MC-Transaction on Biotechnology, 2014, Vol. 6, No. 1, e4

新型灰關聯聚類法應用於心臟疾病之分類

蔡坤龍[1]、李嘉陵[2,*]
[1]銘傳大學資訊學院電子工程學系(中華民國 台灣 桃園)
[2]銘傳大學健康科技學院生物醫學工程學系(中華民國 台灣 桃園)

中文摘要

為提升波多大學心臟病資料庫之分類效能，本研究進而提出新型式的灰關聯聚類法。本研究修改傳統型灰關聯聚類法的處理程序，以四項主要工作流程：(1)建立灰關聯空間，包含灰關聯因子空間及改善型灰關聯測度，本研究將選取六項數值屬性及四項類別屬性等與心臟病高相關的因子，經資料處理後，構成具有可比性的比較向量。(2)產生參考向量，本研究以資料庫中有病與無病類各自的平均值作為各類固定式的參考向量。(3)計算每筆紀錄的兩類灰關聯值。(4)比較兩類灰關聯值而進行分類。實驗結果顯示，分類正確率達 87%、有病捕捉率 75%、有病 F-Measure 等於 0.8371 和分類成本等於 31，上述指標皆優於傳統型，說明本分類法的有效性及進步性。

關鍵字：心臟病、灰關聯、灰聚類、分類指標
通訊作者: 李嘉陵[jllee@mail.mcu.edu.tw]
收稿：2014-8-25 修改：2014-10-21 接受：2014-11-4 線上發表：2014-11-23

緒　論

許多人喜歡精緻美食，但往往容易造成攝取過多的油、鹽和低密度膽固醇等不利於健康的因素，再加上生活規律錯亂、缺乏運動、經常處於各種壓力的情境之中，這些因素極易導致高血壓和心臟病。根據衛生署統計，在台灣近年來高血壓和心臟病一直高居國人十大死因之前幾名，而且每年約有4000人因急性心肌梗塞症而死亡，以死亡率達15%，每年應該至少有2萬6千人罹患急性心肌梗塞症，因此推估以心血管所衍生的疾病將成為我國重要的醫療問題。尤其在高齡化社會，此問題益顯得重要。美國每年約有600萬人曾因胸痛而至急診室求診，其中僅約有1/12的病人是由心電圖的檢查而診斷出急性心肌梗塞[1]，由此可知僅由心電圖的檢查仍有不足之處。一般的醫院中均含有與心血管疾病相關的危險因子的檢測項目，

例如收縮壓、舒張壓、膽固醇含量等，若能設計一平台從中做資料分析與評估，並鑑別出疑似有病的人選，再請醫師進一步以其他較為可靠的方式做正確的診斷與治療，期望能降低國人心臟病的發生率與死亡率，亦不失為一種便宜可行的方案。

在資料分析方面，尤其是需要測定兩種數量級序列之間的關聯性，傳統上是以數理統計的方法，例如相關分析或迴歸分析等[2]，然而其通常要求有大量的數據，而且數據需符合典型的分佈，例如常態分佈，明顯地對於某些少量數據以及多因素的醫學資料就無法滿足統計上的要求，因此，灰色系統(Grey System)理論可望於醫學領域成為另一種選擇。灰色理論於 1979 年提出[3-5]，主要是針對系統的部分訊息已知的小量樣本，不需數據的常態分佈，而且允許含有不確定量的狀況下，提供建模、分析、評估等處理系統的方法，以便了解系統的特徵和行為。灰色理論的方法正好可以彌補統計上的不足，對於小量樣本的醫學資料仍可期望得到可靠的結果。灰色聚類及灰色關聯是灰色理論中兩種重要的方法，灰色聚類是將評估對象對於評估項目的數據樣本進行各個灰類的映射及運算處理，以便了解評估對象的所屬類別，其中各個類別是固定以某些特徵的白化函數表示[4-5]，例如要評估不同地區的水質，需先收集各地區水的硬度、硝酸鹽、硫酸鹽等可表示水質的評估項目，再訂定高標、中標、低標等三種類別的白化函數，即可用灰色聚類法做水質的歸類研究[6-7]。灰色關聯分析是通過灰關聯因子空間的建立及灰關聯測度的運算來序化參考向量與比較向量之間的關聯程度[4-5, 8]，例如利用肝功能檢查項目做分級評估的研究[9]，需先篩選影響肝功能的因子，並選擇各等級的標準參考向量，再計算每位病患的肝功能檢查項目的序列資料與各等級的參考向量之間的關聯程度，而給予病患適當的分級與建議。另外，應用上述灰色關聯的序化結果以及設定適當的閥值，並採用非監督式演算法可將多筆資料分為幾個類別，以此形成灰色關聯聚類法[8, 10-11]。本論文將融合上述兩種聚類法而形成新型式的分類方式，其基本上是修改傳統灰色關聯聚類法，使得不需更新群聚中心、也不用排序及閥值，而是將灰色聚類法中表示類別的白化函數轉換為固定的群聚中心或參考向量，僅以計算與比較灰色關聯值的方式而進行分類，而此群聚中心可由醫學資料庫中依常模數據事先訂定有病或無病之特徵值而產生這兩類別的參考向量，並以新型式的灰關聯測度為依據而發展出符合心臟病資料庫的分類系統。

