具隱私保護之醫療健檢關聯規則探勘

薛夙珍
朝陽科技大學資訊管理系 助理教授
schsueh@cyut.edu.tw

徐韻雯
朝陽科技大學資訊管理所 研究生
s9414607@cyut.edu.tw

摘要

資料探勘的技術已應用於許多的領域中，但資料探勘的過程中卻也增加了侵犯資料隱私的威脅。本論文針對醫療健檢資料作研究，當醫療單位缺乏資料探勘軟體資源或缺少具備資料探勘的分析人員，而需藉助委外單位分析資料時，醫療單位必須對醫療資料做適當的隱私保護(Privacy Preserving)，以保護敏感性資料不被揭露。在方法設計上，著重於如何隱藏敏感性項目集，使釋出的資料庫無法推論出原始的敏感性項目集。本研究基於阻擋技術提出一種未知值取代方式，此方法利用常態分佈的特性均勻取代敏感項目集，以達到隱藏的目的。此外，藉由多餘型樣(Artifact Patterns)及型樣遺失(Misses Cost)二項指標，來衡量提出的隱私保護效能。實驗結果證實，本研究提出的方法既不會產生多餘型樣外，亦能減少型樣遺失。

關鍵詞：資料探勘、關聯規則探勘、隱私保護、阻擋技術、健康檢查資料。

Abstract

Data mining techniques have been exercised in many applications. Nevertheless, the mining process also may cause threats to the invasion of privacy. Health examination records contain private information and should be protected during the mining process. For some small or medium scaled medical institutes, the mining task can be outsourced to achieve the goal of analysis without having to installing the high-cost mining facilities. In this case, the privacy data for outsourcing should be adequately provided to prevent sensitive data from being exposed in advance. In this paper, we focus on the problem of outsourcing association rule mining of healthcare data while preserving the privacy of sensitive data. The original private data is assured not to be inferred from the released mining database. The proposed approach is to transform the original data into a masked form based on the replacement of question-marks with the consideration of normal distributions of sensitive itemsets. In addition, two performance metrics are performed to evaluate the efficiency of our approach. The experimental results show that our method minimizes the cost of missing patterns while producing no artificial patterns for best privacy preserving of association rule mining.

Keywords: Data Mining, Association Rule, Privacy Preserving, Blocking Technique, Health Examination.

1. 前言

隨著資訊科技的快速發展，醫療資訊都已經以電子化儲存，在醫療資訊系統的廣泛使用下，資料庫中的資料量亦隨之快速的累積，蘊藏在其中的資訊也相對的提高。由於醫院的資料庫累積了大量的就醫資料，而這些資料中可能隱藏許多的資訊是尚未發覺的知識，藉由資料探勘的技術可將資料庫中有用的知識挖掘出來，以找出具參考意義的醫學知識。其中，最常使用資料探勘技術之一為關聯規則探勘(Association Rule Mining)，關聯規則探勘主要的概念是找出項目之間的關聯性，目前已有許多醫療資訊利用關聯規則應用，如探討疾病與症狀之間的關聯性、疾病與疾病之間的關聯性等。

探勘軟體的開發以及分析人員培訓需要耗費相當大的時間及成本，此非一般醫療院所皆能負擔。此時，當缺乏資料探勘軟體資源或缺少具備資料探勘的分析人員，院方亦可選擇將內部的醫療資料，委外給外界機構，藉助更專業的能力與資源來協助醫院進行資料探勘分析[4]。其中，在醫療資料最具有豐富的醫療資訊即是健檢資料，健檢資料包括了個人基本資料及檢驗資料，個人基本資料包括了：姓名、性別、身份證字號、地址、電話、病歷史、家族史等，檢驗資料包括了：血液檢查、尿液檢查、生化檢查、胸部X光、B型肝炎等檢查資料。然而，醫療健檢資料藉由委外單位探勘時牽涉隱私的議題，因此，必須對醫療資料做適

當的隱私保護的技術，才能避免個人隱私被揭露，再者，將醫療資料交由委外的資訊公司做關聯規則探勘前，必須要藉由關聯規則隱私保護(Privacy Preserving Association Rule)的方法，來保護敏感資料及資訊不被揭露。

