翁慈宗、楊乃玉(2018)，『潛在狄氏配置簡易貝氏分類器』，
中華民國資訊管理學報，第二十五卷，第一期，頁 54-75。

潛在狄氏配置簡易貝氏分類器

翁慈宗
國立成功大學資訊管理研究所

楊乃玉*
國立成功大學資訊管理研究所

摘要

在資料探勘的分類演算法中，簡易貝氏分類器具有運算效率高且分類正確率佳之優勢，已廣泛應用在許多實務上。由於簡易貝氏分類器係以計算條件機率之方式進行分類預測，所以大部分會加入先驗分配之機制提升分類正確率，且一般係採用狄氏分配或廣義狄氏分配當成先驗分配進行資料屬性可能值機率之參數調整。然而，過去研究對於資料檔中類別值的機率卻未有加入先驗分配之機制，如此可能導致分類正確率之提升有所限制。所以本研究提出潛在狄氏配置簡易貝氏分類器（Latent Dirichlet Allocation Naïve Bayes; LDANB），透過潛在狄氏配置模型，將先驗分配機制加入類別值之機率，進行參數調整，使資料更接近原本真實概念，並藉由 UCI 的 20 個資料檔進行實證研究測試。研究結果顯示使用潛在狄氏配置模型之簡易貝氏分類器優於僅將屬性可能值加入先驗分配之情況，且廣義狄氏分配優於狄氏分配。惟廣義狄氏分配之運算複雜度較高，是故，建議潛在狄氏配置簡易貝氏分類器之先驗分配模式採用屬性可能值為廣義狄氏分配，結合類別值為狄氏分配機制，更能在有限的運算成本下，提升分類正確率。

關鍵字：簡易貝氏分類器、狄氏分配、廣義狄氏分配、潛在狄氏配置

* 本文通訊作者。電子郵件信箱：cherry@mail.hwai.edu.tw
2016/11/23 投稿；2017/07/13 修訂；2017/11/17 接受

Wong,T.T. and Yang,N.Y.(2018),' Latent dirichlet allocation naïve bayes',
Journal of Information Management, Vol. 25, No. 1,pp. 54-75.

Latent Dirichlet Allocation Naïve Bayes

Tzu Tsung Wong
Institute of Information Management, National Cheng Kung University

Nai Yu Yang*
Institute of Information Management, National Cheng Kung University

Abstract

Purpose—Naïve Bayesian classifier is widely employed for classification tasks, because of its computational efficiency and competitive accuracy. The prior distributions of attributes in the naïve Bayesian classifier are implicitly or explicitly assumed to follow either Dirichlet or generalized Dirichlet distributions. However, none of previous studies apply the prior distributions on classes in the naïve Bayesian classifier. The aim of this study is to develop a model based on LDA, called LDANB, that introduces prior distributions for both attributes and classes in the naïve Bayesian classifier so that the performance of this classification method can be improved.

Design/methodology/approach—The prior distributions of both attributes and classes in the naïve Bayesian classifier can be Laplace's estimate, Dirichlet distribution, or generalized Dirichlet distribution. Nine combinations of priors for attributes and classes are explored to investigate their impact on the performance of the naïve Bayesian classifier.

Findings—The experimental results on 20 data sets demonstrate that the LDANB generally has the best classification accuracy when the priors for both attributes and classes are generalized Dirichlet distributions. When computational efficiency is taken into account, Dirichlet prior can be a proper choice for classes.

Research limitations/implications—The multivariate distributions defined on the unit simplex other that Dirichlet and generalized Dirichlet distributions are not considered.

* Corresponding author. Email: cherry@mail.hwai.edu.tw。
2016/11/23 received; 2017/07/13 revised; 2017/11/17 accepted

Practical implications – The procedure for introducing Dirichlet and generalized Dirichlet priors for attributes and classes is proposed to improve the performance of the naïve Bayesian classifier. The experimental results show that assuming priors for both attributes and classes are beneficial.

Originality/value – The LDANB model for introducing priors for the naïve Bayesian classifier is novel, and this model can enhance the competitiveness of the naïve Bayesian classifier in classification tasks.

Keywords: Naïve Bayesian classifiers, Dirichlet distribution, generalized Dirichlet distribution, latent Dirichlet allocation

壹、緒論

在資料探勘的分類演算法中，簡易貝氏分類器（Naïve Bayesian classifiers）具有運算效率高且分類正確率佳之優勢，已廣泛應用在許多實務上，例如：圖像辨識（Keren 2003）、垃圾郵件偵測（Sipahi et al. 2015）、文件情感分類（Perikos & Hatzilygeroudis 2016）、產品缺陷偵測（Addin et al. 2007）和軟體缺陷預測等（Menzies et al. 2007; Turhan & Bener 2009）。以文件分類領域而言，有學者（Tang et al. 2016）提出將簡易貝氏分類器加入類別特定（class-specific）屬性之自動文件分類方法，允許每個類別選擇最重要的屬性，且易於結合現有的屬性挑選標準，研究結果顯示該方法能有效提升文件分類之效能。其次，Alvi 與 Pears（2017）應用簡易貝氏分類器將人臉辨識數據庫劃分為子空間來縮小搜索空間，以改善辨識正確率。再者，簡易貝氏分類器也使用於醫療領域的成人心血管疾病風險程度檢測（Miranda et al. 2016）和藥物引起的骨髓抑制預測（Zhang et al. 2015）、生物科技領域的基因序列（Terribilini et al. 2007）及物種基因預測（Yousef et al. 2006）。

由於簡易貝氏分類器係以貝氏定理為基礎，藉由訓練樣本的學習，計算資料的條件機率，並配合資料屬性間彼此條件獨立的假設（The conditional independent assumption），進行分類預測，所以運算效率快，許多大數據分析也經常使用簡易貝氏分類器進行分析，例如：網頁意見評論分析（Tripathy et al. 2016）。此外，有研究發現簡易貝氏分類器搭配屬性離散化方法進行網路流量分類，相較於支援向量機（support vector machine; SVM）及 *k*-NN 等機器學習方法，有更佳的分類速度及正確率（Zhang et al. 2013）。而且研究發現在藥物引起的骨髓抑制應用上，簡易貝氏分類器亦優於 SVM 及單隱藏層前饋神經網絡（single-hidden-layer feedforward neural network）（Zhang et al. 2015）。另有學者（Trovato et al. 2016）應用簡易貝氏分類器於社交型機器人之人機互動模式，研究發現簡易貝氏分類器可以有效處理小型和不完整的資料檔，並且表現優於 SVM。由此可知，簡易貝氏分類器相較於其他機器學習方法除了有運算效率佳之優勢外，分類正確率亦具有競爭力。

