แบบรายงานวิจัยฉบับสมบูรณ์

"การคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย"

(Optimal Feature Selection for Intrusion Detection System)

โดย

ผศ. ดร. นวลสวาท หิรัญสกลวงศ์

โครงการวิจัยประเภทส่งเสริมนักวิจัย

ที่ได้รับการสนับสนุนด้วยงบประมาณเงินรายได้ ประจำปี 2554

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

กิตติกรรมประกาศ

(Acknowledgement)

รายงานการวิจัยเรื่อง การคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย งานวิจัยนี้ได้รับทุนวิจัยประเภทส่งเสริมนักวิจัยคณะวิทยาศาสตร์ ด้วยงบประมาณเงินรายได้ ประจำปี 2554 ผู้จัดทำขอขอบคุณ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่สนับสนุนทุนวิจัยในครั้งนี้

ผศ. ดร. นวลสวาท หิรัญสกลวงศ์

บทคัดย่อ

ปัจจุบันมีการใช้ internet เป็นจำนวนมาก และจำนวนผู้ใช้มีขนาดเพิ่มขึ้นในลักษณะก้าวกระโดด (exponential) การรักษาความปลอดภัยของข้อมูลถือว่าเป็นปัญหาที่สำคัญยิ่ง ระบบการตรวจจับการบุกรุกในเครือข่ายเป็นระบบที่คอยตรวจพฤติกรรมการใช้ที่ผิดปกติหรือผิดไปจากกฎเกณฑ์ที่ตั้งไว้จึงเป็นระบบที่สำคัญพื้นฐานสำหรับงานรักษาความปลอดภัยของข้อมูล ด้วยปริมาณการใช้ที่เพิ่มขึ้น ปัญหาคือต้องพยายามหาระบบการตรวจจับการบุกรุกในเครือข่ายที่มีการประมวลผลที่รวดเร็ว และมีความผิดพลาดในการแจ้งเตือนน้อยที่สุด ระบบการตรวจจับการบุกรุกในเครือข่ายส่วนใหญ่จะใช้คุณลักษณะทั้งหมดจำนวน 41 ตัวแปรเพื่อการหาวิธีใหม่ๆ สำหรับการจัดประเภทของการบุกรุกในเครือข่ายโดยมีการปรับค่าน้ำหนักของแต่ละคุณลักษณะ แต่มีส่วนน้อยที่เริ่มสนใจคัดเลือกเฉพาะคุณลักษณะที่เด่นเท่านั้นเพื่อการลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพการตรวจจับการบุกรุกในเครือข่าย ซึ่งขั้นตอนวิธีในการคัดเลือกคุณลักษณะเด่นนี้ยังซับซ้อน และประมวลผลช้าอยู่ ผู้วิจัยนำเสนอ factor analysis ในการจัดกลุ่มคุณลักษณะทั้งหมดเหล่านี้ แล้วเลือกเฉพาะลักษณะเด่นในแต่ละกลุ่มมาเป็นตัวแทนมาคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่าย ผลการทดลองกับฐานข้อมูลจำนวน 494,020 รายการ จาก KDD Cup 1999 ปรากฏว่าวิธีที่นำเสนอ เลือกคุณลักษณะเด่นเพียง 13 ตัว จาก 41 ตัว ประมาณ 31.71 % ทำให้ประหยัดเวลาในการประมวลผลโดยยังคงความถูกต้องเทียบเท่าการใช้คุณลักษณะทั้งหมด 41 ตัว

Abstract

The internet and local area networks are growing larger in recent years. As a great variety of people all over the world connecting to the internet, they are unconsciously encountering the number of security threats such as viruses, worms and attacks from hackers. Therefore, intrusion detection is becoming a more and more important technology which follows up network traffic and identifies network intrusion such as anomalous network behaviors, unauthorized network access, and malicious attacks to computer systems. Mostly, researches use 41 features to create new methods for classifying network intrusion. Rarely researches focus on feature selection. This paper uses the knowledge of factor analysis to group features, and then picks the main feature in each group representing for main features. The way reduces only 13 main features from 41 features (31.71 %). It is obvious that the reducing features can reduce processing time. With 494,020 records dataset from KDD Cup 1999, the experimental results show that the novel feature selection gain the accuracy greater than or equal to the accuracy from the whole features.

สารบัญ

สารบัญตาราง

บทที่ 1

บทนำ (Introduction)

