I806425

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I806425 B**

(45)公告日： 中華民國 112 (2023) 年 06 月 21 日

(21)申請案號：111105254 (22)申請日： 中華民國 111 (2022) 年 02 月 14 日

(51)Int. Cl.： ***G06N20/00 (2019.01)*** ***G06N3/08 (2006.01)***

G16H50/20 (2018.01)

(71)申請人：宏碁股份有限公司 (中華民國) ACER INCORPORATED (TW)

新北市汐止區新台五路一段 88 號 8 樓

宏碁智醫股份有限公司 (中華民國) ACER MEDICAL INC. (TW)

新北市汐止區新台五路一段 86 號 7 樓

長庚醫療財團法人基隆長庚紀念醫院 (中華民國) CHANG GUNG MEMORIAL HOSPITAL, KEELUNG (TW)

基隆市麥金路 222 號

財團法人國家衛生研究院 (中華民國) NATIONAL HEALTH RESEARCH INSTITUTES (TW)

苗栗縣竹南鎮 35053 科研路 35 號

(72)發明人：林意淳 LIN, YI-CHUN (TW) ；許銀雄 HSU, YIN-HSONG (TW) ；蔡宗憲 TSAI, TSUNG-HSIEN (TW) ；詹韻玄 CHAN, YUN-HSUAN (TW) ；蔡亭芬 TSAI, TING-FEN (TW) ；徐唯哲 HSU, WEI-CHE (TW) ；葉集孝 YEH, CHI-HSIAO (TW)

(74)代理人：葉璟宗；卓俊傑

(56)參考文獻：

TW	202020887A	CN	101061510B
CN	113435602A	US	2016/0174902A

審查人員：李惟任

申請專利範圍項數：11 項 圖式數：7 共 29 頁

(54)名稱

特徵挑選方法

(57)摘要

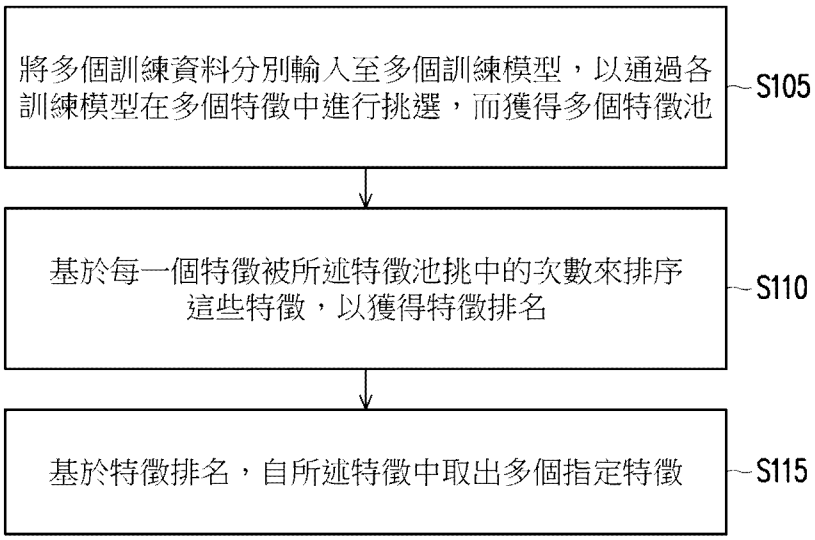
一種特徵挑選方法，包括：將多個訓練資料分別輸入至多個訓練模型，以通過各訓練模型在多個特徵中進行挑選，而獲得多個特徵池；基於每一個特徵被所述特徵池挑中的次數來排序這些特徵，以獲得特徵排名；以及基於特徵排名，自所述特徵中取出多個指定特徵。

A feature selection method is provided, including: inputting a plurality of training data into a plurality of training models to perform selection in a plurality of features through each training model for obtaining multiple feature pools; sorting the features based on a number of times each feature is selected by the feature pools to obtain a feature ranking; and extracting a plurality of designated features from the features based on the feature ranking.

指定代表圖：

符號簡單說明：

S105~S115:特徵挑選方法的步驟



【圖1】

專利案號：111105254

公告本

I806425

【發明摘要】

【中文發明名稱】特徵挑選方法

【英文發明名稱】FEATURE SELECTION METHOD

【中文】一種特徵挑選方法，包括：將多個訓練資料分別輸入至多個訓練模型，以通過各訓練模型在多個特徵中進行挑選，而獲得多個特徵池；基於每一個特徵被所述特徵池挑中的次數來排序這些特徵，以獲得特徵排名；以及基於特徵排名，自所述特徵中取出多個指定特徵。

【英文】A feature selection method is provided, including: inputting a plurality of training data into a plurality of training models to perform selection in a plurality of features through each training model for obtaining multiple feature pools; sorting the features based on a number of times each feature is selected by the feature pools to obtain a feature ranking; and extracting a plurality of designated features from the features based on the feature ranking.

【指定代表圖】圖1。

【代表圖之符號簡單說明】

S105～S115:特徵挑選方法的步驟

【特徵化學式】無

【發明說明書】

【中文發明名稱】特徵挑選方法

【英文發明名稱】FEATURE SELECTION METHOD

【技術領域】

【0001】 本發明是有關於一種模型建構方法，且特別是有關於一種特徵挑選方法。

【先前技術】

【0002】 在醫院看診的過程中，通常醫師會利用抽血取得生理資訊，做為輔助判別疾病的指標。抽血可取得的生理資訊可能有代謝體、基因等體學特徵。過去技術大多數僅考慮單一種體學資料，並利用機器學習等方法來進行特徵挑選。倘若同時考慮多種體學資料，也是將全部體學資料一起加入，再利用機器學習等方法進行特徵挑選。然而，由於體學特徵的數量少則百個多則上萬個，若全部一起挑選去做，就算是透過機器學習去挑選，也相當耗費時間及資源。

【發明內容】

【0003】 本發明提供一種特徵挑選方法，可有效地挑選出最具影響性的特徵。

【0004】 本發明的特徵挑選方法是利用電子裝置在多個特徵中進

行挑選，所述特徵挑選方法包括：將多個訓練資料分別輸入至多個訓練模型，以通過各訓練模型在多個特徵中進行挑選，而獲得多個特徵池；基於每一個特徵被所述特徵池挑中的次數來排序這些特徵，以獲得特徵排名；以及基於特徵排名，自所述特徵中取出多個指定特徵。

【0005】 在本發明的一實施例中，上述通過各訓練模型在所述特徵中進行挑選，而獲得所述特徵池的步驟包括在下述三種挑選方式中的至少其中一種：(1)透過各訓練模型逐一針對單一特徵計算至少一個統計指標，並將統計指標與對應的臨界值進行比對，藉此決定是否選定此特徵至對應的特徵池；(2)透過各訓練模型在所述特徵中執行特徵擷取動作，藉此獲得分別與這些訓練模型對應的多個特徵池；(3)基於多個特徵類型，將這些特徵分類為多個特徵群組，以使得各訓練模型在各特徵群組所包括的特徵中執行特徵擷取動作，藉此獲得各訓練模型分別對應至所述特徵群組的多個特徵池。

【0006】 在本發明的一實施例中，在所述挑選方式(3)中，包括：將各特徵群組對應的特徵池設定為一個特徵集合；基於各特徵被所述特徵池挑中的次數來排序各特徵集合中的各特徵，以獲得各特徵集合的特徵排名；以及基於各特徵群組對應的權重以及特徵排名，自各特徵集合中取出對應數量的指定特徵。在此，各特徵群組對應的權重是基於各特徵群組所包括的特徵數量佔全部特徵數量的比值。

【0007】 在本發明的一實施例中，在所述三種挑選方式中選擇多個的情況下，更包括：針對所述多個挑選方式的每一個，獲得符合對應的指定數量的指定特徵，進而分別獲得對應於多個挑選方式的多個選定特徵組。

