私立東海大學資訊工程與科學研究所

碩士論文

指導教授：許玟斌 博士

植基於規則推導的電腦輔助醫療診斷

Rule Induction on Mining Large Database

研究生：陳世彥

中 華 民 國 九 十 七 年 九 月 二 十 九 日

摘要

我們利用資料探勘技術從大量的病歷資料中進行診斷項目與診斷結果之間的分析比較，藉此獲得綜合數個診斷項目，推導診斷結果的可行性，並進一步探討診斷結果的預測性。本研究發展的雛形系統準確率可達 95%。

在醫療診斷資料中蘊含著許多有價值的資訊，而如何自這些資料中萃取出有用的資訊，資料探勘已儼然成為不可或缺的工具。所謂資料探勘是指從大量資料或大型資料庫中由電腦自動選取一些重要的、潛在有用的資料類型或知識以做為決策分析之參考。目前資料探勘所包含的各種技術已被廣泛的應用在許多領域上，例如，商業交易資料的購物籃分析與資料檔案檢索等。

為使得大量診斷資料能成為醫護人員診斷時更有效率、更精確的輔助，本研究以機率統計與資料探勘技術為基礎，提出電腦輔助疾病診斷系統(Computer-aided Disease Diagnostic System, CDDS)。本系統設計主要分為 3 階段：第 1 階段，首先計算各診斷項目資料間的相關係數(Correlation Coefficient)，去除(Prune)係數較小的診斷項目，以達到精簡龐大資料量之目的。第 2 階段，找出各診斷項目資料之最佳分佈(Distribution)，並藉以產生隨機值以補齊診斷項目中的遺失資料(Missing Value)。第 3 階段，藉由 AND 模組運算產生重要診斷項目與診斷結果間的規則。接著，應用規則推導(Rule Induction)方法中的 J-Measure[22]，計算各規則之資訊獲益(Information Gain，即 J-Information)並保留有用的規則。最後，再佐以澳洲研究機構之甲狀腺診斷資料[24]驗證規則之正確性。根據實驗結果數據，我們提出之方法能依據診斷項目檢查值有效預測診斷結果，也直接證實了運用本方法於輔助醫療診斷之可行性。

關鍵字：維度簡化、遺失資料產生器、規則推導、病歷資料分析。

Abstract

There are lots of valuable information that are hidden in medical databases, however, it is often too tedious or too complicate to discover useful knowledge from them. So that, how to use effective methods to extract information from large medical records has become an important issue today.

The principle of data mining is in sorting through large amount of data and filtering out relevant information. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "the science of extracting useful information from large data sets or databases." To date, data mining techniques have been widely used in many fields such as education and e-commerce, etc.

By applying data mining techniques, we proposed the Computer-aided Disease Diagnostic System (CDDS), which can be used to evaluate the relationship between diagnostic items and diagnosis from a large medical database to induce valuable information, rules, and to predict the diagnoses.

CDDS takes three stages to complete the work: (1) reduces database size by calculating the correlation coefficients between diagnostic items and diagnosing decision, and prune items whose correlation coefficients are small; (2) find the best-fit probability distribution and generate random variates to fill in the missing values among those records; (3) employ AND operations on diagnostic items to generate rules, and calculate J-Information of each rule. Retain rules with higher J-Information and use them to predict the diagnostic.

In our experiment, the ratio of correctness is 95%. As you can see, by applying CDDS, we can not only extract valuable information from medical databases but also provide some aids to those medical professionals in diagnosing diseases.

Keyword: Dimension Reduction、Missing data generator、Rule Induction、Medical data analysis

目錄

圖片目錄

表格目錄

第1章 緒論

1.1. 研究動機

近年來，資訊的應用隨著業界需求，資料處理技術由資料庫資訊系統(Information System)、資料倉儲(Data Warehouse)演變到資料探勘(Data Mining)。所謂資料探勘意指：「由資料中挖掘非顯然的、未知的、潛在的「可能」有用資訊之過程」[26]；Reinschmidt[21]則認為：「資料探勘是從資料庫中萃取有效的、有用的、未知的可理解資訊，能作為決策的依據」。由此可見，妥善運用資料分析的技術將能彙整更多有用的知識以供決策參考。

隨著醫療體系的發展，民眾對醫病之間的關係也愈加地重視，在醫療人員每天必須診斷眾多病患的情況下，如何避免診斷上的疏失及錯誤，是醫療管理必須考量的問題之一[9]。在醫療人員診斷病患所罹患之疾病的過程中，通常以病患口述或顯示之症狀佐以血液等篩檢結果做為診斷的依據，而診斷是否正確有賴於醫療人員的臨床經驗及是否詳細詢問及觀察病患症狀等因素。對於醫療糾紛的產生，以病患認為醫療人員對已知症狀未能做出正確的診斷，而導致病患疾病的延誤治療或惡化最為常見。因此，如何以最少、最有效的幾個因素輔助醫療人員做出準確判斷，並減少誤診及疏忽的可能性，使得醫療糾紛能降至最低，是醫療人員或醫務管理必須思考的問題。

1.2. 研究目的

根據台灣行政院衛生署資料[2]顯示：為了提升醫療水準以創造更高的醫療產值，及節省健保支出費用，避免民眾重複就醫所造成的醫療資源浪費，自 2002 年開始，衛生署就已著手召集各大醫療院所，大力推

行「病歷電子化」，並陸續制訂相關法令規章。例如，委託台灣醫學資訊學會制定的「台灣電子病歷基本格式」。由此可見，儲存病患的診斷資料從傳統紙本病歷轉變成電子病歷(Electronic Medical Record, EMR)乃勢在必行。因此，如何從醫療院所的電子病歷資料庫中找出診斷項目與診斷結果之間的關聯性(Relations)，做出診斷結果預測以提供醫療參考，同時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，已成為如何利用診斷資料的重要研究主題之一。

1.3. 研究流程

在本研究中，我們以資料探勘及統計分析為基礎，提出有效利用病歷資料輔助醫療診斷之系統。本系統主要流程有 3 個階段：第 1 階段，首先計算各診斷項目資料間的相關係數(Correlation Coefficient)，去除(Prune)係數較小的診斷項目，以達到精簡龐大資料量之目的。第 2 階段，找出各診斷項目資料之最佳分佈(Distribution)，並藉以產生隨機值以補齊診斷項目中的遺失資料(Missing Value)。第 3 階段，藉由 AND 模組運算產生重要診斷項目與診斷結果間的規則。接著，應用規則推導(Rule Induction)方法中的 J-Measure[22]，計算各規則之資訊獲益(Information Gain，即 J-Information)並保留有用的規則。最後，再佐以澳洲研究機構之甲狀腺診斷資料驗證規則之正確性。

1.4. 論文結構

本論文結構如下：第 2 章中，我們將對資料探勘技術於醫療之相關研究及背景知識做一完整說明； 第 3 章中，我們將詳細說明主要研究步驟與流程；第 4 章中，我們以澳洲研究機構之甲狀腺診斷資料[24]為例，以 80%-20%比例隨機抽樣，分成訓練資料(Training Data)與測試資

料(Testing Data)來驗證本研究發展方法之可行性；最後，在第 5 章中，我們提出彙整本研究的結論。

第2章 文獻探討及背景知識

2.1. 文獻探討

近年來，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來建立醫學知識，進而找出各種疾病的醫療指引[9]。因此，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。

目前資料探勘所包含的各種技術已被廣泛的應用在許多醫療診斷領域上。例如，藉由醫療資料庫中的歷史資料，運用資料探勘技術和貝氏網路(Bayesian Network)等方法，將資料有規則的一面呈現出來，並找出醫師對同類疾病的醫療行為模式，作為建構臨床路徑(Clinical Pathway)的參考[3]。以線性判別分析(Linear Discriminant Analysis, LDA)、主成分分析(Principle Component Analysis, PCA)結合類神經網路(Artificial Neural Networks, ANN)輔助機器學習鑑別青光眼[6]。透過以關聯法則為基礎之模式，分析牙醫病歷中關聯資訊，建置一知識庫為基礎之牙醫決策支援系統，提供牙醫師診斷之參考[5]。Abdel-fattah 利用多群判別分析(Multi-group Discriminant Analysis, MDA)結合血清檢驗與放射性治療產出一組線性函數(產出新變數)，再透過 ROC 曲線(Receiver Operation Characteristics)分析新變數的解釋能力，來預測 C 型肝炎是否病變為肝硬化[6]。利用關聯規則(Association Rule)之 FP-Growth 演算法，從健保資料檔中快速地發掘疾病間隱藏的關係，所獲得規則包含有關病患罹患某種疾病後，再罹患其他疾病的條件機率，其探勘結果能有效成為醫生與醫療研究人員進行醫學研究的助力[10]。以貝氏網路(Bayesian

Network)、決策樹(Decision Tree)與倒傳遞神經網路(Back Propagation Neural Network, BPN)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[11]。運用 C4.5 決策樹分析法及倒傳式類神經網路，從醫療單位對泛可黴素進行血中藥物濃度監測程序(Therapeutic Drug Monitoring, TDM)監控的歷史案例中，建構出可用以預測泛可黴素在病患身上的作用結果之分類模式，有效協助醫藥人員掌握泛可黴素的療效，降低了可能的醫療資源浪費[1]。

上述相關研究，皆證實資料探勘技術在輔助、甚至預測醫療診斷結果之可行性。因此，我們提出以資料探勘技術為基礎，結合統計分析與規則推導理論，實做一輔助醫療診斷系統。在下面章節中，我們將詳細說明本研究主要流程。

2.2. 知識探索

在 1992 年由 Frawley 等學者首先定義「知識探索」 (Knowledge Discovery, KD) 的內涵，是指從資料中挖掘非顯然的、未知的、潛在的「可能」有用資訊之過程 [26]，要完成知識探索的過程主要包含五個要素：

1. 存放資料本身的資料庫。
2. 被探索的領域相關的知識。
3. 處理過程參與的使用者。
4. 探索知識所使用的方法。
5. 對所探索出的知識的表達與應用。

1996 年 Fayyad 等更進一步的定義資料庫的知識探索(Knowledge Discovery in Databases, KDD)一詞；KDD 是一個指出資料中令人信服的、潛在有用的一個非細瑣流程，其最終目的是瞭解資料的樣式[25]。他將 KDD 的發展與進行分為五個步驟，如圖 1。

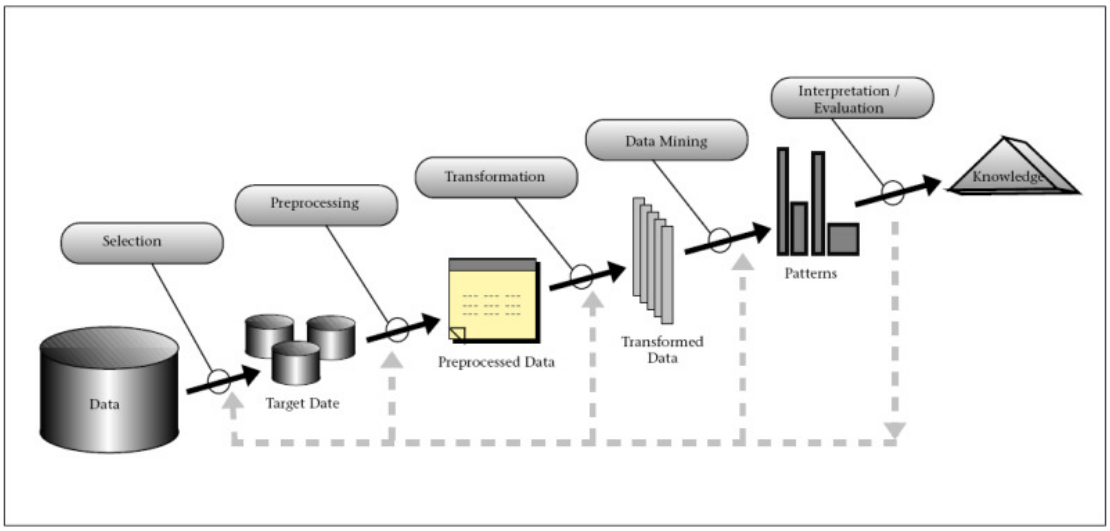

圖 1、知識探索概觀圖

Brachman and Anand[14]又將上述流程細分為九個步驟：

1. 加強對目的領域的應用與知識的瞭解，從使用者的角度清楚的定義進行知識探索的目標。
2. 針對希望探索的資料建立標的的資料庫。
3. 對資料做前置處理，包括雜訊的去除與解釋，收集模組(Model)所必須的資訊、掌握遺失資料的欄位以及定義時間、順序有關的資訊。
4. 資料的歸納與規劃，包括尋找對完成目標有用的資料屬性欄位、應用維度精簡(Dimension Reduction)或轉換的方法(Transformation Methods)簡化資料。
5. 依據第一步驟所定義的目標，選擇適合的資料探勘(Data Mining)方法，如摘要法(Summarization)、分類技術(Classification)、分群技術(Clustering)、迴歸分析法(Regression)等。
6. 進行探索的分析(Analysis)、模組以及假設(Hypothesis)的選擇，包括資料探勘演算法以及尋找資料樣式(Data Patterns)方法的選擇，決定使用的模組(Model)與使用的參數值。
7. 執行探勘尋找所希望的資料樣式，例如，分類的規則(Rule)或決策樹(Decision Tree)、分群後的群組(Clusters)。