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันมีการใช้ internet เป็นจำนวนมาก และจำนวนผู้ใช้มีขนาดเพิ่มขึ้นในลักษณะก้าวกระโดด (exponential) การรักษาความปลอดภัยของข้อมูลถือว่าเป็นปัญหาที่สำคัญยิ่ง ระบบการตรวจจับการบุกรุกในเครือข่ายเป็นระบบที่คอยตรวจพฤติกรรมการใช้ที่ผิดปกติหรือผิดไปจากกฎเกณฑ์ที่ตั้งไว้จะสันนิฐานว่าระบบถูกบุกรุก จึงเป็นระบบที่สำคัญพื้นฐานสำหรับงานรักษาความปลอดภัยของข้อมูล เนื่องด้วยปริมาณการใช้งานบนระบบเครือข่ายได้มีปริมาณที่เพิ่มขึ้นอย่างรวดเร็ว ปัญหาคือต้องพยายามหาระบบการตรวจจับการบุกรุกในเครือข่ายที่มีการประมวลผลที่รวดเร็ว และมีความผิดพลาดในการแจ้งเตือนน้อยที่สุด งานวิจัยระบบการตรวจจับการบุกรุกในเครือข่ายส่วนใหญ่จะใช้คุณลักษณะทั้งหมดจำนวน 41 ตัวแปรในการหาวิธีใหม่ๆ ในการจัดประเภทของการบุกรุกในเครือข่ายโดยมีการปรับค่าน้ำหนักของแต่ละคุณลักษณะ ซึ่งงานวิจัยในระยะหลังเริ่มให้ความสนใจกระบวนการแปลงค่าคุณลักษณะก่อนดำเนินการจัดประเภทของการบุกรุกในเครือข่าย แสดงว่าคุณลักษณะบางตัวอาจเป็นตัวถ่วงต่อระยะเวลาการประมวลผลและลดประสิทธิภาพการตรวจจับการบุกรุกในเครือข่ายลง แต่มีงานวิจัยส่วนน้อยที่เริ่มสนใจคัดเลือกเฉพาะคุณลักษณะที่เด่นเท่านั้นเพื่อการลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพการตรวจจับการบุกรุกในเครือข่าย ซึ่งขั้นตอนวิธีในการคัดเลือกคุณลักษณะเด่นนี้ยังซับซ้อน และประมวลผลช้าอยู่ เปรียบเทียบได้กับการวิเคราะห์ผู้เป็นโรคเบาหวานจากการวิเคราะห์ผลเลือด ขณะที่เลือดประกอบด้วยสารเคมีหลายชนิดมาก จะเลือกพิจารณาเฉพาะสารเคมีตัวที่บ่งบอกภาวะน้ำตาลเท่านั้น การวิจัยการตลาดจะทำการค้นหาตัวแปรที่เป็นปัจจัยที่สำคัญก่อนจะสร้างตัวแบบแก้ปัญหาสิ่งที่สนใจ ดังนั้นการคัดเลือกคุณลักษณะที่เหมาะสม (อาจจะมีจำนวนน้อยกว่า 41 ตัวแปร) จึงเป็นสิ่งที่สำคัญก่อนการหาวิธีจัดประเภทชนิดของการบุกรุกในระบบการตรวจจับการบุกรุกในเครือข่าย ผู้วิจัยต้องการวิจัยค้นหาวิธีที่ดีที่สุดในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่าย

วิเคราะห์องค์ประกอบ (factor analysis) ใช้ในการลดตัวแปรที่ผู้วิจัยสนใจศึกษาโดยอาจกระทำการรวมตัวแปรย่อยๆ ให้เป็นตัวแปรใหญ่ขึ้นมาใหม่ ซึ่งตัวแปรย่อยที่ถูกรวมเป็นตัวแปร

ใหม่ก็จะมีการให้น้ำหนักแก่ตัวแปรย่อยเหล่านั้น หรือการเลือกแต่เฉพาะตัวแปรเด่นที่มีความสำคัญมากสุดในแต่ละกลุ่ม โดยตัดตัวแปรย่อยๆ ในแต่ละกลุ่มออก ซึ่งวิธีการดังกล่าวก็น่าจะสามารถนำมาใช้เป็นขั้นตอนวิธีในการคัดเลือกเฉพาะคุณสมบัติที่เด่น ผู้วิจัยจึงอยากศึกษาว่าวิธีการวิเคราะห์องค์ประกอบเป็นวิธีที่ดีที่สุดในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่ายหรือไม่

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้ต้องการหาวิธีที่ดีที่สุดในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ต้องการขั้นตอนวิธีที่ง่ายไม่ซับซ้อนที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่าย

1.3 สมมติฐานของการศึกษา

วิเคราะห์องค์ประกอบสามารถใช้ในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ขั้นตอนวิธีต้องง่ายไม่ซับซ้อนที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่าย

1.4 ขอบเขตการวิจัย

ศึกษาวิธีการวิเคราะห์องค์ประกอบว่าเป็นวิธีที่ดีที่สุดในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่ายได้หรือไม่ โดยทำการทดลองกับฐานข้อมูลจำนวน 494,020 รายการ จาก KDD Cup 1999 โดยทำการเปรียบเทียบผลทดลองที่ได้จากวิธีที่นำเสนอกับผลการทดลองที่ได้จากวิธีที่ได้จากงานวิจัยที่ผ่านมาพร้อมทั้งเปรียบเทียบผลการทดลองจากคุณสมบัติทั้งหมดทุกตัว

