# 由醫療資料庫發掘有意義之模糊關聯規則

謝楠楨
台北護理學院資訊管理學系

## 摘要

本研究將提出一種適用於醫療資料庫探勘之四階段作業程序，以改善現有關聯規則(association rule)資料探勘研究中常見，如所發掘之關聯規則語意不清晰、關聯規則重複，以及因傳統關聯規則「支持度\信賴度」機制的限制，造成遺失有意義的規則等問題。為使發掘之關聯規則語意清晰，本研究首先運用叢集劃分(cluster partitioning)技術，自動將資料表格中數值資料(quantitative data)的資料欄位，轉換成為口語化述辭(linguistically terms)形式的模糊集合，其後使用自我組織映射圖網路(SOM, self-organizing maps)叢集分析法，依據敏感度分析(sensitivity analysis)所獲得之相對重要資料欄位，以及資料本身特徵，將所有資料區分為數個內部資料特徵相似的叢集，並對各叢集進行關聯規則分析，其後並以模糊相似關聯(fuzzy resemblance relation)概念設計之演算法，將語意近似之重覆關聯規則加以合併。藉由關聯規則之合併，可有效減少發掘關聯規則之數量，且所保留之關聯規則更具資訊表達之完整性(informative)，且更易於醫療領域之解釋及運用。另為判斷關聯規則之可信度，本研究並運用模糊資料庫(fuzzy database)中真實值(truth value)評量方法，保留具較高真實度之關聯規則。最後，我們並使用一真實的疾病醫療資料庫驗證本研究提出的作法。

**關鍵字：**資料探勘、叢集劃分、自我組織映射圖網路、模糊關聯規則、模糊重組關聯、真實值

# Finding Relevant Fuzzy Association Rules from Medical Databases

Nan-Chen Hsieh
Department of Information Management, National Taipei College of Nursing

## Abstract

For data mining applications, association rule can be used to support a decision making process. However, association rule algorithms usually yield a large numbers of rules, and many of the rules may contain redundant, irrelevant information or describe trivial knowledge. In this paper we present a four-stage data mining processes for finding relevant fuzzy association rules from medical database. Fuzzy association rules are especially suitable in medical mining, since they consist of simple linguistically interpretable rules and do not have the drawbacks of symbolic or crisp association rule. In the first phase, the Cluster partitioning technique was used to automatically transform quantitative values into fuzzy linguistically terms. The linguistically terms were modeled by means of fuzzy sets defined in the appropriate attribute domains. Next, a Kohonen self-organizing map (SOM) was used to identify clusters based on shared feature attribute values. The resulting clusters were then classified by feature attributes determined using an Apriori association rule algorithm. Because the association rule algorithm tended to generate large numbers of rules, we present interactive strategies for pruning redundant association rules on the basis of fuzzy resemblance relation to enhance its readability, and evaluate the truth degree of the discovered fuzzy association rules by the truth evaluation mechanism. Finally, we demonstrate our approach on a real disease medical database.

Keywords: Data mining, cluster partitioning, self-organizing map (SOM), fuzzy association rule, fuzzy resemblance relation, truth value

# 壹、簡介

藉由醫療資料分析技術由病患原始資料中，發掘有價值之資訊有助於加強病患健康管理(Chaea et al. 2003)。為達成此目標，許多醫療院所內部之醫院資訊系統(HIS, hospital information system)已收集數量極為可觀，包含病患基本資料、實驗室檢驗數據、醫令等與疾病相關之資料庫，分析此類資料庫可作為瞭解疾病成因與釐定病患健康管理政策之依據。在不同於一般資料分析技術，對於資料進行單純之總結分析(summation analysis)，資料探勘(Chen et al. 1996; Frawley et al. 1991; Lavrac 1999)係為一種自動化的資料分析技術，適合處理發掘資料彼此間規則性(regularity)與固定形式(pattern)的問題，且可產生高度總結形式之有價值知識，是以非常適用於醫療資料庫之探勘應用。

在 Fayyad et al. (1996)所建議之 KDD(Knowledge discovery in database)流程中，主要分為選擇目標資料(selection)、前置處理(preprocessing)、格式轉換(transformation)、資料探勘(data mining)、解釋(interpretation)以及評估(evaluation)等數個階段，在各階段中若處理不適當，即有可能產生問題，影響資料探勘的結果。而現有資料探勘技術中，主要使用的探勘方法有叢集分析(clustering)(Chen et al. 1996; Frawley et al. 1991; Lavrac 1999; Kohonen 1995)、預測分析(predicate) (Chen et al. 1996; Frawley et al. 1991; Lavrac 1999)以及關聯式規則(association rule)(Agrawal et al. 1993; Chen et al. 1996; Frawley et al. 1991; Han & Fu 1995; Lavrac 1999)等技術，本研究主要以叢集分析與關聯式規則為主要資料分析方法，其中叢集分析為依據資料值，將具相似特徵樣本歸屬為同一類別(class)的方法，關聯式規則則是在包含大量資料的資料庫中，將一些資料欄位間隱含之規則性找出來的方法。

然而現有關聯規則於資料探勘實務應用上，存在許多問題，如 Apriori、C4.5、CART 等演算法，在處理數值型態資料(quantitative data)之連續變數(continuous variable)時，往往因資料型態本身的特性，致使會產生數量極為龐大之關聯規則，而且在所發掘之關聯規則中可能存在有許多「重複」(redundancy)以及「無意義」(irrelevant)的關聯規則(Bastide et al. 2000)，以至於造成所發掘之關聯規則難以解釋與應用。此外，傳統關聯規則資料探勘技術，所發掘的規則必需在滿足使用者訂定的最小支持度(support)和信賴度(confidence)門檻值時，此規則才將保留，但在醫療資料庫中某些疾病特徵具特殊性，由傳統「支持度\信賴度」機制，直接對醫療資料庫分析並進行關聯規則取捨，有可能會遺失某些支持度低、信賴度高，但有意義的關聯規則。

為解決此類問題，在(Ordonez et al. 2000)的研究中，嘗試於資料前置處理階段，將數值資料型態之連續變數，以劃分為數個區間(interval)的方法，將其轉換為固定的數個述辭(term)，以減少相對屬性領域(attribute domain)的大小，並以該類述辭取代原數值資料作為資料探勘新的輸入資料，以此作法雖可有效減少資料探勘後之關聯規則數量，但該資料轉換的方法因過於主觀，是以可能轉換後之資料無法確實表達原資料所欲傳達的資訊，造成可能無法發掘蘊藏豐富資訊(informative)的關聯規則。

為解決上述問題，如圖 1 所示，本研究提出一個四階段的資料探勘作業程序，以從醫療資料庫中發掘出具有意義(relevant)的關聯規則。為使最後發掘之關聯規則蘊藏豐富資訊，且接近於人類使用的自然語言，我們首先使用叢集劃分(cluster partitioning)技術(Kaufman et al. 1990; Ng & Han 1994)，計算出資料表格中各數值型態資料欄位的主要中點值(medoid)，並依據各資料欄位所獲得的中點值，將數值型資料自動的轉換為對應口語化述辭(linguistic term)之模糊集合(fuzzy set)(Yager 1984; Yager 1988; Zemankova & Kandel 1985)。其次，以類神經網路之敏感度分析(sensitivity analysis)，以主要病徵為預測變數(predicate variable)，決定相對重要(relative importance)的變數，並以相對重要變數作為後續叢集分析之輸入變數。

其次為避免遺失支持度低、信賴度高，但可能有意義的關聯規則，我們於第二階段叢集分析，使用非監督式(unsupervised)的 Kohonen SOM 演算法，將疾病資料值依據輸入變數資料值的特徵，劃分為數個內部資料特徵相似的叢集，將原始資料劃分為數個叢集的目的，在於避免遺失支持度低但信賴度高的關聯規則。第三階段將 SOM 產生之叢集，分別以 Apriori 關聯規則演算法擷取各叢集的特徵檔案(cluster profile)，由於 Apriori 關聯規則演算法傾向於產生大量的關聯規則，我們提出一種以模糊相似關聯(fuzzy resemblance relation)為基礎的規則刪減(pruning)方法，以將重覆之關聯規則予以合併(merge)，合併後之關聯規則數量將大幅減少，且規則本身的語意更為清晰與精簡。在第四階段，我們另提出有別於「支持度/信賴度」機制(Agrawal et al. 1993)的方法，而以真實值(truth value) (Zadeh 1978; Hsieh 2004)評量技術，輔助判別所發掘關聯規則之可信度。

本研究其餘部分內容如後說明，第二節為本研究問題描述與問題的解決策略；第三節說明真實之膠原疾病資料庫內容，資料預處理以及資料轉換的方法；第四節說明以 SOM 叢集分析技術，解決可能遺失有意義關聯規則的方法；第五節說明叢集特徵檔案的建立，以及以相似關聯為基礎之規則合併方法；第六節提出一關聯規則真實值評估的方法，以輔助判別模糊關聯規則(fuzzy association rule)之可信度；第七節為本研究之結論，並提出未來的研究方向與建議。

