輔仁管理評論

中華民國 99 年 1 月，第十七卷第一期，121-142

疾病診斷異常之偵測：關聯規則之應用

陳垂呈*

(收稿日期：98 年 8 月 17 日；第一次修正：98 年 11 月 18 日；
接受刊登日期：99 年 1 月 7 日)

摘要

本研究以診斷資料爲探勘的資料來源、及以某一病患爲探勘的目標，利用資料探勘的關聯規則分別從以下兩方面偵測病患的疾病診斷是否異常：一是設計一個快速探勘關聯規則的方法，並且關聯規則的前置項目組必須包含於此病患症狀中，根據關聯規則所顯示出的傾向特徵，可判斷此病患是否具有疾病診斷異常的傾向；二是設計一個快速探勘關聯規則的方法，並且關聯規則的前置項目組必須包含於此病患的診斷疾病中，根據關聯規則所顯示出的傾向特徵，可判斷此病患是否具有症狀問診異常的傾向。依據文中所提出之方法，我們設計與建置一個偵測疾病異常診斷的探勘系統。此探勘結果，對臨床經驗不足之醫療人員可以對其避免診斷的疏忽，可以提供非常有用的參考資訊。

關鍵詞彙：資料探勘，關聯規則，疾病，症狀，異常診斷

壹・導論

隨著資訊科技的發展與應用，醫療院所儲存病患的診斷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」。從過去病患的診斷資料中，找出外在顯示症狀與引發疾病之間的關聯性，做爲醫療診斷的參考資訊，並提升醫療的準確性及時效性，降低在診斷疾病過程中的疏忽，是利用診斷資料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，對企業決策分析可提供相當有用的參考資訊，資料探勘技術目前已普遍地應用在許多的領域中 (Han and Kamber, 2006)。本研究將以病患之診斷資料做爲探勘的資料來源，每一筆診斷資料記錄有病患症狀與罹患的疾病，並以某一病患的症狀和診斷罹患的疾病做爲探勘目標，利用關聯規則 (association rules) 發掘

* 作者簡介：陳垂呈，南台科技大學資訊管理系副教授。

那些症狀與那些疾病具有高度的關聯性，以做爲檢核病患疾病診斷是否具有異常傾向的依據。

假設欲探勘的病患症狀爲 X、且診斷罹患疾病爲 Y，X 爲包含一個或以上的症狀項目、Y 爲包含一個或以上的疾病項目，本研究分別從以下兩方面探討此病患疾病診斷是否具有異常的傾向：

1. 偵測病患之疾病診斷是否異常：文中以此病患症狀爲探勘的目標，在偵測的過程中，設計一個快速探勘前置項目組包含於 X 之關聯規則的方法，關聯規則以 $S_1 \rightarrow D_1$ 表示之，其中 $S_1 \subseteq X$，S_1 爲包含一個或以上的症狀項目、D_1 爲包含一個或以上的疾病項目。由於 $S_1 \subseteq X$，顯示以上關聯規則可表現出此病患症狀之罹患疾病的傾向。從關聯規則所顯示出的傾向特徵，找出此病患症狀最可能罹患的疾病項目，並藉此計算診斷此病患罹患的疾病與探勘出最可能罹患之疾病間的相似度，以做爲判斷此病患是否具有疾病診斷異常的依據。

2. 偵測病患之症狀問診是否異常：文中以此病患之診斷疾病爲探勘的目標，設計一個快速探勘前置項目組包含於 Y 之關聯規則的方法，關聯規則以 $D_2 \rightarrow S_2$ 表示之，其中 $D_2 \subseteq Y$，D_2 爲一個或以上的疾病項目、S_2 爲一個或以上的症狀項目。由於 $D_2 \subseteq Y$，顯示以上關聯規則可表現出罹患此疾病所顯示之症狀的傾向性。從關聯規則所顯示的傾向特徵，找出此病患之診斷疾病最可能顯示的症狀項目，並藉此計算此病患症狀與探勘出最可能顯示症狀項目之間的相似度，以做爲判斷此病患是否具有症狀問診異常的依據。

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爲例，設計與建置一個病患疾病診斷的偵測系統。本研究的探勘結果，對臨床經驗不足的醫療人員，提供或輔助其在疾病診斷異常的預警，對降低醫療人員在醫療診斷過程中的疏忽，可以提供非常有用的參考資訊。

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；第參節中設計一個探勘的方法，偵測病患是否具有疾病診斷異常的傾向；第肆節中設計一個探勘的方法，偵測病患是否具有症狀問診異常的傾向；第伍節中依據所提出的方法，設計與建置一個病患疾病診斷的偵測系統，並以南部某一醫學中心的診斷資料爲例，評估所設計之方法的效益；最後，在第陸節中做一結論。

貳・相關研究

隨著資訊技術在醫學上的應用而發展出醫學資訊學 (medical informatics)，其目的是利用資訊技術的輔助，並以病患爲中心、醫療問題爲導向的診斷模式，希望藉由資訊技術的支援來建立醫學知識，進而找出各種疾病的醫療指引 (朱彩屏，2004)。若能有效利用資訊技術於疾病診斷上，做爲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)、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技術可以有效地應用在醫療診斷中，其相關研究有：俞旭昇 (2002) 透過資料探勘的技術，以標準健保資料作爲系統資料的來源，建構出一套醫療領域專門的資料探勘系統，以探究不同疾病之間的關係，可提供未來預防治療的參考；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；吳素英 (2004) 利用資料探勘技術以建構出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；黃勝崇 (2001) 利用關聯規則及分類分析找出病患症狀與疾病之間的關聯性，藉此輔助指引病患就醫、或是協助醫療人員診斷的參考；陳世源 (2000) 從病歷資料中，藉由資料探勘技術找出病例與用藥之間的關聯性，防止健保制度用藥浮濫的問題；吳國禎 (1999) 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分析，藉以證明資料探勘技術可以有效應用於輔助醫療診斷中，甚至診斷的準確率高過人爲的診斷；潘雅雪 (2007) 利用資料探勘技術建構關於精神疾病、糖尿病與腎臟疾病的診斷評估模式。Ye and Keane (1997) 曾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；Lin et al. (2009) 提出一探勘模式可找出關於冠狀心臟疾病資料中較有意義的關聯規則；Palaniappan and Awang (2008) 利用資料探勘技術建構一個智慧型心臟疾病偵測系統：Tsipouras et al. (2008) 提出一個以模糊規則爲基礎 (fuzzy rule-based) 的決策支援系統可自動診斷冠狀動脈疾病。

Agrawal et al. (1993) 首先提出從交易資料中擷取關聯規則來顯示項目之間的關聯性，關聯規則的定義說明如下：假設 I 是所有項目的集合，T 是全部交易資料的集合，一筆交易資料 T_j，$T_j \in T$，是由一個或以上項目所組成的集合，

