การพัฒนาขั้นตอนวิธีในระบบตรวจจับการบุกรุกทางเครือข่ายด้วยเอดาบูทเอ็มวัน

Algorithm Development of Network Intrusion Detection with Adaboost.m1

พลอยพรรณ สอนสุวิทย์

สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร อำเภอเมือง จังหวัดกำแพงเพชร

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) พัฒนาขั้นตอนวิธีในการตรวจจับสิ่งผิดปกติทางเครือข่ายคอมพิวเตอร์ด้วยเทคนิคเอดาบูทเอ็มวัน (Adaboost.m1) และลดมิติด้วยเทคนิคมาตรฐานอัตราส่วนเกน (Gain Ratio) และ 2) เปรียบเทียบประสิทธิภาพของการจำแนก (Classification) ขั้นตอนวิธีที่นำเสนอ และขั้นตอนวิธีอื่นๆของ Supervised Learning การทดลองนี้ได้ใช้ ฐานข้อมูล NSL-KDD ซึ่งเป็นฐานข้อมูลการบุกรุกเครือข่าย และการวิเคราะห์และเปรียบเทียบประสิทธิภาพจะใช้ค่า อัตราผลบวกจริง อัตรา ผลบวกปลอม ค่าความแม่นยำ ค่าความไว ค่าถ่วงดุล และค่าความถูกต้อง

ผลการวิจัยพบว่า 1) การลดมิติของข้อมูลทำให้ได้เฉพาะคุณสมบัติที่สำคัญ เมื่อทำการจำแนกข้อมูลด้วยเอดาบูทเอ็มวัน โดยมีการใช้ทฤษฎีต้นไม้ตัดสินใจ เป็น Weak Learner พบว่ามีประสิทธิภาพในการจำแนกสูงที่สุด มีค่า ความถูกต้องร้อยละ 99.79 2) เมื่อทำการเปรียบเทียบประสิทธิภาพ พบว่า ขั้นตอนวิธีที่นำเสนอ มีประสิทธิภาพดีกว่า การจำแนกที่ลดมิติธรรมดาโดยไม่ได้ใช้เอดาบูทเอ็มวัน และการจำแนกที่ไม่มีการลดมิติ เมื่อเปรียบเทียบระยะเวลาในการประมวลผลพบว่า ขั้นตอนวิธีที่นำเสนอจะใช้เวลาสูงที่สุด เมื่อเทียบกับทุกวิธีเนื่องจากจะต้องมีการสร้างต้นแบบ (Models) จำนวนหลายต้นแบบ เพื่อทำการรวบรวมเสียงข้างมากเป็นคำตอบสุดท้าย (Final Hypothesis) ในการนำไปประยุกต์ใช้งานกับการตรวจสิ่งผิดปกติจับเครือข่ายจริง สามารถเลือกวิธีที่เหมาะสมได้ตามการใช้งาน

คำสำคัญ: ขั้นตอนวิธี Adaboost.m1, ฐานข้อมูล NSL-KDD, การลดมิติ, สิ่งผิดปกติทางเครือข่าย

ABSTRACT

The objectives of this research were 1) to develop and detect network anomaly with Adaboost.m1 technique and conduct dimension reduction with Gain Ratio and 2) to compare the efficiency of classifying proposed algorithm with Supervised Learning algorithms, This experiment used NSL-KDD database, a network intrusion database. True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Recall, f-Measure, and Accuracy were determined for the performance analysis and comparison.

The results of this study were as follows: 1) data dimension reduction resulted in important features. When data were classified with Adaboost.m1 technique and decision tree was used as weal learner, it was found that the for highest classification efficiency, the accuracy was 99.79%, 2) When efficiency was compared, the efficiency of proposed algorithm was better than dimension reduction without Adaboost.m1 technique, and classification technique without dimension reduction, when processing time were compared, it was found that proposed algorithm took the highest time, compared to all methods because it required to create a number of models compared to all methods for voting the final answer. For the application with the network anomaly detection, the appropriate method can be selected according to the need.

Keywords: Adaboost.m1 Algorithm, NSL-KDD Database, Dimension Reduction, Network Anomaly

