(19)中華民國智慧財產局
(12)發明說明書公開本 (11)公開編號：TW 201224812 A1
(43)公開日： 中華民國 101 (2012) 年 06 月 16 日

(21)申請案號：100133243 (22)申請日： 中華民國 100 (2011) 年 09 月 15 日
(51)Int. Cl.： *G06F17/30 (2006.01)*
(30)優先權：2010/10/05 歐洲專利局 10186540.0
(71)申請人：萬國商業機器公司 (美國) INTERNATIONAL BUSINESS MACHINES CORPORATION (US)
美國
(72)發明人：林根菲爾德克里斯多福 LINGENFELDER, CHRISTOPH (DE)；沃斯特麥可 WURST, MICHAEL (DE)；龐貝帕斯卡 POMPEY, PASCAL (FR)
(74)代理人：蔡坤財；李世章
申請實體審查：無 申請專利範圍項數：14 項 圖式數：6 共 45 頁

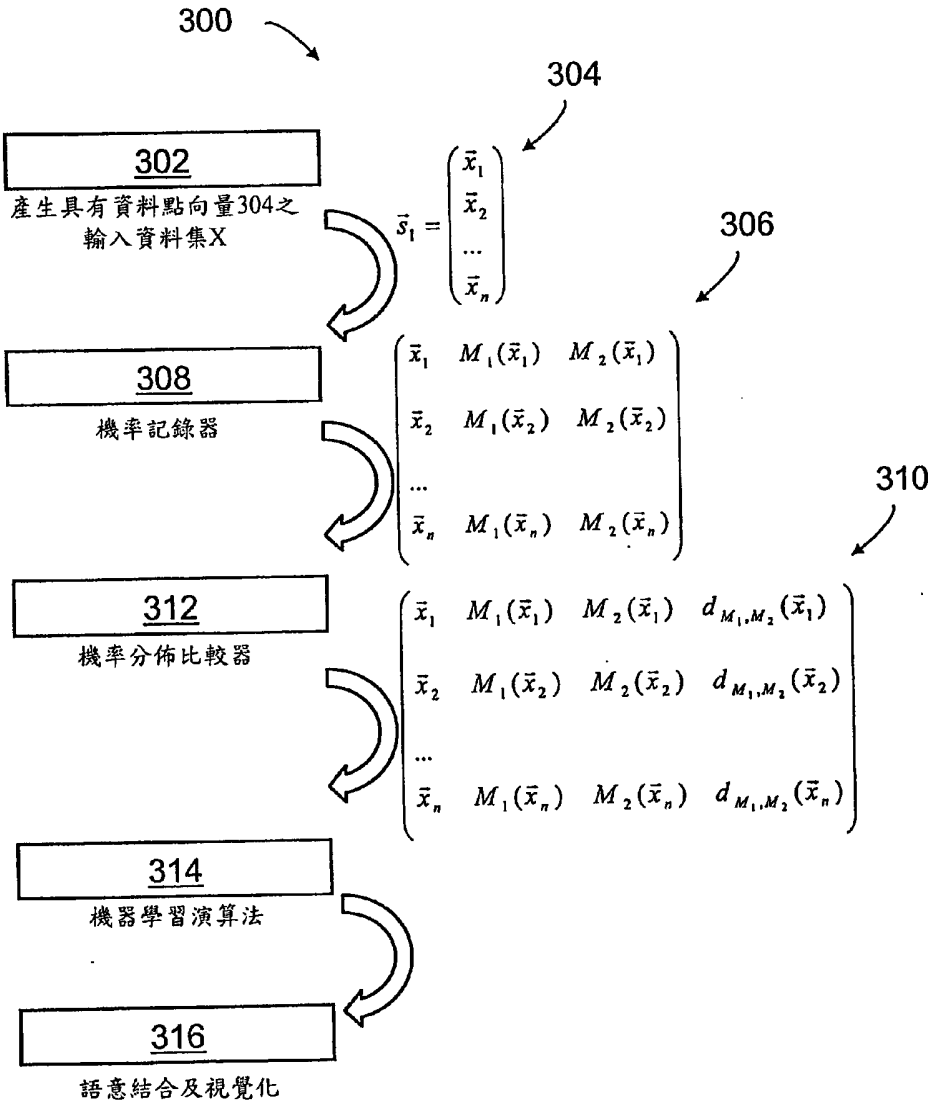
(54)名稱

機率性資料探勘模型比較引擎

PROBABILISTIC DATA MINING MODEL COMPARISON ENGINE

(57)摘要

本發明揭示了一種用於比較第一資料探勘模型與第二資料探勘模型之比較引擎。第一資料探勘模型 M1 表示對第一資料集 D1 執行第一資料探勘任務之結果，且該第一資料探勘模型 M1 提供第一預測值集。第二資料探勘模型 M2 表示對第二資料集 D2 執行第二資料探勘任務之結果，且該第二資料探勘模型 M2 提供第二預測值集。決定該等預測值集之間的關係 R。針對輸入資料集之至少第一記錄，第一機率分佈及第二機率分佈係基於應用於第一記錄之第一資料探勘模型及第二資料探勘模型而形成，該等機率分佈將機率與該等預測值集相關聯。使用第一機率分佈及第二機率分佈及其關係來計算該第一記錄之距離量測 d。至少一個感興趣區域(region of interest)係基於該距離量測 d 來決定。




300：透視圖
302：方塊
304：資料點向量
306：矩陣
308：機率記錄器
310：資料集/矩陣
312：機率性分佈比較器
314：方塊
316：步驟

(19)中華民國智慧財產局
(12)發明說明書公開本 (11)公開編號：TW 201224812 A1
(43)公開日： 中華民國 101 (2012) 年 06 月 16 日

(21)申請案號：100133243 (22)申請日： 中華民國 100 (2011) 年 09 月 15 日
(51)Int. Cl.： *G06F17/30 (2006.01)*
(30)優先權：2010/10/05 歐洲專利局 10186540.0
(71)申請人：萬國商業機器公司 (美國) INTERNATIONAL BUSINESS MACHINES CORPORATION (US)
美國
(72)發明人：林根菲爾德克里斯多福 LINGENFELDER, CHRISTOPH (DE)；沃斯特麥可 WURST, MICHAEL (DE)；龐貝帕斯卡 POMPEY, PASCAL (FR)
(74)代理人：蔡坤財；李世章
申請實體審查：無 申請專利範圍項數：14 項 圖式數：6 共 45 頁

(54)名稱

機率性資料探勘模型比較引擎

PROBABILISTIC DATA MINING MODEL COMPARISON ENGINE

(57)摘要

本發明揭示了一種用於比較第一資料探勘模型與第二資料探勘模型之比較引擎。第一資料探勘模型 M1 表示對第一資料集 D1 執行第一資料探勘任務之結果，且該第一資料探勘模型 M1 提供第一預測值集。第二資料探勘模型 M2 表示對第二資料集 D2 執行第二資料探勘任務之結果，且該第二資料探勘模型 M2 提供第二預測值集。決定該等預測值集之間的關係 R。針對輸入資料集之至少第一記錄，第一機率分佈及第二機率分佈係基於應用於第一記錄之第一資料探勘模型及第二資料探勘模型而形成，該等機率分佈將機率與該等預測值集相關聯。使用第一機率分佈及第二機率分佈及其關係來計算該第一記錄之距離量測 d。至少一個感興趣區域(region of interest)係基於該距離量測 d 來決定。

六、發明說明：

【發明所屬之技術領域】

本發明大體而言係關於一種用於執行資料探勘比較引擎之方法以及一種資料探勘比較引擎。

本發明進一步係關於一種電腦系統、一種資料處理程式及一種電腦程式產品。

【先前技術】

資料探勘為分析處理，該分析處理設計成探索資料（通常，為大量資料）以搜尋變數之間的相同樣式及/或系統關係，隨後藉由將偵測樣式應用至新的資料集來驗證研究結果(findings)。預測型資料探勘為最常用類型之資料探勘且具有最直接應用。資料探勘之製程由三個階段組成：(1)初始探索，(2)用驗證/確認進行模型建立或樣式辨識，及(3)部署，亦即，將模型應用至新的資料以產生預測。

初始探索階段通常從資料準備開始，該資料準備可包含清除資料、資料轉換、選擇記錄子集及在具有大量變數（「欄位或維度」）之資料集之情況下執行一些初步特徵選擇操作以使變數數量進入可管理範圍（取決於考量之統計方法）。隨後，視分析問題之性質而定，此資料探勘處理之第一階段可能涉及簡單選擇回歸(regression)模型之直接預測子(predictor)與使用各種圖解方法及統計方法之精緻探索分析之間的任何活動，以識別最相關變

數且決定可在下一階段中納入考量之模型的複雜性及/或通用性質。

第二階段——用驗證/確認進行模型建立或樣式辨識——涉及考量各種模型及基於各種模型之預測效能來選擇最佳模型(亦即，闡釋考慮中的變化性(variability in question)及產生取樣間之穩定結果)。此舉聽似簡單操作，但實際上此舉有時涉及極精緻製程。存在經開發以實現彼目標之各種技術，該等技術中之許多技術係基於所謂的「模型之競爭評估」，亦即，將不同的模型應用至相同的資料集，隨後比較不同模型之效能以選擇最佳模型。此等技術——通常被視為預測資料探勘之核心——包括：裝袋(bagging)(投票、平均化)、加速(boosting)、堆疊(stacking)(堆疊通用化)及元學習(meta-learning)。

