外，亦能藉此瞭解其具有使用簡便及效率佳等優點；再者，使用等寬度與等頻率離散化後之分類正確率的表現，相較於其他更複雜的方法並無顯著差異 (Hsu et al., 2003)，所以經常應用於簡易貝氏分類器之離散化處理。

反之，倘若訓練樣本的資料量減少時，藉由上述等寬度與等頻率兩種離散化方法進行處理，則可能會造成每個區間中的樣本數過少，而對未來預測的分類正確率造成影響。因此，以Yang and Webb (2009) 提出的比例離散化方法而言，可藉由設定離散化的區間數與區間內的樣本數，達到有效控制離散化的偏差與變異。其主要作法是先將區間數 t 與區間內的樣本數 s 設定為相同的值，回顧第二章之等式 (2.23)。假設 N 為樣本數，透過 $\sqrt{N} = s = t$ ，找出區間數與區間內的樣本數，藉此可增加每個區間的資料筆數，而達到減少離散化的偏差與變異，並增加簡易貝氏分類器預測正確率的穩定性與可靠度。再者，仍以溫度屬性資料為例，使用比例離散化方法進行資料的前置處理。由於資料表的樣本筆數為33筆，所以離散化後的區間數與區間內之樣本數為： $\sqrt{33} = 5.7 \approx 6$ ，而其離散化的結果呈現於表 3.5。

表 3.5 比例離散化之結果

區間	1	2	3	4	5	6
範圍	[65,68]	(68,70]	(70,74]	(74,79]	(79,81]	(81,85]
樣本數	6筆	6筆	5筆	6筆	6筆	4筆

另一方面，在監督式離散化方法的部分，由於必須考慮類別值的分佈情形，所以大多數係透過計算資訊混亂程度來決定切割點，即所謂的 entropy 值，而且當 entropy 值越小時，代表切割後的區間資料欲達到明確之分類結果所需使用的資訊 (bits) 越少。因此，以 Fayyad and Irani (1993) 提出的最小 entropy 離散化方法而言，主要係評估屬性值在排序後，各個相異屬性值之間的候選切割點，其二元分割後的 entropy 值，並選擇最小的 entropy 值當成切割點，將資料離散化成兩個區間；之後，重複上述方式，將連續型屬性資料離散化成多元區間。

假設以表3.1 溫度屬性資料為例，使用最小 entropy 離散化方法處理時，同樣必須先將屬性排序（如表 3.2），再分別計算每個相異屬性值之間候選切割點的 entropy 值，計算方式如第二章等式 (2.24) 與 (2.25) 所示：

$$Ent(S) = - \sum_{i=1}^k P(C_i, S) \log_2(P(C_i, S))$$

$$E(X_i, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

若以屬性值 70 與 71 之間的候選切割點 $(70+71)/2 = 70.5$ 而言，則需先分別計算小於 70.5 和大於 70.5 的區間樣本 entropy 值。其中，小於 70.5 的樣本資料 S_1 包含 12 筆資料，有 5 筆 yes、7 筆 no，所以 entropy 值為：

$$Ent(S_1) = - \frac{5}{12} \log_2 \frac{5}{12} - \frac{7}{12} \log_2 \frac{7}{12} = 0.979 \text{ bits}$$

而大於 70.5 的樣本資料 S_2 ，包含 21 筆資料，有 12 筆yes、 9 筆no，因此其entropy 值為：

$$Ent(S_2) = -\frac{12}{21} \log_2 \frac{12}{21} - \frac{9}{21} \log_2 \frac{9}{21} = 0.985 \text{ bits}$$

是故，使用切割點 70.5 進行離散化後，產生的二元區間之 entropy 值則為：

$$E(X_i, T; S) = \frac{12}{33} \times 0.979 + \frac{21}{33} \times 0.985 = 0.983 \text{ bits}$$

由此可知，最小entropy 離散化方法主要係藉由上述方式，計算每個相異屬性值之間候選切割點的entropy值，選擇最小entropy值的二元切割點，將資料離散化成兩個區間，在此稱為第一階段區間離散化（如圖 3.2 之第一階段數列），並藉此找出第一階段區間之最小 entropy 值的切割點為 77；之後，繼續在第一階段離散化後的二元區間內，使用相同的計算方式，分別尋找第二階段子區間的最小entropy值切割點（如圖 3.2 之第二階段數列）；依此類推，直到符合最小化描述長度原則之切割停止條件為止，而此條件限制如第二章之等式 (2.26) 所示：

$$Gain(X_i, T; S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(X_i, T; S)}{N}$$

綜觀上述，假設以溫度屬性資料為例，在考慮類別值分佈的情況下，透過最小entropy 離散化方法處理的結果，則會將屬性資料離散化成三個區間範圍，如表 3.6 所示。

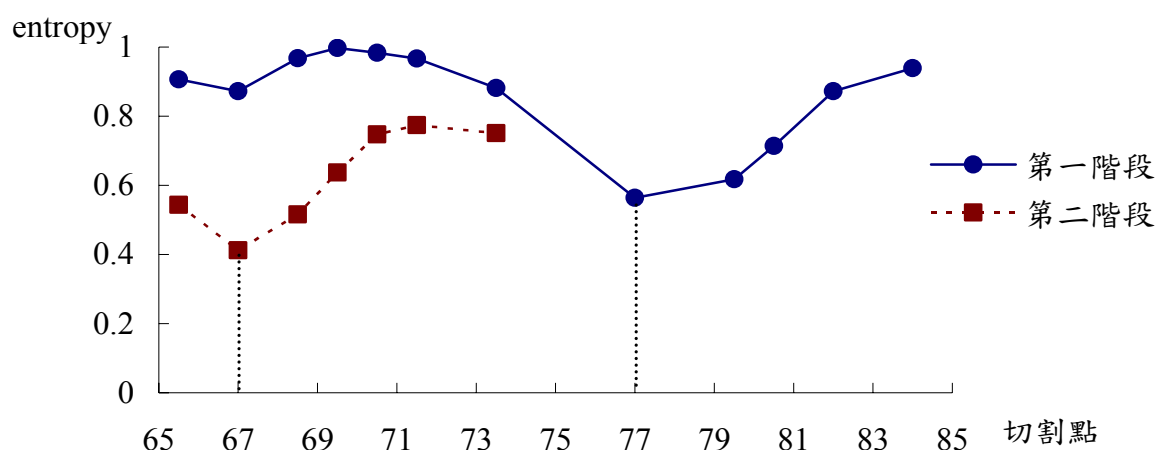


圖 3.2 使用最小 entropy 離散化處理溫度屬性資料

表 3.6 最小 entropy 離散化之結果

區間	1	2	3
範圍	[65,67]	(67,77]	(77,85]
樣本數	4筆	18筆	11筆

此外，由於 Fayyad and Irani (1993) 證明透過最小 entropy 離散化方法所選出的切割點，會產生在兩個不同類別值區間的交界處，因而可減少大量候選切割點的計算。是故，以溫度屬性資料的第一階段區間離散化為例，原本需要計算 12 個候選切割點的 entropy 值，如果只考慮不同類別值區間交界處之候選切割點，則僅需計算 67 和 77 兩個候選切割點，而可省略 10 個候選切割點的評估。倘若在相異屬性值個數眾多時，則能夠藉此節省相當多的運算時間及成本。

3.2 簡易貝氏分類器

簡易貝氏分類器之運作原理在上一章已有詳細介紹，其主要係以貝氏定理為基礎，再加入屬性條件獨立的假設，進行分類預測。因此，假設某一資料檔中有 n 個屬性 X_1, X_2, \dots, X_n ，倘若在給定某筆資料 $x = (x_1, x_2, \dots, x_n)$ 的情況下， x 屬於第 j 個類別值 C_j 的機率，回顧第二章等式 (2.1) 所示，則事後機率為：

$p(C_j | x) = p(x | C_j)p(C_j) / p(x)$ 。其中， $p(x)$ 為資料 x 出現的機率，由於在比較不同類別值之事後機率時，同一筆資料出現的機率皆相等，所以可省略分母的部份，將式子簡化為等式 (2.2)： $p(C_j | x) \propto p(x | C_j)p(C_j)$ ；之後，根據簡易貝氏分類器屬性條件獨立的假設，則可再將等式 (2.2) 展開成等式 (2.3)：

$$\begin{aligned} p(C_j | x) &\propto p(x_1 | C_j) \times p(x_2 | C_j) \times \cdots \times p(x_n | C_j) \times p(C_j) \\ &= \prod_{i=1}^n p(x_i | C_j) \times p(C_j) \end{aligned}$$

是故，簡易貝氏分類器係透過計算在給定某一類別值 C_j 的情況下，資料 x 的各個屬性值 (x_1, x_2, \dots, x_n) 出現之機率 $\prod_{i=1}^n p(x_i | C_j)$ （即概似函數），藉此找出資料 x 在各類別值的事後機率，並判定擁有最大事後機率的類別值 C^* 為資料 x 之類別值。

