การวิเคราะห์รูปแบบการบุกรุกข้อมูลบนเครือข่าย โดยใช้เทคนิคดาต้าไมน์นิง

จิโรจน์ ภาคคีรี* และ กาญจนา วิริยะพันธ์*

บทคัดย่อ

งานวิจัยนี้ใช้เทคนิคดาต้าไมน์นิง (Data mining) เพื่อวิเคราะห์รูปแบบการบุกรุกข้อมูลบนเครือข่าย ชนิดการโจมตีผ่านเครือข่ายแบบปฏิเสธการให้บริการ (Denial of Service) เช่นการโดนผู้บุกรุก(Intruders) ทำการซิงฟลัด (Syn Flood) ซึ่งการใช้เทคนิคทางด้านดาต้าไมน์นิงนั้นจะทำให้วิเคราะห์รูปแบบของแพ็กเก็ตต่างๆที่เข้ามาในระบบนี้ ว่าจะเป็นการถูกโจมตีหรือไม่ ด้วยการอาศัยการจดจำจากรูปแบบ (Pattern) ของแพ็กเก็ต (Packet) ซึ่งหากการป้องกันทำได้เร็วและได้ผลย่อมจะส่งผลอันดีต่อระบบคอมพิวเตอร์เครือข่าย ในปัจจุบันซึ่งนับวันจะทวีความรุนแรงมากขึ้นในเรื่องของความไม่ปลอดภัยในข้อมูล เช่นการโจรกรรมบัตรเครดิต การโจรกรรมข้อมูลในองค์กร เป็นต้น โดยงานวิจัยฉบับนี้ได้ใช้เทคนิคการ คัดแยกกลุ่มข้อมูล (Classification) ซึ่งใช้โมเดลที่แตกต่างกัน 3 ชนิดคือ การใช้แผนผังต้นไม้เพื่อการตัดสินใจ (Decision Tree) การใช้กฎ (Rules) และการใช้นิวรอลเน็ทเวิร์ค (Neural Network) มาเป็นโมเดลในการสอน (Train) ให้กับข้อมูล ซึ่งผลลัพธ์ที่ได้จะเห็นว่าร้อยละของความถูกต้องของข้อมูลในการใช้แผนผังต้นไม้เพื่อการตัดสินใจนั้นจะมีมากกว่านิวรอลเน็ทเวิร์ค อย่างเห็นได้ชัด และค่าเฉลี่ยความผิดพลาด (Mean Square Error) มีค่าน้อยมาก ในอนาคตเมื่อนำไปประยุกต์ใช้ในระบบไอดีเอส (IDS) ก็จะสามารถช่วยระบบความปลอดภัยในเครือข่ายคอมพิวเตอร์ได้ดียิ่งขึ้น

คำสำคัญ : การบุกรุกข้อมูล ผู้บุกรุกเครือข่าย ดาต้าไมน์นิง การคัดแยกกลุ่มข้อมูล

1. ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันเมื่อคำนึงถึงเรื่องความปลอดภัยของคอมพิวเตอร์ มักเป็นการยากในการมองภาพที่ชัดเจนว่า สิ่งใดจะเป็นมาตรวัดของการใช้งานคอมพิวเตอร์ที่มีความปลอดภัย แต่อย่างไรก็ตามสามารถเปรียบเทียบความปลอดภัยของคอมพิวเตอร์กับการรักษาความปลอดภัยสถานที่ ในการรักษาความปลอดภัยสถานที่นั้นนอกจากการจัดบริเวณที่ต้องรักษาความปลอดภัยให้มีรั้วรอบขอบชิด มีกุญแจที่ใช้ล็อกประตูหรือทางเข้าออก สิ่งหนึ่งที่จะขาดไม่ได้คือการจัดให้มีบุคคลหรืออุปกรณ์ที่คอยตรวจสอบ การละเมิดต่ออุปกรณ์หรือเครื่องกีดขวางที่จัดตั้งเพื่อความปลอดภัย ทั้งนี้เนื่องจากอาจมีผู้ไม่หวังดีพยายามบุกรุกโดยทำลายอุปกรณ์หรือเครื่องกีดขวางดังกล่าว ดังนั้นจึงต้องอาศัยระบบที่ใช้ตรวจสอบเมื่อมีการทำลายหรือล่วงล้ำต่ออุปกรณ์ หรือเครื่องกีดขวางที่ได้ติดตั้งไว้อีกชั้นหนึ่ง ตัวอย่างอุปกรณ์ที่ใช้ตรวจสอบเช่น ระบบสัญญาณเตือนขโมยที่ใช้ควบคู่กับรั้วที่แข็งแรง

ไฟร์วอลล์ (Firewall) นับเป็นเครื่องมือหลักที่ใช้ในการป้องกันระบบเครือข่าย เพราะนอกจากจะทำหน้าที่กรองแพ็กเก็ตที่ไม่ต้องการแล้ว ในปัจจุบันยังได้เพิ่มเติมความสามารถโดยยอมให้มีการสร้างส่วนต่อขยาย (Plug-In) ออกไปเพื่อทำหน้าที่ต่าง ๆ ได้ เช่น สแกนไวรัส ตรวจจับผู้บุกรุก ตรวจสอบเนื้อหา (Content Screening) และอื่น ๆ อีกมากมาย แม้ว่าไฟร์วอลล์จะทำหน้าที่ตามที่กล่าวมาแล้วได้เป็นอย่างดี แต่หากมองจากงานที่ต้องการความปลอดภัยอย่างสูง เช่น การเงินการธนาคาร หรืองานทางทหารแล้ว ไฟร์วอลล์ยังไม่สามารถตอบสนองความต้องการได้อย่างเต็มที่ ทั้งนี้เนื่องจากไฟร์วอลล์สามารถป้องกันการโจมตีได้ในระดับโพรโตคอลเท่านั้น แต่ไม่สามารถป้องกันการโจมตีตามรูปแบบได้ จึงได้มีการนำเอาโปรแกรมประเภทตรวจจับผู้บุกรุกเข้าไปผนวกกับ ไฟร์วอลล์ ซึ่งเท่ากับว่าเป็นการไปเพิ่มภาระของไฟร์วอลล์ มากขึ้นไปอีก ดังนั้นในองค์กรขนาดใหญ่ ๆ จึงมักเพิ่มเซิร์ฟเวอร์สำหรับตรวจจับผู้บุกรุกเข้าไปโดยตรงมากกว่า โดยระบบตรวจจับผู้บุกรุกจะทำหน้าที่ตรวจสอบแพ็กเก็ต หรือถ้าจะเปรียบว่าระบบคอมพิวเตอร์คือบ้านหลังหนึ่ง ไฟร์วอลล์จะทำหน้าที่เป็นเซฟที่จะเปิดให้เฉพาะผู้ที่ได้รับอนุญาต ในขณะที่ระบบตรวจจับผู้บุกรุกจะทำหน้าที่เป็นอุปกรณ์ตรวจจับการบุกรุก ซึ่งหากขโมยเข้ามาในบ้าน ไฟร์วอลล์จะไม่ทราบว่าขโมยเข้ามา

