隨機森林運用於白血病基因分類

趙李英記
文化大學國企所
skywithme118@gmail.com

摘要

本研究採用採用隨機森林分類法，是訓練一群的決策樹，對於所輸入資料進行決策投票。本研究所採用白血病128個樣本，每個樣本有12625個基因，最初，使用12625個基因產生不均勻分類，因為不相關連的基因太多，感興趣的基因太少時，分辨率會很低，所以，經過特徵縮減，並檢驗隨基森林分類效果。特徵篩選模式，採用簡單t統計，進行特徵的選擇，縮減其特徵，因而展現出較佳的分類效果。

關鍵字：隨機森林；白血病；留一交叉驗證法

A study of Leukemia gene using Random Forest classifier

Abstract

The Random Forest classifier, is training a group of decision tree, and voting the input data to make decisions. There are 128 Leukemia samples were used in this study. Initially, each sample is with 12625 genes, produced uneven classification. Because there are too many related genes and interesting genes are too small. The discrimination will be low, before the feature reduction test effect of their classification. The feature selection, using a simple t-statistics method, ranked by the relationship between the characteristics and the target attribute. By above gene selection, to reduce those characters, thus was better for classification.

Keywords: Random Forest, Leukemia, LOOCV

1. 前言

隨機森林(Random Forest)演算法是直接使用變數的重要性作為分類的模式，也是一種多分類(Multiclass)、多變數 (Multivariate) 方法[1]。該方法不需要預先指定所需的基因選擇數量，而是以分類器自己的方式去適應選擇基因的數量。隨機森林演算法主要應用多核處理器的方法計算資料集作為分群。隨機森林演算法本質上有偏向確認小基因分

群仍然可以達到良好的預測性能（因此，高度相關的基因將不被選中，因為他們認為是冗餘的基因）。在機器學習領域裡，隨機森林是一種進階的分類方法[2]，其藉由訓練一群決策樹，來對輸入資料進行決策投票。也就是說，隨機森林中的每顆決策樹會對輸入資料進行分類，給予一個類別標籤，當聚集所有決策樹的類別標籤後，即可以得到類別標籤的投票(vote)分布。此時，亦可以定義獲得最多票數的類別為此筆資料的資料類別。對於很多資料集將可以產生高度精確的分類結果，可以處理巨大的屬性個數(維數很高)，它可以估計屬性的重要性，在建構樹的過程中產生的誤差為內部無偏估計一般性錯誤，包括一個好方法估計缺失資料，當很大比例的資料缺失，仍然保持高的正確率，提供一個經驗方法去估計屬性的交互作用，資料集類別不平衡的情況下，它也可以平衡誤差，它可以計算實例之間的相關性，對於聚類，異常檢驗(新穎性檢驗)和使用排列(By scaling) 比較有用的可視化資料。因上述特點，可以用於無標籤資料，進而在執行一個無監督聚類，新穎性檢測和資料可視化的學習是速度很快。

2. 隨機森林分類法

Leo Breiman 和 Adele Cutler[2]發展的隨機森林機器學習演算法推論。是一個包含多個(核)決策樹的分類器，由個別樹輸出各類別的眾數來決定輸出類別。最初是由貝爾實驗室的 Tin Kam Ho(1995 年)所提出的隨機決策森林（random decision forests）演算法則，並使用"Random Forests"這個專用術語申請商標。因此結合 Breimans 的 "Bootstrap aggregating" 想法和 Ho 的"random subspace method" [3]建構決策樹的集合。

根據下列方式建每棵樹演算法：N 表示訓練例子的個數，M 表示變數的數目。我們會被告知一個數 m 使用在一個節點上，做決定時有多少個變數被用來做決定，其中 m 遠小於 M。

可從 N 個訓練案例中，可以重複取樣，如果取樣 N 次，便形成一組訓練集（即自己啟動拔靴法 bootstrap 取樣）。並且可以使用這棵樹來對剩餘預測其類別，並評估其誤差。

