國立勤益科技大學
工業工程與管理系

碩士論文

應用資料採礦技術來預測電影票房之研究

(以美國市場為例)

指導教授：曾懷恩　　博士

研 究 生：楊孟迪

學　　號：49715010

中 華 民 國 九 十 九 年 六 月

Abstraction

Motion pictures are the most important industrial of entertainment. There are more than 4000 movies to be produced every year. American is the biggest market in the world. There were more than 9 billion dollars revenue of box-office in 2004 (Motion Picture American Association) .The box-office of a motion picture could be significant, but the budget of it is often very large. A movie will be under risk without considerable investment. No one can be sure that a movie is going to have successful box-office before it shows up on the screens. Most predicting models of box-office in previous works had to wait until the movie is played in theaters. About 25% of box-office are gained in the first 2 weeks. It is easier to predict the final box-office after knowing the first 2 weeks revenue. This study collected 10 years movie data of American market from 1999 to 2008 and used neural network and decision tree to build predicting models. These models are able to predict how much the box-office of a motion picture will be before manufactory starts to make it. It can help investors to avoid unnecessary lost of making a movie.

Keywords: Predicting Model, Neural Network, Decision Tree.

摘要

隨著生活水平的提高，電影產業也相對的被重視，全世界每年製造約 4000 部電影。然而美國的電影工業非常的發達，是全世界最活躍的市場，光是在 2004 年就有 90 億美元的票房收入(Motion Picture American Association)。不過製作成本常常高達上億美金，雖然票房也可能非常可觀，但是如果若是貿然的投資卻很有可能讓片商血本無歸，因為在電影正式上映前，沒有人可以預測觀眾對於這部影片的喜好程度，也沒人可以保證一定會賣座。經過過去的文獻調查發現，許多研究針對預測做出模型，只是大部分的研究都只是在電影上映後，預測最後總票房的落點。大約 25%的票房是在前兩週上映的時候得到的，一旦確認了前兩週的票房收入要預測最後的票房就相對的容易。本研究將使用資料採礦的方法，建立起過去 10 年電影票房的資料庫，並使用類神經網路(Neural Network)和決策樹(Decision Tree)為工具，製作一個電影製造前期的預測模型，讓片商在製作時可以輸入準備投入製作的資源、特徵，去預測票房的落點，避免投資者的損失。

關鍵字:電影預測模型，資料採礦，類神經網路，決策樹。

目錄

圖目錄

表目錄

一、緒論

1.1 研究背景與動機

電影工業在美國是相當重要的娛樂工業，平均估每部電影估計花費四千萬美金去製作，還有大約兩千萬美金去做行銷宣傳費用。但是卻低於 70%的電影是可以獲利的，大部份的電影上映週數都不會超過 15 周(Eliashberg and Sawhney, 1994)。每部電影的生命週期是如此短暫，並且是獨一無二，所以事先分析好每部電影的組成屬性，在製作前加以分析跟預測將會變的非常的重要。

然而預測票房系統是非常困難，且充滿挑戰，對於每部電影需求的對象都不一樣，這種不確定性，對電影製作的投資者來說充滿了風險。前美國電影協會總裁 Valenti 曾說: ”沒人可以知道一部電影是否會賣座，直到上映，畫面呈現在觀眾眼前的那一刻。” (Valenti , 1978)

即使預測票房系統不是容易的，仍有許多研究針對預測做出模型，只是大部分的研究都只是在電影上映後，預測最後總票房的落點，然而學者發現大約有25%的票房是在上映後兩周內得到的(Litman and Ahn, 1998)，而一旦確認了前兩週的票房收入，要預測最後的票房就相對的容易。然而如果是想要在上映前做預測的話就更加的困難。

為了能讓現今競爭激烈的環境下，投資者能更有目標的去投資電影，能掌握到未來可能得到的票房落點，就能在電影製作前期新增或更改電影的內涵特徵值，預測可能的票房落點，避免過度的投資所造成的損失，能對電影未來上映的走勢更有把握。

本研究目主要有三大動機：

1. 對現有的電影票房紀錄資料進行分析研究，找出賣座和不賣座電影的規則，是否與其所含屬性，如上映日期、導演、電影分級等等有關聯。
2. 以製作片商的觀點來建立模型來預測未來電影的票房，從製作電影準備投入的資訊，當製片人提出了製作影片的提案，即可以此的模型，去判斷以此電影製作提案的元素或是特徵是否會賣座，是否能夠獲利，或是造成片商損失，進而對於其投入做調整，或是否決此提案，減少發生虧損的機率。
3. 本研究針對連續十年美國上映電影的資料建立資料庫(1999~2008)，包含觀眾對於電影喜愛程度，作為評斷是否會是影電影賣座的歷史資訊。並且用不同的資料採礦的方法建立模型，找出效果較佳的模型。

1.2 研究目的

電影在全世界都是很重要的娛樂工業，全世界每年約製造 4000 部電影(Motion Picture Association of America, 2004)。美國的電影工業非常的發達，是全世界最活躍的市場，光是在 2004 年就有 90 億美元的門票收入(Eliashberg et. al ,2005)。而製作成本常常高達上億美金，雖然票房也可能非常可觀，但是如果若是貿然的投資卻很有可能讓片商血本無歸，因為沒有人可以預測觀眾對於這部影片的感覺。

在此研究中將用類神經網路(Neural Network) 還有決策樹 (Decision Tree)方法在製作前期製作模型，讓電影在製作前期的時候能有一個預測的模型，知道投入製作電影的屬性會讓此電影票房的落點在哪裡，需要改變或是加強哪些屬性，減少電影失敗的可能。

決策樹是一個預測模型，代表的是對象屬性與對象值之間的一種映射關係。樹中每個節點表示某個對象，而每個分叉路徑則代表的某個可能的屬性值，而每個葉結點則對應從根節點到葉節點所經歷的路徑所表示的對象的值。決策樹僅有單一輸出，若欲有複數輸出，可以建立獨立的決策樹，以處理不同輸出。

決策樹的優點在於它可以產生易於了解明白的規則，不需要太多計算方式就可以進行分類，且訓練時間短，可提供明確的方向，告知在進行預測的和分類，哪一個變數是最為重要的。

SPSS Clementine 10.1(2006)這套工具結合商業技術可以快速建立預測性模型，進而應用到商業活動中，幫助人們改進決策過程。Clementine 10.1 具有強大的數據挖掘功能和顯著的投資回報率。同那些僅僅著重於模型的外在表現而忽略了數據挖掘在整個業務流程中的應用價值的其它數據挖掘工具相比，Clementine 10.1 其功能強大的數據挖掘算法，使數據挖掘貫穿業務流程的始終，在縮短投資回報週期的同時極大提高了投資回報率。

1.3 研究對象

本研究搜集美國電影近十年(1999~2008)上映電影的資訊，因為美國市場有完整的公開資訊紀錄，故選作本研究的實驗對象。因在進行資料搜集時為 2009 年，故 2008 年以前的資料及票房資訊較為完整，故從 1999 年開始至 2008 年的資訊。搜集 1782 筆資料，其中包含了電影的成本，導演的資訊，上映廳數，電影類型，是否是續集，分級、影星、上映競爭度，還有北美票房成績。資料來源為電影資料庫網站 BoxOffice Mojo [1]以及 IMDb [11]。

1.4 研究流程

本計劃書主要研究架構分為，各章節內容敘述如下：

1. 緒論：　說明研究背景與動機、研究目的、研究範圍與研究流程。
2. 文獻探討：探討有關資料採礦及電影預測的相關研究。
3. 研究方法：本研究將利用決策樹中的 C5.0，和類神經網路的倒傳遞類神經網路(BPN)的方法，使用 Clementine 10.1 去尋找其中的規則，並建立模型。
4. 個案研究：將資料來源選取的投入產出變數，並且利用決策樹與類神經及兩種分析方法來從資料中建立預測模型及對照組，再探討是否準確。
5. 結論與建議：針對個案研究結果做最後評述，並提出未來可以進一步研究的方向。

二、文獻探討

2.1 資料採礦

資料採礦是資料倉儲應用方式中最重要的一種。資料採礦是用來將資料中隱藏的資訊挖掘出來，所以資料採礦其實是所謂的知識發現的一部份，資料採礦使用了許多統計分析與建模的方法，到資料中尋找有用的特徵（Patterns）以及關連性（Relationships）。

資料採礦所得到的知識，來自大量的資料庫，它被應用到許多問題上，諸如醫療系統、混合銷售、電信客戶分析、廣告分析等等議題。其最為重要的貢獻從龐大的資料庫萃取出有意義的資訊及有結構的模式，作為接下來做決策的依據。資料採礦主要具有以下的功能(Fayyad et al. ,1996):

1.預測(Prediction)：由現在的資料庫數據去預測未來數據或是行為。使用工具包括迴歸分析、時間數列分析及類神經網路。

2.摘要(Summarization)：針對資料的子集合，去發現能簡潔描繪資料的方法。

3.分類(Classfication)：將資料項目分配至數個預先定義的集合中，按照分析對象的屬性加以定義，建議類別(Class)。例如將電信用戶區分為高風險流失群、中風險流失群及低風險流失群。可使用決策樹(Decision Tree)、記憶基礎推理(Memory Based Reasoning)。

4.關聯規則(Association Rules)：以此來幫助判斷哪些物品應該放至同一位置，提高購買率。譬如超市 Wal-Mart 就發現，當啤酒跟尿布擺放在一起的時候可以提升銷售量。可以用此功能來設計出具有交叉銷售(Cross Selling)機會的產品群組。

5.估計(Estimation)：由現有的數據資料，估計未知且連續變數。使用的工具為統計方法上之相關分析，迴歸分析及類神經網路。

6.分群(Clustering)：尋找識別資料的有限種類集合或能夠描述資料的類別，和分類最大的差別在於不需要事先定義集合，直到分群結果出現才能解釋其涵義。使用技巧有 K-means 法、Kohonen 法等。

1. Frawley et al. (1991) 資料採礦是由資料中挖掘非顯然的、未知的、潛在”可能”有用資訊過程。
2. Grupe and Owrang (1995) 資料採礦是指從已存在的資料庫中挖掘出新的事實，或者發現專間尚且不知道的新關係。
3. Fayyad et. al (1996)資料採礦是一個確定資料中有效的、從未發現過的與可能有用的，並且能被容易了解其模式的掘取過程。
4. Michael et. al (1997) 資料採礦是為了要發現出有意義的樣型或規則，而必須從大量資料中以自動或是半自動的方式來探索和分析資料。

