第一章緒論

第一節 前言

隨著科技的進步，醫療院所對病歷書寫的方式，由以往的手書寫方式改為電腦媒體的儲存，因而存有患者病歷的醫學資料庫便產生了。根據醫學資料庫，醫生不但能有效掌握病人的資料（包含病人以往的病史和用藥習慣），同時對於臨床醫學和基礎醫學研究都有很大的幫助，這將對於提升醫療品質而言可說有正面的意義。

一般而言，病患間普遍存在一現象，就是到了醫院不知該向哪一科掛診？往往根據病患自己主觀意識隨意找個醫生，在發現不對勁之後，才重新掛另一門診。因此，許多人不管三七二十一，就直接向家醫科掛號，將家醫科醫生充當為『預診』醫生，由家醫科醫生負責病人的轉診，而研究中認為家醫科醫生不應只具有『預診』功能，還需要負責其它事務的進行，面對這樣重覆性的掛錯門診，造成醫療資源的浪費及病情的延誤，引發想建立起一套『自動預診系統』的動機，可以將病患對症狀的描述，透過資料探勘技術，找尋症狀描述與疾病間的關聯，以達到『預診』功能。

隨著醫療系統廣泛使用使得資料庫中的資料日益增加，也因而所隱藏的資訊也愈來愈豐富。若能從現有的病歷資料將數據作分析處理，相信會對找出病症特徵和疾病間的關聯性有所裨益。但是面對如此巨量的資料，若是單靠人工來分析實難有所成效，為了得到些許有用的資訊，專家學者們致力於使電腦的強大運算能力與醫療資訊相結合的研究，此種找出資料中內隱知識的方法統稱 KDD（Knowledge Discovery in Database），或是資料探勘（Data Mining）。

第二節 研究背景

近年來在資料庫領域中，資料探勘是相當熱門的研究主題。資料探勘的目的是要從資料堆中，發掘一些隱性或未知的資訊、知識，以作為決策輔助或者預測的參考。

目前台灣的醫療院所的持續增加，區域等級以上的醫院比十年前增加了許多。舉例來說，以台南市而言，有二家醫學中心及三家區域教學醫院[1]負責七十幾萬人口的醫療服務，但這五家醫院常出現人潮擁擠的現象，在深入探究其原因後發現，主要是病人求診觀念不正確，其次是缺乏病症的知識，而造成病人需要不斷地輾轉求

醫才能得到正確的治療，因而造成醫藥資源的浪費，也增加患者治療的時間及延誤病情。

由於社會的競爭，人們生活的壓力也相對地增加，所以投入工作的時間就增長，這樣長時間所累積下來的身體不適，造成疾病發生率逐漸增加。然而每個人都有這樣的經驗，就是身體有病痛之後，但自己本身缺乏醫學常識，而無法到正確的科別求診，之後還歷經幾次的轉診後才獲得正確的診治，但轉診或者複診期間病情是不等人的，所以增加了病情加劇的風險，也讓自己跟國家負擔額外的醫療成本。但也曾有過相同的病人給不同醫生診斷時，其診斷結果有差異的情形，甚至同一個醫師在不同時段診斷，其結果有時也會有差異，這牽涉到當時醫師的精神、體力、情緒等因素，所造成的誤差，這樣的誤差是可以將其避免的。所以，開始有專家希望透過電腦，將過去醫生對病症所做的診斷記錄儲存於資料庫中，並讓電腦將資料庫中所收集的資料經過篩選、分析、統計及學習後，轉換成診斷病症時有效的知識。因此，本研究希望利用電腦的輔助，幫助病人及其家屬依病情症狀的特徵，自行判斷自己所罹患的疾病，不但能讓患者快速獲得正確治療外，也可輔助醫生做診斷，以提高醫生的診斷正確率。同時也減少許多醫藥資源的浪費，讓全民健康保

險能照顧到每個需要照顧的人。

第三節 研究目的

在前一節探討中，我們可以知道醫學資料庫的資料量，會隨著病患病歷求診而每日劇增，但在醫學資料庫中的資料通常是跨科別的臨床資料，所以在這當中隱藏許多可供醫學研究疾病關聯性的寶貴資料。在這樣龐大的醫學資料量中，可以透過資料探勘的技術，挖掘出有益於病症的分析或者是醫療的診斷，以幫助患者能了解自己的疾病及求診時協助醫生作診斷，間接促使醫療品質提升。因此，本研究預期朝向幾個目標努力：

探索可能滿足本研究需求的演算法。

以目前搜集到的醫學資料庫，推論出疾病間可能的關聯性。

以現有的資訊技術，建置起『自動預診系統』。

第四節 研究流程

本研究希望對醫院門診記錄加以分析，推論出病患主觀描述的症狀與疾病間的關聯性。因此，首先要探討相關研究，以了解研究

中資料探勘的技術的施力點為何？接著分析醫院門診資料庫的資料，以了解看診記錄的特性及結構。利用資料探勘工具的運用，推導出病患主觀描述病情與疾病間的關聯性，再經由專家進行驗證，彙整成有用的知識庫。本論文研究流程圖如下（圖 1-1）：

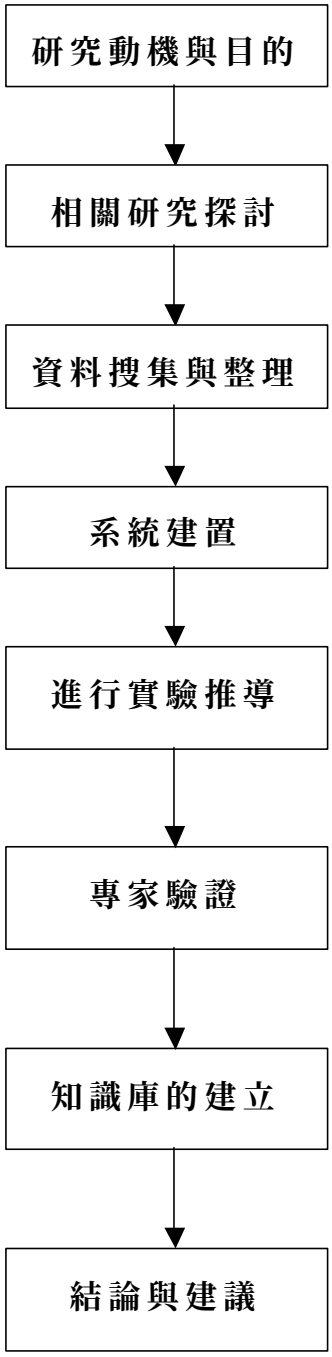

圖 1-1 研究步驟流程圖

本研究流程如下：

1. 研究動機與目的：說明本研究的源起與預期的目標。

2. 相關研究探討：拜讀過去學者對醫學資料庫、資料探勘技術方面的研究，並找出本研究中需要的演算法。

3. 資料的搜集與整理：對於研究中用到的醫學資料庫，進行搜集、前處理及彙整的動作。

4. 系統建置：對於先前相關研究中選定的演算法，利用資訊科技，建立起一套人機介面的『自動預診系統』。。

5. 進行實驗推導：將第 4 步驟中準備好的資料，轉匯進系統中以實際運作。

6. 專家驗證：將前一步驟所得的結果，請專家測試系統，是否與本身專業素養所判斷的結果相合。

7. 知識庫的建立：將先前系統判斷所得的型樣，整理成關聯性的法則。

8. **結論與建議：對本研究的成果做總結，並提供後續研究方向。**

第五節 研究限制

本研究將以資料探勘技術找出病患主觀描述病情與疾病間的關聯性，但在研究中有幾點是無法難以成就的：

1. **前置詞不考慮：如：**no cough**，病人提供這樣的詞句難以界定統一說法，容易使實驗結果產生碎裂性。**

2. **數字詞不考慮：如** a few fever**，**fever 38^{0}C **……，我們都歸類** fever**。**

3. **疾病類別採大分類：如** 150.1,150.2,150.3 **…..都將歸類到** 150**。**

4. **醫學知識的先天性不足：由於資訊與醫療非屬同一學門，故研究中關鍵部份（病患主觀描述病情與疾病間的關聯性）難以達到專家的水準。**

第六節 論文架構

本篇論文共分五章，其結構如下：

第一章 序論：說明本研究的研究背景、研究目的、研究流程、研究限制。

第二章 相關研究探討：主要是介紹 KDD 與資料探勘的關係及資料探勘的相關領域。

第三章 研究方法：依據研究目的及相關研究中的理論基礎，描述本研究中可能的演算法（決策樹與關連式法則）及系統的建置步驟。

第四章 實驗成果：說明本研究中的研究成果。

第五章 結論與後續研究：本研究中結果的討論及後續研究方向的說明。

第二章 文獻探討

第一節 資料庫知識發掘

源起

在商場上，由於企業資料量的爆增，面對這樣龐大的資料庫，若僅將其當成歷史性的資料備份，則是浪費許多儲存成本，並無實質的貢獻；而且資料庫中包含歷年的客戶交易記錄，若能從中找出有用的資訊、知識，將可提昇競爭優勢。以決策支援系統（Decision Supply System）的設計觀點，如何善加利用這些資料庫裡的資料，製作出精準的模式庫以提高決策的品質，即成為管理者在決策時重要職責。而這套自資料庫裡推得出「法則」或者「知識」來製作成模式庫以支援決策的途徑即稱「知識庫知識發掘」[Pieter Adriaans 96]。

KDD 與 Data Mining 的關聯

有些學者會將 KDD 與 Data Mining 混淆， KDD 與 Data Mining 二者的關係是需要釐清的。Data Mining 可視為 KDD 的一部份，KDD 的整個過程包括資料選取（Selection）；再從資料中作前置處理

（Pre-processing），除去不一致或錯誤的資料；然後作資料簡化與轉換工作（Transformation），再經由資料探勘的技術程序使其成為型樣；做回歸分析（Regression Analysis）或找出分類型態（Classification Type）；最後經由解釋或評估成為有用的知識（如圖 2-1）。這些程序是一個循環的關係，一直重複的步驟，最後才得到一些有用的知識。所以 KDD 是一種知識發現的一連串程序，而資料探勘是其中的一個重要程序。