【0008】 在本發明的一實施例中，在分別獲得對應於所述多個挑選方式的多個選定特徵組之後，對這些選定特徵組執行聯集、交集以及差集其中一者來獲得整合特徵池。

【0009】 在本發明的一實施例中，在獲得整合特徵池之後，透過多體學特徵調控途徑分析，查詢多個已知資料庫，以在整合特徵池中挑選出一或多個代表特徵。

【0010】 在本發明的一實施例中，在獲得所述多個代表特徵之後，利用多個測試資料，在分別選用不同的多個特徵數量的代表特徵的情況下來獲得各訓練模型的多個準確率；以及基於這些準確率在所述代表特徵中選出一或多個最終特徵。

【0011】 在本發明的一實施例中，其中基於特徵排名，自所述特徵中取出指定特徵的步驟包括：基於特徵排名，自所述特徵中取出符合指定數量的指定特徵。其中，在獲得特徵池之後，利用多個測試資料，在選用不同的多個特徵數量的特徵的情況下來獲得各訓練模型的多個準確率；基於準確率，在所述訓練模型中選擇其中一個；基於被選擇的其中一個訓練模型的特徵數量與準確率來獲得陡坡圖；以及基於陡坡圖，在這些特徵數量中獲得指定數量。

【0012】 在本發明的一實施例中，在獲得指定特徵之後，透過一多體學特徵調控途徑分析，查詢多個已知資料庫，以在所述指定特徵中挑選出一或多個代表特徵。

【0013】 在本發明的一實施例中，在獲得所述多個代表特徵之後，利用多個測試資料，在分別選用不同的多個特徵數量的代表特徵的情況下來獲得各訓練模型的多個準確率；以及基於準確率在所述代表特徵中選出一或多個最終特徵。

【0014】 本發明的特徵挑選方法是利用電子裝置在多個特徵中進行挑選，本特徵挑選方法包括：透過下述三種挑選方式中的其中一種，將多個訓練資料分別輸入至多個訓練模型，以通過各訓練模型在多個特徵中進行挑選來獲得多個指定特徵，所述挑選方式包括：(1)透過各訓練模型逐一針對單一特徵計算至少一個統計指標，並將統計指標與對應的臨界值進行比對，以自所述特徵中獲得指定特徵；(2)透過各訓練模型在全部特徵中執行特徵擷取動作，藉此獲得分別與這些訓練模型對應的多個特徵池，基於各特徵被特徵池挑中的次數來排序這些特徵，以獲得特徵排名，基於特徵排名，自所述特徵中取出指定特徵；(3)基於多個特徵類型，將全部特徵分類為多個特徵群組，以使得各訓練模型在各特徵群組所包括的特徵中執行特徵擷取動作，藉此獲得各訓練模型分別對應至這些特徵群組的多個特徵池，基於各特徵被特徵池挑中的次數來排序這些特徵，以獲得特徵排名，基於特徵排名，自這些特徵中取出指定特徵。

【0015】 基於上述，本揭露利用多個訓練模型來篩選特徵，再根據各特徵被訓練模型挑選到的次數來進行下一步的篩選，據此，透過多層級的特徵篩選，不僅節省特徵挑選時間也可挑選出最具影響性的特徵，同時維持高準確率。

【圖式簡單說明】

【0016】

圖 1 是依照本發明一實施例的特徵挑選方法的流程圖。

圖 2 是依照本發明一實施例的針對單一特徵的挑選方式的示意圖。

圖 3 是依照本發明一實施例的針對全部特徵的挑選方式的示意圖。

圖 4 是依照本發明一實施例的針對特徵群組的挑選方式的示意圖。

圖 5 是依照本發明一實施例的陡坡圖的示意圖。

圖 6 是依照本發明一實施例的特徵挑選方法的流程圖。

圖 7 是依照本發明一實施例的多體學特徵調控途徑分析的示意圖。

【實施方式】

【0017】 一般而言，血液中可取得的生理資訊可能有代謝體、基因等體學特徵。若能同時考慮多種類型的體學資訊，從不同資訊

面向來協助分析，對於臨床應用將會有很大的幫助，且可提高準確率，並且還可進進一步協助疾病（例如糖尿病、腎臟病等）的預測。另外，若能使用最少的體學特徵來去解釋及判斷生理狀態，將可提高判斷效率。因此，底下提出一種特徵挑選方法，可達到高效率、高準確率、高應用性。底下實施例是透過具有運算功能的電子裝置來實現。例如，可採用伺服器、個人電腦、筆記型電腦、平板電腦等電子裝置來實現，甚至可採用智慧型手機來實現。

【0018】 電子裝置中具有處理器、儲存元件以及通訊元件。處理器例如為中央處理單元（Central Processing Unit，CPU）、物理處理單元（Physics Processing Unit，PPU）、可程式化之微處理器（Microprocessor）、嵌入式控制晶片、數位訊號處理器（Digital Signal Processor，DSP）、特殊應用積體電路（Application Specific Integrated Circuits，ASIC）或其他類似裝置。

【0019】 儲存元件例如是任意型式的固定式或可移動式隨機存取記憶體（Random Access Memory，RAM）、唯讀記憶體（Read-Only Memory，ROM）、快閃記憶體（Flash memory）、硬碟或其他類似裝置或這些裝置的組合。儲存元件中儲存有一或多個程式碼片段所組成，上述程式碼片段在被安裝後，會由處理器來執行以實現下述特徵挑選方法。

【0020】 通訊元件可以是採用區域網路（Local Area Network，LAN）技術、無線區域網路（Wireless LAN，WLAN）技術或行動通訊技術的晶片或電路。區域網路例為乙太網路（Ethernet）。無

線區域網路例如為 Wi-Fi。行動通訊技術例如為全球行動通訊系統（Global System for Mobile Communications，GSM）、第三代行動通訊技術（third-Generation，3G）、第四代行動通訊技術（fourth-Generation，4G）、第五代行動通訊技術（fifth-Generation，5G）等。

【0021】 圖 1 是依照本發明一實施例的特徵挑選方法的流程圖。請參照圖 1，在步驟 S105 中，將多個訓練資料分別輸入至多個訓練模型，以通過各訓練模型在多個特徵中進行挑選，而獲得多個特徵池。訓練模型可以採用不同的多個統計模型或不同的多個機器學習模型來實現。統計模型例如可採用最小絕對值收斂和選擇算子（least absolute shrinkage and selection operator，Lasso）演算法、逐步邏輯回歸（stepwise logistic regression）法、統計檢驗（statistical test）法等。機器學習模型例如採用隨機森林（random forest）演算法、支援向量機（support vector machine，SVM）演算法等。

【0022】 在一實施例中，可基於這些訓練資料來劃分出多個訓練資料集，並逐一將這些訓練資料集輸入至各訓練模型進行訓練，以由訓練模型來挑選具有最強關聯性的特徵。在此，可根據選擇的挑選方式的不同，由一個訓練模型獲得一個特徵池，也可由一個訓練模型來獲得多個特徵池。

【0023】 在本實施例中，可選擇下述三種挑選方式(1)～(3)中的至少其中一種。挑選方式(1)：透過每一個訓練模型逐一針對單一特

徵計算至少一個統計指標，並將統計指標與對應的臨界值進行比對，藉此決定是否選定此特徵至對應的特徵池。所述臨界值為預先設定的固定值，可由訓練模型自行決定。統計指標例如為 P 值、勝算比（odds ration）、相關係數（correlation coefficient）、差異倍數（fold change）等。

【0024】 舉例來說，圖 2 是依照本發明一實施例的針對單一特徵的挑選方式的示意圖。在圖 2 中僅繪示兩個訓練模型 Ms(1)、Ms(2)，然，並不以此為限。在此，基於多個訓練資料來獲得 S 組訓練資料集 TD1～TDS，將這些訓練資料集 TD1～TDS 逐一輸入至訓練模型 Ms(1)～Ms(2)來針對單一個特徵計算一個統計指標。在此以訓練資料集 TD1 使用訓練模型 Ms(1)～Ms(2)來進行說明，其他訓練資料集 TD2~TDS 亦以此類推。以訓練模型 Ms(1)採用統計檢驗（statistical test）而言，利用訓練模型 Ms(1)針對特徵 f1～fn 中的每一個，算出對應的 P 值 p(f1)～p(fn)。之後，將 P 值 p(f1)～p(fn)與對應的臨界值 T1 進行比對，並設定為 P 值≤T1。假設臨界值 T1=0.05，即，挑選 P 值小於或等於 0.05 的特徵至對應的特徵池 Ps(1)。