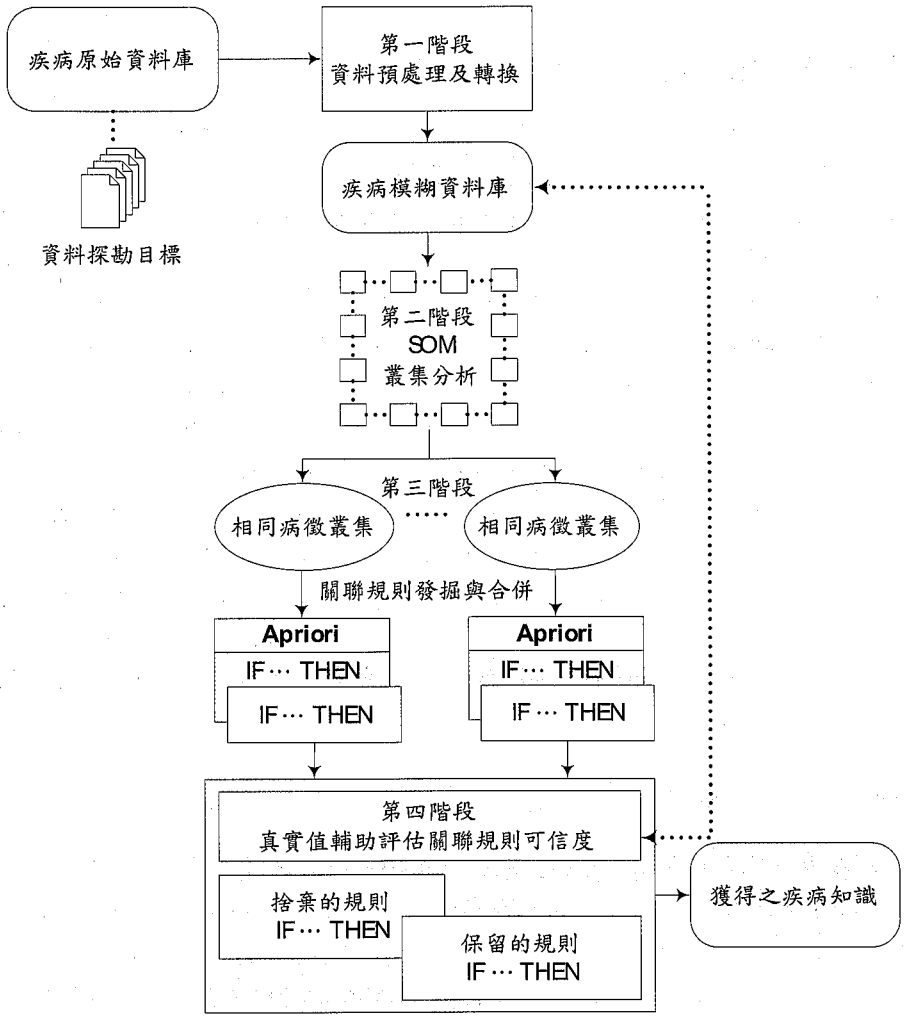



圖 1：本研究醫療資料探勘作業程序

# 貳、問題描述與解決策略

本研究所要探討的是有關於醫療方面之資料探實務應用，然而以現有關聯規則資料探勘技術於醫療資料庫應用上，存在以下問題。

## 一、蘊藏豐富資訊的規則 (Informative Rules)

相較於交易型(transactional)以及製造業(manufacturing)資料庫，醫療(medical)資料庫內容包含有大量的數值型實驗室檢驗數據，以及醫生之主觀診察判斷陳述句(judgment statement)，是以資料內部隱藏有更為豐富的知識且更不易以傳統之資料分析技術取得。資料探勘中之關聯規則分析技術，可由資料庫內擷取出規則型式(rule format)的知識，在人類一般交談與理解(reasoning)的情形下，相較於規則的本身包含有準確的語句(precise term)，不嚴密的語句(imprecise term)除了能表達出蘊藏豐富資訊的規則外，更適合於實務上應用所發掘之規則(Delgado et al. 2001)。

是以如何由醫療資料庫中擷取出，包含人類容易理解之不嚴密形式語句以及蘊藏豐富資訊之關聯規則，為另一相當重要的議題。不同於其他研究主觀的將數值型屬性轉換為模糊集合，在本研究中我們嘗試以叢集劃分技術中之 CLARANS 演算法(Ng & Han 1994)，計算出各數值型態資料欄位，所包含資料之數個資料中點值，並根據所計算出的資料中點值，自動的將各數值型態資料欄位所有的資料值，轉換為對應的數個模糊集合，其後再以轉換後之模糊資料庫，作為後續資料探勘工作之輸入資料，進行關聯規則的發掘，所發掘出關聯規則的型式將包含有人類容易理解的不嚴密語句且蘊藏豐富資訊。

## 二、關聯規則重複問題 (Redundant Rules)

如何發掘出有意義(relevant)且有用(usefulness)的關聯規則，在資料探勘實務應用上為一相當重要的議題，其原因在於運用如 Apriori 等關聯規則演算法，對於真實世界的資料庫進行資料探勘，往往會發掘出數以千計的關聯規則，而且在所發掘的關聯規則中包含有大量重複(redundancy)、無意義(irrelevant)且難以解釋的關聯規則。雖然一般可藉由視覺化的方法(Klemettinen et al. 1994)，觀察並篩選出具相對重要性的關聯規則，以解決上述問題，但仍需要更具效率的重複關聯規則刪減方法。

在模糊理論應用領域中，模糊近似關聯(similarity relation)、模糊鄰近關聯(proximity relation)以及模糊相似關聯(resemblance relation)等，皆可作為判斷模糊集合彼此間相似程度的方法(Zadeh 1978; Zadeh 1984; Zemankova & Kandel 1985)，由於模糊相似關聯直接計算數值間距離的設計，最為適合應用於數值類型之模糊集合。本研究將以模糊相似關聯為基礎，設計一重複關聯規則刪減的方法，原本重覆之關聯規則經刪減後，將合併為具相似真實值之最小數目關聯規則，且所有合併後的關聯規則將具

有最少之規則前項(antecedent)，與最大之規則後項(consequence)。

### 三、決定最引人興趣的關聯規則 (Interestingness Association Rules)

在傳統篩選引人興趣的關聯規則問題解決上，所發掘出的關聯規則必需滿足使用者訂定的最小支持度和信賴度門檻值時，此關聯規則才考慮採用。但於醫療資料庫中，某些疾病特徵具特殊性，亦即縱使少數資料亦可能隱含有用的關聯法則，若直接以完整醫療資料庫內容為目標進行資料探勘，將可能遺失某些有用的關聯規則，所以資料預先分類(pre-classification)為醫療資料探勘之一重要工作。在本研究中，我們以非監督式的 SOM 叢集演算法，預先依照醫療資料之特徵屬性值(feature attribute value)，將原始資料分類為內部資料特徵屬性相似的數個叢集，再進行後續資料探勘作業，以避免因所發掘關聯規則支持度過小而可能遺失的問題。

進一步探討以傳統關聯規則「支持度\信賴度」篩選機制，可能會產生無意義、遺失支持度低但信賴度高規則的問題。例如，若 $A \Rightarrow B$ 為一關聯規則，$A \Rightarrow B$ 之支持度表示為 $P(A,B)$，支持度在於說明同時支持 $A \cup B$ 屬性集合值組(tuple)之個數；$A \Rightarrow B$ 之信賴度表示為 $P(B|A)$，信賴度在於說明 $B$ 相對於 $A$ 之條件機率，亦可表示為 $P(A,B)/P(A)$。然而，若 $B$ 之出現與 $A$ 不直接相關，則依條件機率概念 $P(B)$可以忽略，但此舉將造成 $P(A,B)/P(A)$條件機率值與 $P(B)$相同，亦即縱使關聯規則之信賴度超過預設之門檻值，但有可能所篩選出為無意義之關聯規則。在模糊理論重要應用領域上，模糊關聯規則的產生程序同等於評估模糊口語化述辭(fuzzy linguistically term)的真實程度(truth degree)，在本研究中，我們提出另一種以真實值輔助評估關聯規則可信度，以篩選有用關聯規則的方法。

## 叁、醫療資料庫說明

### 一、原始膠原疾病資料庫

膠原疾病(collagen disease)(Levin et al. 1999; PKDD 2002; Taylor 1999; Zytkow & Gupta 2000)為一致死率甚高之自我免疫系統(auto-immune system)病變，若病患罹患該疾病將會遭受自我抗體(antibody)攻擊，遭受攻擊的人體器官(organ)將逐漸喪失原有功能進而致命，此類複雜的自我免疫系統病變有一明顯的病徵在於血栓(thrombosis)的形成，目前在醫療界的實務應用上，對於膠原疾病本身形成的原因以及與血栓病徵之間隱含的關係，所擁有的知識仍相當有限。本研究所使用之膠原疾病資料庫，取自於 PKDD1999 至 PKDD2002(PKDD 2002)國際知名研討會所公開之真實醫療資料庫，在研討會中已有十餘篇的學術論文對於此公開之膠原疾病資料庫提出研究成果，主要的研究成果分為以適當的資料探勘技術，分析膠原疾病資料庫屬性特徵值之間的規律性(regularity)，以及建立預測(predication)模型瞭解膠原疾病

與血栓病徵形成之間的關係。然而大多數醫療資料探勘的研究，普遍包含有第二節所述各項問題。在本研究中，我們提出一適用於醫療資料探勘的四階段作業程序，並嘗試以此真實膠原疾病醫療資料庫為例，以實例驗證方式說明本研究資料探勘的作法，與如何解決第二節所述的問題。現將本研究所使用之膠原疾病資料庫內容，摘要說明如後。

此原始膠原疾病資料庫共包含有七個資料表格，其中資料表格 PATIENT_INFO(共 1239 筆紀錄)儲存醫院門診病患之個人基本資料；資料表格 DIAGNOSTIC(共 1956 筆紀錄)所儲存的內容為，醫生對於每位病患確認或疑似罹患某種疾病之診斷結果，診斷結果除膠原疾病項目外亦包含有其他疾病項目；資料表格 ANTIBODY_EXAM(共 770 筆紀錄)所儲存內容為檢驗室對於特定病患，在血栓可能形成前或血栓形成期間，對於人體主要抗體等級、血液凝結程度所進行之各項實驗室檢驗數據，在此表格中 aCL_IgG, aCL_IgM, ANA, aCL_IgA 為血栓攻擊人體主要必須檢測之抗體；資料表格 THROMBOSIS(共 198 筆紀錄)所儲存內容為，血栓形成且攻擊人體期間，病患所顯現之病徵；資料表格 LAB_EXAM(共 57,542 筆紀錄)為醫院資訊系統(HIS)近 15 年內，儲存所有病患曾於醫院檢驗室所進行的各種檢驗數據，在此表格中所儲存之檢驗數據並不一定與血栓之形成有關；資料表格 DISEASE(共 65 筆紀錄)所儲存為醫生診斷時，判別所有可能疾病的名稱；資料表格 ANA_PATTERN(共 644 筆紀錄)儲存所有 ANA 抗體之類型。