在醫療資料中使用隱私保護之關聯規則，著重在於敏感規則的保護以及避免錯誤規則的產生，若醫療人員參考到不正確的資訊，可能會危害到病人的生命。Saygin等學者認為在某些情況下，給予使用者一些錯誤或不存在的規則可能會造成無法彌補的過錯，Sayging等學者提出新的隱私保護方法，利用未知值“？” (unknown value) 來取代錯誤的值 “0”或“1” (false value)[16][17]。

然而，Sayging 等學者在[16][17]論文中未提及，如何適當的取代未知值，因此，本篇論文延伸阻擋技術的特性，提出新的取代未知值方法。本研究主要植基於醫療的環境下，將原始的醫療資料庫透過均勻取代阻擋技術使原始資料庫轉換為隱私保護資料庫，藉由委外單位做關聯規則探勘，最後探勘出結果調整原始關聯規則。

本篇論文的架構如下：第壹節緒論，介紹本研究背景與動機、研究的目的。第貳節是相關研究，介紹在資料探勘的隱私保護相關議題一些研究學者的方法。第參節研究方法設計，介紹本篇研究在隱藏程序。第肆節實驗結果，透過亂數產生器，產生常態分配資料，將所提出的隱藏策略應用於具常態分配的醫療資料，以證明方法提出方可行性。第伍節結論及未來研究，為本篇研究做個總結並提及本研究的貢獻和未來的研究方向。

2. 相關研究

2.1 醫療資料探勘

近年來，資料探勘技術之關聯規則探勘應用最廣[8]，利用關聯規則探勘技術在醫療用來挖掘潛藏的醫療知識相關的應用，已受到醫學界及學術界的重視。吳國禎學者[2]藉由大量的資料累積建立醫療判斷系統比傳統的診斷更為精確，相繼也有許多學者探討資料探勘之關聯規則在醫療資訊之應用，黃昱銘學者[5]提出利用醫療資料中探勘症狀與疾病之間的關聯規則，建立診斷系統提供協助醫生就診，且醫院能藉此提供網路預約診療服務。健保局提供研究使用的全民健康保險研究資料庫[1]由中央健康保險局建立一個以學術研究為目的之資訊資料庫，提供給學術單位及非營利機構之學者專家進行醫藥衛生相關研究。李炳雄學者[3]利用全民健康保險研究資料庫，資料探勘之關聯規則技術來偵測醫學資料庫中的異常病歷資料與探究疾病與疾病、疾病與藥品間之關聯性，找出最適合某項病症之藥品或醫令組合，提供醫生診斷時的參考依據；楊燕珠及陳曉芬學者[7]利用全民健康保險研究資料庫，擷取了相關的疾病分類代碼之骨質疏鬆症患者為研究對象，目的為分析骨質疏鬆症患者人口學特性與就醫習慣、骨質疏鬆症與其他疾病的關聯性。盧怡玲學者[6]利用某地區的真實健康檢查資料探勘疾病之間的關聯性。有鑑於此，資料探勘技術應用在醫療資料上，愈受重視及應用，其能降低醫療成本花費，更能提高醫療知識的品質以幫助決策的制定上。

2.2 關聯規則隱私保護

隱私保護之關聯規則探勘研究分類[4]，主要可分為兩個大類探討。第一大類為修改公開內容，主要探討防止原始資料中敏感性資料或藉由資料探勘後得到資訊，所設計的方法為在公開內容前，先行針對敏感性資料保護，其方法有新增、修改、刪除敏感項目。第二大類為多單位合作[13][18]，根據交易被垂直或水平分割在不同地點。

在修改公開內容的方法可細分為二種，第一種探討敏感性規則隱藏(Hiding Sensitive Rule)，指在已知敏感性規則下，事先針對敏感性規則隱藏；第二種探討敏感性項目集隱藏(Hiding Sensitive Itemsets)，指在未知敏感性的規則下，針對某種特定的敏感性項目集隱藏。以下分別介紹大致隱私保護的方法：

Oliveria 等學者[14]提出有關敏感性項目集保護的方法，利用降低敏感性項目集的支持度達到隱藏目的。若交易中包含某個敏感項目集，則表示此筆交易中會出現這個敏感項目集，評估各敏感項目集中的犧牲項目，且根據公開程度計算修改的交易筆數，讓交易不再出現此敏感項目集，避免被使用者挖掘出來。雖然可解決敏感項目集洩露的問題，但卻會產生型樣遺失的問題。