第三階段——部署——涉及使用在上述階段中選為最佳的模型及將該模型應用至新的資料，以產生預期結果的預測或評估。

熟知的資料探勘類別包括叢集分析(cluster analysis)、線性回歸及非線性回歸、分類、規則分析及時間序列分析。分群可定義為在不使用資料之已知結構之情況下發現資料之群組及結構的任務，該資料之群組及結構之成員在某一方面或其他方面「類似」。分類可定義為使待應用於新的資料之已知結構通用化之任務。例如，電子郵件程式可試圖將接收的電子郵件分類為合法郵件或垃圾

郵件。常用演算法包括決策樹學習、最近鄰居(nearest neighbor)、天真貝氏(Naive Bayesian)分類及類神經網路。

回歸分析試圖找出函數，該函數以最小誤差來模組化資料。

關聯規則學習搜尋變數之間的關係。例如，超市可收集顧客購買習性之資料。使用關聯規則學習，超市可決定哪些產品經常被一起購買且可使用此資訊來達成行銷目的。此舉有時稱為購物籃分析(market basket analysis)。

在期望在有限確定性之狀況下顯示可指導決策之知識結構的情況下，資料探勘之概念作為資訊管理工具日益風靡。由於涉及大量資料點之緣故，藉由使用人力技術將不可能進行此舉。

然而，為了有效地使用資料探勘技術，可能需要比較資料探勘模型以從現存資料獲得最佳結果。

資料探勘模型之比較可適用於不同情境。許多應用情境並非具有單個而是多個與此相關的資料探勘模型。一些典型實例為在不同時間點或在不同資料子集中導出的資料探勘模型，例如來自不同生產地點之生產品質資料。另一個常見狀況為以不同類型之資料探勘模型表示具有資料探勘模型之相同資料，以擷取不同資料態樣。在所有此等狀況中，不僅對個別資料探勘模型感興趣，而且亦對個別資料探勘模型之間的相似性及差異感興趣。此種差異可能告知，例如，生產品質及相依性如何隨時間發展；不同類型之資料探勘模型在表示為相同設

施中生產不同產品之方式上如何不同，或者生產設施彼此之間如何不同。

視可用資料量而定，利用人力來比較資料探勘模型可能極昂貴、易出錯且不可行。儘管極其重要，但資料探勘模型之自動比較在實務上尚未被廣泛採納，主要是因為兩個原因：(a)資料探勘模型之自動比較僅允許比較相同且預定義樣式類型之模型，因此，資料探勘模型之自動比較缺乏通用性，從而不可能對大多數其他樣式類型使用該等方法。(b)資料探勘模型之自動比較係基於資料探勘模型之結構，因此，資料探勘模型之自動比較在資料探勘模型之自動比較之表達性方面受到嚴格限制，從而導致通常極難解譯之複雜結果。

文件US 7,636,698揭示一種基於兩個分類資料探勘模型之比較自資料集以產生資料樣式——或資料探勘模型——之方法。揭示了一種架構，該架構用於分析資料探勘模型之資料樣式之樣式轉移並輸出結果。此舉允許比較且描述兩個語意類似分類樣式——或分類探勘模型——之間的差異，且此舉允許分析相同分類模型版本之歷史變化，或此舉允許分析由應用於相同資料的兩個或更多個分類演算法發現之樣式差異。

因此，可能需要一種改進方法及一種用於比較資料探勘模型之引擎，尤其針對資料探勘模型並非屬於相同類別之資料探勘模型之狀況。

【發明內容】

可藉由根據獨立項之一種用於執行資料探勘比較引擎之方法、一種資料探勘比較引擎、一種電腦系統、一種資料處理程式及一種電腦程式產品來解決此需要。

本發明提供一種用於執行資料探勘模型比較引擎以供比較第一資料探勘模型與第二資料探勘模型之方法。該方法包含以下步驟：

提供第一資料探勘模型，該第一資料探勘模型表示對第一資料集執行第一資料探勘任務之結果，該第一資料探勘模型提供第一預測值集；

提供第二資料探勘模型，該第二資料探勘模型表示對第二資料集執行第二資料探勘任務之結果，該第二資料探勘模型提供第二預測值集；

提供該第一預測值集與該第二預測值集之間的關係；

決定輸入資料集；

對決定輸入資料集之至少一第一記錄實施以下步驟：

基於應用於第一記錄之第一資料探勘模型來形成第一機率分佈，該第一機率分佈將機率與該第一預測值集相關聯；

基於應用於第一記錄之第二資料探勘模型來形成第二機率分佈，該第二機率分佈將機率與該第二預測值集相關聯；

使用第一機率分佈及第二機率分佈及其關係來計算第一記錄之距離量測；

基於距離量測決定至少一個感興趣區域(region of interest)，該距離量測係針對決定輸入資料集之記錄計算。

一種資料探勘比較引擎，該資料探勘比較引擎包含：

提供單元，該提供單元適合於提供第一資料探勘模型M1及第二資料探勘模型M2，該第一資料探勘模型M1提供第一預測值集，該第二資料探勘模型M2提供第二預測值集，該第一資料探勘模型M1表示對第一資料集D1執行第一資料探勘任務之結果，而該第二資料探勘模型M2表示對第二資料集D2執行第二資料探勘任務之結果，其中提供單元亦適合於提供該第一預測值集與該第二預測值集之間的關係R；

輸入資料集單元，該輸入資料集單元適合於決定輸入資料集X；

計算單元，該計算單元適合於針對輸入資料集之每一個記錄：

基於應用於該輸入資料集之第一資料探勘模型計算第一機率分佈，該第一機率分佈將機率與該第一預測值集相關聯；

基於應用於該輸入資料集之第二資料探勘模型計算第二機率分佈，該第二機率分佈將機率與該第二預測值集相關聯；

使用第一機率分佈及第二機率分佈及其關係來計算該等記錄之距離量測d。

區域發現單元，該區域發現單元適合於基於距離量測d決定至少一個感興趣區域，該距離量測d係針對輸入資料集之記錄計算。

【實施方式】

**詳細描述**

由於在發明內容部分中描述了方法之普通概念，所以在此處詳述一些更具體內容。至少兩個資料探勘模型M1、M2之比較係基於比較由資料探勘模型提供之預測結果的機率。本文中提到之使用預測結果之機率的比較之一個優點在於，該比較包含該等模型之不確定性，該等模型之不確定性與任何先前技術之資料探勘模型之比較之不確定性不同。

由於資料探勘模型M1、M2係基於原始資料集D1、D2產生，而原始資料集D1、D2可含有不同記錄，所以對輸入資料集X進行M1與M2之比較，該輸入資料集X通常為

此而產生。即使D1等於D2，當比較模型時D1及D2之資料記錄也可能不再可用。

第一資料集D1與第二資料集D2可能為量測資料點之群組。每一個資料點可理解為具有複數個維度之向量，所有維度均具有可能不同類型之資料值。舉例而言，在涉及天氣之量測中，一個資料點可能包含針對以下量之值：溫度、壓力、濕度、風向、風力強度以及日期、時間及地理位置。一群組之此種資料點向量可能為用於發展資料探勘模型之開始位置，例如，如所闡釋，用於預測次日天氣狀況之天氣模型，該天氣模型自該群組之資料點向量開始。

第一資料集D1及第二資料集D2中之資料點（記錄）可含有相同量之值；例如該等資料點（記錄）可表示在兩個不同時間點或兩個不同位置中之整個量測結果集。資料集D1、D2甚至可以為單個相同資料集。然而，第一資料集D1及第二資料集D2中之資料點有可能含有不同數目之量或不同量。例如，第一資料集D1可含有作為第一資料集D1之記錄之一部分之濕度值，而第二資料集D2無濕度值。第一資料探勘模型M1為對第一資料集D1執行第一資料探勘任務之結果，而第二資料探勘模型M2類似地涉及第二資料集D2。為比較資料探勘模型M1及M2，可能需要常用輸入資料集X。此輸入資料集X可基於原始資料

集D1、D2及/或資料探勘模型M1、M2而產生。一個選擇可為選擇來自D1及/或D2之記錄。此輸入資料集之記錄需要含有針對至少彼等量之值，第一資料探勘模型及第二資料探勘模型需要該等值作為輸入。輸入資料集X可與原始資料集D1、D2中之一者或兩者（若D1、D2為相同資料集）相同。

資料探勘模型M1及M2可提供不同預測。若第一資料探勘任務T1及第二資料探勘任務T2不同（例如，分群任務、分類任務、回歸任務），但亦可針對相同任務而出現，例如若M1含有叢集A1、叢集B1及叢集C1，而M2含有A2及B2，則情況如此。

M1提供第一預測值集（例如，數個叢集）而M2提供第二預測值集（例如，不同數目之叢集；基於回歸之預測；標記/類別集）。為能夠比較M1與M2，需要提供該第一預測值集與該第二預測值集之間的關係R。若該第一預測值集與該第二預測值集相同，則此關係R可為恆等關係。在詳細實例中，在下文給出了此種關係之一些其他實例。基於彼等實例，熟習此項技術者很清楚如何形成此種關係。即使探勘任務T1、T2為相同的探勘任務，所得資料探勘模型M1、M2也可能由於資料集D1、D2可能不同或者因為演算法或演算法之參數化可能不同而不同。因此，在本專利申請案中所闡釋之方法亦可用以比較與探