膠原疾病發生主要原因在於血栓的形成，由於 ANTIBODY_EXAM 資料表格儲存的內容為，與血栓形成所有主要抗體之實驗室檢驗數據以及血栓形成嚴重的程度(0：尚未形成，1~3：非常嚴重、嚴重至輕微)，是以此資料表格內容與膠原疾病成因主要相關，絕大部分膠原疾病的研究(Jensen 2001; Levin et al. 1999)，亦僅以各種資料探勘技術分析此資料表格，以瞭解各種抗體數據與血栓形成嚴重程度之間的關係。然而，血栓形成與所有相關抗體程度彼此間的關連係為已知，單純分析此資料表格並無法有效由醫療資料庫中獲得額外的知識。LAB_EXAM 資料表格為病患除血栓抗體檢驗數據外，其他於醫院實驗室所進行過的各種檢驗數據，此類檢驗數據可能與血栓形成有關亦可能無關，由於該資料表格包含超過 15 年的病患檢驗歷史資料，所以在合併(join) ANTIBODY_EXAM 以及 LAB_EXAM 此二資料表格時，僅保留檢驗日期相距 3 年內之資料，以符合疾病與病徵彼此間時間近似關聯的特性。整體資料表格之類別圖(class diagram)如圖 2 所示。

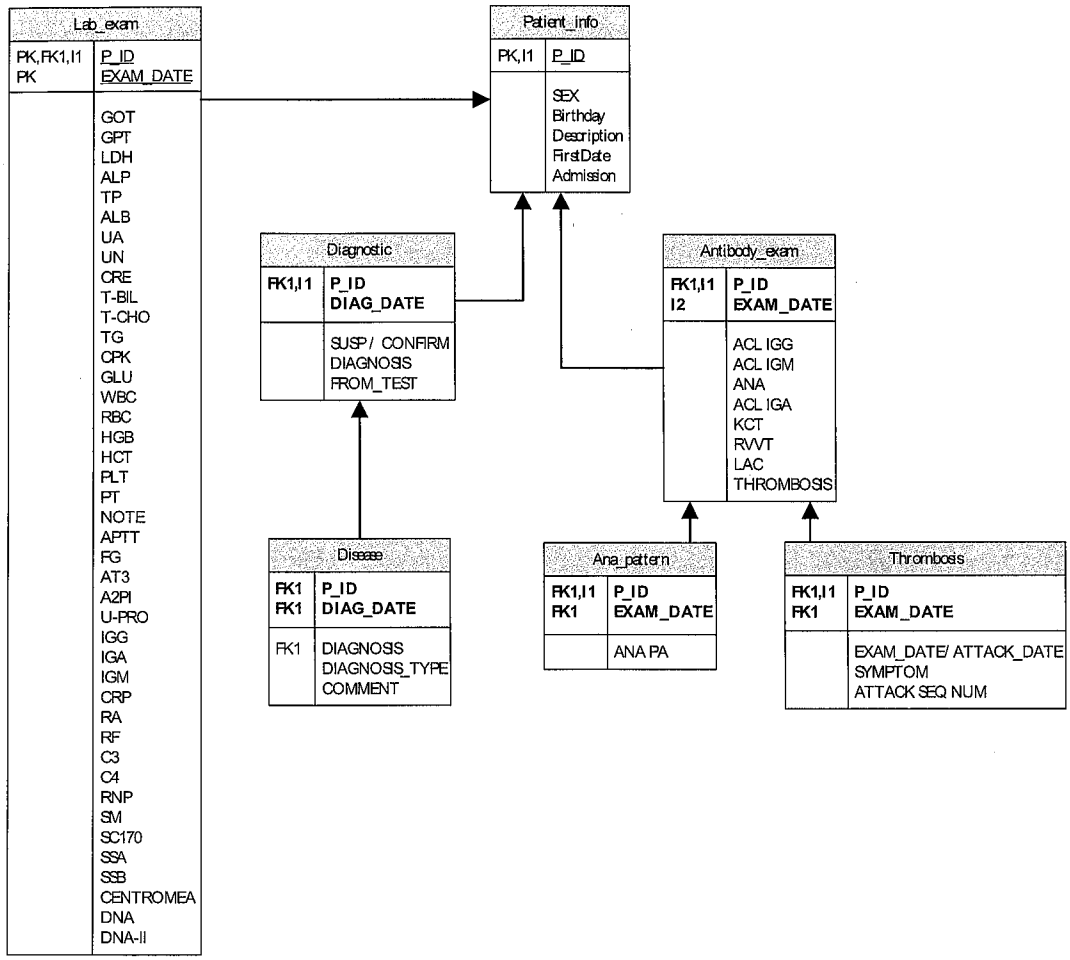



圖 2：膠原疾病資料庫

## 二、相對重要屬性分析

合併後之資料表格共包含有 65 個資料欄位，其中 10 個資料欄位為文字(text)或類別型(categorical)屬性，其餘 55 個資料欄為連續型(continuous)屬性，由於 DIAGNOSTIC 資料表格為醫師診斷病患時直接填寫之症狀描述，是以相對於膠原疾病之辨別率為百分之百，對於後續資料探勘僅能由其中獲得一般性(trivial)知識，並無太大幫助(PKDD 2002)，是以將相關資料欄位刪除。另刪除重複、遺漏值(missing value)過多以及非直接相關之資料欄位，共保留 36 個資料欄位進行後續分析。依據膠原疾病特徵我們先行假設，分析人體血液凝結程度以及多種抗體值將有助於瞭解該疾病的成因，以及將膠原疾病區分為不同的類別，此外性別、年齡以及因時間不同各檢驗數據之變化值，亦可能有助於分析瞭解疾病的形成。我們首先透過病人的病例號碼將各資料表格加以合併為單一資料表格，再對合併後之資料表格進行類神經類型之敏感度分析(sensitivity

analysis)，敏感度分析目的在於找出各項檢驗數據對於血栓凝結、以致於膠原疾病之影響程度，除藉以瞭解先前假設是否正確外，並可以相對重要性(relative importance)較高的變數，作為後續資料探勘主要輸入變數，我們並將各資料欄位值可能對血栓形成影響程度依序由高至低排列，類神經網路敏感度分析結果如表 1 所示。經多次實驗結果驗證，以選取相對重要程度值大於 0.00598 的變數，對於後續叢集分析有最好之分類結果。

表 1：類神經網路之敏感度分析

| Input Layer (no. of neurons) : | 5289 |
|---|---|
| Hidden Layer #1 (no. of neurons) : | 13 |
| Output Layer (no. of neurons) : | 3 |
| Predicted Accuracy : | 97.71% |
| Relative Importance to Thrombosis | |
| ALP | 0.03074 |
| GPT | 0.02747 |
| CRE | 0.02717 |
| UN | 0.02717 |
| UA | 0.02715 |
| TP | 0.02685 |
| LDH | 0.02550 |
| ALB | 0.02550 |
| CRP | 0.01455 |
| TG | 0.01438 |
| IGA | 0.01437 |
| PLT | 0.01433 |
| WBC | 0.01419 |
| IGG | 0.01415 |
| T-CHO | 0.01171 |
| GOT | 0.01169 |
| IGM | 0.01111 |
| GLU | 0.01077 |
| RF | 0.01043 |
| C3 | 0.00965 |
| C4 | 0.00930 |
| Year | 0.00921 |
| T-BIL | 0.00819 |
| RBC | 0.00794 |
| ACL IGA | 0.00751 |
| CPK | 0.00696 |
| HCT | 0.00688 |
| HGB | 0.00684 |
| ACL IGG | 0.00631 |
| ACL IGM | 0.00598 |
| U-PRO | 0.00361 |
| ANA | 0.00302 |
| LAC | 0.00252 |
| KCT | 0.00176 |
| SEX | 0.00171 |
| RVVT | 0.00137 |

## 三、資料轉換

在人類一般交談與理解(reasoning)的情形下，為使由醫療資料庫內擷取出的關聯規則，能表達出蘊藏豐富的資訊，以及符合於實務上應用，規則的本身包含有不嚴密的語句(imprecise term)相較於準確的語句(precise term)將更為適合。關聯規則資料探勘演算法較適合處理布林(Boolean)與類別(categorical)型態的資料欄位，若資料欄位為數值(quantitative)型態，將有發掘之關聯規則數量龐大且規則本身難以解釋的問題，由於膠原疾病資料庫本身大部分為數值型態的資料欄位，於資料探勘前必須先行處理輸入資料。例如在 Srikant 與 Agrawal 的研究中(Srikant & Agrawal 1996)，為了將數值型資料轉換為布林或類別型態之資料，首先將各資料欄位之屬性領域(attribute domain)切割為數個小的區間，其後將鄰近區間逐步合併為較大之區間，合併期間並保持關聯規則足夠之支持度，最後並以合併後區間取代原有數值資料。然而此種對數值資料尖銳區間切割(sharp boundary division)的作法，並不能有效處理鄰近區間數值之辨別，以及不合於人類對於資料直覺判斷涵義的問題。例如，直接以"20 至 28 歲之間"區隔年紀為年輕的做法，並不合於人類的直覺且區隔方式過於主觀。

模糊集合(fuzzy set)可以較為平順(smooth)方式處理鄰近區間之數值，也符合人類對於數值資料處理的直覺，若能將數值資料轉換為以口語化述辭(linguistic term)表示之模糊集合，將有助於產生蘊藏豐富資訊的規則(informative rule)、減少發掘關聯規則的數量與重複性規則的發生，然而於實務應用上如何事先決定合適的模糊集合十分困難，而且定義數值型態欄位合適的模糊集合，將直接影響後續資料探勘作業結果的品質。一般將數值資料轉換為模糊集合的作法，係由使用者或領域專家(domain expert)根據既有知識，定義模糊集合與對應的歸屬函數(membership function)，然而此種作法的缺點在於模糊集合與歸屬函數的定義會因人而異也不符合實務上應用，而且以主觀方式決定數值區間的切割，亦有可能不適用於後續資料探勘作業關聯規則的產生。若能以資料庫中，每個數值型態資料欄位所儲存資料值的分布為基礎，自動的產生模糊集合，應為一較適當的作法。