บทนำ

ระบบตรวจจับการบุกรุก (Intrusion Detection System) เป็นระบบที่สามารถตรวจจับความผิดปกติของคอมพิวเตอร์และสารสนเทศ [1] ที่เกิดขึ้นเนื่องจากการบุกรุกของบุคคลหรือสิ่งอื่นใดที่อาจสร้างความเสียหายแก่ระบบ โดยหากแบ่งประเภทของการตรวจจับการตรวจจับการบุกรุก จะแบ่งได้เป็น 3 ประเภทหลัก [2] คือ Signature- based Detection เป็นการตรวจจับที่วิเคราะห์ความเหมือนของพฤติกรรมการโจมตี โดยจะเปรียบเทียบ Signature โดยตรงแล้วทำการวิเคราะห์ค่าว่าใช่การบุกรุกหรือไม่ และ Anomaly-based Detection จะเป็นการวิเคราะห์พฤติกรรมการโจมตีว่ามีค่าความเบี่ยงเบนไปจากพฤติกรรมปกติที่รู้จำไว้แล้วหรือไม่ หากระบบวิเคราะห์ค่าว่าเบี่ยงเบนไปเกิน Threshold ที่รู้จำ จะตัดสินใจได้ว่าเป็นการบุกรุก ซึ่งมีงานวิจัยที่ได้พัฒนาขั้นตอนวิธีทั้งสองรูปแบบ พบว่ามีข้อดีและข้อเสียที่แตกต่างกันคือ Signature-based Detection จะมีความถูกต้อง (Accuracy)ในการตรวจจับที่สูง ค่าความผิดพลาด (False) ต่ำ และไม่สามารถตรวจจับสิ่งผิดปกติแบบใหม่ๆ ที่ไม่เคยรู้จำมาก่อนได้ แต่ Anomaly-based detection จะมีค่า ความถูกต้องต่ำ ค่าความผิดพลาดที่สูง แต่สามารถตรวจจับสิ่งผิดปกติแบบใหม่ได้ ซึ่งจัดว่าเป็นความต้องการของระบบตรวจจับการบุกรุกที่สำคัญในปัจจุบัน ส่วน Hybrid technique เป็นการรวมความสามารถของ Signature-based Detection และ Anomaly-based detection ซึ่งเสมือนเป็นการรวมเอาความสามารถของทั้งสองแบบไว้ด้วยกัน คือจะเพิ่มประสิทธิภาพในการตรวจจับและลดค่าความผิดพลาด รวมไปถึงสามารถตรวจจับการบุกรุกแบบใหม่ๆได้เช่นกัน ซึ่งในปัจจุบัน ผู้บุกรุก (Intruder) จะพยายามคิดค้นขั้นตอนแบบใหม่ในการบุกรุกระบบเพิ่มขึ้นหลากหลายรูปแบบ โดยขั้นตอนวิธีต่างๆจะพยายามเอาชนะระบบที่องค์กรใช้ป้องกันอยู่ และองค์กรก็ไม่สามารถคาดเดาได้ว่าจะถูกบุกรุกระบบเมื่อใด ด้วยวิธีการใด และมีความร้ายแรงระดับใด จึงมีผู้วิจัยในปัจจุบันพยายามคิดค้นการพัฒนาขั้นตอนวิธีแบบ Hybrid technique มากขึ้นเพื่อเพิ่มขีดความสามารถในการป้องกันรักษาความปลอดภัยให้กับองค์กรให้ได้มากที่สุด

เทคนิคของการทำเหมืองข้อมูล (Data Mining) ใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่รวมถึงได้มีการนำมาใช้วิเคราะห์จำแนกสิ่งผิดปกติทางเครือข่าย โดยได้มีนักวิจัยได้พัฒนาขั้นตอนวิธีในรูปแบบ การจัดกลุ่ม (Clustering) และ การจำแนกประเภท (Classification) [3] โดยขั้นตอนวิธีของการจัดกลุ่มจะมีกลไกการทำงานเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยจะสามารถจัดกลุ่มข้อมูลสิ่งผิดปกติแยกออกจากข้อมูลที่ปกติได้โดยบางส่วนของข้อมูลที่ถูกจัดกลุ่มอาจเป็นการโจมตีแบบใหม่ แต่ในส่วนของการจำแนกประเภท จะให้ระบบเรียนรู้พฤติกรรมก่อนแบบมีผู้สอน (Supervised Learning) จะเป็นการสร้างแบบจำลองที่รู้จำพฤติกรรม (Classification Model) ที่ถูกต้อง ก่อนที่จะตัดสินข้อมูลจริงว่าเป็นการโจมตีระบบหรือไม่ แต่ขั้นตอนวิธีนี้อาจไม่สามารถจำแนกสิ่งผิดปกติแบบใหม่ได้แม่นยำ เนื่องจากเป็นข้อมูลที่ไม่เคยรู้จำและถูกสร้างในแบบจำลอง ดังนั้น จะเห็นได้ว่าอาจไม่มีระบบใดที่สามารถรองรับการบุกรุกได้อย่างแม่นยำ ซึ่งนักวิจัยในปัจจุบันได้พยายามพัฒนาขั้นตอนวิธีแบบใหม่ๆเพื่อเพิ่มประสิทธิภาพในการตรวจจับ ให้ทันต่อการบุกรุกโจมตีที่ไม่อาจคาดการณ์ช่วงเวลาได้ ประกอบกับต้องลดการแจ้งเตือนที่ผิดพลาด (False Alarm) ลงด้วย เพราะอาจหมายถึงเป็นการแจ้งเตือน หรือกีดกันพฤติกรรมการใช้งานปกติ (Normal Behavior) ของผู้ใช้ได้ในชีวิตประจำวันได้ การใช้เทคนิคหนึ่ง Adaboost.m1 เป็นวิธีการที่มีการรวบรวมเสียงข้างมากของคำตอบเพื่อจำแนกข้อมูล ซึ่งมีประสิทธิภาพสูงวิธีการหนึ่ง ซึ่งการนำ Adaboost.m1 มาใช้จำแนก และนำ Gain Ratio มาช่วยลดมิติของข้อมูล ให้เหลือเพียงมิติที่สำคัญทำให้เกิดต้นแบบของขั้นตอนวิธีที่จะสามารถ ตรวจจับสิ่งผิดปกติได้อย่างในยำ มีประสิทธิภาพสูงที่สุดเมื่อเทียบกับทุกวิธี และช่วยลดภาระการประมวลผลข้อมูลที่มีจำนวนมิติมากได้ จึงเป็นต้นแบบขั้นตอนวิธีที่อาจนำไปประยุกต์ใช้จริงกับการตรวจจับการบุกรุกจริงที่มีการวิเคราะห์ข้อมูลปริมาณมากและไม่สามารถคาดเดาความเสียหายได้ในอนาคต