材料與方法

資料庫簡介：

本論文採用葡萄牙波多大學(University of Porto)人工智慧及電腦科學實驗室所提供的心臟病資料庫[12]，共有 270 筆病患資料，其中可區分為患有心臟病 120 位及無心臟病 150 位等兩大類別，而且每筆紀錄包含基本項目，例如年齡、性別，以及醫學檢驗項目，例如收縮壓、血液中膽固醇含量等 13 個項目，包含 6 個數值屬性及 7 個類別屬性，本論文將選取的屬性名稱及其屬性描述如表一及表二所示：

表一、數值屬性

序號	屬性名稱	屬性描述(代號)
1	Age	年齡(AG)
2	Resting_blood_pressure	收縮血壓(BP)
3	Serum_cholestoral	血液中膽固醇含量(SC)
4	Maximum_heart_rate_achieved	最大心跳速率(MHR)
5	Oldpaek	運動心電圖 ST 下降程度(OP)
6	Number_of_major_vessels	冠狀動脈阻塞數(NMV)

表二、類別屬性

序號	屬性名稱	屬性值	屬性描述(代號)
1	Chest Pain	A11 A12 A13 A14	胸痛類型(CP) A11 主體脈剝離引起主動脈瘤急性發作時，所引發的疼痛 A12 心包膜炎引起的胸痛，深呼吸時會使疼痛加重

			A13 官能性神經症所引發的疼痛 A14 心絞痛
2	Resting_electro_ cardiography	A21 A22 A23	休息時心電圖(ECG) A21 正常 A22 輕微異常 A23 異常
3	Exercise_induced_angina	A31 A32	運動是否引發心絞痛(EIA) A31 否 A32 是
4	ST Slope	A41 A42 A43	運動心電圖 ST 間最高斜率(STS) A41 完全正常 A42 連續兩個以上的 ST 段上升超過 1 mm 以上 A43 連續兩個以上的 ST 段下降超過 1 mm 以上

基本數學模型：

灰關聯空間是本論文所需要的基本數學模型，包含下列兩項要點：第一，為設計心臟病的分類系統而收集與其相關的因子集如表一、表二所示，但須具備下列五項特性，才能建立灰關聯因子空間$\{X\}_{[4\text{-}5,\,8]}$：(1)存在性、(2)獨立性、(3)影響性、