現說明模糊集合自動產生的方法，經合併與預先處理後之膠原疾病資料表格共包含 8,645 筆紀錄，我們計劃使用資料分析技術當中的群集劃分(cluster partitioning)演算法，事先找出每個數值型態資料欄位的主要中點值(medoid)，然後再以中點值為基礎，自動的產生每個數值型態資料欄位對應的模糊集合。K-medoids、CLARA 以及 CLARANS(Kaufman & Rousseeuw 1990; Ng & Han 1994)為三個主要的群集劃分演算法，其中 K-medoids 演算法並不適合處理數量大的資料集合，其餘二種演算法適合處理數量大的資料集合，其中 CLARA 以取樣(sampling)、CLARANS 以隨機搜尋(randomize search)的方式找尋中點值，CLARANS 相較於 CLARA 具有較好的執行效能與準確性，所以我們採用 CLARANS 演算法，作為計算數值型態資料欄位中點值的方法。

接下來，再根據 CLARANS 所計算出的 $k$ 個中點值，轉換為該資料欄位對應的 $k$ 個模糊集合(Fu et al. 1998)。例如，某個數值型資料欄位的屬性領域(attribute domain)

範圍由 $v_1$ 至 $v_2$，$\{m_1, m_2,\ldots,m_k\}$為 CLARANS 所計算出該資料欄位的 $k$ 個中點值，則 $(v_1\sim m_1), (m_1\sim m_2),\ldots,(m_k\sim v_2)$為 $k$ 個模糊集合的資料範圍。模糊集合歸屬函數的轉換方式為，資料範圍$(v_1\sim m_1)$，則對應第 1 個模糊集合之歸屬函數為：

$$f_{m_1}(x)=\begin{cases}1 & \text{if } x\le m_1\\ x-m_1/m_1-m_2 & \text{if } m_1<x<m_2\\ 0 & \text{if } x\ge m_2\end{cases},$$

資料範圍$(m_{p-1}\sim m_p)$，$p = 2,\ldots, k\text{-}1$，則對應第 $p$ 個模糊集合之歸屬函數為：

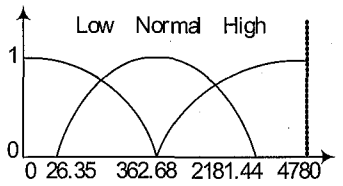
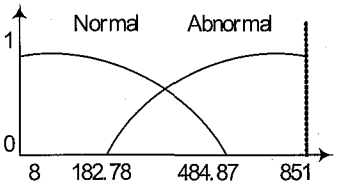
$$f_{m_p}(x)=\begin{cases}0 & \text{if } x\le m_1\\ x-m_{p-1}/m_p-m_{p-1} & \text{if } m_{p-1}<x<m_p\\ 1.0 & \text{if } x=m_2\\ x-m_{p+1}/m_p-m_{p+1} & \text{if } m_p<x<m_{p+1}\\ 0 & \text{if } x\ge m_{p+1}\end{cases},$$

資料範圍$(m_k\sim v_2)$，則對應第 $k$ 個模糊集合之歸屬函數為：

$$f_{m_k}(x)=\begin{cases}0 & \text{if } x\le m_k\\ x-m_k/v_2-m_k & \text{if } m_k<x<v_2\\ 1 & \text{if } x\ge v_2\end{cases}。$$

以抗體 GPT、T-CHO 兩資料欄位為例，GPT 數值型態資料欄位之資料值範圍由 0 至 4780，我們以 CLARANS 演算法計算出 3 個中點值，其值分別為{26.35, 362.28, 2181.44}，則我們可以定義對應於 GPT 欄位三個口語化述辭{Low, Normal, High}之模糊集合；同樣的，T-CHO 數值型態資料欄位之資料值範圍由 8 至 851，我們以 CLARANS 演算法計算出 2 個中點值，其值分別為{182.78, 484.87}，則我們可以定義對應於 T-CHO 欄位二個口語化述辭{Normal, Abnormal}之模糊集合，其後依前述模糊集合界定之範圍，將數值型資料轉換成口語化述辭之模糊集合，結果如表 2 所示。

表 2：GPT 與 T-CHO 之模糊集合

| 欄位 | 中點值 | 模糊集合數值範圍 | 模糊集合之歸屬函數 | 轉換之範例 |
|---|---|---|---|---|
| GPT | 26.35 | Low: 0~362.28 |  | 350 ⇒ (Low,0.03)<br>945 ⇒ (Normal,0.67)<br>2493 ⇒ (High,1) |
| | 362.28 | Normal: 26.35~2181.44 | | |
| | 2181.44 | High: 362.28~4780 | | |
| T-CHO | 182.78 | Normal: 8~484.87 |  | 204 ⇒ (Normal,0.92)<br>512 ⇒ (Abnormal,1) |
| | 484.87 | Abnormal: 182.78~851 | | |

# 肆、自我組織映射圖網路叢集分析

合併後之膠原疾病資料表格包含有病患之血栓抗體特殊檢驗數據，鄰近該血栓檢驗三年內其他非血栓檢驗紀錄，以及病患如年齡、性別等個人特徵資料。在醫療上的經驗，膠原疾病與血栓形成有絕對的關係，但血栓相關抗體一般在醫生高度懷疑病患罹患膠原疾病時才會進行檢驗，所以僅針對血栓抗體檢驗資料表格進行分析，只能獲得各檢驗項目數據值與血栓嚴重程度間的關係，以及間接瞭解造成膠原疾病的因素。如能進一步分析除血栓抗體檢驗資料外其他造成膠原疾病的原因，將有助於瞭解疾病本身以及各項數據與膠原疾病間的關聯性。由於輸入之膠原疾病為高維度(multi-dimensional)的資料表格，可能無法有效率與直接的由其中萃取出有意義的關聯規則。所以在本研究中，我們將資料轉換後之資料表格，事先依輸入資料的特徵加以分類(classification)，並將輸入資料歸納至數個本身資料特徵相似的叢集(cluster)，如此每個叢集內的資料將具有同質性(homogenous)，其後再對各叢集進行分析以瞭解與血栓嚴重程度間的關係，如此由各叢集中萃取出的關聯規則，相較於直接分析原始資料更具意義(Markey et al. 2003)，亦不會有遺失支持度低但信賴度高關聯規則的問題。至於資料的分類，我們採用適合處理高維度資料表格的自我組織映射圖網路(Self-Organizing Maps，SOM)演算法。

SOM 係由 Kohonen (1995)於 1980 年所提出的一種非監督式學習(unsupervised learning)類神經網路模型，SOM 可藉由資料特徵的不同，將性質類似的資料歸類為相同的叢集。在一般應用上，SOM 類神經網路可安排為一二維的矩陣圖形，其中神經單元與輸入向量連結並於學習過程中自動調整觸動權重(synaptic weight)，SOM 學習過程

包含有兩個主要階段，第一階段為粗略評估階段(rough estimation phase)，目的在於取得整體資料樣式(gross data pattern)，第二階段為細部調整階段(fine tuning phase)，目的在於調整 SOM 內部神經單元以獲得更佳的資料描述能力。由於單一隱藏層(hidden layer)的類神經網路，足以描述複雜的系統與提供期望的準確度(Cybenko 1989; Hornik et al. 1989)，所以我們所使用之類神經網路僅包含單一隱藏層。在 SOM 學習階段中，Euclidean 距離係用於計算資料樣式與各神經單元彼此間之相似性，在本研究中，經資料預先處理後資料表格之資料欄位均設定為 SOM 之輸入變數，一但資料開始輸入，SOM 便開始計算資料的輸入向量與各神經單元之距離，距離愈近者，代表其相似程度愈高。因此，在比較所有神經單元與輸入資料的距離後，距離最短者為優勝者，並成為該神經單元中最具代表性的資料。接著 SOM 會根據所有神經單元距離優勝單元者與鄰近神經單元的距離關係，分別對其處理單元的權數做調整。鄰近距離愈大，該鄰近係數則愈小，而權數的調整程度也就比較小，反之亦然。直到所有膠原疾病資料均依上述方式調整完畢，即稱為一個學習循環；每執行一個學習循環鄰近半徑收縮一次，直到各權數均已達穩定狀態為止。

合併後膠原疾病資料表格，我們先將無血栓病徵(Thrombosis=0)之資料刪除，共保留 1,962 筆資料，其後依照敏感度分析選取前 30 個相對重要變數，並以 SOM 進行叢集分析，分析的結果如圖 3，顯示為一 4×4 之二維類神經網路單元圖形，其中包含有 16 個類神經單元的叢集，每個類神經單元中並顯示有所佔資料的筆數、相對於全部資料的比例、以及對應於各血栓嚴重程度所佔的比例。值得注意的是，類神經細胞編號 3~4、8~9、13，所組成的 5 個膠原疾病資料叢集中，血栓嚴重程度皆趨向於 T(Thrombosis)=1，因此可將這些類神經細胞歸屬於「非常嚴重」血栓病徵之病例資料；而類神經細胞編號 12、15~16 中，所組成的 3 個膠原疾病資料叢集中，血栓嚴重程度皆趨向於 T(Thrombosis)=2，因此可將這些類神經細胞叢集「嚴重」病患之病例資料。至於血栓嚴重程度皆趨向於「輕微」T(Thrombosis)=3 者，由於為輕微病徵且資料佔全部資料比例極小，是以並不將作進一步分析。

根據叢集分析概念，具相似屬性特徵值(shared feature attributes)之資料將歸屬於同一類神經單元之叢集，所以將原始資料表格先經 SOM 分類，再將屬性特徵值相似之叢集分別輸出，建立個別的叢集特徵檔案(cluster feature profile)，相較於傳統以全部原始資料進行分析的方法，會獲得更為有用的分析結果。