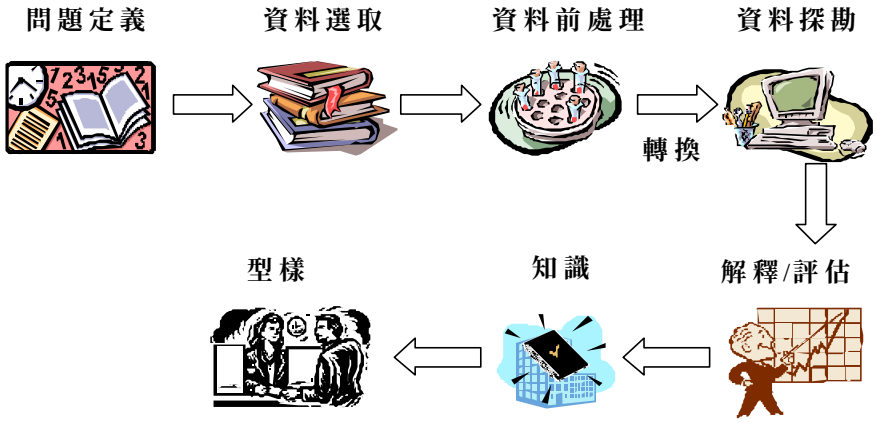

圖 2-1 資料庫知識發掘之程序

第二節 資料探勘

近十年來資料庫技術逐漸成熟，已成功將資料庫技術應用在傳統的企業日常作業中，但資料的累積使得歷史性的資料量日漸地增

加，於是紛紛建立超大容量的資料庫。而這些不斷累積的資料其實都潛藏許多的資訊及商機，例如消費者的購買行為，這些有用的訊息無法直接取得，但若直接使用這些大量的資料，則不符合經濟效益，其參考價值也因而降低。

因此，需要對這些龐大的資料庫作分析處理，以找出其中所隱含的有用資訊，若能從中發現相關的資訊，對企業或者研究單位將產生極大的幫助。一般資料庫系統中並無法提供如此的應用功能，針對資料量多且複雜龐大的資料，必須有一快速且有效的技術來加以分析處理，因此資料探勘（Data Mining）的技術因此因應而生。

資料探勘基本定義

資料探勘，是從大量資料中去發掘潛在有用資料，藉由各種不同的資料分析方式，來得到事前從未知曉的知識或資訊；利用各種資料分析工具，能夠自動的在資料庫中比對並分析資料，最後將獲得的結果呈現出來，以提供決策人員作決策時參考。

是以發現為基底（Discovery -Based）的型樣配對 （Pattern Matching）方式為其主要的方法，這些方法可以處理各種的多維度

資料型態，並找出資料集之中屬於顯性（Dominate）的與例外（Exceptional）的情況。

學者們對資料探勘提出相關的定義，其定義如下：

1. Data Mining 是資料庫知識發掘的核心：

資料庫知識發掘就如同資料探戡的全部處理, 其動作範圍包含從資料的儲存到面對使用者的知識呈現;而資料探勘僅為其核心，是一種從大量已存在資料中，發掘有意義的資訊和知識的處理。（Atre，1996；Rigdon，1997；Greenfeld，1996；Dickey，1996；Imielinsky，1996）

2.Data Mining 是一種自動或半自動的處理：

資料的分析可透過人工智慧或其他統計和數學技術的資訊架構來輔助，而且不需要先建立假設。（Greenfeld，1996；Mena，1996）

3.Data Mining 的結果通常是從前未知且未預測的：

Data Mining 經常找出過去未知和不可預測的結果，但卻不

是最後容易理解的資訊。（Mason，1995；Newing，1996；Krivda，1995）

4.Data Mining 是複雜且需要極大的資料倉儲：

對極大的資料庫而言，複雜式的 Data Mining 技術可以描寫難以偵測的資訊，資料庫在此通常會被轉成資料倉儲（Axelrod，1996）。而資料倉儲與資料探勘有密不可分的關係。

5.操作的人是一般使用者：

Data Mining 允許一般使用者使用簡單的程式及統計觀念從大

量的資料中萃取資訊（Mace，1994；Axelrod，1996）

資料探勘的模型與技術

根據資料探勘的分析方式與產生知識型態，一般而言，資料探勘可有下列五種模型。[14]

1. 分類分析（Classification）：

在專家系統、機器學習、統計學領域中，資料分類是一個行之有年的技術，它屬於一種監督式的學習過程。分類分析是從已知的物件集合中，依據屬性來建立類別（Class），再根據各類別資料的特徵，對於其他未分類或者新的資料來做預測。其目的在於，利用訓練資料（Training Data）中的各種特徵屬性，來建構一個分類器（Classifier）；之後再使用訓練資料特徵屬性相同，但資料內容不相同的測試資料（Testing Data），驗證此分類器是否可以達到使用者可接受的正確率；若正確率高的話，則可用此分類器來預測新資料的分類。當前可用的技術有 ID3、C4.5、迴歸分析等。目前已有的例子，如應用於信用卡申請人風險屬性的分類。

2. 推估（Estimation）：

依據現有的連續性數值的相關屬性資料，以得到某一個未知的值。與分類不同點在於，分類是屬於以離散方式進行資料預測。推估可用的技術有迴歸分析及類神經網路（Neural Network）等。目前已有的例子，如應用於：根據信用卡申請

人之職業、教育程度、年齡等因子，來推估其消費型態。

3. **預測（Prediction）：**

依據某一特定對象屬性的過去觀察值或歷史資料，來推估其未來的值會是多少。例如：由顧客過去刷卡記錄，預測其未來的消費量。目前可用的技術：時間數列分析、迴歸分析及類神經網路等方法。目前已有的例子，如應用於股票未來趨勢走向預測。

4. **關聯分析（Association Rule Analysis）：[9]**

首先，蒐集一組交易記錄，每一交易記錄包含若干交易項目；再則，利用連結分析的方法在交易記錄中找出交易項目的關聯法則（Association Rule）。一般而言，關聯法則會提供下列的描述：「若 A、B 二種交易項目發生，則會發生交易項目 C 的機率為 p，機率值愈高則表示關聯性愈高」。例如，有一客人買三明治和奶茶，則客人同時會買茶葉蛋的機率為 85%。藉由這些購物行為的分析，業者可以將那些具有關聯性的商品放在一起，或是調整物品擺設的位置及改變產品型錄的設計，

或者調整存貨與訂單數量 .....等等，將對於業者的營運會有極大的幫助。目前的應用，如客戶銷售系統上的交叉銷售（Cross Selling）。

5. **群集分析（Clustering Analysis）：**

群集分析指的是將物件或資料分成若干群集的過程，也就是根據物件間的相似性（或不相似性），將所有物件分成若干個群集，使得每個群集內的物件具有高度的相似性，而不同群集間具有高度的不相似性。其目的是要將群集間的差異找出來，同時也要將群集中成員的相似性找出來。如圖 2-2，利用顧客資料的相似性將顧客分為四個群集。

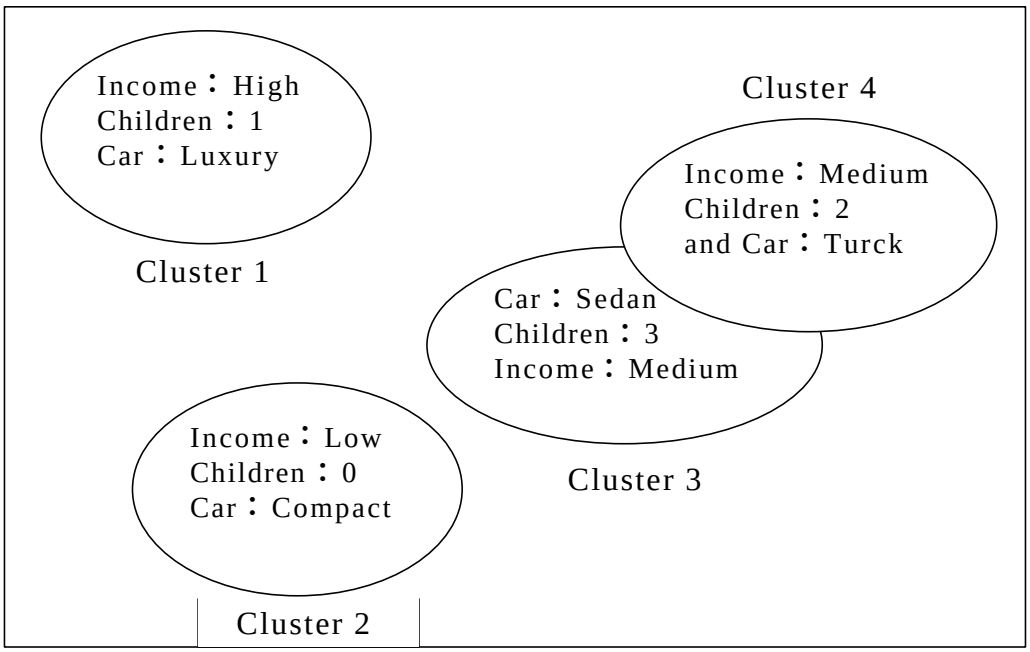

圖 2-2 將顧客分為四個群集[23]

此外，在資料探勘技術的“群集化”，是針對目標資料集合，利用合適的群集化演算法，有效的將資料集分成一個個的群集，使得各個群集的特徵可以被有效的突顯出來，這些特徵即是隱藏在資料中的資訊，利用這些資訊可以協助使用者下達決策。目前所使用的技術有 K-means **法。**

資料探勘之過程

資料探勘的重心莫過於發掘的過程。在下圖（圖 2-3**），將說明資料探勘的程序。**[14]

