系統環境預警研究案
以機器學習及自然語言處理實作

112年12月

電機所博一

張浩祥

簡報大綱

- 緣起
- 系統日誌異常預警相關文獻
- 系統環境預警建模
- 異常預警建模成果
- 結論
- 參考文獻

緣起

- 系統環境異常檢測係現行資訊服務維運監控中至關重要的一環，近年來人工智慧如機器學習、深度學習等技術普遍用於系統維運及系統異常判定等相關應用。

- 本研究係使用之資料以經本行日誌收集平台正規化之微軟日誌**(Windows Event Log)**為主；經由資料清理，篩選出有效日誌文本；再進行日誌資料分析(Pre-Train)，轉換後導入**機器學習**建模應用；同時以**自然語言處理**之深度學習建模，最後再比較模型判斷異常日誌之評估表現。

緣起(續)

序號	日誌文本	正常(0)/異常(1)
1	Certificate Services template security was updated.	0
2	The Per-user audit policy table was created.	0
3	An attempt was made to register a security event source.	0
4	An attempt was made to unregister a security event source.	0
5	The CrashOnAuditFail value has changed.	0
6	Auditing settings on object were changed.	0
7	Special Groups Logon table modified.	0
8	The local policy settings for the TBS were changed.	0
9	The group policy settings for the TBS were changed.	0
10	Resource attributes of the object were changed.	0
11	Per User Audit Policy was changed.	0
12	Central Access Policy on the object was changed.	0
13	An Active Directory replica source naming context was established.	0
14	An Active Directory replica source naming context was removed.	0
15	An Active Directory replica source naming context was modified.	0
16	An Active Directory replica destination naming context was modified.	0
17	Synchronization of a replica of an Active Directory naming context has begun.	0
18	Synchronization of a replica of an Active Directory naming context has ended.	0
19	Attributes of an Active Directory object were replicated.	0
20	Replication failure begins.	1
21	Replication failure ends.	1
22	A lingering object was removed from a replica.	0
23	The following policy was active when the Windows Firewall started.	0

系統日誌異常預警相關文獻

- Lin Yang et al. (2021) 以**非監督式**方法對於日誌的異常檢測，以減少耗時的手動標記，依自我標籤機制評估及結合過往異常的經驗參考，所建模型對即時異常日誌有不錯的預警結果。

- Rakesh Bahadur Yadav et al. (2020) 使用**深度學習**之類神經網絡建模檢測日誌異常，針對即時性資料預警，文獻結論顯示該研究模型有相當程度的預警可靠性。
- Amir Farzad T. and Aaron Gulliver (2020)運用機器學習及自然語言處理偵測日誌異常之模型，多以**非監督式**方法建模為主要趨勢，主要係因**非監督式**方法具有無需通過日誌訊息、排除人為誤植項目、巨量資料處理優點。

- Wibisono et al., 2019與傳統預警方法相較，機器學習及自然語言處理不講求可解釋性、以結果論成敗。

- Markus Wurzenberger et al. (2017) 運用**半監督式學習**於日誌資料大幅增量及聚類之異常檢測，相較於傳統式方法，亦有更高的準確率。

自然語言處理– Bag of Words

- **One-hot encoding**：單詞當成一個維度，若總共有十萬個單詞，那就會有十萬個維度。one-hot encoding 使用稀疏方式儲存，**彼此之間獨立**，常見的技術為BOW(bag of words)， BOW常用於簡單之**輿情分析、影評**。

  - Review 1: This movie is very scary and long
  - Review 2: This movie is not scary and is slow
  - Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good
Review 1	1	1	1	1	1	1	1	0	0	0	0
Review 2	1	1	2	1	1	1	1	1	1	0	0
Review 3	1	1	1	0	0	1	1	0	0	1	1

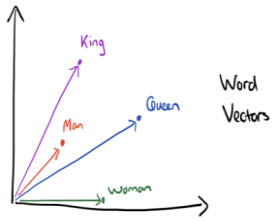
	Lable
Vector of Review 1: [1 1 1 1 1 1 1 0 0 0 0]	0
Vector of Review 2: [1 1 2 0 0 1 1 0 1 0 0]	0
Vector of Review 3: [1 1 1 0 0 0 1 0 0 1 1]	1

自然語言處理(續) – Word Embedding

- **Word Embedding** 通常會已屬於非監督式學習（Unsupervised Learning）實作，無需標記數據，透過 cluster 單詞來學習，自動完成關聯。

- Word Embedding有許多種實作方法，較廣為人知的基本作法為Word2Vec，其源自於 Tomas Mikolov發表論文Efficient Estimation of Word Representations in Vector Space中提到的一種用來表示每個字的方法。