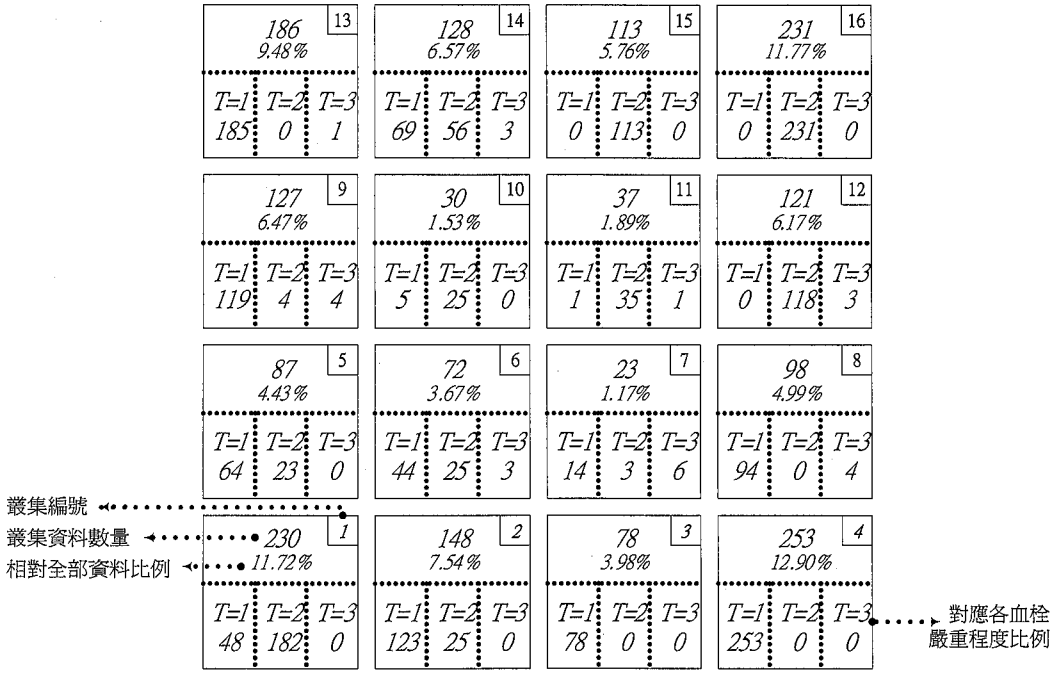



圖 3:自我組織映射圖叢集分析結果

# 伍、模糊關聯規則之發掘與合併

## 一、模糊關聯規則發掘

資料探勘技術中，由資料表格中發掘關聯規則為重要的研究領域，而關聯規則目的在於描述資料表格屬性間另人感到興趣的關係，目前已有相當多的研究提出適合於布林(Boolean)或類別(Categorical)型態資料欄位的關聯規則演算法(Agrawal et al. 1993; Srikant & Agrawal 1994; Srikant & Agrawal 1995)，另有以雜湊(hash)函數、二元樹(binary tree)(Park et al. 1995)輔助加速探勘效率的關聯規則演算法。然而上述演算法並不能直接處理數值(quantitative)類型的資料，Strikant 等人(Srikant & Agrawal 1996)提出將數值型態資料的領域範圍(domain range)，分割為數個區間(interval)的方法，以發掘數值類型之關聯規則，然而以此種方式發掘出之關聯規則在規則型式上並不簡潔，而且規則本身並不容易解釋。不以區間方式切割領域範圍，我們以第三節所述之資料轉換方式將數值型態資料對應為以口語化述辭所表示的模糊集合，以口語化述辭的方式描述關聯規則，規則的本身將更為自然也更易於應用，由於所發掘出的關聯規則包

含有口語化述辭，所以在本研究中我們稱其為"模糊關聯規則"(fuzzy association rule)(Chen & Wei 2002; Delgado et al. 2001; Frawley et al. 1991; Fu et al. 1998; Mitra 2002)。

如何由各叢集萃取出膠原疾病之特徵檔案為本階段主要工作。經由 SOM 叢集分析後，原始膠原疾病資料依據資料特徵相似程度的不同，區分為數個不同的叢集，我們選取血栓嚴重程度顯著趨向於「重度嚴重」(T=1)與「嚴重」(T=2)的叢集進行關聯規則分析，此種以嚴重程度顯著趨向選擇策略的理由在於，同一叢集內所包含的資料將具有相似的特徵。在許多關聯規則演算法中，由於 Apriori 演算法在關聯規則產生過程中使用上具較大彈性，所以我們選擇該演算法作為關聯規則之資料探勘工具。

表 3：叢集 12 發掘之部分模糊關聯規則

| | |
|---|---|
| 規則 1 | PLT = (LOW, 0.11) ⇒ Thrombosis == 2 |
| 規則 2 | PLT = (LOW, 0.27) ⇒ Thrombosis == 2 |
| 規則 3 | PLT = (LOW, 0.32) ⇒ Thrombosis == 2 |
| 規則 4 | PLT = (LOW, 0.83) ⇒ Thrombosis == 2 |
| 規則 5 | ALP = (LOW,1) & GPT = (LOW,0.84) & HCT = (LOW,0.60) ⇒ Thrombosis == 2 |
| 規則 6 | ALP = (LOW,1) & GPT = (LOW,0.74) & HCT = (LOW,0.27) ⇒ Thrombosis == 2 |
| 規則 7 | ALP = (LOW,1) & GPT = (LOW,0.84) & HCT = (LOW,0.60) ⇒ Thrombosis == 2 |
| 規則 8 | ALP = (LOW,0.95) & GPT = (LOW,0.84) & HCT = (LOW,0.73) ⇒ Thrombosis == 2 |
| 規則 9 | TG = (LOW,1) & RBC =(LOW,1)& PLT =(LOW,1) & IGA = (LOW,1)& T-CHO = (Abnormal,1) ⇒ Thrombosis == 2 |
| 規則 10 | RBC =(LOW,1)& PLT =(LOW,0.98) & IGA = (LOW,1) ⇒ Thrombosis == 2 |
| 規則 11 | RBC =(LOW,1)& PLT =(LOW,0.98) & IGA = (LOW,1) & T-CHO = (Abnormal,1) ⇒ Thrombosis == 2 |
| 規則 12 | TG = (LOW,0.99) & RBC =(LOW,1)& PLT =(LOW,1) & IGA = (LOW,1) ⇒ Thrombosis == 2 |
| 規則 13 | TG = (LOW,1) & RBC =(LOW,1) ⇒ Thrombosis == 2 |
| 規則 14 | RBC =(LOW,1)& PLT =(LOW,0.98) & IGA = (LOW,1) & CPK = (NORMAL,1) ⇒ Thrombosis == 2 |
| 規則 15 | RBC =(LOW,0.1)& PLT =(LOW,0.98) & IGA = (LOW,0.75) & CPK = (NORMAL,1) ⇒ Thrombosis == 2 |
| 規則 16 | RBC =(LOW,0.84)& PLT =(LOW,0.98) & IGA = (LOW,0.75) & CPK = (NORMAL,1) ⇒ Thrombosis == 3 |
| 規則 17 | RBC =(LOW,0.84)& PLT =(LOW,0.98) ⇒ Thrombosis == 3 |
| 規則 19 | UA = (NORMAL,0.16) & TG = (High,0.87) ⇒ Thrombosis == 2 |
| 規則 20 | UA = (NORMAL,0.17) & TG = (High,0.87) ⇒ Thrombosis == 2 |
| 規則 21 | UA = (NORMAL,0.22) & TG = (High,0.95) ⇒ Thrombosis == 2 |
| 規則 22 | UA = (NORMAL,0.26) & TG = (High,0.95) ⇒ Thrombosis == 2 |
| .... | |

為簡單說明起見，我們僅以編號 12 叢集解釋關聯規則之處理方法，Apriori 演算法關聯規則篩選參數，設定為 90%以上的信賴度以及 3%資料筆數的支持度，最初共產生 756 條滿足條件之關聯規則，表 3 所列為部份的模糊關聯規則，為決定出「最引人興趣的關聯規則」，所以我們設定較低之支持度，並以下一節所說明之真實值輔助評估關聯規則之可信度。現若假設某一條模糊關聯式規則型式為「ALP=(LOW,1) & GPT=(LOW,0.84) & HCT=(LOW,0.6) ⇒ Thrombosis = 2」，則此關聯規則表示當病徵 ALP 為(LOW,1)、GPT 為(LOW,0.84)且 HCT 為(LOW,0.6)時，則病患會有嚴重程度之血栓症狀，此種模糊關聯規則的型式在語意上，相較於數值關聯規則清晰，可提供醫療人員診斷時適當的協助。

## 二、模糊關聯規則合併

除規則本身不易解釋與應用外，關聯規則演算法另一問題在於產生大量的重複性規則，是以如何找出有意義(relevant)與有用(usefulness)的關聯規則，為另一個必須解決的問題。將數值資料轉換為口語化述辭之模糊集合，可使規則本身具有意義且可藉由屬性領域(attribute domain)數量(cardinality)的縮減，間接減少重複的關聯規則。其他如 Klemettinen et al. (1994)採用以視覺化的方式，簡化發掘關聯規則的數量；在以統計為評量方式的相關研究中，另有以 correlation (Brin et al. 1997)取代傳統信賴度的方法，在 Heckerman (1996)的研究中提出 derivation 評量，藉由關聯規則之支持度與信賴度彼此間之距離，以決定關聯規則是否重複；在以規則型式的相關研究中，一般型式關聯規則(generalized association rules)(Chen & Wei 2002)係以事先定義資料庫屬性間之 taxonomy 關聯結構，並於探勘過程中應用該結構並將隸屬於同一 taxonomy 層級之項目加以合併，藉以解決關聯規則重複問題；Bayardo & Agrawal (1999)所提出之 A-maximal rules 規則的概念則在於保留具最大前項(antecedent)之關聯規則作為規則合併之基礎，Bastide et al. (2000)則提出以 Galois connection 的概念定義頻繁且接近之項目集合(frequent closed itemsets)，逐步合併語意重複之關聯規則。

由於本研究所發掘知識為模糊關聯規則的型式，上述方法並不適合直接用於模糊關聯規則的簡化，所以本研究提出一種依「模糊相似關聯」(fuzzy resemblance relation)(Hsieh 2004; Zemankova & Kandel 1985)，計算關聯規則間相似程度，並將語義相似之模糊關聯規則加以合併(merge)的方法，在第六節中本研究並將提出適合評估模糊關聯規則品質的方法。為說明重複模糊關聯規則之處理方法，我們由表 3 編號 12 叢集之模糊關聯規則中，擷取 10、11 兩條模糊關聯規則作為範例說明。很明顯的，由於規則 10 相對於規則 11 而言，規則 10 具有數目較少的前項(antecedent)與數目較高(或相同)的後項(consequent)，該條規則無法提供使用者額外的資訊，是以規則 10 在語義上為重複的，為保留語義清晰與刪除無意義、重複的關聯規則，最後只有規則 11 必須加以保留，重複關聯規則合併之處理如下說明。

