運用粗糙集合理論於二階段方法中探勘糖尿病臨床用藥知識

黃純敏
雲林科技大學 資訊管理學系
副教授
huangcm@yuntech.edu.tw

林蒼亨
雲林科技大學 資訊管理學系
碩士生
viclin513@yahoo.com.tw

張精哲
雲林科技大學 資訊管理學系
碩士生
g9523717@yuntech.edu.tw

摘要

資料探勘在醫學領域的應用，近十年來已經有許多良好成果，尤其在糖尿病與心血管疾病等慢性病，醫師若能以慢性病的臨床用藥經驗來對於患者作健康照護，則對慢性病患極有幫助。

由於醫學疾病的診斷，各項檢查值與疾病特徵的判斷與人類的思考模式相似，具有不精確的特性，有時是屬性介於二種不同疾病之間而難以區別，具有近似性、邊界之特性。有鑑於此，本研究利用應用粗糙集合理論來得到較佳之規則庫和正確率。

藉由醫師平日之看診習慣及客觀的生化檢查數據，運用二階段方法來探勘醫師平常之診斷模式及用藥的知識，得到有97.5%涵蓋率之8條決策規則。以粗集簡化方法、分解樹、神經網路做分類比較，結果傳統之粗集方法具有最高之89.1%正確率及96%涵蓋率。

關鍵詞：糖尿病、粗集理論、分類回歸樹、資料探勘

1. 前言

資料探勘能夠分析儲存在資料庫中的資料，經由對資料的分類或是分群加以模擬與建構預測模式便可發現其屬性間相關性，本研究以人工智慧之方法及技術應用在臨床用藥探勘上，其中包含模糊分群、決策樹、粗集理論、貪婪演算法等，得以改進診斷速度、正確性、穩定性。

過去已有文獻提供其應用之成功經驗，例如：藉由人工智慧模擬演算法如基因演算法、遺傳規劃、模擬策略、模擬規劃、分類系統、混合系統等六種演算法來解決醫學上包括診斷、預判、圖像、信號處理、規劃、排程等問題[25]。也有學習分類系統與規則庫系統整合運用對流行病監控，然而論效能，C4.5 決策樹則優於此項流行病監控系統[13]。以貝氏分類、類神經網路、決策樹等人工智慧技術，應用於心臟疾病的診斷上，結果發現貝氏分類與決策樹較優且較容易說明預測的結果[15]。

在乳癌動態磁性推論影像之應用，類神經網路則比放射線系統得到更好結果。類神經網路、定理元素分析與信號雜訊比對乳癌與糖尿病的診斷屬性監控，結果類神經網路僅次於信號雜訊比。類神經運用在早期乳癌、肝癌、肺癌等癌症的偵測與診斷。運用機器學習方式對攝護腺癌做預測、肝癌方面有使用類神經網路預測慢性 C 型肝炎帶原的病源與臨床因子其預測正確率高達 95%，比多變量對數迴歸的86.7%正確率更高[2]。

就上述發現目前應用於醫學上大都使用類神經網路與案例庫推理，在糖尿病文獻部分則較少，因此本研究希望透過決策分析科學粗集理論從專家之臨床用藥資料加以訓練學習，進而能產生決策知識規則庫，以輔助專家診斷。

2. 文獻探討

決策支援與知識探勘在醫學上的應用已有許多文章被提出[1, 27, 30]，甚至粗集理論在醫學應用上也逐漸被發展為一可行的模式，Pesonen 等人也以類神網路演算法應用在急性闌尾炎的診斷[8, 9]，Davide Dazzi 等人則利用模糊類神經(Neuro-fuzzy)方法在血糖控制的標準[3]，這些文章中使用了包含自適應共振網路、自組織映射，以及倒傳遞等，而這些方法皆屬於人工智慧的計算方法[8]。尤以粗集理論由 Pawlak[20]在八 0 年代所提出，被應用在特徵分類與辨識上。Carlin[7]等人以粗集工具 Rosetta 應用於醫藥知識庫的分類上，皆已獲得許多成就，林林總總的人工智慧應用已使得機器學習、資料探勘[18]知識的獲取、專家系統與決策系統[28]的模式建立皆漸漸趨於成熟。

2.1 糖尿病之鑑別

臨床上當空腹血糖大於 130 mg/dL 或是二小時飯後血糖大於 200 mg/dL 者定義為糖尿病的危險因子，而非胰島素依賴型也就是糖尿病第 II 型則佔有 95%以上。糖尿病高危險群依定義上規則而做的歷年統計[35]中，吸煙者低密度膽固醇含量高於不吸煙者[11]，而肥胖者容易過度的在內臟堆積脂肪，其中三酸甘油脂(Triglyceride; TG)一項，肥胖者比非肥胖者高，它會囤積在肝臟中，當其囤積超過其輸出之能力，則容易罹患冠狀動脈疾病、心肌梗塞及冠心症等相關疾病；或低密度膽固醇(Low Density Lipoprotein; LDL)，它會攜帶血液中 70%的膽固醇，生產過多或代謝太慢會導致膽固醇的增加，長期血脂過高，由上述判定肥胖與糖尿病和粥狀動脈硬化有正的相關[10]。

根據臨床上的解釋，高密度膽固醇(High Density Lipoprotein ; HDL)是所謂好的膽固醇，HDL 與 LDL 血液中含量產生異常則容易罹患心血管疾病，通常 HDL 的標準值是 35 mg/dL 以上，而 LDL 之標準值為 130 mg/dL 以下，因為 HDL 會吸收游離之膽固醇經由膽汁排出，故又稱好的膽固醇[37]。

所有糖尿病患者中約有 40%上的人患高血脂，而有 80%以上的患者具有三酸甘油脂過高[14]，在臨床上，糖尿病合併之高血脂通常區分為原發性或續發性，而血脂症中則以高膽固醇血脂症最常見，另家族混合性血脂異常，因肥胖與酒精刺激而導致三酸甘油脂血脂症也較常見，而糖尿病之併發其它疾病如，眼、腎臟、心臟疾病等[36]共分為 10 類依國際疾病碼(ICD9)為 250.1-250.10。