【0025】 另外，訓練模型 Ms(2)是用來計算勝算比。利用訓練模型 Ms(2)針對特徵 f1～fn 中的每一個，算出對應的勝算比 r(f1)～r(fn)。之後，將勝算比 r(f1)～r(fn)與對應的臨界值 T2 進行比對，並設定為勝算比>T2。假設臨界值 T2=2，即，挑選挑選勝算比大於 2 的特徵至對應的特徵池 Ps(2)。在其他實施例中，還可進一步

加入第三個或更多的訓練模型來計算各特徵的統計指標，並將其與對應的臨界值進行比對來獲得第三個或更多的特徵池。

【0026】 挑選方式(2)：透過每一個訓練模型在全部特徵中執行特徵擷取動作，藉此獲得分別與這些訓練模型對應的多個特徵池。即，由一個訓練模型針對多個訓練資料集進行訓練來獲得最強關聯性的一組特徵，而獲得此訓練模型對應的特徵池。

【0027】 舉例來說，圖 3 是依照本發明一實施例的針對全部特徵的挑選方式的示意圖。在圖 3 中，使用 X 個訓練模型 M(1)～M(X)。將多個訓練資料集 TD1～TDS 逐一輸入至訓練模型 M(1)～M(X)中的每一個進行訓練，以挑選出具有最強關聯性的特徵。在此以訓練資料集 TD1 使用 X 個訓練模型 M(1)~M(X)來進行說明，其他訓練資料集 TD2~TDS 亦以此類推。利用訓練模型 M(1)在全部特徵 f1～fn 中進行挑選而獲得特徵池 Pm(1)，利用訓練模型 M(2)在全部特徵 f1～fn 中進行挑選而獲得特徵池 Pm(2)，以此類推獲得 X 個特徵池 Pm(1)～Pm(X)。

【0028】 在一實施例中，可根據訓練模型 M(1)～M(X)的準確率來決定特徵池 Pm(1)～Pm(X)所欲挑選的指定數量。例如，基於陡坡圖/手肘法（Elbow method）來決定指定數量。以特徵池 Pm(1)而言，其包括由訓練資料集 TD1~TDS 使用訓練模型 M(1)所得到的 S 個特徵池，再透過計算特徵 f1~fn 被 S 個特徵池挑中的次數進行排名，取出指定數量的特徵。特徵池 Pm(2)～Pm(X)亦以此類推。

【0029】 挑選方式(3)：先基於多個特徵類型，將全部特徵分類為

多個特徵群組，之後，透過每一個訓練模型在各個特徵群組所包括的特徵中執行特徵擷取動作，藉此一個訓練模型可獲得對應至所述多個特徵群組的多個特徵池。例如，以體學特徵而言，體學特徵可分類為代謝體學、基因體學等特徵類型，故多個特徵可分類為代謝體學群組、基因體學群組等。

【0030】 舉例來說，圖 4 是依照本發明一實施例的針對特徵群組的挑選方式的示意圖。在本實施例中，針對一個訓練資料集 TD1 使用 X 個訓練模型 M(1)～M(X)來進行說明，其他訓練資料集 TD2～TDS 亦以此類推。並且，假設基於 N 種特徵類型將全部特徵進行分類而獲得 N 個特徵群組 G(1)～G(N)。。

【0031】 請參照圖 4，針對不同的特徵群組將訓練資料集 TD1 輸入至訓練模型 M(1)～M(X)中的每一個進行訓練。以特徵群組 G(1)而言，訓練資料集 TD1 輸入至訓練模型 M(1)進行訓練，利用訓練模型 M(1)在特徵群組 G(1)中進行挑選而獲得特徵池 P1(G1)，訓練資料集 TD1 輸入至訓練模型 M(2)進行訓練，利用訓練模型 M(2)在特徵群組 G(1)中進行挑選而獲得特徵池 P2(G1)，…，以此類推，獲得特徵池 P1(G1)、P2(G1)、.....、PX(G1)。

【0032】 接著，再分別針對特徵群組 G(2)～特徵群組 G (N)，將訓練資料集 TD1 輸入至訓練模型 M(1)～M(X)進行訓練，來獲得對應的特徵池，結果如表 1 所示。特徵群組 G(1)對應至特徵池 P1(G1)、P2(G1)、.....、PX(G1)；特徵群組 G(2)對應至特徵池 P1(G2)、P2(G2)、.....、PX(G2)等等。而每一個特徵群組對應的多個特徵池

可合併為一個大的特徵池（TD1(G1)～TD1(GN)）。例如，對應於特徵群組 G(1)的 X 個特徵池 P1(G1)～PX(G1)可合併為大的特徵池 TD1(G1)。

【0033】 表 1

特徵群組	特徵池	特徵池
G(1)	P1(G1)、P2(G1)、.....、PX(G1)	TD1(G1)
G(2)	P1(G2)、P2(G2)、.....、PX(G2)	TD1(G2)
.....	.....	.....
G(N)	P1(GN)、P2(GN)、.....、PX(GN)	TD1(GN)

【0034】 而訓練資料集 TD2～TDS 亦如同圖 4 所示的訓練資料集 TD1 來分別針對不同的特徵群組進行訓練，以獲得對應於特徵群組 G(1)～G(N)的大的特徵池 TD1(G1)~TDS(GN)，如表 2 所示。例如，以訓練資料集 TD2 而言，特徵群組 G(1)～G(N)分別對應至大的特徵池 TD2(G1)～TD2(GN)。

【0035】 表 2

訓練資料集	特徵群組 G(1)	特徵群組 G(2)	.....	特徵群組 G(N)
TD1	TD1(G1)	TD1(G2)	.....	TD1(GN)
TD2	TD2(G1)	TD2(G2)	.....	TD2(GN)
…				
TDS	TDS(G1)	TDS(G2)	.....	TDS(GN)

【0036】 返回圖 1，在獲得特徵池之後，在步驟 S110 中，基於每

一個特徵被所述特徵池挑中的次數來排序這些特徵，以獲得特徵排名。例如，假設特徵 f1 被兩個特徵池挑中，則其計數的次數為 2。因此，可基於各特徵池所挑選到的特徵來計數各特徵的次數。

【0037】 之後，在步驟 S115 中，基於特徵排名，自所述特徵中取出多個指定特徵。在一實施例中，可自所述特徵中取出符合指定數量的指定特徵。

【0038】 在此，指定數量可根據訓練模型的準確率來決定。具體而言，可利用多個測試資料，在選用不同的多個數量的特徵的情況下來獲得各訓練模型的多個準確率。測試資料是用於檢測訓練模型。測試資料只會在檢驗訓練模型時使用，用於評估訓練模型的準確率。在獲得各訓練模型在選擇不同數量的特徵的情況下的準確率之後，基於這些準確率來選擇其中一個訓練模型。例如，選擇具有最高準確率的訓練模型。假設分別選定 10 個不同數量的特徵來檢驗訓練模型的準確率，則一個訓練模型會獲得 10 個準確率。X 個訓練模型則包括 10X 個準確率。在 10X 個準確率中找出最高準確率，以選定具有最高準確率的訓練模型。

【0039】 接著，基於被選擇的訓練模型的數量與準確率來獲得陡坡圖/手肘法（Elbow method），如圖 5 所示。圖 5 是依照本發明一實施例的陡坡圖/手肘法的示意圖。請參照圖 5，橫軸表示特徵數量，縱軸表示準確率。而在另一實施例中，在不同特徵數量下，根據模型預測機率畫出操作特徵曲線（receiver operating characteristic curve，ROC），並計算曲線下面積（area under curve，

