利用分群化建置病患罹患疾病探勘系統

陳垂呈[1,*]　楊明憲[1]　陳宗義[2]　李靖平[3]
[1]南台科技大學資訊管理研究所
[2]南華大學電子商務管理系
[3]南台科技大學企業電子化學位學程
*E-mail: ccchen@mail.stut.edu.tw

摘要

本研究以病患之診斷資料為探勘的資料來源，每一筆診斷資料包含病患罹患的疾病與其顯示的症狀，利用資料探勘(data mining)中的分群化(clustering)技術分析病患罹患的疾病傾向。文中以k位病患為探勘的目標，$k\geq 1$，設計一個分群化方法將診斷資料分群成k個群組，每一個群組中的診斷資料具有最大症狀相似度，且k個群組的症狀相似度總和為最大，然後分別從各群組中找出病患最可能罹患的疾病。本研究根據所提出的方法，設計與建置一個病患疾病診斷探勘系統。本系統的探勘結果，對一般民眾自我檢視罹患疾病、及臨床經驗不足醫療人員的疾病診斷，可以提供非常有用的參考資訊。

關鍵詞：資料探勘、分群化、診斷資料、疾病

利用分群化建置病患罹患疾病探勘系統

1. 簡介

隨著資訊技術的發展，醫療院所儲存病患的診斷資料已從傳統紙本病歷轉成電子病歷，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」。從過去病患的診斷資料中，找出外在顯示症狀與引發疾病之間的關聯性，提供醫療診斷的參考資訊，並提升醫療診斷的準確性及時效性，降低診斷疾病過程中的疏忽，是應用診斷資料領域中重要的研究主題之一。

資料探勘(data mining)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，資料探勘目前已普遍應用在許多的領域中(Han and Kamber, 2006)。本研究以病患的診斷資料為探勘的資料來源，每一筆診斷資料記錄病患就醫時的症狀與罹患的疾病，並以k位病患為探勘的目標，$k \geq 1$，利用分群化(clustering)技術發掘那些症狀與罹患疾病具有高度的關聯性，藉此發掘k位病患最可能罹患的疾病。例如病患顯示的症狀有「胸部不適」、「胸悶」、「胸部感到沈重」、或「胸痛」等，則可能罹患心血管的疾病。

假設欲探勘之k位病患症狀分別為$s_1, s_2, \ldots, s_i, \ldots s_k$，$1 \leq i \leq k$，先以$k$位病患症狀分別設定為一群組的中心點，本研究設計一個分群化方法將診斷資料歸屬於與中心點具有最大症狀相似度的群組，並達到整體症狀相似度總和為最大值的分群化目標，然後從群組中分別發掘k位病患最可能罹患的疾病。本研究根據所提出的方法，以南部某一醫學中心的診斷資料為例，設計與建置一個病患疾病診斷探勘系統。本系統的探勘結果，對一般民眾檢視罹患疾病可提供相當有用的參考資訊，疾病預防與即早治療可提供相當有用的預警功效，而臨床經驗不足醫療人員的疾病診斷，也可提供輔助的參考資訊。

本論文的架構如下：下一節中說明資料探勘技術、及其在醫療應用上的相關研究；第3節中以k位病患為探勘的目標，設計一個分群化方法，藉由分群化後之群組所顯示的疾病特徵，做為分析病患最可能罹患之疾病的依據；第4節中以南部某一醫學中心的診斷資料為例，根據所提出的方法，設計與建置一個病患疾病診斷探勘系統；最後，在第5節中做一結論。

2. 相關研究

醫學領域應用資訊技術而發展出醫學資訊學(medical informatics)，其目的是利用

資訊技術的輔助，並以病患為中心、醫療問題為導向的診斷模式，希望藉由資訊技術的支援建立醫學知識，進而找出各項疾病的醫療指引(朱彩屏，2004)。若能有效利用資訊技術於疾病診斷上，做為診斷病患可能罹患之疾病的參考資訊，對病患的治療及疾病的預防將可提供相當大的幫助。

資料探勘是從大量資料中挖掘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)、及預測等(Chen, et al., 1996; Han and Kamber, 2006)，目前已有許多的研究顯示資料探勘可以有效應用在醫療診斷中，其相關研究有：俞旭昇(2002)透過資料探勘技術，以標準健保資料做為系統資料來源，實作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，以提供預防治療的參考；吳素英(2004)利用資料探勘技術建構醫院疾病分類的知識管理系統；唐壽生(2004)利用資料探勘技術於肺結核病患的醫療預測；陳迪祥(2003)利用關聯規則找出罹患疾病彼此之間的發生機率；Ye和Keane (1997)利用複合項關聯規則(association rules with composite items)探討症狀與疾病的關聯性。

