國立台北大學資訊工程學系專題報告

Data Tendency Extraction and Adversarial Attack Resistance for Deep Learning Intrusion Detection Model based on Accurate Tendency

專題組員:覃毅翔、林奐呈、簡毅、吳振瑋

專題編號:PRJ-NTPUCSIE-111-011

執行期間:2022 年 09 月 至 2023 年 06 月

1. 摘要

由於深度學習模型的分類能力，以深度學習模型為基礎的入侵檢測模型已然成為了防禦大量網路攻擊的重要方式。這樣的入侵檢測模型需要依賴著高品質的資料集以訓練出高準確率分類模型。針對這個需求，特徵品質改進的方法是必要的手段，像是特徵選擇和特徵提取都是常用的方法，但這些方法都是針對特徵進行改良，且並未提供能指出所選特徵能提高分類精度的原因。因此本研究提出了一種新型的資料萃取方法，利用特徵與分類之間的關聯，計算並取代出原本的特徵值，不僅可以提供提高分類準確的原因，且提高資料的品質以提高分類的準確率及抗攻擊能力。

2. 簡介

2.1. 研製背景

由於深度學習模型已經展現出顯著的分類能力，以深度學習為基礎的入侵檢測模型已然成為防禦大量網路攻擊的有前景的解決方案。這些新的檢測模型依賴著高品質的資料集，以訓練高精度的分類模型。高品質的資料集代表著網路特徵與攻擊分類標籤有著密切關係，而深度學習模型則可以將這些密切關係作為其參數進行學習，利用這些參數進行對測試資料樣本進行正確的預測。因此高品質的資料集成為準確的深度學習入侵檢測模型的重要條件。

然而網路攻擊資料集的資料非常離散，包含許多不可預測的雜訊，這些特徵使得品質增加變得非常有挑戰性。傳統的特徵品質改進方法通常是針對特徵進行篩選，改良。但是，這些方法始終無法提供令人信服的原因來說明其方法能提高分類精度。

且攻擊者為了逃避基於深度學習的入侵檢測，會利用生成對抗攻擊(generate adversarial attack)來繞過檢測。FGSM(Fast Gradient Sign Method)是一種經典的對抗攻擊，它利用了深度學習的梯度(gradient)機制來誤導分類結果。基於深度學習的入侵檢測採用數字資料集，雖然 FGSM 一開始是為了圖像資料集設計的，但它也同樣適用於數字資料集，因此 FGSM 可以有效地使攻擊者避開入侵檢測模型進行入侵。並且，對抗樣本可以輕易地根據不同程度的梯度進行更改，並且對抗樣本與正常樣本之間的差異很難分辨。因此攻擊者可以輕鬆地生成不同的對抗樣本進行攻擊。

2.2. 相關研究文獻探討

2.2.1. 資料品質改進方式

特徵學習是一種經典的資料品質改進方法方法，用於選擇適合的特徵以進行分類，以提高入侵檢測的準確性。Tabu Search – Random Forest(TS-RF)[1]是一種基於 wrapper 的特徵選擇算法，用於入侵檢測。Tabu Search 在 TS-RF wrapper 中執行特徵搜索和權重分配，而 Random Forest 用做學習法。Aminanto[2]提出了一種特徵選擇方法，稱為 D-FES。它由人工神經網路、支持向量機(support vector machine)和 C4.5 決策樹(decision tree)的三種選擇方法組成。D-FES 從原始特徵和從 stack auto encoder 產生的附加特徵中選擇特徵。CorrCorr[3]是一種基於 multivariate correlation 的網路異常檢測系統特徵選擇方法，它採用 Principal Component Analysis 和 Pearson class label correlation 來選擇特徵。SM. Kasongo[4]提出了一種基於 Extraa Trees 和 information gain 的特徵選擇方法，用於網路入侵檢測。CorrACC[5]是一種 wrapper 技術，通過使用特定的機器學習分類器的 accuracy metric 來選擇有效特徵。這些方法都可以提高準確性，但無法提供直接的證明來解釋改善的效果。

2.2.2. FGSM(Fast Gradient Sign Method)

I.J. Goodfellow 提出了 FGSM[6]，FGSM 的目的是在訓練分類模型時，將其損失函數(loss function)最大化，與平常將損失函數最小化相反。FGSM 利用計算神經網路(neural network)訓練過程中的反向傳播(backpropagation)來計算梯度(gradient)，確認梯度方向，並沿著梯度方向移動，以對輸入的圖像的像素進行微小的修改來生成對抗樣本(adversarial sample)，以顯著地降低模型分類的準確度。