8. 根據探勘的結果，解譯資料樣式與包含之意義。
9. 完成報告與現行的知識做比較，進一步應用所得的知識，改善現行的作業，並擴及其他的系統。

2.3. 資料探勘

「資料探勘」一詞是泛指從巨大的資料庫中粹取，綜合出未知資訊為主軸的複雜活動之通俗的講法，它是資料庫知識探索所有處理程序中的一個步驟 ，另一方面也是指有關於為了現實生活問題所存在的大量資料所作的各種領域的研究或發展的演算法及開發的軟體環境，與 KDD 相同的 DM 的處理通常亦可以分為下列五個步驟：

1. 資料選擇：由資料探勘處理的選擇目標和工具所組合的，辨識所要採掘的資料，然後選擇適合的輸入項屬性和輸出項的資訊來呈現給交付之工作。
2. 資料轉換：包括下列操作：(a)以想要的方法來組織資料(b)轉換資料的形式(如將符號轉為數字)(c)定義新的屬性 縮減資料的幅員(d)去除不必要的部份與主題無關的部份(e)標準化(f)如果適當的話，決定如何處理遺失的資料的策略。
3. 資料的探勘：就步驟的本身而論，接下來是對這些轉換後的資料進行採掘，使用一種或多種的技術來粹取感興趣的樣式。
4. 結果詮釋與驗證(Result Interpretation and Validation)：對所探勘出的結果進一步的解釋建立模型，並應用已知的評估方法及未使用的資料庫中資料來測試它的正確度。
5. 組織所探索到的知識(Incorporation of the Discovered Knowledge)：將結果呈現給決策者，做選擇或決定與目前的認知所潛在的衝突，將被粹取的知識應用於新探勘到的模型。

在進行資料探勘時，依據所要探勘的目的及資料的性質，通常會進行下列幾種工作：

2.3.1. 摘要(Summarization)

主要是對給予的資料產生結構上或特質的概略性描述。它能採取多個形式： 數字(如簡單的統計描述語：平均值、標準差等)，圖表形式，例如，直方圖，散佈圖(Scatter Plots)，或者是 「if-then」形式的規則。它可能在整個資料庫裡或者在選擇的子集裡關於所推測的對象提供描述。

2.3.2. 分群(Clustering）

主要是對給予未知歸類特性的資料中找出彼此特性相似的群組，其目的是要將組與組之間的差異找出來，同時也要將一個組之中的成員的相似性找出來。

2.3.3. 分類(Classification)

是根據一些變數的數值做計算，再依照結果作分類。(計算的結果最後會被分類為幾個少數的離散數值，例如將一組資料分為「可能會回應」或是「可能不會回應」兩類)。分類會用一些已經分類的資料來研究它們的特徵，然後再根據這些特徵對其他未經分類或是新的資料做預測。這些我們用來尋找特徵的已分類資料可能是來自我們的現有的歷史性資料，或是將一個完整資料庫做部份取樣，再經由實際的運作來測試；譬如利用一個大的郵寄對象資料庫的部份取樣來建立一個 分類模型(Classification Model)，以後再利用這個模型來對資料庫的其他資料或是新的資料作預測。

2.3.4. 迴歸分析(Regression)

迴歸分析(Regression）屬於建造一個些許透通的模型的監督式(Supervised）學習問題，其主要目的是做預測，目標是使用一系列的現有數值發展一種能以一個或多個預測變數的數值做為應變數，以預測一個連續數值的可能值的方法，可透過古典或更先進統計方法和以經常在分類任務過程中使用的符號式(Symbolic)方法來進行。

2.3.5. 變動及偏移偵測(Change and Deviation Detection)

這項工作主要是在發現資料的實際內容與被預期的內容（先前所預估的）或標準值間是否有顯著的變動、誤差或偏移，這些變動可以包含時間上的偏差或群組間的差異。

2.3.6. 依賴度(從屬性)模型 (Dependency Modeling)

依賴度模型的問題在於發現一個模型以描述屬性間顯著的的依賴或從屬關係，這些依賴度通常被以「if antecedent is true then consequent is true」的「if-then」規則形式表示。

2.3.7. 時序問題(Temporal Problem)

Time-Series Forcasting 與 Regression 很像，只是它是用現有的數值來預測未來的數值。Time-Series Forcasting 的不同點在於它所分析的數值都與時間有關。Time-Series Forcasting 的工具可以處理有關時間的一些特性，譬如時間的階層性(例如每個星期五個或六個工作天)、季節性、節日、以及其他的一些特別因素如過去與未來的關連性有多少。

2.3.8. 因果關係(Causation Modeling)

這是一個在資料的屬性中發現因果關係的問題，使用一個「if-then」形式的因果規則，表明條件(前項)和規則的當然結果(後項)之間有相互關係。

2.3.9. 關聯規則(Association Rule)

Association Rule 是要找出在某一事件或是資料中會同時出現的東西：項目 A 是某一事件的一部份，則項目 B 也出現在該事件中的機率有 n%。例如，一個顧客買了低脂乳酪以及低脂優酪乳，那麼這個顧客同時也買低脂牛奶的機率是 85%。

2.3.10. 屬性導向歸納法(Attribute Oriented Induction)

屬性導向歸納法是一種以歸納屬性為基礎的資料分析技術，其技術核心為線上資料歸納方法，將相關式表格(Relational Dataset)資料集合中的每一個屬性，檢查其資料的分佈，判斷應歸納到那個相關的抽象層級。

2.3.11. 樣式導向相似性搜尋 (Pattern-Based Similarity Search)

在時間或時間-空間資料庫搜索相似的樣式，經常會應用到兩種查詢類型：(a)物件關聯相似度查詢(Object-relative Similarity Query)，亦即相似度查詢(Similarity Query)或範圍查詢(Range Query)，在所收集到的物件中，尋找使用者指定的範圍或距離中，符合的物件。(b)完全關聯相似度查詢（All-pair Similarity Query），亦即空間聯合（Spatial Join）目標是找到彼此都是在一段使用者指定的範圍或距離內的全部相符的要素。

2.3.12. 資料方塊法 (Data Cube)

資料方塊法一般概念為將經常被要求的高成本計算具體化，尤其是計數(Count)、總計(Sum)、求平均數(Average)、取最大值(Max)等的歸納函數，將歸納後的具體化景觀儲存在一個多重維度資料庫(資料方塊)，可供決策支援、知識發現及其他應用做參考。

2.3.13. 序列樣式探勘 (Sequence Pattern Mining)

在包含時序關係的資料庫中尋找一定數量所支持的序列樣式，主要是找出關聯順序進行行為模式上的預測，例如若 A 事件發生，則 B 事件可能接著會發生。

2.4. 規則推導

規則推導(Rule Induction)的主要涵義是從群訓練案例中尋找出最佳的、正確的、可了解的分類方法的規則[17]。較常見的規則推導方法大致上有：以樹狀推導(Tree Induction)的表達方式，例如，C4.5 演算法；以類神經網路(Neural Network)的各連結(Link)權重(Weight)表達方式；以及 J-Measure[22]等方式。

2.4.1. 決策樹推導

決策樹的推導(Decision tree induction)是一種使用樹狀架構的方法來做分類，節點代表不同的 feature，樹枝為 feature 的值，而樹葉則是不同的分類類別(class label)。

這種方式是先找一個最佳的特徵作為根節點，所有的資料以此根節點為判斷根據，進行分類，分類在每一個分支的資料再選出最佳的特徵作為根節點，再進行分類，形成一棵子樹，如此的過程一直重複，直到在一個分支內的所有資料都屬於同一個類別，推導過程才算結束，這個

最終的分支就會形式樹葉，裡面記載著該樹葉內的資料所屬的類別，這樣就會形式一棵決策樹，如圖 2 所示。

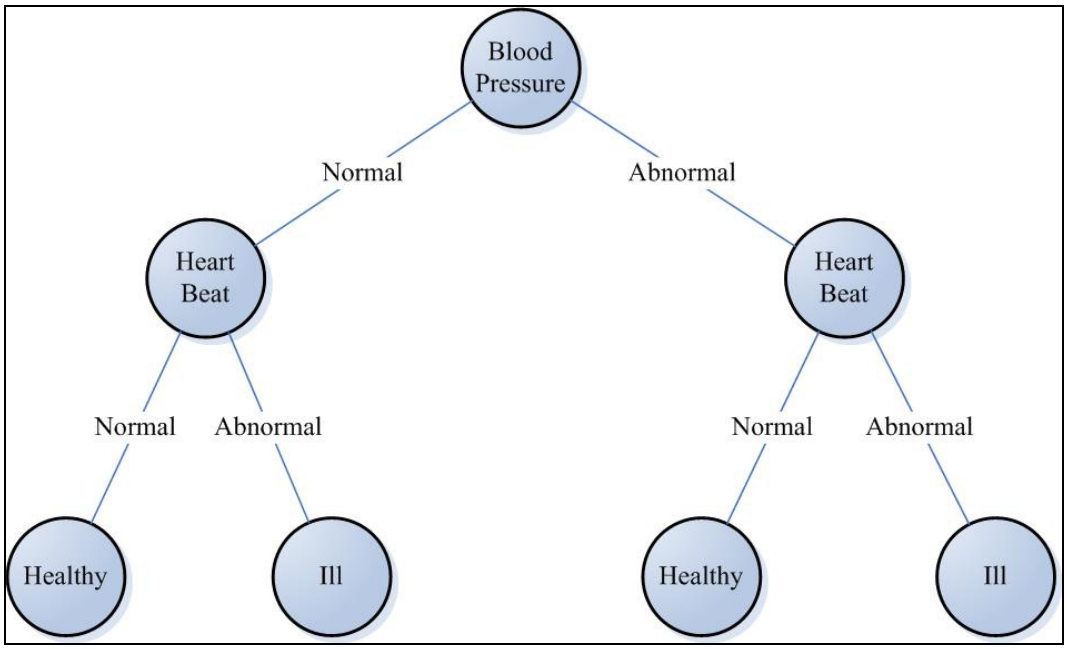

圖 2、決策樹示意圖

樹的大小決定於分支的節點數，希望用最少的節點就可以分出結果，因此原則上希望使分支後的每一個節點，就所要預測的變數而言，同質性越高越好(Homogeneous)，可利用資訊理論(Information Theory)的 Entropy 理論來定義評估標準，同質性高者包含較少資訊，因此 Entropy 比較小。資訊理論主要的觀念是利用互斥資訊(Mutual Information)的原理，計算某資訊對於問題的不確定性(Uncertainty，或稱為熵(Entropy))能夠降低多少。例如，假設一個資料集 N={1, 2, 3, 4}，某人自該資料集中隨機挑選一個數字 ε，而我們該問幾個問題才能知道他選的數字是多少？針對這個問題，若有下列提問：(1) ε 是否為 4？(2) ε 是否為 3 或 4？；從這兩個提問，可以輕易地看出我們能獲得的資訊並不同，提問(2)顯然比提問(1)能給予我們更多有關 ε 的資訊。底下是如何計算決策規則之資訊確定性量測的介紹[13]。

關於此類問題，一般不確定性(Uncertainty)定義如下：

$$H(N)=\sum_{n=1}^{k}-P(n)\log_2 P(n) \tag{1}$$

公式(2)中之$\log_2(X)$為X的以 2 為底的對數值，$P(n)$是$n=\varepsilon$的機率，而k為資料集N內資料的數量。當針對某問題之提問的回答可能的集合為$Q=\{q_1, q_2, ..., q_c\}$時，對ε所剩餘之不確定性(又稱為平均離散條件資訊, Average Discrete Conditional Information)之定義如下：

$$H(N\mid Q)=\sum_{q=q_1}^{q_c}\sum_{n=1}^{k}P(q)P(n\mid q)\log_2 P(n\mid q) \tag{2}$$

其中$P(q)$是在資料集N中回答為q的機率，$P(n\mid q)$是在資料集N中所有回答為q的資料內，$n=\varepsilon$的機率，k為資料集N內資料的數量。接著，針對某問題之提問後資訊獲得(Information Gain)之定義如下：

$$I(N;Q)=H(N)-H(N\mid Q) \tag{3}$$

因此，$I(N;Q)$便可代表針對某問題之某提問所能獲得的資訊。而我們通常把某問題之提問視為“規則(Rule)”，因此，上述的核心觀念便是“針對一個問題去尋找資訊獲得最大的規則”，決策樹即是以此理論發展而成。