5. Kleissner (1998) 資料採礦是一種新的且不斷循環的決策支援分析過程，它能夠從組合在一起的資料中，發現出隱藏價值的知識，以提供企業專業人員參考。
6. Kim (2000) 資料採礦為了提煉出可能有用的知識，用來幫助決策的一系列程。
7. Han and Kamber (2001) 大致而言，資料採礦乃是從儲存在資料庫、資料倉儲或其他資訊儲存體中的大量資料中，發現有意義與效益的特徵。

歸納上述學者的言論，皆認為資料挖掘位在一大群資料中找出有效益的特徵，同時資料挖掘也是知識發掘(KKD, Knowledge Discovery in Databases)過程中的一個重要或不可獲缺步驟。可以了解知識發掘(KKD)為一重要的、確認有效的、潛在有用的程序，最後的目地為了解資料的特徵、型態、趨勢與規則性。依 Fayyad 之定義得知，資料挖掘僅為其流程中一個步驟也屬於最重要的一環，整個過程包括對應用領域的認知、具備相關的專業知識，確認分析目標的資料來源，接著進行資料淨化(Data Selection)、資料整合(Data Transformation)、資料採礦(Data Mining)、型樣評估(Pattern Evaluation)、知識呈現(Knowledge Representation)等七個步驟 (Han and Kamber, 2001) 。

1.資料淨化：實際應用的資料庫，通常會有不完整，或是有雜訊，及不一致的情況產生。資料淨化的工作主要是要將不完整的資料(Missing Value)、去除資料中的離群值，已修正資料的不一致性等等。
2.資料整合：將目標相關的各種資料來源加以整合，例如將不同類型的資料庫整合成資料倉儲。
3.資料選擇：自資料庫或資料倉儲中取得與分析目標相關的資料。
4.資料轉換：執行 Summary 及 Aggregation 的運算，將資料轉換成適合採礦的形式。
5.資料採礦：是一種不斷循環的決策分析過程，運用智慧型方法自資料中萃取出特徵值(Patterns)、關係(Relation)等等。
6.型樣評估：資料採礦所產生的型樣並非完全都是使用者所期待的資訊，因此必須利用一個較客觀的評估方式，確認出真正有意義的知識。
7.知識呈現：使用視覺化的技術或是知識表達的方法，將知識傳達給使用者。

從以上七個步驟來看，資料採礦牽涉到大量的規劃與準備工作，在資料採礦的過程中，有 80%的時間是在於準備資料階段，這部份包含了整理表格內容還有資料轉換的工作。所以資料採礦單純只是知識發覺過程中的一個步驟而已，要完成這個步驟有很多前置工作要先完成。

而資料採礦要實際應用到企業、公司的時候，很重要的還是需要該領域的相

關知識，資料採礦的工具要結合該公司領域，專業的工具、語言去進行分析，才能獲得對於公司真正有用的知識。資料採礦就是希望能從龐大的資料庫找到精華的、前所未知的、可以理解的知識，讓企業能從中獲得利益。

但是在進行資料採礦的過程中，因為很多因素的影響可能導致結果不如預期，資料內容太複雜、資料品質的好壞、資料數量多寡、資料遺漏值、資料取得難易度、資料的時效性等等。所以這表示從資料取得知識充滿了變化，並沒有一個模型可以用到所有問題，所以必須要了解資料項目所代表的資訊，進而可以加以說明跟解釋結果。

2.2 跨產業資料採礦標準作業程序

跨產業資料採礦標準作業程序 CRISP-DM (Cross Industry Standard Process for Data Mining)，其資料採礦作業程序主要是由 NCR Systems Engineering Copenhagen (美國和丹麥)、DaimlerChrysler AG (德國)、SPSS Inc. (美國)、OHRA Verzekeringen en Bank Group B.V (荷蘭)這幾家公司在 1996 年聯合發展而成。CRISP-DM 具有產品中立性，使用上並不受限於特定作業平台。以下將介紹 CRISP -DM 作業程序之六個主要步驟：

1. 了解企業需求 (Business Understanding)：重點主要是以企業的觀點來找出推動此方案的目的，在此步驟要先定義資料採礦問題，並且訂定初步計畫方案。
2. 了解資料特性 (Data Understanding)：主要為收集完整資料，並對收集的資料作初步分析，包括識別資料的質量問題、找到對資料的基本觀察，接著並設立假設前提。
3. 準備資料 (Data Preparation)：主要為篩選資料中各項表格、紀錄以及變數，接著整理經篩選出來的資料，則即可應用於模型選擇工具上。
4. 設計模型 (Modeling)：此步驟著重於選擇並應用一或多種資料採礦技術。
5. 評估 (Evaluation)：主要為分析結果，並證實前一步驟設計的模型是否符合企業所推動方案之目的，以及進一步的決定將來是否繼續採用此一模型。
6. 建置 (Deployment)：此步驟主要是經評估後，若所建立之模型符合企業目標，則將再進一步擬定該模式之推動計畫。

2.3 電影預測相關文獻

Litman (1983) , Litman and Kohl (1989) , Litman and Ahn (1998) 和Elberse andEliashberg (2002) 提出的模型是屬於以經濟跟定性方法，針對會影響到近期上映電影票房收的因子去建立模型。De Silva (1998) , Eliashberg and Sawhney (1994) , Eliashberg et al. (2000); Sawhney & Eliashberg (1996) 提出的行為模型是屬於啟發性的架構，譬如說針對消費者行為的整合(對續集電影的喜愛)所建立的模型。而針對電影不同時期的分群方法，De Silva (1998) , Eliashberg et al. (2000) , Litman (1983) , Litman and Kohl (1989) , 提出的預測方法是在電影上映前，對於票房成功與否去進行預測。而Neelamegham and Chintagunta (1999) , Ravid (1999) Sawhney and Eliashberg (1996) 提出的方法則是在電影上映初期，當電影再戲院正式上映過一周，根據當周的票房收入對於接下來的票房做出預測。電影上映後所做的預測因為具有比較多可以參考的因素，包含影評跟觀眾的喜好程度，所以比較準確。

Ramesh and Dursun (2006)利用統計方法與類神經網路，建立模型並且進行比較，並且發現由類神經網路獲得較佳的結果。本研究將加入不同的輸入項目，加入可以量化星力指標的前三部演員作品票房的成績。還有評量導演的項目，包含導演前三部作品的票房還有影迷的評價。就此，可以更貼近演員和導演在當時期受歡迎的程度。並利用資料採礦技術在這個領域上，使用類神經網路和決策樹為工具，針對在美國上映的 1999~2003 年間，上映的 859 部電影作為本研究的資料庫，並以 2004~2008 年的電影資料作為驗證，最後再進行比較，哪種模型對於票房的預測有較佳的效果

2.4 決策樹 (Decision Tree)

決策樹是一種常用於預測模型的演算法，預測技術乃依據某一特定對象屬性，觀察其過去的行為或歷史資料，藉以推估其未來的值會是多少。 決策樹是同時提供分類和預測常用的方法。藉由一連串問題和規則將資料分類，藉由相似的型態來推測相同的結果，決策樹理論十分適合進行醫學預測及資料分析說明。由於決策樹是將資料依據不同的變數循序來產生分析結果，可藉決策樹分析方法來分析目標特質與異同點。決策樹是資料採礦在各產業最常用的方法之一，因為功能強大且相當受到歡迎的分類和預測工具，引人注意之處在於有規則，運用樹狀圖來表達資料規則的途徑，令可以用文字表達，也可以轉換為 SQL (Structured Query Language)之類的資料庫語言，讓落在特定類別的資料紀錄可被搜尋。決策樹分析模式主要演算法包含有 C5.0、CART(Classification and Regression Trees)分類樹、CHAID(CHi-squared Automatic Interaction Detector)卡方自動分互動偵，與QUEST (Quick, Unbiased , Efficient Statistical Tree)這四種演算法。(牛田一雄 , 高井免 , 木幕大輔 , 2006)

此篇研究中所採用的其中一項演算法為 ID3 演算法的延伸：C5.0。決策樹是以樹狀資料結構為基礎的分類分析方法，其主要的優點在於可產生易被人類瞭解與運用的決策法則。決策樹的建構是利用監督式學習法，從訓練範例集合中，以適當屬性挑選函數，從訓練範例的屬性中挑選出可用以建構決策樹根節點(Root Node)及內部節點(Internal Nodes)的屬性，用以建構決策樹並對訓練範例進行區分的處理。C5.0 是目前最常使用的決策樹分類分析法，C5.0 是學者 Quinlan 改進著名的 ID3 學習演算法而發展出的決策樹歸納學習法。

ID3 為一決策樹歸納技術，在構建決策樹過程中， ID3 以資訊獲利(Information Gain)為依據，選擇最佳的屬性當成決策樹的節點，使得所導致的決策樹為一最簡單(或接近最簡單)的決策樹。資訊獲利(Information Gain)是由以某一屬性為決策樹節點所產生的子決策樹之Entropy與物件集合的Entropy所決定假設訓練資料形成得集合S中有n 種類別Ci , $i = 1,2,3 \ldots n$ ，每個類別的資料個數如下公式(1)：

$$freq = (C_i, S) \tag{1}$$

$|S|$代表S 中所有資料的個數，因此各類別其資料出現機率可表示為

$$\frac{freq(Ci, S)}{|S|} \tag{2}$$

，因此根據訊息理論，各類別的資訊為

$$-\log_2(\frac{freq(Ci, S)}{|S|}) \tag{3}$$

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，為各個類別的資訊量乘上各個類別資料的機率總和為公式(4)：

$$-\sum_{i=1}^{n} (\frac{freq(Ci, S)}{|S|}) \log_2(\frac{freq(Ci, S)}{|S|}) \tag{4}$$

，根據Entropy(S)的計算方式，當集合S 根據某個屬性A 分割成多個子集合$S_1, S_2, S_3, \ldots S_m$時，其分割後所占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和：

$$entropy_x(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times S_i \tag{5}$$

因此集合S經由屬性X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，以公式(6)表示：

$$gain(X) = entropy(S) - entropy_x(S) \tag{6}$$

而ID3 學習系統選擇分類屬性的方法即計算所有屬性的Gain 值，並選擇其中Gain 值最大的做為分類屬性。決策樹以此屬性的屬性質分割成多個訓練子集合，形成多個樹。

各個子樹重複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，在分割成多個子樹直到不能再分為止。ID3 選擇分類屬性的方法對於一般學習問題已經有不錯的結果，但是當分類條件較偏向分出的子集合較多的屬性，其中最特殊的便是當集合S 分割後的子集合都只有一個資料時，其分割後的資訊量為零，因此所或的的資訊量最大。然而此種分割並沒有太大的意義。為了彌補這種缺點，Quinlan 在C4.5 中提出將Gain 正規化的方法以緩和分成過多子集合的效應。正規化的方法是利用將原有的Gain 值除以 Split infomation(X) 的值， 即

$$gainratio(S) = \frac{entropy(S)}{split\_entropy_x(S)} \tag{7}$$

其中

$$split\_entropy_x(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{|S|} \tag{8}$$

