資料分析應用於網路管理

Using Data Analytics in Network Management

[1] 李秋緣 [2] 陳威全 [3,4] 彭賓鈺 [2,5] 李仁鐘

[1]Chou-Yuan Lee [2] Wei-Quan Ting [3,4] Bin-Yu Peng [2,5] Zne-Jung Lee

[1]福州外語外貿學院信息系

[1] Department of Information Technology, Fuzhou University of International Studies and Trade, Fuzhou, China

[2] 華梵大學資訊管理學系

[2] Department of Information Management, Huafan University

[3] 華梵大學機電系

[3]Department of Mechatronic Engineering, Huafan University

[4] 康寧大學資管科

[4]Department of Information Management, University of Kang Ning

[5] 安碁資訊股份有限公司

Acer Cyber Security Inc.

摘要

世界關務組織（World Customs Organization, WCO）通過「國際貿易安全與便捷化標準架構」，行政院也規劃出相應「優質經貿網絡計畫」，由財政部關務署輔導修改通關即用報關軟體，在電子化作業流程下，報關行出現網路管理需求、資訊安全管理問題與設備效能評估。本研究應用網路管理簡單網路管理協定(Simple Network Management Protocol, SNMP)與開放原始碼軟體架構報關行網路管理平台，協助管理廠商與設備、提供網路拓樸與設備狀態查詢。使用多元回歸與決策樹將網站平台收集資料進行分析，上傳量多元回歸分析均方根誤差(Root Mean Square Error, RMSE)為 0.03334、決定係數(Coefficient of Determination, R^2)為 0.9994，下載量多元回歸分析 RMSE 為 0.24833、R^2 為 0.99448。透過決策樹分析發現流量「設備待機」、「日常運作」與「處理器高負載」三種模式。經由結果得知，網站平台能協助使用者建立管理規則與發現可能資訊安全管理問題，也可作為進行設備效能評估時之參考建議。

關鍵字：網路管理、簡單網路管理協定、多元回歸、決策樹

Abstract

World Customs Organization (WCO) passed the “Framework of Standards to Secure and Facilitate Global Trade (WCO SAFE)”. According to the WCO SAFE, Executive also planned “Ubiquitous Economy and Trade Network Plan”. Customs Administration revised the procedure of customs automation software and Customs broker encountered problems such as network management, information security management and the evaluation of equipment performance in electrization operational procedures. In this paper, a customs broker network operation center (CBNOC) is built simple network management protocol (SNMP) and open-source software to provide equipment management, network topology, multiple regression analysis, and decision tree analysis. In multiple regressions, root mean square error (RMSE) and coefficient of determination (R^2) for the flow of upload are 0.03334 and 0.9994, respectively. RMSE and R^2 for the flow of download are 0.03334 and 0.9994, respectively. There are three modes, “standby,” “daily operation” and “cpu busy” that are generated from the flow results of decision tree. From the results, CBNOC could generate decision rules and find the problems of security management for users. Moreover, it could provide the decision-making for the evaluation of equipment performance.

Keywords: Network Management, Simple Network Management Protocol, Multiple Regression, Decision Tree.

1.研究背景與目的

我國行政院為符合 WCO SAFE 標準架構中關務資訊交換規範
[9][10]，而規劃出「優質經貿網絡計畫」計畫[3]。其中計畫內「關港貿單一窗口」與「出口預報貨物資訊」兩項子計畫，為輔導並鼓勵相關資訊服務業者修改通關即用報關系統[1]，因此改變報關行以往傳統的作業模式。本研究以報關行作業為例，提出「報關行網路管理平台」整合網路管理服務需求並收集網路設備中每日產出之網路傳輸量、系統日誌等資料，透過資料探勘演算法技術進行分析，找出資訊安全管理問題與歸納網路管理服務營運平台規則，最後提供企業做為服務營運管理工作效能評估時的參考建議。

2.文獻探討

2.1 簡易網路管理協定

簡易網路管理協定(Simple Network Management Protocol, SNMP)屬於 TCP/IP 協定中應用層(Application Layer)[12]，可針對網路設備如伺服器主機、閘道器(Gateway)、終端機(Terminal)等，收集如中央處理器(CPU)、記憶體(Memory)等用量。目前 SNMP 被用於智慧電網管理結合其他人機介面技術[6]，透過視覺化界面呈現設備目前的設備狀態與歷史耗能[4][5]。