2.4.2. 類神經網路推導

所謂類神經網路(Artificial Neural Network)，它是一種平行計算系統，包含硬體與軟體，它使用大量的相連人工神經元來模仿生物神經網路能力。

為何我們要去模仿生物神經網路？因為我們都知道現今電腦雖然擅長執行高速的複雜計算，而且所得結果具有高度精確與可靠性，但還是有許多工作人腦比電腦強的多，例如：語音辨識、圖形辨識、人臉辨識等…因此我們使用類神經網路來模仿生物神經網路的資訊處理系統。

在現代智慧型控制的領域裡，類神經網路已成為現代智慧型控制的主流，類神經網路(Artificial Neural Network)，或從字面直譯為人工類神經網路，乃指模仿生物神經網路的資料處理系統，其為模仿生物神經網路的能力之計算系統，故使用大量簡單的相連人工神經元，從外界或其他神經元取的資訊後，經過簡單的運算，最後將其結果輸出到外界或其它神經元。總而言之類神經網路即是利用現今電腦的優點—高速處理複雜計算的能力、以彌補其缺點—對於樣本識別和專職決策能力的不足，而在其應用上非常廣泛，幾乎涵蓋了各行各業以及其他相關的應用科學。

類神經網路顧名思義，其網路架構是模仿生物神經網路，整個網路可大致分為三個部分：神經元(又稱處理單元，Processing Element, PE)、層(Layer)、網路(Network)。

神經元其輸入輸出的關係式，可用以下函數式子表示：

$$Y_j = f\left(\sum_i W_{ij} X_i - \theta_j\right)$$

其中 Y_j=模仿生物神經元模型的輸出訊號。

f=模仿生物神經元模型的轉換函數(Transfer Function)，是一個將輸入值乘上權重加總後再經轉換成人工神經元輸出值的數學式。

W_{ij}=模仿生物神經元模型的神經節強度，又稱連結權重(Weight)值。

X_i=模仿生物神經元模型的輸入訊號。

θ_j=模仿生物神經元模型的偏權值。

圖 3 為神經元之示意圖：

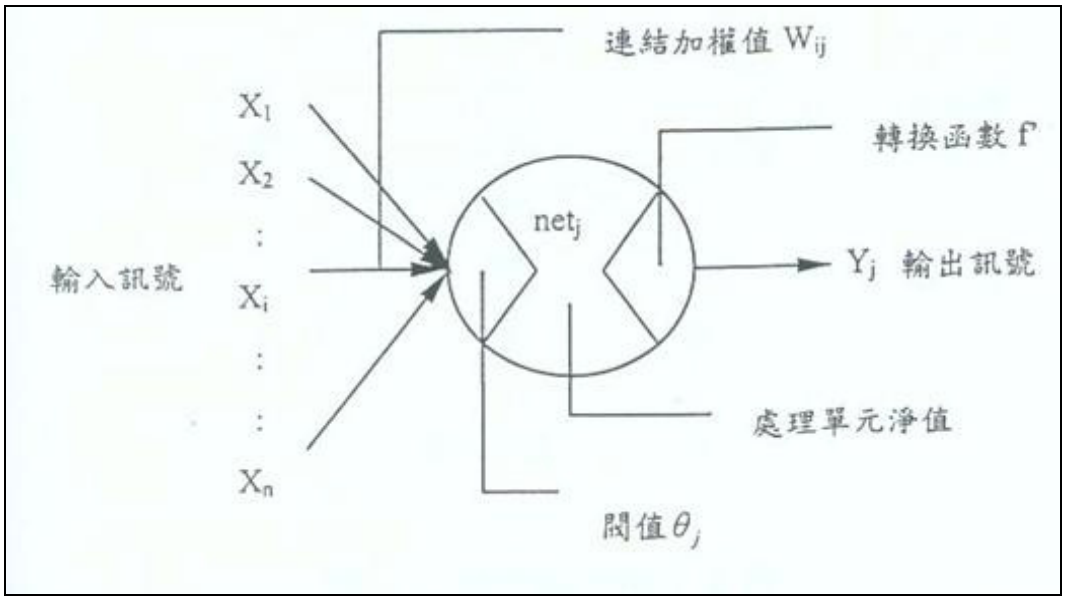

圖 3、神經元示意圖

若干個相同作用的處理單元之集合便形成一個層(Layer)，幾個「層」進行堆疊集合，就成為了「網路」。一個神經網路的架構分為輸入層、隱藏層、輸出層，由下方做說明：

輸入層：用來表現網路的輸入變數，其神經元數目依問題而定。

隱藏層：用來表現輸入神經元間的交互影響，其處理神經元數目並沒有標準的方法以做決定，通常要以試驗的方式來決定其最佳數目，網路可以不只一層隱藏層，也可以沒有隱藏層。

輸出層：用來表現網路的輸出變數，其神經元數目依問題而定。

如同在生物神經網路之中，神經元的強度可視為生物神經網路儲存資訊的所在，神經網路的學習即在調整神經結的強度。類神經網路各處理單元之間則以連接鍵互相連結，整個類神經網路的記憶就存放於這些連接鍵之中，以權重(Weight)來表示。圖 4 為整體類神經網路之示意圖：

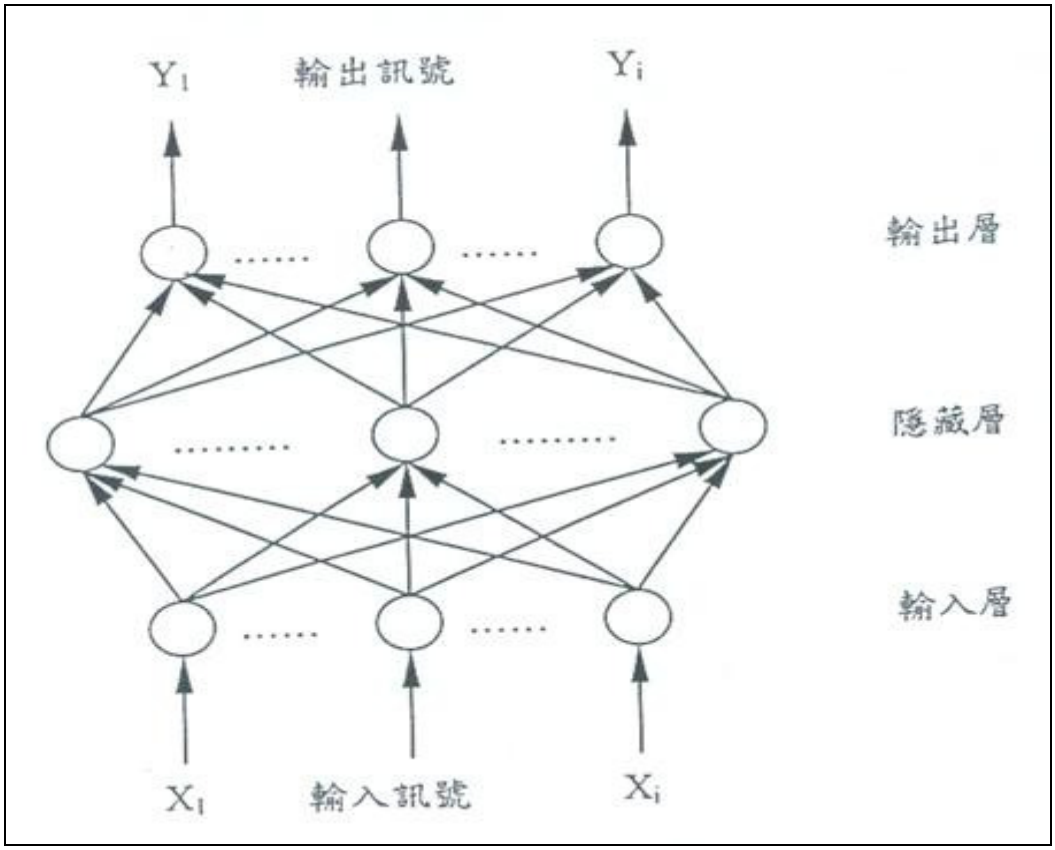

圖 4、類神經網路示意圖

類神經網路的運作過程可分為兩個階段，亦即：(1)學習過程(Learning)：網路依學習演算法，從範例中學習已調整網路連結加權值，使網路的輸出盡可能和期望的輸出值一樣。若網路達到穩定的狀態時，則學習過程即可終止。(2)回想過程(Recall) ：網路依回想演算法，以輸入資料決定輸出資料的過程。

學習演算法基本上都是由「能量函數」(Energy Function) 推導而得。能量函數是用來衡量網路的學習效果，因此網路的學習過程即是使能量函數最小化的過程。學習演算法可分為三類：(1) 監督式學習(Supervised Learning)：網路在學習的過程中，每一筆學輸入的資料都有一個對應的期望輸出值，來監督網路的學習，學習的目標為調整處理單元間的連接權值以降低網路推論輸出值與期望值之間的差距。由於網路需要不斷的學習與調整，監督式學習通常在學習的過程中需要多次的循

環及較長的時間才能夠得到較好的結果。(2) 非監督式學習(Unsupervised Learning)：相對應於監督式學習，在非監督式學習中，網路的每筆輸入樣本並沒有相對應的期望值，網路學習的目的為降低網路優勝單元的連結加權值所構成的向量與輸入向量的距離，以達每個輸出單元的連結加權值向量可以代表一群訓練樣本在空間中的聚類型態。由於非監督式學習較注重於聚類分析， 因此在學習的過程中已包含了分類，學習並不會花費太多的時間。(3) 網路的處理單元狀態變數所組成的向量是用來表示一個式樣(Pattern)。網路學習旨在使從初始狀態變數向量(初始式樣)，經「聯想」迭代所得的最終狀態變數向量(最終式樣)，其與網路所記憶的式樣之一相同或近似。

倒傳遞類神經網路(Back-propagation Neural Network, BPN) 屬於監督式學習網路，適於應用在診斷、預測等問題的實驗分析與探討，是目前類神經網路模式中最具代表性，且應用最廣泛、最普遍的類神經網路之一。其基本原理是利用最陡坡降法(The Gradient Steepest Descent Method) 的觀念，將誤差函數予以最小化。其學習過程通常以一次一個訓練範例的方式進行，直到學習完所有訓練範例，即一個學習回合(Learning Epoch)，一個網路可以訓練範例反覆學習，直到網路的學習達到收斂。

2.4.3. J-Measure

利用互斥資訊(Mutual Information)的原理，計算某資訊對於問題的不確定性(Uncertainty，或稱熵(Entropy))能夠降低多少。但此方法較決策樹優越的地方在於：J-Measure 針對資料集 *N* 中資料區分為數個類別(Class)，再以各類別中的區域(Region)進行計算，而不是單純將 *N* 視為一個類別(Class)；因此，J-Measure 可以計算單一規則(某一類別中的某區域)所獲得的資訊，獲得更佳的推導結果。

第3章 研究步驟與方法

資料探勘的主要流程如圖 5；在確定研究方向或要解決的問題後，進行(1)相關資料(Data)的蒐集，接著(2) 將原始資料分組(Grouping)轉換成較有意義的資訊(Information)，最後(3)在所有資訊中擷取出對我們有用的知識(Knowledge)。

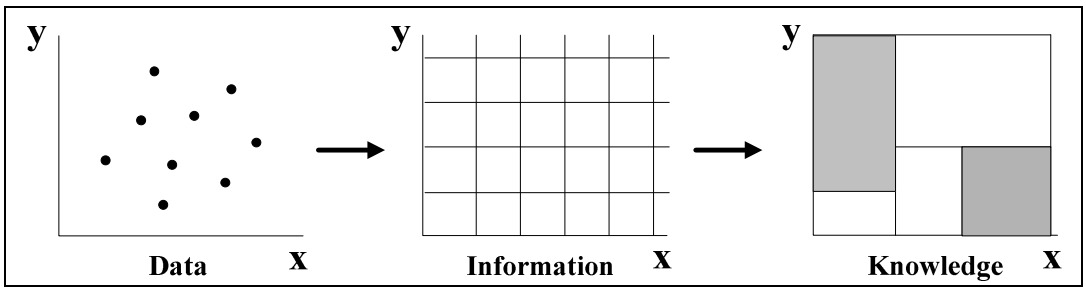

圖 5、資料探勘過程圖

因本研究主要應用資料探勘技術於輔助醫療診斷，遵行上述探勘流程，我們將系統設計如圖 6。本系統流程主要分成 3 階段：(1)資料蒐集部分，透過檔案伺服器或資料庫伺服器取得原始資料； (2)將取得之資料分別藉由相關係數計算模組、分佈配對模組、分組模組及規則產生模組以產生決策規則；(3)將預處理後之資料(規則)透過規則引導模組及結果預測模組的進一步運算，提取出對我們有用的知識(從診斷項目預測結果)。下面我們將詳細介紹上述系統各流程之運算方法與實做方式。

