福祉科技與服務管理學刊
Journal of Gerontechnology and Service Management

研究論文

特徵分析和機器學習方法應用於肝臟疾病檢測

*陳志華[1,2,3] 楊子緯[4] 張訓楨[5] 賴永崧[6]
[1]中華電信研究院 智慧聯網研究所
[2]國立交通大學 資訊管理與財務金融學系
[3]國立交通大學 電機工程學系
[4]中山醫學大學附設醫院 肝膽腸胃科
[5]國立交通大學 應用數學系
[6]國立交通大學 生物科技系暨研究所

摘要

本研究旨在使用臨床上常見的生化檢測數據建構肝臟疾病的預測模式，期望能夠及早篩檢出肝臟疾病患者、及時轉介就醫。本研究使用數據為 UCI (University of California, Irvine)機器學習庫(Machine Learning Repository)所提供之「肝臟疾病資料檔」進行分析，考量 6 個特徵屬性（包含 Alanine aminotransferase (GPT)、Aspartate aminotransferase (GOT)等特徵屬性）判斷是否罹患肝臟疾病。並且，本研究結合臨床經驗，將 GOT 及 GPT 兩項肝功能生化數值欄位進行比值換算，新增兩個特徵屬性 GPT/GOT 和 GOT/GPT 來提升肝臟疾病篩檢正確率。在機器學習方法的部分，本研究使用決策樹、隨機森林、貝氏分類、支援向量機、k 個最近鄰居、以及類神經網路等方法，來進行實作和分析。在研究結果中，運用隨機森林方法，並結合新加入的兩個特徵屬性（即 GPT/GOT 和 GOT/GPT），將可有效將正確率提升至 73.91%，較其他機器學習方法好。因此，未來在篩檢肝臟疾病患者時，可以考慮運用 GPT/GOT 和 GOT/GPT 特徵屬性，以提升篩檢正確率。

關鍵詞：肝臟疾病、機器學習、特徵選取、隨機森林

1. 前言

根據 2014 年衛生統計資料結果顯示：肝和肝內膽管癌位居我國癌症十大死因第二位，死亡率為每十萬人口 34.9 人（衛生福利部，2015）。其中，感染 B 型肝炎或 C 型肝炎病患更是肝癌的高危險群，隨著病毒的潛伏、伺機而動，病患的肝臟由慢性肝炎反應走向肝纖維化、進而肝硬化，最後終至肝癌。幸而目前臨床醫學的進步，透過臨床血液檢查、影像檢查等方式可以及早篩檢出肝癌病患。因此，目前臨床醫師多會建議肝癌的高危險群病患平時定期返診追蹤，期能即早發現、即時處理。

*通訊作者：chihua0826@gmail.com

B 型肝炎、C 型肝炎病患是罹患肝癌的高危險群之一。其中，早期 C 型肝炎的治療主打是以干擾素合併使用抗病毒藥物 ribavirin，但其副作用的多面貌令許多病患裹足不前。近年來 C 型肝炎的抗病毒藥物如雨後春筍般的陸陸續續被研發成功，所以現在的病患僅需服用抗病毒藥物其治癒效果並不亞於干擾素合併療法。更令人讚嘆的是現在的治療藥物副作用已經趨近零，但藥價的昂貴也令人卻步。因此，公共衛生單位、臨床醫師們開始面臨另一個抉擇：面對著昂貴的藥價，是否可以先篩選出有立即治療必要的患者來接受治療，或是提供定期回診追蹤檢查的建議。

有鑑於此，本研究將先採用開放資料 UCI (University of California, Irvine)機器學習庫(machine learning repository)所提供之「肝臟疾病資料檔」(UCI, 1990)，運用該資料所擷取的生化數值和問卷調查數值進行分析。本研究亦結合臨床經驗，將生化數值資訊進行轉化產生新的特徵屬性以協助肝臟疾病篩檢。並且，提出一個特徵屬性分析方法，可分析和比較各個特徵屬性的重要性，藉以篩選重要的特徵屬性。並且，本研究結合機器學習方法來建立分類器，透過分析開放資料建立肝臟疾病篩檢和預測模式，以協助臨床判斷肝臟疾病，並避免肝臟切片手術產生的成本和風險。

本研究共分為五個章節，在第二節中將探討肝臟疾病篩檢作法和機器學習應用於肝臟疾病預測相關的研究背景和文獻探討。第三節說明本研究的研究流程和方法設計原理。第四節分析說明實驗環境與結果，分析各個特徵屬性和肝臟疾病篩檢正確率。最後一節則說明本研究的結論與未來研究方向。

