การตรวจจับการบุกรุกด้วยเทคนิคการจำแนกในการทำเหมืองข้อมูล

Intrusion Detection Using Classification Techniques in Data Mining

ผดุง นันอำไพ[1*] และ จารี ทองคำ[2]

Phadung Nanaumphai[1*] and Jaree Thongkam[2]

คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม[1,2]

Faculty of Informatics at Mahasarakham University[1,2]

E-Mail joshpy@me.com, jaree.thongkam@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) พัฒนาแบบจำลองที่สามารถจำแนกและตรวจจับการบุกรุกในระบบเครือข่าย และ 2) เปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการบุกรุกในระบบเครือข่าย เครื่องมือที่ใช้ในการจำแนกข้อมูลรูปแบบการบุกรุกบนระบบเครือข่าย คือ โปรแกรม WEKA โดยใช้เทคนิคการจำแนกในการทำเหมืองข้อมูล 4 เทคนิคด้วยกันคือ เทคนิค Decision Table, เทคนิค Naïve Bayes, เทคนิค RIPPER และเทคนิค PART decision list เพื่อพัฒนาแบบจำลองและเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการบุกรุกในระบบเครือข่าย ในงานวิจัยนี้ใช้ชุดข้อมูลการบุกรุกระบบเครือข่ายจากฐานข้อมูลความรู้ KDD Cup'99 หลักการ 10-Fold Cross Validation ได้ถูกนำมาใช้ในการแบ่งชุดข้อมูลออกเป็นชุดเรียนรู้และชุดทดสอบ สถิติที่ใช้ในการวิจัย 1) ค่าความแม่นยำ Precision 2) ค่าระลึก Recall และ 3) ค่า F-Measure

ผลการทดลองพบว่า 1) แบบจำลองที่ใช้เทคนิค RIPPER มีค่า Precision มากที่สุดคือ 99.00% และเทคนิค PART decision list มีค่า Precision 98.20% ตามด้วยเทคนิค Decision Table มีค่า Precision 97.50% และเทคนิคที่มีค่าความถูกต้องน้อยที่สุดคือ เทคนิค Naïve Bayes มีค่า Precision 49.40% และ 2) ผลเปรียบเทียบการวิเคราะห์แบบจำลองการจำแนกการตรวจจับการบุกรุกของแต่ละเทคนิค พบว่า แบบจำลองที่ใช้เทคนิค RIPPER ให้ค่าเฉลี่ยทางสถิติเป็นเปอร์เซ็นต์มากที่สุดอย่างมีนัยสำคัญ จากเทคนิคทั้งหมด 4 เทคนิค

คำสำคัญ: การตรวจจับการบุกรุก, เทคนิคการจำแนกข้อมูล, การทำเหมืองข้อมูล, เทคนิคการจำแนกด้วยกฎ

Abstract

The purposes of the research were 1) to develop intrusion detection model and 2) to compare the effectiveness of intrusion detection model. The tools used in classification are WEKA. using four classification techniques including Decision table, Naïve Bayes, RIPPER and PART decision list in data mining. In this thesis, the knowledge database "KDD Cup'99" is used. 10-fold cross validation is employed to divided data into training and testing sets. The statistics used were the percentage, Precision, Recall and F-Measure

The research findings showed that 1) RIPPER has highest precision which is up to 99.00% Then, PART decision list has precision which is up to 98.20%. Follow by Decision Table has highest precision which is up to 97.50%. The lowest precision is Naïve Bayes 49.40%. 2) The Comparative analysis of intrusion detection model showed that RIPPER model has highest average significantly.

Keywords: Intrusion Detection, Classification Techniques, Data Mining, Rule Based Techniques