稱之爲項目組 (itemsets)，若一個項目組包含有 k 個項目，稱之爲 k-項目組 (k-itemsets)，以 $itemset_k$ 表示之，$k \geq 1$。在項目組 X 與 Y 之間有一關聯規則被表示成 $X \rightarrow Y$，X、$Y \subseteq I$ 且 $X \cap Y = \varnothing$，其中 X 稱之爲前置項目組 (antecedent)，而 Y 稱之爲後置項目組 (consequent)。有兩個參數 s 與 c 分別爲支持度 (support) 與信賴度 (confidence)，用來決定關聯規則是否成立。支持度 s 的定義爲：在所有的交易資料中，同時包含有 $X \cup Y$ 的比率值，即 s= (包含有 $X \cup Y$ 之交易資料的數量) / (總交易數量)；信賴度 c 的定義爲：在包含有 X 的交易資料中，也同時包括有 Y 的比率值，即 c= (包含有 $X \cup Y$ 之交易資料的數量) / (包含有 X 之交易資料的數量)。擷取出來的關聯規則，其支持度與信賴度必須大於或等於所指定的最小支持度與最小信賴度，這樣的關聯規則才成立。

探勘關聯規則的過程中主要分成以下兩個階段：首先，找出滿足最小支持度的所有項目組，這些滿足最小支持度的項目組稱之爲高頻項目組 (frequent itemsets)，若某 k-項目組滿足最小支持度，即稱之爲高頻 k-項目組 (frequent k-itemsets)，$k \geq 1$，以 $frequent_k$ 表示之；然後，根據前階段所找出的高頻項目組及以最小信賴度爲條件，計算出所有符合的關聯規則。例如 ABC 爲高頻 3-項目組，假如關聯規則 AB→C 滿足最小信賴度，則此關聯規則成立。

探勘關聯規則的相關研究有：Agrawal and Srikant (1994) 提出一個 Apriori 演算法探勘關聯規則，可減少候選項目組的產生進而提昇執行效率；之後有許多研究提出不同儲存項目組的資料結構，以提昇搜尋項目組的執行效能，其相關研究可參考 Park et al. (1997)、Agarwal et al. (2000)、Han et al. (2004)、Coenen et al. (2004)、Da Silva Camargo and Martins Engel (2002)、Holt and Chung (2002)、Liu et al. (2004)、Li et al. (2005)、Tsay and Chiang (2005)、及 Tsay et al. (2004)。在眾多探勘關聯規則的方法中，Apriori 演算法是最具代表性的方法之一，其具有容易了解與實作的優點，也是後續相關研究中最常被使用來比較評估效能優劣的演算法之一。以下說明 Apriori 演算法擷取關聯規則的步驟：

1.找出 $frequent_{k-1}$，$k>1$，若爲∅，則停止執行。

2.由步驟 1 中組合任兩個有 k-2 項目相同的 $frequent_{k-1}$，形成 $itemset_k$。

3.判斷由步驟 2 所找出的 $itemset_k$ 其所有包括的 $itemset_{k-1}$ 之子集合是否都出現在步驟 1 中，若成立就保留此 $itemset_k$，否則就刪除。

4.再檢查由步驟 3 所擷取的 $itemset_k$ 是否滿足最小支持度，假如符合就成爲 $frequent_k$，否則就刪除。

5.計算 $frequent_k$ 所形成的關聯規則，若滿足最小信賴度，則關聯規則成立。

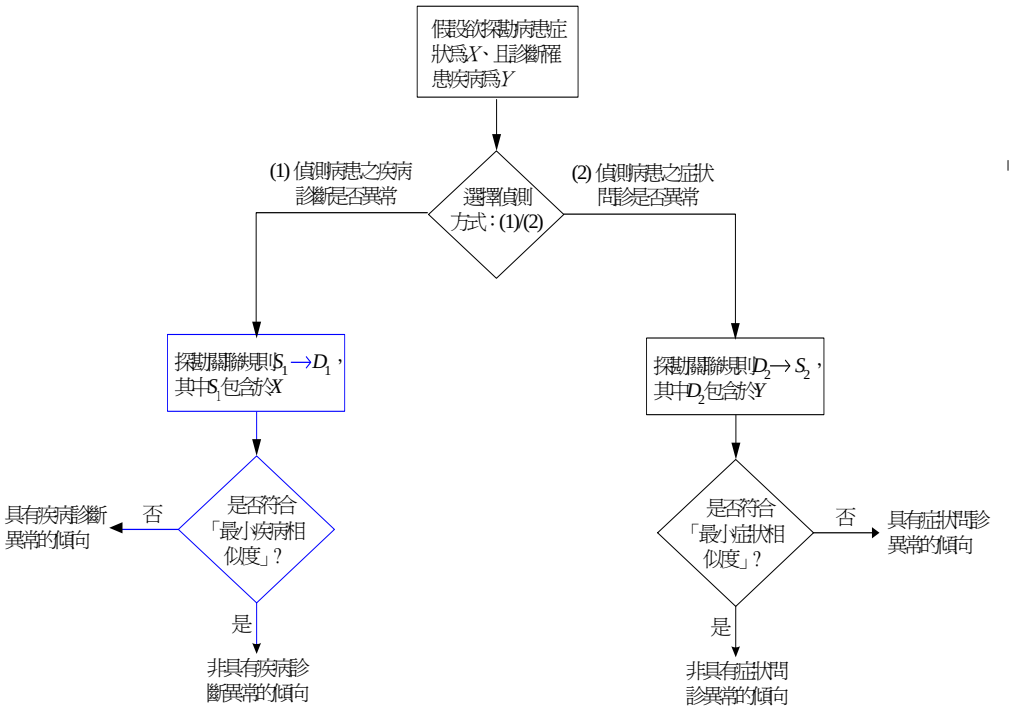
6.跳至步驟 1 找 $frequent_{k+1}$，直到無法產生高頻項目組為止。

本研究以病患每次就醫時之診斷資料為探勘的資料來源，並以某一病患症狀之診斷疾病為探勘的目標，利用關聯規則探討病患疾病診斷是否具有異常的傾向。文中定義診斷資料的格式為[*A*, *B*]，*A* 為包含一個或以上的症狀項目，*B* 為包含一個或以上的疾病項目，即顯示有 *X* 症狀、及罹患有 *Y* 疾病，如表一。每一筆診斷資料包含病患所顯示出的症狀項目、及診斷罹患的疾病項目。

表一　診斷資料格式

診斷資料編號	症狀項目	疾病項目

假設欲探勘的病患症狀為 *X*、且診斷罹患疾病為 *Y*，*X* 為包含一個或以上的症狀項目、*Y* 為包含一個或以上的疾病項目，本研究探討此病患疾病診斷是否具有異常的傾向，其探勘過程如圖一流程圖。