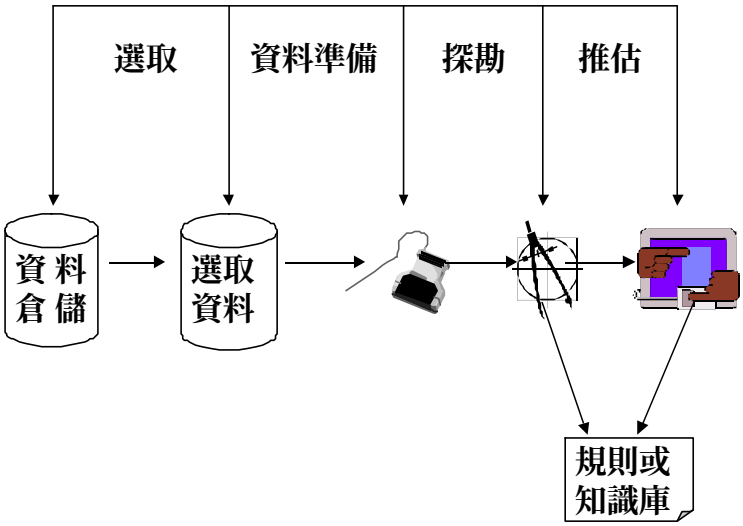

圖 2-3 資料探勘的過程

資料探勘的過程分為以下四個階段[曾詠淑]：

第一階段：選取（設定目標）

分析現有的模型，以確認資料探勘應用的領域，進一步設定此模型的目標及評估準則，並逐一考量影響此模型潛藏的因素。這在企劃過程中有絕對的意義，亦關係著資料探勘是否成功的關鍵。

第二階段：資料的準備

（1）選擇合適資料：在此階段進行資料的蒐集，將大量資料過

濾、去除不一致或錯誤的資料，建立起適用的資料庫。

（2）資料準備過程：根據前一步驟所建立的資料庫，對其分析資料的特質（屬性），以篩選出資料探勘模型中必要的欄位記錄，將其另存於新資料表並選擇必要的開發技術及工具。

（3）資料轉換：依據先前建立的新資料表，做資料轉換的工作。（將資料轉換成資料探勘工具中所需要的格式）。

第三階段：資料探勘工作

此階段為資料探勘真正核心。藉由資料探勘演算法找出隱藏在資料背後的規則、特性、型樣。

第四階段：結果分析

解釋並評估前階段所產生的結果。一般而言，會將結果以圖表的方式呈現出來，讓使用者對於分析結果能深刻了解。

應用

資料探勘可以在大量資料中找出隱藏的有用資訊，以提供企業

高層決策下達輔助之用，因此若可以有效利用其特點於各領域中，其幫助是相當大。以下將列舉目前已應用的領域：[11]

1. 商業行為方面：

A. 客源開發與既有客戶間的良性互動。

B. 品管的提升。

C. 生產成本的降低。

D. 降低產品與原料的庫存。

E. 預測股票的跌漲。

F. 預警信用卡呆帳的可能性。

2. 醫療方面：

A. 療程的改善。

B. 藥品庫存的降低。

C. 醫學影像的分析與處理。

3. WWW 上的應用

A. 使用者在 Web 上的行走型樣。

B. 網頁個人化的建制。

C. 高效果的搜尋引擎。

第三節 關聯法則

資料探勘的技術可依資料庫型態、知識類別、應用層次來加以分類，而其中最被廣泛討論的為關聯法則。關聯法則指出在龐大的資料庫中某一些物件間存有彼此的關聯性 。其起源在於分析市場購物籃資料（Market Basket Data）中之大量商品集合（Itemsets）的關聯程度，因此也稱為市場購物籃分析（Market Basket Analysis)。關連法則經常應用於商業上的購物車分析（Market Basket Analysis)，藉由銷售點系統（Point of Sale；POS）所記錄的消費者購買歷史，可分析出購買商品之間的關聯性，根據這些關聯性，商家可做為進貨或儲貨的依據，或可作為內部商品擺設的參考，藉此組合相關的產品，實行交叉銷售，以提高行銷績效。

關聯法則的目的主要是要找出資料項間的關聯性，其型式為 X→Y，其中 X 與 Y 分別表示在資料庫中不同的資料項目組，其意識即若購買項目集合 X 時，可能會在購買項目集合 Y。若要找出規則 X→Y，則必須先計算項目集合 X 及項目集合 X∪Y 的支持度（Support），這就表示資料庫中有幾筆資料包含此項目集合。接著必須確定 X∪Y 是否為最大集合項；如果是，即可將 X∪Y 的支持度除以 X 的支持度，也就是 Support（X∪Y）/ Support（X），所得值代表 X→Y 的信賴度（Confidence）。若信賴度超過最小信賴度（Minimum Confidence），則關聯法則 X→Y 成立。支持度及信賴度都是由使用者所自訂，一般有效的關聯法則其支持度及可靠度均須在一定程度以上，也就是說關聯法則的支持度及可靠度必須大於或等於使用者所訂的最小限制，則此關聯才被認為是有意義的。

所以要形成 X→Y 之法則需符合二條件：

1. X∪Y 在資料庫中，出現的筆數要比最小限制值還大。

2. 此外其信賴度要符合要求，即 X∪Y 的支持度除以 X 的支持度的值超過最小信賴度，方可成立。

在關聯法則中最常見的演算法為 Apriori，以下將介紹其演算法。

Apriori 演算法介紹

Apriori 演算法在 1994 年由 Agrawal *et al.*所提出，目前已是探討關聯法則時最具代表性的演算法；其後雖有針對各種不同目的或情況所提出的各類演算法，但其大多根據於 Apriori 演算法並加以延伸改進。

於此探討關聯法則的推導，其主要包含兩個步驟：

1. 搜尋資料庫以找出所有的高頻項目組。

2. 利用高頻項目組推導出所有的關聯法則 。

在步驟 1 中，相關項目組在資料庫中出現的頻率必須要大於或等於某一使用者所定之最小支持度限制，滿足此條件限制的便稱之為高頻項目組。若其出現的頻率太低，則其相對地沒有意義，這樣的項目組對使用者來說是可以不需要的，因此可將之加以忽略。使用者比較在意的只是出現次數較多的項目組，在統計學上而言，這也是比較能作為決策參考的。在尋找高頻項目組的過程中，我們必

須重覆地搜尋資料庫，並根據高頻項目組的所有子集合也必為高頻項目組的特性，經由聯結（join）以及刪除（prune）這兩項子步驟的進行以產生新的候選項目組，而後再計算其支持度以濾出高頻項目組。

在步驟 2 中，我們根據由步驟 1 中所得的高頻項目組去找出真正的關聯法則；我們根據可靠度來判斷此關聯法則是否有意義。相同地，可靠度的值也必須要大於或等於使用者所定之最小限制，若其可靠度太低，則雖然其出現的頻率高於使用者所定之最小支持度限制，我們仍認為其是較沒有意義的關聯法則因而不予考量。經由上述之兩個步驟的推導，我們即可探勘出龐大資料庫中有意義的關聯法則。

在 Apriori 演算法中（表 2-1）[30]，第一個掃描（pass）決定了 large 1-itemsets。最後的掃描 k 由兩階段組成，一、是由找出的 Large itemsetsL_{k-1} 使用 Apriori-gen 函數如表 2-2 產生 candidate itemsets C_k，二、是每次掃描整個資料庫去計算 candidate itemsets 的 support。

以下將舉個例子來說明整個關聯法則的產生過程。在 Database D 中，如表 2-3，可以看到{A}在 TID 100，300 中出現，所以他的

support 是 2，{D}只在 TID 100 出現，所以 Support 是 1。假設將最小 Support 的值設為 2，可得到 L_1 如表 2-4，為了發現 large 2-itemsets 的集合，Apriori 使用 $L_1 * L_1$（＊為 join 的運算元）來產生 C_2 itemsets 的 candidate set，結果如表 2-5。相同的方式 C_3 是由 $L_2 * L_2$ 產生的，如表 2-6。

在表 2-5 中{BC}＊{BE}產生{BCE}，{BCE}變成一個 candidate 3-itemsets 放如 C_3 中，再計算{BCE}的 support，之後再也沒有其他從 L_2 出來的 candidate 3-itemsets。我們可以看到{BCE}再 TID 200，300 出現，因此它的 support 是 2。Apriori 接著從 large 3-itemsets L_3 要產生 C_4，因為沒有 candidate 4-itemsets 可以從 L_3 產生，Apriori 結束所有產生 large itemset 的處理而產生關聯規則出來。

表 2-1 Apriori 演算法[1]

```
1.  L1= {large 1-itemsets }
2.  for ( k=2;Lk-1≠0; k++)do begin
3.        Ck= apriori=gen ( Lk-1);
4.        forall transaction ∈ D do begin
5.           Ct = subset (Ck , t );
6.           Forall candidates c ∈ Ct do
7.                    count++;
8.           end
9.           Lk = {c ∈ Ct | c.count≧minsup}
10.          end
11.          Answer= UkLk ;
```

表 2-2 Apriori-gen function[2]

```
Insert into Ck
Select p, item1, p.item2, ..., p.itemk-1, q.itemk-1
From Lk-1 p, Lk-1 q
Where p.item1＝q. item1, .., p.itemk-2＝q. itemk-2, ..,
p. itemk-1＜q. .itemk-1 ;
```