บทนำ

สังคมในปัจจุบันมีความต้องการในการรับและส่งข้อมูลสารสนเทศผ่านทางระบบเครือข่ายคอมพิวเตอร์เป็นจำนวนมาก อีกทั้งยังมีอุปกรณ์ต่าง ๆ ที่ใช้ในการรับและส่งข้อมูลที่แตกต่างกันไป ไม่ว่าจะเป็นเครื่องคอมพิวเตอร์ โทรศัพท์มือถือ ออุปกรณ์อื่น ๆ จากการสำรวจการมีการใช้เทคโนโลยีสารสนเทศและการสื่อสารในครัวเรือน โดยสำนักงานสถิติแห่งชาตินั้นมีค่าสถิติการรับส่งข้อมูลเพิ่มขึ้นทุกปี [1] ดังนั้นสิ่งที่มีความสำคัญและไม่ควรมองข้าม คือความปลอดภัยบนระบบเครือข่ายคอมพิวเตอร์ เนื่องจากในปัจจุบันมีผู้ประสงค์ร้ายและต้องการหาผลประโยชน์ในระบบเครือข่ายคอมพิวเตอร์เป็นจำนวนมาก แม้จะมีอุปกรณ์และเครื่องมือต่าง ๆ ที่สามารถช่วยให้ระบบเครือข่ายคอมพิวเตอร์ปลอดภัยขึ้น แต่ผู้บุกรุกก็ยังมีเทคนิควิธีการและเครื่องมือต่าง ๆ มากมายที่ใหม่ขึ้นและซับซ้อนขึ้นเพื่อหลีกเลี่ยงการตรวจจับ

การนำระบบตรวจจับการบุกรุก (Intrusion Detection System: IDS) เข้ามาใช้ทำให้ช่วยเพิ่มความปลอดภัยในระบบเครือข่ายคอมพิวเตอร์ได้มากขึ้น จึงมีงานวิจัยที่นำเสนอทฤษฎีและเทคนิคต่าง ๆ ที่นำมาใช้ในการวิเคราะห์รูปแบบการบุกรุกหลากหลายรูปแบบ ซึ่งในงานวิจัยนี้จะเปรียบเทียบเทคนิคการตรวจจับการบุกรุกเพื่อวัดประสิทธิภาพ โดยใช้เทคนิคการจำแนก 4 เทคนิคด้วยกันคือ เทคนิค Decision Table เป็นวิธีที่นำมาทดสอบการทำงานร่วมกันของเงื่อนไขที่มีหลายเงื่อนไข มีลักษณะคล้ายกับ Decision Tree แต่จะอยู่ในรูปของตาราง เทคนิค Naïve Bayes เป็นเทคนิคที่นิยมมากในการนำมาใช้จำแนกข้อมูล เนื่องจากมีแบบจำลองที่เข้าใจได้ง่ายและไม่ซับซ้อน เทคนิคนี้ใช้หลักการของความน่าจะเป็น เทคนิค RIPPER (Rule-Based Classification) เป็นอัลกอริทึมที่สามารถสร้างกฎเองได้ โดยเรียนรู้จากข้อมูลที่เตรียมไว้ให้ และเทคนิค PART decision list ซึ่งเป็นอัลกอริทึมที่สามารถจัดการกับข้อมูลที่หายไปและคุณลักษณะทางตัวเลขที่ต่างกันได้ดี โดยการวิจัยครั้งนี้มีจุดประสงค์เพื่อศึกษาข้อดีและข้อเสียในการจำแนกข้อมูลด้วยเทคนิคที่แตกต่างกัน เพื่อให้ได้เทคนิคการจำแนกข้อมูลบนระบบเครือข่ายคอมพิวเตอร์ที่มีประสิทธิภาพ

1. วัตถุประสงค์การวิจัย

1.1 เพื่อพัฒนาแบบจำลองที่สามารถจำแนกและตรวจจับการบุกรุกในระบบเครือข่าย
1.2 เพื่อเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการบุกรุกในระบบเครือข่าย

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

การทำเหมืองข้อมูล [2] เป็นกระบวนการหาประโยชน์หรือความรู้ที่ไม่เคยรู้มาก่อนจากฐานข้อมูลขนาดใหญ่ เช่น รูปแบบ ความสัมพันธ์และกฎ ที่ซ่อนอยู่ภายในฐานข้อมูล เพื่อนำความรู้ที่ได้ไปใช้ประโยชน์และประกอบการตัดสินใจได้ ยกตัวอย่างเช่น การเก็บข้อมูลลูกค้าของห้างสรรพสินค้าโดยการให้ลูกค้านั้นสมัครสมาชิกและนำข้อมูลเหล่านั้นไปกลั่นกรองโดยผ่านเทคนิคการทำเหมืองข้อมูล เพื่อที่จะสามารถจัดโปรโมชั่นให้ลูกค้า และสามารถจัดวางสินค้าที่ลูกค้าชอบซื้อควบคู่กันไว้ใกล้ๆ กัน การจำแนกข้อมูล (Classification) เป็นหนึ่งในเทคนิคการทำเหมืองข้อมูล การจำแนกข้อมูลนั้นเป็นกระบวนการทางสถิติเพื่อจัดประเภท และวิเคราะห์หารูปแบบของกลุ่มข้อมูลใหม่ โดยทำการแบ่งข้อมูลออกเป็น 2 กลุ่มหลัก ๆ คือ ข้อมูลที่ใช้ในการสอน (Training Data) และข้อมูลที่ใช้ในการทดสอบ (Test Set) เพื่อนำผลลัพธ์ที่ได้มาเป็นแบบจำลองหรือโมเดล