1.5 ขั้นตอนการศึกษาและดำเนินงานวิจัย

งานวิจัยนี้มีขั้นตอนการศึกษาและดำเนินงานวิจัยดังนี้

1. ศึกษางานวิจัยที่เกี่ยวข้องด้านการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย
2. ศึกษาการวิเคราะห์องค์ประกอบ
3. ตั้งสมมติฐานโดยคาดว่าวิธีวิเคราะห์องค์ประกอบสามารถใช้ในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ขั้นตอนวิธีต้องง่ายไม่ซับซ้อนที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่าย
4. ฐานข้อมูลจะใช้ฐานข้อมูลจำนวน 494,020 รายการ จาก KDD Cup 1999
5. ทำการเปรียบเทียบผลทดลองที่ได้จากวิธีที่นำเสนอกับผลการทดลองที่ได้จากวิธีที่ได้จากงานวิจัยที่ผ่านมาพร้อมทั้งเปรียบเทียบผลการทดลองจากคุณสมบัติทั้งหมดทุกตัว

1.6 ประโยชน์ที่คาดว่าจะได้รับ

สามารถหาขั้นตอนวิธีใช้ในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย ขั้นตอนวิธีต้องง่ายไม่ซับซ้อนที่สามารถลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพความแม่นยำการตรวจจับการบุกรุกในเครือข่าย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงคุณลักษณะทั่วไปของงานที่เกี่ยวข้องกับระบบการตรวจจับการบุกรุกในเครือข่าย รวมทั้งทฤษฎีการประยุกต์ข้อมูลทางสถิติโดยเฉพาะการวิเคราะห์องค์ประกอบที่สามารถนำมาพัฒนาและประยุกต์ขั้นตอนวิธีในการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับระบบการตรวจจับการบุกรุกในเครือข่าย

2.1 ระบบการตรวจจับการบุกรุกในเครือข่าย

Intrusion Detection คือการกลั่นกรองหรือตรวจสอบข้อมูลกิจกรรมต่าง ๆ ที่อาจจะเป็นภัยคุกคามต่อเครื่องหรือระบบเครือข่ายคอมพิวเตอร์ ดังนั้นระบบตรวจจับผู้บุกรุกก็คือซอฟต์แวร์หรือฮาร์ดแวร์ที่ใช้เพื่อให้บรรลุวัตถุประสงค์ดังกล่าว ซึ่งจะคอยทำหน้าที่เฝ้าตรวจสอบข้อมูลต่าง ๆ ที่วิ่งเข้าและออกบนเครือข่ายคอมพิวเตอร์ และเมื่อมีข้อมูลที่ต้องสงสัยระบบจะระบุว่าเป็นการบุกรุกพร้อมกับรายงานต่อไปยังผู้ดูแลระบบ (Administrator) และดำเนินการตามเงื่อนไขต่าง ๆ ตามที่ผู้ดูแลระบบได้กำหนดไว้ การบุกรุก (Intrusion) คือความพยายามที่จะกระทำการใดๆที่ส่งผลต่อความสมบูรณ์ (Integrity) ความลับ (Confidentiality) และความพร้อมใช้งาน (Availability) ของทรัพยากรต่าง ๆ บนระบบเครือข่าย การบุกรุกจะถูกกระทำโดยผู้ไม่ประสงค์ดีต่อระบบซึ่งอาจหมายถึงแฮกเกอร์ (Hacker), แคร็กเกอร์ (Cracker) หรือบุคคลกลุ่มอื่นที่ไม่หวังดีต่อระบบเครือข่าย โดยในระบบตรวจจับผู้บุกรุกจะเรียกบุคคลกลุ่มดังกล่าวว่า ผู้บุกรุก (Intruder) ซึ่งการบุกรุกของผู้บุกรุกสามารถกระทำได้จากภายในเครือข่าย (Inside) หมายถึงผู้บุกรุกที่มีสิทธิ์ในระบบเครือข่ายภายในที่พยายามจะใช้งานทรัพยากรระบบเกินอำนาจสิทธิ์ที่ตนได้รับ และผู้บุกรุกภายนอกเครือข่าย (Outside) หมายถึงผู้บุกรุกจากภายนอกเครือข่ายที่พยายามจะเข้าใช้งานทรัพยากรระบบทั้งที่ไม่ได้รับสิทธิ์ และบุคคลที่พยายามจะกระทำการใด ๆ ที่ไม่เป็นผลดีต่อระบบเครือข่าย จากภายนอก

ประเภทของระบบตรวจจับผู้บุกรุก

ในการจัดประเภทของระบบตรวจจับผู้บุกรุกสามารถแบ่งได้หลายประเภท โดยใช้หลักเกณฑ์ต่าง เช่น แหล่งข้อมูลที่นำมาใช้วิเคราะห์ หรือแนวทางในการใช้ตรวจจับการบุกรุก

รวมทั้ง ช่วงหรือระยะเวลาที่นำมาใช้เพื่อทำการวิเคราะห์การบุกรุก แต่ส่วนใหญ่แล้วจะแบ่งประเภทของระบบตรวจจับผู้บุกรุกออกเป็น 3 ประเภท ได้แก่ Host-Based Intrusion Detection System (HIDS), Network-Based Intrusion Detection System (NIDS) และ Application-Based Intrusion Detection System (AIDS)

