การตรวจจับการบุกรุกโดยใช้เอดาบูทเอ็มวันและคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์

Anomaly based intrusion detection using Adaboost.m1 and Correlation-based feature selection

พลอยพรรณ สอนสุวิทย์(Ployphan Sornsuwit)[1*] มนตรี ใจแน่น(Montree Jainan)[2]
พิมกาญดา จันดาหังดง(Phimkarnda Jundahuadong)[3]
[1,2] สาขาวิชาคอมพิวเตอร์ธุรกิจ [3] สาขาวิชาการจัดการทั่วไป
คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร
[*]Corresponding author. E-mail: ployphan.en@gmail.com

บทคัดย่อ

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเพื่อพัฒนาขั้นตอนวิธีในการตรวจจับการ บุกรุกเครือข่ายคอมพิวเตอร์ด้วยการใช้ เอดาบูทเอ็มวัน (Adaboost.m1) โดยมีการลดมิติด้วยเทคนิคการคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ และได้เปรียบเทียบประสิทธิภาพของการจำแนกกับวิธีการอื่นๆ รวมไปถึงเปรียบเทียบการจำแนกแบบ 2 กลุ่ม เพื่อยืนยันประสิทธิภาพของขั้นตอนวิธีที่พัฒนา ซึ่งงานวิจัยนี้ได้ใช้ใช้ฐานข้อมูล NSL-KDD เป็นฐานข้อมูลการบุกรุกเครือข่ายคอมพิวเตอร์ โดยจะใช้เป็นข้อมูลสำหรับฝึกสอนและข้อมูลสำหรับทดสอบจากต้นฉบับที่ผู้พัฒนาจัดเตรียมไว้ ผลการวิจัยพบว่าขั้นตอนวิธีที่นำเสนอ มีประสิทธิภาพโดยรวมสูงที่สุด 4 ใน 5 Weak Learner ในกรณี แบบหลายกลุ่ม (Multiclass) จะมี J48 weak learner มีประสิทธิภาพสูงสุด ซึ่งมีค่าความถูกต้องร้อยละ 78.77 และในกรณีแบบ 2 กลุ่ม (Binary Class) จะมี k-NN weak learner มีประสิทธิภาพสูงสุด ซึ่งมีค่าความถูกต้องร้อยละ 80.83 จากการวิจัยนี้พบว่า การลดมิติยังช่วยลดภาระการประมวลผลลงจาก 41 มิติ เหลือเพียง 11 มิติ และเพิ่มประสิทธิภาพการจำแนกได้ดีขึ้นอีกด้วย

คำสำคัญ: การคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ ฐานข้อมูล NSL-KDD เอดาบูทเอ็มวัน

Abstract

The objectives of this research were to develop algorithm for detecting computer network intruders with Adaboost.m1 through the dimensionality reduction conducted with correlation-based feature selection and to compare the efficiency of the classification with other methods as well as binary class classification to assure the efficiency of the developed algorithm. This research used the NSL-KDD as a

computer network intrusion database that would be used as data for instructional training and testing from the original prepared by the developer. The results indicated that the proposed algorithm achieved the highest overall efficiency (4 of 5 weak learner). The multiclass scenario would have J48 weak learner, the highest efficiency. The accuracy value was 78.77%. The binary class would have k-NN weak learner, the highest efficiency. The accuracy value was 80.83. The findings showed that dimensionality reduction could also reduce the burden of processing from 41 to 11 dimensions and enhance the classification efficiency.

Keywords: Correlation-based Feature Selection, NSL-KDD, Adaboost.m1

บทนำ

ปัจจุบันอินเทอร์เน็ตมีบทบาทที่สำคัญต่อชีวิตประจำวัน ทั้งในการใช้งานทางธุรกิจ ทางการศึกษา ทางการแพทย์ และด้านอื่นๆรวมไปถึงใช้ในการจ่ายเงินในการทำธุรกรรมต่างทั้งภาครับและภาคเอกชน โดยใช้เงินอิเล็กทรอนิกส์ โดยได้รับการผลักดันจากภาครัฐ อย่างเป็นรูปธรรม. เมื่อมีการใช้งานอินเทอร์เน็ตมากขึ้น ส่งผลให้ข้อมูลที่มีการส่งผ่านเครือข่ายมีปริมาณมาก และเป็นข้อมูลสำคัญ เช่น ข้อมูลการซื้อขายสินค้าออนไลน์ รหัสบัตรเครดิต ข้อมูลส่วนตัวต่างๆ หรือข้อมูลความลับทางธุรกิจ เป็นต้น ทำให้การป้องกันความปลอดภัยของข้อมูล เป็นสิ่งที่จำเป็นและต้องมีการรักษาความปลอดภัยของข้อมูลตลอดเวลา เพราะหากถูกผู้บุกรุกโจมตีระบบเครือข่าย ณ ช่วงเวลาใดๆ นั่นหมายถึงอาจเกิดความสูญเสีย ที่อาจประเมินค่าไม่ได้ เช่น ผู้บุกรุกดักเก็บข้อมูล (Sniffer) ในวง LAN แล้วนำไปถอดรหัสด้วยโปรแกรมสำเร็จรูป เพื่อหารหัสผ่านของผู้อื่น เป็นต้น