2.2 粗集理論

它是一種新的數學工具，這個方法以人工智慧與認知科學為基礎，它改良 G. Frege 於 1904 年所提出之二元邏輯，因現實上資料的分佈具有不可辨識性，也就是它們並不能被分類出來，Pawlak[20]將這些不可區別的個體都歸入邊界區，又叫做粗糙集(Rough set)或粗集，為簡化方便以下皆簡稱粗集理論。

粗集理論的概念與 Dempster-Shafer 的實證理論有重疊的部份，它們之間的不同在於實證理論以信賴函數作為主要的工具[33]，而 Pawlak[21, 24]又將此一集合定義為上近似和下近似的差，利用上近似與下近似都可由數學公式計算出結果，故含混的元素數目就可以被計算出來。

近來粗集理論使用在很多領域中，尤其在醫學之應用上極具有潛力，特別是當一個疾病為多屬性時，利用粗集理論來簡約，頗符合疾病分類的觀點，而臨床上疾病常大多具有多種屬性，基於不確定且資訊具有不可辨識的特性，當疾病屬性過於複雜而不易分辨，利用粗集理論則具有良好的使用性[29]。
它的基本概念如表 1：

表 1：決策表 Pawlak [20]

	屬性			決策
	頭痛	肌肉痛	體溫	感冒
e1	Y	Y	Normal	No
e2	Y	Y	High	Yes
e3	Y	Y	Very high	Yes
e4	N	Y	Normal	No
e5	N	N	High	No
e6	N	Y	Very high	Yes

粗集理論使用不可區別的關係，例如集合{(e1,e2,e3),(e4,e6)}為二個子集，無法區別彼此，它們明顯的不可區別關係，也叫做等價關係，而不可區別的集合，我們叫它元素集合(elementary set)。

任意二個基礎集合的聯集，被稱作可定義集(definable set)。不可區別關係很容易去找到冗餘的屬性。當元素關係個別存在則任意屬性屬於這個超集，意謂一個集合的屬性和它的超集定義了相同的不可區別關係[12]。

而頭痛和體溫屬性並沒有不可區別關係，肌肉痛與體溫則存在不可區別關係，對於超集而言形成了單點(Singleton)，{頭痛,肌肉痛}與{肌肉痛、體溫}則具有交集關係，因此肌肉痛是一個多餘的屬性。而{頭痛}和{體溫}構成的集合，它沒有多餘的屬性關係，因此又叫做最小集合或獨立(independent)，它是簡約(Reduction)後的屬性集合對於此一資訊表，它具備含糊和不確定的特色，並且能產生了一決策表如表 1。

2.3 粗集理論簡化方法

上述粗集理論以決策知識表來表達不可區別集合，若決策表的近似空間有多個屬性值，粗糙集合提供了一個特徵選取的方法[1, 26]，以降低屬性找到一最佳的屬性副集合來表現分類的資料關係，我們稱它為古典方法，古

典方法必需計算所有屬性之確定因子，通常使用耗費搜尋增加了計算複雜度。Ziarko[34]證明了屬性簡約為一 NP-Hard 問題，因此本研究提出利用模糊分群與決策樹的方法來適應粗集理論簡約。

古典粗糙集簡約以上近似、下近似及邊界來針對不可辨識集做簡約，逼近的結果可以消去多餘屬性。Pawlak 認為所有的知識都建立在論域上，而我們建立一資訊系統有個實體 S=(U,A, V_a, f_a)，這裡：

1、U 是論域含有非空集合的有限物件，

2、A= $\{a_1 a_2 ... a_m\}$ 是屬性 $C \cup D$ 的一個非空集合的有限集合。這裡 C 是條件屬性的集合，D 是決策屬性的集合。

3、V_a 是 a 屬性的領域，每一屬性 $f_a : U \times A \to V_a$ 針對任意 $a \in A$。

4、$f_a : U \times A \to V_a$ 是所有決策函數稱作資訊函式如 $f(x,a) \in V_a$ 對 $\forall a \in A, \forall x \in U$。

不可區別關係：設 (U,A, V_a, f_a)是一個資訊系統，B $\subseteq A$ 且 x $\subseteq U$ 而有任意屬性子集合 B $\subseteq A$，一個二元的不可區別關係，又叫做 B-indiscernibility 關係它的定義如下：

$$IND(B)=\{(x,y) \in U \times U : a(x)=a(y), \forall a \in B\} \quad (1)$$

決策規則：讓 $|C|$ 表示在 U 中所有物件的集合，而 U 含有 C 在資訊系統中的意義，這裡|*|指示一個集合中的基數。若決策規則中是 $C \to D$ 意謂 IF C THEN D，這裡 C= $\{c_1 c_2 ... c_m\}$ 是條件屬性，D= $\{d_1 d_2 ... d_m\}$ 為決策屬性，則在資訊系統中由 $C \to D$ 的支持度將會是：

$$supp(C,D)=card(C \cap D) \quad (2)$$

對任意條件屬性的集合在資訊系統中它的機率被定義為 $\sigma(C) = p(\lfloor C \rfloor)$，而有決策規則 $C \to D$ 的條件機率是

$$\sigma(D/C) = p(|D|/\lfloor C \rfloor) \quad (3)$$

確定因子：決策規則 $C \to D$ 的強度由 $\sigma(C,D) = \frac{supp(C,D)}{card(U)}$，決策規則之確定因子 (certainty factor) 可由 Cer(C,D)表示如:

$$Cer(C,D) = \sigma(D/C) = \frac{supp(C,D)}{card(C)} = \frac{\sigma(C,D)}{\sigma(C)} \quad (4)$$

確定因子被解釋為一條件機率中 y 屬於 D，所給予的 y 屬於 C，它很容易以 Cer(C,D) ∈[0,1]，若 cer(C,D)=1 則所給的決策規則是可決定的或確定決策規則，否則 0<cer(C,D)<1，所給予的決策規則是不可決定的或不確定的決策規則。

涵蓋因子(coverage factor)：由 $C \to D$ 之決策規則可由 Cov(C,D)表達如下：

$$Cov(C,D) = \sigma(C/D) = \frac{supp(C,D)}{card(D)} = \frac{\sigma(C,D)}{\sigma(D)} \quad (5)$$