其次，一般在使用簡易貝氏分類器時，為了避免訓練資料中，有些屬性可能值從未出現在某些類別，而造成概似函數之計算結果產生 0 的情況，大部分會藉由 Laplace's estimate 的方式進行屬性參數之設定，如等式 (3.3) 所示：

$$p(x_i = V_{im} | C_j) = \frac{y_{im} + 1}{y + (k + 1)} \quad (3.3)$$

在上述等式中，屬性值 x_i 為第 i 個屬性的第 m 個可能值 V_{im} ； y_{im} 表示 x_i 與類別值 C_j 同時出現的次數； y 為類別值 C_j 出現之次數；而 $k+1$ 則代表第 i 個屬性的可能值個數。

由此可知 Laplace's estimate 係指在進行分類預測時，先將概似函數分子部份所代表的屬性可能值出現於某類別值之次數加上 1，然後分母的部份再加上該屬性可能值的個數，藉此避免某些屬性的可能值從未出現在特定類別值，造成概似函數分子部分為 0，進而影響分類的結果。

另一方面，資料如果出現遺漏值 (missing value) 的情形，本研究係透過將該資料屬性忽略不計的方式進行處理，意即在計算某筆資料的概似函數 $\prod_{i=1}^n p(x_i | C_j)$ 時，假設該筆資料的第二個屬性有遺漏值的情況，則該筆資料的 $p(x_2 | C_j)$ 會被忽略不列入計算。因此，以原本資料必須計算 n 個屬性的可能值出現之機率而言，倘若其中某一屬性資料出現遺漏值，本研究將會採用只計算 $n-1$ 個屬性的概似函數，進行簡易貝氏分類器之分類預測。

3.3 屬性排序

一般而言，使用簡易貝氏分類器會加入先驗分配，其主要目的係為了使訓練資料的分佈更符合真實之情況，所以必須針對各個屬性的先驗分配進行參數調整。然而，屬性先驗分配的參數調整之先後順序，對未來的分類結果會產生影響。由於優先進行參數調整之屬性具有較大之影響力，因此，在使用屬性先驗分配之前，必須透過屬性排序的方式，決定屬性參數調整之先後順序。

同時，藉由第二章的文獻介紹，雖然已能瞭解目前常見的幾種屬性排序方式與屬性挑選方法，但是整體而言，大多數的屬性排序方法之結果評估差異並不大 (Biesiada et al., 2005)，因此使用上較少會採用高計算複雜度的方法，例如：基於計算屬性之間相互資訊的方法；反之，由 Langley and Sage (1994) 提出的貝氏屬性挑選法，係基於簡易貝氏分類器的概念延伸出的屬性挑選方法，其計算複雜度較低，且能有效提升簡易貝氏分類器之分類表現；再者，張良豪 (2009) 亦提出結合貝氏屬性挑選法與先驗分配的方式進行實驗，並透過研究結果顯示貝氏屬性挑選法能有效提升簡易貝氏分類器之分類正確率，而且分類表現更為穩定。是故，本研究將採用貝氏屬性挑選法，找出屬性先驗分配之參數調整順序。

其次，貝氏屬性挑選法的運作機制，主要係以向前搜尋的方式，透過簡易貝氏分類器評估各屬性之分類正確率，而且每次會挑選一個最佳的屬性放入空集合中，直到屬性子集合的分類正確率下降或不再提升為止，藉此排除相依或冗餘屬性。但是，由於本研究不考慮資料檔中冗餘屬性或相依屬性對於分類正確率的干擾影響，所以在使用貝氏屬性挑選法時，並不會因為挑選出的屬性子集合無法提升分類正確率而停止挑選，反而會繼續進行屬性挑選的步驟，直至所有屬性排序結果皆呈現為止，以找出屬性先驗分配之參數調整順序。因此，本研究使用貝氏屬性挑選法進行屬性排序之流程，茲如圖 3.3 所示。

此外，本研究在使用貝氏屬性挑選法時，亦會先將所有屬性的參數設定為 Laplace's estimate，如上一小節之等式 (3.3) 所示，以避免某些屬性可能值從未出現在特定類別，而造成分類預測之概似函數計算結果為 0 的情況。因此，本研究在使用貝氏屬性挑選法進行屬性排序時，同樣會先將所有屬性可能值之出現次數加上 1，再進行簡易貝氏分類器之分類預測，以找出屬性先驗分配之參數調整順序。

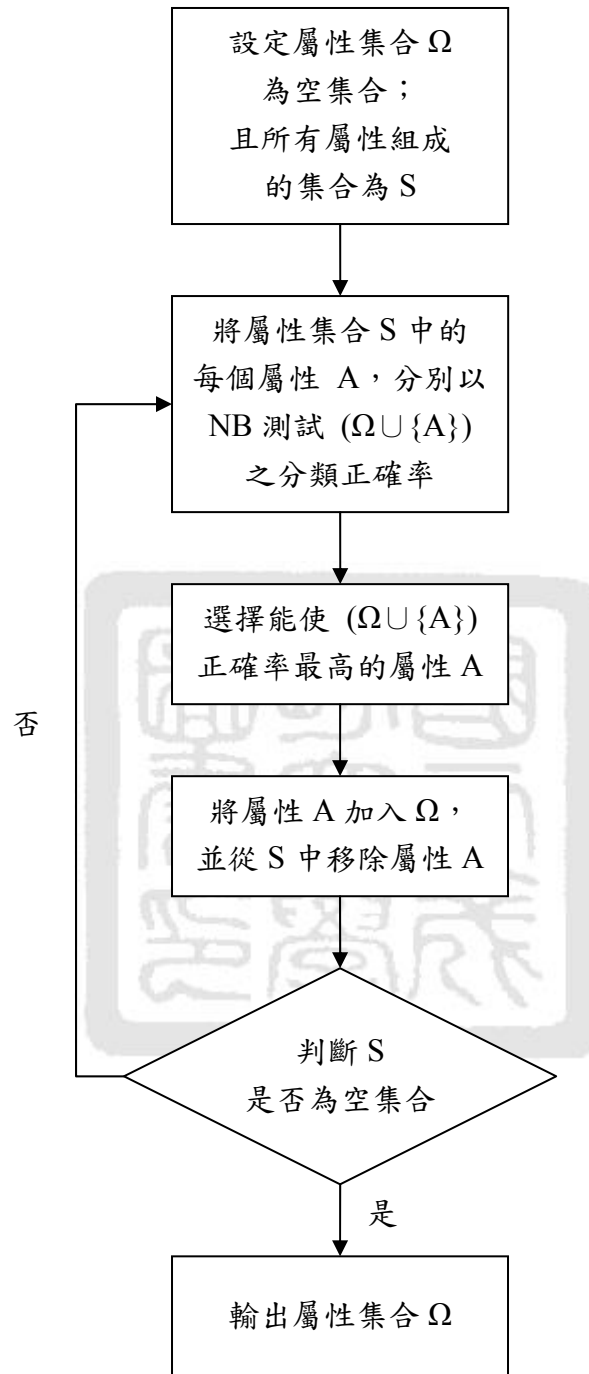


圖 3.3 使用貝氏屬性挑選法進行屬性排序流程圖

3.4 先驗分配之參數調整

經由上一小節的說明，可清楚地瞭解如何應用貝氏屬性挑選法決定屬性先驗分配參數調整之順序，接下來的研究步驟，即開始針對各個屬性的先驗分配進行參數調整。此外，回顧第二章在先驗分配文獻探討的敘述，曾經提及簡易貝氏分類器之先驗分配必須符合變數非負且總和為一的單位體特性，同時具有共軛性質。所以，本研究將選用符合上述條件的狄氏分配與廣義狄氏分配做為簡易貝氏分類器之先驗分配，並在此一小節詳述其參數調整之方式。

3.4.1 狄氏分配

首先，在第二章的文獻探討中，已詳細介紹過狄氏分配之定義，並瞭解其為多變量的機率分配，具有單位體之特性和共軛性質，而且狄氏分配的一般動差函數計算簡單，是故，經常被用來做為簡易貝氏分類器之先驗分配。其次，在使用狄氏分配時，為了運算操作簡便，以及滿足無資訊性 (noninformative) 的限制與降低對事後期望值的影響 (Wong, 2009)，大部分會採用 Laplace's estimate 的方式（如3.2小節的等式(3.3)所示）。其中，由於無資訊性係指在沒有任何資訊的情況下，先驗分配的各個變數皆應給定相同的期望值 $E(\theta_j) = \alpha_j / \alpha$ ，意即 $E(\theta_1) = E(\theta_2) = \dots = E(\theta_{k+1})$ ，並可求得 $\alpha_1 = \alpha_2 = \dots = \alpha_{k+1}$ ；所以使用 Laplace's estimate 的方式，可視同係將狄氏分配之屬性參數 α_i ，($i=1, 2, \dots, k+1$) 設定為1，也相當於 $\alpha_1 = \alpha_2 = \dots = \alpha_{k+1} = 1$ ，即 $D_k(1, 1, \dots, 1; 1)$ 。