(4)可數性、(5)可比性，其中存在性和獨立性是構成數學空間最基本的條件，影響性是指所選因子對於探討本主題具有關鍵性的影響，可數性是指因子數目是有限的，且均可測量而可用數值序列表示，可比性是指不同的因子序列之間可以比較；通常若滿足下列三個條件即可成為可比性的序列：(1)去除測度單位而成無因次性(non-dimension)、(2)數值具同等級性(order)、(3)因子的描述狀態皆為同方向性或同極性(polarization)。由於不同數值屬性之因子常有不同的單位與數量等級，例如膽固醇含量 250 mg/dl 及心跳 70 bpm，兩者單位不同且數值等級有明顯差異，無法比較，然而，實務上我們僅需依每一數值屬性之因子的性質是望大值或望小值而給予適當的極性處理，使得所有因子皆朝向同一目標屬性是有病或無病，即可讓所有數值樣本同時滿足可比性的三個條件了。第二，選取適當的灰關聯測度函數 $\gamma(x_i, x_j)$，以此衡量參考向量和比較向量之間的關聯程度，並進行後續的分類，其中灰關聯測度函數$\gamma(x_i, x_j)$須為正實數，並須滿足下列四個條件：
(1)規範性：當$\gamma(x_i, x_j)$=1 時，兩序列x_i, x_j為完全相關；當$\gamma(x_i, x_j)$=0 時，兩序列x_i, x_j為完全不相關，亦即$0 \leq \gamma(x_i, x_j) \leq 1, \forall i, j$；
(2)對稱性：當序列只有兩組時，$\gamma(x_1, x_2) = \gamma(x_2, x_1)$；
(3)整體性：當序列超過三組時，通常會滿足$\gamma(x_i, x_j) \neq \gamma(x_j, x_i)$；
(4)接近性：序列差$|x_i - x_j|$為$\gamma(x_i, x_j)$主控項，亦即，當$|x_i - x_j|$越小，$\gamma(x_i, x_j)$越大。綜上所述，我們結合灰關聯因子空間{ X }及全體灰關聯映射

$\Gamma = \{\gamma(x_i, x_j), \forall i, j\}$即可構成灰關聯空間{X, Γ}。

在因子集選取方面，本論文參考心臟病的危險因子以及對所有屬性的灰統計結果[13]，在上述資料庫中選取每筆紀錄的六項數值屬性及四項類別屬性等項目構成一個具有特徵性的比較序列，其中數值屬性如表一所示，依序為「年齡(AG)」、「收縮血壓(BP)」、「血液中膽固醇含量(SC)」、「最大心跳率(MHR)」、「運動心電圖 ST 下降程度(OP)」以及「冠狀動脈阻塞數(NMV)」等六項；類別屬性依序為「胸痛類型(CP)」、「休息時心電圖(ECG)」、「運動是否引發心絞痛(EIA)」 以及「運動心電圖 ST 間最高斜率(STS)」等四項。此因子集已滿足存在性、獨立性、影響性、及可數性等，但仍須將不同數值屬性的資料再經相對應的極性處理如下：

望大值目標: $x_i^*(k) = \frac{x_i(k) - \min_k[x_i(k)]}{\max_k[x_i(k)] - \min_k[x_i(k)]}$ (1)

望小值目標: $x_i^*(k) = \frac{\max_k[x_i(k)] - x_i(k)}{\max_k[x_i(k)] - \min_k[x_i(k)]}$ (2)

拒中值目標: $x_i^*(k) = \frac{|x_i(k) - x_o|}{\max_k[x_i(k)] - x_o}$, (3)

其中x_o為目標值，並須滿足$^{max}_{k}[x_i(k)] \geq x_o \geq {}^{min}_{k}[x_i(k)]$。本論文在「收縮血壓(BP)」屬性是用拒中值目標處理，「最大心跳率(MHR)」屬性是用望小值目標處理，其他數值屬性則是用望大值目標處理。由上式(1)-(3)可知，經極性處理之後，數值序列已滿足可比性，因為所有數值序列皆正規化到零與一之間(無因次性、同等級性)，而且每一數值屬性若其值較大，表示有病之傾向(同極性)。另一方面，對於如表二所示之類別屬性須以直觀的尺度加以量化，量化原則同樣是有病之傾向者給予較大數值，本論文給予的尺度值如下：「胸痛類型(CP)」屬性有四種類型，故依序給予[0.1, 0.3, 0.6, 0.9]，「休息時心電圖(ECG)」屬性及「運動心電圖 ST 間最高斜率(STS)」屬性有三種類型，依序給予[0.25, 0.5, 0.75]，「運動是否引發心絞痛(EIA)」屬性有兩種類型，分別給予[0, 1]。另外，雖然性別及糖尿病也是心血管疾病之危險因子，但為限制其數量，我們捨去「性別」與「空腹時血糖值是否大於 120」兩項因子。在選定灰關聯公式方面，本論文為提升分類之成效而提出改善型灰關聯公式，並與傳統型比較，這兩種公式分別敘述如下：