2.4 決策樹

決策樹是經由常見的程序-遞迴劃分(recursive partitioning)而建立的。即所有訓練集中的紀錄，都被放入一個大箱子裡，再將每個資料持續進行反覆劃分的過程。該演算法嘗試把這些資料分離，對每一個欄位使用某一種可能的二元分岔(binary split)。該演算法選擇這種可以把資料分成兩部分的分岔法，分完的資料會比原始資料較為純化。再來將這樣的分岔或分割，應用到每一個新箱子上。這樣的程序會持續到沒有可用的分岔為止。如此，該演算法的核心就是決定初始分岔的規則。

CART(分類迴歸樹)演算法為Brieman et al.(1984)年所發表出來。CART藉由一個單一輸入變數函數，在每一個節點分隔資料，以建構一個二分式決策樹。因此，第一個任務是決定哪一個自變數可成為最好的分隔變數。此處所謂最好分隔的定義是能夠將資料最完善的分配到一個單一類別支配的群體。

Breiman 1984[6] 年提出 CART (Classification And Regression Tree) 以 Gini impurity 做為特徵選取之標準如式 6，這裡 p_i 是類別 i 於 t 之機率，所有次分割的差異被加總，並且讓差異中有最大簡化被選取。

$$I_G(t) = 1 - \sum_{i=1}^{m} p_i^{\,2} \quad (6)$$

CART 的最大優點之一，就是演算法會自動檢驗模型，找出最佳的一般模型。其作法是先建立一個非常複雜的樹，再根據交互檢驗或測試集檢驗的結果，將該複雜樹修剪成最佳的一般決策樹。根據各種修剪版本對測試集資料的效能，來決定修剪的結果，最複雜的樹通常不會是最好的決策樹，因為它有過度學習(over-fitting)的傾向，為了解決因過度學習而使得樣本外效果不佳的情況，有修剪法(pruning)及盆栽法(bonsai technique)之解決方式。

2.5 模糊分群

Fuzzy C-means (FCM)是資料演算法裡，由

隸屬函數的程度決定，每一資料點屬於一個群集。Bezdek[5]在 1973 年提出這個演算法來改善較簡易的 hard C-means。FCM 分割一向量資料 $x_i, i=1,...,n$ 到 c 模糊群裡，並且找到每一群的群集中心如同差異度測量最小值的成本函數。HCM 與 FCM 的差異在應用模糊分割資料點屬於某些群集的程度，由隸屬函數決定在 0 和 1 之間。

$$\sum u_{ij} = 1, \forall_j = 1,...n \qquad (7)$$

在正規化的條件下加總資料集合的隸屬程度。

$$J(U, c_1,..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \qquad (8)$$

這裡 u_{ij} 介在 0 與 1 之間; c_i 是第 i 群的模糊分群中心; $d_{ij} = \|c_i - x_j\|$ 是第 i 點到第 j 點的歐氏距離 (Euclidean distance)。$m \in [1,\infty)$ 是權重指數。而目前已知隸屬函數在 Bezdek 之後一系列的文章發表。[John Sum, Lai-Wan Chan]也提出不同的隸屬函數。

$$u_k(x) = \begin{cases} 1, x = v_k. \\ 0, x = v_i \wedge j \neq k. \\ \left[\sum_{j=1}^{c}\left(\frac{\|x - v_k\|^2}{\|x - v_j\|^2}\right)^{1/(m-1)}\right]^{-1} \text{ otherwise.} \end{cases} \qquad (10)$$

2.6 評估準則

預測模式可靠度

利用假說判斷表(Contingency table)與受測者操作特徵曲線(Receiver Operating Characteristic Curve，ROC)，進一步分析預測模式的可靠度。

1. 假說判斷表以表的每一行表示目標分類，而每一列表示推論分類，表的第j 行第i 列的元素值，表示應當屬於第j 種分類，而被網路推論為第i 種分類的案例數。舉例說明，當建立某一疾病分類的假說判斷表時，將所有的驗證分類依其假說分類成「真」與「偽」兩類，凡其「目標」分類為疾病患者為真，否則為偽；另外再依其推論分類成「正」與「負」兩類，凡模式「推論」分類為疾病患者則為正，反之為負。根據表(2) 可以得到兩個主要的衡量指標，這兩個衡量指標，其值愈大表示此預測模式之效果愈佳。

I. 敏感度(Sensitivity)：衡量此模式能夠正確分類罹患疾病之病患的有效程度，以 TP/(TP+FN) 計算。

II. 鑑別度(Specificity)：衡量此模式能夠正確分類非罹患疾病之病患的有效程度，以 TN/(FP+TN) 計算。

2. 受測者操作特徵曲線(ROC)分析主要應用於決策時的參考依據，藉由觀察 ROC 曲線上的操作點（Operating point），可瞭解決策問題本身的損益平衡狀況，而應用在預測的問題時，可作為評估預測模式的局部性能。利用最佳分類精度來評估分類器的優劣是最常使用的評估標準，針對完成訓練的分類器，如能完全正確地將真實分類(True class)與錯誤分類(False class) 全數分類正確，則此分類器可說是最佳分類器；而True Positive (TP)與 False Positive (FP)是衡量分類器的性能指標。在分類問題中，TP 是指將目標樣本分類的正確樣本數目比率，而 FP 是指非目標樣本分類成目標樣本的錯誤樣本數目比率。TP 又稱為敏感度(Sensitivity) ， 而 (1-FP) 稱為鑑別度(Specificity)。完成訓練的分類器，能夠得到一組(TP、FP)或稱為操作點，而ROC曲線是一條由許多不同的操作點(TP、FP)所組成，藉由曲線的形狀，能夠觀察分類器的損益平衡情形。評估方式是以曲線下方面積（Area Under the ROC Curve；AUC）作為ROC曲線評估的準則，此數值若是愈大則表示ROC曲線愈佳。

表 2：假說判斷表(Contingency table)

推論＼目標		假說	
		真	偽
判斷	正	True Positive (TP)	False Positive (FP)
	負	False Negative (FN)	True Negative (TN)

3.研究內容與方法

本篇文章針對糖尿病提出一個二階段方法的計算模式，根據臨床醫師判斷實驗診斷之結果集如圖 1，再由其中糖化血紅素與血糖之生化檢驗資料做為糖尿病的診斷依據如圖 2，經醫師診斷病症並開立處方之流程，由這個流程建立一個決策模式如圖 3。

