

# Applying Data Mining Techniques to Analyze Match Results from the Association of Tennis Professionals

李尚鴻、林朝源

E-mail: 9707465@mail.dyu.edu.tw

## ABSTRACT

ABSTRACT Nowadays all advanced countries use computers, enterprises have accumulated quite a hung mass of electronic data in more than twenty years in the past. However there are many unpredictable behaviors of Human can ' t be observed and predicated directly by eyes. Using data mining techniques can help us analyzing problem ' s attributions and find out indications of why the event was happened by past experiences. Even though enterprises have adopted a data mining approach to accomplish a number of marketing goals and objectives in business, but data mining techniques have not bee applied widely in domain of sports. Applying data mining techniques can bestead the coach direct athletes training and make strategies of competitions in the future. Therefore, how to apply data mining techniques to dig useful information from huge amount of record data of the past and how we classify complicated behaviors of Human to find out factors which really infected athletes ' behavior are both worth to be researched. In this paper, we are going to use real record data of competitions of the past from ATP(Association of Tennis Professional) for a sample of the data mining research. Using decision tree and association rules techniques of the software of SQL Server to analyze problems and establish predictable models to find out important rules. At last, models will be tested with data of the different time and produces a classified matrix. Using 4 kinds of indexes to analyze values of the classified matrix and AUC(Area Under Curve) to estimate models.

Keywords : Data Mining, Decision Tree, Associated Rules, SQL Server

## Table of Contents

目錄 封面內頁 簽名頁 授權書 iii 中文摘要 v ABSTRACT vi 誌謝 vii 目錄 viii 圖目錄 xi 表目錄 xiii 第一章 緒論 1 1.1 研究背景與動機 1 1.2 研究目的 2 1.3 研究流程 3 第二章 文獻探討 5 2.1 決策樹法學習法相關文獻回顧 5 2.2 關聯規則學習法相關文獻回顧 6 2.3 資料探勘應用於運動領域的相關研究 8 2.4 國內資料探勘應用相關文獻 8 第三章 研究方法 11 3.1 定義目標問題 12 3.2 搜尋所有可得的資料 12 3.3 選擇適合的資料探勘技術 12 3.3.1 SQL軟體內的決策樹演算學習法的原理 13 3.3.2 SQL軟體內的關聯規則原理 15 3.4 分析資料 19 3.5 資料預處理 20 3.5.1篩選有效變數 20 3.5.2連續變數與類別變數轉換處理 21 3.5.3遺漏值處理 21 3.5.4極端值與標準化 22 3.6 建立模型 23 3.7 評估模型 23 第四章 實證分析 26 4.1 實行探勘流程 26 4.1.1 定義目標問題 27 4.1.2搜尋所有可得資料 27 4.1.3選擇適合的資料探勘技術 29 4.1.4資料分析 29 4.1.5資料預處理 30 4.2 選擇最佳模型 31 4.2.1建立第一種資料庫的決策樹模型 31 4.2.2建立第一種資料庫的關聯規則模型 37 4.2.3建立第二種資料庫的決策樹模型 39 4.2.4建立第二種資料庫的關聯規則模型 45 4.2.5兩種資料庫的決策樹模型比較 47 4.2.6兩種資料庫的關聯規則模型比較 48 4.3 分析解釋模型產生的規則 49 4.4 不同資料量與不同時間點資料的模型比較 57 4.5 與其它探礦技術模型的比較 61 第五章 結論與未來建議 64 5.1 研究結論 64 5.2 未來建議 65 參考文獻 67 附錄A 70