[1] （資料來源：Agrawal & Srikant, 1994）

[2] （資料來源：Agrawal & Srikant, 1994）

表 2-3 交易資料庫範例

Database D

TID	Items
100	ACD
200	BCE
300	ABCE
400	BE

表 2-4 pass 1 candidate 和 large item sets 的產生

C1

Item sets	Sup.
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

Scan D →

L1

Item sets	Sup.
{A}	2
{B}	3
{C}	3
{E}	3

表 2-5 pass2 candidate item sets 和 large item set 的產生

C2

Item sets
{AB}
{AC}
{AE}
{BC}
{BE}
{CE}

Scan D →

C2

Item sets	Sup.
{AB}	1
{AC}	2
{AE}	1
{BC}	2
{BE}	3
{CE}	2

L2

Item sets	Sup.
{AC}	2
{BC}	2
{BE}	3
{CE}	2

表 2-6 pass 3 candidate item sets 和 large item sets 的產生

C3

Item sets
{BCE}

Scan D →

C3

Item sets	Sup.
{BCE}	2

L3

Item sets	Sup.
{BCE}	2

第四節決策樹

決策樹是一種語意樹（Semantic Tree），與一般資料結構中的樹一樣（有節點、樹葉等結構），每一節點都被安排一個適當的測試，然後利用該測試結果決定資料將再利用此一節點的哪一棵子樹作為分類條件繼續做決策。

決策樹是從訓練資料中，由上到下產生一個特定的方向，利用某項特性作為節點來分割方向。若所有樣本屬於相同的類別且能獲得辨識，則完成決策樹的分類。新事務由決策樹樹根節點開始測試進行，新事務選擇符合其屬性值的分支，往下移至另一節點，依此遞迴方式繼續進行，直至遇到樹葉為止，則此樹葉就是一個類別。圖 2-4 為決策樹的一個簡單例子。

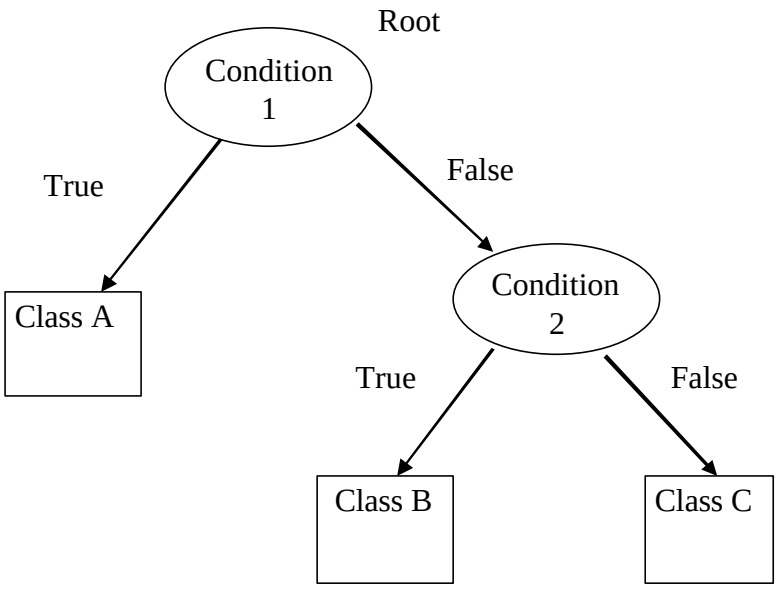

圖 2-4 決策樹

決策樹的歸納演算法如下：

1. **將訓練樣本的所有範例放入決策樹的樹根。**

2. **若此節點不含任何範例或節點中的範例都屬於同一類別時，則此節點成為空樹葉或所有範例共同類別的樹葉；若此節點含有不只一種類別的範例時，則需依據某一評估函數（例如：經驗法則），對範例的所有屬性進行評估，並選出一個適當的屬性。依據此屬性的值將節點中的範例分成 N 部份，每一部分就是連接根節點的一個新節點。此過程**

稱為節點分割（splitting node）。

3. 經過節點分割後，判斷這些新節點是否為樹葉；若否，則以新節點為子樹的樹根來建立新的子樹。

4. 以上步驟經遞迴的方式持續進行，直到所有的新節點都是樹葉為止。經由此種歸納方法獲得的決策樹能夠將訓練資料的範例完全分類。

建構決策樹

本研究所採用決策樹演算法是 C4.5 學習法，它的基本理論是從 ID3（Iterative Dichomizer 3）學習系統改良而來的，而 ID3 是 Quinlan 於 1979 年將 CLS（Concept Learning System）改良而成的演算法則，之後 Quinlan 又在 1993 年提出 C4.5 將 ID3 改良而成的演算法則。

C4.5 學習方法的第一個步驟即是由訓練資料建構決策樹，它的基本構想可追溯至 1950 年代末期的 Hoveland 及 Hunt 二位學者提出的理論。簡單說明如下：假設一組訓練資料所組成的集合 Y 中有 T 種類別，及 $Y=\{C_1, C_2, C_3, \ldots\ldots C_T\}$，在建構決策樹時有三種情形可能發生：[7]

1. 當 Y 中的所有訓練資料都屬於同一種類別 C_j 時，所建構的決策樹就只會有單一樹葉節點，此樹葉節點 C_j 的所有資料。

2. 當 Y 中沒有任何訓練資料時，所建構的決策樹依然只有單一樹葉，則此樹葉代表的類別則由集合 T 以外的訓練資料決定。

3. 當 Y 中包含多種類別的訓練資料時，則將集合 T 根據某種屬性分割成多種子集合 Y_1，Y_2 .....Y_n，每個子集合儘可能包含一種類別的資料。Y 所建構的決策樹會包含一個判斷節點及 n 個分枝，而每一個子集合的訓練資料分別對應 T 的一個分枝。

一般在給予訓練資料時，會有多種決策樹可以正確的將資料分類，如何找出一最小而能正確的分類的決策樹，則取決於分類屬性的考量。由於 C4.5 是由其前身 ID3 學習法改進而來，而在 ID3 中分類屬性的選擇標準稱為 gain，它是以測量資訊量多寡來計算各個類別的資訊量，進而計算出該訓練集合的平均資訊量，也就是所謂的熵值（Entropy），來表達該集合中資料的複雜度。

假設訓練資料形成的集合 Y 中有 n 種類別 Xi，i＝{1，2，3 ..，n}，每個類別的資料個數以（Xi，Y）表示，｜Y｜代表 Y 中所有資料的個數，因此各個類別其資料出現機率可表示為

$$P = \frac{(Xi, Y)}{|Y|}$$

因此根據訊息理論，各個類別的資訊為

$$-\log_2(P)$$

訓練集合中包含各個類別的訓練資料，由各類別的資訊量可以計算出訓練集合的平均資訊量（即 Entropy），為所有各個類別的資訊量乘上各個類別的資訊量乘上各個類別資料的出線機率總和：

$$\inf o(Y) = -(P1 * \log_2(P1) + P2 * \log_2(P2) + ... + Pn * \log_2(Pn))$$

根據 info(Y)的計算方式，當集合 Y 根據某個屬性 A 分割成多個子集合 Y1，Y2，..，Ym 時，其分割後所佔的資訊量等於各個子集合的資訊量乘上各個子集合所佔的比例的總合：

$$\inf o_A(Y) = \sum_{i=1}^{n} \frac{|Yi|}{|Y|} * \inf o(Yi)$$

因此集合 Y 經由屬性 A 分割後所獲得的資訊量則為分割前的資訊量減去分割後的資訊量，表示為：

$$gain(A) = inf\ o(Y) - inf\ o_{A}(Y)$$

決策樹以此屬性的屬性質分割成多個訓練子集合，形成多個子樹。各個子樹重覆上述步驟尚未被選為分類的屬性中在找出 gain 值最大的作為分類屬性，直到分割到不能再分為止。

修剪決策樹

ID3 選擇分類屬性以往都有不錯的表現, 但是除了一個狀況：ID3 在分類時會偏向分出較多子集合的屬性值，而這樣的情況有時會造成 Y 集合分割後的子集合會只有一個資料，其分割後的資訊量為零，所獲得的資訊量最大，但此種分割並不是最佳的決策樹，此種分割也沒有有太大的意義。

因此，Quinlan 為了要解決上述的缺點，於是 Quinlan 在 1993 年提出 C4.5，其中對資訊量再以測試屬性的資訊量作正規化，以減少過多子集合裡只有一個資料或少數個資料的問題產生，並將決策樹修剪為最佳化。如圖 2-5。

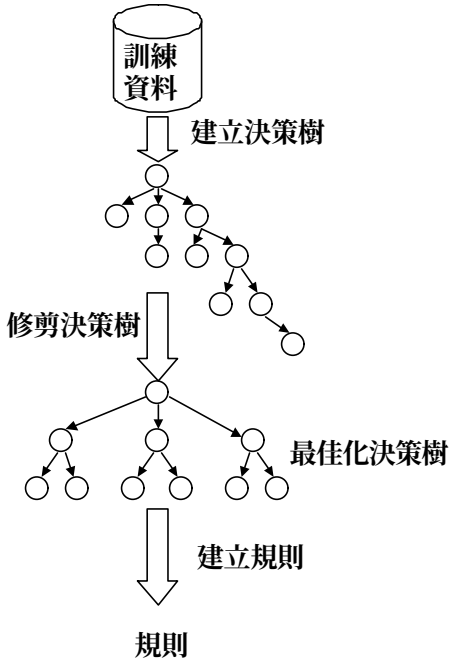

圖 2-5 決策樹建立流程

正規化的方法是將原來的 gain **值除以** split info（A）**的值，即**

gain ratio（A）＝gain（A）/split info（A）

而

$$Split\ \inf o(A) = \sum_{i=1}^{n} \frac{|Yi|}{Y} * \log_2(\frac{|Yi|}{Y})$$

屬性 A **分割後的子集合個數愈多** split info **的值就會愈大，而** gain ratio **的值就會相對地變小。**C4.5 **可利用** split info **改善** ID3 **分類偏向較多子集合的缺點，成為最佳化的決策樹。**

第三章 研究方法與步驟

第一節 研究方法

本研究利用醫療院所的門診資料作為研究的資料來源。此門診資料為病患門診後的就診紀錄，我們將就診紀錄以資料庫型態儲存；再將資料經過篩選後，利用病患對症狀主觀描述出現的次數轉化為機率值，也就是將資料型態由文字描述轉為數值型態；接著用 Apriori 演算法利用機率值的高低找出高頻項目組；再利用決策樹分類方法產生結果，也就是疾病名稱（以 ICD-9 碼表示）。

