系統環境異常預警研究
以機器學習及自然語言處理實作

電機所博一 張浩祥 112.12

一、序言

系統環境異常檢測係現行系統維運監管中至關重要的一環，其中記錄系統運行時之日誌(Log)訊息被廣泛用於異常檢測。傳統上，系統開發、維運或操作人員經常以關鍵字搜索和定義異常規則匹配等方式手動檢查日誌，然而隨著系統進化，其規模及複雜性的增加，日誌的數量呈爆炸式增長，使得人工檢查變得越來越不可行。

本研究收集系統環境中多種日誌，經由多種資料清理，篩選出本案計 631 種有效日誌文本，再依詞袋模型(Bag of Words)、文檔摘要(Term Frequency - Inverse Document Frequency)及停用詞(Stop Words)等文字處理方法進行日誌資料分析，其後將分析後日誌資料導入機器學習建模應用，建立共包含 24 種機器學習預警模型；另外，考量系統日誌中含有許多描述性英文文字及詞語內容，其性質即為人類所用之語言描述，故同時依據資料清理之日誌文本進行自然語言處理(Natural Language Processing)之深度學習建模嘗試，建置理解日誌之描述意涵之自然語言預警模型，最後再比較 25 種模型判斷異常日誌之評估表現。

二、資料清理及分析

(一) 資料清理作業

本研究以分析經本行日誌收集平台正規化之微軟日誌(Windows Event Log)為主。

微軟日誌收集平台為結構化之資料，因來源及樣態不同，導致有許多無效描述之資料，本研究歷經多次無效資料清理，清理之資料如下：

1. 該資料項之列數空白值過多者。
2. 該資料項含有未含文字描述意義者(如代號等)。
3. 該資料項內容文字幾近相同者。

4. 該資料項過多雜訊或代號者。

經資料清理後，本研究實作之有效文字描述資料，計 631 種日誌文本，其中包含正常日誌文本資料計 541 種，異常日誌文本資料計 90 種，作為模型輸入資料。。

(二) 資料分析作業

本研究採取之無監督方法，係依自然語言處理之深度學習進行，其本身已含複雜之模型架構及複雜的數學模型演算法，輸入模型時直接依資料清理後之 631 種日誌文本資料即可，不需額外針對日誌資料進一步分析及轉換。

就機器學習而言，由於模型相對於自然語言處理較為單純，模型輸入前的資料分析更顯重要。考量本研究 631 種日誌文本含有許多描述性英文文字及詞語內容，較適合用文字獨特性、文本與文本間相關性等相關做法進行資料分析，茲列出使用到之資料分析方法如下：

1. 詞袋模型(Bag of Words，簡稱為 BoW)。

2. 文檔摘要(Term Frequency - Inverse Document Frequency，簡稱為 TF-IDF)。

3. 停用詞(Stop Words)。

本研究所用 631 種日誌文本，經由上開資料分析作業之交錯搭配，於機器學習建模資料載入前，產生共計 6 大類文本，各類文本含有 631 種轉換資料，並分別獨立進行建模，分為(1)詞袋模型；(2)詞袋模型及停用詞；(3)詞袋模型及文檔摘要(TF)；(4)詞袋模型、文檔摘要(TF)及停用詞；(5)詞袋模型及文檔摘要(TF-IDF)；(6) 詞袋模型、文檔摘要(TF-IDF)及停用詞。

三、建模評估標準與模型建置

(一) 模型準確率評估標準

如本文第二節所介紹之預警相關文獻，其中多使用 F-Score（亦被稱做 F-measure）作為量測指標，主因係針對二元判斷時，F-Score 挺適合，故採用 F-score 作為本研究之量測指標。

F-Score 是一種量測演算法之精確度常用的指標，經常用來判斷

演算法精確率（precision）和召回率（recall）[1]，F-score 能同時考慮這兩個數值，平衡的反映這個演算法的精確度，

(二) 監督式方法建模

1. 演算法選擇

本研究選用常用之機器學習演算法，共計 4 種：

(1) 貝式分類 (Bayesian Classifier)

(2) 邏輯迴歸(Logistic Regression)

(3) 隨機森林(Random Forest)

(4) 支援向量分類(Support Vector Classifier, SVC)

2. 機器學習建模

以機器學習監督式方法進行的系統環境異常預警建模，在實作細節方面，各項條件定義如下：

(1) 來源資料範圍：631 種日誌文本。

(2) 來源資料分類：資料分析後產生之 6 種獨立文本大類：A.詞袋模型；B.詞袋模型及停用詞；C.詞袋模型及文檔摘要(TF)；D.詞袋模型、文檔摘要(TF)及停用詞；E.詞袋模型及文檔摘要(TF-IDF)；F.詞袋模型、文檔摘要(TF-IDF)及停用詞。