- Word2Vec是google據上開方法實作之NLP概念，特點為將所有的詞向量化，可據以度量詞與詞間之關係，一個詞彙的意義，或許可以用身邊的詞(Context)去表示他。 概念如自己身邊的朋友們，可以反映出自己是個怎麼樣的人一樣。



自然語言處理(續) – BERT(非監督式學習)

- **BERT: 基於變換器的雙向編碼器表示技術**（英語：Bidirectional Encoder Representations from Transformers，BERT）是用於自然語言處理（NLP）的預訓練技術，由Google提出。

- 2018年Google建立並發布了BERT。並利用BERT來更好地理解使用者搜尋語句的語意。

- 最初的英語BERT發布時提供兩種類型的預訓練模型：
  - BERTBASE模型，一個12層，768維，12個自注意頭（self attention head），110M參數的神經網路結構。
  - BERTLARGE模型，一個24層，1024維，16個自注意頭，340M參數的神經網路結構。兩者的訓練語料都是BooksCorpus以及英語維基百科語料，單詞量分別是8億以及25億。

系統環境預警建模－資料整理及分析

- 本研究係使用之資料以經本行日誌收集平台正規化之微軟日誌**(Windows Event Log)**。
- 自然語言之演算法建模實作，必須提供**大量之有效之描述性文字**
- 本案實作資料清理歷程及嘗試：
  - **欄位空值(Null)過多者**
  - **欄位無文字描述意義者**
  - **欄位內容文字幾近相同者**
  - **欄位過多雜訊或代號者**

序號	日誌文本	正常(0)/異常(1)
1	Certificate Services template security was updated	0
2	The Per-user audit policy table was created	0
3	An attempt was made to register a security event source	0
4	An attempt was made to unregister a security event source	0
5	The CrashOnAuditFail value has changed	0
6	Auditing settings on object were changed	0
7	Special Groups Logon table modified	0
8	The local policy settings for the TBS were changed	0
9	The group policy settings for the TBS were changed	0
10	Resource attributes of the object were changed	0
11	Per User Audit Policy was changed	0
12	Central Access Policy on the object was changed	0
13	An Active Directory replica source naming context was established	0
14	An Active Directory replica source naming context was removed	0
15	An Active Directory replica source naming context was modified	0
16	An Active Directory replica destination naming context was modified	0
17	Synchronization of a replica of an Active Directory naming context has begun	0
18	Synchronization of a replica of an Active Directory naming context has ended	0
19	Attributes of an Active Directory object were replicated	0
20	Replication failure begins	1
21	Replication failure ends	1
22	A lingering object was removed from a replica	0
23	The following policy was active when the Windows Firewall started	0

系統環境預警建模(續)– 資料整理及分析

- 本研究進行Bag of Words及衍伸演算法建模之資料分析(Pre-Train)方式：
  - Pure Bag of Words
  - Pure Bag of Words + Stop Words
  - Bag of Words (TF)
  - Bag of Words (TF) + Stop Words
  - Bag of Words (TF/IDF)
  - Bag of Words (TF/IDF) + Stop Words

*TF/IDF: 某一特定文章內的高詞語頻率，以及該詞語在整個文章集合中的低檔案頻率，可以產生出高權重的tf-idf。

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

*Stop Words: 自然語言處理中，為節省存儲空間和提高搜索效率，在資料（或文本）之前或之後會**自動過濾掉某些太常見、普遍並常無代表性之字或詞**，稱為**Stop Words(停用詞)**。

系統環境預警建模(續) – 模型評估標準

- F-score（亦被稱做F-measure）是一種**量測方法的精確度常用的指標**，經常用來判斷演算法**精確率（precision）**和**召回率（recall）**，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。beta =1時的F-score，亦有寫作F1-score。

- 若**precision或recall趨近於0，F-score就會趨近於0**，代表著這個演算法的精確度非常低；**F-score最理想的數值是趨近於1**。

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F - score = \frac{(1 + \beta^2)\, precision \times recall}{\beta^2\, precision + recall}$$