再者，假設 $y = (y_1, y_2, \dots, y_{k+1})$ 為訓練資料中，某屬性的 $k+1$ 個可能值分別出現之次數，隨機向量 $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ 為該屬性的第1~ k 個可能值出現之機率，其服從 k 維的狄氏分配 $\Theta \sim D_k(\alpha_1, \alpha_2, \dots, \alpha_k; \alpha_{k+1})$ ，如果 θ_j 為 Θ 的其中一個變數，而

θ_j 在給定資料屬性 y 的條件下，狄氏分配的期望值如第二章的等式 (2.11)：

$E(\theta_j|y) = (y_j + \alpha_j) / \sum_{i=1}^{k+1} (y_i + \alpha_i)$ ，則可將其分解成下列等式：

$$\begin{aligned} E(\theta_j|y) &= \frac{y_j + \alpha_j}{\sum_{i=1}^{k+1} (y_i + \alpha_i)} = \left[\frac{\sum_{i=1}^{k+1} y_i}{\sum_{i=1}^{k+1} (y_i + \alpha_i)} \right] \times \frac{y_j}{\sum_{i=1}^{k+1} y_i} + \left[\frac{\sum_{i=1}^{k+1} \alpha_i}{\sum_{i=1}^{k+1} (y_i + \alpha_i)} \right] \times \frac{\alpha_j}{\sum_{i=1}^{k+1} \alpha_i} \\ &= \left[\frac{\sum_{i=1}^{k+1} y_i}{\sum_{i=1}^{k+1} (y_i + \alpha_i)} \right] \times \frac{y_j}{\sum_{i=1}^{k+1} y_i} + \left[\frac{\sum_{i=1}^{k+1} \alpha_i}{\sum_{i=1}^{k+1} (y_i + \alpha_i)} \right] \times E(\theta_j) \end{aligned} \quad (3.3)$$

在等式 (3.3) 中， $\sum_{i=1}^{k+1} y_i / \sum_{i=1}^{k+1} (y_i + \alpha_i)$ 和 $\sum_{i=1}^{k+1} \alpha_i / \sum_{i=1}^{k+1} (y_i + \alpha_i)$ 分別表示訓練資料與先驗分配的資訊對於事後分配所佔之比重。是故，若欲瞭解狄氏分配參數設定的比重對分類正確率之影響，則應針對參數 α_i 的設定進行分析。所以，Wong (2009) 透過研究測試不同的參數 α_i 對於狄氏分配之影響，其結果顯示當 α_i 大於60 的情況下，分類正確率有逐漸下降的趨勢。因此，本研究在進行狄氏分配之屬性參數設定時，將採用參數範圍在 [1,60] 之間的整數值進行測試，詳細之參數設定步驟如下所述：

步驟一：藉由屬性排序的結果，選擇最優先調整之屬性，針對其參數值 α_i 進行設定，

亦即測試該屬性在各個可能值參數為 $\alpha_1 = \alpha_2 = \dots = \alpha_{k+1}$ 的條件下，將參數設

定為[1,60]之間整數值的分類正確率，並找出能使分類正確率表現最佳之參

數值，將其設定為 α_i^* 。

步驟二：當第一個屬性的參數值在 $\alpha_1 = \alpha_2 = \dots = \alpha_{k+1} = \alpha_i^*$ 的情況下，重複上步驟，找

出第二個排序後屬性之最佳參數值 α_j^* ，並將其設定為第二個屬性的參數

$\alpha_1 = \alpha_2 = \dots = \alpha_{k+1} = \alpha_j^*$ 。

步驟三：依此類推，在第一個屬性參數為 α_i^* 與第二個屬性參數為 α_j^* 的情況下，繼續尋找其他屬性的狄氏分配最佳參數設定值，藉此提升簡易貝氏分類器之分類正確率。

此外，先驗分配在給定資料屬性 $y = (y_1, y_2, \dots, y_{k+1})$ 的情況下，以連續型屬性資料而言，透過不同的離散化方法進行資料前置處理，可能會將同一筆屬性可能值的資料離散化到不同的區間內，使其產生不同的區間值，亦即期望值中的變數 θ_j 會隨之改變；相對之下，落在每個區間內的屬性可能值出現之次數 y_j 也會有所不同。是故，藉此可瞭解使用不同的離散化方法進行資料之前置處理時，會產生不同的先驗分配之期望值 $E(\theta_j|y)$ 。

再者，藉由 3.2 節所提到的 Laplace's estimate 之等式 (3.3) 可發現，簡易貝氏分類器中的概似函數 $p(x_i|C_j)$ 與先驗分配之期望值 $E(\theta_j|y)$ 所代表的意義相同；換言之，此即表示簡易貝氏分類器在使用先驗分配進行機率估計時，可透過先驗分配的期望值 $E(\theta_j|y)$ 做為概似函數 $p(x_i|C_j)$ 之估計值，進行分類預測，亦即說明在給定類別值 C_j 的情況下，屬性 x 的第 i 個可能值出現之機率，會隨著離散化方法的不同而有所變化。是故，以簡易貝氏分類器而言，採用不同的離散化方法進行資料的前置處理，除了產生的區間數不一樣，會影響先驗分配的維度之外，各個區間內的樣本數亦會隨之改變，進而影響簡易貝氏分類器之分類結果。

3.4.2 廣義狄氏分配

由於廣義狄氏分配為狄氏分配的延伸，所以透過第二章的定義2.2 之介紹，如下所示：

$$f(\Theta) = \prod_{j=1}^k \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j-1} (1 - \theta_1 - \dots - \theta_j)^{\beta_j}$$

可瞭解倘若其參數符合 $\beta_j = \alpha_{j+1} + \beta_{j+1}$, ($j=1,2,\dots,k-1$)，則隨機向量 Θ 會退化成狄氏分配，即 $\Theta \sim D_k(\alpha_1, \alpha_2, \dots, \alpha_k; \beta_k)$ 。藉此可說明廣義狄氏分配要退化成狄氏分配的基本條件，為所有變數的正規化變異數皆相等 (Wong, 2009)；換言之，廣義狄氏分配之隨機變數的信賴度不一定要相同，所以並無等信賴需求；再者，廣義狄氏分配除了第一個變數必須與其他變數為負相關以外，其他變數之間可為正相關或負相關，亦即表示其無負相關需求。由此可知，廣義狄氏分配放寬了狄氏分配之條件限制，以致於更能符合實務上的應用。

其次，倘若以屬性先驗分配之參數調整而言，廣義狄氏分配同樣是透過無資訊性的參數設定方式，亦即藉由參數設定使其變數之期望值皆為 $1/(k+1)$ ，($k+1$ 為該屬性可能值之個數)。再者，回顧第二章廣義狄氏分配分配的變數期望值，如等式 (2.14) 所示：

$$E(\theta_j) = E\left[Z_j \prod_{i=1}^{j-1} (1 - Z_i)\right] = \frac{\alpha_j}{\alpha_j + \beta_j} \prod_{i=1}^{j-1} \frac{\beta_i}{\alpha_i + \beta_i}$$

若期望值均為 $1/(k+1)$ ，則藉此可得知 $GD_k(\alpha_1, \alpha_2, \dots, \alpha_k; \beta_1, \beta_2, \dots, \beta_k)$ 中的參數

$\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$ 有下述之關係：

$$\frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{1}{k+1} \quad (3.4)$$

$$\frac{\alpha_2}{\alpha_2 + \beta_2} \times \frac{\beta_1}{\alpha_1 + \beta_1} = \frac{\alpha_2}{\alpha_2 + \beta_2} \times \frac{k}{k+1} = \frac{1}{k+1} \Rightarrow \frac{\alpha_2}{\alpha_2 + \beta_2} = \frac{1}{k} \quad (3.5)$$

$$\frac{\alpha_3}{\alpha_3 + \beta_3} \times \frac{\beta_2}{\alpha_2 + \beta_2} \times \frac{\beta_1}{\alpha_1 + \beta_1} = \frac{\alpha_3}{\alpha_3 + \beta_3} \times \frac{k}{k+1} \times \frac{k-1}{k} = \frac{1}{k+1} \Rightarrow \frac{\alpha_3}{\alpha_3 + \beta_3} = \frac{1}{k-1} \quad (3.6)$$

藉由等式 (3.4)、(3.5) 與 (3.6) 可歸納出，所有參數皆須遵守如下之關係：

$$\frac{\alpha_i}{\alpha_i + \beta_i} = \frac{1}{k - i + 2}, \quad (i = 1, 2, \dots, k) \quad (3.7)$$

