BPSO-KNN 預測骨質疏鬆症

莊麗月
義守大學
化學工程系
chuang@isu.edu.tw

吳國銓
國立高雄應用科技大學
資訊工程系
1097308101@cc.kuas.edu.tw

張學偉
高雄醫學大學
生物醫學暨環境生物系
changhw@kmu.edu.tw

楊正宏
稻江科技暨管理學院
網路系統學系
國立高雄應用科技大學
電子工程系
chyang@cc.kuas.edu.tw

***摘要*—透過生物資訊學分析單核苷酸多型性在正常個體與病患間的關聯性，這種人類多基因遺傳病變的研究不僅可找出影響骨質疏鬆症、糖尿病或乳癌等常見重大疾病的致病基因，更可針對個人體質不同實施個人化之疾病預防與治療。因此本文利用資料探勘方法，針對骨質疏鬆症資料集進行特徵選取及分類問題。依據本文收集之 304 名患有或非骨質疏鬆症之婦女的數據資料。其中包含年齡、是否已過更年期及十一個可能與骨質疏鬆症相關的單核苷酸多型性，共十三個屬性。我們使用粒子族群最佳化作為特徵選取及參數最佳化，並以 K 最近鄰居法為分類方法，進行骨質疏鬆症之預測，並與其他分類方法做比較。結果顯示，本研究方法對於骨質疏鬆症有 73%以上預測正確率，其正確率不但優於其他分類方法，且能有效挑選出重要性較高的單核苷酸多型性。**

***關鍵詞*—單核苷酸多型性、骨質疏鬆症、資料探勘、粒子族群最佳化、K 最近鄰居法**

一、前言

人類基因體計畫的進行，目的是為完全解讀人類遺傳圖譜，破解人類基因密碼，最終得以解讀基因體核苷酸序列，並註解所有人類基因功能。目前，已步入後基因體世代，主要工作是利用序列探索該資訊涵義。其中，分析人類基因序列變異如研究單核苷酸多型性(Single Nucleotide Polymorphisms, SNPs)對於人類疾病關聯性分析，再加上最近研究顯示 SNP 為人類相關性疾病研究之重要指標[8, 21]。因此，利用 SNP 資料來研究遺傳流行病學的複雜疾病成為近來相當熱門領域。透過生物資訊學分析 SNP 在正常個體與病患間的關聯性，這種人類多基因遺傳病變的研究不僅可找出影響骨質疏鬆症(Osteoporosis)、糖尿病或乳癌等常見重大疾病的致病基因，更可針對個人體質不同實施個人化之疾病預防與治療。

SNP 是由 E. Lander 於 1996 年所提出，被稱為「第三代 DNA 遺傳標記」，在同一物種而不同個體間基因體內某單一核苷酸不相同，其在群體間分佈的對偶基因比率至少大於 1%。SNP 為人體基因中單一個核苷酸改變所造成，故被視為人體基因多樣性主要原因之一。每個人基因組並不完全相同是造成基因差異表現的原因，除了一段基因在染色體上排列位置的改變、插入或刪除核苷酸序列的突變或變異外，不同個體間 SNP 亦是一重要影響因子。

資料探勘中分類問題廣泛應用於各領域，其分類問題乃依據資料變數的值加以計算，並依結果作分類或預測的動作。許多機器學習方法被提出，如：K 最近鄰居法(K Nearest-Neighbor, KNN) [1]、貝氏分類器(Naïve Bayes, NB) [10]、支持向量機(Support Vector Machine, SVM) [17]及隨機森林(Random Forest, RF) [4]，並廣泛應用於各領域，如鑑定決策、影像辨識、電力負載預測、藥物診斷及市場行銷等[24]。而改善分類