AUC），而以 AUC 來作為縱軸。之後，通過陡坡圖/手肘法在特徵數量中獲得指定數量。

【0040】 而在選定挑選方式(3)的情況下，可根據特徵群組將多個特徵池設定為一個特徵集合，之後基於每一個特徵被特徵池挑中的次數來排序各特徵集合中的各特徵，以獲得各特徵集合的特徵排名。以圖 4 而言，將訓練資料集 TD1～TDS 對應於特徵群組 G(1)分別所獲得的 S 個大的特徵池設定為特徵集合 TD(G1)（包括 TD1(G1)～TDS(G1)），並基於特徵集合 TD(G1)來對特徵群組 G(1)中的特徵進行排名，而獲得對應的一組特徵排名 R1。將訓練資料集 TD1～TDS 對應於特徵群組 G(2)分別所獲得的 S 個大的特徵池設定為特徵集合 TD(G2)（包括特徵池 TD1(G2)～TDS(G2)），並基於特徵集合 TD(G2)來對特徵群組 G(2)中的特徵進行排名，而獲得對應的一組特徵排名 R2。以此類推，獲得 N 組特徵排名 R1~RN。之後，基於每一個特徵群組對應的權重以及特徵排名，自各特徵群組中取出對應數量的指定特徵。例如，每一個特徵群組對應的權重是基於每一個特徵群組所包括的特徵數量佔全部特徵數量的比值。假設全部特徵數量為 n 個，特徵類型包括三種，則可分類為三個特徵群組，每一個特徵群組所包括的特徵數量為 n1、n2、n3（n=n1+n2+n3），則其對應權重分別為 n1/n、n2/n、n3/n。可進一步將指定數量乘上對應權重，而自各特徵群組中取出對應數量的指定特徵。

【0041】 圖 6 是依照本發明一實施例的特徵挑選方法的流程圖。

請參照圖 6，在步驟 S605 中，在挑選方式(1)～(3)中選擇至少一種。針對每一種挑選方式，獲得符合對應的指定數量的指定特徵，進而分別獲得對應於所述多個挑選方式的多個選定特徵組。假設選擇挑選方式(2)與挑選方式(3)，則獲得兩個選定特徵組{x1}、{x2}。其中，選定特徵組由圖 1 中步驟 S115 所取出的多個指定特徵所組成。

【0042】 接著，在步驟 S610 中，對選定特徵組{x1}、{x2}執行聯集（{x1}∪{x2}）、交集（{x1}∩{x2}）或差集（{x1}-{x2}或{x2}-{x1}）來獲得整合特徵池{x3}。

【0043】 之後，在步驟 S615 中，透過多體學（multiomics）特徵調控途徑（regulation pathways）分析，查詢多個已知資料庫，以在整合特徵池{x3}中挑選出一或多個代表特徵而獲得另一特徵池{x4}。

【0044】 圖 7 是依照本發明一實施例的多體學特徵調控途徑分析的示意圖。在本實施例中使用的已知資料庫包括：基因資料庫，例如為國家生物技術資訊中心（National Center for Biotechnology Information，NCBI）設置的與生物技術和生物醫學相關的一系列資料庫；代謝資料庫，例如為 MetaCyc 資料庫；基因與蛋白質交互作用資料庫，例如為 BioGRID（Biological General Repository for Interaction Datasets）；基因功能資料庫，例如為 DAVID 資料庫；基因與蛋白質表現量資料庫，例如為人類蛋白質地圖（The Human Protein Atlas）；生醫論文資料庫，例如為 PubMed 資料庫。然，在

此僅為舉例說明，並不以此為限。

【0045】 以基因體學與代謝體學兩個特徵類型而言，查詢基因資料庫來取得被分類至基因體學的特徵對應的基因名稱與相關資訊；並且查詢代謝資料庫來取得被分類至代謝體學的特徵對應的代謝途徑（metabolic pathway）與相關資訊。根據取得的基因名稱進一步查詢基因與蛋白質交互作用資料庫、基因功能資料庫以及基因與蛋白質表現量資料庫，以篩選出現在所述資料庫中的特徵。並且，根據基因名稱及代謝途徑查詢生醫論文資料庫，來取得出現在生醫論文資料庫中的特徵。

【0046】 之後，執行多體學特徵調控途徑分析，即，找出生理機制是哪些基因及代謝物質所引起。一般而言，生理機制是由許多基因互相影響，引起一連串物理及化學反應與代謝物質，且代謝物質又會引發其他反應。而多體學特徵調控途徑分析可找出生理機制是哪些基因及代謝物質所引起。多體學特徵調控途徑分析會考慮生物背後的遺傳變異，結合已知的資料庫，將多體學特徵結合或串聯在一起。故，將多體學特徵調控途徑分析應用於將整合特徵池{x3}中的特徵，可篩選出臨床上有意義的特徵池{x4}。特徵池{x4}中的特徵皆是有關聯性的，例如會相互影響。

【0047】 最後，在步驟 S620 中，自所述代表特徵選出最終特徵。即，衡量特徵池{x4}中其特徵對於疾病的預測表現（例如透過陡坡圖/手肘法），以決定最終的特徵池{x5}，確保特徵池{x5}中的特徵都是臨床上具有意義且對疾病預測的準確率都是醫學上可接受

的。

【0048】 例如，可利用多個測試資料，在分別選用特徵池{x4}中的不同數量的代表特徵的情況下來獲得各訓練模型的多個準確率。之後，基於所述準確率在這些代表特徵中選出一或多個最終特徵。即，利用如圖 5 所述的陡坡圖/手肘法來獲得一數量 F 後，在特徵池{x4}中篩選出最終的特徵池{x5}，其中特徵池{x5}中的特徵數量為 F，再依據圖 4 所取得的特徵排名選取前 F 名得到特徵池{x5}。

【0049】 另外，在另一實施例中，在選定挑選方式(1)的情況下，不需要進行排名，而是直接獲得指定特徵。即，透過每一個訓練模型逐一針對單一特徵計算至少一個統計指標，並將統計指標與對應的臨界值進行比對，以自所述特徵中獲得指定特徵。以圖 2 而言，特徵池 Ps(1)、Ps(2)中的特徵即為指定特徵。

【0050】 另外，在其他實施例中，倘若選定了多個挑選方式中包括挑選方式(1)（針對單一特徵），則可先執行挑選方式(1)做單一特徵，接著，再以挑選方式(1)所挑出來的特徵池去執行其他挑選方式。

【0051】 綜上所述，本揭露利用多個訓練模型來篩選特徵，再根據各特徵被訓練模型挑選到的次數來進行下一步的篩選。據此，可利用最少特徵數量來獲得最好的預測結果

【0052】 另外，本揭露還提供了三種挑選方式(1)～(3)進行初步篩選獲得至少兩個選定特徵組{x1}、{x2}，於初步篩選所獲得的選

定特徵組{x1}、{x2}中進行篩選獲得整合特徵池{x3}，透過多體學特徵調控途徑分析可自整合特徵池{x3}中篩選出特徵池{x4}，再進一步於特徵池{x4}篩選出特徵池{x5}。據此，透過多層級的特徵篩選，不僅節省特徵挑選時間也可挑選出最具影響性的特徵，同時維持高準確率。

【符號說明】

【0053】

f1～fn:特徵

G(1)～G(N):特徵群組

Ms(1)、Ms(2)、M(1)～M(X):訓練模型

p(f1)～p(fn):P 值

Ps(1)、Ps(2)、Pm(1)～Pm(X)、P1(G1)～PX(GN):特徵池

r(f1)～r(fn):勝算比

TD1～TDS:訓練資料集

S105～S115:特徵挑選方法的步驟

S205、S210:步驟

S605～S620:特徵挑選方法的步驟

【發明申請專利範圍】

【請求項1】 一種特徵挑選方法，利用一電子裝置在多個特徵中進行挑選，該特徵挑選方法包括：

將多個訓練資料分別輸入至多個訓練模型，使得每一該些訓練模型分別在該些特徵中進行挑選，而獲得多個特徵池，其中每一該些訓練模型獲得該些特徵池的至少其中一個；

基於每一該些特徵被該些特徵池挑中的次數來排序該些特徵，以獲得一特徵排名；以及

基於該特徵排名，自該些特徵中取出多個指定特徵。

【請求項2】 如請求項1所述的特徵挑選方法，其中每一該些訓練模型分別在該些特徵中進行挑選，而獲得該些特徵池的步驟包括：透過一第一挑選方式、一第二挑選方式以及一第三挑選方式中的至少其中一種，自該些特徵中進行挑選以獲得該些特徵池，其中