是故，只要參數 α_i 已知，即能透過等式 (3.7) 求得 β_i ，並將 α_i 與 β_i 代入廣義狄氏分配的期望值進行計算，如第二章的等式 (2.21) 所示：

$$E(\theta_j | y) = (y_j + \alpha_j) / (\alpha_j + \beta_j + n_j) \prod_{i=1}^{j-1} [(\beta_i + n_{i+1}) / (\alpha_i + \beta_i + n_i)], \quad \text{其中 } n_i = y_i + y_{i+1} + \dots + y_k + y_{k+1},$$

$i = 1, 2, \dots, k+1$ ；之後，再用此期望值做為簡易貝氏分類器 $p(x|C_j)$ 的機率估計值。

由此可知，廣義狄氏分配與狄氏分配在進行參數設定時，所考慮的情況幾乎相同，亦即僅需針對不同的參數值 α_i 進行調整設定，即可瞭解該參數設定值是否可提升簡易貝氏分類器之分類正確率。因此，本研究進行廣義狄氏分配之參數設定步驟茲如下述：

步驟一：依照屬性排序的結果，選擇最優先調整之屬性，然後針對其第一個屬性可能值的參數 α_1 進行設定，測試其為 $[1, 60]$ 之間的整數值，並計算出相對應的 β_1 ，尋找能使分類正確率最佳的 α_1 參數值，將其設定為 α_1^* 。

（備註：此階段該屬性之其他可能值的參數 $\alpha_i, (i = 2, 3, \dots, k)$ 皆設定為 1）

步驟二：在設定完 $\alpha_1 = \alpha_1^*$ 後，則繼續針對該屬性的第二個可能值 α_2 進行參數設定，同樣測試當 α_2 為 $[1, 60]$ 之間的整數值，並計算出相對應的 β_2 ，選擇能使分類正確率最佳的 α_2 參數值，將其設定為 α_2^* 。

步驟三：重複步驟一和步驟二之參數設定方式，設定該屬性其餘可能值的參數 $\alpha_i, (i = 3, 4, \dots, k)$ 之最佳值，且同樣計算出相對應的 β_i^* ，以求得該屬性

$GD_k(\alpha_1, \alpha_2, \dots, \alpha_k; \beta_1, \beta_2, \dots, \beta_k)$ 之分類正確率。

步驟四：當第一個屬性之參數設定為最佳廣義狄氏分配的情況下，重複步驟一至步驟三之方式，依序設定第二個屬性，以及其他屬性的所有可能值參數，藉此找出每個屬性的最佳廣義狄氏分配參數。

綜觀上述，本研究之目的主要係期望藉由不同的離散化方法進行測試比較，分析透過不同的離散化方法進行資料的前置處理，搭配簡易貝氏分類器以狄氏分配或廣義狄氏分配當做先驗分配的情況下，對於分類正確率所產生之影響，並藉此找出是否有特定的離散化方法，適合某種先驗分配，進而能提升簡易貝氏分類器之分類正確率。

3.5 結果評估方式

藉由上述說明，已能瞭解基於本研究目的所設計之研究方法，其詳細之運作流程與步驟。因此，在最後一小節將介紹本研究實際進行資料測試時，所採用之結果評估方式。

本研究採用的結果評估指標將著重於簡易貝氏分類器預測之分類正確率，並透過 f -fold 交互驗證的方式進行分析。其主要係將資料檔內的資料樣本分成 f 個 folds，再以 $f-1$ 個 folds 當成訓練資料，進行分類學習，其餘的 1 個 fold 的資料做為測試資料進行測試，而且各個 fold 皆會依序當作測試資料進行分類預測，亦即表示每個資料檔皆會產生 f 次預測的分類正確率。因此，最後會取其平均值作為分類正確率之評估指標。

此外，為了在進行 f -fold 交互驗證時，能控制每個 fold 中的資料樣本數不會少於 30 筆資料（以符合統計上具有意義之樣本數），進而產生較可靠與穩定之分類結果。是故，本研究將選用 5-folds 交互驗證的方式進行分類結果之評估，同時亦不考

慮使用資料樣本數少於150筆的資料檔進行測試，藉此避免每個 fold 中的資料筆數少於 30 筆，而影響最後簡易貝氏分類器之分類結果評估。



第四章 實證研究

藉由第三章研究方法之闡述，已清楚說明如何使用本研究挑選的四種離散化方法：等寬度、等頻率、比例與最小entropy離散化方法處理連續型態的屬性資料，以及簡易貝氏分類器的先驗分配之屬性參數設定方式。因此本章將延續前述之模式，實證美國加州大學歐文分校 (UCI) 機械學習資料存放站 (Asuncion and Newman, 2007) 上，23個含有連續型態屬性資料的資料檔。並利用程式開發軟體 Visual Basic 6.0 撰寫四種不同的離散化程式，將連續型態屬性資料進行離散化的前置處理。之後再放入具先驗分配的簡易貝氏分類器進行分類預測，最終可透過分析每個資料檔的分類正確率，評估是否有特定的離散化方法特別適用於簡易貝氏分類器的先驗分配為狄氏分配或廣義狄氏分配之情況。是故，本章第一節將先介紹各資料檔的特性；第二節會顯示各資料檔中連續型屬性離散化後的屬性可能值個數；然後，第三節則呈現各資料檔採用四種離散化方法時，在簡易貝氏分類器先驗分配為狄氏分配與廣義狄氏分配的情形下，產生的分類正確率之差異；而第四節主要會針對本章之研究結果進行小結整理。

4.1 資料檔特性

首先，表4.1主要對於本研究所選用的23個資料檔之特性進行介紹，包括各資料檔的資料筆數、連續或離散型態的屬性個數，以及類別值個數等。其次，承上一章節所述，本研究將採用 5-folds 交互驗證的方式進行分類結果之評估，為了避免每個 fold 中的資料筆數少於 30 筆，僅會選擇資料樣本數大於150筆的資料檔進行測試。再者，由於本研究目的主要為，評估連續型態屬性資料經過不同離散化方法之前置處理後，對於具先驗分配的簡易貝氏分類器之分類正確率的影響性，所以進行實證的資料檔，必須選擇包含連續型態屬性的資料檔。其中，有12個資料檔為純連續型態屬性的資料檔，另外11個資料檔則為含有連續型態與離散型態的混合型資料檔。同時，為了不讓資料檔中原先存在的離散型態屬性資料影響最後之分類結果，本研究於進行分

類預測前，會先刪除混和型資料檔中，原本已有的離散型態屬性資料（如表4.1標示括弧的屬性個數），而僅針對連續型態的屬性資料進行測試。雖然可能會造成原先資料檔的分類正確率下降，但是如此卻更能突顯不同離散化方法對於最終分類正確率產生之影響性。

表 4.1 資料檔特性

編號	資料檔名稱	資料筆數	連續型屬性 個數	離散型屬性 個數	類別值個數
1	annealing	898	6	(32)	5
2	blood	748	4	(0)	2
3	breast2	569	30	(0)	2
4	cleve	303	6	(7)	2
5	cmc	1473	2	(7)	3
6	crx	690	6	(9)	2
7	ecoli	336	5	(2)	8
8	flags	194	10	(18)	8
9	german	1000	7	(13)	2
10	glass	214	9	(0)	6
11	hepatitis	155	6	(13)	2
12	ionosphere	351	33	(0)	2
13	iris	150	4	(0)	3
14	liver disorder	345	6	(0)	2
15	mfeat	2000	3	(3)	10
16	newthyroid	215	5	(0)	3
17	pima	768	8	(0)	2
18	sick-euthyroid	3163	7	(18)	2
19	vehicle	843	18	(0)	4
20	vowel	990	10	(2)	11
21	waveform	5000	21	(0)	3
22	wine	178	13	(0)	3
23	yeast	1484	8	(0)	10

4.2 離散化之屬性可能值個數

透過第三章的介紹，已瞭解本研究所採用的四種離散化方法之運作方式。其中，以三種非監督式離散化方法而言，等寬度與等頻率離散化方法必須先決定離散化後的區間個數，所以本研究將採用一般常見的參數 $t=10$ 進行離散化，亦即將原本連續型態的屬性資料，離散化成 10 個屬性可能值。由此可知，使用等寬度與等頻率離散化後，在每個資料檔中產生的屬性可能值個數皆相同。再者，於比例離散化方法的部分，其運作原理主要為權衡區間數與區間內的樣本數，因而離散化後的屬性可能值個數，會隨著資料檔的資料筆數多寡有所增減。在此也透過表 4.2 彙整使用上述三種非監督式離散化方法所產生之屬性可能值個數。