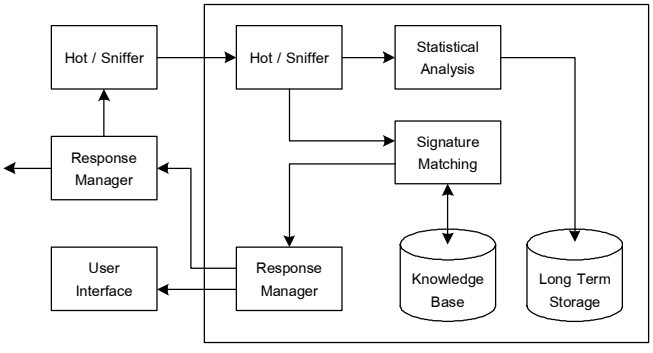
* *ภาควิชาการจัดการเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

หากเซฟนั้นสร้างมาดีพอ สิ่งที่อยู่ในเซฟก็จะปลอดภัย แต่หากเซฟสร้างมาไม่ดีหรือขโมยมีความเก่งกาจมาก ก็อาจจะเปิดเซฟออกมาได้

ระบบตรวจจับผู้บุกรุกที่ดีจะต้องตรวจสอบแพ็กเก็ตที่วิ่งไปมาในระบบเครือข่ายเพื่อหาแพ็กเก็ต ที่มีลักษณะไม่ปกติ นอกจากนั้นยังต้องวิเคราะห์ว่า ความไม่ปกตินั้นเป็นความพยายามที่จะเจาะระบบหรือไม่ เนื่องจากผู้ที่เจาะระบบคือมนุษย์ ดังนั้นการที่ซอฟต์แวร์ตัวหนึ่งจะสามารถรู้เท่าทันมนุษย์ว่ากำลังทำการเจาะระบบอยู่ จึงเป็นเรื่องที่ท้าทายอย่างยิ่ง จากปัญหาของความปลอดภัยของคอมพิวเตอร์ที่เกิดจากการบุกรุกของผู้ที่ไม่ปรารถนาดีต่อระบบเครือข่ายดังกล่าว จึงเป็นสาเหตุจูงใจให้ผู้จัดทำงานวิจัยมีแนวคิดในการใช้เทคนิคดาต้าไมน์นิงเพื่อช่วยวิเคราะห์รูปแบบการบุกรุก และเป็นแนวทางสำหรับผู้ใช้งานระบบเครือข่ายที่สนใจศึกษาในระบบรักษาความปลอดภัยในเครือข่ายคอมพิวเตอร์ (Security in Computer Networking) ต่อไป

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

ระบบตรวจจับผู้บุกรุกเครือข่าย (Intrusion Detection System) หรือไอดีเอส (IDS) คือ ระบบที่ทำหน้าที่ตรวจจับการบุกรุกและวิเคราะห์ข้อมูลที่อยู่บนเครือข่ายภายในและเครือข่ายภายนอกว่ามีพฤติกรรมที่เป็นความเสี่ยงและก่อความเสียหายต่อระบบงานภายในองค์กรหรือไม่ ซึ่งขัดกับข้อบังคับและเจตจำนงในการใช้งาน ส่งผลต่อความปลอดภัยของระบบคอมพิวเตอร์หรือเครือข่ายคอมพิวเตอร์ 3 ประการ คือ ความถูกต้อง (Integrity) ความน่าเชื่อถือได้ (Confidentiality) และสภาพพร้อมใช้งาน (Availability) โดยระบบไอดีเอสนี้จะมีระบบแจ้งเตือนให้กับผู้ดูแลระบบทราบ และหยุดพฤติกรรมดังกล่าวทันที โดยส่วนประกอบของระบบตรวจจับผู้บุกรุกเครือข่ายมีลักษณะดังภาพที่ 1



ภาพที่ 1 ส่วนประกอบของระบบ IDS

2.1 ประเภทของการบุกรุกในระบบเน็ตเวิร์ค [1]