一般而言，醫師診斷之生化檢驗資料，常只針對其中某些項目做檢驗，如針對高血壓患者測量其血液中脂肪含量；針對疑似糖尿病患者，測量其血液中血糖濃度。

因此在檢驗資料不完全的情形下，有必要初期先建立一模式，利用無監督學習的方法，針對檢驗結果建立一初步的規則庫，做為開立處方治療的決策。

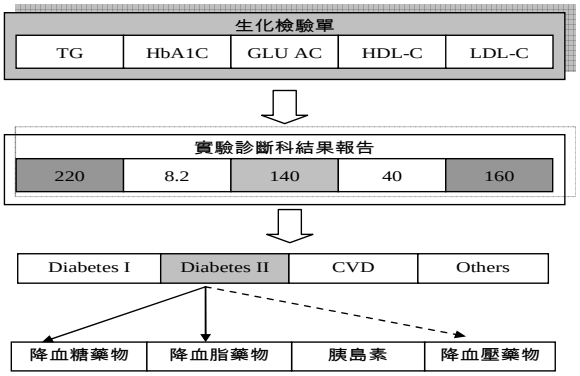

圖 2 醫師臨床鑑別診斷與用藥之流程

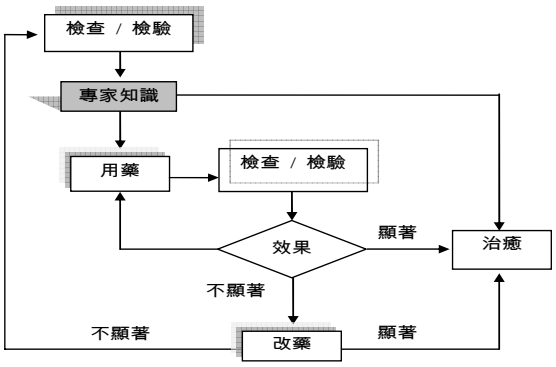

圖 1 醫學專家(醫師)之用藥知識

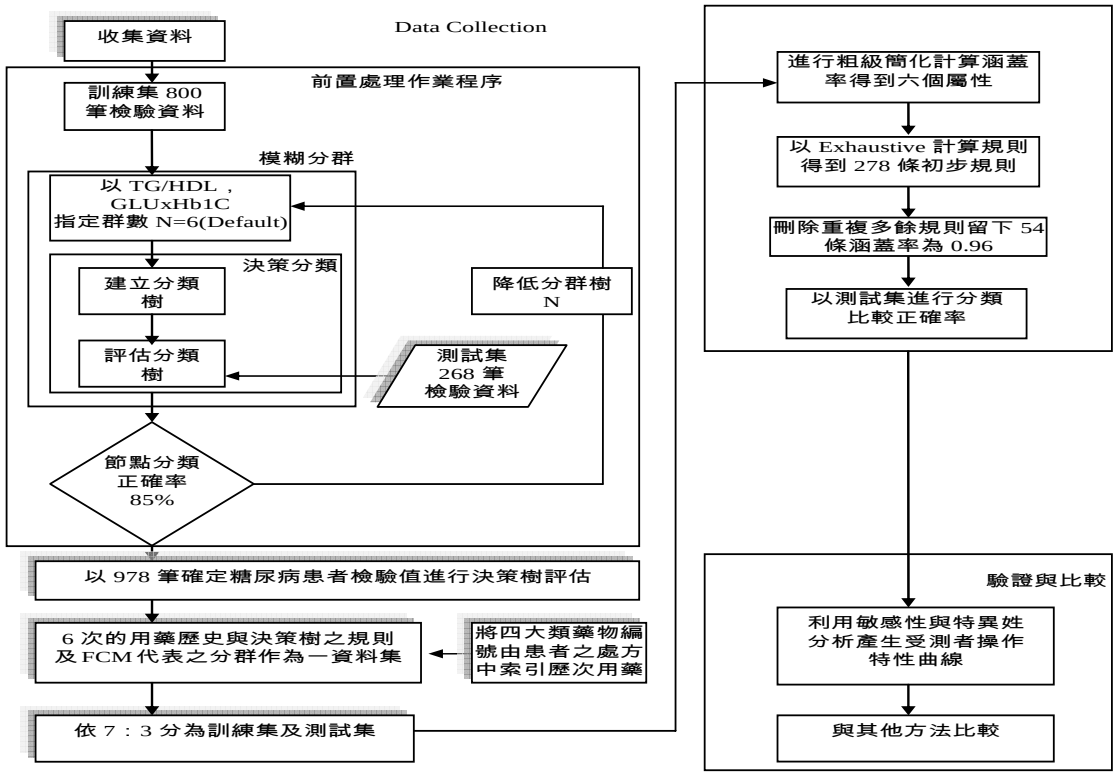

圖 3 研究模式

3.1 適應的決策規則

Zadeh 提出運算方式包括神經網路、模糊邏輯、等人工智慧方法， Yeh 等人提出階層式運算方法針對資料的屬性，去區別小血球性貧血已獲得一些成果[31]，而醫學資料常需要專家知識，針對不同的屬性提供有效的建議，而糖尿病之醫學知識常需藉助於臨床醫師之經驗。基於此種運算應用，本研究提出一適應的探勘方法如圖 3，利用一般生化檢查之資料集，透過分類探索可能的檢驗規則庫，臨床醫師以檢查做為前期的疾病推論，接著以患者健康狀態反應藥物動力學的效果，反應在患者另

次的檢驗結果上。

區別函數(Discriminant function)應用於血液檢查可得知檢驗值之趨勢，Yeh 等人利用 CBC 項目之檢驗值，以 Fuzzy C-means 驗證出良好的分群效果[31]，其中 TG 與 HDL 的比值反應出 LDL 隨 TG 上升而上升，針對 1068 筆檢驗資料中正常值比較其比值，p=0.001，而 GluAC 與 HbA1C 之比值常能反應糖尿病的特性。