勘任務T1、T2有關之資料探勘模型M1、M2，該等探勘任務T1、T2為相同的探勘任務。

此外，資料探勘模型M1與資料探勘模型M2之比較使用M1及M2指派給預測值的機率，該等預測值與輸入資料集記錄有關。機率之比較允許進行資料探勘模型之比較，該等資料探勘模型結構不同或涉及不同探勘任務（例如，使用多項式之不同最高次序之兩個回歸模型；分群模型及回歸模型）。

吾等再參閱原始資料集D1、D2。資料點向量之每一個數字維度具有兩個極值，亦即，最高值及最低值。所有極值定義感興趣定義域(domain of interest)。因而，資料集之所有資料點向量均處於感興趣定義域內。在數學上，感興趣定義域定義所有資料點向量之空間或僅定義資料集之空間。熟習此項技術者知道如何將此定義延伸至分類維度。

資料探勘模型一旦經定義，即亦可產生全新的資料集，該全新的資料集遵循原始資料集之規則。若原始資料點向量不再可用，則此舉可適用。在此類情況下，可使用特定資料探勘模型來產生新的資料集。然而，此處亦可能，資料點向量中之一些資料點向量可處於邊界之外，例如，資料點向量中之一些資料點向量具有一或更多維度之資料值，其中一值略處於相應極值之外。同時，彼等資料點向量可能有資格用於比較資料探勘模型。因此，輸入資料集X可含有（一些或所有）來自原始資料集

D1、D2之記錄，及/或該輸入資料集X可含有基於資料探勘模型M1、M2而產生之記錄。在一些情況下，輸入資料集X可等於D1或D2（或者所有三個集合可能均相等）。

輸入資料集X中之每一個資料點向量$\vec{x}_i$，可由下式表示

$$\vec{x}_i = \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_k \end{pmatrix} \quad (1)$$

值$v_1$、$v_2$、……、$v_k$均可表示天氣模型中以上所述之值，亦即，$v_1$可表示溫度值，$v_2$可為壓力值等。在其他應用中，可使用其他值。

輸入資料集X可由k乘n矩陣表示，該k乘n矩陣包含n個資料點向量，其中每一個資料點具有k個維度，意謂k個不同無向量之測量值(scalar measurement values)已被看作為每一個資料點：

$$\begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \cdots \\ \vec{x}_n \end{pmatrix} \quad (2)$$

通用模型可被理解為矩陣，其中維度係為n乘(k+2)，其中M1及M2為兩個待比較之資料探勘模型且每一個資料探勘模型M1、M2為一個量提供一個預測：

$$\begin{pmatrix} \xi_1 & M_1(\xi_1) & M_2(\xi_1) \\ \xi_2 & M_1(\xi_2) & M_2(\xi_2) \\ \cdots & & \\ \xi_n & M_1(\xi_n) & M_2(\xi_n) \end{pmatrix} \quad (3)$$

每一個值$M_a(V_{i,1}, V_{i,2}, \ldots, V_{i,k})$可為一個機率分佈$P_a$，其中「a」等於1或2，且其中「i」表示第i個資料點向量。對分群及分類而言，此分佈為離散型機率分佈，對回歸而言，此分佈亦可為連續機率分佈。

在更複雜的情況下，該等機率分佈並非基於兩個模型之相同可能值集或相同數字範圍集。在此情況下，通用模型亦含有關係R，以匹配該等分佈。舉例而言，此舉可由個別可能值之笛卡爾乘積(Cartesian product)之聯合分佈表示。

可假定，在簡單的天氣模型中，在第一列中位於行k+1處之矩陣值$M_1(\xi_1)$表示次日天氣狀況之機率分佈。可能之天氣狀況可包括「晴朗」、「多雲」或「下雨」。例如，機率分佈可表示為{60%晴朗、30%多雲、10%下雨}。

根據另一個資料探勘模型$M_2$，預測分佈$M_2(\xi_1)$可為{80%晴朗、10%多雲、10%下雨}。在最簡單方式中，距離量測可為機率值對之間的絕對差之總和，亦即，|0.6-0.8|+|0.3-0.1|+|0.1-0.1|=0.4。替代性差異量測包括庫貝克-李柏(Kullback-Leibler)距離。

當處理兩個分佈中之不同可能值時，一個較佳實施例為，將$M_1(\xi_1)$轉換為$M_2(\xi_1)$之值，或反之亦然。可藉由自聯合分佈計算條件機率且將該等條件機率乘以由兩個原始模型計算之分佈來達成此舉。

例如，假定 $M_1(\xi)$={60%晴朗、30%多雲、10%下雨}且$M_2(\xi)$={30%晴、20%局部多雲、30%陰天、20%小雨、0%大雨}。明顯地，該等模型不僅進行不同的預測，而且使用不同的術語。往往在該等術語之間可存在一對一映射關係，但通常，更複雜之關係為必需的。以下矩陣圖示兩個不同組之術語意義如何重疊（關係R）：

| | 晴 | 局部多雲 | 陰天 | 小雨 | 大雨 |
|---|---|---|---|---|---|
| 晴朗 | 0.8 | 0.2 | 0 | 0 | 0 |
| 多雲 | 0 | 0.3 | 0.6 | 0.1 | 0 |
| 下雨 | 0 | 0 | 0.1 | 0.5 | 0.4 |

例如，當術語「多雲」係由第一模型使用時，該術語對應於「局部多雲」之可能性為30%、「陰天」之可能性為60%及小雨之可能性為10%。現在，模型M1之預測可由矩陣乘法轉換為國外術語，如$M_1'(\xi)$={48%晴、21%局部多雲、19%陰天、8%小雨、4%大雨}。現在，可使用$M_1'(\xi)$及$M_2'(\xi)$以常規方式計算距離，該$M_1'(\xi)$及$M_2(\xi)$之機率分佈具有相同集合。

若涉及數字預測，則關係R可使用區間而非個別數字。

與關係R共用之矩陣(3)可命名為通用模型。對熟習此項技術者而言，很清楚，通用模型不應該與其他已知資料探勘模型混合。基本上，通用模型表示資料集（亦即，資料點向量）及待比較之不同資料探勘模型之預測。可能需要關係R來匹配機率分佈之元素。

在下一步驟中，可建立維度為n乘(k+3)之矩陣(4)，其中行k+3保持矩陣(4)之位置k+1及k+2處之兩個值之間的距離 $d_{M_1,M_2}(\xi)$ ：

$$\begin{pmatrix} \xi_1 & M_1(\xi_1) & M_2(\xi_1) & d_{M_1,M_2}(\xi_1) \\ \xi_2 & M_1(\xi_2) & M_2(\xi_2) & d_{M_1,M_2}(\xi_2) \\ \cdots & & & \\ \xi_n & M_1(\xi_n) & M_2(\xi_n) & d_{M_1,M_2}(\xi_n) \end{pmatrix} \quad (4)$$

如上文已論述的，此距離量測 $d_{M_1,M_2}$ 可為簡單的分佈比較，但在更複雜的狀況下，此距離量測 $d_{M_1,M_2}$ 可涉及使用關係R在機率分佈之間進行映射。

亦很清楚，通用模型可使用不同數目之矩陣。作為一個實例，第一矩陣可含有輸入資料集X、第二矩陣可含有來自M1之預測、第三矩陣可含有來自M2之預測、第四矩陣可含有關係R且第五矩陣可含有距離d。

最後，可定義感興趣定義域之子集S。此舉可視作選擇不同列之矩陣(4)。根據上一行之值使用該等列之矩陣(4)

之分類處理可能為有益的。預定義範圍中之彼等資料點向量（例如，具有所有「行(k+3)值」中之最大20%之列）可定義具有類似差異特性（例如，兩個比較資料探勘模型之間的最高距離量測）之資料點集。視覺化技術經常由於大量維度而不能充分地描述此等個別資料點集。

因此，並非僅分類矩陣，可能較佳地，將預測資料探勘演算法應用於（預期）距離值中，以形成「距離模型」，該資料探勘演算法識別樣式。此種模型通常可含有區域描述，該等區域連接於輸入維度之空間中。隨後，感興趣區域可為具有模型間之最大平均距離之區域。然而，亦可使用其他距離量測基準，例如，具有最低距離量測或任何其他預定義值範圍之距離量測之區域。

使用涉及可應用之學習演算法的視覺化技術，可視覺化以此方式決定之一或更多個區域。具有兩個或三個獨立變數之視覺化可能較直接。涉及超過三個維度之視覺化可能需要額外映射。

在本申請案之情境下，遵循以下表達慣例：

資料探勘：術語資料探勘可表示分析處理，該分析處理設計為探索大量資料以搜尋變數之間的一致樣式及/或系統關係，隨後藉由將偵測樣式應用至新的資料子集來驗證研究結果。因此，資料探勘通常可表示為自資料提取樣式之處理。資料探勘日益成為一種重要工具，以將資料轉換為資訊。通常，資料探勘不僅用於各種概況

分析實踐中，諸如生產品質控制及最佳化、欺詐偵測、監控，而且用於行銷及科學發現中。

資料探勘可用以揭露資料樣式，但資料探勘往往僅執行於資料之取樣上。若取樣並未較好地表示資料之較大部分主體，則探勘處理可能無效。若該等樣式並未出現於正被「探勘」之取樣中，則資料探勘可能不會發現可出現於資料之較大部分主體中之樣式。不能發現樣式可導致不同生產設施間之一些爭端。因此，資料探勘並非極簡單的，但若收集到具充分代表性之資料取樣，則該資料探勘可適用。在特定資料集中發現特定樣式並非一定意謂在提取取樣之較大資料中之其他地方發現樣式。處理中之一個重要部分為驗證及確認其他資料取樣之樣式。

