巨量資料分析技術與應用

Term Project Report

第 12 組

309551*** 朱○毅

309551*** 王○憲

309551*** 張玟棋

309553*** 陳○安

組員分工與各自執行細項

(Team members & Task allocation)

309551*** 朱◯毅：SVM、Presentation、Report

309551*** 王◯憲：Naive Bayes、Presentation、Report

309551*** 張玟棋：Preprocessing、Analysis、RandomForest、Presentation

309553*** 陳◯安：Logistic Regression、Presentation、Report

一、 計畫目標的問題

(Target problem)

DoS 為一種嘗試對目標系統（如網站、應用程式等）進行惡意攻擊的行為，通常攻擊者會產生大量的封包或請求，最終使得目標系統無法負荷。而 DDoS 中第一個 D 意指「分散式」，即攻擊者會使用多個盜用或受控的來源來產生 DoS 攻擊。

防範 DDoS 的方法主要有兩種，一種是以流量限制。先預留一點頻寬或系統資源，當流量高過門檻時，對整體流量進行限制，但這個方法的缺點是正常的使用者也會因此受到影響；另一種方法則是過濾出有問題的 IP，直接將進行惡意攻擊的來源加入黑名單。我們希望透過巨量資料分析的方式找出進行 DDoS 的特徵，藉此將 DDoS 造成的傷害降到最低。

二、 選用的資料集描述與觀察

(Descriptions of selected datasets, including the characteristics in terms of Big Data)

我們選用的資料集名稱是 DDoS Dataset，來源為 Kaggle 網站 (https://www.kaggle.com/devendra416/ddos-datasets)。這個資料集的大小為 3.84GB，總共有 85 個欄位，最後一個欄位是 label，其餘 84 個欄位中包含了 Source 和 Destination 的 IP 以及 Port、Protocol、Total Forward Packet 等等，其中

除了 IP 和 Timestamp 之外的資料皆是純數字。另外，在資料集中有遭受 DDoS 攻擊和沒有遭受 DDoS 攻擊的資料比例是 2 ： 8。

關於這個資料集是否符合巨量資料 3V 的特性，我們認為，此 DDoS Dataset 中總共有 760 多萬筆資料，資料量夠大，而我們必須從高達 84 種的屬性中找出數據之間的關聯性，因此也有符合資料的多元性。但由於此資料集並沒有與相關系統串聯，所以資料內容是固定的，不具有即時性。