2. 文獻探討

在本節中將先介紹肝臟疾病和臨床篩檢作法，再探討運用機器學習方法預測肝臟疾病的相關研究文獻，分述如下。

2.1 肝臟疾病和臨床篩檢作法

臨床上有所謂肝病三部曲：肝臟纖維化→肝硬化→肝癌，臨床醫師必須確立病患肝臟正處於哪個疾病階段，再與病患及家屬進行溝通討論以得到適合病患的個人化醫療，是目前臨床上所努力的方向(Wu & Lee, 2006)。因此，瞭解病患肝臟纖維化程度是為首要條件。目前臨床上主要確立肝臟纖維化程度的檢查仍是使用肝臟切片的侵入檢查，同時隨著疾病的進展，病患可能 3-5 年甚至更密集的反覆接受肝臟切片檢查，而非終身僅此一次。此外，嚴重肝硬化病患同時會合併血小板低下問題，更徒增病患接受該項檢查的風險。因此，臨床醫師在面臨病患拒絕或無法進行肝臟切片檢查的當下，如何評估、判斷並與病患共同進行醫療決策的討論亦是一大難題。

2.2 機器學習應用於肝臟疾病預測

近幾年隨著醫學的進步，臨床上的醫療處置亦隨著新知的公告而時時做調整。其中，肝臟疾病的進展更是明顯。並且，近幾年陸續有學者提出運用各種機器學習方法進行肝臟疾病篩檢和預

測。例如 Eslam 等人(2016)則分別收集 derivation cohort（C 型肝炎 1992 人）、independent cohort (C 型肝炎 1242 人)及 validation cohort（分別為 NAFLD 488 人及 B 型肝炎 555 人），共計 4,277 人的人口學資料、生化基本檢測及基因等數據，以決策樹(decision tree)方法來建立預測模式，並由實驗結果可知此預測模式可以取得病患肝臟纖維化程度之預測值。然而，此此研究需有完整的生化基本檢測及基因等數據才能達到高準確率。而 2015 年 Konerman 等人使用隨機森林(Random Forest)方法來建構比目前更為精準的預測模式。期望臨床上 C 型肝炎病患透過定期回診、定期血液檢查等方式取得其主要之預測變項連續數值，再使用該預測模式來輔助臨床醫師們與病患討論是否接受相關治療之決策(Konerman et al., 2015)。有鑑於此，由過去的文獻中主要可以了解到決策樹和隨機森林在肝臟疾病預測上有良好的表現，並且本研究主要將運用既有的開放資料集合來進行肝臟疾病分析，以期降低臨床實驗的成本。

3. 研究方法

在本節中將說明本研究的研究流程與步驟，並提出特徵分析方法，再介紹本研究將使用的機器學習方法，並依此來建立分類器，以應用於肝臟疾病篩檢和預測，分述如下。

3.1 研究流程與步驟

本研究的研究流程與步驟為取得 UCI 機器學習庫所提供之「肝臟疾病資料檔」、分析各個特徵屬性、建立機器學習分類器。

在 UCI 機器學習庫所提供之「肝臟疾病資料檔」的資料筆數共 345 筆，每一筆皆有 7 個特徵屬性，且無任何缺漏值，如表 1 所示。資料表中的 7 個特徵屬性分別為 Mean corpuscular volume (MCV)、Alkaline phosphotase (ALK-P)、Alanine aminotransferase (GPT)、Aspartate aminotransferase (GOT)、Gamma-glutamyl transpeptidase (rGT)、Drinks、以及 Selector，其中第 7 個屬性 Selector 為是否有罹患肝臟疾病的識別特徵屬性，分別如表 2 所示。後續將參考這些特徵屬性設計機器學習方法之分類器，建立肝臟疾病篩檢和預測模式。

表 1. 資料表內容說明

項目	內容
Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer, Real
Area	Life
Number of Instances	345
Number of Attributes	7
Missing Values	No
Date Donated	1990-05-15

表 2. 資料欄位說明

項次	欄位	說明	Alcohol(+)
1	Mean corpuscular volume (MCV)	MCV>100：酒精使用、B12 或葉酸缺乏 MCV<80：地中海貧血、缺鐵性貧血	↑, MCV > 100 (Macrocytic)
2	Alkaline phosphotase (ALK-P)	消化道疾病、膽道疾病、肝臟疾病、腎病、骨頭病變、懷孕	↑
3	Alanine aminotransferase (GPT)	肝臟疾病	↑
4	Aspartate aminotransferase (GOT)	肝臟疾病、骨骼肌、心肌受損	↑, GOT > 2xGPT
5	Gamma-glutamyl transpeptidase (rGT)	酒精使用、藥物、膽道疾病	↑
6	Drinks	飲酒量（個人飲酒習慣）	
7	Selector	是否有罹患肝臟疾病	