可代表集合透過屬性A 分割的子集合個數指標，分割後的子集合個數越多Split Information 的值就會越大，相對的Gain Ratio 的值就偏小。因此利用Split Information 使得C5.0學習系統改善了ID3分類偏向多子集合的缺點。

另外C5.0改進了ID3無法處理連續屬性和處理遺失屬性的問題。如果訓練案例T中的所有屬性值按照順序排好，表示如下：{v_1、v_2…v_m}。則可用下列代表新的分類屬性

$$\frac{v_i + v_{i+1}}{2} \tag{9}$$

至於處理遺失屬性資料的問題，則是採用下列公式：

$$gain(X) = \frac{K}{T} \times (entropy(S) - entropy_x(S)) \tag{10}$$

T代表了Training Set，而K代表了從T中扣除掉有遺失屬性資料的集合。所以說就是算出正常資料的比例再乘以Gain(X)。兩種演算法比較如表 1. ID3與C5.0比較表格

表 1. ID3 與 C5.0 比較表格

演算法	分割規則	修剪規則	處裡連續屬性
ID3	Entropy、Gain ratio	錯誤預估率	否
C5.0	Gain Ratio	錯誤預估率	可

2.5 類神經網路

葉怡成(1997)再應用類神經網路一書中介紹，類神經網路在控制領域中，若要精確的分析輸入與輸出之間的關係，則必須將系統藉由數學的方式作成模型，類神經網路的一個優點在於並不需要瞭解系統的數學模型為何，而直接以類神經網路取代系統的模型，一樣可以得到輸入與輸出之間的關係。類神經網路利用其平行分散處理、學習及引申的能力，被廣泛應用於各類型的分類以及預測問題上，例如醫學監測、流量控制、生態環境、電力負載、旅客行為、財經資訊，以及時間序列模擬等問題。

類神經網路要能正確運作，則必須透過訓練的方式，讓類神經網路反覆的學習，直到每個輸入項都能正確的對應到所需要的輸出，因此在類神經網路學習前，必須建立出一個訓練樣本，使類神經網路在學習的過程中有一個參考，訓練樣本的建立來自於實際系統輸入與輸出或是以往的經驗，類神經網路的工作性能與訓練樣本友直接的關係，若訓練樣本不正確、太少或是太相似，類神經網路的工作區間與能力將大打折扣。

類神經網路的優點：

1. 類神經網路可以建構非線性的模型，透過訓練法則的調整，可以更有彈性的模擬真實世界資料的複雜映射關係，應用領域相當的廣泛。
2. 類神經網路有良好的推廣性，對於未知的輸入亦可得到正確的輸出。
3. 類神經網路可以接受不同種類的變數作為輸入，適應性強。
4. 類神經網路具有優異的容錯能力，因其平行分散架構使然。

類神經網路之模型十分繁多，在大多的研究中，主要是運用倒傳遞類神經網路模型 (Back-Propagation Neural Network)及自我組織映射類神經網路模型(Self-Organizing Map)兩種。本研究所使用的是Back-Propagation Neural Network (BPN)模型，它是一種監督式學習的類神經網路，屬於層狀前饋式網路架構(Layered Feed-Forward Network)的非線性轉換函數為雙彎曲函數(Sigmoid Function)和雙曲線正切函數(Tan H Function)，適合用來做預測和診斷，是最被廣為應用的類神經網路。BPN基本元來乃是利用最陡坡降法(The Gradient Steepest Descent Method)的觀念，採用倒傳遞式學習演算法(Back-Propagation Learning Algorithm)，將錯誤的訊號以回饋方式修正網路上的連結權重，使誤差含數最小化之下，調整網路權值成為一最適合權重，使演算的輸出值能夠最接近目標的輸出值。圖 1為倒傳遞類神經網路模型 (BPN)的結構圖。

BPN 因為具有能夠處裡複雜的非線性數值，而且具有隱藏層，可以處表現輸入處理單位元間的交互影響。所以研究將使用 BPN 作為探索電影各特徵值跟其票房表現的工具。

本研究預測模式不使用統計模型的原因是，不用去設置參數建立線性的模型，憑藉著決策樹和類神經的網路輸入，可以直接找到輸入項所對應的輸出，而且可以處裡連續型變數的輸入項，Samesh and Dursun (2006)的文章中，它的輸入項只只有一項是連續型變數，而本研究有五項的連續型變數，所以更適合用決策樹和類神經網路去做預測。

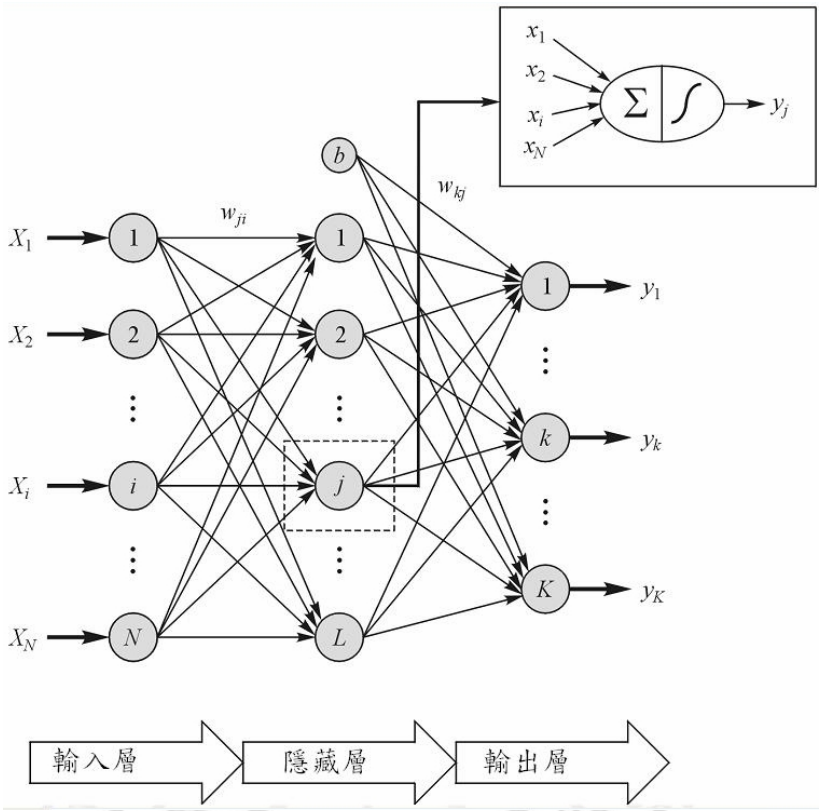

圖 1 倒傳遞類神經網路模型 (BPN)結構圖

三、研究方法

3.1 研究架構

本研究架構如圖 2所示，首先確定研究動機，是要建立起有效預測電影票房的模型。採用資料採礦的技術，並閱讀有關電影票房、預測的文獻。決定研究的範圍是以近十年的北美上映電影的資料作為本研究的Database。蒐集並整理資料，並且建立起可供Clementine執行的資料庫，以資料採礦的兩種方法分別建立模型並且測試模型的精準度，回顧結果探討哪個模型具有較好的預測準確率。

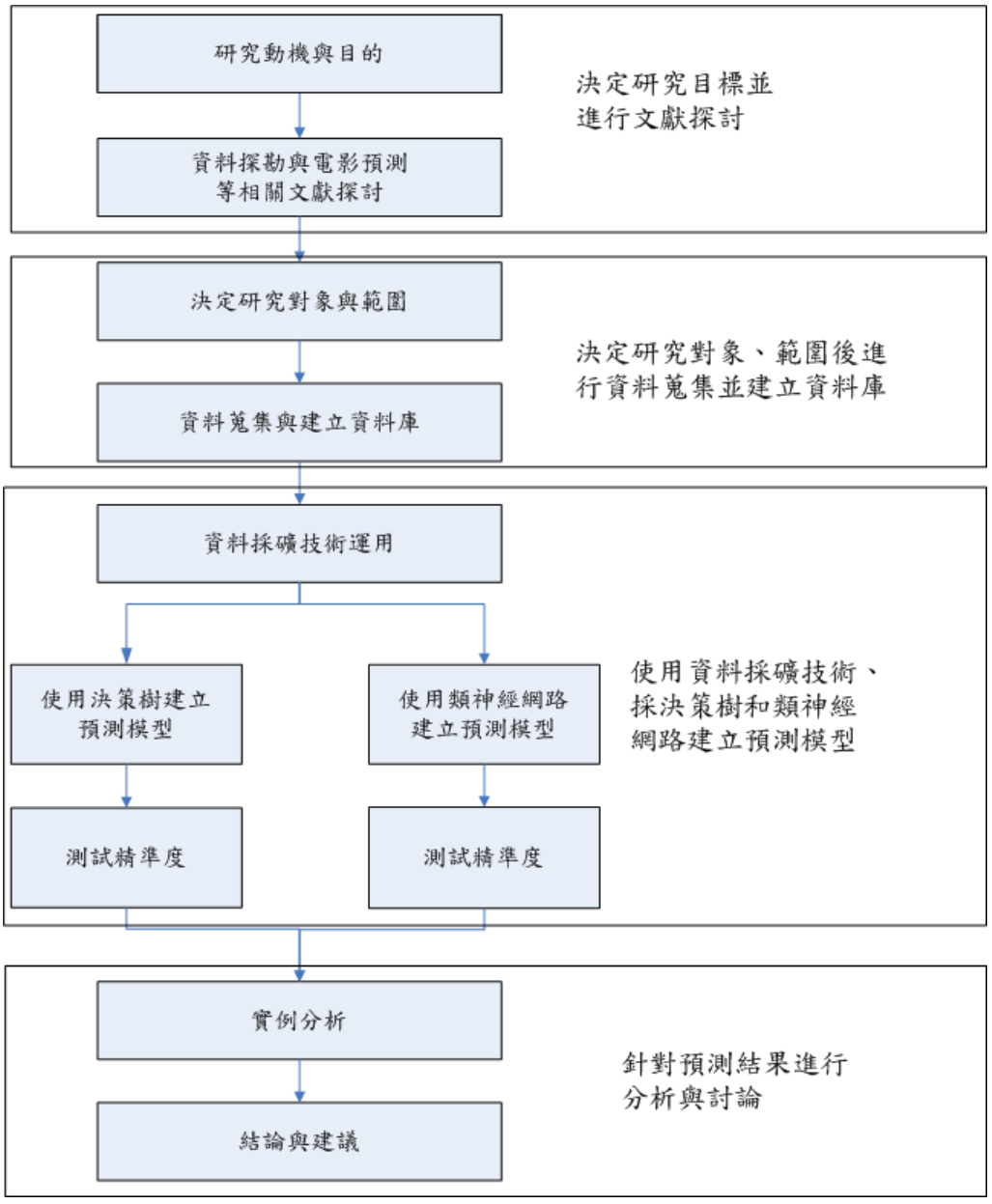

圖 2.研究架構圖

3.2 個案研究方法

資料採礦的應用很廣泛，適用在各類型的產業，面對不同的環境下，會遇到不一樣的問題，譬如說在面對客戶習性問題，要去了解客戶習慣購買的產品？客戶的特徵屬性為何？客戶回流率？這些特性都可以用資料採礦來協助分析，找尋解答。若是能很完整了解原始資料，會使資料採礦分析後的結果更加完整。