對於每一個節點而言，隨機選擇 m 個基於此點上的變數。根據這 m 個變數，計算其最佳的分割方式。其過程中，每棵樹都會完整成長而不會剪枝（Pruning）（這有可能在建完一棵正常樹狀分類器後會被採用）。

綜合上述，隨機森林中每顆決策樹的建構依據三個準則：

(一) 所有訓練資料過程中，皆以隨機抽樣 n 筆的資料，做為建決策樹的訓練集。對於不同的樹所選取可以重複的的訓練資料。

(二) 倘若每筆資料皆含有 M 個觀察值，隨機選取 m 種觀察值時(條件 m<<M)，做為決策參數，因此以此做為建樹，並在其各節點尋找最好分割的依據。同時在建每顆樹時，皆需選取相同的 m 值。

(三) 每一顆樹都讓其成長到最大，而不做任何的修剪(pruning)。

最佳化目標隨機森林訓練，有兩個主要因素，即每一顆樹的分類能力與任意兩顆樹的相關性。直觀上，當每棵樹具有較低錯誤率，成長出較為強健(strength)時，隨機森林整體分類能力就會上升。倘若兩顆樹的相關性同時增加時，其互補的能力就降低了，錯

誤率也就可能提高。前述 m 值的設定，會強烈影響樹的強健度，對於樹與樹之間的相關性。為了減少 m 值，個別決策樹可選用的特徵值會減少，樹的強健度也跟著下降，同時樹與樹之間容許的差別亦會減少，彼此間的相關性亦同時下降。

利用機器學習資料探勘，快速找出最適合的 m 值，訓練隨機森林時，藉由計算 OOB (out-of-bag)袋外錯誤率，調整 m 值的大小直到收斂。假設某一筆資料位於所有沒選中訓練資料的樹，所得到的類別標籤集合為 J，則在 J 中分類錯誤的比例即是 OOB 袋外錯誤率。在隨機森林的訓練，首先採用自己啟動拔靴法(Bootstrap)的方法，取出不同樣本資料來建每棵樹。每次建樹時，保留三分之一的資料未被使用，則這些資料即可被視為測試資料。將這些測試資料置入剛建好的樹中，可得到一系列 OOB 的類別標籤，並將這些標籤置入 J 中。訓練完全部的樹後，每筆訓練資料大約會被森林中三分之一的樹測試過。由 J 算出 OOB 的錯誤率，即可以衡量此隨機森林的訓練結果。

隨機森林$\{h(x,\theta k), k=1,....\}$是樹型分類器的集合，利用原 CART 分類器演算法建構出沒有剪枝的分類迴歸；其中 x 為實驗輸入變數，$\{\theta k\}$是獨立同分佈的隨機向量，隨機向量 θk 決定單一決策樹的生長過程。隨機森林與 CART 皆採用基尼索引值(Gini index)，如公式(1)進行屬性分裂。D 是包含 n 個樣本；屬性的機率為 $p(j/t)$。

$$Gini(D)=1-\sum_{j=1}^{n} p(j/t)_j^2 \qquad (1)$$

當 Gini(D)為 0 時，此節點處所有記錄都屬於同一類別，可能得到最大且有用的資訊；當 Gini(D)最大時，此節點所有的屬性呈現均勻分佈時，僅能表示得到最小且有用資訊。如果集合成分 l 的部分，則進行分裂的基尼索引值(Gini split)，如公式(2)。

$$Gini_{split}(D)=\sum_{i=1}^{l}\frac{n_i}{n}Gini(i) \qquad (2)$$

l 是子節點個數， ni 為子節點屬i 範圍之樣本數，n 為節點母體樣本數。利用公式(2)得出最小$Gini\ (D)\ split$ ，選擇作為此節點處分裂的標準進行屬性分裂。隨機森林分類方法是以Bagging法進行，將輸入變數x以簡單多數投票(vote)法得到分類準則再用隨機森林輸出。Bagging法是利用機率分配方式在資料集合中重覆產生。Bagging分類法處理方式是， x 為輸入變數； y 為目標變數，不斷迭代得出變數的分類準則。如表(1)。