ระบบตรวจจับการบุกรุก (Intrusion Detection System) คือระบบที่มีการติดตามเหตุการณ์ที่เกิดขึ้นในระบบคอมพิวเตอร์หรือเครือข่าย แล้วจึงวิเคราะห์รูปแบบว่าเป็นการบุกรุกหรือไม่ (Hung J.L., Chun H. R., Ying C. L. and Kuan Y. T., 2017, pp. 16-24) จัดเป็นระบบที่มีความจำเป็นต่อการป้องกันความปลอดภัยเครือข่ายในปัจจุบัน จึงได้มีงานวิจัยมากมาย (Huiwen, W., Jie, G. and Shanshan, W., 2017, pp. 130-139) (Abdulla, A. A. and Mamun B. I. R., 2017, pp. 135–152) (Soo, Y. J. Bong, K. J. , Seonho, C. and Dong, H. J., 2016, pp. 9-17) (Dewan, M. F., Li Z., Chowdhury M. R., M.A. and H., R. S., 2014, pp. 1937–1946) (Aditi N., Basant, T. and Vivek, T., 2016, pp. 26-31) ที่ศึกษาการพัฒนาขึ้นตอนวิธีในการตรวจจับการบุรุก เพื่อเพิ่มประสิทธิภาพในการตรวจจับ ซึ่งงานวิจัยเหล่านั้น ต้องการเพิ่มประสิทธิภาพของขั้นตอนวิธีให้สูงขึ้น แต่อาจยังมีค่าการแจ้งเตือนที่ผิดพลาด (False Alarm) สูงอยู่บ้าง. งานวิจัยกลุ่มหนึ่ง (Michael, G., Kumaravel, A., Chandrasekar A., 2015, pp. 2455-2459) (Mirza, M. B., Mian,

M. A., El-Sayed, M. E., 2017, pp. 120–126) (Yali, Y., Georgios, K., Dieter, H., 2016, pp. 111-114) ที่ได้ศึกษาการตรวจจับการบุกรุก โดยใช้หลักการของการเรียนรู้แบบกลุ่ม (Ensemble) มาเพิ่มประสิทธิภาพในการตรวจจับ ซึ่งพบว่ามีประสิทธิภาพสูง เนื่องจากโดยหลักการของ Boosting จะมีการเรียนรู้ (Training) หลายต้นแบบ (Model) เพื่อทำการรวบรวมเสียงข้างมาก (Vote) ในการพยากรณ์คำตอบ จึงทำให้มีโอกาสในการตอบถูกสูง