本研究是針對美國近十年上映電影的票房資料作為研究對象，使Clementine 10.1 做為分析的工具。本研究分析流程將採用上一章所介紹，跨產業資料採礦標準作業程序CRISP-DM來進行電影票房賣座因子的討論與分析(圖 3)。在資料了解階段，首先將進行原始資料的收集，理解收集到的原始資料，標記資料的問題，去發現可能有的隱藏資訊去做假設。

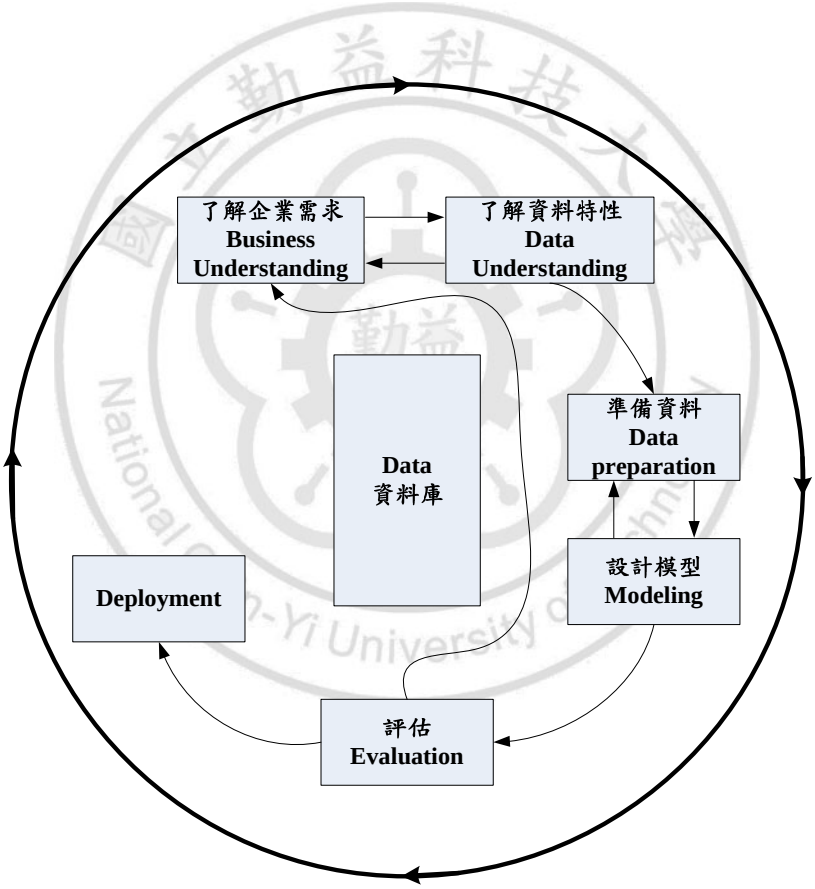

圖 3. CRISP-DM 流程圖

3.2.1 了解企業需求 (Business Understanding)

開始挖掘知識之前，最先同時也是最重要的要求就是了解資料和業務問題。如果事先沒有這種瞭解，不管用任何演算法，無論方法有多複雜，即使夠提供有價值的結果，也難以使人信賴。缺少了這些背景知識，就沒辦法明確定義要解決的問題，不能為資料採礦而準備資料，也很難正確的解釋所得到的結果。要想充

分發揮資料採礦的價值，必須要對研究目標有一個清晰明確的定義，即決定到底想要什麼目標。譬如說想提高直接郵件推銷的用戶回應時，想做的可能是"提高用戶回應率"，也可能是"提高一次用戶回應的價值"，要解決這兩個問題而建立的模型幾乎是完全不同的，必須做出選擇。有效的問題定義還應該包含一個對知識發現專案得到結果進行衡量的標準。

1.決定商業目標：理解電影工業的產業背景及未來發展，仔細研究電影工業的需求。一般來說一部電影上映之後，製作電影的片商就是期待它能在市場上獲利。此目標確定以後，就可以從電影上映的資料中採礦出有用的規則，能在初始製作階段即能找出將會影響結果的重要特徵屬性，匯入本研究所建立的模型，就可根據規則預測該電影的總收益，若是結果不理想即可在製作的時候就做修正，像是增加製作成本，改變上映日期等等。
2.決定資料採礦目標：在訂定目標後，本研究將以資料採礦技術來達成目標，研究是否在電影產業中是否存在賣座電影的規則？如何找出這樣特殊的行為模式？

3.2.2 了解資料特性 (Data Understanding)

資料是資料採礦的原始材料，在 CRISP-DM 的此階段，強調了需要清楚資料的來源是什麼，有什麼樣的特徵。這個過程包括資收集原始資料、描述資料、探索資料、及證實資料的質量。在此階段對於電影工業，要有基本的認識，取得資料前要先進行初步規劃、確定商業目標等計畫。
1.初步收集資料：從網站上的公開資訊中 IMDb.com 和 boxoffice-mojo.com 收集近十年的在美國上映的票房資訊，包括其中的相關屬性，對資料內容有瞭解與認識並做處理。本研究取得包含電影的票房、演員、還有成本資料等以進行分析。

2.描述資料：由電影工業中，定義出電影相關屬性資料。資料一共蒐集到 1782 筆的資料長度，以票房總收入做為決策屬性。在每筆電影資料內(請見表 2.原始收集資料證明(以 1999 年 10 筆為例) 總筆數 1782)，包含有 1.年份(Year)、2.電影名稱(Film)、3.上映月份(Month)、4.製作成本(Budget)、5.導演(Director)、6.導演前三部電影評價(Last 3 IMDb Grade Average)、7.導演前三部電影票房(Last 3 Box-office Average)、8.首週上映廳數(Open Screens)、9.類型(Genre)、10.是否為續集電影(Sequel)、11.分級(Rating)、12.演員A(Performer A)、13.演員A前三部電影票房(A Performer Last 3)、14.演員B(Performer B)、15.演員B前三部電影票房(B Performer Last 3)、16.電影總票房收入(Box office)，此表格是以美金做為單位。本研究所用的變數是參考過去對於電影票房預測的研究(Elberse and Eliashberg, 2002; Litman, 1983; Litman and Kohl, 1989; Neelamegham and Chintagunta, 1999; Ravid,1999; Sawhney and Eliashberg, 1996; Sochay, 1994)。

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Year	**Film**	**Month**	**Budget**	**Director**	**Last 3 IMDB Grade Average**	**Last 3 Boxoffice Average**	**Opening Screens**	**Genre**	**Sequel**	**Rating**	**Performer A**	**A Last 3**	**Performer B**	**B Last 3**	**Box office**
1999	AT FIRST SIGHT	January	60,000,000	Irwin Winkler	5.9	22,136,973	1,814	Romance	N	PG-13	Val Kilmer	67,042,389	Mira Sorvino	24,542,278	22,365,133
1999	VIRUS	January	75,000,000	John Bruno	0	0	2,018	Sci-Fi	N	R	Jamie Lee Curtis	21,474,119	0	0	14,036,005
1999	In Dreams	January	30,000,000	Neil Jordan	7.1	37,865,613	1,670	Thriller	N	R	Annette Bening	47,895,526	0	0	12,017,369
1999	VARSITY BLUES	January	16,000,000	Brian Robbins	5.5	13,158,679	2,121	Drama	N	R	Jon Voight	52,871,014	James Van Der Beek	0	52,894,169
1999	GLORIA	January	30,000,000	Sidney Lumet	5.8	10,970,453	1,527	Drama	N	R	Sharon Stone	43,422,203	0	0	4,197,729
1999	SHE'S ALL THAT	January	10,000,000	Robert Iscove	0	0	2,222	Romantic Comedy	N	PG-13	Freddie Prinze Jr.	56,104,201	Rachael Leigh Cook	0	63,366,989
1999	PAYBACK	February	90,000,000	Brian Helgeland	0	0	2,729	Drama	N	R	Mel Gibson	114,088,374	0	0	81,526,121
1999	Simply Irresistible	February	6,000,000	Mark Tarlov	0	0	1,642	Romantic Comedy	N	PG-13	Sarah Michelle Gellar	72,219,395	0	0	4,398,989
1999	BLAST THE PAST	February	35,000,000	Hugh Wilson	5.8	26,500,000	2,542	Romantic Comedy	N	PG-13	Brendan Fraser	55,826,645	0	40,537,403	26,511,114
1999	Wild Wild West	July	170,000,000	Barry Sonnenfeld	6.7	123,714,552	3,342	Comedy	N	PG-13	Will Smith	222,620,438	Kevin Kline	29,245,397	113,804,681

表 2.原始收集資料證明(以 1999 年 10 筆為例) 總筆數 1782

在圖 4 中所包含的資訊包含網頁中間的總票房(Domestic Total Gross: 113,804,681)，上映日期(Release Date: June 30 ,1999)，電影類型(Genre :Western Comedy)，分級(MPAA Rating: PG-13)，製作成本(Production Budget:$170 million)，導演(Barry Sonnenfeld)，演員(Will Smith and Kevin Kline)。圖 5則表示的是該電影在IMDb網站的頁面，可以看到觀眾對於此電影的評分為 4.3 (UserRating:4.3/10)。

製作成本(Budget)：

製作一部電影所花費的總成本，包含演員的薪水還有，劇組人員的花費，特效處理費用，還有宣傳等花費。此輸入項是連續型變數。

導演前三部電影評價(Last 3 IMDb Grade Average)：

IMDb(International Movie Data Base)創建於 1990 年, 是當今世界最大的網上電影資料庫，收錄近 20 萬部影片，40 萬演員和 4 萬導演的信息。它有一個評分系統，由影迷自己來打分，平均每月有高達 2000 萬電影愛好者造訪，所以 IMDb 被認為是權威的影片評分，很多電影雜誌都應用 IMDb 的評論和評分，具有很大參考價值，分數高也代表影迷對於電影的一種肯定。評分的標準是從 1~10 分。故本研究用此來代表觀眾對於一部電影的喜好程度