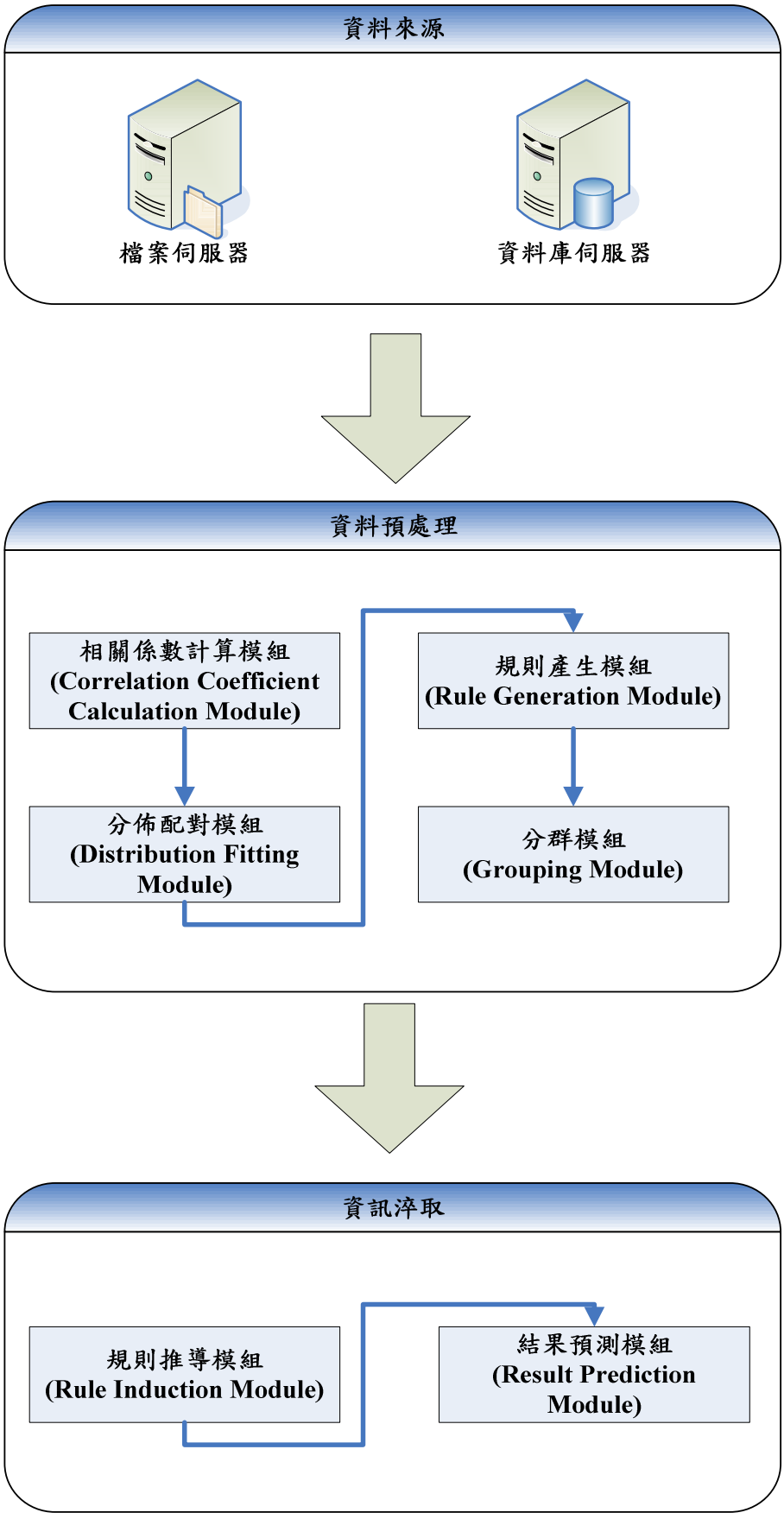
資料來源
檔案伺服器
資料庫伺服器
資料預處理
相關係數計算模組
(Correlation Coefficient Calculation Module)
規則產生模組
(Rule Generation Module)
分佈配對模組
(Distribution Fitting Module)
分群模組
(Grouping Module)
資訊萃取
規則推導模組
(Rule Induction Module)
結果預測模組
(Result Prediction Module)

圖 6、系統流程圖

3.1. 相關係數與維度刪減

在資料探勘過程中，常遇到因資料數量龐大，而導致探勘時效率不彰的窘境。於是，如何能夠有效地簡化(Simplify)或修剪(Prune)欲處理的大量資料且不失去原本資料隱含的資訊，也是一個資料探勘重要的研究方向。

資料修剪(Data Pruning)的觀念主要是刪除(1)與欲探勘結果較不相關，或(2)可能誤導探勘結果的資料，以減少整體的資料量與計算量，並增加最後探勘結果的精確度。我們採用的資料庫中，屬性種類高達 23 種，其前十筆如表 1，表中空白地方為沒有資料(null)。

表 1、澳洲研究機構之甲狀腺診斷資料(前十筆)

age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	I131 treatment	query hypothyroid	query hyperthyroid	lithium	goitre	tumor	hypopituitary	psych	TSH	T3	TT4	T4U	FTI	TBG	diagnoses
29	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.3						0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.6	2	128				0
41	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0						11	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						26	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						36	1
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						26	0
77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						21	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	3	116				0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.2	2	76				0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.9	2	83				0

而其屬性名稱與資料型態如表 2。若我們能有效精簡屬性項目，對整體探勘效率將會大有助益。

表 2、屬性名稱與資料型態

Name	Type	Name	Type	Name	Type
sex	Boolean	lithium	Boolean	T4U	Numeric
on thyroxine	Boolean	goitre	Boolean	FTI	Numeric
query on thyroxine	Boolean	tumor	Boolean	TBG	Numeric
on antithyroid medication	Boolean	hypopituitary	Boolean		
sick	Boolean	psych	Boolean		
pregnant	Boolean	diagnoses	Boolean		
thyroid surgery	Boolean	age	Numeric		
I131treatment	Boolean	TSH	Numeric		
query hypothyroid	Boolean	T3	Numeric		
query hyperthyroid	Boolean	TT4	Numeric		

3.1.1. 相關係數

針對數值(Numeric)型態的資料，我們採用相關係數的方式來做資料修剪。所謂相關係數，是衡量兩數值變數之線性關係強度及方向(正、負)的參數。樣本相關係數 r 之計算公式[17]如下：

$$r_{X,Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}} \tag{4}$$

公式(1)中，X, Y 為兩數值變數，$i = 1, 2, \ldots, n$，其中 $\overline{X}, \overline{Y}$ 為 X 與 Y 的平均數。經過上式計算後的各屬性與目標屬性相關係數表如表 3。接著我們將進行維度刪減的動作。

表 3、各屬性與目標屬性(Diagnoses)相關係數表

Name	diagnoses	Name	diagnoses
age	0.028	lithium	-0.028
sex	-0.089	goitre	-0.031
on thyroxine	0.071	tumor	0.025
query onthyroxine	0	hypopituitary	0.025
on antithyroid medication	0.01	psych	-0.072
sick	-0.011	TSH	0.24
pregnant	0.113	T3	-0.035
thyroid surgery	-0.006	TT4	0.122
I131treatment	0.005	T4U	0.095
query hypothyroid	0.055	FTI	0.108
query hyperthyroid	0.038	TBG	0.53

3.1.2. 維度刪減

因為相關係數與各數值變數的度量單位無關，故資料庫中各屬性間的相關係數能夠用來當作比較依據。本研究中，我們將以每個屬性與診斷結果(Diagnoses)屬性的相關係數為依據進行屬性修剪。例如，我們挑選屬性與診斷結果間相關係數較大者，作為後續資料處理的依據。又，因為大部分屬性型態為布林值(Boolean)，隱含的資訊量不足，故本模型先將非布林型態的屬性刪除之後得到：TSH、T3、TT4、T4U、FTI 與 TBG 六項，接著，再取出相關係數較大(>0.1)者，最後屬性剩下：TSH、TT4、FTI 與 TBG。因此，原本 23 項屬性便刪減為 4 屬性，精簡了規則產生所需的計算量。同時，我們使用主成分分析，針對所有屬性做一次計算，其結果如下。

表 4、主成分分析表

Component Matrix(a)					
	Component				
	1	2	3	4	5
age	-0.026	-0.437	0.436	-0.108	-0.292
sex	-0.2	-0.423	-0.361	0.086	0.097
on thyroxine	0.382	0.095	0.485	0.056	0.296
query on thyroxine	0.07	0.064	0.058	0.748	-0.076
on antithyroid medication	-0.012	0.344	-0.167	-0.094	-0.066
sick	-0.052	-0.184	0.098	0.021	-0.576
pregnant	0.176	0.554	-0.189	-0.006	-0.104
thyroid surgery	-0.046	0.174	0.142	0.116	0.164
I131 treatment	0.003	0.017	0.236	-0.102	0.182
query hypothyroid	-0.062	0.111	0.471	-0.066	0.216
query hyperthyroid	0.082	0.477	-0.089	-0.155	-0.107
lithium	-0.019	0.044	-0.096	-0.024	0.261
goitre	-0.023	0.085	-0.171	0.005	-0.035
tumor	0.062	0.295	-0.052	-0.019	-0.417
hypopituitary	0.013	0.05	0.011	0.742	0.002
psych	-0.008	-0.188	-0.457	0.008	0.405
TSH	-0.529	0.257	0.327	0.002	0.187
T3	-0.013	-0.021	-0.052	-0.094	0.025
TT4	0.885	0.024	-0.007	-0.039	0.038
T4U	0.015	0.1	-0.049	0.032	0.053
FTI	0.848	-0.199	0.004	-0.022	0.021
TBG	0.026	0.015	-0.091	-0.028	0.043

由表 4，我們發現各屬性所隱含資訊之重要性幾乎與相關係數相同，也再次驗證利用相關係數法來做維度刪減的可行性。

3.2. 產生遺失資料

一般而言，當我們準備分析一個資料庫/資料集(Data Set)時，通常會先對該資料庫資料集做所謂“資料預處理(Data Preprocessing)”的動作，確保資料的完整性(Completeness)、寂靜性(Noiseless)以及一致性(Consistence)。其中一項在解決資料完整性時常遇到的問題便是遺失資

料(如表 1 中的空白地方)。這類資料通常被標示為“空白(blank)”、“NaN(Not a number)”或者“？”。而解決遺失資料的方法大致上有下面 5 種：

1. 忽略該筆資料(此方法當遺失資料太多時，例如，超過所有資料的一半，最後的探勘結果可能會不具說服力)。
2. 以手動方式為遺失資料填入數值(此方法太過困難及主觀，最後的探勘結果可能會不具公信力)。
3. 利用該遺失資料之屬性內的所有資料平均(Mean)或者中位數(Median)取代之(此方法當該屬性的資料中有離群點(Outlier)存在時，最後探勘結果可能會不具合理性)。
4. 利用迴歸(Regression)分析或者決策樹(Decision Tree)等方法預測該遺失資料的數值(此方法只能針對有出現過的資料進行預測，當面臨從無出現過的資料時，最後的探勘結果可能會不具精確性)。
5. 在該屬性的所有資料中尋找最佳分佈配對(Best Distribution Fitting)，並以該分佈與其機率密度函數(Probability Density Function)產生隨機數值取代之。(此方法需先求出分佈及機率密度函數，其求解過程通常十分繁複)。

上述 5 點中，每點皆有其缺陷。經我們評估後，第 5 種方法因為使用分佈與機率的概念來產生遺失資料，較其他方法具備公信力、合理性及精確性。因此在本研究所使用之資料庫的 5000 筆病歷資料(Record)中，我們將以尋找最佳分佈配對並產生隨機資料的方式解決各屬性內的遺失資料。

下列 3 個步驟為本研究解決遺失資料之流程：

1. 取得某屬性資料內之最大與最小值，並制訂組距，繪製直方圖。
2. 使用 EasyFit®[19]軟體，計算該些資料之各種可能分佈及參數後，利用 Kolmogorov Smirnov[14]檢定之結果尋找最佳分佈。