HIDS คือระบบนี้จะทำงานอยู่บนเครื่องคอมพิวเตอร์แต่ละเครื่อง โดยจะทำการตรวจสอบข้อมูลที่ผ่านเข้าและออกคอมพิวเตอร์แต่ละเครื่องเพื่อหาว่าโปรแกรมหรือผู้ใช้คนใดที่ทำให้เกิดการบุกรุกขึ้นบนระบบ

NIDS คือระบบนี้จะทำหน้าที่ตรวจจับข้อมูลทั้งหมดที่มีการไหลเข้าและออกบนระบบเครือข่ายแล้วทำการวิเคราะห์ว่ากิจกรรมใดบนเครือข่ายเป็นการบุกรุกหรือพยายามที่จะบุกรุกหรือไม่ โดยอาศัยค่าต่าง ๆ อาทิเช่น ปริมาณข้อมูลบนเครือข่าย, ลักษณะของเพ็กเก็จที่ส่งเข้ามาในเครือข่าย เป็นต้น

AIDS ระบบนี้จะทำงานคล้ายกับ HIDS แต่จะมีส่วนของการรวบรวมข้อมูลจากการทำงานของโปรแกรมประยุกต์ที่ทำงานบนเครื่องคอมพิวเตอร์เพิ่มเข้ามา เพื่อใช้สำหรับการวิเคราะห์และตรวจสอบหาพฤติกรรมที่น่าสงสัยหรือผิดปกติที่เกิดขึ้น

แนวทางในการตรวจจับการบุกรุก

แนวทางในการใช้ในการตรวจสอบการบุกรุกมีอยู่ 2 แนวทาง ได้แก่ Anomaly Detection และ Misuse Detection โดยแนวทางแรกสำหรับ Anomaly Detection นั้นเป็นการตรวจจับและค้นหาพฤติกรรมต่าง ๆ ของผู้ใช้ที่เปลี่ยนแปลงไปจากสภาวะการใช้งานปกติ ตัวอย่างเช่นการที่ผู้ใช้ส่วนใหญ่จะมีการเข้าถึงข้อมูลหรือทรัพยากรต่าง ๆ เฉพาะช่วงกลางวันเท่านั้น แต่ถ้าหากวันใดได้มีการใช้งานในช่วงกลางคืนของผู้ใช้บางคนก็จะถูกสงสัยว่าเป็นการบุกรุก เป็นต้น ขณะที่แนวทางที่สองสำหรับ Misuse Detection จะตรวจจับโดยอาศัยรูปแบบที่ถูกกำหนดหรือสร้างขึ้นไว้แล้วว่าพฤติกรรมรูปแบบใดถูกระบุว่าเป็นผู้บุกรุก ซึ่งรูปแบบดังกล่าวจะถูกนำมาเปรียบเทียบค้นหากับเหตุการณ์ต่างๆที่เกิดขึ้นในระบบซึ่งถ้าพฤติกรรมใดที่เกิดขึ้นตรงกันกับรูปแบบพฤติกรรมที่ได้กำหนดไว้ก็จะถูกระบุว่าเป็นการบุกรุก แนวทางนี้เป็นแนวทางที่จะต้องอาศัยผู้เชี่ยวชาญทางด้านนี้เป็นผู้กำหนดรูปแบบพฤติกรรมการบุกรุกไว้ล่วงหน้าแล้ว ดังนั้นแนวทางนี้อาจจะถูกเรียกอีกชื่อหนึ่งว่า Signature-Based Detection

งานวิจัยระบบการตรวจจับการบุกรุกในเครือข่ายส่วนใหญ่จะใช้คุณลักษณะทั้งหมด จำนวน 41 ตัวแปรในการหาวิธีใหม่ๆ เพื่อจัดประเภทของการบุกรุกในเครือข่ายโดยมีการปรับค่าน้ำหนักของแต่ละคุณลักษณะ ซึ่งงานวิจัยในระยะหลังเริ่มให้ความสนใจกระบวนการแปลงค่าคุณลักษณะก่อนดำเนินการจัดประเภทของการบุกรุกในเครือข่าย แสดงว่าคุณลักษณะบางตัวอาจเป็นตัวถ่วงต่อระยะเวลาการประมวลผลและลดประสิทธิภาพการตรวจจับการบุกรุกในเครือข่ายลง แต่มีงานวิจัยส่วนน้อยที่เริ่มสนใจคัดเลือกเฉพาะคุณลักษณะที่เด่นเท่านั้นเพื่อการลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพการตรวจจับการบุกรุกในเครือข่าย ซึ่งขั้นตอนวิธีในการคัดเลือกคุณลักษณะเด่นนี้ยังซับซ้อน

2.2 การวิเคราะห์องค์ประกอบ (Factor Analysis)