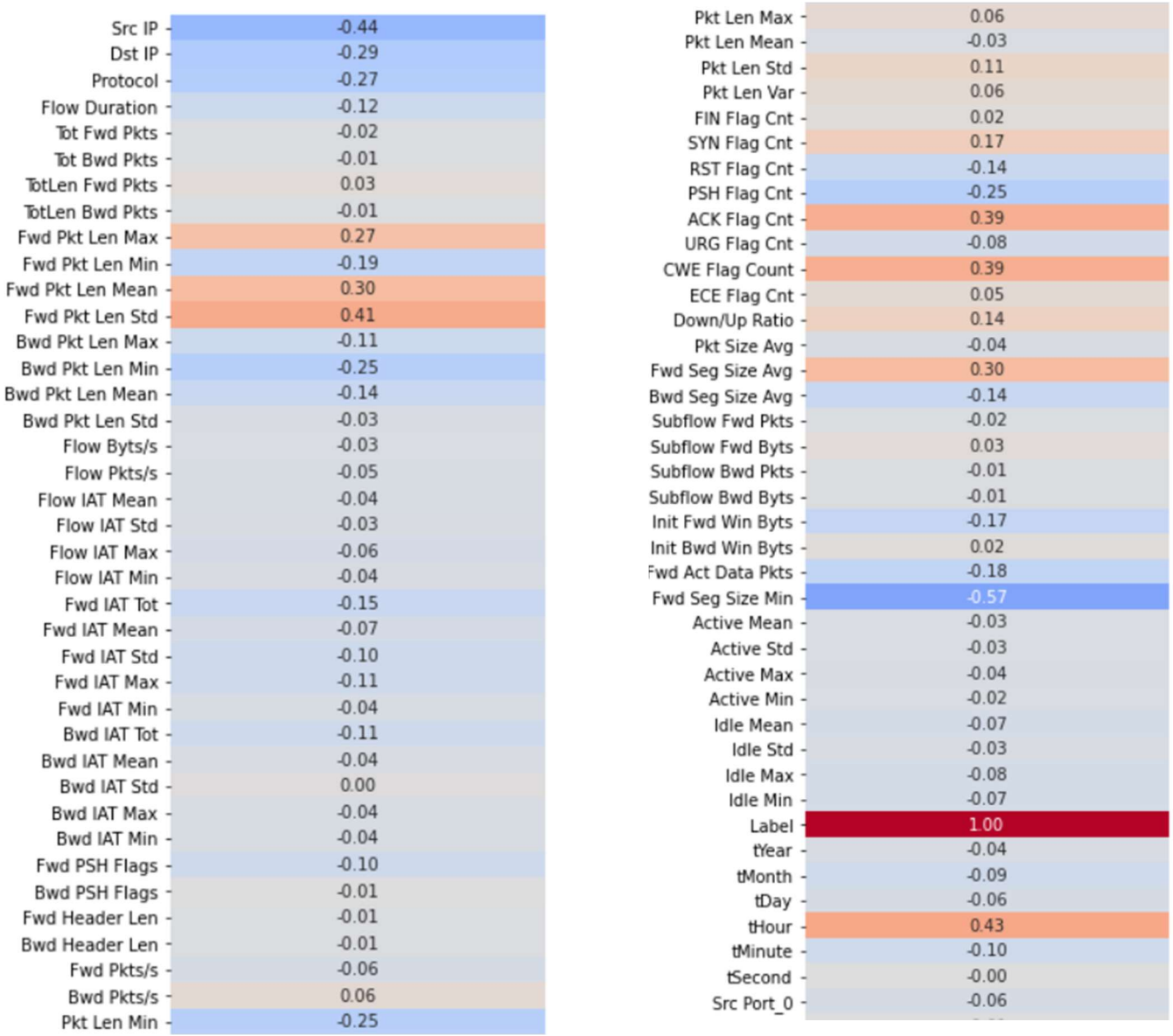
三、 針對問題設計的分析流程

(Analysis workflow)

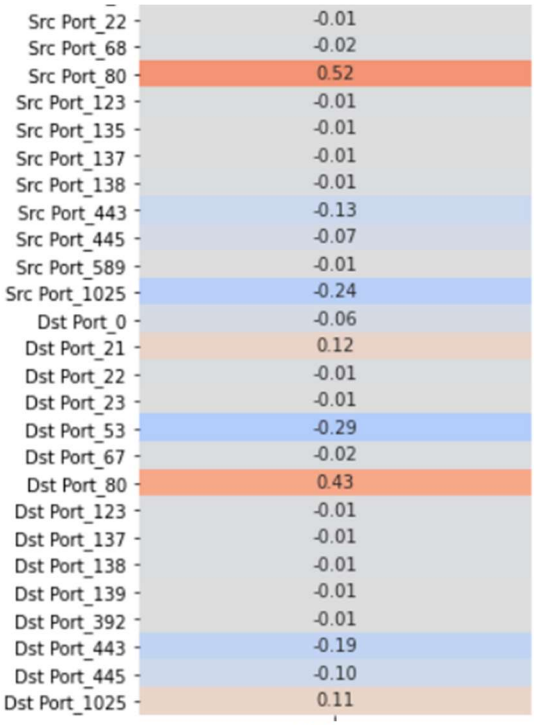
主要分成 4 個步驟：「Preprocessing」、「Analysis」、「Model Training」和「Test & Evaluation」。