在該第一挑選方式中，每一該些訓練模型對應至一個特徵池，該第一挑選方式包括：透過每一該些訓練模型逐一針對該些特徵中的單一特徵計算至少一統計指標，並將該統計指標與對應的一臨界值進行比對，藉此決定是否挑選該單一特徵至每一該些訓練模型對應的特徵池；

在該第二挑選方式中，每一該些訓練模型對應至一個特徵池，該第二挑選方式包括：透過每一該些訓練模型在該些特徵中執行一特徵擷取動作，藉此在該些特徵中挑選多個至每一該些訓

練模型對應的該些特徵池；

在該第三挑選方式中，每一該些訓練模型具有對應至多個特徵類型的多個特徵池，該第三挑選方式包括：基於該些特徵類型，將該些特徵分類為多個特徵群組，透過每一該些訓練模型在每一該些特徵群組所包括的特徵中執行該特徵擷取動作，藉此在每一該些特徵群組中挑選多個特徵至每一該些訓練模型所包括的對應至每一該些特徵群組的特徵池。

【請求項3】 如請求項2所述的特徵挑選方法，其中透過該第三挑選方式獲得的該些特徵池之後，更包括：

將每一該些特徵群組對應的該些特徵池設定為一個特徵集合；

基於每一個特徵被該些特徵池挑中的次數來排序各特徵集合中的各特徵，以獲得各特徵集合的該特徵排名；以及

基於每一該些特徵群組對應的一權重以及該特徵排名，自各特徵集合中取出對應數量的指定特徵，

其中，每一該些特徵群組分類對應的該權重是基於每一該些特徵群組分類所包括的特徵數量佔全部特徵數量的比值。

【請求項4】 如請求項2所述的特徵挑選方法，其中在該第一挑選方式、該第二挑選方式以及該第三挑選方式中選擇多個的情況下，更包括：

針對該第一挑選方式、該第二挑選方式以及該第三挑選方式的每一個，獲得符合對應的指定數量的該些指定特徵，進而分別

獲得對應於該第一挑選方式、該第二挑選方式以及該第三挑選方式的多個選定特徵組。

【請求項5】 如請求項4所述的特徵挑選方法，其中在分別獲得對應於該第一挑選方式、該第二挑選方式以及該第三挑選方式的該些選定特徵組之後，更包括：

對該些選定特徵組執行一聯集、一交集以及一差集其中一者來獲得一整合特徵池。

【請求項6】 如請求項5所述的特徵挑選方法，其中在獲得該整合特徵池之後，更包括：

透過一多體學特徵調控途徑分析，查詢多個已知資料庫，以在該整合特徵池中挑選出一或多個代表特徵。

【請求項7】 如請求項6所述的特徵挑選方法，其中在獲得所述一或多個代表特徵之後，更包括：

利用多個測試資料，在分別選用不同的多個特徵數量的該些代表特徵的情況下來獲得每一該些訓練模型的多個準確率；以及

基於該些準確率在該些代表特徵中選出一或多個最終特徵。

【請求項8】 如請求項1所述的特徵挑選方法，其中基於該特徵排名，自該些特徵中取出該些指定特徵的步驟包括：

基於該特徵排名，自該些特徵中取出符合一指定數量的該些指定特徵，

其中，在獲得該些特徵池之後，更包括：

利用多個測試資料，在選用不同的多個特徵數量的特徵的情

況下來獲得每一該些訓練模型的多個準確率；

基於該些準確率，在該些訓練模型中選擇其中一個；

基於被選擇的其中一個訓練模型的該些特徵數量與該些準確率來獲得一陡坡圖；以及

基於該陡坡圖，在該些特徵數量中獲得該指定數量。

【請求項9】 如請求項1所述的特徵挑選方法，其中在獲得該些指定特徵之後，更包括：

透過一多體學特徵調控途徑分析，查詢多個已知資料庫，以在該些指定特徵中挑選出一或多個代表特徵。

【請求項10】 如請求項9所述的特徵挑選方法，其中在獲得所述多個代表特徵之後，更包括：

利用多個測試資料，在分別選用不同的多個特徵數量的該些代表特徵的情況下來獲得每一該些訓練模型的多個準確率；以及

基於該些準確率在該些代表特徵中選出一或多個最終特徵。

【請求項11】 一種特徵挑選方法，利用一電子裝置在多個特徵中進行挑選，該特徵挑選方法包括：

透過一第一挑選方式、一第二挑選方式以及一第三挑選方式中的其中一種，將多個訓練資料分別輸入至多個訓練模型，使得每一該些訓練模型分別在該些特徵中進行挑選以獲得多個特徵池；

基於每一該些特徵被該些特徵池挑中的次數來排序該些特徵，以獲得一特徵排名；以及

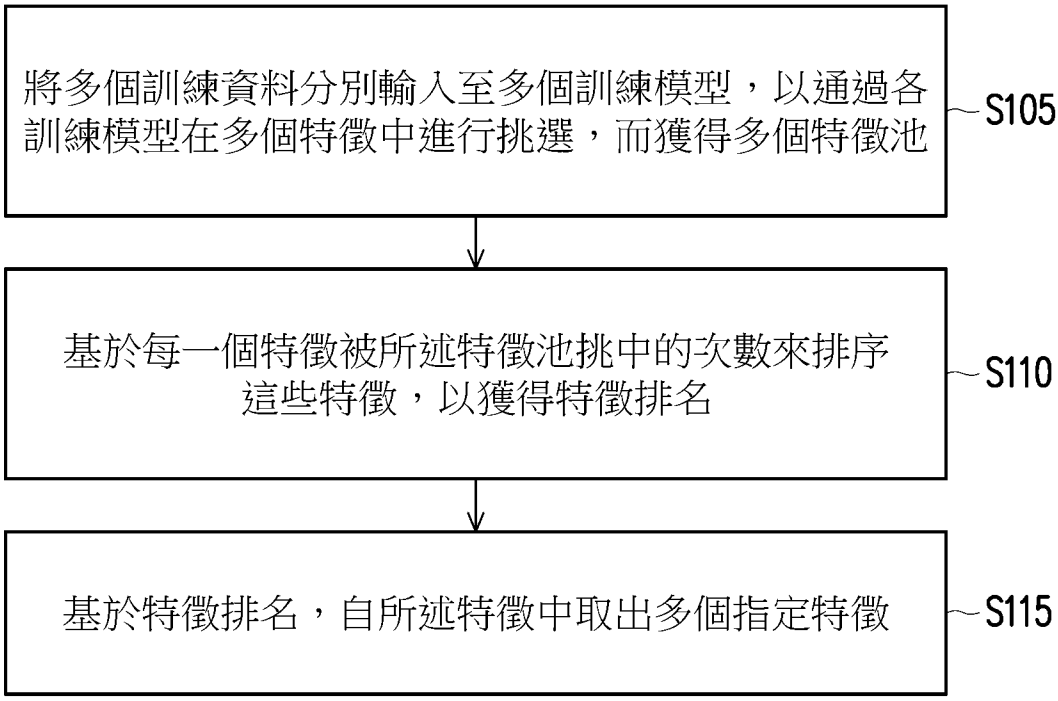
基於該特徵排名，自該些特徵中取出多個指定特徵，

其中，在該第一挑選方式中，每一該些訓練模型對應至一個特徵池，該第一挑選方式包括：透過每一該些訓練模型逐一針對該些特徵中的單一特徵計算一統計指標，並將該統計指標與對應的一臨界值進行比對，以決定是否挑選該單一特徵至每一該些訓練模型對應的特徵池；