## REFERENCES

- 參考文獻 [1] 王堅斌, (民92), 使用分類結合法與決策樹判別忠誠顧客, 碩士論文, 成功大學資訊管理學系。  
[2] 尹相志, (民96), SQL Server 2005 Data Mining 資料探礦, 悅知文化。  
[3] 林艷, (民90), 以資料挖礦法則以資料挖礦法則發掘Tree- Based系統效能型態—以ERP系統為例, 碩士論文, 中央大學資訊管理學系。  
[4] 呂金品, (民90), 運用決策樹分類方法估計小額放款倒帳率 兼論放款策略, 碩士論文, 中山大學財務管理學系。  
[5] 吳聲弘, (民92), 漸增式跨階層次高效率關聯規則探勘, 碩士論文, 南台科技大學資訊管理學系。  
[6] 張璋倫, (民89), 應用資料挖掘學習方法探討顧客關係管理問題, 碩士論文, 輔仁大學資訊管理學系。  
[7] 廖聖傑, (民92), 從學習歷程檔案建構決策樹以支援網路教學, 碩士論文, 中山大學資訊管理學系。  
[8] 陳秀如, (民92), 效率之關聯規則探勘演算法暨應用其探採具多重支持度之加權數量關聯規則, 碩士論文, 南台科技大學資訊管理學系。  
[9] 陳俊堯, (民92), 在異動資料庫中有效率地動態探勘關聯規則, 碩士論文, 南台科技大學資訊管理學系。  
[10] 蔡家昌, (民91), 應用決策樹歸納法探討台灣行動電話市場區隔, 碩士論文, 南台科技大學資訊管理學系。

- [11] 蔡智政, (民91), 應用CART決策樹與資料視覺技術於低良率晶圓成因探討, 碩士論文, 南台科技大學資訊管理學系。
- [12] Chen – Yueh Chen & Yi-Hsiu Lin, (2006), A New Market Research Approach in Sport-Data Mining, *The Sport Journal*, 46-51.
- [13] Chidanand Apte' and Sholom Weiss, (2005), Data mining with decision trees and decision rules, *Future Generation Computer Systems*, 197-210.
- [14] Di Dong and Rafael A. Calvo, (2007), Integrating Data Mining Processes within the Web Environment for the Sports Community, *Integration Technology, IEEE International*, 658-662.
- [15] Emekci, F., Sahin, O.D., Agrawal, D. and El Abbadi, A., (2007), Privacy preserving decision tree learning over multiple parties, *Data & Knowledge Engineering*, 348-361.
- [16] Jie Sun and Hui Li, (2008), Data mining method for listed companies' financial distress prediction, *Knowledge-Based Systems*, 1-5.
- [17] Jiuyong Li and Jason Jones, (2006), Using multiple and negative target rules to make classifiers more understandable, *Knowledge-Based Systems*, 438-444.
- [18] Islam, M.Z. and Brankovic, L., (2005), DETECTIVE: a decision tree based categorical value clustering and perturbation technique for preserving privacy in data mining, *Industrial Informatics, IEEE International Conference*, 701-708.
- [19] Khaing, M.M. and Thein, N., (2006), An efficient association rule mining for XML data, *SICE-ICASE, International Joint Conference*, 5782-5786.
- [20] Osama K. Solieman, *Data Mining in Sports: A Research Overview*, MIS Masters Project.
- [21] Ramakrishnan Srikant and Rakesh Agrawal, (1996), Mining quantitative association Rules in large Relational tables, *Proceedings of the 1996 ACM SIGMOD international conference on Management of data table of contents*, Montreal, Quebec, Canada, 1-12.
- [22] Shintani, T., (2006), Mining association rules from data with missing values by database partitioning and merging, *IEEE/ACIS, International Workshop on Component-Based Software Engineering, Software Architecture and Reuse.*, 193-200.
- [23] Xinwang Liu and Qingli Da, (2005), A decision tree solution considering the decision maker's attitude, *Fuzzy Sets and Systems*, 437-454.
- [24] Zhang Hong, Zhang Bo, Kong Ling-Dong and Cai Zheng-Xing, (2007), Generalized association rule mining algorithms based on data cube, *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, 803-808.
- [25] <http://www.atptennis.com/1/en/home/>