另一方面，為了提升簡易分類器之分類正確率，大部分會採用狄氏分配（Dirichle distribution）或廣義狄氏分配（generalized Dirichlet distribution）當成先驗分配（prior distribution）進行資料屬性可能值機率之參數調整（Wong 2009）。而加入先驗分配的目的，主要係為了能有更接近原始資料概念的訓練樣本，藉由先驗參數之調控，使資料的分佈更趨近於真實情況。此外，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，因此，對應的分配必須符合變數非負且總和為一的性質，由於多變量中的狄氏分配與廣義狄氏分配擁有此特性，所以可將其假設為簡易貝氏分類器之先驗分配。

惟過去研究僅於資料屬性出現之機率加入先驗分配進行參數調整，而資料檔中類別值的機率卻未有加入先驗分配之機制，如此可能導致分類正確率之提升有

所限制。所以本研究目的為提出一個新的方法，透過潛在狄氏配置模型（Latent Dirichlet Allocation model），除了將先驗分配機制應用於屬性值，亦於類別值之機率加入先驗分配，進行參數調整，使資料分佈更接近原本真實概念，以提升簡易貝氏分類器之分類正確率。再者，傳統潛在狄氏配置模型之先驗分配係以狄氏分配為主，以實務應用而言，狄氏分配之等信賴需求與負相關需求條件限制可能過於嚴苛（Wong 2009）；相對之下，廣義狄氏分配放寬了狄氏分配之條件限制，僅限制第一個變數與其他變數之間必須為負相關，其他變數則無此限制，也無正規化變異數相等之限制。是故，本研究亦加入廣義狄氏分配進行混合模型測試，如此，在實務上可更廣泛地應用，同時能提升簡易貝氏分類器之分類正確率。

本文章的架構如下：第二節說明簡易貝氏分類器與潛在狄氏配置模型之原理與機制；第三節的研究方法，會闡述本研究提出的潛在狄氏配置簡易貝氏分類器之運作及先驗分配之參數調整方式；第四節的實證研究將呈現本研究實驗結果；最後，第五節內容為結論和未來建議。

貳、簡易貝氏分類器與潛在狄氏配置

簡易貝氏分類器係貝氏分類器的一種，藉由計算條件機率進行分類預測，因此，運算效率高且分類正確率佳。其次，一般使用簡易貝氏分類器時，對於資料屬性會加入先驗分配之機制，但資料檔中類別值的機率卻未有加入先驗分配，如此可能導致分類正確率之提升有所限制。所以本研究透過潛在狄氏配置模型，將先驗分配機制混合應用於屬性值與類別值之機率，使資料分佈更接近原本真實概念，以提升簡易貝氏分類器之分類正確率。本節會介紹簡易貝氏分類器加入先驗分配之運作原理，以及潛在狄氏配置模型之架構。

一、簡易貝氏分類器

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 $X_1, X_2, \ldots, X_n$，其中一筆資料 $\mathbf{x}=(x_1, x_2, \ldots, x_n)$ 屬於第 j 個類別值 C_j 的機率為：

$$p(C_j \mid \mathbf{x}) = \frac{p(C_j \cap \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid C_j)p(C_j)}{p(\mathbf{x})} \tag{1}$$

上述等式(1)中的 $p(C_j \mid \mathbf{x})$ 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；其次，$p(\mathbf{x} \mid C_j)$ 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；而 $p(\mathbf{x})$ 則為資料 **x** 出現的機率。由於在計算同一筆資料時，$p(\mathbf{x})$ 是固定的，因此等式(1)可以簡化為：

$$p(C_j \mid \mathbf{x}) \propto p(\mathbf{x} \mid C_j)p(C_j) \tag{2}$$

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** 的類別值判定為 C^*。

同樣地，簡易貝氏分類器係以貝氏分類法則為主，再加入屬性條件獨立的假設作為分類的依據，亦即在給定某類別值的情況下，各屬性之間必須彼此互相獨立。因此，以等式(2)為基礎，再加入條件獨立之假設，則可將式子展開成：

$$\begin{aligned} p\left(C_j \mid \mathbf{x}\right) &\propto p\left(x_1 \mid C_j\right) \times p\left(x_2 \mid C_j\right) \times \cdots \times p\left(x_n \mid C_j\right) \times p\left(C_j\right) \\ &= \prod_{i=1}^{n} p\left(x_i \mid C_j\right) \times p\left(C_j\right) \end{aligned} \tag{3}$$

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(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 & 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。

再者，大部分簡易貝氏分類器會加入先驗分配之機制以提升分類正確率，主要係為了能有更接近原始資料概念的訓練樣本，藉由先驗參數之調控，使資料的分佈更趨近於真實情況。其次，簡易貝氏分類器是透過計算屬性可能值的條件機率進行分類預測，因此對應的分配必須符合變數非負且總和為一的單位體（unit simplex）性質，由於多變量中的狄氏分配與廣義狄氏分配擁有此特性，所以可將其假設為簡易貝氏分類器之先驗分配。而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），最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，而且具有共軛性質（conjugate property）。

定義 1

隨機向量 $\mathbf{\Theta} = (\theta_1, \theta_2, ..., \theta_k)$ 滿足 $\theta_1 + \theta_2 + \cdots + \theta_k \le 1$ 與 $\theta_j \ge 0,\ j = 1, 2, ..., k$；參數 $\alpha_j > 0,\ j = 1,\ 2,\ ...,\ k+1$ 且 $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_{k+1}$，其機率密度函數為：

$$f\left(\mathbf{\Theta}\right) = \frac{\Gamma\left(\alpha\right)}{\prod_{j=1}^{k+1} \Gamma\left(\alpha_j\right)} \prod_{j=1}^{k} \theta_j^{\alpha_{j-1}} \left(1 - \theta_1 - \theta_2 - \cdots - \theta_k\right)^{\alpha_{k+1}-1} \tag{4}$$

上述等式(4)中的隨機向量 $\mathbf{\Theta}$ 服從 k 維的狄氏分配，以 $\mathbf{\Theta} \sim D_k\left(\alpha_1, \alpha_2, ..., \alpha_k;\ \alpha_{k+1}\right)$ 表示。其次，大部分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，造成概似函數之計算結果產生 0 的情況，一般會藉由拉普拉斯估計（Laplace's estimate）的方式進行屬性參數之設定（Cestnik & Bratko 1991），如此相當於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 $D_k\left(1, 1, ..., 1; 1\right)$。

另一方面，在服從狄氏分配的隨機向量中，任兩變數之間必為負相關。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）。此外，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，其信賴水準也都會一樣，稱為等信賴需