另一方面，在監督式離散化方法的應用，由於必須考慮類別值分佈的情形。因此採用最小 entropy 離散化方法進行資料處理時，必須先依類別值的分佈情形，計算候選切割點之 entropy 值，才能決定最終的切割點及離散化後的屬性可能值個數。是故，使用最小 entropy 離散化方法時，可能會造成即使在同一資料檔中，每個連續型態屬性進行離散化後，產生之屬性可能值個數有所不同的情形，茲如表 4.3 所示。

然而，從表 4.3 的結果可發現，在大多數的資料檔中，使用最小 entropy 離散化方法相較於另外三種非監督式的離散化方法，產生之屬性可能值個數有偏少的情形。甚至在某些資料中，最小 entropy 離散化後產生的屬性可能值個數僅有 1 或 2 個可能值，如：資料檔編號為 2、4、6、8、9 和 14 的資料檔。

表 4.2 使用三種非監督式離散化方法產生之屬性可能值個數

編號	資料檔名稱	等寬度	等頻率	比 例
1	annealing	10	10	29
2	blood	10	10	27
3	breast2	10	10	23
4	cleve	10	10	17
5	cmc	10	10	38
6	crx	10	10	26
7	ecoli	10	10	18
8	flags	10	10	13
9	german	10	10	31
10	glass	10	10	14
11	hepatitis	10	10	12
12	ionosphere	10	10	18
13	iris	10	10	12
14	liver disorder	10	10	18
15	mfeat	10	10	44
16	newthyroid	10	10	14
17	pima	10	10	27
18	sick-euthyroid	10	10	56
19	vehicle	10	10	29
20	vowel	10	10	31
21	waveform	10	10	70
22	wine	10	10	13
23	yeast	10	10	38

表 4.3 使用最小 entropy 離散化方法產生之屬性可能值個數

屬性	資料檔編號																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	2	2	4	2	3	2	3	2	2	3	1	2	3	1	9	5	2	3	3	9	1	3	3
2	3	2	2	1	3	2	2	2	2	2	2	4	2	1	8	4	4	2	4	9	4	3	3
3	4	2	4	1		2	3	1	1	2	1	5	3	1	13	3	1	3	5	3	5	2	3
4	9	1	4	2		2	3	1	1	3	1	4	3	1	3	2	1	4	3	4	5	2	3
5	4		2	2		2	2	1	1	1	3	6		2	7	4	3	3	3	5	7	2	2
6	2		3	2		2		1	1	4	2	3		1	6		2	1	4	2	7	3	2
7			4					2	1	4		5					2	1	5	3	6	4	1
8			4					2		2		5					2		5	3	6	2	4
9			3					1		1		4							5	3	5	2	
10			1					1				5							5	2	6	3	
11			4									5							4		5	4	
12			1									6							7		6	3	
13			4									4							4		6	4	
14			4									5							2		6		
15			1									5							2		5		
16			3									6							1		6		
17			3									3							5		6		
18			3									6							2		5		
19			2									3									4		
20			2									5									3		
21			4									5									1		
22			3									5											
23			4									3											
24			4									5											
25			2									3											
26			4									3											
27			3									3											
28			4									5											
29			3									3											
30			2									5											
31												3											
32												5											
33												5											

4.3 離散化之先驗分配測試結果

本小節主要呈現連續型態屬性資料經由上述四種離散化方法處理後，於簡易貝氏分類器先驗分配為最佳狄氏分配或最佳廣義狄氏分配之分類正確率測試結果，並藉此分析不同的離散化方法是否有其適合的先驗分配。首先，表 4.4 為四種離散化方法在使用簡易貝氏分類器時，僅透過拉普拉斯估計，亦即將先驗分配之參數皆設定為 1 的情況下，進行分類預測之結果。

表 4.4 離散化後使用拉普拉斯估計的簡易貝氏分類器之分類正確率

編號	資料檔名稱	等寬度	等頻率	比 例	entropy
1	annealing	83.41%	78.39%	82.37%	87.73%
2	blood	76.90%	75.65%	76.66%	75.32%
3	breast2	93.98%	93.85%	93.65%	95.53%
4	cleve	73.58%	72.63%	71.75%	75.94%
5	cmc	50.10%	49.72%	48.77%	49.87%
6	crx	70.48%	78.08%	77.94%	78.78%
7	ecoli	81.58%	83.61%	81.14%	84.86%
8	flags	37.52%	39.85%	41.24%	35.28%
9	german	70.49%	69.47%	69.13%	69.92%
10	glass	59.35%	71.32%	69.01%	74.88%
11	hepatitis	83.38%	80.54%	80.41%	83.06%
12	ionosphere	91.44%	89.80%	89.78%	90.93%
13	iris	93.19%	92.05%	91.08%	93.64%
14	liver disorder	63.27%	60.79%	60.09%	63.36%
15	mfeat	44.66%	52.31%	48.67%	51.22%
16	newthyroid	91.42%	96.94%	97.65%	96.54%
17	pima	76.10%	73.92%	74.66%	77.38%
18	sick-euthyroid	94.17%	94.31%	94.77%	95.64%
19	vehicle	61.24%	60.64%	62.08%	62.28%
20	vowel	67.44%	67.09%	59.71%	68.42%
21	waveform	80.42%	80.91%	80.09%	81.60%
22	wine	96.14%	97.66%	96.07%	98.77%
23	yeast	58.17%	56.81%	52.44%	58.93%
分類正確率平均值		73.84%	74.62%	73.88%	76.08%
正確率表現最佳之資料檔個數		5	1	2	15

在表 4.4 中，粗體字部分顯示，以該資料檔的測試結果而言，其為分類正確率表現最佳之離散化方法。由此可知，監督式的最小 entropy 離散化方法之分類正確率普遍優於其他三種非監督式的離散化方法。其在 23 個資料檔中，有 15 個資料檔的分類正確率優於其他離散化方法，且平均值亦勝過其他離散化方法。後續，本研究則進一步測試使用最佳狄氏分配或最佳廣義狄氏分配為先驗分配時，簡易貝氏分類器之分類正確率，並將結果呈現於表 4.5 及表 4.6。

表 4.5 離散化後於最佳狄氏分配之分類正確率

編號	資料檔名稱	等寬度	等頻率	比 例	entropy
1	annealing	83.75%	78.43%	82.37%	88.00%
2	blood	77.45%	76.02%	77.91%	75.32%
3	breast2	94.56%	95.28%	95.29%	96.09%
4	cleve	75.91%	75.25%	72.70%	77.38%
5	cmc	51.28%	50.44%	49.31%	50.24%
6	crx	70.84%	79.38%	79.33%	79.07%
7	ecoli	82.50%	83.61%	82.07%	85.82%
8	flags	39.42%	41.79%	41.60%	36.94%
9	german	71.06%	70.45%	70.21%	69.92%
10	glass	63.52%	73.38%	71.40%	75.94%
11	hepatitis	84.09%	86.04%	81.73%	84.13%
12	ionosphere	92.20%	91.21%	90.43%	91.58%
13	iris	95.31%	92.67%	92.67%	96.70%
14	liver disorder	64.66%	64.33%	62.44%	63.36%
15	mfeat	45.10%	52.70%	50.00%	51.66%
16	newthyroid	92.22%	97.46%	98.82%	96.57%
17	pima	76.71%	75.82%	76.31%	78.81%
18	sick-euthyroid	94.65%	94.71%	95.70%	95.91%
19	vehicle	63.08%	62.05%	64.68%	63.17%
20	vowel	68.97%	67.65%	62.10%	69.32%
21	waveform	80.98%	81.38%	81.01%	81.77%
22	wine	96.68%	98.22%	97.25%	98.77%
23	yeast	58.53%	57.68%	52.97%	59.07%
分類正確率平均值		74.93%	75.91%	75.14%	76.76%
正確率表現最佳之資料檔個數		4	4	3	12

表 4.6 離散化後於最佳廣義狄氏分配之分類正確率

編號	資料檔名稱	等寬度	等頻率	比 例	entropy
1	annealing	84.32%	80.14%	82.47%	88.16%
2	blood	78.04%	77.67%	79.08%	75.32%
3	breast2	94.70%	95.06%	94.55%	96.03%
4	cleve	76.86%	75.39%	74.89%	77.38%
5	cmc	52.29%	51.10%	53.09%	50.58%
6	crx	71.97%	80.36%	81.26%	79.07%
7	ecoli	83.08%	85.00%	84.19%	86.02%
8	flags	41.30%	44.03%	46.34%	36.94%
9	german	71.66%	70.45%	71.17%	69.92%
10	glass	64.11%	76.57%	73.99%	78.20%
11	hepatitis	85.04%	85.15%	85.13%	84.93%
12	ionosphere	91.70%	90.74%	90.37%	91.58%
13	iris	96.68%	95.10%	93.92%	96.70%
14	liver disorder	67.33%	67.49%	66.56%	63.36%
15	mfeat	45.90%	54.44%	52.13%	54.22%
16	newthyroid	92.35%	98.64%	99.19%	97.83%
17	pima	77.16%	75.77%	77.69%	78.48%
18	sick-euthyroid	94.87%	94.65%	95.94%	95.91%
19	vehicle	62.82%	62.31%	66.01%	64.17%
20	vowel	70.10%	70.59%	65.76%	71.08%
21	waveform	81.67%	81.90%	81.94%	82.11%
22	wine	98.22%	98.22%	97.37%	98.77%
23	yeast	59.81%	59.24%	56.38%	59.13%
分類正確率平均值		75.74%	76.96%	76.93%	77.21%
正確率表現最佳之資料檔個數		3	3	7	10