Preprocessing 顧名思義，為資料的初步處理。當資料讀進來時，我們會先刪除含有 null 的 row，之後開始處理 Timestamp、IP、Port 這 3 個 column。Timestamp 依其格式切成「年」、「月」、「日」、「小時」、「分鐘」、「秒」這 6 個 column；IP 則是將其數值化，即將「A.B.C.D」化成「A*256*256*256 + B*256*256 + C*256 + D」；而 Port 因為數字相近的 Port 不一定有相近的功能，我們覺得將其當單純的數值資料會有問題，這樣造成 DDoS 的 Port 附近的其他 Port 有可能也會被誤判成高風險的 Port，故作 One-Hot Encoding，而 1025 以上的 Port 不一定為固定應用所用，故統一化作 1025，當成不明應用的 Port。接著刪除不必要的 column，如原本的 Timestamp、Port 等，以及「Fwd Byts/b Avg」、「Fwd Pkts/b Avg」等不會有任何數值變化的 column，最後對每個 column 作 Normalization（MinMaxScalar）。

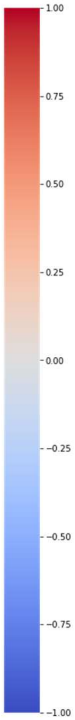
Analysis 主要是要做資料的降維，由於 feature 太多，後續的 Model Training 會變得很花時間，因此我們要做資料的降維。我們利用 Pyspark 的內建函式，算出現有的 feature 與 Label 之間的相關係數（結果如下頁的圖一~三，圖四為塗色的標準，用以判斷何者的相關係數較高），並挑選相關係數高的 feature 作為我們要使用的 feature（挑選標準為相關係數的絕對值大於 0.2），最後選出了 18 個 feature。


Src IP -0.44
Dst IP -0.29
Protocol -0.27
Flow Duration -0.12
Tot Fwd Pkts -0.02
Tot Bwd Pkts -0.01
TotLen Fwd Pkts 0.03
TotLen Bwd Pkts -0.01
Fwd Pkt Len Max 0.27
Fwd Pkt Len Min -0.19
Fwd Pkt Len Mean 0.30
Fwd Pkt Len Std 0.41
Bwd Pkt Len Max -0.11
Bwd Pkt Len Min -0.25
Bwd Pkt Len Mean -0.14
Bwd Pkt Len Std -0.03
Flow Byts/s -0.03
Flow Pkts/s -0.05
Flow IAT Mean -0.04
Flow IAT Std -0.03
Flow IAT Max -0.06
Flow IAT Min -0.04
Fwd IAT Tot -0.15
Fwd IAT Mean -0.07
Fwd IAT Std -0.10
Fwd IAT Max -0.11
Fwd IAT Min -0.04
Bwd IAT Tot -0.11
Bwd IAT Mean -0.04
Bwd IAT Std 0.00
Bwd IAT Max -0.04
Bwd IAT Min -0.04
Fwd PSH Flags -0.10
Bwd PSH Flags -0.01
Fwd Header Len -0.01
Bwd Header Len -0.01
Fwd Pkts/s -0.06
Bwd Pkts/s 0.06
Pkt Len Min -0.25
Pkt Len Max 0.06
Pkt Len Mean -0.03
Pkt Len Std 0.11
Pkt Len Var 0.06
FIN Flag Cnt 0.02
SYN Flag Cnt 0.17
RST Flag Cnt -0.14
PSH Flag Cnt -0.25
ACK Flag Cnt 0.39
URG Flag Cnt -0.08
CWE Flag Count 0.39
ECE Flag Cnt 0.05
Down/Up Ratio 0.14
Pkt Size Avg -0.04
Fwd Seg Size Avg 0.30
Bwd Seg Size Avg -0.14
Subflow Fwd Pkts -0.02
Subflow Fwd Byts 0.03
Subflow Bwd Pkts -0.01
Subflow Bwd Byts -0.01
Init Fwd Win Byts -0.17
Init Bwd Win Byts 0.02
Fwd Act Data Pkts -0.18
Fwd Seg Size Min -0.57
Active Mean -0.03
Active Std -0.03
Active Max -0.04
Active Min -0.02
Idle Mean -0.07
Idle Std -0.03
Idle Max -0.08
Idle Min -0.07
Label 1.00
tYear -0.04
tMonth -0.09
tDay -0.06
tHour 0.43
tMinute -0.10
tSecond -0.00
Src Port_0 -0.06