ระบบตรวจจับการบุกรุก [3] คือซอฟต์แวร์หรือฮาร์ดแวร์ที่ถูกนำมาใช้ในการตรวจสอบข้อมูลที่ถูกรับและส่งภายในระบบเครือข่ายคอมพิวเตอร์ (Traffic) โดยระบบตรวจจับการบุกรุกจะทำหน้าที่ในการวิเคราะห์รูปแบบของข้อมูลย่อยที่ถูกรับและส่ง (Packet) เพื่อหารูปแบบและความผิดปกติที่มีพฤติกรรมเข้าข่ายการบุกรุกระบบเครือข่าย และจะทำการแจ้งเตือนไปยังผู้ดูแลระบบให้ตรวจสอบเพื่อป้องกันและแก้ไขได้อย่างทันท่วงที

เทคนิคตารางตัดสินใจ (Decision Table) [4] เป็นวิธีที่นำมาทดสอบการทำงานร่วมกันของเงื่อนไขที่มีหลายเงื่อนไข มีลักษณะคล้ายกับต้นไม้ตัดสินใจ (Decision Tree) แต่จะอยู่ในรูปของตาราง ซึ่งตารางดังกล่าวจะประกอบไปด้วยเงื่อนไข (Conditions) และการกระทำ (Actions)

เทคนิคการจำแนกข้อมูลด้วย Naïve Bayes [5] เป็นวิธีที่ได้รับความนิยมในการนำมาใช้จำแนกข้อมูล เนื่องจากมีแบบจำลองที่เข้าใจได้ง่ายและไม่ซับซ้อน เทคนิคนี้ใช้หลักการของความน่าจะเป็น

เทคนิค RIPPER (Rule-Based Classification) [6] เป็นเทคนิคที่พัฒนาจากเทคนิค IRIP สามารถสร้างกฎเองได้ โดยการเรียนรู้จากข้อมูลที่เตรียมไว้ให้กฎที่สร้างขึ้นจะอยู่ในรูปของ if then else มีขั้นตอนหลักอยู่ 3 ขั้นตอน ขั้นตอนแรกคือการสร้างกฎเริ่มต้น (Building) จะแบ่งเป็น 2 กระบวนการคือ กระบวนการเจริญเติบโต (Growth) โดยกระบวนการนี้จะทำการเพิ่มจำนวนกฎให้เหมาะสมกับข้อมูล จากนั้นจะตัดกฎที่ไม่ไม่จำเป็นหรือกฎที่ลดประสิทธิภาพในการเรียนรู้ออก (Pruning) ขั้นตอนที่ 2 คือขั้นตอนการเพิ่มประสิทธิภาพ (Optimization) โดยจะมีการเพิ่มคุณลักษณะให้แต่ละกฎ ขั้นตอนที่ 3 คือ ขั้นตอนการลบกฎออกจาก Rule set และเลือกเฉพาะกฎที่ดีที่สุดเก็บไว้

เทคนิค PART Decision List [7] เป็นเทคนิคที่พัฒนาจาก C4.5 และ RIPPER โดยรวมทั้ง 2 เทคนิคเข้าด้วยกัน มีจุดเด่นคือ สามารถเรียนรู้กฎได้เองเหมือนเทคนิค RIPPER และสามารถจัดการกับข้อมูลที่หายไปและคุณลักษณะทางตัวเลขที่ต่างกันได้ดี

ฐานข้อมูลคามรู้ KDD Cup’99 [8] เป็นชุดข้อมูลที่ใช้ในการทดสอบระบบความปลอดภัยของเครือข่ายคอมพิวเตอร์ใช้ในการแข่งขัน The Third International Knowledge Discovery and Data Mining เป็นการเก็บข้อมูลการโจมตีที่ Lincoln Laboratory ของสถาบัน Massachusetts Institute of Technology ใช้ระยะเวลาในการเก็บข้อมูล 9 สัปดาห์