本研究收集三個資料集，第一個資料集收集一般生化檢驗值 1068 筆，這個資料集含有五種檢驗值結果(TG/HDL/LDL/Gluac/HbA1c)，並且都屬於心血管疾病、糖尿病、高血壓之檢驗值，但不包含診斷結果值。第二個資料集為 978 個糖尿病患者之檢驗結果值，但只含有上述五項檢驗之部份結果。第三個結果集為 978 個糖尿病患者之六次用藥記錄，原始資料對應到藥物編號，其中共有五種疾病治療藥物，包含高血脂、高血壓、心血管疾病、糖尿病、心臟病等。

3.2 生化檢驗資料集之初步分群

將檢驗資料做為初步分群的概念，在於模擬臨床醫師將檢驗值做為疾病之初步鑑別，提供一簡單的判斷，但與實驗診斷科提供之正常值或危險值不同，在於文章中使用無監督學習做為與臨床用藥資料集之間的映射關係。

初步分群中，本研究定義了適應的決策判斷規則，針對第二個資料集之涵蓋率為 95%以上，針對第一個資料集之測試集之正確率至少 85%以上，本章節初步使用 Fuzzy C-means 做為分群，此分群演算法與其他分群演算法比較有底下優點：

1. 降低導數最佳化過程造成之時間成本，如神經網路[31]。
2. 降低迭代演算法中如 SOM(Self-Organization Map)學習率參數主觀性。
3. 提供一個可修正分群數之觀察結果。

此一初步結果，N=5 時未剪除樹達 94%之正確率。輸入第二個資料集進行決策樹評估，在剪除了 5 層子樹 (Prune Tree)後，仍然可達到 86%的分群正確率。

此法在分群後，還能保有原始資料的訊息，可供後續之研究。不像一般集群分析法在分群後就喪失了原始資料的訊息，藉以做為本篇研究中粗集理論之決策屬性。

3.3 研究設計

此一研究模式含有六次的醫師處方資料，其中我們搜集該院中醫師可能開立之藥物，分為四大類，含有糖尿病治療藥物如胰島素及降血糖藥物，高血壓治療藥物，心血管治療藥物及高血脂治療藥物，並統計其使用次數，同類藥中屬於相同學名而有不同商品名之藥物也一併列入，加以編號，1-10 為治療高血脂藥物，11-30 為治療高血壓藥物，31-39 為治療糖尿病藥物，40-55 為治療心血管疾病藥物。

而每一屬性都先以 Fuzzy C-Mean 做為分群，使得資料離散為 4 種型式，這時分群之結果，沒有特定之解釋意義，只能做為輸入變

表 3 粗集資料集

用藥一	用藥二	用藥三	用藥四	用藥五	用藥六	屬於群七	規則編號八
1	1	3	1	3	4	1	5
2	1	1	1	4	4	3	2

數，用藥資料屬性 1 到屬性 6 為藥物分群如表 3，屬性 7 為 978 個患者之檢驗值所對映到初步分群的編號，屬性 8 為適應的決策樹之決策規則做為決策屬性，978 個患者之檢驗資料針對決策樹評估所對映之規則編號。組合成有 7 個屬性與一個決策屬性的資料集如表 3：

每次用藥之原始資料，本章以初步分群中之檢驗資料分群數 N=4 做為分群目標。

用藥的過程中，臨床醫師可能因為針對藥物作調整，原因是患者對藥物過敏或者不適用，如孕婦對其它降血糖藥物可能導致對懷孕過程的傷害，而替代以 AMARYL 等藥物，故每一 fold 之藥物都有可能隨之改變。由六次的藥物治療的分群結果做為屬性，可反應患者用藥過程改變的狀況。

3.4 建立糖尿病患者之初步分群

檢驗值中，臨床醫師對第二型患者，也就是膽固醇與糖尿病患者判斷指標項目為血糖與糖化血紅素。

本研究係針對全部樣本皆為糖尿病患者，通常第二型依比例為 90%，以 TG/HDL 可反應出 LDL 皆為異常故於本研究可忽略，而

Glu AC 與 HbA1c 之乘積越大者表現出血糖異常，故由 TG/HDL 之比值與 GluAC 與 HbA1c 之乘積，做為 Fuzzy C-means 之輸入。根據圖 4，我們由流程中得到在 N=4 時得出之決策樹規則經由測試集輸入後評估正確率，比較分群數，6,5,4,3 之正確率其步驟如下：

步驟一、 預先計算出資料集中之 TG/HDL 比值與 Glu AC 與 HbA1c 之乘積，做為輸入。

步驟二、 指定分群數 N 為由 6,5,4….遞減。

步驟三、 3.5 節之分類評估步驟，評估分類之結果集之正確率。

由分群結果如圖 4，四個子集合之項目中，第二群之 TG 值最高平均高達 1821 mg/dL，正常值通常為 160mg/dL，HbA1c 之平均值為 9.0667 mg/dL，而第四群則血糖平均值 270 mg/dL 為正常值 120 mg/dL，第三群為 HbA1c 高或者 Glu AC 高。這裡第二群和第四群是兩群差異呈兩極化。

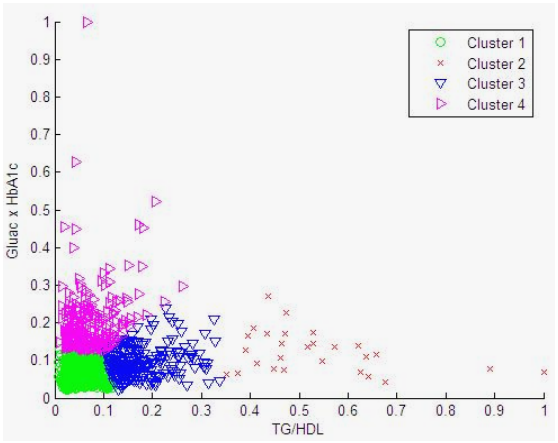

圖 4 Fuzzy C-means 分群結果

3.5 評估決策分群效度

由上述分群之結果，做為決策樹之訓練目標，將分群所得之結果逐一將每一個患者所屬之群的編號，做為決策樹之監督者，分類樹的優點在於分類的結果為一符號或類別，而不需要屬性為數值的分類[6]。在識別每個節點所得之結果，很容易被人的概念所接受。當分類節點太多時，CART 分類出一個最大可能的”樹”，而後再藉由修剪所產生的小樹，我們利用分類誤差的正確率來確認最佳的模型。本研究由足夠的 1068 筆樣本中，利用較多測試樣本 800 筆做為訓練集，而 268 筆做為測試集，我們利用底下的程序來剪除樹，N 由 1 開始剪除，並重覆測試步驟如下：