求（equal-confidence requirement）（Wong 2009）。而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 $\mathbf{y} = (y_1, y_2, ..., y_{k+1})$ 為資料屬性的 k+1 個可能值分別出現的次數，且隨機向量 $\boldsymbol{\Theta} = (\theta_1, \theta_2, ..., \theta_k)$ 為該屬性第 1~k 個可能值出現的機率，其服從 k 維的狄氏分配 $\boldsymbol{\Theta} \sim D_k(\alpha_1, \alpha_2, ..., \alpha_k; \alpha_{k+1})$，然後概似函數 $\mathbf{y}|\boldsymbol{\Theta}$ 服從多項式分配，則 $\boldsymbol{\Theta}$ 在給定 $\mathbf{y}$ 的機率密度函數 $f(\boldsymbol{\Theta}|\mathbf{y})$ 會與 $p(\mathbf{y}|\boldsymbol{\Theta}) \times p(\boldsymbol{\Theta})$ 成正比。由於狄氏分配具有共軛性質，因此 $\boldsymbol{\Theta}$ 在給定 $\mathbf{y}$ 的條件下，仍服從狄氏分配，即 $(\boldsymbol{\Theta}|\mathbf{y}) \sim D_k(\alpha_1 + y_1, \alpha_2 + y_2, ..., \alpha_k + y_k; \alpha_{k+1} + y_{k+1})$（Wong 1998）。如果 θ_j 為 $\boldsymbol{\Theta}$ 的其中一個變數，則 θ_j 在給定 $\mathbf{y}$ 的條件下，期望值為：

$$E(\theta_j|\mathbf{y}) = \frac{y_j + \alpha_j}{\sum_{i=1}^{k+1}(y_i + \alpha_i)} \tag{5}$$

因此藉由等式(5)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(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。

然而，以實務應用而言，狄氏分配之條件限制過於嚴苛，所以Connor 與 Mosimann（1969）將狄氏分配一般化，推導出廣義狄氏分配。其主要條件限制係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，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地應用。其定義如下所示：

定義 2

隨機向量 $\boldsymbol{\Theta} = (\theta_1, \theta_2, ..., \theta_k)$ 滿足 $\theta_1 + \theta_2 + \cdots + \theta_k \le 1$ 與 $\theta_j \ge 0,\ j = 1, 2, \ldots, k$；參數 α_j、β_j、λ_j 符合 $\alpha_j > 0,\ j = 1, 2, ..., k$、$\beta_j > 0,\ j = 1, 2, ..., k$、$\lambda_k = \beta_k - 1$ 與 $\lambda_j = \beta_j - \alpha_{j+1} - \beta_{j+1},\ j = 1, 2, ..., k-1$，且機率密度函數為：

$$f(\boldsymbol{\Theta}) = \prod_{j=1}^{k} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j - 1}(1 - \theta_1 - \cdots - \theta_j)^{\lambda_j} \tag{6}$$

在等式(6)中的隨機向量 $\boldsymbol{\Theta}$ 服從 k 維的廣義狄氏分配，表示為 $\boldsymbol{\Theta} \sim GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$。透過定義2可得知，廣義狄氏分配與狄氏分配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。而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；其次，倘若其參數符合 $\beta_j = \alpha_{j+1} + \beta_{j+1},\ j = 1, 2, ..., k-1$ 時，所有變數的正規化變異數皆會相同，則隨機向量 $\boldsymbol{\Theta}$ 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。

再者，若將廣義狄氏分配當作簡易貝氏分類器之先驗分配時，則係假設 $\mathbf{y} = (y_1, y_2, ..., y_{k+1})$ 為資料屬性的 k+1 個可能值分別出現的次數，且隨機向量 $\boldsymbol{\Theta} = (\theta_1, \theta_2, ..., \theta_k)$ 服從 k 維的廣義狄氏分配，即 $\boldsymbol{\Theta} \sim GD_k(\alpha_1, \alpha_2, ..., \alpha_k; \beta_1, \beta_2, ..., \beta_k)$，

然後概似函數 $\mathbf{y}|\mathbf{\Theta}$ 服從多項式分配，而 $\mathbf{\Theta}$ 在給定 $\mathbf{y}$ 的機率密度函數 $f(\mathbf{\Theta}|\mathbf{y})$ 會與 $p(\mathbf{y}|\mathbf{\Theta})\times p(\mathbf{\Theta})$ 成正比。由此可知，廣義狄氏分配亦具有共軛性質，所以 $\mathbf{\Theta}$ 在給定 $\mathbf{y}$ 的條件下，仍然會服從廣義狄氏分配，即 $(\mathbf{\Theta}|\mathbf{y}) \sim GD_k(\alpha_1+y_1, \alpha_2+y_2, ..., \alpha_k+y_k; \beta_1+\sum_{i=2}^{k+1} y_i, \beta_2+\sum_{i=3}^{k+1} y_i, ..., \beta_k+y_{k+1})$（Wong 1998）。而且，倘若 θ_j 為 $\mathbf{\Theta}$ 的其中一個變數，令 $n_i = y_i + y_{i+1} + \cdots + y_k + y_{k+1}$，$i = 1,2,...,k+1$，則 θ_j 在給定 $\mathbf{y}$ 的條件下，其期望值為：

$$E\left(\theta_j|\mathbf{y}\right) = \frac{\alpha_j + y_j}{\alpha_j + \beta_j + n_j} \prod_{i=1}^{j-1} \frac{\beta_i + n_{i+1}}{\alpha_i + \beta_i + n_i} \tag{7}$$

而第 k+1 個變數 $\theta_{k+1} = 1 - \theta_1 - \theta_2 - \cdots - \theta_k$ 的期望值為：

$$E\left(\theta_{k+1}|\mathbf{y}\right) = \prod_{i=1}^{k} \frac{\beta_i + n_{i+1}}{\alpha_i + \beta_i + n_i} \tag{8}$$

藉由等式(7)或(8)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(3)找出擁有最大事後機率的類別值，當成簡易貝氏分類器預測資料之類別值。

二、潛在狄氏配置

潛在狄氏分配（latent Dirichlet allocation; LDA）係由 Blei, Ng 與 Jordan（2003）提出的一種的機率模型，最初用於發掘文件（document）之特定主題（topic），因此又稱為主題模型，經常應用於資訊檢索、文字探勘或機器學習領域（Azzopardi et al. 2004; Zhou et al. 2009）。然而，LDA 並不限於文件資料之應用，也可使用於影像處理或醫療保險數據上的診斷預測與用藥預測（Niebles et al. 2008; Lu et al. 2016）