Dasseni 等學者[12]根據關聯式規則的支持度(Support)及信心度(Confidence)門檻值提出三種隱藏敏感規則的方法。第一種作法是從支持度著手，藉由減少前項的敏感項目其低於設定之最小支持度(Minimal Support Threshold)，而第二及第三種作法是從信心度著手，分別藉由增加規則的前項支持度及減少規則的後項支持度來降低敏感規則的信心度使其

低於最小於最小信心度的門檻值(Minimal confidence Threshold)。雖然可解決敏感規則洩露的問題，但卻會產生型樣遺失和多餘型樣等問題。

Agrawal 等學者[9]提出重建技術的概念，重建技術在隱私防護之關聯規則探勘中，主要目的為保護資料傳送方之原始資料內容，同時，讓資料接收方挖掘出隱藏於原始資料內容中有用的型樣。

Atallah 等學者[10]已證實隱私保護的最佳化問題是一個NP-hard問題。因此，目前學者所提方法中均存在部分潛在問題，例如隱藏不完全、規則遺失、多餘型樣等，主要因為資料探勘與隱私保護之二者之間的平衡取捨不易衡量。

本研究所採用的是Sayging 等學者[16][17]所提出的阻擋技術，方法流程與Dasseni等學者所提方法相似，差異為阻擋技術在修改程序上使用未知值“？”來取代，原先的高頻項目集出現以“1”表示改為不出現以“0”表示那麼相對關聯規則的支持度值將會降低，將使的規則不再具有敏感性，然而，可能會造成錯誤的規則產生，因此，在方法設計上他們介紹一個新的文字符號“？”，代表沒有任何的資訊知道這一個項目集是否出現。

3.研究方法

3.1 方法架構

本研究方法之架構主要有三個程序，分別為轉換階段(前處理)、探勘階段及調整階段(後處理)，如圖 1 所示。院方將資料傳送前先行透過轉換程序對原始交易資料庫進行轉換後，將其傳送至分析單位。分析單位將轉換後醫療資料庫進行關聯規則探勘，探勘完成後即傳回資料傳送方。最後，院方人員藉由調整程序對探勘結果進行轉換，即可得知相近似原始之關聯規則。

(1) 轉換程序

由於醫療資料庫含有敏感性資料內容以及隱藏在其中的敏感性資訊，故醫療單位在將原始資料庫交於委外單位探勘前，先行透過轉換階段(前處理程序)對敏感性資料作保護，再將轉換後的資料庫傳送委外單位進行探勘。程序中藉助於轉換機制之使用將原始資料庫轉換後資料庫，本研究著重在於轉換機制設計。

(2) 探勘程序

分析單位使用關聯規則探勘方法 Apriori 演算法，針對醫療單位傳送的資料庫做關聯探勘，探勘的過程中不斷調整適當的支持度及信心度，最後，將探勘後結果產生報表回傳醫院單位。

(3) 調整程序

分析單位在轉換後資料庫作關聯規則探勘，所探勘出來的結果已受保護。當分析單位傳回探勘結果，院方透過調整機制(後處理程序)，即可得知正確之關聯規則。

3.2 方法介紹

方法設計主要針對醫療型資料庫作設計，假設原始醫療資料庫由 m 筆醫療紀錄所組成，每一筆醫療紀錄內容包括醫療編號(MID)及醫療資料包括了個人基本資料及健檢項目 n 個欄位(a1~an)。其中，每筆醫療資料具有三種欄位特性：

(a) 二位元資料型態(如性別、婚姻狀況等)。
(b) 固定長度型態(如年齡、身高、體重等)。
(c) 非固定長度型態(如收縮壓、舒張壓、白血球、血小板檢驗值等)。

三種欄位可依照區間的範圍決定不同的切割區間，以下分別介紹不同欄位的區間切割方式，二位元資料型態的區間介於 0/1 之間(男或是女、已婚或是未婚)；而固定長度型態的區間切割可依照資料屬性切割區間範圍，年齡可分為四大區間：兒童、青年、中年、老年等；而非固定長度型態的區間以檢驗值可細分為偏高、略偏高、正常、略偏低、偏低等。