分群化是將物件根據相似度來進行分群，關於分群化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(Han and Kamber, 2006)。Berry 和 Linoff (1997)曾描述：「一開始想對資料進行分析、了解資料意涵並描繪出最好的利用方式，分群化分析(cluster analysis)是一個很好的方法」。本研究將修改分割式分群化的方法，做為分群化診斷資料的方法依據。

眾多分割式分群化演算法中較著名的有PAM(Partitioning Around Medoids) (Kaufman and Rousseeuw, 1990)、k-means (Alsabti, et al.,1997; Dubes and Jain, (1988)及CLARANS (Ng and Han,1994)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。以下介紹PAM演算法的分群化步驟。

PAM演算法由Kaufman 和 Rousseeuw (1990)所提出，為了將全部物件分群成k個群組，$k \geq 1$，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，就依據相似度決定非medoid物件是屬於那一個群組，其相似度表示物件彼此之間的距離(Euclidean distance)，$d(O_a, O_b)$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，並跳至步驟(2)。
(4) 否則停止執行，已完成分群。

本研究以病患每次就醫時之診斷資料為探勘的資料來源，並以k位病患為探勘的目標，k≥1，利用分群化方法探討k位病患症狀最可能罹患的疾病。文中定義診斷資料的格式為{S, D}，S為包含一個或以上的症狀項目，D為包含一個或以上的疾病項目，例如診斷資料{a, X}，即顯示症狀a其罹患疾病X，其中a⊆S、X⊆D。每一筆診斷資料包含病患所顯示的症狀項目、及診斷罹患的疾病項目。

3. 發掘病患症狀最可能罹患之疾病

本章節以病患每次就醫之診斷資料為探勘的資料來源，並以k位病患為目標，k≥1，k位病患症狀分別以$s_1, s_2, \ldots, s_i, \ldots, s_k$表示之，$1 \le i \le k$，利用分群化方法將診斷資料分群化成$k$個群組，藉由群組所顯示的疾病特徵，做為發掘病患最可能罹患之疾病的依據。本章節共分為兩小節如下：第3.1節中設計一個分群化方法發掘病患最可能罹患的疾病；第3.2節中以一實例做說明。

3.1 分群化方法

文中定義以下症狀相似度做為診斷資料歸屬於那一群組的依據：

症狀相似度={診斷資料∩群組中心點}的症狀項目數量/群組中心點的症狀項目數量。

例如一筆診斷資料為{abc, XYZ}，一個群組中心點的症狀項目為{acdef}，其中{a, b, c, d, e}為症狀項目集合，{V, W, X, Y, Z}為疾病項目集合，則症狀相似度=2/5=40%。計算診斷資料與各群組中心點之間症狀相似度，然後將診斷資料歸屬於症狀相似度最大的群組中。在每次分群化之後計算整體症狀相似度的總和，若目前分群化的整體症狀相似度總和大於之前的分群，則將目前分群中心點取代之前的中心點。

由於欲發掘k位病患最可能罹患的疾病，文中先以k位病患症狀$s_1, s_2, \ldots, s_k$分別設定為一群組的中心點，依據症狀相似度的大小將診斷資料T_j歸屬於群組中，分別以s_1-群組、s_2-群組、…、s_k-群組表示之，$1 \le j \le m$，表示有m筆的診斷資料，分群化的過程可表示為：

Clustering ($s_1, s_2, \ldots, s_k$) {
  if k=1 {
    計算診斷資料 T_j 與 s_1 之間的症狀相似度，
    若大於等於所設定的最小症狀相似度，則將診斷資料 T_j 歸屬於 s_1-群組;
    break;
  }
  else {
    以 k 位病患症狀 $s_1, s_2, \ldots, s_i, \ldots, s_k$ 分別設定為群組中心點;
    d_1=計算分群後的整體症狀相似度的總和;
    $d=d_1$; /*表示分群化之後減之前整體症狀相似度總和的差值*/
    while d>0 {
      挑選任一診斷資料，其與 s_i 之間的症狀相似度大於等於所設定的最小症狀相似度，取代原先 s_i-群組的中心點，$1 \le i \le k$;
      計算分群化之後的整體症狀相似度的總和;
      d_2=選出整體症狀相似度總和為最大值的中心點組合;
      $d=d_2-d_1$;
      if $d \le 0$ {
        保留之前的分群;
        break;
      }
      else {
        將目前分群中心點取代之前的中心點;
        $d_1=d_2$;
        }
      }
    完成分群;
}