การรบกวนขัดขวางการทำงานของเครื่องคอมพิวเตอร์หรืออุปกรณ์เครือข่าย ทำให้เครื่องคอมพิวเตอร์หรือเครือข่ายคอมพิวเตอร์ไม่สามารถให้บริการได้ ตัวอย่างการโจมตีชนิดนี้ เช่น SYN Flood, PING of Death, Land Attack, Teardrop attack, ICMP Flood หรือแม้แต่โปรแกรมประเภท virus หรือ worm ที่มีอยู่และแพร่หลายบนอินเทอร์เน็ต ในกรณีที่มีการบุกรุกโจมตีระบบด้วย SYN Flood นั้นจะมีชื่อเรียกอีกอย่างหนึ่งว่า การโจมตีแบบ Dos (Denial of Service Attack) หรือการโจมตีด้วยการทำให้ Server ปฏิเสธการให้บริการ จะส่งผลให้ Server ต้องหยุดให้บริการไปชั่วระยะหนึ่ง ซึ่งจะเป็นปัญหาอย่างยิ่งสำหรับ เว็บไซต์ (Website) ที่มีมูลค่าการทำธุรกรรมสูง ๆ ซึ่งการโจมตีเช่นนี้เกิดขึ้นได้โดยอาศัยช่องโหว่หรือความอ่อนแอภายในของโปรโตคอล TCP/IP ที่ใช้ในระบบอินเทอร์เน็ต โดย Hacker จะเข้าครอบครองแบนด์วิธของระบบเครือข่าย (Bandwidth Consumption) ที่สนใจ แม้ว่า Hacker จะปฏิบัติงานอยู่ ณ เครือข่ายอื่น ๆ ก็ตาม ซึ่ง สามารถ “Flood” Network Connection ไปที่เครื่องเหยื่อเป้าหมายได้ เพื่อที่ทำให้ Hacker ได้รับ Bandwidth ที่มากขึ้น หรืออีกกรณีหนึ่งที่ Hacker สามารถจะเพิ่มศักยภาพของการโจมตีด้วย Dos ได้ โดยการทำให้ไซต์อื่น ๆ เข้าร่วมในการ Flood Network Connection ของเครื่องเหยื่อเป้าหมาย ในกรณีนี้จะทำให้ทรัพยากรหมดไป (Resource Starvation) เพราะ Hacker จะพยายามเข้าครอบครองทรัพยากรที่อยู่บนเครื่อง Server แทนที่จะเข้าครอบครองทรัพยากรบนเครือข่าย

2.2 ดาต้าไมน์นิง (Data Mining)

ดาต้าไมน์นิง เป็นกระบวนการเพื่อกลั่นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ [2] โดยมองที่ความสัมพันธ์ของข้อมูล แนวโน้มของข้อมูลต่างๆ เพื่อให้สามารถนำข้อมูลที่กลั่นกรองได้นำไปใช้ประโยชน์ เป็นข้อมูลสนับสนุนในการตัดสินใจในเรื่องต่าง ๆ ต่อไป ซึ่งมีด้วยกัน 5 รูปแบบ คือ

2.2.1 Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูลโดยค้นหาความสัมพันธ์ หรือความเชื่อมโยงของข้อมูลทั้งสองชุด หรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน

2.2.2 Classification and Prediction การจำแนกประเภทและการทำนาย ใช้ค้นหาโมเดลที่อธิบายข้อมูลแต่ละประเภทได้ โดยการนำเสนออาจอยู่ในรูปแบบ Decision-tree, Classification Rule, Neural Network ซึ่งผู้ใช้ทำนายค่าบางอย่างที่ไม่รู้ หรือค่าที่หายไปในฐานข้อมูล

การแก้ปัญหาชนิดแบบแบ่งกลุ่มข้อมูล (Techniques for Classification Problems) คือ การที่พยายามสรุป Class Label ให้กับ Data Record ตัวอย่างเช่น โรงพยาบาล มีระบบที่จะสามารถบอกได้ว่า ผู้ป่วยเป็นโรคอะไร โดยการกรอกอาการ เช่น อุณหภูมิร่างกาย มีไข้หรือไม่ อาการไอ ระบบจะทำนายว่าผู้ป่วยมีความน่าจะเป็นที่จะเป็นโรคอะไรมากที่สุด สิ่งที่ต้องคำนึงถึง คือ วิธีการที่จะได้ กล่องตรวจสอบนี้ที่สามารถทำงานร่วมกันระหว่าง Class ที่มีอยู่แล้วกับข้อมูลใหม่ที่จะนำมาทำนาย ปัญหาประเภทนี้ เรียกว่า Classification [4] ประกอบด้วย

ส่วนที่ 1 : Predicted Class คือการที่จะ Associate Class Labels ให้กับข้อมูลใหม่ สิ่งที่จำเป็นก็คือ ต้องมี File Record หรือ Class Labels ว่าจะกำหนดอย่างไร เช่น โรงพยาบาลกำหนด Class Labels เป็นการเจ็บป่วย ดังนี้ 1) ไข้หวัด 2) ท้องเสีย 3) ปวดฟัน มี 3 อาการ ไว้ในระบบ เมื่อมีคนไข้รายใหม่เข้ามา กรอกอาการป่วยเข้าไปในระบบ ระบบจะทำนายว่าน่าจะเป็นโรคใดมากที่สุด จาก Class Label ที่ได้กำหนดไว้

ส่วนที่ 2 : Training Data คือ การเก็บข้อมูลจากประสบการณ์ของโรงพยาบาล จากตัวอย่างคนไข้จำนวนมากๆ โดยพิจารณาอาการ และโรคที่เป็น เช่น การมีไข้ อุณหภูมิเท่าใด อาการไอ อาการปวดท้อง ปวดฟัน เป็นต้น

ส่วนที่ 3 : Algorithm ในการสร้าง Decision tree สรุปสิ่งสำคัญในการแก้ปัญหานี้ คือ

1) กำหนดคลาสเลเบล (Class Label) เป็นการกำหนดแอทริบิวต์ (Attribute) ให้กับค่าที่ต้องการจะวัด หรือ พยากรณ์

2) การหาชุดข้อมูลการเรียนรู้ (Training Data) ซึ่งจะได้แก่ ชุดข้อมูล (Data sets) ซึ่งจะมีชุดข้อมูลที่สำคัญและเป็นข้อมูลทั้งหมดที่จะนำมาทำ ดาต้าไมน์นิง

3) อัลกอริทึม (Algorithm) หรือกระบวนการคิดในการเรียนรู้และทดสอบข้อมูล ที่ใช้ในการสร้างคือรูปแผนผังต้นไม้เพื่อการตัดสินใจ (Decision Tree) โดยในกระบวนการของคัดแยกกลุ่มข้อมูล (Classification) มี 2 ขั้นตอนที่สำคัญ คือ