เป็นการวิเคราะห์หลายตัวแปรเทคนิคหนึ่งเพื่อการสรุปรายละเอียดของตัวแปรหลายตัว หรือเรียกว่าเป็นเทคนิคที่ใช้ในการลดจำนวนตัวแปรเทคนิคหนึ่งโดยการศึกษาถึงโครงสร้างความสัมพันธ์ของตัวแปร และสร้างตัวแปรใหม่เรียกว่า องค์ประกอบ โดยองค์ประกอบที่สร้างขึ้นจะเป็นการนำตัวแปรที่มีความสัมพันธ์กันหรือมีความร่วมกันสูงมารวมกันเป็นองค์ประกอบเดียวกัน ส่วนตัวแปรที่อยู่คนละองค์ประกอบมีความร่วมกันน้อย หรือไม่มีความสัมพันธ์กันเลย ในเทคนิคนี้จะใช้ค่าสัมประสิทธิ์สหสัมพันธ์วัดความสัมพันธ์ ระหว่างตัวแปร ดังนั้นการวิเคราะห์องค์ประกอบจึงเป็นเทคนิคการลดจำนวนตัวแปร จากจำนวนตัวแปรมากๆ ให้เหลือเพียงไม่กี่ปัจจัย หรือ ตัวแปร

2.3 งานวิจัยที่เกี่ยวข้อง

งานวิจัยทาง NIDS ก็พยายามพัฒนาขั้นตอนวิธี เพื่อเพิ่มค่าความแม่นยำในการทำนาย โดยงานวิจัยส่วนใหญ่เลือกใช้ data mining และ machine learning มีดังนี้

คุณ Chi-Hoon Lee และคณะ [12] ได้นำเสนองานวิจัยเรื่อง “Network Intrusion Detection through Genetic Feature Selection” งานวิจัยนี้เสนอวิธีใหม่ในการคัดเลือกคุณสมบัติร่วมสูงสุดระหว่าง รูปแบบพฤติกรรมปกติและรูปแบบพฤติกรรมที่เป็นการบุกรุกระบบเครือข่ายงาน โดยงานวิจัยนี้นำเสนอวิธีการคัดเลือกคุณสมบัติที่ดีที่สุดจากการใช้หลักการคัดเลือกพันธุกรรม ซึ่งสามารถเพิ่มระดับอัตราความแม่นยำของระบบการตรวจจับการบุกรุกเครือข่าย ซึ่งได้ทำการทดลองกับชุดข้อมูล KDD CUP 1999 จาก MIT Lincoln Labs ผลการ

ทดลองได้เลือกคุณลักษณะเด่นจำนวน 21 ตัว จากทั้งหมด 41 ตัว (คิดเป็น 51.22%) แต่ข้อเสียของการใช้หลักการคัดเลือกพันธุกรรมในการคัดเลือกคุณลักษณะเด่นใช้เวลาในการประมวลผลนานมาก

คุณ Kok-Chine Khor และคณะ [24] ได้นำเสนองานวิจัยเรื่อง "A Feature Approach for Network Intrusion Detection" งานวิจัยนี้เสนอวิธีใหม่ในการคัดเลือกคุณสมบัติโดยใช้ขบวนการวิธีทาง data mining ที่ซับซ้อน สามารถเพิ่มระดับอัตราความแม่นยำของระบบการตรวจจับการบุกรุกเครือข่าย ซึ่งได้ทำการทดลองกับชุดข้อมูล KDD CUP 1999 จาก MIT Lincoln Labs ผลการทดลองได้เลือกคุณลักษณะเด่นจำนวน 7 ตัว จากทั้งหมด 41 ตัว (คิดเป็น 17.07%) แต่ข้อเสียคือไม่สามารถตรวจจับการบุกรุกชนิด R2L ได้

บทที่ 3
ขั้นตอนวิธีที่นำเสนอ

การวิจัยนี้ใช้เทคนิคการวิเคราะห์องค์ประกอบ (Factor Analysis) โดยใช้โปรแกรม SPSS for Windows ในการวิเคราะห์ข้อมูล ดังนี้

1. ทำการตรวจสอบว่าตัวแปรต่างๆ มีความสัมพันธ์กันหรือไม่ หากมีความสัมพันธ์กันอย่างมีนัยสำคัญจึงควรใช้เทคนิคการวิเคราะห์องค์ประกอบ ซึ่งในที่นี้ใช้สถิติ KMO (ค่า KMO หรือ Kaiser-Meyer-Olkin ปกติจะอยู่ที่ค่า 0 ถึง 1 หากมีค่า KMO สูงที่ใกล้ 1 ก็ควรใช้เทคนิคการวิเคราะห์องค์ประกอบ หาก ค่า KMO <0.5 ก็ไม่ควรใช้เทคนิคการวิเคราะห์องค์ประกอบ)
2. ทำการสกัดปัจจัย เพื่อหาจำนวนปัจจัย (Factor) ที่สามารถใช้แทนข้อมูลทั้งหมดได้ โดยใช้วิธี Principal Component Analysisหรือ PCA
3. หมุนแกนปัจจัย เพื่อให้ค่า Factor loading ของตัวแปรมีค่ามากขึ้นหรือลดลงจนกระทั่งทำให้ทราบว่าตัวแปรนั้นๆ ควรอยู่ในปัจจัยใดในการวิจัยนี้จะใช้ Orthogonal rotation
4. ตัวแปรแรกในแต่ละปัจจัยจะเป็นตัวแปรที่มีคุณลักษณะเด่น ที่ได้รับเลือกเป็นตัวแทน
5. เมื่อได้คุณลักษณะเด่นก็นำคุณลักษณะเหล่านั้นไปจัดประเภทด้วย อัลกอริทึม C4.5
6. ทำการเปรียบเทียบค่า Truth positive rate ที่ได้กับ ค่า Truth positive rate จากวิธีต่างๆ เช่นจากข้อมูลทั้งหมด (full features) จากวิธีของคุณ Chi-Hoon Lee และคณะ (ซึ่งจะย่อว่า GA) และจากวิธีของคุณคุณ Kok-Chine Khor และคณะ(ซึ่งจะย่อว่า Kok)