導演是決定電影風格的很重要的因素，一個導演會不會說故事，對電影重大的影響。然而電影評價的好壞與票房的高低卻不一定相關，像是知名導演大衛芬奇(David Fanchy)的鬥陣俱樂部(Fight Club, 1999)，在 IMDb 上獲得的 8.8 分，歷史排名第 16，但是票房卻只有三千七百萬美元。而由 Will Smith 主演的飆風戰警(Wild Wild West)，總票房在美國超過一億一千萬元，可是在 IMDb 的評價卻只有 4.3，可見得這部電影雖然賣座，但是觀眾看過以後並不喜歡。

選用前三部電影的平均，是因為希望這個數字能隨著時間有所變化，如果一位導演連續幾部電影都不受好評，也會影響影迷對於導演的信任。像是曾經拍過膾炙人口的靈異第六感(The Sixth Sense , 1999)的印度籍導演 M. Night Syamalan，曾經被人形容才華洋溢的出世奇才，靈異第六感在 IMDb 獲得 8.2 分的高評價，隔年的靈異象限(Signs ,2002)也有 7.2 不錯的成績，近年的幾部作品陰森林(The Village , 2002)，水中的女人(Lady in the Water , 2006)，破天荒(The Happening , 2008)成績分別是 6.6、5.8、5.2，由於他的作品的品質每況愈下，觀眾也開始對於這位導演產生懷疑，而他作品也不再被觀眾信任。故以前三部電影的平均，就是希望藉此衡量觀眾近期對於這位導演的評價。此輸入項是連續型變數。

導演前三部電影票房(Last 3 Boxoffice Average)：

然而一部電影的評價，不能完全影響到電影的收入，像是導演 Michael Bay 擅長拍攝大場面的動作片，像是變形金剛(Transformers , 2007)，珍珠港(Pearl Harbor , 2001)，世界末日(Armageddon , 1998)，在他的電影裡總是充滿爆破和火花，雖然沒有太多的藝術價值，但是觀眾卻很享受在聲光效果的刺激，進戲院觀賞他的電影，他的每部作品幾乎都是破億美元的票房收入。所以衡量一位導演應該也該包括他對於票房的影響力。與導演評價一樣選用前三部的成績，做為近期表現的衡量指標。此輸入項是連續型變數。

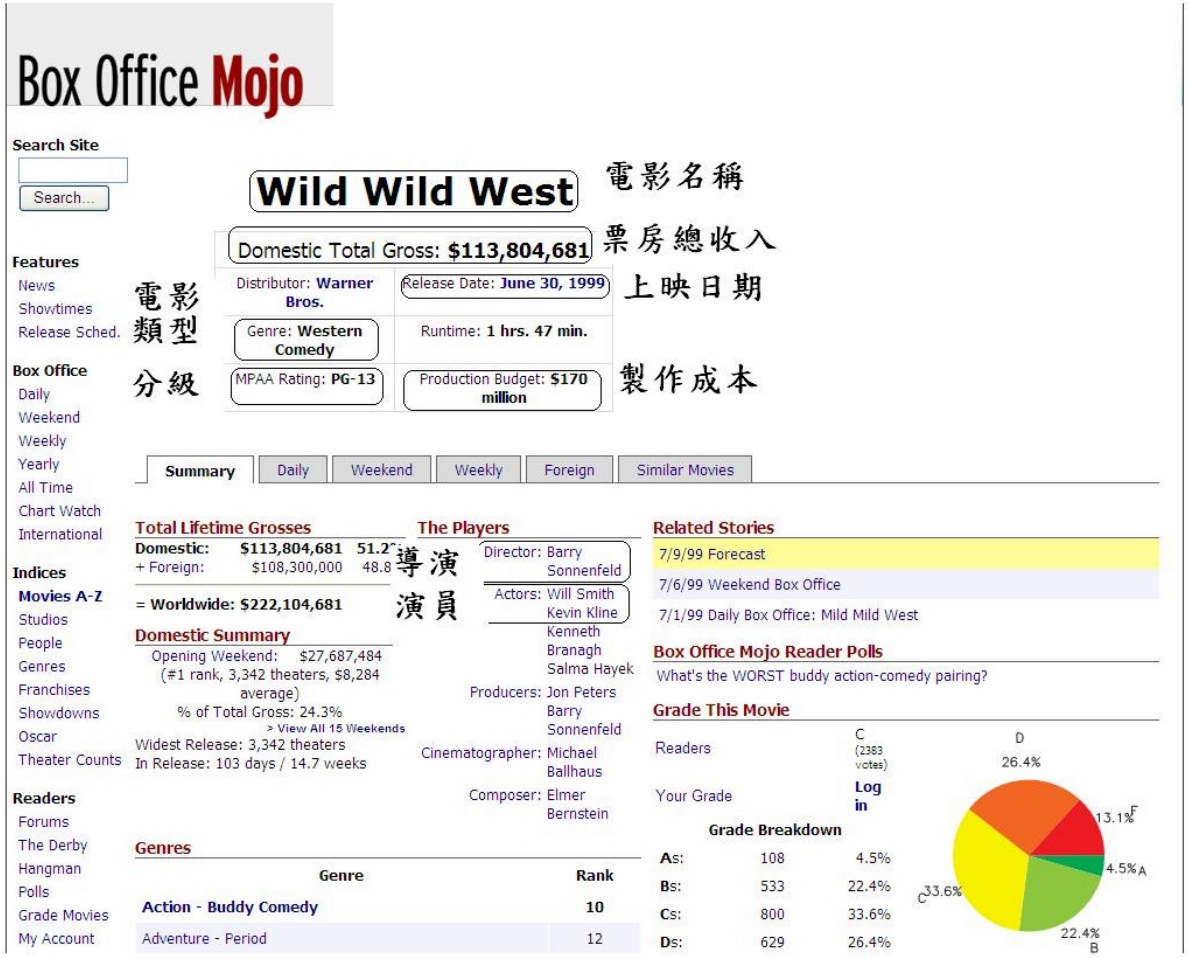

圖 4.BOXOFFICE MOJO網頁資訊（http://www.boxofficemojo.com/movies/?id=wildwildwest.htm）

圖 5.IMDb網頁資訊 (http://www.imdb.com/title/tt0120891/)

演員前三部電影票房(A Performer Last 3)：

一部電影有什麼樣的演員，也是是否能吸引觀眾進場觀賞很重要的因素，一個演員對於觀眾是否有吸引力，本研究將用該演員拍攝該電影前，上映的前三部電影票房成績作為依據，作為衡量該演員近期星力指數的指標。如同導演，觀眾對於同一位演員的吸引力也會隨著他的作品好壞而有所改變。

本研究取該電影最具代表性，或是該電影主打的主角作為代表，如表 2.所示，以電影 Wild Wild West 為例，演員 A 是 Will Smith，他前三部作品的成績是 222,620,438 美金，演員 B 是 Kevin Kline，該演員前三部作品的成績是 29,245,397。若是該電影只有一個代表性演員，則只取演員 A，演員 B 則以 0 來代表，如表 12 中的電影 PAYBACK，Mal Gibson 是唯一主打的演員，所以就只以他的成績來代表演員 A 的項目，演員 B 因為沒有演員作為代表，所以以 0 表示。若是該電影只有新進演員的話也以 0 來代表該位置，同樣的若是紀錄片的話也是以 0 代表演員的欄位。此二輸入項是連續型變數。

首周上映廳數(Open Screens)：

過去的研究發現，前兩周上映的票房成績大約占一部電影收入的 25%，所以本研究將採用電影首周上映的廳數，來判斷上映時的規模是否會影響到票房的收入。雖然每間電影院的影廳，不見得座位都是相同，但是上映規模的廳數，還是可以決定可以進場觀眾數目的多寡。一部電影的上映，片商該採取什麼策略，到底是小規模上映贏取口碑？還是全國同時間大規模上映？上映的規模應該大到怎麼樣的數字？如表 12 第八行 Opening Screens 所示，影廳的數目介於 1~4000 之間的連續型變數。

分級(Rating)：

電影分級制度是以美國電影協會(MPAA)的標準去區分，以內容的暴力程度、裸露尺度、成人語言，區分為 G (普遍級)，PG (建議家長陪同觀賞) , PG-13(13 歲以上得以觀賞) 和 R (17 歲以上得以觀賞) ，這 4 個等級。

類型(Genre)：

本研究將電影內容區分為 13 個類別，分別為 Action(動作), Comedy(喜劇), Drama(劇情), Action Comedy(動作喜劇), Animation(動畫), Family(家庭),Horror(恐怖), Romantic Comedy(浪漫喜劇), Romance(愛情), Sci-Fi(科幻),Thriller(驚悚), War(戰爭), Music(音樂)。

是否為續集電影(Sequel)：

使用 Yes 表示為續集電影，No 表示否。藉此資訊來了解是否續集電影對於電影票房是否會有造成影響。

3.2.3 準備資料 (Data Preparation)

將資料分類好以後，需要準備用於採礦的資料。準備過程包含了選擇，清理，重構，整合資料。

1.選擇資料：以資料採礦為目標作為選擇分析資料的準則，考慮所選擇及不必要的屬性。本研究挑選出電影的上映月份的競爭性、製作成本、導演前三部電影的評價、導演前三部作品的平均票房、上映時的廳數、電影的類型、是否為續集電影、電影分級、演員A的前三部電影票房、演員B的前三部票房及該電影最後的票房收益。如表 3 原始資料經過轉換後表格所示。

2.清理資料：為配合分析所需格式，藉由 Excel 軟體輔助建立此資料庫，展開

表 3 原始資料經過轉換後表格

No.	Month	製作成本	導演前三部電影在 IMDb 評價	導演前三部電影平均票房	上映廳數	電影型態	續集電影	分級	演員 A 前三集電影票房	演員 B 前三集電影票房	票房總收益
01	Low	5,879,254	5.9	22,136,973	1,814	Romance	N	PG-13	67,042,389	24,542,278	O4
02	Low	8,668,740	0	0	2,018	Sci-Fi	N	R	21,474,119	0	O3
03	Low	17,154,061	7.1	37,865,613	1,670	Thriller	N	R	47,895,526	0	O3
04	Low	17,689,177	5.5	13,158,679	2,121	Drama	N	R	52,871,014	0	O5
05	Low	44,125,987	5.8	10,970,453	1,527	Drama	N	R	43,422,203	0	O2
06	Low	4,979,469	0	0	2,222	Romantic Comedy	N	PG-13	56,104,201	0	O5
07	Low	21,882,043	0	0	2,729	Drama	N	R	114,088,374	0	O6
08	Low	5,879,254	0	0	1,642	Romantic Comedy	N	PG-13	72,219,395	0	O2
09	Low	8,668,740	5.8	26,500,000	2,542	Romantic Comedy	N	PG-13	55,826,645	40,537,403	O4
10	Low	17,154,061	5.1	15,161,047	2,275	Sci-Fi	N	PG-13	28,064,325	54,054,418	O4