傳統型灰關聯係數公式[4-5, 8]：

$$\gamma_k\left(x_{oi}(\mathrm{k}), x_j(\mathrm{k})\right) = (\delta_{min} + \xi\,\delta_{max}) / (\delta_j(\mathrm{k}) + \xi\,\delta_{max}) \quad (4a)$$

$$\delta_j(\mathrm{k}) = \left|x_{oi}(\mathrm{k}) - x_j(\mathrm{k})\right| \quad (4b)$$

$$\delta_{min} = min_{\forall j} min_{\forall k}\, \delta_j(\mathrm{k}) \quad (4c)$$

$$\delta_{max} = max_{\forall j} max_{\forall k}\, \delta_j(\mathrm{k}) \quad (4d)$$

i=1, 2，j=1, 2, 3,…, 54，k=1, 2, 3,…, 10

其中$x_{oi}(\mathrm{k})$代表無病類或有病類的參考向量的第 k 個元素，$x_j(\mathrm{k})$代表無病類或有病類的比較向量的第 k 個元素，$\delta_j(\mathrm{k})$為$x_{oi}(\mathrm{k})$及$x_j(\mathrm{k})$之間差的絕對值，$\xi \in (0, 1]$為辨識係數。因此，任一比較序列x_j與參考序列x_{oi}之間的灰關聯值定義為各元素的灰關聯係數的加權平均如下所示：

$$\gamma(x_{oi}, x_j) = \sum_{k=1}^{m} \alpha_k \gamma_k\left(x_{oi}(\mathrm{k}), x_j(\mathrm{k})\right) \quad (5)$$

其中α_k為權重值，且須滿足 $\sum_{k=1}^{m} \alpha_k = 1$。

改善型灰關聯係數公式：

$$\gamma_k\left(x_{oi}(k),x_j(k)\right)=\exp\left(-\frac{\delta_j(k)^2}{2\xi^2}\right), \quad (6)$$

i=1, 2，j=1, 2, 3,…, 54，k=1, 2, 3,…, 10

其中 exp()為指數函數，$\xi\in(0,1]$為辯識係數，$\delta_j(k)$之定義與(4)式相同，由此定義的灰關聯測度函數$\gamma_k\left(x_{oi}(k),x_j(k)\right)\in(0,1]$，滿足上述灰關聯測度的規範性、對稱性、整體性、和接近性等四個條件，因此，本論文提出的改善型灰關聯係數可以做為灰關聯的量測值。同樣地，任一比較序列x_j與參考序列x_{oi}之間的灰關聯值亦如(5)式的加權平均。在產生參考向量方面，我們在灰關聯空間內以資料庫中有病與無病類各自的平均值作為各類固定式的參考向量。然後，即可將測試樣本中的每筆紀錄，依據選定的公式(4)-(5)或(5)-(6)分別計算和有病及無病類的參考向量的灰關聯值，並取其較大者為分類的依據，即可將多筆且有多項屬性的序列分類到各類之中。

另一方面，為評估分類之效能，本研究首先須將分類結果建立如表三之混合矩陣，

表三、分類結果之混合矩陣

	預測無病	預測有病
實際無病	AA	AP
實際有病	PA	PP

其中AA表示實際無病而且預測亦是無病，PP表示實際有病而且預測亦是有病，這兩項皆表示得到正確的分類；AP表示實際無病但預測卻是有病，PA表示實際有病但預測卻是無病，亦即AP與PA兩項皆是得到錯誤的分類。其次，檢驗下列五項分類效能指標：

（1）分類正確率 CA=(AA+PP)/(AA+AP+PA+PP)

（2）有病精確率 P=PP/(AP+PP)

（3）有病捕捉率 R=PP/(PA+PP)

（4）有病 F-Measure=2*P*R/(P+R)

（5）分類錯誤成本 Cost=AP+PA*5

其中公式之代號如上表三所示，這些指標的意義說明如下：

1. 分類正確率(CA)：表示被正確分類的樣本數佔全部樣本數之比率，此指標越大表示越好。
2. 有病精確率(P)：表示被預測為有病的樣本數中，真正有病的比率，此指標越大越好。
3. 有病捕捉率(R)：表示實際有病的樣本數中，被正確分類的比率，此指標越大越好。有病捕捉率相對於有病精確率是較具重要性的指標。
4. 有病 F-Measure：綜合了有病精確率(P)和有病捕捉率(R)而合成一項重要的衡量指標，任一項指標較低，皆會使 F-Measure 較低，亦即唯有同時提高有病精確率和有病捕捉率，才能得到較高的 F-Measure，此指標越大越好。
5. 分類錯誤成本(Cost)：表示每次預測錯誤時所要付出的成本，其中以實際有病卻分類錯誤者(PA)較為嚴重，也表示實際上所需付出的成本較高，此指標是越小越好。