資料探勘模型：資料探勘模型可表示一群組以某種方式相關之資料點之抽象模型或數學模型。該模型可允許基於某一機率來預測待成為資料區域之一部分的資料點。可使用不同的資料探勘模型，例如，回歸模型、分群模型、分類模型、規則決定模型等。熟習此項技術者知道何時使用何種模型。有時，不同的模型可應用於相同的資料集。此舉可能為不易看出何種資料探勘模型可能為最合適的資料探勘模型之情況。尤其在彼等狀況下，具有一種方法及一種引擎來比較在不同區域資料集中之不同資料探勘模型之行為可能有利。

區域：區域可表示連接的數學空間，該連接的數學空間具有給定資料集之資料點向量之定義邊界。因此，完整的資料集定義數學資料點空間。邊界可視作超曲面。

通用模型：通用模型可表示類似於矩陣(3)之矩陣，該矩陣包含所有資料點向量及應用於個別資料點向量之資料探勘模型的結果。該通用模型進一步包含預測機率分佈之成員之間的關係。

距離量測：距離量測可表示與輸入值相同之資料點向量的不同資料探勘模型結果之間的差異。在最簡單方式中，距離量測可為機率值對之間的絕對差之總和。替代性差異量測包括庫貝克-李柏距離。距離量測亦可包括使用預測之間的關係來延伸自兩個模型返回之分佈，以便分佈經定義於相同值上（可為聯合分佈）。

在數字狀況下，機率分佈可具有不同形式，如高斯(Gaussian)分佈曲線、卜瓦松(Poisson)分佈曲線或其他分佈曲線。為易於比較且為建立距離量測，機率分佈曲線可接近於約為期望值之區域中，該等期望值由不同資料探勘模型產生，且此種近似值之平均值隨後可進行比較且可建立距離量測。以彼方式建立之距離函數可被稱作機率性距離量測。

上述用於執行資料探勘比較引擎及相關資料探勘比較引擎之方法可提供一些優點。

發明方法可允許資料探勘模型之比較，該等資料探勘模型與待比較之原始資料探勘模型之結構無關。該方法

亦可確保比較方法在統計學上保持相關且可能不會被資料探勘模型之初始結構偏置。此外，該方法可允許進行比較處理結果之彈性學習及視覺化。

因此，提供了一種用於比較資料探勘模型而不管資料探勘模型之類型的通用方法。除比較由兩個不同專家建立之兩個資料探勘模型之外，發明方法亦可允許若干其他應用。一個應用可為隨時間追蹤資料探勘模型。可按有規律的間隔（例如，每月）建立資料探勘模型。隨後，可能比較基於更新資料集建立之資料探勘模型如何與源自上個月之資料探勘模型相關。

另一個常見應用可為，針對整個資料集之不同子集建立資料探勘模型，例如，在生產類型環境中僅來自一周中之特定一天的生產資料。隨後，此應用可顯示生產品質如何及為何可在整周中有所不同，例如，在週一，一個產品之生產品質較低，而在一周中另一天其他產品之生產品質更低。

資料探勘技術用於許多科學研究領域及不同行業領域。一些實例為：

(a)已簡要提及之天氣預測模型。

(b)針對來自不同生產地點或在不同時間之類似產品之生產品質比較。

(c)備用零件可靠性比較：針對相關資料探勘模型之輸入值可包含：零件年齡、生產地點、產品用途、在下一定義時間段中之故障機率。

(d)油及天然氣探索：發現自然資源之機率取決於許多輸入變數。不同的資料探勘模型可傳送不同的結果。比較可揭露針對更佳探索結果之較高變化。

(e)信用行業：資料探勘模型可用以基於信用卡使用者之不同特性，諸如年齡、收入、家庭狀態、早期信用、年收入等來預測信用風險。

(f)購買行為分析及預測：資料探勘模型可用以基於歷史購買樣式及其他個人基準來預測顧客之購買行為。

(g)在針對噴發預測之火山研究中之震測法：此資料探勘模型比較之應用可增強在火山附近生活之人們之安全性。

針對所有彼等資料探勘應用領域及其他資料探勘應用領域，不同資料探勘模型之間的比較方法可能極有益，以產生更佳預測。

用於執行資料探勘模型比較引擎之方法的不同實施例在異於上述應用領域之應用領域中可能有利。

在該方法之一個實施例中，機率性距離量測為比較單變數機率性分佈之量測。此舉意謂僅一個變數被視為機率性分佈之相依變數(dependent variable)。此舉可能使距離量測之計算更可行且更易於解譯。在多變數機率性分佈中，不同資料探勘模型之間的差異解譯更複雜，且視覺化亦變得更複雜。在該方法之另一個實施例中，單變數機率性距離量測可能為L1度量。

在該方法之又一個實施例中，使用者可能關心距離量測之預定義值範圍。預定義值範圍可能為所有區域之距離量測之最低值或者為所有區域之距離量測之最高值。然而，可能可指定其他預定義距離量測範圍。通常，熟習此項技術者可選擇具有待比較或待視覺化之資料探勘模型間之最高差異（亦即，最大距離量測）的一或更多個子區域。

如上所述，資料探勘模型M1、M2可各自單獨地屬於包含分類、回歸及分群之群組中之一個資料探勘任務。此處，由於可比較不同類型之資料探勘模型，故本發明方法之優點可能又變得清晰了。熟習此項技術者理解：分類、回歸及分群為完全不同的資料探勘方法，該等資料探勘方法可能不能直接比較。

在該方法之另一個實施例中，該方法亦可包含統計測試，從而確保統計顯著距離量測。結果之統計顯著性為計算距離量測純粹靠機會出現之機率且為總體中之機率，取樣自該總體中提取，而無此種差異存在。使用較少技術術語，吾等可說，結果之統計顯著性可說明結果「真實」度（在「總體代表」之意義上而言）。更技術性地，p值(p-value)表示結果可靠性之減少指數。p值愈高，吾人愈不相信區域中之觀察距離量測可為距離量測之可靠指示。具體而言，p值表示可涉及接收距離量測之計算值為有效時之錯誤機率，亦即，p值表示可涉及接收距離量測之計算值為「總體代表」之錯誤機率。例如，p值為

0.05可指示發現於子區域中之距離量測為「僥倖猜對(fluke)」之機率為5%。在許多研究及工業應用領域中，習慣將p值為0.05作為「可接受邊線」誤差位準。

此外，電腦或電腦系統可包含如上所述之資料探勘比較引擎，及參閱用於執行資料探勘比較引擎之方法以比較量測資料點向量。

應注意，實施例可採取整體硬體實施例、整體軟體實施例或含有硬體元件及軟體元件兩者之實施例的形式。在較佳實施例中，在軟體中實施本發明，該軟體包括（但不限於）韌體、常駐軟體及微碼(microcode)。

在一個實施例中，提供了一種用於在資料處理系統中執行之資料處理程式，該資料處理程式包含軟體代碼部分，以使程式在資料處理系統上執行時執行如上所述之方法。資料處理系統可為電腦或電腦系統。

此外，實施例可採取電腦程式產品之形式，該電腦程式產品可自提供程式碼以供或結合電腦或任何指令執行系統來使用的電腦可使用或電腦可讀取媒體存取。為達成此描述之目的，電腦可使用或電腦可讀取媒體可為任何可包含儲存、通訊、傳播或輸送該程式以供或結合指令執行系統、設備或裝置使用之構件的設備。

媒體可為傳播媒體之電子系統、磁鐵系統、光學系統、電磁系統、紅外系統或半導體系統。電腦可讀取媒體之實例可包括半導體記憶體或固態記憶體、磁帶、隨機存取記憶體(RAM)、唯讀記憶體(ROM)、硬磁碟及光碟。光

碟之當前實例包括唯讀光碟記憶體(CD-ROM)、讀寫光碟(CD-R/W)、DVD及藍光光碟。

亦應注意，已參照不同標的描述了本發明之實施例。特定言之，已參照方法類請求項描述了一些實施例，同時參照設備類請求項描述了其他實施例。然而，除非以其他方式告知，否則熟習此項技術者將自上述及以下描述瞭解到，不僅屬於一類標的之特徵之任何組合，而且涉及不同標的之特徵之間的（尤其，方法類請求項之特徵及設備類請求項之特徵之間的）任何組合皆被認為待揭示於本文件中。

本發明之上文定義之態樣及其他態樣可自下文待描述之實施例之實例顯而易見，且參照實施例之實例予以闡釋，但本發明並非局限於該等態樣。

**示例性實施例之詳細描述**

在下文中，將給出圖式之詳細描述。所有圖式說明皆為示意性。首先，將描述用於執行資料探勘比較引擎之發明方法之方塊圖。隨後，將描述方法及資料探勘比較引擎之實施例。

**第1圖**圖示用於執行資料探勘比較引擎以比較量測資料點向量之發明方法100的方塊圖。方法可包含以下步驟：選擇性地提供第一資料集D1及第二資料集D2(102)，該等第一資料集D1及第二資料集D2包含量測資料點向量。第一資料探勘模型M1及第二資料探勘模型M2(102)為對D1、D2實施資料探勘任務T1、T2之結果。在一