อรนุช พันโท และมนต์ชัย เทียนทอง [9] ได้เสนอการเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบข้อมูลการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล โดยใช้วิธีการ 10–fold validation แลใช้เทคนิคการจำแนกข้อมูล 3 เทคนิคคือ Bayes, Decision Tree, และ Rules-Based ผลปรากฏว่าการจำแนกข้อมูลด้วยเทคนิค Decision Tree มีประสิทธิภาพสูงที่สุดคือ 82.78%

ปรีชา สมหวัง และศิริวัฒน์ โทศิริกุล [10] ได้นำเสนอแนวคิดในการตรวจจับการใช้งานคอมพิวเตอร์ในทางที่ผิด โดยใช้การวิธีการวิเคราะห์องค์ประกอบหลักเพื่อคัดแยกข้อมูลและใช้ฟัซซี ซีมีน เพื่อจัดกลุ่ม ผลการทดลองมีความแม่นยำ 81.48% และมีความผิดพลาดในการตรวจจับ 18.52%

ปวีณา ชัยวนารมย์ [11] ได้เสนอการพยากรณ์การเกิดความเครียดในหลายระดับด้วยเทคนิคการทำเหมืองข้อมูล และใช้อัลกอริทึมในการทำเหมืองข้อมูลจำนวน 6 อัลกอริทึมในการสร้างแบบจำลอง คือ Bayesian Network, Naïve Bayesian, Decision Tree, Decision Table, Partial Rules (PART) และ Multilayer Perceptron (MLP) จากการทดสอบพบว่า แบบจำลองที่เหมาะสมในการพยากรณ์ความเครียดคือแบบจำลองของอัลกอริทึม Multilayer Perceptron (MLP) ซึ่งมีค่าความถูกต้องเท่ากับ 81% ค่าความแม่นยำเท่า 0.81 ค่าความระลึกเท่ากับ 0.81 และค่าความเหวี่ยงเท่ากับ 0.81

ณัฐวุฒิ ปั้นรูป และอัฐพร กิ่งบู [12] ได้เสนอวิธีการจำแนกข้อมูลให้ดีขึ้นโดยมีการรวมเทคนิคจัดกลุ่มเคมีนและเทคนิคจำแนกข้อมูลเอสวีเอ็มแบบหลายกลุ่มทำงานร่วมกัน เลือกใช้ชุดข้อมูลจากฐานข้อมูลความรู้ KDD Cup’99 ผลปรากฏว่า เทคนิคที่ได้ผลดีที่สุดคือเทคนิค KMMSVM มีค่าความถูกต้องอยู่ที่ 99.32% ตามด้วยเทคนิคการจัดกลุ่มเคมีน ซึ่งมีค่าความถูกต้องอยู่ที่ 99.19%

ธนกร มีหินกอง [13] ได้เสนอเทคนิคการตรวจหาการบุกรุกเป็นระบบที่ใช้ในการตรวจหาผู้ที่บุกรุกเข้ามาในเครือข่ายคอมพิวเตอร์เพื่อมุ่งทำลายระบบหรือขโมยข้อมูล ด้วยเทคนิคเมืองข้อมูลวิเคราะห์ด้วยกฎความสัมพันธ์จากโครงข่ายประสาทเทียม ผลจากการทดลองพบว่า ระบบการตรวจหาการบุกรุกที่ได้พัฒนาขึ้นนี้สามารถรายงานผลได้อย่างรวดเร็ว โดยมีค่าความเที่ยงที่ 97.4 %

ไพชยนต์ คงไชย [14] ได้เสนอการพัฒนาขั้นตอนวิธีเพื่อจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัดเพื่อเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูล โดยเปรียบเทียบอัลกอริทึม CCFAR กับอีก 9 อัลกอริทึม คือ C4.5, RIPPER, OneR, CBA, GARC, OAC, FURIA, CFAR และ CFARC ผลปรากฏว่า อัลกอริทึม CCFAR มีความเหมาะสมของกฎอยู่ในอันดับที่ 1 คือ 0.9104 ซึ่งมีค่าความเหมาะสมของกฎแตกต่างจากอัลกอริทึม CFARC ที่เป็นอันดับ 2 ไม่มาก