1. วัตถุประสงค์การวิจัย

1. เพื่อพัฒนาขั้นตอนวิธีในการตรวจจับสิ่งผิดปกติทางเครือข่ายคอมพิวเตอร์ ด้วยเทคนิคเอดาบูทเอ็มวันและลดมิติด้วยเทคนิค Gain Ratio

2. เพื่อเปรียบเทียบประสิทธิภาพของการจำแนก (Classification) ขั้นตอนวิธีที่นำเสนอ และขั้นตอนวิธีอื่นๆของ Supervised Learning

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

เกรียงไกร [4] ได้ศึกษาระบบตรวจจับการบุกรุกเครือข่ายด้วยวิธี Naïve Bayes ในการวิเคราะห์เครือข่ายคอมพิวเตอร์ สำนักหอสมุด มหาวิทยาลัยเชียงใหม่ โดยใช้โปรแกรม WEKA Mining ในการวิเคราะห์ประสิทธิภาพ. การทดลองใช้ฐานข้อมูล KDD Cup'1999 ซึ่งมีประเภทของการบุกรุกอยู่ 37 ประเภทย่อย จัดอยู่ใน 5 กลุ่มใหญ่ ได่แก่ Dos Probe R2L U2R และ Normal (ปกติ) ผลการจำแนกข้อมูลพบว่าในส่วนข้อมูลที่เป็นตัวเลขนั้น ทำนายออกมาได้ไม่ดีที่เท่าควร จึงแบ่งช่วงข้อมูลให้เป็นตัวเลขเชิงกลุ่ม จึงมีประสิทธิภาพที่สูงขึ้น มีค่าความผิดพลาดที่ยอมรับได้ สามารถนำไปใช้ป้องกันระบบได้ แต่ผู้ที่พัฒนาต้องมีความรู้เรื่องพฤติกรรมการบุกรุกเครือข่ายแต่ละประเภทค่อนข้างดีด้วย

Yasmen [5] ได้ศึกษาการตรวจจับการบุกรุกทางเครือข่าย โดยมีการคัดเลือกคุณลักษณะสำคัญ ด้วย Correlation-based Feature Selection (CFS) และ Information Gain (IG) หลักการของการคัดเลือกคุณลักษณะสำคัญในงานวิจัยนี้คือ จะคัดเลือกคุณลักษณะสำคัญด้วย CFS ก่อน แต่เนื่องจาก CFS ไม่สามารถยืนยันได้ว่าคุณลักษณะที่เลือกมาเป็นคุณลักษณะที่เกี่ยวข้องกันสูง จึงใช้ IG ในการคัดเลือกลักษณะอีกครั้ง ซึ่งในลักษณะที่จะได้นำไปใช้ในขั้นตอนการทดลอง จะเป็นลักษณะจากขั้นตอน IG ร่วมกับลักษณะที่ได้จากขั้นตอน CFS ทำให้สามารถลกจำนวนคุณลักษณะสำคัญจาก 41 เหลือเพียง 15 คุณลักษณะ และเมื่อนำไปทดลองด้วยการใช้ Adaboost.M1 โดยมี Naïve Bayes เป็น Weak learner พบว่ามีความสามารถในการตรวจจับและจำแนกประเภทสิ่งผิดปกติได้ดีขึ้นสำหรับการจำแนกแบบ 5 กลุ่มของประเภทสิ่งผิดปกติร่วมกันข้อมูลปกติ

