利用資料探勘技術於醫療院所輔助病患就診科別之研究

陳垂呈　　戴良安　董志源　韓志賢　王筱薇
南台科技大學資訊管理系　　南台科技大學資訊管理研究所
ccchen@mail.stut.edu.tw　{n9090023, n9290012, m9390105}@webmail.stut.edu.tw

摘要

在本篇論文中，我們以病患每次就醫之就診資料為探勘的資料來源，每一筆就診資料包含有病患症狀與其就診科別，並以目前某一病患症狀*X*為探勘目標，*X*為包含有*k*個症狀項目所形成的項目組，*k*≥1，利用分群化(clustering)技術從以下兩方面來找出病患症狀與就診科別之間的關聯性：一是以此一病患症狀*X*為一群組的中心點，將與中心點滿足最小相似度的就診資料，歸屬於*X*-群組中，然後從*X*-群組中找出出現次數最大的就診科別，做為輔助此一病患症狀*X*就診之科別項目的依據；二是以科別為群組的中心點，將包含中心點之科別的就診資料，歸屬於同一群組中，然後分別從各群組中找出最常出現的*k*個症狀，再將此結果與此一病患症狀*X*做相似度的計算，我們將具有最大症狀相似度之群組的科別，做為輔助此一病患症狀*X*就診之科別項目的依據。我們根據所提出的方法，設計與建置一個輔助病患就診科別的指引系統。此探勘結果，對提昇醫療院所的服務品質、及對病患有效就診並降低延誤就醫的風險，都可以提供非常有用的參考資訊。

關鍵詞：資料探勘、分群化、症狀、科別

一、簡介

隨著醫療體系的日益發展，民眾對醫病關係也相對的更加重視，在民眾就醫時往往衍生出許多的醫病問題，其中又以病患通常未具有醫療專業知識，當患病前往醫療院所求診時，往往無法依其症狀來判斷應該看診那一科別，是最為常見的問題之一，其結果可能導致病患有延誤就醫的風險。因此，如何在病患求診時，輔助病患症狀就診之科別項目的指引，以降低延誤就醫的可能性，並提昇有效醫療的服務品質，即成為醫療人員必須思考的問題之一。

藉著醫療院所的資訊化，儲存病患的就診資料已從傳統紙本病歷轉變成電子病歷，根據美國電子病歷學會(Computer-based Patient Record Institute, CPRI)的描述：「關於個人終其一生之健康狀態及醫療照護的電子化資訊，電子病歷將取代紙本病歷以符合臨床應用、行政管理、醫學教育、研究調查及其他合法需求的主要醫療資料來源」。從過去病患的就診資料中，找出病患症狀與就診科別之間的關聯性，做為醫療就診的參考資訊，以提升醫療的準確性及時效性，並降低病患延誤就醫的風險，是利用就診資料重要的研究主題之一。

資料探勘(data minig)是從大量資料中挖掘潛在有用的資訊與知識，以做為決策分析的參考資訊，資料探勘技術目前已普遍地應用在各領域中[4]。在本篇論文中，我們以病患之就診資料為探勘的資料來源，每一筆就診資料記錄有病患症狀與就診的科別，並以目前某一病患症狀 *X* 為探勘目標，*X* 為包含有 *k* 個症狀項目所形成的項目組，*k*≥1，利用分群化(clustering)技術分別從以下兩方面來探討如何輔助此一病患就診科別的指引：

(1) 以此一病患症狀為群組的中心點：我們設定此一病患症狀 *X* 為群組的中心點，並將與中心點滿足最小相似度的就診資料，歸屬於同一群組中，稱之為 *X*-群組。然後從 *X*-群組中計算出現次數最大的就診科別，做為輔助此一病患症狀 *X* 就診之科別項目的依據。

(2) 以科別為群組的中心點：我們分別設定各科別項目為群組的中心點，並將包含中心點之科別的就診資料，歸屬於同一群組中，然後分別從各群組中找出最常出現的 k 個症狀，再將此結果與此一病患症狀 X 做相似度的計算。我們將具有最大症狀相似度之群組中心點的科別項目，做為輔助此一病患症狀 X 就診之科別項目的依據。

我們根據所提出的方法，設計與建置一個輔助病患就診科別的指引系統。此探勘結果，對病患選擇正確的就診科別、進而達到有效醫療並降低延誤就醫的風險，並對提昇醫療院所的服務品質及避免重複就診之醫療資源的浪費，都可以提供非常有用的參考資訊。