圖一　探勘病患疾病診斷是否具有異常傾向的流程圖

參・偵測病患疾病診斷是否異常

本研究以病患每次就醫之診斷資料爲探勘的資料來源，並以某一病患症狀之診斷疾病爲偵測的目標，探勘關聯規則其前置項目組包含於此病患症狀中，以判斷此病患是否具有疾病診斷異常的傾向。

一、探勘方法

假設欲探勘之病患症狀爲 X、且被診斷罹患疾病爲 Y，X 爲包含一個或以上的症狀項目、Y 爲包含一個或以上的疾病項目，文中必須找出以下形式的關聯規則：

$S_1 \rightarrow D_1$, $S_1 \subseteq X$，S_1 爲包含一個或以上的症狀項目、D_1 爲包含一個或以上的疾病項目，$S_1 \cup D_1$ 是高頻項目組。

其顯示出的傾向爲：若病患顯示症狀 S_1，則會有罹患疾病 D_1 的傾向。由於 $S_1 \subseteq X$，表示以上關聯規則可表現出此病患症狀之罹患疾病的傾向特徵，且若 S_1 愈相近於 X，則關聯規則愈能反映出此病患症狀之罹患疾病的傾向特徵，其罹患疾病爲 D_1 的傾向性也愈強。我們藉由關聯關規則的傾向特徵，可做爲判斷此病患是否具有疾病診斷異常傾向的依據。

爲了配合探勘的需要及避免計算與 X 無關的項目組，我們修改 Apriori 演算法，直接組合 X 中的症狀項目與疾病項目而形成項目組，並判斷這些項目組是否爲高頻項目組。探勘的過程說明如下：

1.從原始診斷資料庫 D_1 中，找出 X 中及疾病項目中的 $frequent_1$，而且必須至少各包含有一項，否則停止執行。若診斷資料未包含 X 中任一症狀項目，則刪除之，並形成新的診斷資料庫 D_2。

2.由步驟 1 中，組合包含於 X 中之任一 $frequent_1$ 與包含於疾病項目中之任一 $frequent_1$ 而形成 $itemset_2$。從診斷資料庫 D_2 中檢查 $itemset_2$ 是否滿足最小支持度，假如符合就成爲 $frequent_2$，否則就刪除。在掃瞄診斷資料庫 D_2 的過程中，若診斷資料未包含 $itemset_2$，則刪除之，並形成新的診斷資料庫 D_3。

3.找出所有的 $frequent_{k-1}$，$k>2$，並形成新的診斷資料庫 D_k。

4.由步驟 3 中，組合任兩個有 k-2 項目相同的 $frequent_{k-1}$，形成 $itemset_k$。

5.判斷由步驟 4 所找出的 $itemset_k$，其所包括的所有子集合 $itemset_{k-1}$ 是否都有出現在步驟 3 中 (若 $itemset_{k-1} \cap X = \varnothing$、或 $itemset_{k-1} \cap$ 疾病項目$=\varnothing$，則不列入考慮)，假如成立就保留此 $itemset_k$，否則就刪除。

6.從診斷資料庫 D_k 中檢查由步驟 5 所找出的 $itemsert_k$ 是否滿足最小支持度，假如符合就成爲 $frequent_k$，否則就刪除。在掃瞄診斷資料庫 D_k 的過程中，若診斷資料未包含 $itemset_k$，則刪除之，並形成新的診斷資料庫 D_{k+1}。

7.計算 $frequent_k$ 所形成的關聯規則，其形式爲：

$$S_1 \rightarrow D_1,\ S_1 \subseteq X，\{S_1 \cup D_1\} \in frequent_k。$$

8.跳至步驟 3 繼續找出 $frequent_{k+1}$，直到無法產生高頻項目組爲止。

從以上演算法的步驟 2 開始所擷取的 $frequent_k$，必定爲 $frequent_k \cap X \neq \varnothing$，計算高頻項目組所形成的關聯規則 $S_1 \rightarrow D_1$、且 $S_1 \subseteq X$，若滿足最小信賴度，則關聯規則成立。藉由關聯規則 $S_1 \rightarrow D_1$ 所顯示出的傾向特徵，文中做成以下定義：

關聯規則 $S_1 \rightarrow D_1$，且 $S_1 \subseteq X$，則 D_1 爲此病患症狀最可能傾向罹患的疾病，且 S_1 愈相近於 X，則罹患疾病 D_1 的傾向性也愈強。

然後計算此病患症狀最可能罹患的疾病與診斷疾病 Y 之間的疾病相似度，其定義如下：

找出 S_1 最相近於 X 的關聯規則 $S_1 \rightarrow D_1$，疾病相似度=$\{(D_1 \cap Y)$ 的項目個數$\}$ / Y 的項目個數。

若計算出的疾病相似度滿足所給定的臨界值，則顯示此病患爲非具有疾病診斷異常的傾向；否則顯示出具有疾病診斷異常的傾向。

在探勘的計算過程中，文中只擷取高頻項目組 $frequent_k$、且 $frequent_k \cap X \neq \varnothing$，如此將避免計算未包含 X 任一項的項目組。在每次掃瞄診斷資料庫以判斷 $itemset_k$，$k>1$，是否爲 $frequent_k$ 的過程中，藉由刪除未包含 $itemset_k$ 的診斷資料，將可大幅減少診斷資料數量。基於以上計算的改進，以文中所設計的探勘方法將比原先 Apriori 演算法，會更有效率找到所要的關聯規則「症狀 (包含於欲探勘的病患症狀))→疾病」。

二、實例說明

表二爲一診斷資料庫 D_1，其包含 6 筆的診斷資料，其中{A, B, C, D, E}爲症狀項目的集合，{X, Y, Z}爲疾病項目的集合，{T_1, T_2, T_3, T_4, T_5, T_6}爲診斷資料的集合。假設欲探勘的病患症狀爲 BE、且被診斷罹患疾病爲 X，設定最小支持度爲 30% (即最小支持數量爲 1.8)、最小信賴度爲 50%、及疾病相似度臨界值爲 60%。以下說明偵測此病患是否具有疾病診斷異常的傾向。

