應用簡單貝氏分類於失智預診斷與應用程式界面之開發

蘇偉亮[1] 洪子倫[1] 藍祚鴻[2, 3] 郭登堯[1]

[1]逢甲大學 應用數學系

[2]國立陽明大學 醫學系

[3]臺中榮民總醫院 精神部

摘要 —— 失智症是一種因腦部病變或損傷所導致的漸進性認知功能退化，而且退化的幅度遠高於一般正常老化的速度。失智症的發生，受影響的除了病患與他們的家人，對於國家的財政也影響巨大。早期發現失智症，就能給予相應的治療與照顧。本研究使用 Naive Bayes 分類作爲建模的方法，訓練後的模型準確率約為 82%，最後再配合 Matlab GUIDE 的工具，嵌入該模型成爲一個應用程式界面，方便操作。未來透過這樣一個失智症的篩檢系統，讓醫事人員能快速進行失智症的初步診斷。

關鍵字：失智症、Naive Bayes 、GUIDE

一、 簡介

失智症是由於大腦中的損傷或疾病所引起認知功能退化的一種病變，即大腦的退化速度比同齡的正常人更快。失智症不是一種病，而是影響腦部功能的症候群，其中阿茲海默症最爲常見。它會影響記憶、方向、學習、語言等方面，如果病情嚴重，可能會伴隨著行爲障礙、人格變化、幻覺等，進而影響到日常的生活與工作的表現。

失智症可分爲可逆和不可逆兩種情況。其中可逆的情況只占不到 10%，即只有不到 10%的失智症患者可透過適當的治療來逆轉病情，而阿茲海默症就是屬於不可逆的情況。根據世界衛生組織 (WHO) 的統計，目前全世界約有 4700 萬失智症患者，而每年約有 990 萬的新增病例，即每 3 秒就會出現一個新的病例。據預測，2030 年的失智症患者總數將達到 7500 萬人，而 2050 年將會高達 1.32 億人。其中生活在中低收入國家的居民佔大比例。可以相信這龐大的社會成本、醫療費用和非正式護理費用將會對各國財政影響甚大[1]。

雖然失智症主要影響老年人，但這並不是老年化的正常現象。有一部分的病患是因爲家族遺傳而患有失智症，而他們大部分多在 65 歲前被發現，稱爲早發性失智。失智症的發生除了病患自身需承受負擔外，還有照顧他們的家人也將面臨巨大的壓力，

其中包括情感及經濟上的壓力。失智症的發生會使得病患認知能力衰退，這將影響病患失去基本的日常生活能力。比起老年患者，早發性失智患者在家庭經濟的壓力影響更爲明顯。此外，如果失智症患者沒有得到家人的妥善照顧和長期支持，甚至可能會造成病情的惡化。

本研究採用機器學習中常見的監督式學習——簡單貝氏分類器 (Naive Bayes Classifiers)作爲初步鑑別失智症的方法。簡單貝氏分類最大的好處就是假設各特徵因子是彼此獨立的，而簡單貝氏分類所需要的參數可利用最大概似估計方法 (Maximum Likelihood Estimation)來求出，進而建構出一個機率模型，然後利用 Matlab 的使用者界面開發環境 (GUIDE)，將訓練完成的模型嵌入，並設計成一個應用程式界面，方便醫事人員進行初步失智篩檢[2]。

爲了建構失智症的預測模型，我們需要收集與失智症相關的危險因子作爲訓練模型的特徵因子。本研究的特徵因子有：性別、年齡、教育程度 (受教育的年數)、高血壓、心臟病、糖尿病與腦中風，並加上常用於評估認知障礙的簡易智能量表 (MMSE)。而我們也利用臨床失智評估量表 (CDR)來作爲模型的預測目標，探討該模型的準確性。

由於臨床數據不易收集，爲了提高準確度，後續工作將持續進行數據的收集與訓練，並將訓練完成的模型嵌入我們設計的應用程式界面中，使得相關的醫事人員能方便使用，並作爲一種早期預防的輔助診斷工具。