灰關聯聚類的具體程序：

本研究修改傳統灰關聯聚類法的處理程序，而成為下列四項主要工作流程，以便完成資料庫之病症分類：(1)建立灰關聯空間，包含上述經資料處理後而構成因子空間的比較序列及選取改善型灰關聯測度。(2)產生固定的參考向量，本研究以資料庫中有病與無病類各自的平均值作為各類的參考向量。(3)計算每筆紀錄和兩類參考向量之間的灰關聯值。(4)比較每筆紀錄的兩類灰關聯值，取其大者為分類的依據。本研究隨機挑選 30 筆無病與 24 筆有病之混和資料作為測試樣本，並以此樣本做為比較不同方法的差異，本分類法的具體程序詳述如下：

(1) 讀取資料庫中有病類別之資料。
(2) 作極性處理：同上述數值屬性依其性質而分別以公式(1)-(3)處理；類別屬性依其屬性值而分別給予量化，處理之後，將資料矩陣命名為 T_p。
(3) 產生有病參考向量：在 T_p 矩陣中，以屬性值的平均形成有病參考向量，存放到 Cent 矩陣的第一列。
(4) 讀取資料庫中無病類別之資料。
(5) 作極性處理：同步驟 2 之方式處理，並將資料矩陣命名為 T_a。
(6) 產生無病參考向量：同步驟 3 之方式處理 T_a，得無病參考向量，存放到 Cent 矩陣的第二列。
(7) 讀取混和測試資料。
(8) 作極性處理：同步驟 2 之方式處理，並將處理後的資料矩陣命名為 D_t。
(9) 計算灰關聯值：將 D_t 矩陣中每一列向量以下列方式計算灰關聯值：
    (i)選取 Cent 矩陣第一列的有病參考向量
    (ii)求 D_t 中每一列向量與參考向量之差的絕對值，得一個 Del 差值矩陣
    (iii)將 Del 差值矩陣中每個元素代入公式(4)或(6)，得灰關聯係數矩陣 R

(iv)將灰關聯係數矩陣 R 以列方向求其平均值，得灰關聯行向量 a

(v)選取 Cent 矩陣第二列的無病參考向量，重複步驟(ii)~(iv)，得灰關聯行向量 b

(vi)將 a, b 兩類灰關聯行向量分別存放到 R_g 矩陣的第一行(預測有病)和第二行(預測無病)

(10) 分類：逐筆比較 R_g 矩陣中每筆紀錄的兩類灰關聯值的相對大小，取其較大者給予該序列分類至其所代表之類別，得如同表三之混合矩陣。

(11) 評估分類效能：將混合矩陣之數據以上述五項分類效能指標計算，以這些指標評估分類系統之效能。

結　　果

為證實本論文提出的灰關聯聚類法的有效性，本研究採用同一測試樣本用以比較傳統型及改善型灰關聯聚類的差異，其中「收縮血壓(BP)」的中值設為 110 mmHg，兩者的測試結果分別說明如下：

傳統型灰關聯聚類的測試結果：將測試樣本依上述演算法進行處理、計算、分類及評估，在計算灰關聯時是使用傳統型公式(4)和(5)，辨識係數設為 0.365，樣本分析結果如下： 兩類樣本的參考向量分別為 Cent =

有病類的參考向量: [0.7335 0.6252 0.5300 0.2703 0.3924 0.2829 0.8594 0.4410 0.2865 0.6042]

無病類的參考向量: [0.6860 0.5956 0.4586 0.1137 0.1000 0.2111 0.5604 0.3583 0.0875 0.4806]