3.2 特徵屬性分析方法

本研究運用機率模型提出一個特徵屬性分析方法，以協助分類器挑選合適的特徵屬性。有鑑於在分類上，觀察同一個特徵屬性值於兩個類別的機率密度，當不同類別的機率密度交集面積越大時，代表資料的重疊程度越高，所以資料越無法區分到正確的類別；而當不同類別的機率密度交集面積越小時，代表資料的重疊程度越小，所以資料越有機會區分到正確的類別。因此，本研究採用此核心精神分析每個特徵屬性其在不同類別間的機率密度交集面積，並依交集面積由小到大排序以找出最重要的特徵屬性。

以 UCI 機器學習庫所提供之「肝臟疾病資料檔」為例，Selector-1 為有肝臟疾病、Selector-2 為無肝臟疾病。GOT 特徵屬性之機率交集面積為 86.9%（如圖 1 所示），而 Drinks 特徵屬性之機率交集面積為 80.0%（如圖 2 所示）。由此分析結果可知，觀察 GOT 特徵屬性的 Selector-1 類別和 Selector-2 類別的資料有 86.9%重疊在一起，重疊率較高，因此運用 GOT 特徵屬性將可能無法明確區分出是屬於 Selector-1 類別或 Selector-2 類別。而觀察 Drinks 特徵屬性在這兩個類別的分類上，資料分佈相較於 GOT 特徵屬性其重疊率較低，將更有助於肝臟疾病篩檢和預測。

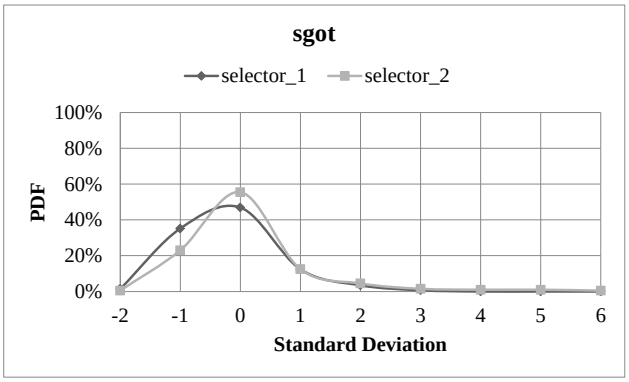

圖 1. GOT 特徵屬性之機率交集面積

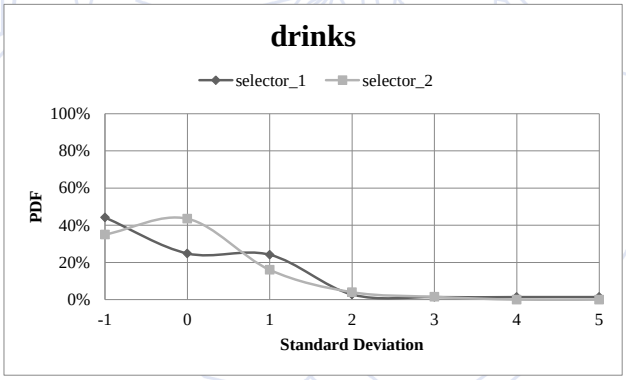

圖 2. Drinks 特徵屬性之機率交集面積

3.3 機器學習方法

本研究採用的機器學習方法有決策樹(decision tree)、隨機森林(random forest)、貝氏分類(naive bayes)、支援向量機(support vector machine)、k 個最近鄰居(k nearest neighbors)、以及類神經網路(neural network)等 6 種方法來實作分類器，以建立肝臟疾病篩檢和預測模式，分述如下。

3.3.1 決策樹

決策樹主要將獨立分析每個特徵屬性，並計算該特徵屬性值在每個類別分類上的混亂程度，藉以評估挑選該特徵屬性可以得到多少資訊量。計算每個特徵屬性其所能得到的資訊量來計算出每個特徵屬性的重要程度，並挑出最重要的特徵屬性作為根節點。當挑選出根節點後，再依根節點特徵屬性的值來區分出不同的葉節點，再分別計算每個葉節點之每個特徵屬性的重要程度，挑選出該階段最重要的特徵屬性作為該節點的特徵屬性。依此類推，將能依資料集合內容建立出一棵決策樹分類器，並可根據輸入的資料進行分類和預測(Quinlan, 1987)。

3.3.2 隨機森林

隨機森林的核心精神為建立多個決策樹分類器，再由個別的決策樹分類器得到的分類結果取眾數決定。其作法主要將先隨機挑選資料集合中部分或全部的特徵屬性，並隨機挑選部分或全部的資料建立決策樹分類器。其中，每個決策樹分類器將由不同的特徵屬性組合，不同的資料子集合組合所建立，故每個決策樹分類器將可從不同的資料面向來進行分類和預測，藉以提升分類的正確率(Ho, 1998)。