ชุดข้อมูลที่ใช้ทดสอบเป็นชุดข้อมูลที่เป็นมาตรฐาน KDD CUP 1999 จาก MIT Lincoln Labsซึ่งฐานข้อมูลมีจำนวน 494,020 รายการ มีตัวแปรคุณลักษณะทั้งหมด 41 ตัวแปร เป็นชุดข้อมูลที่ได้ถูกใช้และถูกอ้างถึงมากที่สุดในงานวิจัยด้านการตรวจจับการบุกรุกเครือข่าย ซึ่งฐานข้อมูลดังกล่าวนอกจากจะมีคุณลักษณะ 41 ตัวแปรแล้ว ยังมีตัวแปรสุดท้ายเป็นตัวแปรที่บอกประเภทซึ่งมีค่าเป็นปกติ หรือชนิดการบุกรุกดังนี้ normal, Dos, Probing, R2L, U2R โดยรายละเอียดของตัวแปรคุณลักษณะแสดงไว้ในตารางที่ 1

ตารางที่ 1: รายละเอียดตัวแปรคุณลักษณะทั้งหมด 41 ตัว

ตัวแปร	ชื่อคุณลักษณะ	ตัวแปร	ชื่อคุณลักษณะ
V1	duration	V2	protocol_type
V3	service	V4	flag
V5	src_bytes	V6	dst_bytes
V7	land	V8	wrong_fragment
V9	urgent	V10	hot
V11	num_failed_logins	V12	logged_in
V13	num_compromised	V14	root_shell
V15	su_attempted	V16	num_root
V17	num_file_creations	V18	num_shells
V19	num_access_files	V20	num_outbound_cmds
V21	is_host_login	V22	is_guest_login
V23	count	V24	serv_count
V25	serror_rate	V26	srv_serror_rate
V27	reerror_rate	V28	srv_rerror_rate
V29	same_srv_rate	V30	diff_srv_rate
V31	srv_diff_host_rate	V32	dst_host_count
V33	dst_host_srv_count	V34	dst_host_same_srv_rate
V35	dst_host_diff_srv_rate	V36	dst_host_same_src_port_rate
V37	dst_host_srv_diff_host_rate	V38	dst_host_serror_rate
V39	dst_host_srv_serror_rate	V40	dst_host_rerror_rate
V41	dst_host_srv_rerror_rate		

ตาราง 2: รายละเอียดประเภทของการบุกรุกเครือข่าย

ประเภทการบุกรุก	รายละเอียด
DoS	back, land, Neptune, pod, smurf, teardrop
Probing	ipsweep, nmap, portsweep, satan
U2R	buffer_overflow, loadmodule, perl, rootkit
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

บทที่ 4

ผลการทดลอง

การทดลองทำการตรวจสอบว่าตัวแปรต่างๆ มีความสัมพันธ์กันหรือไม่ปรากฏว่าได้ค่า KMO (Kaiser-Meyer-Olkin) = 0.851 การทดสอบ Bartlett's Test ให้ค่า Sig. =0.000 แสดงว่าข้อมูลมีความสัมพันธ์กัน ควรใช้เทคนิคการวิเคราะห์องค์ประกอบ (Factor Analysis) ได้ เมื่อสกัดปัจจัยและหมุนแกนปัจจัยสามารถแสดงรายละเอียดได้ดังตารางที่ 3

ตารางที่ 3: ตารางแสดงรายละเอียดปัจจัยที่ได้หลังหมุนแกนปัจจัย

ปัจจัยที่	ค่า Eigen value	ตัวแปรที่อยู่ในปัจจัยเดียวกันเรียงตามค่า PCA
1	5.68	V27, V28, V40,V41, V37
2	4.47	V35,V36,V1
3	3.15	V13,V16,V15,V14
4	2.215	V24,V23
5	2.067	V2
6	1.812	V29
7	1.619	V26,V25
8	1.393	V22,V10
9	1.333	V39,V38
10	1.182	V11,V9
11	1.035	V31
12	1.015	V17
13	1.003	V5,V18

