應用資料探勘探討資訊揭露透明度

程守雄
建國科技大學資訊管理系
shcheng@ctu.edu.tw

摘要

在資本市場，投資人的決策必須依賴及時可靠的資訊作判斷。資訊揭露透明度對投資人的影響是最直接的，尤其是資訊越透明，投資者決策能力越強。雖然官方機構每年固定會公佈資訊揭露透明度評鑑結果，但卻在隔年的六月才公佈評鑑結果。所以評鑑結果對於廣大的投資大眾而言，並不能在當年度當作投資準則。本研究主要運用資料探勘技術，建置一套預測及分類資訊揭露透明度的方法。本研究採用決策樹為基礎的資料探勘技術，探討台灣股票市場的資訊揭露透明度分類模型。實證結果顯示本研究所提出的分類預測模型可以精準的預測資訊揭露透明度。本研究所提出的方法是直接且有效率的。

關鍵字：資料探勘、資訊揭露、決策樹

一、緒論

1997 年亞洲發生金融風暴、2001年美國安隆公司(Enron)、世界通訊(WorldCom)、默克藥廠(Merck)等相繼爆發了震驚全世界的會計交易醜聞。為了加強公司治理，重建投資人對資本市場的信心，經濟合作暨發展組織(Organization for Economic Cooperation and Development，簡稱OECD)於1999 年起，連續三年年會，都以探討亞洲企業之公司治理為主要議題，並規範出公司治理原則，包括了股東權利、公平對待股東、董事會責任、資訊揭露和透明度以及公司治理跟利害關係人角色。其中，資訊揭露及透明度的加強對投資人的影響是最直接的，尤其是在資本市場，投資人的決策須依賴及時可靠的資訊作判斷；資訊越透明，投資者決策能力越強，資源分配越有效率，因此各國紛紛提出健全公司治理機制及資訊揭露評鑑系統的補牢措施。根據Mckinsey & Company (2000)的調查，投資人對於公司治理良好的公司，願意多付出10%到30%的溢價來投資。Healy and Palepu(2001)認為，公司年度報表與資訊揭露的需求乃是起因於內部管理者與外部投資人間存在著資訊不對稱。為解決此種資訊不對稱，由臺灣證券交易所及證券櫃買中心委託中華民國證券暨期貨市場發展基金會(簡稱證期會) 自2003年開始每年辦理上市櫃公司資訊揭露評鑑。評鑑內容係根據前一年受評公司輸入公開資訊觀測站揭露的資訊為主，不包括報章雜誌等媒體之報導，但受評公司針對媒體報導所作的澄清性揭露

(以輸入公開資訊觀測站為限)及公司網站所揭露之資訊亦列入。第一屆與第二屆資訊揭露評鑑結果分為兩類，分別為透明與不透明，評鑑結果從第三屆開始有所改變，依所有上市櫃公司受評成績高低區分為A+級、A級、B級、C級以及C-級。

雖然台灣證期會(Securities and Futures Institute)每年固定會公佈評鑑結果，但卻在隔年的六月才公佈評鑑結果；所以評鑑結果對於廣大的投資大眾而言，並不能當作當年度的投資準則，只能被動的接受這些陳舊的資料。因此，發展一套可以精確的預測及分類資訊透明度的方法，是一個重要且有迫切需要的研究題目。然而，目前探討這個題目相關的研究還非常稀少。因此，本研究將利用資料探勘的技術，提出一套可以精確預測及分類資訊透明度的方法。相信這個預測及分類方法，將可以幫助投資者主動且及時的分類和預測上市櫃公司所有類型的透明度，而不必等待官方隔年六月才出爐的報告。

2. 研究方法

2.1 條件屬性與決策屬性

本研究條件屬性共分為股權結構、董事會組成、管理當局、財務指標等四個構面，資訊揭露程度為決策屬性， 17 項屬性定義如下：

1. 資訊揭露透明度：公司資訊揭露透明度的高低，本研究中設定，A+級：5，A級：4，B級：3，C級：2，C-級：1。
2. 管理者持股比率：總經理、副總經理，重要部門經理及協理，負責公司重要決策之高階主管持股總數佔公司流通在外股數比率。
3. 董監事持股比率：董監事持股總數佔公司流通在外股數比率。
4. 獨立董監事席次比率：獨立董監事席次佔董事會總席次之比率。
5. 大股東持股比率：大股東（持股數佔公司流通在外股數比率10%以上）持股總數佔公司流通在外股數比率。
6. 家族持股比率：董監事及經理人之配偶、未成年子女及利用他人名義持有股數佔公司流通在外股數比率。
7. 機構投資人持股比率：國內金融機構及證券投資信託基金持股總數佔公司流通在外股數比率。
8. 外資持股比率：僑外之機構投資人、法人、證券投資信託基金、自然人等持股總數佔公司流通在外股數比率。
9. 官方持股比率：政府機構持股總數佔公司流通在外股數比率。
10. 法人持股比率：公司法人及其他法人等以法人形式持有公司持股總數佔公司流通在外股數比率。
11. 負債比率：負債總額／資產總額。
12. 公司規模：資產總額取自然對數。
13. 董事長是否兼任總經理：為一虛擬變數，有為1，無則為0。