3.3.3 貝氏分類

貝氏分類方法主要將先考慮每個特徵屬性其劃分到每個類別的條件機率，再計算每個特徵屬性組合的情況下，其條件機率乘積後每個類別的機率值，選擇出機率最高的類別作為分類結果(Tang et al., 2016)。在貝氏分類方法的使用上，由於在不同特徵屬性組合時，主要將彼此的條件機率直接進行乘積，隱含假設每個條件機率在組合時其特徵屬性為獨立關係，故當特徵屬性間的相依程度高時，將可能導致較大的誤差。

3.3.4 支援向量機

支援向量機的核心精神為建構一個或多個高維度的超平面，並以此超平面作為分類的邊界，用以區分不同類別的資料點。在訓練資料點與超平面（即分類邊界）的距離稱之為間隔(margin)，並在此方法中主要將讓間隔越大越好，如此越能把資料區分清楚，減少分類誤差。因此，支援向量機將運用一些數學模型計算來找出間隔最大的超平面，以作為資料分類邊界(Cortes & Vapnik, 1995)。

3.3.5 k 個最近鄰居

k 個最近鄰居方法的作法為將新的資料與每一筆歷史資料進行比對，計算與每一筆資料的相似度或距離，挑出最相似或最接近的 k 筆資料，再判斷這 k 筆資料所對應的類別，以眾數的方式決定新資料所屬類別(Altman, 1992)。在 k 個最近鄰居方法的使用上，主要有參數 k 和相似度算法或距離算法需要設定，並且將可能因 k 值不同，而導致不同的分類結果。因此，在本研究中將採用不同的 k 值組合進行分析和驗證。

3.3.6 類神經網路

類神經網路主要將建立輸入層、隱藏層、輸出層的人工神經網路，並將每個輸入的資料特徵屬性作為一個神經元，且每個輸出的資料作為一個神經元。在一個或多個隱藏層中建立多個神經元，再把每個神經元間進行連結，為每個連結建立權重值。最後，依資料集合來調整每個權重值，以讓輸入值經由激勵函數(activation function)和加權計算後，可以得到預期的輸出值。而當與輸出值不同時可以再回饋調整權重值，不斷讓類神經網路進行機器學習和調整，直到類神經網路計算收斂為止(LeCun et al., 2015)。

4. 實驗環境與結果

在本節中將分別各特徵屬性重要性、分類結果分析、以及結合新建因子分類結果分析進行詳細探討。

4.1 特徵屬性分析

在研究中主要以 Selector 欄位（是否有罹患肝臟疾病）特徵屬性將資料分為兩類，Selector-1 為有肝臟疾病、Selector-2 為無肝臟疾病。對將每個特徵屬性值分別以這兩類進行區分後，統計其機率密度函數，並分析每個特徵屬性中兩個類別的機率交集面積。當交集面積越高，則代表資料重疊性越高，則該特徵屬性應用在此分類上的效果則越不顯著。如果交集面積越小，則代表資料重疊性越低，則該特徵屬性應用在此分類上將可以得到越好的分類正確率。MCV、ALK-P、GPT、GOT、rGT、Drinks 之 6 個特徵屬性的機率交集面積結果分別如圖 3-圖 8 和表 3 所示，並由分析結果可知每個屬性的機率交集面積皆在 80%以上，故每個屬性單獨分析時將無任何一個屬性能顯著區分出目標類別資料，而相對比較下 Drinks 和 rGT 這兩個屬性對分類上將較具有影響力。

此外，本研究有鑑於臨床經驗中 GPT/GOT 和 GOT/GPT 可作為判斷肝病的參考依據，因此再將 GPT 和 GOT 兩個因子進行組合和綜合判斷，新建兩個因子 GPT/GOT 和 GOT/GPT。這兩個新建因子的機率交集面積分析結果分別如圖 9-圖 10 和表 4 所示，並且可以得知新建因子的交集面積較小，將更有助於資料分類。