表 1　Bagging 分類方法

x	0.1	0.2	0.3	0.4	0.8	0.9
y	1	1	1	-1	-1	-1

由表 3.2，Bagging 分類法進行 vote 後顯示，當 $x \le 0.3$ 時 y=1； $x > 0.3$ 時 y= -1。以此規則，經過不斷迭代找出最終的決策準則。經由此種規則，經過不斷迭代結果找出最

終的決策準則。Bagging 分類錯誤稱為，袋外錯誤率 OOB error (out of bagging error,)。隨機森林分析袋外錯誤率過程，是從所有訓練資料中，隨機抽樣 n 筆資料，作為建構決策樹訓練集。隨機森林建構多棵決策樹中，不同的樹選取的訓練集可以重覆的。經過整合後，得出 OOB error 公式(3)，袋外錯誤率。P 為樹間平均相關程度，S 為樹的分類能力(strength)。

$$OOB \quad error \le \frac{P(1-S^2)}{S^2} \qquad (3)$$

從生成決策樹中整合得出袋外錯誤率。整合後，隨機森林產生的袋外錯誤率，將其給予上限，使決策樹分類能力擁有平均分類效果。因此利用分類法邊界(margin)公式所衡量機率值，如公式(4)。

$$M(X,Y) = P(Y'_\theta = Y) - \max_{z \neq Y}(Y'_\theta = Z) \qquad (4)$$

Y′為隨機選取變數 θ 所建立的分類法之預測類別 X。邊界(margin)越高表示 X 資料分類準確率越高。

機森林分類具有特殊優點，即使是非正規化特徵(non-normalized character) 目前使用於隨機森林分類的機器學習的分類法，能達到最佳分類效果的分類器。但是，特別要須注意，隨機森林所輸出的投票百分比(機率)，並不等於基因表現性狀可能性的機率分布。每一棵決策樹都有其自我決策機制，此百分值暗示了基因是表現性狀決策的信心度。這裡真正須重視的是各種性狀得票百分比邊際限度(margin)，一個大的邊際限度即表示一個相對可信的決策結果。

自 Breimen(2001)與 Díaz-Uriarte and Andrés (2006) 研究顯示，隨機森林進行分類與特徵選取時，有四個重要參數：mtryFactor、c.sd、ntree 與 vars. drop. frac。四個參數是影響隨機森林建構後之結果。

1. mtry Factor：是隨機森林在變數投入形成決策樹，在內部節點以隨機方式進行分裂的數量，利用隨機分裂方式降低變數相依性。在 Breimen(2001)隨機森林中原始設定為 n ，n 為 資料變數的數量。
2. c.sd：是設定建構一完整決策樹，是以規則導向或分類導向。當 c.sd 設定為 0 為分類導向；1 為規則導向。
3. ntree：為決策樹生長的多寡。ntree 的大小與袋外錯誤率有負相關的關係；當 ntree 越大時，袋外錯誤率逐漸降低，最後錯誤率呈現平穩狀態。
4. vars. drop. frac：是隨機森林在特徵選取中，化減變數的比例，而且在每次迭代中利用此參數進行化減變數，找出顯著變數。Breimen(2001)原始設定為 0.2。

3. 研究方法

3.1 研究對象

本研究解決基因採集樣本的困難，雖然在少量 128 個樣本中，提出的本演算法，驗證確實足以勝任基因分類的工作。

3.1. 研究對象

針對白血病病患資料來自來 Chiaretti, Li, Gentleman, Vitale, Vignetti, Mandelli, Ritz, and Foa (2004)從急性淋巴細胞白血病人取得急性淋巴細胞白血病組織，共有白血病 128 個樣本，每個樣本有 12625 個基因[4]。