以下分別介紹轉換程序、探勘程序、調整程序之運作步驟如下：

(1) 轉換程序：

步驟 1：院方人員設定量化區間，並將每個欄位對應至所屬的量化值。
區間切割方式是非固定長度，依照醫療檢驗值分佈的特性，院方人員可以依照醫療知識自行設定區間範圍。

步驟 2：院方人員選擇欲隱藏的敏感項目。

步驟 3：設定敏感項目取代未知值“？”比例均勻取代。

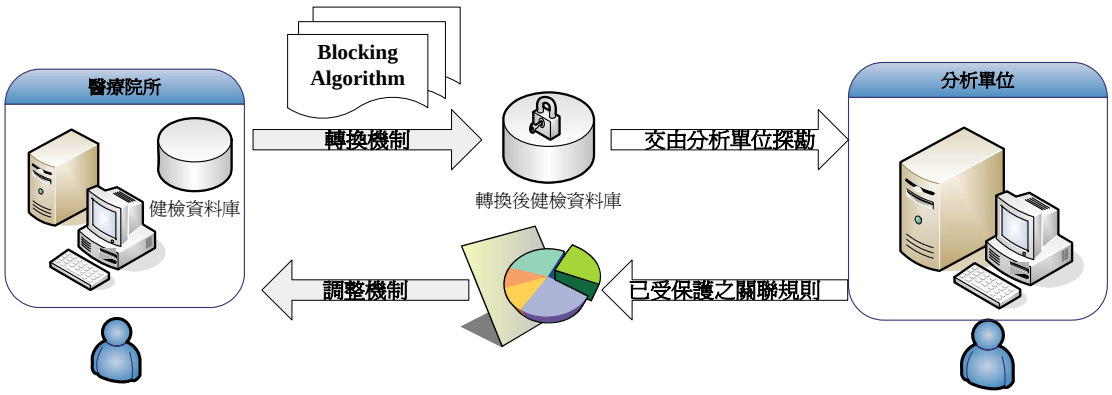

圖 1：以阻擋技術為基礎之隱私保護程序概圖

均勻取代的目的，使用阻擋技術取代的過程，為了要保持原始資料的常態分佈，因此，在取代未知值時針對每個敏感欄位之屬性值的比例來取代。

均勻取代的公式如下：

$$\text{公式}\quad n_i = n \times \frac{P_i}{\sum_1^K P_i},\ \text{i=1,2,…,k} \qquad (1)$$

其中，每個變數代表的意義如下：

n_i：計算每個屬性值需要取代多少個未知值。

n：取代未知值的數值。

P_i：屬性值的比例。

步驟 4：計算後的屬性值使用隨機(Random)的方式取代。

步驟 5：產生轉換後的隱私保護資料庫。

步驟 6：將隱私保護資料庫交由分析單位作關聯規則探勘。

(2) 探勘程序：

假設分析單位使用Apriori演算法作關聯規則探勘[8]。

步驟 1：設定最小支持度及最小信心度。(由分析單位設定)

步驟 2：依據自訂最小支持度，濾除候選項目集支持度小於最小支持度之關聯規則，重覆步驟 2 直到沒有新的候選項目集產生。

步驟 3：依據使用者自訂最小信心度，濾除關聯規則信心度小於最小信心度的規則，重覆步驟 3 直到沒有新的關聯規則產生為止。

步驟 4：分析單位不斷測試最小支持度及最小信心度的適當值。

步驟 5：分析單位將產生的關聯規則報表回傳給院方單位。

(3) 調整程序：

院方人員在解讀規則時真正的支持度及信心度比探勘規則時支持度及信心度所隱藏含意為接近實際值但是比探勘出來的數值更高，也就是分析單位探勘後的關聯規則支持度及信心度代表意義更大，原因為我們在轉換程序時針對敏感性欄位做阻擋技術取代原始資料，使得探勘後的支持度降低。