問題方法陸續被提出，其中較熱門的方法為特徵選取。特徵選取主要目的為：1.在每筆資料中，有些特徵會影響到分類的正確性或是多餘的，經特徵選取後反而可能會提升分類的正確性；2.於特徵集合 D 中挑選子集合 $d(D>d)$作分類辨識，在減少維度的情況下可節省運算時間，因此在分類辨識前，特徵選取將會是一個非常重要的前置處理。特徵選取於分類在最佳化的標準下，由一個原始的特徵集合 D，經過挑選利於分類的辨識後，產生一個特徵子集合 d，以作為機器學習、圖形辨識、調適性控制[16]等用途。而特徵選取方法大致上可分為兩種：1)Filters：主要是針對單一特徵進行重要性的測量，將所有重要性較高的特徵組合成特徵子集合，較常用的方法有資訊增益及交互資訊等[2]。2)Wrappers：藉由持續的加入或刪除特徵，利用演算法中的目標函數評估特徵集合，再利用演算法中的特性搜尋最佳的子集合。近年來，有許多不同類型的特徵選取方法被發展出來，包含窮舉演算法、分枝界限演算法[15]、循序搜尋演算法、基因演算法[16]及粒子族群最佳化(Particle swarm optimization, PSO)[5]等。

本研究收集 304 個患有骨質疏鬆症及非骨質疏鬆症樣本，其中包含年齡、是否已過更年期及十一個可能與骨質疏鬆症相關 SNP，共十三個屬性。以現有樣本利用資料探勘分類問題之方法，以達到預測效果。為分析所有樣本中各個屬性與骨質疏鬆症之關聯性或重要性，亦加入特徵選取方法挑選出較重要屬性，進行分類預測。因此本文以 K 最近鄰居法作為分類方法，粒子族群最佳化作為特徵選取方法，並最佳化 K 最近鄰居法之重要參數 K，使粒子族群最佳化達到特徵選取及參數最佳化之目的，並獲得較好的分類效果。最後，我們利用 Weka [24]中現有分類方法進行驗證及比較。以下將詳述本文的研究方法：粒子族群最佳化、K 最近鄰居法與粒子族群於特徵選取及參數最佳化。

二、研究方法

(一) 粒子族群最佳化

粒子族群最佳化[11]由 Eberhart 和 Kennedy 兩位學者提出，藉著觀察鳥群和魚群在自然界中覓食習性，引發構想設計出一套演化式最佳化演算法。在一個社會化的族群中，每一個個體的行為不但會受其過去經驗和認知的影響，也同時受到整體社會行為影響。在粒子族群最佳化中每個體在搜尋空間中擁有各自的位置 X 及速度 V，根據自我經驗與族群行為，進行機率式的搜尋策略。在粒子族群最佳化中，每個粒子均視為一個解，各自搜尋解經驗中，個體最佳經驗稱之為 *pBest*，而在所有粒子中最佳經驗稱之為 *gBest*。根據這兩種經驗來決定飛行速度及移動方向，並決定所在位置。然而為了使粒子族群最佳化能解離散問題，由實數編碼改為二進制編碼，稱二進制粒子族群最佳化(Binary Particle Swarm Optimization, BPSO)[12]。在 BPSO 中，假設 N_{dim} 為問題空間維度(即搜尋空間 $\Re^{N_{\dim}}$)，有 P 個粒子，每個粒子有其位置 $X_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$及速度 $V_i = \{v_{i1}, v_{i2}, \ldots, v_{id}\}$，其中 $i = 1, 2, \ldots, P$，$d = 1, 2, \ldots, N_{dim}$，$x_{id} \in \{0, 1\}$，$v_{id} \in [-v_{max}, v_{max}]$。其更新公式如下：

$$V_{id}^{new} = w \cdot V_{id}^{old} + c_1 \cdot rand_1(\ ) \cdot (pBest - X_{id}^{old}) + c_2 \cdot rand_2(\ ) \cdot (gBest - X_{id}^{old}) \quad (1)$$

$$S(V_{id}^{new}) = \frac{1}{1+e^{-V_{id}^{new}}} \quad (2)$$

$$if\ (rand_{id} < S(V_{id}^{new})) \quad than \quad X_{id}^{new} = 1, \quad else \quad X_{id}^{new} = 0 \quad (3)$$

其中 V_{id}^{new} 代表粒子更新後速度，V_{id}^{old} 代表粒子更新前速度。X_{id}^{new} 代表粒子更新後位置，X_{id}^{old} 則代表粒子更新前位置。$(pBest - X_{id}^{old})$ 是粒子本身最佳位置和粒子當前所在位置間之距離，