	判斷為真	判斷不為真
事實上為真	TP	FN
事實上不為真	FP	TN

異常Precision: Log在模型判斷為異常預警的情況，實際真的異常的機率。

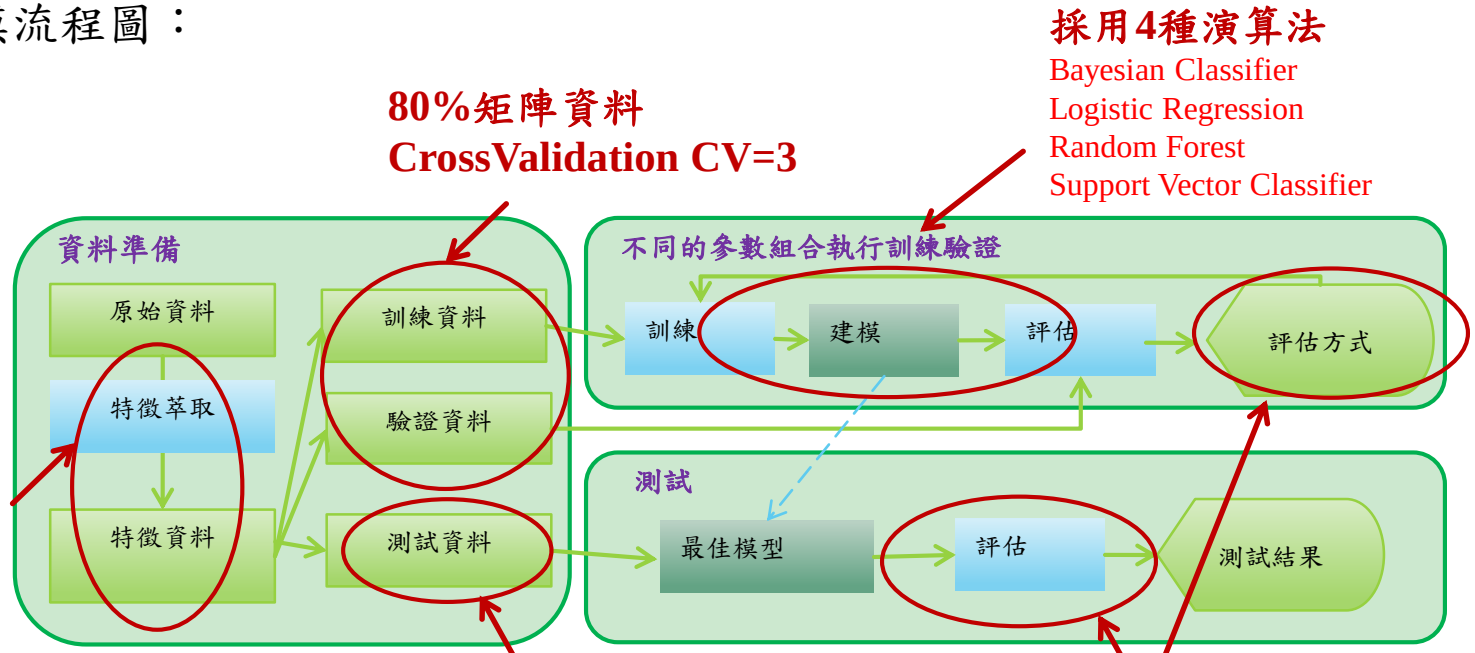
異常Recall: Log實際為異常預警的情況下，模型也偵測出異常的機率。

系統環境預警建模(續)－機器學習(監督式學習)建模

- 本研究建模流程：
- **Training Data/Validation Data**: 80% of feature
- **Test Data**: 20% of feature **Validation**: Cross Validation CV fold = 3
- **Label**: Information, Warning/Alert label
- **Machine Learning Algorithm**:
  - Bayesian Classifier
  - Logistic Regression
  - Random Forest
  - Support Vector Classifier
- **Accuracy**: F1-Score
- ***共計24種output模型**