表二　診斷資料庫 D_1

診斷資料編號	症狀項目	疾病項目
T1	ABD	XZ
T2	BE	X
T3	ACE	Y
T4	AC	Y
T5	BCE	XY
T6	B	X

擷取包含 BE 中任何項目之高頻項目組的過程如下：

$itemset_1$
B
E
X
Y
Z

掃瞄 D_1 →

$itemset_1$	出現次數
B	4
E	3
X	4
Y	3
Z	1

≥1.8 →

$frequent_1$	出現次數
B	4
E	3
X	4
Y	3

刪除未包含 BE 中任一症狀項目的診斷資料 →

診斷資料庫 D_2

診斷資料編號	症狀項目	疾病項目
T_1	ABD	XZ
T_2	BE	X
T_3	ACE	Y
T_5	BCE	XY
T_6	B	X

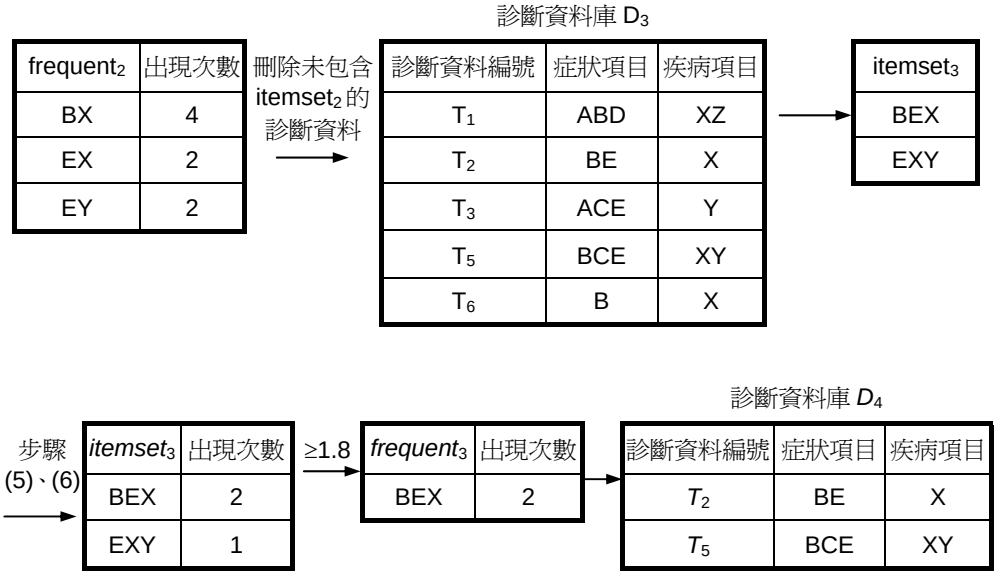
→

$itemset_2$
BX
BY
EX
EY

掃瞄 D_2 →

$itemset_2$	出現次數
BX	4
BY	1
EX	2
EY	2

≥1.8 →



無 4-項目組。

以高頻 3-項目組 BEX 爲例，計算形成的關聯規則 BE→X，其信賴度爲 2/2=100%，滿足最小信賴度，關聯規則成立，並且此關聯規則之前置項目組的症狀項目是最相近於此病患症狀。此關聯規則所顯示的傾向特徵爲：若病患症狀爲 BE，則會有罹患疾病 X 的傾向。然後計算疾病相似度={(X∩X) 的項目個數}/1=100%，滿足所給定的疾病相似度臨界值。經由以上計算的結果，可判斷出此病患爲非具有疾病診斷異常的傾向。

肆・偵測病患症狀問診是否異常

病患就診的過程中，醫療人員大多會以病患描述或顯示之症狀做爲初步診斷的依據，因此醫療人員是否能確實清楚了解病患的症狀，往往是影響其診斷罹患疾病是否正確的重要因素之一。此章節仍以病患每次就醫之診斷資料爲探勘的資料來源，並以某一病患之診斷疾病爲偵測的目標，探勘關聯規則其前置項目組包含於此病患的診斷疾病中，以判斷此病患是否具有症狀問診異常的傾向。

一、探勘方法

假設欲探勘之病患症狀爲 X、且被診斷罹患疾病爲 Y，X 爲一個或以上的症狀項目、Y 爲一個或以上的疾病項目，文中必須找出以下形式的關聯規則：

$D_2 \rightarrow S_2$, $D_2 \subseteq Y$，D_2 爲一個或以上的疾病項目、S_2 爲一個或以上的症狀項目，$D_2 \cup S_2$ 是高頻項目組。

以上關聯規則所顯示的傾向特徵爲：若病患被診斷罹患疾病爲 D_2，則會有顯示症狀爲 S_2 的傾向。由於 $D_2 \subseteq Y$，表示此關聯規則可表現出此病患罹患疾病所顯示之症狀的傾向特徵，且當 D_2 愈相近於 Y，則關聯規則愈能反映出此病患罹患疾病所顯示之症狀的傾向特徵，其顯示症狀 S_2 的傾向性也愈強。因此，藉由以上形式之關聯規則所顯示的傾向特徵，可做爲判斷此病患是否具有症狀問診異常傾向的依據。

爲了配合探勘的需要及避免計算與 Y 無關的項目組，文中修改前一節中所描述的演算法，直接組合 Y 中的疾病項目與症狀項目而形成項目組，並判斷這些項目組是否爲高頻項目組。探勘的過程說明如下：

1.從原始診斷資料庫 D_1 中，找出 Y 中及症狀項目中的 $frequent_1$，而且必須至少各包含有一項，否則停止執行。若診斷資料未包含 Y 中任一疾病項目，則刪除之，並形成新的診斷資料庫 D_2。

2.由步驟 1 中，組合包含於 Y 中之任一 $frequent_1$ 與包含於症狀項目中之任一 $frequent_1$ 而形成 $itemset_2$。從診斷資料庫 D_2 中檢查 $itemset_2$ 是否滿足最小支持度，假如符合就成爲 $frequent_2$，否則就刪除。在掃瞄診斷資料庫 D_2 的過程中，若診斷資料未包含 $itemset_2$，則刪除之，並形成新的診斷資料庫 D_3。

3.找出所有的 $frequent_{k-1}$，$k>2$，並形成新的診斷資料庫 D_k。

4.由步驟 3 中，組合任兩個有 k-2 項目相同的 $frequent_{k-1}$，形成 $itemset_k$。

5.判斷由步驟 4 所找出的 $itemset_k$，其所包括的所有子集合 $itemset_{k-1}$ 是否都有出現在步驟 3 中 (若 $itemset_{k-1} \cap Y = \varnothing$、或 $itemset_{k-1} \cap$ 症狀項目$=\varnothing$，則不列入考慮)，假如成立就保留此 $itemset_k$，否則就刪除。

6.從診斷資料庫 D_k 中檢查由步驟 5 所找出的 $itemsert_k$ 是否滿足最小支持度，假如符合就成爲 $frequent_k$，否則就刪除。在掃瞄診斷資料庫 D_k 的過程中，若診斷資料未包含 $itemset_k$，則刪除之，並形成新的診斷資料庫 D_{k+1}。