3.3 例子說明

本研究以列舉十筆健檢資料為例(如表 1 所示)，進行阻擋技術隱私保護關聯規則探勘。以下將分別說明：轉換程序、探勘程序、調整程序之運作方法：

(1) 轉換程序：

步驟 1：將每個檢查項目對應至所屬性值(如表 2 所示)。以第一筆醫療資料為例，其年齡、性別、收縮壓、舒張壓、血紅素，各別的屬性值為「女」、「兒童」、「略偏低」、「正常」、「略偏低」，經量化關聯表(如表 3 所示)對應的值為「2」、「1」、「4」、「3」、「4」。

表 1：原始健檢資料庫

健檢號碼	性別	年齡	收縮壓	舒張壓	血紅素
A0001	Female	5	86	60	8
A0002	Male	8	120	72	10
A0003	Male	12	148	80	12
A0004	Male	18	150	86	11
A0005	Female	25	78	90	14
A0006	Female	28	128	86	6
A0007	Male	32	167	90	18
A0008	Male	35	162	88	20
A0009	Female	56	90	68	15
A0010	Male	60	130	78	16

表 2：量化對應表

性別對應表

性別	屬性 / 屬性值
Male	性別(男性) / 1
Female	性別(女性) / 2

年齡對應表

範圍	屬性 / 屬性值
11 歲以下	兒童 / 1
12-23 歲	青年 / 2
23-30 歲	成年 / 3
31-50 歲	中年 / 4
51 歲以上	老年 / 5

收縮壓對應表

範圍	屬性 / 屬性值
151 以上	偏高 / 1
141～150	略偏高 / 2
90～140	正常 / 3
80～89	略偏低 / 4
79 以下	偏低 / 5

舒張壓對應表

範圍	屬性 / 屬性值
101 以上	偏高 / 1
91～100	略偏高 / 2
60～90	正常 / 3
80～89	略偏低 / 4
75 以下	偏低 / 5

血紅素對應表(性別『男性』)

範圍	屬性 / 屬性值
21 以上	偏高 / 1
18～20	略偏高 / 2
12～17	正常 / 3
10～11	略偏低 / 4
9 以下	偏低 / 5

血紅素對應表(性別『女性』)

範圍	屬性 / 屬性值
19 以上	偏高 / 1
16～18	略偏高 / 2
10～15	正常 / 3
8～10	略偏低 / 4
7 以下	偏低 / 5

表 3：量化資料庫

健檢號碼	性別	年齡	收縮壓	舒張壓	血紅素
A0001	2	1	4	3	4
A0002	1	1	3	3	4
A0003	1	2	2	4	3
A0004	1	2	2	4	4
A0005	2	3	5	3	3
A0006	2	3	3	4	2
A0007	1	4	1	3	2
A0008	1	4	1	4	2
A0009	2	5	3	5	3
A0010	1	5	3	5	1

步驟 2：院方人員選擇欲隱藏的敏感項目。
院方人員選擇檢驗資料的『舒張壓』、『血紅素』做為欲隱藏敏感項目欄位。

步驟 3：設定敏感項目取代未知值(Unknown)“？”的比例均勻取代。

公式 $n_i = n \times \frac{P_i}{\sum_1^K P_i}$,i=1,2,…,k (1)

其中，每個變數代表的意義如下：

n_i：計算每個屬性值需要取代多少個未知值。

n：取代未知值的數值。

P_i：每個屬性值的比例。

(計算出結果四捨五入)

(1)收縮壓設定取代 40%的未知值 (量化區間)

(a)「偏高」取代未知值的比例 4*(1/10) = 0.
(代表在收縮壓偏高的屬性取代 0 個“？”)

(b)「略偏高」取代未知值的比例 4*(1/10) = 0

(c)「正常」 取代未知值的比例 4*(4/10) = 2

(d)「略偏低」取代未知值的比例 4*(3/10) = 1

(e)「偏低」 取代未知值的比例 4*(2/10) = 1

(2)血紅素設定取代 40%的未知值

(a)「偏高」 取代未知值的比例 4*(1/10) = 0

(b)「略偏高」取代未知值的比例 4*(3/10) = 1

(c)「正常」 取代未知值的比例 4*(3/10) = 1

(d)「略偏低」取代未知值的比例 4*(3/10) = 1

(e)「偏低」 取代未知值的比例 4*(0/10) = 0

步驟 4：依照循環(Round Robin)、隨機(Random)方式取代，如表 4 所示。
(循環取代設定 2)