觀察表 4.5 可得知，在使用最佳狄氏分配的情況下，四種離散化方法的表現仍以監督式的最小 entropy 離散化方法表現較好，另外三種非監督式的離散化方法表現較差。然而，在表 4.6 中卻發現，使用比例離散化搭配最佳廣義狄氏分配的分類正確率平均值，雖然並未優於最小 entropy 離散化方法，但是其在各資料檔中的表現，相較於表 4.5，則由原本僅於 3 個資料檔中優於其他離散化方法，提升至 7 個資料檔表現最佳。是故，為能更清楚地呈現此四種離散化方法對於簡易貝氏分類器在搭配不同先

驗分配的模式下所產生之差異性，本研究亦彙整四種離散化方法在搭配最佳狄氏分配時，與原本僅使用拉普拉斯估計的簡易貝氏分類器，其分類正確率的差異值，如表 4.7 所示；而表 4.8 則呈現這些離散化方法在採用最佳廣義狄氏分配為先驗分配時，相對於使用最佳狄氏分配的情況下，分類正確率產生之差異值。

表 4.7 離散化後使用最佳狄氏分配相較於僅加入拉普拉斯估計的簡易貝氏分類器之分類正確率差異值

編號	資料檔名稱	等寬度	等頻率	比 例	entropy
1	annealing	0.34%	0.04%	0.00%	0.27%
2	blood	0.55%	0.37%	1.25%	0.00%
3	breast2	0.58%	1.43%	1.64%	0.56%
4	cleve	2.33%	2.62%	0.95%	1.44%
5	cmc	1.18%	0.72%	0.54%	0.37%
6	crx	0.36%	1.30%	1.39%	0.29%
7	ecoli	0.92%	0.00%	0.93%	0.96%
8	flags	1.90%	1.94%	0.36%	1.66%
9	german	0.57%	0.98%	1.08%	0.00%
10	glass	4.17%	2.06%	2.39%	1.06%
11	hepatitis	0.71%	5.50%	1.32%	1.07%
12	ionosphere	0.76%	1.41%	0.65%	0.65%
13	iris	2.12%	0.62%	1.59%	3.06%
14	liver disorder	1.39%	3.54%	2.35%	0.00%
15	mfeat	0.44%	0.39%	1.33%	0.44%
16	newthyroid	0.80%	0.52%	1.17%	0.03%
17	pima	0.61%	1.90%	1.65%	1.43%
18	sick-euthyroid	0.48%	0.40%	0.93%	0.27%
19	vehicle	1.84%	1.41%	2.60%	0.89%
20	vowel	1.53%	0.56%	2.39%	0.90%
21	waveform	0.56%	0.47%	0.92%	0.17%
22	wine	0.54%	0.56%	1.18%	0.00%
23	yeast	0.36%	0.87%	0.53%	0.14%
分類正確率平均值		1.09%	1.29%	1.27%	0.68%
正確率表現最佳之資料檔個數		3	7	11	2

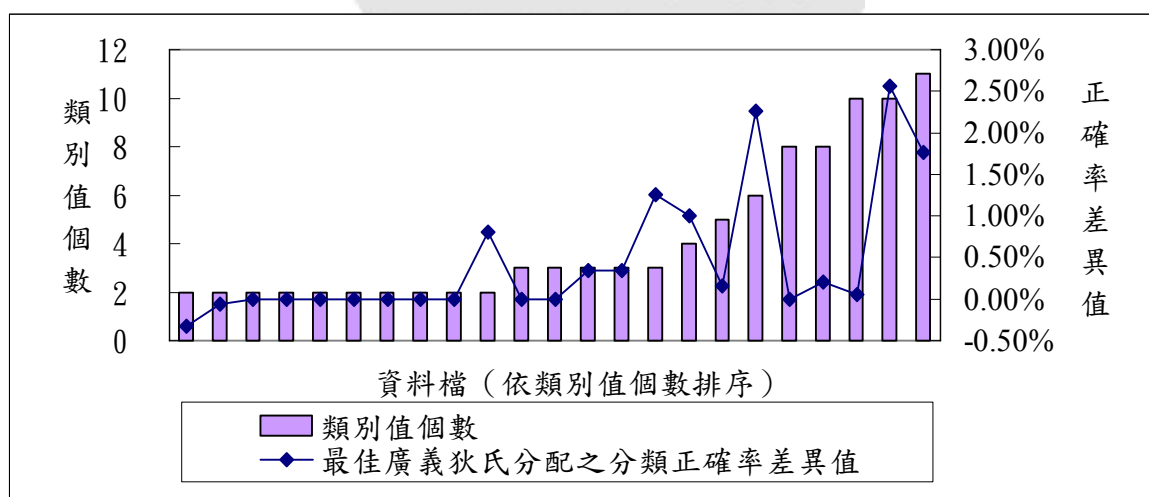
表 4.8 離散化後使用最佳廣義狄氏分配相較於最佳狄氏分配之分類正確率差異值

編號	資料檔名稱	等寬度	等頻率	比 例	entropy
1	annealing	0.57%	1.71%	0.10%	0.16%
2	blood	0.59%	1.65%	1.17%	0.00%
3	breast2	0.14%	-0.22%	-0.74%	-0.06%
4	cleve	0.95%	0.14%	2.19%	0.00%
5	cmc	1.01%	0.66%	3.78%	0.34%
6	crx	1.13%	0.98%	1.93%	0.00%
7	ecoli	0.58%	1.39%	2.12%	0.20%
8	flags	1.88%	2.24%	4.74%	0.00%
9	german	0.60%	0.00%	0.96%	0.00%
10	glass	0.59%	3.19%	2.59%	2.26%
11	hepatitis	0.95%	-0.89%	3.40%	0.80%
12	ionosphere	-0.50%	-0.47%	-0.06%	0.00%
13	iris	1.37%	2.43%	1.25%	0.00%
14	liver disorder	2.67%	3.16%	4.12%	0.00%
15	mfeat	0.80%	1.74%	2.13%	2.56%
16	newthyroid	0.13%	1.18%	0.37%	1.26%
17	pima	0.45%	-0.05%	1.38%	-0.33%
18	sick-euthyroid	0.22%	-0.06%	0.24%	0.00%
19	vehicle	-0.26%	0.26%	1.33%	1.00%
20	vowel	1.13%	2.94%	3.66%	1.76%
21	waveform	0.69%	0.52%	0.93%	0.34%
22	wine	1.54%	0.00%	0.12%	0.00%
23	yeast	1.28%	1.56%	3.41%	0.06%
分類正確率平均值		0.80%	1.05%	1.79%	0.45%
正確率表現最佳之資料檔個數		2	3	15	3

藉由表 4.7 與表 4.8 的結果可發現，比例離散化方法不僅在搭配先驗分配為最佳狄氏分配時，有 11 個資料檔相較於其他三種離散化方法，能提升更多的分類正確率；甚至在使用最佳廣義狄氏分配的情況下，改善分類正確率的幅度，亦優於另外三種離散化方法。是故，本研究建議在選用比例離散化方法進行連續型態資料處理時，倘若能配合先驗分配為最佳廣義狄氏分配之模式，對於分類正確率的提升會更有幫助。

4.4 小結

綜觀以上實驗數據可發現，以監督式的最小 entropy 離散化方法而言，使用最佳狄氏分配的分類正確率提升幅度，雖然不及其他離散化方法，但仍有所提升。然而，其在搭配先驗分配為最佳廣義狄氏分配時，相較於最佳狄氏分配的模式，卻有 12 個資料檔，無法提升分類正確率的表現。因此，本研究藉由分析上述表 4.1、表 4.3 與表 4.8 的關連性，並且回顧第二章廣義狄氏分配的參數調整原理後，推測最小 entropy 離散化方法會出現此現象之原因，最主要為最小 entropy 離散化方法相較於另外三種離散化方法產生的屬性可能值個數較少，以致於當離散化後最大屬性可能值個數少於 3 個時，使用最佳廣義狄氏分配的分類表現會與最佳狄氏分配的分類結果相同。由此得知，使用最小 entropy 離散化時，可藉其離散化後的最大屬性可能值個數，做為是否需要搭配最佳廣義狄氏分配，以提升分類正確率的依據之一。同時，本研究亦發現，由於最小 entropy 離散化方法是一種監督式的離散化方法，所以即使其分類正確率提升之多寡，與資料檔的類別值個數，並無呈現正相關的影響，但仍可藉實驗結果發現，在大部分的情況下，當資料檔的類別值個數少於 3 個類別值時，搭配最佳廣義狄氏分配的分類正確率會與最佳狄氏分配相同，茲如圖 4.1 所示。