本篇論文的架構如下：下一節中，我們說明資料探勘技術、及其在醫療應用上的相關研究；第三節中，我們以某一病患症狀為群組的中心點，設計一個分群化方法來輔助此一病患就診之科別項目的指引依據；第四節中，我們以各科別項目為群組的中心點，設計一個分群化方法來輔助病患就診之科別項目的指引依據；第五節中，我們依據所提出的方法，設計與建置一個輔助病患就診科別的指引系統；最後，我們在第六節中做一結論。

二、相關研究

資料探勘是在大量的資料中找出潛藏有用的資訊與知識，其可完成以下任務或是更多：關聯規則 (association rules) 、分群 (clustering) 、分類 (classification) 、次序相關分析 (sequential pattern analysis)等[5]，在疾病診斷應用上，可藉由發掘病患症狀與疾病之間的關聯性，做為診斷病患可能罹患之疾病的參考資訊，以便進行有效的治療及預防。目前資料探勘技術已普遍地應用在醫療診斷中，其相關研究有：[1]從病歷資料著手，尋找病例與用藥之間的關係，並希望藉由資料探勘的技術，防杜健保制度中用藥浮濫的問題；[2]透過資料探勘的技術，以標準健保資料作為系統資料的來源，實作出一套醫療領域專門的資料探勘系統，藉以探究不同疾病之間的關係，以提供未來預防治療的參考；[3]以貝氏網路、決策樹與倒傳遞神經網路等演算法針對乳部腫瘤、中醫舌診影像與糖尿病健康管理紀錄進行處理，藉以證明資料探勘技術可以用於輔助醫生診斷的用途上，甚至診斷的準確率高過人為的診斷。

分群化是將物件根據相似度來進行分群，關於分群化的研究，主要可分為以下幾種：分割式 (partitioning) 、階層式 (hierarchical) 、格子基礎 (grid-based) 、密度基礎 (density-based) 與模型基礎 (model-based) 等幾種[4]。在本篇論文中，我們將修改分割式分群化的方法，做為分群化交易資料的方法依據。

在眾多分割式分群化演算法中，較著名的有PAM[6]、k-means[7, 8]及CLARANS[9]等，其目的是分群成使用者所指定的k個群組，此分割方式可將每一物件歸屬於最相似的群組中。以下我們介紹PAM(Partitioning Around Medoids)演算法的分群化步驟。

PAM演算法由Kaufman and Rousseeuw[6]所提出，為了將全部物件分群成k個群組，PAM的方法是先為每個群組決定一個代表物件(representative objects)，此代表物件稱之為*medoid*，一旦把k個medoids選定之後，就依據相似度來決定非medoid物件是屬於那一個群組，其相似度是以物件彼此之間的距離(Euclidean distance)來表示，$d(O_a, O_b)$表示物件O_a與O_b之間的距離。例如O_i為medoid，而O_j為非medoid物件，如果$d(O_j, O_i)=min\{d(O_j, O_e)\}$， O_e表示所有的medoids，則O_j歸屬於O_i群組。

對任一個非medoid物件O_j而言，當一個medoid O_i被一個非medoid物件O_h取代時，所造成的改變成本C_{jih}定義如下：

$$C_{jih} = d(O_j, O_m) - d(O_j, O_n)$$

O_m 表示以 O_h 取代 O_i 之後，與 O_j 有最大相似度(最短距離)的 medoid；

O_n 表示以 O_h 取代 O_i 之前，與 O_j 有最大相似度(最短距離)的 medoid。

以 O_h 取代 O_i 成為 medoid 之後，所造成的總改變成本為：

$$TC_{ih} = \sum_{j} C_{jih}$$

若 $TC_{ih}>0$ 時，表示以 O_h 取代 O_i 之後的總距離比取代前大，則 O_i 將不會被 O_h 所取代。以 TC_{ih} 為衡量依據，PAM演算法說明如下：

Algorithm PAM

(1) 任意選取 k 個物件做為 medoids。

(2) 對所有 O_i 與 O_h 之組合，計算出其 TC_{ih}，其中 O_i 表示任一個的 medoid，O_h 表示任一個非 medoid 物件。

(3) 選擇出 TC_{ih} 為最小值的 O_i 與 O_h 配對；假如 $TC_{ih}<0$，則以 O_h 取代 O_i 成為 medoid，並跳至 (2)。

(4) 否則停止執行，已完成分群。

在本篇論文中，我們將修改分群化技術來做為探勘就診資料的方法依據，並以某一病患症狀為探勘的目標，從以下兩方面來探討輔助此一病患症狀之就診科別的指引：一是以此一病患症狀為中心點；二是以各科別項目為中心點。接下來，我們定義一些名詞如下：