7.計算 $frequent_k$ 所形成的關聯規則，其形式爲：

$$D_2 \rightarrow S_2,\ D_2 \subseteq Y，\{D_2 \cup Y_2\} \in frequent_k。$$

8.跳至步驟 3 繼續找出 $frequent_{k+1}$，直到無法產生高頻項目組爲止。

從以上演算法的步驟 2 開始所擷取出的 $frequent_k$，必定爲 $frequent_k \cap Y \neq \varnothing$，計算高頻項目組所形成的關聯規則 $D_2 \rightarrow S_2$、且 $D_2 \subseteq Y$，若滿足最小信賴度，則關聯規則成立。藉由關聯規則 $D_2 \rightarrow S_2$ 所顯示出的傾向特徵，文中做成以下定義：

關聯規則 $D_2 \rightarrow S_2$，且 $D_2 \subseteq Y$，則 S_2 爲此病患疾病最可能傾向顯示的症狀，當 D_2 愈相近於 Y，則顯示症狀爲 S_2 的傾向性也愈強。

藉由以上關聯規則所顯示的傾向特徵，計算此病患疾病最可能顯示的症狀與問診症狀 X 之間的症狀相似度，其定義如下：

找出 D_2 最相近於 Y 的關聯規則 $D_2 \rightarrow S_2$，症狀相似度= $(S_2 \cap X)$ 的項目個數 /X 的項目個數。

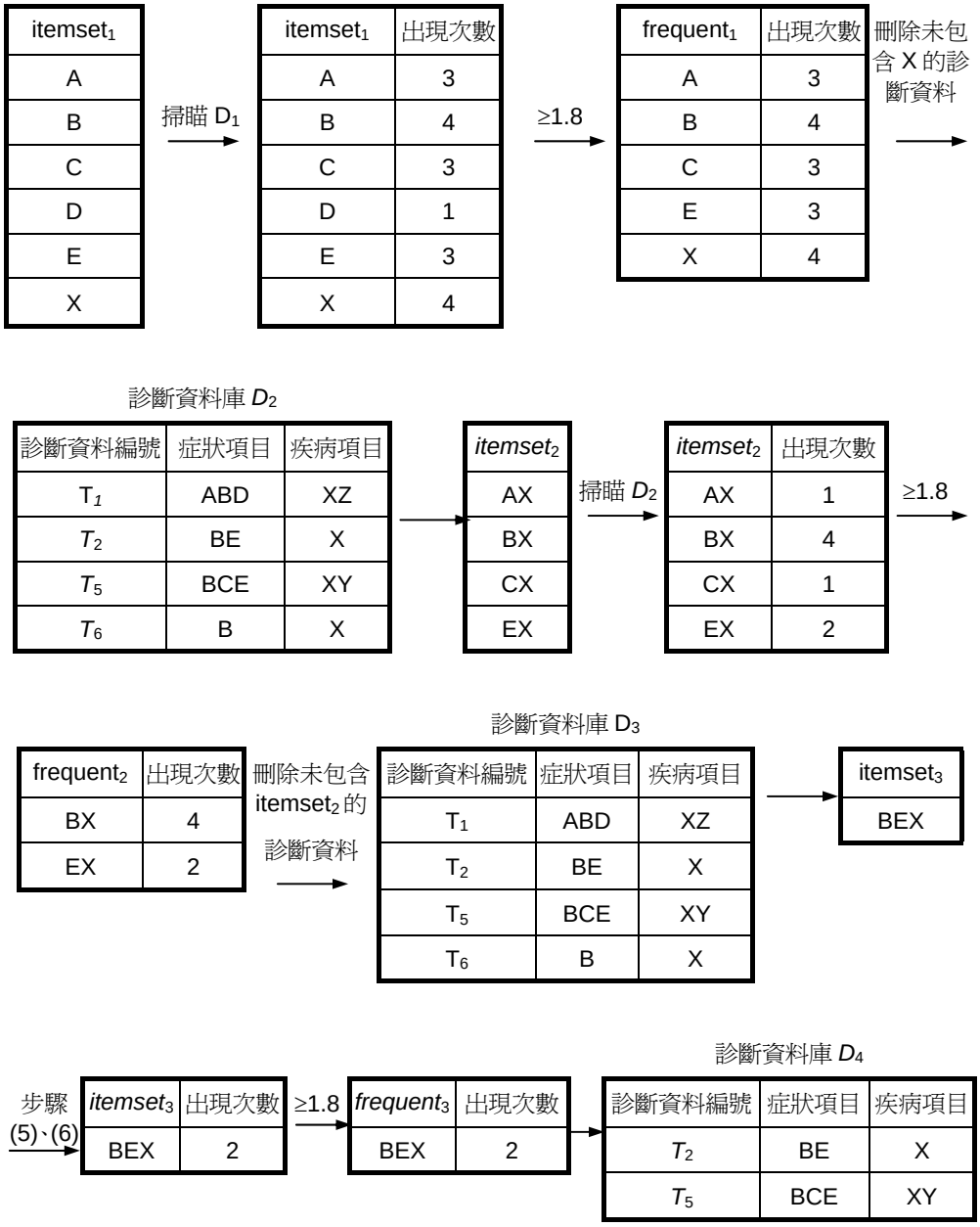
若計算出的症狀相似度滿足所給定的臨界值，則顯示此病患疾病的症狀問診，符合診斷資料庫中所找出的顯示症狀傾向，即定義以上情況爲非具有症狀問診異常的傾向；否則定義爲具有症狀問診異常的傾向。

在探勘的計算過程中，文中只擷取高頻項目組 $frequent_k$，且 $frequent_k \cap Y \neq \varnothing$，如此將避免計算未包含 Y 任一項的項目組。在每次掃瞄診斷資料庫以判斷判斷 $itemset_k$，$k>1$，是否爲 $frequent_k$ 的過程中，藉由刪除未包含 $itemset_k$ 的診斷資料，將可大幅減少診斷資料數量。基於以上計算的改進，文中所設計的探勘方法將比原先 Apriori 演算法，會更有效率找到所要的關聯規則「疾病 (包含於欲探勘之病患被診斷的罹患疾病)→症狀」。

二、實例說明

文中仍以表 2 之診斷資料庫 D_1 爲例，假設欲偵測病患症狀爲 BE、且被診斷罹患疾病爲 X，設定最小支持度爲 30% (即最小支持數量爲 1.8)、最小信賴度爲 50%、及症狀相似度臨界值爲 60%。以下說明偵測此病患是否具有症狀問診異常的傾向。

擷取包含 X 之高頻項目組的過程如下：



無 4-項目組。

以高頻 3-項目組 BEX 爲例，計算形成的關聯規則 X→BE，其信賴度爲 2/4=50%，滿足最小信賴度，關聯規則成立，並且此關聯規則之前置項目組爲疾病 X。此關聯規則所顯示的傾向特徵爲：若病患罹患疾病爲 X，則會有顯示症狀 BE 的傾向。然後計算症狀相似度={(BE∩BE) 的項目個數}/2=100%，滿足所給定的症狀相似度臨界值。因此，經由以上計算的結果，可判斷出此病患