Sunila และ Ritu [6] ได้ทดสอบเปรียบเทียบขั้นตอนวิธีแบบไม่มีผู้สอน กับการรู้จำรูปภาพ ได้แก่ขั้นตอนวิธี K-Means, Hierarchical และ Make Density Based Clustering การทดลองได้ทดลองด้วยซอฟต์แวร์ WEKA Mining ผลการทดลองพบว่า K-Means มีการใช้ระยะเวลาในการประมวลผลต่ำที่สุด และมีความถูกต้องในการจำแนกข้อมูลดีที่สุด มากกว่า 2 ขั้นตอนวิธีที่ได้เปรียบเทียบ

Leena และ Shuwesh [7] ได้ใช้หลักการของ Incremental SVM (ISVM) ซึ่งขั้นตอนวิธีที่ทดสอบจะเป็นแบบ Binary SVM แต่เนื่องจากข้อมูลที่ทดลองเป็น Intrusion Detection ซึ่งมีหลายกลุ่มประเภทสิ่งผิดปกติ ขั้นตอนวิธีจึงต้อง Mapping ไปยังมิติที่สูงกว่าด้วยฟังก์ชัน Kernel จึงจะสามารถแยกกลุ่มข้อมูลได้ด้วยเส้นตรง ผลการทดลองพบว่าขั้นตอนวิธีที่นำเสนอสามารถเพิ่มความถูกต้องในการจำแนกข้อมูล ลดระยะเวลาในการฝึกสอนข้อมูลได้ ทำให้สามารถจัดการกับข้อมูล Intrusion ขนาดใหญ่ได้

จากข้อมูลดังกล่าวจะพบว่า มีการใช้ขั้นตอนวิธีการเรียนรู้ของเครื่องในการตรวจจับสิ่งผิดปกติที่หลากหลายวิธี โดยทั้งมีการลดมิติ และพัฒนาขั้นตอนวิธีร่วมกันเพื่อเป้าหมายในการเพิ่มประสิทธิภาพของการจำแนก ซึ่งการพัฒนาขั้นตอนวิธีในการตรวจจับจึงเป็นสิ่งที่น่าสนใจและสำคัญต่อประเด็นการป้องกันความปลอดภัยของระบบคอมพิวเตอร์ ในการใช้งานปกติ และทำธุรกรรมใดๆตลอดเวลาในยุคปัจจุบัน

วิธีดำเนินการวิจัย

1. เครื่องมือการวิจัย

1.1 โปรแกรมที่ใช้ในการวิเคราะห์ใช้โปรแกรมคอมพิวเตอร์ในการวิเคราะห์ค่าสถิติการพยากรณ์ ด้วยโปรแกรม WEKA Mining

1.2 เทคนิคที่ใช้ในการลดมิติข้อมูลได้แก่ Gain Ratio และเทคนิคที่ใช้ในการจำแนกข้อมูลได้แก่ Adaboost.m1 และเทคนิคของการเรียนรู้ด้วยเครื่อง (Machine Learning) ที่ใช้เป็น Weak Learner ทั้ง 5 เทคนิค ซึ่งจะใช้ในการเปรียบเทียบประสิทธิภาพอีกด้วย ได้แก่ ทฤษฎีต้นไม้ตัดสินใจ (Decision Tree) ทฤษฎีความน่าจะ

เป็นแบบนาอีฟเบย์ (Artificial Neural Network: Naive Bayes) ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbor: k-NN) โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron: MLP) และ วิธีเทคนิคซัพพอร์ตเวคเตอร์แมชชีนส์ (SVM)

2. กลุ่มเป้าหมาย

งานวิจัยนี้ได้ใช้ NDL-KDD ซึ่งเป็นฐานข้อมูลการบุกรุกเครือข่าย โดยการทดลองจะแบ่งข้อมูลออกเป็น 10 ส่วนสำหรับขั้นตอนการฝึกสอน (Training) และขั้นตอนการทดสอบ (Training) ประสิทธิภาพของระบบ (10-fold cross-validation) โดยฐานข้อมูล NDL-KDD มีรายละเอียดดังต่อไปนี้

ตารางที่ 1 รายละเอียดข้อมูลที่ใช้ในการทดลอง

ประเภทการบุกรุก	จำนวนข้อมูล
Normal	13449
Dos	9234
Probe	2289
R2L	209
U2R	11
รวม	25192