從 128 個樣本中，選取基因個數太少時，會遺漏重要的致病性基因，基因選取個數太多時，又會造成演算過程過於冗長又費時，同時維度太大產生過度擬合(overfitting)的效果。Baolin (2005)解釋，由於微陣列數據「大 P 小的 N」，小量的樣本量和大量的變數(基因)，提議利用收縮的理想方法等的一些特設 (ad-hoc)方法縮小維度，可以幫助防止過度擬合(overfiting)和產生更可靠的估計，並在實證研究證明是有用的[5]。

3.2. 研究限制

資料探勘研究結果僅對本研究樣本技術所使用的演算法與資料有效。

3.3. 效能評估

由於資料樣本筆數過少，本研究使用最繁瑣並且精確留一交叉驗證法的準確率，做為本研究的效能評估。

3.4. 分析工具

DNA 微陣列基因表現圖譜是目前基因分類最普遍使用的分析工具，DNA 微陣列技術探勘關鍵利用兩個 RNA 樣本加上兩個不同的螢光染劑(如 Cy3 是綠色，Cy5 是紅色)時，兩個 RNA 樣本能在相同的 DNA 晶片上雜合，紅點顯示條件 1 的基因表現，綠點是條件 2 的基因表現，黃點是基因在這兩種條件下的表現。

3.5. 基因表現圖譜與特徵選取

基因表現圖譜，使用 Affymetrix 公司型號 HGU95AV2 的基因晶片微陣列的基因表現圖譜如表 2。

表 2　基因為陣列資料

基因　樣本	01005	・・	類別
1000_at	7.59732	・・	BCR/ABL
1001_at	5.04619	・・	NEG
1002_f_at	3.90046	・・	BCR/ABL

・・・	・・・	・・	・・・

3.6. 基因選擇資料

特徵選取(feature selection)目的是選擇重要基因，在資料探勘過程中是一個很重要的技巧，分類器藉由前置處理選取少數部分適當的特徵(feature)後，可以有效的提升分類正確率，也可以加速整個學習效率，而當基因微陣列上數以萬計的基因被當做特徵看待時其實就是選取重要基因的問題。學者希望藉由基因選取篩選出具代表性的基因集，將雜訊(noise)或者與分類結果不相關(irrelevant)的基因給剔除掉，往後也可以進一步的利用這些基因集來預測未知的樣本類別。基因分類學習的過程中只要少數幾個基因就足夠進行後續的分類工作。因此如何在為數眾多的基因中挑選出較具代表性的基因集是非常重要且具有挑戰性的任務。

4. 隨機森林分類優缺點

4.1. 隨機森林的優點如下：

(一) 運算本質類同的隨機森林，是由一群決策樹組成，故能有效處理大量的輸入資料。
(二) 隨機森林在訓練過程中，會主動尋找適當的 m 值，即使每一筆含有數以千計特徵變數的資料，也可被接受。
(三) 經由統計計算每一個變數在隨機森林中所有樹上的使用率，並可以間接估出變數在分類的重要性。
(四) 藉由計算 OOB 的錯誤率，隨機森林也可達到最佳化，同時，訓練資料產生的偏差也可以被限制。因此，對於未見過的資料常常能保持強健的鑑別力。
(五) 當資料漏失(Missing Data)的情況發生，可以適當拋棄部分的決策樹，使隨機森林能繼續正常分類。
(六) 藉由權重，各類別 OOB 的錯誤率，對於目標類別不平衡的訓練資料，可以算出對於類別平衡的誤差。
(七) 資料的分布和特徵的相關性，隨機森林如同決策樹，不需有任何前提的假設。

4.2. 隨機森林的缺點如下：

(一) 資料集必須防範可能過擬合(overfitting)使用時，尤其對於一些有噪訊(noise)的資料集更是如此。
(二) 不能處理大量的不相干特徵和集成熵所減小的決策樹。
(三) 更容易選擇一個隨機決策的邊界，卻不是減小決策邊界中的熵。導致更大有可行性的整合，最初看來這可能像是一個優點，但是其在計算時會導致從訓練的時候向評估的時候偏移，很多應用來說這是一個缺點。