在進行攻擊檢測時，通常會使用原始樣本訓練分類模型，並利用訓練好的模型對對抗樣本進行分類，生成對抗樣本的的標籤(label)。我們通過反向傳播來計算梯度的 FGSM 公式如下：

$$\eta = \varepsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

在這項公式中 θ 表示權重(weight)參數，x 表示原始數據，y 表示 x 的分類標籤(class label)。J(θ, x, y)表示神經網路的損失函數。∇_x表示偏微分(partial derivative)，即反向傳播中使用的梯度值。Sign 函數用來確認梯度的方向。當 U > 0 時，Sign(U) = 1，當 U < 0 時，Sign(U) = -1。Epsilon 代表學習率(learning rate)，也是一個偏移值(offset value)。根據公式和神經網路的計算結果，可以計算出一個偏移值，或稱之為擾動(perturbation)。通過將擾動添加到原始樣本，可以生成對抗樣本。在對抗攻擊中可以使用 Epsilon 改變攻擊強度，較高的 Epsilon 對原始樣本的擾動較明顯。

2.3. 目標

為了逃避基於深度學習的入侵檢測模型，攻擊者透過 FGSM 生成對抗性攻擊樣本以繞過檢測，且為了對抗當前的攻擊檢測方式，攻擊者可以通過不同程度的攻擊生成對抗樣本。因此，目前的對抗性攻擊檢測方法不足以偵測對抗性攻擊。為了檢測對抗性攻擊，我們將提出新的對抗性攻擊檢測模型。

同時，基於深度學習的入侵檢測模型需要高品質的資料集，特別是涉及離散(discrete)及 noises issue 的數值樣本。特徵選擇可以改善資料集的品質，但無法解決 noises issue 及提供明確的指示。且目前的資料選擇方法無法替代當前的隨機選擇，因此對於檢測模型來說，針對這方面的改善方法會成為一個非常理想的解決方案。

3. 專題進行方式

3.1. Accurate Class Tendency(ACT)

我們提出了一個用於入侵檢測數值資料集的一種特徵萃取方法。為了評估資料樣本的品質，我們提出了對於資料樣本中每個特徵值的 Accurate Class Tendency(ACT)。根據樣本的分類標籤，與該標籤相關的特徵值與資料集中所有特徵值的比例表示特徵值朝向該標籤的可能性。

所以我們參考了使用特徵來篩選、作為分類依據的決策樹模型以及相關演算法，CART[7]及 C4.5[8]。

決策樹是一種常用的監督式學習演算法，用於分類和回歸問題。在決策樹中，每個內部節點表示一個特徵，每個分支代表一個特徵值，而每個葉子節點則代表一個分類標籤或回歸值。建立決策樹的過程是將資料集分割成越來越小的子集，直到子集內所有樣本都屬於同一個類別或具有相似的屬性。

CART(Classification and Regression Trees)算法是決策樹的一種演算法，可以用於分類和回歸問題。在 CART 算法中，每個節點都是二元的，即只有兩個分支。CART 算法選擇最優的特徵和最優的分割點來進行分割，通常使用基尼指數(Gini index)來評估特徵的重要性。

C4.5 算法是 ID3 算法的改進版，同樣用於分類問題。C4.5 算法使用信息增益比(Information Gain Ratio)作為特徵選擇的標準，避免了 ID3 算法中對取值較多的特徵有所偏好的問題。此外，C4.5 算法還可以處理缺失值的情況，並且在樹的建立過程中可以進行剪枝，以避免過度擬合(Overfitting)的問題。

綜上所述，決策樹演算法是常用的監督式學習演算法之一，可以用於分類和回歸問題。CART 算法和 C4.5 算法都是常見的決策樹演算法，使用不同的特徵選擇標準以及不同的分支方式。而 P 值是從這個想法延伸而來，將 CART、C4.5 算法中每個特徵的每個特徵值對應到子樹中類別的分佈情況進一步用來衡量資料集特徵的重要性，將這個分佈的比例定義為 P。

由於特徵值的 P 值在決策樹中已被廣泛使用，並證實是作為分類的有效依據，反映出了各特徵在資料集中分布離散的現況。所以我們由 P 值出發，假設我們有一個的資料集，標籤(label)被二分為無惡意(positive)和惡意(negative)兩種，分為 1 跟 0。首先，P 值為一個特徵值對該類別的傾向，也就是該特徵值出現在無惡意標籤(positive label)中的比例。對於資料集中的一個特徵值X_{ij}，N 代表資料集中的總樣本數，$N_{X_{ij}}$代表在資料集內X_{ij}的樣本數，$N_{X_{ij}}^{label}$代表該特徵值所對應到的特定 label 樣本數。則$P_{X_{ij}}$定義為：

$$P_{X_{ij}} = \frac{N_{X_{ij}}^{label}}{N_{X_{ij}}}$$

其中在 label 為惡意和非惡意的二分情況下，label 可以是 0 表示該樣本的標籤為無惡意(positive)， label 可以是 1 表示該樣本的標籤為惡意(negative)。這樣產生出的 P 值範圍會在 0 到 1 之間，P 值越大代表該特徵值的傾向越靠近該類別。因為 label 有兩個類別 0 和 1，因此會有兩個 P 值，兩者相加為 1。這裡取較大的 P 值，來代表$P_{X_{ij}}$。