(3) 來源資料分群：分 80%訓練、驗證資料及 20%測試資料；其中正常日誌文本、異常日誌文本為隨機分配。

(4) 採用前述 4 種演算法獨立建模：A.貝式分類；B.邏輯迴歸；C.隨機森林；D.支援向量分類。

(5) 各演算法建模採用預設參數設定。

(6) 建模驗證採用交叉驗證(Cross Validation)之 3 次平均(CV=3)。

(7) 各模型評估標準採用 F1-Score。

[1] 精確率(precision)和召回率(recall)用於機器學習模型辨識、資訊索引和分類中判斷是否準確之性能指標。精確率為模型判斷為實例項目中，事實上真的為實例項目之比例；召回率為事實上為實例項目中，真正被模型判斷出來的比例。

因來源資料計 6 大類，每類 4 種演算法之建模，共計有 24 種預警模型，建模流程如圖 1。

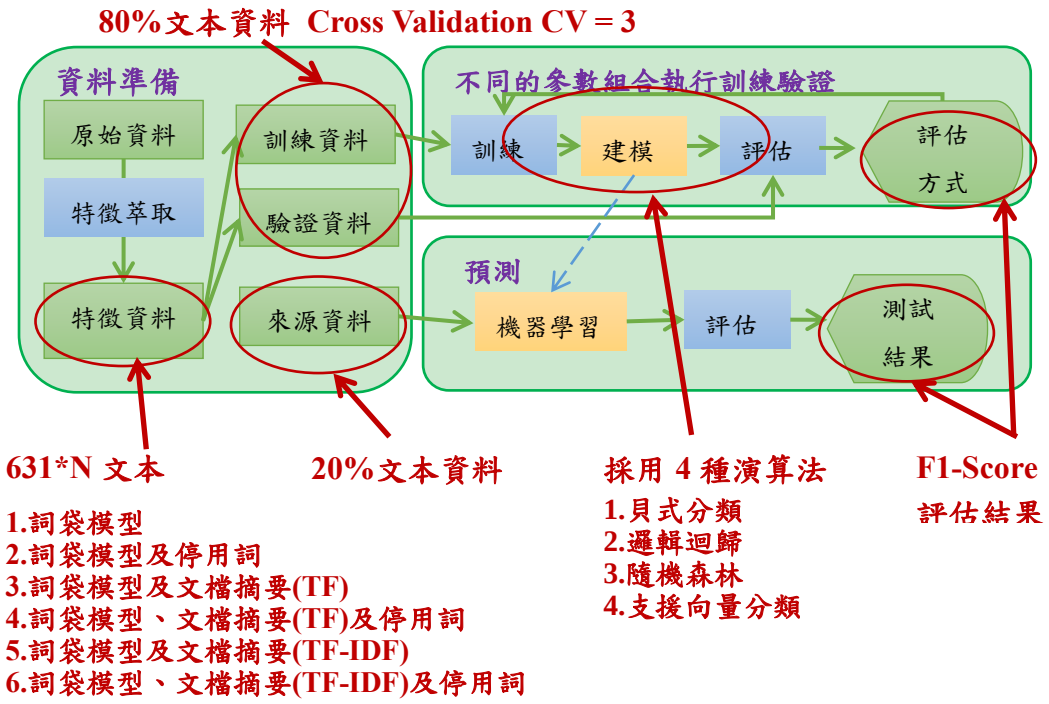

圖 1：機器學習建模流程圖

(三) 無監督方法建模

1. 模型選擇

本研究採用 Google 於發布名稱為「基於變換器的雙向編碼器表示技術」(Bidirectional Encoder Representations from Transformers，簡稱 BERT)之自然語言處理中可理解各式搜尋語句及了解人類言語溝通之技術，該模型之一般模型計有 12 個神經系統層、每層 768 種維度；各種維度中包含 12 個自我訓練神經元(Self-Attention Head)，總共有 1 億 1 百多萬參數的神經網路結構。