第二階段資料處理作業，過濾、修正對應的資料類型。移除過於繁複的導演、演員姓名、電影名稱。

3.建立資料：對於上映月，定義 6 月,11 月為高競爭性(High)，5 月,7 月,12 月為中競爭性(Middle)，其他剩下的月份為低競爭性(Low)，憑此來判斷上映月份的競爭性對票房收入是否會有影響。(Krider and Weinberg ,1998; Litman and Kohl ,1989 ,Radas and Shugan, 1998; Sochay , 1994)。輸出變數為該電影最終票房收益，本研究依其票分區分為九個等級，分別是票房小於一百萬美金，以 O1 來表示；一百萬至一千萬美金，以 O2 表示；一千萬至兩千萬美金，以 O3 表示；兩千萬至四千萬美金，以 O4 表示；四千萬至六千五百萬美金，以 O5 表示；六千五百萬至一億美金以，O6 表示；一億至一億五千萬美金，以 O7 表示，一億五千萬至兩億美金，以 O8 表示；超過兩億美金以 O9 表示(Ramesh and Dursun , 2006)。

4.格式化資料：對於資料製作格式化的轉換，配合分析工具的使用，但並改變初始資料表示的意義，利用工具將資料一一編排，方便建構資料採礦的模型。

3.2.4 設計模型 (Modeling)

對設計模型來說，要記住最重要的事是它是一個反復的過程。需要仔細考察不同的模型以判斷哪個模型對該商業問題最有效用。在尋找好的模型的過程中學到的東西會啓發研究者修改資料，甚至改變最初對問題的定義。這部份包括了選擇模型的技巧、安排測試計劃、建模與模型評估。

1.選擇模型技術：本研究採用決策樹 C5.0 還有類神經網路 BNP 來做為執行工具。
2.規劃測試：實際建立模型前，先將之前的資料有效的統整，再將其匯入 Clementine 10.1 資料採礦分析軟體中，產生一套檢驗的程序或是機制，以確保模型的品質和有效性。
3.建立模型：經由資料採礦軟體 Clementine 10.1 來分析，本研究利用來決策樹以及類神經網路來進行分析各群組的特徵屬性，接著使用驗證資料來做更進一步的分析。
4.模型選擇確定：這個階段目標就是以商業角度立場，經由分析的資料結果，對兩個模型進行評估。

訓練和測試資料採礦模型需要把資料至少分成兩個部分：一個用於模型訓練，另一個用於模型測試。如果使用不同的訓練和測試集，那麼模型的準確度就很難使人信服。用訓練集將模型建立出來之後，就可以先在測試集資料先試驗一次，此模型在測試集上的預測準確度就是一個很好的方向，它說明如果將來與訓練集和測試集類似的資料用此模型預測時，正確的百分比會有多大。這並不能保證模型的正確性，只是說相似的資料用此模型會得出相似的結果。

3.2.5 評估 (Evaluation)

選擇了模型，接著就應準備好以結果去對商業目標作為評估。這階段要進行評估結果、回顧資料採礦的過程、確認下一個步驟。

1.評估結果：此步驟是為了檢驗分析結果有無符合資料採礦鎖定地的目標，是否能符合此研究要預測票房收益的目的，例如從決策樹分析後，是否能準確預測製作成本是否會直接影響到票房的好壞。

2.檢視流程：針對採礦的各項步驟進行回顧，確定是否會影響到結果的目標重要變數，是否已經匯入到分析結果中，或是匯入的資料太複雜，造成資料分散沒辦法取得規則。

3.決定下一個步驟：根據評估結果與回顧過程決定是否要繼續作業或是結束。

3.2.6 部署(Deployment)

最後這個階段是要把從此專案得到新知識，應用到實際電影票房預測的商業運作過程，進而解決研究一開始所面對的問題，是否能夠達到預測的效果。

四、個案研究

4.1 問題描述

近年來，美國電影市場的總收益，一直都以每年 3.4%的速率穩定的成長，在 2008 年的年度總票房已經來到了 96 億 3 千萬美元，12 個主要的製片公司，無不想要在這競爭激烈的市場佔有更好的收益，為了追求更高的票房，吸引觀眾進場，在製作電影常常毫不手軟的砸下高額的製作費，有些電影能如期吸引眾多觀眾進場，有些卻出乎意料的市場反應冷淡，造成製作公司蒙受巨大損失。本研究希望從過去十年的電影資料中，從繁複的資訊當中利用資料採礦技術找到規則，避免錯誤的投資。

本研究資料庫，是包含了自 1999 年至 2008 年十年之中的 1782 部電影，並將之區分為兩個部份，1999 年至 2003 年這五年當中的 859 部電影是作為模型學習的樣本，2004 年至 2008 年這五年中的 923 部電影則是用來做為模型驗證的樣本。其中的資料包含了競爭性、製作成本、電影類型、導演前三部電影的評價、導演前三部作品的平均票房、上映時的廳數、是否為續集電影、電影分級、演員 A 的前三部電影票房、演員 B 的前三部電影的票房。希望藉由這些資料能了解電影票房落點的模式，找出規則，讓製片公司的投資能更有效率。

4.2 資料理解

利用資料採礦中的決策樹、倒傳遞類神經網路來建立模型，並利用 Clementine 10.1 來進行各項的資料研究。首先要先了解所搜集到的電影資料的基本型態及其屬性特徵。

表 4.每部電影中的各項特徵屬性 是每部電影的特徵屬性，資料筆數一共是 1782 筆，資料蒐集的時間是從 1999 年至 2008 年十年間在美國上映的電影。其中將前五年的 859 筆資料作為模型學習的樣本，後五年的 923 筆資料做為測試模型精準度的樣本。前八項是輸入項，其中上映時期的競爭性 (Competitiveness) 、電影類型 (Genre)、是否為續集電影 (Sequel)、電影分級 (Rating)這四項分別是離散變數，而製作成本 (Budget)、導演前三部電影的評價(Director Last 3 IMDb Grade Average)、導演前三部電影的平均票房(Director Last 3 Box-office Average)、首周上映時的廳數(Screen)則是連續型變數。No.9 是決策變數，也是輸出項目，為離散變數。

表 4.每部電影中的各項特徵屬性

No.	獨立變數的名稱	變數數目	變數值
1	競爭性 (Competitiveness)	3	高,中,低
2	製作成本 (budget)		連續型變數
3	電影類型 (Genre)	13	動作,喜劇,劇情,動作喜劇,動畫,家庭,恐怖,浪漫喜劇,愛情,科幻,驚悚,戰爭,音樂
4	導演前三部電影的評價 (Director Last 3 IMDb Grade Avg.)		連續型變數(1~10)
5	導演前三部電影的平均票房 (Director Last 3 boxoffice Avg.)		連續型變數
6	首周上映時的廳數 (Screen)		連續型變數
7	是否為續集電影 (Sequel)	2	Yes , No
8	電影分級 (Rating)	4	G , PG , PG-13 , R
9	票房收入 (box-office)	9	O1,O2, O3, O4, O5, O6, O7, O8, O9

4.3 執行結果

將學習資料 823 筆，使用Clementine，以決策樹與類神經網路建立起電影賣座票房模型。得到了圖 6的規則和圖 7的樹狀圖。

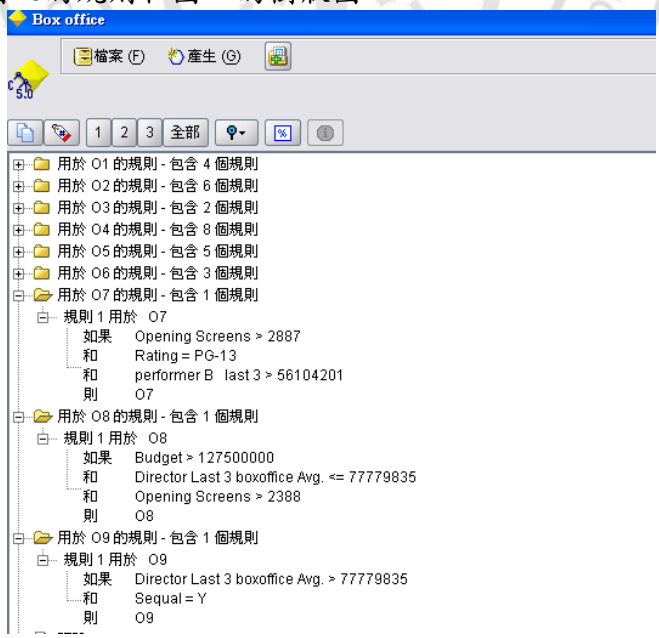

圖 6.以決策樹 C 5.0 所找出的規則

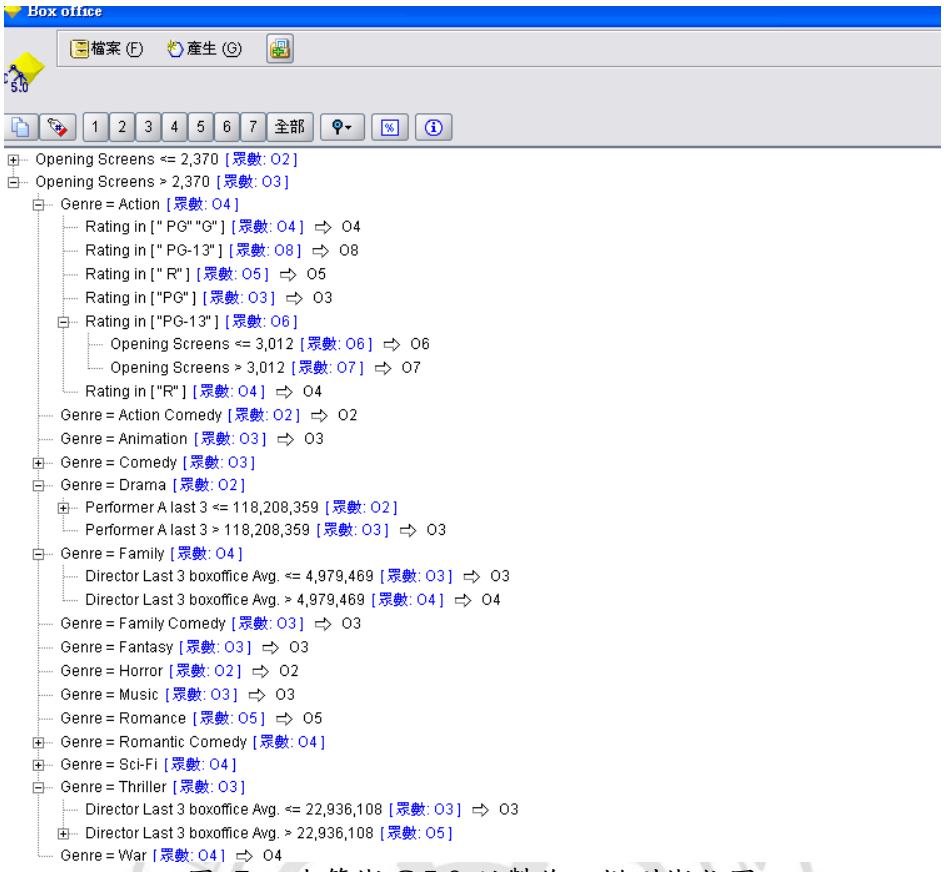

圖 7 以決策樹 C 5.0 所製作之模型樹狀圖

圖 8表示在Clementine 10.1 軟體中建立類神經網路模型，並以決策樹表現規則的執行流程圖。類神經網路中，隱藏層中的規則，利用執行後的結果，以決策樹CART去做表示，利用決策樹的表示方法去找尋規則。