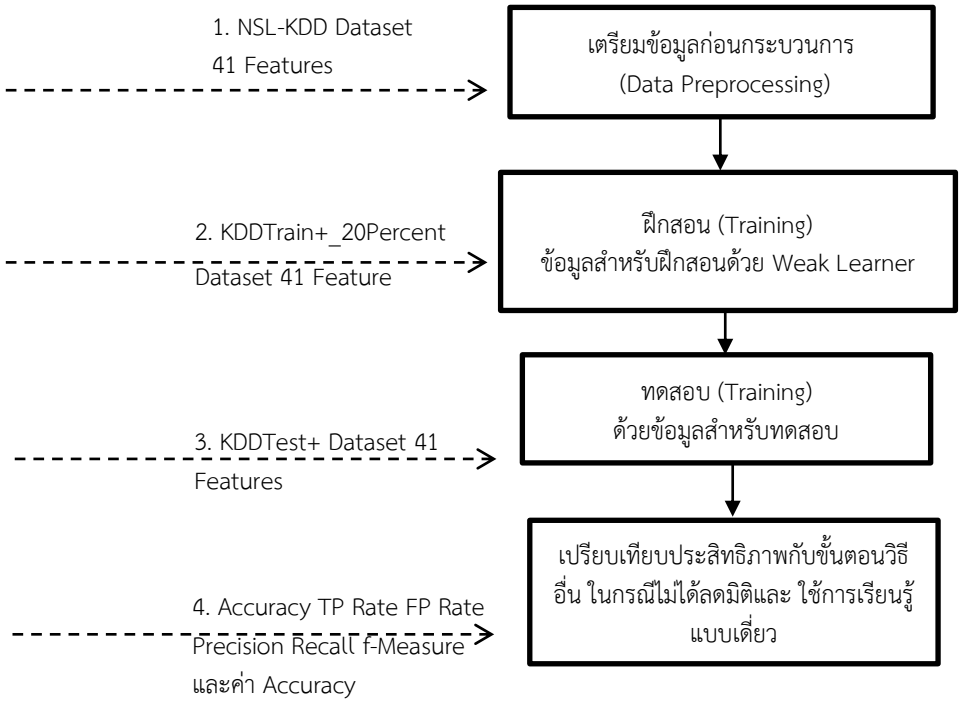
งานวิจัยนี้ ผู้วิจัยจึงมีวัตถุประสงค์เพื่อพัฒนาขั้นตอนวิธีในการตรวจจับการบุกรุก ด้วยขั้นตอนวิธีเอดาบูทเอ็มวัน (Adaboost.m1) และขั้นตอนการลดมิติข้อมูลด้วยเทคนิคการคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ (Correlation-based Feature Selection) ในการเพิ่มประสิทธิภาพการตรวจจับการบุกรุกเครือข่ายคอมพิวเตอร์ และเปรียบเทียบประสิทธิภาพของการจำแนก (Classification) กับขั้นตอนวิธีอื่นๆที่ไม่ลดมิติ ใช้การเรียนรู้แบบเดี่ยว (Single Learner) และ การจำแนกแบบ 2 กลุ่ม

วัตถุประสงค์ของการวิจัย

เพื่อพัฒนาขั้นตอนวิธีในการตรวจจับการบุกรุกเครือข่ายคอมพิวเตอร์ด้วยเอดาบูทเอ็มวัน

วิธีดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยเชิงปริมาณ ที่ต้องการพัฒนาขั้นตอนวิธีในการตรวจจับการบุกเครือข่ายคอมพิวเตอร์ ด้วยการใช้ เอดาบูทเอ็มวัน และลดมิติด้วยเทคนิคการคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ มีขั้นตอนดังต่อไปนี้



รูปที่ 1 แสดงขั้นตอนวิธีการทดลอง

1. ประกอบไปด้วย 4 ขั้นตอนดังต่อไปนี้

1.1 ขั้นตอนการเตรียมข้อมูลก่อนกระบวนการ (Data Preprocessing) โดยการดาวน์โหลดจากเว็บไซต์ (University of New Brunswick, 2009) ข้อมูลที่ใช้ ในการทดลองจะใช้ฐานข้อมูล NSL-KDD ซึ่งเป็นฐานข้อมูลการบุกรุกเครือข่ายที่เป็นรุ่นของการปรับปรุงมาจาก KDD Cup 99 โดยจะลดจำนวนข้อมูลที่ซ้ำซ้อนลงและมีจำนวนประเภทการบุกรุกต่างๆทั้งในข้อมูลฝึกสอนและข้อมูลทดสอบเหมาะสมขึ้น. เนื่องจากฐานข้อมูลที่ใช้งานประกอบไปด้วย 22 ประเภทย่อยของการบุกรุก และ 1 ประเภทของ Normal ขั้นตอนนี้ ได้ทำการจัดกลุ่มประเภทย่อยๆของการบุกรุก ให้เป็นกลุ่มหลัก 5 กลุ่มหลัก ได้แก่ Normal Dos Probe R2L และ U2R. จากนั้นจึงทำการลดมิติของข้อมูลด้วยเทคนิคการคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ (Correlation-based feature Selection: CFS) เพื่อให้เหลือเพียงคุณลักษณะที่มีความสัมพันธ์กันสูงเท่านั้นมาใช้งาน