$(gBest - X_{id}^{old})$ 是截至目前迭代為止之粒子最佳位置和粒子當前所在位置間之距離。w 為慣性權重，一般介於 0.9 到 1.2 之間。c_1 和 c_2 為學習因數，一般均設定為 2 [22]。$rand_1(\,)$、$rand_2(\,)$、$rand_{id}$ 皆為介於 0 至 1 之間均勻分佈的隨機變數。在更新公式裡，$S(\,)$為一個 Sigmoid 函數，粒子移動速度 V 為粒子位置 X 改變為 1 或 0 的機率，因此經由 Sigmoid 函數之轉換後，使 V 介於 0 到 1 之間的機率數值，若使 $v_{max} = 6$ 可使 v_i 介於 0.9975 至 0.0025 之間，再藉由隨機產生之亂數 $rand_{id}$ 來判斷粒子位置是否改變，如公式(3)所示，其粒子族群最佳化之虛擬碼如下所示。

Begin
  Initialize particle swarm by randomly
  While (stopping criterion is not met)
    Evaluate fitness of particle swarm
    Update *pBest* and *gBest*
    Update *X* and *V* of particle swarm
  Next generation until stopping criterion
End

(二) K 最近鄰居法

K 最近鄰居法是由 Fix 和 Hodges 在 1951 年所提出的[6]。在 K 最近鄰居法的訓練資料中，每一個資料點都依照本身的特徵維度被定義在一個 D-維的空間中，而 K 所代表的是測試資料在 D-維空間裡所尋找的 K 個最近的鄰居，K 最近鄰居法的分類效果就是受到這 K 個鄰居的數量影響[6]，而鄰居的計算方式是根據歐幾里德距離(或是馬式距離)，利用此種方式針對測試資料與訓練資料進行相似性的量測，統計 K 個鄰居中各類別出現的頻率，將測試資料的類別定義為出現頻率最高之類別，藉此達到分類的目的。假設 m 筆訓練資料(x_i, y_i)以及測試資料 $\boldsymbol{x}$，其中 $i = 1, 2, \ldots, n$，n 為資料量；y_i 為資料 x_i 之類別；$\boldsymbol{x}$ 為特徵向量。距離的量測值定義如下：

$$d(x,\ x_i) = \sqrt{\sum_{j=1}^{d}(x_j - x_{ij})^2} \tag{4}$$

其中，d 為特徵向量的維度。最近鄰居法規則為 $\mathrm{nnr}(\boldsymbol{x}) = y_k$，其中 $k = \arg\min_i d(\boldsymbol{x}, \boldsymbol{x}_i)$。當最近鄰居法參數 K > 1 則利用投票策略。例如當 K = 3 時，計算出測試資料與訓練資料中三個最小的距離量測值。假設類別 A 包含兩個最小的距離量測值，而類別 B 只有一個，則判斷為 A 類別，其 K 最近鄰居法之虛擬碼如下所示。

Begin
  For i = 1 to number of test set
    For j = 1 to number of train set
      Calculating distance of test with train set
    Next j
  Next i
  For k = 1 to number of parameter K
    Determine class of test set by vote strategy
  Next k
  Determine the classification accuracy
End

(三) 粒子族群於特徵選取及參數最佳化

本研究主要提出以粒子族群最佳化找出最佳特徵集合及 K 最近鄰居法之參數 K，利用 K 最近鄰居法所計算出預測正確率作為適應函數值。以下詳細介紹本研究的方法：方法之流程及架構、粒子編碼、族群初始化及適應函數。