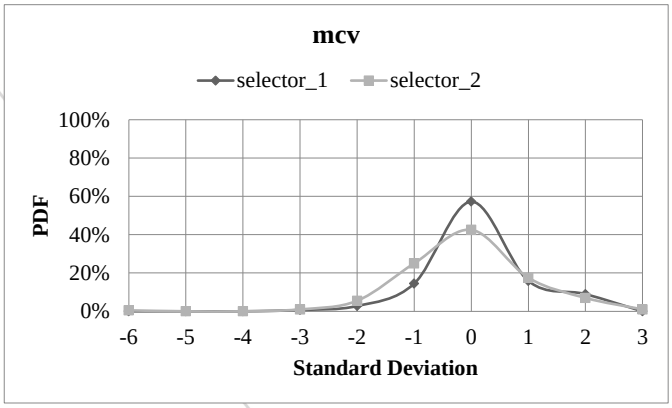

圖 3. MCV 特徵屬性之機率交集面積

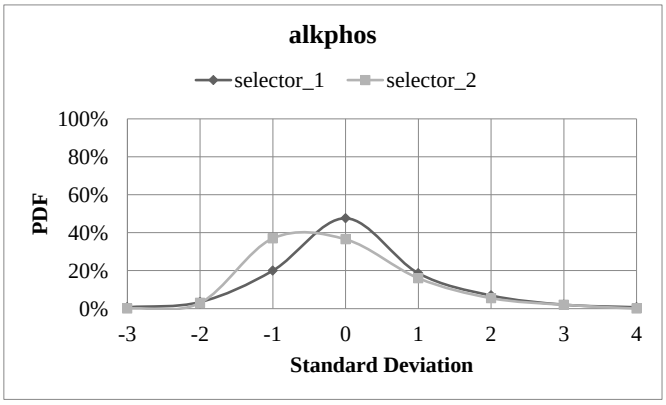

圖 4. ALK-P 特徵屬性之機率交集面積

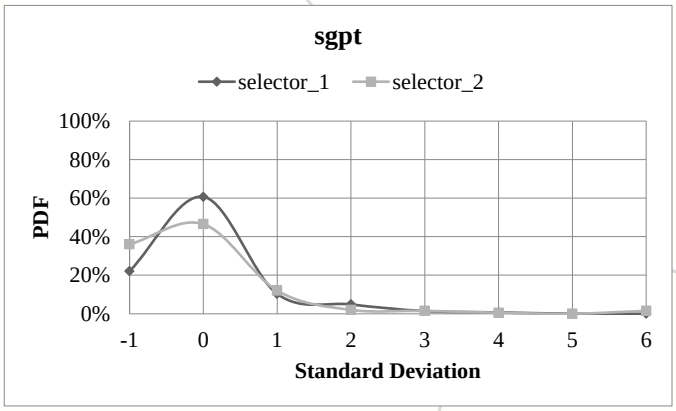

圖 5. GPT 特徵屬性之機率交集面積

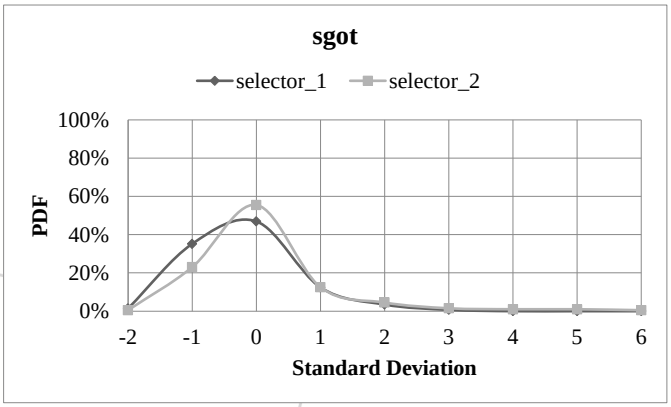

圖 6. GOT 特徵屬性之機率交集面積

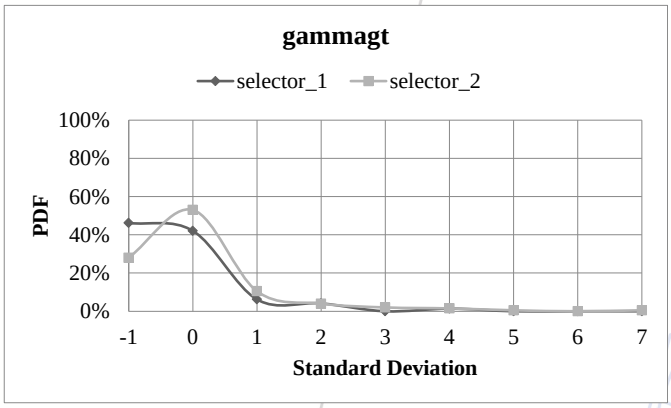

圖 7. rGT 特徵屬性之機率交集面積

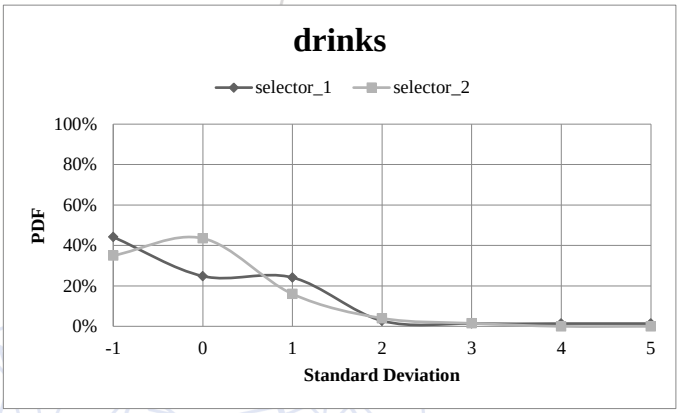

圖 8. Drinks 特徵屬性之機率交集面積

表 3. 各因子交集面積分析結果

因子	交集面積	排名
MCV	83.3%	5
ALK-P	83.0%	4
GPT	82.8%	3
GOT	86.9%	6
rGT	81.7%	2
Drinks	80.0%	1