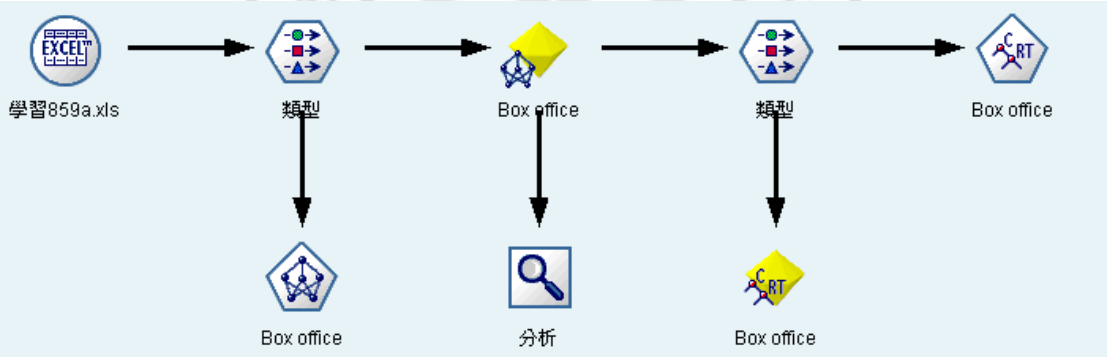

圖 8. 以類神經網路建立模型的流程

可以從中歸納出表 5.由決策樹所歸納出來的規則中，幾項較顯著的規則，如表 5 所說明

表 5.由決策樹所歸納出來的規則

	規則	落點	舉例說明
1	If 上映廳數小於 1561 and 分級制度為 R and 製作成本小於 36,000,000 時	O1	PUNKS (2001)
2	If 導演前三部票房小於 77,779,835 and 上映廳數大於 2,388 and 演員 A 前三部作品大於 129,734,803	O3	21 Grams (2003)靈魂的重量 The Life of David Gale (2003) 鐵案疑雲
3	If 競爭性低 and 製作成本小於 48,000,000 and 分級制度 R and 製作成本大於 27,000,000 and 非續集電影 and 上映廳數大於 2,388	O4	Valentine (2001) 致命砍人節 The Pianist (2002) 鋼琴師
4	If 上映廳數小於 1,722 and 製作成本大於 57,000,000 and 分級制度 PG	O5	Message in a Bottle (1999) 瓶中信 Shanghai Noon (2000) 西域威龍
5	If 當導演前三部票房大於 82,314,371 and 製作成本大於 105,000,000 and 非續集電影	O6	Master and Commander :The Far Side of the World (2003) 怒海爭鋒
6	If 導演前三部票房小於 77,779,835 and 上映廳數多於 2,388 and 製作成本大於 127,500,000	O8	The Perfect Storm (2000) 天搖地動
7	If 當導演前三部票房大於 77,779,835 and 續集電影	O9	The Lord of the Rings: The Return of the King (2003) 魔戒 3 :王者歸來 Star Wars: Episode II - Attack of the Clones (2002) 星戰二部曲:複製人進攻

表 6.各輸入項的重要性

各輸入項的重要性	
分級	0.269
上映時的競爭力	0.149
分級	0.143
續集電影	0.129
上映廳數	0.086
導演前 3 部電影在 IMDb 的評價	0.066
演員 A 前三部電影票房	0.044
導演前三部電影票房	0.045
製作成本	0.045
演員 B 前三部電影票房	0.029

表 6.各輸入項的重要性是以類神經網路製作的模型中，分析得到的各輸入項對於模型的重要性，其中最重要的項目為電影的類型，原因是電影類型是觀眾決定是否要進場欣賞的主要因素，像是動作喜劇或是浪漫愛情類型的電影比較容易吸引大部分想要進場放鬆心情，而像是記錄片類性的電影就只能吸收到想要觀賞特殊題材的觀眾。

其次是上映時的競爭力(3.2.2)，而一部電影選定上映日期也是很重要，在競爭力高的時候上映的電影，賣座的機率較高，因為那時候觀眾進場的人數多，可是相對的上映的電影也是很多，是兵家必爭之地，若是在製作前期的預測，結果不理想，製作時候的策略就應該改變預計上映的日期，選擇別的月份。

分級制度，一部電影的分級制度，決定了進場觀眾年齡層的大小，一部普通級的電影，能吸引到的客層就是比較寬廣。是否是為續集電影，近年來續集電影有如雨後春筍，因為觀眾對於續集電影趨之若鶩，若是前部作品受到歡迎，續集電影就是票房的保證。

演員 B 是最不重要的輸入項，代表觀眾還是重視電影最受注目的影星。而成本則是因為不同的電影有不同的製作策略，譬如說小成本製作的電影預期的收入，如一千萬製作的電影，預期可能只要能達到兩千萬就算是成功，所以製作成本不是會影響到電影是否賣座的重要因素。

將已經建立的模型，代入驗證用的 923 筆資料去做測試。在本研究將所有的電影最後的票房收入分成九個層級，從好的賣座電影 O9，到最差的票房最差的 O1，透過建立的模型去做預測，如果得到的落點跟實際電影票房落點一樣，那就把它定義為命中。若是預測落點是與實際落點相同或是相鄰的話，就將它定義為鄰近，如預測電影票房將會落在 O6，實際票房是落在 O5、O7 或是在 O6 上，則 O5、O6、O7 都稱之為鄰近。表 7、表 8 分別以決策樹與類神經網路所預測出來的結果，行是代表電影實際賣座票房落點的分群，列的數字代表的是模型所預測落點的分群。相交的部份就是預測命中實際的分群。表格下方則是預測命中率，還有預測相鄰分群的比率。

決策樹C5.0 所建立的模型(表 7 以決策樹C 5.0 所預測之結果)所得到的預測結果，共有 345 部電影被完全命中，完全命中的機率是 37.3%。而一共有 707 部電影預測結果為鄰近，預測命中鄰近的機率為 72.2%。而以類神經網路所建立的模型(表

8. 以類神經網路(BPN)所預測之結果)，所得到的預測結果，一共有 322 部電影被完全命中，完全命中的機率是 34.9%。，一共有 680 部電影是為鄰近，而預測命中鄰近的命中率為 73.7%。

表 9則是將兩種預測方法做比較，針對命中率的平均數、標準差、中位數做比較，可以發現決策樹的準確度較類神經網路的模型為精準。但是標準差類神經網路模型就略勝決策樹的模型。

表 7 以決策樹 C 5.0 所預測之結果

		實際分群									平均
		O1	**O2**	**O3**	**O4**	**O5**	**O6**	**O7**	**O8**	**O9**	
預測落點	**O1**	**9**	7	10	7	2	0	1	0	0	
	O2	4	**35**	29	11	7	1	3	1	0	
	O3	4	19	**71**	49	15	7	4	1	1	
	O4	1	6	36	**79**	29	8	2	0	2	
	O5	0	1	20	34	**65**	36	6	4	1	
	O6	3	5	9	20	36	**32**	13	1	0	
	O7	2	3	5	6	9	10	**29**	3	9	
	O8	0	1	3	0	3	12	10	**11**	19	
	O9	0	0	2	1	0	7	5	9	**14**	
	總數	23	77	185	207	166	113	73	30	46	
	命中	0.391	0.455	0.384	0.381	0.391	0.283	0.397	0.367	0.304	0.373
	鄰近	0.565	0.792	0.735	0.782	0.783	0.69	0.712	0.767	0.717	0.748

表 8. 以類神經網路(BPN)所預測之結果

		實際分群									平均
CC		**O1**	**O2**	**O3**	**O4**	**O5**	**O6**	**O7**	**O8**	**O9**	
預測落點	**O1**	**7**	10	15	3	1	0	3	0	0	
	O2	7	**24**	20	15	8	1	0	1	0	
	O3	3	21	**56**	21	15	4	4	1	1	
	O4	4	13	61	**85**	27	10	4	0	0	
	O5	1	3	19	48	**59**	31	3	1	1	
	O6	0	2	9	17	43	**41**	15	5	8	
	O7	1	3	5	8	23	17	**24**	5	7	
	O8	0	0	0	5	5	9	12	**13**	16	
	O9	0	1	0	5	2	0	8	4	**13**	
	總數	23	77	185	207	166	113	73	30	46	
	命中	0.304	0.312	0.303	0.41	0.355	0.363	0.329	0.43	0.283	0.349
	鄰近	0.61	0.714	0.741	0.744	0.777	0.788	0.699	0.733	0.63	0.737

表 9.兩種預測方法的比較：命中率

	決策樹		類神經網路	
分群	命中	鄰近	命中	鄰近
O1	39.1%	56.5%	30.4%	61%
O2	45.5%	79.2%	31.2%	71.4%
O3	38.4%	73.5%	30.3%	74.1%
O4	38.1%	78.2%	41%	74.4%
O5	39.1%	78.3%	35.5%	77.7%
O6	28.3%	69%	36.3%	78.8%
O7	39.7%	71.2%	32.9%	69.9%
O8	36.7%	76.7%	43%	73.3%
O9	30.4%	71.7%	28.3%	63%
平均數	37.3%	74.8%	34.9%	73.7%
標準差	5.13%	6.67%	4.77%	5.73%
中位數	38.1%	73.5%	32.9%	73.3%

4.4 實驗結果討論

經過實驗後，本研究所建立的模型，在預測電影的落點的準確性達到了超過 70%的鄰近命中率，其中完全命中的機率也超過 35%。所建立的兩個模型中，決策樹 C5.0 得到了 37.3%的命中率和 74.8%的鄰近預測，與類神經網路 34.9%的命中率和 73.7%的命中率。證明決策樹在製作電影票房的預測上，較類神經網路有較佳的預測結果。