爲非具有症狀問診異常的傾向。

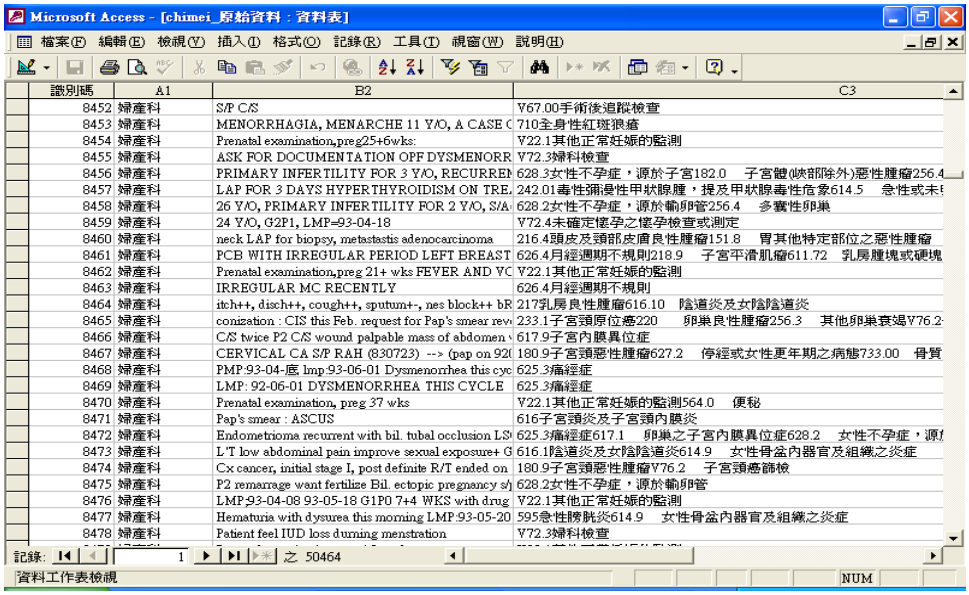
伍・病患疾病診斷偵測系統

文中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的偵測系統，表三爲偵測系統的開發平台。

表三　系統開發平台

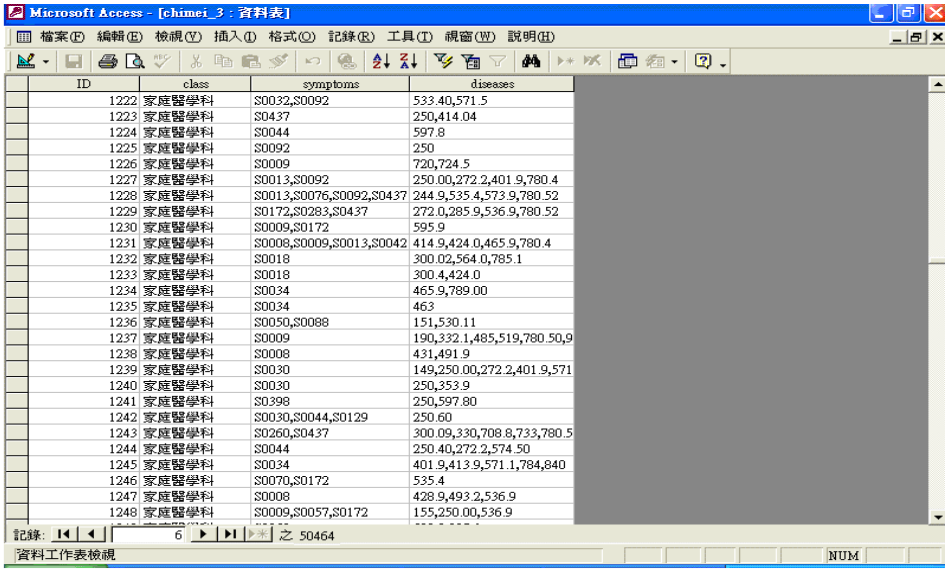
作業系統	Windows XP Professional Edit
CPU	AMD K-7 1.3GHz
主記憶體	512M SDRAM
設計語言	ASP、VB Script、Java Script
資料庫	Microsoft Access 2002

本研究以南部某一醫學中心之病患每次就醫的診斷資料爲例，診斷資料從 2004/4/1 到 2004/4/7 共計 50464 筆，以做爲所設計之探勘方法的資料來源。以前面 50000 筆之診斷資料做爲探勘計算的訓練資料，然後以最後 464 筆診斷資料做爲探勘計算的驗證資料。圖二爲診斷資料的原始資料，這些原始資料是以病患每次就醫爲一個記錄儲存，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。



圖二　原始診斷資料

首先必須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 碼爲 493.9。另外，在探勘計算之前須從症狀描述中篩選出較重要的症狀字詞並進行編碼，分別以 S0001, S0002, S0003 等依次進行編碼。在圖三中分別以編碼後之疾病碼及症狀碼來替代原始診斷資料中的疾病名稱及症狀描述。



圖三　編碼後的診斷資料

本研究以前面 50000 筆的診斷資料做爲探勘的訓練資料，探勘出某些症狀項目最可能罹患的疾病項目、及某些疾病項目最可能顯示的症狀項目，以下說明所建置的病患疾病診斷偵測系統，其在探勘訓練資料的執行過程。在偵測過程中爲了讓使用者了解各參數設定值的意義，以期輸入較有效度的值，因此我們分別增加解釋的文字方塊，其內容分別爲：「最小支持度」的文字方塊說明是「值的大小會影響診斷資料中所找出的樣本數目，可依專業判斷設定所需要的最小支持度值，進而提昇偵測的客觀性」、「最小信賴度」的文字方塊說明是「值的大小會影響症狀與疾病之間的關聯強度，可依專業判斷設定所需要的最小信賴度值，進而提昇偵測的客觀性」、及「相似度臨界值」的文字方塊說明是「值的大小會影響評估探勘結果與診斷疾病 (或是問診症狀) 之間的精確性，可依專業判斷設定所需要的相似度臨界值，進而提昇偵測的客觀性」。

圖四爲點選「偵測病患診斷的罹患疾病是否異常」功能的探勘畫面，分別在「病患症狀」欄位中輸入欲探勘之病患所顯示出的症狀項目、在「診斷疾

病」欄位中輸入所診斷罹患的疾病項目、分別在「最小支持度」及「最小信賴度」欄位中輸入數值，並在「相似度臨界值」欄位中填入數值，以做爲判斷計算疾病相似度的臨界值。經由第參節所描述之探勘方法的計算過程，可在「偵測結果」欄位中顯示出偵測的結果，如圖五。