2.2 多元回歸

回歸分析(Regression Analysis)最早是由英國 Francis Gal-ton 提出在統計學上屬於一種資料分析模式，屬於資料探勘中預測方法，是分析依變數與多個自變數之間關係的直線方程式。假定一條多元回歸模型中有 1 個依變數 Y 與 n 個自變數 X ，若依變數與自變數之間呈現線性關係時[2]，其公式為：

$$Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + \cdots + C_nX_n \quad (1)$$

其中C_0為殘差(Residuals)，$C_1, C_2, C_3, \cdots C_n$ 為回歸係數(Coefficients)，當C_1 對應X_1 且固定$X_1, X_2, X_3, \cdots X_n$的情境下，每當$X_1$ 產生變化，則會影響對 Y 的應變量。

2.3 分類回歸樹

分類回歸樹(Classification and Regression Tree, CART)於 1984 年由美國 Brieman 所提出[7]，其特性為可以分析離散型變數與連續型變數，使用二元分割規則進行歸納與分析，其中分支方法使用 *Gini* 係數(Gini Coefficient)作為分支的依據[15]，其公式如下所示[8][13]：

$$Gini(\boldsymbol{S}) = \mathbf{1} - \sum_{j=2}^{n} \boldsymbol{p_j^2} \quad (2)$$

$$\boldsymbol{Gini_A(S)} = \frac{|\boldsymbol{S_1}|}{|\boldsymbol{S}|}\boldsymbol{Gini_A(S_1)} + \frac{|\boldsymbol{S_2}|}{|\boldsymbol{S}|}\boldsymbol{Gini_A(S_2)} \quad (3)$$

$$\Delta\boldsymbol{Gini(A)} = \boldsymbol{Gini(S)} - \boldsymbol{Gini_A(S)} \quad (4)$$

其中當每種預測值在該節點中之出現頻率為p_j，再利用欄位A將資料集S分割為S_1與S_2，計算欄位A內S_1及S_2所建構成兩組子集合資料。當選擇最大不純度之降低值或 *Gini* 係數$Gini_A(S)$為最小時，其欄位則會作為分支點。

2.4 均方根誤差

均方根誤差(Root Mean Square Error, RMSE)是一種常用於了解實際值與預測值之間差距的公式，可做為判斷預測模型的正確性與精確度，RMSE 是由均方誤差(Mean Square Error, MSE)公式導出後開根號，MSE 公式如下[11]：

$$MSE = \frac{\sum(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}}}{\boldsymbol{n}} \quad (5)$$

其中$\sum(\hat{y}_i - y_i)^2$為實際值與預測值的誤差平方和。均方根誤差之公式則為：

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}}}{\boldsymbol{n}}} \quad (6)$$

2.5 決定係數

決定係數(Coefficient of Determination)又稱為 R^2，透過自變量佔依變量的總變異量(Variance)百分比，來表示模型的解釋力程度[14]。假設資料集有 n 筆資料($Y_1, Y_2, \ldots Y_n$)，對應的預測值分別為($F_1, F_2, \ldots F_n$)，殘差則定義為 $E_i = Y_i - F_i$，其 R^2 公式如下[16]：

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad (7)$$

$$R^2 \equiv 1 - \frac{\sum_i (Y_i - F_i)^2}{\sum_i (Y_i - \bar{Y})^2} \quad (8)$$

其中$\bar{Y}$為平均觀察值，$\sum_i (Y_i - F_i)^2$為實際值與預測值的殘差平方和，$\sum_i (Y_i - \bar{Y})^2$則為總平方和。

3.系統架構

本平台系統架構圖如圖所示。報關行網路管理平台主要提供功能包括：管理廠商與設備、設備狀況、資料分析與網路拓樸。

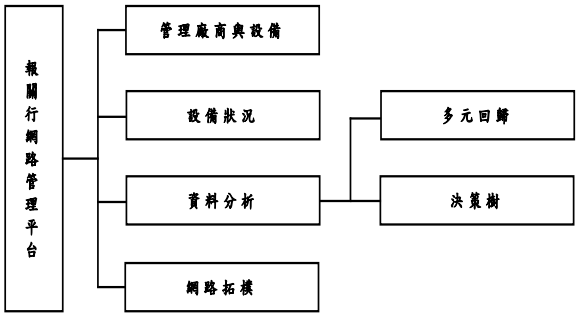

圖 1: 研究架構圖

本研究所使用的程式開發語言為 Python 程式語言，平台建置使用 Django Framework，使用 Microsoft SQL Server 資料庫來管理與儲存收集的設備 SNMP 資料。簡列平台研究開發時採用的工具與環境如下：