การสร้างโมเดล (Model Building for Training set) การสร้างโมเดลนี้เป็นการสอนชุดข้อมูลให้มีการเรียนรู้ตามอัลกอริทึมที่ได้กำหนดเอาไว้

การใช้โมเดล (Model Testing for Testing set) การใช้โมเดลเป็นการใช้ทดสอบกับข้อมูลที่จะต้องทดสอบ (Test set) ซึ่งจะเป็นการทำนายค่าผลลัพธ์ (Output) ซึ่งผลลัพธ์ที่ได้จะเป็นการถูกทำนายค่าจากอัลกอริทึมที่ได้เรียนรู้จากข้างต้นแล้ว

การใช้แผนผังทางเลือกเพื่อการตัดสินใจ (Decision Tree) นั้นมีอยู่หลายประเภทด้วยกัน ซึ่งที่ใช้เปรียบเทียบมีอยู่ 3 ชนิด ได้แก่

1) นิวรอลเน็ทเวิร์ค (Neural Network) [4] มีพื้นฐานมาจากแบบจำลองการทำงานของสมองมนุษย์ และใช้ได้ดีกับปัญหา Classification, Regression และ Clustering เทคนิคนี้มักถูกเรียกว่า “black box” เนื่องจากการทำงานมีความซับซ้อนมากกว่าเทคนิคอื่นๆ ค่อนข้างมาก โดยโครงสร้างการทำงานของโครงข่ายประสาทเทียม (Neural Network) ประกอบด้วยชั้นของข้อมูลเบื้องต้น 3 ชั้น คือชั้นของข้อมูลป้อนเข้า (Input Layer) ชั้นซ่อน (hidden layer) และชั้นผลลัพธ์ (Output Layer) สามารถแบ่งประเภทของโครงข่ายประสาทเทียมตามความซับซ้อน หรือจำนวนชั้นของข้อมูลในชั้นซ่อน เช่น โครงข่ายประสาทเทียมอย่างง่าย(Single Layer Neural Network) [5] เป็นโครงข่ายที่มีชั้นซ่อนเพียงหนึ่งยูนิต เวกเตอร์ตัวแปรด้านนอกและไบแอส (bias) หรือเพอร์เซพตรอนชั้นเดียว (single-layer perceptron) เพอร์เซพตรอน (perceptron) เป็นข่ายงานประสาทเทียมประเภทหนึ่ง ที่ได้รับการพัฒนาขึ้นใน ค.ศ. 1957 โดยแฟรงค์ โรเซนบลัทท์ ที่ Cornell Aeronautical Laboratory [6] เพอร์เซพตรอนประกอบด้วยเซลล์ประสาทเทียมหรือนิวรอนอย่างน้อยหนึ่งชั้น อินพุตจะถูกส่งตรงไปยังเอาท์พุตโดยผ่านชุดค่าน้ำหนัก

2) กฎของริปเปอร์ [7] ถูกสร้างโดย Cohen ในปี 1995 โดยริปเปอร์ ถูกประกอบด้วย 2 เฟส คือ เฟสแรกทำการระบุกฎเริ่มต้น และเฟสที่สองจะระบุค่า post-process rule optimization โดยข้อมูลที่ถูกการเรียนรู้ (Training data) จะถูกแบ่งไปเป็น growing set และ pruning set ดังนั้น อัลกอริทึมนี้จะสร้างกฎความสัมพันธ์ใน greedy fashion ในขณะที่สร้างกฎริปเปอร์นั้น จะหาค่าที่ดีที่สุดสำหรับgrowing set ใน rule space ซึ่งจะอธิบายได้จาก BNF หลังจากได้ growing set ก็จะทำการ pruning ข้อมูลทันที เมื่อเสร็จแล้วจะได้กลุ่มตัวอย่างที่เหมือนกันออกมาที่ครอบคลุมกฎของ training set จากนั้นก็จะถูกลบทิ้ง ซึ่ง training data ที่เหลือ จะถูกแบ่งใหม่อีกครั้งหนึ่งหลังจากถูกเรียนรู้ตามกฎแล้ว เพื่อช่วยแก้ปัญหาที่เกิดมาจากการแบ่งแยกกลุ่มที่ผิดพลาด [8] ซึ่งกระบวนการนี้จะกระทำจนกระทั่งผลเป็นที่น่าพอใจ

3) อัลกอริทึม J48 Decision Tree หรือ อัลกอริทึมของ C4.5 ซึ่งเป็นอัลกอริทึมอันหนึ่งที่ใช้สร้าง decision tree

ที่พัฒนาโดย Ross Quinlan [9] โดย C4.5 นี้ คือส่วนขยายที่เพิ่มเติมมาจาก อัลกอริทึม ID3 Decision tree โดยโครงสร้างนี้สามารถถูกใช้สำหรับการคัดแยกกลุ่มข้อมูล (classification) และ ด้วยเหตุผลอันนี้ C4.5 ถูกเรียกใช้บ่อยๆ สำหรับตัว statistical classifier ในส่วนของอัลกอริทึม C4.5 สร้าง decision trees มาจากกลุ่มของ training data เหมือนกับ ID3 ที่ใช้หลักการของ Information Entropy [10] โดย C4.5 ใช้ความถูกต้องของแต่ละแอทริบิวต์ของข้อมูล เพื่อใช้เป็นการตัดสินใจแบ่งกลุ่มข้อมูลไปยังกลุ่มย่อย ๆ ซึ่ง C4.5 จะพิจารณาตรวจสอบ normalized information gain (ความแตกต่างในentropy) ผลลัพธ์จากการเลือกแอทริบิวต์สำหรับการแบ่งกลุ่มข้อมูล โดยคุณลักษณะด้วยค่า normalized information gain ที่สูงสุดนั้นคือหนึ่งในการสร้างการตัดสินใจ ผู้จัดทำไม่สามารถบอกได้ว่าวิธีการหรือเทคนิคใดดีที่สุด ดังนั้นการทำวิจัยจะต้องทำการทดลองสร้างโมเดลหลายๆ ครั้งแล้วนำมาเปรียบเทียบกันว่าโมเดลใดใช้ทำนายค่า (Prediction)ได้ดีที่สุด