步驟 5：將原始資料庫轉換成隱私保護資料庫，如表 5 所示。

步驟 6：將隱私保護資料庫交由分析單位做關聯規則探勘及設定最小支持度及最小信心度。

表 4：阻擋技術均勻取代

健檢號碼	性別	年齡	收縮壓	舒張壓	血紅素
A0001	2	1	4	2	4
A0002	1	1	3	2	**? (4)**
A0003	1	2	2	2	3
A0004	1	2	**? (2)**	2	4
A0005	2	3	5	2	**? (3)**
A0006	2	3	**? (3)**	2	2
A0007	1	4	1	2	**? (2)**
A0008	1	4	**? (1)**	2	2
A0009	2	5	3	2	3
A0010	1	5	**? (3)**	2	1

表 5：隱私保護資料庫

健檢號碼	性別	年齡	收縮壓	舒張壓	血紅素
A0001	2	1	4	2	4
A0002	1	1	3	2	?
A0003	1	2	2	2	3
A0004	1	2	?	2	4
A0005	2	3	5	2	?
A0006	2	3	?	2	2
A0007	1	4	1	2	?
A0008	1	4	?	2	2
A0009	2	5	3	2	3
A0010	1	5	?	2	1

(2) 探勘程序：

分析單位關聯探勘結果：

當最小支持度 = 10%及最小信心度= 80%，有 10 條關聯規則等。

[support, confidence]

Rule 1:收縮壓(略偏高)→性別(男性)[10%,100]

Rule 2:性別(女性)→收縮壓(略偏低) [10%,100]

Rule 3:收縮壓(略偏低)→性別(女性) [10%,100]

Rule 4:性別(女性)→血紅素(略偏低) [10%,100]

Rule 5:收縮壓(略偏高)→血紅素(正常) [10%,100]

Rule 6:收縮壓(略偏低)→血紅素(略偏低)[10%,100]

Rule 7:性別(男性),收縮壓(略偏高)→血紅素(正常) [10%,100]

Rule 8:性別(男性),血紅素(正常)→收縮壓(略偏高) [10%,100]

Rule 9:收縮壓(略偏高),血紅素(正常)→性別(男性) [10%,100]

Rule 10:收縮壓(略偏高)→性別(男性),血紅素(正常) [10%,100]

等….

(3) 調整程序：

當院方人員在解讀探勘後的關聯規則結果，假設以挑選第一條關聯規則：收縮壓(略偏高)→性別(男性)之支持度為 10、信心度為 100，經過隱私保護後關聯規則真正代表意義為收縮壓(略偏高)→性別(男性)支持度超過 10 以上，也就是當分析單位探勘後的關聯規則其支持度及信心度所隱藏含意接近實際值但是比探勘出來的數值更高。

4.實驗分析

4.1 實驗環境

為了測試所提出的隱藏策略，實驗的機器是含 AMD Athlon 64 X2 3800+ 2GHz 的處理器、記憶體為 1GB DDRAM、使用的作業系統為 Windows XP、使用的程式開發語言為 C。

4.2 實驗設計

在建立測試資料方面採用亂數產生器之常態分配產生測試資料，產生了資料集 6 個項目集，其中包括了二位元資料型態(性別、婚姻狀況)、固定長度型態(年齡、身高)及非固定長度(收縮壓、血紅素)，紀錄總筆數為 10K 筆。

4.3 評估分析

本方法採用Rizvi等人提出的σ^+與σ^-指標[15]，分別評估方法對多餘項目集與遺失項目集問題之影響程度，基於關聯規則係由高頻項目集所推導而知，故實驗以轉換後交易資料庫的高頻項目集所構成集合(R)與原始交易資料庫的高頻項目集形成之集合(F)差異，作為衡量潛在問題的影響程度，詳細公式如(2)與(3)。