門診紀錄的背景簡介

每個人都有到看病的經驗，當身體感到不適時會到醫院掛號看病，在門診室內醫生透由患者自己訴說病情症狀的，再佐以目視患者的情況或者檢驗報告，最後依照經驗來判斷患者是罹患何種疾病，這一串看診的程序都會記錄在患者的病歷表上。

近年來，有許多醫療院所開始推動電子化病歷，因此也就有醫療院所將門診病患看診的過程記錄在電腦裡，而這些門診紀錄的資料內容包含了「患者看診日」、「患者看診科別」、「患者症狀描述」、

「疾病名稱及對應的 ICD-9 碼」（The International Classification of Disease，Ninth Revision）[2][5]、「患者檢驗檢查項及報告」、「處方簽」等欄位，在本研究中必須參考使用的欄位有「患者看診科別」、「患者症狀描述」、「疾病名稱及對應的 ICD-9 碼」等項，其各項所代表意義如下：

1. **患者看診科別：**當患者門診掛號時，患者必須要先確定要掛哪一科別的門診，當完成掛號手續後在電腦上顯示出患者此次掛號的科別，但是科別為了資料輸入的方便，是以代號來表示，本研究中資料門診部分科別代號如表 3-1。

表 3-1 部分門診科別表

科別代號	科 別 名 稱
01	內科
02	外科
03	皮膚科
04	眼科
05	耳鼻喉科
06	牙科

續表 3-1 部分門診科別表

07	婦產科
08	胸腔內科
09	胃腸科
10	小兒科
15	泌尿科
16	骨科
40	復健科

2. **患者症狀描述：**病患在看診時，將自己的症狀特徵（例如有一些鼻塞、喉嚨痛）告訴醫生，醫生再將患者口述的症狀紀錄下來，其記錄可當病患的往後的病史內容，如表 3-2 門診資料表範例中的「主觀描述」欄位的資料為即為患者對症狀的主觀描述。

表 3-2 門診資料表範例

病歷號碼	科別	主觀描述	疾病名稱 1	疾病代碼 1	疾病名稱 2	疾病代碼 2
0318765	01	Cough with Whitish sputum	急性喉炎及氣管炎	464		

1439654	04	Blurred vision OU	白內障	366	視覺不良	368.13
0289755	05	Nasal OBS for Many Days	急性支氣管炎	466.0	過敏性鼻炎	477
0089741	10	Cough More Running Nose For Days	支氣管肺炎	485	未明示之氣喘	493.9
1008360	10	Cough with Sputum，Nasal Obtruction，Rinorrhear and Fever for days	支氣管肺炎	485	急性咽炎	462

3. **疾病名稱及對應的 ICD-9 碼：門診中，醫生根據患者症狀的主觀的描述，再依照醫生的經驗或者檢驗報告判斷病患罹患的疾病，並將疾病名稱及對應的** ICD-9 **碼（健保局規定醫療院所申報健保給付時，疾病名稱要以** ICD-9 **碼來申報，例如：黴菌病的** ICD-9 **碼為** 117.9**）記錄下來，其記錄可作為日後病患的病史。表** 3-2 **門診資料表範例中的「疾病名稱** 1**」「疾病名稱** 2**」為患者的疾病名稱;「疾病代碼** 1**」「疾病代碼** 2**」為患者的疾病名稱** ICD-9 **代碼。**

第二節 研究流程步驟

以下將介紹本研究中的研究流程圖：

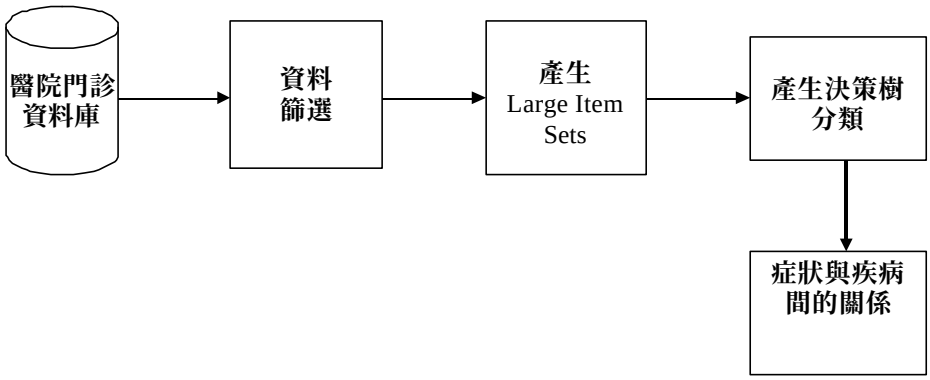

圖 3-1 研究流程圖

首先，取得醫院門診的資料庫，並對其施以前處理再進入資料篩選的階段。挑選出研究中的必要性欄位將其匯整，以便於進行下階段的機率值轉化。利用統計軟體將相同症狀出現的次數，轉化成[0,1]的機率值。在這群機率值中，挑選出高頻的部份，成為 Large Itemset。找出 Large Itemset 後，給定一組信賴度及支持度的值，以找出有意義的關聯法則。在找出關聯法則後，就可以得知症狀間的關聯性為何。將以上的結果入下一步驟，利用 C4.5 作決策樹的分類。

最後，C4.5 的系統會推論出症狀與疾病間的絕對關聯性。再對研究流程進行細部解說。

1. 醫院門診資料庫的選取

研究中的病患原始資料原本以 Oracle 資料庫的型式存放，為了方便本研究的使用，所以將資料轉匯至 Acess 資料庫。其步驟為：將 Oracle 資料庫的檔案匯出成為文字檔，再把文字用匯入的方式存入 Acess 與 Excel 中。

病患資料庫的欄位有「編號」、「看診日期」、「主觀描述」、「客觀描述」、「疾病名稱 1」~「疾病名稱 3」及「疾病代碼 1」~「疾病代碼 3」等 10 個欄位（如表 3-3 所示）。

表 3-3 病患資料庫原始欄位說明

欄位名稱	編號	看診日期	主觀描述	客觀描述
意義說明	資料筆數	看診日期	病患本身主觀的描述	醫生對病患症狀的描述

欄位名稱	疾病名稱 1~疾病名稱 3	疾病代號 1~疾病代號 3
意義說明	所患疾病的中文名稱	所患病症的代碼（ICD-9）

2. 資料的篩選

在資料庫中，可以發現到病例資料的「主觀描述」的欄位常有 Null 值的出現，此乃因於國內醫療制度中，對於病歷記錄僅認可手寫的，對於電子化的病歷在法律上未有其合法性，連醫療院所向健保局申請健保給付，也必須要附上病患的手寫病歷影本，健保局才認同病患診療過程的合法性。因此，若醫療院所未強制醫生必須將看診記錄鍵入電腦，就會使門診資料庫中的病患看診記錄產生缺漏的現象，也就是出現 Null 值的由來。

表 3-4 無效門診資料表範例

病歷號碼	科別	主觀描述	疾病名稱 1	疾病代碼 1	疾病名稱 2	疾病代碼 2
2588765	01	Cough， Sputum，Chest tightness	支氣管炎及氣喘	493.90		
1496954	01		高血壓性心臟疾病	402.90		
0025877	01	Dizzness	眩暈	780.4	睡眠障礙	307.40
0147980	01	Cough More Running Nose For Days	支氣管肺炎	485	未明示之氣喘	493.9
1200087	01	Hiccup，Sneezing	消化不良	536.8	唇炎	528.5
1118936	01		頭痛	784.0	心絞痛	413
0234024	01	Palitation with Chest Pain	狹心症	413.9		
1865712	09	Rlq Pain，Soft Stool Passage	胃腸炎	558.9	急躁大腸症候群	564.1
1145578	09	General Malaise，Poor Appetite	消化不良	536.8		

由於『主觀描述』這個欄位是病患對於自己病情主觀的描述，對於研判需分類到何種科別而言是主要的因素，，所以『主觀描述』該欄位是空白的話，這一筆病患資料對本研究而言其參考價值就降低，因此本研究將此筆資料刪除。舉例而言，在表 3-4 中，第二筆和第六筆病患門診紀錄資料裏，『主觀描述』的欄位是空白的，因此將這兩筆資料給予刪除。

3. 產生 Large Item Sets

資料篩選後，從資料庫中將記錄中的『主觀描述』排序並付予每一個文字一個代號，計算每一個文字出現的頻率；將頻率高於平均值的部份留下作為研究用，並將這部份出現頻率較高的文字，以代號方式帶入病患門診資料中的『主觀描述』欄位，並利用 Apriori 演算法獲得 Large Item Sets。整個過程如圖 3-2：

A. 計算『主觀描述』欄位裡每一文字出現頻率

由於『主觀描述』欄位裡都是描述性質的資料，例如“Weakness no fever，no hematuria”，將欄位裡的文字分割使得每個文字獨立出現，並計算每個文字出現的頻率，表 3-5 為其結