5. 輸出結果

由 OOB 袋外資料（out-of-bag）估計模型的錯誤率（OOB-error），作為參數選擇依據，並且利用袋外錯誤率增減變化，對輸入自變數的重要性進行排序。RF 的輸出組合方法，兩種方法分別針對資料有簡單多數投票法分類和單棵樹輸出結果的平均做迴歸。

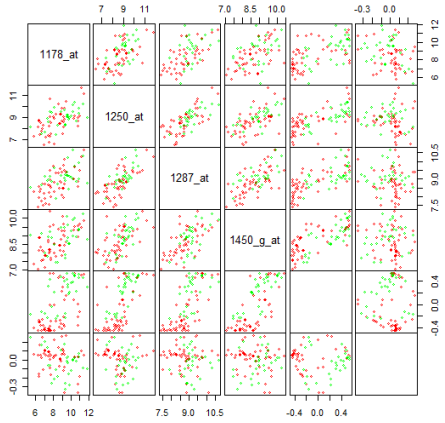

圖 1　隨機森林資料: Predictors and MDS of Proximity Based on RandomForest

Type of random forest: classification ntree: 500 No. of variables tried at each split: 3

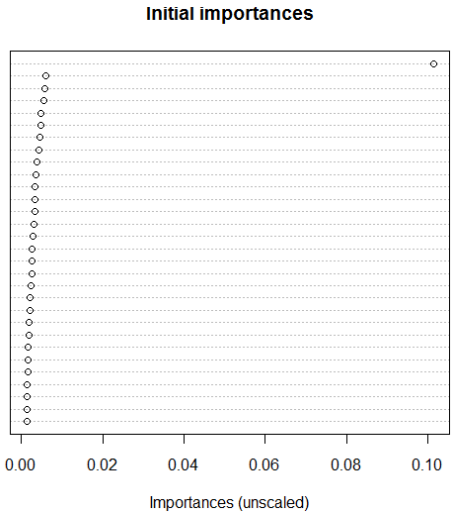

圖 2　使用隨機森林 OOB 錯誤率選擇變數

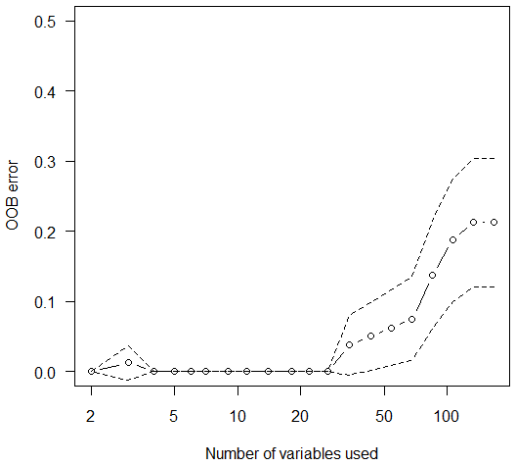

圖 3　從隨機森林啟動變數選擇步驟

圖 2. 隨機森林向後減少 ntree =5000 ; mtryFactor= 1　選擇使用 12 個變數: "31863_at" "33433_at"　"34372_at"　"35039_at"　"37707_i_at" "37736_at"　"39762_at" "40431_at"　"40859_at"　"41411_at"　"911_s_at"　"943_at"

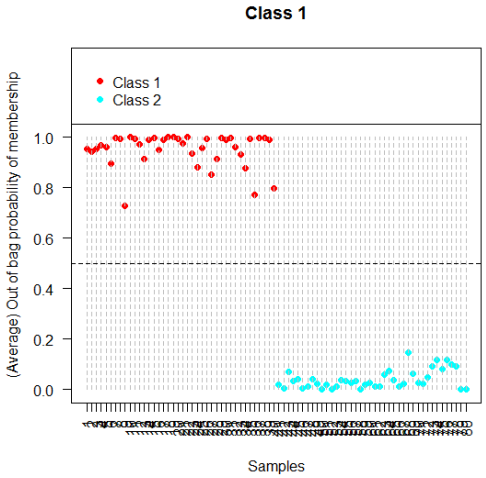

圖 4　第一分類 OOB 隸屬函數百分比 (平均)