3. ขั้นตอนการดำเนินการวิจัย

ประกอบไปด้วย 5 ขั้นตอนดังต่อไปนี้

3.1 ขั้นตอนการเตรียมข้อมูลก่อนกระบวนการ (Data Preprocessing) โดยการดาวน์โหลดจากเว็บไซต์ [8] เนื่องจากฐานข้อมูลที่ใช้งานประกอบไปด้วย 22 ประเภทย่อยของการบุกรุก และ 1 ประเภทของ Normal ขั้นตอนนี้ ได้ทำการจัดกลุ่มประเภทย่อยๆของการบุกรุก ให้เป็นกลุ่มหลัก 5 กลุ่มหลัก ยกตัวอย่างเช่น Neptune จัดเป็นกลุ่ม Dos หรือ Portsweep จัดเป็นกลุ่ม Probe Multihop จัดเป็นกลุ่ม R2L และ buffer_overflow จัดเป็นกลุ่ม U2R เป็นต้น

3.2 ทำการลดมิติของข้อมูลด้วยเทคนิคมาตรฐานอัตราส่วนเกน จากทั้งหมด 41 คุณลักษณะ (Features) ให้เหลือน้อยลงเพื่อให้เหลือเพียงคุณลักษณะที่สำคัญต่อการจำแนกและลดภาระการประมวลผลลง

3.3 วิเคราะห์ข้อมูลที่จัดเตรียมด้วยเทคนิคเอดาบูทเอ็มวัน ด้วย Weak learner ทั้ง 5 เทคนิค โดยกำหนดให้ขั้นตอนการเรียนรู้สร้าง 50 ต้นแบบ สำหรับการรวบรวมเสียงข้างมาก เพื่อให้ได้คำตอบสุดท้าย ของการจำแนกข้อมูล

3.4 โดยเปรียบเทียบประสิทธิภาพของการจำแนกของ Weak learner ทั้ง 5 เทคนิค และเปรียบเทียบกับการจำแนกที่ไม่ได้ลดมิติแล้วใช้เอดาบูทเอ็มวัน รวมถึงเปรียบเทียบเวลาที่ใช้ในการประมวลผล

3.5 สรุปผลการศึกษา

4. สถิติที่ใช้ในการวิจัย การวิเคราะห์ประสิทธิภาพของการจำแนก ได้พิจารณาเปรียบเทียบค่าต่างๆดังต่อไปนี้ ค่าอัตราผลบวกจริง (True Positive Tate: TP Rate) อัตราผลบวกปลอม (False Positive Rate: FP Rate) ค่า ความแม่นยำ (Precision) ค่าความไว (Recall) ค่าถ่วงดุล (f-Measure) และค่าความถูกต้อง (Accuracy)

ผลการวิจัย

1. ผลพัฒนาขั้นตอนวิธีในการตรวจจับสิ่งผิดปกติทางเครือข่ายคอมพิวเตอร์ ด้วยเทคนิคเอดาบูทเอ็มวัน และลดมิติด้วยเทคนิค Gain Ratio

เมื่อนำข้อมูลมาลดมิติ และประมวลด้วย เอดาบูทเอ็มวัน ดังขั้นตอนวิธีที่นำเสนอ พบว่า ลดจำนวนมิติจาก 42 มิติ เหลือเพียง 23 มิติ (Cut off=0.15) มีค่าความถูกต้องและเวลาในการจำแนกข้อมูลเมื่อมีการใช้ Weak Learner ทุก Weak Learner สูงที่สุดเมื่อเทียบกับทุกวิธีการ ยกเว้นเทคนิคซัพพอร์ตเวคเตอร์แมชชีนส์ที่มีค่าเท่ากัน และใช้เวลาในการประมวลผลที่มากกว่าวิธีอื่นๆเช่นกัน แสดงดังตารางที่ 2

ตารางที่ 2 แสดงการเปรียบเทียบค่าความถูกต้องและเวลาในแต่ละขั้นตอนวิธี

จำนวน Feature	ขั้นตอนวิธี	ความถูกต้อง (Accuracy) (%)	เวลา (s)
41Features	J48	78.81	0.35
	Naive Bayes	75.88	1.16
	k-NN	74.25	59.17
	MLP	72.36	0.28
	SVM	43.07	40.09
23 Features	J48	99.54	0.68
	Naive Bayes	86.38	0.05
	k-NN	98.90	0
	MLP	95.90	21.6
	SVM	53.37	30.88
23Features+Adaboost .m1	J48	99.79	12.18
	Naive Bayes	99.49	6.41
	k-NN	98.99	4338.34
	MLP	99.28	1619.31
	SVM	53.37	138.87

2. ผลการเปรียบเทียบประสิทธิภาพของการจำแนก (Classification) ขั้นตอนวิธีที่นำเสนอ และขั้นตอนวิธีอื่นๆ ของ Supervised Learning