$$\sigma^+ = \frac{|R-F|}{|F|} \times 100 \qquad (2)$$

$$\sigma^- = \frac{|F-R|}{|F|} \times 100 \qquad (3)$$

其中，

R：轉換後交易資料庫中高頻項目集之集合；

F：原始交易資料庫中高頻項目集之集合。

4.4 實驗結果

實驗中首先測試不同的取代比例利用此演算法的所需 CPU 的執行時間如圖 2 所示，隨著未知值取代遞增，CPU 的執行時間亦逐漸上升。

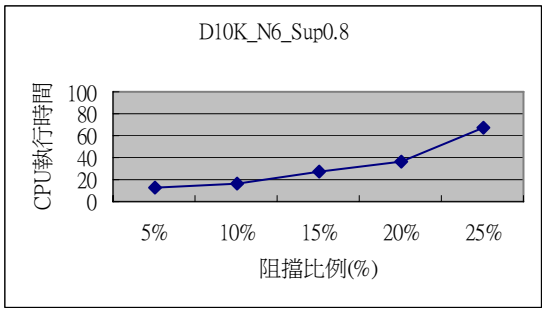

圖 2 CPU 執行效能結果

接著測試不同隱藏策略對資料庫所造成的型樣遺失結果如圖 3，型樣遺失指經由資料的調整而使原始的項目集消失的數量，由圖所示，隨著支持度之下降，項目集遺失個數亦逐漸增加。原因為當支持度門檻設定愈低，條件較易符合下即會挖掘較多之項目集，未知值取代個數為 200 個。

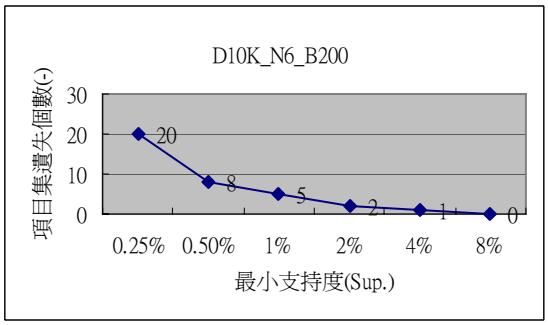

圖 3 支持度對項目遺失結果

最後針對在不同支持度下所產生的多餘型樣結果如圖 4，多餘型樣，指經由隱私保護後使原始的項目集增加的數量，由圖所示，隨著支持度之增加，亦不會增加多餘型樣項目個數。原因為本論文提出的方法不針對原始資料庫做新增及修改內容，亦不會產生多餘型樣，可避免錯誤規則產生。

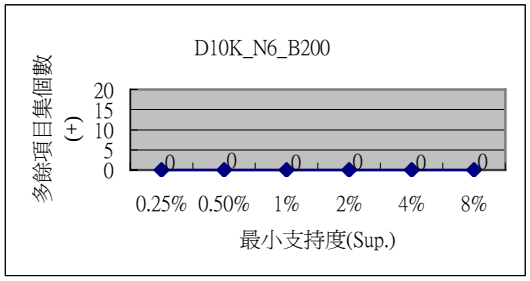

圖 4 支持度對多餘項目結果

5.結論及未來研究

本文針對醫療資料之關聯規則隱私保護，在保護醫療資料下進行關聯規則探勘。為避免錯誤規則產生，使用均勻取代未知值方法來取代真實值，能保有醫療資料特性常態分配，並且本研究分別評估隱私保護之型樣遺失、多餘型樣，結果證實本研究不會產生多餘型樣以及型樣遺失能降至最少。

本研究的貢獻為：

(一) 提出一個適用於醫療資料型態的關聯規則探勘隱私保護方法。

(二) 依照醫療檢驗值分佈特性，提出非固定長度區間的切割方式。

由於關聯規則只是資料探勘其中之一技術，還有其他的技術如分類、群聚、循序特徵、時間序列等，同樣也會面臨相同的隱私保護問題，因此，如何將隱私保護的議題應用到這些技術上，是未來可以研究的領域。此外，像醫院之間的健檢資料彙整包括水平整合或垂直整合等使用資料探勘，也必須考慮多單位的隱私保護資料探勘，是未來可以延伸研究的方向。