果範例：

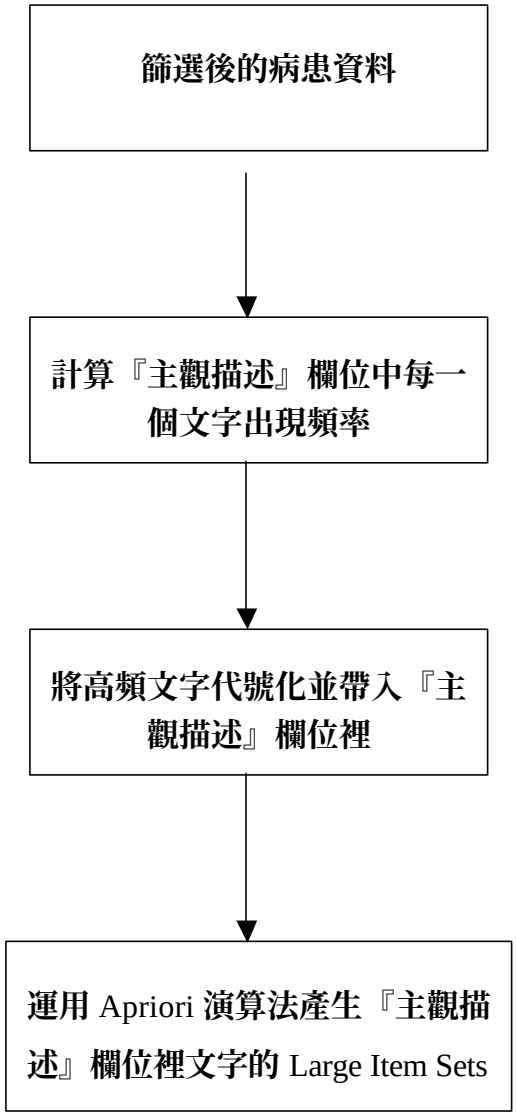

圖 3-2 Large Item Sets 產生過程

表 3-5 計算文字頻率

文　字	出現頻率
AC	21
AFTER	128
ALCOHOL	96
AREA	3
BP	156
BW	96
HEADAHE	2
INNER	2
NEASEA	4
STRESS	7

B. 將高頻文字代號化並帶入『主觀描述』欄位裡

統計出文字的出現頻率後，再計算總出現頻率的平均值。若出現頻率大於平均值者，就將該文字留下；若出現頻率小於平均值者，就則該文字刪除，這是因為其出現率太低，則參考價值較小沒有留下的參考價值。將剩下的文字予以排序並依序給予代號，例如排序第一的文字其代號為 1；排序第二的代號為 2；……依此類推，最後將文字代號帶入『主觀描述』欄位裡，其結果範例如表 3-6：

C. 運用 Apriori 演算法產生『主觀描述』欄位裡文字的 Large Item Sets

當我們得到『主觀描述』的文字描述，將其代號化表示後，再將其帶入修改後的 Apriori 演算法中去做處理，並將其支持度設為 0.1%，其結果即為 Large Item Sets。請見表 3-7 為其結果範例。

表 3-6 病患主觀描述文字代號資料範例

患者病歷	主 觀 描 述
0028965	36 204 52 99，308 176
1013187	36 38，311 7
0298541	52 308 171，177 93
0362111	93，254，291
1354693	254

表 3-7 代號後的 Large Item Sets 範例

Large Item Sets					
304	(0.9%)	**52 99**	(1.4%)	**205 273**	(1.0%)
210	(0.9%)	**52 253**	(2.4%)	**205 1**	(0.9%)
76	(0.9%)	**52 273**	(0.9%)	**308 253**	(2.2%)
23	(0.9%)	**177 93**	(2.3%)	**99 236**	(1.5%)
265	(0.9%)	**36 204**	(5.3%)	**99 56**	(1.0%)
52 177	(1.2%)	**205 44**	(1.1%)	**44 275**	(2.0%)
52 308	(4.0%)	**205 84**	(1.4%)	**44 14**	(1.7%)
275 14	(1.4%)	**52 308 253**	(2.0%)		

4. 決策樹分類產生

本研究利用病患主觀描述症狀的情形，建立決策樹以推論出病患所罹患的疾病。以病歷中敘述症狀的文字為基礎，選取其高頻項目組（以代號表示）與疾病代碼 ICD-9 作為訓練資料，並分割屬性為「有」或「沒有」兩種。再利用 C4.5 演算法來建立決策樹，分割節點為病患主觀描述文字高頻項目組的代號，決策樹最底層的樹葉節點即為疾病代碼 ICD-9，。

由門診資料中，可以發現每一病患的門診紀錄有 6 個欄位，是作為記錄病患的疾病名稱和疾病代碼的用途（如表 3-8），我們取『疾病名稱 1』和『疾病代號 1』這兩個欄位的資料，因為這兩個欄位是記錄病患主要的疾病名稱和代號，但質化的輸入值是電腦所無法處理的，必須要選取，因此最後只選擇『疾病代碼 1』的資料以作為訓練之用。基於研究的限制，本研究對於疾病分類採用大分類：如 150.1，150.2，150.3....都將歸類到 150 這個類別之下。

表 3-8 病患門診資料的疾病名稱和疾病代碼之欄位表

病歷號碼	疾病名稱 1	疾病代碼 1	疾病名稱 2	疾病代碼 2	疾病名稱 3	疾病代碼 3
0318765	急性喉炎及氣管炎	464				
1439654	白內障	366	視覺不良	368.13	消化不良	536.8
0289755	急性支氣管炎	466.0	過敏性鼻炎	477		
0089741	支氣管肺炎	485	未明示之氣喘	493.9	心絞痛	413
1008360	支氣管肺炎	485	急性咽炎	462		

我們將 62 組 Large Item Sets 與『疾病代碼 1』欄位裡的 ICD-9 碼利用程式執行後的結果來探討『疾病代碼 1』欄位裡的疾病對這 62 組 Large Item Sets 是否有產生屬性的關係, 如有對某組 Large Item Sets 有產生屬性的話會在這組 Large Item Sets 下方產生『1』的結果；若對此組沒產生屬性的話會在這組 Large Item Sets 下方產生『0』的結果。其結果範例如表 3-9：

表 3-9 屬性產生表

Large Item Sets / ICD-9	304	210	52 99	52 305 253
460	0	0	0	1
401	0	1	0	1
250	1	0	1	0
413	0	0	0	0
464	1	0	0	0
490	0	0	1	0

最後將病患主觀描述文字中高頻項目組的代號和疾病代碼的資料當輸入值。以 1 或 0 作為分割屬性的區別，1 代表『有』；0 代表『沒有』。，帶入 C4.5 的程式工具後即可得到最佳化決策樹。如圖 3-3 所示。

5. 症狀與疾病間的關係

決策樹中的各個節點的數字都有其意涵。例如 112＝0（112 其文字為 hand）是指「hand」文字若「沒有」出現的話，則進行下一階段 210（其文字為 pc）分割節點再作判斷；直到最後第二行 11＝1：244 是指若「hand」文字「有」出現，則推論出病人罹患疾病代碼為 244（甲狀腺功能不足症）疾病。

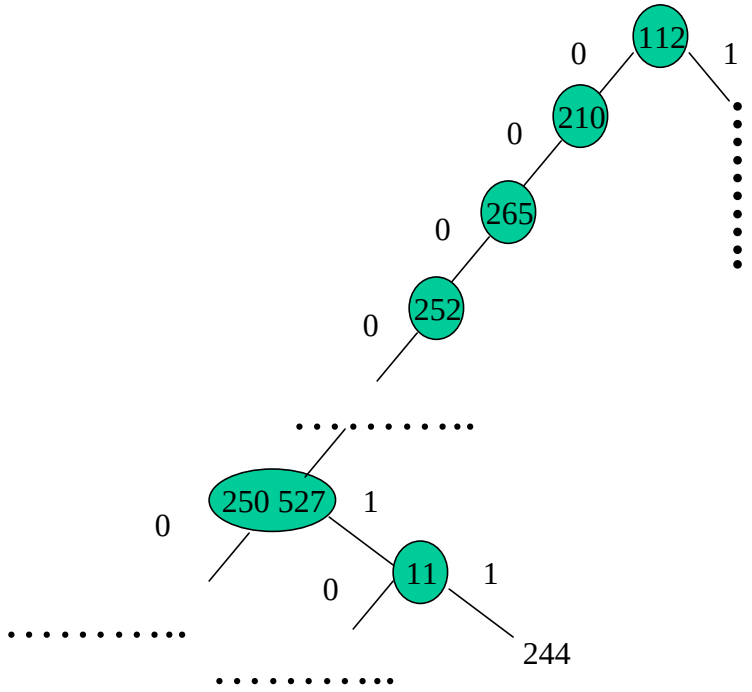

圖 3-3 決策樹範例

第四章 實驗成果

本研究資料來源為南部某區域教學醫院的內科門診記錄，共取得一個月份 15984 筆門診記錄進行資料篩選。首先，利用『主觀描述』裡文字描述來切割出各個獨立的單字，並將單字依序給予代號且利用 Apriori 演算法獲得 Large Item Sets。其次，選取『疾病代碼 1』的資料，並將其淨化為大分類的疾病分類碼。最後，將淨化的疾病代碼與『主觀描述』裡文字（代碼化）的 Large Item Sets 當成輸入值，執行 C4.5 演算法獲得最佳化的決策樹。以下分別描述實驗之結果。

資料篩選

資料篩選過程主要是將實驗所需的資料整理完備，將無效資料或不完整的資料給予去除，根據圖 4-1 的資料篩選流程圖所示，從醫院得到 15984 筆資料，經篩選後得到有效資料為 15286 筆。並從有效資料裡將『疾病代碼 1』中的 ICD-9 碼轉換為大分類的 ICD-9 碼，其結果如表 4-1，4-2。

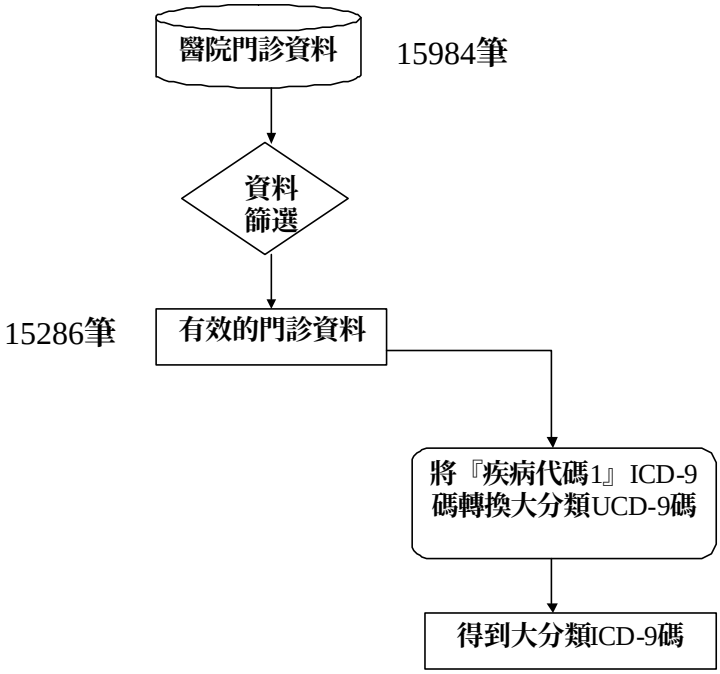

圖 4-1 資料篩選流程圖

表 4-1『疾病代碼 1』原始 ICD-9 碼

疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1
460	234.9	782.3	592.0	592.0	493.0	719.45
401.1	726.0	413	464	427	401.9	250.90
386	490	533.90	493.9	571.4	413.9	582.9
250	413	274	250.60	486.5	490	558.9
250.4	493	413.9	250.90	250.90	493.90	402.8
528.9	412	427	465.0	465.0	402.90	780
789.0	571.40	401.1	582.9	627.2	465.0	571.40