是故，本研究歸納出採用最小 entropy 離散化方法時，較適合加入最佳廣義狄氏分配以提升分類正確率的情況，有兩項要素分別為：一、離散化後產生的屬性可能值個數，需有大於 2 個可能值的屬性；二、資料檔的類別值個數要大於 2 個類別值之情況。由於大部分的資料檔必須同時符合此兩項要素，才能讓最小 entropy 離散化方法在搭配最佳廣義狄氏分配時，有提升分類正確率的空間；換言之，一般無法具備此兩項要素的資料檔，則僅需使用最佳狄氏分配即可。再者，本研究也將 23 個資料檔分為兩類群組，呈現其在搭配不同先驗分配之分類正確率差異值。第一類群組：資料檔的類別值個數，以及最小 entropy 離散化後之最大屬性可能值個數均大於 2，如圖 4.2 所示；第二類群組：除了第一類群組以外的其他資料檔，如圖 4.3 所示。

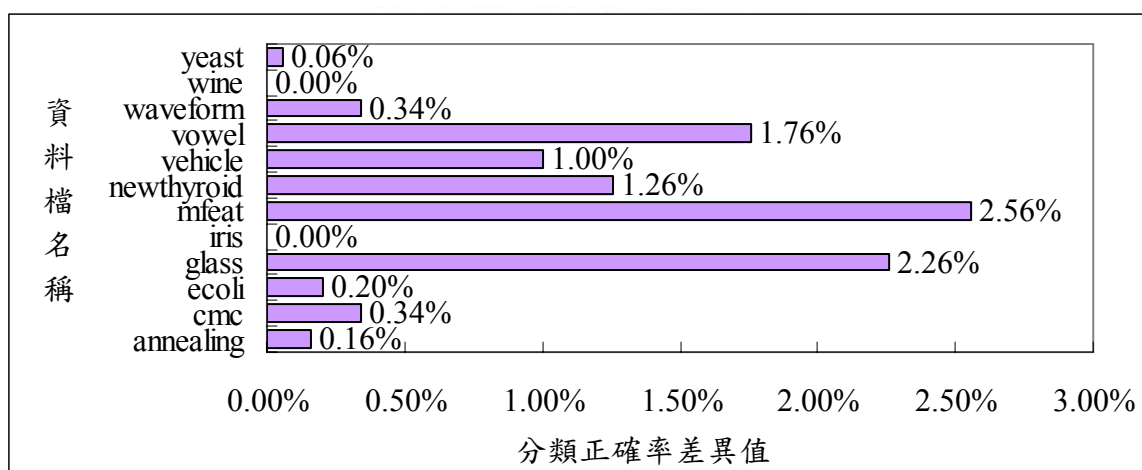


圖 4.2 第一類群組資料檔於最佳廣義狄氏分配和最佳狄氏分配之分類正確率差異值

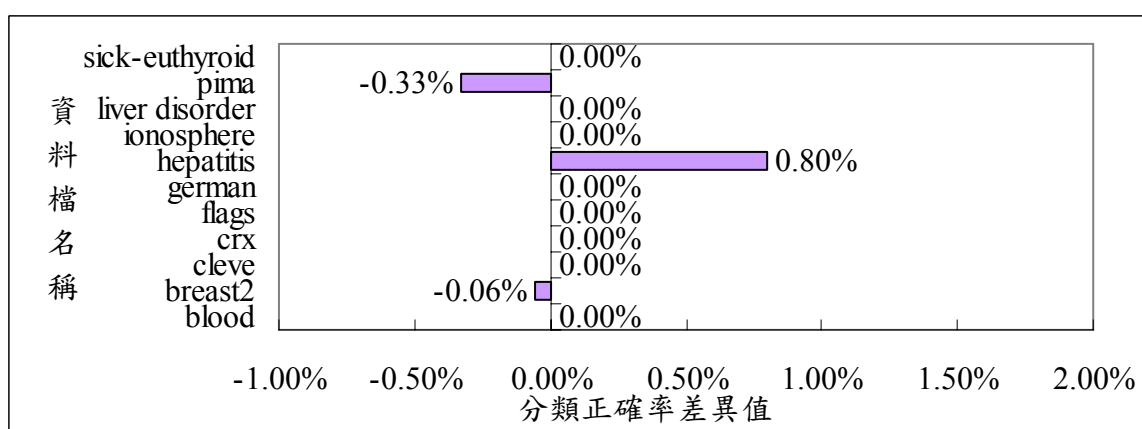


圖 4.3 第二類群組資料檔於最佳廣義狄氏分配和最佳狄氏分配之分類正確率差異值

藉由圖 4.2 與 4.3 可發現，在第一類群組中，最佳廣義狄氏分配相較於最佳狄氏分配的分類正確率會有較多的提升；反之，在第二類群組中，最佳狄氏分配與最佳廣義狄氏分配的分類正確率幾乎沒有差異性。如此亦符合上述本研究對於最小 entropy 離散化方法需搭配最佳廣義狄氏分配之建議。其次，本研究也將四種離散化方法搭配不同先驗分配模式下，23 個資料檔的分類正確率平均值進行彙整，如圖 4.4 所示。

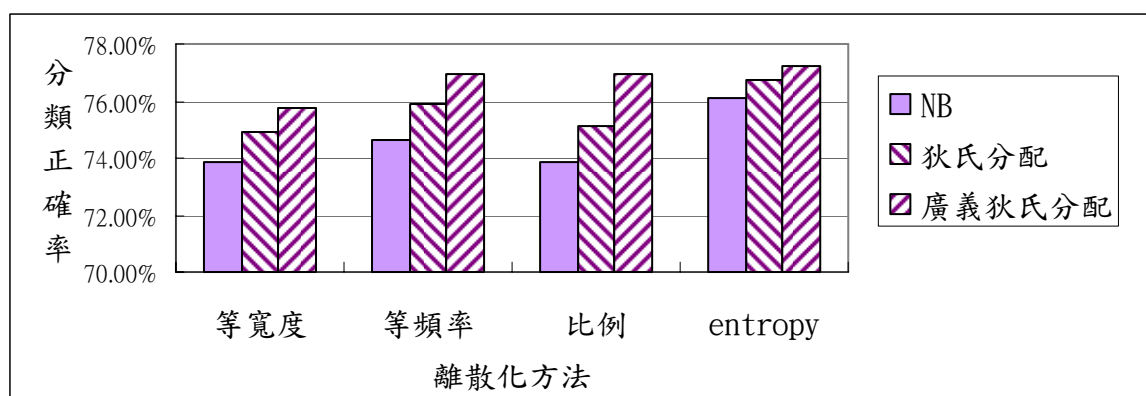


圖 4.4 四種離散化方法搭配不同先驗分配模式之分類正確率平均值

觀察圖 4.4 可得知，在各種模式下，皆以最小 entropy 離散化方法的分類正確率平均值表現較佳，但是其在搭配最佳廣義狄氏分配時，正確率的提升幅度最小。所以，本研究除了歸納上述兩項要素，為造成最小 entropy 離散化方法在搭配最佳廣義狄氏分配時，相較於最佳狄氏分配之分類正確率無法提升的主因外，本研究也進一步推測其他的影響因子，可能為使用最小 entropy 離散化方法的分類正確率已高達某一程度，因此對於分類正確率的提升空間有所限制。亦或原先的資料檔內，即已存在某些干擾屬性，使其在實證結果中，反而有少數資料檔搭配最佳廣義狄氏分配的分類正確率低於最佳狄氏分配之情況發生。

相較之下，其他三種非監督式的離散化方法：等寬度、等頻率與比例離散化，在搭配先驗分配為最佳廣義狄氏分配的模式下，對於分類正確率的提升較有助益。而且以比例離散化提升的幅度較為顯著。因此，本研究建議在選用此三種非監督式離散化方法時，可搭配先驗分配為最佳廣義狄氏分配之模式，藉此提升其分類正確率。

第五章 結論與建議

一般而言，使用簡易貝氏分類器在進行連續型態屬性資料預測前，會選用運算複雜度較低的離散化方法作為資料的前置處理工具。主要原因為不同離散化方法對於分類正確率的影響並無顯著差異。再者，大部分使用簡易貝氏分類器時，亦會假設各個屬性值出現之機率服從某先驗分配，以提升分類正確率。因此，本研究期望透過實證研究結果，評估上述兩種改善簡易貝氏分類器之分類正確率的方法，是否有其較適合的搭配模式，能更進一步提升分類正確率的表現。