2.2.3 การจัดกลุ่มข้อมูล (Cluster analysis) ต้องมีความคล้ายกันมากที่สุด

2.2.4 การหาค่าผิดปกติที่เกิดขึ้น (Outlier analysis) หรือข้อมูลบางอย่างไม่น่าจะเป็นจริงได้

2.2.5 การวิเคราะห์แนวโน้ม (Trend and evolution analysis)

3. วิธีการดำเนินการวิจัย

3.1 ศึกษาปัญหาและความต้องการของระบบ เพื่อนำมาเป็นข้อมูลในการวิเคราะห์และออกแบบพัฒนาในขั้นตอนต่อไป ผู้พัฒนาจึงได้ทำการวิเคราะห์รูปแบบ การบุกรุกข้อมูลบนเครือข่ายด้วยการใช้เทคนิคทางดาต้าไมน์นิง โดยใช้อัลกอริทึม (Algorithm) ของโครงข่ายประสาทเทียม (Neural Network) แผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree) และกฎความสัมพันธ์ของริปเปอร์ (Ripper Rule) มาทำการเทียบเคียงหาประสิทธิภาพของค่าความผิดพลาดอันที่จะเกิดขึ้นได้ ซึ่งค่าความผิดพลาดที่จะยอมรับได้นั้น คือ ค่า Mean Square Error (MSE) โดยตามปกติ ค่า MSE นั้นจะต้องมีค่าน้อยมากๆ ทั้งนี้เพื่อความแม่นยำในการทำนายค่า โดยวิเคราะห์ลักษณะของชุดข้อมูล (Datasets) ซึ่งชุดข้อมูลที่ใช้เป็นข้อมูลที่ได้จากฐานข้อมูลความรู้ Knowledge Discovery in Database (KDD) Cup data ซึ่งเป็นชุดข้อมูลในปี 1999 ซึ่งชุดข้อมูลชุดนี้ได้มาจากความร่วมมือของ โครงการวิจัยและพัฒนาเพื่อการทหารของประเทศสหรัฐอเมริกา (Defense Advanced Research Projects Agency : DARPA) ซึ่งร่วมมือกับทางมหาวิทยาลัยเมตซาซูเซตส์ สหรัฐอเมริกาชุดข้อมูลนี้ถูกสร้างจากการจำลองการโจมตีของผู้บุกรุกจาก U.S. Air Force local area network ตั้งขึ้นที่ Lincoln Labs โดยข้อมูลนั้นมีระยะเวลาในการจัดทำนานถึง 9 สัปดาห์ จากการเก็บข้อมูลจากแพ็กเก็ต TCP ผ่านโปรแกรม Tcp dump ประกอบด้วยข้อมูลขนาดใหญ่มาก (Gigabytes) ซึ่งมี แอทริบิวต์มากถึง 41 ตัวที่ได้จาก แพ็กเก็ต TCP (Raw TCP packet) รวมถึงชนิดของโปรโตคอลซึ่งมีค่า"TCP","ICMP"และ"UDP" ซึ่งเป็นแอทริบิวต์ที่มีความต่อเนื่องเป็นในลักษณะข้อความ (Nominal) ซึ่งจะมีสถานะ (Label) กำกับไว้เสมอในแต่ละบรรทัด (Record) ว่าข้อมูลชุดนี้มีสถานะปกติ (Normal) หรือว่าเป็นชุดข้อมูลที่ถูกโจมตี (malicious) ทางผู้จัดทำได้คัดเลือก ชุดข้อมูลที่มีมากถึง 5 ล้านบรรทัด (Record) ออกมาประมาณ 10% เท่านั้นเพื่อสะดวกในการจัดทำชุดข้อมูลเรียนรู้ (Training set) และชุดข้อมูลทดสอบ(Testing set)

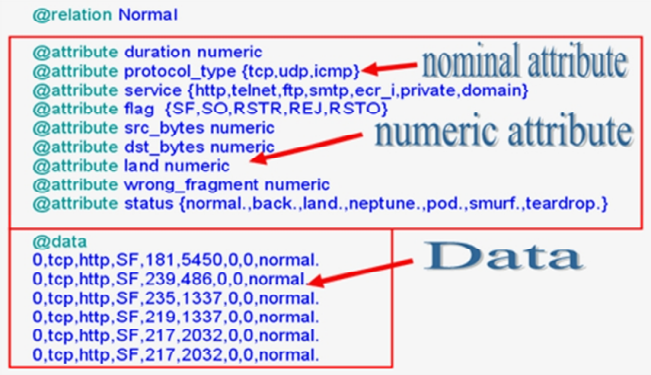
ชุดข้อมูลที่ได้จากเว็บไซต์ (KDD Website) นี้ อยู่ในรูปของเครื่องกลเรียนรู้ (Machine Learning Pattern) โดยแอทริบิวต์ตัวขวามือสุดคือ คลาสแอทริบิวต์ (Class attribute) ที่เป็นตัวบอกสถานะว่าถูกโจมตี (attack) หรือไม่ ซึ่งหากไม่ได้ถูกโจมตีจะมีค่าเป็นปกติ (normal) โดยข้อมูลดังกล่าวข้างต้นมีค่าของฟีเจอร์ (feature) ในแต่ละแอทริบิวต์ (attribute) เนื่องจากการโจมตีแบบปฏิเสธการให้บริการ (Denial of Service) มีชื่อเรียกการโจมตีแบบต่าง ๆ กัน ถึง 6 แบบ คือ แบคดอร์ (back door) แลนด์ (Land Attack) เนปจูน (Neptune) ปิงค์ออฟเดท (Ping of Death) พอด (pod) สเมิร์ฟ (smurf) และเทียร์ดรอป (teardrop) นั้นจะมีแอทริบิวต์ที่จำเป็นต่อการวิเคราะห์ เพียงแค่ 8 หลักเท่านั้น ดังนั้นจะต้องทำการสกัดข้อมูล (Extract) ออกมา 9 หลัก โดยรวมหลักสุดท้ายหลักที่ 42 ซึ่งเป็นสถานะ (status) ดังตารางที่ 1