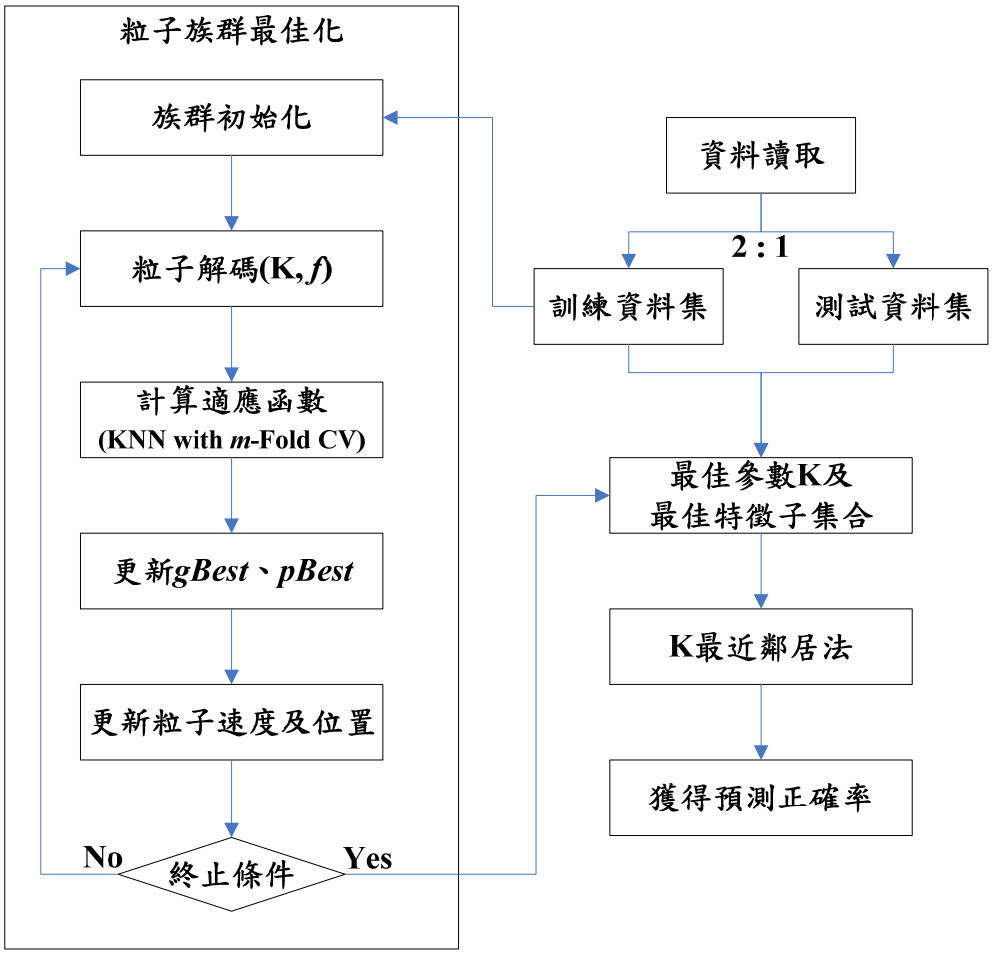

圖 1 方法流程圖

1. 流程及架構

粒子族群最佳化找出最佳特徵子集合及 K 最近鄰居法參數 K，基本流程如圖 1 所示，其流程簡述如下：

第一部份 - 資料處理

將資料隨機分成訓練及測試資料集(訓練與測試資料比為 2：1)，由訓練資料集利用演算法透過 *m*-Fold 交叉驗證法(*m*-Fold cross validation, *m*-Fold CV) [23]進行正確率評估，找出最佳特徵子集合及 K 最近鄰居法參數 K。之後再利用測試資料驗證演算法效能。

第二部份 - 演算法

1) 族群初始化：隨機產生粒子的位置及速度。

2) 粒子解碼：將粒子位元解碼，粒子前 n_k 個位元(其中 $n_k \in \mathbf{N}$)，代表 K 最近鄰居法之參數 K (即 $K \in \{1, 3, \ldots, (2\times 2^{n_k}-1)\}$)；其餘位元為特徵，其解碼方式即當位元為 0 時表示此特徵未選取，反之則表示此特徵被選取。

3) 計算適應函數：以 *m*-Fold 交叉驗證法利用 K 最近鄰居法獲得之正確率作為適應函數值。

4) 更新 *pBest* 及 *gBest*：粒子目前位置之適應值，與本身及群體最佳值比較，若當前解比本身最佳解好，則當前解為 *pBest*。若此 *pBest* 為所有群體最佳解，則當前解為 *gBest*。

5) 更新粒子目前位置：利用公式(1)~(3)更新粒子之速度及位置。

6) 停止條件：當迭代次數達設定次數則停止；否則跳到步驟 3，直到符合停止條件為止。

7) 最佳參數及特徵：經過粒子族群最佳化後，會得到一組最佳化解，其中包含 K 及特徵子集合。

2. 粒子編碼

由於考慮 K 最近鄰居法在參數 K 會因為資料分佈的關係，而使分類效果不同。因此我們在編碼上除了資料特徵外，另外亦加入 K 最佳鄰居法之參數 K，每一個粒子之編碼方式如圖 2 所示。