**定義一**：假設 $A_R = \{A_1,\ldots,A_n\}$為一資料欄位之屬性集合，且$\alpha_i$，$i = 1,\ldots,n$ 為相對於各屬性之門檻值(threshold value)。現若 $A_P$、$A_Q$，且 $A_P \subseteq A_R$、$A_Q \subseteq A_R$為描述二模糊關聯規則"前項"(antecedent) 或"後項"(consequent)之屬性集合，且 $x_1$ 與 $x_2$ 為描述任意兩條模糊關聯規則之資料庫值組(tuple)，則此模糊關聯規則之前項(後項)稱為"$\alpha$-近似" (表示 $A_P \approx A_Q$ )若且唯若

$E(x_1[A_i], x_2[A_i])$ = true, and

$\mu_{EQ}(\mu(x_1[A_i]), \mu(x_2[A_i])) \geq \alpha_i$, for all $A_i \in A_P$ and $A_Q$,

其中$\mu(x_1[A_i])$表示值組 $x_1$ 於屬性 $A_i$ 之歸屬函數值(membership value)，$E$ 為一同等述詞(equality predicate)，模糊相似關聯 $EQ$ 可以定義為

$$\mu_{EQ}(x,y) = \begin{cases} 0 & x \neq y \\ (1 - abs(\mu(x) - \mu(y))) & x = y \end{cases} 。$$

模糊關聯規則之前項或後項若為"$\alpha$-近似"，表示其語意近似，所以可以進一步合併。本研究主要目的在於保留有用、有意義但非重複之模糊關聯規則，一條模糊關聯規則重複的條件為存在其他具有相同型式，且模糊關聯規則之前項或後項為$\alpha$-近似之模糊關聯規則。

**定義二**：假設一模糊關聯規則 $r: A \Rightarrow B$ 為非重複之模糊關聯規則，若且唯若，不存在另一條模糊關聯規則 $r': A' \Rightarrow B'$，$A' \subseteq A$、$B \subseteq B'$且 $A \approx A'$、$B \approx B'$。

根據定義一與定義二，所有發掘之模糊關聯規則，將歸屬於同類之$\alpha$-近似「相似類別」(resemblance class)內，而歸屬於同一相似類別內之所有模糊關聯規則，可進一步合併為單一之模糊關聯規則，合併後之模糊關聯規則，將具有數量最多之前項與數量最少之後項，且合併後之規則語義更為清晰並易於應用。在實際應用時，可將叢集趨向於血栓病徵「非常嚴重」之 3~4、8~9、13 等 5 個叢集，以及叢集趨向於血栓病徵「嚴重」之 12、15~16 等 3 個叢集，所產生之模糊關聯規則特徵檔案分別匯整，再進行規則之合併作業，藉此以獲得更完整與精簡之知識，以及綜合分析處理叢集間特徵檔案之可能隱含關係。

現以叢集 12 產生之模糊關聯規則特徵檔案，說明合併的方法。根據模糊關聯規則之型式，合併型式可分為五類，在表 3 中模糊關聯規則經合併後結果如表 4 所示(假設各屬性$\alpha$-近似之容忍值均設定為 0.5)，合併可能型式如下說明。

(1)模糊關聯規則之前項、後項型式相同，且二模糊關聯規則為$\alpha$-近似趨近，則可合併。合併後，模糊關聯規則之歸屬函數值，取原來二規則中模糊集合歸屬函數之最小值。例如表 3 規則 1、2、3，其前項、後項型式皆相同，且模糊關聯規則為$\alpha$-近似趨近，因此可合併為表 4 之規則 1。

(2)模糊關聯規則之前項、後項型式相同，但歸屬函數值不合於容忍值，則不可合併。例如表 3 規則 1、2、3 與規則 4 之前項、後項型式皆相同，但 PLT α-近似差距值，μEQ(0.83,0.30) < 0.5，因此規則 4 不可與規則 1、2、3 合併。

(3)模糊關聯規則的前項型式不同、後項型式相同，但一規則之前項包含在另一規則之前項內，且歸屬函數值均合於容忍值，則可加以合併，合併後保留前項較長之模糊關聯規則。例如表 3 中規則 9、10、11、12、13 其後項型式相同，雖規則之前項型式均不相同，但規則 10、11、12、13 的前項包含於規則 9 之前項內，且其α-近似之差距值，均合於容忍值，因此 10、11、12、13 四條規則可合併至規則 9，合併結果如表 4 規則 4 所示。

(4)模糊關聯規則前項型式相同、後項型式不同，則可加以合併；合併後，後項成為"互斥或"(inclusive-or)型式之模糊關聯規則。例如表 3 規則 14、16，雖然後項型式相同，但 Thrombosis 屬性值分別為 2 及 3 並不相同，但前項型式相同且 $\alpha$-近似差距值皆合容忍值，此種規則合併方式，需將規則後項合併為「or」型式之規則；合併後結果如表 4 規則 5 所示。

表 4：叢集 12 中合併後之部分模糊關聯規則

| | |
|---|---|
| 規則 1 | PLT = (LOW, 0.11) ⇒ Thrombosis == 2 |
| 規則 2 | PLT = (LOW, 0.83) ⇒ Thrombosis == 2 |
| 規則 3 | ALP = (LOW, 0.95) & GPT = (LOW,0.74) & HCT = (LOW,0.27) ⇒ Thrombosis == 2 |
| 規則 4 | TG = (LOW,0.99) & RBC =(LOW,1) & PLT =(LOW,0.98) & IGA = (LOW,1) & T-CHO = (Abnormal,1) ⇒ Thrombosis == 2 |
| 規則 5 | RBC =(LOW,0.84) & PLT =(LOW,0.98) & IGA = (LOW,0.75) & CPK = (NORMAL,1) ⇒ Thrombosis == (2 ∨ 3) |
| 規則 6 | UA = (NORMAL,0.16) & TG = (High,0.87) ⇒ Thrombosis == 2 |
| .... | |

# 陸、以真實值評估模糊關聯規則真實度

在本節中，我們提出一種以真實值評量模糊關聯規則，以決定最引人興趣關聯規則(interestingness association rules)的方法，此方法可輔助解決傳統關聯規則支持度\信賴度篩選機制，可能會產生無意義、遺失支持度低但信賴度高規則的問題。真實值(truth value)為模糊資料庫領域中評量(evaluate)資料庫查詢的主要概念，在傳統資料庫中，因為資料欄位之屬性值均為單一且明確之資料值(single precise value)，所以任何值組(tuple)對應於資料庫查詢的真實值，只有「確實為真」(1, definite true)以及「確實為假」(0, definite false)兩種可能。在模糊資料庫中(Zadeh 1978)，資料庫查詢之真實值可以 $\Pi(Truth(t)) = F$ 之機率分配(possibility distribution)來表示，其中Π表示機率分配函數，$Truth(t)$為值組 $t$ 之真實值，$F$ 則為用以表示真實值語意之模糊集合。

以模糊邏輯(fuzzy logic)為基礎的計算方式，可提供更為複雜的如"大多數(most)"、"非常(very)"等口語化量辭述句(linguistically quantified proposition)真實值的

解釋與評估。根據 Yager (1984 & 1988)所陳述，一個語言摘要(linguistic summary)係為一口語化量辭述句，用以描述該敘述句，相對於一模糊資料庫(fuzzy database)中部份值組真實值的方法，此方法非常適用於由資料庫中發掘未知的知識。其次，對於口語化量辭述句真實值評量的方法，Zadeh (1984)基於模糊邏輯的基本理論，提出一種以計算方式評估口語化量辭述句真實值的方法；Yager (1984 & 1988)則提出以 OWA 及競爭族聚(competitive aggregative)概念為基礎的另一種真實值評量方法。兩種方法之間主要的差異在於，Zadeh 的方法在於表達口語化量辭述句之總真實值；而 Yager 的方法則關著於個別模糊敘述句真實值之呈現。

由於 Zadeh 所提出之真實值評估方法，在概念上較直覺且計算方式簡單(Heckerman 1996)，因此本研究採用 Zadeh 的計算方式，來決定口語化量辭述句相對於模糊資料庫之真實度。Zadeh 口語化量辭述句型式可表示為：

$Q$ $Y$'s are $F$,

其中 $Y$ 為一模糊集合(fuzzy set)，$F$ 為 $Y$ 之模糊子集合，$Q$ 為一口語化量詞。該述句真實值的評估，在於決定模糊集合 $Y$ 中元素(element)之歸屬函數值滿足模糊子集合 $F$ 之程度。例如，「大多數($Q$)的病患若其血液凝結情況為嚴重($Y$)，則其血液檢驗之 UA 值為(NORMAL,0.16)而且 TG 值為(High,0.87)($F$)」為模糊關聯規則 UA = (NORMAL,0.16) & TG = (High,0.87) ⇒ Thrombosis == 2，相對於模糊資料庫之口語化量辭述句。

如 3.3 節資料轉換說明，原始之膠原疾病資料庫，將轉換為以下型式之模糊關聯式資料庫(fuzzy relational database)。

**定義三** (Hsieh 2004)：模糊關聯資料庫綱要(fuzzy relational database schema)為一有序之屬性串列 R(A1,...,An,μr)，若 r 為相對於 R 之模糊關聯(fuzzy relation)，則模糊關聯可定義為：

$$r = \{(t,\mu_r(t)) \mid t \in Dom(A_1) \times \ldots \times Dom(A_n) \wedge \mu_r(t) \in [0,1]$$

$$\wedge\ \mu_r(t) = min(\mu(t[A_1]), \ldots, \mu(t[A_n]))\ \},$$

其中 $Dom(A_i)$為屬性 $A_i$ 之屬性領域，$\mu(t[A_i])$為值組$(t,\mu_r(t))$對應於屬性 $A_i$ 之歸屬函數值，$\mu_r(t)$則為值組$(t,\mu_r(t))$之歸屬函數值。□