วิธีดำเนินการวิจัย

1. เครื่องมือการวิจัย

1.1 โปรแกรม WEKA ใช้ในการจำแนกข้อมูลรูปแบบการบุกรุกบนระบบเครือข่าย

1.2 เทคนิคที่ใช้ในการจำแนกข้อมูล 4 เทคนิค ได้แก่ เทคนิคการจำแนกข้อมูล Decision Table เทคนิคการจำแนกข้อมูล Naïve Bayes เทคนิคการจำแนกข้อมูล RIPPER และเทคนิคการจำแนกข้อมูล PART Decision List

2. ขั้นตอนการดำเนินการวิจัย

2.1 ทำความเข้าใจและศึกษาการบุกรุกบนระบบเครือข่ายคอมพิวเตอร์รูปแบบต่าง ๆ และศึกษาเทคนิคการจำแนกในการทำเหมืองข้อมูลเพื่อหาเทคนิคที่เหมาะสมในการจำแนกรูปแบบการบุกรุกบนระบบเครือข่ายคอมพิวเตอร์ โดยศึกษาจากเอกสารและงานวิจัยที่เกี่ยวข้อง

2.2 ทำความเข้าใจข้อมูล ผู้วิจัยได้คัดเลือกข้อมูลจากฐานข้อมูลความรู้ KDD Cup’99 ซึ่งเป็นชุดข้อมูลที่ใช้ในการทดสอบระบบความปลอดภัยของเครือข่ายคอมพิวเตอร์ มีข้อมูลจำนวน 494,202 เรคคอร์ด แอททริบิว 41 แอททริบิว และคลาสผลลัพธ์ 23 คลาส แสดงดังตารางที่ 1

ตารางที่ 1 แสดงชื่อและประเภทของแอททริบิว

ที่	ชื่อ	ประเภท	ที่	ชื่อ	ประเภท
1	duration	continuous	22	is_guest_login	discrete
2	protocol_type	discrete	23	count	continuous
3	service	discrete	24	srv_count	continuous
4	flag	discrete	25	serror_rate	continuous
5	src_bytes	continuous	26	srv_serror_rate	continuous
6	dst_bytes	continuous	27	rerror_rate	continuous
7	land	discrete	28	srv_rerror_rate	continuous
8	wrong_fragment	continuous	29	same_srv_rate	continuous
9	urgent	continuous	30	diff_srv_rate	continuous
10	hot	continuous	31	srv_diff_host_rate	continuous
11	num_failed_logins	discrete	32	dst_host_count	continuous
12	loged_in	discrete	33	dst_host_srv_count	continuous
13	lnum_compromised	continuous	34	dst_host_same_srv_rate	continuous
14	lroot_shell	discrete	35	dst_host_diff_srv_rate	continuous
15	lsu_attempted	discrete	36	dst_host_same_src_port_rate	continuous
16	lnum_root	continuous	37	dst_host_srv_diff_host_rate	continuous
17	lnum_file_creations	continuous	38	dst_host_serror_rate	continuous
18	lnum_shells	continuous	39	dst_host_srv_serror_rate	continuous

ที่	ชื่อ	ประเภท	ที่	ชื่อ	ประเภท
19	lnum_access_files	continuous	40	dst_host_rerror_rate	continuous
20	lnum_outbound_cmds	continuous	41	dst_host_srv_rerror_rate	continuous
21	is_host_login	discrete	42	label (Class)	discrete

2.3 การคัดลือกข้อมูลผู้วิจัยได้ลบคลาสผลลัพธ์ออกให้เหลือจำนวน 13 คลาส

2.4 สร้างแบบจำลองโดยใช้โปรแกรม WEKA และเลือกใช้เทคนิค 4 เทคนิค คือ Decision Table, Naïve Bayes, RIPPER และ PART decision list

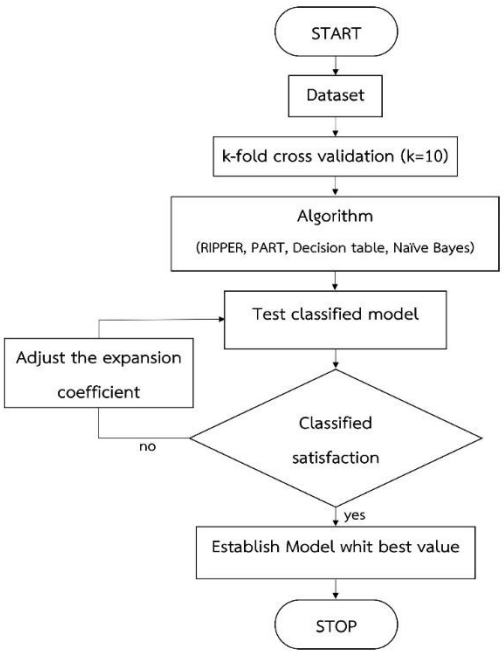
2.5 ประเมินผลและเปรียบเทียบประสิทธิภาพแบบจำลอง ค่าที่ใช้ในการวัดประสิทธิภาพคือค่าทางสถิติ ได้แก่ Precision, Recall และ F-Measure