1. 程式語言：Python 3.5.3
2. 平台建置環境：Django 1.10.8
3. 資料庫：Microsoft SQL Server 2008R

本研究資料來源為華梵大學智慧型實驗室實驗電腦與合作廠商測試電腦。「處理器核心數、處理器使用量、虛擬記憶體可用量、虛擬記憶體使用量、實體記憶體可用量、實體記憶體使用量、儲存硬碟可用量、儲存硬體使用量」在 SNMP 協定中為每 60 秒更新一次；「網路上傳量與網路下載量」則為每 10 秒更新一次，透過自動化程式定時收集目標設備資料，資料取得後儲存於資料庫中

4.系統成果展示

平台首頁圖如圖 2 所示。本研究建置出一個報關行網路管理平台。使用者在進行加入廠商以及加入設備功能後，可使用查看網路拓樸、設備狀況以及查看資料分析中的多元回歸分析結果與決策樹分析結果。

圖 2: 平台首頁

多元回歸分析結果頁面圖如圖 3 所示。左側功能列為所有已分析的設備 IP 位置。若使用者選擇查看設備 IP 後，網站平台會透過視覺化界面與資料呈現 IP 位置的上傳量中位數、上傳量預測值、下載量中位數、下載量預測值，供使用者比較實際值與預測值的差異，協助發現不正常的流量值。

圖 3: 多元回歸分析結果

決策樹分析結果頁面圖如圖 4 所示。該頁面會呈現目前該小時時段的上傳量分析樹狀規則圖以及下載量分析樹狀規則圖，供使用者作為判別該時段內影響上傳量與下載量之重要影響因素。

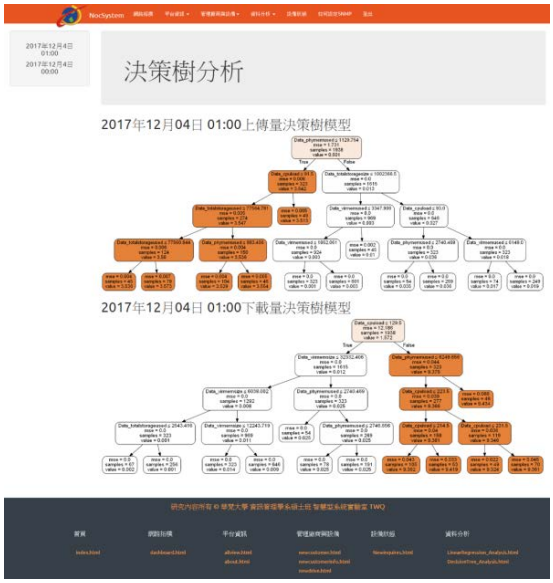

圖 4: 決策樹分析結果

4.1 多元回歸分析結果

本研究依據欄位說明敘述資料以每小時執行一次進行多元回歸分析，取當下時間前 23 小時作為訓練資料，第 24 小時做為測試資料，透過多元回歸函數建立模型得出上傳量預測方程式。上傳量預測方程式取自 2017 年 11 月 23 日 0 時上傳量預測方程式係數表如表 1 所示。建立模型結果與成效透過測試資料分析後上傳量預測方程式均方根誤差為 0.03334、決定係數為 0.9994。

表 1: 2017 年 11 月 23 日 0 時上傳量預測方程式係數表

項目	變數	係數	備註
0	intercept	-1.32209126131	截距
1	Data_cpucore	2.42434709927	處理器核心數
2	Data_cpuload	0.000282617853941	處理器使用量
3	Data_virmemsize	-0.0000196194664783	虛擬記憶體可用量
4	Data_virmemused	0.0000419291818093	虛擬記憶體使用量
5	Data_phymemsize	0.000916843124622	實體記憶體可用量
6	Data_phymemused	-0.0000163019175467	實體記憶體使用量
7	Data_totalstoragesize	0.0000184566564465	儲存硬體可用量
8	Data_totalstorageused	-0.000000238398990513	儲存硬體使用量