然而，藉由第四章的實驗結果可得知，本研究選用的四種離散化方法：等寬度、等頻率、比例及最小 entropy 離散化方法，在搭配不同先驗分配的情況下，誠如之前文獻所述，雖然對於各資料檔的分類正確率影響差異並不大，但是在大部分資料檔中，仍發現最小 entropy 離散化方法的分類正確率普遍高於其他離散化方法；反之，由於最小 entropy 離散化方法的計算複雜度，相對其他三種非監督式的離散化方法複雜許多。所以本研究建議，倘若在不考慮運算複雜度的情況下，可採用最小 entropy 離散化方法，否則依然可選擇其他計算較簡便的離散化方法。

其次，本研究亦發現，最小 entropy 離散化方法在搭配先驗分配為最佳廣義狄氏分配時，相較於最佳狄氏分配的分類正確率提升幅度，不如另外三種非監督式離散化方法顯著。因此，本研究推測其原因可能為，由於某些資料檔本身的類別值個數偏少，以及使用最小 entropy 離散化後，產生的屬性可能值個數也偏少，進而造成使用最小 entropy 離散化方法搭配最佳廣義狄氏分配時，其分類正確率的提升空間不大。是故，本研究建議僅在資料檔類別值個數及離散化後最大屬性可能值個數兩者皆偏多的情況下，才考慮使用最小 entropy 離散化方法搭配先驗分配為最佳廣義狄氏之模式，否則僅需採用最佳狄氏分配即可。

相較之下，另他三種非監督式的離散化方法，在大部分的情況，搭配先驗分配為最佳廣義狄氏分配的模式時，對分類正確率的提升較有助益。其中，更以比例離散化方法配合最佳廣義狄氏分配的模式，對於簡易貝氏分類器之正確率的提升較為顯著。但是，在另一方面亦由於比例離散化方法所產生的屬性可能值個數，會隨著資料檔的筆數增加而遞增，以致於在大型資料檔中，若搭配最佳廣義狄氏分配的模式時，進行參數調整的運算複雜度也隨之增加。因此，在大型資料檔中，較不建議使用比例離散化方法搭配最佳廣義狄氏分配之模式。

除此之外，本研究於決定先驗分配參數調整之順序時，由於僅採用簡易貝氏屬性挑選法做為依據，所以先驗分配的參數調整順序已被固定，如此可能限制了原本參數調整的空間。未來倘若能搭配其他不同的屬性挑選方式，作為參數調整順序的組合，或是當進行分類預測時，再配合上屬性挑選的方法進行測試，藉此排除資料檔中某些干擾或冗餘屬性，相信對於簡易貝氏分類器的分類正確率之提升會更有助益。

參考文獻

中文

張良豪 (2009)，利用貝氏屬性挑選法與先驗分配提升簡易貝氏分類器之效能，國立成功大學工業與資訊管理學系碩士班碩士論文。

英文

Addin, O., Sapuan, S. M., Mahdi, E., and Othman, M. (2007). A naïve Bayes classifier for damage detection in engineering materials. *Materials and Design*, 28(8), 2379-2386.

Aitchison, J. (1985). A general class of distributions on the simplex. *Journal of the Royal Statistical Society Series B*, 47(1), 136-146.

Asuncion, A. and Newman, D.J. (2007). UCI machine learning repository <http://www.ics.uci.edu/~mllearn/MLRepository.html>. Irvine, CA: University of California, School of Information and Computer Science.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural-net learning. *IEEE Transactions on Neural Networks*, 5(4), 537-550.

Bier, V. M. and Yi, W. (1995). A Bayesian method for analyzing dependencies in precursor data. *International Journal of Forecasting*, 11(1), 25-41.

Biesiada, J., Duch, W., Kachel, A., Maczka, K., and Palucha, S. (2005). Feature ranking methods based on information entropy with Parzen window. *International Conference on Research in Electrotechnology and Applied Informatics*, 109-118, Katowice, Poland.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. *Proceedings of the 5th European Working Session on Learning on Machine Learning*, 164-178, Porto, Portugal.

Cestnik, B. and Bratko, I. (1991). On estimating probabilities in tree pruning. *Proceedings*

of the 5th European Working Session on Learning on Machine Learning, 138-150, Porto, Portugal.

Chen, K., Kurgan, L., and Rahbari, M. (2007). Prediction of protein crystallization using collocation of amino acid pairs. *Biochemical and Biophysical Research Communications*, 355(3), 764-769.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64, 194-206.

Domingos, P. and Plazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero one loss. *Machine Learning*, 29, 103-130.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Proceedings of the 12th International Conference on Machine Learning*, 194-202, San Francisco, Morgan Kaufmann.

Elomaa, T. and Rousu, J. (2003). On decision boundaries of naïve Bayes in continuous domains. *Proceedings of the 7th European Conference on Knowledge Discovery in Databases*, 2838, 144-155, Berlin, Heidelberg.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022-1027, Chambery, France.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11, 63-91.

Hsu, C. N., Huang, H. J., and Wong, T. T. (2000). Why discretization works for naïve Bayesian classifiers. *Proceedings of the 17th International Conference on Machine Learning*, 309-406.

Hsu, C. N., Huang, H. J., and Wong, T. T. (2003). Implications of the Dirichlet assumption

- for discretization of continuous attributes in naïve Bayesian classifiers. *Machine Learning*, 53, 235-263.
- Huang, J. J., Cai, Y. Z., and Xu, X. M. (2008). A parameterless feature ranking algorithm based on MI. *Neurocomputing*, 71, 1656-1668.
- John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Proceedings of the 11th International Conference on Machine Learning*, 121-129.
- Kerber, R. (1992). Chimerge: Discretization for numeric attributes. *Proceedings of the 10th International Conference on Artificial Intelligence*, 123-128.
- Keren, D. (2003). Recognizing image style and activities in video using local features and naïve Bayes. *Pattern Recognition Letters*, 24, 2913-2922.
- Kohavi, R. and Sahami, M. (1996). Error-based and entropy-based discretization of continuous features. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 114-119.
- Kwak, N. and Choi, C. H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 13(1), 143-159.
- Langley, P. and Sage, S. (1994). Induction of selective Bayesian classifiers. *Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence*, 399-406.
- Liang, J., Yang, S., and Winstanley, A. (2008). Invariant optimal feature selection: a distance discriminant and feature ranking based solution. *Pattern Recognition*, 41, 1429-1439.
- Lopez de Mantaras, R. (1991). A distance-based attribute selecting measure for decision tree induction. *Machine Learning*, 6, 81-92.
- Menzies, T., Greenwald, J., and Frank, A. (2007). Data mining static code attributes to learn defect predictors. *IEEE Transactions on Software Engineering*, 33(1), 2-13.

- Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226-1238.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes in C*. Cambridge University Press, Cambridge.
- Robnik, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53, 23-69.
- Setiono, R. and Liu, H. (1995). Chi2: feature selection and discretization of numeric attributes. *Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence*, 388-3913.
- Spector, P. (1990). *An Introduction to S and S-PLUS*. Duxbury Press, Belmont.
- Sridhar, D. V., Bartlett, E. B., and Seagrave, R. C. (1998). Information theoretic subset selection. *Computers in Chemical Engineering*, 22, 613-626.
- Terribilini, M., Sander, J. D., Lee, J. H., Zaback, P., Jernigan, R. L., Honavar, V., and Dobbs, D. (2007). *Nucleic Acids Research*, 35, 578-584.
- Turhan, B. and Bener, A. (2009). Analysis of naive Bayes' assumptions on software fault data: an empirical study. *Data and Knowledge Engineering*, 68, 278-290.
- Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York.
- Wong, T. T. (1998). Generalized Dirchlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, 97, 165-181.
- Wong, T. T. (2009). Alternative prior assumptions for improving the performance of naive Bayesian classifiers. *Data Mining and Knowledge Discovery*, 18, 183-213.
- Yang, Y. and Webb, G. I. (2002a). Non-disjoint discretization for naïve Bayes classifiers. *Proceedings of the 19th International Conference on Machine Learning*, 666-673.

- Yang, Y. and Webb, G. I. (2002b). A comparative study of discretization methods for naïve Bayes classifiers. *Proceedings of the Pacific Rim Knowledge Acquisition Workshop*, 159-173.
- Yang, Y. and Webb, G. I. (2009). Discretization for naïve Bayes learning: managing discretization bias and variance. *Machine Learning*, 74, 39-74.
- Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C., and Showe, M. K. (2006). Combining multi-species genomic data for microRNA identification using a naïve Bayes classifier. *Bioinformatics*, 22(11), 1325-1334.