若以文件分類應用而言，LDA 模型架構為假設每篇文件係由多個主題所組成，所以文件被視為潛在主題的隨機混合體。每個主題會包含許多字彙（word）。換言之，文件是由多個主題所構成的機率分配；每一個主題則是由多個字彙所構成的機率分配，而且 LDA 係藉由狄氏分配調控主題與字彙的機率，使其更趨近真實情況，並以此來計算該文件隸屬於某主題之機率。由於每篇文件中的主題皆有其狄氏分配機率，不同的文件對應的主題分配有所不同，因此也可藉由主題分配情形來判斷文件的相似度（Blei et al. 2003）。

由上述可知 LDA 模型與簡易貝氏分類器的機率分配皆係透過先驗機率參數進行控制，惟兩者最大差異在於 LDA 模型係同時在主題與字彙之機率加入先驗參數，而簡易貝氏分類器僅針對屬性可能值之機率加入先驗參數，但對於類別值之機率並無先驗分配之參數調控機制。

參、研究方法

由於簡易貝氏分類器的運作方式，主要係透過計算資料屬性值分佈於各類別的機率進行分類預測，處理的資料屬性以離散型態資料為主，所以在此先假設資料皆為離散型態。而本節首先係介紹本研究提出之潛在狄氏配置簡易貝氏分類器之模型架構，接著說明如何運用不同的先驗分配機制進行參數調整，所會使用的先驗分配包括狄氏分配與廣義狄氏分配。

一、潛在狄氏配置簡易貝氏分類器

由於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 $x_1, x_2, \ldots, x_n$ 出現之機率$\prod_{i=1}^{n} p\left(x_i \mid C_j\right) \times p(C_j)$，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。因此，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中，而造成概似函數之計算結果產生 0 的情況，大部分會藉由拉普拉斯估計的方式進行屬性參數之設定，如等式(9)所示：

$$p\left(x_i = V_{im} \mid C_j\right) = \frac{y_{im} + 1}{y + \left(k + 1\right)} \tag{9}$$

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，此即為採用狄氏分配為先驗分配，且參數設定為1。

然而，在過去研究中，簡易貝氏分類器之先驗分配機制僅用於資料屬性可能值機率之參數調整，對於資料檔中類別值的機率卻未有加入先驗分配之機制，如此可能導致分類正確率之提升有所限制。所以本研究提出潛在狄氏配置簡易貝氏分類器（Latent Dirichlet Allocation Naïve Bayes; LDANB），透過潛在狄氏配置模型，將先驗分配機制加入類別值之機率，進行參數調整，使資料更接近原本真實概念，圖 1 為 LDANB 模型示意圖。

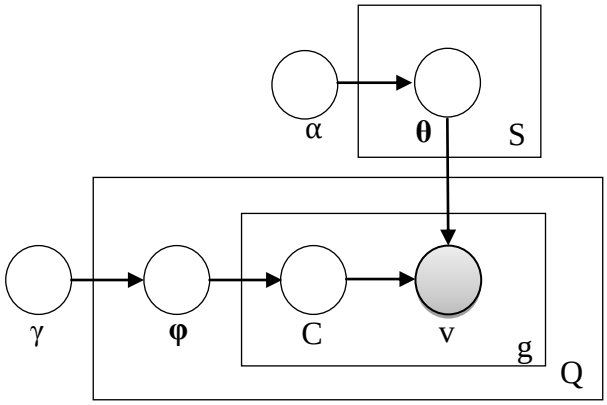

圖 1：LDANB 模型示意圖

圖 1 的 γ 為資料檔中類別值之先驗參數，意即相對於 LDA 模型中主題之先驗參數；而α為資料屬性可能值之先驗參數，亦相對 LDA 模型中各字彙之先驗參數。**φ** 為資料檔中類別值之分配，可設定為 $\boldsymbol{\varphi}\sim D_k(\gamma_1, \gamma_2, \ldots, \gamma_k; \gamma_{k+1})$ 或 $\boldsymbol{\varphi}\sim GD_k(\gamma_1, \gamma_2, \ldots, \gamma_k; \delta_1, \delta_2, \ldots, \delta_k)$ ；**θ** 為資料屬性可能值之分配，可設定為 $\boldsymbol{\theta}\sim D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$或 $\boldsymbol{\theta}\sim GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$，C 代表類別值，Q 為資料檔中的資料筆數，S 為屬性個數，g 為屬性可能值個數，而 v 則為每筆資料中的屬性可能值。

二、先驗分配參數調整

大部分簡易貝氏分類器會以狄氏分配做為先驗分配，主要係因狄氏分配具有單位體之特性和共軛性質，而且各個變數的動差（moment）計算較簡單。其次，使用狄氏分配設定參數時，為了運算操作簡便，一般會假設是無資訊性的（noninformative）。無資訊性係指在沒有任何資訊的情況下，各個隨機變數出現的機率都相同，因此如果符合無資訊性，則期望值 $E(\theta_j)=\alpha_j/\alpha$ 都會相等，代表 $E(\theta_1)=E(\theta_2)=\cdots=E(\theta_{k+1})$，亦表示狄氏分配之各個屬性參數都會相同，即 $\alpha_1=\alpha_2=\cdots=\alpha_{k+}$；所以使用拉普拉斯估計的方式，可視同係將狄氏分配之屬性參數 α_i 設定為 1，$i=1,2,\ldots,k+1$，即 $\boldsymbol{D}_k\left(\mathbf{1},\mathbf{1},\ldots,\mathbf{1};\mathbf{1}\right)$（Wong 2009）。

再者，假設 $\mathbf{y}=(y_1, y_2, \ldots, y_{k+1})$ 為訓練資料中，類別值 C_j 資料之某屬性的 $k+1$ 個可能值分別出現之次數，隨機向量 $\boldsymbol{\Theta}=(\theta_1, \theta_2, \ldots, \theta_k)$ 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 $\boldsymbol{\Theta}\sim D_k\left(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1}\right)$，如果 θ_j 為 $\boldsymbol{\Theta}$ 的其中一個變數，而 θ_j 在給定資料屬性 **y** 的條件下，可藉由上一節之等式 (5) $E\left(\theta_j|\mathbf{y}\right)=(y_j+\alpha_j)/\sum_{i=1}^{k+1}(y_i+\alpha_i)$，分解得到 $\sum_{i=1}^{k+1}y_i/\sum_{i=1}^{k+1}(y_i+\alpha_i)$ 和，$\sum_{i=1}^{k+1}\alpha_i/\sum_{i=1}^{k+1}(y_i+\alpha_i)$，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。

然而，一般資料檔與文件分類資料檔不同，由於一般資料檔中，每筆資料係互相獨立，所以無法估算相關的變異數與共變異數；相較之下，文件分類可從資料檔中，在任何給定類別值的情況，估計相異字出現機率之變異數和任一對相異字出現機率之間的共變異數（Wong 2014）。因此，以一般資料檔而言，先驗分配的最佳參數無法從資料檔中直接估計，進而需採用不同的參數進行測試，以找出有助於分類正確率提升之參數。而 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。所以本研究在進行先驗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。