經由上述分群化步驟可將診斷資料歸屬於最適合的群組，並達到整體症狀相似度總和為最大值的目標，並且在挑選群組新中心點時，新中心點與原中心點之間的症狀相似度必須滿足所設定的「最小症狀相似度」，如此將可在分群化過程中保留群組的獨特性。在s_i-群組中計算各項疾病出現的比率值為：各項疾病出現在s_i-群組中的數量/s_i-群組包含的診斷資料數量，然後將比率值最大的疾病項目，視為病患症狀s_i最可能罹患的疾病。藉由s_i-群組的疾病傾向特徵，文中定義病患症狀s_i最可能罹患的疾病如下：

病患症狀s_i最可能罹患的疾病：在s_i-群組中出現比率值為最大的疾病項目。

根據以上的探勘計算，即可分別發掘*k*位病患最可能罹患的疾病。在實際的應用中，對於設定「最可能罹患的疾病」的疾病項目數量，可依據應用上的需要而彈性調整，例如在群組中出現比率值為最大的前*r*項疾病，$r \geq 1$。

3.2 實例說明

文中以一實例說明發掘病患症狀最可能罹患之疾病的探勘過程，表1為一診斷資料庫*D*，其包含4筆的診斷資料，其中{a, b, c, d, e}為症狀項目的集合，{V, W, X, Y, Z}為疾病項目的集合，{T_1, T_2, T_3, T_4}為診斷資料的集合。假設欲探勘之病患症狀分別為a及be，最小症狀相似度為50%。

表 1 診斷資料庫 *D*

診斷資料編號	症狀項目	疾病項目
T_1	ad	VX
T_2	acd	XY
T_3	bce	WXZ
T_4	bcde	WYZ

首先設定病患症狀 a 及 be 分別為群組的中心點，經由演算法 *Clustering()*的計算，可得到以下兩個群組：

a-群組={T_1, T_2 }及 be-群組={T_3, T_4}

在 a-群組中計算各項疾病出現的比率值為：V=1/2=50%, X=2/2=100%, Y=1/2=50%，可發掘病患症狀 a 最可能罹患的疾病為 X。在 be-群組中計算各項疾病出現的比率值為：W=2/2=100%, X=1/2=50%, Y=1/2=50%, Z=2/2=100%，可發掘病患症狀 be 最可能罹患的疾病為 WZ。

4. 病患疾病診斷探勘系統

本研究將前面章節所描述的探勘方法，設計與建置一個病患疾病診斷探勘系統，表2為系統的開發平台。

表2 系統開發平台

作業系統	Windows XP Professional Edition
CPU	AMD K-7 1.3GHz
主記憶體	512M SDRAM
設計語言	ASP、VB Script
資料庫	Microsoft Access 2002

文中以南部某一醫學中心之病患每次就醫的診斷資料為例，診斷資料從2004/4/1到2004/4/7共計50464筆，以做為所設計之探勘方法的資料來源，其中以前面50000筆之診斷資料作為探勘計算的訓練資料，並以最後464筆診斷資料做為探勘計算的驗證資料。圖1為診斷資料的原始資料，這些原始資料是以病患每次就醫為一個記錄儲存，每一筆診斷資料包含有就醫時的「科別」、「症狀」、及「疾病」等欄位資料。