1.2 ฝึกสอนข้อมูลที่จัดเตรียม (KDDTrain+_20Percent) ด้วยเทคนิคเอดาบูทเอ็มวันโดยใช้ Weak learner ทั้ง 5 เทคนิคของ Supervised Learning ได้แก่ Naïve Bayes, Decision Tree, Multilayer Perceptron, k-NN และ SVM โดยกำหนดให้ ขั้นตอนกาเรียนรู้สร้าง 50 ต้นแบบสำหรับการ Vote เพื่อให้ได้คำตอบสุดท้าย ของการจำแนกข้อมูล

1.3 ทดสอบข้อมูลจากข้อมูลสำหรับทดสอบ (KDDTest+) ที่ได้จัดเตรียมไว้

1.4 เปรียบเทียบประสิทธิภาพของการจำแนกของ Weak learner ทั้ง 5 เทคนิค แบบหลายกลุ่ม (Multiclass) และ แบบสองกลุ่ม (Binary) รวมไปถึงเปรียบเทียบกับการจำแนกที่ไม่ได้ลดมิติและการเรียนรู้แบบเดี่ยว (Single Learner)

2. เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการประมวลผลการลดมิติ ฝึกสอนข้อมูล ทดสอบข้อมูล และวิเคราะห์ประสิทธิภาพ ได้แก่โปรแกรม WEKA Mining 3.8.1

3. สถิติที่ใช้ในการวิเคราะห์ข้อมูล

การวิเคราะห์ประสิทธิภาพของการจำแนก ได้พิจารณาเปรียบเทียบค่าต่างๆดังต่อไปนี้ ค่า True Positive Rate (TP Rate) ค่า False Positive Rate (FP Rate) ค่า Precision ค่า Recall ค่า f-Measure และค่า Accuracy วิเคราะห์จากตาราง Confusion Matrix (Gulshan, K, 2014)

ตารางที่ 1 Confusion Matrix

Predict	**Actual**	
	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{f} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

เมื่อ

TP (True Positive) = อัตราผลบวกจริง

TN (True negative) = อัตราผลบวกปลอม

FP (False Positive) = อัตราผลลบจริง

FN (False Negative) = อัตราผลลบปลอม
Precision = ค่า ความแม่นยำ
Recall = ค่าความไว
F-Measure = ค่าถ่วงดุล
Accuracy = ค่าความถูกต้องโดยรวม

ผลการวิจัย

ผลการทดลองในขั้นตอนเตรียมข้อมูลก่อนกระบวนการ พบว่ามีจำนวนข้อมูลสำหรับฝึกสอนและจำนวนข้อมูลสำหรับทดสอบแยกตามประเภทดังตารางที่ 2

ตาราง 2 จำนวนของข้อมูลแยกตามประเภท

ประเภท	ข้อมูลฝึกสอน	ข้อมูลทดสอบ
Normal	13,449	9,711
Dos	9,234	7,460
Probe	2,289	2,421
R2L	209	2,885
U2R	11	67
รวม	25,192	22,544

จากนั้นได้คัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ จากทั้งหมด 41 คุณลักษณะ (Feature) เหลือเพียง 11 คุณลักษณะ ได้แก่ 1. servcies 2. flag 3. src_bytes 4. dst-bytes 5. logged_in 6. root_shell 7. srv_serror_rate 8. same_srv_rate 9. diff_srv_rate 10. dst_host_srv_diff_host_rate และ 11. dst_host_serror_rate.

จากนั้น เมื่อทำการฝึกสอนด้วยเอดาบูทเอ็มวัน โดยสร้างต้นแบบ จำนวน 50 ต้นแบบสำหรับทุกๆ Weak Learner แล้วจึงทดสอบค่าประสิทธิภาพของการจำแนก เปรียบเทียบกับวิธีการที่ลดมิติและเปรียบเทียบกับการเรียนรู้แบบเดี่ยวที่ไม่ได้ลดมิติ แสดงผลการเปรียบเทียบค่า Accuracy ดังตารางที่ 3

ตารางที่ 3 แสดงการเปรียบเทียบค่าความถูกต้องของการจำแนกหลายกลุ่ม

จำนวน Feature	**ขั้นตอนวิธี**	**ความถูกต้อง (%)**
41Features	J48	77.66
	Naive Bayes	**70.95**
	k-NN	73.69
	MLP	72.27
	SVM	68.21
11 Features	J48	75.45
	Naive Bayes	40.66
	k-NN	75.33
	MLP	73.89
	SVM	43.07
ขั้นตอนวิธีที่นำเสนอ 11 Features+Adaboost.m1	J48	**78.77**
	Naive Bayes	45.30
	k-NN	**76.99**
	MLP	**74.14**
	SVM	**73.45**