4.2 決策樹分析結果

本研究進行決策樹分析，依據欄位說明敘述資料以每小時執行一次資料分析，取當下時間前 23 小時作為訓練資料，第 24 小時做為測試資料，經多次運算並調整相關之參數值以最大深度(Max depth)設定為 4，形成葉節點上的最小樣本數量(Min samples leaf)為 45，輸出變量為上傳量，建立規則模型得出上傳量決策樹。

上傳量決策樹規則取至 2017 年 11 月 23 日 0 時，上傳量規則表如表 2 所示。共計獲得 16 條規則的決策樹建立模型結果與成效透過測試資料分析後獲得均方根誤差為 0.03201，決定係數為 0.99944。

表 2: 2017 年 11 月 23 日 0 時上傳量決策樹規則表

編號	規則	結果
1	Data_totalstoragesize <= 994567.875 Data_cpuload <= 4.5 Data_virmemused <= 1273.531	Flow_outflow =0.004
2	Data_totalstoragesize <= 994567.875 Data_cpuload <= 4.5 Data_virmemused > 1273.531	Flow_outflow =0.006
3	Data_cpuload <= 68.5 Data_totalstoragesize <= 994567.875 Data_cpuload > 4.5 Data_phymemused <= 1264.057	Flow_outflow =0.013
4	Data_cpuload <= 68.5 Data_totalstoragesize <= 994567.875 Data_cpuload > 4.5 Data_phymemused > 1264.057	Flow_outflow =0.012
5	Data_cpuload <= 68.5 Data_totalstoragesize> 994567.875 Data_totalstorageused<= 115426.047	Flow_outflow =0.024
6	Data_cpuload <= 68.5 Data_totalstoragesize> 994567.875 Data_totalstorageused<= 123801.469 Data_totalstorageused > 115426.047	Flow_outflow =0.028
7	Data_totalstoragesize> 994567.875 Data_totalstorageused > 123801.469 Data_cpuload <= 0.5	Flow_outflow = 0.023
8	Data_cpuload <= 68.5 Data_totalstoragesize> 994567.875 Data_totalstorageused > 123801.469	Flow_outflow = 0.022

	Data_cpuload > 0.5	
9	Data_cpuload > 68.5 Data_totalstorageused<= 163583.625 Data_phymemused <= 1272.906	Flow_outflow =3.407
10	Data_cpuload > 68.5 Data_totalstorageused<= 163583.625 Data_phymemused <= 1277.312 Data_phymemused > 1272.906	Flow_outflow =3.437
11	Data_cpuload > 68.5 Data_phymemused >1277.312 Data_totalstorageused <= 78322.016	Flow_outflow =3.394
12	Data_cpuload > 68.5 Data_totalstorageused<= 163583.625 Data_phymemused >1277.312 Data_totalstorageused > 78322.016	Flow_outflow =3.415
13	Data_cpuload > 68.5 Data_totalstorageused> 163583.625 Data_phymemused <= 6860.562 Data_virmemused <= 11918.0	Flow_outflow =0.032
14	Data_cpuload > 68.5 Data_totalstorageused> 163583.625 Data_phymemused <= 6860.562 Data_virmemused > 11918.0	Flow_outflow =0.017
15	Data_cpuload > 68.5 Data_totalstorageused> 163583.625 Data_phymemused > 6860.562 Data_cpuload <= 117.5	Flow_outflow =0.044
16	Data_totalstorageused> 163583.625 Data_phymemused > 6860.562 Data_cpuload > 117.5	Flow_outflow =0.109

經過分析後發現規則 1 至規則 8 儘管處理器使用量小於 68.5 且設備的儲存空間規格是小於等於或大於 994567.875(約 1TB)，上傳量的變化最大也只到 0.028，這表示不論設備為業務設備或單一功能的設備下，皆為設備待機時的網路上傳流量。造成此上傳流量現象可能為與交換器(Switch)或路由器(Router)之間的封包，或是作業系統中已預設的工作排程功能所造成的流量。規則 9 至規則 12 則發現當處理器使用量大於 68.5，儲存空間的使用量介於 78322.016 與 163583.625 之間時，上傳量最高會達 3.437。這表示為設備正在進行多工運作，造成此上傳流量現象可能為日常作業運作，包含使用遠端桌面功能進行工作、網路大量查詢或是執行需要網路傳輸的背景作業。規則 13 至規則 16 當處理器大於 68.5 時上傳流量最大卻只到 0.109，這表示設備正處於忙碌的運算，但卻不會造成過高的上傳量，可以將設備與上班時段交叉比對，可得知設備是否於正常的上班時段執行作業。