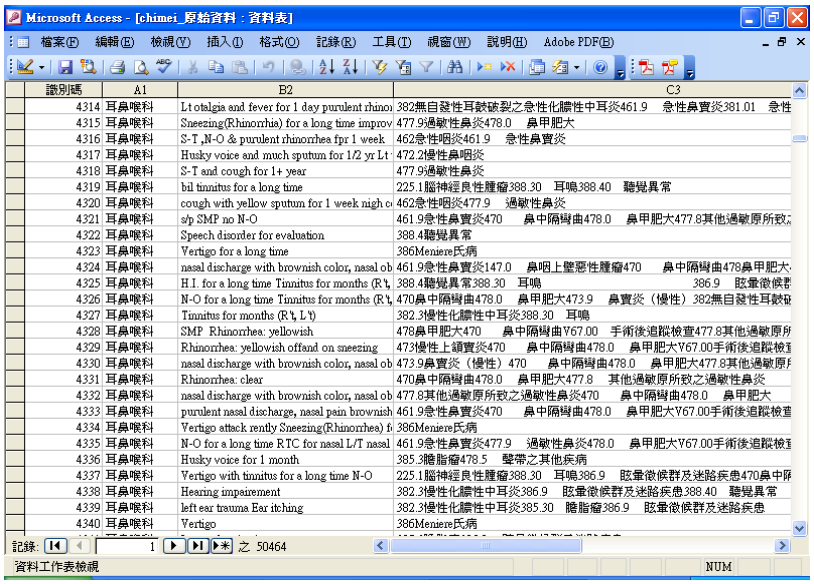

圖1 原始診斷資料

探勘過程中必須先分別對診斷資料中的症狀描述及疾病名稱進行編碼，由於疾病名稱已可利用ICD-9-CM碼(The International Classification of Disease, 9th Revision, Clinical Modification)進行編碼，例如氣喘，其ICD-9-CM碼為493.9。至於症狀名稱編碼，則從症狀描述中篩選出較重要的症狀字詞進行編碼，並分別以S0001, S0002, S0003等依次進行編碼。在圖2中分別以編碼後之疾病碼及症狀碼替代原始診斷資料中的疾病名稱及症狀描述。

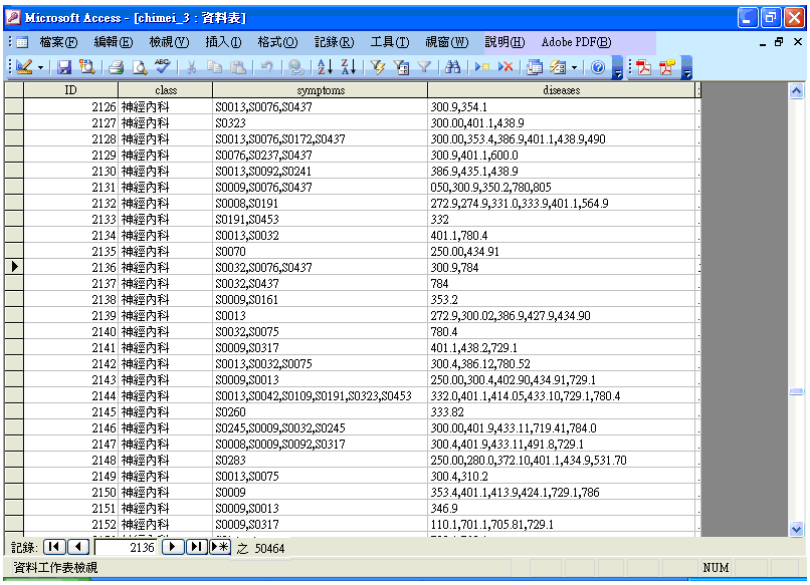

圖2 編碼後的診斷資料

本研究以前面50000筆的診斷資料做為探勘的訓練資料，以發掘病患症狀最可能罹患的疾病，以下說明所建置的病患疾病診斷探勘系統，其在探勘訓練資料的執行過程：圖3的探勘畫面中在「症狀」欄位輸入欲探勘之病患症狀項目，經由第3節所描述的計算過程，可在「最可能罹患的疾病」欄位中顯示出探勘的結果，如圖4。

圖3 發掘病患症狀最可能罹患之疾病的執行畫面

圖4 發掘病患症狀最可能罹患之疾病的結果畫面

本研究以剩餘464筆診斷資料做為探勘的驗證資料，評估在前面訓練資料中所探勘之結果的成效。文中從剩餘464筆診斷資料中，利用第3節的探勘方法，分別輸入這些診斷資料中出現的症狀項目，以發掘病患症狀最可能罹患的疾病，然後再檢查這些診斷資料中的疾病項目是否包含「最可能罹患的疾病」，若有即定義診斷資料可反應出系統所發掘之最可能罹患的疾病，否則定義為未能反應出最可能罹患的疾病。圖5中驗證評估在不同診斷資料數量的情況下，可反應出最可能罹患的疾病之診斷資料數量，從圖中顯示，系統可穩定驗證出最可能罹患的疾病之診斷資料數量。