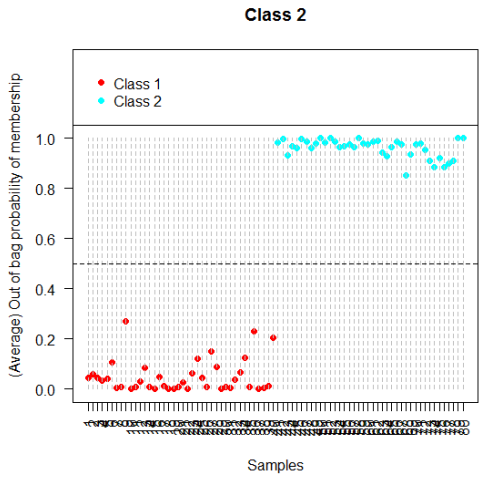

圖 5　第二分類 OOB 隸屬函數百分比(平均)

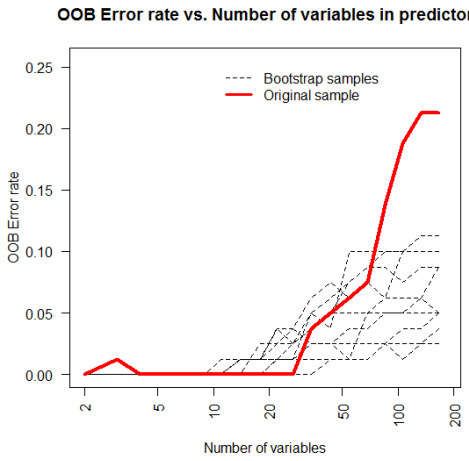

圖 6　預測 OOB 錯誤率 VS. 變數個數

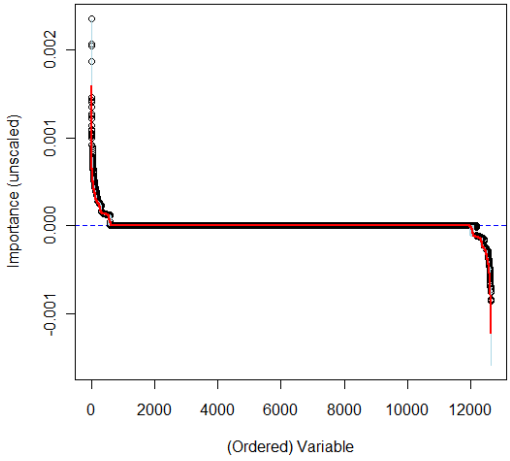

圖 7　隨機森林變數重要性排列分類標籤

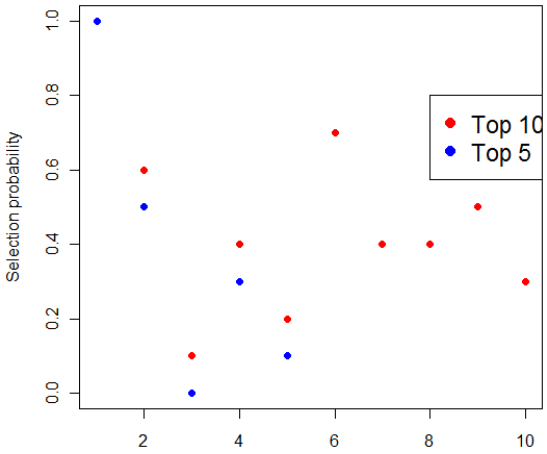

圖 8　隨機森林變數重要性的選擇百分比圖

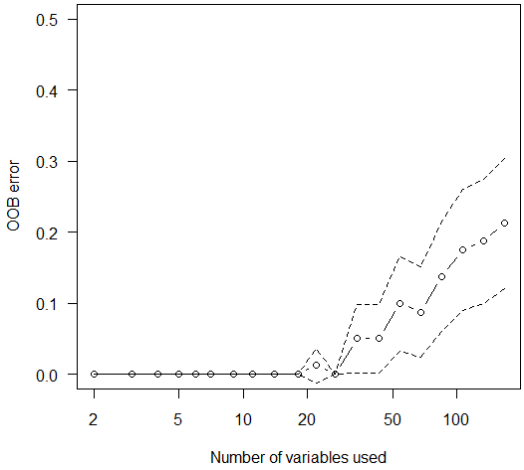

圖 9　從隨機森林啟動變數選擇步驟

圖 8 隨機森林向後減少 ntree =500 ; mtryFactor= 1 選擇使用 6 個變數:"32262_at" "34372_at" "37707_i_at" "41775_at" "911_s_at" "960_g_at"

6. 圖片、表格說明及方程式

留一交叉驗證法(leave-one-out cross-validation)方法在有 n 個樣本的情況下，把每一筆資料單獨抽出來當作測試集，其他的 n-1 筆資料則為訓練集合，且留一交叉驗證法在樣本過少的狀況下容易造成過度配適的問題[5]。分類準確率(Classification Correct Rate)：CR = (TPN + TNN)/N。其中 Np、Nn、和 N 分別為：

根據實驗的預測結果我們先導初四個統計數據分別為：
(一)腫瘤組織– 正確辨識個數（True Positive Number, TPN），真陽性。
(二)正常組織– 正確辨識個數(True Negative Number, TNN)，真陰性。
(三)腫瘤組織– 錯誤辨識個數(False Negative Number, FNN)，偽陰性。
(四)正常組織– 錯誤辨識個數(False Positive Number, FPN），偽陽性。
在利用上述四個數目可定義下列準則：
(一)偵測率(Detection Rate)：DR = TPN /Np
(二)錯誤警示率(False Alarm Rate)：FAR=FPN/Nn
(三)分類準確率(Classification Correct Rate)：CR = (TPN + TNN)/N。
其中 Np、Nn、和 N 分別為：
(一)Np 代表正樣本(腫瘤組織)的個數；
(二)Nn 代表負樣本(正常組織)的個數。
(三)N 代表全部樣本的個數。N = Np + Nn。
本研究使用留一法驗證，分成二群，驗證其正確率為 84.81%。

7. 結論