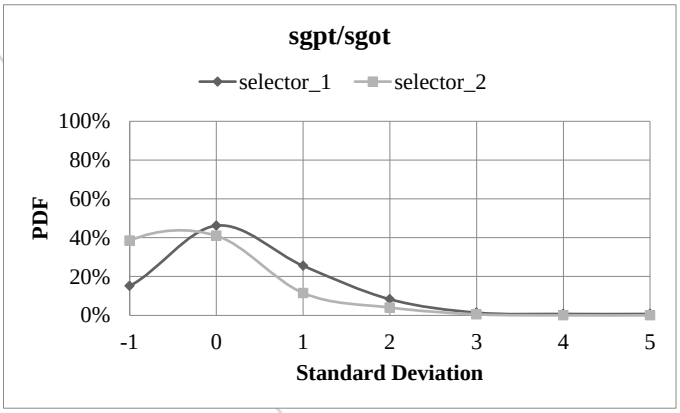

圖 9. GPT/GOT 機率交集面積

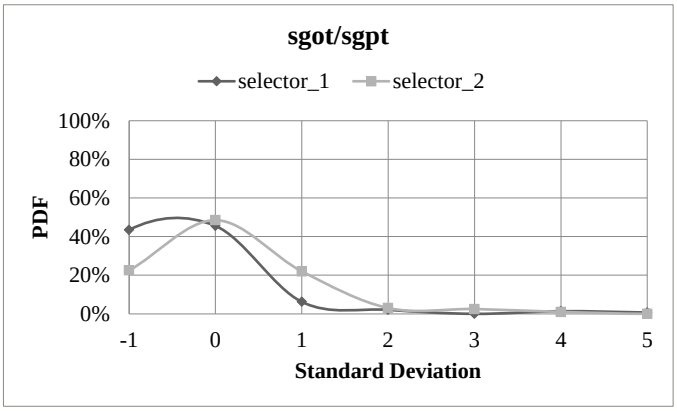

圖 10. GOT/GPT 機率交集面積

表 4. 各因子交集面積分析結果

因子	交集面積	排名
MCV	83.3%	7
ALK-P	83.0%	6
GPT	82.8%	5
GOT	86.9%	8
rGT	81.7%	4
Drinks	80.0%	3
GPT/GOT	74.2%	1
GOT/GPT	77.3%	2

4.2 肝臟疾病篩檢結果分析

在本研究中分別採用決策樹、隨機森林、貝氏分類、支援向量機、k 個最近鄰居以及類神經網路進行分類器實作，並且皆採用 R 語言及相關套件分別實作這些分類器，如表 5 所示。

表 5. 分類器和套件與參數設定

分類器	套件名稱	參數設定
決策樹	party	
隨機森林	randomForest	ntree: 1000
貝氏分類	e1071	
支援向量機	e1071	
k 個最近鄰居	class	k: 1~7
類神經網路	neuralnet	Learning rate: 0.1 Threshold: 0.01

其中，k 個最近鄰居分類器的分析上與參數 k 值有關，因此在本研究中實作 k 值為 1~7 不同的情境，並分析每個情境的正確率，如表 6 所示。由實驗結果可知，當 k 值為 6 時，將可以得到較高的正確率 66.09%，後續分析上將採用此正確率進行比較。

表 6. k 個最近鄰居分類器之 F1-Measure 和正確率分析

k	F1-Measure	正確率（標準差）
1	59.89%	60.87% (4.66%)
2	59.90%	60.58% (4.28%)
3	59.89%	61.45% (1.79%)
4	63.60%	64.93% (4.56%)
5	63.59%	65.22% (3.09%)
6	64.63%	66.09% (3.09%)
7	63.84%	65.51% (1.48%)

此外，類神經網路分類器的分析上與網路結構有關，因此在本研究中實作不同的網路結構，分別採用 1 個隱藏層和 2 個隱藏層，並在隱藏層中建立不同的神經元數量，再分析每個情境的正確率，如表 7 所示。由實驗結果可知，當結構為 6-1 時，將可以得到較高的正確率 71.68%，後續分析上將採用此正確率進行比較。

表 7. 類神經網路分類器之 F1-Measure 和正確率分析

Network Structure	F1-Measure	正確率（標準差）
6-1	69.72%	71.68% (3.09%)
6-1-1	65.73%	69.28% (2.05%)
6-2-1	69.28%	70.43% (3.76%)
6-3-1	69.06%	70.72% (1.79%)
6-4-1	65.95%	66.67% (2.69%)
6-5-1	65.32%	66.96% (1.88%)
6-6-1	67.16%	68.12% (2.05%)
6-6-6-1	60.65%	62.32% (3.50%)