**定義四**(Hsieh 2004)：假設"Q{t1,...,tn} are F"為一口語化量辭述句，{t1,...,tn}為模糊資料庫中之值組，則該述句真實值計算方式為：

$$Truth(Q\{t_1,\ldots,t_n\} \text{ are } F) = \mu_Q\left(\tfrac{1}{n}\sum_{i=1}^{n}\mu_F(t_i)\right),$$

其中 $\mu_F(t_i)$為 $F$ 對應於$(t_i,\mu(t_i))$之個別真實值(individual truth value)，$\frac{1}{n}\sum_{i=1}^{n}\mu_F(t_i)$ 為 $F$ 對應於模糊資料庫之整體平均真實值(average truth value)；當模糊集合 $F = F_1 \vee \ldots \vee$

$F_m$，則$\mu_F(t_i) = \max_{j=1}^{m}(\mu_{F_j}(t_i))$，當模糊集合 $F = F_1 \wedge \ldots \wedge F_m$，則$\mu_F(t_i) = \min_{j=1}^{m}(\mu_{F_j}(t_i))$；$Q$ 為使用者定義評估真實程度之模糊集合，例如"大多數(most)"之模糊集合可定義為：

$$\mu_{most}(x) = \begin{cases} 1 & \text{for} \quad x \geq 0.85 \\ 2x - 0.7 & \text{for} \quad 0.35<x<0.85 \\ 0 & \text{for} \quad x \leq 0.35 \end{cases}。$$

現說明如何應用真實值評估模糊關聯規則，以由各叢集所發掘之規則集合中，決定出最引人興趣的關聯規則。現以表 4 中規則 4 為範例說明真實值計算方式，首先由模糊資料庫中選擇出所有隸屬於編號 12 叢集的資料，共計 121 筆資料(部分資料如表 5 所示)，其次再以口語化量辭述句真實值計算的方式，計算規則 4 與所有叢集內資料之真實值，步驟如下：

(1)將規則 4 與編號 12 叢集內資料，計算出所有 121 筆資料之個別真實值。例如，紀錄編號 1 之個別真實值為 min(min(0.85,0.99), min(1,1), min(0.8,0.98), min(1,0.6), min(0.95,1))=0.6; 第 2 筆資料中，由於模糊集合 TG=(Normal,0.0)可對應為 TG=(Low,0.8)，是以紀錄編號 2 之個別真實值為 0.63。

(2)計算模糊集合 $F$，對應於模糊資料庫之整體平均真實值。假設關聯規則篩選容忍值為 0.60，根據定義四方式計算，最後獲得規則 4 整體平均真實值為 0.67，由於此規則真實值大於容忍值，是以此規則之可信度合於接受範圍。

(3)計算口語化量辭述句之真實度。假設"大多數"(most)模糊集合之歸屬函數如定義四中所述，則規則 4 口語化量辭述句之真實度為$\mu_{most}(0.67)=0.64$，所以我們可以獲得規則 4「大多數(*most*)的病患若其血液凝結情況為嚴重($Y$)，則其血液檢驗之 TG=(LOW,0.99)且 RBC=(LOW,1)且 PLT=(LOW,0.98)且 IGA=(LOW,1)且 T-CHO=(Abnormal,1) ($F$)」真實程度為 0.64 之結論。

表 5：模糊資料庫部份內容

| No. | | IGA | T-CHO | IGM | PLT | TG | RBC | | HGB | SOM-X | SOM-Y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | … | (Low,0.60) | (Abnormal,0.95) | (Normal,0.22) | (Low,0.80) | (Low,0.85) | (Low,1) | … | (Low,1) | 3 | 2 |
| 2 | … | (Low,0.76) | (Abnormal,0.64) | (Normal,0.33) | (Low,0.63) | (Normal,0.00) | (Low,1) | … | (Low,1) | 3 | 2 |
| 3 | … | (Normal,0.47) | (Abormal,0.96) | (Normal,0.08) | (Low,0.99) | (Low,0.31) | (Low,1) | … | (Low,1) | 3 | 2 |
| 4 | … | (Normal,0.10) | | (Normal,0.69) | (Normal,0.18) | | (Normal,0.00) | … | (Low,1) | 3 | 2 |
| 5 | … | (Normal,0.48) | (Abnormal,0.72) | (Normal,0.39) | (Low,0.34) | (Low,0.23) | (Low,1) | … | (Low,1) | 3 | 2 |
| 6 | … | (Low,0.35) | (Abormal,0.87) | (Normal,0.00) | (Normal,0.54) | (Low,0.39) | (Low,1) | … | (Low,1) | 3 | 2 |
| 7 | … | (Normal,0.10) | (Normal,0.60) | (Low,1) | (Low,0.46) | (Low,0.60) | (Low,1) | … | (Low,1) | 3 | 2 |
| 8 | … | (Normal,0.34) | (Normal,0.54) | (Low,1) | (Normal,0.64) | (Low,0.55) | (Low,1) | … | (Low,1) | 3 | 2 |
| 9 | … | (Low,0.97) | (Abormal,0.64) | (Normal,0.11) | (Low,0.71) | (Low,0.83) | (Normal,0.04) | … | (Low,1) | 3 | 2 |
| 10 | … | (Low,1) | (Abormal,1) | (Low,1) | (Low,0.80) | (Low,0.60) | (Low,1) | … | (Low,1) | 3 | 2 |
| 11 | … | (Low,0.67) | (Normal,0.31) | (Low,1) | (Low,1) | (Low,0.34) | (Low,1) | … | (Low,1) | 3 | 2 |
| 12 | … | (Low,1) | (Abormal,0.68) | (Low,1) | (Low,0.77) | (Normal,0.29) | (Low,1) | … | (Low,1) | 3 | 2 |
| … | … | … | … | … | … | … | … | … | … | … | … |

# 柒、結論

因應資訊科技的進步，如何從龐大的資料中擷取出有價值的資訊或知識，已成為重要的議題。目前資料探勘已經應用到許多領域，然而資料探勘在醫療資訊上卻少見相關研究，本研究以真實膠原疾病資料庫為例，試圖應用資料探勘中之叢集分析與關聯式規則技術，尋找尚未明確發現不為人知且令人感興趣之膠原疾病要素，從而歸納其間發生的必然性，則可讓醫療人員診斷時給予協助，進而提昇醫療品質。本研究提出適用於醫療資料庫新的資料探勘作業程序，為使發掘規則蘊藏豐富資訊，首先利用模糊劃分技術自動將數值資料轉換成口語化述辭之模糊集合。為避免遺失支持度低、信賴度高但有意義之關聯規則，第二階使用 SOM 類神經網路叢集分析法將具相似疾徵之樣本予以分類。第三階段運用關聯規則技術進行關聯規則資料探勘，再應用模糊相似關聯將語義重覆之關聯規則加以合併，其後並運用真實度評量方法評量模糊關聯規則，以保留滿足預先設定真實值之模糊關聯規則。本研究所提出之資料探勘方法，除可有效解決如第二節所述，現有資料探勘常見問題外，並可發掘有意義之模糊關聯規則，且保留之模糊關聯規則語意更為清晰且亦於運用。本研究貢獻除在實務應用上提供一醫療資料探勘建議程序外，並提出醫療資料預先處理之方法，及關聯規則合併及真實度評量方法之應用，以解決傳統關聯規則於實務的應用上，產生大量、無意義、重複之關聯規則等問題。為使發掘之模糊關聯規則更有意義，未來研究方向將以模糊相似關聯概念重新設計關聯規則演算法，以及產生前項或後項中，包含有或(or)條件型式之模糊關聯規則。

# 參考文獻

1. Agrawal R., Imielinski T., and Swami A., Mining association rules between sets of items in large databases, *ACM SIGMOD International Conference*, Washington D.C., May 1993, pp. 207-216.
2. Bastide Y., Pasquier N., Taouil R., Stumme G., and Lakhal L., Mining minimal non-redundant association rules using frequent closed item sets, *Lecture Notes in Computer Science*, 1861, 2000.
3. Bayardo R.J. and Agrawal R., Mining the most interesting rules, *Proc. KDD Conference*, 1999, pp.145-154.
4. Brin S., Motwani R. and Silversterin C., Beyond market baskets: Generalizing association rules to correlation, *Proc. SIGMOD conference*, 1997, pp.265-276.
5. Chaea Y.M., Kima H.S., Tarkb K.C., Parkb H.J., Hoa S.H., Analysis of healthcare quality indicator using data mining and decision support system, *Expert Systems with Applications*, 24 ,2003, pp. 167–172.