5.結論與建議

本研究在雲端技術、整合網路管理 SNMP 協定下，提供網路拓樸、設備狀態功能，且加入資料探勘中多元回歸預測與決策樹預測。其中上傳量多元回歸均方根誤差為 0.03334、決定係數為 0.9994。上傳量決策樹發現「設備待機」、「日常運作」與「處理器高負載」三種模式。藉由本平台可協助使用者可降低客戶端、網路設備端與資料分析端所帶來的問題，透過資料分析觀點找出網路管理問題與歸納出服務營運平台規則，規則所代表的意義也可做為設備工作效能評估時的參考建議。未來研究可建議至少取得一周資料來建立模組，新增其他網路協定日誌、新增其他分

析演算法、改善或新增功能與模擬測試，能納入更多直接反應設備狀態的因素加入分析，強化多元回歸函數與預測結果的準確性。

參考文獻

中文參考文獻

[1] 中央通訊社，「大容電腦等 10 家資訊服務業者通過新通關系統驗證」，http://www.cna.com.tw/postwrite/Detail/126137.aspx，發表日期：2013/05/15，檢索日期：2017/11/25。

[2]方世榮、張文賢，統計學導論，第六版，華泰文化事業股份有限公司，台灣台北市，2010 年 7 月，第 543-545 頁。

[3] 財政部關務署，「優質經貿網絡計劃」，https://web.customs.gov.tw/News_Content.aspx?n=1F7726156BC53908&sms=62F2C4D35690CE93&s=43E064D213EABC6E，發表日期：2013/11/12，檢索日期：2017/11/25。

[4]賴瑄，「基於 SNMP 之資通設備能源管理之研究」，國立暨南國際大學資訊管理學系碩士論文，2015 年 7 月。

[5]顏晟陞，「基於 SNMP 之 Hadoop 雲端運算平台」，國立暨南國際大學資訊管理學系碩士論文，2014 年 8 月。

[6]鐘揮雄，「應用於智慧電網管理之通訊協定 SIP_SNMP 效能量測研究」，國立暨南國際大學資訊管理學系碩士論文，2012 年 7 月。

英文參考文獻

[7] B. Leo, F. Jerome, C.J. Stone and R.A. Olshen, Classification and Regression Trees, 1 edition, Chapman and Hall, London, UK, January 1984.

[8] C. Lidia and V. Paolo, “The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini,” The Journal of Economic Inequality, Vol. 10, Issue 3, pp.421-443, September 2012.

[9] D. Tweddle, “Logistics, security and compliance: the part to be played by Authorized Economic Operators (AEOs) and data management,” World Customs Journal, Vol. 2(1), pp. 101-105, April 2008.

[10] European Commission, “Authorized Economic Operators–GUIDELINES,” https://ec.europa.eu/taxation_customs/sites/taxation/files/resources/documents/customs/policy_issues/customs_security/aeo_guidelines_en.pdf, publication date: 2016/03/11, access date: 2017/11/25.

[11] H. Susan, “Root Mean Square Error,” http://statweb.stanford.edu/~susan/courses/s60/split/node60.html, publication date: 2000/11/28, access date: 2017/11/25.

[12] Internet Engineering Task Force (IETF), “RFC 1157- A Simple Network Management Protocol (SNMP),” https://www.ietf.org/rfc /rfc1157.txt, publication date: 1990/05/31, access date: 2017/11/25.

[13] J.K. Kim, H. S. Song, T. S. Kim, and H. K. Kim, “Detecting The Change of Customer Behavior Based on Decision Tree Analysis,” Expert Systems, Vol. 22, pp.193-205, September 2005.

[14] L.D. Jay, Probability and Statistics for Engineering and the Sciences, 8th edition, Cengage Learning, Boston, USA, January 2011.

[15] R.A. Johnson and D.W. Wichern, Applied Multivariate Statistical Analysis, 6 edition, Pearson, London, UK, April 2007.

[16] O.R. John, G.P. Sastry, A.D. David, Applied Regression Analysis: A Research Tool, 2 edition, Springer, Berlin, Germany, April 2001.