二、 文獻回顧

a) 特徵因子

失智症是一種綜合症狀，主要是認知功能受損所導致。然而認知功能受損的原因，可能是疾病所引發或生活環境所導致。過去的研究顯示，年齡與失智症存在關聯性，失智症的發生率會隨著年齡的增長而增加[3-5]。在性別方面，女性在失智症的發生率比男性高一些。研究也顯示，女性在阿茲海默症的發生率較高，而男性在血管性失智的發生率較高[6]。

另一個失智的危險因子就是教育程度。由於生活環境的影響，較低的教育成果可能會有更高的患病機會[7]。此外，受過高等教育的人們可有效的降低患有失智症的可能[8,9]。

血管性失智是阿茲海默症外最常見的失智症。主要是因爲中風而導致大腦血液供應不足所致，因而導致腦細胞死亡，造成大腦皮層的損傷，使得記憶、注意力、思維、語言等方面受影響。根據研究指出，血管性失智的危險因子包括年齡[3,4,6,10]、高血壓[10-14]、糖尿病[10,12,13,15]、心血管疾病[10,14,16]和腦血管疾病[12,16,17]。

簡易智能量表 (MMSE)是用於檢查認知衰退的問卷測驗。也因爲它執行簡單，所以 MMSE 已經成爲臨床和研究領域廣泛使用的認知篩檢工具。醫生可讓患者透過完成 MMSE，來記錄且瞭解患者對治療反應的成果。

根據以上的研究報告，我們選定了 8 個危險因子作爲我們預測模型中的特徵因子，包括：性別 (Gender)、年齡 (Age)、受教育的年數 (Edu)、高血壓 (HighBP)、心臟病 (HD)、糖尿病 (DM)、腦中風(CVA)和簡易智能量表(MMSE)。

b) Naive Bayes Model

假定有一筆由 m 個特徵因子所組成的數據 $x = \{f_1, f_2, \dots, f_m\}$ ，在 n 個類別 $C = \{c_1, c_2, \dots, c_n\}$ 中將其分類。可表示為一個條件機率式子：

$$p(c_i|x) = \frac{p(c_i) * p(x|c_i)}{p(x)} = \frac{p(c_i) * p(f_1, f_2, \dots, f_m|c_i)}{p(f_1, f_2, \dots, f_m)} \quad , \quad i = 1,2, \dots, n$$

目標是找出 $p(c_i|x)$的最大值，如果 $p(c_k|x) = \max\{p(c_1|x), p(c_2|x), \dots, p(c_n|x)\}$ ，則 $x \in c_k$ 。因分母是一個常數，所以只要關心分子的部分，即要找出 $p(c_i) * p(x|c_i)$ 的最大值。在 Naive Bayes 的基礎假設下，各特徵因子是條件獨立，所以有：

$$\begin{aligned} p(c_i|x) \propto p(c_i) * p(x|c_i) &= p(c_i) * p(f_1|c_i) * p(f_2|c_i) * \dots * p(f_m|c_i) \\ &= p(c_i) * \prod_{k=1}^{m} p(f_k|c_i) \end{aligned}$$

其中 $p(c_i)$ 為第 i 個類別的先驗機率（已知），而 $p(f_k|c_i)$ 為在第 i 個類別中，第 k 個特徵因子的機率密度函數。而 $p(f_k|c_i)$ 服從的分配所需的參數可利用最大概似估計方法來求出[2,18]。

最後將收集到的數據，根據上述過程進行訓練，便可建構出 Naive Bayes Model 來進行分類。實際上，Naive Bayes 的方法也常廣泛應用在文字信息檢索中。因爲它可快速地訓練出模型進行分類，且分類效果也表現良好[19]。

三、 研究過程

a) 分層隨機抽樣

本研究共募集到 169 名受試者的資料，其中包括了 8 個特徵因子的資料和用來作爲預測因子的 CDR，如表 1。CDR 有 5 種評分結果，CDR 分數為 0（無失智）和 0.5（可能失智）代表無失智症(class 1：無失智) ，CDR 分數為 1（輕度），2（中度）和 3（重度）代表有失智症(class 2：失智確診) ，見表 1。