จากตารางที่ 3 จะพบว่า ขั้นตอนวิธีที่งานวิจัยนำเสนอ เมื่อลดมิติแล้วใช้เอดาบูทเอ็มวัน ซึ่งมี J48 Naïve Bayes k-NN และ SVM เป็น Weak Learner มีค่า Accuracy สูงกว่าการประมวลผลแบบเดี่ยวของทุกวิธีดังกล่าวทุกวิธี และยังสูงกว่าการไม่ลดมิติอีกด้วย ยกเว้น Naïve Bayes จากผลการทดลองพบว่า J48 มีค่า Accuracy สูงที่สุดคือ 78.77% แสดงค่า Confusion Matrix ดังตารางที่ 4 ดังนี้

ตารางที่ 4 แสดงค่า Confusion Matrix

	จำแนกเป็น				
กลุ่มจริง	**Normal**	**Dos**	**R2L**	**Probe**	**U2R**
Normal	**9351**	83	11	266	0
Dos	1752	**5698**	0	10	0
R2L	2401	215	**16** **6**	103	0
Probe	496	132	4	**1789**	0
U2R	62	0	0	0	**5**

เมื่อวิเคราะห์ค่าประสิทธิภาพโดยละเอียดของทุกประเภทการบุกรุก แยกตามวิธีการทดลองที่เปรียบเทียบ แสดงค่าประสิทธิภาพ TP Rate FP Rate Precision Recall และ f-Measure ดังตารางที่ 5-7

ตารางที่ 5 เปรียบเทียบค่า TP Rate FP Rate Precision Recall และ f-Measure ของขั้นตอนวิธีที่นำเสนอ

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	f-Measure
Decision Tree					
Normal	0.969	0.333	0.687	0.969	0.804
Dos	0.83	0.014	0.966	0.83	0.893
R2L	0.232	0.001	0.965	0.232	0.374
Probe	0.606	0.013	0.849	0.606	0.707
U2R	0.224	0	0.833	0.224	0.353
Naïve Bayes					
Normal	0.282	0.1	0.682	0.282	0.399
Dos	0.707	0.021	0.943	0.707	0.808
R2L	0.002	0.001	0.147	0.002	0.003
Probe	0.892	0.069	0.61	0.892	0.725
U2R	0.537	0.415	0.004	0.537	0.008
k-NN					
Normal	0.921	0.272	0.719	0.921	0.808
Dos	0.833	0.065	0.864	0.833	0.848
R2L	0.132	0.019	0.511	0.132	0.209
Probe	0.742	0.017	0.842	0.742	0.789
U2R	0.299	0	0.645	0.299	0.408
MLP					
Normal	0.977	0.386	0.657	0.977	0.786
Dos	0.717	0.017	0.954	0.717	0.819
R2L	0	0	0	0	0
Probe	0.776	0.031	0.752	0.776	0.764
U2R	0	0	0	0	0
SVM					

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	f-Measure
Normal	0.968	0.422	0.635	0.968	0.767
Dos	0.737	0.017	0.954	0.737	0.831
R2L	0.001	0.001	0.143	0.001	0.001
Probe	0.687	0.015	0.849	0.687	0.759
U2R	0	0	0	0	0

ตารางที่ 6 เปรียบเทียบค่า TP Rate FP Rate Precision Recall และ f-Measure ของวิธีการลดมิติ

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	f-Measure
Decision Tree					
Normal	0.963	0.367	0.665	0.963	0.787
Dos	0.764	0.029	0.93	0.764	0.839
R2L	0.058	0.001	0.917	0.058	0.108
Probe	0.739	0.019	0.825	0.739	0.78
U2R	0.075	0	1	0.075	0.139
Naïve Bayes					
Normal	0.194	0.149	0.497	0.194	0.279
Dos	0.736	0.054	0.87	0.736	0.797
R2L	0.001	0.001	0.158	0.001	0.002
Probe	0.72	0.021	0.802	0.72	0.759
U2R	0.627	0.454	0.004	0.627	0.008
k-NN					
Normal	0.925	0.325	0.683	0.925	0.786
Dos	0.833	0.069	0.856	0.833	0.844
R2L	0.023	0.001	0.802	0.023	0.044
Probe	0.708	0.016	0.841	0.708	0.768
U2R	0.224	0	0.625	0.224	0.33
MLP					
Normal	0.985	0.375	0.665	0.985	0.794
Dos	0.716	0.014	0.962	0.716	0.821
R2L	0	0	0	0	0
Probe	0.724	0.043	0.671	0.724	0.696
U2R	0	0	0	0	0
SVM					