其次，廣義狄氏分配同樣係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為該屬性可能值個數。而廣義狄氏分配的變數期望值（Wong 1998）如等式(10)所示：

$$E(\theta_j) = \frac{\alpha_j}{\alpha_j + \beta_j} \prod_{i=1}^{j-1} \frac{\beta_i}{\alpha_i + \beta_i} \tag{10}$$

在期望值皆相等的條件下，可得知$GD_k(\alpha_1, \alpha_2, ..., \alpha_k; \beta_1, \beta_2, ..., \beta_k)$中的參數皆須遵守如下之關係：

$$\frac{\alpha_i}{\alpha_i + \beta_i} = \frac{1}{k - i + 2}\ ,\ i = 1, 2, ..., k \tag{11}$$

是故，只要參數α_i已知，即能透過等式(11)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，如上一節之等式 (7) 所示：$E(\theta_j|\mathbf{y}) = (\alpha_j + y_j)/(\alpha_j + \beta_j + n_j) \prod_{i=1}^{j-1} [(\beta_i + n_{i+1})/(\alpha_i + \beta_i + n_i)]$，其中$n_i = y_i + y_{i+1} + \cdots + y_k + y_{k+1}$，$i = 1, 2, ..., k+1$；之後，再用此期望值做為簡易貝氏分類器$p(\mathbf{x}|C_j)$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。以下分別說明本研究提出之潛在狄氏配置簡易貝氏分類器的類別值與屬性值之狄氏分配與廣義狄氏分配參數設定步驟。

LDANB類別值之狄氏分配參數設定步驟：

此即原本簡易貝氏分類器在計算類別值C_j的機率$p(C_j)$時，加入先驗參數γ_i，並測試該參數為$\gamma_1 = \gamma_2 = \cdots = \gamma_{t+1}$的條件下，參數設定為[1,60]之間整數值的分類正確率，之後，找出能使分類正確率表現最佳之參數值，將其設定為γ_i^*。

LDANB類別值之廣義狄氏分配參數設定步驟：

步驟一：依序針對各類別值參數γ_i進行設定，首先將第一個類別值的參數γ_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δ_1，尋找能使分類正確率最佳的γ_1參數值，將其設定為γ_1^*，而此時其他尚未設定之類別值參數$\gamma_i, i = 2, 3, ..., t$皆設定為1。

步驟二：在設定完$\gamma_1 = \gamma_1^*$後，依序針對該第二個類別值γ_2進行參數設定，同樣測試當γ_2為[1,60]之間的整數值，並計算出相對應的δ_2，選擇能使分類正確率最佳的γ_2參數值，將其設定為γ_2^*。

步驟三：重複步驟一和步驟二之參數設定方式，設定其餘類別值的參數γ_i，$i = 3, 4, ..., t$之最佳值，且同樣計算出相對應的δ_i^*。

LDANB屬性值之狄氏分配參數設定步驟：

步驟一：在類別值之先驗參數值設定後，依序將屬性之參數值α_i進行設定，亦即測試該屬性在各個可能值參數為$\alpha_1 = \alpha_2 = \cdots = \alpha_{k+1}$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_1^*，此時其他屬性之可能值參數皆設定為1。

步驟二：當第一個屬性的參數值在$\alpha_1 = \alpha_2 = \cdots = \alpha_{k+1} = \alpha_i^*$的情況下，重複上述步驟，找出第二個屬性之最佳參數值α_j^*，並將其設定為第二個屬性的參數 $\alpha_1 = \alpha_2 = \cdots = \alpha_{k+1} = \alpha_j^*$。

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。

LDANB屬性值之廣義狄氏分配參數設定步驟：

步驟一：在類別值之先驗參數值設定後，依序將第一個屬性之可能值的參數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^*，而此時該屬性之其他可能值參數α_i，$i = 2, 3, ..., k$皆設定為1。

步驟二：當設定完成$\alpha_1 = \alpha_1^*$後，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α_i，$i = 3, 4, ..., k$之最佳值，且同樣計算出相對應的 β_i^*，以求得該屬性$GD_k(\alpha_1, \alpha_2, ..., \alpha_k; \beta_1, \beta_2, ..., \beta_k)$之分類正確率。

步驟四：在第一個屬性之所有可能值參數皆設定為最佳廣義狄氏分配的情況下，再重複步驟一至步驟三之方式，依序設定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。

此外，本研究亦同時比較屬性值與類別值使用不同先驗分配模式之分類正確率，總共分為九種模式，如表1所示：

表1：本研究比較之九種先驗分配模式

		屬性值先驗分配參數設定方式		
		拉普拉斯估計	狄氏分配	廣義狄氏分配
類別值先驗分配參數之設定方式	無	模式一	模式二	模式三
	狄氏分配	模式四	模式五	模式六
	廣義狄氏分配	模式七	模式八	模式九

本研究使用 LDANB 進行表 1 中的九種先驗參數設定時，係以類別值為優先進行參數設定，之後才調整屬性先驗分配之參數，以找出該模式下最佳分類正確率之先驗參數。以模式五為例說明，即當類別值與屬性值之先驗分配皆為狄氏分配之模式。首先，會先設定類別值之先驗參數$\gamma_1 = \gamma_2 = \cdots = \gamma_{t+1} = 1$，然後再進行屬性值之狄氏分配參數設定，如同上述 LDANB 屬性值之狄氏分配參數設定步驟，會測試各屬性值參數在[1,60]之間整數值的分類正確率，找出能使分類正確率表現最佳之參數值；之後，才繼續測試當類別值之先驗參數為$\gamma_1 = \gamma_2 = \cdots = \gamma_{t+1} = 2$時，各屬性值之狄氏分配參數，同樣找出各屬性值參數在[1,60]之間整數值的最佳分類正確率；依此類推，繼續測試當類別值之狄氏參數為[3,60]之間整數值，以及在該類別值之參數值條件下具狄氏參數之屬性值。最後，找出 LDANB 在模式五的情況下，類別值與屬性值之最佳先驗分配參數值及其分類正確率。

三、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 等份交互驗證（f-fold cross validation）的方式進行分析。其主要係將資料檔內的資料樣本隨機分成 f 等份，再以 f-1 等份當成訓練資料，進行分類學習，其餘的 1 個等份的資料做為測試資料進行測試，而且各個等份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。最後會取其平均值作為分類正確率之評估指標。

為了在進行f等份交互驗證時，能控制每個等份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，本研究將選用5等份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，藉此避免每個等份中的資料筆數少於30筆，而影響最後簡易貝氏分類器之分類結果評估。