步驟一、隨機挑選 800 筆資料建立分類樹，設定剪除樹的層 N=0。

步驟二、剪除子樹，由 N=1 開始減除。

步驟三、將 268 筆剩餘資料測試，進行正確率分析。

步驟四、 以第二個資料集 978 位糖尿病患者之不完整生化檢驗結果，測試其規則涵蓋率為 95%以上，若未達則遞增 N 並重覆步驟二，否則停止。

其中原始分類樹共有 10 個 level，Prune 5 個 level 後得到此一結果，如圖 5，共有 8 條規則。

其中 LDL、HDL 都在樹的 Prune 程序中被剪除，根據 NCEP 及 ADA 兩個機構已證明，糖尿病症其皆有患高血脂與心血管疾病之危險因子[17]，故本研究把 HDL 與 LDL 視為已

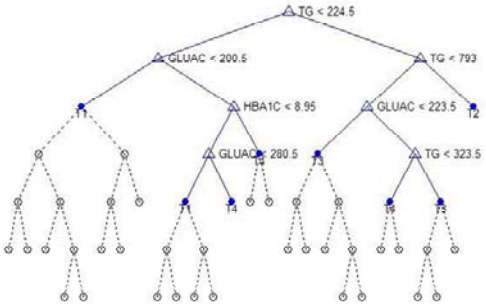

圖 5 分類樹結果

存在之因子，而用藥之結果與 HDL 和 LDL 之間的關係可以忽略。

在研究模式中，由於臨床醫師對於檢驗值之結果常基於長期患者之檢驗經驗，有時只驗 TG 或 HbA1C，或者 HDL 一項，故無法如第一個 1068 筆成人健檢資料集，提供一完整的項目檢驗值，因此有必要透過此一初步的分群模式，來得到 978 筆檢驗值中的所對應的規則。

經過 5 次的修剪子樹，本研究初步獲得 8 條規則如圖 6，當 N=2 時有最高的正確率；但根據圖 1，臨床醫師診斷患者，不全然相信檢驗值之結果，依照臨床醫師之經驗判斷者約為 80%~85%。

因此欲更加確定糖尿病之確認，必需配合藥物之試驗，及藥物之服法、頻次所收到之效果，這些都必需加以考量，因此下一階段將利用 978 名糖尿病患者之某次檢驗資料值，經過此 8 條規則評估結果，涵蓋率達 99.5%，只有 1 個患者對應不到規則，因此最後之決策屬性將有 9 個也就是規則 1-8 及 0，而此 0 規則代表無法區分。

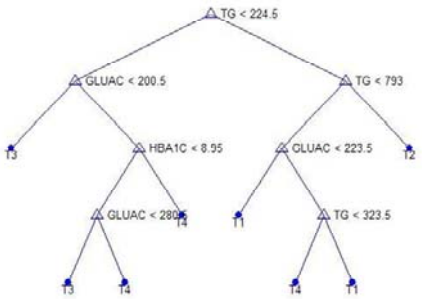

圖 6 分類樹修剪子樹後之結果

4. 實驗結果與驗證

本章節將驗證二階段方法模式在糖尿病患者藥物知識的探勘，利用四種方法加以比較，每一種方法都利用二階段方法模式之前處理決策規則為相同之資料集，這四種方法包含粗集典型方法、LTF-C、分解樹(Decomposition Tree)、前向倒傳遞神經網路，而以涵蓋率及分類正確率加以評估比較。

4.1 以粗集理論分類用藥

本實驗使用 RSES 2 軟體做為分析工具，RSES 是由粗集理論創立者 Pawlak[19, 22, 23] 所屬的波蘭 Warsaw 大學所開發。

本研究之第二階段模式之資料集初步分為二個子集，依 7:3 比例劃分。傳統粗集方法必需先針對屬性加以簡化，若為連續資料，則必需加以離散化後再簡化，本實驗研究收集每位患者之六次用藥，做為條件屬性，加以分群做為粗集方法之輸入，以取代透過人工離散資料之主觀性。其步驟如下：

步驟一、分割 978 個患者之前處理資料集，依 7:3 比產生訓練集與測試集，決策模式產生之規則集，做為決策屬性。

步驟二、利用 Exhaustive 演算法簡化屬性。

步驟三、產生決策規則並計算確定因子，屬性支持度(Support)及涵蓋因子，以刪除多餘的規則數。

步驟四、以測試集測試步驟三產生之簡化規則集，計算正確率與涵蓋率。

本實驗經過 RSES 以耗費演算法共得到 278 條規則，並簡約了一個屬性，其正域為 0.595，該屬性為 978 個患者不完整之生化檢驗值的分群編號，與決策樹之 8 條規則成對，屬於該資料集的邊界，因此計算涵蓋因子與確認因子後，得到一有 54 條決策規則之規則集，以隨機投票方式選取之訓練集進行分類，結果正確率達 93.9%，涵蓋率為 96.3%，使用測試集分類結果針對實際規則進行預測，其涵蓋率與正確率如表 4：

以測試集進行分類結果之正確率為 89.1% 涵蓋率則達 96.6%，錯誤分類數為 41。達到設定之效果，並且與其他的方法比較。

4.2 受測者操作特性曲線(Receiver Operation characteristic Curves)

醫學相關之研究常以敏感度與特異度做為衡量研究效果之工具[16, 31]本研究分類用藥映射到生化結果的規則，以敏感度和特異度做為受測者操作特性曲線來表示分類結果之正確率，橫軸為特異度，縱軸為敏感度，而分類結果常常有太多的假正(false positive)與假負(false negative)之錯誤分類，其中敏感度與特異

表 4 54 條決策規則之分類結果(試測集)