6. Chen M.S., Han J. and Yu P.S., Data mining: An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8, 1996, pp. 866-883.
7. Chen G. and Wei Q., Fuzzy association rules and the extended mining algorithms, Information Sciences 147, 2002, pp. 201–228.
8. Cybenko G., Approximating by super positions of a sigmoidal function, *Mathematical Control Signal Systems*, 2, 1989, pp. 303-314.
9. Delgado M., Sánchez D., Martín-Bautista M.J., and Vila M.A., Mining association rules with improved semantics in medical databases, *Artificial Intelligence in Medicine*, 21, 2001, pp.241-245.
10. Fayyad U., Piatetsky-Shapiro G. and Smyth P., From data mining to knowledge discovery in databases, *AI Magazine*, 17, 1996, pp. 37-54.
11. Frawley W.J., Piatetsky-Shapiro G. and Matheus C.J., Knowledge discovery in databases: an overview, AAAI/MIT Press, 1991, pp. 1-27.
12. Fu A., Wong M., Sze S., Wong W., and Yu W., Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes, *Proceedings of International Symposium on Intelligent Data Engineering and Learning*, Hong Kong, 1998, pp. 263-268.
13. Han J. and Fu Y., Discovery of multiple-level association rule from large databases, *Proceedings VLDB conference*, 1995, pp. 420-431.
14. Heckerman D., Bayesian networks for knowledge discovery, *Advances in Knowledge Discovery and Data Mining*, 1996, pp. 273-305.
15. Hornik K., Stinchcombe M. and White H., Multilayer feedforward networks are universal approximations, *Neural Networks*, 2, 1989, pp. 336-359.
16. Hsieh N.C., Handling indefinite and maybe information in logical fuzzy relational databases, *International Journal of Intelligent Systems*, 19, 3, March, 2004. 257-276.
17. Jensen S., Mining medical data for predictive and sequential patterns, *PKDD 2001 Discovery Challenge on Thrombosis Data*, 2001, pp.11-21.
18. Lavrac N., Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, 16 ,1999, pp. 3–23.
19. Kaufman L. and Rousseeuw P.J., Finding Groups in Data: An introduction to cluster analysis. John Wiley & Sons, 1990.
20. Klemettinen M., Mannila H., Ronkainen P., Toivonen H., and Verkamo A.I.. Finding interesting rules from large sets of discovered association rules. *Proceeding CIKM conference*, November 1994, pp. 401-407.
21. Kohonen T., The self-organizing map, *Berlin: Springer*, 1995.
22. Levin B., Meidan A., Cheskis A., Gefen O., and Vorobyov, PKDD99 Discovery Challenge-Medical Domain, *Workshop Notes on Discovery Challenge*, 1999, pp. 55-57.
23. Markey M.K., Lo J.Y., Tourassi G.D., Floyd Jr.C.E., Self-organizing map for cluster

analysis of a breast cancer database, *Artificial Intelligence in Medicine*, 27, 2003, pp. 113-127.

24. Mitra S., Data mining in soft computing framework: A survey, *IEEE Transactions on Neural Networks*, 13 (1), January 2002.
25. Ng R.T. and Han J., Efficient and effective clustering methods for spatial data mining, I*n Proceeding 20th International Conference on Very Large Databases,* 1994, pp. 144-155.
26. Ordonez C., Santana C.A., and Braal L. de, Discovering interesting association rules in medical data, In *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery* (DMKD 2000), 2000, pp. 78-85.
27. Park J.S., Chen M-S., Yu P.S., An effective hash-based algorithm for mining association rules, *Proceedings of ACM SIGMOD*, 1995, pp. 175-186.
28. PKDD2002, 6th European Conference on Principles and Practice of Knowledge Discovery in Databases, August 2002, Helsinki, Finland, http://ecmlpkdd.cs.helsinki.fi/.
29. Srikant R. and Agrawal R., Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, 1994, pp. 487-499.
30. Srikant R. and Agrawal R., Mining generalized association rules, *Proceedings of the 21th VLDB Conference*, 1995, pp. 407-419.
31. Srikant R. and Agrawal R., Mining quantitative association rules in large relational tables, in *Proceedings of the ACM SIGMOD International Conference*, Montreal, Canada, June 1996, pp. 1-12.
32. Taylor C.C., PKDD'99 Discovery Challenge: Medical Data Set, *Workshop Notes on Discovery Challenge*, 1999, pp. 59-64.
33. Yager R.R., General multiple-objective decision functions and linguistically quantified statements, *International Journal of Man-Machine Studies*, 21, 1984, pp. 389-400.
34. Yager R.R., On ordering weighted averaging aggregation operations in multicriteria decision-making, *IEEE Transactions on System, Man, Cybernetics*, 18, 1988, pp. 183–190.
35. Zadeh L.A., A computational approach to fuzzy quantifiers in natural languages, *Computers Mathematics with Applications*, 9, 1984, pp. 149–184.
36. Zadeh L.A., Fuzzy sets as a basis for theory of possibility, *Fuzzy Sets and Systems*, 1, 1978, pp. 3-28.
37. Zytkow J., Gupta S., Guide to Medical Data on Collagen Disease and Thrombosis, *PKDD 2001 Discovery Challenge on Thrombosis Data*, 2000.
38. Zemankova M. and Kandel A., Implementing Imprecise in Information Systems, Information Sciences, 37, 1985, pp. 107-141.

# 附錄(PKDD 2003; Zytkow & Gupta 2000)

# 原始膠原疾病資料庫欄位說明

表格 1：PATIENT_INFO

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| Sex | 性別 | char(1) |
| Birthday | 出生日期 | date |
| Description Date | 病患資料首次建檔日期 | date |
| First Date | 病患首次門診日期 | date |
| Admission | 門診後決定：'+'住院；'~'門診追蹤 | char(1) |

表格 2：DIAGNOSIS

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| DIAG_DATE | 診察日期 | date |
| SUSP/CONFIRM | 疾病狀態：疑似(~)/確認(+) | char(1) |
| DIAGNOSIS | 診斷病徵 | varchar(32) |
| FROM_TEST | 本紀錄資料來源檔案；<br>'DT'=TSUM_A；'ST'=TSUM_B | char(2) |

表格 3：ANTIBODY_EXAM

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| EXAM_DATE | 血栓病徵發生與開始攻擊之日期 | date |
| ACL IGG | ACL_IGG 抗體檢測值 | number |
| ACL IGM | ACL_IGM 抗體檢測值 | number |
| ANA | ANA 抗體檢測值 | number |
| ACL IGA | ACL_IGA 抗體檢測值 | number |
| KCT | KCT 血液凝結度測試;'+'超越正常範圍,'~'正常值 | varchar(1) |
| RVVT | RVVT 血液凝結度測試；'+'超越正常範圍，'~'正常值 | varchar(1) |
| LAC | LAC 血液凝結度測試;'+'超越正常範圍,'~'正常值 | varchar(1) |
| THROMBOSIS | 血栓嚴重程度；0：無血栓病徵，1：非常嚴重，2：嚴重，3：輕微 | number |

表格 4：ANA_PATTERN

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| EXAM_DATE | 檢驗日期 | date |
| ANA PA | 'P' ⇒ Peripheral Pattern [DNA] {SLE, PSS, MCTD, DLE, SjS, DILE}<br>'H' ⇒ Homogeneous Pattern [DNP-Histon, Histon] {SLE,DLE,DILE,SjS,PSS,RA}<br>'S' ⇒ Speckled Pattern [ENA,NAPA] {SLE,MCTD,SjS,PSS,PM/DM,DEL,RA}<br>'N' ⇒ Nucleolar Pattern [Nucleolar body] {PSS,SLE,SjS}<br>'D' ⇒ Discrete Speckled [Centromere] {CREST,PSS} | varchar(1) |

表格 5：THROMBOSIS

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| EXAM_DATE | 檢驗日期 | date |
| EXAM_DATE/ ATTACK_DATE | 血栓開始攻擊之日期 | date |
| SYMPTOM | 血栓攻擊期間觀察到的病徵 | varchar(32) |
| ATTACK SEQ NUM | 範圍由 1-5 之數字；表示攻擊的次數<br>0 ⇒ Symptom<br>1 ⇒ Symptom1<br>2 ⇒ Symptom2<br>3 ⇒ Symptom3<br>4 ⇒ Symptom4<br>5 ⇒ Symptom5 | number |

表格 6：LAB_EXAM

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| EXAM_DATE | 檢驗日期 | date |
| GOT | Normal Range : N < 60 | number |
| GPT | Normal Range : N < 60 | number |
| LDH | Normal Range : N < 500 | number |
| ALP | Normal Range : N < 300 | number |
| TP | Normal Range : 6.0 < N < 8.5 | number |
| ALB | Normal Range : 3.5 < N < 5.5 | number |
| UA | Normal Range : N > 8.0 (Male)<br>N > 6.5 (Female) | number |
| UN | Normal Range : N < 30 | number |
| CRE | Normal Range : N < 1.5 | number |

| | | |
|---|---|---|
| T-BIL | Normal Range : N < 2.0 | number |
| T-CHO | Normal Range : N < 250 | number |
| TG | Normal Range : N < 200 | number |
| CPK | Normal Range : N < 250 | number |
| GLU | Normal Range : N < 180 | number |
| WBC | Normal Range : 3.5 < N < 9 | number |
| RBC | Normal Range : 3.5 < N < 6 | number |
| HGB | Normal Range : 10 < N < 17 | number |
| HCT | Normal Range : 29 < N < 52 | number |
| PLT | Normal Range : 100 < N < 400 | number |
| PT | Normal Range : N < 14 | number |
| NOTE | 備註 | varchar(10) |
| APTT | Normal Range : N < 45 | number |
| FG | Normal Range : 150 < N < 450 | number |
| AT3 | Normal Range : 70 < N < 130 | number |
| A2PI | Normal Range : 0 < N < 30 or TR | number |
| U-PRO | Normal Range : 3.5 < N < 9 | number |
| IGG | Normal Range : 900 < N < 2000 | number |
| IGA | Normal Range : 80 < N < 500 | number |
| IGM | Normal Range : 40 < N < 400 | number |
| CRP | Normal Range : N < 1 or N = '+', '', '±' | varchar(4) |
| RA | Normal Range : N = '+', '', '±' | varchar(4) |
| RF | Normal Range : N < 20 | number |
| C3 | Normal Range : N > 35 | number |
| C4 | Normal Range : N > 10 | number |
| RNP | Normal Range : N = '+', '', '±' | varchar(4) |
| SM | Normal Range : N = '+', '', '±' | varchar(4) |
| SC170 | Normal Range : N = '+', '', '±' | varchar(4) |
| SSA | Normal Range : N = '+', '', '±' | varchar(4) |
| SSB | Normal Range : N = '+', '', '±' | varchar(4) |
| CENTROMEA | Normal Range : N = '+', '', '±' | varchar(4) |
| DNA | Normal Range : N < 8 | number |
| DNA-II | Normal Range : N < 8 | number |

表格 7：DISEASE

| 欄位名稱 | 說明 | 資料型態 |
|---|---|---|
| P_ID | 病患編號 | varchar(32) |
| DIAG_DATE | 檢驗日期 | date |
| DIAGNOSIS | 檢驗名稱 | varchar(32) |
| DIAGNOSIS_TYPE | 檢驗型態<br>'C' ⇒ Collagen disease<br>'D' ⇒ Non-collagen disease<br>'N' ⇒ No diagnosis<br>'O' ⇒ Observation | varchar(1) |
| COMMENT | 備註 | varchar(128) |