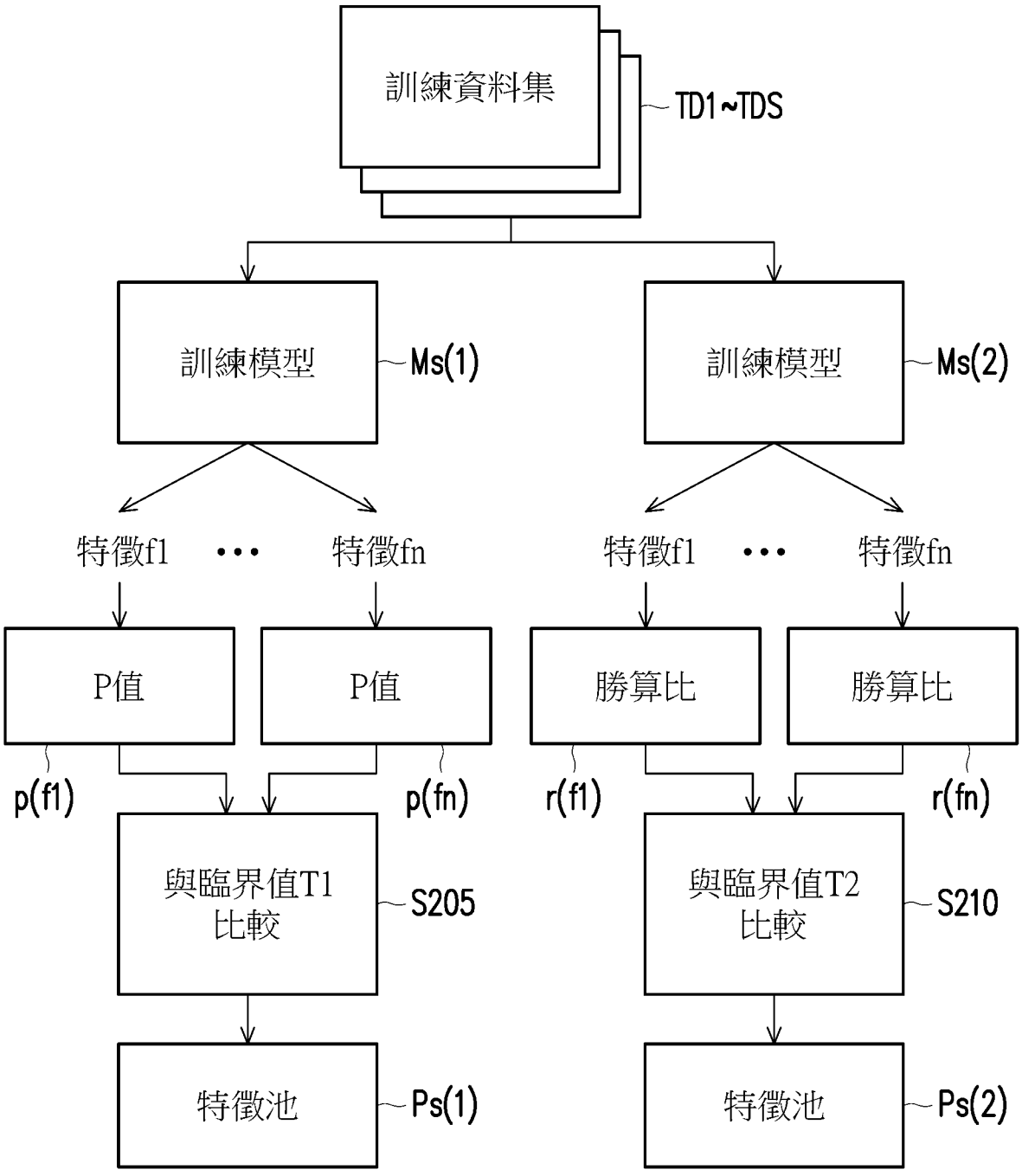
在該第二挑選方式中，每一該些訓練模型對應至一個特徵池，該第二挑選方式包括：透過每一該些訓練模型在該些特徵中執行一特徵擷取動作，藉此在該些特徵中挑選多個至每一該些訓練模型對應的特徵池；

在該第三挑選方式中，每一該些訓練模型具有對應至多個特徵類型的多個特徵池，該第三挑選方式包括：基於該些特徵類型，將該些特徵分類為多個特徵群組，透過每一該些訓練模型在每一該些特徵群組所包括的特徵中執行該特徵擷取動作，藉此在每一該些特徵群組中挑選多個特徵至每一該些訓練模型所包括的對應至每一該些特徵群組的特徵池。

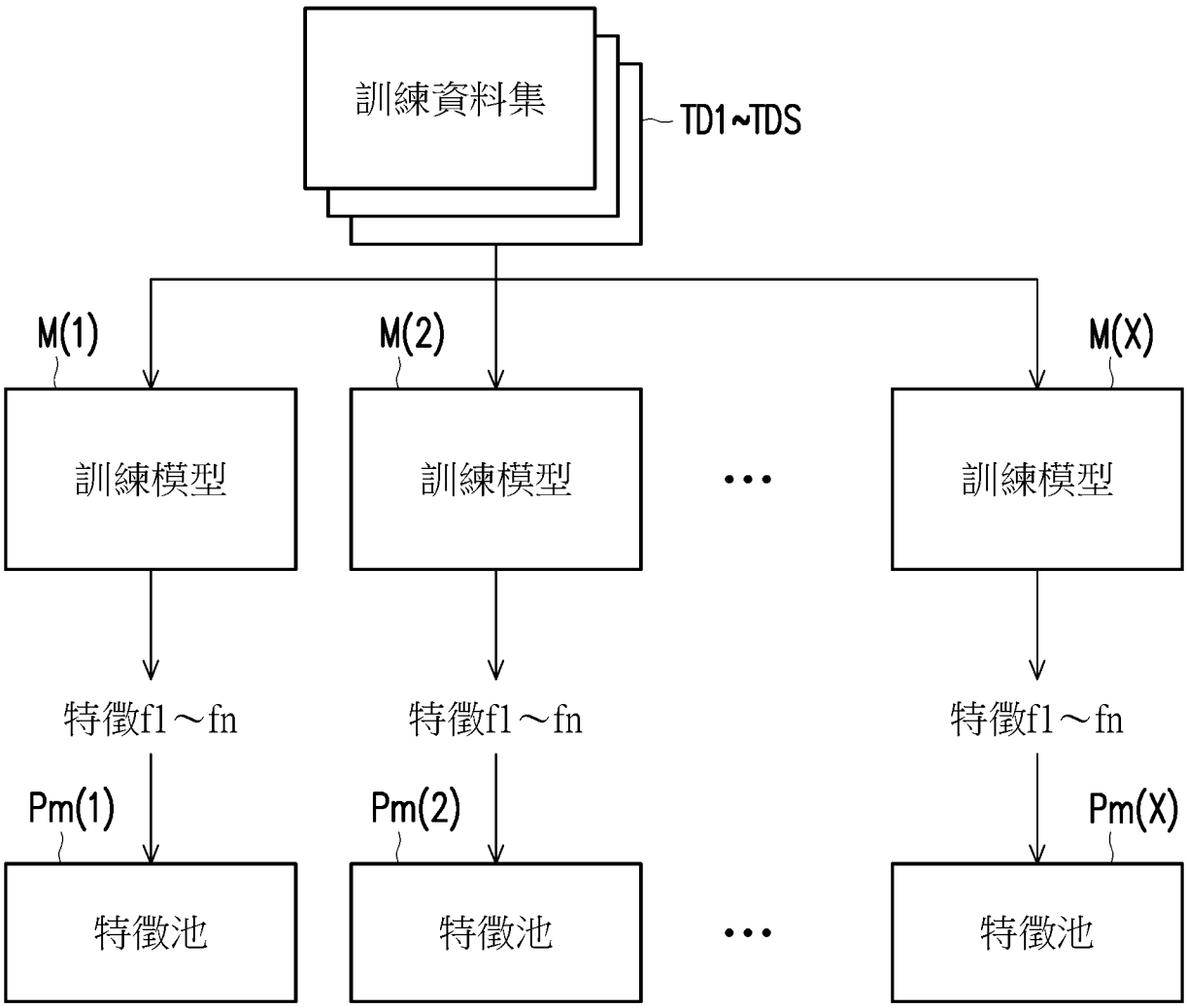
【發明圖式】


將多個訓練資料分別輸入至多個訓練模型，以通過各訓練模型在多個特徵中進行挑選，而獲得多個特徵池
S105
基於每一個特徵被所述特徵池挑中的次數來排序這些特徵，以獲得特徵排名
S110
基於特徵排名，自所述特徵中取出多個指定特徵
S115