些實施例中，在102中僅提供M1、M2即足夠。方塊104可圖示決定 M1及M2之預測值之間的關係R。方塊106可圖示決定輸入資料集X。在方塊108中，資料探勘模型M1、M2應用於輸入資料集X之記錄。方塊110可圖示使用距離函數d(x, M1, M2)來計算距離量測。最後，方塊112可表示決定及顯示至少一個感興趣區域，該區域具有滿足預定義條件之距離量測值。

使用此方法可解決以下任務：公司可生產產品P，該產品P可在若干個生產地點生產。部分生產可能不合格，且公司可在每一個生產地點及在若干個時間點形成資料探勘模型，以預測成品是否可能不合格。為進行預測，可進行數個量測，從而產生資料點向量，諸如室溫、濕度、一天之時間、一周之某天、生產帶之速率、由加熱單元所消耗之能量等。

使用預測模型，可偵測故障之潛在原因，且在更複雜原因之情況下，有可能當一或更多潛在重要情況發生時，在故障出現之前反應。

為偵測生產地點之間的差異，公司可能欲比較各個資料探勘模型且可決定可能在一個地點導致故障但在另一個地點並無負效應之情況。類似地，可偵測隨時間之變化，例如，一個資料探勘模型可在替換某個機器之前形成，且另一個資料探勘模型在替換某個機器之後形成。

**第2圖**可圖示發明方法之步驟之更詳細的方塊圖200。作為資料探勘模型M1 206及資料探勘模型M2 208之輸入

值$\vec{x}_i$ 204，可能有兩個選擇可用。輸入值可能為量測資料點向量，或者可使用現存資料探勘模型M1(206)或資料探勘模型M2(208)或者使用該兩者來產生新的資料點向量。記錄產生器或資料點向量產生器202可圖示彼狀況。

在下一步驟中，由M1之機率記錄器210使用輸入值或資料點向量。另外，M2亦由機率記錄器212執行機率記錄。第一機率分佈214及第二機率分佈216為機率記錄之結果。接著，機率性分佈比較器218計算距離量測 $d_{M_1,M_2}(\vec{x}_i)$ 220。

在另一個步驟222中，可檢查在感興趣區域中，預測中之差異是否基於距離量測在統計學上顯著。距離亦可與預定義值範圍相比且由預定義值範圍過濾。最後，可向使用者顯示符合預定義值範圍之區域(224)，以進行進一步動作及決定。

**第3圖**圖示用於執行資料探勘比較引擎之方法細節的不同透視圖300。方塊302可圖示記錄產生器產生具有資料點向量$\vec{x}_i$ 304之輸入資料集X之步驟，其中「i」可自1延續至n。每一個資料點向量304可具有k個維度，意謂對每一個向量而言k個個別值或資料元素可能相關。自資料集X及所提供之資料探勘模型M1、M2開始，機率記錄器308可產生組合記錄資料集，該組合記錄資料集由矩陣306表示。此組合記錄資料集可命名為通用模型。此矩陣可包含k+2行，亦即，原始資料點向量以及資料探勘模型M1及資料探勘模型M2應用至資料點向量之結果。該等應用位於矩陣306中，分別處於行k+1及行k+2中。通常，應

注意，通用模型亦包含M1之預測與M2之預測之間的關係R。

接著，可要求機率性分佈比較器312計算比較後之資料集310。另外，在矩陣310中，距離量測$d_{M,M_1(x)}$可包括於行k+3中。方塊314可表示機器學習演算法，以決定原始輸入資料集X之子區域，該等子區域可由所有資料點向量之空間建立。最後，步驟316可表示一或更多子區域之語意結合及視覺化，該一或更多子區域符合兩個比較後之資料探勘模型間之差異的預定義值範圍。通常，可顯示具有最大差異之一或更多子區域。然而，熟習此項技術者可理解，亦可選擇具有其他特性之子區域。

**第4圖**圖示根據本發明之一實施例之資料探勘比較引擎400的方塊圖。可提供提供單元402，該提供單元402適合於選擇性地提供量測資料點向量之第一資料集D1及第二資料集D2。提供單元亦可提供第一資料探勘模型M1及第二資料探勘模型M2。關係決定單元404可決定M1提供之第一預測值集與M2提供之第二預測值集之間的關係。此功能亦可為單元402之一部分。輸入資料集單元406可決定輸入資料集X，該輸入資料集X在一些狀況下可為如上所述之資料集D1、D2中之一者。另外，計算單元408可適合於基於第一資料探勘模型及第二資料探勘模型以及輸入資料集計算通用模型。第二計算單元可適合於基於第一資料探勘模型M1與第二資料探勘模型M2之間的通用模型，使用距離函數d(x, M1, M2)計算距離量測。第

一計算單元及第二計算單元可提供為單個計算單元。區域發現單元410可適合於基於距離量測值發現輸入資料集X內之區域。此外，決定及顯示單元412可適合於決定且顯示至少一個區域，該區域具有距離量測之預定義值範圍。

在第1圖及第4圖之情境下，應注意，可計算距離函數d(x, M1, M2)，其中x可為任何量測資料點向量。或者，亦可使用其他資料點向量，該等資料點向量可位於感興趣定義域內或在感興趣定義域附近，例如，使用第一資料探勘模型M1或第二資料探勘模型M2產生之彼等資料點向量。

**第5圖**圖示如何決定可顯示區域之流程圖。最初，清空工作之最佳區域集(502)。隨後，可使用完整的資料集作為第一最佳區域(504)。以下兩個步驟可重複直至可滿足終止基準為止(510)：a)選擇具有最低距離量測之區域，隨後將此區域分離為兩個或兩個以上子區域(506)；b)儲存元組(tuple)作為所得工作之最佳區域集之一部分，其中元組可包含區域，該等區域由該等區域之邊界及相應距離量測界定。

最後，彼等元組可自工作集移除，該等元組可能不能圖示任何統計相關性(512)。另外，在顯示之前，可移除具有低於預定義閾值之距離量測之彼等元組(514)。

事實上，本發明之實施例可在任何類型之適於儲存及/或執行程式碼的電腦上實施，而不管所使用之平臺。例

如，如**第6圖**所展示，電腦系統600可包括：一或更多處理器602，其中每個處理器具有一或更多核心；相關記憶體元件604；內部儲存裝置606（例如，硬碟、光學驅動機，諸如光碟驅動機或數位視訊光碟(DVD)驅動機、快閃記憶棒等）；以及大量其他元件及當今電腦典型之功能（未圖示）。記憶體元件604可包括：主記憶體，該主記憶體在實際執行程式碼期間使用；以及快取記憶體，該快取記憶體提供暫時儲存至少一些程式碼或資料以減少必須自外部大容量儲存器616取得代碼以供執行的次數。可藉助於匯流排系統618來將電腦600內之元件與相應配接器鏈接在一起。另外，資料探勘模型比較引擎400可附接至匯流排系統618。

電腦系統600亦可包括輸入構件，諸如鍵盤608、滑鼠610或麥克風（未圖示）。此外，電腦600可包括輸出構件，諸如監測器612（例如，液晶顯示器(LCD)）、電漿顯示器、發光二極體顯示器(LED)或陰極射線管(CRT)監測器）。電腦系統600可連接至網路（例如，區域網路(LAN)、廣域網路(WAN)），諸如網際網路或任何其他類似類型之網路，包括經由網路介面連接614之無線網路。此舉可允許耦接至其他電腦系統。熟習此項技術者將理解，存在許多不同類型之電腦系統，且上述輸入構件及輸出構件可採用其他形式。一般而言，電腦系統600可至少包括為實踐本發明之實施例所必需之最小處理、輸入及/或輸出構件。

此外，熟習此項技術者將理解，上述電腦系統600之一或更多元件可位於遠端位置且經由網路連接至其他元件。此外，可在具有複數個節點的分佈系統上實施本發明之實施例，其中本發明之每一個部分可位於分散式系統內之不同節點上。在本發明之一個實施例中，節點對應於電腦系統。或者，節點可對應於具有相關實體記憶體之處理器。或者，節點可對應於具有共享記憶體及/或資源的處理器或智能手機。

此外，執行本發明之實施例之軟體指令可儲存於電腦可讀取媒體上，諸如光碟(CD)、磁片、磁帶或任何其他電腦可讀取儲存裝置。

儘管已相對於有限數目之實施例描述了本發明，但受益於本揭示案之熟習此項技術者將理解，可設計不脫離如本文中所揭示之本發明之範疇的其他實施例。因此，本發明之範疇應僅由所附申請專利範圍限制。

亦應注意，術語「包含」並不排除其他元素或步驟，且「一」或「一個」並不排除複數。同樣，可組合結合不同實施例描述之元素。亦應注意，不應將申請專利範圍中之元件符號理解為限制元素。