$X_k^1 \ldots X_k^i \ldots X_k^{n_k}$	$X_f^1 \ldots X_f^i \ldots X_f^{n_f}$

圖 2 粒子編碼示意圖

$X_k^1 \sim X_k^{n_k}$ 為參數 K 之編碼；$X_f^1 \sim X_f^{n_f}$ 為特徵之編碼。其中 n_k 是參數編碼位元的長度，在本文設為 3 位元(即 K∈{1, 3, …, 15})。而 n_f 則是特徵數的位元長度，依據資料的形態而有所變動，如本文資料特徵數為 13。

3. 族群初始化

依設定的族群數 P 及編碼長度 l，利用隨機方式產生 P 個粒子(即 P 個解)，產生之位元字串由{0, 1}所組成，每個粒子之初始速度 V 為隨機產生[0, 1]間的數值。

4. 適應函數

利用每個粒子位置，取出資料的特徵子集合及 K 最近鄰居法之參數 K，利用訓練資料集以 m-Fold 交叉驗證法進行訓練及預測，最後獲得預測正確率作為每個粒子的適應函數值，如下所示。

$$fitness(x_{id}) = Accuracy_{\text{KNN with m-Fold cross validation}} \quad (5)$$

三、結果與討論

(一) 實驗描述

本研究實驗資料共收集 50 名停經前(平均年齡 43 歲)和 254 名停經後婦女(平均年齡 59 歲)參與了這項研究(停經後婦女的定義是超過 6 個月沒有月經的發生或年齡超過五十歲) [13]。樣本屬性分別為年齡、是否停經及 11 個 SNP (如表 1 所示)共 13 個特徵。而 SNP 基因型(genotype)為字母型態，其資料型態轉換由表 1 可知，例如 SNP_1 基因型裡 TT = 1、CT = 2、CC = 3。

實驗中，資料隨機分成訓練及測試資料集(訓練、測試資料比為 2：1)，在訓練過程中，透過 m-Fold 交叉驗證法進行正確率評估。m-Fold 交叉驗證法將資料以隨機方式平均分成 m 個部份，將 m 個部份中每一部份獨立做為測試集合，而其餘 m-1 個部份做為訓練集合，當訓練集合訓練出支持向量機的雛型時，再利用測試集合進行評估，如此交叉驗證後即可獲得正確率之評估標準。一般 m 值依資料樣本數為依據，若資料樣本數過小，m 設定愈大愈好，如此可使訓練樣本數變多[19]，本文 m 值設定為 10。

(二) 正確率評估

本研究採用醫學診斷最常使用的評估方式，分別為：陽性猜中率(Positive hit rate)，即敏感度(Sensitivity)、陰性猜中率(Negative hit rate)，即特異度(Specificity)及正確率(Accuracy rate)。若正確預測出有病稱真陽性(True Positive, *TP*)，然而，當預測沒病但實際上有病則稱偽陰性(False Negative, *FN*)。相對地，若正確預測出沒病稱真陰性(True Negative, *TN*)，當預測有病但實際上沒病則稱假陽性(False Positive, *FP*)。在資料探勘領域裡，一般正確率之公式如下：

$$Accuracy\ rate = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)$$

表 1 SNP 資料型態

SNP	Chromosome	Gene (location)	rs number	Genotype		
				1	2	3
1	6	TNFα-857	rs1799724	TT	TC	CC
2	19	TGFβ1-509	rs1800469	TT	TC	CC
3	1	Osteocalcin	rs1800247	CC	CT	TT
4	6	TNFα-308	rs1800629	AA	AG	GG
5	11	PTH (*BstB Ⅰ*)	rs6254	GG	AG	AA
6	11	PTH (*Dra Ⅱ*)	rs6256	AA	AC	CC
7	2	IL1_ra[b]	VNTR[a]	A1A1	A1A2	A1A4
8	6	HSP70 hom	rs2227956	CC	CT	TT
9	6	HSP 70-2	rs1061581	GG	AG	AA
10	7	CTR	rs1801197	CC	CT	TT
11	14	BMP-4	rs17563	CC	CT	TT

Legends: [a] Variable number tandem repeats; [b]IL1_ra genotype: A1, 410 bp; A2, 240bp; and A4, 325 bp; Data source [13].