表 4-2『疾病代碼 1』ICD-9 碼大分類後

疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1	疾病代碼 1
460	234	782	592	592	493	719
401	726	413	464	427	401	250
386	490	533	493	571	413	582
250	413	274	250	486	490	558
250	493	413	250	250	493	402
528	412	427	465	465	402	780
789	571	401	582	627	465	571

尋找 Large Item Sets

研究中有效樣本共計有 15286 筆，其資料欄位格式如下所示：

表 4-3 有效資料病患門診紀錄欄位表

欄位名稱	病歷號碼	看診日期	主觀描述	客觀描述
意義說明	病患的病歷號碼	看診日期	病患本身主觀的描述	醫生對病患症狀的描述

欄位名稱	疾病名稱 1~疾病名稱 3	疾病代碼 1~疾病代碼 3
意義說明	病患所患疾病的中文名稱	病患所患疾病名稱的代碼（ICD-9）

由於『主觀描述』這個欄位是病患對於自己症狀特徵的描述，對本研究而言是非常重要的參考項目，因此我們將此欄位的資料擷取下來（如表 4-4)，並對欄位內的文字描述資料給予切割，讓每個文字形成獨立的單字，並計算每個文字出現的次數，研究中發現切割後共有 3563 個獨立文字出現，每個文字平均出現次數為 21.77 次，故將平均值定為 22 次。若文字出現次數小於 22 次則將資料刪除；若文字出現次數大於或等於 22 次則保留，結果共有 312 筆的文字出現次數大於或等於 22 次，之後將這 312 筆資料都各給一個代號，以代號來代表此文字，其內容如表 4-5：

表 4-4 『主觀描述』欄位裏原始資料

病歷號碼	主觀描述
0029875	BP 125/ 80 P 83 COUGH FOR A MONTH WITH NIGHT WHEEZING
1274551	bp 140 / 84 p 87 , ant chest tight without cold sweating , no oppresive sensation
1698475	bp 140 / 84 p 87 , ant chest tight without cold sweating , no oppresive sensation
0007896	BP 151/ 80 BS 462
1007845	BP 99/ 68 P66, OCCASIONAL ELEVATION OF SUGER , PALPITATION, EXERTION CHEST TIGHT SOME PALE
1335487	BS 265
1235526	BS 257
1097854	NUCHAL CREPITUS WHEN TURN ARROUND BP 153/104
1200092	R'T LOWER ABD PAIN , APPENDIX REMOVED 6 YEARS AGO
0204096	BP 135 /82 BS 220

表 4-5 高頻文字

代號	文 字	次數	代號	文 字	次數	代號	文 字	次數
1	ABD	500	26	BITEMPERAL	45	51	CONTUSION	40
2	ABDOMEN	144	27	BITTER	51	52	COUGH	4880
3	AC	57	28	BLACK	100	53	CRUMPING	28
4	ACID	278	29	BLEEDING	50	54	DATA	99
5	AFTER	153	30	BLOOD	102	55	DAY	34
6	AGGRAVATED	22	31	BLOODY	32	56	DAYS	600
7	AGO	114	32	BLURRED	40	57	DEEP	37
8	ALCOHOL	33	33	BLUURED	27	58	DENY	117
9	ALL	25	34	BODY	138	59	DIARR	106
10	AMOUNT	69	35	BOTH	70	60	DIARRHEA	233
11	AND	448	36	BP	3399	61	DIFFUSE	60
12	ANKLE	31	37	BREATH	43	62	DISCHARGE	70
13	ANOREXIA	24	38	BS	1071	63	DISCOMFORT	257
14	ANT	614	39	BUT	146	64	DIZZI	790
15	ANXIETY	196	40	BW	267	65	DIZZINESS	480
16	APPETIDE	220	41	BY	22	66	DM	48
17	APPETITE	149	42	CANCER	43	67	DONE	25
18	AREA	86	43	CASE	21	68	DOWN	56
19	ARM	34	44	CHEST	1621	69	DRINKING	37
20	ASTHMA	35	45	CHESTTIGHT	23	70	DRIP	57
21	AT	137	46	CHILL	61	71	DRUG	56
22	ATTACK	22	47	CHILLNESS	102	72	DRY	28
23	BACK	298	48	CM	119	73	DUE	35
24	BEFORE	38	49	COLD	241	74	DULL	30
25	BILATERAL	66	50	CONSTIPATION	281	75	DURATION	118

續表 4-5 高頻文字(2)

代號	文 字	次數	代號	文 字	次數	代號	文 字	次數
76	DYSPNEA	452	101	FROM	39	126	HOARSENESS	71
77	DYSURIA	34	102	FRONTAL	23	127	HUNGER	87
78	EAR	59	103	FULLNESS	326	128	HX	40
79	ECHO	36	104	FUNCTION	33	129	HYPOCHONDRIC	58
80	EDEMA	328	105	GENERAL	369	130	HYPOGASTRIC	43
81	ENDOSCOPE	22	106	GOOD	86	131	IMPROVED	221
82	ENLARGED	38	107	GOT	41	132	IN	368
83	EPI	27	108	GPT	42	133	INCREASE	55
84	EPIGASTRIC	953	109	GREENISH	21	134	INJ	270
85	ERYTHEMA	22	110	HAD	144	135	INJ.	31
86	ESP	73	111	HALF	23	136	INSOMNIA	297
87	EXERTION	127	112	HAND	307	137	INTERMITTENT	37
88	EXT	22	113	HAPPENED	24	138	INTOLERANCE	226
89	EYE	38	114	HB	23	139	ITCHING	137
90	FACE	72	115	HBSAG	25	140	JOINT	98
91	FACIAL	21	116	HEADACHE	941	141	KG	37
92	FELL	35	117	HEARING	74	142	KNEE	75
93	FEVER	1671	118	HEART	55	143	LAST	105
94	FH	28	119	HEAT	192	144	LEFT	389
95	FINGER	57	120	HEMAPTYSIS	25	145	LEG	349
96	FLANK	135	121	HEMATURIA	36	146	LEGS	122
97	FLASH	31	122	HICCUP	47	147	LIE	32
98	FOOD	42	123	HIGH	31	148	LIMB	38
99	FOR	1674	124	HIP	24	149	LITTLE	212
100	FREQ	138	125	HISTORY	82	150	LIVER	87

續表 4-5 高頻文字(3)

代號	文 字	次數	代號	文 字	次數	代號	文 字	次數
151	LOCAL	36	176	NIGHT	553	201	ORTHOPNEA	21
152	LONG	122	177	NO	3931	202	OUT	82
153	LOOSE	100	178	NOCTURIA	25	203	OVER	94
154	LOSS	351	179	NODE	21	204	P	1881
155	LOW	242	180	NORMAL	31	205	PAIN	2991
156	LOWER	326	181	NOT	242	206	PALLOR	27
157	LT	43	182	NOW	121	207	PALPITATION	589
158	L'T	38	183	NP	41	208	PASSAGE	177
159	LUNG	25	184	NSAID	41	209	PATIENT	48
160	MALAISE	465	185	NTG	33	210	PC	297
161	MARKED	33	186	NUCHAL	205	211	PERIUMBILICA	32
162	MASS	150	187	NUMBNESS	136	212	PH	56
163	MEAL	71	188	OBSTR.	25	213	POLYDIPSIA	26
164	MENSE	22	189	OBSTRUCTION	439	214	POLYURIA	54
165	MILD	296	190	OCCASION	28	215	POOR	343
166	MINIMAL	131	191	OCCASIONAL	70	216	POST	115
167	MONTH	154	192	OCCIPITAL	26	217	POSTPRANDIAL	31
168	MONTHS	75	193	OF	274	218	PR	28
169	MORE	207	194	OFF	165	219	PRANDIAL	26
170	MORNING	105	195	OK	25	220	PUD	26
171	MUCH	133	196	ON	135	221	RADIATED	26
172	MULTIPLE	54	197	ONCE	39	222	RASH	36
173	NASAL	673	198	ONE	34	223	RECENT	33
174	NAUSEA	209	199	ONLY	42	224	RECENTLY	36
175	NECK	236	200	ONSET	29	225	REGURGITATIO	212

續表 4-5 高頻文字(4)