類別	4	5	1	7	0	6	2	8	3 物件數	正確率	涵蓋率
4	12	0	0	0	0	0	0	0	3	21	0.8	0.714
5	0	60	0	0	0	0	0	0	0	60	1	1
1	0	0	174	0	0	0	0	0	0	174	1	1
7	0	4	0	0	0	0	0	0	0	4	0	1
0	0	0	0	0	1	0	0	0	0	1	1	1
6	8	0	0	0	0	1	0	0	1	13	0.1	0.769
2	0	0	13	0	0	0	0	0	0	13	0	1
8	0	0	0	0	0	0	0	4	0	4	1	1
3	2	0	0	0	0	0	0	0	1	4	0.333	0.75
陽性率	0.55	0.94	0.93	0	1	1	0	1	0.2			

度如下：

$$\text{Sensitivity} = \frac{\text{true positives}}{(\text{true positives} + \text{false negatives})}$$

$$\text{Specificity} = \frac{\text{true negatives}}{(\text{true negatives} + \text{false postives})}$$

受測者操作特性曲線以 Sensitivity 與 1-Specificity 做為座標軸，而 Specificity 為正確率，輸入計算後的值產生如圖 7：

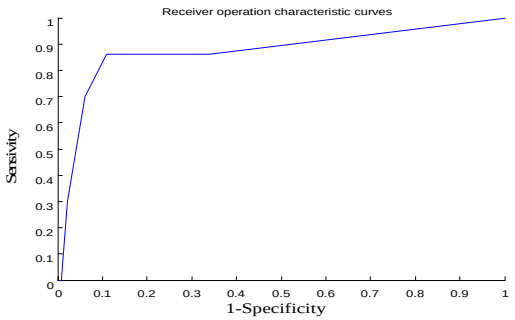

圖 7 受測者操作特性曲線

4.3 實驗結果與其他方法比較

許多文章已經發表關於粗集理論做為特徵選取與分類[6, 26, 32]而神經網路在醫學資料的分類目前也成為趨勢[11, 16, 31]，這些計算方法在醫學資料的應用包括藥物動力學、檢驗資料之疾病診斷都具有相當的效果，目前應用粗集理論於糖尿病用藥之研究仍然不足，因此本研究以三種模式比較其正確率與涵蓋率，其中粗集理論優於 LTF-C(Local Transfer Function Classifier)網路與解構樹[4]與前向倒傳遞之神經網路 (Feed Forward back propagation ;FFBP) 方法。

比較結果發現，解構樹分類正確率為 0.89 而涵蓋率為 0.929，錯誤分類數含假正與假負 (False positives, False Negatives) 共有 39 個分類錯誤數。LTF-C 之正確率 0.82 但涵蓋率為 100%而其分類錯誤數達 38 個。神經網路使用 RSES 提供之 LTF-C 模式，其中 LTF-C 架構同於放射狀基礎網路，訓練集有 13 個神經元，測試 20000 個迭代，正確率為 0.82，涵蓋率為 1，共分類錯誤達 38 筆。本研究以訓練集使用前向式倒傳遞，訓練標準為平均平方根誤差 (RMSE)為 0.005，訓練 684 個樣本再套以測試集結果，正確率為 0.77，涵蓋率為 1，共分類錯誤達 62 筆。

5. 結論

過去疾病之預測在統計上乃運用對數迴歸(Logistic Regression)的方法來進行，本研究捨統計方法而就人工智慧的技術與方法，乃因在人工智慧的資料探勘技術與方法就功能上更能夠找出可能的組合。

粗集理論是一個無監督學習的分類器，它的優點在於藉由資料的特性，產生知識的規則庫，通常規則越多，正確率也越高，但人類的思考通常基於不確定或不精確，故有效的知識簡約方法，可以簡化這些規則成為正確高規則少之有效知識。

資料探勘的技術與方法，舉凡運用在醫學上對醫師的醫療輔助、預防醫學或是疾病養護創新應用等。神經網路雖然是良好的分類器，但其缺點主要在於它是一個黑箱，無法提供模式化結果的規則集，LTF-C 與 FFBP 無法萃取規則知識庫，而分解樹只能提供部份之分類規則，而良好的分類，由研究結果發現 LTF-C 神經網路在分類錯誤數，低於粗集，然而多個分類項中，LTF-C 有 6 類與解構樹有 5 類無法分類，正確率為 0，意謂其敏感性高，而特異性低，粗集只有兩類之正確率為 0，而分解樹之分類正確率略低於粗集，但涵蓋率低於粗集，本研究期望有較高的涵蓋率與較高之正確率，粗集於本研究模式中為一最佳之分類器，來針對糖尿病用藥知識之探索。

從研究結果得到糖尿病用藥在臨床上對患者之健康照護有莫大的影響，其主要的貢獻有：

1. 適應的模式能提供臨床醫師在一般生化檢查危險值判斷提供更進一步的用藥依據。
2. 資訊系統不再只侷限於處方開立與資料查詢，也能整合有效的的知識規則，能提供良好的用藥服務。
3. 提供日後相關糖尿病合併高血脂、心血管疾病、高血壓之用藥推論架構。

本研究在未來的發展，可以朝幾個方向：

1. 關於其它疾病檢驗項目的區別。
2. 糖尿病相關併發症之研究。

參考文獻

[1] Abdel-Aal, R. E., "Improved classification of medical data using abductive network committees trained on different feature subsets," ***Computer Methods and Programs in Biomedicine***, vol. 80, pp. 141-153, 2005.

[2] Abdolmaleki, P., Buadu, L. D., and Naderimansh, H., "Feature Extraction and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging Using Artificial Neural Network," ***Cancer Letters***, vol. 171, pp. 183-191, 2001.

[3] Apte, C. and Weiss, S. M., "Data Mining with decision trees and decision rules," ***Future Generation Computer Systems***, vol. 13, pp. 197-210, 1997.

[4] Banzan, J. G., Szczuka, M. S., Wonjna, A., and Wojnarski, M., "On the Evolution of Rough Set Exploration System," in ***Lecture Notes in Computer Science***. vol. 3066: Springer Berlin / Heidelberg, pp. 592-601, 2004.

[5] Bezdek, J. C., "Fuzzy mathematics in pattern classification," in ***Apply Mathematic center***. vol. PhD Ithaca: Cornell University, 1973.

[6] Breiman, L., Friedman, J., Olshen, R., and Stone, C., "Classifications and Regression Trees," in ***Wadsworth*** Monterrey, 1984.