肆、實證研究

簡易貝氏分類器係以計算條件機率進行分類預測，因此處理的資料屬性以離散型態資料為主。一般在處理連續型態的屬性資料時，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 & Sahami 1996）。所以本研究使用常見的等寬度離散化方法（equal width discretization）進行離散化，並將連續型屬性資料之區間範圍分為 10 個區間，即 10-bin，此即把連續型態的資料離散化成 10 個屬性可能值。本節將先介紹實證研究所使用的資料檔特性，之後，呈現本研究使用的 LDANB 的九種模式測試比較結果。

一、資料檔介紹

本研究選用美國加州大學歐文分校機器學習資料存放站（UCI Machine Learning Repository）（Lichman 2013）的 20 個資料檔進行測試比較，如表 2 所示。在此 20 個資料檔之中，資料筆數最少為 150 筆，最多為 5473 筆，而且屬性值與類別值個數皆大於 2 個。

表 2：資料檔介紹

資料檔	資料筆數	屬性值個數	連續型屬性	離散型屬性	類別值個數
annealing	898	18	6	12	5
car	1728	6	0	6	4
cardiotocographic	2126	21	21	0	10
dermatology	366	34	33	1	6
ecoli	336	7	5	2	8
flags	194	28	10	18	8
glass	214	9	9	0	6
image segmentation	2310	18	18	0	7
iris	150	4	4	0	3
leaf	340	14	14	0	30
mammographic	959	5	1	4	6
mfeat	2000	6	3	3	10
newthyroid	215	5	5	0	3
page	5473	10	10	0	5
seeds	210	7	7	0	3
user	403	5	5	0	4
vertebral	310	6	6	0	3
vowel	990	12	10	2	11
waveform	5000	21	21	0	3
yeast	1484	8	8	0	10

二、潛在狄氏配置簡易貝氏分類器實證研究

本小節將呈現LDANB的各種先驗分配模式之實驗結果。表3為原本簡易貝氏分類器屬性加入先驗分配為拉普拉斯估計、狄氏分配與廣義狄氏分配之資料檔測試結果，分別為模式一、模式二和模式三，其類別值為尚未加入先驗分配機制。

表 3：模式一、模式二與模式三之測試結果

資料檔	模式一	模式二	模式三
annealing	93.59%	93.62%	**94.48%**
car	85.43%	85.78%	**86.02%**
cardiotocographic	72.93%	75.04%	**76.01%**
dermatology	97.81%	**98.29%**	**98.29%**
ecoli	81.62%	82.28%	**83.66%**
flags	61.70%	64.78%	**67.53%**
glass	59.35%	63.52%	**64.11%**
image segmentation	88.47%	90.00%	**90.26%**
iris	93.19%	95.93%	**96.68%**
leaf	51.75%	56.26%	**58.71%**
mammographic	76.57%	78.17%	**78.30%**
mfeat	69.77%	69.92%	**70.88%**
newthyroid	91.42%	92.22%	**92.35%**
page	91.89%	92.16%	**92.44%**
seeds	89.55%	**92.04%**	91.97%
user	85.59%	87.86%	**88.07%**
vertebral	72.90%	74.83%	**76.83%**
vowel	65.35%	67.36%	**68.70%**
waveform	80.42%	80.98%	**81.67%**
yeast	58.17%	58.53%	**59.81%**
分類正確率平均值	78.37%	79.98%	**80.84%**
勝出數	0	2	**19**

從表 3 結果可發現，當屬性值加入三種先驗分配模式時，係以廣義狄氏分配之平均分類正確率表現最佳，在 20 個資料檔中有 19 個資料檔皆係如此。

其次，表 4 為 LDANB 在類別值加入狄氏先驗分配後，屬性為三種先驗分配之測試結果。由表 4 中可發現，在類別值加入狄氏先驗分配條件下，屬性值之先驗分配仍以搭配廣義狄氏分配的模式表現最佳，且分類正確率平均值都比表 3 的相對模式更高，顯示在類別值加入狄氏分配可提高分類正確率表現。

表 4：模式四、模式五與模式六之測試結果

資料檔	模式四	模式五	模式六
annealing	94.20%	94.20%	**95.09%**
car	88.06%	88.06%	**89.74%**
cardiotocographic	72.84%	75.86%	**76.54%**
dermatology	97.81%	**98.29%**	**98.29%**
ecoli	82.26%	83.49%	**84.46%**
flags	62.20%	**68.31%**	67.53%
glass	59.20%	64.58%	**66.76%**
image segmentation	88.47%	90.29%	**90.65%**
iris	93.81%	95.93%	**97.32%**
leaf	53.26%	59.64%	**61.54%**
mammographic	76.57%	78.20%	**78.49%**
mfeat	69.77%	70.06%	**71.14%**
newthyroid	93.70%	93.70%	**94.20%**
page	92.16%	92.18%	**92.62%**
seeds	89.55%	92.04%	**92.31%**
user	85.81%	87.91%	**88.77%**
vertebral	74.30%	76.73%	**79.19%**
vowel	65.35%	69.27%	**69.61%**
waveform	80.42%	80.98%	**81.72%**
yeast	58.59%	59.05%	**60.42%**
分類正確率平均值	78.92%	80.94%	**81.82%**
勝出數	0	2	**19**

之後，表 5 為 LDANB 在類別值加入廣義狄氏先驗分配後，屬性為三種先驗分配模式之分類正確率。研究結果顯示在大部分的資料檔中，表 3、表 4 與表 5 具有一致性的表現，皆為屬性值加入廣義狄氏分配的情況會有較高的分類正確率，且表 5 相較於表 4 而言，在分類正確率平均值上亦可再次獲得提升。藉此可證明透過 LDANB 模式將類別值導入先驗分配之機制，再搭配原本既已使用先驗分配之屬性，在大部分資料檔中，確實有助於分類正確率之提升，而且類別值之先驗分配使用廣義狄氏分配優於狄氏分配。

表 5：模式七、模式八與模式九之測試結果

資料檔	模式七	模式八	模式九
annealing	94.61%	94.78%	**95.55%**
car	87.27%	87.87%	**88.87%**
cardiotocographic	73.18%	76.59%	**77.31%**
dermatology	97.81%	**98.29%**	**98.29%**
ecoli	82.83%	83.45%	**84.72%**
flags	62.38%	67.78%	**69.07%**
glass	60.32%	64.58%	**67.64%**
image segmentation	88.47%	90.33%	**90.81%**
iris	93.81%	95.93%	**97.32%**
leaf	52.91%	60.64%	**62.22%**
mammographic	76.89%	78.46%	**78.60%**
mfeat	69.82%	70.17%	**71.09%**
newthyroid	93.14%	93.70%	**95.00%**
page	92.24%	92.32%	**92.66%**
seeds	89.55%	**92.04%**	91.97%
user	86.08%	88.40%	**89.35%**
vertebral	74.41%	76.73%	**79.21%**
vowel	65.35%	69.27%	**69.85%**
waveform	80.42%	80.98%	**81.74%**
yeast	59.15%	58.88%	**60.26%**
分類正確率平均值	79.03%	81.06%	**82.08%**
勝出數	0	2	**19**