3. สถิติที่ใช้ในการวิจัย

3.1 ค่าความแม่นยำ Precision คำนวณจากจำนวนข้อมูลที่จำแนกได้ถูกต้องของกลุ่มนั้นหารด้วยจำนวนของข้อมูลที่ถูกจำแนกว่าเป็นกลุ่มนั้นทั้งหมด โดยค่า Precision นั้นจะเป็นการวัดความสามารถของแบบจำลองโดยการขจัดข้อมูลที่ไม่เกี่ยวข้องออกไป ผลลัพธ์จะบ่งบอกว่าสามารถขจัดการจำแนกประเภทที่ผิดพลาดได้มากน้อยเพียงใด

3.2 ค่าระลึก Recall คำนวณจากจำนวนข้อมูลที่จำแนกได้ถูกต้องของกลุ่มนั้นหารด้วยจำนวนของข้อมูลที่มีอยู่จริงในกลุ่มนั้น โดยค่า Recall นั้นจะเป็นการวัดความสามารถของแบบจำลองว่าการจำแนกข้อมูลที่เกี่ยวข้องออกมานั้นมีประสิทธิภาพมากน้อยเพียงใด

3.3 ค่า F-Measure เป็นการวัดค่าความแม่นยำ Precision และค่าระลึก Recall พร้อมกันของแบบจำลองขั้นตอนการสร้างแบบจำลอง แสดงดังภาพที่ 1

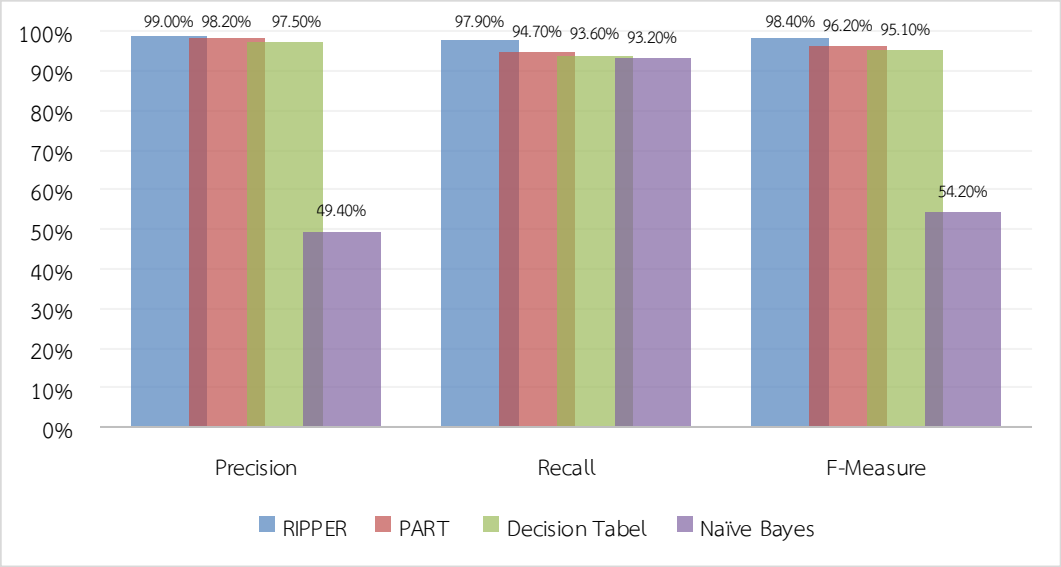


ภาพที่ 1 ผังงานขั้นตอนการสร้างแบบจำลอง

ผลการวิจัย

1. ผลการพัฒนาแบบจำลอง

ผู้วิจัยได้ดำเนินการพัฒนาแบบจำลองการตรวจจับการบุกรุกด้วยเทคนิคการจำแนกในการทำเหมืองข้อมูลด้วย เทคนิค Decision Table, เทคนิค Naïve Bayes, เทคนิค RIPPER และเทคนิค PART decision list ได้ผลลัพธ์ แสดงดังภาพที่ 2