3.2 การเตรียมข้อมูลสำหรับทำดาต้าไมน์นิง (Data preparation) ขั้นตอนการเตรียมข้อมูลก่อนที่จะส่งไปเข้าสู่กระบวนการหาอัลกอริทึมและวิเคราะห์ ซึ่งขั้นตอนนี้สำคัญมาก เนื่องมาจากเป็นขั้นตอนที่ป้อนอินพุตเข้าไปในระบบหรือโปรแกรม โดยโปรแกรมที่ใช้ในการทำงานวิจัยในครั้งนี้ ผู้จัดทำได้เลือกใช้โปรแกรม Weka เวอร์ชั่น 3.4 ซึ่งเป็นซอฟต์แวร์ด้านการทำเหมืองข้อมูลที่ได้รับการยอมรับและแพร่หลายในต่างประเทศ โดยโปรแกรมจะต้องถูกจัดเตรียมแฟ้มงาน (File) ให้อยู่ใน

ตารางที่ 1 รายละเอียดของแอทริบิวต์

Attribute หลักที่	ชื่อ (Feature Name)	รายละเอียด	ชนิด (Type)
1	Duration	จำนวนวินาทีในการเชื่อมต่อ	ตัวเลข
2	Protocol_type	ชนิดของโปรโตคอล เช่น TCP, UDP	ตัวอักษร
3	Service	ชนิดของการให้บริการ เช่น http, ftp	ตัวอักษร
4	Flag	สถานะ normal หรือ error ของ connection	ตัวอักษร
5	Src_bytes	จำนวน data bytes ของต้นทาง	ตัวเลข
6	Dst_bytes	จำนวน data bytes ของปลายทาง	ตัวเลข
7	Land	= 1 ถ้ามาจาก host/port เดียวกัน , = 0 อื่นๆ	ตัวเลข
8	Wrong_Fragment	จำนวนของ “wrong” fragment	ตัวเลข
42	Status (CLASS)	ชื่อของการโจมตี / normal = ไม่ถูกโจมตี	ตัวอักษร

รูปแบบ ARFF (ARFF Format) ดังภาพที่ 2 ซึ่งจะมีลักษณะแฟ้มงานคล้ายกับ CSV Format ซึ่งสามารถทำการแปลงไฟล์ (convert) ในโปรแกรมประเภทสมุดงาน (Spread Sheet) ได้



ภาพที่ 2 รายละเอียดของไฟล์ ARFF

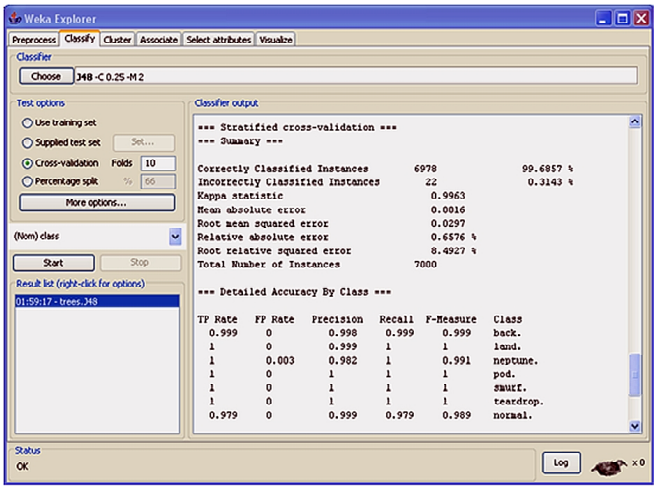
3.3 การสร้างโมเดลระบบและการสอนข้อมูล (Building Models and Train Datasets) การเรียนรู้แบบมีการควบคุม (Supervised Learning) เป็นการเรียนรู้ซึ่งต้องมีชุดข้อมูลสำหรับการเรียนรู้ (Training Data) ซึ่งมีทั้งชุดข้อมูลที่เป็นอินพุต และเอาต์พุต

3.3.1 ขั้นตอนการสร้างโมเดลแผนภาพต้นไม้เพื่อการตัดสินใจ (Building a decision tree) โดยขั้นตอนนี้อยู่ในบริเวณของ Explorer ซึ่งจะทำการเลือกชนิดของการทำไมน์นิงในรูปแบบของ decision tree ซึ่งอันดับแรกจะใช้คลิกเลือกที่แท็บ Classify หลังจากนั้น คลิกที่>>Choose >>Trees >> J48

3.3.2 ขั้นตอนการสร้างโมเดลกฎริปเปอร์แบ่งแยกข้อมูล (Building a Ripper Rule) โดยขั้นตอนนี้อยู่ในบริเวณของ Explorer ซึ่งจะทำการเลือกชนิดของการทำไมน์นิงในรูปแบบของ decision tree ซึ่งอันดับแรกจะใช้คลิกเลือกที่แท็บ Classify หลังจากนั้น คลิกที่ >>Choose >>Rules>>JRip ซึ่งในขั้นตอนนี้ก็จะคล้ายกับในข้อ 3.3.1 ที่ผ่านมา

3.3.3 ขั้นตอนการสร้างโมเดลโครงข่ายประสาทเทียม (Neural Network) คลิกที่ >> Choose >> Function >> MultilayerPerceptron ซึ่งในขั้นตอนนี้ก็จะคล้ายกับในข้อ 3.3.1 ที่ผ่านมา