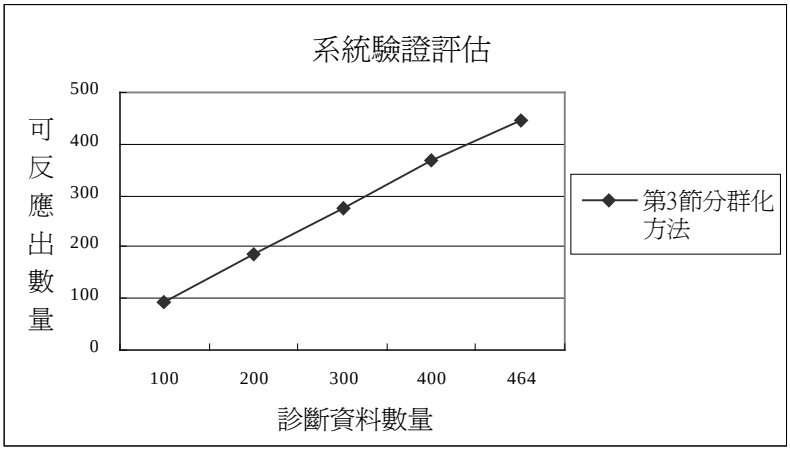

圖5 驗證最可能罹患的疾病之診斷資料數量

5. 結論

病患每次就醫的病歷中會記錄「就診科別」、「患者症狀陳述」、「疾病中文名稱及其對應的ICD-9-CM碼」、及「用藥處方簽」等欄位資料，這些病歷資料中隱藏醫療人員在疾病診斷的智慧、經驗與知識，若能善加管理與運用，必定可以挖掘出很多有用

的知識與資訊。本研究以病患每次就醫的診斷資料為探勘的資料來源，以*k*位病患症狀為探勘的目標，*k*≥1，設計一個方法分群化診斷資料，然後從群組中找出疾病傾向特徵，藉此可發掘病患症狀最可能罹患的疾病。文中所設計的分群化方法，除了保留原先PAM演算法的精神，在每次分群化過程中挑選取代原中心點的診斷資料，也具備群組本身的獨特性，並根據所提出的方法設計與建置一個病患疾病診斷探勘系統。本研究成果對一般民眾於疾病自我檢視、及輔助臨床經驗不足醫療人員的疾病診斷，必定可以提供相當有用的參考資訊。

參考文獻

1. 朱彩屏，2004，資料探勘在醫療資料庫之研究-以疝氣臨床路徑為例，國立中正大學資訊管理研究所碩士論文。
2. 吳素英，2004，資料探勘技術應用於知識管理系統之建構—以醫院疾病分類管理為例，國立中正大學資訊管理研究所碩士論文。
3. 唐壽生，2004，資料探勘技術應用於肺結核病患完治的預測，國立中正大學資訊管理研究所碩士論文。
4. 陳迪祥，2003，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。
5. 俞旭昇，2002，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。
6. Alsabti, K., Ranka, S. and Singh, V., “An Efficient K-Means Clustering Algorithm,” Proceedings of the PPS/SPDP Workshop on High Performance Data Mining, 1997.
7. Berry, M. J. A. and Linoff, G. S., Data Mining Techniques for Marketing, Sales, and Customer Support, New York: John Wiley, 1997.
8. Chen, M. S., Han, J. and Yu, P. S., “Data Mining: an Overview from a Database Perspective,” IEEE Transactions on Knowledge and Data Engineering, 8(6), 1996, pp. 866-883.
9. Dubes, R. C. and Jain, A. K., Algorithms for Clustering Data, Prentice Hall, 1988.
10. Han, J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd Ed., Morgan Kaufmann, 2006.
11. Kaufman, L. and Rousseeuw, P. J., Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.
12. Ng, R. T. and Han, J., “Efficient and Effective Clustering Methods for Spatial Data Mining,” Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 144-155.
13. Ye, X. and Keanem J. A., “Mining Association Rules with Composite Items,” Proceedings of IEEE Internaltional Conference on Systems, Man, and Cybernetics, 1997.

Using Clustering Techniques to Build Mining System of Diagnoses Diseases for Patients

Chui-Cheng Chen[1, *] Ming-Xian Yang[1] Tsung-Yi Chen[2] Jing-Ping Lee[3]
[1]Institute of Information Management, Southern Taiwan University
[2]Department of Electronic Commerce Management, Nanhua University
[3]Electronic Business Program, Southern Taiwan University
*E-mail: ccchen@mail.stut.edu.tw

Abstract

This paper uses diagnostic data as the source of mining, and each diagnostic data includes a patient's diagnosed diseases and symptoms. Clustering technique in data mining is used to analyze tendentiousness of a patient's diagnosed diseases. Let *k* patient as the target of mining, $k \geq 1$, we present a clustering method to cluster diagnostic data to *k* groups with the maximum similarity of symptoms. The most possible diagnosed diseases of the patients' symptoms are found from each group. According to the presented method, a mining system of diagnoses diseases for patients is designed and built. The results of mining can provide very useful information for self-diagnose diseases of people and diagnose diseases of inexperience hospital staffs.

Keywords: data mining, clustering, diagnostic data, disease