จากตารางที่ 3 จะเห็นได้ว่าชุดข้อมูลดังกล่าวที่มีค่า Eigen value มากกว่า 1 สามารถสกัดได้เป็น 13 ปัจจัย ตัวแปรที่อยู่ในปัจจัยเดียวกันเรียงตามค่า PCA โดยเรียงจากค่ามากไปน้อย ผู้วิจัยมีความคิดว่าข้อมูลตัวแปรที่อยู่ในปัจจัยเดียวกันย่อมมีความคล้ายกันสูง ข้อมูลต่างปัจจัยกันย่อมมีความสัพพันธ์กันน้อย ดังนั้นผู้วิจัยจึงทำการเลือกตัวแปรตัวแรกของทุกปัจจัยมาเป็นคุณลักษณะเด่น ดังนั้นคุณลักษณะเด่นที่ได้รับการคัดเลือกมีทั้งหมด 13 ตัวแปรดังนี้ V27, V35, V13, V24, V2, V29, V26, V22, V39, V11, V31, V17, และ V5

ตารางที่ 4 : แสดงรายละเอียดคุณลักษณะของแต่ละวิธี

คุณลักษณะ	รายละเอียดชื่อคุณลักษณะที่ได้รับเลือก	จำนวนคุณลักษณะ
Full Features	duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, urgent, hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, is_host_login, is_guest_login, count, serv_count, serror_rate, srv_serror_rate, reerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate,	41
GA approach	service, flag, wrong_fragment, hot, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_srv_serror_rate, duration, src_bytes, srv_serror_rate, land, urgent, su_attempted, num_root, num_shells, num_access_files, isguest_login	21
Kok approach	service, dst_bytes, logged_in, count, dst_host_count, root_shell, dst_host_rerror_rate	7
Factor Analysis approach วิธีที่นำเสนอ	reerror_rate, dst_host_diff_srv_rate, num_compromised, serv_count, protocol_type, same_srv_rate, srv_serror_rate, is_guest_login, dst_host_srv_serror_rate, num_failed_logins, srv_diff_host_rate, num_file_creations, src_bytes	13

ทำการสุ่มแบ่งชุดข้อมูลเป็น 3 ชุดมีรายละเอียดดังตารางที่ 5 ผลการทดลองแสดงในตารางที่ 6, 7, และ 8 ตามลำดับ

ตารางที่ 5: รายละเอียดของข้อมูล 3 ชุด

ชื่อชุดข้อมูล	จำนวนข้อมูล (records)
Dataset-1	186,745
Dataset-2	49,438
Dataset-3	25,419

ตารางที่ 6: ผลการทดลองกับชุด Dataset 1

ประเภทการบุกรุก	ค่าความถูกต้อง (Truth positive rate)			
	คุณลักษณะทั้งหมด	วิธี GA	วิธีคุณ Kok	วิธีที่นำเสนอ
Normal	99.97	99.95	99.81	99.96
Dos	99.96	97.21	99.39	99.96
Probing	97.83	100.00	78.26	100
R2L	12.03	6.01	0.00	12.08
U2R	15.38	12.82	23.08	23.08
ทุกประเภท	99.80	97.32	99.25	99.8

ตารางที่ 7: ผลการทดลองกับชุด Dataset 2

ประเภทการบุกรุก	ค่าความถูกต้อง (Truth positive rate)			
	คุณลักษณะทั้งหมด	วิธี GA	วิธีคุณ Kok	วิธีที่นำเสนอ
Normal	99.48	99.30	99.11	99.36
Dos	99.95	97.10	99.34	99.96
Probing	99.08	99.31	78.39	98.61
R2L	9.91	13.51	0.00	13.51
U2R	15.38	12.82	23.08	23.08
ทุกประเภท	99.58	97.30	98.82	99.57

ตารางที่ 8: ผลการทดลองกับชุด Dataset 3

ประเภทการบุกรุก	ค่าความถูกต้อง (Truth positive rate)			
	คุณลักษณะทั้งหมด	วิธี GA	วิธีคุณ Kok	วิธีที่นำเสนอ
Normal	99.70	99.44	99.22	99.66
Dos	99.96	97.28	99.38	99.94
Probing	99.67	99.33	78.00	99.33
R2L	15.00	5.00	0.00	20.00
U2R	15.38	12.82	23.08	23.08
ทุกประเภท	99.50	97.31	98.66	99.51

จากผลการทดลองจะเห็นได้ว่าค่าความถูกต้องมีค่าที่ใกล้เคียงกับเมื่อใช้คุณลักษณะทั้งหมด 41 ตัว ดังนั้นผลการทดลองยืนยันว่าสามารถใช้เทคนิคการวิเคราะห์องค์ประกอบมาคัดเลือกคุณลักษณะเด่นได้ ซึ่งขบวนการไม่ยุ่งยาก และไม่ซับซ้อน สามารถประมวลผลได้เร็วด้วยโปรแกรม SPSS