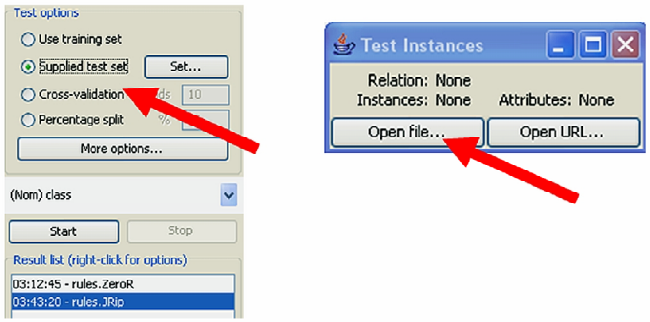
3.4 การทดสอบข้อมูลและปรับปรุงแก้ไข (Test Datasets and improvement) เป็นขั้นตอนที่เป็นหัวใจของการทำดาต้าไมน์นิง เพราะเป็นขั้นตอนสุดท้ายของการประมวลผลเพื่อนำผลลัพธ์ ที่อยู่ในรูปของข้อความ (Text Output) และ แผนภาพ (Graphics Output) เช่น แผนภาพต้นไม้ (Tree) เป็นต้น นำไปแสดงผล ซึ่งผลลัพธ์ที่เป็นข้อความ (Text Output) นั้นจะอยู่บริเวณขอบขวาของจอภาพ ซึ่งสามารถทำการคัดลอกได้ จะต้องใช้การไฮไลต์และกดปุ่ม Ctrl+C เท่านั้น ส่วนการจะดูแผนภาพต้นไม้ นั้นก็สามารถคลิกเมาส์ขวาที่บริเวณ Result list แล้วเลือก Visualize Tree ได้



ภาพที่ 3 หน้าจอสดงผลลัพธ์จากการทำไมน์นิง (Classifier Output)

ในส่วนของการนำไฟล์ทดสอบ (Testing dataset) เข้ามาในระบบที่ได้ผ่านการเรียนรู้ (Training sets) เป็นที่เรียบร้อยแล้วนั้นสามารถทำได้ โดยนำเมาส์คลิกที่ Test Options >> Supplied test set จากนั้น กดปุ่ม Set เพื่อนำทางไปสู่ (Browse) การเลือกชื่อไฟล์เป็นอันดับถัดไป ต่อจากนั้นคลิกส์ที่ Open file แล้วเลือกไฟล์ ARFF ที่ต้องการจะทำการทดสอบ (Testing) แล้วจึงกดปุ่ม start เพื่อทดสอบข้อมูล (Test datasets) โดยขั้นตอนการเลือก

ไฟล์ทดสอบ แสดงดังภาพที่ 4



ภาพที่ 4 การเลือกไฟล์ (Test Sets) ที่จะนำมาทดสอบกับโมเดลที่ได้ทำการสอนไว้แล้ว

4. ผลการดำเนินงาน

การวิเคราะห์รูปแบบการบุกรุกข้อมูลบนเครือข่ายโดยใช้เทคนิคดาต้าไมน์นิง ซึ่งได้ทำการนำข้อมูลมาทำการเรียนรู้และทดสอบแล้วนั้นก็จะมีค่าทางสถิติออกมาหลายค่า เช่นค่า Root Mean Square Error ,TP,FP รวมถึง ค่า Confusion Matrix และนอกจากนั้น ยังมีการแยกข้อมูลออกเป็นแผนภาพต้นไม้เพื่อการตัดสินใจอีกด้วย ซึ่งจะแบ่งผลลัพธ์เป็นดังนี้

การให้ข้อมูลมีการเรียนรู้ (Train) ถูกแบ่งออกเป็น 3 ประเภท โดยจะแบ่งเป็นอัลกอริทึม J.48 Decision Tree, JRip Rules และ Neural Network เป็นตัวคัดแยก(Classifier) หลังจากการเรียนรู้เสร็จแล้วก็เปรียบเสมือนว่า ได้ทำการจัดโมเดลข้อมูลให้กับข้อมูลไม่เหมือนกัน 3 ประเภท ซึ่งจะมีผลต่อการ Test และ เปอร์เซ็นต์ของความถูกต้องในการจัด class อย่างแน่นอน โดยการเรียนรู้ครั้งนี้ได้คัดเลือกจำนวน 7,000 บรรทัด (Records) ที่มีค่า class กระจายกันไป โดยให้มีความหลากหลายแบบสุ่ม (Cross Validation) จำนวนครั้งที่สลับ (Fold) = 10 ซึ่งผลลัพธ์ที่ได้สามารถสรุปได้ดังตารางที่ 2

ตารางที่ 2 การเปรียบเทียบค่า MSE และ ค่าความถูกต้องของ Classified Instance

สิ่งที่เปรียบเทียบ	J48 Decision Tree	JRip Rule	Neural Network
Correctly Classified Instances	99.998%	99.15%	14.1714%
Incorrectly Classified Instances	0.002%	0.85%	85.8286%
MSE	0.0297	0.0512	0.35

จากตารางที่ 2 ผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิงโดยใช้โมเดลในการเรียนรู้ด้วยอัลกอริทึมโครงข่ายประสาทเทียมมีค่าความความผิดพลาดสูงมากกว่าโมเดลการเรียนรู้ชนิดอื่นอย่างเห็นได้ชัด กล่าวคือ ค่าความผิดพลาด (MSE) ที่ได้จากโมเดลนิวรอลเน็ทเวิร์ค (Neural Network) มีค่าสูงมากที่สุด คือ 0.35 ในขณะที่ JRip Rule และ J48 Decision Tree มีค่าเพียง 0.0512 และ 0.0297 ตามลำดับ

