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

覃毅翔、林奐呈、吳振瑋、簡毅

01 摘要

在機器學習模型訓練之前，我們需要對資料品質進行提升，以應對資料的離散性。傳統的方法包括了特徵選擇和特徵提取，然而這些方法未能提出分類精度提高的原因。

為了解決這個問題，我們提出了一種新的資料萃取方式，增強模型的準確率和抗攻擊能力。相較於其他方法，這種新的資料萃取方式在訓練模型時更為高效，也能夠更好地應對資料的離散性，對於訓練出更優秀的機器學習模型具有重要意義。

02 Accurate Class Tendency

首先，我們在網路事件資料集中選擇要萃取的特徵以及特徵值，接著利用該特徵值出現次數計算出良性紀錄以及惡性紀錄分別的出現比例。

計算出的比例即是該特徵值對於此分類標籤的 **P** 值。

我們接著將 **ACT** 值定義為 **P** 值減去 **0.5**，因為網路事件分類就只有良性事件與惡性事件，符合二分類的原則。又二分類的基礎機率值為 **0.5** ，因此這樣的方式可以直接用正負號來判斷是否該特徵值與分類標籤呈現正相關。

最後將 **ACT** 值替換回資料集中，並建立一個 **Map** 表，完成 **ACT** 資料萃取。

ip	port_num	
0.221	...	...
0.45	...	...
0.221	...	...

ip	ACT
203.0.113.100	0.221
198.51.100.1	0.45

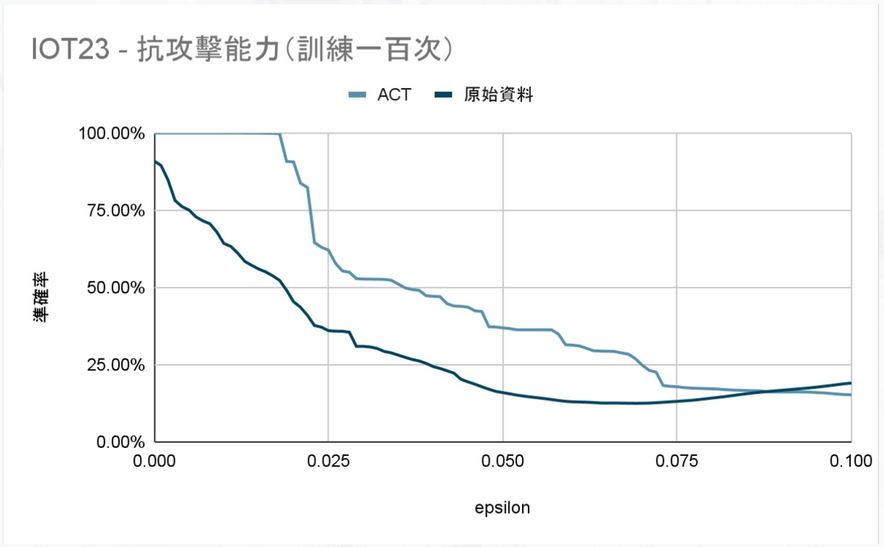
03 實驗過程

我們以兩種資料集用不同的品質提升方法後分別訓練出辨識模型，並用不同的檢測方式來檢驗 **ACT** 是否能成功提升辨識準確率以及提升抗攻擊能力。

04 成果比較

	string encoding + normalize	ACT
NB15	77.4336%	99.1909% WIN
IOT23	89.9315%	99.9979% WIN

同資料集以**不同的品質改進方式**，所訓練出的模型辨識**準確度**



經過 ACT 資料萃取後，**準確率下降幅度**會比原始資料來的**緩慢**

05 總結

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。