- $S=\{s_1, s_2, \ldots, s_a\}$，是全部症狀項目的集合，共有 a 項。
- $D=\{d_1, d_2, \ldots, d_b\}$，是全部科別項目的集合，共有 b 項。
- $T=\{T_1, T_2, \ldots, T_j, \ldots T_m\}$，為全部就診資料的集合，共 m 筆，其中 T_j 表示第 j 筆就診資料，$1\le j\le m$。
- 就診資料 T_j 之格式為 $T_j=[X, Y]$，$X\subseteq S$、X 為一項或以上症狀項目所組成的項目組，$Y\subseteq D$，Y 為一項或以上科別項目所組成的項目組，即病患症狀為 X、及就診科別為 Y，如表 1。

表 1 就診資料格式

就診資料編號	症狀項目	科別項目

三、以病患症狀為中心點輔助病患就診之科別項目

在此一章節中，我們以病患每次就醫時之就診資料做為探勘的資料來源，每一筆就診資料包含有病患的症狀項目及就診的科別項目，並以目前某一病患症狀為探勘目標，我們設計一個方法來分群化就診資料，並從分群化後之群組所顯示出的傾向特徵，做為輔助此一病患就診之科別項目的指引依據。此章節共分為兩小節如下：第一小節中，我們說明以某一病患症狀為群組之中心點的分群化過程；第二小節中，我們以一實例做說明。

(一) 以某一病患症狀為中心點分群化就診資料

我們設計一個簡單、快速的分群化方法，以某一病患症狀為群組的中心點，然後將與中心點具有滿足最小症狀相似度的就診資料，歸屬於同一群組中。假設 T_{j1} 及 T_{j2} 為兩筆就診資料，我們定義兩筆就診資料之間的症狀相似度為：

症狀相似度 t = (就診資料 T_{j1} 與 T_{j2} 之間有相同症狀項目的個數)/(就診資料 T_{j1} 的症狀項目個數)，當 t 愈大，表示就診資料 T_{j2} 包含有愈多與 T_{j1} 相同的症狀項目。

我們將兩筆就診資料中的症狀項目直接進行比較計算，可以有效率地得到兩筆就診資料 T_{j1} 與 T_{j2} 之間的相似度。我們定義一函數 *fetch-item*(T_j, i_1) 表示可以擷取就診資料 T_j 中第 i_1 個的症狀項目。例如，T_1={ABC, XY}，A、B、C∈症狀項目，X、Y∈科別項目，則 *fetch-item*(T_1, 2)=B。每一就診資料中所包含的症狀項目及科別項目，都已事先由小到大

的排序過，例如 A<B<C 及 X<Y，因此計算兩筆就診資料 T_{j1} 與 T_{j2} 之間的症狀相似度，可表示成以下的演算法：

```
Float Per-Same-S-Item(Tj1, Tj2) {
   int same_item=0; /*相同症狀項目的數量變數*/
   int i1=i2=1;     /*表示就診資料 Tj1 中第 i1 個症狀
                      項目、及 Tj2 中第 i2 個症狀項目*/
   while (fetch-item(Tj1, i1) <> ∅) and
                          (fetch-item(Tj2, i2) <> ∅) {
      if (fetch-item(Tj1, i1) == (fetch-item(Tj2, i2) {
         same_item++;
         i1++;
         i2++;
      }
      elseif (fetch-item(Tj1, i1) > (fetch-item(Tj2, i2)
         i2++;
      else i1++;
   }
   return same_item/|Tj1|;  /*|Tj1|為就診資料 Tj1 的
                              症狀項目個數*/
}
```

例如，T_{j1}={BCE, XY}及 T_{j2}={ABD, YZ}，A、B、C、D、E∈症狀項目，X、Y、Z∈科別項目，經由上述演算法的計算，其症狀相似度=1/3=33%

假設目前欲探勘之病患症狀為 *X*，*X* 為一個或以上症狀所形成的集合，設定 *X* 為一群組的中心點，依據之前所定義的症狀相似度，並設定一個「最小症狀相似度」，來將與 *X* 具有滿足此條件的就診資料 T_j 歸屬於同一群組，稱之為 *X*-群組，$1 \le j \le m$，表示共有 *m* 筆的就診資料，分群化的過程可表示成以下演算法：

```
Clustering-1(X) {
   for (j=1; j≤m; j++)
      if Per-Same-S-Item(X, Tj) ≥最小症狀相似度
         Tj∈X-群組;
}
```