參考文獻

[1] 全民健康保險研究資料庫，http://www.nhri.org.tw/nhird/，2003。

[2] 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，2000。

[3] 李炳雄，***資料探勘在醫學資料庫之應用－以全民健康保險學術研究資料庫為例***，國立台北大學資訊管理研究所碩士論文，2003。

[4] 梁嘉鴻，***具隱私防護之關聯規則探勘研究***，朝陽科技大學資訊管理系碩士論文，2004。

[5] 黃昱銘，***有效率地探勘疾病和病症之複合項關聯規則***，南台科技大學資訊管理所碩士論文，2004。

[6] 盧怡伶，***應用資料探勘技術分析健康檢查資料***，樹德科技大學資訊管理所碩士論文，2004。

[7] 楊燕珠、陳曉芬，***應用資料挖礦技術於全民健康保險研究資料庫-以骨質疏鬆症為例***，2004 電子商務與數位生活研討會論文集，2004。

[8] Agrawal, R., Imielinski, T., and Swami, A., "Mining Association Rules between Sets of Items in Large Databases," ***Proceedings of the ACM SIGMOD International Conference on Information and Management of Data***, Vol. 22 No. 2, pp. 207-216, 1993.

[9] Agrawal, R., and Srikant, R., "Privacy-Preserving Data Mining," ***Proceedings of the ACM SIGMOD International Conference on Management of Data***, pp. 439-450, 2000.

[10] Atallah, M. J., Bertino, E., Elmagarmid, A. K., Ibrahim, M. and Verykios, V. S., "Disclosure Limitation of Sensitive Rules," ***Proceedings of the IEEE Workshop on Knowledge and Data Engineering Exchange***, pp. 45-52, 1999.

[11] Clifton, C. and Marks, D., "Security and Privacy Implications of Data Mining," ***Proceedings of the ACM SIGMOD Workshop on Data Mining and Knowledge Discovery***, pp. 15-19, 1996.

[12] Dasseni, E., Verykios, V. S., Elmagarmid, A. K., and Bertino, E., "Hiding Association Rules by Using Confidence and Support," ***Proceedings of the 4th International Workshop on Information Hiding***, pp. 369-383, 2001.

[13] Kantarcioglou, M. and Clifton, C., "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," ***Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Engineering (RIDE)***, pp. 24-31, 2002.

[14] Oliveira, S. R. M. and Zaiane O. R., "Privacy Preserving Frequent Itemset Mining," ***Proceedings of IEEE International Conference on Data Mining (ICDM)***, pp. 43-54, 2002.

[15] Rizvi, S. and Haritsa, J., "Maintaining Data Privacy in Association Rule Mining," ***Proceedings of the 28th International Conference on Very Large Data Bases (VLDB)***, pp. 682-693, 2002.

[16] Saygin, Y., Verykios V. S. and Clifton C., "Using Unknowns to Prevent Discovery of Association Rules," ***ACM SIGMOD Record***, Vol. 30, No. 4, pp.45-54, 2001.

[17] Saygin, Y., Verykios, V. S., and Elmagarmid, A. K., "Privacy Preserving Association Rules Mining," ***Proceedings of the International Workshop on Research Issues in Data Engineering (RIDE)***, pp. 151-158, 2002.

[18] Vaidya, J. and Clifton, C., "Privacy Preserving Association in Vertically Partitioned Data," ***Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***, pp.639-644, 2002.

附錄一

健康檢查資料範圍意義[6]

症狀		過輕	正常	過重	肥胖
體重標準測量		<-10%	-10%-10%	10%-20%	>20%
症狀	指標	性別	偏低	正常	偏高
血液常規檢查	紅血球(RBC)	男	< 4.0	4.0-6.0	> 6.0
		女	< 3.5	3.5-5.5	> 5.5
	血紅素(HGB)	男	< 12	12-17	> 17
		女	< 10	10-15	> 15
	血球容積(HCT)	男	< 28	28-50	> 50
		女	< 35	35-47	> 47
	白血球(WBC)		< 4.0	4.0-10.0	> 10.0
	血小板(PLT)		< 300	300-400	> 400
血壓	收縮壓(BPC)		< 90	90-140	> 140
	舒張壓(BPG)		< 60	60-90	> 90
腎功檢查	尿素氮(BUN)		< 8	8-25	> 25
	肌酸酐(CR)		< 0.6	0.6-1.5	> 1.5
肝功能檢查	丙酮酸轉氨基胺(GPT)		< 5	5-35	> 35
	苯酮酸轉氨基胺(GOT)		< 8	8-40	> 40
總膽固醇	膽固醇(CHOL)		< 130	130-250	> 250