พบว่า ขั้นตอนวิธีที่นำเสนอมีค่าประสิทธิภาพสูงที่สุดในทุกๆสถิติของการวิเคราะห์แสดงดังตารางที่ 3 ตารางที่ 4 และตารางที่ 5

ตารางที่ 3 แสดงการเปรียบเทียบการจำแนกด้วยขั้นตอนวิธีต่างๆโดยไม่ได้ลดมิติและไม่ได้ใช้เอดาบูทเอ็มวัน

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	F-Measure
Decision Tree					
Normal	0.968	0.334	0.687	0.968	0.804
Dos	0.83	0.014	0.966	0.83	0.893
Probe	0.234	0	0.987	0.234	0.378
R2L+U2R	0.611	0.013	0.847	0.611	0.71
Naive Bayes					
Normal	0.907	0.261	0.724	0.907	0.805
Dos	0.706	0.044	0.888	0.706	0.786

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	F-Measure
Probe	0.264	0.006	0.861	0.264	0.404
R2L+U2R	0.932	0.064	0.636	0.932	0.756
			k-NN		
Normal	0.976	0.4	0.648	0.976	0.779
Dos	0.753	0.018	0.953	0.753	0.842
Probe	0.027	0	0.9	0.027	0.053
R2L+U2R	0.646	0.019	0.805	0.646	0.716
			MLP		
Normal	0.978	0.443	0.626	0.978	0.763
Dos	0.735	0.014	0.962	0.735	0.833
Probe	0	0	0	0	0
R2L+U2R	0.553	0.017	0.8	0.553	0.654
			SVM		
Normal	1	1	0.431	1	0.602
Dos	0	0	0	0	0
Probe	0	0	0	0	0
R2L+U2R	0	0	0	0	0

ตารางที่ 4 แสดงการเปรียบเทียบการจำแนกด้วยขั้นตอนวิธีต่างๆโดยได้ลดมิติเทคนิคมาตรฐานอัตราส่วนเกน และไม่ได้ใช้เอดาบูทเอ็มวัน

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	F-Measure
			Decision Tree		
Normal	0.997	0.005	0.996	0.997	0.996
Dos	0.998	0.001	0.998	0.998	0.998
Probe	0.841	0	0.939	0.841	0.887
R2L+U2R	0.987	0.001	0.989	0.987	0.988
			Naive Bayes		
Normal	0.846	0.074	0.929	0.846	0.886
Dos	0.916	0.023	0.959	0.916	0.937
Probe	0.477	0.014	0.225	0.477	0.306
R2L+U2R	0.794	0.08	0.497	0.794	0.611
			k-NN		
Normal	0.992	0.013	0.988	0.992	0.99
Dos	0.995	0.002	0.996	0.995	0.996
Probe	0.823	0.001	0.883	0.823	0.852
R2L+U2R	0.965	0.003	0.973	0.965	0.969
			MLP		
Normal	0.991	0.076	0.938	0.991	0.963
Dos	0.959	0.003	0.995	0.959	0.976
Probe	0.009	0	1	0.009	0.018

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	F-Measure
R2L+U2R	0.865	0.004	0.954	0.865	0.907
SVM					
Normal	1	1	0.534	1	**0.696**
Dos	0	0	0	0	**0**
Probe	0	0	0	0	**0**
R2L+U2R	0	0	0	0	**0**

ตารางที่ 5 แสดงการเปรียบเทียบการจำแนกด้วยขั้นตอนวิธีที่นำเสนอ

ประเภทการบุกรุก	TP Rate	FP Rate	Precision	Recall	F-Measure
Decision Tree					
Normal	0.999	0.003	0.997	0.999	**0.998**
Dos	0.999	0	0.999	0.999	**0.999**
Probe	0.914	0	0.98	0.914	**0.946**
R2L+U2R	0.995	0	0.996	0.995	**0.996**
Naive Bayes					
Normal	0.996	0.006	0.995	0.996	**0.996**
Dos	0.999	0.002	0.997	0.999	**0.998**
Probe	0.886	0.001	0.924	0.886	**0.905**
R2L+U2R	0.983	0.001	0.991	0.983	**0.987**
k-NN					
Normal	0.991	0.01	0.991	0.991	**0.991**
Dos	0.996	0.002	0.996	0.996	**0.996**
Probe	0.859	0.001	0.883	0.859	**0.871**
R2L+U2R	0.972	0.003	0.97	0.972	**0.971**
MLP					
Normal	0.995	0.009	0.993	0.995	**0.994**
Dos	0.998	0.001	0.998	0.998	**0.998**
Probe	0.85	0.001	0.93	0.85	**0.888**
R2L+U2R	0.975	0.002	0.979	0.975	**0.977**
SVM					
Normal	1	1	0.534	1	**0.696**
Dos	0	0	0	0	**0**
Probe	0	0	0	0	**0**
R2L+U2R	0	0	0	0	**0**