將收集到的資料隨機抽取 118 位（70%）作為訓練集，51 位（30%）作為測試集。由於無失智與失智確診人數差異很大，所以若採用簡單隨機抽樣，很可能使得訓練集內的無失智與失智確診人數比例過於偏頗，而無法代表全體樣本。因此，最好的方式是採用分層隨機抽樣，按無失智與失智確診的人數比例約 23：77（i.e. $\frac{39}{169}:\frac{130}{169}$）來抽取。無失智部分從 39 個樣本中隨機抽出 27 個（i.e. $118 * 0.23$），而失智確診部分從 130 個樣本中隨機抽出 91 個（i.e. $118 * 0.77$）。此$118（ = 27 + 91）$筆資料便構成我們需要的訓練集。建立好訓練集和測試集後，便可開始訓練模型。

	Variable	**Subgroup**	**Frequencies/Means**
Demographic Factor	Gender	Male Female	87 (51.48%) 82 (48.52%)
	Age (years)		78.85 ± 9.27
	Edu (years)		7.14 ± 5.19
Mental Status	MMSE		17.12 ± 6.84
Disease History	HighBP	HighBP Non-HighBP	53 (31.36%) 116 (68.64%)
	HD	HD Non-HD	44 (26.04%) 125 (73.96%)
	DM	DM Non-DM	39 (23.08%) 130 (76.92%)
	CVA	CVA Non-CVA	24 (14.2%) 145 (85.8%)
Predictive Target	CDR	0,0.5 (Class 1) 1,2,3 (Class 2)	39 (23.08%) 130 (76.92%)

表 1　169 名病患之統計資料

b) 訓練模型

假設收集到的樣本數據類別 1 與 2 皆服從高斯分佈，各類別的機率密度函數所需要的參數，即平均值和標準差，利用最大概似估計方法來求出，再計算出各類別的先驗機率後，經過一些運算，模型便訓練完成。

隨後我們將利用測試集來計算該模型的準確性、敏感性與特異性。同時我們將重複做 10000 次的抽樣實驗，找出抽樣分佈的平均值（i.e. 準確性），並建立對應的模型。

c) Matlab 使用者界面環境

利用 Matlab 的 GUIDE 開發應用程式界面，接著將我們訓練好的模型嵌入，以建立一個失智症篩檢系統。設計的界面裏，我們將 8 個特徵因子作爲使用者輸入的參數。輸入完畢，執行“檢測”鍵，即可預測受測者是否有失智的可能。

四、 研究結果

我們利用測試集來檢測模型的準確率，並重複 10000 次的實驗，圖 1 所示為 10000 次隨機實驗的準確率之直方圖，其中準確率最高可達到約 98%。該直方圖也呈現高斯分佈的現象，所以我們可知道約 95%的機率，準確率會落在±2個標準差內，即 72%到 92%之間。則當樣本夠大時，預測準確率高達 82%。