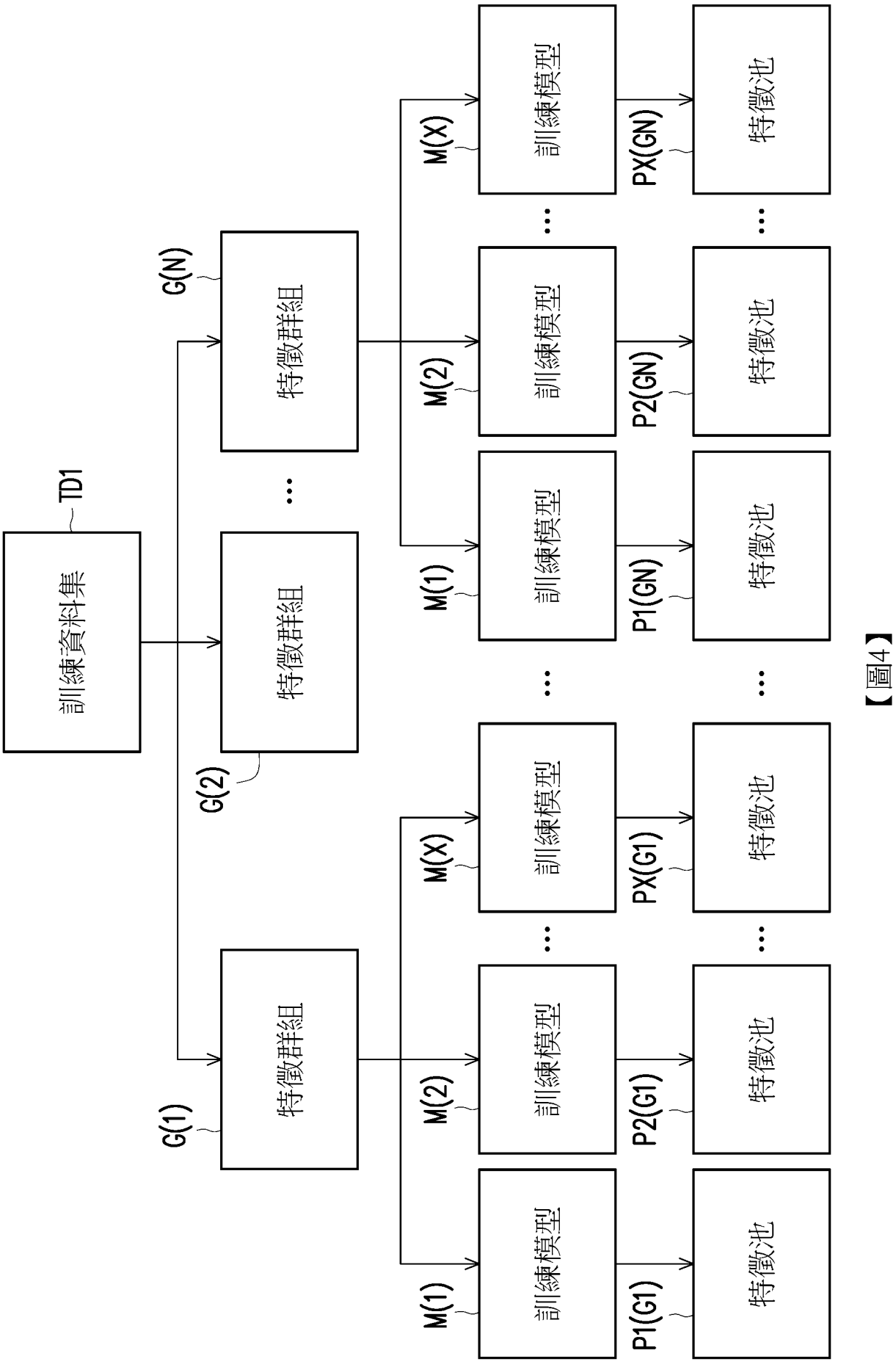
【圖1】


訓練資料集
TD1~TDS
訓練模型
Ms(1)
訓練模型
Ms(2)
特徵f1
特徵fn
特徵f1
特徵fn
P值
P值
勝算比
勝算比
p(f1)
p(fn)
r(f1)
r(fn)
與臨界值T1
比較
S205
與臨界值T2
比較
S210
特徵池
Ps(1)
特徵池
Ps(2)

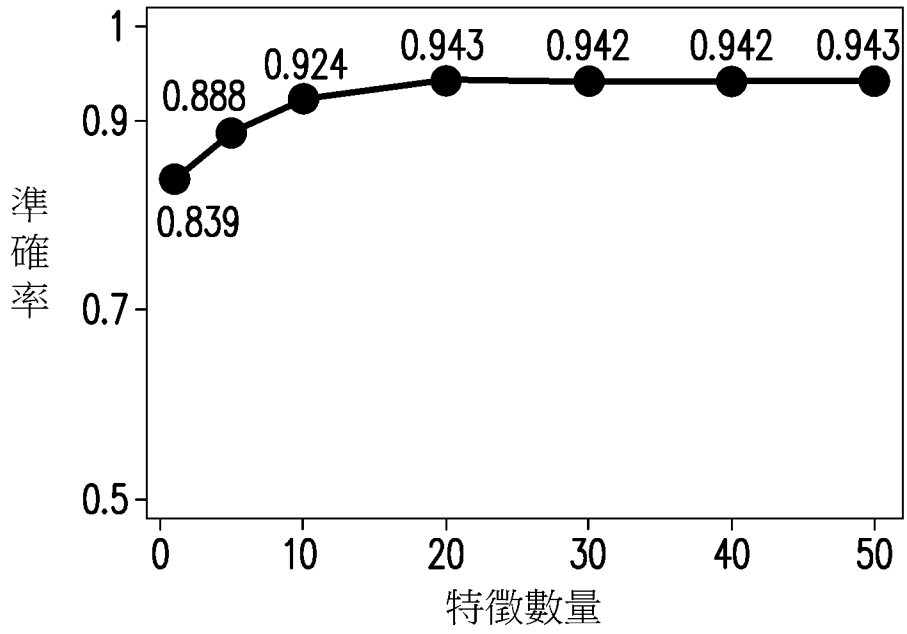
【圖2】


訓練資料集
TD1~TDS
M(1)
M(2)
M(X)
訓練模型
訓練模型
訓練模型
特徵f1～fn
特徵f1～fn
特徵f1～fn
Pm(1)
Pm(2)
Pm(X)
特徵池
特徵池
特徵池