系統環境預警建模(續)－機器學習(監督式學習)建模

本研究建模流程圖：



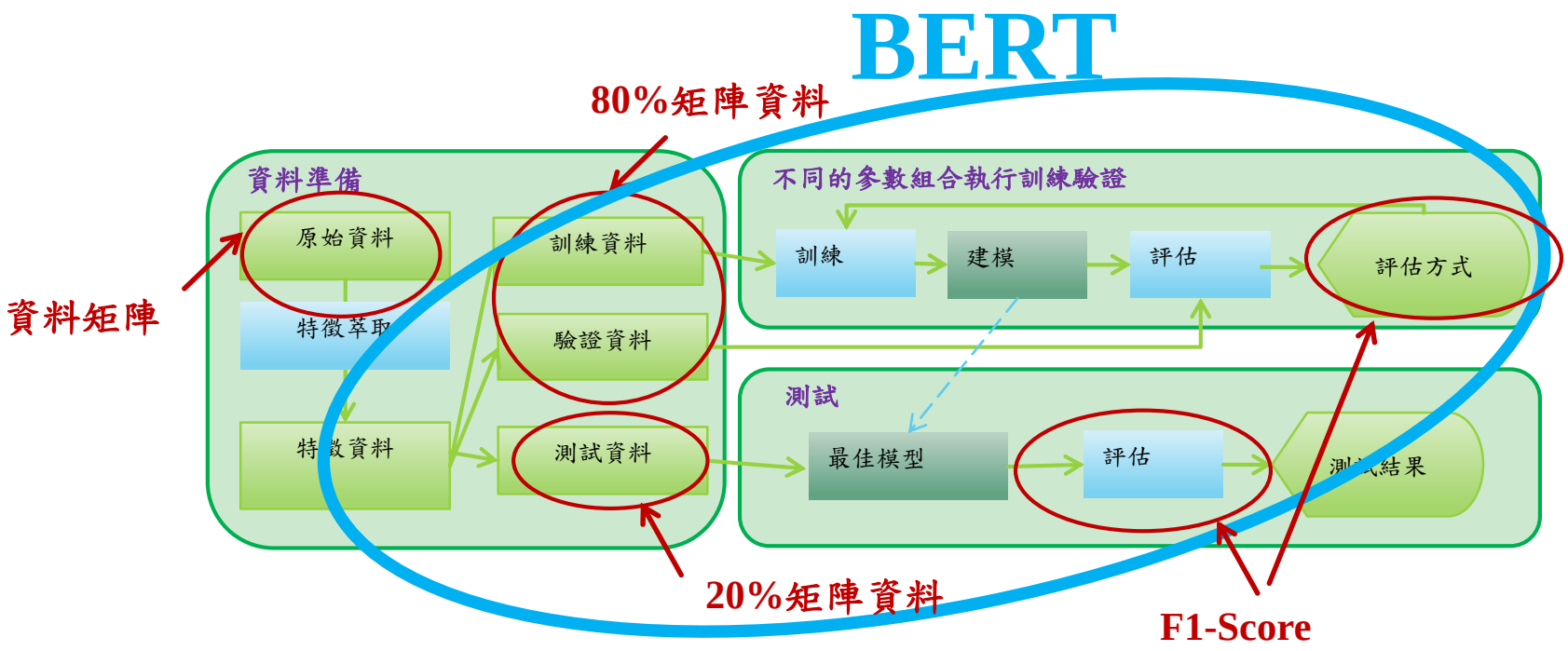
Pure Bag of Words
Pure Bag of Words + Stop Words
Bag of Words (TF)
Bag of Words (TF) + Stop Words
Bag of Words (TF/IDF)
Bag of Words (TF/IDF) + Stop Words

系統環境預警建模(續) –
深度學習(非監督式學習)建模 - BERT

- BERT建模：
  - Training & Validation Data/Test Data: 80%/20%
  - Activation Function: Softmax
  - Loss Function: Cross Entropy
  - Accuracy: F1 Score
  - Optimizer: Gradient Decent (Adam)
  - Back Propagation: 24 + 1(result)

系統環境預警建模(續)－深度學習(非監督式學習)建模 - BERT

本研究建模流程圖：



異常預警建模成果

- 各系統預警模型之F1-Score數值

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**			

異常預警建模成果(續)

- 本研究各模型綜合評估F1-Score之4項計算標準，**自然語言BERT模型之F1-Score準確率可達到0.92**，顯然優於以機器學習演算法所建之所有模型，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。

- 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。

結論

- **本研究系統日誌預警，自然語言處理之非監督式方法表現較機器學習等監督式方法。**
  - BERT本身為Google已完整訓練好之模型(Well-Trained Model)，內部已含複雜之類神經結構，且以處理描述性文字或語言為主，運行時所耗之軟硬體資源比重亦高。

- **自然語言處理模型已有一定成熟度的發展。**
  - 人工智慧及深度學習等實作可多考量採用已有之模型，如同站在巨人的肩膀上，可事半功倍。

結論(續)

- **自然語言處理等深度學習建模，係注重成果是否有效，而非模型本身之可解釋性。**
  - 自然語言處理為深度學習實作方法，係屬黑盒子理論，其模型內部運作內容無法完全透明化，模型之可解釋性較低，預警是否達到效果之重要關鍵係成果，非認定模型可解釋性及日誌之因果關係，亦非講求時序重要性。

- **若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。**
  - 本研究所用預警之文檔日誌資料僅631筆，用於建模略顯不足；另外觀察各模型針對實際異常之判斷，預警表現最好之BERT模型也僅達到F1-Score之7成準確率。
  - Atscale(2019)大數據成熟度調查報告顯示，機器學習、深度學習於資料分析時能若能結合多項專業知識、領域深入或聘用專家分析資料，模型預警的準確率，亦可更貼近實情。

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.
- 李弘毅(2021), “BERT、Self Attention及Word Embedding,” URL: https://speech.ee.ntu.edu.tw/~hylee/ml/2021-spring.html