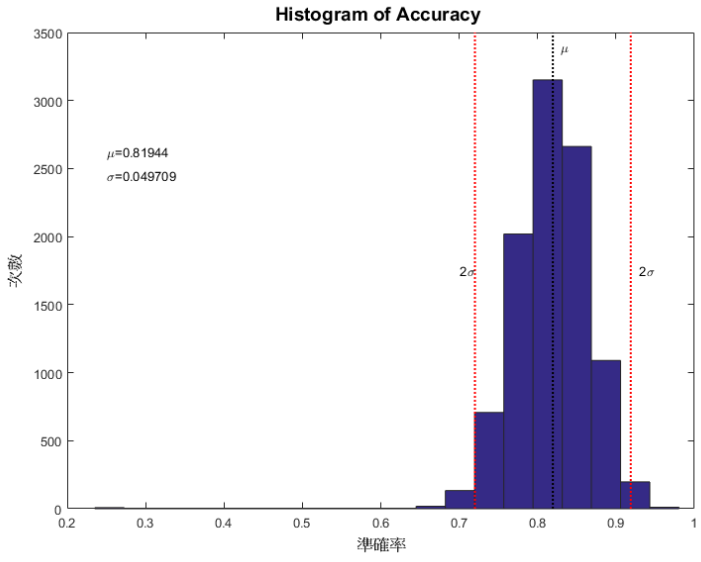

圖 1　　10000 次隨機實驗的準確率之直方圖

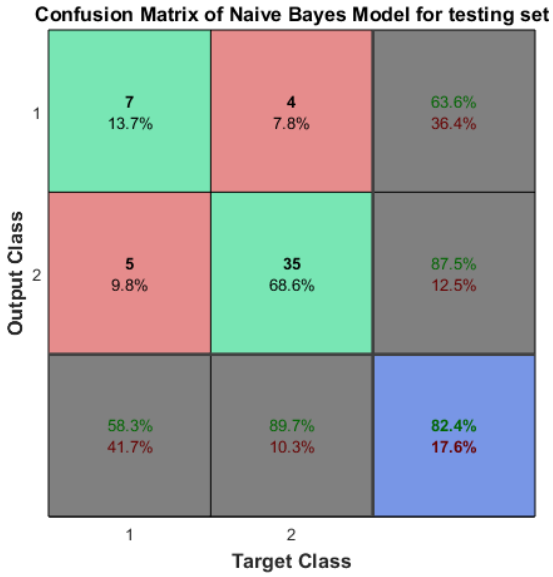

圖 2　　測試集的 confusion matrix

圖 2 是利用準確率 82.4%的模型，x 軸代表實際類別，y 軸代表預測類別，其中類別 1 為無失智症，類別 2 為有失智症。從圖中我們可知道該測試集的準確性為 82.4%，敏感性為 89.7%，而特異性為 58.3%。

利用 Matlab 的 GUIDE 工具，製作成圖形使用者界面以方便使用者進行失智症的篩檢。圖 3 為失智預診斷應用程式界面，而圖 4 與圖 5 分別是檢查結果為無失智症與有失智症所呈現的結果。

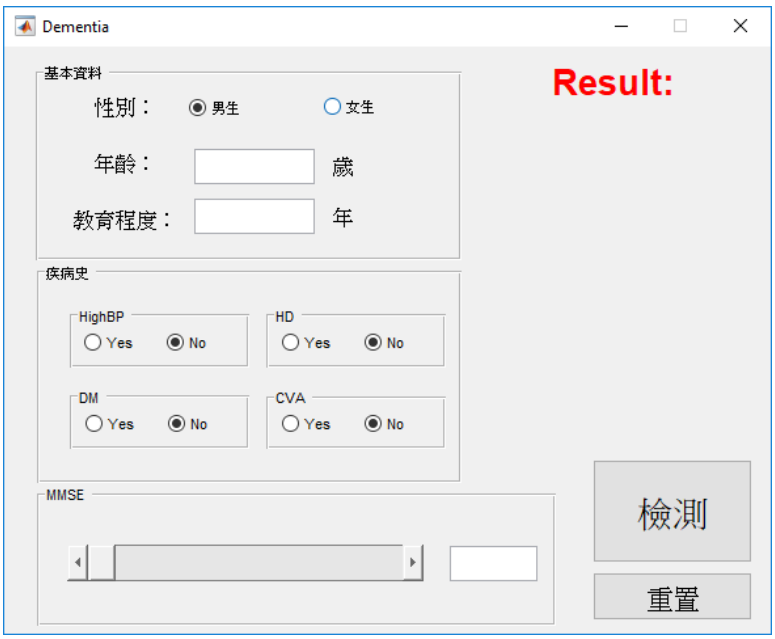

圖 3　　失智預診斷應用程式界面

圖 4　　預診斷結果為無失智之可能

圖 5　　預診斷結果為有失智之可能

五、 結論

由於臨床資料收集不易，加上 CDR 各個分數級別的人數差異很大，因此本研究的模型只能用來預測是否有失智症之可能。未來我們可以收集更多的資料來作改善，希望模型可完全用來預測 CDR 的評分。同時我們也可增加其他與失智症有關的危險因子，嘗試改善預測模型的準確性。

失智症的發生不僅患者與他們的家人受影響，對於國家、社會、財政與經濟也一樣。我們希望透過一個易使用的應用程式來幫助醫事人員進行快速的檢測。提早發現失智症，給予相對應的治療與照顧，對於患者而言絕對是一件喜訊。

六、 參考文獻

[1] http://www.who.int/mediacentre/factsheets/fs362/en/

[2] Konstantinos Koutroumbas and Sergios Theodoridis, Pattern Recognition, 4th Edition, 2009, 59-61.

