Machine Learning - Internet Firewall Data

Using K-Nearest Neighbors, Decision Tree, Random Forest and XGBoost

Prawit Kamchaiya 6510422004

Tada Nualsanit 6510422011

Titiwat Tanasuthisaree 6510422015

Kanya Meekaew 6510422016

Chalermwong Saleepattana 6510422029

Graduate School of Applied Statistics

National Institute of Development Administration

Machine Learning Mini-Project Report

1) Summarize the target paper

Target paper เป็นการประยุกต์การทำงานของ machine learning โดยวิธีการ Support vector machine (SVM) ในการพยากรณ์ว่าข้อมูล Firewall log ควรที่จะทำ action อะไร โดย พิจารณา kernel ของ model SVM ทั้งหมด 4 แบบ ได้แก่ SVM Linear, SVM Polynomial, SVM RBF, SVM Sigmoid ซึ่งผลลัพธ์ของการพยากรณ์มีทั้งหมด 4 กลุ่มข้อมูลได้แก่ Allow, Deny, Drop, Reset-Both และพิจารณา Feature ของการทำโมเดลโดยใช้ 4 ข้อมูล Port, Byte, Packets, Time โดยวัดผลของการพยากรณ์ด้วยค่า Precision, Recell, F1, ROC Curves

Method	F1 Score	Precision	Recall
SVM Linear	75.4	67.5	85.3
SVM Polynomial	53.6	61.8	47.4
SVM RBF	76.4	63.0	97.1
SVM Sigmoid	74.8	60.3	98.5

Table 1: ผลการประเมินโมเดล SVM Linear, SVM Polynomial, SVM RBF และ SVM Sigmoid จาก paper

2) Reproduced process (demonstrate step by step) (*should use pipeline)

1. Data collection
   ในการศึกษาครั้งนี้ใช้ Data set จาก UCI ประกอบด้วย ข้อมูลทั้งหมด 65532 แถว จำนวนคอลัมน์ 12 คอลัมน์ Attribute Information: Source Port, Destination Port, NAT Source Port, NAT Destination Port, Action, Bytes, Bytes Sent, Bytes Received, Packets, Elapsed Time (sec),pkts_sent, pkts_received โดยมีเป้าหมายคือคอลัมน์ Action

2. Data Cleaning
   เนื่องจากข้อมูลมีความสมบูรณ์ ไม่มีค่าว่าง (Null) ,ไม่พบความผิดปกติของข้อมูล เช่น Outlier, การใส่หลักตัวเลขของ Port ผิด

3. EDA
   - Feature selection: Features ทั้งหมดมี 11 Features โดยคัดเลือก Destination Port, NAT Source Port, NAT Destination Port , Bytes, pkts_sent , Elapsed Time (sec) เพื่อนำไปใช้ในการสร้างโมเดล เนื่องจากมีแนวโน้มที่จะสามารถจำแนกแต่ละประเภทของคอลัมน์เป้าหมาย (Action) ได้ เนื่องจากการสังเกตความสัมพันธ์ของกราฟ Pair plot ระหว่าง Features ทั้งหมดเทียบกับ Action พบว่ามีข้อมูลบางคอลัมน์มีการกระจายตัวแบบเป็นกลุ่มก้อนตามแต่ละ Action จึงตั้งสมมติฐานว่าข้อมูลเหล่านี้มีความสัมพันธ์ของกับเป้าหมาย
   - Imbalanced data: ข้อมูลของคอลัมน์เป้าหมาย (Action) แบ่งเป็น 4 ประเภท ได้แก่ Allow 37,640 แถว, Drop 14,987 แถว, Deny 12,851 แถว และ Reset-both 54 แถว ซึ่งพบว่าข้อมูลมีลักษณะไม่สมดุล (Imbalanced data) ส่งผลให้ต้องเตรียมวิธีเพื่อที่จะสกัดข้อมูลมาใช้แบบเกิดความเอนเอียงของคำตอบ (Bias) น้อยที่สุด

4. Training (Cross Validation)
   - การเตรียมข้อมูลเพื่อใช้การทดสอบโมเดล โดยแบ่งข้อมูลสำหรับ Train : Test เป็น 70 : 30
   - วิธีการจัดการของ Imbalanced data โดยใช้ Oversampling ด้วยวิธีการ smote เพื่อให้คอลัมน์เป้าหมาย (Action) แต่ละประเภทมีจำนวนข้อมูลสำหรับใช้ในการเรียนรู้เท่ากันเพื่อลดโอกาสการเกิด Bias ของข้อมูลคะแนนได้รับการพิสูจน์โดยการเปรียบเทียบค่า ROC ก่อนและหลังการ Oversampling
   - ในการศึกษาครั้งนี้ใช้โมเดล 4 โมเดล ได้แก่ K-Nearest Neighbors, Decision Tree, Random Forest และ XGBoost และเลือกโมเดลที่ให้ผลลัพธ์ค่า Accuracy, F1-Score และ AUC มากที่สุด
   - การปรับค่า Parameters โดยใช้เครื่องมือ GridsearchCV จาก Library สำหรับโมเดล KNearest Neighbors, Decision Tree และ XGBoost และ RandomizedsearchCV สำหรับ Random Forest
   - นำโมเดลที่ทำการปรับ Parameters แล้วไป Train ด้วยชุดข้อมูลที่เตรียมไว้