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	f-Measure
Normal	1	0.99	0.433	1	0.604
Dos	0	0	0	0	0
R2L	0	0.002	0	0	0
Probe	0	0.004	0.011	0	0.001
U2R	0	0	0	0	0

ตารางที่ 7 เปรียบเทียบค่า TP Rate FP Rate Precision Recall และ f-Measure ของวิธีการเรียนรู้แบบเดี่ยว

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	f-Measure
Decision Tree					
Normal	0.961	0.3	0.708	0.961	0.816
Dos	0.829	0.029	0.934	0.829	0.878
R2L	0.003	0	0.833	0.003	0.007
Probe	0.81	0.037	0.725	0.81	0.765
U2R	0.209	0	0.875	0.209	0.337
Naïve Bayes					
Normal	0.853	0.243	0.726	0.853	0.785
Dos	0.705	0.044	0.888	0.705	0.786
R2L	0.087	0.003	0.812	0.087	0.157
Probe	0.9	0.063	0.631	0.9	0.742
U2R	0.328	0.064	0.015	0.328	0.029
k-NN					
Normal	0.978	0.412	0.642	0.978	0.775
Dos	0.75	0.019	0.952	0.75	0.839
R2L	0.024	0	0.933	0.024	0.047
Probe	0.599	0.018	0.802	0.599	0.686
U2R	0.075	0	0.833	0.075	0.137
MLP					
Normal	0.975	0.443	0.625	0.975	0.762
Dos	0.706	0.013	0.964	0.706	0.815
R2L	0	0	0	0	0
Probe	0.644	0.018	0.808	0.644	0.716
U2R	0	0	0	0	0

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	f-Measure
		SVM			
Normal	0.983	0.529	0.584	0.983	0.733
Dos	0.646	0.011	0.966	0.646	0.774
R2L	0	0	0	0	0
Probe	0.418	0.01	0.831	0.418	0.556
U2R	0	0	0	0	0

จากผลการวิเคราะห์ค่าประสิทธิภาพที่พบว่า เมื่อพิจารณาค่า f-measure จะเห็นได้ว่าวิธีการนำเสนอ โดยภาพรวม ส่วนมากมีค่าสูงกว่าทุกๆวิธีการที่เปรียบเทียบเมื่อแยกตามรายกลุ่มของการบุกรุก นั่นหมายถึง สามารถจำแนกประเภทของ การบุกรุกได้ดีที่สุดโดยภาพรวม ซึ่งสอดคล้องกับค่า Accuracy ที่นำเสนอ

เมื่อเปรียบเทียบประสิทธิภาพกับฐานข้อมูลเดิม แต่เป็นกรณี 2 ประเภทของพฤติกรรม ได้แก่ ประเภทปกติ (Normal Type) และประเภทไม่ปกติ หรือ ประเภทการบุกรุก (Abnormal Type) พบว่า ขั้นตอนวิธีที่นำเสนอ มีค่าความถูกต้องสูงที่สุด เช่นกันเมื่อเทียบกับวิธีการอื่นๆ แสดงดังตารางที่ 8

ตารางที่ 8 แสดงการเปรียบเทียบค่าความถูกต้องของการจำแนกแบบ 2 กลุ่ม

จำนวน Feature	ขั้นตอนวิธี	ความถูกต้อง (%)
41Features	J48	**79.98**
	Naive Bayes	76.30
	k-NN	75.60
	MLP	75.33
	SVM	47.37
11 Features	J48	77.51
	Naive Bayes	75.09
	k-NN	78.33
	MLP	75.48
	SVM	62.96
ขั้นตอนวิธีที่นำเสนอ 11 Features+Adaboost.m1	J48	79.37
	Naive Bayes	**76.55**
	k-NN	**80.83**