## 【圖式簡單說明】

現將結合以下圖式僅以舉例之方式描述本發明之較佳實施例：

第1圖圖示用於執行資料探勘比較引擎之發明方法之方塊圖。

第2圖圖示發明方法之步驟之更詳細的方塊圖。

第3圖圖示用於執行資料探勘比較引擎之方法細節之不同透視圖。

第4圖圖示根據本發明之一實施例之資料探勘比較引擎的方塊圖。

第5圖圖示如何決定可顯示區域之流程圖。

第6圖圖示根據本發明之一實施例之包含資料探勘比較引擎的電腦系統。

【主要元件符號說明】

| | | | |
|---|---|---|---|
| 100 | 方法 | 102 | 方塊 |
| 104 | 方塊 | 106 | 方塊 |
| 108 | 方塊 | 110 | 方塊 |
| 112 | 方塊 | 200 | 方塊圖 |
| 202 | 記錄產生器/資料點向量產生器 | 204 | 輸入值 |
| 206 | 資料探勘模型 M1 | 208 | 資料探勘模型 M2 |
| 210 | 機率記錄器 | 212 | 機率記錄器 |
| 214 | 第一機率分佈 | 216 | 第二機率分佈 |
| 218 | 機率性分佈比較器 | 220 | 距離量測 |
| 222 | 步驟 | 224 | 步驟 |
| 300 | 透視圖 | 302 | 方塊 |
| 304 | 資料點向量 | 306 | 矩陣 |
| 308 | 機率記錄器 | 310 | 資料集/矩陣 |
| 312 | 機率性分佈比較器 | 314 | 方塊 |
| 316 | 步驟 | 400 | 資料探勘比較引擎 |
| 402 | 提供單元 | 404 | 關係決定單元 |
| 406 | 輸入資料集單元 | 408 | 計算單元 |
| 410 | 區域發現單元 | 412 | 決定及顯示單元 |
| 502 | 步驟 | 504 | 步驟 |
| 506 | 步驟 | 508 | 步驟 |
| 510 | 步驟 | 512 | 步驟 |
| 514 | 步驟 | 600 | 電腦系統 |
| 602 | 處理器 | 604 | 記憶體元件 |
| 606 | 內部儲存裝置 | 608 | 鍵盤 |
| 610 | 滑鼠 | 612 | 監測器 |
| 614 | 網路介面連接 | 616 | 外部大容量儲存器 |
| 618 | 匯流排系統 | | |

201224812

# 發明專利說明書

（本說明書格式、順序，請勿任意更動，※記號部分請勿填寫；惟已有申請案號者請填寫）

※ 申請案號：100133243

※ 申請日期：100年9月15日　　※IPC 分類：

G06F17/30 2006.01

## 一、發明名稱：（中文/英文）

機率性資料探勘模型比較引擎/PROBABILISTIC DATA MINING MODEL COMPARISON ENGINE

## 二、中文發明摘要：

本發明揭示了一種用於比較第一資料探勘模型與第二資料探勘模型之比較引擎。第一資料探勘模型M1表示對第一資料集D1執行第一資料探勘任務之結果，且該第一資料探勘模型M1提供第一預測值集。第二資料探勘模型M2表示對第二資料集D2執行第二資料探勘任務之結果，且該第二資料探勘模型M2提供第二預測值集。決定該等預測值集之間的關係R。針對輸入資料集之至少第一記錄，第一機率分佈及第二機率分佈係基於應用於第一記錄之第一資料探勘模型及第二資料探勘模型而形成，該等機率分佈將機率與該等預測值集相關聯。使用第一機率分佈及第二機率分佈及其關係來計算該第一記錄之距離量測d。至少一個感興趣區域(region of interest)係基於該距離量測d來決定。

## 三、英文發明摘要：

Comparison engine for comparing a first data mining model and a second data mining model is disclosed. A first data mining model Ml represents results of a first data mining task on a first data set Dl and provides a set of first prediction values. A second data mining model M2

represents results of a second data mining task on a second data set D2 and provides a set of second prediction values. A relation R is determined between said sets of prediction values. For at least a first record of an input data set, a first and second probability distribution is created based on the first and second data mining models applied to the first record, said probability distributions associating probabilities with said sets of prediction values. A distance measure d is calculated for said first record using the first and second probability distributions and the relation. At least one region of interest is determined based on said distance measured.

## 七、申請專利範圍：

1. 一種用於執行一資料探勘模型比較引擎以供比較一第一資料探勘模型與一第二資料探勘模型之方法，該方法包含以下步驟：

提供一第一資料探勘模型M1，該第一資料探勘模型M1表示對該第一資料集D1執行一第一資料探勘任務之結果，該第一資料探勘模型提供一第一預測值集；

提供一第二資料探勘模型M2，該第二資料探勘模型M2表示對該第二資料集D2執行一第二資料探勘任務之結果，該第二資料探勘模型提供一第二預測值集；

提供該第一預測值集與該第二預測值集之間的一關係R；

決定一輸入資料集X；

對該決定輸入資料集之至少一第一記錄實施以下步驟：

基於應用於該第一記錄之該第一資料探勘模型來形成一第一機率分佈，該第一機率分佈將機率與該第一預測值集相關聯；

基於應用於該第一記錄之該第二資料探勘模型來形成一第二機率分佈，該第二機率分佈將機率與該第二預測值集相關聯；

使用該等第一機率分佈及第二機率分佈以及該關係來計算該第一記錄之一距離量測d；

基於該距離量測d決定至少一個感興趣區域，該距離量測d係針對該輸入資料集之記錄計算。

2. 如請求項1所述之方法，其中該第一資料探勘模型及該第二資料探勘模型結構不同。

3. 如請求項1或請求項2所述之方法，其中該第一資料探勘任務不同於該第二資料探勘任務。

4. 如請求項1或請求項2所述之方法，其中該第一資料探勘任務及該第二資料探勘任務各自單獨地為以下中之一者：分類、回歸、分群。

5. 如請求項1或請求項2所述之方法，其中使用一機率性距離量測來計算該距離量測。

6. 如請求項5所述之方法，其中該距離量測為一L1度量。

7. 如請求項6所述之方法，該方法包含以下步驟：使至少一個感興趣區域對一使用者視覺化。

8. 如請求項7所述之方法，其中該至少一個感興趣區域由該輸入資料集之記錄組成，該輸入資料集之記錄具有滿足一預定義基準之一距離量測。

9. 如請求項8所述之方法，其中該方法亦包含一統計測試，從而確保基於該距離量測之一統計學上的顯著差異。

10. 一種資料探勘比較引擎(400)，該資料探勘比較引擎(400)包含：

一提供單元(402)，該提供單元(402)適合於提供一第一資料探勘模型M1及一第二資料探勘模型M2，該第一資料探勘模型M1提供一第一預測值集，該第二資料探勘模型M2提供一第二預測值集，該第一資料探勘模型M1表示對一第一資料集D1執行一第一資料探勘任務之結果，而該第二資料探勘模型M2表示對一第二資料集D2執行一第二資料探勘任務之結果，其中該提供單元亦適合於提供該第一預測值集與該第二預測值集之間的一關係R；

一輸入資料集單元(406)，該輸入資料集單元(406)適合於決定一輸入資料集X；一計算單元(408)，該計算單元(408)適合於針對該輸入資料集之每一個記錄：

基於該第一資料探勘模型計算一第一機率分佈，該第一機率分佈將機率與該第一預測值集相關聯；

基於該第二資料探勘模型計算一第二機率分佈，該第二機率分佈將機率與該第二預測值集相關聯；

使用該等第一機率分佈及第二機率分佈以及該關係來計算該等記錄之一距離量測d；一區域發現單元(410)，該區域發現單元(410)適合於基於該距離量測d決定至少

一個感興趣區域，該距離量測d係針對該輸入資料集之記錄計算。

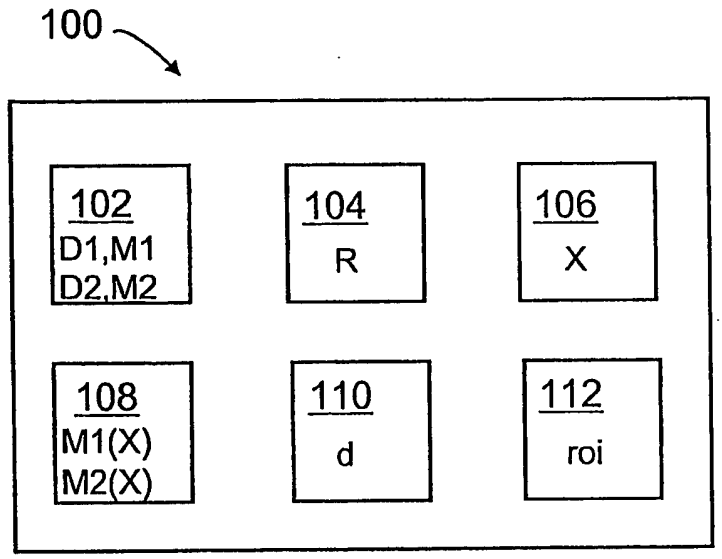
11. 如請求項10所述之資料探勘比較引擎，該資料探勘比較引擎進一步包含一決定及顯示單元(412)，該決定及顯示單元(412)適合於基於該距離量測之一預定義值範圍來顯示至少一個感興趣區域。

12. 一種電腦系統(600)，該電腦系統(600)包含如請求項10至請求項11中任一項所述之資料探勘比較引擎(400)。

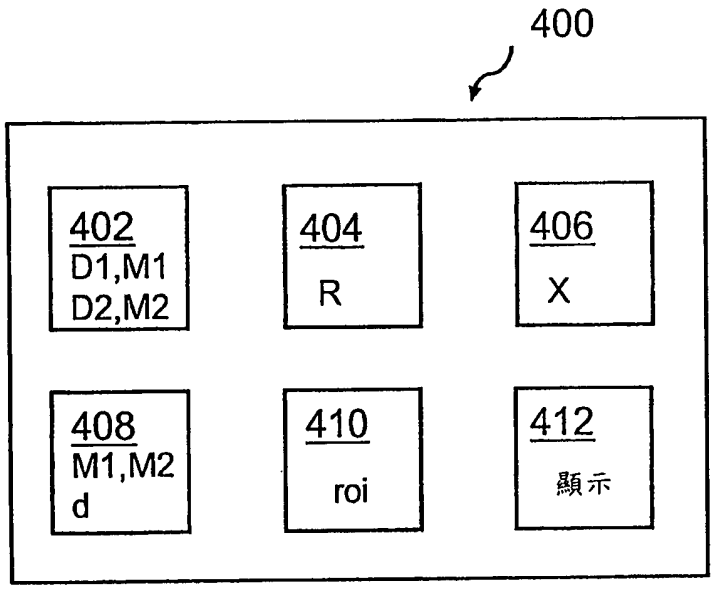
13. 一種用於在一資料處理系統(600)中執行之資料處理程式，該資料處理程式包含軟體代碼部分，以在該程式在一資料處理系統(600)上執行時執行一種如前述請求項1至請求項9中任一項所述之方法。