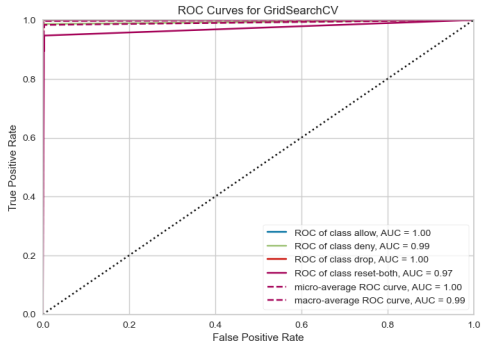
5. Testing
   - นำโมเดลที่ทำการTrain แล้วมาทดสอบด้วยชุดข้อมูลสำหรับการ Test (30% ของข้อมูลทั้งหมดจากการ Split)

6. Evaluation
   - ประเมิน โดยเปรียบคอลัมน์เป้าหมายของชุดข้อมูลทดสอบ(y_test) กับค่าที่ได้จากการพยากรณ์ออกมาจากโมเดล โดยพิจารณา ค่า Accuracy, F1-Score และ AUC ที่ได้

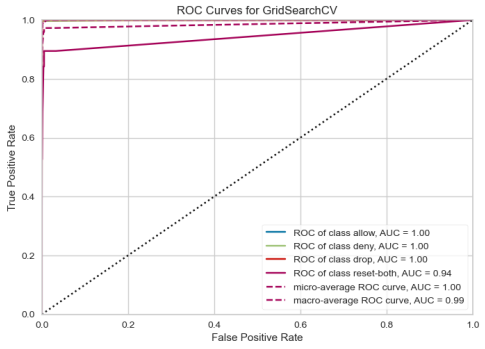
3) Show and discuss your results (figures, table, etc.)

Method	F1 Score	Precision	Recall	Accuracy
K-Nearest Neighbors	99.6	99.8	99.5	99.5
Decision Tree	99.4	99.8	99.0	99.0
Random Forest	99.8	99.9	99.8	99.8
XGBoost	99.7	99.9	99.7	99.7

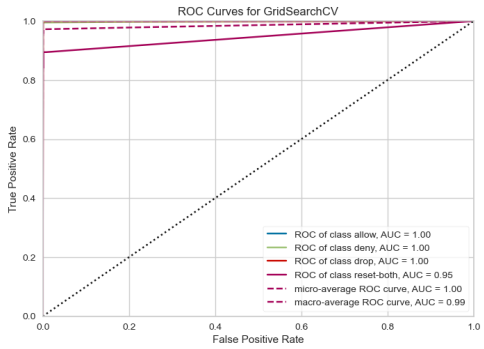
Table 2: ผลการประเมินโมเดล K-Nearest Neighbors, Decision Tree, Random Forest และ XGBoost



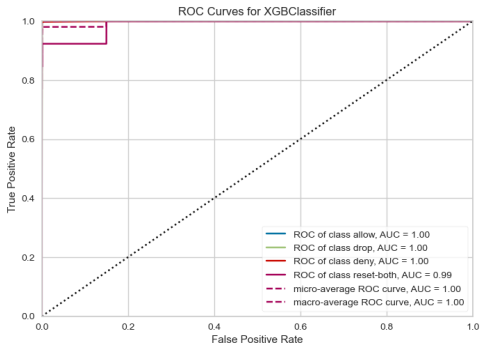
(a) ROC curve of K-Nearest Neighbors



(b) ROC curve of Decision Tree



(c) ROC curve of Random Forest



(d) ROC curve of XGBoost

Figure 1: กราฟแสดง ROC และ คะแนน AUC ของ K-Nearest Neighbors, Decision Tree, Random Forest และ XGBoost

จากตารางที่ 2 สามารถเปรียบเทียบ F1-Score ,Precision , Recall และ Accuracy ของทั้ง 4 โมเดล มีลักษณะไปในทางเดียวกันคือโมเดลที่ดีที่สุดคือ Random Forest รองลงมาคือ XGBoost , K-Nearest Neighbors และ Decision Tree ตามลำดับ

จากรูปที่ 1 จากกราฟ ROC ของแต่ละโมเดลเมื่อคำนวณหาพื้นที่ใต้กราฟ (AUC) เฉลี่ยของแต่ละโมเดล พบว่า XGBoost มีคะแนนดีที่สุด นอกจากนี้ยังพบว่าคะแนน AUC ในส่วนประเภท Reset - both มีคะแนนมากกว่าโมเดลอื่น รองลงมาคือ KNN, Random Forest, Decision Tree ตามลำดับ

4) Improvement over the reference paper

- การแก้ไขความไม่สมดุลของข้อมูลด้วยวิธี Oversampling ก่อนนำไปเข้ากระบวนการเรียนรู้
- การเลือก Features เพิ่มเติมโดยพิจารณาจากแนวโน้มและความสัมพันธ์ของข้อมูลในแต่ละ Features
- การเลือก Parameters ที่เหมาะสมของแต่ละโมเดลโดยใช้ GridsearchCV และ RandomizedsearchCV