例如，假設所設定的最小症狀相似度為 70%，則與 *X* 具有 70%或以上症狀相似度的就診資料，就歸屬於 *X*-群組中。經由上述演算法的分群化步驟，即可將具有滿足最小症狀相似度的就診資料歸屬於 *X*-群組中。經由分群化之後，可在 *X*-群組中計算出現次數最大的科別項目，其定義說明如下：

病患就診科別項目=*max*{在 *X*-群組中各科別項目出現的次數}。

藉由從分群化之後的 *X*-群組中，找出出現次數最大的科別項目，以做為輔助此一病患就診之科別項目的指引依據。

(二) 實例說明

我們以一實例來說明輔助某一病患就診之科別項目的探勘過程，表 2 為一就診資料庫，其包含有 4 筆的就診資料，其中{A, B, C, D, E}表示所有症狀項目的集合，{X, Y, Z}表示所有科別項目的集合，{T_1, T_2, T_3, T_4}表示所有就診資料的集合。假設目前欲探勘之病患症狀為 CE，設定最小症狀相似度為 60%。

表 2 就診資料庫

就診資料編號	症狀項目	科別項目
T_1	ABD	XZ
T_2	BE	X
T_3	ACE	Y
T_4	BCE	XY

我們以此一病患症狀 CE 為一群組的中心點，經由 *Clustering-1* 演算法的計算，可得到以下的 CE-群組：

CE-群組={T_3, T_4}

在 CE-群組中出現次數最大的科別項目=*max*{X=1/2, Y=2/2}= Y。因此，藉由從分群化之後的 CE-群組中，找出 Y 為輔助此一病患就診之科別項目的

指引依據。

四、以科別項目為中心點輔助病患就診之科別項目

在此一章節中，我們仍以病患每次就醫時之就診資料做為探勘的資料來源，並以目前某一病患症狀為探勘目標，我們以各科別項目為群組中心點，設計一個分群化方法，從分群化後之各群組所顯示出的傾向特徵，做為輔助此一病患就診之科別項目的指引依據。此章節共分為兩小節如下：第一小節中，我們說明以各科別項目為群組之中心點的分群化過程；第二小節中，我們以一實例做說明。

(一) 以科別項目為中心點分群化就診資料

我們設計一個簡單且快速的分群化方法，以各科別項目為群組的中心點，然後將包含有中心點之科別項目的就診資料，歸屬於同一群組中。假設共有 $d_1, d_2, \ldots, d_b$ 等 b 個科別項目，目前欲探勘之病患症狀為 X，X 為一個或以上症狀項目所形成的集合且其個數為 k，我們設定各科別項目 d_i 為群組的中心點，$1 \le i \le b$，將包含有中心點 d_i 的就診資料歸屬於 d_i-群組中。就診資料 T_j 共有 m 筆，$1 \le j \le m$，分群化的過程可表示成以下演算法：

```
Clustering-2() {
   for (j=1; j≤m; j++)
      for (i=1; i≤b; i++)
         if (di ⊆ Tj)    Tj∈ di-群組;
}
```

例如，假設就診資料包含有科別 d_2，則將此就診資料歸屬於 d_2-群組中。經由上述演算法的分群化步驟，即可將包含有各科別項目之就診資料歸屬於各科別的群組中。經由分群化之後，可在各科別群組中計算出現次數最大的前 k 個症狀項目，其定義說明如下：

各科別群組的症狀項目= *max* 前 k 個{在 d_i-群組中各症狀項目出現的次數/在 d_i-群組中就診資料的總筆數}，$1 \le i \le b$。

然後，我們再找出此一病患症狀 X 與各科別群組的症狀項目之間症狀相似度的最大者，其定義如下：

症狀相似度= *max*{(病患症狀 $X \cap$ 各 d_i-群組的症狀項目，$1 \le i \le b$)的項目個數/k}。

藉由計算具有症狀相似度最大者的科別群組中，我們即以此群組之科別項目，以做為輔助此一病患就診之科別項目的指引依據。

(二) 實例說明

我們仍以表 2 之就診資料庫為例，假設目前欲探勘之病患症狀為 BE，其輔助此一病患就診之科別項目的探勘過程說明如下。

我們以各科別項目為群組的中心點，經由 *Clustering-2* 演算法的計算，可得到以下的科別群組：

X-群組={T_1, T_2, T_4}
Y-群組={T_1, T_3}
Z-群組={T_1}

計算各科別群組中症狀出現最大的前 2 項為：

X-群組= *max* 前 2 個{A=1/3, B=3/3, C=1/3, D=1/3, E=2/3}= BE。
Y-群組= *max* 前 2 個{A=2/2, B=1/2, C=1/2, D=1/2, E=1/2}
= AB or AC or AD or AE。
Z-群組= *max* 前 2 個{A=1/1, B=1/1, D=1/1}
= AB or AD or BD。