再者，本研究比較 LDANB 模式在屬性先驗分配機制相同時，類別值導入不同之先驗分配機制對於分類正確率的影響，如圖 2 所示。第一群為屬性先驗分配皆使用拉普拉斯估計，包含：模式一為類別值不導入先驗分配，模式四為類別值導入狄氏分配，模式七為類別值導入廣義狄氏分配模式。第二群為屬性先驗分配皆使用狄氏分配，包含：模式二為類別值不導入先驗分配，模式五為類別值導入狄氏分配，模式八為類別值導入廣義狄氏分配模式。第三群為屬性先驗分配皆使用廣義狄氏分配，包含：模式三為類別值不導入先驗分配，模式六為類別值導入狄氏分配，模式九為類別值導入廣義狄氏分配模式。

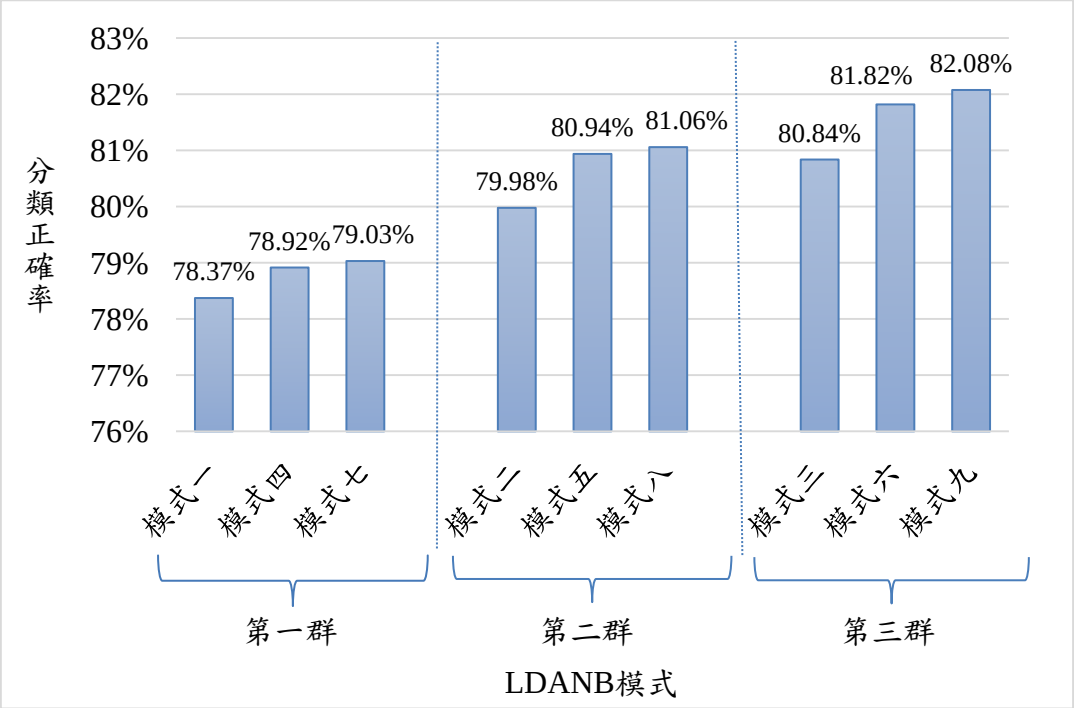

圖 2：LDANB 九種模式之分類正確率平均值

從圖 2 可得知，雖然此三群皆以類別值導入廣義狄氏分配的分類正確率較佳，但類別值導入狄氏分配相較於不導入先驗分配的分類正確率平均值上升幅度，大於類別值導入廣義狄氏分配相較於狄氏分配之分類正確率上升幅度。

由於使用不同運算平台與程式撰寫方式之差異皆會影響各模式實際執行時間。若以簡易貝氏分類器完整執行一次資料檔中每筆資料分類之運算時間為基準，亦即使用模式一之運算時間為 1 單位，進行運算成本比較，九種先驗分配模式之運算成本如表 6 所示。

在表 6 中，n 代表屬性個數，而屬性可能值個數為 k+1，類別值個數則以 t+1 表示，測試先驗分配之參數範圍在[1,60]之間的整數值，並假設在同一資料檔中，各個屬性之屬性可能值皆相同，且忽略廣義狄氏分配相較於狄氏分配計算期望值之運算成本進行比較。以模式九的運算成本而言，係指屬性使用廣義狄氏分配時，類別值亦導入廣義狄氏分配，其運算成本為 60nk 乘以 60t，則為 3600nkt。藉此可知，模式九相較於類別值僅導入狄氏分配之模式六，運算成本會增加 t 倍。

表 6：九種先驗分配模式之運算成本

模式一	模式二	模式三
1	60n	60nk
模式四	模式五	模式六
60	3600n	3600nk
模式七	模式八	模式九
60t	3600nt	3600nkt

以 annealing 資料檔為例，其有 5 個類別值，若屬性先驗分配使用廣義狄氏分配，而類別值亦導入廣義狄氏分配，代表類別值之先驗分配維度 t=4，相較於類別值導入狄氏分配之運算時間，在使用相同的程式語言與運算平台的情況下，大約增加 4 倍，但其分類正確率之提升幅度卻有限。是故，當屬性之先驗分配模式相同時，類別值導入廣義狄氏分配相較於狄氏分配之運算複雜度較高，且會受類別值個數影響。因此若考量運算成本，建議可使用屬性為廣義狄氏分配，搭配類別值導入狄氏分配之 LDANB 模式六。

伍、結論及未來研究

大部分使用簡易貝氏分類器時，會在概似函數之屬性值加入先驗分配，以提升分類正確率。而且一般會採用狄氏分配或廣義狄氏分配當成先驗分配進行資料屬性可能值機率之參數調整。然而，過去研究對於資料檔中類別值的機率卻未有加入先驗分配之機制，如此可能導致分類正確率之提升有所限制。所以，本研究提出潛在狄氏配置簡易貝氏分類器，藉由潛在狄氏配置模型，將先驗分配機制同時加入類別值與屬性值之中，進行參數調整，使資料更接近原本真實概念。實證研究結果顯示，使用潛在狄氏配置模型之簡易貝氏分類器優於僅將屬性可能值加入先驗分配之情況，而且使用廣義狄氏分配會優於狄氏分配之分類正確率。惟廣義狄氏分配之運算複雜度較高。是故，建議潛在狄氏配置簡易貝氏分類器之先驗分配模式採用屬性可能值為廣義狄氏分配，搭配類別值為狄氏分配之機制，更能在有限的運算成本下，提升分類正確率。除此之外，由於一般資料檔中可能存在冗餘屬性而影響分類結果，並增加運算成本，未來建議使用潛在狄氏配置簡易貝氏分類器時，可藉由屬性挑選方法進行前置處理，以排除冗餘屬性干擾，更能有效提升潛在狄氏配置簡易貝氏分類器之分類正確率。