樣本經測試之結果如表四所示，並以此計算分類效能指標如表五所示。

表四、傳統型灰關聯聚類的測試結果

	預測無病	預測有病
實際無病	30	0
實際有病	9	15

表五、傳統型灰關聯聚類的效能指標

分類正確率 CA	有病精確率 P	有病捕捉率 R	有病 F-Measure	分類成本 Cost
83%	100%	62.5%	0.7692	45

改善型灰關聯聚類的測試結果：使用上述同一組參考向量及測試樣本，並依照上述演算法進行處理、計算、分類及評估，此時在計算灰關聯時是使用改善型公式(6)和(5)，辨識係數ξ同樣設為 0.365，對同一測試樣本之測試結果如表六，並以此計算分類效能指標如表七所示。

表六、改善型灰關聯聚類的測試結果

	預測無病	預測有病
實際無病	29	1
實際有病	6	18

表七、改善型灰關聯聚類的效能指標

分類正確率 CA	有病精確率 P	有病捕捉率 R	有病 F-Measure	分類成本 Cost
87%	94.7%	75%	0.8371	31

討　　論

由表五及表七可知，雖然傳統型在有病精確率的指標上有良好的成果(100%)，但是改善型在此指標亦有不錯的結果(94.7%)；除此之外，在其他更有相對重要性的指標上，改善型的分類正確率有 87%、有病捕捉率 75%、有病 F-Measure 等於 0.8371 和分類成本等於 31 等指標皆優於傳統型。然而，為何有病捕捉率的重要性比有病精確率高呢？因為影響有病精確率的主要因素是在於實際無病被誤判為有病的人數(AP)，而影響有病捕捉率的主要因素在於實際有病被誤判為無病的人數(PA)，由此可知，若是實際無病被誤判的結果可能是虛驚一場，但若是實際有病被誤判的結果可能要付出慘痛代價；這問題也可從分類成本的定義上知道：為何要賦予實際有病卻分類錯誤者(PA)較大的權重？故以此觀之，改善型灰關聯聚類法仍是優於傳統型。雖然如此，本研究另從有病 F-Measure 的定義及其數據了解：應該要同時提高有病精確率和有病捕捉率兩者，才能提升整體的分類正確率。因此，就全體效能指標而論，傳統型和改善型皆有其限制，未來應研究如何突破此限制。

若將前述分類效能指標改為以無病症為主的指標，則在分類正確率和分類成本的公式是一樣的，僅在另外三項指標應修正如下：

(1)無病精確率 P=AA/(PA+AA)

(2)無病捕捉率 R=AA/(AP+AA)

(3)無病 F-Measure=2*P*R/(P+R)

本研究同樣根據表四及表六的分類結果，分別計算上述無病分類效能指標，得其結果如下表八及表九。由此可知，若以無病分類效能指標比較兩型分類法，傳統型轉變為在無病捕捉率方面優於改善型，然而在其他指標上，仍然是改善型較佳。除此之外，本研究發現無病精確率和有病捕捉率互為正相關，以及無病捕捉率和有病精確率也是互為正相關，尤其是當正確預測無病人數和有病人數相等時(AA=PP)，無病精確率即等於有病捕捉率，以及無病捕捉率等於有病精確率，亦即說明無病分類效能指標和有病分類效能指標具有相關性，故選擇分類效能指標時，兩者只取其一即可，但是仍建議以有病分類效能指標為主要考量，因為如上所述：若是實際有病被誤判可能會得到嚴重後果。

表八、傳統型分類結果的無病效能指標

分類正確率 CA	無病精確率 P	無病捕捉率 R	無病 F-Measure	分類成本 Cost
83%	76.9%	100%	0.8695	45

表九、改善型分類結果的無病效能指標

分類正確率 CA	無病精確率 P	無病捕捉率 R	無病 F-Measure	分類成本 Cost
87%	82.9%	96.7%	0.8923	31

本研究對資料庫中的紀錄進行取樣時，發現某些記錄有奇異情形，例如有一條冠狀動脈被阻塞或是有心絞痛等情形，但卻被歸類為無心臟病，這會影響分類的正確率。本研究考慮每人的生理參數可能存在極大的差異性，以及尊重資料庫提供者的決定，因此，本論文的分類結果是並未將那些奇異資料排除。若依據資料的領域分析，應將那些資料標記為無效而去除，如此則可以大大提升分類效能。