再計算各科別群組與此一病患症狀之間的症狀相似度為：

X-群組= (BE∩BE)/2= 100%
Y-群組= (BE∩AB)/2 or (BE∩AE)/2= 50%

Z-群組= (BE∩AB)/2 or (BE∩BD)/2= 50%

具有最大症狀相似度的科別群組為 X，因此，藉由計算分群化之後的各科別群組之症狀相似度，找出 X 為輔助此一病患就診之科別項目的指引依據。

五、輔助病患預診系統之實作

我們將前面章節所描述的探勘方法，應用到輔助病患就診科別之指引系統的實作上，以C#為撰寫的程式語言。在不失一般性的條件下，假設症狀項目全部有26項，分別以A, B, C, …, Z來表示之，科別項目全部有10項，分別以1, 2, 3, …, 9, 0來表示之，並以亂數隨機產生每一筆就診資料，每一筆就診資料包含有最多7個症狀項目與最多4個科別項目，共產生500筆就診資料，以下為此一系統的探勘執行過程。

圖1為此一系統的就診資料，包含「就診資料編號」、「症狀」及「科別」等欄位資料。

圖 1 就診資料

圖 2 表示探勘畫面，其中包含有兩項功能選項：「以科別為群組的中心點」與「以某一病患症狀為群組的中心點」。假設目前點選「以科別為群組的中心點」的功能，並在「輸入病患症狀」欄位中輸入病患症狀。假設目前輸入之病患症狀為 AB，經由第四章節演算法的探勘過程，可在「群組」欄位中顯示出各科別群組所包含的就診資料，並在「就診科別指引」欄位中顯示出探勘的結果，如圖 2。

圖 2 以科別為群組之中心點的探勘執行畫面

圖 3 表示點選「以某一病患症狀為群組的中心點」的功能，並在「輸入病患症狀」欄位中輸入病患症狀、及在「輸入最小症狀相似度」欄位中輸入相似度值。假設目前輸入之病患症狀為 AB，最小症狀相似度為 50%，經由第三章節演算法的探勘過程，可在「群組」欄位中顯示出 AB-群組所包含的就診資料，並在「就診科別指引」欄位中顯示出探勘的結果，如圖 3。

圖 3 以病患症狀為群組之中心點的探勘執行畫面

六、結論

隨著醫療的發展，醫療科別的分工也愈趨精細，在病患就診時，往往不知其症狀較合適於那一科別看診，如此結果不僅造成醫療資源的浪費，病患也無法得到有效的醫療，甚至可能導致病情延誤的情況。在本篇論文中，我們以病患每次就醫之就診資料為探勘的資料來源，並以目前某一病患症狀為探勘目標，設計兩個簡單且快速的分群化方法，分別從以下兩方面來找出病患症狀與就診科別之間的關聯性：一是以此一病患症狀為一群組的中心點；二是以科別為群組的中心點。我們從以上兩個方法所建立的群組中，分別找出與此一病患症狀最具有關聯性的科別項目，藉此做為輔助此一病患症狀應就診那一科別的依據。此探勘結果，對病患有效就診並降低延誤就醫的風險、及提昇醫療院所的服務品質，都可以提供非常有用的參考資訊。我們根據所提出的方法，設計與建置一個輔助病患就診科別的指引系統。

七、參考文獻

1. 陳世源，資料採礦技術在病例與藥品關連性之研究，國立中山大學資訊管理研究所碩士論文，1999。
2. 俞旭昇，以資料探勘技術發掘疾病隱藏關係之研究，國立暨南國際大學資訊管理研究所碩士論文，2002。
3. 吳國禎，資料探索在醫學資料庫之應用，中原大學醫學工程研究所碩士論文，1999。
4. J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.
5. M. S. Chen, J. Han and P. S. Yu, “Data Mining: An Overview from a Database Perspective,” *IEEE Trans. on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 866-883, 1996.
6. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
7. R. C. Dubes and A. K. Jain, *Algorithms for Clustering Data*, Prentice Hall, 1988.
8. K. Alsabti, S. Ranka, and V. Singh, “An Efficient K-Means Clustering Algorithm,” *PPS/SPDP Workshop on High Performance Data Mining*, 1997.
9. R. T. Ng and J. Han, “Efficient and Effective Clustering Methods for Spatial Data Mining,” *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 144-155, 1994.