จำนวน Feature	ขั้นตอนวิธี	ความถูกต้อง (%)
	MLP	**75.90**
	SVM	**78.55**

สรุปและอภิปรายผล

สรุปผลการวิจัย

งานวิจัยนี้ ได้พัฒนาขั้นตอนวิธีในการตรวจจับการบุกรุก ด้วยการเพิ่มประสิทธิภาพในการจำแนกข้อมูลการบุกรุกให้มีประสิทธิภาพเพิ่มขึ้น ในการทดลองได้ลดมิติของข้อมูลด้วยเทคนิคการคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ แล้วจำแนกข้อมูลด้วยเอดาบูทเอ็มวัน ซึ่งจะใช้ Weak learner ที่แตกต่างกันในการเปรียบเทียบ เพื่อหาขั้นตอนวิธีที่มีประสิทธิภาพสูงสุด จากนั้นจึงเปรียบเทียบกับขั้นตอนวิธีที่มีการลดมิติ และเปรียบเทียบกับการเรียนรู้แบบเดี่ยวที่ไม่ได้ลดมิติ ผลการทดลองพบว่า ขั้นตอนวิธีที่นำเสนอโดยใช้ J48 Weak Learner มีประสิทธิภาพสูงที่สุด และสูงกว่าขั้นตอนวิธีอื่นๆทั้งภาพรวมและแยกวิเคราะห์ตามประเภทการบุกรุกแต่ละประเภท ยกเว้น Naïve Bayes ในส่วนของการทดลองแบบ 2 กลุ่ม พบว่ามีผลการทดลองเป็นไปในทางเดียวกันคือ ขั้นตอนวิธีที่นำเสนอมีประสิทธิภาพโดยรวมสูงที่สุด ยกเว้น J48 weak leaner ซึ่งแม้ว่ามีบาง Weak Learner ที่ไม่ได้มีประสิทธิภาพสูงสุด แต่การลดมิติก็สามารถลดภาระการประมวลผลลงไปได้ โดยลดลงจาก 41 มิติ เหลือเพียง 11 มิติ จึงเหมาะแก่การนำไปพัฒนาต่อยอดให้ใช้งานได้จริงในการตรวจจับการบุกรุกเครือข่ายคอมพิวเตอร์ในอนาคต

อภิปรายผล

1. เพื่อพัฒนาขั้นตอนวิธีในการตรวจจับการ บุกรุกเครือข่ายคอมพิวเตอร์ด้วยการใช้เอดาบูทเอ็มวัน วันและลดมิติด้วยเทคนิคการคัดเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ มีประสิทธิภาพในการจำแนกข้อมูลการบุกรุกเครือข่ายได้ดีที่สุดในภาพรวมเมื่อทดสอบกับ Weak Learner ในทุก Weak learner โดยเมื่อใช้ J48 เป็น Weak Learner มีค่าความถูกต้องสูงที่สุด เป็นร้อยละ 78.77 ยกเว้น Naïve Bayes ไม่ได้มีค่าความถูกต้องสูงสุด เนื่องจากขั้นตอนวิธีของเอดาบูท ใช้เทคนิคของการ Vote ค่า Weight ในทุกๆคำตอบของแต่ละต้นแบบ เพื่อให้ได้คำตอบสุดท้าย (Final Hypothesis) ทำให้ Weak Learner ใดๆ สามารถเรียนรู้ได้ 50 ต้นแบบแล้วจึง Vote ได้คำตอบที่ถูกต้องที่สุดตามความสามารถในการจำแนก แต่ Naïve Bayes อาจเป็นขั้นตอนวิธีการคิดค่าความน่าจะเป็น ซึ่งต้นแบบเดี่ยว อาจไม่ได้มีค่าความถูกต้องสูง ทำให้เมื่อเกิดการ Vote ในหลายๆต้นแบบจึงเกิดการคำนวณค่าจาก Weight ของคำตอบที่ไม่ถูกต้องแล้วผลการจำแนกจึงต่ำลงไปได้ แต่แม้ไม่ได้มีค่าความถูกต้องที่สูงที่สุด ก็ยังคงมีค่าไกล้เคียงกัน และสามารถลดภาระการประมวลผลในด้าน

จำนวน feature ที่ลดลงมากได้จาก 41 Features เหลือเพียง 11 Features การวิจัยนี้ สอดคล้องกับงานวิจัยของ Wei L.i และ QingXia Li (2010).