3. 利用最佳分佈產生隨機數值以補齊遺失資料。

下面是我們利用最佳分佈所產生的隨機變數修正原始資料後的結果。

表 5、遺失資料修正表

Before Filling Missing Value						After Filling Missing Value				
TSH	TT4	FTI	TBG	diagnoses		TSH	TT4	FTI	TBG	diagnoses
			11	0		14.0966	134.026	123.902	11	0
			26	0		5.89702	97.3913	107.516	26	0
			36	1		1.90474	123.241	75.3877	36	1
			26	0		1.2642	122.676	132.27	26	0
			21	0		0.91469	124.205	57.8129	21	0
			36	1		1.33908	95.8903	98.8349	36	1
70	3.9	5	28	1		70	3.9	5	28	1
			36	1		10.5152	71.8251	89.4685	36	1
			23	0		1.79822	107.783	140.614	23	0
	145	144	26	0		0.75405	145	144	26	0

3.3. 資料分組

為繼續精簡資料數量，減少運算時間，我們以編製直方圖所使用的組數替代某屬性內之資料進行分組(Grouping)動作，例如，我們編製直方圖時將屬性 A 之資料分為 16 組，則將數值落於第 1 組距內的資料編號為 A1，落於第 2 組距內的資料編號為 A2，以此類推。如此一來，若原本某屬性內有 5000 筆各不相同的數值資料，即 5000 個組別；經此分組方法後，將精簡為 16 組，大幅縮短了後續 AND 模組的計算時間。表 6 為經過遺失資料修正後的資料與分組後資料對照表。

表 6、資料分組結果

TSH	TT4	FTI	TBG	diagnoses		TSH	TT4	FTI	TBG	diagnoses
14.10	134.03	123.90	11.00	0		2	8	7	1	0
5.90	97.39	107.52	26.00	0		0	5	6	3	0
1.90	123.24	75.39	36.00	1		0	7	4	5	1
1.26	122.68	132.27	26.00	0		0	7	7	3	0
0.91	124.21	57.81	21.00	0		0	7	3	2	0
1.34	95.89	98.83	36.00	1		0	5	5	5	1
70.00	3.90	5.00	28.00	1		10	0	0	3	1
10.52	71.83	89.47	36.00	1		1	4	5	5	1
1.80	107.78	140.61	23.00	0		0	6	8	3	0
0.75	145.00	144.00	26.00	0		0	8	8	3	0

3.4. 規則產生與 J-Measure 應用

經上述“資料預處理”的過程後，我們現在已經得到精簡而具代表性的資料。我們將使用這些資料來產生規則，並根據產生的規則與測試資料來做甲狀腺腫瘤的預測。

3.4.1. 規則產生

因為我們的最終目的在透過診斷項目預測診斷結果，所以我們需要比較具體化(Specific)的規則，才能根據診斷項目確切判斷其結果，而如果使用 XOR、OR 等運算來產生規則，將會使產生的規則較一般化，較不具代表性；反之，透過 AND 運算能產生較明確，較具體的規則。故規則產生的過程，我們採用 AND 運算來完成。假設經過 3.1 節及 3.3 節之資料刪減程序後，剩餘的屬性為 A、B、C、D、E 及 F，而 F 為目標屬性，經過 AND 運算，我們產生之規則如表 7，其中 n-D Rule 代表第 n 個重要項目集合(large itemset)，$1 \le n \le 5$。而本實驗資料所產生的規則於表 11。

表 7、經過 AND 運算後得到之規則表

1-D Rule	A->F	B->F	C->F	D->F	E->F					
2-D Rule	A,B->F	A,C->F	A,D->F	A,E->F	B,C->F	B,D->F	B,E->F	C,D->F	C,E->F	D,E->F
3-D Rule	A,B,C->F	A,B,D->F	A,B,E->F	A,C,D->F	A,C,E->F	A,D,E->F	B,C,D->F	B,C,E->F	B,D,E->F	C,D,E->F
4-D Rule	A,B,C,D->F	A,B,C,E->F	A,B,D,E->F	A,C,D,E->F	B,C,D,E->F					
5-D Rule	A,B,C,D,E->F									

3.4.2. J-Measure

就決策樹相關技術之理論基礎是將資料集 N 視為一個類別(Class)，而不是針對 N 中資料區分為數個類別(Class)，再以類別中的區域(Region)進行計算，因此無法計算單一規則(某一類別中的某區域)所獲得的資訊。有鑑於此，本研究採用針對單一規則設計的 J-Measure[22]，其作法詳細說明如下：

若以類別及區域的觀念取代將整體資料及視為單一類別的作法，則公式(4)中 $I(N,Q)$ 可視為 $I(C,R)$ ，其中 $C \subseteq N$ 且 $R \subseteq Q$。若將 $I(C,R)$ 轉換為期望值之呈現方式，公式(4)可改寫為：

$$\begin{aligned} I(C,R) &= H(C) - H(C \mid R) \\ &= E_{C_i,R_j}\left[\log_2 \frac{P(C_i \mid R_j)}{P(C_i)}\right] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{k} P(R_j)P(C_i \mid R_j)\log_2 \frac{P(C_i \mid R_j)}{P(C_i)} \end{aligned} \tag{5}$$

其中 n 為區域的數目，k 為類別數目，$P(C \mid R)$ 為在類別 C 中區域 R 出現的機率，而 $P(C)$ 為 C 在整體資料中出現的機率。

在利用規則推導方法所計算得到的規則中，通常包含兩部分：資料適合度(Data Fit)及意識適合度(Mental Fit)；資料適合度通常指某筆資料符合某規則之機率，即一個計算數值。而意識適合度通常為某規則對於整體資料的適合程度，即該規則適不適宜用來當作評斷整體資料的一個標準。在 J-Measure 中之資料適合度(j-information)定義如下：

$$j(C;R)=\sum_{i=1}^{k}P(C_i\mid R_j)\log_2\frac{P(C_i\mid R_j)}{P(C_i)} \tag{6}$$

將公式(6)帶入公式(5)，則我們可以得到公式(7)：

$$I(C;R)=E_{R_j}[j(C;R_j)] \tag{7}$$

根據公式(7)， 意識適合度(J-Information $J(C;R_j)$)，如下式：

$$J(C;R_j)=P(R_j)\,j(C;R_j) \tag{8}$$

最後，因為一個規則主要是提供關於自身所屬的類別(C_m)及其互補類別(not C_m)的資訊，因此公式(8)可以成為：

$$\begin{aligned}J(C;R_j)=P(R_j)P(C_m\mid R_j)\log_2\frac{P(C_m\mid R_j)}{P(C_{m)}}\\ +P(R_j)(1-P(C_m\mid R_j))\log_2\frac{1-P(C_m\mid R_j)}{1-P(C_{m)}}\end{aligned} \tag{9}$$

本研究中，便是利用公式(9)來計算 3.4.1 節中產生的各規則所能獲得的資訊(即，J-Information)，並挑選出 J-Information 較大的規則當作主要規則，進一步根據測試資料的診斷項目預測診斷結果。

下面一節中，我們將以實驗方式比較 J-Measure、決策樹及類神經網路三種不同模型之預測準確度優劣，並驗證本研究方法之可行性。

第4章 實驗分析與探討

4.1. 實驗環境

於硬體方面，我們採用 Intel® Core™2 Quad CPU Q6600 2.4GHz。基於開發速度及便利性，我們使用 Python 及 C#為主要的程式語言。而在統計分析應用工具上，我們選擇較具公信力的 EasyFit®[19]與 SPSS®[23]。最後，在實驗資料部份，我們以澳洲研究機構之甲狀腺診斷資料[24]。

4.2. 實驗分析

本實驗中，分別比較了本論文提出之輔助醫療診斷系統、決策樹以及類神經網路三種模型。各模型之結果將於下面小節中詳細敘述。

4.2.1. 輔助醫療診斷系統

原本資料庫中共有 5000 筆資料與 23 項屬性。表 8 為使用相關係數刪減後的剩餘診斷項目與其相關係數。

表 8、診斷項目與相關係數表

		TSH	TT4	FTI	TBG	diagnoses
TSH	Correlation	1	-0.290289741	-0.272957959	-0.117757396	0.252924823
	N	4683	4643	4434	32	4683
TT4	Correlation	-0.290289741	1	0.734792749	-0.025716308	0.096213405
	N	4643	4653	4441	42	4653
FTI	Correlation	-0.272957959	0.734792749	1	-0.200290398	0.101728131
	N	4434	4441	4444	41	4444
TBG	Correlation	-0.117757396	-0.025716308	-0.200290398	1	0.529773894
	N	32	42	41	349	349
diagnoses	Correlation	0.252924823	0.096213405	0.101728131	0.529773894	1

在繪製值方圖時，我們將資料分為 16 組。表 9 為屬性 TSH、TT4、FTI 及 TBG 之最小與最大值、組距及遺失資料統計表。

表 9、最小最大值及組距及遺失資料數量表

Name	Min	Max	Range	Missing Value Count
TSH	0.01	103	6.436875	317
TT4	2	257	15.9375	347
FTI	1.3999	274	17.03750625	556
TBG	0.1	114	7.11875	4651