14. 一種儲存於一電腦可使用媒體上的電腦程式產品，該電腦程式產品包含電腦可讀取程式構件，以使一電腦(600)在該程式於該電腦(600)上執行時執行一種如請求項1至請求項9中任一項所述之方法。

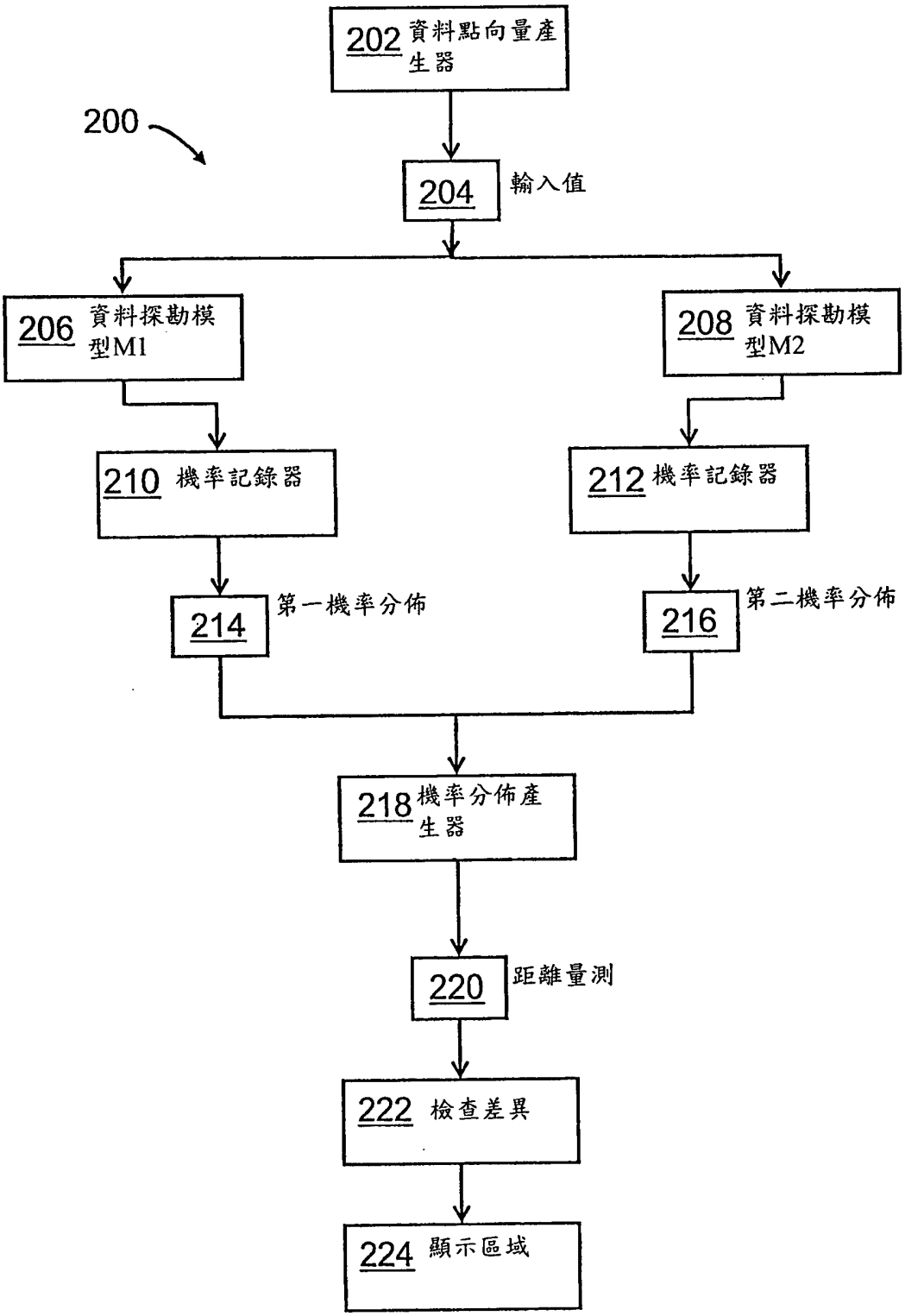
八、圖式：


100
102
D1,M1
D2,M2
104
R
106
X
108
M1(X)
M2(X)
110
d
112
roi

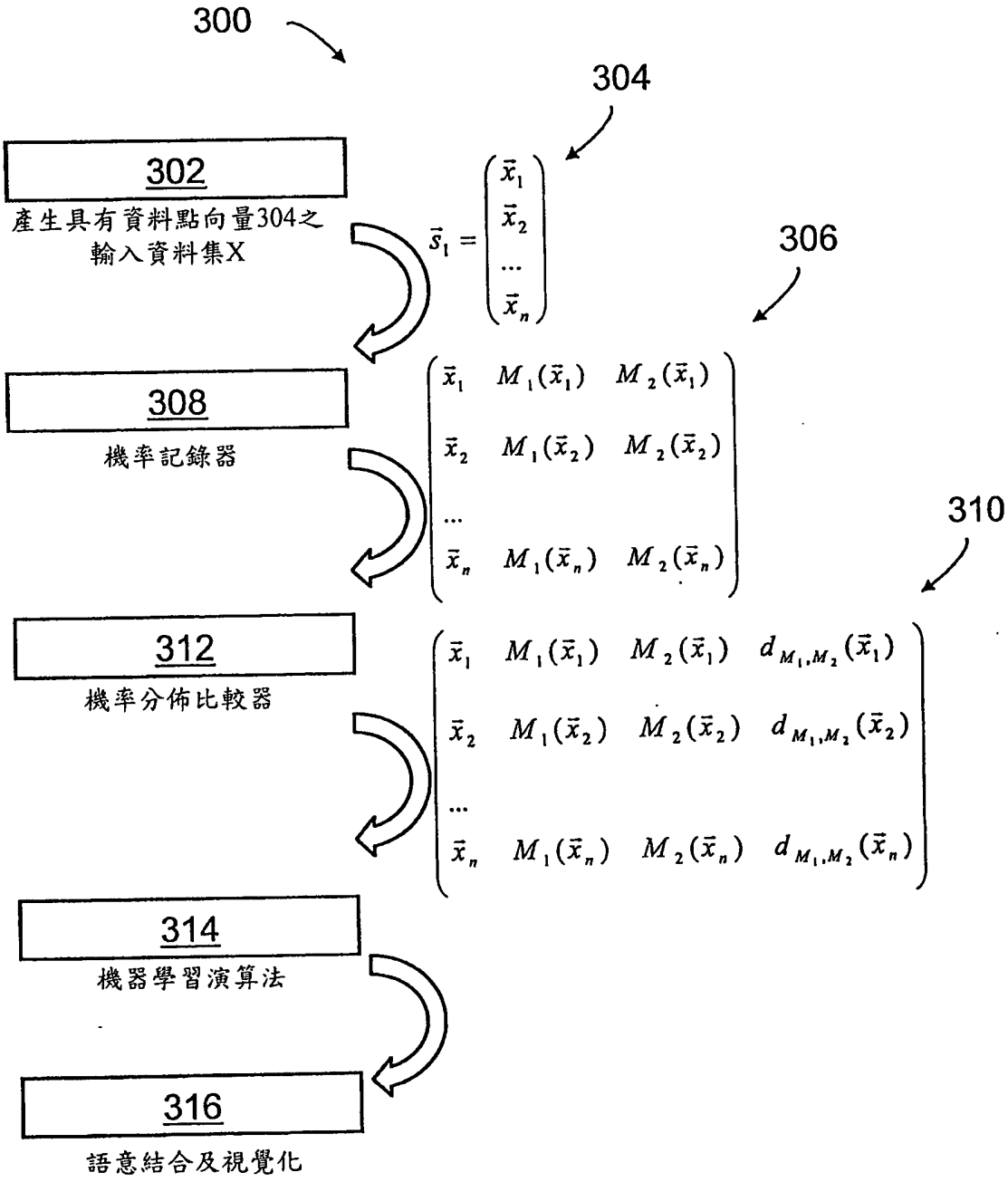
第1圖


400
402
D1,M1
D2,M2
404
R
406
X
408
M1,M2
d
410
roi
412
顯示

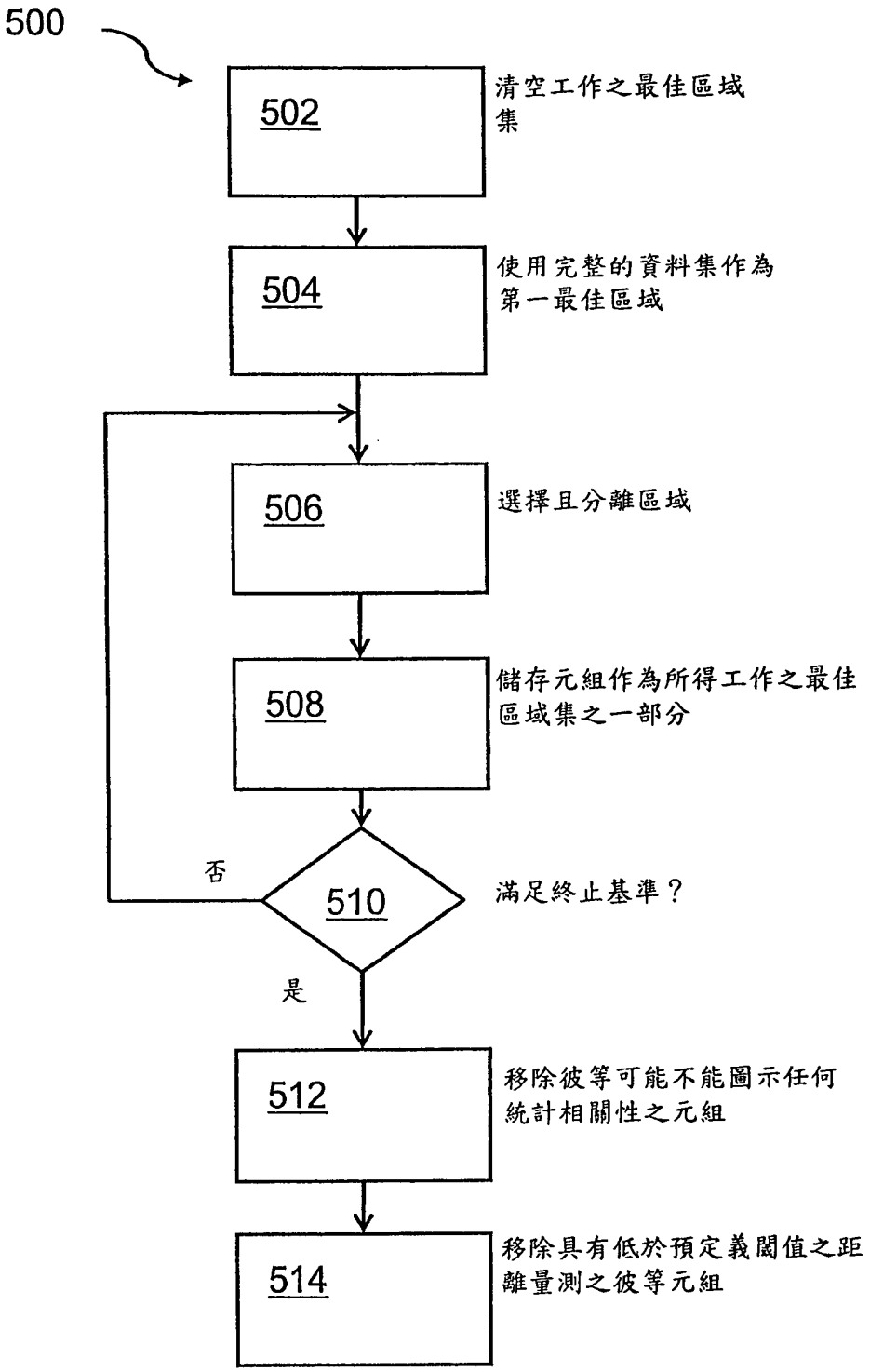
第4圖


200
202 資料點向量產生器
204 輸入值
206 資料探勘模型M1
208 資料探勘模型M2
210 機率記錄器
212 機率記錄器
214 第一機率分佈
216 第二機率分佈
218 機率分佈產生器
220 距離量測
222 檢查差異
224 顯示區域

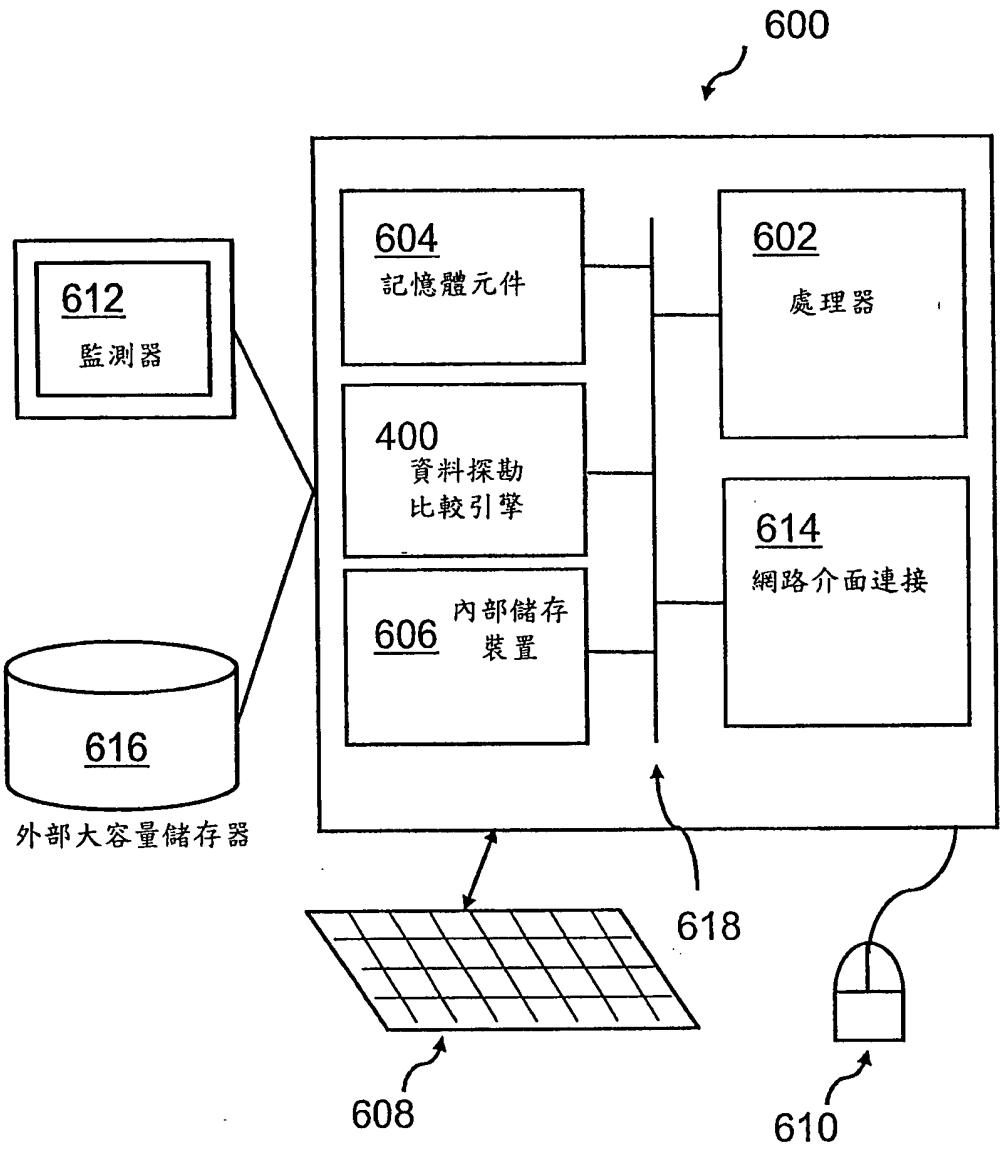
第2圖


300
304
302
產生具有資料點向量304之
輸入資料集X
$\vec{s}_1 = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \cdots \\ \vec{x}_n \end{pmatrix}$
306
308
機率記錄器
$\begin{pmatrix} \vec{x}_1 & M_1(\vec{x}_1) & M_2(\vec{x}_1) \\ \vec{x}_2 & M_1(\vec{x}_2) & M_2(\vec{x}_2) \\ \cdots & & \\ \vec{x}_n & M_1(\vec{x}_n) & M_2(\vec{x}_n) \end{pmatrix}$