其可評估分類器的正確率[24]。對於分類器而言以 *TP*、*FP* 較為重要[26]，而敏感度與特異度則是區分分類器對於有病或沒病的效果。敏感度是正確預測有病的比例，其公式為 $P(T^+|D^+) = TP/(TP+FN)$。特異度是正確預測沒病的比例，其公式為 $P(T^-|D^-) = TN/(TN+FP)$。

(三) 結果

本研究利用 Weka 中現有分類器，包含 K 最近鄰居法、C4.5 演算法、支持向量機、隨機森林及貝氏分類器，資料以 2：1 比例隨機分成訓練及測試資料進行預測，每個方法均實驗 10 次，取平均進行結果比較。表 2 為本研究實驗結果，由表可知在不同的訓練資料集，K 最近鄰居法獲得最佳參數 K、特徵數及分類結果均不同。其中平均預測正確率為 73.53±3.13 (n = 10)，表示本研究利用年齡、更年期與否及十一個 SNP，即可獲得 73%左右的骨質疏鬆症之預測正確率。表 3 顯示出本研究方法結果優於其他分類方法，其中相較於 K 最近鄰居法之參數 K 設定為 1、3 及 5，本研究方法除省去手動設定參數 K，經特徵選取的挑選後，能獲得較好的結果。

(四) 討論

對於 K 最近鄰居法參數 K 之最佳選擇完全根據資料集的分佈。一般而言，較大 K 值能減少資料在分類上雜訊的干擾，但較小 K 值對於分類上的界限較為明顯。此外 Ghosh [7]指出最佳的 K 值取決於特定的資料集，需利用訓練資料集進行觀測而得知。另一方面，K 最近鄰居法之時間複雜度為 $O(Kn \log n)$，可知參數 K 將直接影響 K 最近鄰居法的執行效率，因此本文將粒子族群最佳化之粒子編碼加入參數 K。相較於其他最佳化方法，粒子族群最佳化的優點除了演化方式簡單且易實現外，其搜尋範圍廣且快速收斂。其快速收斂的特性可彌補高運算量的支持向量機，而搜尋範圍廣能使特徵選取及參數最佳化獲得更有效的搜尋。此外，粒子族群最佳化有幾個重要的參數，其中包含族群數

表 2 BPSO 於特徵選取及參數最佳化之實驗結果

實驗編號	BPSO-KNN				
	最佳 K	特徵數	敏感度	特異度	正確率
1	9	6	72.22	77.27	75.49
2	5	8	81.25	75.71	77.45
3	13	6	68.75	70.00	69.61
4	15	7	71.79	80.95	77.45
5	5	6	66.67	76.39	73.53
6	7	2	71.79	76.19	74.51
7	11	6	53.33	80.56	72.55
8	7	5	82.14	72.97	75.49
9	7	5	65.63	72.86	70.59
10	3	6	64.86	70.77	68.63
平均 ± 標準差 (n = 10)			69.84 ± 8.31	75.37 ± 3.74	73.53 ± 3.13

表 3 各分類器應用於骨質疏鬆症資料集之預測結果

分類器	敏感度	特異度	正確率
1NN	55.38	69.30	64.35
3NN	60.85	69.59	66.96
5NN	68.42	70.25	69.80
C4.5	62.02	73.28	68.63
SVM	**76.02**	68.53	69.61
RF	66.78	72.80	70.18
NB	69.24	73.17	71.76
BPSO-KNN	69.84	**75.37**	**73.53**

Legends: (1) KNN: K Nearest-Neighbor; (2) SVM: Support Vector Machine; (3) RF: Random Forest; (4) NB: Naïve Bayes; (5) BPSO-KNN: our propose approach.

P_{size}、慣性權重 w、學習因子 c_1、c_2 以及迭代次數 I_{size}，其中族群數若設定太大會造成運算時間過於冗長，反之則無法在解空間找到最佳解，文獻[3]提出粒子數設定為 50 可獲得較好的結果，因此本研究中族群數設定為 50；w、c_1、c_2 的參數則是影響粒子族群最佳化的收斂效果，若設定過大會造成粒子移動的速度過快，導致無法找到最佳解；反之若設定過小，則使粒子移動過慢，使得若要找到最佳解則需花費冗長的運算時間，依文獻建議各設定為 $w = 1.0$，$c_1 = c_2 = 2$ [22]。最後迭代次數設為 100，配合 K 最近鄰居法做為正確率的評估方法。數據顯示，100 次迭代已可達到收斂效果。對於分類問題中，特徵選取及參數最佳化可能有 Overfitting