14. 研發比例：研發費用／資產總額。
15. 產業別：為一虛擬變數，電子業為1，其他產業為0。
16. 每股盈餘：（稅後淨利-特別股股利）/流通在外股數
17. 總資產報酬率=稅後淨利/資產總額

2.2 決策樹分類模型

ID3為最早使用的決策樹演算法之一，ID3的主要核心是以遞迴的方式將訓練資料作切割。在每一次產生節點時，某些輸入的訓練子集將取出測試，以資訊獲取量來當作測試，在選取過後，將以具有最大資訊獲取量的值當作分支的節點，接下來依照其遞迴的動作選取下一個分支節點，直到每一個訓練資料都屬於一個分類之中或是符合某個滿足條件。C4.5是ID3的延伸方法，改善了ID3產生過多子集合，而每個子集合僅包含少數資料的問題，並且具備處理連續數值型屬性、雜訊的處理，另外也兼具修剪樹的能力。C4.5 在決策樹的每個節點上使用資訊獲取量來選擇測試屬性，選擇具有最高資訊獲取量(或最大熵壓縮)的屬性作為當前節點的測試屬性。

本研究採用C5.0來產生決策樹並且預測結果類別。C5.0是延續C4.5的演算法架構，不同的是C5.0提供了在許多應用上較受歡迎的規則集，將分類的條件一一 以規則形式表達，增加了閱讀分類規則的可讀性。C5.0可用來處理數值性或是名目性欄位的資料，分析結果可用容易理解的決策樹或是若-則的關係呈現。C5.0與C4.5不同之處在於C5.0可以處理更多種資料型態，如日期、時間、時間戳記、序列性的離散型資料…等資料型態。將C5.0決策樹分類演算方法說明如下：

(1) 模型參數

C5.0 不限制只能做二元分割，這是與其它決策樹分類法不同的地方。先根據資訊獲取量來決定分割，建構一棵完整的決策樹，再針對每一個內部節點依定義的錯誤預估率來作決策樹修剪的動作，選擇資訊增大量最大的屬性為分割屬性。

(2) 分支演算法

根據資訊獲取量來決定分割的預測變數，假設資料集 S 包含二個類別 P 與 N，而其中 P 類別有 p 筆資料，N 類別有 n 筆資料，所含的資訊量表

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \quad (1)$$

若以預測變數 A 得出的分類數有 v 筆，其中包含了 P 類別 p_i 個與 N 類別 n_i 個，那麼以預測變數 A 的熵為

$$E(A) = \sum_{i=1}^{V}\frac{p_1+n_1}{p+n}I(p_1,n_1) \quad (2)$$

，在此以 A 為分割節點可得到的資訊增加量為

$Gain(A) = I(p,n) - E(A)$ ，而 C 5.0 選擇資訊增加量最大的屬性為分割節點。

(3)決策樹停止成長規則

A. 節點內只剩下一種值。
B. 節點內的預測變數的數目已經與所有目標變數的數目皆相同時。
C. 樹的分割層數達到預設的最大值。
D. 母節點內的資料個數已小於預設可分割的值。
E. 子節點內的資料個數已小於預設

的值。

F. 修剪決策樹

採用錯誤預估率作判斷條件，由底部往上測試每樹葉節點，若被分類於某一子樹的所有訓練資料個數為T，其中有E個訓練資料分類錯誤，而錯誤預估率就是由E/T評估，當有新的資料被測試時，則利用E/T判斷結果為不正確的機率有多高。若以一個樹葉節點代替後所得到的錯誤預估率比原先來得低則將修剪為單一樹葉節點，否則就保留原來的結構。

3.實證分析

3.1 資料來源

本研究以臺灣證券交易所及證券櫃買中心委託證基會（http://www.sfi.org.tw/EDIS/） 所建置的「資訊揭露評鑑系統」於2009年6月所公佈的「第六屆上市櫃公司資訊揭露評鑑系統評鑑結果」中接受評鑑的上市櫃公司為樣本。樣本資料來源為台灣經濟新報資料庫(TEJ) （http://www.tej.com.tw/）及公開資訊觀測站 http://mops.twse.com.tw/mops/web/index。

3.2 實驗結果

C5.0 規則歸納演算法，是透過解釋欄位跟輸出欄位間的關係，用遞迴方式將資料分割成子集後建立出決策樹並導出決策樹規則，嘗試解釋資料中不同部分跟輸出欄位或結果的關係。本研究所建立的分類預測模型，實驗結果主要有下列三項：

(1) 各條件屬性重要性：

表 1 為各條件屬性重要性。由表一的結果，顯示影響資訊透明度評鑑結果最重要的屬性為公司規模，依序為負債比率、法人持股比率、外資持股比率等。

表 1 各條件屬性重要性

條件屬性	重要性
產業別	0
研發比例	0
董事長是否兼任總經理	0.01
官方持股比率	0.02
大股東持股比率	0.03
每股盈餘	0.04
管理者持股比率	0.04
家族持股比率	0.04
董監事持股比率	0.05
機構投資人持股比率	0.08
總資產報酬率	0.08
獨立董監事席次比率	0.08
外資持股比率	0.09
法人持股比率	0.11
負債比率	0.14
公司規模	0.2