310
312
機率分佈比較器
$\begin{pmatrix} \vec{x}_1 & M_1(\vec{x}_1) & M_2(\vec{x}_1) & d_{M_1,M_2}(\vec{x}_1) \\ \vec{x}_2 & M_1(\vec{x}_2) & M_2(\vec{x}_2) & d_{M_1,M_2}(\vec{x}_2) \\ \cdots & & & \\ \vec{x}_n & M_1(\vec{x}_n) & M_2(\vec{x}_n) & d_{M_1,M_2}(\vec{x}_n) \end{pmatrix}$
314
機器學習演算法
316
語意結合及視覺化

第3圖


500
502
清空工作之最佳區域集
504
使用完整的資料集作為第一最佳區域
506
選擇且分離區域
508
儲存元組作為所得工作之最佳區域集之一部分
510
滿足終止基準？
否
是
512
移除彼等可能不能圖示任何統計相關性之元組
514
移除具有低於預定義閾值之距離量測之彼等元組

第5圖


600
604
記憶體元件
602
處理器
612
監測器
400
資料探勘
比較引擎
614
網路介面連接
606
內部儲存
裝置
616
外部大容量儲存器
618
608
610

第6圖

四、指定代表圖：

(一)本案指定代表圖為：第（ 3 ）圖。

(二)本代表圖之元件符號簡單說明:

300 透視圖
302 方塊
304 資料點向量
306 矩陣
308 機率記錄器
310 資料集/矩陣
312 機率性分佈比較器
314 方塊
316 步驟

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

無