表 4 BPSO 訓練及測試正確率之實驗結果比較

實驗編號	訓練樣本數 (case / ctrl.)	測試樣本數 (case / ctrl.)	訓練正確率	測試正確率
1	127 / 75	61 / 41	78.74	75.49
2	129 / 73	59 / 43	78.33	77.45
3	129 / 73	59 / 43	80.17	69.61
4	126 / 76	62 / 40	76.21	77.45
5	123 / 79	65 / 37	78.86	73.53
6	129 / 73	59 / 43	76.81	74.51
7	116 / 86	72 / 30	78.26	72.55
8	129 / 73	59 / 43	78.29	75.49
9	126 / 76	62 / 40	77.17	70.59
10	129 / 73	59 / 43	78.76	68.63
平均 ± 標準差 (n = 10)			78.16 ± 1.15	73.53 ± 3.13

問題的存在。Overfitting 問題出現在計算較密集或複雜的演算法，評估上可能造成偏差，然而在這種情況將會影響預測效果[18]。另外，如果訓練資料的屬性(數據)過於接近，亦使分類器預測品質下降。當沒有足夠的訓練樣本，足以訓練分類器，則分類器學習的概念則會被覆蓋。這個問題普遍存在於現實生活中數據雜訊較多的樣本[14]。為了避免 Overfitting 問題，近來有一些方法被提出，如交叉驗證法(cross validation)、資料規則化(regularization)、訓練提早中斷(early termination)及重新取樣(resampling) [20, 25]。然而，最好解決 Overfitting 問題的方法是取得大量的訓練樣本。在本文，我們利用 *m*-Fold 交叉驗證法避免 Overfitting 問題。最後由表 4 可得知，粒子族群於特徵選取及參數最佳化之訓練及測試結果，並無明顯落差，亦可說明本研究方法未落入 Overfitting 問題。

在資料集裡利用既有樣本之年齡、是否已過更年期及十一個可能與骨質疏鬆症相關 SNP，以特徵選取方法選出這些重要特徵。對於特徵選取問題而言，13 個特徵共有 $2^{13} = 8192$ 種組合。然而在不同的特徵組合及不同的 K 之設定，會獲得不同的結果。因此本文參考文獻[9]染色體的設計，將參數設為搜尋解的空間，以最佳化演算法取代參數手動設定，在本文搜尋空間共 $2^{13+3} = 65536$ 種組合，而我們利用粒子族群最佳化進行搜尋，共 $P_{size} \times I_{size} = 5000$ 組解即可得較佳的結果，而省去使用暴力演算法消耗多餘時間。在本文所獲得之預測正確率，乃利用有限樣本以及與骨質疏鬆症有關聯性的 SNP，進行機器學習訓練與測試所獲得。其結果可供生物學家參考，倘若配合臨床實驗證明、骨質疏鬆症資訊相關的搜集及更多可用的樣本，可使本研究方法更強健、穩固且更可靠。

四、結論

本研究利用 11 個可能與骨質疏鬆有關聯的 SNP 做為實驗，目的在於驗證所提出之方法能有效獲得較佳預測能力及 SNP 挑選。結果顯示，本研究方法有 73%以上預測正確率，且優於其他方法。我們希望此項成果可以供往後醫

學預測骨質疏鬆症或對於 SNP 挑選的使用。針對未來研究方向，將持續與生物學家合作，利用有效的機器學習方法進行其他疾病的預測或取得更多 SNP 資料，尋找對人類疾病具有幫助及意義的資訊。

五、參考文獻

[1] D. W. Aha, D. Kibler and M. K. Albert, “Instance-Based Learning Algorithms”, Machine Learning, Vol. 6, pp.37-66, 1991.

[2] R. Battiti, “Using mutual information for selecting features in supervised neural net learning”, IEEE Transactions on Neural Networks, Vol. 5, pp.537-550, 1994.

[3] D. Bratton and J. Kennedy, “Defining a Standard for Particle Swarm Optimization”, Proceedings of the 2007 IEEE Swarm Intelligence Symposium, Honolulu, Hawaii, USA, pp.120-127, 2007.