圖四　偵測病患疾病診斷是否異常之探勘執行畫面

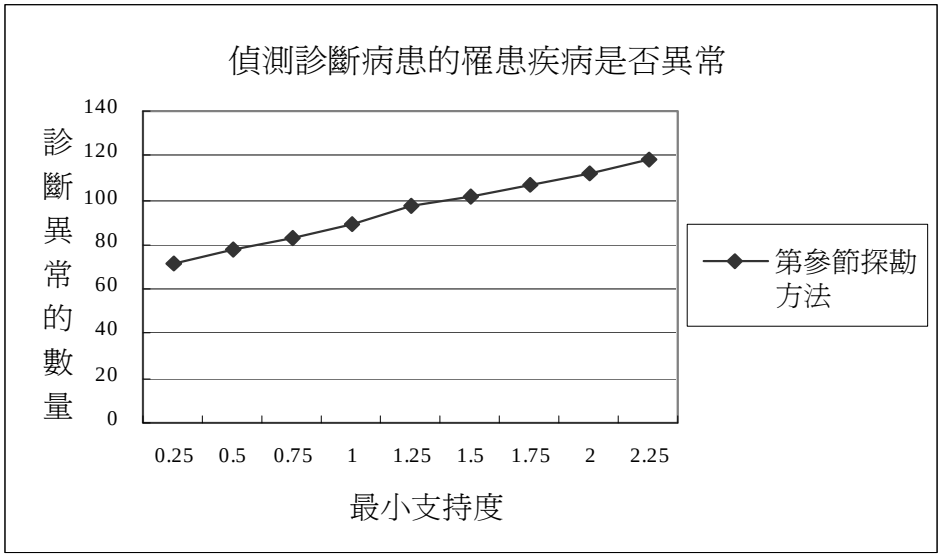
圖五　偵測病患疾病診斷是否異常之探勘結果畫面

圖六爲點選「偵測問診病患的顯示症狀是否異常」功能的探勘畫面，分別在「病患症狀」欄位中輸入病患所顯示的症狀項目、在「診斷疾病」欄位中輸入所診斷的罹患疾病項目、分別在「最小支持度」及「最小信賴度」欄位中輸入數值，並在「相似度臨界值」欄位中填入數值，以做爲判斷計算症狀相似度的臨界值。經由第肆節所描述之探勘方法的計算過程，可在「偵測結果」欄位中顯示出探勘的結果，如圖七。

圖六　偵測問診病患之顯示症狀是否異常的執行畫面

圖七　偵測問診病患之顯示症狀是否異常的結果畫面

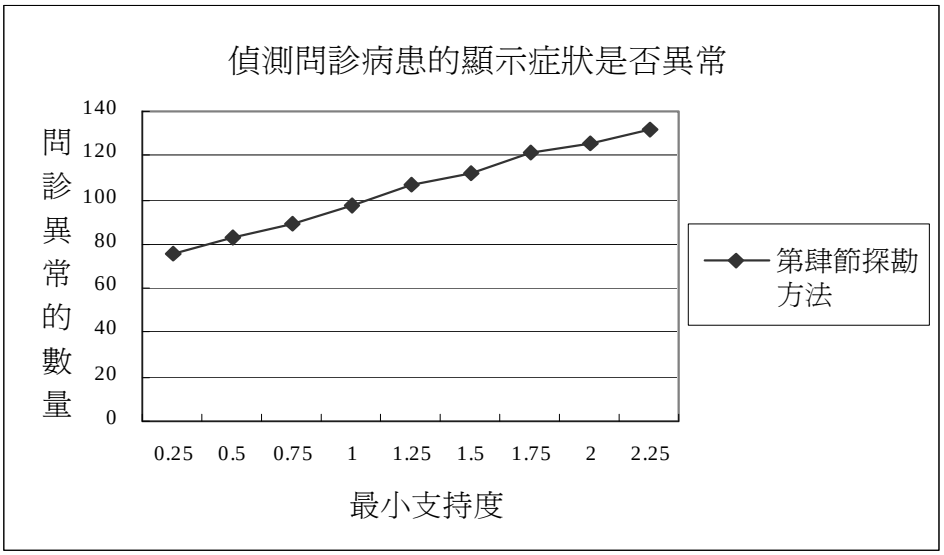
本研究以其餘 464 筆診斷資料做爲偵測的驗證資料，以評估在前面訓練資料中所探勘之結果的成效。在考量病患被診斷之罹患疾病是否具有異常的傾向，利用圖四的偵測執行畫面，從驗證資料中分別輸入每一筆診斷資料的症狀項目、診斷的疾病項目、最小支持度、最小信賴度、及相似度臨界值，然後進行偵測的驗證。在圖八中，設定最小信賴度爲 60%及相似度臨界值 (疾病相似度的臨界值) 爲 60%，評估在不同最小支持度的情況下，出現疾病診斷異常之診斷資料的數量。從圖八中顯示，當最小支持度設定值較小時，偵測系統可探勘出的關聯規則，其可包含較多的症狀項目個數及疾病項目個數，因此偵測出病患症狀所罹患的疾病與診斷疾病之間會有較小的不一致性，其診斷異常出現的比例值也會較小。由於最小支持度設定值的大小，會影響訓練資料中所找出的樣本數目，可藉由專業判斷設定所需要的最小支持度值，進而提昇驗證的客觀性。



圖八　偵測診斷病患之罹患疾病的異常數量

在考量病患被問診之顯示症狀是否具有異常的傾向，利用圖六的偵測執行畫面，從驗證資料中分別輸入每一筆診斷資料的症狀項目、診斷的疾病項目、最小支持度、最小信賴度、及相似度臨界值，然後進行偵測的驗證。在圖九中，設定最小信賴度爲 60%及相似度臨界值 (症狀相似度的臨界值) 爲 60%，評估在不同最小支持度的情況下，出現症狀問診異常之診斷資料的數量。從圖九中顯示，當最小支持度設定值較小時，偵測系統可探勘出的關聯規則，其可包含較多的疾病項目個數及症狀項目個數，因此偵測出病患被診斷罹

患的疾病其顯示症狀與問診症狀之間會有較小的不一致性，其問診異常出現的比例值也會較小。同樣地，最小支持度設定值的大小，會影響訓練資料中所找出的樣本數目，可藉由專業判斷設定所需要的最小支持度值，進而提昇驗證的客觀性。



圖九　偵測問診病患之顯示症狀的異常數量

陸・結論與未來研究

病患就診的醫療過程中，可能因醫療人員疏失而導致誤診是醫療糾紛中最爲常見的因素之一，避免及降低醫療診斷上的疏失，是醫療過程的管理上必須思考及注重的問題。在醫療院所的門診病歷中記錄有「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與知識，若能善加管理與運用，對醫療人員避免於疾病診斷過程中的疏失，必定可以提供相當有用的參考資訊。