這樣的結果可以使想要製作電影的片商，在製作影片的前端，就可以根據即將製作電影所內含的特徵質，帶入所建立的模型去進行落點預測，若是結果不盡好就可以在製作前做變更，例如增加製作成本，更改分級制度等等，以期上映時能得到觀眾的喜愛，讓製片降低虧損的風險，本研究針對提電影製作提供了一個有效的電影預測模型。

以電影實例來做說明，2006 年上映的邁阿密風雲(表 10)，根據本研究所得到的規則進行預測，得到的結果是落在O6，介於六千五百萬美金至一億美元，而實際上的票房收入是六千三百多萬美元，是在O5 區間，是屬於鄰近，根據製作成本一億兩千萬美元來看，是不理想的票房，必須有所調整。而 2006 年所上映的王者天下(表 11)，知名導演Ridley Scott，還有超高預算，但是根據預測結果將是會略在O6，而實際上的票房更是只有四千七百多萬美元。如這兩部電影的製作成本所示，預測票房都較成本低，若是能在製作前其進行預測，將可以趁早更改或是取消拍攝計畫，以避免片商蒙受重大的損失。

表 10.邁阿密風雲上映資訊

Miami Vice 邁阿密風雲 (2006)　預測票房落點 O6　實際票房 63,450,470 (O5)								
成本	競爭度	分級	類型	導演前三部票房成績	導演前三部電影評價	上映廳數	演員 A 前三部電影票房	演員 B 前三部電影票房
				Michael Mann	Michael Mann		Colin Farrell	Jamie Foxx
120,000,000	中	R	動作	53,694,551	7.1	3,021	6,155,916	56,552,653

表 11 王者天下上映資訊

Kingdom of heaven 王者天下 (2005)預測票房落點 O6 實際票房：47,398,413 (O5)								
成本	競爭度	分級	類型	導演前三部票房成績	導演前三部電影評價	上映廳數	演員 A 前三部電影票房	演員 B 前三部電影票房
				Ridley Scott	Ridley Scott		Orlando Bloom	
130,000,000	中	R	戰爭	103,534,469	7.2	3,216	45,505,026	0

本研究找出了各種票房區間的規則，並且用兩種不同的方法建立預測模型，經過檢驗後找出了決策樹的模型具有較佳的預測結果。

五、結論與建議

5.1 結論

看電影已經是現代人生活中很重要的休閒行為，2009 年在北美就有 105 億美元的票房總收入，而每年還穩定的以 3%的幅度在成長。為了能在市場上有更好的收益，各製片廠無不想辦法來吸引觀眾進場，但是隨著製作成本的節節高升，卻不一定能得到比較好的票房收入。該如何有效的投資製作電影，避免蒙受的損失，並能獲得利潤，一直是各片商想追求的。

本研究與 Ramesh and Dursun (2006)差別，在於搜集了近十年的電影資料作為資料庫並且進行驗證。在輸入項加入了衡量導演和演員近期作品的票房平均值，用數字動態量化觀眾對於劇組的喜好程度。在輸入變數使用較多的連續變數。利用資料採礦的技術，針對 859 部 1999 年到 2003 年，五年間在北美上映的電影，進行資料分析，並分別以決策樹 C5.0)和類神經網路 BPN 建立預測模型，再以 2004 年至 2008 年，五年內上映的 923 部電影去做為模型的驗證的樣本。本研究搜集的電影資訊中，以數字量化演員和導演過去的表現，具體變現出觀眾對於他們喜愛的程度。實驗的結果，決策樹找到了七項規則，並有 37.3%的機率可以預測到電影票房所落在的區間，74.8%可以預測到鄰近區間的區塊。而類神經網路的模型得到了 34.9%的完全命中率與 73.7%的鄰近命中率。決策樹得到的結果較類神經網路來的精準，代表決策樹在分析預測電影票房上有較好的效果。

面對每部電影的獨特性，利用資料採礦的技術，分析過去成功失敗電影的歷史資訊，建立的模型提供了欲製作新作品的片商，可以在衡量一個電影的企劃的時候，將其中的特徵質，代入已建立的模型，去判斷該電影票房收入可能會落在哪個區間，若是不理想或是有虧損的可能，可以在製作前取消該拍攝計畫，或是更改內容，像是分級制度、製作成本、演員選擇等，再代入模型檢視是否比較好的結果。如此即可讓投資更有保障，片商也減少虧損的風險。

5.2 建議

1. 電影工業未來可能將會走入另外一個更強調視覺感官的領域，隨著拍攝技術的逐漸上升，電影阿凡達，3D 技術的空前成功，而 3D 電影的票價實質上比較，可以提高票房收入。所以未來的研究可以加入這個投入項去做衡量。
2. 未來的研究可以針對特定時間性的歷史數據去做資料採礦分析，如在聖誕節、新年間所上影的電影的屬性去做探討，是否會增加預測的精準性。
3. 增加每部電影的類型數目，一部電影劇情所含的元素可能很多元，若可以加以獨立區分，以二進位編碼方式去做輸入項，如表 12所示，若電影同時含有動作、愛情、喜劇、冒險等元素，則可以用此表精確的表示出來。
4. 本研究是針對美國電影市場的消費特性去做的分析與探討，若是台灣能夠建立起完整的電影上映資料，就能進行資料採礦的動作，進而分析台灣消費者的習性，讓國片的製作能有個依據，對於普遍資金缺乏的台灣電影產業將會是一個助力。

表 12.電影中類型所含屬性表

動作	愛情	喜劇	驚悚	恐怖	劇情	科幻	史詩	戰爭	漫畫	懸疑	青春	兒童	歌舞	溫馨	劇情	冒險	災難	動畫
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

參考文獻

[1] BoxOffice Mojo (http://www.boxofficemojo.com/)

[2] De Silva, I. Consumer Selection of Motion Pictures, appeared in The Motion Picture Mega-Industry by B. Litman. Allyn & Bacon Publishing, Inc.: Boston, MA. (1998).

[3] Eliashberg, J., Junker, J. J., Sawhney, M. S., & Wierenga, B. MOVIEMOD: An implementable decision support system for prerelease market evaluation of motion pictures. Marketing Science, 19(3), 226–243. (2000).

[4] Eliashberg, J. and Sawhney, M.S. "Modeling Goes to Hollywood: Predicting Individual Differences in movie enjoyment." Management science, Vol 40, issue.9, September ,1151-1173. (1994).

[5] Eliashberg, J. "The Film Exhibition Business: Critical Issues, Practice and Research." In Charles C. Moul (Editor). A Concise Handbook of Movie Industry Economics. New York City, New York: Cambridge University Press, 138-162. (2005).

[6] Elberse, A. and J. Eliashberg "The Drivers of Motion PicturePerformance: The Need to Consider Dynamics, Endogeneity and Simultaneity, to appear in the Proceedings of the Business and Economic Scholars Workshop in Motion Picture Industry Studies",Florida Atlantic University, pp. 1–15. (2002)

[7] Frawley, W.J.m Paitetsky: Shapiro, G. and Matheus ,C.J. ,"Knowledge Discovery in Databases: An Overview Knowledge Discovery in Database," California , AAAI/MIT Press , 1-30. (1991)

[8] Fayyad, U.M., "Dataming and Knowledge Discovery: Marketing Sense out of Data", IEEE Expert , Vol. 11, No.5 , 20-25.(1996)

[9] Grupe, F.H. and Owang,M.M. , "Databse Mining Discovering New knowledge and Cooperative Advantage," Information Systems Management, Vol.12 No.4, 26-31 (1995)

[10] Han, J and Kamber, M., 2001, Data Mining Concepts and techniques, Morgan Kaufmann.

[11] IMDb (http://www.imdb.com)

[12] Krider, R. E., Weinberg, C. B. "Competitive dynamics and the introduction of new products: The motion picture timing game." Journal of Marketing Research, 35(1), 1–15. (1998)

[13] Kim, Sung-Min , Jong-Dal Kim, Jeong-Hee Hong, Do-Won Nam, Don-Ha Lee , Jeon-Young Lee , " A System for Association Rule Finding from an Internet Portal Site," (2000)

[14] Kleissner, C. , "Data mining for the enterprise," In Proceedings of the Thrity-First Hawaii International Conference on, Volume 7 ,295-304. (1998)

[15] Litman, B. R., Ahn, H. "Predicting financial success of motion pictures: The early 90's experience." In B. Litman (Ed.), The motion picture mega-industry. Boston: Allyn and Bacon Publishing. (1998)

[16] Litman, B. R., and Kohl, A. "Predicting financial success of motion pictures: The 80's experience." The Journal of Media Economics, 2(1),35–50. (1989).

[17] Litman, B. R. "Predicting success of theatrical movies: An empirical Study". Journal of Popular Culture, 16(9), 159–175. (1983).

[18] Michael , J.A. and Linoff, G., "Data mining Technique: for Marketing , Sales and Customer Support," Wiley Computer Publishing, New York. (1997)

[19] MPAA. MPAA Economic Review. (2004)

[20] Neelamegham, R., and Chintagunta, P. “A Bayesian Model to forecast new product performance in domestic and international markets”.Marketing Science, 18(2), 115–136. (1999).
[21] Radas, S., and Shugan, S. M. “Seasonal marketing and timing new product introductions.” Journal of Marketing Research, 35(3), 296–315. (1998).
[22] Ramesh and Dursun “Predicting box-office success of motion pictures with neural networks”, Exper System with Applications 30 , 243-254(2006)
[23] Ravid, S. A. Information, blockbusters, and stars: A study of the film industry. Journal of Business, 72(4), 463–492. (1999).
[24] Sawhney, M. S., & Eliashberg, J. A parsimonious model for forecasting gross box-office revenues of motion pictures. MarketingScience, 15(2), 113–131. (1996).
[25] Sochay, S. “Predicting the performance of motion pictures.” The Journal of Media Economics, 7(4), 1–20. (1994).
[26] S. Basuroy, S. Chatterjee, and S. A. Ravid, “How critical are critical reviews? The box office effects of film critics, star power and budgets,” Journal of Marketing, vol. 67, pp. 103–117, (2003).
[27] SPSS Inc. (2006). Clementine® 10.1. In-Database Mining Guide.
[28] Valenti, J. “Motion Pictures and their impact on society in the year 2000”, speech given at the Midwest Research Institute , Kansas City, April 25, P.7. (1978)
[29] 牛田一雄、高井免、木幕大輔原著 陳耀茂編審，”資料採礦利用 Clementine 使用手冊”：鼎茂書局。(2006)
[30] 葉怡成，類神經網路模式應用與實作(第七版：儒林圖書有限公司)，2000。