經過 EasyFit®[19]計算後各屬性之直方圖與最佳分佈如圖 7 到圖 10。

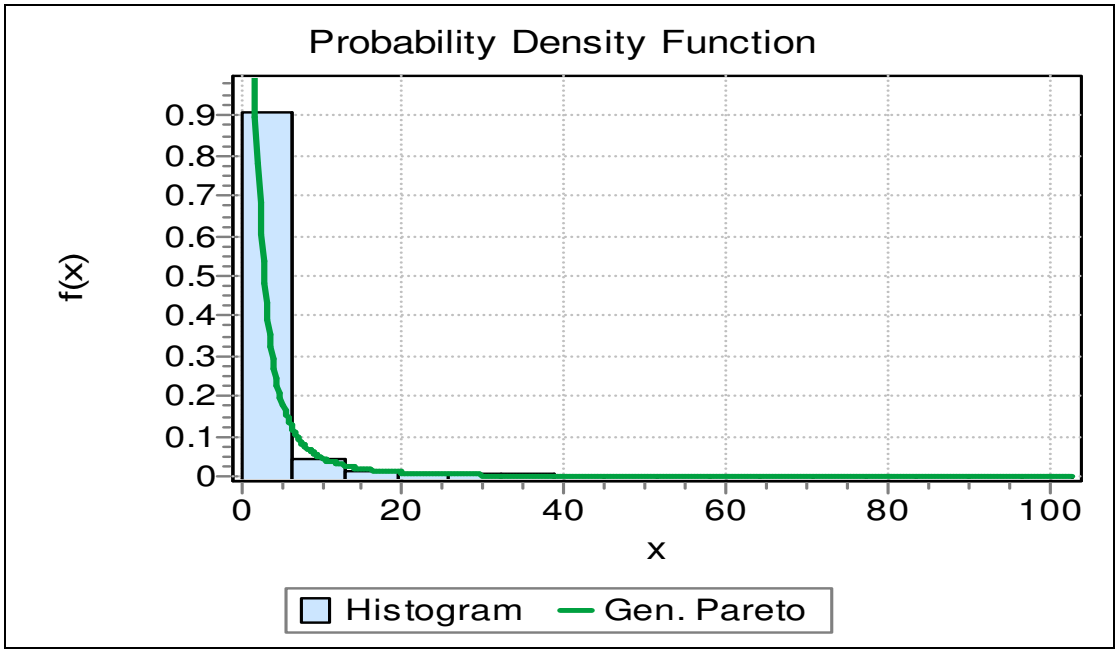

圖 7、TSH 最佳分配圖

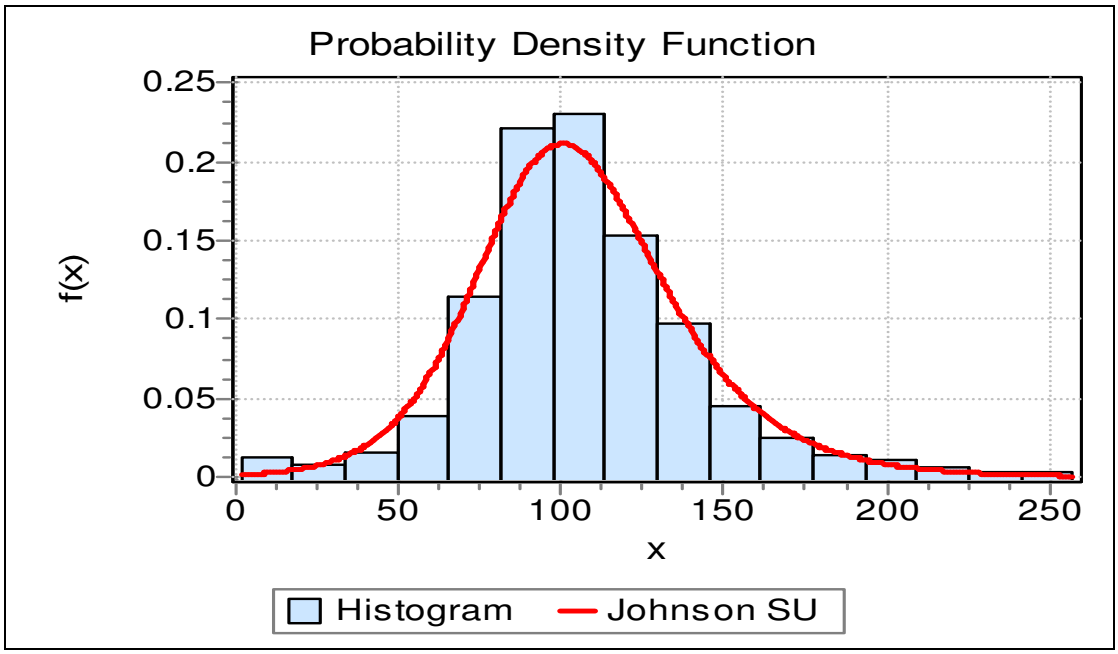

圖 8、TT4 最佳分配圖

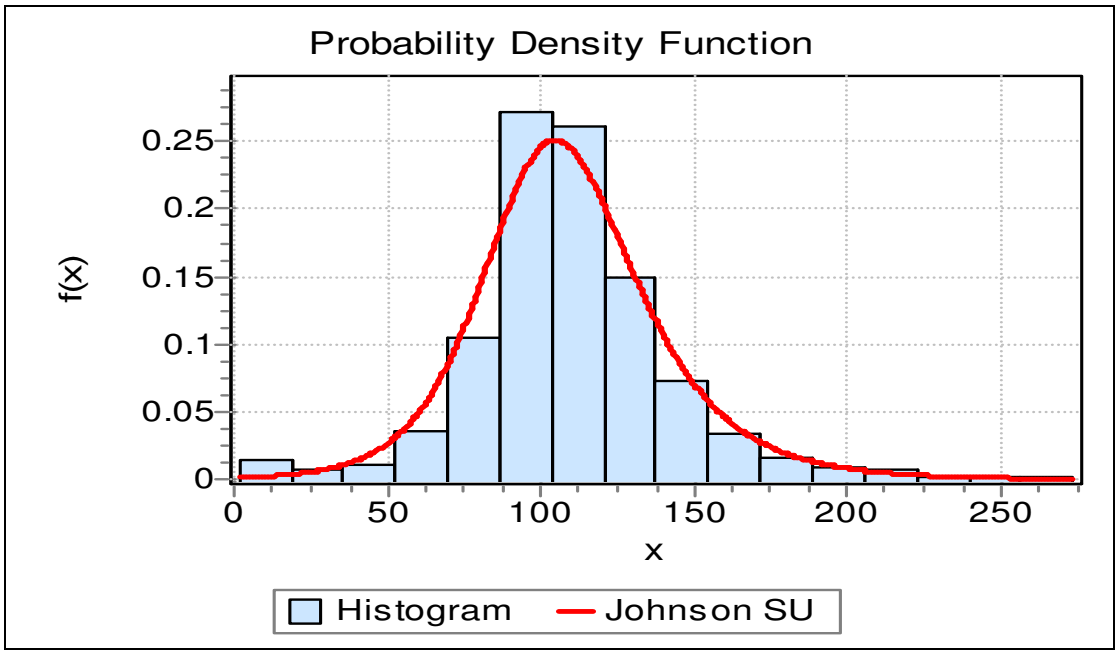

圖 9、FTI 最佳分配圖

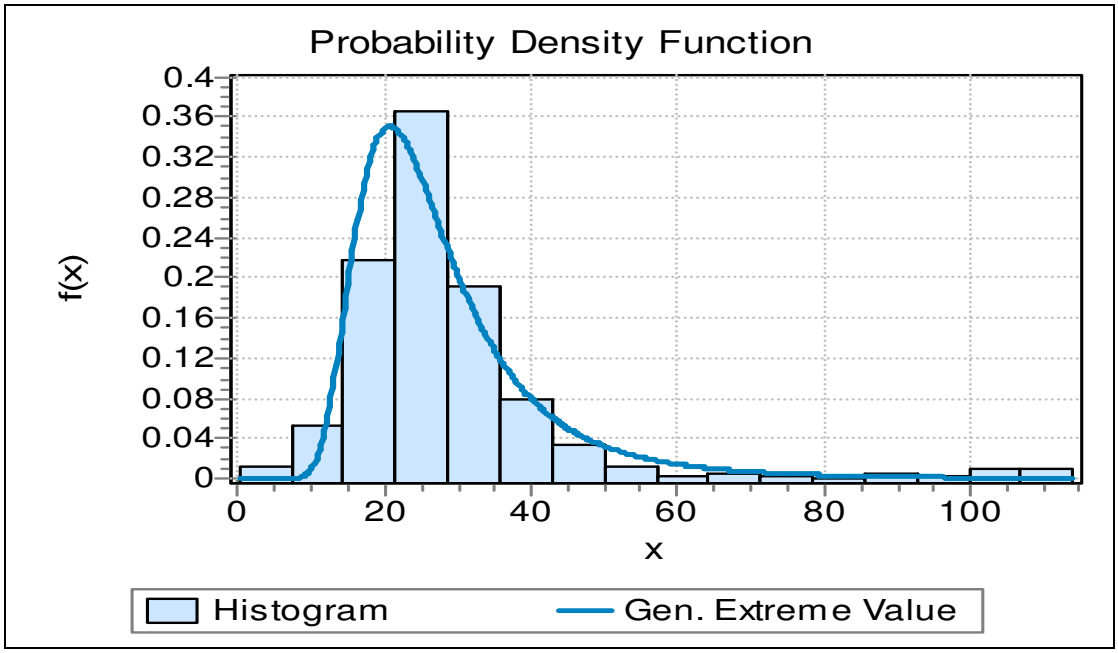

圖 10、TBG 最佳分配圖

各分配之相關參數如表 10：

表 10、屬性最佳分佈與參數表

Name	Best Distribution	Parameters			
TSH	Gen. Pareto	k=0.642945	σ=1.1944	μ=-0.03091	
TT4	Johnson SU	γ=-0.630443	δ=1.99268	λ=57.6398	ζ=85.6671
FTI	Johnson SU	γ=-0.414067	δ=1.66687	λ=44.179	ζ=96.4283
TBG	Gen. Extreme Value	k=0.206357	σ=7.61126	μ=22.0326	

依據最佳分佈與參數估計值產生隨機數值取代遺失資料後，我們利用 AND 運算產生候選規則及計算每條規則之 J-Information，表 11 為候選規則前 10 筆的運算結果。

表 11、規則及 J-Information 表

No.	Rule		J-Information
1	IF TSH=1 AND TBG=5	THEN diagnoses=1	0.008739
2	IF TSH=2 AND TT4=4	THEN diagnoses=1	0.007599
3	IF TSH=2 AND FTI=4	THEN diagnoses=1	0.007219
4	IF TT4=9 AND FTI=10	THEN diagnoses=1	0.007219
5	IF TSH=0 AND TT4=9 AND FTI=10	THEN diagnoses=1	0.007219
6	IF TSH=2 AND TBG=2	THEN diagnoses=1	0.006839
7	IF TT4=1 AND FTI=1	THEN diagnoses=1	0.006839
8	IF TSH=1 AND TT4=4 AND FTI=4	THEN diagnoses=1	0.006839
9	IF TSH=1 AND FTI=5 AND TBG=3	THEN diagnoses=1	0.006839
10	IF TSH=6	THEN diagnoses=1	0.005699

我們從整體資料 5000 筆隨機抽取 4000 筆(80%)為訓練資料集，其餘的 1000 筆(20%)為測試資料。這 1000 筆測試資料再依據訓練資料探勘的規則計算預測準確率，並重複實驗 50 次。準確率結果如圖 11。