จากตารางที่ 5 พบว่า ขั้นตอนวิธีที่งานวิจัยนำเสนอมีประสิทธิภาพสูงที่สุด เมื่อพิจารณาจากค่า f-Measure นอกจากนี้ เมื่อพิจารณาแยกตามประเภทของการบุกรุกพบว่า สามารถจำแนกในแต่ละประเภทได้ดีที่สุดเช่นกัน ยกเว้นเทคนิคซัพพอร์ตเวคเตอร์แมชชีนส์ ที่มีประสิทธิภาพไม่แตกต่างกัน เมื่อมีการใช้เอดาบูทเอ็มวันและไม่ใช้เอดาบูทเอ็มวัน หากพิจารณาผลการทดลองโดยละเอียดจะพบว่า Weak Learner ที่มีประสิทธิภาพในการจำแนกสูงที่สุดคือ Decision Tree ซึ่งสามารถจำแนกการบุกรุกประเภท Dos ได้ถึง 0.999 รองลงมาคือ Normal

R2L+U2R และ Probe ตามลำดับ และมีค่าความถูกต้องถึงร้อยละ 99.79 รวมไปถึงใช้เวลาในการประมวลผลเพียง 12.18 วินาที

อภิปรายผลการวิจัย

1. ขั้นตอนวิธีในการตรวจจับสิ่งผิดปกติทางเครือข่ายคอมพิวเตอร์ ด้วยเทคนิคเอดาบูทเอ็มวัน และลดมิติด้วยเทคนิค Gain Ratio พบว่าเป็นขั้นตอนวิธีที่มีประสิทธิภาพสูงในการจำแนกข้อมูลการบุกรุกเครือข่ายคอมพิวเตอร์ เนื่องจากเทคนิคเอดาบูทเอ็มวัน จะมีการสร้างต้นแบบจำนวน 50 ต้นแบบ จากนั้นจึงนำมารวบรวมเสียงข้างมากเพื่อหาคำตอบว่าข้อมูลจัดอยู่ประเภทใด ซึ่งขั้นตอนการสร้างต้นแบบ จะมีการทายตำตอบและได้ค่าน้ำหนัก (Weight) ที่ได้จากการคำนวณทุกๆคำตอบ ส่งผลให้ในขั้นตอนการรวบรวมเสียงข้างมาก จะได้คำตอบสุดท้ายที่มีโอกาสถูกต้องสูง จึงเป็นข้อแตกต่างจากขั้นตอนวิธีทั่วไปตรงที่ได้หลายต้นแบบ มาช่วยหาคำตอบ สอดคล้องกับงานวิจัยของ Zhenyu และ Xiaoyao [9] เรื่อง Research on Adaboost.M1 with Random Forest พบว่า การใช้เทคนิคเอดาบูทเอ็มวัน มีประสิทธิภาพสูงในการตรวจจับการบุกรุก มีค่าความถูกต้องร้อยละ 80.84

2. เมื่อเปรียบเทียบประสิทธิภาพกับการจำแนกข้อมูลด้วยเทคนิคของการเรียนรู้ด้วยเครื่อง และเปรียบเทียบกับการลดมิติแล้วจำแนกด้วยเทคนิคของการเรียนรู้ด้วยเครื่อง พบว่า ขั้นตอนวิธีที่นำเสนอมีประสิทธิภาพสูงที่สุด รองลงมาคือ การลดมิติแล้วจำแนกด้วยเทคนิคของการเรียนรู้ด้วยเครื่อง และ การจำแนกข้อมูลด้วยเทคนิคของการเรียนรู้ด้วยเครื่อง ตามลำดับ ทำให้เห็นได้ว่า การลดมิติส่งผลให้มีประสิทธิภาพสูงขึ้น เนื่องจากใช้เพียงคุณลักษณะที่สำคัญและมีความเกี่ยวข้องกันเท่านั้นในการคำนวณ ในส่วนของการลดมิติแล้วใช้เทคนิคเอดาบูทเอ็มวัน ซึ่งมีประสิทธิภาพสูงสุด จึงเสมือนเป็นการเลือกใช้ข้อมูลเฉพาะมิติที่สำคัญ มาประมวลผลด้วยวิธีการที่มีการรวบรวมเสียงข้างมาก จากคำตอบที่มาจาก 50 ต้นแบบ ทำให้ขั้นตอนวิธีดังกล่าว มีประสิทธิภาพสูงที่สุดเมื่อเทียบกับทุกๆวิธี