圖一



圖二



圖三

圖四

Model Training 的部分，我們將資料隨機切割成 8：2，8 成為 Training data、2 成為 Testing data，然後將經前述處理的資料丟進我們挑選的 Model，我們使用了 Logistic Regression、SVM、Random Forest、Naive Bayes 等，前三者在文獻中的表現優異，而 Naive Bayes 表現不好，作為對照組。

Test & Evaluation 的部分，我們將資料丟進前述訓練好的 4 個 Model，並使用 Confusion matrix 來作為我們判斷 Model 好壞的標準，最後依 TP、TN、FP、FN 算出 Accuracy、Precision、Recall、F1-score。

對於這個資料的 VOLUME 與 VARIETY 的特性問題，我們是先以 1/1000 的資料來進行上述的步驟來減少每次重新讀取、處理的時間成本，當除了 Naive Bayes 的最終數據皆高於標準時，我們就將寫好的程式碼應用在整份資料集上，而我們藉由相關係數找出有用的 feature，使最後丟進 Model 的資料從 84 維降成了 18 維，降低了 Model Training 的時間。

四、 分析結果

(Analysis results)

表 1 各個 Model 的表現比較

	NB	SVM	LR	RF	目標
Accuracy	0.909	0.970	0.995	0.998	0.85
Precision	0.939	0.899	0.988	0.988	0.80
Recall	0.665	0.928	0.983	0.999	0.85
F1-score	0.779	0.913	0.986	0.994	X

選取的 18 個 feature：

「Src IP」、「Dst IP」、「Protocol」、「Fwd Pkt Len Max」、「Fwd Pkt Len Mean」、「Fwd Pkt Len Std」、「Bwd Pkt Len Min」、「Pkt Len Min」、「PSH Flag Cnt」、「ACK Flag Cnt」、「CWE Flag Count」、「Fwd Seg Size Avg」、「Fwd Seg Size Min」、「tHour」、「Src Port_80」、「Src Port_1025」、「Dst Port_53」、「Dst Port_80」

五、 過程中遭遇的挑戰與討論

(Discussions)

由於我們挑選出來的 feature 有些是會動態改變的，例如最大值、最小值、總數等等的資料，因此當我們使用此資料集建立出的模型時，需要等到流量或是頻寬達到一定的門檻後，才能判斷現有的流量是否為 DDoS 攻擊。

另一個值得探討的問題是根據我們參考的論文，Naive Bayes 的各項表現最多只會到 60%左右，而我們使用 Naive Bayes 做出來的結果只有 recall 比較低，其他項皆高於 90%。雖然我們使用的資料集和論文中使用的資料集不同，但不免還是有些疑慮。在經過仔細地檢查後，推論是因為這個資料集的特徵過於明顯，所以才導致我們實作出的各項表現都特別優異。

參考文獻

1. M. S. Elsayed, N. -A. Le-Khac, S. Dev and A. D. Jurcut, "DDoSNet: A Deep-Learning Model for Detecting Network Attacks," *2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Cork, Ireland, 2020, pp. 391-396, doi: 10.1109/WoWMoM49955.2020.00072
2. Lopez, Alma D., Asha P. Mohan, and Sukumaran Nair. "Network Traffic Behavioral Analytics for Detection of DDoS Attacks." *SMU Data Science Review* 2.1 (2019): 14.