ภาพที่ 2 กราฟเปรียบเทียบผลการวิเคราะห์ค่า Precision, Recall และ F-Measure

จากภาพที่ 2 พบว่า แบบจำลองการจำแนกการตรวจจับการบุกรุกที่ใช้เทคนิค RIPPER มีค่าเฉลี่ยเป็นเปอร์เซ็นต์มากที่สุด คือ ค่าเฉลี่ย Precision เท่ากับ 99%, ค่าเฉลี่ย Recall เท่ากับ 97.90% และค่าเฉลี่ย F-Measure เท่ากับ 98.40% และเทคนิค PART decision list มีค่าเฉลี่ย Precision เท่ากับ 98.20%, ค่าเฉลี่ย Recall เท่ากับ 94.70% และค่าเฉลี่ย F-Measure เท่ากับ 96.20% ตามด้วยเทคนิค Decision Table มีค่าเฉลี่ย Precision เท่ากับ 97.50%, ค่าเฉลี่ย Recall เท่ากับ 93.60% และค่าเฉลี่ย F-Measure เท่ากับ 95.10% และเทคนิคที่มีค่าเฉลี่ยทางสถิติน้อยที่สุดคือเทคนิค Naïve Bayes มีค่าเฉลี่ย Precision เท่ากับ 49.40% ค่าเฉลี่ย Recall เท่ากับ 93.20% และค่าเฉลี่ย F-Measure เท่ากับ 54.20%

2. ผลการเปรียบเทียบประสิทธิภาพ

ผู้วิจัยได้เปรียบเทียบแบบจำลองการจำแนกการตรวจจับการบุกรุกที่ใช้เทคนิค Decision Table เทคนิค Naïve Bayes, เทคนิค RIPPER และเทคนิค PART Decision list โดยสถิติที่ใช้ในการเปรียบเทียบเทคนิคทั้ง 4 เทคนิคคือ Precision Recall และ F-Measure แสดงดังตารางที่ 2

ตารางที่ 2 แสดงการเปรียบเทียบค่าทางสถิติของแบบจำลองทั้ง 4 เทคนิค

Algorithms	Precision (%)	Recall (%)	F-Measure (%)
RIPPER	99.00	97.90	98.40
PART decision list	98.20	94.70	96.20
Decision Table	97.50	93.60	95.10
Naïve Bayes	49.40	93.20	54.20

จากตารางที่ 2 ผลเปรียบเทียบการวิเคราะห์แบบจำลองการจำแนกการตรวจจับการบุกรุกของแต่ละเทคนิคพบว่าแบบจำลองที่ใช้เทคนิค RIPPER ให้ค่าเฉลี่ยทางสถิติเป็นเปอร์เซ็นต์มากที่สุด อย่างมีนัยสำคัญ จากเทคนิคทั้งหมด 4 เทคนิค

อภิปรายผลการวิจัย

การพัฒนาแบบจำลองเพื่อเปรียบเทียบการจำแนกรูปแบบการบุกรุกบนระบบเครือข่ายคอมพิวเตอร์ประกอบด้วย 5 ขั้นตอนได้แก่ 1) ขั้นตอนการทำความเข้าใจและศึกษาการบุกรุกบนระบบเครือข่ายคอมพิวเตอร์ 2) ขั้นตอนการทำความเข้าใจข้อมูล 3) ขั้นตอนการคัดเลือกข้อมูล 4) ขั้นตอนการสร้างแบบจำลอง 5) ขั้นตอนการประเมินผลและเปรียบเทียบประสิทธิภาพแบบจำลอง ผลการวิจัยพบว่า การจำแนกรูปแบบการบุกรุกนั้นเหมาะที่จะนำไปพัฒนาเป็นระบบตรวจจับการบุกรุกบนระบบเครือข่าย เนื่องจากสามารถจำแนกข้อมูลที่มีจำนวนมากได้อย่างมีประสิทธิภาพ สอดคล้องกับปรีชา สมหวัง และศิริวัฒน์ โทศิริกุล [11] ได้นำเสนอแนวคิดในการตรวจจับการใช้งานคอมพิวเตอร์ในทางที่ผิด ผลการทดลองมีความแม่นยำ 81.48%