本研究以某一病患症狀之診斷疾病做爲偵測的目標，利用關聯規則分別從以下兩方面偵測病患疾病診斷是否具有異常的傾向：一是從診斷資料中找出此病患症狀最可能罹患的疾病項目，藉此判斷此病患是否具有疾病診斷異常的傾向；二是從診斷資料中找出此病患疾病最可能顯示的症狀項目，藉此判斷此病患是否具有症狀問診異常的傾向。在許多利用資料探勘技術於醫療管理的相

關研究中，已顯示資料探勘確實可以有效應用在醫療診斷的輔助上 (陳世源，1999；陳迪祥，2003；朱彩屏，2004；唐壽生，2004；黃勝崇，2001；Ye and Keanem, 1997)。因此，本研究的探勘結果，對醫療人員在診斷病患疾病錯誤的預警、及降低在問診病患症狀的疏忽，都可以提供非常有用的參考資訊。

本研究目前僅以病歷記錄中的「症狀」及「疾病」兩項欄位資料、及如何利用關聯規則偵測病患的疾病診斷和症狀問診，是否具有異常傾向的探勘過程做探討，對於未來的相關研究有：

1.經由醫療臨床的專業實驗，或模擬不同的設定值，以取得「最小支持度」、「最小信賴度」、及「相似度臨界值」等參數較合適的起始建議值。

2.有效分析與運用病歷資料中的其他欄位項目，例如科別、姓別、年齡、及職業等資料。

3.有效擷取病患之症狀陳述中的症狀項目、及對症狀項目進行標準編碼化。

4.探討其他資料探勘技術應用在此研究問題的可行性。

5.改良偵測系統的設計與建置，並進行臨床實際應用的驗證分析。

參考文獻

朱彩屏，「資料探勘在醫療資料庫之研究－以疝氣臨床路徑爲例」，國立中正大學資訊管理研究所碩士論文，2004 年。

吳素英，「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爲例」，國立中正大學資訊管理研究所碩士論文，2004 年。

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 年。

唐壽生，「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，2004 年。

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，2000 年。

陳迪祥，「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，2003 年。

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 年。

黃勝崇，「資料探勘應用於醫療院所輔助病患看診指引之研究」，南華大學資訊管理研究所碩士論文，2001 年。

潘雅雪，「資料探勘技術於疾病診斷之應用」，國立暨南國際大學資訊管理研究所碩士論文，2007 年。

Agarwal, R., Aggarwal, C. & Prasad, V. V. V., "A Tree Projection Algorithm for Generation of Frequent Itemsets", *Journal of Parallel and Distributed Computing*, Vol. 63(3), 2000, pp. 350-371.

Agrawal, R., Imielinski, T. & Swami, A., "Mining Association Rules between Sets of Items in Very Large Database", Proceedings of the ACM SIGMOD Conference on Management of Data, 1993, pp. 207-216.

Agrawal, R. & Srikant, R., "Fast Algorithms for Mining Association Rules in Large Database", Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487-499.

Chen, M. S., Han, J. & Yu, P. S., "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8(6), 1996, pp. 866-883.

Coenen, F., Leng, P. & Ahmed, S., "Data Structure for Association Rule Mining T-trees and P-trees", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16(6), 2004, pp. 774-778.

Da Silva Camargo, S. & Martins Engel, P., "MiRABIT: A New Algorithm for Mining Association Rules", Proceedings of the 22nd International Conference of the Chilean Computer Science Society (SCCC'02), 2002, pp. 162-166.

Han, J. & Kamber, M., "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann, 2006.

Han, J., Pei, J., Yin, Y. & Mao, R., "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", *Data Mining and Knowledge Discovery*, Vol. 8(1), 2004, pp. 53-87.

Holt, J. D. & Chung, S. M.,. "Mining Association Rules Using Inverted Hashing and Pruning", *Information Processing Letters*, Vol. 83, 2002, pp. 211-220.

Li, Z. C., He, P. L. & Lei, M., "A High Efficient AprioriTid Algorithm for Mining Association Rule", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, 2005, pp. 1812-1815.

Lin, Z. K., Yi, W. G., Lu, M. Y., Liu, Z. & Xu, H., "Correlation Research of Association Rules and Application in the Data about Coronary Heart Disease", Proceedings of the International Conference of Soft Computing and Pattern Recognition (SOCPAR '09), 2009, pp. 143-148.

Liu, P. Q., Li, Z. Z. & Zhao, Y. L., "Effective Algorithm of Mining Frequent Itemsets for Association Rules", Proceedings of the Third International Conference on Machine Learning and Cybernetics, 2004, pp. 1447-1451.

Palaniappan, S. & Awang, R., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", Proceedings of IEEE International Conference on Computer Systems and Applications, 2008, 108-115.

Park, J. S., Chen, M. S. & Yu, P. S., "Using a Hash-based Method with Transaction Trimming for Mining Association Rules", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 9(5), 1997, pp. 813-825.

Tsipouras, M. G.., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K. & Michalis, L. K., "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and

Fuzzy Modeling", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 12(4), 2008, pp. 447-458.

Tsay, Y. J. & Chang-Chien, Y. W., "An Efficient Cluster and Decomposition Algorithm for Mining Association Rules", *Information Sciences*, Vol. 160, 2004, pp. 161-171.

Tsay, Y. J. & Chiang, J. Y., "CBAR: An Efficient Method for Mining Association Rules", *Knowledge-Based Systems*, Vol. 18, 2005, pp. 99-105.

Ye, X. & Keanem, J. A., "Mining Association Rules with Composite Items", Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, 1997, pp. 1367-1372.

An Application of Association Rules in Detection of Careless Diagnoses of Diseases

CHUI-CHENG CHEN *

ABSTRACT

This paper uses diagnostic data as the source of mining. We let a patient to be as the target of mining, and use association rules of data mining to detect careless diagnosis of the patient's diseases from two aspects: one is to propose a fast method to mine association rules whose antecedents are contained in the patient's symptoms, and we detect whether the diseases diagnosed is carelessness or not according to the characteristics of the association rules; the other one is to propose a fast method to mine association rules whose antecedents are contained in the patient's diseases diagnosed, and we detect whether the symptoms inquired is carelessness or not according to the characteristics of the association rules. A mining system is designed and constructed to detect careless diagnoses of diseases based on the both methods. The results of detecting can provide very useful information to avoid careless diagnoses of diseases for inexperience hospital staffs.

Keywords: data mining, association rule, disease, symptom, careless diagnosis

* Chui-Cheng CHEN, Associate Professor, Department of Information Management, Southern Taiwan University.