3. เมื่อเปรียบเทียบเวลาที่ใช้ในการประมวลผลพบว่า เมื่อใช้เทคนิคเอดาบูทเอ็มวัน ทำให้มีเวลาที่สูงแตกต่างจากขั้นตอนวิธีอื่นๆมาก เนื่องมาจากเทคนิคเอดาบูทเอ็มวัน ต้องสร้างต้นแบบจำนวน 50 ต้นแบบเพื่อนำมารวบรวมเสียงข้างมาก จึงเสมือนกับต้องประมวลผลมากกว่าขั้นตอนวิธีอื่นๆหลายสิบเท่า ทั้งนี้ ยังมีผลการทดลองเมื่อมีการใช้ Decision Tree เป็น Weak Learner พบว่า มีค่าประสิทธิภาพสูงที่สุดและใช้เวลาไม่มากนักเมื่อเทียบกับวิธีอื่นๆ ซึ่งเหมาะแก่การนำต้นแบบนี้ไปพัฒนาต่อยอด เพื่อให้เป็นขั้นตอนวิธีที่สามารถตรวจจับการบุกรุกบนอุปกรณ์รักษาความปลอดภัยบนเครือข่ายคอมพิวเตอร์จริงได้ในอนาคต

ข้อเสนอแนะ

งานวิจัยนี้เป็นการศึกษาการลดมิติของข้อมูลที่มีมิติมากๆ และใช้ขั้นตอนวิธีที่มีการสร้างต้นแบบมากเพื่อช่วยรวบรวมเสียงข้างมากหาคำตอบ หากนำไปประยุกต์ใช้จริง อาจต้องดูขีดความสามารถและข้อจำกัดของอุปกรณ์ดักเก็บข้อมูลบนเครือข่ายว่า สามารถมีมิติข้อมูลต่างๆตามงานวิจัยนี้หรือไม่ และ หากต้องการความเร็วในการประมวลผล จะต้องปรับจำนวนของต้นแบบให้เหมาะสมมากขึ้นโดยยังคงไว้ซึ่งประสิทธิภาพที่สูงดังเดิม

การวิจัยในครั้งต่อไปควร พัฒนาขั้นตอนวิธีให้สามารถตรวจจับการบุกรุกชนิดใหม่ๆ ที่ไม่เคยพบมาก่อนได้ในข้อมูลสำหรับการฝึกสอน เพื่อให้เสมือจริงกับการตรวจจับการบุกรุกปัจจุบันที่มีโอกาสพบสิ่งผิดปกติแบบใหม่ๆเกิดขึ้นได้เสมอ

เอกสารอ้างอิง

[1] Zarpelão, Bruno Bogaz., Mianib,Rodrigo Sanches., Kawakania,Cláudio Toshio., and Alvarengaa, Sean Carlisto. (2017). A survey of intrusion detection in Internet of Things. *Journal of Network and Computer Applications*, 84(C), 25-37.

[2] Buczak, Anna L. and Guven, Erhan. (2017). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE COMMUNICATION SYRVEYS & TUTORIALS*, 18(2), 1153-1176.

[3] Stefanowski, Jerry. (2009). *Data Mining - Clustering*. สืบค้นจาก http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

[4] เกรียงไกร ชัยมินทร์. (2557). การตรวจจับการบุกรุกเครือข่ายสำหรับสำนักหอสมุด มหาวิทยาลัยเชียงใหม่ โดยใช้ตัวจำแนกข้อมูลนาอีฟเบส์. *ปริญญาวิทยาศาสตร์มหาบัณฑิต*, มหาวิทยาลัยเชียงใหม่. เชียงใหม่.

[5] Wahba, Yasmen, ElSalamouny, Ehab, ElTaweel, Ghada. (2015). Improving the Performance of Multi-class Intrusion Detection System using Feature Reduction. *IJSCI International Journal of Computer Science*, 12(3), 255-262.

[6] Godara, Sunila and Yadav, Ritu. (2013). PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHM FOR CHARACTER RECOGNITION USING WEKA TOOL. *International Journal of Advance Computer and Mathematical Sciences*, 4(1), 119-123.

[7] Bramhe, Leena and Shukla, Shuwesh. (2013). A Novel Approach for Improve detection Rate in Anomaly based Intrusion Detection System. *International Journal of IT, Engineering and Applied Science Research (IJIEASR)*, 2(5), 40-44.

[8] Lashkari, Arash Habibi. (2015). NSL-KDD dataset. สืบค้นจาก http://www.unb.ca/cic/datasets/nsl.html

[9] Zhang, Zhenyu and Xie, Xiaoyao. (2010). Research on Adaboost.M1 with Random Forest. *2nd International Conference on Computer Engineering and Technology*, 1(7), 647-652.