บทที่ 5

สรุปผลการทดลอง

ผลการทดลองยืนยันว่าสามารถใช้เทคนิคการวิเคราะห์องค์ประกอบมาคัดเลือกคุณลักษณะเด่นกับข้อมูลการบุกรุกเครือข่ายด้วยชุดข้อมูลที่เป็นมาตรฐาน KDD CUP 1999 จาก MIT Lincoln Labs ฐานข้อมูลมีจำนวน 494,020 รายการ ได้ ซึ่งขบวนการไม่ยุ่งยาก และไม่ซับซ้อน สามารถประมวลผลได้เร็วด้วยโปรแกรม SPSS คุณลักษณะเด่นที่ได้รับการคัดเลือกมีเพียง 13 ตัวจากทั้งหมด 41 ตัว เพียงประมาณ 31.71 % ทำให้ประหยัดเวลาในการประมวลผล โดยยังคงความถูกต้องเทียบเท่าการใช้คุณลักษณะทั้งหมด 41 ตัว

บรรณานุกรม

[1] S. Hansman and R. Hunt. A Taxonomy of network and computer attacks. Computers & Security. 2005, 24, 31-43.

[2] S. Hettich, S.D. Bay. The UCI KDD Archive. Irvine, CA: University of California, Department of Information and Computer Science, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.

[3] R. H. Gong, M. Zulkernine, and P. Abolmaesumi. A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection. *Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05)*. 2005.

[4] Y. Bai and H. Kobayashi. Intrusion Detection Systems: Technology and Development. *Proceeding of the 17th International Conference on Advanced Information Networking and Applications (AINA'03)*. 2003.

[5] J. S. Han and B. Cho. Detecting intrusion with rule-based integration of multiple models. Computer & Security. 2003, 22,613-623.

[6] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan. Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project. *DARPA Information Survivability Conference*. 2000.

[7] S. Mukkamala and A. H. Sung. A comparative study of techniques for intrusion detection. *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*. 2003.

[8] G. John, R. Kohavi, and Pfleger. Irrelevant features and the subset selection problem. *Int. Conf. on Machine Learning, Morgan Kaufman, San Francisco*. 1994, 121-129.

[9] I.H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

[10] C. Lui, T. Fu, and T. Cheung. Agent-based Network Intrusion Detection System Using Data Mining Approaches. *Proccedings of the Third International Conference on Information Technology and Application (ICITA'05)*. 2005

[11] W. Xuren, H. Famei, and X. Rongsheng. Modeling Intrusion Detection System by Discovering Association Rule in Rough Set Theory Framework. *International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*. 2006

[12] C. Lee, S. Shin, and J. Chung. Network Intrusion Detection Through Genetic Feature Selection. *Proceedings of the Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06)*. 2006

[13] W. Siedlecki, and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letter*. 1989

[14] K. Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. 1995, 2, 1137-1143.

[15] W. Hu, J. Li, and J. Shi. Optimal Evaluation of Feature Selection in Intrusion Detection Modeling. *Proceeding of the 6th world congress on Intelligent Control and Automation,*

Dalian, China. June 21- 23 2006.

[16] W. Xuren, H. Famei, and X. Rongsheng. Modeling Intrusion Detection System by Discovering Association Rule in Rough Set Theory Framework. *International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*. 2006.

[17] L. Wang, Y. Zhang, and J. Feng. On the Euclidean Distance of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[18] E. Bigdeli. and Z. Bahmani. Comparing Accuracy of Cosine-based Similarity and Correlation-Based Similarity Algorithms in Tourism Recommender Systems. *Proceeding of The 4th IEEE International Conference on Management of Innovation and Technology (ICMIT)*, 2008.

[19] A. Karnik, S. Goswami. and R. Guha. Detecting Obfuscated Viruses Using Cosine Similarity Analysis. *Proceedings of the First Asia International Conference on Modelling & Simulation (AMS'07)*. 2007.

[20] R. Yeh, C. Liu, B. Shla, Y. Cheng, and Y. Huwang. Imputing manufacturing material in data mining. *Springer Science+Business Media, LLC*. 2007.

[21] J. Wang, Q. Yang, and D. Ren. An intrusion detection algorithm based on decision tree technology. *Asia-Pacific Conference on Information Processing*. 2009.

[22] J. Du and W. Guo. Data Mining on Patient Data. IEEE, 2005.

[23] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel. Feature Ranking Selection and Discretization. *Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Istanbul*. June 2003, pp, 251-254.

[24] K. Khor, C. Ting, and S. Amnuaisuk. A Feature Selection Approach for Network Intrusion Detection. *2009 International Conference on Information Management and Engineering*. 2009

[25] A. Suebsing and N. Hiransakolwong. Feature Selection Using Euclidean Distance and Cosine Similarity for Intrusion Detection Model. *Asian Conference on Intelligent Information and Database Systems (ACIIDS 09)*. 2009.

[26] A. Suebsing and N. Hiransakolwong. Euclidean-based Feature Selection for Network Intrusion Detection. *Proceedings of 2009 International Conference on Machine Learning and Computing*. 2009.

[27] A. Conklin, G. White, C. Cothren, D. Williams and R. Davis. Principles of computer Security: Security+ and Beyond. McGraw-Hill, 2005.

[28] J. Ross Quinlan. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, CA, 1992.