誌謝

本研究之經費由科技部編號 MOST 106-2410-H-006-020 之計畫所贊助。

參考文獻

Addin, O., Sapuan, S.M., Mahdi, E. and Othman, M. (2007), 'A naïve Bayes classifier for damage detection in engineering materials', *Materials and Design*, Vol. 28, No. 8, pp. 2379-2386.

Aitchison, J. (1985), 'A general class of distributions on the simplex', *Journal of the Royal Statistical Society Series B*, Vol. 47, No. 1, pp. 136-146.

Alvi, F.B. and Pears, R. (2017), 'A composite spatio-temporal modeling approach for age invariant face recognition', *Expert Systems with Applications*, Vol. 72, pp. 383-394.

Azzopardi, L., Girolami, M. and Van Rijsbergen, C.J. (2004), 'Topic based language models for ad hoc information retrieval', *IEEE International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, pp. 3281-3286.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), 'Latent Dirichlet allocation', *Journal of Machine Learning Research*, Vol. 3, No. 4-5, pp. 993-1022.

Cestnik, B. and Bratko, I. (1991), 'On estimating probabilities in tree pruning', *Proceedings of the 5th European Working Session on Learning on Machine Learning (EWSL 91)*, Porto, Portugal, March 6-8, pp. 138-150.

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*, Vol. 64, No. 325, pp. 194-206.

Domingos, P. and Plazzani, M. (1997), 'On the optimality of the simple Bayesian classifier under zero one loss', *Machine Learning*, Vol. 29, No. 2-3, pp. 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 (ICML 1995)*, California, July 9-12, pp. 194-202.

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

Keren, D. (2003), 'Recognizing image "style" and activities in video using local features and naive Bayes'. *Pattern Recognition Letters*, Vol. 24, No. 16, pp. 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 (KDD 1996), Oregon, August 2-4, pp.114-119.

Lichman, M. (2013), 'UCI Machine Learning Repository', *Irvine, CA: University of California, School of Information and Computer Science*, available at http://archive.ics.uci.edu/ml.

Lu, H.M., Wei, C.P. and Hsiao, F.Y. (2016), 'Modeling healthcare data using multiple-channel latent Dirichlet allocation', *Journal of Biomedical Informatics*, Vol. 60, pp. 210-223.

Menzies, T., Greenwald, J. and Frank, A. (2007), 'Data mining static code attributes to learn defect predictors', *IEEE Transactions on Software Engineering*, Vol. 33, No. 1, pp. 2-13.

Miranda, E., Irwansyah, E., Amelga, A.Y. and Maribondang, M.M. (2016), 'Detection of cardiovascular disease risk's level for adults using naive Bayes classifier', *Healthcare Informatics Research*, Vol. 22, No. 3, pp. 196-205.

Niebles, J.C., Wang, H. and Fei-Fei, L. (2008), 'Unsupervised learning of human action categories using spatial-temporal words', *International Journal of Computer Vision*, Vol. 79, No. 3, pp. 299-318.

Perikos, I. and Hatzilygeroudis, I. (2016), 'Recognizing emotions in text using ensemble of classifiers', *Engineering Applications of Artificial Intelligence*, Vol. 51, pp. 191-201.

Sipahi, D., Dalkılıç, G. and Ozcanhan, H. (2015), 'Detecting spam through their Sender Policy Framework records', *Security Communication Networks*, Vol. 8, No. 18, pp. 3555-3563.

Tang, B., He, H., Baggenstoss, P.M. and Kay, S. (2016), 'A Bayesian classification approach using class-specific features for text categorization', *IEEE Transactions on Knowledge and Data Engineering*, Vol.28, No.6, pp. 1602-1606.

Terribilini, M., Sander, J.D., Lee, J.H., Zaback, P., Jernigan, R.L., Honavar, V. and Dobbs, D. (2007), 'RNABindR: a server for analyzing and predicting RNA-binding sites in proteins', *Nucleic Acids Research*, Vol. 35, pp. 578-584.

Tripathy, A., Agrawal, A. and Rath, S.K. (2016), 'Classification of sentiment reviews using n-gram machine learning approach', *Expert Systems with Applications*, Vol. 57, pp. 117-126.

Trovato, G., Chrupala, G. and Takanishi, A. (2016), 'Application of the naive Bayes classifier for representation and use of heterogeneous and incomplete knowledge in social robotics', *Robotics*, Vol. 5, No. 1, pp. 6-26.

Turhan, B. and Bener, A. (2009), 'Analysis of naive Bayes' assumptions on software fault data: an empirical study'. *Data and Knowledge Engineering*, Vol. 68, No. 2,

pp. 278-290.

Wong, T.T. (1998), ‘Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, Vol. 97, No. 2-3, pp. 165-181.

Wong, T.T. (2009), ‘Alternative prior assumptions for improving the performance of naive Bayesian classifiers’, *Data Mining and Knowledge Discovery*, Vol. 18, No. 2, pp. 183-213.

Wong, T.T. (2014), ‘Generalized Dirichlet priors for Naïve Bayesian classifiers with multinomial models in document classification’, *Data Mining Knowledge Discovery*, Vol. 28, No. 1, pp. 123-144.

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 Naive Bayes classifier’, *Bioinformatics*, Vol. 22, No. 11, pp. 1325-1334.

Zhang, H., Yu, P., Zhang, T.G., Kang Y.L., Zhao, X., Li, Y.Y., He, J.H. and Zhang, J. (2015), ‘In silico prediction of drug-induced myelotoxicity by using Naïve Bayes method’, *Molecular Diversity*, Vol. 19, No. 4, pp. 945-953.

Zhang, J., Chen, C., Xiang, Y. and Zhou, W. (2013), ‘Internet traffic classification by aggregating correlated Naive Bayes predictions’, *IEEE Transactions on Information Forensics and Security*, Vol. 8, No. 1, pp. 5-15.

Zhou, S., Li, K. and Liu, Y. (2009), ‘Text categorization based on topic model’, *International Journal of Computational Intelligence Systems*, Vol. 2, No. 4, pp. 398-409.