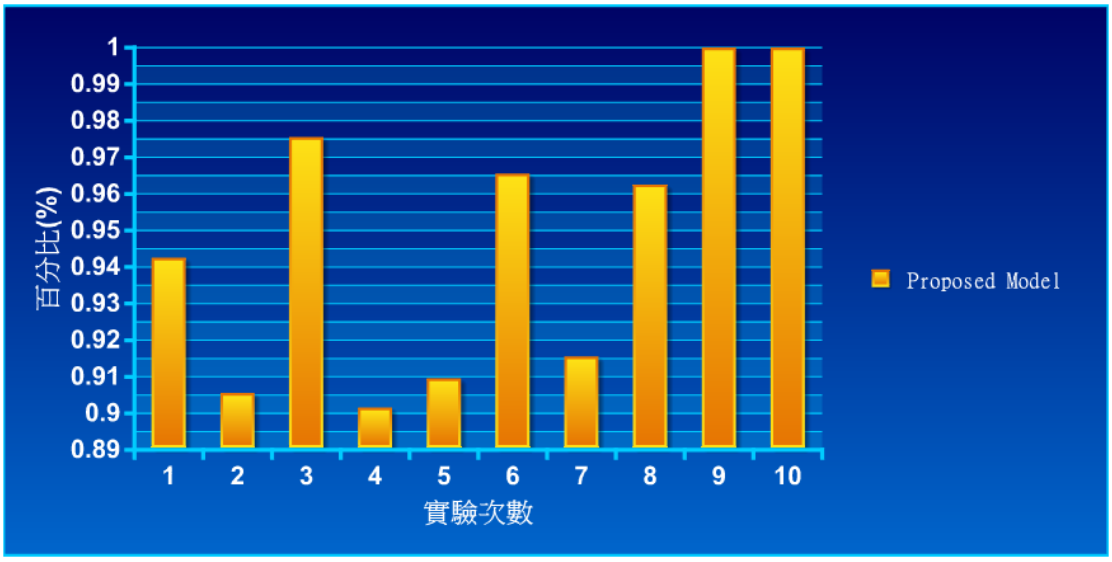

圖 11、本論文提出之模型實驗次數與準確率

實驗結果顯示，本系統前 10 次實驗平均預測準確率為 0.9474，而 50 次實驗之平均為 0.94928，確實驗證了本研究方法的可行性。

此外，為比較本研究提出之模型的優劣，我們依照上述實驗環境與設定，分別以決策樹與類神經網路兩種理論為基礎，進行了相同的實驗，其產出與結果將於下兩節中說明。

4.2.2. 決策樹模型

決策樹模型的實驗，我們以現有工具[23]，結合手邊的資料，首先得出該決策樹的模型圖，如圖 12 所示。

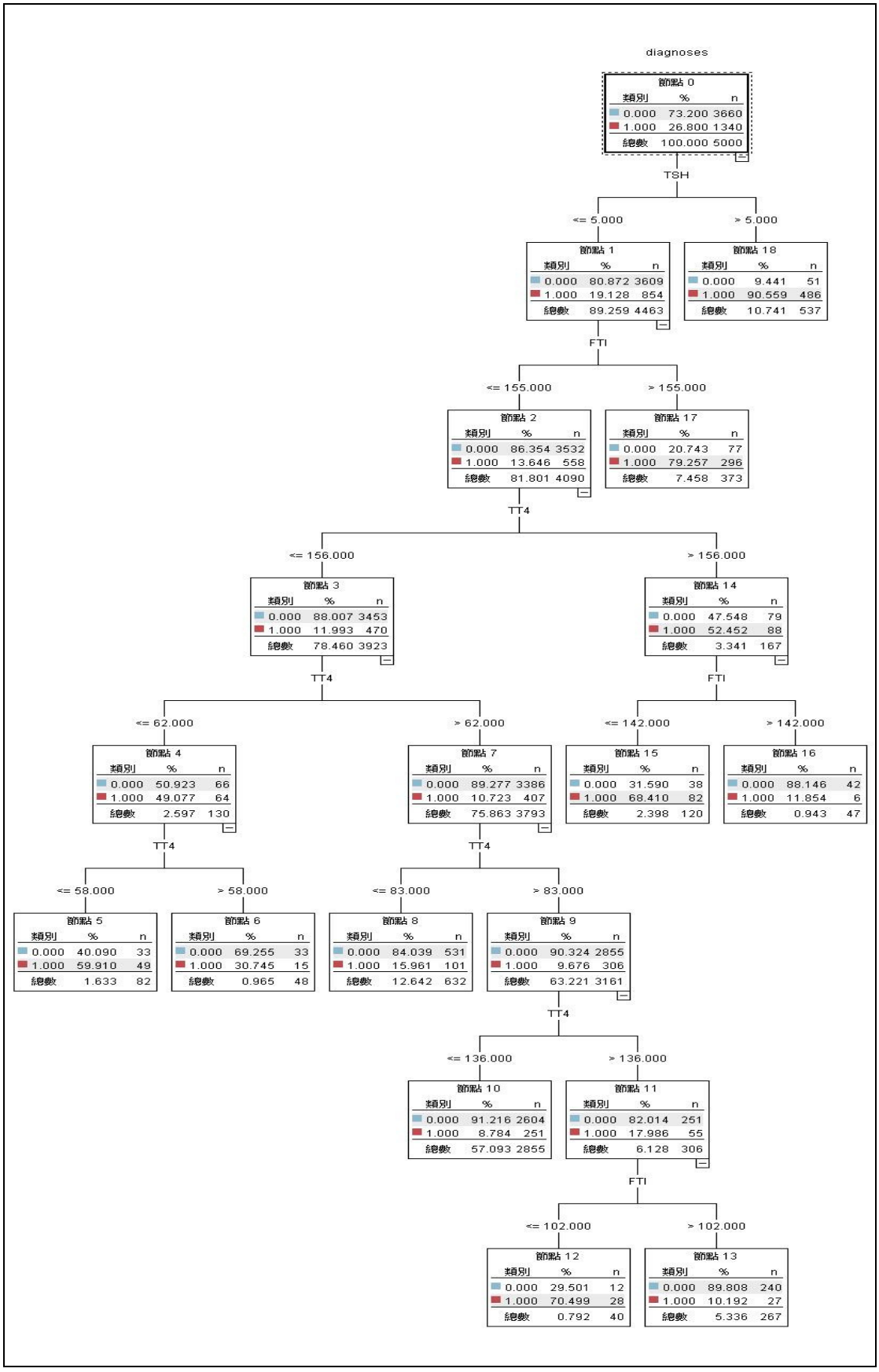
diagnoses
節點 0
類別 % n
0.000 73.200 3660
1.000 26.800 1340
總數 100.000 5000
TSH
<= 5.000
> 5.000
節點 1
類別 % n
0.000 80.872 3609
1.000 19.128 854
總數 89.259 4463
節點 18
類別 % n
0.000 9.441 51
1.000 90.559 486
總數 10.741 537
FTI
<= 155.000
> 155.000
節點 2
類別 % n
0.000 86.354 3532
1.000 13.646 558
總數 81.801 4090
節點 17
類別 % n
0.000 20.743 77
1.000 79.257 296
總數 7.458 373
TT4
<= 156.000
> 156.000
節點 3
類別 % n
0.000 88.007 3453
1.000 11.993 470
總數 78.460 3923
節點 14
類別 % n
0.000 47.548 79
1.000 52.452 88
總數 3.341 167
TT4
FTI
<= 62.000
> 62.000
<= 142.000
> 142.000
節點 4
類別 % n
0.000 50.923 66
1.000 49.077 64
總數 2.597 130
節點 7
類別 % n
0.000 89.277 3386
1.000 10.723 407
總數 75.863 3793
節點 15
類別 % n
0.000 31.590 38
1.000 68.410 82
總數 2.398 120
節點 16
類別 % n
0.000 88.146 42
1.000 11.854 6
總數 0.943 47
TT4
TT4
<= 58.000
> 58.000
<= 83.000
> 83.000
節點 5
類別 % n
0.000 40.090 33
1.000 59.910 49
總數 1.633 82
節點 6
類別 % n
0.000 69.255 33
1.000 30.745 15
總數 0.965 48
節點 8
類別 % n
0.000 84.039 531
1.000 15.961 101
總數 12.642 632
節點 9
類別 % n
0.000 90.324 2855
1.000 9.676 306
總數 63.221 3161
TT4
<= 136.000
> 136.000
節點 10
類別 % n
0.000 91.216 2604
1.000 8.784 251
總數 57.093 2855
節點 11
類別 % n
0.000 82.014 251
1.000 17.986 55
總數 6.128 306
FTI
<= 102.000
> 102.000
節點 12
類別 % n
0.000 29.501 12
1.000 70.499 28
總數 0.792 40
節點 13
類別 % n
0.000 89.808 240
1.000 10.192 27
總數 5.336 267

圖 12、決策樹模型

圖 13 是使用決策樹模型搭配測試資料所得到的實驗結果。我們可以發現，前十次平均預測準確率只有 0.8833，與本實驗所提出之模型有顯著的落差。

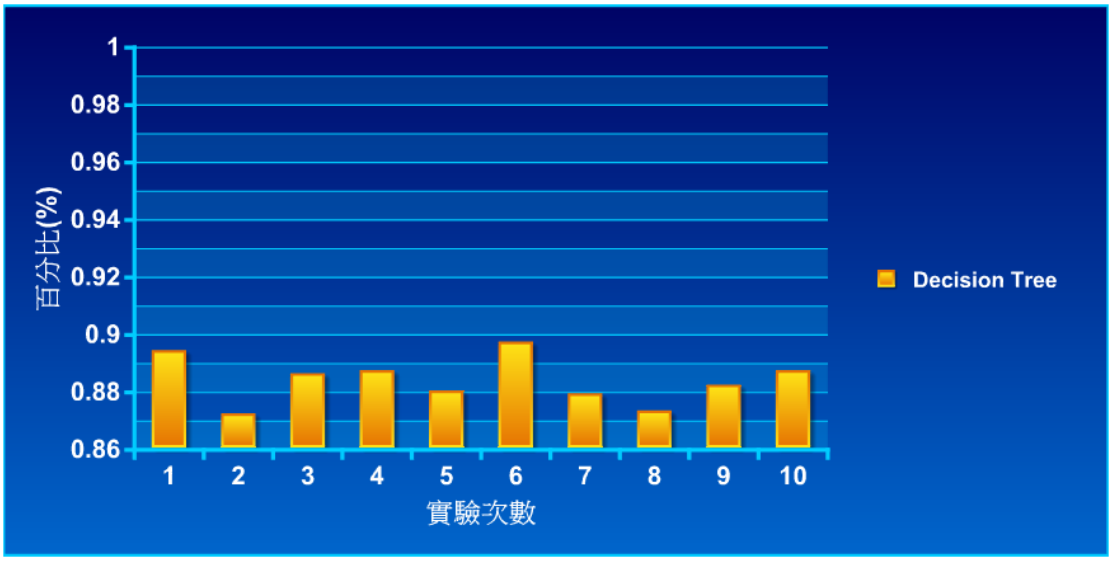

圖 13、決策樹實驗次數與準確率

4.2.3. 類神經網路模型

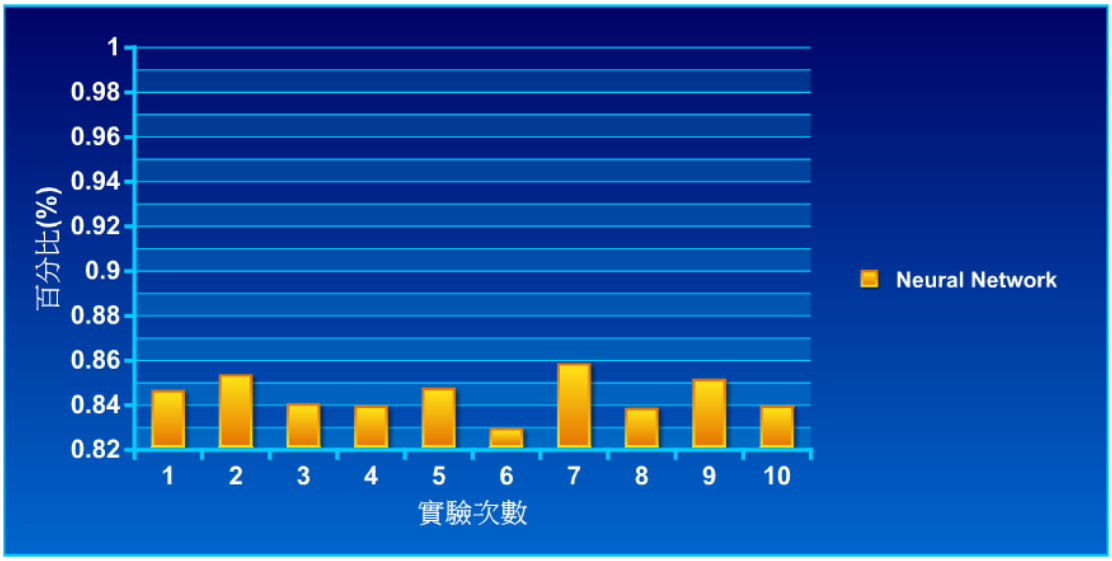

圖 14、類神經網路實驗次數與準確率

圖 14 是類神經網路模型的實驗結果。前十次平均預測準確率只有 0.844556，比起決策樹模型，又有明顯的落差。最後，我們綜合比較三種方法的準確率。結果如圖 15。

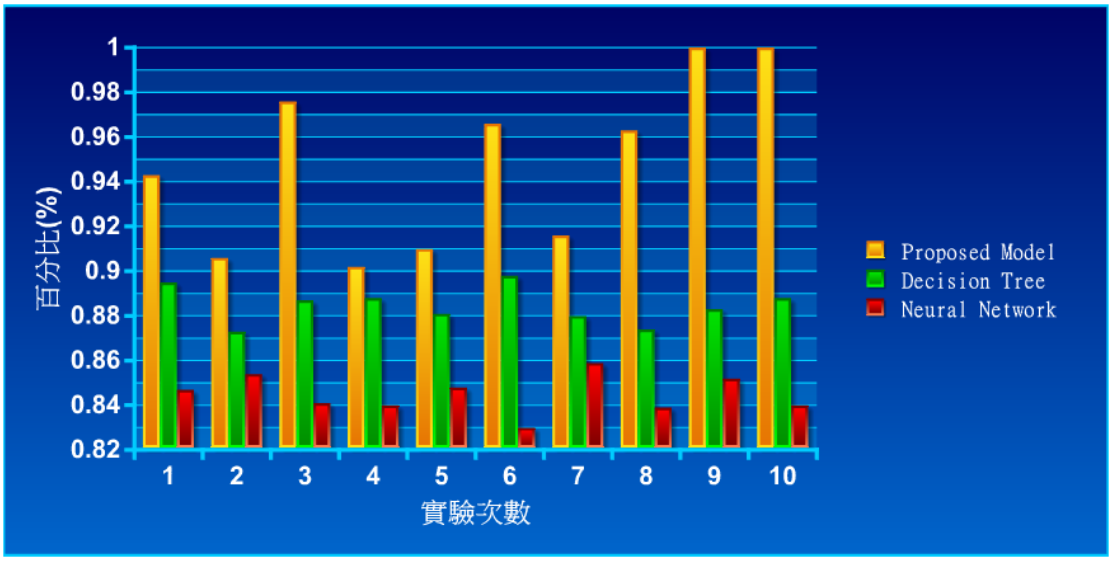

圖 15、所有模型準確率綜合比較圖

圖 15 中可以明顯看出，相較於其他模型，本模型具有十分卓越的預測準確率。

第5章 結論與未來研究

在病患就診的醫療過程中，因醫療人員疏失而導致誤診是醫療糾紛中最為常見的因素之一，如何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的問題。從大量的病歷資料中，擷取出醫療人員在診斷疾病及問診症狀的智慧與知識，對醫療人員避免於診斷及問診過程中的疏失，必定可以提供相當有用的參考資訊。

在本研究中，我們藉由相關係數的計算，尋找出資料中較可能有用的資訊，再透過最佳分佈的隨機數值替代遺失資料。最後應用本研究定義的模組運算與 J-Measure 來產生規則，預測率高達 0.94928。由此可見，本研究的探勘結果，對醫療人員在診斷疾病錯誤的預警、及降低在診斷疾病過程中的疏忽，都可以提供非常有用的參考資訊。

在本實驗中，因模組運算後的規則數量龐大，規則計算所花費的時間平均為 10-20 分鐘，加上程式使用之資料結構與資料型態均以物件(Object)裝箱(Boxing)，運算時需先拆箱(Un-boxing，即轉換為數值，例如，int、float 等型態)，導致產生許多不必要的效能支出(Performance Overhead)。故如何改善目前實驗效能，縮短規則資訊獲得之計算時間乃是本研究未來重要的研究方向之一。

此外，本研究之目的主要在於：設計一泛用模型以探勘大型資料庫(例如，病歷資料庫、連鎖商場銷售資料庫等)，但本論文中尚未對其他大型資料庫做完整測試，驗證本模型之正確性，故如何以測試各種大型資料庫樣本的方式驗證本模型之可行性也是本研究未來的重點方向。