(2)分類預測規則：

透過 C5.0 規則歸納演算法，可以獲得每一種資訊揭露透明度的分類預測規則，我們以資訊揭露透明度為『A+』分類預測規則，說明如下：

規則 1: 如果 負債比率 <= 82.460 且 公

司規模 <= 13.369 且 總資產報酬率 <= 11.350 且 產業別 = 1 且 外資持股比率 > 0.670 且 大股東持股比率 > 4.440 且 獨立董監事席次比率 <= 0.364 且 大股東持股比率 <= 7.400 則資訊揭露透明度為 A＋

規則 2: 如果 公司規模 > 16.746 且 管理者持股比率 <= 1.580 且 董事長是否兼任總經理 = 0 且 機構投資人持股比率 > 7.150 且 機構投資人持股比率 <= 9.030 則資訊揭露透明度為 A＋

規則 3: 如果 公司規模 > 16.746 且 管理者持股比率 <= 1.580 且 董事長是否兼任總經理 = 0 且 機構投資人持股比率 > 7.150 且 家族持股比率 <= 44,063 則資訊揭露透明度為 A＋

規則 4: 如果 公司規模 > 16.746 且 管理者持股比率 <= 1.580 且 董事長是否兼任總經理 = 0 且 機構投資人持股比率 > 7.150 且 家族持股比率 <= 44,063 且 公司規模 <= 17.863 則資訊揭露透明度為 A＋

規則 5: 如果 公司規模 > 16.746 且 管理者持股比率 <= 1.580 且 董事長是否兼任總經理 = 0 且 機構投資人持股比率 > 7.150 且 家族持股比率 <= 44,063 且 公司規模 <= 17.863 且 外資持股比率 <= 20 則資訊揭露透明度為 A＋

規則 6: 如果 負債比率 > 82.460 且 法人持股比率 <= 39.210 且 法人持股比率 > 11.350 則資訊揭露透明度為 A＋

規則 7: 如果 負債比率 > 82.460 且 法人持股比率 <= 39.210 且 法人持股比率 > 11.350 且 法人持股比率 <= 18.380 則資訊揭露透明度為 A＋

規則 8: 如果 負債比率 > 82.460 且 法人持股比率 <= 39.210 且 法人持股比率 > 18.380 且 外資持股比率 <= 21.760 則資訊揭露透明度為 A＋

(3) 分類預測的準確性：

本研究分類預測的結果與評鑑系統評鑑結果的比較如表 2。在樣本數 1118 家上市櫃公司中，以 75%的樣本數為訓練資料集，以 25%的樣本數為驗證資料集。驗證結果顯示 A+級公司預測的準確率可以達到 94%，A 級公司預測的準確率可以達到 98%，B 級公司預測的準確率可以達到 99%，C 級公司預測的準確率可以達到 90%。然而 C-級公司預測的準確率卻只有 50%，探討其準確率較低的原因為列為 C-級公司本身公告的資訊本來就比較不透明，所以本研究所能獲得的資訊也相對比較不準確。但以整體來看，整體的準確率仍然可以達到 98%。結果顯示，本研究方法確實可以相當有效的預測。

表 2 公告結果與預測結果的比較

透明度類別	A+級公司數目	A 級公司數目	B 級公司數目	C 級公司數目	C-級公司數目
訓練樣本數	27	243	479	79	9
驗證樣本數	10	82	160	26	3
訓練準確率	95%	98%	99%	92%	53%
驗證準確率	94%	98%	99%	90%	50%

4.結論

本研究主要以決策樹演算法為基礎建立了一套可以準確預測資訊揭露透明度的分類模型。從公司股權結構、董事會組成、管理當局、財務指標等四個構面建立可以精確預測及分類資訊透明度的方法，影響資訊揭露透明度的條件屬性，以及預測及分類資訊透明度的方法的規則。並由實例的驗證，驗證本方法是快速且有效率的。

參考文獻

[1] Forker, J. (1992). Corporate governance and disclosure quality. Accounting and Business Research, 22, 111－124.

[2] Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed.). Amsterdam: Morgan Kaufmann.

[3] Kim, Y. S., Street, W. N., & Menczer, F. (2006). Optimal ensemble construction via meta-evolutionary ensembles. Expert Systems with

[4] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1,81－106.

[5] Roiger, R. J., & Geatz, M. W. (2003). Data Mining: A Tutorial-based Primer. Boston: Addison Wesley.

[6] Sandeep, A., Amra, B., & Liliane, B. (2002). Measuring transparency and disclosure at firm-level in emerging markets. Emerging Markets Review, 3, 325－337.

[7] Tsai, Y. C., Cheng, C. H., & Chang, J. R. (2006). Entropy-based fuzzy rough classification approach for extracting classification rules. Expert Systems with Applications, 31(2), 436－443.

[8] Verrecchia, A. (1983). Discretionary disclosure. Journal of Accounting and Economics, 5, 179－194.