綜合比較每個分類器，決策樹、隨機森林、貝氏分類、支援向量機、k 個最近鄰居、以及類神經網路這 6 個分類器應用於肝病分類上，正確率分別為 61.16%、71.30%、55.36%、72.46%、66.09%、以及 71.68%，如表 8 所示。由實驗結果可知，支援向量機在此資料集中的表現最佳，可提供 72.46& 的正確率，可採用支援向量機進行肝病判斷。

表 8. 各個分類器之 F1-Measure 和正確率分析

分類器	F1-Measure	正確率（標準差）
決策樹	53.63%	61.16% (4.73%)
隨機森林	69.67%	71.30% (3.09%)
貝氏分類	55.20%	55.36% (5.74%)
支援向量機	70.05%	72.46% (1.08%)
k 個最近鄰居	64.63%	66.09% (3.09%)
類神經網路	69.72%	71.68% (3.09%)

4.3 結合新建因子分類結果分析

在本研究中，加入臨床經驗之 GPT/GOT 和 GOT/GPT 兩個新建因子，並在前面章節已分析 GPT/GOT 和 GOT/GPT 的機率交集面積較小，預期將有助於肝病分類。在此節中，分別考慮 MCV、ALK-P、GPT、GOT、rGT、Drinks、GPT/GOT、以及 GOT/GPT，共 8 個資料屬性，並採用決策樹、隨機森林、貝氏分類、支援向量機、k 個最近鄰居、以及類神經網路這 6 個分類器進行分類和實驗比較，如表 9 所示。其中，k 個最近鄰居和類神經網路的參數設定和正確率分析，分別描述於附錄 A 和附錄 B。將表 8 和表 9 進行比較，由實驗數據可知，加入臨床經驗之新建因子 GPT/GOT 和 GOT/GPT 後，每個分類器正確率平均提升 2~3%，並且將此表 8 和表 9 兩組正確率以 t 檢定進行成對比較測試後，可得 p-value 為 10%，故加入新建因子後平均值有顯著差異，所以加入新建因子將能提升分類正確率。因此，未來在判斷肝病上將應同時考量 GPT/GOT 和 GOT/GPT 兩個因子。

表 9. 各個分類器之 F1-Measure 和正確率分析

分類器	F1-Measure	正確率(標準差)
決策樹	60.39%	64.64% (1.79%)
隨機森林	72.69%	73.91% (0.71%)
貝氏分類	66.10%	66.38% (2.28%)
支援向量機	71.26%	73.04% (1.42%)
k 個最近鄰居	64.19%	65.80% (1.79%)
類神經網路	72.20%	73.62% (0.41%)

5. 結論與未來研究

有鑑於目前臨床上要確立肝臟纖維化程度的檢查仍是使用肝臟切片的侵入檢查，同時隨著疾病的進展，病患可能 3-5 年甚至更密集的反覆接受肝臟切片檢查，將可能造成肝病患者較大的醫療成本和手術風險。因此，本研究主要使用臨床上常見的生化檢測數據建構肝臟疾病的預測模式，期望能夠及早篩檢出肝臟疾病患者、即時轉介就醫。在本研究中，主要使用 UCI 機器學習庫之「肝臟疾病資料檔」進行分析和驗證，並提出特徵屬性分析方法和結合臨床經驗，將 GOT 及 GPT 兩

項肝功能生化數值欄位進行比值換算，新增兩個特徵屬性 GPT/GOT 和 GOT/GPT。最後，本研究再結合機器學習方法，實作分類器進行肝臟疾病篩檢和預測。由實驗結果顯示，運用隨機森林方法，並結合新加入的兩個特徵屬性(即 GPT/GOT 和 GOT/GPT)，將可有效將正確率提升至 73.91%，較其他機器學習方法好。因此，未來在篩檢肝臟疾病患者時，可以考慮運用 GPT/GOT 和 GOT/GPT 特徵屬性，以提升篩檢正確率。

本研究目前主要採用常見的機器學習方法進行實驗和分析，用以證實新建之 GPT/GOT 和 GOT/GPT 特徵屬性可協助臨床提高篩檢正確率，但在研究中未修改機器學習方法。在未來研究上，將可考慮深入分析資料分佈狀況，並改良現有的機器學習方法，例如：可考慮依隨機森林方法的核心精神，隨機產生多個支援向量機分類器，並可以分析每種特徵屬性的排列組合，藉以找出最佳特徵屬性組合，以期提升肝臟疾病篩檢和預測正確率。此外，由於目前 UCI 機器學習庫所提供的資料僅包含有 6 個特徵屬性，且未具備年齡、性別等特徵屬性，因此未來可以取得更多不同的臨床資料和生化基本檢測及基因來進行實證分析，進行肝臟疾病篩檢和預測模式的測試，並針對不同年齡層（如銀髮族）提供健康促進實施建議。