結　　論

本論文以新型式灰關聯聚類法完成心臟病資料庫的分類。本分類法的所有處理程序皆須映射至灰關聯空間內進行。本論文經實證結果顯示傳統型灰關聯聚類法在有病精確率的指標上優於改善型在此指標的結果，但在其他重要指標上，例如分類正確率、有病捕捉率、有病 F-Measure 和分類成本等指標，改善型皆優於傳統型，由此足以證明本論文提出的改善型灰關聯聚類的有效性及進步性。本論文亦分析無病分類效能指標而發現無病和有病分類效能指標具有相關性，故兩種分類效能指標擇一即可，但是仍建議以有病分類效能指標為主。本研究又從 F-Measure 的數據知道：傳統型和改善型皆有其限制，故未來應以突破此限制為研究目標，使得更能提升整體的分類效能指標。

參考文獻

[1] 何逸儒，病理檢驗醫學，台北:力大，2002。

[2] Hines WW, Montgomery DC, Goldsman DM, Borror CM: Probability and Statistics in Engineering. John Wiley & Sons; 2003.

[3] Deng JL: The control problems of grey systems. Systems & Control Lett 1982, 5:288-294.

[4] 鄧聚龍，灰色系統理論與應用，台北:高立，1999。

[5] 溫坤禮，灰色理論，台北:五南，2009。

[6] Liu L, Zhou JZ, An XL, Yang L, Liu SQ: Improvement of the grey clustering method and its application in water quality assessment. Wavelet Analysis and Pattern Recognition 2007, 2:907-911.

[7] Zhu C, and Liu Q: Evaluation of water quality using grey clustering. Second International Workshop on Knowledge Discovery and Data Mining 2009, 803-805.

[8] 翁慶昌、陳嘉欉、賴宏仁，灰色系統基本方法及其應用，台北:高立，2001。

[9] 陳曉瑩，應用灰關聯度於肝功能檢查結果分析之研究暨電腦工具箱之研發，建國科技大學自動化工程系 2009 碩士論文。

[10] Wong CC, and Chen CC: Data clustering by grey relational analysis. J. Grey System 1998, 10:281-288.

[11] Wong CC, and Lai HR: A new grey relational measurement. J. Grey System 2000, 12:341-346.

[12] http://www.liacc.up.pt/

[13] 李嘉陵、蔡坤龍，應用灰色統計於心臟病資料屬性分析，銘傳健康科技學刊 2010，1:1-6。

A Novel Grey Relational Clustering Methodology Applied to Classification of Heart Diseases

Steven Tsai[1], Jia-Ling Lee[2, *]

[1] Department of Electronic Engineering, School of Information Technology, Ming-Chuan University, 333 Taoyuan, Taiwan, R.O.C.,

[2] Department of Biomedical Engineering, School of Health Technology, Ming-Chuan University, 333 Taoyuan, Taiwan, R.O.C.

Abstract

A novel grey relational clustering methodology is to be proposed in this paper in order to have advanced in classifying the heart-disease database of university of Porto. We modify the processing flow of a traditional grey relational clustering methodology into four main subprocesses in the following: (1) To construct a grey relational space, including the grey relational factor space and improved grey relational measure. We will choose ten factors highly related with heart diseases, including 6 numerical attributes and 4 ordinal attributes. These factors via data processing can become a comparison vector with comparable property. (2) To generate two kinds of reference vectors in the grey relational space. We use the arithmetic mean of each kind of those patients in the database as their respective reference vectors, which remain fixed in the whole period of data processing. (3) To compute two kinds of grey relational measures between two reference vectors and each comparison vector for each record. (4) To compare those measures for each record, and then to classify each record into either one of the two classes of those patients in the database. The evaluated trial shows that the rate of accurate classification can attain 87%, and the rate of getting heart diseases can attain 75%, F-Measure index equals 0.8371, and the classification cost equals 31. It is clear that the novel classification method proposed in this paper is superior to the traditional one. Hence, this fully has demonstrated that the proposed classification scheme in this paper is effective and progressive in classifying those patients in the database.

Keyword: Heart disease, Grey relational, Grey clustering, Classification index
Corresponding author: Jia-Ling Lee [jllee@mail.mcu.edu.tw]
Received 8-25-2014 / Revised 10-21-2014 / Accepted 11-4-2014 / Online published 11-23-2014

MC-Transaction on Biotechnology, 2014, Vol. 6, No. 1, e4