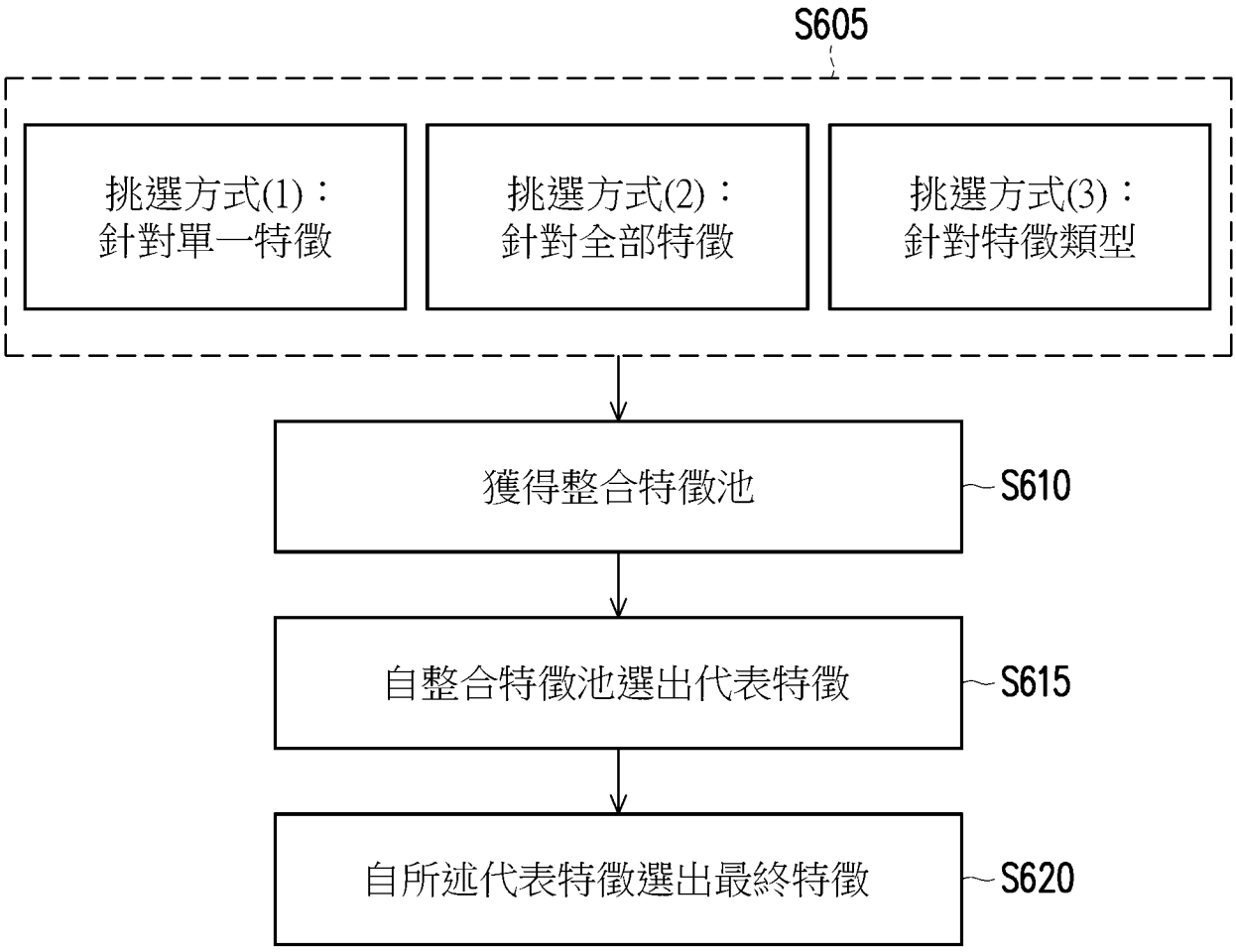
【圖3】


訓練資料集
TD1
特徵群組
G(1)
特徵群組
G(2)
特徵群組
G(N)
M(1)
訓練模型
P1(G1)
特徵池
M(2)
訓練模型
P2(G1)
特徵池
M(X)
訓練模型
PX(G1)
特徵池
M(1)
訓練模型
P1(GN)
特徵池
M(2)
訓練模型
P2(GN)
特徵池
M(X)
訓練模型
PX(GN)
特徵池

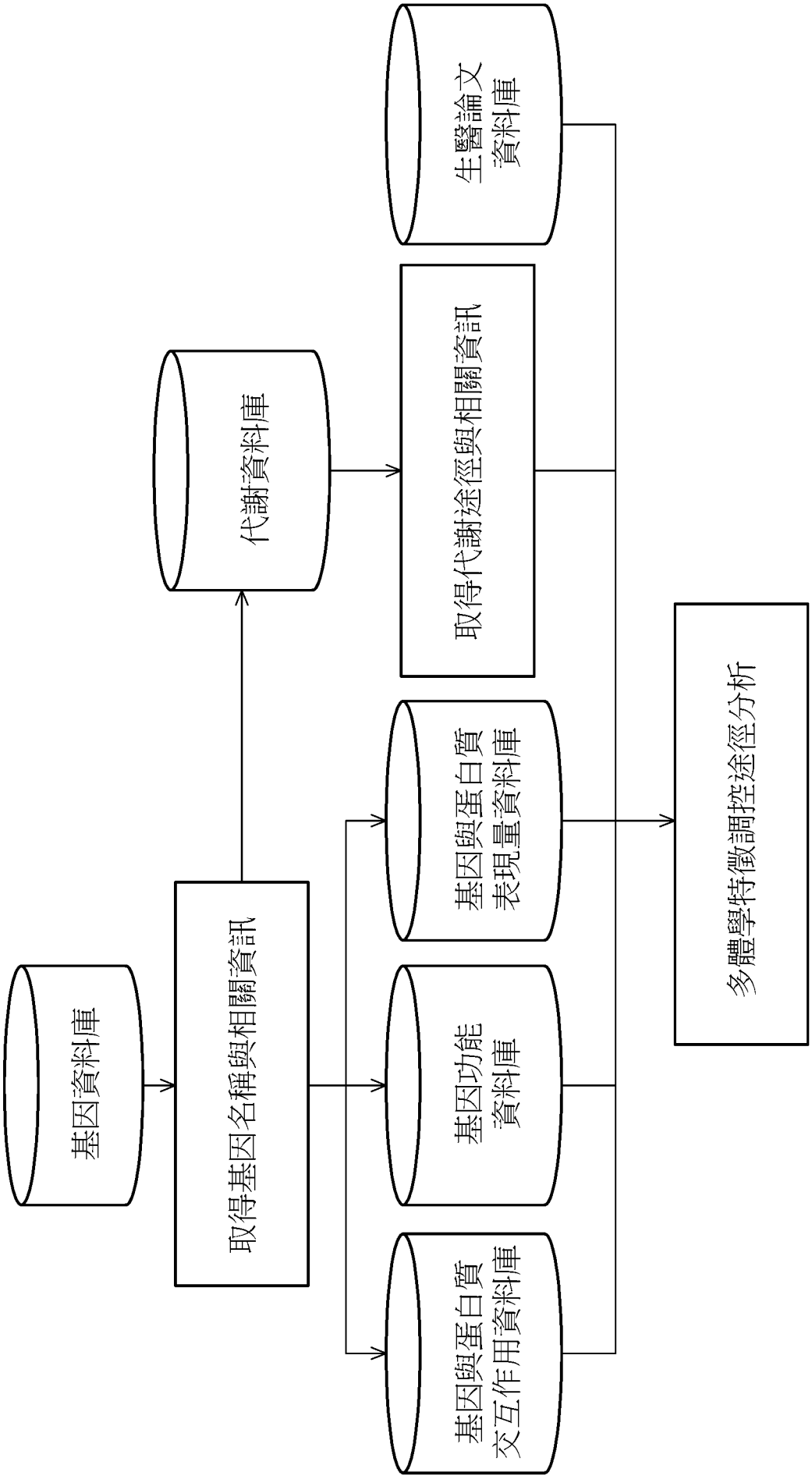
【圖4】


1
0.9
0.7
0.5
準確率
0.888
0.924
0.943
0.942
0.942
0.943
0.839
0
10
20
30
40
50
特徵數量

【圖5】


S605
挑選方式(1)：
針對單一特徵
挑選方式(2)：
針對全部特徵
挑選方式(3)：
針對特徵類型
獲得整合特徵池
S610
自整合特徵池選出代表特徵
S615
自所述代表特徵選出最終特徵
S620

【圖6】


基因資料庫
取得基因名稱與相關資訊
代謝資料庫
取得代謝途徑與相關資訊
生醫論文資料庫
基因與蛋白質表現量資料庫
基因功能資料庫
基因與蛋白質交互作用資料庫
多體學特徵調控途徑分析

【圖7】