參考文獻

1. Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185.
2. Cortes, C.; Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.
3. Eslam, M.; Hashem, A.M.; Romero-Gomez, M.; Berg, T.; Dore, G.J.; Mangia, A.; Chan, H.L.; Irving, W.L.; Sheridan, D.; Abate, M.L.; et al. (2016). FibroGENE: A gene-based model for staging liver fibrosis. *Journal of Hepatology, 64*(2), 390–398.
4. Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.
5. Konerman, M.A.; Zhang, Y.; Zhu, J.; Higgins, P.D.; Lok, A.S.; Waljee, A.K. (2015). Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology, 61*(6), 1832–1841.
6. LeCun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature, 521, 436–444.
7. Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies, 27*(3), 221–234.
8. Tang, B.; He, H.; Baggenstoss, P.M.; Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering, 28*(6), 1602–1606
9. UCI (1990). UCI Machine Learning Repository: Liver Disorders Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Liver+Disorders
10. Wu, Y.M.; Lee, P.H. (2006). Hepatocellular Carcinoma. *Formosan Journal of Medicine, 10*(4), 482–487.
11. 衛生福利部(2015)。103 年國人死因統計結果。
    Retrieved from http://www.mohw.gov.tw/news/531349778

附錄 A. k 個最近鄰居分類器之 F1-Measure 和正確率分析

本節主要描述在結合新建因子後，k 個最近鄰居分類器實作 *k* 值為 1~7 不同的情境，分析每個情境的正確率，如表 10 所示。由實驗結果可知，當 *k* 值為 7 時，將可以得到較高的正確率 65.80%，在 4.3 節的實驗結果中將採用此正確率進行比較。

表 10. k 個最近鄰居分類器之 F1-Measure 和正確率分析

k	F1-Measure	正確率（標準差）
1	60.22%	61.16% (4.28%)
2	58.37%	59.71% (3.91%)
3	59.89%	61.45% (1.79%)
4	59.65%	60.87% (1.88%)
5	63.59%	65.22% (3.09%)
6	63.24%	64.93% (2.87%)
7	64.19%	65.80% (1.79%)

附錄 B. 類神經網路分類器之 F1-Measure 和正確率分析

本節主要描述在結合新建因子後，類神經網路分類器實作不同的網路結構，分別採用 1 個隱藏層和 2 個隱藏層，並在隱藏層中建立不同的神經元數量，再分析每個情境的正確率，如表 11 所示。由實驗結果可知，當結構為 8-2-1 時，將可以得到較高的正確率 73.62%，在 4.3 節的實驗結果中將採用此正確率進行比較。

表 11. 類神經網路分類器之 F1-Measure 和正確率分析

Network Structure	F1-Measure	正確率（標準差）
8-1	68.79%	71.30% (0.00%)
8-1-1	69.71%	72.46% (3.20%)
8-2-1	72.20%	73.62% (0.41%)
8-3-1	70.55%	72.17% (1.42%)
8-4-1	66.49%	67.54% (1.08%)
8-5-1	66.08%	67.25% (1.64%)
8-6-1	67.82%	68.70% (1.88%)
8-6-6-1	60.65%	62.32% (3.50%)

Feature Extraction and Machine Learning Methods for Liver Disease Detection and Prediction

*Chen, C.-H.[1, 2,3], Yang, T.-W.[4], Chang, H.-C.[5], Lai, Y.-S.[6]
[1] Telecommunication Laboratories, Chunghwa Telecom Co., Ltd.
[2] Department of Information Management and Finance, National Chiao Tung University
[3] Department of Electrical and Computer Engineering, National Chiao Tung University
[4] Chung Shan Medical University Hospital
[5] Department of Applied Mathematics, National Chiao Tung University
[6] Department of Biological Science and Technology, National Chiao Tung University

Abstract

This study focuses on liver disease detection and prediction models that use biochemistry data to medically screen patients. The liver disorders data set of UCI (University of California, Irvine) machine learning repository which includes six features (e.g., alanine aminotransferase (GPT) and aspartate aminotransferase (GOT)) is adopted to detect and predict liver disease. Furthermore, this study considers clinical experiences to analyze GPT and GOT data and generate two features – GPT/GOT and GOT/GPT – for the improvement of liver disease detection and prediction accuracy. For the evaluation of machine learning methods, this study uses and implements decision tree, random forest, naive Bayes, support vector machine, k nearest neighbors, and neural network. Experimental results show that the accuracy of liver disease detection and prediction can be improved up to 73.91% by using random forest with the two proposed features. Therefore, the two proposed features can be adopted into liver disease detection and prediction models for future clinical trials.

Keywords: liver disease, machine learning, feature extraction, random forest