2. 自然語言處理深度學習建模

無監督方法之系統環境異常預警建模之實作流程如圖 2，各項建模細節條件定義如下：

(1) 來源資料範圍：631 種日誌文本。

(2) 來源資料分群：分 80%訓練、驗證資料及 20%測試資料；其中正常日誌文本、異常日誌文本為隨機分配。

(3) 模型之參數及各階段函式選擇，皆採用預設設定。

(4) 模型評估標準採用 F1-Score。

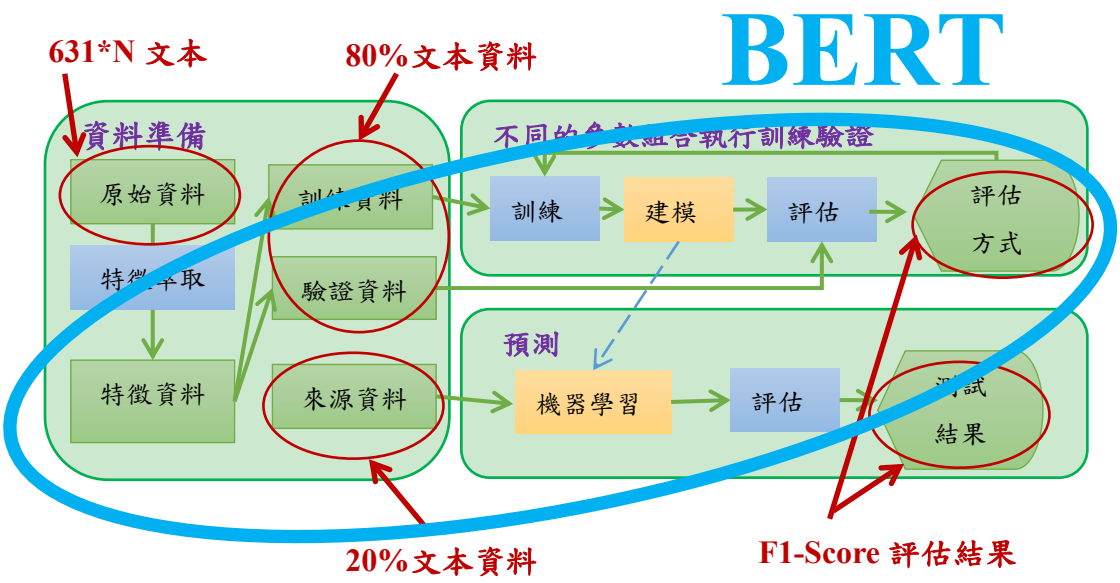

圖 2：自然語言處理深度學習建模流程圖

四、模型實作及結果比較

(一) 各模型預警準確率

本研究各模型建模成果如表 1，其中自然語言 BERT 模型之 F1-Score 準確率可達到 0.92，顯然優於所有機器學習所建之演算法，原因可能為 BERT 模型本身包含許多複雜之類神經結構，其模型本身設計時，即以處理描述性文字或語言為主，且模型架構所耗之軟硬體運行資源亦較高。

資料分析\演算法	貝式分類	隨機森林	邏輯迴歸	支援向量分類
詞袋模型	0.76	0.73	0.75	0.46
詞袋模型及停用詞	0.72	0.76	0.74	0.46
詞袋模型及文檔摘要(TF)	0.57	0.75	0.60	0.46

詞袋模型、文檔摘要(TF)及停用詞	0.59	**0.79**	0.61	0.46
詞袋模型及文檔摘要(TF/IDF)	0.55	0.78	0.67	0.46
詞袋模型、文檔摘要(TF/IDF)及停用詞	0.58	0.72	0.66	0.46
自然語言處理深度學習模型(BERT)	**0.92**			

表 1：各系統預警模型之 F1-Score 數值

另外觀察 24 種機器學習演算模型中，以隨機森林、文檔摘要(TF)及停用詞模型之準確率最高，F1-Score 為 0.79；其餘模型中，貝式分類和邏輯迴歸演算法下加入文檔摘要，無論是否有包含 IDF 或僅有 TF，預警表現反而遜於未加入者；此外，支援向量分類演算法下所有模型預警表現皆不及其他模型。

七、結論

(一) 系統日誌預警，自然語言處理之無監督方法表現較監督式方法好。

(二) 自然語言處理等深度學習建模研究，係注重成果而非可解釋性。

(三) 提供更多有效日誌及結合專業知識分析資料，可提高預警表現。

參考文獻

Wibisono, Okiriza, Hidayah Dhini Ari, Anggraini Widjanarti, Alvin Andhika Zulen, and Bruno Tissot (2019), “The Use of Big Data Analytics and Artificial Intelligence in Central Banking,” IFC Bulletin No 50.

Lin Yang, Junjie Chen, Zan Wang, Weijing Wang, Jiajun Jiang, Xuyuan Dong and Wenbin Zhang (2021), “Semi-Supervised Log-Based Anomaly Detection via Probabilistic Label Estimation”, 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), PP.1448-1460.

Rakesh Bahadur Yadav, P Santosh Kumar and Sunita Vikrant Dhavale (2020), “A Survey on Log Anomaly Detection using Deep Learning”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), PP.1215-1220.

Markus Wurzenberger, Florian Skopik, Max Landauer, Philipp Greitbauer, Roman Fiedler and Wolfgang Kastner (2017), “Incremental Clustering for Semi-Supervised Anomaly Detection applied on Log Data”, Proceedings of the 12th International Conference on Availability, Reliability and Security.

Amir Farzad T and Aaron Gullivert (2020), “Unsupervised log message anomaly detection”, ICT Express, Volume 6, Issue 3, PP.229-237.

Atscale (2019), “2018 Big Data Maturity Survey,” URL: https://cdn2.hubspot.net/hubfs/488249/AtScale_2018MaturitySurveyReport.pdf