代號	文 字	次數	代號	文 字	次數	代號	文 字	次數
226	RESPONSE	32	251	SORENESS	325	276	TIGHTNESS	96
227	RHIN	587	252	SORETHROAT	568	277	TIME	45
228	RHINORREA	21	253	SPUTUM	1907	278	TIMES	220
229	RHINORRHEA	568	254	STILL	160	279	TINNITUS	134
230	RN	53	255	STOOL	711	280	TIRED	217
231	RT	71	256	STREAK	25	281	TO	218
232	R'T	453	257	SUBSTERNAL	81	282	TOE	22
233	SCANTY	141	258	SUDDEN	33	283	TOLD	28
234	SENSATION	65	259	SWALLOWING	44	284	TRANSFER	30
235	SEROUS	29	260	SWEATING	190	285	TREMOR	264
236	SEVERAL	536	261	SWELLING	60	286	ULCER	44
237	SEVERE	177	262	SYMPTOM	31	287	UP	51
238	SHOULDER	48	263	SYNCOPE	32	288	UPPER	57
239	SINCE	72	264	TAKING	29	289	URI	95
240	SIZE	120	265	TARRY	291	290	URINARY	132
241	SKIN	102	266	TASTE	43	291	URINE	120
242	SLEEP	78	267	TEMPERAL	41	292	VERTIGO	50
243	SLEEPY	37	268	TENDERNESS	182	293	VISION	63
244	SLIGHT	22	269	THAN	23	294	VOIDING	96
245	SNEEZING	239	270	THE	309	295	VOMITING	296
246	SO	22	271	THIRST	155	296	WAKED	21
247	SOB	524	272	THIS	55	297	WALKING	21
248	SOFT	39	273	THROAT	982	298	WATERY	229
249	SOME	147	274	THYROID	57	299	WEAK	23
250	SORE	252	275	TIGHT	754	300	WEAKNESS	95

續表 4-5 高頻文字(5)

代號	文 字	次數	代號	文 字	次數	代號	文 字	次數
301	WEEK	107	305	WHEN	245	309	WITHOUT	86
302	WEEKS	90	306	WHILE	63	310	YEAR	39
303	WEIGHT	108	307	WHITISH	100	311	YEARS	95
304	WHEEZING	592	308	WITH	2008	312	YELLOWED	48

從表 4-5 可看出每個文字都有一個代號，此代號則是用來代表研究記錄的文字，例如代號 301 是指”WEEK”；代號 311 是指”YEARS”。接著將這 312 個高頻文字的代號帶入『主觀描述』裡，把這欄位裏描述文字的部份以代號替之，若『主觀描述』裡描述的文字是屬於高頻文字，其文字代號才會出現，若其描述文字不是屬於高頻文字，其文字代號不會出現，結果範例內容如表 4-6：

表 4-6 主觀描述文字內容代號化後

病 歷 號 碼	主 觀 描 述
0029875	36 204 52 99 167 , 308 176
1274551	36 204 , 14 44 275 309 49 260 , 177
1698475	36 204 , 14 44 275 309 49 260 , 177
0007896	36 38
1007845	36 , 191 193 , 207 , 87 44 275 249
1335487	38
1235526	38
1097854	186 305 , 36
1200092	156 1 205 , 311 7
0204096	36, 38

我們將文字代號化表示後，其用意是質化的輸入值轉化為數值性資料，以便將資料輸入程式利用電腦快速運算的特性獲得分類的預測。再將代號化的記錄輸入修改後的 Apriori **演算法中去運算，並將其支持度設為** 0.1%**，其運算完後的結果即為『主觀描述』的** Large Item Sets**。其運算結果共得到** 62 **組** Large Item Sets**，內容如表** 4-7**：**

表 4-7 Large Item Sets

52 (13.5%)	116 (2.4%)	252 (1.3%)	304 (0.9%)	205 44 (1.1%)
177 (10.6%)	275 (2.2%)	160 (1.3%)	210 (0.9%)	205 84 (1.4%)
36 (9.9%)	64 (2.1%)	247 (1.2%)	76 (0.9%)	205 273 (1.0%)
205 (6.7%)	173 (2.0%)	56 (1.2%)	23 (0.9%)	205 1 (0.9%)
308 (5.9%)	255 (1.8%)	144 (1.1%)	265 (0.9%)	308 253 (2.2%)
204 (5.5%)	14 (1.8%)	132 (1.1%)	52 177 (1.2%)	99 236 (1.5%)
99 (4.9%)	236 (1.6%)	105 (1.1%)	52 308 (4.0%)	99 56 (1.0%)
44 (4.7%)	176 (1.5%)	145 (1.0%)	52 99 (1.4%)	44 275 (2.0%)
253 (4.2%)	207 (1.5%)	215 (1.0%)	52 253 (2.4%)	44 14 (1.7%)
93 (3.7%)	1 (1.5%)	156 (1.0%)	52 273 (0.9%)	275 14 (1.4%)
38 (3.1%)	65 (1.4%)	270 (0.9%)	177 93 (2.3%)	52 308 253 (2.0%)
84 (2.8%)	227 (1.3%)	112 (0.9%)	36 204 (5.3%)	44 275 14 (1.4%)
273 (2.8%)	11 (1.3%)			

決策樹產生

本研究利用病患主觀描述症狀的文字記錄，建立決策樹以推論

出病患所罹患的疾病。以病歷中敘述症狀的文字為基礎，選取其高頻項目組（以代號表示）與疾病代碼 ICD-9 作為訓練資料，並分割屬性為「有」或「沒有」兩種。再利用 C4.5 演算法來建立決策樹，分割節點為病患主觀描述文字高頻項目組的代號，決策樹最底層的樹葉節點即為疾病代碼 ICD-9。

在前面，我們已獲得『主觀描述』的 Large Item Sets 即『疾病代碼 1』裡大分類的 ICD-9 碼的資料。將『主觀描述』和『疾病代碼 1』的資料當輸入值。以 1 或 0 作為分割屬性的區別，1 代表『有』；0 代表『沒有』，帶入 C4.5 的程式工具後即可得到最佳化決策樹。其結果內容如表 4-8（完整決策樹詳見附錄一）：

本研究的資料來源是某區域教學醫院內科系的門診資料，因此研究中產生的決策樹其節點（病患主觀描述的文字）和樹葉（患者罹患疾病的 ICD-9 碼）所出現的文字或者疾病名稱大多是在內科門診時出現頻率較高的文字敘述或疾病名稱，因此本研究所產生的決策樹應可輔助內科系門診診斷時參考輔助使用。

表 4-8 決策樹

```
112 = 0:
|   210 = 1: 250 (242.0/16.2)
|   210 = 0:
|   |   265 = 0:
|   |   |   252 = 0:
|   |   |   |   255 = 0:
|   |   |   |   |   105 = 0:
|   |   |   |   |   |   65 = 0:
|   |   |   |   |   |   |   38 = 0:
|   |   |   |   |   |   |   |   205 1 = 0:
|   |   |   |   |   |   |   |   |   84 = 0:
|   |   |   |   |   |   |   |   |   |   207 = 0:
|   |   |   |   |   |   |   |   |   |   |   76 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   52 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   |   116 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   173 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   273 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   23 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   205 44 = 0:[S1]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   205 44 = 1:[S2]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   23 = 1:[S3]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   273 = 1:[S4]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   173 = 1:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   304 = 1: 464 (2.0/1.8)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   304 = 0:[S5]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   116 = 1:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   99 56 = 0:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   275 14 = 0:[S6]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   275 14 = 1:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   204 = 0:[S7]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   204 = 1:[S8]
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   99 56 = 1:
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   177 = 0: 250 (2.0/1.8)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   177 = 1: 386 (4.0/1.2)
```

第五章 結論與後續研究

第一節 結論

本研究採用醫療院所的門診資料作為研究資料來源，運用資料探勘技術中的決策樹分類與 Apriori 演算法以獲得文字敘述的 Large Item Sets，並推導出病人主觀描述病情與疾病間的關係。在研究中，利用 C4.5 演算法來自動產生最佳化的決策樹，再利用決策樹節點作為判斷，最後可以得到疾病名稱的預測結果。因此我們可以說：將 **Apriori 與 C4.5 演算法應用於醫學資料庫的資料探勘是可行的。**

研究中我們發現在內科門診資料中『主觀描述』裡的文字描述中以 **COUGH、BP、NO、PAIN** 四個文字出現頻率最高，表示在內科門診裡，病患常以 COUGH 和 PAIN 來敘述身體某部位的不舒服；BP（血壓）記錄最多，也表示量血壓在內科門診裡是常做的動作；而 NO 文字出現表示醫生在記錄病患描述症狀時，其用詞習慣以用 NO 這個文字頻率最多，例如 NO FEVER，NO COUGH YESTERDAY....等等。

本研究的結果可以幫助病患依自己的症狀特徵找出可能罹患的

疾病，或者協助醫生門診時作為診斷的參考，而使醫療品質提高讓民眾能獲得正確的醫療。

第二節 後續研究及建議

本研究還有未怠仍須再努力，所以在此提出幾點作為後進學者繼續研究的方向：

1. 利用病患門診的資料，配合 Apriori 演算法與 C4.5 演算法來推導出病患症狀特徵描述、疾病和藥物治療三者之間的關係。例如：病患症狀特徵描述是拉肚子、腹部脹氣，可利用 Apriori 演算法與 C4.5 演算法來推導出病患可能罹患的疾病是大腸急躁症，建議服用的藥物是 Duspataline 與 Gascon。這樣的話就可將 Apriori 演算法與 C4.5 演算法應用於整個門診醫療流程。

2. 本研究的結果希望可以透過人機介面呈現給使用者以方便操作，讓病患自己或醫生依症狀的特徵，透過電腦交談式的畫面而得知病患所罹患的疾病。

3. 研究的過程中，『主觀描述』欄位裡有許多文字描述存

在不確定, 例如 Has Little Fever、Cough with Little Sputum ..... 等等，這些不確定的描述對於症狀的鑑定會產生干擾，因此希望在後續研究可結合灰色（Fuzzy）理論將資料量化，以提高本實驗結果的精確度。

4. 在實驗過程中，基於本實驗的限制，將疾病代碼採以大範圍的分類，例如 150.1、150.2、150.3 .....都歸類到同一類 150 中，但是每個疾病代碼都有自己疾病名稱存在，後續研究如能將詳細的疾病代碼帶入實驗，應該會使結果的精確度提高。

5. 由於資訊與醫療非屬同一學門，故醫療知識並不是很充足，因此在研究中有些關鍵部份的判斷或認知上會有所偏差，進而影響實驗結果的準確性，因此建議後續學者在研究上多充實醫療知識或具有相關背景之學養，以期能提高研究的可信度。