第6章 參考文獻

中文部分

1. 尤春惠(2004)，資料探勘在用藥安全上得應用：預測泛可黴素在腎衰竭病患上的用量適當性，國立中山大學資訊管理學系碩士在職專班畢業論文。
2. 台灣電子病歷交換基本格式，http://emr.doh.gov.tw/
3. 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。
4. 吳國禎(2000)，資料探索在醫學資料庫之應用，私立中原大學醫學工程學系碩士班畢業論文。
5. 李政宏(2004)，以知識庫為基礎之牙醫決策支援系統，國立屏東科技大學資訊管理系碩士班碩士學位論文。
6. 林文燦(2006)等，應用資料探勘技術提升急診醫學檢傷分類之一致性-以台灣某醫學中心急診醫學部為例，中華民國品質學會第 42 屆年會暨第 12 屆全國品質管理研討會，勤益技術學院。
7. 陳世源(2000)，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。
8. 陳垂呈(2006)，應用資料探勘技術發掘最適性之線上拍賣競標者，開南管理學院運籌研究集刊第一期，pp. 1-14。
9. 陳垂呈(2007)，利用資料探勘技術輔助疾病診斷是否異常，國立屏東教育大學資訊科學應用期刊第二期。
10. 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。
11. 黃美玲等(2006)，類神經網路輔助醫療診斷分類模式之建構，中華民國品質學會第 42 屆年會暨第 12 屆全國品質管理研討會，勤益技術學院。

西文部分

12. Agrawal, R. et al. (1993), Database mining: a performance perspective, IEEE Transaction on Knowledge and Data Engineering, 1993, pp. 914-925
13. Berthold, M. M. (2007), Intelligent Data Analysis: An Introduction (2ed), Springer, 2007
14. Brachman, R. and Anand, T. (1996), The Process of Knowledge Discovery in Database: A Human-Centered Approach. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif., AAAI Press, pp37-58
15. Chakravarti, Laha, and Roy. (1967), Handbook of Methods of Applied Statistics Volume I, John Wiley and Sons, 1967, pp. 392-394.
16. Chen, M. S., J. Han, and P. S. Yu. (1996), Data mining: An overview from a database perspective, IEEE Trans. on Knowledge and Data Engineering, 1996, pp. 866-883.
17. Clark, P. (1990), Machine learning: Techniques and recent developments, In A. R. Mirzai, editor, Artificial Intelligence: Concepts and Applications in Engineering, Chapman and Hall, 1990, pp. 65-93
18. Devore, J. L. (1999), Probability and Statistics for Engineering and the Sciences (5ed), Brooks/Cole Publishing Company, 1999.
19. Mathwave. 2008, EasyFit®, http://www.mathwave.com/
20. Milton, J. S., and J. C. Arnold. (2002), Introduction to probability and statistics: Principles and Applications for Engineering and the Computing Science (4ed), McGraw-Hill Science/Engineering/Math, 2002.

21. Reinschmidt, J.,Gottschalk, H., Kim, H., and Zwietering, D., (1999), Intelligent Miner for Data: Enhance Your Business Intelligence, IBM Internation Technical Support Organization.
22. Smyth, P. Goodman. (1992), An information theoretic approach to rule induction from databases, IEEE Transactions on Knowledge and Data Engineering, 1992, pp.301-306.
23. SPSS Inc. 2008, SPSS, http://www.spss.com/
24. Thyroid Disease Database. 1987, http://ftp.ics.uci.edu/pub/machine-learning-databases/thyroid-disease/
25. U.M. Fayyad, G., Piatetsky-Shapiro, and P., Smyth (1996), From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining, U.M. Fayyad et al., eds., AAAI/MIT Press, Menlo Park, Calif., pp1-34
26. William, J. Frawley, Gregory Piatetsky-Shapiro, and Christopher, J. Matheus (1992), Knowledge Discovery in Database – An Overview, AI Magazine, 1992, pp.57-70.

附錄(實驗資料前 100 筆)

age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery
29	F	f	f	f	f	f	f
29	F	f	f	f	f	f	f
41	F	f	f	f	f	f	f
36	F	f	f	f	f	f	f
32	F	f	f	f	f	f	f
60	F	f	f	f	f	f	f
77	F	f	f	f	f	f	f
28	F	f	f	f	f	f	f
28	F	f	f	f	f	f	f
28	F	f	f	f	f	f	f
54	F	f	f	f	f	f	f
42	F	f	f	f	f	f	f
51	M	t	f	f	f	f	f
51	F	f	f	f	f	f	f
37	F	f	f	f	f	f	f
16	M	f	f	f	f	f	f
54	F	f	f	f	f	f	f
43	M	f	f	f	f	f	f
63	F	t	f	f	t	f	f
36	F	f	f	f	f	f	f
40	F	f	t	f	f	f	f
40	F	f	f	f	f	f	f
40	F	f	f	f	f	f	f
77	F	f	f	f	f	f	f
77	?	f	f	f	f	f	f
77	F	f	f	f	f	f	f
77	F	f	f	f	f	f	f
51	F	f	f	f	f	f	f
75	F	f	f	f	f	f	f
56	M	f	f	f	f	f	f
42	M	f	f	f	f	f	f

85	M	f	f	f	f	f	f
41	M	f	f	f	f	f	f
71	F	t	f	f	f	f	f
67	M	f	f	f	f	f	f
67	M	f	f	f	f	f	f
51	F	f	f	f	f	f	f
51	F	f	f	f	f	f	f
28	F	f	f	f	f	f	f
55	F	t	f	f	f	f	f
32	F	t	f	f	f	f	f
61	F	t	f	f	f	f	f
46	M	f	f	f	f	f	f
41	F	f	f	f	f	f	f
44	F	t	f	f	f	f	f
51	F	f	f	f	t	f	f
29	M	f	f	f	f	f	f
41	F	f	f	f	f	f	f
82	M	f	f	f	f	f	f
64	F	t	f	f	f	f	f
70	F	f	f	f	f	f	f
54	F	f	t	f	f	f	f
33	F	f	f	f	f	f	f
59	F	t	f	f	f	f	f
44	M	f	f	f	f	f	f
44	F	f	f	f	f	f	f
44	M	f	f	f	f	f	f
53	F	f	f	f	f	f	f
41	F	f	f	f	f	f	f
52	F	f	f	f	f	f	f
59	M	f	f	f	f	f	f
49	F	f	f	f	f	f	f
44	F	f	f	f	t	f	f
44	M	f	f	f	f	f	f
46	F	f	f	f	f	f	f
35	F	f	f	f	f	f	f
56	F	f	f	f	f	f	f

54	F	f	f	f	f	f	f
54	F	f	f	f	f	f	f
54	M	f	f	f	f	f	f
54	F	f	f	f	f	f	f
61	M	f	f	f	t	f	f
40	F	t	f	f	f	f	f
48	M	t	f	f	f	f	f
64	F	f	f	f	f	f	f
27	F	f	f	f	f	f	f
69	F	f	f	f	f	f	f
69	F	f	f	f	f	f	f
60	F	t	f	f	f	f	f
27	F	f	f	f	f	f	f
27	F	f	f	f	f	f	f
51	F	f	f	f	f	f	f
76	M	f	f	f	f	f	f
52	F	t	f	f	f	f	f
73	F	t	f	f	f	f	f
68	M	t	f	f	f	f	f
66	M	t	f	f	f	f	f
66	F	f	f	f	f	f	f
30	M	f	f	f	f	f	f
30	F	f	f	f	f	f	f
30	M	t	f	f	f	f	f
28	M	f	f	f	f	f	f
88	F	f	f	f	f	f	f
38	F	f	f	f	f	f	f
66	F	t	f	f	f	f	f
58	F	f	f	f	f	f	f
21	F	f	f	f	f	f	f
56	M	t	f	f	f	f	f
60	F	t	f	f	f	f	f
38	F	f	f	f	f	f	f
I131 treatment	query hypothyroid	query hyperthyroid	lithium	goitre	tumor	hypopituitary	psych
f	t	f	f	f	f	f	f
f	f	f	f	f	f	f	f

f	f	t	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	t	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	t	t	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	t	f	f	f	f	f	f

f	f	f	f	f	f	f	f
f	t	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	t	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	t	f	f	f	f	f	f
f	f	t	f	f	f	f	f
f	f	t	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	t	f	f	f	f	f
t	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	t	f	f	f	f	f
f	f	t	f	f	f	f	f
f	f	f	f	f	f	f	f
f	t	f	f	f	f	f	f
f	f	t	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f

f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	t
f	t	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	t	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
t	f	f	f	f	f	f	f
f	f	f	f	f	f	f	f
f	f	f	f	t	f	f	f
f	t	f	f	f	f	f	f
f	f	f	f	f	f	f	f

TSH	T3	TT4	T4U	FTI	TBG	diagnoses
0.3	?	?	?	?	?	-
1.6	1.9	128	?	?	?	-
?	?	?	?	?	11	-
?	?	?	?	?	26	-
?	?	?	?	?	36	S
?	?	?	?	?	26	-
?	?	?	?	?	21	-
0.7	2.6	116	?	?	?	-
1.2	1.8	76	?	?	?	-

1.9	1.7	83	?	?	?	-
1.9	2.3	133	?	?	?	-
1	1.8	105	?	?	?	-
0.5	?	?	?	?	?	-
0.7	2.4	116	?	?	?	-
?	2.9	?	?	?	?	-
2.6	?	?	?	?	?	-
1	2	122	?	?	?	-
?	2.1	116	?	?	?	-
68	?	48	1.02	47	?	F
1.5	2.4	90	1.06	85	?	-
?	?	79	0.94	84	?	-
1.2	2.3	104	1.08	96	?	-
5.9	2.1	88	0.84	105	?	-
0.05	2.4	107	1.13	95	?	-
4	2	126	?	?	?	-
0.4	?	113	1.07	106	?	-
0.8	2.4	150	?	?	?	-
0.05	2.1	93	0.87	106	?	-
0.05	1.6	157	0.89	176	?	AK
0.2	1.6	80	0.62	129	?	-
3	2	91	0.91	100	?	-
0.05	0.1	47	0.68	69	?	-
0.05	1.6	39	1	39	?	R
0.05	?	126	1.38	91	?	I
0.2	?	71	0.79	90	?	-
0.05	2	88	0.95	93	?	-
1.2	2.1	104	1.57	66	?	-
0.3	?	111	0.92	121	?	-
?	2.1	86	0.92	93	?	-
9.599999	2.4	136	1.48	92	?	M
0.05	2	163	0.94	173	?	N
1.9	2.1	118	1.1	93	?	-
?	1.4	82	0.7	117	?	-
140	?	33	1.07	31	?	F
0.6	1.8	134	1.13	113	?	-

1.5	1.2	60	0.89	67	?	-
1.7	2.4	102	1.01	101	?	-
0.05	1.5	97	1.05	92	?	-
0.4	1.3	115	0.91	126	?	-
2.5	1.6	86	0.96	90	?	-
0.05	2.6	132	1.07	123	?	-
?	?	116	0.78	149	?	-
0.4	2	114	1.01	113	?	-
6.8	2.5	94	1.4	68	?	M
?	?	90	1.05	86	?	-
0.4	1.8	94	?	?	?	-
1	2.7	118	?	?	?	-
0.7	1.3	145	?	?	?	-
0.2	1.5	106	?	?	?	-
0.2	1.9	87	0.66	132	?	-
2.1	1.8	104	0.86	121	?	-
1.2	2.2	80	0.76	105	?	-
1.1	1.5	105	0.9	117	?	-
0.8	2.5	120	?	?	?	-
?	?	121	1	121	?	-
?	2.2	?	?	?	?	-
0.05	2.5	152	1.16	131	?	-
0.05	1.8	79	0.68	116	?	-
0.7	2.1	98	?	?	?	-
0.8	2.4	109	1.12	97	?	-
0.05	1.6	114	0.92	124	?	-
9.799999	1.2	114	0.84	136	?	G
0.05	?	139	0.98	142	?	-
3.7	2.8	106	1	106	?	-
0.4	2	104	1.04	100	?	-
0.05	2.8	131	1.26	104	?	-
0.1	3.2	112	?	?	?	-
0.35	2.4	64	?	?	?	-
0.35	1.8	109	0.83	131	?	-
90	0.4	7.5	0.94	7.5	?	F
1.3	2.2	104	0.97	107	?	-

1.2	1.7	68	0.93	73	?	-
0.5	1.6	97	0.88	110	?	-
0.05	?	133	1.02	130	?	-
?	?	99	0.95	104	?	-
2.7	1.3	64	0.7	97	?	-
1.6	2.1	92	1.04	88	?	-
2.9	1.6	115	0.9	128	?	-
1.7	1.9	120	0.98	122	?	-
0.25	1.8	95	0.87	102	?	-
0.6	1.8	100	?	?	?	-
0.2	2.4	106	?	?	?	-
0.2	0.4	98	0.73	134	?	K
0.2	2.4	184	1.13	163	?	R
2.1	?	81	1.29	63	?	I
5.8	1.7	86	0.91	95	?	-
0.2	?	95	1.04	91	?	-
?	?	87	1.3	67	?	I
0.35	1.8	109	0.83	131	?	-
?	?	86	1	86	?	-