[4] L. Breiman, “Random forests”, Machine Learning, Vol. 45, pp.5-32, 2001.

[5] L.-Y. Chuang, H.-W. Chang, C.-J. Tu and C.-H. Yang, “Improved binary PSO for feature selection using gene expression data”, Computational Biology and Chemistry, Vol. 32, pp.29-38, 2008.

[6] E. Fix and J. Hodges, “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”, Technical Report. USAF School of Aviation Medicine, Randolph Field, TX., 1951.

[7] A. K. Ghosh, “On optimum choice of k in nearest neighbor classification”, Computational Statistics & Data Analysis, Vol. 50, pp.3113-3123, 2006.

[8] I. C. Gray, D. A. Campbell and N. K. Spurr, “Single nucleotide polymorphisms as tools in human genetics”, Human Molecular Genetics, Vol. 9, pp.2403-2408, 2000.

[9] C.-L. Huang and C.-J. Wang, “A GA-based feature selection and parameters optimization for support vector machines”, Expert Systems with Applications, Vol. 31, pp.231-240, 2006.

[10] G. H. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers”, Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, pp.338-345, 1995.

[11] J. Kennedy and R. Eberhart, “Particle swarm optimization”, Proceedings of the 1995 IEEE International Conference on Neural Networks, Vol.4, pp.1942-1948, 1995.

[12] J. Kennedy and R. C. Eberhart, “A discrete binary version of the particle swarm algorithm”, 1997 IEEE International Conference on Systems, Man, and Cybernetics, Orlando, FL, USA, pp.4104-4108, 1997.

[13] G.-T. Lin, H.-F. Tseng, C.-K. Chang, L.-Y. Chuang, C.-S. Liu, C.-H. Yang, C.-J. Tu, E.-C. Wang, H.-F. Tan, C.-C. Chang, C.-H. Wen, H.-C. Chen and H.-W. Chang, “SNP combinations in chromosome-wide genes are associated with bone mineral density in Taiwanese women”, Chinese Journal of Physiology, Vol. 51, pp.32-41, 2008.

[14] J. Loughrey and P. Cunningham, Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets, Research and Development in Intelligent

Systems XXI, 2005, pp.33-43.

[15] P. M. Narendra and K. Fukunaga, “A Branch and Bound Algorithm for Feature Subset Selection”, IEEE Transactions on Computers, Vol. C-26, pp.917-922, 1977.

[16] I.-S. Oh, J.-S. Lee and B.-R. Moon, “Hybrid genetic algorithms for feature selection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, pp.1424-1437, 2004.

[17] J. C. Platt, “Fast training of support vector machines using sequential minimal optimization”, Advances in kernel methods: support vector learning, B. Schölkopf, C.J.C. Burges, and A.J. Smola, eds., Cambridge, MA, MIT Press, pp.185-208, 1999.

[18] J. Reunanen, I. Guyon and A. Elisseeff, “Overfitting in Making Comparisons Between Variable Selection Methods”, Journal of Machine Learning Research, Vol. 3, pp.1371-1382, 2003.

[19] S. L. Salzberg, “On comparing classifiers: Pitfalls to avoid and a recommended approach”, Data Mining and Knowledge Discovery, Vol. 1, pp.317-327, 1997.

[20] C. Schaffer, “Overfitting avoidance as bias”, Machine learning, Vol. 10, pp.153-178, 1993.

[21] B. S. Shastry, “SNP alleles in human disease and evolution”, Journal of Human Genetics, Vol. 47, pp.561-566, 2002.

[22] Y. Shi and R. C. Eberhart, “A modified particle swarm optimizer”, Proceedings of the IEEE International Conference on Evolutionary Computation, Anchorage, AK, USA, pp.69-73, 1998.

[23] M. Stone, “Cross-validatory choice and assessment of statistical predictions”, Journal of the Royal Statistical Society, 36, pp.111-147, 1974.

[24] I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, San Francisco: Morgan Kaufmann, 2005.

[25] D. H. Wolpert, “On overfitting avoidance as bias”, Technical Report SFI-TR-92-03-5001, Santa Fe Institute, 1993.

[26] K. Woods and K. W. Bowyer, “Generating ROC curves for artificial neural networks”, IEEE Transactions on Medical Imaging, Vol. 16, pp.329-337, 1997.