จากผลการทดลองนี้สรุปได้ว่า โมเดลในการเรียนรู้และการทดสอบที่ควรใช้คือ J48 Decision Tree เพราะให้ค่าพยากรณ์ที่มีความถูกต้องสูงถึง 99.998 % และมีค่าความผิดพลาดต่ำที่สุดเมื่อเปรียบเทียบกับโมเดลชนิดอื่นๆ รองลงมาคือโมเดล JRip Rule ซึ่งให้ค่าความถูกต้อง 99.15% และมีความผิดพลาด 0.85% ในขณะที่โมเดลของอัลกอริทึมนิวรอลเน็ทเวิร์ค มีค่าความถูกต้องเพียง14.1714% และมีความผิดพลาดสูงถึง 85.8286 % โมเดลอัลกอริทึมนิวรอลเน็ทเวิร์คจึงไม่เหมาะสมกับการนำมาเรียนรู้ (Training) และทดสอบ (Testing) ข้อมูลชุดนี้

5. สรุป อภิปรายผล และข้อเสนอแนะ

การจัดทำวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์รูปแบบการบุกรุกข้อมูลบนเครือข่าย ชนิดการโจมตีผ่านเครือข่ายแบบปฏิเสธการให้บริการ (Denial of Service) ซึ่งการทำดาต้าไมน์นิงนั้นจะทำให้รู้ว่าแพ็กเก็ตต่างๆ ที่เข้ามาในระบบนี้จะเป็นการถูกโจมตีหรือไม่ ด้วยการสังเกตจากรูปแบบ (Pattern) ของแพ็กเก็ต ที่ผ่านเข้ามาในระบบโดยหาความสัมพันธ์ของข้อมูลออกมาเป็นกฎต่างๆ ด้วยเทคนิคทางดาต้าไมน์นิง ก็จะสามารถค้นพบร่องรอยหรือกระบวนการที่จะค้นหาการบุกรุกโจมตีได้ ซึ่งเทคนิคทางดาต้าไมน์นิงชนิด การเรียนรู้แบบมีการควบคุม (Supervised Learning) นี้สามารถนำไปประยุกต์ใช้งานในด้านอื่นๆได้อีกด้วย

จากการทดลองหาโมเดลในการทดสอบชุดข้อมูลที่เหมาะสมในโปรแกรม Weka พบว่าการวิเคราะห์โดยใช้อัลกอริทึมโครงข่ายประสาทเทียมจะมีค่าความผิดพลาดมากกว่าโมเดลชนิดอื่น โดยเปรียบเทียบค่าจากโมเดลการคัดแยก จึงกล่าวได้ว่าอัลกอริทึมนิวรอลเน็ทเวิร์คไม่เหมาะสมที่จะมาทำการวิเคราะห์ในข้อมูลชุดนี้ สำหรับโมเดลอัลกอริทึมที่เหมาะสมมากกว่า คือ โมเดลอัลกอริทึมแผนภาพต้นไม้เพื่อการตัดสินใจ (J48 Decision Tree) เพราะค่าความผิดพลาดที่เกิดจากการทดสอบมีค่าน้อยกว่ามาก และเปอร์เซ็นต์ความถูกต้องข้อมูลมีค่ามากกว่าโมเดลอัลกอริทึมชนิดอื่นๆ อย่างมาก

การทำนายค่าแอทริบิวต์ของข้อมูลนี้ จะมีความสามารถ

ปรับปรุงการเรียนรู้ได้ด้วยตัวเอง และสามารถปรับส่วนต่างๆ ให้เข้ากับลักษณะข้อมูลได้ เช่น สามารถกำหนดชนิดของโครงข่าย ฟังก์ชันการสอน ฟังก์ชันการเรียนรู้ และอื่นๆ อีกมากมาย ซึ่งในโมเดลของ J48 Decision Tree นั้นหากทำให้มีการทดสอบข้อมูล โดยจำนวนรอบ (Fold) สูงก็จะทำให้ค่าความผิดพลาดลดน้อยลงจนถึงระดับหนึ่งจะคงที่

งานวิจัยนี้ใช้การพยากรณ์สถานะ (Class) ว่าเป็นการถูกบุกรุกหรือไม่ ซึ่งข้อมูลที่ได้นั้น บางแอทริบิวต์ (attribute) ล้าสมัยแล้ว เนื่องจากปัจจุบันได้มี ผู้บุกรุกระบบ(Intruders) คิดค้นวิธีการโจมตีแบบใหม่ๆ อยู่ตลอดเวลา จึงทำให้ ระบบป้องกันที่ได้จากการคำนวณนั้น ไม่ทันสมัย จึงเสนอแนะให้หน่วยงานหรือองค์กรต่างๆ ปรับปรุงระบบป้องกันการบุกรุกผ่านเครือข่ายให้มากขึ้น

6. เอกสารอ้างอิง

[1] จอร์จ เคิร์ส,โจเอล สแคมเบรย์. “HACKING EXPOSED ปิดทางแฮกเกอร์.” แปลโดย เอกสิทธิ์ วิริยจารี กรุงเทพมหานคร : สำนักพิมพ์ซีเอ็ดยูเคชั่นจำกัด (มหาชน), 2544.

[2] พนิดา ยืนยงสวัสดิ์. “การพยากรณ์ปริมาณการใช้ยาโดยใช้โครงข่ายประสาทเทียม.” สารนิพนธ์ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2547.

[3] Wenke Lee,Salvatore J. Stolfo,Philip K. Chan,Wei Fan “Real Time Data Mining-based Intrusion Detection” Columbia University,NY 10027 ,2001.

[4] Zheng Zhang,Jun Li,Jay Jorgenson “Neuron Networks in Statistical Anomaly Intrusion Detection”, Network Security Solutions,NJ 07059 , USA ,2002.

[5] สุทธิ ขำศรี. “ระบบการพยากรณ์ยอดขายสินค้าโดยใช้โครงข่ายประสาทเทียม” ปัญหาพิเศษ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2548.

[6] Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386-408.

[7] W. Lee, K. W. Mok, and S. J. Stolfo. Mining sequential patterns: Techniques, visualization, and applications. submitted for publication, August 1998.

[8] W. W. Cohen. Fast effective rule induction. In Machine Learning: the 12th International Conference, Lake Taho, CA, 1995. Morgan Kaufmann.

[9] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[10] J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996.

[11] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunninghan, and M. Zissman. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition, January 2000.