[3] Ritchie K. and Kildea D., Is senile dementia "age-related" or "ageing-related"? -- evidence from meta-analysis of dementia prevalence in the oldest old, The Lancet, 1995 346(8980), 931-934.

[4] Sujuan Gao, Hugh C. Hendrie, Kathleen S. Hall and Siu Hui, The Relationships Between Age, Sex, and the Incidence of Dementia and Alzheimer Disease: A Meta-analysis, Arch. Gen. Psychiatry, 1998 55(9), 809-815.

[5] Kaarin J Anstey, Richard A Burns, Carole L Birrell, David Steel, Kim M Kiely and Mary A Luszcz, Estimates of probable dementia prevalence from population-based surveys compared with dementia prevalence estimates based on meta-analyses, BMC Neurology, 2010 10(62).

[6] Benjamin J Sadock and Virginia A Sadock, Kaplan & Sadock's Concise textbook of clinical psychiatry (3rd ed.). Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins, 2008, p. 52.

[7] Margaret Gatz, James A. Mortimer, Laura Fratiglioni, Boo Johansson, Stig Berg, Ross Andel, Michael Crowe, Amy Fiske, Chandra A. Reynolds, Nancy L. Pedersen, Accounting for the relationship between low education and dementia: A twin study, Physiology & Behavior 2007 92, 232-237.

[8] Valenzuela MJ, Sachdev P., Brain reserve and dementia: a systematic review, Psychol Med. 2006 36(4):441-54.

[9] L Fratiglioni, HX Wang, Brain reserve hypothesis in dementia, J. Alzheimers Dis., Volume 12, Number 1 / 2007 11-22.

[10] Réjean Hébert, Joan Lindsay, René Verreault, Kenneth Rockwood, Gerry Hill, and Marie-France Dubois, Vascular Dementia Incidence and Risk Factors in the Canadian Study of Health and Aging, Stroke 2000 31, 1487-1493.

[11] Vera Novak and Ihab Hajjar, The relationship between blood pressure and cognitive function, Nat. Rev. Cardiol. 2010 7, 686-698.

[12] R. Sahathevan, A. Brodtmann, and G. A. Donnan, Dementia, stroke, and vascular risk factors; a review, Int. J. Stroke 2012 7(1), 61-73.

[13] Philip B. Gorelick, Risk Factors for Vascular Dementia and Alzheimer Disease, Stroke 2004 35, 2620-2622.

[14] John S. Meyer, Gaiane M. Rauch, Ronald A. Rauch, Anwarul Haque, and Kate Crawford, Cardiovascular and Other Risk Factors for Alzheimer's Disease and Vascular Dementia, Annals of the New York Academy of Sciences 2000 903, 411-423.

[15] Linda B. Hassing, Boo Johansson, Sven E. Nilsson, Stig Berg, Nancy L. Pedersen, Margaret Gatz, and Gerald McClearn, Diabetes Mellitus Is a Risk Factor for Vascular Dementia, but Not for Alzheimer's Disease: Population-Based Study of the Oldest Old, Int. Psychogeriatr. 2002 14(3), 239-248.

[16] Gustavo C Román, Vascular dementia may be the most common form of dementia in the elderly, J. Neurol. Sci. 2002 203-204, 7-10.

[17] Salka S. Staekenborg, Wiesje M. van der Flier, Elisabeth C.W. van Straaten, Roger Lane, Frederik Barkhof, Philip Scheltens, Neurological Signs in Relation to Type of Cerebrovascular Disease in Vascular Dementia, Stroke 2008 39, 317-322.

[18] https://en.wikipedia.org/wiki/Naive_Bayes_classifier/

[19] Domingos P., Pazzani M.“Beyond independence: Conditions for the optimality of the simple Bayesian classifier,”Machine Learning,Vol. 29,pp. 103–130,1997.