2. ประสิทธิภาพการจำแนกโดย ละเอียดแยกตามประเภท พบว่าสามารถจำแนกประเภทของการบุกรุกได้ดีที่สุด โดยภาพรวมซึ่งมีค่า f-Measure สูงที่สุด เป็นจำนวนที่มากที่สุด ดังตาราง 4-6 อาจมีบางประเภทการบุกรุกที่จำแนกได้ดีกว่ากับขั้นตอนวิธีอื่นที่ทำการเปรียบเทียบ เนื่องจากบางประเภทของการบุกรุกมีจำนวนมากในฐานข้อมูล เช่น Dos หรือบางประเภทมีจำนวนน้อย เช่น R2L และ U2R ทำให้เกิดปัญหาข้อมูลไม่สมดุล (Class Imbalance) โดยจำนวนกลุ่มที่มีปริมาณน้อยอาจไม่สามารถจำแนกได้เลย และบางกลุ่มที่มีจำนวนมากกว่า อาจมีรูปแบบ (Pattern) ที่คล้ายคลึงกันกับประเภทอื่นๆ ซึ่งอาจต้องใช้จำนวนมิติที่มาก ทั้ง 41 มิติ ในการจำแนกข้อมูลที่จำแนกออกจากกันบางบางประเภท.

3. เมื่อทดลองแบบ 2 กลุ่มพบว่า ขั้นตอนวิธีที่นำเสนอ มีประสิทธิภาพในการจำแนกข้อมูลการบุกรุกเครือข่ายได้ดีที่สุดในภาพรวม ซึ่งเมื่อใช้ MLP เป็น Weak Learner มีค่าความถูกต้องสูงที่สุด เป็นร้อยละ 80.33 เมื่อทดสอบกับ Weak Learner ในทุก Weak learner ยกเว้น J48 เนื่องจาก ไม่ได้มีค่าความถูกต้องสูงที่สุด เนื่องจากเป็นขั้นตอนวิธีการสร้างต้นไม้ (Tree) ซึ่งหากมีจำนวนของ Feature ที่มากกว่าในการคำนวนค่า Information Gain (IG) อาจทำให้โครงสร้าง Tree ที่สมบูรณ์กว่า พยากรณ์คำตอบสุดท้ายที่ถูกต้องได้มากกว่า.

ข้อเสนอแนะ

งานวิจัยนี้มีขั้นตอนการลดมิติเพื่อเพิ่มประสิทธิภาพ ซึ่งในการพัฒนาต่อยอดควรพัฒนาความเร็วในการประมวลผลเพื่อให้สามารถนำไปใช้ได้จริงในการตรวจจับการบุกรุกบนเครือข่ายขนาดใหญ่

เอกสารอ้างอิง

Abdulla, A. A. and Mamun B. I. R. (2017). A survey of intrusion detection systems based on ensemble and hybrid classifiers. ScienceDirect, 65(1), 135–152

Aditi N., Basant, T. and Vivek, T. (2016). Improving Accuracy for Intrusion Detection through Layered Approach Using Support Vector Machine with Feature Reduction, *WIR '16 Proceedings of the ACM Symposium on Women in Research 2016* (pp. 26-31). India: Indor.

Dewan, M. F., Li Z., Chowdhury M. R., M.A. and H., R. S. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(1), 1937–1946

Gulshan, K. (2014). Evaluation Metrics for Intrusion Detection Systems - A Study. International Journal of Computer Science and Mobile Applications, 2(11), 11-17

Hung J.L., Chun H. R., Ying C. L. and Kuan Y. T. (2017). Intrusion detection System: A Comprehensive review. *Journal of Computer and Application*, 36(1), 16-24

Huiwen, W., Jie, G. and Shanshan, W. (2017). An effective intrusion detection framework based on SVM with feature augmentation. *Knowledge-Based Systems*, 136(1), 130-139

Michael, G., Kumaravel, A., Chandrasekar A. (2015) Detection of malicious attacks by Meta classification algorithms. *Advanced Networking and Applications*, 6(5), 2455-2459

Mirza, M. B., Mian, M. A., El-Sayed, M. E. (2017). AdaBoost-based artificial neural network learning. *Neurocomputing*, 248 (1), 120–126

Soo, Y. J. Bong, K. J. , Seonho, C. and Dong, H. J. (2016). A multi-level intrusion detection method for abnormal network behaviors. *Journal of Network and Computer Applications*, 62(1), 9-17

University of New Brunswick. (2009). NSL-KDD dataset. Retrieved Dec 3, 2016, from http://www.unb.ca/cic/datasets/nsl.html

Wei, L. and QingXia, L. (2010). Using Naive Bayes with AdaBoost to Enhance Network Anomaly Intrusion Detection, *ICINIS '10 Proceedings of the 2010 Third International Conference on Intelligent Networks and Intelligent Systems* (pp. 486-489).China: China

Yali, Y., Georgios, K., Dieter, H. (2016). A Novel Semi-Supervised Adaboost Technique for Network Anomaly Detection, *MSWiM '16 Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems* (pp. 111-114). Malta: Malta.