[7] Carlin, U. S., Komorowski, J., and Ohrn, A., "Rough set analysis of medical datasets in a case of patients with suspected acute appendicitis," in ***Workshop on intelligent Data Analysis in Medicine and Pharmacology***, 1998.

[8] Dazzi, D., Taddei, F., Gavarini, A., Uggeri, E., Negro, R., and Pezzarossa, A., "The control of blood glucose in the critical diabetic patient A neuro-fuzzy method," ***Journal of Diabetes and its Complications***, vol. 15, pp. 80-87, 2001.

[9] Garcia, N. L., Val'kov, V., Khrustalev, B., Karpenko, M., Shkuryaeva, V., Kim, H., and Koehler, G., "Theory and Practice of Decision Tree Induction," ***Omega International Journal Management Science***, vol. 23, pp. 637-652(16), 1995.

[10] Grover, S., "Evaluating the benefits of treating dyslipidemia: the importance of diabetes as a risk factor," ***The American Journal of Medicine***, vol. 115, pp. 122-128, 2003.

[11]Haffner, S., "Rationale for New American Diabetes Association Guidelines: Are National Cholesterol Education Program Goals Adequate for the Patient with Diabetes Mellitus?," ***The American Journal of Cardiology***, vol. 96, pp. 33-36, 2005.

[12] Hanley, J. A. and McNeil, B. C., "The meaning and use of the Area Under a Receiver Operating Characteristic Curve," ***Radiology***, vol. 143, pp. 29-36, 1982.

[13] Holmes, J., Durbin, D., and Winston, F., "The Learning Classifier System: an Evolutionary Computation Approach to Knowledge Discovery in Epidemiologic Surveillance," ***Intelligence in Medicine***, vol. 19, pp. 53-74, 2000.

[14] Kohavi, R. and Frasca, B., "Useful Feature Subsets and Rough Set Reducts," in ***Third International Workshop on Rough Sets and Soft Computing***, 1994.

[15] Kononenko, I., "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective," ***Artificial Intelligence in Medicine***, vol. 23, pp. 89-109, 2001.

[16] Lisboa, P. J. G., "A review of evidence of health benefit from artificial neural networks in medical intervention," ***Neural Networks***, vol. 15, pp. 11-39, 2002.

[17] Nesto, R., "Correlation between cardiovascular disease and diabetes mellitus: current concepts," ***The American Journal of Medicine***, vol. 116, pp. 11-22, 2004.

[18] Ohrn, A., Komorowski, J., Skowron, A., and Synak, P., "The Design and Implementation of a knowledge Discovery Toolkit Based on Rough Sets," in ***The Rosetta system manual***, pp. 24, 1998.

[19] Pawlak, Z., "Probability, Truth and Flow Graph," ***Electronic Notes in Theoretical Computer Science***, vol. 82, 2003.

[20] Pawlak, Z., "Rough Sets," ***Informational Journal of Information and Computer Sciences***, vol. 11, pp. 341-356, 1982.

[21] Pawlak, Z., ***Rough Sets-Theoretical Aspects of Reasoning about Data***: Kluwer Academic Publishers, 1988.

[22] Pawlak, Z., "Rough sets, decision algorithms and Bayes' theorem," ***European Journal of Operational Research***, vol. 136, pp. 181-189, 2002.

[23] Pawlak, Z., Busse, J. G., Slowinski, R., and Ziarko, W., "Rough sets," ***Communication of the ACM***, vol. 38(11), 1995.

[24] Pawlak, Z., Wong, S. K. M., and Ziarko, W., "Rough Sets: Probabilistic versus Deterministic Approach," ***International Journal of Man-Machine Studies***, vol. 29, pp. 81 - 95, 1988.

[25] Pena-Reyes, C. A. and Sipper, M., "Evolutionary Computation in Medicine: An Overview," ***Artificial Intelligence in Medicine***, vol. 19, pp. 1-23, 2000.

[26] Pesonen, E., Eskelinen, M., and Juhola, M., "Comparison of different neural network algorithms in the diagnosis of acute

appendicitis," ***International Journal of Bio-Medical Computing***, vol. 40, pp. 227-233, 1996.

[27] Questier, F., Arnaut-Rollier, I., and Walczak, B., "Massart.,Application of rough set theory to feature selection for unsupervised clustering," ***Chemometrics and Intelligent Laboratory Systems***, vol. 63, pp. 155-167, 2002.

[28] Questier, F., Arnaut-Rollier, I., Walczak, B., and Massart, D. L., "Application of rough set theory to feature selection for unsupervised clustering," ***Chemometrics and Intelligent Laboratory Systems***, vol. 63, pp. 155-167, 2002.

[29] Sarasin, F. P., "Decision analysis and its application in clinical medicine," ***European Journal of Obstetrics & Gynecology and Reproductive Biology***, vol. 94, pp. 172-179, 2001.

[30] Walczak, B. and Massart, D., "Tutorial Rough sets theory," ***Chemometrics and Intelligent Laboratory Systems***, vol. 47, pp. 1-16, 1999.

[31] Yeh, J.-S. and Cheng, C.-H., "Using hierarchical soft computing method to discriminate microcyte anemia," ***Expert systems with applications***, vol. 29, pp. 515-524, 2005.

[32] Zhang, L.-H., Zhang, G.-h., Lang, Y.-U., Zhang, J., and Bai, Y.-C., "Intrusion detection using rough set classification," ***JZhejiang Univ SCI***, vol. 5(9), pp. 1076-1086, 2004.

[33] Zhu, W. and Wang, F.-Y., "Reduction and axiomization of covering generalized rough sets," ***Information Sciences***, vol. 152, pp. 217-230, 2003.

[34] Ziarko, W., Piatetsky-Shapiro, G., and Frawley, W. J., "The discovery, analysis, and representation of data dependencies in databases," in ***Knowledge discovery in database*** Cambridge: AAAI/MIT Press, pp. 213-228, 1990.

[35] 行政院衛生署, "台灣地區每十萬人口死亡率按主要死亡原因," 台灣行政院衛生署報告 2004.

[36] 蕭志彥 and 王明誠, "降血脂藥物與慢性腎病," in ***腎臟透析***. vol. 14, 2004.

[37] 蘇大成, "膽固醇與心血管健康," in ***台大醫訊***, 2003, pp. 35-36.