ข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเทคนิคการจำแนก เป็นส่วนหนึ่งในการทำเหมืองข้อมูล ถึงแม้แบบจำลองนั้นมีประสิทธิภาพ แต่ในปัจจุบันนั้นมีรูปแบบการบุกรุกระบบเครือข่ายหลากหลายและซับซ้อนมากขึ้น การนำระบบตรวจจับการบุกรุกไปใช้ในระบบเครือข่ายสามารถทำให้ระบบมีความปลอดภัยมากขึ้น แต่ก็ยังไม่สามารถตรวจจับรูปแบบการบุกรุกทั้งหมดได้

การวิจัยในครั้งนี้ผู้วิจัยได้ใช้ฐานข้อมูลความรู้ KDD Cup’99 ในงานวิจัยซึ่งเป็นชุดข้อมูลที่ถูกสร้างขึ้นหลายปีแล้ว ปัจจุบันนั้นมีรูปแบบการโจมตีและบุกรุกระบบเครือข่ายรูปแบบใหม่ ๆ เกิดขึ้น อาจทำให้แบบจำลองของระบบตรวจจับการบุกรุกในงานวิจัยนี้ไม่สามารถตรวจจับการบุกรุกดังกล่าวได้ จึงแนะนำให้มีการวิจัยในครั้งต่อไป

เอกสารอ้างอิง

[1] สำนักงานสถิติแห่งชาติกระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม. (2561). *การสำรวจการมีการใช้เทคโนโลยีสารสนเทศและการสื่อสารในครัวเรือน.* สืบค้นจาก http://www.nso.go.th/sites/2014/DocLib13/ด้านICT/เทคโนโลยีในครัวเรือน/2561/ict61-CompleteReport-Q1.pdf

[2] เอกสิทธิ์ พัชรวงศ์ศักดา. (2556). *คู่มือการใช้งาน Weka Explorer เบื้องต้น*. กรุงเทพฯ: เอเชียดิจิตอลการพิมพ์.

[3] ธนกร มีหินกอง และประสงค์ ปราณีตพลกรัง. (2555). *ระบบตรวจหาการบุกรุก*. กรุงเทพฯ: สถาบันวิทยาการสารสนเทศ มหาวิทยาลัยศรีปทุม.

[4] ธนาวุฒิ เอื้อชัยกุล. (2551). *การสร้างชุดคำสั่งบีเพลสำหรับกฎธุรกิจจากตารางการตัดสินใจ*. (วิทยานิพนธ์ปริญญามหาบัณฑิต) จุฬาลงกรณ์มหาวิทยาลัย: กรุงเทพฯ.

[5] เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *An Introduction to Data Mining Techniques*. กรุงเทพฯ: เอเชียดิจิตอลการพิมพ์.

[6] Anil Rajput. (2011). J48 and JRIP Rules for E-Governance Data. *International Journal of Computer Science and Security,* (5).

[7] Frank Eibe & Witten H. Ian. (1998). Generating Accurate Rule Sets Without Global Optimization. *(the Fifteenth International Conference on Machine Learning).*

[8] KDD Cup. (1999). *Data*. Retrieved from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[9] อรนุช พันโท และมนต์ชัย เทียนทอง. (2557). การเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล. วารสารเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏอุบลราชธานี, 4(1), pp. 1-11.

[10] ปรีชา สมหวัง และศิริวัฒน์ โทศิริกุล. (2553). ระบบตรวจจับการใช้งานคอมพิวเตอร์ในทางที่ผิด. *National Conference on Information Technology,* NCIT2010. pp. 409-414.

[11] ปวีณา ชัยวนารมย์. (2558). *การพัฒนาแบบจำลองเพื่อพยากรณ์การเกิดความเครียดในหลายระดับด้วยเทคนิคการทำเหมืองข้อมูล*. (รายงานการวิจัย). กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์.

[12] ณัฐวุฒิ ปั้นรูป และอัฐพร กิ่งบู. (2558). Data Classification by K-Means and Multi-Class SVM for Intrusion Detection System. *National Conference on Information Technology, 7.*

[13] ธนกร มีหินกอง. (2558). สถาปัตยกรรมความรู้ด้านความมั่นคงปลอดภัยไซเบอร์เพื่อสนับสนุนระบบตรวจหาการบุกรุกแบบปรับตัวด้วยเทคนิคกฎความสัมพันธ์. *วารสารวิชาการพระจอมเกล้าพระนครเหนือ,* 25(2), 277-288.

[14] ไพชยนต์ คงไชย. (2557). *การพัฒนาขั้นตอนวิธีเพื่อจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด*. (วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต) มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.