隨機森林分類方法多分類、多變數分類法，使用 OOB error 當作減少標準，實現變數減少，並且達到最少重要變數(從隨機森林回傳其重要性)，而且計算資料集分群主要

利用多核處理器，不需要預先指定的基因選擇數量，由自己去適應選擇的基因數量。隨機森林演算法是有偏向驗證小的基因分群，仍然可以達到良好的預測性能（因此，高度相關的基因將不被選中，因為他們認為是冗餘的基因）。

本研究雖未得到很優秀的正確率，但是對於大量特徵的基因表現圖譜，仍不失是一個很好的分類工具。

參考文獻

[1] R Díaz-Uriarte, SA De Andres, “Gene selection and classification of microarray data using random forest” - BMC bioinformatics, 2006.

[2] L. Breiman and A. Cutler, “Random Forests,” *Machine Learning* Vol. 45, October, 2001, pp5-32.

[3] T. K. Ho, “The random subspace method for constructing decision forests”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, 1998, pp. 832-844.

[4] S. Chiaretti, , X. Li, R Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz and R. Foa, “Gene expression profile of adult T-cell acute lymphocylic leukemia identifies dixtinct subsets of patients with different response to therapy and survival,” *Blood*, Vol. 103, No. 7, 2004, pp. 2771-2778.

[5] J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” *IEEE J. Quantum Electron.*, submitted for publication.

[6] B. Wu, ”Differential gene expression detection and sample classification using penalized linear regression models”，*Bioinformatics* 2005 Vol. 22, No 4, pp 472-476。