因為 0.5 是二元分類的基礎機率值，我們提出類別傾向 ACT 值，來反映了 P 值與基礎機率的差異。我們定義 ACT 值為 P 值減去 0.5，來計算和基礎機率值 0.5 的距離。若 ACT 值為正值，代表該值類別分布傾向和 label 相同，反之若和 label 相反，ACT 值則為負值。若 ACT 值為 0，P 值和基礎機率值相同，類別傾向為中立。這樣可以利用正負號更清楚的判別該特稱值的傾向。

$$ACT_{x_{ij}} = P_{x_{ij}} - 0.5$$

最後，利用 ACT 值將資料集中的特徵值進行替換以進行，並對所有特值都進行相同操作。

3.2. 利用 DNN 進行驗證

深度神經網路(DNN, Deep Neural Net Work)是一種深度學習網路，常用來處理數值相關的任務。深度神經網路主要由輸入層，隱藏層，及輸出層所構成。

本次專題中，我們分別會對 IOT23[2]及 NB15[3]兩個資料集，分別計算 ACT 值並利用 DNN 對其進行驗證。網路架構的部分，我們使用三層隱藏層，每個隱藏層都包含一層 Linear，一層 BatchNormalize，一層 LeackyReLu。由於我們要將其作為二分類器觀察分類傾向，因此我們利用 Sigmoid 作為輸出層，將其輸出限制在 0 到 1 之間。 最後，分別利用兩個資料集的原始資料以及經過計算 ACT 的資料對 DNN 進行訓練，並觀察 ACT 是否能成功地將訓練的準確度提高。

3.3. 利用不同 Epsilon 的 FGSM 攻擊對模型進行驗證

由於攻擊者可以利用更改 eps 進行不同程度的 FGSM 攻擊，因此我們利用不同程度 eps 的 FGSM 攻擊來驗證，是否經過 ACT 值能有效降低攻擊的程度。

我們使用了 IoT-23 做為測試的資料集，以下是我們處理資料給模型使用的方法:

(1) 用 chunksize 分批讀入 labeled 檔: 這樣處理是因為避免超過記憶體，並將各輸出檔案分開避免寫檔速度過慢，並將不需訓練的 UID 和 history 欄位刪除和把缺值給補 0。也就是根據設定欄位將原始的 raw_dataset 輸出成我們常見的 csv 檔，在後續方便讀寫。

(2) 做 onehot-encoding: 為資料預處理的部分，使用 encode 的方始使特徵數擴大到 60 個，再將某些欄位的數值調整成常規的 10 進位以便計算，例如: IPv4、IPv6 的位址。

(3) 標準化資料: 將資料集進行標準化，把特定欄位內的數值縮放至

0.1 之間。

(4) 訓練資料: 將上述資料都進行整理後即可使用 Tensorflow 中的 keras api 來跑 FGSM 的機器學習，模型中的程式碼除了針對資料集個欄位的運算式以外，對照上面的概念表達公式，我們將它寫成這行程式碼放入訓練模型中:

adv_data = k.sign(gradients[0]) * eps + X_train

最後分別利用 FGSM 對 IOT23 原始資料集以及經由 ACT 值處理後的資料集進行攻擊。觀察是否 ACT 值能有效的提高模型抗攻擊的能力。

4. 主要成果與評估

此次研究成果包含了是否 ACT 值能夠有效提高模型的分類準確度，以及提高模型的抗攻擊能力。

在分類準確度上，從圖 1 及圖 2 中可以看到，不管是 NB15 還是 IOT23，經由 ACT 值處理後的資料集所訓練出來的模型都明顯有較高的分類精度。且從圖 3 及圖 4 中可以觀察到在前 10 個訓練週期，經由 ACT 值所訓練出來的模型就已經有非常高的分類精度，表示了 ACT 值能使得模型只需較少的訓練週期，就能獲得較高的準確率。

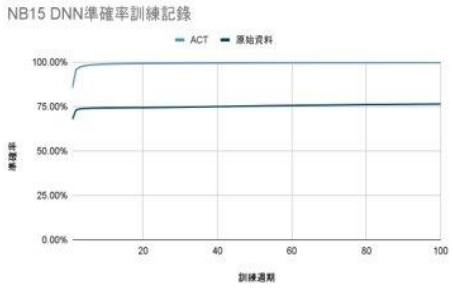

圖 1. NB15 ACT vs. 原始資料 DNN 分類精度

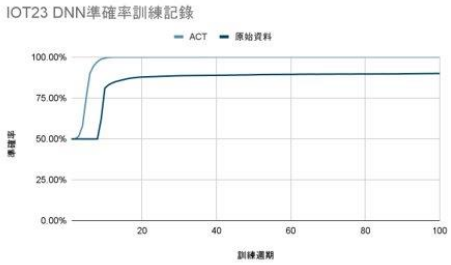

圖 2. IOT23 ACT vs. 原始資料 DNN 分類精度

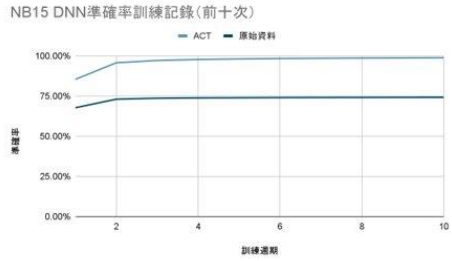

圖 3. NB15 ACT vs. 原始資料 DNN 分類精度 (10 個訓練週期)

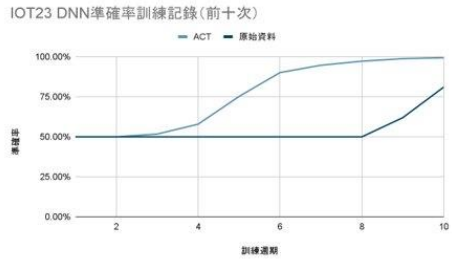

圖 4. IOT23 ACT vs. 原始資料 DNN 分類精度 (10 個訓練週期)

在抗攻擊能力的部分，從圖 5 中可以觀察到，經由 ACT 值處理後的資料集相較於原始資料集準確率下降的幅度有明顯的降低。並且，經由 ACT 值訓練的資料集可以在承受較多的擾動才會使得分類精度降到一定的程度，這也驗證了 ACT 值能使分類模型提高抗攻擊的能力。

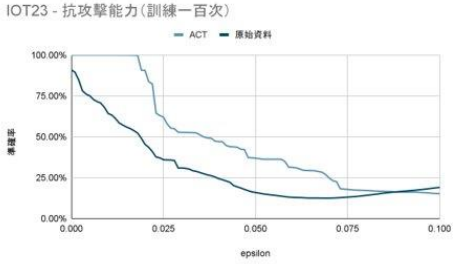

圖 5. 經 FGSM 攻擊之準確度

5. 結語與展望

本次專題我們成功提出了一種新型的特徵萃取方法，利用特徵值與分類標籤的關聯，計算出 ACT 值將特徵值進行處理，能利用簡單的方式有效的提高分類的準確度，並且縮減訓練的時間，這是之前從未出現過的方法。

也期許在未來能利用特徵與標籤關聯相關的方法，進而開發出更多不同的資料品質提升方法。

6. 銘謝

感謝教授在這一年中不管是在研究方法的建議，以及實驗資源的提供，都讓我們在專題的進行上更加順利。也謝謝實驗室的學長姊，在實驗中的協助，讓我們的專題能夠如期完成。

7. 參考文獻

[1] Szegedy, et al., “Intriguing properties of neural networks”, arXiv preprint arXiv:1312.6199, 2013.

[2] A. Nazir and RA. Khan, “A novel combinatorial optimization based feature selection method for network intrusion detection”, Computers & Security, vol.102, 102164, 2021

[3] M. E. Aminanto, R. Choi, H. C. Tanuwidjaja, P. D. Yoo, and K. Kim, "Deep Abstraction and Weighted Feature Selection for Wi-Fi Impersonation Detection," IEEE Transactions on Information Forensics and Security, vol. 13, no. 3, pp. 621-636, 2018.

[4] F. Gottwalt, E. Chang and T. Dillon, “CorrCorr: A feature selection method for multivariate correlation network anomaly detection techniques”, Computers & Security, vol. 83, pp. 234-245, 2019

[5] SM .Kasongo and Y. Sun, “A deep learning method with wrapper based feature extraction for wireless intrusion detection system”, Computers & Security, vol. 92, 101752, 2020.

[6] M. Shafiq, et al., “IoT malicious traffic identification using wrapper-based feature selection mechanisms”, Computers & Security, vol. 94, 101863, 2020.

[7] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, IEEE, 2015.

[8] https://www.stratosphereips.org/datasets-iot23

[9] https://research.unsw.edu.au/projects/unsw-nb15-dataset